Chapter UNIT-IV

4

BASIC STATISTICS

§ 4.1. MEASURES OF CENTRAL TENDENCY

A very important objective of statistical analysis is to get one single value that describes the entire mass of data given in a frequency distribution. This single value usually occurs in the middle or central part of the entire mass of data and called central value or an 'average' or the expected value of the variable. The word average is very familiar in our day-to-day life. For example, an average student, he gets average salary, the average life of an Indian is 70 years etc.

In statistics average is a technique which reduces mass data into a single value and that value contains all characteristics of the whole mass of data. Averages are called 'Measures of Central Tendency'. Thus measures of central tendency is a technique which is widely used in describing and comparing the natures and composition of a series or a frequency distribution.

§ 4.2. DEFINITION OF AVERAGE OR MEASURES OF CENTRAL TENDENCY

Average has been defined differently by various authors from time to time. Some of the definitions are given below :

(i) According to Kellog and Smith, "An average is sometimes called a measure of central tendency, because individual values of the variable usually cluster around it."

(ii) According to Leabo, "The average is sometimes described as a number which is typical of the whole group."

(iii) According to Prof. R. A. Fisher, "The Inherent ability of human to grasp in its entirety a large body of numerical data compels us to seek relatively few constants that will adequately describe the data."

(iv) According to Croxton and Cowden, "An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of the data, it is also called a measure of central value."

(v) According to A. L. Bowley, "Averages are statistical constant which enable us to comprehend in a single effort the significance of the whole."

(vi) According to Simpson and Kafka, "A measure of central tendency is a typical value around which other figures congregate."

(viii) According to Clark, "Average is an attempt to find one single figure to describe whole of figures."

§ 4.3. OBJECTIVES OF AVERAGES OR MEASURES OF CENTRAL TENDENCY

The collection of statistical data is a difficult task and to understand these data is even more difficult. Thus, it is essential that data should be in such a form that every person can understand it. The summarized data is easy to understand and summarization can be done with the help of averages. By averages the complexity of the data can be removed and a single value is obtained. In this ways averages serve the following objectives :

(i) It presents a brief picture of the mass data. With the help of average, the mass information can be understood and grasped easily.

(ii) Averages are very useful for making comparisons. In this way, averages can be used for making comparative study of two or more series of data.

(iii) Averages are very useful in statistical analysis. These are widely used in dispersion, skewness, regression, etc.

(iv) Averages are very useful for making decision in planning activities in various fields.

§ 4.4. CHARACTERISTICS OF AN IDEAL MEASURE OF CENTRAL TENDENCY

An average is a single representative value of the mass of complex data. Thus it must have the following main characteristics (suggested by G. U. Yule) :

(i) It should be rigidly defined.

(ii) It should be easy to understand and easy to calculate.

(iii) It should be based on all observations.

(iv) It should be least (minimum) affected by fluctuations of sampling.

(v) It should be capable of further algebraic treatment.

(vi) It should not be affected by extreme values of the observations.

§ 4.5. TYPES OF MEASURES OF CENTRAL TENDENCY

There are various types of measures of central tendency, but we can broadly classify these into two types :

(A) Mathematical Averages

(B) Positional Averages

Now we shall discuss these two in details :

(A) Mathematical Averages. Mathematical averages further can be classified into three types of averages :

(1) Arithmetic Mean or Mean (A. M.)

(i) Simple A. M.

(ii) Weighted A. M.

(2) Geometric Mean (G. M.)

(3) Harmonic Mean (H. M.)

(B) Positional Averages :

1. Median
2. Quartiles
3. Deciles
4. Percentiles
5. Mode

◈ § 4.6. ARITHMETIC MEAN

Usually to find arithmetic mean is very simple. To find the arithmetic mean, add the values of all terms and then divide this sum by the number of terms. The quotient is the arithmetic mean. There are three methods to find the mean :

(i) Direct method,

(ii) Short-cut method,

(iii) Step deviation method.

◈ § 4.6.1. ARITHMETIC MEAN BY DIRECT METHOD

(1) In a series of individual observations. In individual series the arithmetic mean is obtained by dividing the sum of all terms by the number of terms. *i.e.*, if the values of variate x be $x_1, x_2, ..., x_n$ then the arithmetic mean (A. M.) is defined by

$$\mathbf{M} \text{ or } \mathbf{A.M.} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

[$\because$ number of terms $x_1, x_2, \ldots, x_n$ is n]

or $$\mathbf{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ or } \mathbf{A.M.} = \frac{\Sigma x}{n} \quad \ldots(1)$$

Weighted arithmetic mean. Sometimes it happens that the variate values are not of equal importance. Suppose $w_1, w_2, ..., w_n$ are the weights assigned to the n values $x_1, x_2, \ldots, x_n$ as measures of their importance, then the **weighted arithmetic mean** ($\bar{x}_w$) is defined by

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

or $$\bar{x}_w = \frac{\Sigma wx}{\Sigma w}. \quad \ldots(2)$$

Remark. The **'weight'** means a numerical multiplier that is assigned to each value of the variate and indicate its relative importance. The weights can not be treated as frequencies while frequencies can be treated as weights.

(2) Arithmetic mean in a discrete series. If $x_1, x_2, ..., x_n$ are n values of the variate x and x_1 occurs f_1 times, x_2 occurs f_2 times, ..., x_n occurs f_n times *i.e.*, if the frequency distribution is as follows :

x :	x_1	x_2	$x_3 \dots x_n$
y :	f_1	f_2	$f_3 \dots f_n$

then their arithmetic mean M (or A. M.)

$$= \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + f_2 + \dots + f_n} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i \qquad \left[\because \sum_{i=1}^{n} f_i = N\right]$$

or $$\mathbf{M = \frac{\Sigma fx}{N}}. \qquad \dots(3)$$

(3) Arithmetic mean in a grouped or continuous series. In this case the arithmetic mean is given by the above formula (3) where x's denote mid-value of the class intervals. For example, for the class 30–40, the value of $x = \frac{1}{2}(30 + 40)$ *i.e.*, 35.

ILLUSTRATIVE EXAMPLES

Example 1. (a) ***The marks obtained by nine students in a paper of Statistics are as follows*** **52, 75, 40, 70, 43, 40, 65, 35, 48.**

Calculate the mean.

Solution. By direct method A. M. (*i.e.*, M) is given by

$$M = \frac{\Sigma fx}{n}$$

$$= \frac{52 + 75 + 40 + 70 + 43 + 40 + 65 + 35 + 48}{9} = 468/9 = 52. \quad \textbf{Ans.}$$

Example 1. (b) ***Find the arithmetic mean (A. M.) of first n natural numbers.***

Solution. Here x : 1 2 3 ... n.

$\therefore$ By direct method, the required arithmetic mean (M) is given by

$$M = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1 + 2 + 3 + \dots + n}{n}$$

$$= \frac{\frac{1}{2} n (n + 1)}{n} = \frac{1}{2}(n + 1). \quad \textbf{Ans.}$$

Example 1. (c) ***Show that A. M. of first n natural numbers is*** $\frac{1}{2}(n + 1)$.

Solution. See Exp. 1 (b) above.

Example 2. (a) ***Find the mean height of the students from the following frequency distribution :***

Height (in inches)	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**
No. of Students	**1**	**6**	**10**	**22**	**21**	**17**	**14**	**5**	**3**	**1**

Solution. By direct method

Height (in inches) x	No. of Students f	fx
64	1	64
65	6	390
66	10	660
67	22	1474
68	21	1428
69	17	1173
70	14	980
71	5	355
72	3	216
73	1	73
	$\Sigma f = N = 100$	$\Sigma f x = 6813$

$\therefore$ mean height (M) is given by

$$M = \frac{1}{N}\Sigma fx = \frac{6813}{100} = 68 \cdot 13 \text{inches.}$$ **Ans.**

Example 2. (b) ***Compute the arithmetic mean of the marks from the following table :***

Marks	0–10	10–20	20–30	30–40	40–50	50–60
No. of students	12	18	27	20	17	6

Solution. By directed method

Marks	Mid Value x	No. of students f	fx
0–10	5	12	60
10–20	15	18	270
20–30	25	27	675
30–40	35	20	700
40–50	45	17	765
50–60	55	6	330
Total		$N = 100$	$\Sigma fx = 2800$

Required A. M. (M) is given by

$$M = \frac{1}{N}\Sigma fx = \frac{2800}{100} = 28.$$ **Ans.**

Example 3. ***The percentage of marks of a student in an examination is English 75, Statistics 60, Mathematics 59, Physics 55, Chemistry 63.***

Find the weighted arithmetic mean if the weights given to the subjects are 2, 1, 3, 3 and 1 respectively. Find also the simple mean.

Solution.

Subjects	% of marks x	Weights w	wx
English	75	2	150
Statistics	60	1	60
Mathematics	59	3	177
Physics	55	3	165
Chemistry	63	1	63
Total	$\Sigma x = 312$	$\Sigma w = 10$	$\Sigma wx = 615$

$\therefore$ Weighted A. M.

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{615}{10} = 61 \cdot 5 \text{ marks}$$

and $$\text{A. M. } (\bar{x}) = \frac{\Sigma x}{n} = \frac{312}{5} = 62 \cdot 5 \text{ marks.}$$

◈ § 4.6.2. SHORT-CUT METHOD

This method is used to make the calculations simple. This method is based on the fact that the sum of the deviations of variates x from the actual mean M is zero *i.e.*, $\Sigma (x - M) = 0$.

Let A be any assumed mean (or any assumed number), ξ the deviation of the arithmetic mean, then we have

$$\frac{1}{n} \Sigma f \xi = \frac{1}{N} \Sigma f(x - A) \qquad [\because \ \xi = x - A]$$

$$= \frac{1}{N} \Sigma fx - A \cdot \frac{1}{N} \Sigma f$$

$$= \frac{1}{N} \Sigma fx - A \cdot \frac{1}{N} \cdot N \qquad [\because \ \Sigma f = N]$$

$$= M - A$$

or $$\mathbf{M = A + \frac{\Sigma f \xi}{N}}.$$

It is the required formula.

In fact any point can serve the purpose of assumed mean. But to facilitate the calculations, usually the value of x corresponding to the middle part of the distribution is taken as assumed mean.

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the arithmetic mean (A. M.) of the following by direct and short cut methods both :***

Height in cm.	**219**	**216**	**213**	**210**	**207**	**204**	**201**	**198**	**195**
Mean	**2**	**4**	**6**	**10**	**11**	**7**	**5**	**4**	**1**

Solution.

Height in cm. x	f	fx	$\xi = x - A$, $A = 207$	$f\xi$
219	2	438	12	24
216	4	864	9	36
213	6	1278	6	36
210	10	2100	3	30
207	11	2277	0	0
204	7	1428	−3	−21
201	5	1005	−6	−30
198	4	792	−9	−36
195	1	195	−12	−12
Total	$N = 50$	$\Sigma fx = 10377$		$\Sigma f\xi = 27$

By direct method A. M. is given by

$$M = \frac{1}{N}\Sigma fx = \frac{10377}{50} = 207 \cdot 54 \text{ cm.}$$

By short cut-method. Let assumed mean $A = 207$

$$M = A + \frac{1}{N} \cdot \Sigma f\xi = 207 + \frac{27}{50} = 207 + 0 \cdot 54 = 207 \cdot 54 \text{ cm.}$$

Example 2. ***Compute the mean of the following by direct and short cut methods both :***

Class	20–30	30–40	40–50	50–60	60–70
Frequency	8	26	30	20	16

Solution.

Class	Mid-Value x	f	fx	$\xi = x - A$, $A = 45$	$f\xi$
20–30	25	8	200	−20	−160
30–40	35	26	910	−10	−260
40–50	45	30	1350	0	0
50–60	55	20	1100	10	200
60–70	65	16	1040	20	320
Total		$N = 100$	$\Sigma fx = 4600$		$\Sigma f\xi = 100$

By direct method $M = \frac{1}{N} . \Sigma fx = \frac{4600}{100} = 46.$

By short cut method. Let assumed mean $A = 45.$

$$M = A + \frac{\Sigma f\xi}{N} = 45 + \frac{100}{100} = 46.$$

§ 4.6.3. STEP DEVIATION METHOD

If in a frequency table, the class-intervals have equal width, say i, then it is convenient to use another formula known as step-deviation **formula**, to make the calculations simple.

Let $u = \frac{x - A}{i}$, then $x = A + iu.$

$\therefore$ $\Sigma fx = \Sigma f(A + iu) = A\,\Sigma f + i\,\Sigma fu$

$\Rightarrow$ $\frac{\Sigma fx}{\Sigma f} = A + i . \frac{\Sigma fu}{\Sigma f}.$

$\therefore$ $\mathbf{M = A + i . \frac{\Sigma fu}{N}.}$

It is the required formula.

Remark 1. The above three methods namely direct method, short-cut method and step-deviation method for calculating arithmetic mean are applicable to any type of series (*i.e.*, individual, discrete and grouped).

Remark 2. To add or subtract a constant in the value of the variate is called **change of origin** [See § 4.6.3, short-cut method].

To divide or multiply a value by a constant is called **change of scale**. Thus if we take $u = \frac{x - A}{i}$, then it is called **change of origin and scale both** [See § 4.6.3, step-deviation method].

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the mean of the following frequency distribution :***

Marks	:	10	20	30	40	50	60	70	80
No. of students	:	15	35	60	84	96	127	198	250

Solution. Taking assumed mean, $A = 50$, and step deviation (*i.e.*, class interval), $i = 10$, the table by step deviation method is as follows :

Marks x	No. of students (frequency) f	x − A (x − 50)	$u = \frac{x - A}{i}$ (i = 10)	fu
10	15	− 40	− 4	− 60
20	35	− 30	− 3	− 105
30	60	− 20	− 2	− 120
40	84	− 10	− 1	− 84
50	96	0	0	0
60	127	10	1	127
70	198	20	2	396
80	250	30	3	750
Total	$N = 865$	—	—	$\Sigma fu = 904$

Arithmetic mean marks M are given by

$$M = A + i \cdot \frac{\Sigma fu}{N} = 50 + 10 \times \frac{904}{865}$$

$$= 50 + 10 \cdot 45 = 60 \cdot 45.$$

Example 2. ***Compute the mean of the following frequency distribution:***

Class	0–11	11–22	22–33	33–44	44–55	55–66
Frequency	9	17	28	26	15	8

Solution. By step deviation method

Class	Mid-Value x	f	$\xi = x - A$ (A = 38·5)	$u = (x - A)/i$ i = 11	fu
0–11	5·5	9	− 33	− 3	− 27
11–22	16·5	17	− 22	− 2	− 34
22–33	27·5	28	− 11	− 1	− 28
33–44	38·5	26	0	0	0
44–55	49·5	15	11	1	15
55–66	60·5	8	22	2	16
Total		$N = 103$			$\Sigma fu = -58$

Let the assumed mean $A = 38 \cdot 5$, then

$$M = A + i \frac{\Sigma fu}{N} = 38 \cdot 5 + \frac{11(-58)}{103}$$

$$= 38 \cdot 5 - \frac{638}{103} = 38 \cdot 5 - 6 \cdot 194 = 32 \cdot 306.$$

Example 3. ***Find the missing frequency from the following data :***

No. of tablets	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36	36–40
No. of persons cured	11	13	16	14	?	9	17	6	4

It is being given that, 19·9 is the average number of tablets for being cured.

Solution. Let the assumed mean A be 18 and the missing frequency be f'. Then

No. of tablets	No. of persons cured f	Mid-Value x	$\xi = x - A$	$u = \frac{x-a}{i}$	fu
4–8	11	6	−12	−3	−33
8–12	13	10	−8	−2	−26
12–16	16	14	−4	−1	−16
16–20	14	18	0	0	0
20–24	f'	22	4	1	f'
24–28	9	26	8	2	18
28–32	17	30	12	3	51
32–36	6	34	16	4	24
36–40	4	38	20	5	20
Total	$Nf = 90 + f'$				$\Sigma fu = 38 + f'$

$\therefore$ Arithmetic mean by step deviation formula is

$$M = A + i\,\frac{\Sigma fu}{N}$$

$$\Rightarrow \quad 19\cdot 9 = 18 + 4\,\frac{38 + f'}{90 + f'} \qquad [\because\ M = 19\cdot 9]$$

$$\Rightarrow \quad (19\cdot 9 - 18)(90 + f') = 152 + 4f'$$

$$\Rightarrow \quad (1\cdot 9)(90 + f') = 152 + 4f'$$

$$\Rightarrow \quad (4 - 1\cdot 9)\,f' = 171 - 152$$

$$\Rightarrow \quad (2\cdot 1)\,f' = 19.$$

$$\therefore \quad f' = 19/\,2\cdot 1 = 9\cdot 05.$$

Example 4. ***Find the average marks of the students from the following table.***

Marks	No. of students	Marks	No. of students
Above 0	80	*Above* 60	28
Above 10	77	*Above* 70	16
Above 20	72	*Above* 80	10
Above 30	65	*Above* 90	8
Above 40	55	*Above* 100	0
Above 50	43		

Solution. Here we are given the cumulative frequency of distribution. Therefore we shall first change it into simple frequency distribution. Thus

Marks	No. of students
0–10	3 → (80 – 77)
10–20	5 → (77 – 72)
20–30	7 → (72 – 65)
30–40	10 → (65 – 55)
40–50	12 → Similarly find remaining frequencies
50–60	15
60–70	12
70–80	6
80–90	2
90–100	8

Hence the frequency table can be arranged as follows :

Let A = assumed mean = 55,

i = class interval = 10

Marks	**Mid-Value x**	**f**	$\xi = x - 55$	$u = \frac{x - 55}{10}$	**fu**
0–10	5	3	– 50	– 5	– 15
10–20	15	5	– 40	– 4	– 20
20–30	25	7	– 30	– 3	– 21
30–40	35	10	– 20	– 2	– 20
40–50	45	12	– 10	– 1	–12
50–60	55	15	0	0	0
60–70	65	12	10	1	12
70–80	75	6	20	2	12
80–90	85	2	30	3	6
90–100	95	8	40	4	32
Total		$N = 80$			$\Sigma fu = -26$

By step deviation method, the average marks (*i.e., A. M.*) are given by

$$M = A + i\frac{\Sigma fu}{N} = 55 + 10\frac{(-26)}{80}$$

$$= 55 - 3 \cdot 25 = 51 \cdot 75.$$

§ 4.7. PROPERTIES OF ARITHMETIC MEAN

Property 1. *The algebraic sum of the deviations of all the variates from their arithmetic mean is zero.*

Proof. Let $X_1, X_2, ..., X_i, ..., X_n$ be the values of the variates and let their corresponding frequencies be $f_1, f_2, ..., f_i, ..., f_n$ respectively.

Let x_i be the deviation of the variate X_i from the mean M, where $i = 1, 2, ... n$.

Then

$$x_i = X_i - M, \quad i = 1, 2, ..., n.$$

$$\therefore \quad \sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{n} f_i (X_i - M) = \sum_{i=1}^{n} f_i X_i - \sum_{i=1}^{n} f_i M$$

$$= \sum_{i=1}^{n} f_i X_i - M \sum_{i=1}^{n} f_i \quad [\text{since } M \text{ is the same for all values of } i]$$

$$= M \sum_{i=1}^{n} f_i - M \sum_{i=1}^{n} f_i \qquad \left\{ \because \ M = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} \right\}$$

$$= 0.$$

Proved.

Property 2. *If every value of the variate is increased by the same constant `a', then the arithmetic mean is also increased by 'a'.*

Proof. Let the given frequency distribution be :

x	:	x_1	x_2	...	x_n
f	:	f_1	f_2	...	f_n.

Then the arithmetic mean M is given by

$$M = \frac{\Sigma fx}{\Sigma f}. \qquad ...(1)$$

Now let every value of the variate x be increased by the same constant 'a', so that the frequency distribution becomes

$y = x + a$	:	$x_1 + a$	$x_2 + a$	...	$x_n + a$
f	:	f_1	f_2	...	f_n

If $\overline{M}$ is the arithmetic mean of the new frequency distribution, then

$$\overline{M} = \frac{\Sigma fy}{\Sigma f} \Rightarrow \overline{M} = \frac{\Sigma f(x + a)}{\Sigma f}$$

$$= \frac{\Sigma fx}{\Sigma f} + \frac{a \, \Sigma f}{\Sigma f}.$$

$$\therefore \quad \overline{M} = M + a.$$

Property 3. *Arithmetic mean is not independent of the change of origin and scale.*

Proof. Change of origin and scale is defined is Remark 2, § 4.6.

Let the given frequency distribution be

x	:	x_1	x_2	...	x_n
f	:	f_1	f_2	...	f_n

Then, the arithmetic mean $\overline{M}$ is given by

$$\overline{M} = \frac{\Sigma fx}{\Sigma f}. \qquad ...(1)$$

By change of origin and scale (see § 4.6.3, step-deviation method), let

$$u = \frac{x - a}{i}$$

[where a is any arbitrary point and i is the width of interval]

In this case, let $\overline{u}$ be the arithmetic mean, then

$$\overline{u} = \frac{\Sigma fu}{\Sigma f}$$

$$\Rightarrow \quad \overline{u} = \frac{\Sigma f\left(\frac{x-a}{i}\right)}{\Sigma f} = \frac{1}{i}\frac{\Sigma f(x-a)}{\Sigma f}$$

$$= \frac{1}{i}\left(\frac{\Sigma fx}{\Sigma f} - \frac{a\,\Sigma f}{\Sigma f}\right) = \frac{1}{i}(\overline{M} - a) \qquad \text{[using (1)]}$$

This proves the statement.

Property 4. *The sum of the squares of the deviations of all the variates taken about their mean is minimum.*

Proof. Let the given frequency distribution be as follows :

x :	x_1	x_2	...	x_i	...	x_n
f :	f_1	f_2	...	f_i	...	f_n

Let A be any arbitrary point and let U be the sum of the squares of the deviations of the given value from A, then

$$U = \sum_{i=1}^{n} f_i (x_i - A)^2 \qquad ...(1)$$

Differentiating (1) w. r. t. A, we have

$$\frac{dU}{dA} = \sum_{i=1}^{n} 2f_i(x_i - A)(-1) = -2\sum_{i=1}^{n} f_i(x_i - A)$$

and

$$\frac{d^2U}{dA^2} = -2\sum_{i=1}^{n} f_i(-1) = 2\sum_{i=1}^{n} f_i$$

The conditions for U to be minimum are

$$\frac{dU}{dA} = 0 \text{ and } \frac{d^2U}{dA^2} = +\text{ve}$$

Now

$$\frac{dU}{dA} = 0 \Rightarrow -2\sum_{i=1}^{n} f_i(x_i - A) = 0$$

$$\Rightarrow \sum_{i=1}^{n} f_i x_i = A\sum_{i=1}^{n} f_i$$

$$\Rightarrow A = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = M$$

where M is the arithmetic mean for the variates x_i, $i = 1, 2, \ldots, n$.

Also $\dfrac{d^2U}{dA^2} = +\text{ve}$ for all values of A.

Hence U is minimum when $A = M$.

§ 4.8. COMPOUND ARITHMETIC MEAN

Property. *If $M_1, M_2, \ldots, M_k$ are arithmetic means of k distributions whose corresponding frequencies are $n_1, n_2, \ldots, n_k$ respectively, then the mean M, called compound arithmetic mean of the whole distribution is given by*

$$M = \frac{n_1M_1 + n_2M_2 + \ldots + n_kM_k}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{1}{N}\sum_{i=1}^{n} n_r M_r$$

where $N = n_1 + n_2 + \ldots + n_k$.

Proof. Let $x_{11}, x_{12}, \ldots, x_{1n_1}$ be n_1 observations in the first distribution, $x_{21}, x_{22}, \ldots, x_{2n_2}$ be n_2 observations in the second distribution, ..., and $x_{k1}, x_{k2}, \ldots, x_{kn_k}$ be n_k observations in the kth distribution, then

$$M_1 = \frac{1}{n_1}(x_{11} + x_{12} + \ldots + x_{1n_1})$$

or $$M_1 = (\Sigma x_{11})/n_1 \text{ (say)}$$

$\therefore$ $$\Sigma x_{11} = n_1M_1$$

$$M_2 = \frac{1}{n_2}(x_{21} + x_{22} + \ldots + x_{2n_2})$$

$$= \frac{1}{n_2}\Sigma x_{21} \text{ (say)}$$

$\therefore$ $$\Sigma x_{21} = n_2M_2$$

$$M_k = \frac{1}{n_k}(x_{k1} + x_{k2} + \ldots + x_{kn_k})$$

$$= \frac{1}{n_k}\Sigma x_{k1} \text{ (say)}$$

$\therefore$ $$\Sigma x_{k1} = n_kM_k$$

Now $$M = \frac{\Sigma x_{11} + \Sigma x_{21} + \ldots + \Sigma x_{k1}}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{n_1 M_1 + n_2 M_2 + n_k M_k}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{1}{N} \sum_{r=1}^{k} n_r M_r$$

where $N = n_1 + n_2 + \ldots + n_k$.

ILLUSTRATIVE EXAMPLES

Example 1. ***Show that the arithmetic mean of first n natural numbers whose weights are equal to the corresponding number is equal to*** $\frac{1}{3}(2n+1)$.

Solution. We have

x	:	1	2	3	...	n
w	:	1	2	3	...	n

$$\therefore \quad \bar{x}_w = \frac{\Sigma xw}{\Sigma w} \qquad \ldots(1)$$

Now $$\Sigma xw = 1 \cdot 1 + 2 \cdot 2 + \ldots + n \cdot n$$

$$= 1^2 + 2^2 + \ldots + n^2 = \Sigma n^2$$

$$= \frac{n(n+1)(2n+1)}{6}$$

and $$\Sigma w = 1 + 2 + 3 + \ldots + n$$

$$= \Sigma n = \frac{n(n+1)}{2}.$$

Putting values in (1), we get

$$\bar{x}_w = \frac{n(n+1)(2n+1)}{6} \times \frac{2}{n(n+1)} = \frac{2n+1}{3}.$$

Example 2. ***Show that the weighted mean of first n natural numbers whose weights are equal to the squares of the corresponding number is equal to*** $\frac{3n(n+1)}{2(2n+1)}$.

Solution. We have

x	:	1	2	...	n
w	:	1^2	2^2	...	n^2

Now $$\Sigma xw = 1^3 + 2^3 + 3^3 + \ldots + n^3 = \Sigma n^3$$

$$= \left\{\frac{n(n+1)}{2}\right\}^2$$

$$\Sigma w = 1^2 + 2^2 + 3^2 + \ldots + n^2 = \Sigma n^2$$

$$= \frac{n(n+1)(2n+1)}{6}.$$

$$\therefore \quad \bar{x}_w = \frac{\Sigma xw}{\Sigma w}$$

$$= \frac{n^2(n+1)^2}{4} \times \frac{6}{n(n+1)(2n+1)} = \frac{3n(n+1)}{2(2n+1)}.$$

Example 3. (a) ***Evaluate the arithmetic mean of the series*** $1, 3, 3^2, 3^3, \ldots, 3^n$.

Solution. Here, the number of observations $= n + 1$.

$$\therefore \quad \text{Arithmetic mean} = \frac{1 + 3 + 3^2 + \ldots + 3^n}{n+1}$$

$$= \frac{1 \cdot \frac{(3^{n+1} - 1)}{3 - 1}}{(n+1)} \qquad \text{[sum of Geometric progression]}$$

$$= \frac{3^{n+1} - 1}{2(n+1)}.$$

Example 3. (b) ***Show that the arithmetic mean of the series*** $1, 2, 2^2, \ldots, 2^n$ ***is*** $\frac{2^{n+1} - 1}{n+1}$.

Solution. Proceed as example 3 (a) above.

Example 4. ***The mean age of a combined group of men and women is 30 years. If the mean age for the group of men is 32 and that of women is 27, find out the percent age of the men and women in the group.***

Solution. Let N_1 = number of men,

M_1 = mean age of men

N_2 = number of women,

M_2 = mean age of women

$N = N_1 + N_2$

M = mean age of combined group of men and women.

Thus $M = 30$ years,

$M_1 = 32$ years,

$M_2 = 27$ years.

Now using the formula [see § 4.8]

$$NM = N_1 M_1 + N_2 M_2$$

$$\Rightarrow \quad (N_1 + N_2)\, 30 = 32 N_1 + 27 N_2$$

$$\Rightarrow \quad 3N_2 = 2N_1 \Rightarrow \frac{N_1}{N_2} = \frac{3}{2} \Rightarrow N_1 : N_2 = 3 : 2$$

Percent age of men in the group $= \frac{N_1}{N_1 + N_2} \times 100$

$= \frac{3}{5} \times 100 = 60.$

Percent age of women in the group $= \frac{2}{5} \times 100 = 40.$

Example 5. ***The arithmetic means of three sets are 25, 10 and 15 whose corresponding number of observations are 200, 250 and 300 respectively. Find the combined arithmetic mean.***

Solution. Let $M_1 = 25, M_2 = 10, M_3 = 15$

$N_1 = 200, N_2 = 250, N_3 = 300$

$\therefore \quad N = N_1 + N_2 + N_3 = 750.$

The combined arithmetic mean M is given by

$NM = N_1M_1 + N_2M_2 + N_3M_3$ [see § 4.8]

i.e., $750M = 200 \times 25 + 250 \times 10 \times 300 \times 15$

$\Rightarrow \quad M = (5000 + 2500 + 4500)/\ 750 = 16.$

Example 6. (a) ***The mean marks of 100 students were found to be 40, later on it was discovered that a score of 53 was misread as 83. Find the corrected mean corresponding to the corrected score.***

Solution. Let n = number of students = 100

M = incorrect mean marks of 100 students = 40.

So that incorrect total marks of 100 students = $nM = 100 \times 40 = 4000.$

Total of corrected score (marks) = 4000 – incorrect score + correct score

$= 4000 - 83 + 53 = 3970.$

$\therefore$ Corrected mean $= \frac{3970}{100} = 39 \cdot 7$ marks.

Example 6. (b) ***The mean of 200 items was 50. Later on, it was discovered that two items were misread as 92 and 8 instead of 192 and 88. Find out the correct mean.***

Solution. Here $n = 200$, incorrect mean $M = 50.$

Incorrect total value of

$\Sigma x_i = nM = 200 \times 50 = 10000.$

Total corrected value of Σx_i = 10000 – incorrect values + correct values

$= 10000 - (92 + 8) + (192 + 88) = 10180.$

$\therefore$ Correct mean $= \frac{\text{corrected value of } \Sigma x_i}{n}$

$= \frac{10180}{200} = 50 \cdot 9.$

Example 7. ***Prove that*** $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0,$

where $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n.$

Solution. Since $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n.$

$\therefore \quad x_1 + x_2 + \ldots + x_n = n\bar{x}$

$\Rightarrow \quad x_1 + x_2 + \ldots + x_n = \bar{x} + \bar{x} + \ldots$ upto n terms

$\Rightarrow \quad (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0.$

Example 8. ***The frequencies of values 0, 1, 2, ..., n of a variable are given by***

$$q^n,\ {}^nC_1 q^{n-1} p,\ {}^nC_2 q^{n-2} p^2, \ldots,\ p^n$$

where $p + q = 1$ ***Show that the mean is np.***

Solution. Here the frequency distribution is

x	:	0	1	2	...	n
f	:	q^n	${}^nC_1 q^{n-1} p$	${}^nC_2 q^{n-2} p^2$	...	p^n

$\therefore$ Mean $= \frac{\Sigma fx}{\Sigma f}$

$= \frac{q^n(0) + {}^nC_1\, q^{n-1}\, p(1) + {}^nC_2 q^{n-2}\, p^2(2) + \ldots + p^n(n)}{q^n + {}^nC_1 q^{n-1} p + \ldots + p^n}$

$= \frac{{}^nC_1 q^{n-1} p + 2 \cdot {}^nC_2 q^{n-2} p^2 + \ldots + np^n}{(q + p)^n}$

$= nq^{n-1} p + 2 \cdot \frac{n(n-1)}{2!} q^{n-2} p^2 + \ldots + np^n$ [$\because\ p + q = 1$]

$= np\,[q^{n-1} + {}^{n-1}C_1 q^{n-2} p + \ldots + p^{n-1}]$

$= np\,(q + p)^{n-1} = np.$ [$\because\ q + p = 1$]

Example 9. ***The mean of 6 observations is 12. If each observation is divided by 3, find the new mean.***

Solution. Sum of six observation $= nM = 6 \times 12 = 72$

when each observation is divided by 3.

New mean $= \frac{\frac{1}{3}(x_1 + x_2 + \ldots + x_6)}{6} = \frac{1}{3} \times \frac{72}{6} = 4.$

Example 10. ***The mean monthly salary paid to 75 workers in a company is Rs. 1420. The mean salary of 25 of them is Rs. 1350 and that of 30 others is Rs 1425. What is the mean salary of the remaining ?***

Solution. Here $n = n_1 + n_2 + n_3.$

$\therefore \quad 75 = 25 + 30 + n_3 \Rightarrow n_3 = 20.$

Let $\bar{x}_1 = 1350, \bar{x}_2 = 1425, \bar{x}_3 = ?, \bar{x} = 1420.$

We know that

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} \quad [\text{see } \S 4.8]$$

$$\Rightarrow \quad 1420 = \frac{25 \times 1350 + 30 \times 1425 + 20 \times \bar{x}_3}{75}$$

$$\Rightarrow \quad 33750 + 42750 + 20\bar{x}_3 = 106500$$

$$\Rightarrow \quad 20\bar{x}_3 = 30000 \Rightarrow \bar{x}_3 = 1500.$$

Hence the mean salary of the remaining 20 workers is Rs. 1500.

Example 11. *If* $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{x}_2 = \frac{1}{n}\sum_{i=2}^{n+1} x_i$, *and* $\bar{x}_3 = \frac{1}{n}\sum_{i=3}^{n+2} x_i$, *then show that*

(a) $\bar{x}_2 = \bar{x}_1 + \frac{1}{n}(x_{n+1} - x_1)$, *and* *(b)* $\bar{x}_3 = \bar{x}_2 + \frac{1}{n}(x_{n+2} - x_2)$.

Solution. We have

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

$$\bar{x}_2 = \frac{1}{n}\sum_{i=2}^{n+1} x_i = \frac{1}{n}(x_2 + x_3 + \ldots + x_n + x_{n+1})$$

$$\bar{x}_3 = \frac{1}{n}\sum_{i=3}^{n+2} x_i = \frac{1}{n}(x_3 + x_4 + \ldots + x_n + x_{n+1} + x_{n+2}).$$

Now $\quad \bar{x}_2 - \bar{x}_1 = \frac{1}{n}(x_{n+1} - x_1)$

$$\Rightarrow \quad \bar{x}_2 = \bar{x}_1 + \frac{1}{n}(x_{n+1} - x_1).$$

And, $\quad \bar{x}_3 - \bar{x}_2 = \frac{1}{n}(x_{n+2} - x_2)$

$$\Rightarrow \quad \bar{x}_3 = \bar{x}_2 + \frac{1}{n}(x_{n+2} - x_2).$$

Example 12. ***If every value of the variate is multiplied by the same constant 'a', then the arithmetic mean is also multiplied by 'a'.***

Solution. Let the frequency distributive be :

x	:	x_1	x_2	...	x_n
f	:	f_1	f_2	...	f_n

Then the arithmetic mean is given by

$$M = \frac{\Sigma fx}{\Sigma f}. \quad \ldots(1)$$

If every value of the variate x is multiplied by the same constant 'a', then the new values of variate are $ax_1, ax_2, \ldots, ax_n$. Then new arithmetic mean

$$= \frac{f_1(ax_1) + f_2(ax_2) + \ldots + f_n(ax_n)}{\Sigma f}$$

$$= \frac{a(f_1x_1 + f_2x_2 + \ldots + f_nx_n)}{\Sigma f} = \frac{a\,\Sigma fx}{\Sigma f} = aM.$$

EXERCISE 4 (A)

1. Mark obtained by 9 Students in statistics are given below :
   52 75 40 70 43 40 65 35 48
   Calculate the arithmetic mean.

2. Calculate the arithmetic mean of the following distribution :

Variate	:	6	7	8	9	10	11	12
Frequency	:	5	8	10	12	7	6	4

3. Calculate the arithmetic mean of the following distribution :

Variate	:	5	10	15	20	25	30	35	40	45	50
Frequency	:	20	43	75	67	72	45	39	9	8	6

4. Calculate the mean of the following distribution :

Heights in cms.	:	65	66	67	68	69	70	71	72	73
Number of Plants	:	1	4	5	7	11	10	6	4	2

5. Find the mean of the following distribution :

Class	:	0–10	10–20	20–30	30–40	40–50
Frequency	:	31	44	139	29	10

6. Find the mean of the following distribution :

Class	:	0–7	7–14	14–21	21–28	28–35	35–42	42–49
Frequency	:	19	25	36	72	51	43	28

7. Following is the frequency distribution of yield of cane in tons per acre. Calculate the mean

Class	Frequency	Class	Frequency
35–40	7	60–65	42
40–45	8	65–70	42
45–50	12	70–75	15
50–55	26	75–80	17
55–60	32	80–85	9

8. Calculate the mean of the following frequency distribution :

Class	Frequency	Class	Frequency
45–50	2	70–75	11
50–55	3	75–80	7
55–60	5	80–85	2
60–65	7	85–90	3
65–70	9	90–95	1

9. Calculate the mean of the following frequency distribution :

Class	:	0–11	11–22	22–33	33–44	44–55	55–66
Frequency	:	9	17	27	24	15	8

10. If the arithmetic average of the following frequency distribution is 7·85, find the missing term :

Daily wages in Rupees	:	5	6	7	10	12	15
No. of Labourers	:	10	?	13	8	5	4

11. Find the mean from the following data :

Marks	No. of Students	Marks	No. of Students
Below 10	5	Below 60	60
Below 20	9	Below 70	70
Below 30	17	Below 80	78
Below 40	29	Below 90	83
Below 50	45	Below 100	85

12. The mean of n observations $x_1, x_2, \ldots, x_n$ is $\bar{x}$. What is the new mean if each observation is divided by k.

13. If the mean of n numbers of a series is M, the sum of first $(n-1)$ numbers is k, find the value of the last number.

14. If x, y, u and v are variables and a, b, h and k are constants such that $u = \dfrac{x-a}{h}$ and $v = \dfrac{y-b}{h}$, show that

(a) $\bar{x} = a + h\bar{u}$ (b) $\bar{y} = b + k\,\bar{v}$.

15. If there are two variables u and v in which the values u_i corresponds to the value v_i for each i, and a new variable $z = au + bv$ is formed, show that $\bar{z} = a\bar{u} + b\bar{v}$, where $\bar{u}, \bar{v}, \bar{z}$ denote the arithmetic means of the variables u, v, z respectively.

16. If $\bar{x}_w$ is the weighted mean of x_i's with weights w_i's prove that

$$\left[\sum_{i=1}^{n} w_i\right]\left[\sum_{i=1}^{n} w_i\,(x_i - \bar{x}_w)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i\,w_j\,(x_i - x_j)^2,$$

where $\Sigma\, w_i \neq 0$.

17. What is the new arithmetic mean if each value is (i) increased by a constant a and (ii) multiplied by a constant b.

ANSWERS

1. 52 **2.** 8·81 **3.** 22·2 **4.** 69·18 cms. **5.** 22·75
6. 26·5 **7.** 61·84 Tons. **8.** 68·9 **9.** 32·23 **10.** 15·05
11. 48·41 **12.** $\bar{x}/k$ **13.** $nM - k$.

§ 4.9. MEDIAN

The *median* is defined as the measure of the central term, when the given terms (*i.e.*, values of the variate) are arranged in the ascending or descending order of magnitudes. In other words the *median* is that value of the variate for which total of the frequencies above this value is equal to the total of the frequencies below this value.

Due to **Corner,** "The **median** is that value of the variable which divides the group into two equal parts one part comprising all values greater, and the other all values less than the median".

For example. The marks obtained, by seven students in a paper of Statistics are 15, 20, 23, 32, 34, 39, 48 the maximum marks being 50, then the **median** is 32 since it is the value of the 4th term, which is situated such that the marks of 1st, 2nd and 3rd students are less than this value and those of 5th, 6th and 7th students are greater than this value.

§ 4.10. COMPUTATION OF MEDIAN

(a) Median in individual series. Let n be the number of values of a variate (*i.e.*, number of terms *i.e.*, total of all frequencies).

Two cases arises :

Case I. If n is **odd** then the value of $\dfrac{1}{2}(n+1)$th term gives the median.

Case II. If n is **even** then there are two central terms *i.e.*, $\dfrac{1}{2}$ nth and $\left(\dfrac{1}{2}n + 1\right)^{\text{th}}$. The mean of these two values gives the median.

(b) Median in discrete series. First of all we write the values of the variate (*i.e.*, the terms) in ascending or descending order of magnitudes, then find the cumulative frequencies. Now find median as in case I or II of (a) above.

(c) Median in continuous series (or grouped series). In this case, the median (M_d) is computed by the following formula

$$\mathbf{M}_d = l + \frac{\frac{1}{2}\mathbf{N} - \mathbf{F}}{\mathbf{f}} \times \mathbf{i}. \qquad \ldots(1)$$

Where M_d = median
l = lower limit of median class
N = total frequency
F = total of all frequencies before median class
f = frequency of median class
i = class width of median class.

Sometimes the following formula (2) is used to find the median

$$\mathbf{M}_d = l_1 + \frac{\mathbf{m} - \mathbf{F}}{\mathbf{f}}(l_2 - l_1). \qquad \ldots(2)$$

Where M_d = median
l_1 = lower limit of median class
l_2 = upper limit of median class
m = size of middle item
F = total of all frequencies before median class
f = frequency of median class.

ILLUSTRATIVE EXAMPLES

Example 1. *According to the census of 1991, following are the population figure, in thousands, of 10 cities :*

1400, 1250, 1670, 1800, 700, 650, 570, 488, 2100, 1700.

Find the median.

Solution. Arranging the terms in ascending order.

488, 570, 650, 700, 1250, 1400, 1670, 1700, 1800, 2100.

Here $n = 10$, therefore the median is the mean of the measure of the 5th and 6th terms.

Here 5th term is 1250 and 6th term is 1400.

$$\therefore \quad \text{Median } (M_d) = \frac{1250 + 1400}{2} \text{ Thousands}$$

$$= 1325 \text{ Thousands.}$$

Example 2. ***Below are given the marks obtained by a batch of 15 students in a certain test in Mathematics and English :***

Roll No. of Students	Marks in Maths.	Marks in English	Roll No. of Students	Marks in Maths.	Marks in English
1	46	42	9	30	47
2	20	24	10	61	42
3	41	38	11	50	55
4	43	35	12	63	54
5	25	30	13	45	52
6	54	45	14	56	47
7	47	58	15	58	43
8	36	50			

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

Solution. Here total marks obtained in Maths.

$= 675 =$ total marks obtained in English

$\therefore$ means of marks in Maths. = means of marks in English

$= 675/15 = 45$ marks.

Thus the method of mean can not be used to find the level of knowledge of the students.

Therefore, in order to find the subject in which the level of knowledge is higher, we shall find the medians of both the series, then the subject having the median value higher, will be the subject in which the level of knowledge of the students is higher.

Hence rearranging the two series in ascending order of magnitudes :

Serial No.	Marks in Maths.	Marks in English	Serial No.	Marks in Maths.	Marks in English
1	20	24	9	47	47
2	25	30	10	50	50
3	30	35	11	54	52
4	36	38	12	56	54
5	41	40	13	58	55
6	43	42	14	61	58
7	45	43	15	63	62
8	46	45			

Now median marks in Maths. = measure of $\left(\frac{15+1}{2}\right)^{th}$ term

= measure of 8^{th} term = 46

Median marks in English = measure of $\left(\frac{15+1}{2}\right)^{th}$ term

= measure of 8^{th} term = 45.

Since the median marks in Maths. are more than those in English, hence the level of knowledge in Maths. is higher.

Example 3. (a) ***Find the median from the following data***

Class	**0–6**	**6–12**	**12–18**	**18–22**	**22–24**	**24–30**	**30–36**	**36–42**
Frequency	**5**	**11**	**25**	**20**	**15**	**18**	**12**	**6**

Solution. Here the class intervals are not equal. To compute median we do not require either to make it equal or any other corrections.

Class	Frequency	Cumulative Frequency
0–6	5	5
6–12	11	16
12–18	25	41
18–22	20	61
22–24	15	76
24–30	18	94
30–36	12	106
36–42	6	112
Total	112	—

Median $= \frac{N}{2}$th term $= \frac{112}{2}$th term $= 56$th term.

Clearly this term is situated in the class 18–22. Hence the median class is 18–22.

Here $l = 18$, $N = 112$, $F = 41$, $f = 20$, $i = 4$.

$$\therefore \quad \text{Median } (M_d) = l + \frac{\frac{1}{2}N - F}{f} \times i$$

$$= 18 + \frac{56 - 41}{20} \times 4$$

$$= 18 + \frac{15 \times 4}{20} = 18 + 3 = 21.$$

Example 3. (b) ***Find the median for the following distribution :***

Wages in Rs.	0–10	10–20	20–30	30–40	40–50
No. of workers	22	38	46	35	20

Solution. We shall calculate the cumulative frequencies.

Wages in Rs.	No. of Workers f	Cumulative Frequencies (c. f.)
0–10	22	22
10–20	38	60
20–30	46	106
30–40	35	141
40–50	20	161

Here $N = 161$. Therefore median is the measure of $\frac{1}{2}(N + 1)$th term *i.e.*, $\frac{1}{2}(161 + 1)^{th}$ term *i.e.*, of 81th term. Clearly 81th term is situated in the class 20–30. Thus 20–30 is the median class. Consequently.

$$\therefore \quad \text{Median } (M_d) = l + \frac{\frac{1}{2}N - F}{f} \times i$$

$$= 20 + \frac{\frac{1}{2} \times 161 - 60}{46} \times 10 = 20 + \frac{205}{46} = 20 + 4 \cdot 46 = 24 \cdot 46.$$

If we apply the following formula, then

$$\text{Median} = l_1 + \frac{m - F}{f}(l_2 - l_1)$$

$$= 20 + \frac{81 - 60}{46} \times (30 - 20) = 20 + \frac{21 \times 10}{46} = 24 \cdot 57.$$

Example 4. ***The following table gives the frequency distribution of married women by age at marriage :***

Age (in years)	Frequency	Age (in years)	Frequency
15–19	53	40–44	9
20–24	140	45–49	5
25–29	98	50–54	3
30–34	32	55–59	3
35–39	12	60 *and above*	2

Calculate the median and interpret the result.

Solution.

Age (in years) (class in exclusive form)	Frequency f	Cumulative Frequency (C.F.)
14·5–19·5	53	53
19·5–24·5	140	193
24·5–29·5	98	291
29·5–34·5	32	323
34·5–39·5	12	335
39·5–44·5	9	344
44·5–49·5	5	349
49·5–54·5	3	352
54·5–59·5	3	355
59·5 and above	2	357
Total	$N = \Sigma f = 357$	

Here $N = 357$. Therefore, the median is the measure of $\frac{1}{2}(N + 1)$th term *i.e.*, of $\frac{1}{2}(357 + 1)$th term *i.e.*, of 179th term which lies in the class 19·5–24·5.

Thus 19·5–24·5 is the median class.

Consequently

$$\text{Median } (M_d) = l + \frac{\frac{1}{2}N - F}{f} \times i$$

$$= 19 \cdot 5 + \frac{178 \cdot 5 - 53}{140} \times 5 = 19 \cdot 5 + \frac{125 \cdot 5}{28}$$

$$= 19 \cdot 5 + 4 \cdot 48 = 23 \cdot 98 = 24 \text{ years (nearly).}$$

It tells us that nearly 50% of the women are married between the ages 15 years and 24 years and another 50% after they have reached the age of 24 years.

Example 5. *An incomplete frequency distribution is given below :*

Class	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequencies	12	30	?	65	?	25	18

Find the missing frequencies when the total of frequencies is 229 and the median is 46.

Solution. Let the missing frequencies of class 30–40 be f_1 and that of 50–60 be f_2. Let us calculate the cumulative frequencies.

Class	Frequencies f	Cumulative Frequencies (C. F.)
10–20	12	12
20–30	30	42
30–40	f_1	$42 + f_1$
40–50	65	$107 + f_1$
50–60	f_2	$107 + f_1 + f_2$
60–70	25	$132 + f_1 + f_2$
70–80	18	$150 + f_1 + f_2$
Total	$N = \Sigma f = 229$ (given)	

Here $N = 150 + f_1 + f_2 = 229 \Rightarrow f_1 + f_2 = 79.$ ...(1)

We are given that the value of median is 46, which lies in the class 40–50. Therefore, the median class is 40–50.

$\therefore$ $N = 229, l = 40, f = 65, F = 42 + f_1, i = 10.$

Now median $(M_d) = l + \dfrac{\frac{1}{2}N - F}{f} \times i$

$$\Rightarrow \quad 46 = 40 + \frac{\frac{1}{2} \times 229 - (42 + f_1)}{65} \times 10$$

$$\Rightarrow \quad 46 - 40 = \frac{2(114 \cdot 5 - 42 - f_1)}{13}$$

$$\Rightarrow \quad 6 \times 13 = 145 - 2f_1$$

$$\Rightarrow \quad 2f_1 = 145 - 78 = 67 \Rightarrow f_1 = 33 \cdot 5.$$

Since the frequency f_1 cannot be a rational number, hence $f_1 = 34$. Substituting this value of f_1 in equation (1), we get $f_2 = 45$.

Hence the missing frequencies are 34, 45.

Example 6. *Find the median of the following frequency distribution :*

Marks	No. of students	Marks	No. of students
Less than 10	15	*Less than 50*	106
Less than 20	35	*Less than 60*	120
Less than 30	60	*Less than 70*	125
Less than 40	84		

Solution. The cumulative frequency distribution table :

Class (Marks)	Frequency f (No. of students)	Cumulative Frequency (C.F.)
0–10	15	15
10–20	20	35
20–30	25	60
30–40	24	84
40–50	22	106
50–60	14	120
60–70	5	125
Total	$N = 125$	

Median = measure of $\left(\dfrac{125 + 1}{2}\right)$th term

= measure of th term63th term.

Clearly 63th term is situated in the class 30–40.

Thus median class = 30 – 40.

$\therefore$ Median $M_d = l + \dfrac{\frac{1}{2}N - F}{f} \times i$

$$= 30 + \frac{\frac{125}{2} - 60}{24} \times 10 = 30 + \frac{25}{24} = 30 + 1 \cdot 04 = 31 \cdot 04.$$

§ 4.11. PARTITION VALUES

If the values of the variate are arranged in ascending or descending order of magnitudes then we have seen above that median is that value of the variate which divides the total frequencies in two equal parts. Similarly the given series can be divided into four, ten and hundred equal parts. The values of the variate dividing into four equal parts are called **Quartile**, into ten equal parts are called **Decile** and into hundred equal parts are called **Percentile**.

Quartiles :

Definition. The values of the variate which divide the total frequency into four equal parts, are called **quartiles**. That value of the variate which divides the total frequency into two equal parts is called median. The **lower quartile** or **first quartile** denoted by Q_1 divides the frequency between the lowest value and the median into two equal parts and similarly the **upper quartile** (or **third quartile**) denoted by Q_3 divides the frequency between the median and the greatest value into two equal parts. The formulas for computation of quartiles are given by

$$Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i, \quad Q_3 = l + \frac{\frac{3}{4}N - F}{f} \times i$$

where l, F, N, f, i have the same meaning as in the formula for median.

Deciles :

Definition. Those values of the variate which divide the total frequency into ten equal parts are called **deciles**. The formulas for computation are given by

$$D_1 = l + \frac{(1/10)\,N - F}{f} \times i, \quad D_2 = l + \frac{(2/10)\,N - F}{f} \times i, \text{ etc.}$$

Percentiles :

Definition. Those values of the variate which divide the total frequency into hundred equal parts, are called **percentiles**. The formulas for computation are :

$$P_1 = l + \frac{(1/100)\,N - F}{f} \times i,$$

$$P_2 = l + \frac{(2/100)\,N - F}{f} \times i \text{ etc.}$$

In general, $$P_r = l + \frac{(r/100)\,N - F}{f} \times i, \; 1 \le r \le 100.$$

◈ § 4.11. (A) LOCATION OF MEDIAN, QUARTILES AND PARTITION VALUES BY GRAPHS

In order to locate, graphically, the median, quartiles or partition values, we draw 'less than' type cumulative frequency curve (Ogive) or polygon.

To locate median. Firstly mark a point corresponding to $\frac{N}{2}$ along y-axis, from this point draw a straight line parallel to x-axis meeting the polygon at the point P, say. From P draw a perpendicular PQ on the x-axis, meeting the x-axis at Q. Then the distance OQ (*i.e.*, abscissa of Q) of the point Q from the origin O is the required median.

To locate quartiles. To locate Q_1 and Q_3 mark points $\frac{N}{4}$ and $\frac{3N}{4}$ respectively on y-axis and proceed as in median above.

To locate deciles. To locate $D_1, D_2, \ldots$, mark points $\frac{N}{10}, \frac{2N}{10}, \ldots$ on the y-axis and proceed as in median above.

To locate percentiles. To locate $P_1, P_2, \ldots$, mark points $\frac{N}{100}, \frac{2N}{100}, \ldots$ on the y-axis and proceed as in median above.

Remark 1. The median, quartiles etc. can also be located by drawing 'more than' type cumulative frequency polygon. In this case mark points $N/4$ for Q_1, $2N/4$ for Q_2 (or median), $3N/4$ for Q_3 on y-axis.

Remark 2. If we draw two ogives (*i.e.*, 'less than' type and 'more than' type) with the same scale on the same graph, then these two ogives will intersect at a point P, say. The abscissa of the point P is the median of the given frequency distribution.

Example. What are two types of ogives and define median by the help of them. **[Indore 1984]**

The method will be clear from the following example.

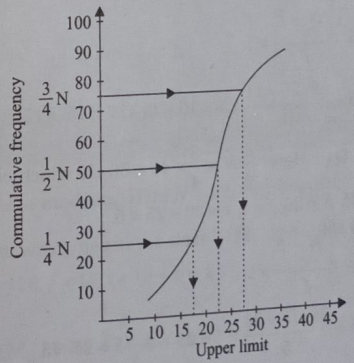
Example. ***The marks obtained by 100 students in statistics are given by***

Marks	:	**0–5**	**5–10**	**10–15**	**15–20**	**20–25**	**25–30**	**30–35**	**35–40**
No. of students	:	**4**	**6**	**10**	**10**	**25**	**22**	**18**	**5**

Draw a cumulative frequency polygon and locate quartiles.

Solution. For 'less than' cumulative frequency polygon, we have

Upper limit	:	5	10	15	20	25	30	35	40
Cumulative frequency	:	4	10	20	30	55	77	95	100



From the graph, we have

$$Q_1 = 17 \cdot 5, Q_2 = M_d = 24, Q_3 = 29 \cdot 5.$$

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the lower and upper quartiles, fourth decile and 70th percentile for the following distribution :***

Marks group	No. of students	Marks group	No. of students
5–10	5	25–30	5
10–15	6	30–35	4
15–20	15	35–40	2
20–25	10	40–45	2

Solution. First we make the cumulative frequency table :

Class	Frequency	Cumulative Frequency	Class	Frequency	Cumulative Frequency
5–10	5	5	25–30	5	41
10–15	6	11	30–35	4	45
15–20	15	26	35–40	2	47
20–25	10	36	40–45	2	49

(i) To compute Q_1. Here $N = 49, \frac{1}{4}N = \frac{1}{4} \times 49 = 12 \cdot 25$ which clearly lies in 15–20. Thus 15–20 is lower quartile class.

$\therefore \quad l = 15, F = 11, f = 15, i = 20 - 15 = 5.$

$$\therefore \quad Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i$$

$$= 15 + \frac{12 \cdot 25 - 11}{15} \times 5 = 15 + 0 \cdot 417 = 15 \cdot 417.$$

(ii) To compute Q_3. Here $\frac{3}{4}N = \frac{3}{4} \times 49 = 36 \cdot 75$ which clearly lies in the class 25–30. Thus $l = 25, F = 36, f = 5, i = 30 - 25 = 5.$

$$\therefore \quad Q_3 = l + \frac{\frac{3}{4}N - F}{f} \times i$$

$$= 25 + \frac{36 \cdot 75 - 36}{5} \times 5 = 25 + 0 \cdot 75 = 25 \cdot 75.$$

(iii) To compute D_4. Here $\frac{4}{10}N = \frac{4}{10} \times 49 = 19 \cdot 6$, which clearly lies in the class 15–20. Thus $l = 15, F = 11, f = 15, i = 5.$

$$\therefore \quad D_4 = l + \frac{\frac{4}{10}N - F}{f} \times i$$

$$= 15 + \frac{19 \cdot 6 - 11}{15} \times 5 = 15 + 2 \cdot 87 = 17 \cdot 87.$$

(iv) To compute P_{70}. Here $\frac{70}{100}N = \frac{7}{10} \times 49 = 34 \cdot 3$ which clearly lies in the class 20–25. Thus $l = 20, F = 26, f = 10, i = 5.$

$$\therefore \quad P_{70} = l + \frac{\frac{70}{100}N - F}{f} \times i$$

$$= 20 + \frac{34 \cdot 3 - 26}{10} \times 5 = 20 + 4 \cdot 15 = 24 \cdot 15.$$

Example 2. ***Calculate the median, lower quartile and upper quartile for the following data :***

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

Solution. Here the class intervals are unequal and therefore, arranging the frequencies as follows :

Class	Frequency f	Cumulative Frequency (C. F.)
0–4	4	4
4–8	6 + 8 = 14	18
8–12	12	30
12–16	5	35
16–20	4	39
	$N = \Sigma f = 39$	

(i) To Calculate median. Here $N = 39$, which is odd, $\therefore$ median $= \frac{1}{2}(N + 1)$th term *i.e.*, 20th term which lies in the class 8–12. So 8–12 is the median class.

Here $l = 8, F = 18, f = 12, i = 12 - 8 = 4.$

$$\therefore \quad \text{Median } (M_d) = l + \frac{\frac{1}{2}N - F}{f} \times i$$

$$= 8 + \frac{\frac{1}{2} \times 39 - 18}{12} \times 4 = 8 + \frac{1}{3}(19 \cdot 5 - 18) = 8 + 0 \cdot 5 = 8 \cdot 5.$$

(ii) To evaluate lower quartile Q_1.

Here lower quartile term $= \frac{1}{4}(N+1)$th term *i.e.*, $\frac{1}{4}(39+1)$th term *i.e.*, 10th term which clearly lies in the class 4–8.

$\therefore \quad l = 4,\ F = 4,\ f = 14,\ i = 4$

$$\therefore \quad Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i = 4 + \frac{\frac{1}{4} \times 39 - 4}{14} \times 4$$

$$= 4 + \frac{23}{14} = 4 + 1 \cdot 643 = 5 \cdot 643.$$

(iii) To compute upper quartile.

Here upper quartile term $= \frac{3}{4}(N+1)$th term *i.e.*, $\frac{3}{4}(39+1)$ *i.e.*, 30th term which clearly lies in the interval 8–12.

$\therefore \quad l = 8,\ f = 12,\ F = 18,\ i = 4.$

$$\therefore \quad Q_3 = l + \frac{\frac{3}{4}N - F}{f} \times i = 8 + \frac{\frac{3}{4} \times 39 - 18}{12} \times 4$$

$$= 8 + \frac{117 - 72}{12} = 8 + 3 \cdot 75 = 11 \cdot 75.$$

EXERCISE 4 (B)

1. (a) Find the median of the following :

   20, 18, 22, 27, 25, 12, 15.

   (b) Below are given the heights in inches of an Indian Hockey Eleven. Compute the median height :

   65, 67, 69, 61, 60, 65, 66, 70, 71, 62, 72.

   (c) The daily wages of 10 workers in a factory are 4, 6, 9, 12, 11, 8, 5, 10, 11, 8. Find the median.

2. Find the median of the following distribution :

Measure :	3	5	7	9	11	13	15
Frequency :	7	3	12	28	10	9	6

3. Find the median from the following table :

Marks	No. of students	Marks	No. of students
0–10	2	40–50	35
10–20	18	50–60	22
20–30	30	60–70	6
30–40	45	70–80	3

4. Find the median from the following table :

Class	20–25	25–30	30–35	35–40	40–45	45–50
Frequency	18	44	102	160	57	19

5. Below are given the marks obtained by a batch of 25 students in a certain test in Mathematics and English :

Roll No. of Students	Marks in Maths.	Marks in English	Roll No. of Students	Marks in Maths.	Marks in English
1	29	36	14	47	44
2	65	30	15	60	85
3	33	38	16	30	20
4	45	39	17	32	32
5	51	64	18	52	25
6	72	50	19	54	55
7	48	46	20	56	28
8	33	15	21	58	53
9	42	42	22	49	35
10	25	10	23	38	40
11	28	72	24	40	62
12	35	33	25	46	58
13	46	80			

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

6. Calculate the median, lower and upper quartiles, third decile and 60th percentile for the following distributions :

Class	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency	5	8	7	12	28	20	10	10

7. Calculate the median, quartiles, 7th decile and 82th percentile for the following distribution :

Wages in Rs.	0–10	10–20	20–30	30–40	40–50
No. of workers	22	38	46	35	25

8. Calculate the quartiles from the following data :

Weekly Wages in Rs.	35	36	37	38	39	40	41	42
No. of workers	14	20	42	54	45	19	7	9

9. Compute the quartiles and median from the following table :

Income	No. of persons
Less than Rs. 30	69
Between Rs. 30 and less than 40	167
Between Rs. 40 and less than 50	207
Between Rs. 50 and less than 60	65
Between Rs. 60 and less than 70	58
Between Rs. 70 and less than 80	27
Rs. 80 and above	10
Total	603

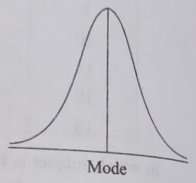
ANSWERS

1. (a) 20 (b) 66 inches (c) Rs. 8·5
3. 36665. In Maths.
6. Median = 46·43, $Q_1 = 34·17$, $Q_3 = 57·5$, $D_3 = 38·3$, $P_{60} = 50$.
7. M_d = Rs. 24·57 Q_1 = Rs. 14·87, $Q_3 = 34·43$, D_7 = Rs. 32·11, P_{82} = Rs. 37·66.
8. $Q_1 = 37·38$, $Q_3 = 39·44$
9. M_d = Rs. 43·2 Q_1 = Rs. 34·9, Q_3 = Rs. 51·7.

§ 4.12. MODE

The word 'Mode' is formed from the French word 'La Mode' which means 'in fashion'. According to Dr. A. L. Bowle 'the value of the graded quantity in a statistical group at which the numbers registered are most numerous, is called the mode or the position of greatest density or the predominant value'.

According to other statisticians, we have 'The value of the variable which occurs most frequently in a distribution is called the mode'.



"The mode of a distribution is the value at the point around the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values".

Definition.

The **mode** is that value (or size) of the variate for which the frequency is maximum or the point of maximum frequency or the point of maximum density. In other words, the mode is the maximum ordinate of the ideal curve which gives the closest fit to the actual distribution.

Methods to Compute the Mode :

(I) When the values (or measures) of all the terms (or items) are given. In this case the mode is the value (or size) of the term (or item) which occurs most frequently.

For example. *Find the mode from the following sizes of shoes*

Size of shoes 3, 4, 2, 1, 7, 6, 6, 7, 5, 6, 8, 9, 5.

Solution. Firstly writing the individual observations in a discrete series as follows :

Size of shoes	1	2	3	4	5	6	7	8	9
Frequency	1	1	1	1	2	3	2	1	1

Here maximum frequency is 3 whose term value is 6. Hence the mode is modal size number 6.

(II) For a **frequency distribution**, the computation of mode is done by the formula given by

$$\text{Mode}\,(M_0) = l + \frac{f_1}{f_{-1} + f_1} \times i. \qquad ...(1)$$

But we shall use the following formula, which is more accurate than the above formula (1) :

$$\text{Mode}\,(M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i \qquad ...(2)$$

where [*for both formula* (1) and (2)]

l = lower limit of modal class,

f = frequency of modal class,

f_{-1} = frequency of the class just preceding to the modal class,

f_1 = frequency of the class just following of the modal class,

i = class interval.

For example. *Find the mode of the following distribution* :

Class	0–7	7–14	14–21	21–28	28–35	35–42	42–49
Frequency	19	25	36	72	51	43	28

Solution. Here maximum frequency 72 lies in the class 21–28. Thus 21–28 is the modal class.

$\therefore \quad l = 21,\ f = 72,\ f_{-1} = 36,\ f_1 = 51,\ i = 7.$

$\therefore \quad \text{Mode}\,(M_0) = l + \dfrac{f_1}{f_{-1} + f_1} \times i$

$$= 21 + \frac{51}{36 + 51} \times 7 = 21 + \frac{357}{87} = 21 + 4·103 = 25·103.$$

Now using formula (2), we have

$$\text{Mode}\,(M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 21 + \frac{72 - 36}{144 - 36 - 51} \times 7 = 21 + \frac{84}{19} = 21 + 4·4 = 25·4.$$

Remarks. (i) The formula given by (2) [or by (1)] is used in case of equal class intervals.

(ii) If the class intervals are not equal (*i.e.*, the distribution is asymmetrical) then we convert the given series into class intervals of equal magnitude and then evaluate the mode. If in some cases, this conversion is not possible then, we use the formula given in § 4.13.

(iii) If in some case, the formula given by (2) fails then use formula given by (1).

(iv) Mode can be graphically located by drawing a histogram. See § 4.13 (A).

(III) Mode by method of grouping. It is an elaborate method of determining mode in case of grouped frequency data. This method is usually applied in the cases when there are two maximum frequencies against two different size of items. This method is also applied in the cases when it is possible that the effect of neighbouring frequencies on the size of item (of maximum frequency) may be greater. The method is as follows :

Firstly the items are arranged in ascending or descending order and corresponding frequencies are written against them.

The frequencies are then grouped in two and then in threes and then in fours (if necessary). In the first stage of grouping, they are grouped (*i.e.*, frequencies are added) by taking, first and second, third and fourth, ..., . In the next stage, similar

grouping is done by taking second and third, fourth and fifth, ..., . After it, the frequencies are added in threes. The frequencies are added in following two ways :

1. (i) First and second, third and fourth, fifth and sixth, seventh and eighth, ...
   (ii) Second and third, fourth and fifth, ...
2. (i) First, second and third; fourth, fifth and sixth, ...
   (ii) Second, third and fourth; fifth, sixth and seventh, ...
   (iii) Third, fourth and fifth; sixth seventh and eighth, ...

Now the items with maximum frequencies are selected and the item which contains the maximum is called the mode. For illustration see following example 1.

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the mode from the following :***

Size of Item	4	5	6	7	8	9	10	11	12	13
Frequency	2	5	8	9	12	14	14	15	11	13

[Jabalpur 1988]

Solution. From the given data we observe that size 11 has the maximum frequency 15, but it is possible that the effect of neighbouring frequencies on the size of the item may be greater. Thus it may happen that the frequencies of size 10 or 12 may be greater and 11 may not remain mode. We shall apply the **method of grouping.**

Size of Items	Frequency					
	I	II	III	IV	V	VI
4	2	7				
5	5		13	15		
6	8	17			22	
7	9		21			29
8	12	26		35		
9	14		28		40	
10	14	29				43
11	15		26	40		
12	11	24			39	
13	13					

We have used brackets against the frequencies which have been grouped.

Now we shall find the size of the item containing maximum frequency :

Column	Size of item having maximum frequency
I	11
II	10, 11
III	9, 10
IV	10, 11, 12
V	8, 9, 10
VI	9, 10, 11

Here size 8 occurs 1 time, 9 occurs 3 times, 10 occurs 5 times, 11 occurs 4 times, 12 occurs 1 time.

Since 10 occurs maximum number of times (5 times).

Hence the required mode is .

Example 2. ***Compute the mode of the following distribution :***

Class-intervals :	**0–7**	**7–14**	**14–21**	**21–28**	**28–35**	**35–42**	**42–49**
Frequency :	**19**	**25**	**36**	**72**	**51**	**43**	**28**

Solution. Here maximum frequency 72 lies in the class-interval 21–28. Therefore 21–28 is the modal class.

$\therefore \quad l = 21,\ f = 72,\ f_{-1} = 36,\ f_1 = 51,\ i = 7$

$$\text{Mode } (M_0) = l + \frac{f_1}{f_{-1} + f_1} \times i$$

$$= 21 + \frac{51}{36 + 51} \times 7 = 21 + \frac{357}{87}$$

$$= 21 + 4 \cdot 103 = 25 \cdot 103.$$

Again using the other formula (which is more accurate) for the mode, we have

$$\text{Mode } (M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 21 + \frac{72 - 36}{144 - 36 - 51} \times 7 = 21 + \frac{252}{57}$$

$$= 21 + \frac{84}{19} = 21 + 4 \cdot 42 = 25 \cdot 42.$$ **Ans.**

Example 3. ***Find the mode of the following :***

Marks	**1–5**	**6–10**	**11–15**	**16–20**	**21–25**	**26–30**	**31–35**	**36–40**	**41–45**
No. of Students	**7**	**10**	**16**	**32**	**24**	**18**	**10**	**5**	**1**

Solution. Here maximum frequency 32 lies in the class 16—20. Therefore 16—20 is the modal class.

$\therefore$ $l = 16,\ f = 32,\ f_{-1} = 16,\ f_1 = 24,\ i = 4.$

$\therefore$ Required mode

Mode $$(M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 16 + \frac{32 - 16}{2 \times 32 - 16 - 24} \times 4 = 16 + \frac{16}{24} \times 4$$

$$= 16 + 2 \cdot 67 = 18 \cdot 67 \text{ (approx.)}.$$

Example 4. ***Compute the mode of the following data :***

Mean Value	15	20	25	30	35	40	45	50	55
Frequency	2	22	19	14	3	4	6	1	1

[Jabalpur 1994; Sagar 98]

Solution. First of all we shall change the given data into a continuous series. Here the class interval is 5. Half of class interval is 2·5. Now substract 2·5 from each mean value to get the lower limit and add 2·5 to each mean value to get the upper limit. Thus the given discrete series transforms to the original continuous series, which is given below :

Mean Value	Class	Frequency	Mean Value	Class	Frequency
15	12·5–17·5	2	40	37·5–42·5	4
20	17·5–22·5	22	45	42·5–47·5	6
25	22·5–27·5	19	50	47·5–52·5	1
30	27·5–32·5	14	55	52·5–57·5	1
35	32·5–37·5	3			

Here the maximum frequency 22 lies in the class 17·5–22·5.

So 17·5–22·5 is the modal class.

$\therefore$ $l = 17 \cdot 5,\ f = 22,\ f_{-1} = 2,\ f_1 = 19,\ i = 5.$

$\therefore$ Required mode $$M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 17 \cdot 5 + \frac{22 - 2}{2 \times 22 - 2 - 19} \times 5$$

$$= 17 \cdot 5 + \frac{100}{23} = 17 \cdot 5 + 4 \cdot 35 = 21 \cdot 85 \text{ (approx)}.$$

Example 5. ***Find the missing frequencies from the following table :***

Wages (in Rs.)	:	0–20	20–40	40–60	60–80	80–100
Number of workers	:	10	?	30	?	14

Total number of workers = 94,

Value of mode = 54.

Solution. Let the missing frequencies of the classes 20–40 and 60–80 be a and b respectively. Then

Class (wages in Rs.)	:	0–20	20–40	40–60	60–80	80–100
Frequency	:	10	a	30	b	14

Total frequency $N = 54 + a + b$

$94 = 54 + a + b$ [$\because$ $N = 94$ (given)]

$\therefore$ $a + b = 40.$ ...(1)

$\Rightarrow$

Since mode (M_0) is 54 (given), so modal class is 40–60. Thus

$$l = 40, f = 30,\ f_{-1} = a,\ f_1 = b,\ i = 20.$$

We have

$$M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$\Rightarrow$ $$54 = 40 + \frac{30 - a}{60 - a - b} \times 20$$

$\Rightarrow$ $$14\,(60 - a - b) = 20\,(30 - a)$$

$\Rightarrow$ $$7b - 3a = 120.$$ ...(2)

Solving (1) and (2), we have $a = 16,\ b = 24.$

Example 6. ***Find the mode and median from the following data***

Marks	No. of students	Marks	No. of students
0–7	12	32–39	15
8–15	11	40–47	06
16–23	16	48–55	04
24–31	18		

Solution. Proceed yourself.

EXERCISE 4 (C)

1. Find the mode from the following data :

   15, 25, 23, 27, 40, 25, 23, 25, 20, 21, 25.

2. Find the mode of the following data :

Measure (x)	2	3	4	5	6	7	8	9	10	11	12	13
Frequency (f)	3	8	10	12	16	14	10	8	17	5	4	1

3. Find the mode for the following frequency distribution :

Height in cms.	52–55	55–58	58–61	61–64
Frequency	15	20	25	30

4. Compute the mode from the following table :

Measure	8	9	10	11	12	13	14	15
Frequency	5	6	8	7	9	8	9	6

5. The following table gives the distribution of wages in a factory :

Wages (in Rs.)	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of Workers	6	10	10	16	12	8	7

Compute the mode.

6. Find the mode from the following table :

Class	0–5	5–10	10–15	15–20	20–25	25–30	30–35
Frequency	1	2	10	4	10	9	2

7. Find the mode from the following table :

Marks	No. of students	Marks	No. of students
0–10	2	40–50	35
10–20	18	50–60	20
20–30	30	60–70	6
30–40	45	70–80	3

8. Find the mode from the following table :

Measure	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79
Frequency	1	2	6	6	12	8	5	3

[Hint. First of all change the given inclusive series into exclusive series *i.e.*, 0–95, 95–195, 195–295 etc.]

9. Find the mode of the following series :

Measure	Frequency	Measure	Frequency
5	48	13	52
6	52	14	41
7	56	15	57
8	60	16	63
9	63	17	52
10	57	18	48
11	55	19	40
12	50		

10. Find the mode of the following data :

Class	0–5	5–10	10–15	15–20	20–25
Frequency	25	15	8	5	2

[Hint. When modal class is first class interval then f_{-1} is zero]

11. Find the mode of the following data :

Class	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	12	21	28	33	54	65	77

[Hint. If modal class is last class interval, then f_1 is zero]

12. Find the mode of the following data :

Class	0–5	5–7	7–9	9–10	10–13	13–15	15–16	16–19	19–20	20–21	21–25
Frequency	2	3	4	2	10	5	3	4	3	2	3

[Hint. Re-arrange in equal class intervals 0–5, 5–10, 10–15, etc.]

13. Find the mode from the following data :

Mean Value	1	2	3	4	5	6
Frequency	2	4	10	8	5	3

14. Find the mode from the following data :

Marks	No. of students	Marks	No. of students
Above 0	80	Above 60	28
Above 10	77	Above 70	16
Above 20	72	Above 80	10
Above 30	65	Above 90	8
Above 40	55	Above 100	0
Above 50	43		

15. Find the mode from the following table :

Marks	No. of students	Marks	No. of students
Below 10	15	Below 50	96
Below 20	35	Below 60	127
Below 30	60	Below 70	198
Below 40	84	Below 80	250

ANSWERS

1. 25 **2.** 16 **3.** 58·75 cm. **4.** 12 **5.** Rs. 36
6. 24·28 **7.** 36 **8.** 35·5 **9.** Size 6 **10.** 3·56
11. 35·67 **12.** 12·727 **13.** 3·25 **14.** 55 **15.** 66·78.

§ 4.13. EMPIRICAL RELATION BETWEEN MEDIAN AND MODE

For moderately asymmetrical distribution (or for asymmetrical curve), the relation

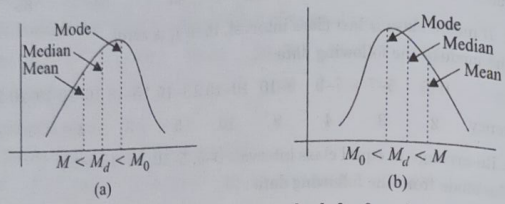
$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}),$$

approximately holds. In such a case, first evaluate mean and median and then mode is determined by

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}.$$

If in the asymmetrical curve, the area on the left of mode is greater than the area on the right [see figure (a)], then in this case

mean < median < mode, *i.e.*, $(M < M_d < M_0)$



If in the asymmetrical curve, the area on the left of mode is less than the area on the right [see fig. (b)] then in this case

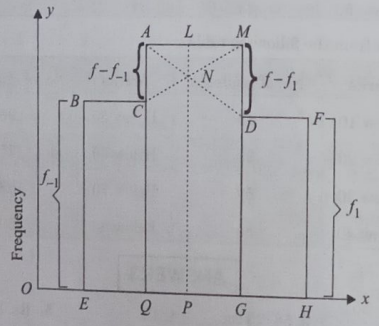
mode < median < mean, *i.e.* $(M_0 < M_d < M)$.

For a symmetrical curve (*i.e.* for a symmetrical distribution) the mean, median and mode coincide.

◈ § 4.13. (A) LOCATION OF MODE BY GRAPH

Mode can be located graphically by drawing a histogram for a given grouped frequency distribution.

We draw histogram for this given grouped frequency distribution. Consider three adjacent rectangles with the tallest rectangle (modal class) in the middle.



(Class-intervals)

In this figure, the rectangle *EBCQ* preceeds the modal class and the rectangle *GDFG* follows the modal class. Suppose *AD* and *CM* intersect at *N* and then draw perpendicular *NP* from *N* on the *x*-axis to meet *x*-axis at *P*. Then *P* determines the mode and is given by the length *OP*.

Now by the properties of similar triangles, we have

$$\frac{AL}{LM} = \frac{AN}{ND} = \frac{CN}{NM}$$

$$\Rightarrow \quad \frac{AL}{AL + LM} = \frac{AN}{AD} = \frac{CN}{CM} = 1 - \frac{NM}{CM}$$

$$\Rightarrow \quad \frac{AL}{i} = \frac{LN}{MD} = 1 - \frac{LN}{AC}$$

$$\Rightarrow \quad \frac{AL}{i} = \frac{LN}{f - f_1} = 1 - \frac{LN}{f - f_{-1}}$$

$$\Rightarrow \quad \frac{AL}{i} = \frac{f - f_{-1}}{f - f_1 + f - f_{-1}}$$

$$\Rightarrow \quad M_0 - l = \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i \qquad [\because AL = OP - OQ = M_0 - l]$$

$$\Rightarrow \quad M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i.$$

Remark. This method of graphical representation of mode can be used in case of unequal class intervals.

◈ § 4.14. GEOMETRIC MEAN

If $x_1, x_2, \ldots, x_n$ are n values of the variate x, none of which is zero, then their geometrical mean G is defined by

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

If $f_1, f_2, \ldots, f_n$ are the frequencies of $x_1, x_2, \ldots, x_n$ respectively, the geometric mean G is given by

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

Where $N = f_1 + f_2 + \ldots + f_n$.

Taking log of (1), we get

$$\log G = \frac{1}{N} [f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n]$$

or
$$\log G = \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i \qquad \ldots(2)$$

or
$$\log G = \frac{1}{N} \sum_{1}^{n} f \log x. \qquad \ldots(3)$$

Note. Above formula (2) shows that the log of G is the weighted arithmetic mean of different values of $\log x_i$.

Mathematical Characteristics of Geometric Mean :

Property 1. *If $G_1, G_2, \ldots, G_n$ be the geometric means of n series whose total frequencies are $N_1, N_2, \ldots, N_n$ then the geometric mean G of the combined series (of all those series) whose total frequency is $N_1 + N_2 + \ldots + N_n = N$, is given by*

$$G = (G_1^{N_1} \, . \, G_2^{N_2} \, . \, \ldots \, G_n^{N_n})^{1/N}$$

or
$$\log G = \frac{1}{N} (N_1 \log G_1 + N_2 \log G_2 + \ldots + N_n \log G_n)$$

or
$$N \log G = \sum_{i=1}^{n} N_i \log G_i.$$

Proof. Let the n given series be as follows :

First	Second	.....	nth
$x_{11}\ f_{11}$	$x_{21}\ f_{21}$	.....	$x_{n1}\ f_{n1}$
$x_{12}\ f_{12}$	$x_{22}\ f_{22}$	.....	$x_{n2}\ f_{n2}$
⋮	⋮		⋮
$x_{1k_1}\ f_{1k_1}$	$x_{2k_2}\ f_{2k_2}$		$x_{nk_n}\ f_{nk_n}$
Total N_1	N_2		N_n

By the definition of Geometric mean

$$G = \left(x_{11}^{f_{11}}\ x_{12}^{f_{12}} \dots x_{1k_1}^{f_{1k_1}}\right)^{1/N_1}$$

or $$G_1^{N_1} = \left(x_{11}^{f_{11}}\ x_{12}^{f_{12}} \dots x_{1k_1}^{f_{1k_1}}\right)$$

= Product of all terms of first series.

Similarly $$G_2^{N_2} = \left(x_{21}^{f_{21}}\ x_{22}^{f_{22}} \dots x_{2k_2}^{f_{2k_2}}\right)$$

= Product of all series of 2nd series.

$$G_n^{N_n} = \left(x_{n1}^{f_{n1}}\ x_{n2}^{f_{n2}} \dots x_{nk_n}^{f_{nk_n}}\right)$$

= Product of terms of nth series.

$\therefore$ Geometrical mean of combined series is given by

G = (Product of all terms of first series × Product of all terms of 2nd series × ... × product of all terms of nth series)$^{1/N}$

or $$G = \left(G_1^{N_1}\ G_2^{N_2} \dots G_n^{N_n}\right)^{1/N}$$

or $$G^N = G_1^{N_1}\ G_2^{N_2} \dots G_n^{N_n}$$

Taking log, we get

$$N \log G = \sum_{i=1}^{n} N_i \log G_i .$$

Property 2. *The geometric mean of the ratios of observations of two series is equal to the ratio of geometric means of these two series.*

Proof. Let $x_{11}, x_{12}, \dots, x_{1n}$ and $x_{21}, x_{22}, \dots, x_{2n}$ be two given series with their geometric means G_1 and G_2 respectively

$$G_1 = (x_{11}\ x_{12} \dots x_{1n})^{1/n},\ G_2 = (x_{21}\ x_{22} \dots x_{2n})^{1/n}.$$

Now the series formed by the ratios of observations is $\frac{x_{11}}{x_{21}}, \frac{x_{12}}{x_{22}}, \dots, \frac{x_{1n}}{x_{2n}}$.

Its geometric mean G is given by

$$G = \left(\frac{x_{11}}{x_{21}} \cdot \frac{x_{12}}{x_{22}} \dots \frac{x_{1n}}{x_{2n}}\right)^{1/n}$$

or $$G = \frac{(x_{11}\ x_{12} \dots x_{1n})^{1/n}}{(x_{21}\ x_{22} \dots x_{2n})^{1/n}} = \frac{G_1}{G_2}.$$

Above result may be written a

$$\log G = \log\ G_1 - \log G_2.$$

Property 3. *The geometric mean of the product of two different series (sets) is equal to the product of their geometric means.*

Proof. Let $\{x_1, x_2, \dots, x_m\}$ and $\{y_1, y_2, \dots, y_n\}$ be the two sets of observations and let G_1, G_2 be their geometric means respectively.

The total number of observations in the product series (or set) are mn. Now

$$G_1 = (x_1 x_2 \dots x_m)^{1/m},\ G_2 = (y_1 y_2 \dots y_n)^{1/n}.$$

The geometric mean G of the product is given by

$$G = [(x_1 y_1 \dots x_1 y_n)(x_2 y_2 \dots x_2 y_n) \dots (x_m y_1 \dots x_m y_n)]^{1/mn}$$

$$= [x_1^n\ (y_1 y_2 \dots y_n) \cdot x_2^n\ (y_1 y_2 \dots y_n) \dots x_m^n\ (y_1 y_2 \dots y_n)]^{1/mn}$$

$$= [(x_1 x_2 \dots x_m)^n\ (y_1 y_2 \dots y_n)^m]^{1/mn}$$

$$= [(G_1^m)^n\ (G_2^n)^m]^{1/mn} = [G_1^{mn}\ G_2^{mn}]^{1/mn}.$$

$\therefore$ $$G = G_1 \cdot G_2.$$

Property 4. *If G is the geometric mean of the series $x_1, x_2, \dots, x_p, x_{p+1}, \dots, x_n$ where $x_1, x_2, \dots, x_p$ are each less than G and $x_{p+1}, x_{p+2}, \dots, x_n$ are each greater than G, then*

$$G^n = (x_1\ x_2 \dots x_p\ x_{p+1} \dots x_n)$$

$$G^p \cdot G^{n-p} = (x_1\ x_2 \dots x_p)(x_{p+1} \dots x_n)$$

$$\frac{G}{x_1} \cdot \frac{G}{x_2} \dots \frac{G}{x_p} = \frac{x_{p+1}}{G} \dots \frac{x_n}{G}$$

$$\frac{x_1}{G} \cdot \frac{x_2}{G} \dots \frac{x_p}{G} = \frac{G}{x_{p+1}} \dots \frac{G}{x_n}.$$

Property 5. *If each value (or item) of a series is replaced by its geometric mean then the product of the values of series remains unaltered.*

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the geometric mean of 4, 8, 16, 32, 64.***

Solution. Here G. M. $= (4 \times 8 \times 16 \times 32 \times 64)^{1/5}$

$$= (16 \times 16 \times 16 \times 16 \times 16)^{1/5}$$

$$= [(16)^5]^{1/5} = 16.$$ Ans.

Example 2. ***Compute the geometric mean of the following distribution :***

Marks	**0–10**	**10–20**	**20–30**	**30–40**
No. of students	**5**	**8**	**3**	**4**

Solution. Here

Class	Mid-Value x	Frequency f	Log $\log_{10} x$	Product f log x
0–10	5	5	0·6990	3·4950
10–20	15	8	1·1761	9·4088
20–30	25	3	1·3979	4·1937
30–40	35	4	1·5441	6·1764
		$N = \Sigma f = 20$		$\Sigma f \log x = 23 \cdot 2739$

$$\therefore \quad \log G = \frac{1}{N} \Sigma f \log x = \frac{23 \cdot 2739}{20} = 1 \cdot 1637,$$

$$\therefore \quad G = \text{anti-log } (1 \cdot 1637) = 12 \cdot 58 \text{ marks.} \qquad \text{Ans.}$$

Example 3. ***Find G. M. from the following table :***

Marks	11	12	13	14	15
No. of students	3	7	8	5	2

Solution. Here

Marks x	No. of students f	log x	f log x
11	3	1·0414	3·1242
12	7	1·0792	7·5544
13	8	1·1139	8·9112
14	5	1·1461	5·7305
15	2	1·1761	2·3522
	$N = \Sigma f = 25$		$\Sigma f \log x = 27 \cdot 6725$

$\therefore$ If G be the geometric mean, then

$$\log G = \frac{1}{N} \Sigma f \log x = \frac{1}{25} \times 27 \cdot 6725 = 1 \cdot 1069$$

$$\therefore \quad G = \text{anti-log } 1 \cdot 1069 = 12 \cdot 79 \text{ marks.} \qquad \text{Ans.}$$

Example 4. ***If g and G are the geometric means of two series having n and N terms respectively, then find the geometric mean of the combined series.***

Solution. Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_N$ be the two given series. Then

$$g = (x_1\, x_2 \ldots x_n)^{1/n}, \quad G = (y_1\, y_2 \ldots y_N)^{1/N}.$$

If G' is the G. M. of the combined series, then

$$G' = (x_1, x_2 \ldots x_n \,.\, y_1, y_2 \ldots y_N)^{1/(n+N)}$$

$$G' = (g^n \,.\, G^N)^{1/(n+N)} = G^{n/(n+N)} \,.\, G^{N/(n+N)}$$

or in other form :

$$\log G' = [n \log g + N \log G] / (n + N).$$

Example. 5. ***In a frequency table, the ratio of upper boundary to the lower boundary has a constant value B. Show that the geometric mean G can be represented by the following formula :***

$$\log G = \log A + \frac{\log B}{N} \sum_{i=1}^{n} f_i (i - 1)$$

where A is the mid value of first class interval.

Solution. Let

Class	Mid Value	Frequency
$x_1 - x_2$	$\frac{x_1 + x_2}{2} = X_1$	f_1
$x_2 - x_3$	$\frac{x_2 + x_3}{2} = X_2$	f_2
$x_3 - x_4$	$\frac{x_3 + x_4}{2} = X_3$	f_3
.....	.....	.....
.....	.....	.....
$x_i - x_{i+1}$	$\frac{x_i + x_{i+1}}{2} = X_i$	f_i
.....	.....	.....
.....	.....	.....
		Total = N

Given : $\frac{x_2}{x_1} = \frac{x_3}{x_2} = \frac{x_4}{x_3}, \ldots = \frac{x_{i+1}}{x_i} = \ldots = B$ ...(1)

$$\frac{x_1 + x_2}{2} = A \qquad \ldots(2)$$

From (1),

$$x_2 = Bx_1, \; x_3 = Bx_2 = B^2 x_1, x_4 = B^3 x_1, \ldots$$

$$x_i = B^{i-1} x_1, x_{i+1} = B^i x_1, \ldots.$$

From (2),

$$x_1 = \frac{2A}{1+B}, \; x_2 = \frac{2AB}{1+B}, \; x_3 = \frac{2AB^2}{1+B}, \ldots, x_i = \frac{2AB^{i-1}}{1+B}, \ldots$$

Now $\log G = \frac{1}{N}\sum_{i=1}^{n} f_i \log X_i$

$= \frac{1}{N}\sum_{i=1}^{n} f_i \log \frac{x_i + x_{i+1}}{2}$ $\left[\because X_i = \frac{x_i + x_{i+1}}{2}\right]$

$= \frac{1}{N}\sum_{i=1}^{n} f_i \log \left[\frac{AB^i - 1}{(1+B)}(1+B)\right]$

$= \frac{1}{N}\Sigma f_i \log [(AB^{i-1})]$

$= \frac{1}{N}\sum_{i=1}^{n} f_i [\log A + (i-1)\log B]$

$\therefore \quad \log G = \log A + \frac{\log B}{N}\sum_{i=1}^{n} f_i (i-1).$

§ 4.15. HARMONIC MEAN

The **Harmonic mean** of a series of values is the reciprocal of the arithmetic mean of their reciprocals. Thus if $x_1, x_2, \ldots, x_n$ (none of them being zero) is a series and H is its harmonic mean then

$$\frac{1}{H} = \frac{1}{n}\left[\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right]$$

or $$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\Sigma\frac{1}{x}}.$$

If $f_1, f_2, \ldots, f_n$ be the frequencies of $x_1, x_2, \ldots, x_n$ (none of them being zero) then harmonic mean H is given by

$$\frac{1}{H} = \frac{1}{N}\sum_{i}^{n}\left(\frac{f}{x}\right) \text{ where } N = \Sigma f_i.$$

Mathematical Properties of Harmonic Mean :

Property 1. *For two observations x_1 and x_2, we have*

$$AH = G^2$$

where A = arithmetic mean, H = harmonic mean and G = geometric mean.

Proof. By definition

$$A = \frac{x_1 + x_2}{2}, \quad H = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \frac{2x_1 x_2}{x_1 + x_2}, \quad G = \sqrt{x_1 x_2}$$

$$\therefore \quad A \times H = \frac{x_1 + x_2}{2} \cdot \frac{2x_1 x_2}{x_1 + x_2} = x_1 x_2 = G^2.$$

Property 2. *For n positive observations $A \geq G \geq H$.*

The sign of equality will hold if and only if the values of all observations are same.

Proof. Let $x_1, x_2, \ldots, x_n$ be n positive observations.

$\therefore \quad A = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad G = (x_1 x_2 \ldots x_n)^{1/n}$

and $$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}.$$

To prove $A \geq G$

We have $(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0$

$\Rightarrow \quad x_1 + x_2 - 2\sqrt{x_1 x_2} \geq 0$

$\Rightarrow \quad \frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$ ...(1)

$\Rightarrow \quad \frac{\frac{x_1 + x_2}{2} + \frac{x_3 + x_4}{2}}{2} \geq \sqrt{\left\{\left(\frac{x_1 + x_2}{2}\right)\left(\frac{x_3 + x_4}{2}\right)\right\}} \geq \sqrt{\{(x_1 x_2)\sqrt{(x_3 x_4)}\}}$

$\Rightarrow \quad \frac{x_1 + x_2 + x_3 + x_4}{4} \geq (x_1 x_2 x_3 x_4)^{1/4}.$ ...(2)

Similarly,

$$\frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8}{8} \geq (x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8)^{1/8}. \quad \ldots(3)$$

Relations (1), (2) and (3) shows clearly that if $A \geq G$ is true for $n = 2$ then it is also true for $n = 2^2, 2^3, \ldots, 2^m$, where m is a positive integer.

Now to prove $A \geq G$ for other values of n, we proceed as follows :

Suppose $x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_2^m$ are n observations, where $k < 2^m, n = 2^m$.

Let $A = \frac{x_1 + x_2 + \ldots + x_k}{k}, \quad G = (x_1 x_2 \ldots x_k)^{1/k}$

and $x_{k+1} = x_{k+2} = \ldots = x_2^m = A$

Now, from above discussions, we have

$$\frac{x_1 + x_2 + \ldots + x_k + x_{k+1} + \ldots + x_2^m}{2^m} \geq (x_1 x_2 \ldots x_k\ x_{k+1} \ldots x_2^m)^{1/2^m}$$

$\Rightarrow \quad \frac{kA + (2^m - k)A}{2^m} \geq (G^k A^{2^m - k})^{1/2^m}$

$\Rightarrow \quad \frac{kA + 2^m A - kA}{2^m} \geq \left(\frac{G^k A^{2^m}}{A^k}\right)^{1/2^m}$

$\Rightarrow \quad A^{2^m} = \frac{G^k A^{2^m}}{A^k} \Rightarrow A^k G^k$

$\Rightarrow \quad A \geq G.$ ...(4)

To prove $G \geq H$

i.e., if $$(x_1 x_2 \ldots x_n)^{1/n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

i.e., if $\left(\frac{1}{y_1\, y_2 \ldots y_n}\right)^{1/n} \geq \frac{n}{y_1 + y_2 + \ldots + y_n}$, where $x_i = \frac{1}{y_i}$

or $\frac{y_1 + y_2 + \ldots + y_n}{n} \geq (y_1\, y_2 \ldots y_n)^{1/n}$

or $A_y \geq G_y$ where A_y = A.M. of y,

G_y = G.M. of y which is true by virtue of equation (4).

Hence $G \geq H$. ...(5)

∴ From (4) and (5), we get

$A \geq G \geq H$.

ILLUSTRATIVE EXAMPLES

Example 1. ***Find the harmonic mean of* 4, 8, 16, 32.**

Solution. If H is the required harmonic mean, then

$$H = \frac{4}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}}$$

$$= \frac{4}{\frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}} = \frac{400000}{46875} = 8 \cdot 533.$$ Ans.

Example 2. ***Find the harmonic mean of the marks obtained in a class test, given below :***

Marks :	11	12	13	14	15
No. of students :	3	7	8	5	2

Solution.

Marks x	Frequency f	$\frac{1}{x}$	$f \times \frac{1}{x}$
11	3	0·0909	0·2727
12	7	0·0833	0·5831
13	8	0·0769	0·6152
14	5	0·0714	0·3570
15	2	0·0667	0·1334
	$N = \Sigma f = 25$		$\Sigma \frac{f}{x} = 1 \cdot 9614$

∴ Required harmonic mean is given by

$$\text{H. M.} = \frac{\Sigma f}{\Sigma f \times \frac{1}{x}} = \frac{15}{1 \cdot 9614} = \frac{250000}{19614} = 12 \cdot 746 \text{ marks.}$$ Ans.

Example 3. ***Find the harmonic mean of* :**

Class (Marks)	0–10	10–20	20–30	30–40	40–50
No. of students	4	5	11	6	4

Solution.

Class Marks	Mid-value x	Frequency	$\frac{1}{x}$	$f \times \frac{1}{x}$
0–10	5	4	0·20000	0·80000
10–20	15	5	0·06667	0·33335
20–30	25	11	0·04000	0·44000
30–40	35	6	0·02857	0·17142
40–50	45	4	0·02222	0·08888
		$N = 30$		$\Sigma f \times \frac{1}{x} = 1 \cdot 83365$

Required H. M. is given by

$$\text{H. M.} = \frac{N}{\Sigma f \times \frac{1}{x}} = \frac{30}{1 \cdot 83365} = 16 \cdot 36 \text{ marks.}$$ Ans.

Example 4. ***Calculate the average speed of a car running at the rate of 20 kilometers per hour during the first 30 kilometers; at 25 kilometers per hour during the second 30 kilometers and at 30 kilometers per hour during the third 30 kilometers.***

Solution. In the questions related to the average speed, harmonic mean is most suitable. Therefore, here we shall compute the harmonic mean of 20, 25, 30.

$$\therefore \quad \frac{1}{H} = \frac{\frac{1}{20} + \frac{1}{25} + \frac{1}{30}}{3}$$

$$= \frac{0 \cdot 05000 + 0 \cdot 04000 + 0 \cdot 03333}{3}$$

$$= \frac{0 \cdot 12333}{3} = 0 \cdot 04111.$$

$$\therefore \quad H = \frac{1}{0 \cdot 04111} = 24 \cdot 32 \text{ kilometers per hour.}$$ Ans.

Example 5. ***Find the average rate of increase in population which in the first decade has increased 15 percent, in the next 22 percent and in the third 44 percent.***

Solution. In such problems geometric mean is most suitable average. Therefore, here we shall find the geometric mean G of 15, 22, and 44

$$\log G = \frac{\log 15 + \log 22 + \log 44}{3}$$

$$= \frac{1 \cdot 1761 + 1 \cdot 3424 + 1 \cdot 6434}{3}$$

$$= \frac{4 \cdot 1619}{3} = 1 \cdot 3873$$

G = Antilog $1 \cdot 3873 = 24 \cdot 40\%$.

Example 6. ***A cyclist goes to college from his house at the rate of 10 miles/hour and returns back at the rate of 15 miles/hour. Find the average speed of cyclist.***

Solution. Average speed $= \dfrac{\text{Total distance}}{\text{Total time}}$ miles/hour.

Let the distance of college from the house $= d$ miles.

$\therefore$ Time taken to reach the college $= d / 10$ hours

and time taken to reach the house back $= d / 15$ hours.

$\therefore$ Average speed $= \dfrac{d + d}{\frac{d}{10} + \frac{d}{15}} = \dfrac{2}{\frac{1}{10} + \frac{1}{15}} = 12$

$\therefore$ Average speed = 12 miles/hours.

Clearly the average speed is the harmonic mean of 10 and 15.

Example 7. ***A cyclist goes to college from his house with a speed 8 km/h return back with the speed 10 km/h. Find the average speed of the cyclist.***

Solution. Average speed = harmonic mean of 8 and 10

$$= \frac{2}{\frac{1}{8} + \frac{1}{10}} = \frac{160}{18} = 8 \cdot 89 \text{ km/h.}$$

Example 8. ***Explain with reasons, that the following statement is 'true' or 'false':***

A car runs between two stations. The car goes with uniform speed of 40 km/h. and returns back with uniform speed of 60 km/h. His average speed is 50 km/h.

Solution. We have

average speed = harmonic mean of 40 and 60

$$= \frac{2}{\frac{1}{40} + \frac{1}{60}} = 48 \text{ km/h.}$$

Hence the given statement is 'false'.

Example 9. ***If x_1 and x_2 are two observations, then show that A.M. > G.M. > H.M.***

Solution. See § 4.5, property 2. Here take only x_1 and x_2.

EXERCISE 2 (D)

1. Find the geometric mean of 2, 6, 18, 54, 162.
2. Find the geometric mean from the following table :

Variate	:	5	7	9	11	13	15
Frequency	:	1	2	3	5	11	9

3. Below are given the marks obtained by the students in a class test, find the geometric mean :

Marks	:	6	7	8	10	12	15
Frequency	:	20	15	12	8	4	2

4. Find the harmonic mean for the following frequency distribution :

Class	:	0–10	10–20	20–30	30–40
Frequency	:	5	8	3	4

5. Find the harmonic mean for the following frequency distribution :

Class	:	40–50	50–60	60–70	70–80	80–90	90–100
Frequency	:	19	15	36	72	51	43

6. Find the harmonic mean for the following frequency distribution :

Wages (in Rs.)	:	40–50	50–60	60–70	70–80	80–90	90–100
No. of workers	:	12	10	15	17	8	3

ANSWERS

1. 18
2. 11·86
3. 7·603
4. 11·312
5. 71·4
6. Rs. 61·9.

MISCELLANEOUS EXAMPLES

Example 1. ***'In finding the arithmetic mean of a set of readings on a thermometer, it does not matter whether we measure the temperature in centigrade or Fahrenheit degrees, but that in finding the geometric mean it does matter'. Comment.***

Solution. Let $C_1, C_2, \ldots, C_N$ be a set of N thermometric readings in Centigrade degrees and the corresponding reading in Fahrenheit be $F_1, F_2, \ldots, F_N$.

We have $F_r = 32 + \dfrac{9}{5} C_r, \quad r = 1, 2, 3, \ldots, N.$

Now A. M. of N reading in centigrade

$$\overline{M}_C = \frac{C_1 + C_2 + \ldots + C_N}{N}$$

A. M. in Fahrenheit degrees $= \dfrac{1}{N} \sum_{1}^{N} F_r$

or $$\overline{M}_F = \frac{1}{N} \sum_{r=1}^{N} \left(32 + \frac{9}{5} C_r\right) = \frac{1}{N} \left(32 N + \frac{9}{5} \sum_{r=1}^{N} C_r\right)$$

$$= 32 + \frac{9}{5} \cdot \frac{(C_1 + C_2 + \ldots + C_N)}{N} = 32 + \frac{9}{5} \overline{M}_C.$$

Thus in finding the A. M. it does not matter that the readings are taken in centigrade or Fahrenheit.

Again

Geometric mean in Centigrade

$$= G_C = G.M. \text{ of } \dots C_1, C_2, \dots C_N$$
$$= (C_1\, C_2 \dots C_N)^{1/N}$$

and geometric mean in Fahrenheit

$$= G_F = (F_1\, F_2 \dots F_n)^{1/N}$$
$$= \left[\left(32 + \frac{9}{5} C_1\right)\left(32 + \frac{9}{5} C_2\right) \dots \left(32 + \frac{9}{5} C_n\right)\right]^{1/N}.$$

The expression on R. H. S. cannot be expressed as a single function $C_1, C_2, \dots, C_n$ of G_C. Thus if geometric mean in Fahrenheit is known it cannot be computed in centigrade till we have readings in centigrades.

Example 2. ***The monthly incomes of 10 families are given by the following table :***

A	B	C	D	E	F	G	H	I	J	Total
85	70	15	75	500	20	45	250	40	36	1136

Compute the arithmetic mean, geometric mean and harmonic mean of the above incomes. Which of the above three means represents the data most suitable ? Give reasons.

Solution. Here arithmetic mean,

$$M = \frac{\Sigma x_i}{n} = \frac{1136}{10} = 113 \cdot 6 \text{ Rs.}$$

Geometric mean,

$$G = (85 \times 70 \times 15 \times 75 \times 500 \times 20 \times 45 \times 250 \times 40 \times 36)^{1/10}$$

$$\Rightarrow \quad \log G = \frac{1}{10} [\log 85 + \log 70 + \log 15 + \log 75 + \log 500 + \log 20 + \log 45 + \log 250 + \log 40 + \log 36]$$

$$= \frac{1}{10} [1 \cdot 9294 + 1 \cdot 8451 + 1 \cdot 1761 + 1 \cdot 8751 + 2 \cdot 6980 + 1 \cdot 3010 + 1 \cdot 6532 + 2 \cdot 3979 + 1 \cdot 6021 + 1 \cdot 5563]$$

$$= \frac{1}{10} \times 18 \cdot 0342 = 1 \cdot 8034 = \log 63 \cdot 59$$

$$\Rightarrow \quad G = 63 \cdot 59 \text{ Rs.}$$

And harmonic mean

$$H = \frac{10}{\frac{1}{85} + \frac{1}{70} + \frac{1}{15} + \frac{1}{75} + \frac{1}{500} + \frac{1}{20} + \frac{1}{45} + \frac{1}{250} + \frac{1}{40} + \frac{1}{36}}$$

$$= \frac{10}{[0 \cdot 01176 + 0 \cdot 0143 + 0 \cdot 0667 + 0 \cdot 0133 + 0 \cdot 002 + 0 \cdot 05 + 0 \cdot 0222 + 0 \cdot 004 + 0 \cdot 025 + 0 \cdot 0278]}$$

$$= \frac{10}{0 \cdot 23706} = 42 \cdot 18 \text{ (approx.)}$$

Clearly the geometric mean represents the data most suitable since out of 10 families, 8 families have their income less than the arithmetic mean while the harmonic mean represents the lower income group only.

Example 3. *The expenditure of 100 families are given below :*

Expenses in Rs. :	**0–10**	**10–20**	**20–30**	**30–40**	**40–50**
No. of families :	**14**	**?**	**27**	**?**	**15**

The mode for the distribution is 24. Find the missing frequencies.

Solution. Let F_1 and F_2 be the unknown frequencies of classes 10–20 and 30–40 respectively.

Exp. in Rs.	No. of families (f)	Cumulative frequency
0–10	14	14
10–20	F_1	$14 + F_1$
20–30	27	$41 + F_1$
30–40	F_2	$41 + F_1 + F_2$
40–50	15	$56 + F_1 + F_2$
	$N = \Sigma f = 56 + F_1 + F_2$	

$$\therefore \quad N = 100 \Rightarrow 56 + F_1 + F_2 = 100$$
$$\Rightarrow F_1 + F_2 = 44. \qquad \dots(1)$$

Here mode is 24 which clearly lies in the class 20–30 and so 20–30 is the modal class. Therefore from the formula, the mode M_0 is given by

$$M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

where $l = 20$, $f = 27$, $f_{-1} = F_1$, $f_1 = F_2$ and $i = 10$.

$$\therefore \quad 24 = 20 + \frac{27 - F_1}{54 - F_1 - F_2} \times 10$$

$$\Rightarrow \quad 3F_1 - 2F_2 = 27. \qquad \dots(2)$$

Solving (1) and (2), we have

$$F_1 = 23, \; F_2 = 21.$$

Example 4. ***Expenditure of some families are given below :***

Expenditure (in Rs.) :	**0–10**	**10–20**	**20–30**	**30–40**	**40–50**
Number of families :	**14**	**?**	**27**	**?**	**15**

median and mode of the distribution are 25 and 24 respectively. Calculate unknown frequencies.

Also determine the arithmetic average of the data.

Solution. Let the missing frequencies of the classes 10–20 and 30–40 be a and b respectively. Then

Class Expenditure (in Rs.)	Frequency f (No. of families)	Cumulative frequency
0–10	14	14
10–20	a	$14 + a$
20–30	27	$41 + a$
30–40	b	$41 + a + b$
40–50	15	$56 + a + b$
Total	$N = 56 + a + b$	

Total frequency $N = 56 + a + b$.

Now median $(M_d) = 25$ (given). So median class is 20–30. We have

$$M_d = l + \frac{\frac{1}{2}N - F}{f} \times i$$

i.e., $$M_d - l = \frac{N - 2F}{2f} \times i.$$

$$\therefore \quad 25 - 20 = \frac{56 + a + b - 2(14 + a)}{2 \times 27} \times 10$$

or $$5 \times 54 = 10\,[28 - a + b]$$

or $$27 = 28 - a + b$$

or $$a - b = 1. \qquad \ldots(1)$$

Again Mode = 24 (given). So Modal class is 20–30. We have

$$M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i.$$

$$\therefore \quad 24 = 20 + \frac{27 - a}{2 \times 27 - a - b} \times 10$$

or $$(24 - 20)(54 - a - b) = 10(27 - a)$$

or $$4(54 - a - b) = 10(27 - a)$$

or $$3a - 2b = 27. \qquad \ldots(2)$$

Solving (1) and (2), we get

$$a = 25,\ b = 24.$$

$\therefore$ Unknown frequencies are 25 and 24 respectively.

II part. To compute Arithmetic mean.

Class	Mid value x	Frequency f	fx
0–10	5	14	70
10–20	15	25	375
20–30	25	27	675
30–40	35	24	840
40–50	45	15	675
Total		$\Sigma f = 105$	$\Sigma fx = 2635$

$\therefore$ Arithmetic mean $M = \frac{\Sigma fx}{\Sigma f} = \frac{2635}{105} = 25 \cdot 1.$

Note. Since median and mode are given, so mean can also be determined by using the formula

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

$\Rightarrow$ $2\,\text{Mean} = 3\,\text{median} - \text{mode}$

$\Rightarrow$ $2\,\text{Mean} = 3 \times 25 - 24$

$\Rightarrow$ $\text{Mean} = 25 \cdot 5.$

Example 5. ***Find the quartiles from the following data :***

Class :	0–9	10–19	20–29	30–39	40–49	50–59	60–69
Frequency :	2	8	17	35	20	12	5

Solution. Changing the given exclusive series into inclusive series, we have

Class	Frequency f	Cumulative frequency
05–95	3	3
95–195	8	11
195–295	17	28
295–395	35	63
395–495	20	83
495–595	12	95
595–695	5	100
Total	$N = 100$	

Hence $N = 100,\ N/4 = 25,\ 2N/4 = 50,\ 3N/4 = 75.$

Clearly $N/4$th *i.e.*, 25th terms lies in the class 195–295. Thus first quartile class is 195–295. Hence first quartile Q_1 is given by

$$Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i \qquad \text{[formula]}$$

$\therefore$ Here $l = 19 \cdot 5, \frac{1}{4}N = 25, F = 11, f = 17, i = 29 \cdot 5 - 19 \cdot 5 = 10.$

$$\therefore \quad Q_1 = 19 \cdot 5 + \frac{25 - 11}{17} \times 10$$

$$= 19 \cdot 5 + \frac{14 \times 10}{17} = 19 \cdot 5 + 8 \cdot 24 = 27 \cdot 74.$$

Now $2N/4$ *i.e.*, 50th term, lies in the class 295–395.

$\therefore$ From $$Q_2 = l + \frac{\frac{2}{4}N - F}{f} \times i.$$

For the 2nd quartile. We have

$l = 29 \cdot 5,\ 2N/4 = 50.$

$\therefore$ $F = 28,\ f = 35,\ i = 39 \cdot 5 - 29 \cdot 5 = 10$

Also $Q_2 = 29 \cdot 5 + \dfrac{50 - 28}{35} \times 10$

$\therefore$ $= 29 \cdot 5 + \dfrac{22 \times 10}{35} = 29 \cdot 5 + 6 \cdot 28 = 35 \cdot 78.$

Again $3N/4$th *i.e.*, 75th term lies in 395–495.

$\therefore$ From $Q_3 = l + \dfrac{\frac{3}{4}N - F}{f} \times i$

For 3rd quartile, we have

$\therefore$ $l = 39 \cdot 5,\ 3N/4 = 75,\ F = 63,\ f = 20,\ i = 49 \cdot 5 - 39 \cdot 5 = 10.$

$\therefore$ $Q_3 = 39 \cdot 5 + \dfrac{75 - 63}{20} \times 10$

$= 39 \cdot 5 + \dfrac{12 \times 10}{20} = 39 \cdot 5 + 6 = 45 \cdot 5.$

Example 6. ***From the following frequency distribution, compute 4th decile, 7th decile, 40th. percentile and 94th percentile***

Class :	0–4	4–8	8–12	12–14	14–18	18–20	20–25	≥ 25
Frequency :	10	12	18	7	5	8	4	6

Solution. The frequency table :

Class	Frequency f	Cumulative frequency
0–4	10	10
4–8	12	22
8–12	18	40
12–14	7	47
14–18	5	52
18–20	8	60
20–25	4	64
Above 25	6	70
Total	$N = 70$	

Here $N = 70,\ \dfrac{4N}{10} = 28,\ \dfrac{7N}{10} = 49,\ \dfrac{40N}{100} = \dfrac{4N}{10} = 28,\ \dfrac{94N}{100} = 65 \cdot 8.$

$\therefore$ 4th decile = 40th. percentile.

Since $4N/10 = 28$ which lies in the class 8–12. Hence in the formula

$$D_4 = l + \frac{(4N/10) - F}{f} \times i,$$

Putting $l = 8,\ 4N/10 = 28,\ F = 22,\ f = 18,\ i = 12 - 8 = 4$, we have

$$D_4 = 8 + \frac{28 - 22}{18} \times 4 = 8 + \frac{6 \times 4}{18} = 8 + 1 \cdot 33 = 9 \cdot 33$$

$P_{40} = D_4 = 9 \cdot 33$

Here $7N/10 = 49$ which lies in the class 14–18. Hence in the formula

$$D_7 = l + \frac{\frac{7N}{10} - F}{f} \times i.$$

Putting $l = 14,\ \dfrac{7N}{10} = 49,\ F = 47,\ f = 5,\ i = 18 - 14 = 4$, we have

$$D_7 = 14 + \frac{49 - 47}{5} \times 4 = 14 + \frac{2 \times 4}{5} = 14 + 1 \cdot 6 = 15 \cdot 4.$$

Here 94th percentile lies in the class 25–30 (the last class-interval, supposing the upper limit to be 30). Hence in the formula

$$P_{94} = l + \frac{\frac{94N}{100} - F}{f} \times i.$$

Putting $l = 25,\ \dfrac{94N}{100} = 65 \cdot 8,\ F = 64,\ f = 6,\ i = 30 - 25 = 5,$ we have

$$P_{94} = 25 + \frac{65 \cdot 8 - 64}{6} \times 5 = 25 + \frac{1 \cdot 8 \times 5}{6} = 25 + 1 \cdot 5 = 26 \cdot 5.$$

Example 7. ***The heights of 70 students of a class are given by the following table. Find their mode.***

Height (in cm.)	120–124	125–129	130–134	135–139	140–144	145–149	150–154
No. of Students	2	5	8	15	20	10	5

Solution. Converting the exclusive series into inclusive series, we have

Class	Frequency
1195–1245	2
1245–1295	5
1295–1345	8
1345–1395	15
1395–1445	25
1445–1495	10
1495–1545	5

By inspection we see that the maximum frequency is 25 which lies in the class 1395–1445. Thus it is the modal class. Hence in the formula of mode.

$$M_0 = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i.$$

Median $= \frac{1}{2}N$th term $= 38$th term.

$$\therefore \quad M_d = l + \frac{\frac{1}{2}N - F}{f} \times i = 20 \cdot 5 + \frac{38 - 35}{11} \times 4$$

$$= 20 \cdot 5 + \frac{3 \times 4}{11} = 20 \cdot 5 + 1 \cdot 09 = 21 \cdot 59.$$

$$\therefore \text{ Mode } (M_0) = 3M_d - 2M$$

$$= 3(21 \cdot 59) - 2(23 \cdot 69) = 64 \cdot 77 - 47 \cdot 38 = 17 \cdot 39.$$ Ans.

Example 10. ***A variate takes values*** $1, r, r^2, \ldots, r^{n-1}$ ***each with frequency unity. Show that***

$$A = \frac{(1 - r^n)}{n(1 - r)}, \quad G = r^{(n-1)/2} \quad \textit{and} \quad H = \frac{n(1 - r)r^{n-1}}{1 - r^n}.$$

Verify $(i)\ AH = G^2$, $(ii)\ A > G > H$.

Solution. Here

$$\Sigma x = 1 + r + r^2 + \ldots + r^{n-1} = \frac{1 - r^n}{1 - r}.$$

$$\therefore \quad A = \frac{\Sigma x}{n} = \frac{(1 - r^n)}{n(1 - r)} \qquad \ldots(1)$$

$$G = [1 \cdot r \cdot r^2 \ldots r^{n-1}]^{1/n}$$

$$= [r^{1 + 2 + \ldots + (n-1)}]^{1/n}$$

$$= \left[r^{\frac{1}{2}(n-1)n}\right]^{1/n} = r^{(n-1)/2} \qquad \ldots(2)$$

$$H = \frac{n}{\frac{1}{1} + \frac{1}{r} + \frac{1}{r^2} + \ldots + \frac{1}{r^{n-1}}}$$

$$= \frac{nr^{n-1}}{r^{n-1} + r^{n-2} + \ldots + r + 1}$$

$$= \frac{n r^{n-1}}{(1 - r^n)/(1 - r)} = \frac{n(1 - r)r^{n-1}}{(1 - r^n)} \qquad \ldots(3)$$

$$\therefore \quad AH = \frac{(1 - r^n)}{n(1 - r)} \cdot \frac{n(1 - r)r^{n-1}}{(1 - r^n)}$$

$$= r^{n-1} = [r^{(n-1)/2}]^2 = G^2. \qquad \ldots(4)$$

Now $A > G$.

If $\dfrac{(1 - r^n)}{n(1 - r)} > r^{(n-1)/2}$

i.e., if $1 + r + r^2 + \ldots + r^{n-1} > nr^{(n-1)/2}$

i.e., if $(1 - 2r^{(n-1)/2} + r^{n-1}) + (r - 2r^{(n-1)/2} + r^{n-2}) + \ldots > 0$

i.e., if $(1 - r^{(n-1)/2})^2 + (r^{1/2} - r^{(n-2)/2})^2 + \ldots > 0$

which is true, $A > G.$ $\qquad \ldots(5)$

Again $G > H.$

If $r^{(n-1)/2} > \dfrac{n(1 - r)r^{n-1}}{(1 - r^n)}$

i.e., if $\dfrac{1 - r^n}{1 - r} > r^{(n-1)/2}$

which is true by (5). Hence $G > H.$ $\qquad \ldots(6)$

$\therefore$ From (5) and (6),

$$A > G > H.$$

Example 11. ***Show that for positive observations, the arithmetic mean (A. M.), the geometric mean (G. M.) and the harmonic mean (H. M.) are particular cases of pth. root of arithmetic mean of pth. powers of the observations.***

Solution. Let the variate x has the following n positive observations $x : x_1\ x_2\ x_3 \ldots x_n$.

and $$W(p) = \left(\frac{x_1^p + x_2^p + \ldots + x_n^p)}{n}\right)^{1/p} \qquad \ldots(1)$$

Case I. Let $p = 1$, then from (1), we have

$$W(p) = \left(\frac{x_1 + x_2 + \ldots + x_n)}{n}\right) = \text{A.M.}$$

Case II. Let $p = -1$, then (1) becomes

$$W(p) = \left(\frac{x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1})}{n}\right)^{-1}$$

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

Case III. Let $P \to 0$, then the limiting value of $W(p)$ is evaluated as follows :

Taking log of both sides of (1), we get

$$\log W(p) = \frac{1}{p} \log \left(\frac{x_1^p + x_2^p + \ldots + x_n^p}{n}\right)$$

$$\therefore \lim_{p \to 0} \log W(p) = \lim_{p \to 0} \frac{\log\left(\frac{x_1^p + x_2^p + \ldots + x_n^p}{n}\right)}{p} \qquad \left[\frac{0}{0} \text{ form}\right]$$

$$= \lim_{p \to 0} \frac{\frac{n}{x_1^p + x_2^p + \ldots + x_n^p} \cdot \left\{\frac{x_1^p \log x_1 + \ldots + x_n^p \cdot \log x_n}{n}\right\}}{1}$$

[By L. Hospital's Rule]

$$= \frac{n}{(1+1+\ldots+1)} \cdot \left\{ \frac{\log x_1 + \ldots + \log x_n}{n} \right\}$$

$$= \frac{\log x_1 + \log x_2 + \ldots + \log x_n}{n} = \log (G.M.)$$

$$\therefore \lim_{p \to 0} W(p) = \text{G.M.}$$

Example 12. ***If the variates x_i's, whose weights are w_i's weighted mean $\bar{x}_w$ then prove the following relation :***

$$\left(\sum_{i=1}^{n} w_i \right) \left(\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \right) = \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_i w_j (x_i - x_j)^2$$

where $\sum_{i=1}^{n} w_i \neq 0$.

Solution. R.H.S. $= \frac{1}{2} \sum_i \sum_j w_i w_j (x_i - x_j)^2$

$$= \frac{1}{2} \sum_i \sum_j w_i w_j (x_i^2 - 2x_i x_j + x_j^2)$$

$$= \frac{1}{2} \left[\left(\sum_i w_i x_i^2 \right) \left(\sum_j w_j \right) - 2 \left(\sum_i w_i x_i \right) \left(\sum_j w_j x_j \right) + \left(\sum_j w_j x_j^2 \right) \left(\sum_i w_i \right) \right]$$

$$= \frac{1}{2} \left[2 \left(\sum w_i x_i^2 \right) \left(\sum_i w_i \right) - 2 \left(\sum w_i x_i \right)^2 \right] .$$

Since i and j both take values from 1 to n. Hence we may write

$$\left(\sum_i w_i x_i^2 \right) \left(\sum_j w_j \right) = \left(\sum_j w_j x_j^2 \right) \left(\sum_i w_i \right) = \sum \left(w_i x_i^2 \right) \left(\sum_i w_i \right)$$

and $$\left(\sum_i w_i x_i \right) \left(\sum_j w_j x_j \right) = \sum_i (w_i x_i)^2$$

$$\text{R.H.S.} = \left(\sum_i w_i \right) \left[\left(\sum_i w_i x_i^2 \right) - \sum_i w_i \left(\frac{\sum_i w_i x_i}{\sum_i w_i} \right)^2 \right]$$

$$= \left(\sum_{i=1}^{n} w_i \right) \left[\sum_{i=1}^{n} w_i x_i^2 - \sum_{i=1}^{n} w_i \cdot \bar{x}_w^2 \right]$$

$$= \left(\sum_i w_i \right) \left[\sum_i w_i x_i^2 - \bar{x}_w^2 \sum_i w_i \right]$$

$$= \left(\sum_i w_i \right) \left[\left(\sum_i w_i \right) \left\{ \frac{\sum_i w_i x_i^2}{\sum_i w_i} - \bar{x}_w^2 \right\} \right]$$

$$= \left(\sum_i w_i \right) \left[\left(\sum_i w_i \right) \left\{ \frac{\sum_i w_i x_i^2}{\sum_i w_i} - 2\bar{x}_w^2 + \bar{x}_w^2 \right\} \right]$$

$$= \left(\sum_i w_i \right) \left[(\Sigma w_i) \left\{ \frac{\sum_i w_i x_i^2}{\sum_i w_i} - 2\bar{x}_w \cdot \frac{\sum_i w_i x_i}{\sum_i w_i} + \bar{x}_w^2 \right\} \right]$$

$$\left\{ \because \bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i} \right\}$$

$$= \left(\sum_i w_i \right) \left[\sum_i w_i x_i^2 - 2\bar{x}_w \sum_i w_i x_i + \bar{x}_w^2 \sum_i w_i \right]$$

$$= \left(\sum_i w_i \right) \left[\sum_i w_i \left(x_i^2 - 2\bar{x}_w x_i + \bar{x}_w^2 \right) \right]$$

$$= \left(\sum_i w_i \right) \left[\sum_i w_i (x_i - \bar{x}_w)^2 \right]$$

$$= \left(\sum_{i=1}^{n} w_i \right) (\Sigma w_i (x_i - \bar{x}_w)^2) = \text{L.H.S.}$$

Example 13. ***Compute the mean, median and mode for the following data :***

Marks less than :	5	10	15	20	25	30	35	40
Frequency :	2	7	14	27	48	64	72	75

Solution. Clearly we are given the cumulative frequency distribution. So first of all we shall convert it to the simple one.

Let A = Assumed mean = 22·5, i = class interval = 5.

Marks	Mid Value x	f	c.f.	$\xi = x - 22 \cdot 5$	$u = \frac{\xi}{5}$	fu
0–5	2·5	2	2	−20	−4	−8
5–10	7·5	5	7	−15	−3	−15
10–15	12·5	7	14	−10	−2	−14
15–20	17·5	13	27	−5	−1	−13
20–25	22·5	21	48	0	0	0
25–30	27·5	16	64	5	1	16
30–35	32·5	8	72	10	2	16
35–40	37·5	3	75	15	3	9
Total		$\Sigma f = 75 = N$				$\Sigma fu = -9$

(i) Mean :

$$M = A + i\,\frac{\Sigma fu}{\Sigma f} = 22 \cdot 5 + 5 \cdot \frac{(-9)}{75}$$

$$= 22 \cdot 5 - 0 \cdot 6 = 21 \cdot 9 \text{ marks.}$$

(ii) Median : Here $N = 75$.

$\therefore$ Median = Measure of $\frac{1}{2}(N+1)$th term *i.e.*, 38th term.

Since 38th term lies in the class 20–25. Therefore the median class is 20–25.

$l = 20, F = 27, m = 38, f = 21, l_1 = 20, l_2 = 25.$

$$\therefore \text{Median } (M_d) = l + \frac{m - F}{f} \times (l_2 - l_1)$$

$$= 20 + \frac{38 - 27}{21} \times 5 = 20 + 2 \cdot 619 = 22 \cdot 6 \text{ (approx.)}$$

(iii) Mode : Here the maximum frequency 21 lies in the class 20–25 and so 20–25 is the modal class.

$\therefore \quad l = 20, f = 21, f_{-1} = 13, f_1 = 16, i = 5.$

$$\therefore \text{Mode } (M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 20 + \frac{21 - 13}{2 \times 21 - 13 - 16} \times 5 = 20 + \frac{8}{10} \times 5 = 23 \cdot 077 \text{ (approx.).}$$

◈ § 4.16. MERITS AND DEMERITS OF VARIOUS AVERAGES

Arithmetic Average or Mean :

Merits (or Advantages)

(i) It is well defined.
(ii) It is based on all observations.
(iii) It can be easily calculated from the given data.
(iv) Its general nature is readily comprehensive. In fact it gives weight to all items in direct proportion of their sizes.
(v) It is least affected by fluctuations of sampling. In most of the cases it is not affected by fluctuations of sampling.
(vi) It lends itself to algebraic treatment and due to this reason it is used in higher statistical analysis.
(vii) It is always definite *i.e.*, it is never indeterminate.
(viii) In the computation of arithmetic mean, the arraying of data is not required (as is required in median). Also the grouping of data is not required (as required in mode).
(ix) Its truthness can be checked by Charlier's check while it is not possible in other means.

Demerits (Disadvantages)

(i) It sometimes may not be represented in the actual data *e.g.*, average number of deaths in a town at 564 per day.
(ii) It gives undue weight to bigger items of a given series *e.g.*, a millionaire will drag up the average income of a locality where the majority consists of ill-paid persons.

(iii) It sometimes gives wrong statements *e.g.*, let the net profits of two trades A and B be as given below :

Year	Trade 'A'	Trade 'B'
1989	Rs. 4000	Rs. 8000
1990	Rs. 6000	Rs. 6000
1991	Rs. 8000	Rs. 4000

The arithmetic mean of both the trades in Rs. 6000. It concludes that the conditions of both the trades are same. But the trade 'A' is progressive and the trade 'B' is regressive.

(iv) It cannot be computed in the cases when one or more items of the data are not given while median and mode can be computed.
(v) It cannot be computed if the extreme ends are open *e.g.*, 'above 100' or 'below 5'.
(vi) It is not easy to locate by inspection as median and mode can be done.
(vii) It cannot be used with qualitative data *e.g.*, honesty, wisdom etc. which are incapable of numerical measurement.

Median :

Merits.

(i) Similar to mean, it is easily understood, easily calculated, well defined and based upon all observations.
(ii) It can be exactly located.
(iii) It can be located by inspection.
(iv) It can be computed in the cases where the end intervals are open.
(v) It does not give undue weight to abnormally large or small values of the variables.
(vi) The size of the median can never be affected greatly by the addition of a few more items.
(vii) It gives best results in the study of direct qualitative measurements *e.g.*, beauty, virtue, intelligence, wisdom etc.

Demerits.

(i) The median does not lend itself to algebraic treatment *i.e.*, the median of the sum or difference of corresponding observations of two or more series is not the sum or difference of the medians of the component series.
(ii) In order to find median, the data are to be arranged in ascending or descending order or operation which involves considerable time and work.
(iii) It may sometimes be indefinite in a discrete series or when the number of items in the median class is large.
(iv) The median may not be represented by the actual data.
(v) It may sometimes be located at the point where the frequency may be quite small, which shows the typical feature of the distribution may not be represented by the median.

Mode :

Merits.

(i) It is easily understood and easily calculated.
(ii) It can be very easily found from the graph.
(iii) It depends on all values of the variate.
(iv) It can be computed where the end values are open.
(v) It is easily located in some cases.

Demerits.

(i) It is not well defined. A clearly defined mode does not always exist.
(ii) Its value may not be well defined.
(iii) Its algebraic treatment is not easy except in continuous distributions.

Geometric Mean :

Merits.

(i) It is based on all the observations.
(ii) Geometric mean lends itself to algebraic treatment.
(iii) It gives weight to each item.
(iv) It is not suitable average in the cases where the ratios are given.

Demerits.

(i) Its calculation is comparatively difficult.
(ii) It vanishes even if a single value of the variate is zero.
(iii) It may become imaginary for negative values of the variate.
(iv) It cannot be computed when the intervals are open.

Harmonic Mean :

Merits.

(i) It is based on all the observations.
(ii) It lends itself to algebraic treatment.
(iii) It is most suitable mean in the cases like calculating the average speed of a train when the speed for different parts of the distance is given in distance per unit time (kilometers per hour, say).

Demerits.

(i) It is difficult to calculate and difficult to understand.
(ii) A very high weightage is given to small values of the variate.

EXERCISE 4 (E)

1. Find out the average height of a plant in a certain garden from the following data. What is the median height and how far does it differ from the mode ?

Heights (in cms.)	No. of plants	Heights (in cms.)	No. of plants
66	1	71	1
67	2	72	2
68	4	73	1
69	3	74	1
70	2	75	1

2. From the following distribution find the mean, median and mode.

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

3. (a) The marks obtained by 76 students in a certain examination are given below, find its median and mode.

Marks	0–10	10–20	20–30	30–40	40–50
Frequency	5	9	14	30	18

(b) Compute the mode from the following table :

Size	Frequency	Size	Frequency
0–9	1	40–49	12
10–19	2	50–59	8
20–29	6	60–69	5
30–39	6	70–79	3

4. From the table :

Weekly income (Rs.)	0–10	10–20	20–30	30–40	40–50
No. of persons	17	20	30	18	15

(a) Find the mean, median and mode.
(b) What is the percentage of the persons earning more than Rs. 25 per week.

5. Compute the median, lower and upper quartiles, 4th decile and 40th percentile from the following distribution :

Marks	No. of Students	Marks	No. of Students
0–4	10	14–18	5
4–8	12	18–20	8
8–12	18	20–25	4
12–14	7	25 and above	6

6. Find the arithmetic mean for the following data :

Age (less than years) :	1	2	3	4	5	6
No. of persons :	15	32	51	78	97	100

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

Wages (in Rs) :	0–10	10–20	20–30	30–40	40–50
No. of workers :	22	38	46	35	20

8. Find the mode of the following series :

Age (less than years) :	1	2	3	4	5	6
No. of persons :	15	32	51	78	97	100

Is the median most suitable mean ? Discuss.

9. The marks obtained by 49 students of a class are given by the following frequency table. Compute the three quartiles :

Marks	No. of Students	Marks	No. of Students
5–10	5	25–30	2
10–15	6	30–35	4
15–20	15	35–40	2
20–25	10	40–45	2

10. A variable takes values $a, ar, ar^2, \ldots, ar^{n-1}$ each with frequency unity. If A, G, H are arithmetic mean, geometric mean and harmonic mean respectively, show that

$$A = \frac{a(1-r^n)}{n(1-r)}, \quad G = ar^{(n-1)/2},$$

$$H = \frac{an(1-r)r^{n-1}}{1-r^n}, \quad AH = G^2.$$

Also show that $A \geq G \geq H$, equality sign holds when $n = 1$.

11. A distribution $x_1, x_2, \ldots, x_r, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_r, \ldots, f_n$ is transformed into the distribution $y_1, y_2, \ldots, y_r, \ldots, y_n$ with the same corresponding frequencies by the relation $y_r = ax_r + b$, where a and b are constants.
Show that the mean, median and mode of the new distribution are given in terms of the first distribution by the same transformation.

12. Discuss in short the different measures of central tendency and write their merits and demerits.

13. What do you mean by the measures of central tendency ? Write down the names of different measures. Give the method of computing any two and discuss their merits and demerits.

14. Discuss the merits and demerits of A. M., G. M. and H. M.

15. Discuss the merits and demerits of mean, median and mode.

16. How is A. M. affected if every value of the variable is (a) increased by a constant a and (b) multiplied by some constant b.

17. What do you mean by the measure of central tendency ? Explain merits and demerits of different measures of central tendency.

ANSWERS

1. $M_0 = 68$ cm., $M_d = 69$ cm., $M = 69\frac{8}{9}$ cm.
2. $M = 47{\cdot}95$, $M_d = 48{\cdot}35$, $M_0 = 48{\cdot}57$
3. (a) $M_d = 33{\cdot}3$, $M_0 = 34{\cdot}3$ (b) $M_0 = 45{\cdot}5$
5. $M_d = 10{\cdot}9$, $Q_1 = 6{\cdot}5$, $Q_3 = 18{\cdot}25$, $D_4 = 9{\cdot}3$, $P_{40} = 10$ **6.** $2{\cdot}77$
7. Rs. 2458, Rs. 247, Rs. 2445 **8.** 35
16. (a) Decreased by a, (b) Multiplied by b.

EXERCISE 4 (F)

Objective Type Questions

I. Match the following :

(a) A. M.	(i) $l + \frac{\frac{1}{2}N - F}{f} \times i$
(b) G. M.	(ii) $l + \frac{f_1}{f_{-1} + f_1} \times i$
(c) H. M.	(iii) $l + \frac{\frac{3}{4}N - F}{f} \times i$
(d) Median	(iv) $l + \frac{\frac{1}{4}N - F}{f} \times i$
(e) Mode	(v) $(x_1 \cdot x_2 \ldots x_n)^{1/n}$
(f) Lower Quartile	(vi) $\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$
(g) Upper Quartile	(vii) $\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$
	(viii) $l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$.

II. Select the correct answer from the four given alternative answers :

1. The arithmetic mean of the numbers 1, 2, ..., n is :
   (a) $\frac{n+1}{2}$ (b) $\frac{n(n+1)}{2}$
   (c) $\frac{n(n+1)^2}{4}$ (d) $\frac{n(n+1)(2n+1)}{6}$.

2. A variable x takes values 1, 2, ..., n each with frequency unity. The mean of the distribution is :
   (a) $\frac{n(n+1)}{2}$ (b) $\frac{n}{2}$
   (c) $\frac{n+1}{2}$ (d) None of these.

3. The most stable measure of central tendency is :
   (a) Mean (b) Median
   (c) Mode (d) None of these.

4. Geometric mean of 2, 4, 16, 32 is :
   (a) 6 (b) 7
   (c) 8 (d) 9.

5. The mean of a set having 7 observations is 10 and the mean of a set having 3 observations is 5. The mean of the combined set is :
(a) 10 (b) 8·5
(c) 7·5 (d) None of these.

6. The A. M. of the series $1, 2, 2^2, ..., 2^n$ is :
(a) $\frac{2n-1}{n+1}$ (b) $\frac{2n+1}{n+1}$
(c) $\frac{2^{n-1}-1}{n+1}$ (d) $\frac{2^{n+1}-1}{n+1}$.

7. The daily wages of three workers are Rs. 6, 9 and 13·5. The G. M. of their wages is :
(a) Rs. 8 (b) Rs. 9
(c) Rs. 10·5 (d) Rs. 11.

8. If the frequency distribution :

Daily Wages (Rs.) :	5	6	7	10	12	15
No. of Workers :	10	?	13	8	5	4

has arithmetic mean 7·85, then the missing term (?) is :
(a) 10·5 (b) 10·05
(c) 15·5 (d) 15·05.

9. Which of the following average is mathematical average :
(a) A. M. (b) Median
(c) Mode (d) Partition values.

10. Which of the following is not mathematical average :
(a) A. M. (b) G. M.
(c) H. M. (d) Median.

11. In moderately asymmetrical distributions, the empirical relation between mean, median and mode is :
(a) Mean – Mode = Mean – Median
(b) Mean – Mode = 2 (Mean – Median)
(c) Mean – Mode = 3 (Mean – Median)
(d) Mean – Mode = 3 (Median – Mean).

12. In symmetrical distribution, the relation between mean, median and mode is :
(a) Mean = Median = Mode (b) Mean > Median > Mode
(c) Mean > Mode > Median (d) Mean < Median < Mode.

13. The weighted mean of first n natural numbers with weights equal to the corresponding number is :
(a) $\frac{1}{2}(n+1)$ (b) $\frac{1}{2}(2n+1)$
(c) $\frac{1}{3}(2n+1)$ (d) $\frac{1}{6}(n+1)(2n+1)$.

14. The median of the data 83, 54, 78, 64, 90, 59, 67, 72, 70, 73 is :
(a) 70 (b) 71
(c) 72 (d) 73.

15. The mode in 7, 12, 8, 5, 6, 4, 9, 10, 8, 9, 7, 9, 6, 5, 9 is :
(a) 7 (b) 8
(c) 9 (d) 12.

16. The arithmetic mean of a series of n items is $\bar{x}$. If the sum of first $(n-1)$ items is k then nth item is :
(a) $n\bar{x} - yk$ (b) $(\bar{x}/n) + (1/k)$
(c) $n\bar{x} - k$ (d) $n\bar{x} + k$.

17. If A, G and H be arithmetic, geometric and harmonic means of a series of n $(n > 1)$ items, then :
(a) $A < G < H$ (b) $A > G > H$
(c) $A = G = H$ (d) None of these.

18. A variable takes values 0, 1, 2, ..., n with frequencies q^n, $\frac{n}{1}q^{n-1}p$, $\frac{n(n-1)}{2}q^{n-2}p^2, \ldots, p^n$ (the terms in binomial expansion), then the mean is :
(a) n (b) np
(c) nq (d) npq.

19. If G_1 and G_2 are the geometric means of two series of n_1 and n_2 terms respectively, then the geometric mean of the combined series is :
(a) $(n_1 \log G_1 - n_2 \log G_2)/(n_1 - n_2)$
(b) $(n_1 \log G_1 + n_2 \log G_2)/(n_1 + n_2)$
(c) anti log $\{(n_1 \log G_1 - n_2 \log G_2)/(n_1 - n_2)\}$
(d) anti log $\{(n_1 \log G_1 + n_2 \log G_2)/(n_1 + n_2)\}$.

20. If G_1 and G_2 are the geometric means of two series x_i and y_i $(i = 1, 2, \ldots, n)$ respectively, then the geometric mean of (x_i/y_i) is :
(a) G_1/G_2 (b) Anti log (G_1/G_2)
(c) $n(\log G_1 - \log G_2)$ (d) Anti log $\{(\log G_1 - \log G_2)/(2n)\}$.

21. The relation between A. M. (A), G. M. (G) and H. M. (H) is :
(a) $AH > G^2$ (b) $AH < G^2$
(c) $AH = G^2$ (d) None of these.

22. Which of the following is true for all sets of data :
(a) A. M. ≤ Median ≤ Mode (b) A. M. ≥ Median ≥ Mode
(c) A. M. = Median = Mode (d) None of these.

23. The point of intersection of the ogives of 'less than' and 'greater than' represents :
(a) Mean (b) Median
(c) Geometric Mean (d) None of these.

24. That value of the data, which divides the group of data into two equal parts such that the values in one part are each less than and the values in the other part are each greater than this value, is :
(a) Mean (b) Median
(c) Mode (d) None of these.

25. For asymmetrical curves, the empirical relation between mean, median and mode is :

(a) Mean – Median = 3 (Mean – Mode)

(b) 3 (Mean – Median) = Mean – Mode

(c) Mean + Mode = 3 (Mean – Median)

(d) Mean – Mode = 3 (Mean + Mode)

26. The mean of first n odd natural numbers is :

(a) n (b) n^2

(c) $\frac{1}{2}n$ (d) $\frac{1}{2}(n+1)$.

27. Which of the following is true :

(a) Mean = 3 Median – 2 Mode (b) Median = 3 Mode – 2 Mean

(c) Mode = 3 Median – 2 Mean (d) Mode = Mean + Median.

ANSWERS

I. (a) (vi) (b) (v) (c) (vii) (d) (i) (e) (i) and (viii) (f) (iv) (g) (iii).

II.

1. (a)	**2.** (c)	**3.** (a)	**4.** (c)	**5.** (b)	**6.** (d)	**7.** (b)	**8.** (d)
9. (a)	**10.** (d)	**11.** (c)	**12.** (a)	**13.** (c)	**14.** (b)	**15.** (c)	**16.** (c)
17. (b)	**18.** (b)	**19.** (d)	**20.** (a)	**21.** (c)	**22.** (d)	**23.** (b)	**24.** (b)
25. (b)	**26.** (a)	**27.** (c).					

Chapter 5

UNIT-IV

MEASURES OF DISPERSION AND SKEWNESS, MOMENTS AND KURTOSIS

§ 5.1. DISPERSION OR VARIATION

The averages give an idea of central tendency of the given distribution but it is necessary to know how the variates are clustered around or scattered away from the average. To explain it more clearly consider the works of two typists who typed the following number of pages in 6 working days of a weak :

	Mon.	Tues.	Wed.	Thu.	Fri.	Sat.	Total Page
I typist :	15	20	25	25	30	35	150
II typist :	10	20	25	25	30	40	150

We see that each of the typist I and II typed 150 pages in 6 working days and so the average in both the cases is 25. Thus there is no difference in the average, but we know that in the first case the number of pages varies from 15 to 35 while in the second case the number of pages varies from 10 to 40. This denotes that the greatest deviation from the mean in the first case is 10 and in the second case it is 15 *i.e.*, there is a difference between the two series. The variation of this type is termed **scatter** or **dispersion** or **spread.**

Definition. *The degree to which numerical data tend to spread about an average value is called variation or dispersion or spread of the data.*

§ 5.2. DESIDERATA FOR A SATISFACTORY DISPERSION

To obtain a satisfactory dispersion, we require the following essential requisities :

(i) It should be easily calculated, rigidly defined and based on all observations.

(ii) It should be readily comprehensive and amenable to algebraic treatment.

(iii) It should be least affected by fluctuations.

§ 5.3. MEASURES OF DISPERSION

Various measures of dispersion or variation are available, the most common are the following :

1. Range.
2. Quartile deviation or Semi-interquartile range.
3. Average deviation or mean deviation.
4. Standard deviation.

§ 5.4. THE RANGE

It is the simplest possible measure of dispersion. *The range of a set of numbers (data) is the difference between the largest and the least numbers in the set (i.e., values of the variable).* If this difference is small then the series of numbers is supposed regular and if this difference is large then the series is supposed to be irregular.

For Example. In the example of § 5.1 above, the range for typist I is $35 - 15 = 20$ and that for typist II is $40 - 10 = 30$.

Demerits :

(i) It depends only on the extreme values of the variable.

(ii) It is subject to fluctuations of considerable magnitude from sample to sample.

(iii) If a giant or dwarf is included in the series representing the heights of persons, then the range will considerably change.

§ 5.5. QUARTILE DEVIATIONS (OR SEMI-INTERQUARTILE RANGE)

Definition. The inter-quartile range of a set of data is defined by

$$\textbf{Inter-quartile range} = Q_3 - Q_1$$

where Q_1 and Q_3 are respectively the first and third quartiles for the data.

Semi-interquartile range (or quartile deviation) is denoted by Q and is defined by

$$Q = \frac{1}{2}(Q_3 - Q_1)$$

where Q_1 and Q_3 have the same meaning as given above.

The semi-interquartile range is a better measure of dispersion than the range and is easily computed. Its drawback is that it does not take into account all the items.

The interdecile range is defined by $D_9 - D_1$ where D_1 and D_9 are the first and the ninth deciles of a set of data. The 80% of total frequency lie in this range.

The 10–90 percentile range is defined by $P_{90} - P_{10}$, where P_{10} and P_{90} are the 10th and 90th percentiles of a set of data.

§ 5.6. AVERAGE DEVIATION OR MEAN DEVIATION

Definition. The average (or mean) deviation about any point M, of a set of N numbers $x_1, x_2, \ldots, x_N$ is defined by

$$\textbf{Mean Deviation (M. D.)} = \delta_m = \frac{1}{N}\sum_{i=1}^{N} |x_i - M|$$

$$= \frac{1}{N}\Sigma|x - M|$$

where M is the mean or median or mode according as the mean deviation from the mean or median or mode is to be computed, $|x_i - M|$ represents the absolute (or numerical) value. Thus $|-5| = 5$.

If $x_1, x_2, \ldots, x_k$ occur with frequencies $f_1, f_2, \ldots, f_k$ respectively, then the mean deviation (δ_m) is defined by

$$\delta_m = \frac{1}{N}\sum_{j=1}^{k} f_j |x_j - M| = \frac{1}{N}\Sigma f|x - M|.$$

Mean deviation depends on all the values of the variables and therefore it is a better measure of dispersion than the range or the quartile deviation. Since signs of the deviations are ignored (because all deviations are taken positive), some artificiality is created.

In, case of grouped frequency distribution the mid-values are taken as x.

ILLUSTRATIVE EXAMPLES

Example 1. ***Find the mean deviation from the arithmetic mean of the following distribution :***

Marks :	0–10	10–20	20–30	30–40	40–50
No. of Students :	5	8	15	16	6

Solution. Let assumed mean $(A) = 25$ and $i = 10$

Class	Mid Value x	Frequency f	$u = \frac{x - A}{i}$	fu	x − M	f | x − M |
0–10	5	5	−2	−10	−22	110
10–20	15	8	−1	−8	−12	96
20–30	25	15	0	0	−2	30
30–40	35	16	1	16	8	128
40–50	45	6	2	12	18	108
Total		$N = \Sigma F = 50$		$\Sigma fu = 10$		$\Sigma f\|x - M\| = 472$

$\therefore$ Arithmetic mean $M = A + \frac{\Sigma fu}{N} \times i = 25 + \frac{10}{50} \times 10 = 27.$

$\therefore$ The required mean deviation from arithmetic mean

$$\delta_m = \frac{\Sigma f\,|\,x - M\,|}{N} = \frac{472}{50} = 9 \cdot 44.$$

Example 2. ***Find the mean deviation from the mean from the following series:***

Age (Less then) :	10	20	30	40	50	60	70	80
No. of persons :	15	30	53	75	100	110	115	125

Solution.

Age (less than)	Class	Cumulative frequency (c. f.)	Frequency f	Mid-Value x	$u = \frac{x - 45}{10}$	fu	x − M	f | x − M |
10	0–10	15	15	5	− 4	− 60	− 30·16	452·40
20	10–20	30	15	15	− 3	− 45	− 20·16	302·40
30	20–30	53	23	25	− 2	− 46	− 10·16	233·68
40	30–40	75	22	35	− 1	− 22	− 0·16	3·52
50	40–50	100	25	45	0	0	9·84	246·00
60	5 0–60	110	10	55	1	10	19·84	190·84
70	60–70	115	5	65	2	10	29·84	149·20
80	70–80	125	10	75	3	30	39·84	398·40
	Total		$N = \Sigma F = 125$			$\Sigma fu = -123$		$\Sigma f\,\|x - M\| = 1976{\cdot}44$

Let assumed mean = $A = 45$ and $i = 10$.

$\therefore$ A.M. $(M) = A + \dfrac{\Sigma fu}{N} \times i$

$$= 45 + \frac{(-123) \times 10}{125} = 45 - 9 \cdot 84 = 35 \cdot 16.$$

$\therefore$ The required mean deviation (M. D.) is given by

$$\delta_m = \frac{\Sigma f\,|\,x - M\,|}{N} = \frac{1976 \cdot 44}{125} = 15 \cdot 8.$$

Example 3. ***Find the mean deviation from the arithmetic mean for the following data, which represents the circumference for the necks of a set of students :***

Mid-Value (in cm) x :	30	31·5	33	34·5	36	37·5	39	40·5
No. of Students f :	4	19	30	63	66	29	18	1

Solution.

x	f	$\xi = x - A$	$f\xi$	x − M	f | x − M |
30	4	− 6·0	− 24·0	− 5·0	20·0
31·5	19	− 4·5	− 85·5	− 3·5	66·5
33	30	− 3·0	− 90·0	− 2·9	60·0
34·5	63	− 1·5	− 94·5	− 0·5	31·5
36	66	0	0·0	1·0	66·0
37·5	29	1·5	43·5	2·5	72·5
39	18	3·0	54·0	4·0	72·0
40·5	1	4·5	4·5	5·4	5·5
	$N = \Sigma f = 230$		$\Sigma f\zeta = -1192$		$\Sigma f\,\|x - M\| = 394$

Let the assumed mean $(A) = 36$, then arithmetic mean

$$(M) = A + \frac{\Sigma f\xi}{N} = 36 + \frac{(-192)}{230}$$

$$= 36 - 0 \cdot 835 = 35 \cdot 165 = 35\,\text{cm.} \quad \text{(approx.)}$$

$\therefore$ The required mean deviation (*M. D.*) is given by

$$\delta_m = \frac{\Sigma f\,|\,x - M\,|}{N} = \frac{394}{230} = 1 \cdot 713.$$

Example 4. ***Find the average deviation about median for the following distribution :***

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

Solution.

x	f	c. f.	$x - M_d$	$f\,\|x - M_d\|$
6	4	4	− 18	72
12	7	11	− 12	84
18	9	20	− 6	54
24	18	38	0	0
30	15	53	6	90
36	10	63	12	120
42	5	68	18	90
	$N = \Sigma f = 68$			$\Sigma f\,\|x - M_d\| = 510$

Here $N = 68$ which is even. The two mid terms are $\frac{1}{2}N$th *i.e.*, 34th and $\left(\frac{1}{2}N+1\right)^{\text{th}}$ th *i.e.*, 35th term. From the table it is clear that both of these terms lie in cumulative frequency 38. Hence the median is 24 *i.e.*, $M_d = 24$. So the required average deviation is given by

$$\delta_{M_d} = \frac{\Sigma f \mid x - M_d \mid}{N} = \frac{510}{68} = 7 \cdot 5.$$

Example 5. ***Compute the semi-interquartile range of the marks of 63 students in Mathematics given below :***

Marks Group	No. of Students	Marks Group	No. of Students
0–10	5	50–60	7
10–20	7	60–70	3
20–30	10	70–80	2
30–40	16	80–90	2
40–50	11	90–100	0

Solution.

Marks Group	Frequency f	Cumulative Frequency c. f.
0–10	5	5
10–20	7	12
20–30	10	22
30–40	16	28
40–50	11	49
50–60	7	56
60–70	3	59
70–80	2	61
80–90	2	63
90–100	0	63
	$N = \Sigma f = 63$	

To Calculate lower Quartile Q_1. Here $N = 63$. So $\frac{1}{4}(N+1)$th *i.e.*, 16th students lies in the marks group 20–30. Thus lower quartile class is 20–30.

$$\therefore \quad Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i = 20 + \frac{15 \cdot 75 - 12}{10} \times 10 = 23 \cdot 75.$$

Similarly $Q_3 = 40 + \frac{47 \cdot 25 - 38}{11} \times 10 = 48 \cdot 4.$

$$\therefore \quad \text{Semi-inter quartile range} = \frac{1}{2}(Q_3 - Q_1)$$

$$= \frac{1}{2}(48 \cdot 4 - 23 \cdot 75) = 12 \cdot 32 \quad \text{Marks.}$$

EXERCISE 5 (A)

1. The marks obtained by four students are : 25, 25, 45, 55. Find the mean deviation from the arithmetic mean.
2. Daily income of five workers are : (in Rs.) 30, 40, 45, 50, 55. Find the mean deviation about median.
3. Compute the mean deviation about the median of the following series :

Measure :	4	6	8	10	12	14	16
Frequency :	2	4	5	3	2	1	4

4. Find the mean deviation from the arithmetic mean of the following data :

x :	56	63	70	77	84	91	98
f :	3	6	14	16	13	6	2

5. Find the mean deviation from the mean of the following data :

(i)

Class :	0–10	10–20	20–30	30–40	40–50
Frequency :	2	8	10	3	4

(ii)

Class :	0–6	6–12	12–18	18–24	24–30
Frequency :	8	10	12	9	5

6. Find the mean deviation from the mean of the following data :

Class :	140–150	150–160	160–170	170–180	180–190	190–200
Frequency :	4	6	10	12	9	4

7. Find the mean deviation from the mean and also about the median of the following data :

Income per month (in Rs.) :	20–30	30–40	40–50	50–60	60–70
No. of families :	120	201	105	75	25

8. Find the mean deviation about the 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

9. Compute the mean deviation about the mean, meadian and mode for the following data :

Measure :	4	6	8	10	12	14	16
Frequency :	2	1	3	6	4	3	1

10. Compute the mean deviation about the mean, median and mode of the following data :

Class :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
f :	3	8	15	20	25	10	9	6	4

11. Compute the mean deviation from the arithmetic mean, median and mode for the following frequency distribution :

Class :	140–150	150–160	160–170	170–180	180–190	190–200
frequency :	4	6	10	18	9	3

12. The income of a person for 12 months is given by the series :

120 150 151 151 157 158 160 161 162 175 175

Find the quartile dispersion of the income.

ANSWERS

1. 19 **2.** 7 **3.** 3·24 **4.** 76·53 **5.** (i) 8·614 (ii) 6·3

6. $M_0 = 174$, M. D. = 11·22 **7.** Rs. 9·2, Rs. 9·1

8. 9·56 **9.** 2·44, 2·4, 2·4 **10.** 14·992, 15·072, 15·00

11. 10·56, 10·24, 9·86

◈ § 5.7. STANDARD DEVIATION, ROOT-MEAN SQUARE DEVIATION

Root-mean Square Deviation :

Definition. It is defined as the positive square root of the mean of the squares of the deviations from an origin A.

It is denoted by s, thus

$$s = +\sqrt{\left\{\frac{1}{N}\Sigma f(x-A)^2\right\}}.$$

Mean square deviation. It is denoted by s^2 and is defined as the mean of the squares of the deviations from an origin A. Thus

$$s^2 = \frac{1}{N}\Sigma f(x-A)^2.$$

Remark. The origin A may be taken any arbitrary point A.

Standard Deviation :

Definition. Standard deviation (or S. D.) is the positive square root of the arithmetic mean of the square deviations of various values from their arithmetic mean M. It is usually denoted by σ.

Thus
$$\sigma = +\sqrt{\left\{\frac{1}{N}\Sigma f(x-M)^2\right\}}.$$

Notes 1. When the deviation is calculated from the arithmetic mean M, then root mean square deviation becomes standard deviation.

2. The square of the standard deviation σ^2 is called **variance.**

3. The quantity s^2 is said to be **second moment** about the value A and is denoted by μ_2'.

4. The variance σ^2 is called the **second moment about the mean** M and is denoted by μ_2.

◈ § 5.8. RELATION BETWEEN STANDARD DEVIATION AND ROOT-MEAN SQUARE DEVIATION

Consider a frequency distribution (discrete series)

x :	x_1	x_2	...	x_n
f :	f_1	f_2	...	f_n

Let A be the assumed mean and M the arithmetic mean. Also suppose $M - A = d$. Then

$$\sigma^2 = \frac{1}{N}\Sigma f(x-M)^2 = \frac{1}{N}\Sigma f(x - A - \overline{M-A})^2$$

$$= \frac{1}{N}\Sigma f(X-d)^2 \text{ where } X = x - A,\ d = M - A$$

$$= \frac{1}{N}\Sigma f X^2 - 2d\cdot\frac{1}{N}\Sigma f X + d^2\cdot\frac{1}{N}\Sigma f$$

$$= \frac{1}{N}\Sigma f(x-A)^2 - 2d\cdot\frac{1}{N}\Sigma f(x-A) + d^2 \qquad [\because \Sigma f = N]$$

$$= \frac{1}{N}\Sigma f(x-A)^2 - 2d\left(\frac{1}{N}\Sigma fx - \frac{1}{N}\Sigma fA\right) + d^2$$

$$= \frac{1}{N}\Sigma f(x-A)^2 - 2d(M-A) + d^2$$

$$= \frac{1}{N}\Sigma f(x-A)^2 - d^2 = s^2 - d^2$$

Hence
$$s^2 = \sigma^2 + d^2. \qquad \ldots(1)$$

Relation (1) shows that s is least when $d = 0$ *i.e.*, $A = M$ and the least value of s is equal to σ.

In other words the standard deviation is the least possible root mean square deviation.

Remark. Since $d^2 > 0$, always, therefore, from (1), we have

$$s^2 > \sigma^2$$

i.e., mean square deviation about any point A is greater than variance.

Example. *Show that the standard deviation is the least possible root mean square deviation.*

◈ § 5.9. SHORT CUT METHOD FOR CALCULATING OF STANDARD DEVIATION

We know that

$$\sigma^2 = \frac{1}{N}\Sigma f(x-M)^2 = \frac{1}{N}\Sigma f(x - A - \overline{M-A})^2$$

$$= \frac{1}{N}\Sigma f(\xi - \overline{M-A})^2, \text{ where } \xi = x - A$$

$$= \frac{1}{N}\Sigma f\xi^2 - 2(M-A)\cdot\frac{1}{N}\Sigma f\xi + (M-A)^2\cdot\frac{1}{N}\Sigma f$$

$$= \frac{1}{N}\Sigma f\xi^2 - \left(\frac{\Sigma f\xi}{N}\right)^2 \qquad \left[\because\ M - A = \frac{\Sigma f\,\xi}{N},\ \Sigma f = N\right]$$

$$\therefore \quad \sigma = +\sqrt{\left\{\frac{1}{N}\Sigma f\xi^2 - \left(\frac{\Sigma f\xi}{N}\right)^2\right\}}.$$

◈ § 5.10. STEP-DEVIATION METHOD

If $u = \dfrac{x-A}{h} = \dfrac{\xi}{h}$, then $\xi = uh$.

$\therefore$ By short cut method

$$\sigma^2 = \frac{1}{N}\Sigma f u^2 h^2 - \left(\frac{\Sigma f u h}{N}\right)^2$$

or

$$\sigma^2 = h^2\left\{\frac{1}{N}\Sigma f u^2 - \left(\frac{\Sigma f u}{N}\right)^2\right\}.$$

$$\therefore \quad \sigma = h\sqrt{\left[\left\{\frac{1}{N}\Sigma f u^2 - \left(\frac{\Sigma f u}{N}\right)^2\right\}\right]}.$$

◈ § 5.11. EMPIRICAL RELATION BETWEEN MEASURES OF DISPERSION

$$\text{Mean Deviation} = \frac{4}{5}\,(\text{Standard Deviation})$$

$$\text{Semi-inter quartile range} = \frac{2}{3}\,(\text{Standard Deviation}).$$

◈§ 5.12. COEFFICIENT OF DISPERSION

Definition. Ratio σ / M is defined as coefficient of Dispersion, where σ is standard deviation and M is the arithmetic mean.

For **quartile deviation** the coefficient of dispersion (Q. D.) is defined by

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

The **Coefficient of Variation** (or C.V.) is defined by

$$\text{C. V.} = \frac{\sigma}{M} \times 100.$$

ILLUSTRATIVE EXAMPLES

Example 1. *Find the variance of a series, in which the values of x are the natural numbers* 1, 2, ..., *N and frequency of each is* 1.

Solution. Here

x:	1	2	3	...	N
f:	1	1	1	...	1

$$\Sigma x = 1 + 2 + 3 + \ldots + N = \frac{1}{2}N(N+1)$$

$$\Sigma x^2 = 1^2 + 2^2 + 3^2 + \ldots + N^2 = \frac{1}{6}N(N+1)(2N+1).$$

$\therefore$ Variance, $\sigma^2 = \dfrac{\Sigma x^2}{N} - \left(\dfrac{\Sigma x}{N}\right)^2$

$$= \frac{N(N+1)(2N+1)}{6N} - \left\{\frac{N(N+1)}{2N}\right\}^2$$

$$= \frac{(N+1)}{12}\,[2(2N+1) - 3(N+1)]$$

$$= \frac{1}{12}(N+1)(N-1) = \frac{1}{12}(N^2-1).$$

And Standard Deviation, $\sigma = \sqrt{\left(\dfrac{N^2-1}{12}\right)}$.

Example 2. ***Calculate the meanm variance and S. D. for a series in which the values of x are*** $1^2, 2^2, 3^2, \ldots, n^2$.

Solution. Here x: 1^2 2^2 3^2 ... n^2

$$\Sigma x = 1^2 + 2^2 + \ldots + n^2 = \frac{1}{6}n(n+1)(2n+1)$$

$$\Sigma x^2 = 1^4 + 2^4 + \ldots + n^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

$\therefore$ $$\text{Mean} = \frac{\Sigma x}{n} = \frac{\frac{1}{6}n(n+1)(2n+1)}{n} = \frac{1}{6}(n+1)(2n+1)$$

Variance, $\sigma^2 = \left(\dfrac{\Sigma x^2}{n}\right) - \left(\dfrac{\Sigma x}{n}\right)^2$

$$= \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30n} - \left\{\frac{n(n+1)(2n+1)}{6n}\right\}^2$$

$$= \frac{(n+1)(2n+1)}{6}\left[\frac{3n^2+3n-1}{5} - \frac{(n+1)(2n+1)}{6}\right]$$

$$= \frac{(n+1)(2n+1)}{180}\,[8n^2+3n-11]$$

$\therefore$ S. D., $\sigma = \sqrt{\left[\dfrac{(n+1)(2n+1)(8n^2+3n-11)}{180}\right]}$.

Example 3. ***Calculate the S. D. and coefficient of variation (C.V.) for the following table :***

Class :	**0–10**	**10–20**	**20–30**	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**
Frequency :	**5**	**10**	**20**	**40**	**30**	**20**	**10**	**5**

Solution. We prepare the following table for the computation of S. D.

Class	Mid-value x	f	$u = \frac{x-55}{10}$	fu	fu^2
0–10	5	5	−3	−15	45
10–20	15	10	−2	−20	40
20–30	25	20	−1	−20	20
30–40	35	40	0	0	0
40–50	45	30	1	30	30
50–60	55	20	2	40	80
60–70	65	10	3	30	90
70—80	75	5	4	20	80
		$N = \Sigma f = 140$		$\Sigma fu = 65$	$\Sigma fu^2 = 385$

Let assumed mean = 35 = A (say) and $h = 10$

$$\therefore \quad \text{A. M.,} \quad M = A + h\frac{\Sigma fu}{N} = 35 + 10\left(\frac{65}{140}\right)$$

$$= 35 + 4 \cdot 64 = 39 \cdot 64$$

$$\text{S. D.,} \quad \sigma = h\sqrt{\left[\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2\right]}$$

$$= 10\sqrt{\left[\frac{385}{140} - (0 \cdot 464)^2\right]}$$

$$= 10\sqrt{[2 \cdot 75 - 0 \cdot 215]} = 10\sqrt{(2 \cdot 535)} = 10 \times 1 \cdot 59 = 15 \cdot 9$$

$$\text{C. V.} = \frac{\sigma}{M} \times 100 = \frac{15 \cdot 9}{39 \cdot 64} \times 100 = 40 \cdot 11\%. \qquad \text{Ans.}$$

Example 4. ***Find the standard deviation for the following data which represents the wages of 320 workers :***

Wages (In Rs.)	No. of Workers	Wages (In Rs.)	No. of Workers
70–80	12	110–120	50
80–90	18	120–130	45
90–100	35	130–140	20
100–110	42	140–150	8

Solution. We have the following table for the computation of S. D.

Wages Class	Mid-Value x	f	$u = \frac{x-105}{10}$	fu	fu^2
70–80	75	12	−3	−36	108
80–90	85	18	−2	−36	72
90–100	95	35	−1	−35	35
100–110	105	42	0	0	0
110–120	115	50	1	50	50
120–130	125	45	2	90	180
130–140	135	20	3	60	180
140–150	145	8	4	32	128
		$N = \Sigma f = 230$		$\Sigma fu = 125$	$\Sigma fu^2 = 753$

Let assumed mean = 105 = A say and $i = 10$.

$$\sigma = i\sqrt{\left\{\frac{1}{N}\Sigma fu^2 - \left(\frac{1}{N}\Sigma fu\right)^2\right\}}$$

$$= 10\sqrt{\left\{\frac{753}{230} - \left(\frac{125}{230}\right)^2\right\}} = \text{Rs. } 17 \cdot 3.$$

Example 5. ***Find the standard deviation of the following series :***

Marks (above) :	**0**	**10**	**20**	**30**	**40**	**50**	**60**	**70**
No. of Students :	**100**	**90**	**75**	**50**	**25**	**15**	**5**	**0**

Solution. By direct method. Arranging the given series in class intervals, we have

Class	Mid-Value x	f	fx	x − M	$(x-M)^2$	$f(x-M)^2$
0–10	5	10	50	−26	676	6760
10–20	15	15	225	−16	256	3840
20–30	25	25	625	−6	36	900
30–40	35	25	875	4	16	400
40–50	45	10	450	14	196	1960
50–60	55	10	550	24	576	5760
60–70	65	5	325	34	1156	5780
Total		$N = 100$	$\Sigma fx = 3100$			25400

$\therefore$ Mean $(M) = \dfrac{\Sigma fx}{N} = \dfrac{3100}{100} = 31.$

$\therefore$ Required S. D. is given by

$$\sigma = \sqrt{\left\{\frac{\Sigma f(x-M)^2}{N}\right\}} = \sqrt{\left(\frac{25400}{100}\right)} = 15 \cdot 94.$$

By short-cut Method.

Class	Mid-Value x	f	$\xi = x - A$ $(A = 35)$	$f\xi$	$f\xi^2$
0–10	5	10	−30	−300	9000
10–20	15	15	−20	−300	6000
20–30	25	25	−10	−250	2500
30–40	35	25	0	0	0
40–50	45	10	10	100	1000
50–60	55	10	20	200	4000
60–70	65	5	30	150	4500
Total		$N = 100$		−400	27000

Let assumed mean $(A) = 35.$

$\therefore$ Mean $(M) = A + \dfrac{\Sigma f\,\xi}{N}$

$$= 35 + \frac{(-400)}{100} = 35 - 4 = 31$$

$$\text{S. D., } \sigma = \sqrt{\left\{\frac{\Sigma f\xi^2}{N} - \left(\frac{\Sigma f\xi}{N}\right)^2\right\}}$$

$$= \sqrt{\left\{\frac{27000}{100} - \left(\frac{-400}{100}\right)^2\right\}}$$

$$= \sqrt{(270 - 16)} = \sqrt{(254)} = 15 \cdot 94.$$

By Step-deviation Method :

Class	Mid-Value x	f	$u = \dfrac{x - 35}{10}$	fu	fu^2
0–10	5	10	−3	−30	90
10–20	15	15	−2	−30	60
20–30	25	25	−1	−25	25
30–40	35	25	0	0	0
40–50	45	10	1	10	10
50–60	55	10	2	20	40
60–70	65	5	3	15	45
Total		$N = 100$		$\Sigma fu = -40$	$\Sigma fu^2 = 270$

Here $A = 35,\ h = 10.$

$\therefore$
$$\sigma = h\sqrt{\left\{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2\right\}}$$

$$= 10\sqrt{\left\{\frac{270}{100} - \left(\frac{-40}{800}\right)^2\right\}}$$

$$= 10\sqrt{(2 \cdot 7 - 0 \cdot 16)} = 10\sqrt{(2 \cdot 54)} = 15 \cdot 94.$$

Note. Students are required to solve the problem by any one method.

Example. 6. ***Find the mean deviation about the mean and standard deviation of the series $a, a + d, \ldots, a + 2nd$. Also show that the standard deviation is greater than the mean deviation.***

Solution. The given series $a,\ a + d,\ a + 2d, \ldots,\ a + nd, \ldots a + 2nd$ is arithmetical series of $2n + 1$ terms

$$\text{Mean } (M) = \frac{a + (a + d) + (a + 2d) + \ldots + (a + 2nd)}{2n + 1}$$

$$= \frac{1}{2n+1}\left[(2n + 1)\,a + (1 + 2 + 3 + \ldots + 2n)\,d\right]$$

$$= \frac{1}{2n+1}\left[(2n + 1)\,a + \frac{2n\,(2n + 1)}{2}\,d\right]$$

$$= a + nd.$$

$\therefore$ Mean deviation about mean is given by

$$\delta_M = \frac{1}{2n+1}\sum_{r=0}^{2n} |\,(a + rd) - M\,|$$

$$= \frac{1}{2n+1}\sum_{r=0}^{2n} |\,(a + rd) - (a + nd)\,|$$

$$= \frac{d}{2n+1}\sum_{r=0}^{2n} |\,r - n\,|$$

$$= \frac{2d}{2n+1}\{n + (n - 1) + \ldots + 2 + 1\}$$

$$= \frac{2d}{2n+1} \cdot \frac{n\,(n + 1)}{2} = \frac{n\,(n + 1)\,d}{2n + 1}$$

and
$$\sigma^2 = \frac{1}{2n+1}\sum_{r=0}^{2n} \{(a + rd) - M\}^2$$

$$= \frac{1}{2n+1}\sum_{r=0}^{2n} \{(a + rd) - (a + nd)\}^2$$

$$= \frac{2d^2}{2n+1}\left[n^2 + (n - 1)^2 + \ldots + 2^2 + 1^2\right]$$

$$= \frac{2d^2}{2n+1} \cdot \frac{n(n+1)(2n+1)}{6}$$

$$= \frac{n(n+1)\,d^2}{3}$$

$$\sigma = d\sqrt{\left\{\frac{n(n+1)}{3}\right\}}.$$

Now S. D. > M. D.

if $d\sqrt{\left(\frac{n(n+1)}{3}\right)} > \frac{n(n+1)\,d}{2n+1}$

i.e., if $\frac{n(n+1)}{3} > \frac{n^2(n+1)^2}{(2n+1)^2}$

i.e., if $(2n+1)^2 > 3n(n+1)$

i.e., if $n^2 + n + 1 > 0$, which is ture as $n > 0$.

Hence S. D. > Mean deviation.

Example 7. ***Find the standard deviation of the following two series. Which of these shows more variation***

***Series A*: 192, 288, 236, 229, 184, 260, 348, 291, 330, 243,**

***Series B*: 83, 87, 93, 109, 124, 126, 126, 101, 102, 108,**

Solution. Here

Series A			Series B		
x	$\xi = x - A$ (A = 260)	ξ^2	x	$\xi = x - A$ (A = 105)	ξ^2
192	−68	4624	83	−22	484
288	28	784	87	−18	324
236	−24	576	93	−12	144
229	−31	961	109	4	16
184	−76	5776	124	19	361
260	0	0	126	21	441
348	88	7744	126	21	441
291	31	961	101	−4	16
330	70	4900	102	−3	9
243	−71	289	108	3	9
	$\Sigma\xi = 1$	$\Sigma\xi^2 = 26615$		$\Sigma\xi = 9$	$\Sigma\xi^2 = 2245$

For Series A.

Arithmetic mean $(M) = A + \frac{\Sigma\xi}{n}$

$$= 260 + 1/10 = 260 + 0{\cdot}1 = 260{\cdot}1$$

S. D. is given by $\sigma = \sqrt{\left\{\frac{\Sigma\xi^2}{n} - \left(\frac{\Sigma\xi}{n}\right)^2\right\}}$

$$= \sqrt{\left\{\left(\frac{26615}{10}\right) - \left(\frac{1}{10}\right)^2\right\}}$$

$$= \sqrt{(2661{\cdot}5 - 0{\cdot}01)} = \sqrt{(2661{\cdot}49)} = 51{\cdot}6$$

$$\text{C. V.} = \frac{\sigma}{M} \times 100 = \frac{51{\cdot}6}{260{\cdot}1} \times 100 = 19{\cdot}8.$$

For Series B.

$$\text{Mean } (M) = A + \frac{\Sigma\xi}{n}$$

$$= 105 + \frac{9}{10} = 105 + 0{\cdot}9 = 105{\cdot}9$$

And S. D. is given by

$$\sigma = \sqrt{\left\{\frac{\Sigma\xi^2}{n} - \left(\frac{\Sigma\xi}{n}\right)^2\right\}}$$

$$= \sqrt{\left\{\left(\frac{2245}{10}\right) - \left(\frac{9}{10}\right)^2\right\}}$$

$$= \sqrt{(224{\cdot}5 - 0{\cdot}81)} = \sqrt{(223{\cdot}69)} = 14{\cdot}96.$$

$$\therefore \quad \text{C. V.} = \frac{\sigma}{M} \times 100 = \frac{14{\cdot}96}{105{\cdot}9} \times 100 = 14{\cdot}1.$$

Since the coefficient of variation (C. V.) of series A is more than that of series B. Hence the series A has more variation.

Example 8. ***The details of runs gained by two batsmen A and B in different innings are as follows*:**

***A*:**	**24**	**79**	**31**	**114**	**14**	**02**	**68**	**01**	**110**	**07**
***B*:**	**05**	**18**	**42**	**53**	**09**	**47**	**52**	**17**	**81**	**56**

Which of the two player is better run scorer ?

Solution. To compute the coefficient of variation (C. V.) of runs of the batsman *A* :

x	$x - \bar{x}_A$	$(x - \bar{x}_A)^2$
24	−21	441
79	34	1156
31	−14	196
114	69	4761
14	−31	961
02	−43	1849
68	23	529
01	−44	1936
110	65	4225
07	−38	1444
Total 450	–	17498

Here $\Sigma x = 450,\ n = 10.$

$\therefore$ Mean $= \bar{x}_A = \dfrac{450}{10} = 45$ Runs.

$$\sigma_A = \sqrt{\left[\frac{\Sigma (x - \bar{x}_A)^2}{n}\right]} = \sqrt{\left(\frac{17498}{10}\right)} = 41 \cdot 83 \text{ Runs.}$$

$\therefore$ C. V. $= V_A = \dfrac{\sigma_A}{\bar{x}_A} \times 100 = \dfrac{41 \cdot 83}{42} \times 100 = 92 \cdot 96\%.$

To compute coefficient of variation (C. V.) of B.

x	$x - \bar{x}_B$	$(x - \bar{x}_B)^2$
05	−33	1089
18	−20	400
42	4	16
53	15	225
09	−29	841
47	9	81
52	14	196
17	−21	441
81	43	1849
56	18	324
Total 380	–	5462

Here $\Sigma x = 380,\ n = 10.$

$\therefore$ $\bar{x}_B = \dfrac{380}{10} = 38$ Runs.

$$\sigma_B = \sqrt{\left[\frac{\Sigma (x - \bar{x}_B)^2}{n}\right]} = \sqrt{\left(\frac{5462}{10}\right)} = 23 \cdot 37 \text{ Runs.}$$

$\therefore$ C. V. $= V_B = \dfrac{\sigma_B}{\bar{x}_B} \times 100 = \dfrac{23 \cdot 37}{38} \times 100 = 61 \cdot 50\%.$

(i) Since $\bar{x}_A > \bar{x}_B$, therefore, A is a better batsman.

(ii) Here $V_B \leq V_A$, therefore, the value of dispersion in the scoring of B is less than that of A *i.e.*, B is more consistent.

Example 9. ***The number of goals scored by two teams in a football match are as follows :***

No. of goals scored in a Football match	No. of Football matches	
	Team A	Team B
0	**15**	**20**
1	**10**	**10**
2	**07**	**05**
3	**05**	**04**
4	**03**	**02**
5	**02**	**01**
Total	**42**	**42**

Compute the coefficients of variations of the teams and find that which team has more consistency ?

Solution. For team A :

x	f_A	$f_A x$	$f_A x^2$
0	15	0	0
1	10	10	10
2	7	14	28
3	5	15	45
4	3	12	48
5	2	10	50
Total	42	61	181

Mean, $\bar{x}_A = \dfrac{\Sigma f_A x}{\Sigma f_A} = \dfrac{61}{42} = 1 \cdot 45$ Goals

S. D. $\sigma_A = \sqrt{\left\{\dfrac{\Sigma f_A x^2}{\Sigma f_A} - \left(\dfrac{\Sigma f_A x}{\Sigma f_A}\right)^2\right\}}$

$$= \sqrt{\left\{\frac{181}{42} - (1 \cdot 45)^2\right\}}$$

$$= \sqrt{(4 \cdot 3095 - 2 \cdot 1025)} = \sqrt{(2 \cdot 2070)} = 1 \cdot 48 \text{ Goals.}$$

$\therefore$ Coefficient of variation for A is given by

$$V_A = \frac{\sigma_A}{\bar{x}_A} \times 100 = \frac{1 \cdot 48}{1 \cdot 45} \times 100 = 102 \cdot 07\%$$

For team B :

x	f_B	$f_B x$	$f_B x^2$
0	20	0	0
1	10	10	10
2	5	10	20
3	4	12	36
4	2	8	32
5	1	5	25
Total	42	45	123

Mean, $\quad x_B = \dfrac{\Sigma f_B x}{\Sigma f_B} = \dfrac{45}{42} = 1 \cdot 07$ Goals

S. D., $\sigma_B = \sqrt{\left\{\dfrac{\Sigma f_B x^2}{\Sigma f_B} - \left(\dfrac{\Sigma f_B x}{\Sigma f_B}\right)^2\right\}}$

$$= \sqrt{\left\{\frac{123}{42} - (1 \cdot 07)^2\right\}}$$

$$= \sqrt{(2 \cdot 9286 - 1 \cdot 449)} = \sqrt{(1 \cdot 7837)} = 1 \cdot 34 \text{ Goals.}$$

C. V., $V_B = \dfrac{\sigma_B}{\bar{x}_B} \times 100$

$$= \frac{1 \cdot 34}{1 \cdot 07} \times 100 = 125 \cdot 23\%.$$

Since $V_A < V_B$. Hence team A has more consistency.

Example 10. ***The mean deviation for a series is 15. Deduce the maximum possible quartile deviation.***

Solution. Here M. D. = 15

$$\text{Q. D.} = \frac{5}{6} \text{ M. D.} = \frac{5}{6} \times 15 = 12 \cdot 5.$$

Hence the maximum possible quartile deviation is $12 \cdot 5$.

Example 11. ***If mean and standard deviation of a variable x are m and σ respectively, then compute the mean and standard deviations of $(ax + b)/c$, where a, b and c are constants.***

Solution. Let $y = (ax + b)/c$ Again let $\bar{y}$ and σ_y be the mean and standard deviations respectively of y. Then

$$\bar{y} = \frac{1}{N} \Sigma \left\{\frac{f(ax + b)}{c}\right\} = \frac{1}{c}\left[a \cdot \frac{\Sigma fx}{N} + b \cdot \frac{\Sigma f}{N}\right]$$

$$= \frac{1}{c}[am + b] \qquad [\because \Sigma f = N, (\Sigma fx)/N = m]$$

and

$$\sigma_y^2 = \frac{1}{N} \Sigma f(y - \bar{y})^2 = \frac{1}{N} \Sigma f\left(\frac{ax + b}{c} - \frac{am + b}{c}\right)^2$$

$$= \frac{a^2}{c^2} \Sigma f(x - m)^2 = \frac{a^2}{c^2} \sigma^2.$$

$\therefore \quad \sigma_y = |a/c| \sigma.$

Example 12. *Show that if the variable takes the values $0, 1, 2, \ldots, n$ with frequencies proportional to the binomial coefficients ${}^nC_0, {}^nC_1, {}^nC_2, \ldots, {}^nC_n$ respectively, then the mean is $\frac{1}{2}n$, the mean square deviation about $x = 0$ in $\frac{1}{4}n(n+1)$, and the variance is $\frac{1}{4}n$. Also find the standard deviation. Verify that variance is half of the mean.*

Solution. The frequency distribution is :

x :	0	1	2	...	n
f :	$k \cdot {}^nC_0$	$k \cdot {}^nC_1$	$k \cdot {}^nC_2$	...	$k \cdot {}^nC_n$

where k is constant of proportionality.

Here $\quad \Sigma f = k({}^nC_0 + {}^nC_1 + \ldots + {}^nC_n)$

$$= k(1 + 1)^n = k \cdot 2^n$$

$$\Sigma fx = \sum_{r=1}^{n} k \cdot {}^nC_r \cdot r$$

$$= k \sum_{r=1}^{n} \frac{n!}{r!(n-r)!} \cdot r = k \sum_{r=1}^{n} \frac{n(n-1)!}{(r-1)!(n-r)!}$$

$$= kn \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} = kn \sum_{r=1}^{n} {}^{n-1}C_{r-1}$$

$$= kn(1 + 1)^{n-1} = rn \cdot 2^{n-1}$$

And, $\quad \Sigma fx^2 = \sum_{r=0}^{n} k \cdot {}^nC_r \cdot r^2$

$$= k \sum_{r=0}^{n} {}^nC_r \cdot [r(r-1) + r] \qquad [\text{Putting } r^2 = r(r-1) + r]$$

$$= k \cdot \sum_{r=0}^{n} r(r-1) \cdot \frac{n!}{r!(n-r)!} + k \sum_{r=0}^{n} r \cdot {}^nC_r$$

$$= kn(n-1)\sum_{r=2}^{n}\frac{(n-2)!}{(r-2)!(n-r)!} + kn \cdot 2^{n-1}$$

$$= kn(n-1)\sum_{r=2}^{n} {}^{n-2}C_{r-2} + kn \cdot 2^{n-1}$$

$$= kn(n-1)(1+1)^{n-2} + kn \cdot 2^{n-1}$$

$$= kn(n-1)2^{n-2} + kn \cdot 2^{n-1}$$

$$= kn \cdot 2^{n-2}(n-1+2) = kn(n+1)2^{n-2}.$$

Thus arithmetic mean,

$$M = \frac{\Sigma fx}{\Sigma f} = \frac{kn \cdot 2^{n-1}}{k \cdot 2^n} = \frac{n}{2}.$$

Mean square deviation about the origin is

$$s^2 = \frac{\Sigma fx^2}{\Sigma f} = \frac{kn(n+1) \cdot 2^{n-2}}{k \cdot 2^n} = \frac{1}{4}n(n+1)$$

variance, $\sigma^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2 = \frac{n(n+1)}{4} - \left(\frac{n}{2}\right)^2 = \frac{n}{4}.$

Standard deviation, $\sigma = \sqrt{\left(\frac{n}{4}\right)} = \frac{1}{2}\sqrt{n}.$

Also $\frac{\text{variance}}{\text{mean}} = \frac{(n/4)}{(n/2)} = \frac{1}{2}$

i.e., variance $= \frac{1}{2}$ (mean).

Example 13. ***Find the mean and standard deviation for the following frequency distribution.***

Age (in years)	:	**10–20**	**20–30**	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**
Frequency	:	**4**	**8**	**10**	**16**	**12**	**6**	**4**

Solution.

Age (in years)	Mid-value x	Frequency f	$u = \frac{x-45}{10}$	fu	x – M	$f(x-M)^2$
10–20	15	4	– 3	– 12	– 30	3600
20–30	25	8	– 2	– 16	– 20	3200
30–40	35	10	– 1	– 8	– 10	800
40–50	45	16	0	0	0	0
50–60	55	12	1	12	10	1200
60–70	65	6	2	12	20	2400
70–80	75	4	3	12	30	3600
Total		$\Sigma f = N = 60$		$\Sigma fu = 0$		$14800 = \Sigma f(x-M)^2$

Let the assumed mean, $A = 45$

Size of class, $i = 10$.

Arithmetic mean, $M = A + \frac{\Sigma fu}{N} \times i$

$$= 45 + \frac{0}{60} \times 10 = 45.$$

Standard deviation σ is given by

$$\sigma = \sqrt{\left[\frac{1}{N}\Sigma f(x-M)^2\right]} = \sqrt{\frac{14800}{60}} = \sqrt{246 \cdot 06} = 15 \cdot 71.$$

EXERCISE 5 (B)

1. (a) Compute the standard deviation for the set of numbers 3, 4, 9, 11, 13, 6, 8, 10.
   (b) Establish the formulae to find arithmetic mean and standard deviation from an assumed mean and compute them for the following series :

Mid-Value (in Rs.)	:	6	8	10	12	14	16	18
Frequency	:	1	4	8	15	14	6	2

2. Compute the standard deviation of monthly income of a worker from the following :

Month	:	1	2	3	4	5	6	7	8	9	10
Monthly income (in Rs.)	:	145	148	150	154	155	155	156	156	159	162

3. Calculate the standard deviation for the following frequency :

x	:	25	35	45	55	65	75	85
f	:	3	61	132	153	140	51	2

4. Calculate the mean, standard deviation, median and mode for the following data :

Mid Value	:	15	20	25	30	35	40	45	50	55
Frequency	:	2	22	19	14	3	4	6	1	1

5. Calculate the standard deviation for the following data :

x	:	0	10	20	30	40	50	60	70	80
f	:	150	140	100	80	80	70	30	14	0

6. Calculate the standard deviation for the following distribution :

Measure	:	5	10	15	20	25	30	35
Frequency	:	2	7	11	15	10	4	1

7. Calculate the standard deviation for the following frequency distribution :

x	:	2	3	4	5	9	10	12	13	15
f	:	25	37	44	59	68	43	30	32	12

8. Calculate the mean and standard deviation for the following series :

Age (Less than)	:	10	20	30	40	50	60	70	80
No. of persons	:	15	30	53	75	100	110	115	125

9. Goals scored by two teams A and B in football season were as follows :

No. of goals scored in a match	0	1	2	3	4
Team A :	27	9	8	5	4
Team B :	17	9	6	5	4

Which team may be considered the more consistent ?

10. The scores of two batsman in different innings are as given below :

A :	12	115	6	73	7	19	119	36	84	29
B :	47	12	76	42	4	51	37	48	13	0

Which of the two is more scorer ? Which of the two is more consistent ?

ANSWERS

1. (a) 3·3166 (b) $M = 12{\cdot}52, \sigma = 2{\cdot}6$

3. 4·8

4. $M = 27{\cdot}8473, \sigma = 8{\cdot}85, M_d = 25{\cdot}65, M_0 = 21{\cdot}84$

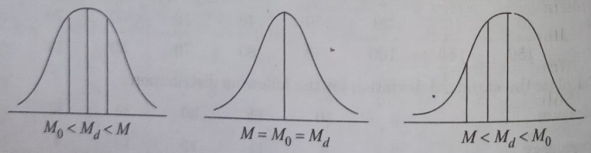
5. 19·577 **6.** 6·708 **7.** 3·76

9. *B*

10. Mean for $A = 50$
Mean for $B = 33$, *A* gets more run
C.V. for $A = 83{\cdot}66\%$, C.V. for $B = 70{\cdot}9\%$
B is more consistent.

§ 5.13. SKEWNESS

By skewness in same frequency distribution we mean the lack in symmetry. [*If the frequencies are symmetrically distributed about the mean, then the distribution is called* **symmetrical,** *in other words, a distribution is called symmetrical when the values equidistant from the mean have equal frequencies.*] Skewness is also termed as **asymmetry**. Skewness denotes the tendency of a distribution to depart from symmetry. According to Simpson, "Skewness or asymmetry is the attribute of a frequency distribution that extends further on one side of the class with the highest frequency that on the other".



We know that for a symmetrical distribution the mean, median and mode coincide. Therefore, skewness in a distribution is shown when these three averages do not coincide. Skewness indicates that the frequency curve has a longer tail on one side of the average. When the frequency curve has a longer tail on right side, the skewness is called positive. When the frequency curve has a longer tail on left side, the skewness is called negative. In other words, the skewness is positive if $M_0 < M_d < M$ and negative if $M < M_d < M_0$, where M, M_d and M_0 are mean, median and mode respectively.

Measure of Skewness :

There are two measures to measure the skewness.

(i) First coefficient of skewness. It is also known as Bowley's coefficient of skewness and is defined as

$$\textbf{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = J_Q$$

where Q_1 and Q_3 are lower and upper quartiles respectively and M_d is median. Clearly this measure is based on the fact that in a skew curve, the median does not lie half way between Q_1 and Q_3. This formula for coefficient of skewness is used when mode is well defined.

(ii) Second coefficient of skewness. It is also called Karl Pearson's coefficient of skewness and is defined as

$$\textbf{Coeff. of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{M - M_0}{\sigma} = J.$$

If mode is not well defined, then

$$\textbf{Coeff. of skewness} = \frac{3(M - M_d)}{\sigma}$$

[using empirical relation : $M - M_0 = 3(M - M_d)$]

Clearly this measure is based on the fact that mean and mode are not coincident.

Note that both of the above coefficients are pure numbers since both the numerator and denominator have the same dimensions.

First coefficient lies between – 1 and 1 and that second coefficient lies between – 3 and 3.

Note. Bowley's coefficient of skewness is also called Quartile coefficient of skewness.

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute the Bowley's coefficient of skewness for the following frequency distribution :***

Marks :	0–10	10–20	20–30	30–40	40–50
No of students :	2	7	10	5	3

Solution.

Marks	Frequency	C. F.
0—10	2	2
10—20	7	9
20—30	10	19
30—40	5	24
40—50	3	27

Lower quartile, $Q_1 = N/4$th term $= \frac{27}{4}$th term $= 6{\cdot}75$th term.

$$\therefore \quad Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i = 10 + \frac{6{\cdot}75 - 2}{7} \times 10$$

$$= 10 + \frac{4{\cdot}75 \times 10}{7} = 10 + 6{\cdot}79 = 16{\cdot}79 \text{ marks.}$$

Median, $M_d = N/2$th term $= 13{\cdot}5$th term

$$\therefore \quad M_d = l\frac{\frac{1}{2}N - F}{f} \times i = 20 + \frac{13{\cdot}5 - 9}{10} \times 10$$

$$= 20 + 4{\cdot}5 = 24{\cdot}5 \text{ marks}$$

$Q_3 = 3N/4$th term $= 20{\cdot}25$th term.

$$\therefore \quad Q_3 = l + \frac{\frac{3}{4}N - F}{f} \times i = 30 + \frac{20{\cdot}25 - 19}{5} \times 10$$

$$= 30 + 2{\cdot}5 = 32{\cdot}5 \text{ marks.}$$

$\therefore$ Bowley's coefficient of skewness

$$J_Q = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{32{\cdot}5 + 16{\cdot}79 - 2 \times 24{\cdot}5}{32{\cdot}5 - 16{\cdot}79} = \frac{0{\cdot}29}{15{\cdot}79} = 0{\cdot}02. \qquad \text{Ans.}$$

Example 2. ***Compute the mean, mode, standard deviation and coefficient of skewness for the following data :***

Age Below :	10	20	30	40	50	60
No. of persons :	15	32	51	78	97	109

Solution. Arranging the given data in class intervals, the calculation table is :

Year	Mid-Value x	f	$u = \frac{x - 35}{10}$	fu	fu^2
0—10	5	15	−2	−45	135
10—20	15	17	−2	−34	68
20—30	25	19	−1	−19	19
30—40	35	27	0	0	0
40—50	45	19	1	19	19
50—60	55	12	2	24	48
Total		$N = 109$		$\Sigma fu = -55$	$\Sigma fu^2 = 289$

Let the assumed mean $A = 35$ and $i = 10$.

$$\therefore \quad \text{Mean } (M) = A + \frac{\Sigma fu}{N} \times i$$

$$= 35 + \frac{-55}{109} \times 10 = 29{\cdot}95.$$

Standard deviation σ is given by

$$\sigma = i\sqrt{\left\{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2\right\}} \qquad \text{[See § 5.13]}$$

$$= 10 \times \sqrt{\left\{\frac{289}{109} - \left(\frac{55}{109}\right)^2\right\}} = 15{\cdot}49.$$

Mode $\quad (M_0) = l + \frac{f_1}{f_{-1} + f_1} \times i = 30 + \frac{19}{19 + 19} \times 10$

The coefficient of skewness is $= \frac{M - M_0}{\sigma} = \frac{29{\cdot}95 - 35}{15{\cdot}49} = -0{\cdot}32.$ **Ans.**

Example 3. ***Compute the quartiles from the following data and then compute the Bowlay's coefficient of skewness***

Wages (in Rs.) :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Number of workers :	20	45	85	160	70	55	35	30

Solution. The calculation table is :

Wages (in Rs.)	Frequency (f)	Cumulative Frequency (C.F.)	
0—10	20	20	
10—20	45	65	
20—30	85	150	$\leftarrow Q_1$
30—40	160	310	$\leftarrow M_d$
40—50	70	380	$\leftarrow Q_3$
50—60	55	435	
60—70	35	470	
70—80	30	500	
Total	$N = 500$		

Here $N = 500$ and class interval $i = 10$.

(i) For lower quartile Q_1, we have

$$\frac{1}{4}N = \frac{500}{4} = 125,$$

which clearly lies in 20—30. Thus 20—30 is the lower quartile class.

$$\therefore \quad Q_1 = l + \frac{\frac{1}{4}N - F}{f} \times i \qquad \text{[Here } F = 65,\ f = 85,\ i = 10\text{]}$$

$$= 20 + \frac{125 - 65}{85} \times 10 = 27 \cdot 059.$$

(ii) For median M_d

$$\frac{1}{2}N = \frac{500}{2} = 250,$$

which clearly lies in 30—40. Thus 30—40 is the median class.

$$\therefore \quad M_d = l + \frac{\frac{1}{2}N - F}{f} \times i \qquad \text{[Here } F = 150,\ f = 160,\ i = 10\text{]}$$

$$= 30 + \frac{250 - 150}{160} \times 10 = 36 \cdot 25.$$

(iii) For upper quartile Q_3

$$\frac{3N}{4} = \frac{3}{4} \times 500 = 375,$$

which clearly lies in 40—50. Thus 40—50 is the upper quartile class.

$$\therefore \quad Q_3 = l + \frac{\frac{3N}{4} - F}{f} \times i \qquad \text{[Here } F = 310,\ f = 70,\ i = 10\text{]}$$

$$= 40 + \frac{375 - 310}{70} \times 10 = 49 \cdot 286.$$

(iv) Bowley's coefficient of skewness

$$J_Q = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{49 \cdot 286 + 27 \cdot 059 - 2 \times 36 \cdot 25}{49 \cdot 286 - 27 \cdot 059} = \frac{3 \cdot 845}{22 \cdot 227} = 0 \cdot 173.$$

Example 4. ***Compute the coefficient of skewness for the following:***

Wages (in Rs.) :	45	55	65	75	85	95	105	115
No of persons :	35	40	48	100	125	87	43	22

Solution. The calculation table is as follows :

Wages	f	$\xi = x - A$ $(A = 7 \cdot 5)$	$f\xi$	$f\xi^2$
45	35	−3	−105	315
55	40	−2	−80	160
65	48	−1	−48	48
75	100	0	0	0
85	125	1	125	125
95	87	2	174	348
105	43	3	129	387
115	22	4	88	352
Total	$N = 500$		$\Sigma fu = 283$	1735

Let the assumed mean $A = 7 \cdot 5$.

$$\therefore \quad \text{Mean } (M) = A + \frac{\Sigma f\xi}{N} = 7 \cdot 5 + \frac{283}{500}$$

$$= 7 \cdot 5 + 0 \cdot 566 = 8 \cdot 066.$$

Since the value of variate, corresponding to maximum frequency 125, is 85.

Hence mode $(M_0) = 8 \cdot 5$.

The standard deviation σ is given by

$$\sigma = \sqrt{\left\{\frac{\Sigma f\xi^2}{N} - \left(\frac{\Sigma f\xi}{N}\right)^2\right\}} = \sqrt{\left\{\frac{1735}{500} - \left(\frac{283}{500}\right)^2\right\}}$$

$$= \sqrt{(3 \cdot 47 - 0 \cdot 32)} = \sqrt{(3 \cdot 15)} = 1 \cdot 77.$$

$$\therefore \text{ Coefficient of skewness is } = \frac{M - M_0}{\sigma} = \frac{8 \cdot 066 - 8 \cdot 5}{1 \cdot 77} = \frac{-0 \cdot 434}{1 \cdot 77} = -0 \cdot 245.$$

Example 5. ***Find the coefficient of skewness for the following distribution :***

Variate	Frequency	Variate	Frequency
0—5	2	20—25	21
5—10	5	25—30	16
10—15	7	30—35	8
15—20	13	35—40	3

Solution. The calculation table is :

Class	Mid-Value x	f	$\xi = x - A$ $(A = 17 \cdot 5)$	$f\xi$	$f\xi^2$
0—5	2·5	2	−15	−30	450
5—10	7·5	5	−10	−50	500
10—15	12·5	7	−5	−35	175
15—20	17·5	13	0	0	0
20—25	22·5	21	5	105	525
25—30	27·5	16	10	160	1600
30—35	32·5	8	15	120	1800
35—40	37·5	3	20	60	1200
Total		$N = 75$		330	6250

Let the assumed mean $A = 17 \cdot 5$.

$$\therefore \quad \text{Mean } (M) = A + \frac{\Sigma f\xi}{N} = 17 \cdot 5 + \frac{330}{75} = 17 \cdot 5 + 4 \cdot 4 = 21 \cdot 9.$$

Clearly modal class is 20–25. Thus

$$l = 20, \ f_{-1} = 13, \ f = 21, \ f_1 = 16, \ i = 5.$$

$$\therefore \quad \text{Mode } (M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 20 + \frac{21 - 13}{2 \times 21 - 13 - 16} \times 5 = 20 + \frac{8}{13} \times 5$$

$$= 20 + \frac{40}{13} = 20 + 3 \cdot 08 = 23 \cdot 08.$$

$$\text{S. D. } (\sigma) = \sqrt{\left\{ \frac{\Sigma f\xi^2}{N} - \left(\frac{\Sigma f\xi}{N} \right)^2 \right\}}$$

$$= \left\{ \frac{6250}{75} - \left(\frac{330}{75} \right)^2 \right\} = \sqrt{(83 \cdot 33 - 19 \cdot 36)} = \sqrt{(63 \cdot 95)} = 8 \text{ (approx).}$$

$$\therefore \quad \text{Coefficient of skewness} = \frac{M - M_0}{\sigma} = \frac{21 \cdot 9 - 23 \cdot 08}{8}$$

$$= -\frac{1 \cdot 18}{8} = -0 \cdot 147.$$

Example 6. ***Calculate the coefficient of skewness for the following distribution :***

Marks above	:	0	10	20	30	40	50	60	70	80
No. of students	:	150	140	100	80	80	70	30	14	0

Solution. The given series is cumulative frequency series. Also mode is ill-defined. Hence the alternative formula namely coefficient of skewness $= 3(M - M_d)/\sigma$ will be used

Marks	Mid-Value x	f	$u = \frac{x - A}{i}$ (A = 45, i = 10)	fu	fu^2	c.f.
0—10	5	10	−4	−40	160	10
10–20	15	40	−3	−120	360	50
20–30	25	20	−2	−40	80	70
30–40	35	0	−1	0	0	70
40–50	45	10	0	0	0	80
50–60	55	40	1	40	40	120
60–70	65	16	2	32	64	136
70–80	75	14	3	42	126	150
Total		$N = 150$		−86	830	–

Let the assumed mean $A = 45, \ i = 10$.

$$\therefore \quad \text{Mean } (M) = A + \frac{\Sigma fu}{N} \times i$$

$$= 45 + \frac{(-86)}{150} \times 10 = 45 - 5 \cdot 73 = 39 \cdot 27.$$

Since $N = 150$ (even). Hence $\frac{1}{2} N$th *i.e.*, $\frac{1}{2}$ (150)th *i.e.*, 75th term lies in class 40–50. Hence the median class is 40–50.

$$\therefore \quad \text{Median } (M_d) = l + \frac{\frac{1}{2} N - F}{f} \times i$$

$$= 40 + \frac{75 - 70}{10} \times 10 = 40 + 5 = 45.$$

$$\text{S.D. } (\sigma) = i \times \sqrt{\left\{ \frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N} \right)^2 \right\}}$$

$$= 10 \times \sqrt{\left\{ \frac{830}{150} - \left(\frac{-86}{150} \right)^2 \right\}}$$

$$= 10 \times \sqrt{(5 \cdot 33 - 0 \cdot 33)}$$

$$= 10 \times 2 \cdot 24 = 22 \cdot 4.$$

$$\text{Coeff. of skewness} = \frac{3(M - M_d)}{\sigma} = \frac{3(39 \cdot 27 - 45)}{22 \cdot 8}$$

$$= \frac{3 \times (-5 \cdot 73)}{22 \cdot 8} = \frac{-17 \cdot 19}{22 \cdot 8} = -0 \cdot 75.$$

Example 7. ***Calculate the Karl Pearson's coefficient of skewness for the following frequency distribution :***

Class	:	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	:	5	20	10	0	5	20	8	7

Solution. Clearly mode is ill-defined (since two classes have maximum frequency 20). So Karl Pearson's coeff. of skewness

$$= \frac{3 \text{ (Mean - Median)}}{\text{standard deviation}} = \frac{3(M - M_d)}{\sigma}.$$

Let assumed mean

$$A = 17 \cdot 5, \ i = 5, \ u = (x - A)/i.$$

Thus the calculation table is :

Class	f	Mid-Value x	$u = \frac{x - 17.5}{5}$	fu	fu^2	c.f.
0–5	5	25	−3	−15	45	5
5–10	20	75	−2	−40	80	25
10–15	10	125	−1	−10	10	35
15–20	0	175	0	0	0	35
20–25	5	225	1	5	5	40
25–30	20	275	2	40	80	60
30–35	8	325	3	24	72	68
35–40	7	375	4	28	112	75
Total	75	–	–	32	404	–

$$\text{Mean } M = A + \frac{\Sigma fu}{\Sigma f} \times i$$

$$= 17.5 + \frac{32}{75} \times 0.5 = 17.5 + 2.13 = 19.63.$$

$$\text{Median } M_d = \frac{1}{2}N\text{th term} = 37.5\text{th term}$$

$$\therefore \quad M_d = l + \frac{\frac{1}{2}N - F}{f} \times i$$

$$= 20 + \frac{37.5 - 35}{5} \times 5 = 22.5$$

$$\text{S. D.,} \quad \sigma = \sqrt{\left[\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2\right]} \times i$$

$$= \left[\frac{404}{75} - \left(\frac{32}{75}\right)^2\right] \times 5$$

$$= \sqrt{5.387 - 0.182} \times 5$$

$$= \sqrt{(5.205)} \times 5 = 2.28 \times 5 = 11.4.$$

$\therefore$ Karl Pearson's coeff. of skewness $= 3(M - M_d)/\sigma$

$$= 3(19.63 - 22.5)/11.4 = -0.76.$$

Hence the given frequency distribution is negatively skew.

EXERCISE 5 (C)

1. Find the coefficient of skewness of the two groups given below and point out which distribution is more skew ?

Marks :	55–58	58–61	61–64	64–67	67–70
Group A :	12	17	33	18	11
Group B :	20	22	25	13	7

[**Hint :** Find Bowley's coefficient of skewness]

2. For the following table of marks obtained by ten candidates, find the coefficient of variation and skewness :

Statistics :	25	50	45	30	70	42	36	38	34	60
Mathematics :	10	70	50	20	95	55	42	60	48	80

[Vikram 1993]

3. Compute the coefficient of skewness from the following data :

Age at the birth of first child :	13	14	15	16	17	18	19	20	21	22	23	24	25
Number of married women :	37	162	343	390	256	433	161	355	65	85	49	46	40

[Vikram 1993]

4. Calculate arithmetic mean, standard deviation and skewness from the following data :

Measure :	3	4	5	6	7	8	9
Frequency :	3	7	22	60	85	32	8

Calculate the Karl Pearson's coefficient of skewness for the following data :

5.

Marks :	0–10	10–20	20–30	30–40	40–50	50–60
No. of Candidates :	5	12	18	38	20	7

6.

Class :	0–4	4–8	8–12	12–16	16–20	20–24
Frequency :	5	7	10	15	8	4

7.

Wages (in Rs.) :	70–80	80–90	90–100	100–110
No. of workers :	12	18	35	42
Wages (in Rs.) :	110–120	120–130	130–140	140–150
No. of workers :	50	45	20	8

8.

Class :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency :	30	40	50	48	24	162	132	14

9.

Height (above in inches) :	58	59	60	61	62	63	64	65	66	67
No. of persons :	100	98	90	76	57	35	20	11	4	0

10.

Class :	0–6	6–12	12–18	18–24	24–30	30–36
Frequency :	12	24	38			

11.

Cultivation (per acre) in quintals :	20–24	25–29	30–34	35–39	40–44	45–49
No. of Forms :	6	20	40	70	35	9

12.

Class	Frequency	Class	Frequency
0—10	10	40—50	16
10—20	12	50—60	14
20—30	18	60—70	5
30—40	25		

ANSWERS

1. $-0{\cdot}02, -0{\cdot}22, B$
2. Statistics : 305, 069 ; Mathematics ; 458, 006
3. $-0{\cdot}11$
4. $M = 6{\cdot}6, \sigma = 1{\cdot}7, J = -0{\cdot}24$
5. $J = -0{\cdot}24$
6. $J = -0{\cdot}26$
7. $J = -0{\cdot}326$
8. $J = -0{\cdot}6$
9. $J = -0{\cdot}48$
10. $J = -0{\cdot}036$
11. $J = -0{\cdot}19$
12. $J = -0{\cdot}04$

◈ § 5.14. MOMENTS

Definition. For any frequency distribution the nth moment about any point A, is defined as the arithmetic mean of the nth powers of the deviations from the point A and is denoted by μ'_n. Thus

$$\mu'_n = \frac{1}{N}\Sigma f(x-A)^n \quad \text{where } N = \Sigma f.$$

The nth moment about the arithmetic mean M is denoted by μ_n and is defined by

$$\mu_n = \frac{1}{N}\Sigma\, f(x-M)^n$$

μ_n is also known as nth central moment.

Let

x	:	x_1	x_2	...	x_n
f	:	f_1	f_2	...	f_n

be a discrete frequency distribution.

For $n = 0, 1, 2$, we have (from above definitions)

$$\mu_0' = \mu_0 = \frac{1}{N}\Sigma f = 1$$

$$\mu_1' = \frac{1}{N}\Sigma f(x-A) = \frac{1}{N}\Sigma\, fx - A = M - A = d \text{ (say)}$$

$$\mu_1 = \frac{1}{N}\Sigma f(x-M) = \frac{1}{N}\Sigma\, fx - M = 0$$

$$\mu_2' = \frac{1}{N}\Sigma f(x-A)^2 = \frac{1}{N}\Sigma\, f(x-M+M-A)^2$$

$$= \frac{1}{N}\Sigma\, f(x-M)^2 + (M-A)^2 = \sigma^2 + d^2$$

$$\mu_2 = \sigma^2.$$

◈ § 5.15. RELATION BETWEEN CENTRAL MOMENTS AND MOMENTS ABOUT ANY POINT

We know that

$$\mu_n = \frac{1}{N}\Sigma f(x-M)^n$$

$$= \frac{1}{N}\Sigma f(x-A-\overline{M-A})^n$$

$$= \frac{1}{N}\Sigma f\{(x-A)-\mu_1'\}^n \qquad [\because \mu_1' = M - A]$$

By binomial theorem, we have

$$\{(x-A)-\mu'\}^n = (x-A)^n - {}^nC_1(x-A)^{n-1}\mu'_1 + \ldots + (-1)^n(\mu'_1)^n$$

$$\therefore \mu_n = \frac{1}{N}\Sigma f(x-A)^n - {}^nC_1\mu'_1\frac{1}{N}\Sigma f(x-A)^{n-1} + \ldots + (-1)^n(\mu'_1)^n\frac{1}{N}\Sigma f$$

or $$\mu_n = \mu'_n - {}^nC_1\mu_1\mu'_{n-1} + {}^nC_2(\mu'_1)^2\mu'_{n-2} - \ldots + (-1)^n(\mu'_1)^n. \quad \ldots(1)$$

Putting $n = 2,\ 3,\ 4$ in (1), we get

$$\mu_2 = \mu'_2 - 2\mu'_1\mu'_1 + (\mu'_1)^2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 3(\mu'_1)^2\mu'_1 - (\mu'_1)^3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 4(\mu'_1)^3\mu'_1 + (\mu'_1)^4$$

$$= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4.$$

Also we can prove that

$$\mu'_n = \mu_n + {}^nC_1\mu_{n-1}\mu'_1 + {}^nC_2\mu_{n-2}(\mu'_1)^2 + \ldots + (\mu'_1)^n. \quad \ldots(2)$$

Putting $n = 2, 3, 4$ in (2), we get

$$\mu'_2 = \mu_2 + 2\mu'_1\mu_1 + (\mu'_1)^2 = \mu_2 + (\mu'_1)^2 \qquad [\because \mu_1 = 0]$$

$$\mu'_3 = \mu_3 + 3\mu'_1\mu_2 + 3(\mu'_1)^2\mu_1 + (\mu'_1)^3$$

$$= \mu_2 + 3\mu'_1\mu_2 + (\mu'_1)^3 \qquad [\because \mu_1 = 0]$$

$$\mu'_4 = \mu_4 + 4\mu'_1\mu_3 + 6(\mu'_1)^2\mu_2 + 4(\mu'_1)^3\mu_1 + (\mu'_1)^4$$

$$= \mu_4 + 4\mu_3\mu'_1 + 6\mu_2(\mu'_1)^2 + (\mu'_1)^4$$

◈ § 5.16. EFFECT OF CHANGE OF ORIGIN AND SCALE ON MOMENTS

Let u be a new variable which is connected with x by the following relation

$$u = \frac{x-A}{h} \quad i.e.,\ x - A = hu.$$

$\therefore\ \bar{x} - A = h\,\bar{u}$ where $\bar{x}$ and $\bar{u}$ are the means of x-series and u-series respectively.

$\therefore\quad x - \bar{x} = h\,(u - \bar{u}).$

$\therefore\quad \mu'_n = \frac{1}{N}\Sigma f(x-A)^n = \frac{1}{N}\Sigma f h^n u^n = h^n\frac{1}{N}\Sigma f u^n$

and $$\mu_n = \frac{1}{n}\Sigma f(x-\bar{x})^n = h^n\frac{1}{N}\Sigma f(u-\bar{u})^n.$$

Hence nth moment of u-variate is h^n times the nth moment of u-variate.

§ 5.17. SHEPPARD'S CORRECTION OF MOMENT

When moments are calculated in case of class-intervals, it is supposed that the frequency of each class-interval is centred at the mid-point of the respective class. Hence there is possibility of some error in the values of the moments. To remove these errors, W. F. Sheppard established some formulae. These are called **Sheppard's Correction** and are given below :

$$\mu_1 \text{ (Corrected)} = \mu_1 = 0 \text{ (no need of correction)}$$

$$\mu_2 \text{ (Corrected)} = \mu_2 - \frac{1}{12}h^2$$

$$\mu_3 \text{ (Corrected)} = \mu_3 \text{ (no correction)}$$

$$\mu_4 \text{ (Corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$$

where h is the width of the class-intervals.

Remark. If σ_1^2 is the value of variance which is obtained from the grouped data and σ^2 be the corrected value, then

$$\sigma^2 = \sigma_1^2 - \frac{1}{12}h^2.$$

Example 1. *The values of second and fourth moments are :*

$$\mu_2 = 88{\cdot}75,\ \mu_3 = -131{\cdot}25 \text{ and } \mu_4 = 25445{\cdot}3125.$$

Calculate the corrected moments when the class-interval is 10.

Solution. (i) μ_2 (corrected) $= \mu_2 - \frac{h^2}{12} = 88{\cdot}75 - \frac{10^2}{12}$

$$= 88{\cdot}75 - 8{\cdot}33 = 80{\cdot}42.$$

(ii) μ_3 (corrected) $= -131{\cdot}25$ (No correction).

(iii) μ_4 (corrected) $= \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$

$$= 25445{\cdot}3125 - \frac{88{\cdot}75 \times 100}{2} + \frac{7 \times 10000}{240}$$

$$= 25445{\cdot}3125 - 4437{\cdot}5 + 291{\cdot}667 = 21299{\cdot}4785.$$

Example 2. *Show that if a range of six times the standard deviation covers at least* 18 *class-intervals. Sheppard's correction will make a difference of less than* 0·5 *percent in the incorrected value of the standard deviation.*

Solution. Let h be the width of the class-interval. Let σ'^2 be the corrected variance and σ^2 the variance computed from the grouped data. Then according to the given conditions :

$$6\sigma \geq 18h \Rightarrow h \leq \frac{1}{3}\sigma. \qquad \ldots(1)$$

From Sheppard's correction, we have

$$\sigma'^2 = \sigma^2 - (1/12)h^2$$

$$\Rightarrow \quad \sigma'^2 \geq \sigma^2 - (1/108)\sigma^2 \qquad \text{[use (1)]}$$

$$\Rightarrow \quad \sigma' \geq \sigma\left(1 - \frac{1}{108}\right)^{1/2}$$

$$\Rightarrow \quad \sigma' \geq \sigma\left(1 - \frac{1}{216}\right) \qquad \text{[Neglecting other terms]}$$

$$\therefore \quad \sigma - \sigma' \leq \frac{\sigma}{216} < \frac{\sigma}{200}$$

$$\Rightarrow \quad \sigma - \sigma' < (0{\cdot}5/100)\,\sigma \Rightarrow \sigma - \sigma' < 0{\cdot}5\% \text{ of } \sigma.$$

§ 5.18. CHARLIER'S ACCURACY CHECK

Charlier gave following identities which are found useful in checking the values of Σfx, Σfx^2, mean, σ, moments etc.

$$\Sigma f(x+1) = \Sigma fx + \Sigma f = \Sigma fx + N$$

$$\Sigma f(x+1)^2 = \Sigma fx^2 + 2\,\Sigma fx + N$$

$$\Sigma f(x+1)^3 = \Sigma fx^3 + 3\,\Sigma fx^2 + 3\,\Sigma\, fx + N$$

$$\Sigma f(x+1)^4 = \Sigma fx^4 + 4\Sigma\, fx^3 + 6\,\Sigma fx^2 + 4\,\Sigma fx + N$$

Thus to calculate Σfx^n, find $\Sigma f(x+1)^n$.

§ 5.19. KARL PEARSON'S α, β AND γ COEFFICIENTS

Karl Pearson gave the following eight coefficients calculated from the moments, which are defined as :

Alpha Coefficients	Beta Coefficients	Gamma Coefficients
$\sigma_1 = \frac{\mu_1}{\sigma} = 0$	$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$	$\gamma_1 = \pm\sqrt{\beta_1}$ $\gamma_2 = \beta_2 - 3$
$\sigma_2 = \frac{\mu_2}{\sigma^2} = 0$	$\beta_2 = \frac{\mu_4}{\mu_2^2}$	$= \frac{\mu_4 - 3\mu_2^2}{\mu_2^3}$
$\sigma_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$		
$\sigma_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$		

The sign of γ_1 depends on μ_3. If μ_3 is positive, γ_1 is positive. If μ_3 is negative, γ_1 is negative.

§ 5.20. COEFFICIENTS OF SKEWNESS BASED ON MOMENTS

When there is symmetrical distribution, all the moments of odd order about the arithmetic mean (*i.e.*, μ_1, μ_3, μ_5 etc.) vanish. If the values of these coefficients do not vanish then there is skewness in the frequency distribution.

According to Karl Pearson the coefficients of skewness are exactly given by the following formulae :

First Coefficient of Skewness $= \dfrac{\mu_2}{\sqrt{\mu_2^3}} = \sqrt{\beta_1} = \alpha_3 = \gamma_1$.

Second Coefficient of Skewness $= \dfrac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\beta_2 - 6\beta_1 - 9)}$. **[Jiwaji 1990; Bhopal 90]**

If skewness in the series is very small then second coefficient of skewness should be used.

◈ § 5.21. KURTOSIS

In Greek language kurtosis means 'bulgines'. Kurtosis indicates the nature of the vertex of the curve. Several Statisticians defined kurtosis. Some of these definitions are :

"In statistics, kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve".

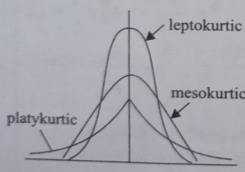
"A measure of kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat-topped".

Karl Pearson in 1905 defined following three types of curves :

1. Normal Curve or Mesokurtic Curve. A curve which is neither flat nor peaked is called a normal curve or meso-kurtic curve. For such type of curve we have $\beta_2 = 3$ and $\gamma_2 = 0$.

2. Leptokurtic Curve. A curve which is more peaked than the normal curve is called leptokurtic curve. For such type of curve, we have $\beta_2 > 3$ and $\gamma_2 > 0$.

3. Platykurtic Curve. A Curve which is more flatter than the normal curve is called platykurtic curve. For such type of curve, we have $\beta_2 < 3$ and $\gamma_2 < 0$.



Measure of Kurtosis :

Second and fourth moments are used to measure kurtosis. Karl Pearson gave the following formula to measure kurtosis :

Kurtosis or $\beta_2 = \mu_4 / \mu_2^2$.

To measure kurtosis, γ_2 is used and it is given by the following formula :

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2}.$$

Deductions. (1) If $\gamma_2 = 0$, the curve is *normal*.

(2) If $\gamma_2 > 0$, the curve is *leptokurtic*.

(3) If $\gamma_2 < 0$, the curve is *platykurtic*.

ILLUSTRATIVE EXAMPLES

Example 1. (a) ***For any frequency distribution, prove that :***

(i) $\beta_2 - \beta_1 - 1 \geq 0$ ***or*** $\beta_2 \geq \beta_1 + 1$.

(ii) $\beta_2 \geq \beta_1$.

(iii) $\beta_2 \geq 1$.

Solution. Let the given frequency distribution be

x :	x_1	x_2	...	x_n
f :	f_1	f_2	...	f_n

So that $\Sigma f = N$. If $\bar{x}$ be the A. M. then

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}.$$

$\therefore$ Deviation $u_i = x_i - \bar{x}$.

$$\mu_r = \frac{\sum_{i=1}^{n} f_i u_i^r}{N}, \quad r = 1, 2, 3, 4.$$

If a, b, c are any constant and $u_i = x_i - \bar{x}$, then we know that the following relation always holds :

$$\Sigma f_i (a u_i^2 + b u_i + c)^2 \geq 0$$

$$a^2\,\Sigma f u_i^4 + b^2\,\Sigma f_i u_i^2 + c^2\,\Sigma f_i + 2ab\,\Sigma f_i u_i^3 + 2ac\,\Sigma f_i u_i^2 + 2bc\,\Sigma f_i u_i \geq 0$$

$$\Rightarrow \quad a^2 \cdot \frac{\Sigma f_i u_i^4}{N} + b^2 \cdot \frac{\Sigma f_i u_i^2}{N} + c^2 \cdot \frac{\Sigma f_i}{N} + 2ab \cdot \frac{\Sigma f_i u_i^3}{N} + 2ac \cdot \frac{\Sigma f_i u_i^2}{N} + 2bc \cdot \frac{\Sigma f_i u_i}{N} \geq 0$$

$$\Rightarrow \quad a^2 \cdot \mu_4 + b^2 \cdot \mu_2 + c^2 \cdot \mu_0 + 2ab \cdot \mu_3 + 2ac \cdot mu_2 + 2bc \cdot \mu_1 \geq 0.$$

Therefore, by the property of 2nd order homogeneous equation, we have :

$$\begin{vmatrix} \mu_4 & \mu_3 & \mu_2 \\ \mu_3 & \mu_2 & \mu_1 \\ \mu_2 & \mu_1 & \mu_0 \end{vmatrix} \geq 0$$

$$\Rightarrow \quad \begin{vmatrix} \mu_4 & \mu_3 & \mu_2 \\ \mu_3 & \mu_2 & 0 \\ 0 & 0 & 1 \end{vmatrix} \geq 0 \qquad [\because\ \mu_1 = 0,\ \mu_0 = 1]$$

$$\Rightarrow \quad \mu_4 (\mu_2 - 0) - \mu_3 (\mu_3 - 0) + \mu_2 (0 - \mu_2^2) \geq 0$$

$$\Rightarrow \quad \mu_4 \mu_2 - \mu_3^2 - \mu_2^3 \geq 0 \quad \Rightarrow \quad \frac{\mu_4}{\mu_2^2} - \frac{\mu_3^2}{\mu_2^3} - 1 \geq 0$$

$$\therefore \quad \beta_2 - \beta_1 - 1 \geq 0 \qquad \left[\because \beta_1 = \frac{\mu_3^2}{\mu_2^3} \geq 0\right]$$

$\Rightarrow$ (i) $\beta_2 \geq \beta_1 + 1$

(ii) $\beta_2 > \beta_1$.

(iii) $\beta_2 \geq 1$.

Example 1. (b) ***Show that for discrete distribution*** $\beta_2 > 1$.

Solution. We have that β_2 is given by

$$\beta_2 = \mu_4 / \mu_2^2.$$

Now we are to prove that $\beta_2 > 1$

i.e., $\mu_4 / \mu_2^2 > 1$, *i.e.*, $\mu_4 > \mu_2^2$

i.e., $$\frac{1}{N}\Sigma f(x-M)^4 > \left\{\frac{1}{N}\Sigma f(x-M)^2\right\}$$

where M is the mean of the given series and $N = \Sigma f$

i.e., $$\frac{1}{N}\Sigma f y^4 > \left\{\frac{1}{N}\Sigma f y^2\right\}^2, \text{ where } x - M = y$$

i.e., if $N\,\Sigma f y^4 > \{\Sigma f y^2\}^2$

i.e., if $(f_1 + f_2 + \ldots + f_n)(f_1 y_1^4 + f_2 y_2^4 + \ldots + f_n y_n^4)$

$$> (f_1 y_1^2 + f_2 y_2^2 + \ldots + f_n y_n^2)^2$$

i.e., if $f_1 f_2 (y_1^2 - y_2^2)^2 + f_1 f_3 (y_1^2 - y_3^2)^2 + \ldots > 0$

which is true since $f_1, f_2, \ldots$ are essentially positive and the terms within the brackets are squared. Hence the expression on the left hand side is essentially positive. Consequently $\beta_2 > 1$.

Example 2. (a) ***Calculate the first and second moments about zero for the observations :***

3, 8, 11, 12, 20.

Solution.

						Total
x :	3	8	11	12	20	54
x^2 :	9	64	121	144	400	738

$\therefore$ First moment about zero,

$$\mu_1' = \frac{1}{n}\Sigma x = \frac{54}{5} = 10 \cdot 8$$

and second moment about zero,

$$\mu_2' = \frac{1}{n}\Sigma x^2 = \frac{738}{5} = 147 \cdot 6.$$

Example 2. (b) ***Calculate*** μ_1, μ_2 ***and*** μ_3 ***from the following series :***

x :	0	1	2	3	4	5	6	7	8
f :	1	9	26	59	72	52	29	7	1

Solution. Let the assumed mean $A = 4$. We have

x	f	x − A	f(x − A)	$f(x-A)^2$	$f(x-A)^3$
0	1	−4	−4	16	−64
1	9	−3	−27	81	−243
2	26	−2	−52	104	−208
3	59	−1	−59	59	−59
4	72	0	0	0	0
5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
Total	$N = 256$		−7	507	−37

$$\therefore \quad M = A + \frac{\Sigma f(x-A)}{N} = 4 + \frac{(-7)}{256}\ 4 - 0 \cdot 021 = 3 \cdot 973$$

$$\mu'_1 = \frac{1}{N}\Sigma f(x-A) = \frac{-7}{256} = -0 \cdot 027$$

$$\mu'_2 = \frac{1}{N}\Sigma f(x-A)^2 = \frac{507}{226} = 1 \cdot 980$$

$$\mu'_3 = \frac{1}{N}\Sigma f(x-A)^3 = \frac{-37}{256} = -0 \cdot 145.$$

$$\therefore \quad \mu_1 = \frac{1}{N} = \Sigma f(x-M) = \frac{1}{N}\Sigma fx - M\frac{\Sigma f}{N} = M - M = 0$$

$$\mu_2 = \mu'_2 - \mu'^2_1 = 1 \cdot 980 - (-0 \cdot 027)^2 = 1 \cdot 980 - 0 \cdot 001 = 1 \cdot 979.$$

$$\mu_3 = \mu'_3 - 3\mu'_2\,\mu'_1 + 2\mu'^3_1$$

$$= -0 \cdot 145 - 3(1 \cdot 980) \cdot (-0 \cdot 027) + 2(-0 \cdot 027)^3$$

$$= -0 \cdot 145 + 0 \cdot 160 - 0 \cdot 00004$$

$$= 0 \cdot 0115.$$

Example 2. (c) ***Calculate the first four central moments from the following data :***

Class :	0–10	10–20	20–30	30–40	40–50
Frequency :	1	3	5	7	4

Solution. To facilitate the calculations, we take the origin at 25 and then unit is 10. Let $u = \dfrac{x - 25}{10}$.

Class	f	Mid-Value x	$u = \frac{x-25}{10}$	fu	fu^2	fu^3	fu^4
0–10	1	5	−2	−2	4	−8	16
10–20	3	15	−1	−3	3	−3	3
20–30	5	25	0	0	0	0	0
30–40	7	35	1	7	7	7	7
40–50	4	45	2	8	16	32	64
Total	$N = 20$	–	–	10	30	28	90

Here $a = 95$, $h = 10$.

Hence moments of variate u about the origin are :

$$m'_1 = \frac{\Sigma fu}{\Sigma f} = \frac{10}{20} = 0 \cdot 5$$

$$m'_2 = \frac{\Sigma fu^2}{\Sigma f} = \frac{30}{20} = 1 \cdot 5$$

$$m'_3 = \frac{\Sigma fu^3}{\Sigma f} = \frac{28}{20} = 1 \cdot 4$$

$$m'_4 = \frac{\Sigma fu^4}{\Sigma f} = \frac{90}{20} = 4 \cdot 5$$

From § 5.16, we know that if m_1, m_2, m_3, m_4 are central moments of the variate u then central moments of the variate x are hm_1, h^2m_2, h^3m_3 and h^4m_4. Hence

$$\mu_1 = 0$$

$$\mu_2 = m_2h^2 = \{m'_2 - (m'_1)^2\}h^2$$

$$= \{1 \cdot 5 - (0 \cdot 5)^2\} \times 10^2 = 1 \cdot 25 \times 100 = 125$$

$$\mu_3 = m_3h^3 = [m'_3 - 3m'_2\, m'_1 + 2(m'_1)^3] \times h^3$$

$$= [1 \cdot 4 - 3(1 \cdot 5)(0 \cdot 3) + 2(0 \cdot 5)^3] \times 10^3$$

$$= [1 \cdot 4 - 2 \cdot 25 + 0 \cdot 250] \times 1000 = -0 \cdot 6 \times 1000 = -600$$

$$\mu_4 = m_4h^4 = [m'_4 - 4m'_3\, m'_1 + 6m'_2\,(m'_1)^2 - 3(m'_1)^4] \times h^4$$

$$= [4 \cdot 5 - 4(1 \cdot 4)(0 \cdot 5) + 6(1 \cdot 5)(0 \cdot 5)^2 - 3(0 \cdot 5)^4] \times 10^4$$

$$= [4 \cdot 5 - 2 \cdot 8 + 2 \cdot 25 - 0 \cdot 1875] \times 1000 = 3 \cdot 7625 \times 1000 = 37625.$$

Example 3. ***Calculate the first four moments about the mean of the distribution; also calculate*** β_1 ***and*** β_2***, x-values in centimeters are the mid-values of intervals :***

x:	20	25	30	35	40	45	50
f:	5	38	65	92	70	40	10

Solution. To facilitate the calculation, we take the origin at 35 and then unit is 05. Let $\xi = \frac{x - 3 \cdot 5}{0 \cdot 5}$.

x	f	ξ	$f\xi$	$f\xi^2$	$f\xi^3$	$f\xi^4$
20	5	−3	−15	45	−135	405
25	38	−2	−76	152	−304	608
30	65	−1	−65	65	−65	65
35	92	0	0	0	0	0
40	70	1	70	70	70	70
45	40	2	80	160	320	640
50	10	3	30	90	270	810
Total	$N = 320$		24	582	156	2598

In the changed scale, we have

$$\mu'_1 = \frac{\Sigma f\xi}{N} = \frac{24}{320} = 0 \cdot 075$$

$$\mu'_2 = \frac{\Sigma f\xi^2}{N} = \frac{582}{320} = 1 \cdot 81875$$

$$\mu'_3 = \frac{\Sigma f\xi^3}{N} = \frac{156}{320} = 0 \cdot 4875$$

$$\mu'_4 = \frac{\Sigma f\xi^4}{N} = \frac{2598}{320} = 8 \cdot 11875.$$

$$\therefore \quad \mu_2 = \mu'_2 - \mu'^2_1 = 1 \cdot 813125$$

$$\mu_3 = \mu'_3 - 3\mu'_2\,\mu'_1 + 2\mu'^3_1 = 0 \cdot 0791$$

$$\mu_4 = \mu'_4 - 4\mu'_3\,\mu'_1 + 6\mu'^2_1\,\mu'_2 - 3\mu'^4_1 = 8 \cdot 033$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0 \cdot 0791)^2}{(1 \cdot 8131)^3} = 0 \cdot 001$$

$$\beta_3 = \frac{\mu_4}{\mu_2^2} = \frac{(0 \cdot 033)}{(1 \cdot 8131)^2} = 2 \cdot 44.$$

From § 3.16, in the original scale, we have

$$\mu_2 = 1 \cdot 8131 \times (0 \cdot 5)^2 = 0 \cdot 4533 \qquad [\because\ h = 5]$$

$$\mu_3 = (0 \cdot 0791) \times (0 \cdot 5)^3 = 0 \cdot 00989$$

$$\mu_4 = (8 \cdot 033) \times (0 \cdot 5)^4 = 0 \cdot 5021.$$

β_1 and β_2 remain unaltered since unit portions (scale portions) in numerator and denominator cancel.

Example 4. ***The first four moments of a distribution about the value 4 of the variate are – 15, 17, – 30 and 108, calculate the moments about the mean.***

Calculate also the moments about the origin. Also find β_1 ***and*** β_2.

State whether the distribution is leptokurtic or platykurtic.

Solution. We are given that

$$\mu'_1 = \frac{1}{N}\Sigma f(x-4) = -1\cdot 5 \quad \ldots(1)$$

$$\mu'_2 = \frac{1}{N}\Sigma f(x-4)^2 = 17 \quad \ldots(2)$$

$$\mu'_3 = \frac{1}{N}\Sigma f(x-4)^3 = -30 \quad \ldots(3)$$

and

$$\mu'_4 = \frac{1}{N}\Sigma f(x-4)^4 = 108 \quad \ldots(4)$$

From (1), $\frac{1}{N}\Sigma f\,x-4 = -1\cdot 5$.

$\therefore$ Mean $= 4 - 1\cdot 5 = 2\cdot 5$

$\mu_1 = 0$ (always)

$\mu_2 = \mu'_2 - \mu'^2_1$

$= 17 - (-1\cdot 5)^2 = 17 - 2\cdot 25 = 14\cdot 75$

$\mu_3 = \mu'_3 - 3\mu'^2_2\,\mu'_1 + 2\mu'^3_1$

$= -30 - 3\times 17(-1\cdot 5) + 2(-1\cdot 5)^3 = 39\cdot 75$

$\mu_4 = \mu'_4 - 4\mu'_3\,\mu'_1 + 6\mu'_2\,\mu'^2_1 + 3\mu'^4_1$

$= 108 - 4\times(-30)\times(-1\cdot 5) + 6\times 17\times(-1\cdot 5)^2 - 3(-1\cdot 5)^4$

$= 142\cdot 3125.$

Let the first four moments about the origin be $\nu_1, \nu_2, \nu_3\ \nu_4$ respectively. Then we have

$\nu_1 = M - A = 2\cdot 5 - 0 = 2\cdot 5$

$\nu_2 = \mu_2 + d^2$, where $d = M - A$

$= 14\cdot 75 + (2\cdot 5)^2 = 14\cdot 75 + 6\cdot 25 = 21$

$\nu_3 = \mu_3 + 3\mu_2\,\nu_1^2 + \nu_1^3$ [$\because$ Here $\mu'_1 = \nu_1$]

$= 39\cdot 75 + 3\times 14\cdot 75\times 2\cdot 5 + (2\cdot 5)^3 = 166$

$\nu_4 = \nu_4 + 4\mu_3\,\nu_1 + 6\mu_1\,\nu_1^2 + \nu_1^4$

$= 142\cdot 3125 + 4(39\cdot 75)\times 2\cdot 5 + 6\times 14\cdot 75\times(2\cdot 5)^2 + (2\cdot 5)^4$

$= 142\cdot 3125 + 3\cdot 975 + 553\cdot 125 + 39\cdot 0625 = 1132.$

Now $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{(39\cdot 75)^2}{(14\cdot 75)^3} = 0\cdot 4924$

and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{142\cdot 3125}{(14\cdot 75)^2} = 0\cdot 6541 < 3.$

Here the curve (*i.e.*, distribution) is platykurtic.

Example 5. ***The first four moments of a distribution about the value 4 of the variable are*** $-1\cdot 5$, 17, -30 ***and*** 108. ***Do you feel that there is some error in their calculation ?***

Solution. To see the consistency of the given moments, we shall have to check the following three relations :

(a) $\mu_2 \geq 0$, (b) $\beta_2 > \beta_1$, (c) $\beta_2 \geq 1$.

From example 4 (above), we have

$\mu_2 = 14\cdot 75 > 0$

$\beta_2 = 0\cdot 6541$ and $\beta_1 = 0\cdot 4924$ $\therefore\ \beta_2 > \beta_1$

and $\beta_2 = 0\cdot 6541 < 1$.

Since the value of β_2 is less than 1, hence there is some error in the calculation of the given moments.

EXERCISE 5 (D)

1. For a frequency distribution, prove that
   (a) Kurtosis is greater than 1.
   [**Hint** : $\beta_2 > 1$].
   (b) The coefficient of skewness is numerically less than 1.
   [**Hint** : $\beta_1^2 < 1$].
   (c) $\beta_2 > \beta_1$.

2. (a) Calculate the 2nd moment about 15 for the following frequency distribution :

x :	5	10	15	20	25
f :	1	2	1	2	4

   (b) Calculate the first four central moments

x :	0	1	2	3	4	5	6
f :	15	38	55	82	60	40	10

   (c) Write down the values of μ_2 and μ_4 in terms of μ_1', μ_2', μ_3' and μ_4' in usual notations.
   Write also the formulae for β_1 and β_2.
   (d) Comment the following statement :
   "The shape of a distribution is known from its first four moments".

3. Calculate the measure of kurtosis for the following distribution :

Marks	3–15	15–25	25–35	35–45	45–55
No. of candidates	1	3	5	7	4

4. Find the first four moments about the mean for the following data :

Mid-Value :	1	2	3	4	5	6	7	8	9
Frequency :	1	6	13	25	30	22	9	5	2

5. The following table represents the height of a batch of 100 candidates. Calculate kurtosis :

Height (in cms.) :	59	61	63	65	67	69	71	73	75
No. of Candidates :	0	2	6	20	40	20	8	2	0

6. The first three moments of distribution about the value 2 of the variable are 1, 16 and – 40. Show that mean = 3, variance = 15, $\mu_3 = -86$. Show also that the moments about $x = 0$ are 3, 24 and 76.
7. The first three moments of a distribution about the value 3 of the variable are 2, 10 and – 30. Show that the first three moments about $x = 0$ are 5, 31 and 141. Show that mean = 5, variance = 6 and $\mu_3 = -74$ of the distribution.
8. The first four moments of a distribution about the value 5 of a variable are 2, 20, 40 and 50. Obtain the various characteristics, as far as possible, of this distribution on the basis of the information given.
9. The first four central moments of a distribution are 0, 2·5, 0·7 and 18·75. Calculate the skewness and kurtosis.
10. The first four moments of a distribution about the value 0 of a variable are – 0·20, – 1·76, – 2·36 and 1088. Find the moments about the mean. Measure also the kurtosis.
11. For a distribution mean is 10, variance is 16, γ is + 1 and the value of β_2 is 4. Find the first four moments about the origin.

ANSWERS

2. (a) 65
3. 1·67
4. $\mu_1 = 0, \mu_2 = 2{\cdot}448, \mu_3 = 6{\cdot}675, \mu_4 = 18{\cdot}334$
5. $\beta_2 = 3{\cdot}16$, curve is Leptokurtic.
8. $M = 7, \sigma = 4, \mu_1 = 0, \mu_2 = 16, \mu_3 = -64, \mu_4 = 162, \beta_1 = 1, \beta_2 = 0{\cdot}63, \gamma_1 = 1, \gamma_2 = 2{\cdot}37$.
9. 0·18, 3; Curve is Mesokurtic.
11. $\mu'_1 = 10, \mu'_2 = 116, \mu'_3 = 1544, \mu'_4 = 23184$.

ILLUSTRATIVE EXAMPLES ON SKEWNESS AND KURTOSIS

Example 1. ***Define the skewness and kurtosis of a distribution. The first four moments about the working mean 28·5 of a distribution are 0·294, 7·144, 42·409 and 454·98. Calculate the moments about the mean. Also evaluate β_1, β_2 and comment upon the skewness and kurtosis of the distribution.***

Solution. Given $\mu_1' = 0{\cdot}294, \mu_2' = 7{\cdot}144, \mu_3' = 42{\cdot}409, \mu_4' = 454{\cdot}98$.

To calculate moments about the mean.

$\mu_1 = 0$ (about)

$\mu_2 = \mu_1' - (\mu_1')^2 = 7{\cdot}144 - (0{\cdot}294)^2 = 7{\cdot}0576$

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

$= 42{\cdot}409 - 3 \times 7{\cdot}144 \times 0{\cdot}294 + 2(0{\cdot}294)^3 = 36{\cdot}1588$

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

$= 454{\cdot}98 - 4 \times 42{\cdot}409 \times 0{\cdot}294 + 6 \times 7{\cdot}144 \times (0{\cdot}294)^2 - 3(0{\cdot}294)^4$

$= 408{\cdot}7896$.

To calculate β_1, β_2

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(36{\cdot}1588)^2}{(7{\cdot}0576)^3} = 3{\cdot}7193$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{408{\cdot}7896}{(7{\cdot}0576)^2} = 8{\cdot}2070.$$

Now coefficient of skewness $= \sqrt{\beta_1} = \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \dfrac{36{\cdot}1588}{\sqrt{(7{\cdot}0576)^3}} = 1{\cdot}9285 = +\text{ve}.$

Since $\sqrt{\beta_1}$ is positive, therefore the distribution is positively skewed.

Again $\beta_2 = 8{\cdot}2070 > 3$ *i.e.*, kurtosis > 3, therefore the curve is Leptokurtic.

Example 2. ***The μ_2 and μ_4 for a distribution are found to be 4 and 48. Discuss the kurtosis of the distribution.***

Solution. Here $\mu_2 = 4, \mu_4 = 48$.

$$\therefore \quad \text{Kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{48}{(4)^2} = \frac{48}{16} = 3.$$

Since $\beta_2 = 3$, therefore the curve is Mesokurtic.

Example 3. ***The first four central moments of a distribution are 0, 1·5, 0·6 and 2·15. Test the kurtosis of the distribution.***

Solution. Here $\mu_1 = 0, \mu_2 = 1{\cdot}5, \mu_3 = 0{\cdot}6, \mu_4 = 2{\cdot}15$.

$$\therefore \quad \text{Kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2{\cdot}15}{(1{\cdot}5)^2} = 0{\cdot}9556 < 3.$$

Since $\beta_2 < 3$, hence the curve is Platykurtic.

Example 4. ***The standard deviation of a symmetric distribution is 4. What must be the value of μ_4 in order that the distribution be***

(i) Mesokurtic (ii) Leptokurtic (iii) Platykurtic.

Solution. Given standard deviation $\sigma = 4$.

We know that $\mu_2 = \sigma^2 \Rightarrow \mu_2 = 16$

Also $\text{kurtosis} = \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{\mu_4}{(16)^2} = \dfrac{\mu_4}{256}$

Now the distribution, will be

(i) *Mesokurtic* if $\beta_2 = 3$ *i.e.*, $\dfrac{\mu_4}{256} = 3$ *i.e.*, $\mu_4 = 768$

(ii) *Leptokurtic* if $\beta_2 > 3$ *i.e.*, $\dfrac{\mu_4}{256} > 3$ *i.e.*, $\mu_4 > 768$

(iii) *Platykurtic* if $\beta_2 < 3$ *i.e.*, $\dfrac{\mu_4}{256} < 3$ *i.e.*, $\mu_4 < 768$

Example 5. ***The fourth moment about mean of a frequency distribution is 768. What must be value of its standard deviation in order that the distribution be***

(i) Leptokurtic (ii) Mesokurtic (iii) Platykurtic.

Solution. Given $\mu_4 = 768$.

$$\therefore \quad \text{Kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = \frac{768}{\sigma^4} \qquad [\because \mu_2 = \sigma^2]$$

Now the distribution, will be

(i) *Leptokurtic if* $\beta_2 > 3 \Rightarrow \dfrac{768}{\sigma^4} > 3 \Rightarrow \sigma^4 < \dfrac{768}{3}$

$$\Rightarrow \sigma^4 < 256 \Rightarrow \sigma^4 < (4)^4 \Rightarrow \sigma < 4$$

(ii) *Mesokurtic if* $\beta_2 = 3 \Rightarrow \dfrac{768}{\sigma^4} = 3 \Rightarrow \sigma = 4$

(iii) *Platykurtic if* $\beta_2 < 3 \Rightarrow \dfrac{768}{\sigma^4} < 3 \Rightarrow \sigma > 4$.

Example 6. ***Find the measure of skewness and kurtosis based on moments for the following distribution and draw your conclusion :***

Marks	5–15	15–25	25–35	35–45	45–55
No. of Students	1	3	5	7	4

Solution. Let $A = 30$. Here $h = 10$. So let $u = \dfrac{x - A}{h} = \dfrac{x - 30}{10}$

Class	Frequency f	Mid-value x	$u = \frac{x-30}{10}$	fu	fu^2	fu^3	fu^4
5–15	1	10	−2	−2	4	−8	16
15–25	3	20	−1	−3	3	−3	3
25–35	5	30	0	0	0	0	0
35–45	7	40	1	7	7	7	7
45–55	4	50	2	8	16	32	74
	$N = \Sigma f = 20$			$\Sigma fu = 10$	$\Sigma fu^2 = 30$	$\Sigma fu^3 = 28$	$\Sigma fu^4 = 90$

The first four moments about $x = A = 30$

$$\mu_1' = \frac{h}{N} = \Sigma fu = \frac{10 \times 10}{20} = 5$$

$$\mu_2' = \frac{h^2}{N} = \Sigma fu^2 = \frac{100 \times 30}{20} = 150$$

$$\mu_3' = \frac{h^3}{N} = \Sigma fu^3 = \frac{1000 \times 28}{20} = 1400$$

$$\mu_4' = \frac{h^4}{N} = \Sigma fu^4 = \frac{1000 \times 90}{20} = 45000.$$

To Calculate Central Moments. $\mu_1 = 0$ (always)

$$\mu_2 = \mu_2' - (\mu_1')^2 = 150 - (5)^2 = 125$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = -600$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 37625.$$

Skewness. Moment Coefficient of skewness $= \gamma_1 = \dfrac{\mu_3}{\mu_2^3} = \dfrac{-600}{\sqrt{(125)^3}} = -0.4293.$

Kurtosis. $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3725}{(125)^2} = 2.408.$

Conclusion. Since γ_1 is negative, therefore the distribution is negatively skewed.

Since $\beta_2 < 3$, therefore the distribution is Platykurtic.

EXERCISE 5 (E)

1. Find the mean and variance for the distribution in which the values 1, 2, 3, ..., n, and the frequency of each is unity.
2. (a) Show that if the variable takes the values 0, 1, 2, ..., n with frequencies proportional to the binomial coefficients ${}^nC_0$, ${}^nC_1$, ${}^nC_2$, ..., ${}^nC_n$ respectively, then the mean is $\frac{1}{2}n$, the mean square deviation about $x = 0$ is $\frac{1}{4}n(n+1)$ and the variance is $\frac{1}{4}n$.

   (b) The marks obtained by the students are :

   25, 50, 45, 30, 70, 42, 36, 38, 34, 60.

   Find the Coefficient of skew.
3. For a frequency distribution of marks in History of 200 candidates (grouped in intervals 0–5, 5–10, ... etc.), the mean and S. D. were found to be 40 and 15. Later it was discovered that the score 43 was misread as 53 in obtaining the frequency distribution. Find the corrected mean and S. D. corresponding to the corrected distribution. (S. D. stands for standard deviation.)
4. The deviation of distribution is measured from a value differing from the mean of the distribution by x. Show that if x is plotted against the corresponding mean square deviation, the points lie on a parabola.
5. The first three uncorrelated moments about the origin are given by

$$\mu_1' = \frac{n+1}{12}, \quad \mu_2' = \frac{(n+1)(2n+1)}{6}, \quad \mu_3' = \frac{n(n+1)^2}{4}.$$

Examine the skewness of the data.

6. What do you understand by skewness and kurtosis ? Explain in brief.
7. What do you mean by variance ? Establish a relation between standard deviation and root mean square deviation.
8. Find the coefficient of skewness of the two groups given below and point out which distribution is more skew ?

Marks :	55–58	58–61	61–64	64–67	67–70
Group A :	12	17	33	18	11
Group B :	20	22	25	13	7

[**Hint :** Find Bowley's coefficient of skewness]

9. For the following table of marks obtained by ten candidates, find the coefficient of variation and skewness :

Statistics :	25	50	45	30	70	42	36	38	34	60
Mathematics :	10	70	50	20	95	55	42	60	48	80

10. Compute the coefficient of skewness from the following data :

Age at the birth of first child :	13	14	15	16	17	18	19	20	21	22	23	24	25
Number of married women :	37	162	343	390	256	433	161	355	65	85	49	46	40

11. Calculate arithmetic mean, standard deviation and skewness from the following data :

Measure :	3	4	5	6	7	8	9
Frequency :	3	7	22	60	85	32	8

Calculate the Karl Pearson's coefficient of skewness for the following data :

12.

Marks :	0–10	10–20	20–30	30–40	40–50	50–60
No. of Candidates :	5	12	18	38	20	7

13.

Class :	0–4	4–8	8–12	12–16	16–20	20–24
Frequency :	5	7	10	15	8	4

14.

Wages (in Rs.) :	70–80	80–90	90–100	100–110
No. of workers :	12	18	35	42
Wages (in Rs.) :	110–120	120–130	130–140	140–150
No. of workers :	50	45	20	8

15.

Class :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency :	30	40	50	48	24	162	132	14

16.

Height (above in inches) :	58	59	60	61	62	63	64	65	66	67
No. of persons :	100	98	90	76	57	35	20	11	4	0

17.

Class :	0–6	6–12	12–18	18–24	24–30	30–36
Frequency :	12	24	38			

18.

Cultivation (per acre) in quintals :	20–24	25–29	30–34	35–39	40–44	45–49
No. of Forms :	6	20	40	70	35	9

19.

Class	Frequency	Class	Frequency
0—10	10	40—50	16
10—20	12	50—60	14
20—30	18	60—70	5
30—40	25		

20. The first four moments of a distribution about the value 4 of the variable are $-1{\cdot}5, 17, -30$ and 108. Discuss the skewness and kurtosis of the distribution.
21. The first four moments of a distribution about the mean are $0, 120, -10, 45000$. Test the kurtosis of the distribution.
22. The first four central moments of a distribution are $0, 2{\cdot}50, 0{\cdot}70$ and $18{\cdot}75$. Calculate the coefficients of skewness and kurtosis.
23. The first four moments of a distribution about the value 0 of a variable are $-0{\cdot}20, 1{\cdot}76, -2{\cdot}36$ and $10{\cdot}88$. Find the first four central moments. Also find $\beta_1, \gamma_1, \beta_2$ and γ_2.
24. Define the coefficients of Skewness and Kurtosis. The first four moments of a distribution about the value 4 of the variable are $-1{\cdot}5, 17, -30$ and 108. Find the moments about the otigin. State whether the distribution leptokurtic or platykurtic.

[**Hint.** See example 8 after § 5.14. We have $\beta_2 = 0{\cdot}6541 < 3$, hence the distribution is platykurtic.]

ANSWERS

1. $\frac{1}{2}(n+1)$, $\sigma^2 = (1/12)(n+1)(4n^2 - n - 3)$. **3.** 3995. 1497
5. Since $H_3 = 0$, No skewness in data.
8. $-0{\cdot}02, -0{\cdot}22, B$ **9.** Statistics : 30·5, 0·69; Mathematics : 45·8, 0·06
10. $-0{\cdot}11$ **11.** $\bar{x} = 6{\cdot}6, \sigma = 1{\cdot}7, J = -0{\cdot}24$ **12.** $J = -0{\cdot}24$
13. $J = -0{\cdot}26$ **14.** $J = -0{\cdot}326$ **15.** $J = -0{\cdot}6$
16. $J = -0{\cdot}48$ **17.** $J = -0{\cdot}036$ **18.** $J = -0{\cdot}19$
19. $J = -0{\cdot}04$ **20.** $\gamma_1 = 0{\cdot}7017$, so positively skewed; $\beta_2 = 0{\cdot}6541$, platykurtic.
21. $\beta_2 = 3{\cdot}125$, leptokurtic.
22. $\gamma_1 = 0{\cdot}18$ positively skewed; $\beta_2 = 3$, the curve in mesokurtic.
23. $\mu_1 = 0$, $\mu_2 = 1{\cdot}72$, $\mu_3 = -1{\cdot}32$, $\mu_4 = 9{\cdot}4192$, $\beta_1 = 0{\cdot}3424$, $\gamma_1 = -0{\cdot}5852$, $\beta_2 = 3{\cdot}1839$, $\gamma_2 = 0{\cdot}1839$.
24. $\mu_1 = 0$, $\mu_2 = 14{\cdot}75$, $\mu_3 = 39{\cdot}75$, $\mu_4 = 142{\cdot}3125$; $\nu_1 = 2{\cdot}5$, $\nu_2 = 21$, $\nu_3 = 106$, $\nu_4 = 1132$; Platykurtic.

EXERCISE 5 (E)

Objective Type Questions

1. The formula for the measure of Quartile Deviation is :
   (a) $\frac{1}{4}(Q_3 + Q_1)$ (b) $\frac{1}{4}(Q_3 - Q_1)$
   (c) $\frac{1}{2}(Q_3 + Q_1)$ (d) $\frac{1}{2}(Q_3 - Q_1)$.

2. The relation between M. D. and S. D. about the mean in a discrete series when all variates are not equal is :
   (a) M. D. = S. D. (b) M. D. ≥ S. D.
   (c) M. D. < S. D. (d) M. D. ≤ S. D.

3. The empirical relation between the measures of dispersion is :
   (a) M. D. $= \frac{3}{4}$ (S. D.) (b) M. D. $= \frac{4}{3}$ (S. D.)
   (c) M. D. $= \frac{4}{5}$ (S. D.) (d) M. D. $= \frac{5}{4}$ (S. D.).

4. The empirical relation between the measures of dispersion is :
   (a) semi-inter quartile range $= \frac{2}{3}$ S. D.
   (b) semi-inter quartile range $= \frac{3}{4}$ S. D.
   (c) semi-inter quartile range $= \frac{4}{5}$ S. D.
   (d) semi-inter quartile range $= \frac{5}{6}$ S. D.

5. For a series, the relation between mean deviation (M. D.) and quartile deviation (Q. D.) is :
   (a) Q. D. $= \frac{5}{6} \times$ M. D. (b) M. D. $= \frac{5}{6} \times$ Q. D.
   (c) Q. D. $= \frac{3}{4} \times$ M. D. (d) M. D. $= \frac{3}{4} \times$ Q. D.

6. The mean deviation for any series is 15, then the maximum possible value of quartile deviation is :
   (a) 10·5 (b) 11·5
   (c) 12·5 (d) 13·5.

7. The relation between root mean square deviation (s) and standard deviation (σ) for any frequency distribution is :
   (a) $\sigma^2 = s^2 + d^2$ (b) $s^2 = \sigma^2 + d^2$
   (c) $\sigma^2 = s.d$ (d) $s^2 = \sigma.d.$
   where $d = M - A$, M = mean, A = assumed mean.

8. For every frequency distribution, the standard deviation is :
   (a) unique and always exists (b) unique but does not always exist
   (c) not unique but always exist (d) not unique and does not always exist.

9. The least value of root mean square deviation is :
   (a) mean deviation (b) standard deviation
   (c) quartile deviation (d) none of these.

10. If every variate x of a set of observations is multiplied by a non-zero constant k and thus a new variate z is obtained, then which of the following relations is true :
   (a) $\sigma_z = \sigma_x$ (b) $\sigma_z = k\sigma_x$
   (c) $\sigma_x = k\sigma_z$ (d) None of these.

11. If every variate of a set of observations is multiplied by a constant greater than 1 (unity), then the variance of the resultant variate will :
   (a) remain unaltered (b) increase
   (c) decrease (d) none of these.

12. By change of origin, the standard deviation will :
   (a) remain unaltered (b) increase
   (c) decrease (d) none of these.

13. The standard deviation :
   (a) is effected by change of origin but uneffected by change of scale
   (b) is uneffected by change of origin but effected by change of scale
   (c) is effected by change of both origin and scale
   (d) is uneffected by change of both origin and scale.

14. If variates x are transformed to variates u by the relation $u = \frac{x-a}{h}$, then the relation between their standard deviations is :
   (a) $\sigma_x = h\sigma_u$ (b) $\sigma_u = h\sigma_x$
   (c) $\sigma_x = a + h\sigma_u$ (d) $\sigma_u = a + h\sigma_x$.

15. The root mean square deviation is least when deviations are measured from :
   (a) mean (b) median
   (c) mode (d) 0.

16. The mean deviation is least when deviations are measured from :
   (a) mean (b) median
   (c) mode (d) 0.

17. The formula for the measure of coefficient of variation is :
   (a) $\frac{\sigma}{M}$ (b) $\frac{M}{\sigma}$
   (c) $\frac{\sigma}{M} \times 100$ (d) $\frac{M}{\sigma} \times 100$
   The symbols have their usual meaning.

18. In symmetric frequency distribution, upper and lower quartiles are at equal distances :
   (a) from mean (b) from median
   (c) from mode (d) from standard deviation.

19. Mean, Mode and Standard deviation of a frequency distribution are 41, 45 and 8 respectively, then Pearson's coefficient of skewness is :
   (a) −1·5 (b) −0·5
   (c) 0·5 (d) none of these.

20. The limit for Bowley's quartile coefficient of skewness is :
   (a) ±3 (b) 0 and 3
   (c) ±1 (d) ±∞.

21. The limit for Pearson's coefficient of skewness is :
(a) ± 3 (b) 0 and 3
(c) ± 1 (d) $\pm \infty$.

22. In a frequency curve of scores, it is found that mode is greater than mean. It shows that the distribution is :
(a) symmetric (b) negative Skewness
(c) positive Skewness (d) normal.

23. Kurtosis for a frequency distribution is :
(a) greater than 1 (b) less than 1
(c) equal to 1 (d) equal to 0.

24. For a discrete distribution :
(a) $\beta_2 = 0$ (b) $\beta_2 < 1$
(c) $\beta_2 = 1$ (d) $\beta_2 > 1$.

25. The measure of kurtosis is :
(a) $\beta_2 = 0$ (b) $\beta_2 = 2$
(c) $\beta_2 = 3$ (d) $\beta_2 = 4$.

26. For a symmetric distribution :
(a) $\mu_2 = 0$ (b) $\mu_2 > 0$
(c) $\mu_3 > 0$ (d) $\mu_3 = 0$.

27. For a symmetric distribution :
(a) $\beta_1 = 0$ (b) $\beta_1 < 0$
(c) $\beta_1 > 0$ (d) $\beta_1 = 1$.

28. If n is a natural number, then for a symmetric distribution, we have :
(a) $\mu_{2n} < 0, \mu_1 = 0$ (b) $\mu_{2n} = 0, \mu_1 = 0$
(c) $\mu_{2n+1} \sqrt{0}, \mu_1 = 0$ (d) $\mu_{2n+1} = 0, \mu_1 = 0$.

29. For a frequency distribution, we always have :
(a) $\beta_2 = \beta_1$ (b) $\beta_2 > \beta_1$
(c) $\beta_2 < \beta_1$ (d) $\beta_2 > \beta_1^2$.

30. If mean and mode of a frequency distribution have the same value, then coefficient of skewness is :
(a) 0 (b) greater than 0
(c) less than 0 (d) none of these.

31. The coefficient of skewness is numerically less than :
(a) the mean deviation (b) root mean square deviation
(c) standard deviation (d) 1.

32. The statement "the variance is equal to the second central moments" :
(a) is always true (b) is sometimes true
(c) is never true (d) not clear.

33. If for a frequency distribution $\beta_2 = 3$, then the curve is :
(a) normal (b) leptokurtic
(c) platykurtic (d) none of these.

34. If for a frequency distribution $\beta_2 < 3$, then the curve is :
(a) normal (b) leptokurtic
(c) platykurtic (d) none of these.

35. If $\beta_2 > 3$, then the curve is :
(a) normal (b) leptokurtic
(c) platykurtic (d) none of these.

36. The standard deviation of a distribution is 5. If the distribution be mesokurtic, then the fourth central moment (μ_4) is :
(a) 3 (b) greater than 1,875
(c) 1,875 (d) less than 1,875.

37. If r be range and σ be the S. D. of a set of n observations then we have the relation :
(a) $\sigma \leq r$ (b) $\sigma \geq r$
(c) $\sigma = r$ (d) none of these.

38. If σ^2 be the corrected value of variance and σ_1^2 be its calculated value then Sheppard's correction for grouping is :
(a) $\sigma_1^2 = \sigma^2 - (h^2/2)$(b) $\sigma^2 = \sigma_1^2 - (h^2/2)$
(c) $\sigma_1^2 = \sigma^2 - (h^2/12)$ (d) $\sigma^2 = \sigma_1^2 - (h^2/12)$.

39. If h be the class-interval then Sheppard's correction for fourth central moment is :
(a) μ_4 (corrected) $= \mu_4$ (calculated) $- \frac{1}{2} h^2 \mu_2$ (calculated) $- \frac{7}{240} h^2 \ldots$
(b) μ_4 (corrected) $= \mu_4$ (calculated) $+ \frac{1}{2} h^2 \mu_2$ (calculated) $- \frac{7}{240} h^2 \ldots$
(c) μ_4 (corrected) $= \mu_4$ (calculated) $- \frac{1}{2} h^2 \mu_2$ (calculated) $+ \frac{7}{240} h^2 \ldots$
(d) no correction in μ_4.

40. The S. D. of first n natural numbers is :
(a) $\sqrt{\left[\frac{(n+1)(2n+1)}{12}\right]}$ (b) $\sqrt{\left[\frac{(n+1)(2n-1)}{12}\right]}$
(c) $\sqrt{\left(\frac{(n^2+1}{12}\right)}$ (d) $\sqrt{\left(\frac{(n^2-1}{12}\right)}$.

41. If σ_x is standard deviation of a series of variates x and σ_u the standard deviation of the corresponding series of variates $u = \frac{ax+b}{c}$, then :
(a) $\sigma_u = a\sigma_x$ (b) $\sigma_u = \frac{1}{c}\sigma_x$
(c) $\sigma_u = \frac{a}{c}\sigma_x$ (d) $\sigma_u = \left|\frac{a}{c}\right|\sigma_x$.

42. The formula for skewness in terms of the coefficient of Karl Pearson is :
(a) Skewness $= \frac{\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2 - 6\beta_1 - 9)}$ (b) Skewness $= \frac{\sqrt{\beta_1}(\beta_2+3)}{2(6\beta_2 - 5\beta_1 + 9)}$
(c) Skewness $= \frac{\sqrt{\beta_1}(\beta_2+5)}{2(9\beta_2 - 5\beta_1 - 9)}$ (d) Skewness $= \frac{\sqrt{\beta_2}(\beta_2+3)}{2(5\beta_1 - 6\beta_2 + 9)}$.

43. Quartile Deviation is :
(a) $Q_3 - Q_1$ (b) $Q_3 + Q_1$
(c) $\frac{Q_3 - Q_1}{2}$ (d) $\frac{Q_3 + Q_1}{2}$.

44. If the S. D. of the variate x is σ_x, then S. D. of $ax + b$ is :

(a) $a\sigma_x$ (b) $a^2 + b^2$

(c) $b + \sigma_x$ (d) $a\sigma_x + b$.

45. If μ_2 be the second moment about the mean $\bar{x}$, then :

(a) $\mu_2 = \bar{x}$ (b) $\mu_2 = \sigma$

(c) $\mu_2 = \sigma^2$ (d) $\mu_2 = \bar{x}^2$.

46. Formula for coefficient of variation is :

(a) $V = \frac{\sigma}{M}$ (b) $V = \frac{M}{\sigma}$

(c) $V = \frac{\sigma}{M} \times 100$ (d) None of these.

ANSWERS

1. (d)	**2.** (d)	**3.** (c)	**4.** (a)	**5.** (a)	**6.** (c)	**7.** (c)	**8.** (a)
9. (b)	**10.** (b)	**11.** (c)	**12.** (a)	**13.** (b)	**14.** (a)	**15.** (a)	**16.** (b)
17. (c)	**18.** (d)	**19.** (b)	**20.** (c)	**21.** (a)	**22.** (b)	**23.** (a)	**24.** (d)
25. (c)	**26.** (d)	**27.** (a)	**28.** (d)	**29.** (b)	**30.** (a)	**31.** (a)	**32.** (a)
33. (a)	**34.** (c)	**35.** (b)	**36.** (c)	**37.** (c)	**38.** (d)	**39.** (c)	**40.** (d)
41. (d)	**42.** (a)	**43.** (a)	**44.** (a)	**45.** (c)	**46.** (c).		

Chapter 6

UNIT-IV

PROBABILITY DISTRIBUTION

§ 6.1. INTRODUCTION

Definition. *When frequency distributions of some universes are not based on actual observations or experiments, but can be derived mathematically from certain pre-determined hypothesis, then such distributions are said to be theoretical distributions.*

For example. Four coins are tossed 80 times, then according to the principle of probability, the expected frequency distribution is obtained as given in the following table :

No. of heads	Probability	Expected frequency distribution in 80 tosses
0	1/16	$80 \times (1/16) = 5$
1	4/16	$80 \times (4/16) = 20$
2	6/16	$80 \times (6/16) = 30$
3	4/16	$80 \times (4/16) = 20$
4	1/16	$80 \times (1/16) = 5$
Total	1	80

Type of Theoretical Distribution. There are several types of theoretical distributions, but the following two types of theoretical distribution are usually used in statistics :

1. **Discrete Distribution (Discrete Probability Distribution).**
   (a) Binomial Distribution,
   (b) Poisson Distribution.

2. Continuous Distribution (Continuous Probability Distribution).

(a) Normal Distribution (b) Exponential Distribution
(c) Gamma Distribution (d) Beta Distribution
(e) Rectangular Distribution.

§ 6.2. BINOMIAL DISTRIBUTION* (OR BERNOULLI'S DISTRIBUTION)

Suppose a random experiment is performed repeatedly. Let E be an event. We shall call the occurrence of the event E a 'success and its non-occurrence a failure'. If p denotes the probability of a success and q denotes the probability of its failure; then $p + q = 1$. Let the event E be tried n times, where n is finite. The hypotheses for binomial distribution are :

(i) All the trials are independent, *i.e.*, the result of one trial will not affect the results of succeeding trials.

(ii) The number (n) of trials is finite.

(iii) The probability p of successes is the same in every trial.

The number of successes in n trials may be 0, 1, 2, 3, ..., r, ..., n and is clearly a random variate. Suppose that the set of n trials is repeated N times, where N is very large. Obviously out of these N sets there will be a few sets with no success, a few sets with one success, a few sets with two successes, a few sets with 3 successes, ..., etc.

Now suppose that first r trials are successes and the remaining $n - r$ trials are failures. Its probability is $p^r\ q^{n-r}$. Clearly it is also the probability for r successes and $n - r$ failures which may occur in any order. But we are to consider all the possible cases where any r trials are successes and thus r can be chosen out of n in ${}^nC_r$ mutually exclusive ways.

Hence the probability $P(r)$ of r successes in a series of n independent trials is given by the following formula [using theorem of total probability] :

$$P(r) = {}^nC_r\ p^r q^{n-r}.$$

Definition. A random variable X is said to follow Bernoulli or binomial distribution if it takes only non-negative values and its probability mass function is given by the following formula :

$$P(X = r) = \begin{cases} {}^nC_r\ p^r q^{n-r} & ;\ r = 0, 1, 2, \ldots, n,\ q = 1 - p \\ 0 & ;\ r \neq 0, 1, 2, \ldots, n. \end{cases}$$

Here, the two independent constants n and p (or q) in the distribution are called **parameters** of distribution.

Since we have considered N sets, each of n trials, therefore the number of sets with r successes $= N \cdot {}^nC_r\ p^r q^{n-r}$ or $N \cdot P(r)$. Hence we obtain the following frequency distribution :

No. of successes :	0	1	2	...	r	...	n
No. of sets :	Nq^n	$N \cdot {}^nC_1\ pq^{n-1}$	$N \cdot {}^nC_2\ p^2q^{n-2}$	...	$N \cdot {}^nC_r\ p^r q^{n-r}$	...	Np^n

*Binomial Distribution was discovered by James Bernoulli (1654–1705) in 1700.

Thus we see that, for N sets of n trials, the frequencies of 0, 1, 2, ..., r, ..., n successes are the successive terms of the following expression :

$$N \cdot q^n + N \cdot {}^nC_1\ pq^{n-1} + N \cdot {}^nC_2\ p^2q^{n-2} + \ldots + N \cdot {}^nC_r\ p^r q^{n-r} + \ldots + N \cdot p^n$$

i.e.,

$$N\,[q^n + {}^nC_1\ pq^{n-1} + {}^nC_2\ p^2q^{n-2} + \ldots + {}^nC_r\ p^r q^{n-r} + \ldots + p^n]$$

which is the binomial expansion of $N \cdot (q + p)^n$. It is called the **Binomial frequency distribution.**

Remark 1. In the above binomial distribution n and p (or q) are said to be parameters.

Remark 2. If $p = q = \frac{1}{2}$, the binomial distribution is called **symmetrical distribution,** otherwise it is called **skew distribution.**

Example 1. *Show that the probability of at most r successes in n trials is* $\sum_{t=0}^{r} {}^nC_t\ p^t q^{n-t}$.

Example 2. *Show that the probability of at least r successes in n trials is* $\sum_{t=r}^{n} {}^nC_t\ p^t q^{n-t}$.

§ 6.3. CONSTANTS OF BINOMIAL DISTRIBUTION

Moments about the origin. We know that binomial distribution is given by

$$(q + p)^n = q^n + {}^nC_1\ q^{n-1}p + {}^nC_2\ q^{n-2}\ p^2 + \ldots + {}^nC_r\ q^{n-r}\ p^r + \ldots + p^n$$

No. of successes :	0	1	2	...	r	...	n
No. of sets : (Frequency f)	q^n	${}^nC_1\ pq^{n-1}$	${}^nC_2\ p^2q^{n-2}$	...	${}^nC_r\ p^r q^{n-r}$	...	p^n

We take arbitrary origin at 0 success (*i.e.* at no success).

(i) First moment about the origin.

$$\mu_1' = \sum_{r=0}^{n} r \cdot {}^nC_r\ p^r q^{n-r} \quad \text{where } q + p = 1$$

$$= \sum_{r=1}^{n} n \cdot {}^{n-1}C_{r-1}\ p^r q^{n-r}$$

$$\left[\because\ r \cdot {}^nC_r = \frac{r \cdot n!}{(n-r)!\,r!} = \frac{n(n-1)!}{(n-r)!\,(r-1)!} = n \cdot {}^{n-1}C_{r-1}\right]$$

$$= np \sum_{r=1}^{n} {}^{n-1}C_{r-1}\ p^{r-1} q^{(n-1)-(r-1)}$$

$$= np\,[q^{n-1} + {}^{n-1}C_1\ q^{n-2}\ p + {}^{n-1}C_2\ q^{n-3}p^2 + \ldots + p^{n-1}]$$

$$[\because\ q + p = 1]$$

$$= np(q + p)^{n-1} = np$$

$\therefore$ Mean $= \mu_1' / 1 = np / 1 = np$.

(ii) Second moment about the origin :

$$\mu_2' = \sum_{r=0}^{n} r^2 \cdot {}^nC_r\ p^r q^{n-r} = \sum_{r=0}^{n} \{r(r-1)+r\}\ {}^nC_r\ p^r q^{n-r}$$

$$= \sum_{r=0}^{n} r(r-1)\ {}^nC_r\ p^r q^{n-r} + \sum_{r=0}^{n} r \cdot {}^nC_r\ p^r q^{n-r}$$

$$= \sum_{r=0}^{n} n(n-1)\ {}^{n-2}C_{r-2}\ p^r q^{n-r} + np$$

$$= n(n-1)\ p^2 \sum_{r=2}^{n} {}^{n-2}C_{r-2}\ p^{r-2}\ q^{(n-2)-(r-2)} + np$$

$$= n(n-1)\ p^2 (q+p)^{n-2} + np, \text{ where } q+p=1$$

$$= n(n-1)\ p^2 + np = np\ [np + (1-p)] = np\ [np+q] \quad [\because\ q = 1-p]$$

or $\mu_2' = npq + n^2p^2$.

(iii) Third moment about the origin :

$$\mu_3' = \sum_{r=0}^{n} r^3 \cdot {}^nC_r\ p^r q^{n-r}$$

$$= \sum_{r=0}^{n} \{r(r-1)(r-2) + 3r(r-1) + r\}\ {}^nC_r\ p^r q^{n-r}$$

$$= n(n-1)(n-2)\ p^3 (q+p)^{n-3} + 3n(n-1)\ p^2 (q+p)^{n-2} + np$$

$$= n(n-1)(n-2)\ p^3 + 3n(n-1)\ p^2 + np.$$

(iv) Fourth moment about the origin :

$$\mu_4' = \sum_{r=0}^{n} r^4 \cdot {}^nC_r\ p^r q^{n-r}$$

$$= \sum_{r=0}^{n} \{r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r\} \cdot {}^nC_r\ p^r q^{n-r}$$

$$= n(n-1)(n-2)(n-3)\ p^4 (q+p)^{n-4} + 6n(n-1)(n-2)\ p^3 (q+p)^{n-3} + 7n(n-1)\ p^2 (q+p)^{n-2} + np$$

$$= n(n-1)(n-2)(n-3)\ p^4 + 6n(n-1)(n-2)\ p^3 + 7n(n-1)\ p^2 + np.$$

Moments About the Mean (Central Moments) :

(v) First moment about the mean $= \mu_1 = 0$ (always)

(vi) Second moment μ_2 about the mean is given by

$$\mu_2 = \mu_2' - \mu_1'^2 = npq + n^2p^2 - n^2p^2 = npq$$

or variance $= \mu_2 = npq$.

S.D. $= \sqrt{npq}$.

(vii) Third moment about the mean is given by

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= np\ [\{(n-1)(n-2)\ p^2 + 3(n-1)\ p + 1] - 3(n^2p^2 + npq) \cdot np + 2(np)^3$$

$$= np\ [\{(n-1)(n-2) - 3n^2 + 2n^2\}\ p^2 + \{3(n-1) - 3nq)\ p + 1]$$

$$= np\ [(-3n+2)\ p^2 + 3(n-1)\ p - 3np(1-p) + 1] \quad [\because\ p = 1-q]$$

$$= np(2p^2 - 3p + 1) = np(1-2p)(1-p)$$

$$= np(1-p)\{1-2(1-q)\} \quad [\because\ p = 1-q]$$

$$= npq(2q-1) = npq\ [q-(1-q)] = npq(q-p).$$

(viii) Fourth moment about the mean is given by

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

$$= [n(n-1)(n-2)(n-3)\ p^4] + 6n\ [(n-1)(n-2)\ p^3 + 7n(n-1)\ p^2 + np] - 4\ [n(n-1)(n-2)\ p^3 + 3n(n-1)\ p^2 + np] \times np + 6\{n(n-1)\ p^2 + np\}\ n^2p^2 - 3(np)^4$$

$$= p^4[(n-1)(n-2)(n-3) - 4n^2(n-1)(n-2) + 6n^3(n-1) - 3n^3] + p^3\ [6n(n-1)(n-2) - 12n^2(n-1) + 6n^2] + p^2\ [6n(n-1) - 4n^2] + np$$

$$= p^4\ [3n^2 - 6n] + p^3\ [12n - 6n^2] + p^2\ [3n^2 - 7n] + pn$$

$$= 3n^2(p^4 - 2p^3 + p^2) - n\ (6p^4 - 12p^3 + 6p^2 - np^2 + np)$$

$$= 3n^2p^2\ (p^2 - 2p + 1) - 6np^2\ (1 - 2p + p^2) + np(1-p)$$

$$= 3n^2p^2\ (1-p)^2 - 6np^2\ (1-p)^2 + npq = 3n^2p^2q^2 - 6np^2q^2 + npq$$

$$= 3n^2p^2q^2 + npq(1-6pq) = npq\ [1 + 3(n-2)\ pq].$$

Karl Pearson's Coefficients :

(ix) $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{[npq(q-p)]^2}{(npq)^3} = \dfrac{(q-p)^2}{npq} = \dfrac{(1-2p)^2}{npq}.$

(x) $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3n^2p^2q^2 + npq\,(1-6pq)}{(npq)^2} = 3 + \dfrac{(1-6pq)}{npq}.$

(xi) $\gamma_1 = \sqrt{\beta_1} = \dfrac{q-p}{\sqrt{(npq)}}.$

(xii) $\gamma_2 = \beta_2 - 3 = \dfrac{1-6pq}{npq}.$

§ 6.4. RENOVSKY FORMULA

Theorem. *The recurrence relation for the moments of Binomial distribution is given by*

$$\mu_{r+1} = pq\left(nr\mu_{r-1} + \frac{d\,\mu_r}{dp}\right)$$

where μ_r is the rth moment about the mean.

Proof. We know that

$$\mu_r = \sum_{x=0}^{n} p(x)\,(x-\mu_1')^r \quad \ldots(1)$$

where $p(x)$ is the probability distribution.

For binomial distribution $\mu_1' = np$. Putting values in (1), we get

$$\mu_r = \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x}\,(x-np)^r$$

$$= \sum_{x=0}^{n} {}^nC_x\, p^x\,(1-p)^{n-x}(x-np)^r \qquad [\because\ q = 1-p]$$

Differentiating w.r.t. 'p', we get

$$\frac{d\mu_r}{dp} = \sum_{x=0}^{n} {}^nC_x\, xp^{x-1}\,(1-p)^{n-x}\,(x-np)^r$$

$$+ \sum_{x=0}^{n} {}^nC_x\, p^x\,(n-x)\,(1-p)^{n-x-1}\,(-1)\,(x-np)^r$$

$$+ \sum_{x=0}^{n} {}^nC_x\, p^x(1-p)^{n-x}\,r(x-np)^{r-1}\,(-n)$$

$$= \sum_{x=0}^{n} {}^nC_x\, xp^{x-1}\,q^{n-x}\,(x-np)^r$$

$$- \sum_{x=0}^{n} {}^nC_x\,(n-x)\,p^x q^{n-x-1}\,(x-np)^r$$

$$- \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x}\,(x-np)^{r-1}\,nr \quad [\because\ q = 1-p]$$

$$= \sum_{x=0}^{n} {}^nC_x\, p^{x-1}\,q^{n-x-1}\,(x-np)^r\ [xq-(n-x)p]$$

$$- nr \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x}\,(x-np)^{r-1}$$

$$= \sum_{x=0}^{n} {}^nC_x\, p^{x-1}\,q^{n-x-1}\,(x-np)\,[x(q+p)-np] - nr\mu_{r-1}. \quad [\text{using (1)}]$$

Multiply both sides by pq and use $p+q=1$, we get

$$pq\frac{d\mu_r}{dp} = \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x}(x-np)^r(x-np) - npqr\,\mu_{r-1}$$

$$= \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x}(x-np)^{r+1} - npqr\,\mu_{r-1}$$

$$= \mu_{r+1} - npqr\,\mu_{r-1}$$

or
$$\mu_{r+1} = qp\left(nr\,\mu_{r-1} + \frac{d\mu_r}{dp}\right). \quad \ldots(2)$$

Cor. *Obtain μ_2, μ_3, μ_4 from Renovsky formula.*

Proof. Putting $r = 1, 2, 3$ in the Renovsky formula (2) successively, we get

$$\mu_2 = pq\,[n\,.\,1\,.\,\mu_0 + d\mu_1/dp] = pq\,(n+0) = npq$$

= variance $\qquad [\because\ \mu_0 = 1 \text{ and } \mu_1 = 0]$

$$\mu_3 = pq\,[n\,.\,2\mu_1 + d\mu_2/dp]$$

$$= pq\left[2n\,.\,0 + nq + np\frac{dq}{dp}\right]$$

$$\left[\because\ \frac{d\mu_2}{dp} = nq + np\,.\,\frac{dq}{dp} = nq - np \text{ as } q = 1-p\right]$$

$$= npq\,(q-p)$$

$$\mu_4 = pq\,[n\,.\,3\mu_2 + d\mu_3/dp] = pq[3n\,.\,np(1-p) + n(6p^2 - 6p + 1)]$$

$$= npq\,[1 - 6p(1-p) + 3npq]$$

$$= 3n^2p^2q^2 + npq(1-6pq).$$

ILLUSTRATIVE EXAMPLES

Example 1. ***Criticise the following statement :***
'For any binomial distribution mean is 5 and S.D. is 3.'

Solution. We are given that

$$\text{Mean} = np = 5 \text{ and S.D.} = \sqrt{npq} = 3$$

$$\therefore \quad npq = 9 \Rightarrow 5q = 9 \Rightarrow q = 9/5 > 1$$

But q cannot be greater than 1, since it is a probability. Hence the given statement is false.

Example 2. ***A perfectly cubical die is thrown a large number of times in sets of 8. The occurrence of 5 or 6 is called a success. In what proportion of the sets do you expect 3 successes ?***

Solution. We have $N = 8$.

p = The probability of getting 5 or 6 = 2/6 = 1/3

$$\therefore \quad q = 1 - \frac{1}{3} = \frac{2}{3}.$$

$$\therefore \quad \text{The binomial distribution is} = \left(\frac{2}{3} + \frac{1}{3}\right)^8.$$

Therefore, the number of sets in which 3 successes are expected

$$= N\left[{}^8C_2\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^5\right] = N\,.\,\frac{56\times 32}{243\times 27}.$$

Hence the percentage $= N\,\dfrac{56\times 32}{243\times 27}\times\dfrac{100}{N} = 27.31\%$.

Example 3. ***Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six ?***

Solution. We know that when a die is thrown, the probability to show a five or six = 2/6 = 1/3 = p (say).

$$\therefore \quad q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}.$$

The probability to show a five or six in at least three dice

$$= \sum_{x=3}^{6} p(x) = p(3) + p(4) + p(5) + p(6),$$

where $p(x)$ is the probability to show five or six

$$= {}^6C_3\, q^3 p^3 + {}^6C_4\, p^4 q^2 + {}^6C_5\, p^5 q + {}^6C_6\, p^6$$

$$= {}^6C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 + {}^6C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 + {}^6C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right) + {}^6C_6 \left(\frac{1}{3}\right)^6$$

$$= \frac{1}{(3)^6}\left[\frac{6\times 5\times 4}{3\times 2\times 1}(2)^3 + \frac{6\times 5}{2\times 1}(2)^2 + 6(2) + 1\right]$$

$$= \left(\frac{1}{729}\right)[160 + 60 + 12 + 1] = \frac{233}{729} = p \text{ (say)}.$$

$\therefore$ The required number $= np = 729(233/729) = 233$.

Example 4. ***The number of males in each 106 eight pig litters was found and they are given by the following frequency distribution :***

Number of males per liter	:	0	1	2	3	4	5	6	7	8	Total
Frequency	:	0	5	9	22	25	26	14	4	1	106

Assuming that the probability of an animal being male or female is even, i.e., $p = q = \frac{1}{2}$ and frequency distribution follows the binomial law, calculate the expected frequencies.

Solution. Here the frequency distribution $= 106\left(\frac{1}{2} + \frac{1}{2}\right)^8$ [using $N(q+p)^n$]

Therefore the expected frequencies are the respective terms of this expansion which are respectively :

$$0\cdot 4,\ 3\cdot 3,\ 11\cdot 6,\ 23\cdot 2,\ 29,\ 23\cdot 2,\ 11\cdot 6,\ 3\cdot 3,\ 0\cdot 4.$$

Example 5. ***In litters of 4 mice the number of litters which contained 0, 1, 2, 3, 4 females were noted. The figures are given in the table below :***

Number of female mice	:	0	1	2	3	4	Total
Number of litters	:	8	32	34	24	5	103

If the chance of obtaining a female in a single trial is assumed constant, estimate this constant of unknown probability. Find also the expected frequencies.

Solution. Here the mean of observations

$$= \frac{8\times 0 + 32\times 1 + 34\times 2 + 24\times 3 + 5\times 4}{8 + 32 + 34 + 24 + 5}$$

$$= \frac{32 + 68 + 72 + 20}{103} = \frac{192}{103} = 1\cdot 864.$$

$\because$ $n = 4$ and np = mean $= 1\cdot 864$.

Thus $$p = \frac{0\cdot 864}{4} = 0\cdot 466.$$

$\therefore$ $$q = 1 - p = 1 - 0\cdot 466 = 0\cdot 534.$$

Hence the expected frequencies are the respective terms of the binomial expansion of $103(0\cdot 534 + 0\cdot 466)^4$.

Example 6. ***The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to these data :***

x	:	0	1	2	3	4	5	6	7	8	9	10	Total
f	:	6	20	28	12	8	6	0	0	0	0	0	80

Solution. Here we have

$$n = 10,\ N = \Sigma f = 80.$$

$\therefore$ $$\text{A. M.} = \frac{\Sigma f x}{N} = \frac{1\times 20 + 2\times 28 + 3\times 12 + 4\times 8 + 5\times 6}{80} = \frac{174}{80}$$

But $$\text{mean} = np = \frac{174}{80} \Rightarrow 10p = \frac{174}{80} \Rightarrow p = \frac{174}{800} = 0\cdot 2175$$

$\therefore$ $$q = 1 - p = 1 - 0\cdot 2175 = 0\cdot 7825.$$

Hence the binomial distribution to be fitted to the given data is

$$80\,(0\cdot 7825 + 0\cdot 2175)^{10} \qquad [\text{using } N(q+p)^n]$$

and from this expansion the successive frequencies of 0, 1, 2, ..., 10 successes are respectively $6\cdot 9,\ 19\cdot 1,\ 24\cdot 0,\ 17\cdot 8,\ 8\cdot 6,\ 2\cdot 9,\ 0\cdot 7,\ 0\cdot 1,\ 0,\ 0,\ 0$, *i.e.*, 7, 19, 24, 18, 9, 3, 1, 0, 0, 0.

Example 7. (a) ***Assuming that half the population are consumers of chocolate, so that the chance of an individual being a consumer is $\frac{1}{2}$ and assuming that 100 investigators each take 10 individuals to see whether they are consumers, how many investigators would you expect to report that three people or less were consumers ?***

Solution. Here p = probability of an individual being a consumer $= \frac{1}{2}$ (given).

$\therefore$ $$q = 1 - \frac{1}{2} = \frac{1}{2}.$$

Again, we have $n = 10$, $N = 100$.

$\therefore$ The binomial distribution is $100\left(\frac{1}{2} + \frac{1}{2}\right)^{10}$

$\therefore$ The required number of investigators to report that 3 or less than 3 persons are consumers of chocolate

$$= \text{first four terms in the expansion } 100\left(\frac{1}{2} + \frac{1}{2}\right)^{10}$$

$$= 100\left[\left(\frac{1}{2}\right)^{10} + {}^{10}C_1\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + {}^{10}C_2\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + {}^{10}C_3\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3\right]$$

$$= 100\left(\frac{1}{2}\right)^{10}\left[1 + 10 + \frac{10\times 9}{2\times 1} + \frac{10\times 9\times 8}{3\times 2\times 1}\right]$$

$$= 100\times\left(\frac{1}{2}\right)^{10}[1 + 10 + 45 + 120]$$

$$= \frac{100}{2^{10}}\times 176 = \frac{25}{2^8}\times 2^4\times 11 = \frac{275}{16} = 17 \text{ nearly.}$$

Example 7. (b) ***In sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain***

(i) at least 3 defective parts ?

(ii) none defective ?

Solution. Mean number of defective parts $= 2 = np = 20p$ $[\because\ n = 20]$

Therefore, the probability of a defective parts $p = \dfrac{2}{20} = 0 \cdot 1$.

$\therefore$ The probability of a part being non-defective part $q = 1 - 0 \cdot 1 = 0 \cdot 9$

(i) Thus the probability of at least 3 defective parts in a sample of 20

= 1 – (the probability that either none, or one, or two parts are non-defective)

$$= 1 - [{}^{20}C_0 (0 \cdot 9)^{20} + {}^{20}C_1 (0 \cdot 1)(0 \cdot 9)^{19} + {}^{20}C_2 (0 \cdot 1)^2 (0 \cdot 1)^{18}]$$

$$= 1 - \left[(0 \cdot 1)^2 + 20 \times 0 \cdot 1 \times 0 \cdot 9 + \frac{20 \times 19}{2} \times (0 \cdot 1)^2\right](0 \cdot 9)^{18}$$

$$= 1 - 4 \cdot 51 \times (0 \cdot 9)^{18} = 0 \cdot 0 \cdot 323.$$

Hence the required number of samples which have at least three defective parts out of 1000 such samples

$= 1000 \times 0 \cdot 323 = 323.$ Ans.

(ii) The probability of none defective parts in a sample of 20

$= q^{20} = (0 \cdot 9)^{20} = 0 \cdot 122.$

Hence the required number of samples which have none defective parts out of 1000 such samples

$= 1000 \times 0 \cdot 122 = 122.$

Example 8. ***A variate assumes the values 0, 1, 2, 3, ..., n, whose frequencies are proportional to the binomial coefficients***

$$1, \binom{n}{1}, \binom{n}{2}, \binom{n}{3}, \ldots, \binom{n}{n}.$$

Show that variance is half of mean.

Solution. Here

x:	0	1	2	3	...	n
f:	1	${}^nC_1$	${}^nC_2$	${}^nC_3$	...	${}^nC_n$

Thus, the sum of the frequencies

$$= 1 + {}^nC_1 + {}^nC_2 + {}^nC_3 + \ldots + {}^nC_n = (1 + 1)^n = 2^n.$$

$\therefore$ For the values 0, 1, 2, 3, ..., n the probabilities respectively are

$$\frac{1}{2^n}, \frac{{}^nC_1}{2^n}, \frac{{}^nC_2}{2^n}, \ldots, \frac{{}^nC_n}{2^n}$$

i.e., $$\left(\frac{1}{2}\right)^n, {}^nC_1\left(\frac{1}{2}\right)^n, {}^nC_2\left(\frac{1}{2}\right)^n, \ldots, {}^nC_n\left(\frac{1}{2}\right)^n.$$

This is a binomial distribution $\left(\dfrac{1}{2} + \dfrac{1}{2}\right)^n$, *i.e.*, $(q + p)^n$ in which $p = q = \dfrac{1}{2}$.

Now, we know that

$$\text{Mean} = np = n\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)n$$

$$\text{Variance} = npq = n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \cdot n$$

$\therefore$ $$\text{Variance} = \left(\frac{1}{2}\right)(\text{mean}).$$

Example 9. ***An irregular six-faced die is thrown, and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws would you expect it to give no even number ?***

Solution. Consider that the probability of getting an even number is p.

$\therefore$ The probability of getting 5 even numbers in 10 throws $= {}^{10}C_5\ p^5 q^5$,

and the probability of getting 4 even numbers in 10 throws $= {}^{10}C_4\ p^4 q^6$.

$\therefore$ $${}^{10}C_5\ p^5 q^5 = 2 \times {}^{10}C_4\ p^4 q^6$$

$\Rightarrow$ $$\left(\frac{3}{5}\right) p = q \Rightarrow \left(\frac{3}{5}\right)(1 - q) = q,\ q = \left(\frac{3}{8}\right).$$

Thus the required number of times of not getting an even number in 10,000 sets of 10 throws $= 10{,}000\,(3/8)^{10} = 1$ nearly.

Example 10. ***Find the most probable number of successes in a series of n independent trials, the probability of successes in each trial being p.***

Solution. Here we have to find the number of successes which have a greater probability than any other.

Thus the probability of r successes is greater than or equal to $r - 1$ or $r + 1$ successes if

$$P(r - 1) \le P(r) \ge P(r + 1)$$

i.e., $${}^nC_{r-1}\ p^{r-1} q^{n-r+1} \le {}^nC_r\ p^r q^{n-r} \ge {}^nC_{r+1}\ p^{r+1} q^{n-r-1}$$

i.e., if $$\frac{r}{n - r + 1} \cdot \frac{q}{p} \le 1 \ge \frac{n - r}{r + 1} \cdot \frac{p}{q}$$

i.e., if $rq \le np - rp + p$ and $rp + q \ge np - rp$

i.e., if $np - q \le r \le np + p$

i.e., if $(n + 1)\ p - 1 \le r \le np + p$ $[\because\ q = 1 - p]$

Now the following two cases arise :

Case 1. If $(n + 1)\ p = k$ (an integer), then the probability will increase till $r = k$ and it will be same for $r = k - 1$ and after that it will decrease.

Case 2. If np = an integer + a fraction, then the probability is maximum when r = the integral part of $(np + p)$.

Note. The most probable number of successes is the mode of the binomial distribution and its value is $np + p$.

Example 11. ***If np be a whole number, then show that the mean of binomial distribution coincides with the greatest term.***

Solution. The frequency of r successes will be greater than $r-1$ successes if

$${}^nC_r\ p^r q^{n-r} > {}^nC_{r-1}\ p^{r-1}q^{n-r+1}$$

i.e., if $$\frac{p}{r} > \frac{q}{n-r+1}$$

i.e., if $$r < np + p. \qquad ...(1)$$

Also the frequency of r successes is greater than the frequency of $r+1$ successes, if

$$r > np - q. \qquad ...(2)$$

From (1) and (2) it is clear that if np is an integer, then $r = np$ gives the greatest term, which is the mean of the binomial distribution.

Example 12. ***The probability of a head in a single tossing of a biased coin is 3/5. Find the most probable number of heads and the mean number of heads in 99 tossings of a coin.***

Solution. Let the number of heads in 99 tossings be $X = 0, 1, 2, \ldots, 99$

$$\therefore \quad P(X = x) = {}^nC_x\ p^x q^{n-x}, \text{ where } p = 3/5,\ n = 99$$

$$= {}^{99}C_x\ (3/5)^x (2/5)^{99-x}, \quad x = 0, 1, 2, \ldots, 99$$

$$\therefore \quad \text{Mean} = np = 99 \times (3/5) = 59 \cdot 4.$$

Now $(n+1)\ p = (99+1)\ (3/5) = 60$ (an integer).

Hence $ex = np + p = 60$ and $x = np + p - 1 = 59$ are the most probable numbers of heads, where

$$P(X = 59) = {}^{99}C_{59}\ (3/5)^{59} (2/5)^{40}$$

and $$P(X = 60) = {}^{99}C_{60}\ (3/5)^{60} (2/5)^{39} \text{ are equal.}$$

Example 13. ***Find the binomial distribution whose mean is 4 and variance is 3. Also find its mode.***

Solution. Here $$\text{Mean}, np = 4 \qquad ...(1)$$

and $$\text{Variance}, npq = 3 \qquad ...(2)$$

Dividing (2) by (1), we get

$$q = 3/4,$$

$$\therefore \quad p = 1 - q = 1 - (3/4) = 1/4.$$

$$\therefore \quad \text{From (1)}, \quad n = \frac{4}{p} = \frac{4}{(1/4)} = 16.$$

Thus the required binomial distribution is

$$N(q+p)^n = N\left(\frac{3}{4} + \frac{1}{4}\right)^{16}$$

$$= N\left[\left(\frac{3}{4}\right)^{16} + {}^{16}C_1\left(\frac{3}{4}\right)^{15}\left(\frac{1}{4}\right) + {}^{16}C_2\left(\frac{3}{4}\right)^{14}\left(\frac{1}{4}\right)^2 + \ldots + {}^{16}C_r\left(\frac{3}{4}\right)^{16-r}\left(\frac{1}{4}\right)^r + \ldots + \left(\frac{1}{4}\right)^{16}\right].$$

Now mode of this binomial distribution = integal part of $(np + p)$

$$= \text{integral part of}\left(4 + \frac{1}{4}\right) = 4.$$

Example 14. ***If $p = \frac{1}{4}$ and $n = 7$, then compute the mode of binomial distribution.***

Solution. Here $np + p = 7 \times \frac{1}{4} + \frac{1}{4} = 2$, which is a positive integer. Hence the given binomial distribution $(q+p)^n$ has two modes and they are :

$$np + p = 2 \quad \text{and} \quad np + p - 1 = 1.$$

Example 15. ***By differentiating the following identity with respect to p and then multiplying by p***

$$\sum_{r=0}^{n} {}^nC_r\ p^r q^{n-r} = (q+p)^n, \quad q = 1 - p$$

show that $$\mu_1' = np.$$

Solution. Given that

$$\sum_{r=0}^{n} {}^nC_r\ p^r (1-p)^{n-r} = (q+p)^n \qquad ...(1)$$

Differentiating (1) with respect to p, we have

$$\sum_{r=0}^{n} {}^nC_r\ [rp^{r-1}(1-p)^{n-r} - p^r(n-r)(1-p)^{n-r-1}]$$

$$= n(q+p)^{n-1}\left(\frac{dq}{dp} + 1\right) = 0 \quad [\because\ q = 1 - p \Rightarrow dq/dp = -1]$$

Now multiplying by p,

$$\sum_{r=0}^{n} {}^nC_r\ rp^r q^{n-r} - \sum_{r=0}^{n} {}^nC_r\ (n-r)\ p^{r+1} q^{n-r-1} = 0$$

or $$\sum_{r=0}^{n} {}^nC_r\ p^r q^{n-r} \cdot r = np \sum_{r=0}^{n-1} {}^{n-1}C_r\ p^r q^{n-r-1} = np(p+q)^{n-1}$$

$$\therefore \quad \mu_1' = np. \qquad [\because\ p + q = 1]$$

Example 16. ***In a precision bombing attack there is a 50% chance that any one bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs must be dropped to give a 99% chance or better of completely destroying the target ?***

Solution. We have $p = 50/100 = 1/2$.

The meaning of 99% chance or better is that the probability must be greater than $0 \cdot 99$.

If n bombs are dropped, then out of them 2 may succeed, 3 may succeed and so on.

$$\therefore \quad {}^nC_2\left(\frac{1}{2}\right)^n + {}^nC_3\left(\frac{1}{2}\right)^n + \ldots + {}^nC_n\left(\frac{1}{2}\right)^n \geq 0 \cdot 99$$

i.e., $$\left(\frac{1}{2}\right)^n [{}^nC_2 + {}^nC_3 + \ldots + {}^nC_n] \geq 0 \cdot 90$$

or $$\frac{2^n - n - 1}{2^n} \geq 0 \cdot 99$$

or $$1-\frac{1+n}{2^n}\geq 1-0\cdot 01 \text{ or } \frac{1+n}{2^n}\leq 0\cdot 01$$

or $$100+100n-2^n\leq 0$$

or $$2^n\geq 100n+100.$$

By trial method, $n=11$ will give this inequality.

Example 17. ***The mean and variance of a binomial distribution $P(X, n, p)$ are 4 and 4/3 respectively. Find $P(X\geq 2)$ and the probability of two successes.***

Also find $P(x>2)$ and $P(x\geq 3)$.

Solution. Let $P(X=x)={}^nC_x\ p^xq^{n-x},\ x=0,1,2,\ldots,n.$

Here Mean, $np=4$ ...(1)

and Variance, $npq=\frac{4}{3}$.

$$\therefore\quad q=\frac{npq}{np}=\frac{\frac{4}{3}}{4}=\frac{1}{3}.$$

So $$p=1-q=1-\frac{1}{3}=\frac{2}{3}.$$

From (1), $$n=\frac{4}{p}=\frac{4}{\frac{2}{3}}=6.$$

Thus, we have :

(i) The probability of two successes *i.e.*,

$$P(X=2)={}^6C_2\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^4=\frac{6\times 5\times 2^2}{1\times 2\times 3^6}=\frac{20}{243}.$$

(ii) The probability of two and more than two successes *i.e.*,

$$P(X\geq 2)=1-P(X<2)=1-\sum_{x=0}^{1}{}^6C_x\left(\frac{2}{3}\right)^x\left(\frac{1}{3}\right)^{6-x}2$$

$$=1-\left[{}^6C_0\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^{6-0}+{}^6C_1\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^{6-1}\right]$$

$$=1-\left[\left(\frac{1}{3}\right)^6+6\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^5\right]=1-\frac{13}{729}=\frac{716}{729}.$$

(iii) The probability of more than two successes *i.e.*,

$$P(X>2)=1-P(X\leq 2)=1-\sum_{x=0}^{2}{}^6C_x\left(\frac{2}{3}\right)^x\left(\frac{1}{3}\right)^{6-x}$$

$$=1-\left[\left(\frac{1}{3}\right)^6+6\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^5+\frac{6\cdot 5}{1\cdot 2}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^4\right]$$

$$=1-\frac{73}{729}=\frac{656}{729}.$$

(iv) The probability of 3 or more than 3 successes *i.e.*, $P(x\geq 3)=P(X>2)=\frac{656}{729}$.

Example 18. ***Let X be a binomially distributed random variable with mean 10 and variance 5. Show that***

(a) $P(X>6)=\left(\frac{1}{2}\right)^{20}\sum_{r=7}^{20}{}^{20}C_r$. (b) $P(3<X<12)=\left(\frac{1}{2}\right)^{20}\sum_{r=4}^{11}{}^{20}C_r$.

Solution. Here mean $np=10$ and variance, $npq=5$.

$$q=\frac{npq}{np}=\frac{1}{2},$$

$$\therefore\quad p=1-q=\frac{1}{2},\quad n=\frac{10}{p}=20$$

$$\therefore\quad P(X=x)={}^{20}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{20-x},\ x=0,1,2,\ldots,20$$

and

$$={}^{20}C_x\left(\frac{1}{2}\right)^{20},\ x=0,1,2,\ldots,20.$$

(a) $$P(X>6)=\sum_{r=7}^{20}{}^{20}C_r\left(\frac{1}{2}\right)^{20}=\left(\frac{1}{2}\right)^{20}\sum_{r=7}^{20}{}^{20}C_r.$$

(b) $$P(3<X<12)=\sum_{r=4}^{11}{}^{20}C_r\left(\frac{1}{2}\right)^{20}=\left(\frac{1}{2}\right)^{20}\sum_{r=4}^{11}{}^{20}C_r.$$

Example 19. ***The following results were obtained when 80 batches of 10 seeds were allowed to germinate on a wet paper $\beta_1=\frac{2}{3}$ and $\beta_2=\frac{8}{3}$. Find the binomial distribution.***

Solution. For binomial distribution, we have

$$\beta_1=\frac{(q-p)^2}{npq}\quad\text{and}\quad\beta_2=3+\frac{1-6pq}{npq}.$$

Here $$\frac{(q-p)^2}{npq}=\frac{2}{3}\qquad\left[\because\ \beta_1=\frac{2}{3}\right]\qquad\text{...(1)}$$

and $$3+\frac{1-6pq}{npq}=\frac{8}{3}\qquad\left[\because\ \beta_2=\frac{8}{3}\right]$$

i.e., $$\frac{1-6pq}{npq}=\frac{8}{3}-3=-\frac{1}{3}.\qquad\text{...(2)}$$

Dividing (1) by (2), we have

$$\frac{(q-p)^2}{1-6pq}=-2\Rightarrow(q-p)^2=-2+12pq$$

$$\Rightarrow\quad(1-2p)^2=-2+12p(1-p)\qquad[\because\ q=1-p]$$

$$\Rightarrow\quad 16p^2-16p+3=0\Rightarrow(4p-1)(4p-3)=0\Rightarrow p=\frac{1}{4}\text{ or }p=\frac{3}{4}.$$

When $p=\frac{1}{4}$, then $q=1-\frac{1}{4}=\frac{3}{4}$. So (1) gives

$$\frac{\left(\frac{3}{4}-\frac{1}{4}\right)^2}{n\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)}=\frac{2}{3} \Rightarrow n=2.$$

Binomial frequency distribution $= N(q+p)^n = 80\left(\frac{3}{4}+\frac{1}{4}\right)^2$.

The required binomial distribution $=(q+p)^n=\left(\frac{3}{4}+\frac{1}{4}\right)^2$

i.e., $$P(X=x)={}^2C_x\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{2-x}, \quad x=0,1,2.$$

Similarly, when $p=\frac{3}{4}$, we have

$$P(X=x)={}^nC_x\left(\frac{3}{4}\right)^x\left(\frac{1}{4}\right)^{2-x}, \quad x=0,1,2.$$

Example 20. ***Prove that in the binomial distribution difference of mean and mode is not greater than unity.***

Solution. We have mean $= np$ (for binomial distribution).

From Example 10 and 11 above, we have three possibilities for mode :

(i) If np is a positive integer then mode and mean are equal.

$\therefore$ $|\text{mean} - \text{mode}| = 0 < 1.$

(ii) If $np + p$ is a positive integer then mode is $np + p$.

$\therefore$ $|\text{mean} - \text{mode}| = |np - np - p| = p < 1$

(iii) If $np + p$ is not a positive integer, then

$$np + p - 1 \le r \le np + p \Rightarrow p - 1 \le r - np \le p$$

$\Rightarrow$ $-p \le \text{mode} - \text{mean} \le p \Rightarrow |\text{mode} - \text{mean}| \le 1.$

Example 21. ***If a coin is tossed N times where N is very large even number, show that the probability of getting exactly*** $\frac{1}{2}N - p$ ***heads and*** $\frac{1}{2}N + p$ ***tails is approximately*** $\left(\frac{2}{\pi N}\right)^{1/2} e^{-2p^2N}$.

Solution. We know that the probability of getting a head in a single throw is $\frac{1}{2}$ and the probability of getting a tail is also $\frac{1}{2}$.

$\therefore$ The probability of getting 0, 1, 2, ..., heads in N trials are the successive terms in the binomial expansion of $\left(\frac{1}{2}+\frac{1}{2}\right)^N$.

Thus we may take $N = 2k$, where k is a very large number since N is a very large even number.

The probability of getting $\left(\frac{1}{2}N - p\right)$ heads and $\left(\frac{1}{2}N + p\right)$ tails in N tosses is

$$f_p = {}^NC_{N/2-p}\left(\frac{1}{2}\right)^{\frac{1}{2}N-p}\left(\frac{1}{2}\right)^{\frac{1}{2}N+p} = {}^NC_{N/2-p}\left(\frac{1}{2}\right)^N = {}^{2k}C_{k-p}\left(\frac{1}{2}\right)^{2k}. \quad ...(1)$$

Again the probability of getting $\frac{1}{2}N$ heads and $\frac{1}{2}N$ tails in N tosses is

$$f_0 = {}^{2k}C_k\left(\frac{1}{2}\right)^{2k} \quad ...(2)$$

Dividing (1) by (2), we get

$$\frac{f_p}{f_0}=\frac{{}^{2k}C_{k-p}}{{}^{2k}C_k}=\frac{k!\,k!}{(k+p)!\,(k-p)!}.$$

By using Stirling's formula, *viz.*, $n! = \sqrt{2\pi}\,e^{-n}\,n^{n+\frac{1}{2}}$ (for very large n), we have

$$\frac{f_p}{f_0}=\frac{[\sqrt{(2\pi)}\,e^{-k}\,k^{k+1/2}]\,[\sqrt{(2\pi)}\,e^{-k}\,k^{k+1/2}]}{[\sqrt{(2\pi)}\,e^{-(k+p)}\,(k+p)^{k+p+1/2}]\,[\sqrt{(2\pi)}\,e^{-(k-p)}\,(k-p)^{k-p+1/2}]}$$

$$=\frac{1}{\left(1+\frac{p}{k}\right)^{k+p+1/2}\left(1-\frac{p}{k}\right)^{k-p+1/2}}=\frac{1}{A}\text{ (say)}$$

$\therefore$ $$\log_e A=\left(k+p+\frac{1}{2}\right)\log_e\left(1+\frac{p}{k}\right)+\left(k-p+\frac{1}{2}\right)\log_e\left(1-\frac{p}{k}\right)$$

$$=\left(k+p-\frac{1}{2}\right)\left(\frac{p}{k}-\frac{1}{2}\frac{p^2}{k^2}+\ldots\right)+\left(k-p+\frac{1}{2}\right)\left(-\frac{p}{k}-\frac{1}{2}\frac{p^2}{k^2}\ldots\right)$$

$$=\frac{p^2}{k}+0\left(\frac{1}{k^2}\right).$$

$\therefore$ $A = e^{p^2/k} = e^{2p^2/N}$ nearly.

$\therefore$ $f_p = f_0\,e^{-2p^2/N}$.

Now $$f_0=\left(\frac{1}{2}\right)^{2k}\frac{(2k)!}{(k)!\,(k)!}=\left(\frac{1}{2}\right)^{2k}\frac{\sqrt{(2\pi)}\,e^{-2k}\,(2k)^{2k+1/2}}{[\sqrt{(2\pi)}\,e^{-k}\,k^{k+1/2}]^2}=\left(\frac{1}{\pi k}\right)^{1/2}=\left(\frac{2}{\pi N}\right)^{1/2}$$

$\therefore$ $$f_p=\left(\frac{2}{\pi N}\right)^{1/2}e^{-2p^2N}.$$

Example 22. ***Show that if two symmetrical binomial distributions of degree n (the same number of observations) are so superposed that the rth term of one coincides with the (r + 1)th term of the other, the distribution formed by adding superposed terms is a symmetrical binomial distribution of degree (n + 1).***

Solution. For symmetrical distribution $p = q = \frac{1}{2}$.

If the same number of observations be N, then the binomial distribution of order n is $N\left(\frac{1}{2}+\frac{1}{2}\right)^n$,

i.e., $$N\left[{}^nC_0\left(\frac{1}{2}\right)^n + {}^nC_1\left(\frac{1}{2}\right)^{n-1}\cdot\frac{1}{2} + \ldots + {}^nC_{r-1}\left(\frac{1}{2}\right)^{n-r+1}\cdot\left(\frac{1}{2}\right)^{r-1} + {}^2C_r\left(\frac{1}{2}\right)^{n-r}\cdot\left(\frac{1}{2}\right)^r + \ldots + {}^nC_n\left(\frac{1}{2}\right)^n\right].$$

Hence the total frequency distribution :

$$N = \left(\frac{1}{2}\right)^n N[{}^nC_0 + {}^nC_1 + \ldots + {}^nC_{r-1} + {}^nC_r + \ldots + {}^nC_n]. \quad \ldots(1)$$

In the above distribution, another similar distribution is superposed whose $(r+1)$th term coincides with the rth term of the other, *i.e.*,

$$N = \left(\frac{1}{2}\right)^n N[{}^nC_0 + {}^nC_1 + \ldots + {}^nC_{r-1} + {}^nC_r + \ldots + {}^nC_n] \quad \ldots(2)$$

So, the resultant distribution :

$$2N = \left(\frac{1}{2}\right)^n N[{}^nC_0 + ({}^nC_1 + {}^nC_0) + ({}^nC_2 + {}^nC_1) + \ldots + ({}^nC_r + {}^nC_{r-1}) + \ldots + ({}^nC_n + {}^nC_{n-1}) + {}^nC_n]$$

$$= \left(\frac{1}{2}\right)^n N[{}^{(n+1)}C_0 + {}^{(n+1)}C_1 + {}^{(n+1)}C_2 + \ldots + {}^{(n+1)}C_r + \ldots + {}^{(n+1)}C_n + {}^{(n+1)}C_{n+1}]$$

$$[\because {}^nC_0 = {}^{(n+1)}C_0,\ {}^nC_r + {}^nC_{r+1} = {}^{(n+1)}C_{r+1},\ {}^nC_n = {}^{(n+1)}C_{n+1}]$$

or $$N = \left(\frac{1}{2}\right)^{n+1} N[{}^{(n+1)}C_0 + {}^{(n+1)}C_1 + \ldots + {}^{(n+1)}C_r + \ldots + {}^{(n+1)}C_{n+1}]$$

$$= N\left(\frac{1}{2}+\frac{1}{2}\right)^{n+1}.$$

Hence the resultant distribution is a symmetrical binomial distribution of order $(n+1)$.

EXERCISE 6 (A)

1. Show that the measure of skewness of the binomial distribution is given by $\frac{q-p}{(npq)^{1/2}}$ and its kurtosis is $3 + \frac{1-6pq}{npq}$.

   [**Hint.** Measure of the skewness is $\sqrt{\beta_1}$ and of kurtosis β_2].

2. Is the statement "the mean and variance of binomial distribution are respectively 6 and 9" true?

3. The mean of binomial distribution is 20 and its standard deviation is 4. Determine its total frequency.

4. (a) What do you understand by binomial distribution ? Find the constants of binomial distribution.

   (b) Calculate the mean and standard deviation of binomial distribution.

5. Determine the mode of a binomial distribution with $p = \frac{1}{4}$ and $n = 7$.

6. Ten coins are tossed 1024 times and the following frequencies observed. Compare these frequencies with the expected frequencies :

x :	0	1	2	3	4	5	6	7	8	9	10
y :	2	10	38	106	188	257	226	128	59	7	3

7. Seven coins are tossed and the number of heads are noted. This experiment is repeated 128 times and the following distribution is obtained :

x :	0	1	2	3	4	5	6	7	Total
y :	7	6	19	35	30	23	7	1	128

   Fit the binomial distribution considering the coin is unbiased.

8. The following data due to Weldon show the results of throwing 12 dice 4096 times, a throw of 4, 5 or 6 being called a success

Successes	Frequency	Successes	Frequency
0	—	7	847
1	7	8	537
2	60	9	257
3	198	10	71
4	430	11	11
5	731	12	—
6	948	—	—
		Total	4096

   Fit a binomial distribution and calculate the expected frequencies. Compare the actual mean and standard deviation with those of the expected distribution.

9. One card was drawn from a pack of playing cards and then replaced and the pack is shuffled. This was replaced 10 times and the number of black suit cards was noted. One thousand results were obtained in this way and are given below, the number of black cards being denoted by x :

x :	0	1	2	3	4	5	6	7	8	9	10	Total
y :	2	8	46	116	211	243	208	119	40	7	9	1000

   What distribution would you expect to apply ? Calculate the theoretical frequencies.

10. Let X follow a binomial distribution $b(x, n, p)$ and r is a non-zero integer. If rth moment about the origin is

$$\mu_r' = E(X^r),$$

prove that $\mu_{r+1}' = npq\mu_r' + p(1-p)\frac{d\mu_r'}{dp}$.

11. Show that

(i) $b(n, p, k) = b(n, 1-p; n-k)$;

(ii) $\sum_{k=r}^{n} b(n, p; k) = 1 - \sum_{k=n-r+1}^{n} b(n, 1-p; k)$

(iii) $b(n+1, p; k) = pb(n, p; k-1) + qb(n, p; k)$.

[**Hint.** (i) $b(n, 1-p; n-k) = {}^nC_{n-k}(1-p)^{n-k}p^{n-(n-k)}$.

(ii) $\sum_{k=r}^{n} b(n, p; k) = \sum_{k=r}^{n} b(n, 1-p; n-k) = \sum_{k=0}^{n-r} b(n, 1-p; k)$]

12. Find mean, variance and r_1 for binomial distribution.

13. The probability of entering student in chartered accountant will graduate is 0.4. Determine the probability that out of 5 students :

(i) none (ii) one and (iii) at least one will graduate.

ANSWERS

2. False **3.** 100 **5.** 201

6. 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1.

7. $128(0.517 + 0.483)^7$; 1, 8, 23, 36, 33, 19, 6, 1.

8. Expected frequencies 1, 12, 66, 220, 495, 792, 924, 792, 495, 220, 66, 1;
Expected mean = 6; Real mean = 6.139; Expected σ = 1.732; Real σ = 1.712.

13. (i) 0.08 (ii) 0.06 (iii) 0.92.

◈ § 6.5. POISSON'S DISTRIBUTION

The Poisson's distribution is a particular limiting form of the binomial distribution when p (or q) is very small and n is large so that the average number of successes np is a finite constant m (say).

We know that, in the binomial distribution, the probability of r successes is given by

$$b(r, n, p) = {}^nC_r\ p^r q^{n-r}$$

$$= {}^nC_r\ p^r(1-p)^{n-r} \qquad [\because\ p+q=1]$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!}p^r\left(1-\frac{np}{n}\right)^{n-r}$$

$$= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)}{r!} \times \frac{(np)^r}{\left(1-\frac{np}{n}\right)^r}\left(1-\frac{np}{n}\right)^n$$

$\therefore$ $P(r)$ = the probability of r successes in Poisson's distribution

$$= \lim_{p\to 0,\ n\to\infty,\ np=m} b(r; n, p)$$

$$= \lim_{n\to\infty} \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)}{r!} \times \frac{m^r}{\left(1-\frac{m}{n}\right)^r}\left(1-\frac{m}{n}\right)^n$$

$$= \frac{m^r \cdot e^{-m}}{r!} \left[\because \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = \lim_{n\to\infty}\left\{\left(1-\frac{m}{n}\right)^{-nm}\right\}^{-m} = e^{-m}\right.$$

$$\left.\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^r = (1-0)^r = 1\right]$$

It is called *Poisson's distribution.*

Therefore, the chances of $0, 1, 2, \ldots, r$ successes are respectively $e^{-m}, \frac{me^{-m}}{1!}, \frac{m^2e^{-m}}{2!}, \ldots, \frac{m^re^{-m}}{r!}$.

$\therefore$ **The limiting form of binomial distribution *i.e.* $(q+p)^n$ where $p\to 0$, $n\to\infty$ so that $np = m$ is called Poisson's distribution.**

Definition. The probability distribution of a random variable x is called Poisson's distribution if x can assume non-negative values only and its distribution is given by

$$P(x=r) = \begin{cases} \frac{e^{-m}m^r}{r!}, & r = 0, 1, 2, \ldots \\ 0, & r \neq 0, 1, 2, \ldots \end{cases}$$

Note 1. m is known as parameter of Poisson's distribution.

Note 2. Following are some examples of Poisson variate :

(a) The number of deaths in a city in one year by a rat disease or by heart attack or by cancer.

(b) The number of telephone calls per minute in a switch board.

(c) The number of suicides in a city in one year.

(d) The number of accidents in a district in one year.

(e) The number of misprints on a page of a book.

(f) The number of defective blades in a packet of 100 blades.

(g) The number of cars passing through a certain street in time t.

§ 6.6. CONSTANTS OF THE POISSON'S DISTRIBUTION

Moments about the Origin

(i) First moment about the origin :

$$\mu_1' = \sum_{r=0}^{\infty} e^{-m}\frac{m^r}{r!}.r = \sum_{r=0}^{\infty}\frac{e^{-m}m^r}{(r-1)!} = e^{-m}\left(m + \frac{m^2}{1!} + \frac{m^3}{2!} + \ldots\right)$$

or $$\mu_1' = m\,.\,e^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots\right) = me^{-m}e^{m} = m$$

i.e., mean $= \mu_1' = m$.

(ii) Second moment about the origin :

$$\mu_2' = \sum_{r=0}^{\infty} e^{-m}\frac{m^r}{r!}r^2 = \sum_{r=0}^{\infty} e^{-m}\frac{m^r}{r!}\{r(r-1)+r\}$$

$$= \sum_{r=0}^{\infty}\frac{e^{-m}m^r}{(r-2)!} + \sum_{r=0}^{\infty}\frac{e^{-m}m^r}{(r-1)!}$$

$$= \sum_{r=0}^{\infty}\frac{m^2e^{-m}m^{r-2}}{(r-2)!} + \sum_{r=0}^{\infty}\frac{me^{-m}m^{r-1}}{(r-1)!}$$

$$= m^2e^{-m}\sum_{r=0}^{\infty}\frac{m^{r-2}}{(r-2)!} + me^{-m}\sum_{r=0}^{\infty}\frac{m^{r-1}}{(r-1)!}$$

$$= m^2e^{-m}e^{m} + me^{-m}e^{m} = m^2 + m.$$

(iii) Third moment about the origin :

$$\mu_3' = \sum_{r=0}^{\infty} e^{-m}\frac{m^r}{r!}r^3 = \sum_{r=0}^{\infty}\left[e^{-m}\frac{m^r}{r!}\{r(r-1)(r-2) + 3r(r-1) + r\}\right]$$

$$= m^3 + 3m^2 + m.$$

(iv) Fourth moment about the origin :

$$\mu_4' = \sum_{r=0}^{\infty} e^{-m}\frac{m^r}{r!}r^4$$

$$= \sum_{r=0}^{\infty}\left[e^{-m}\frac{m^r}{r!}\{r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r\}\right]$$

$$= m^4 + 6m^3 + 7m^2 + m.$$

Moments about the Mean

(v) First moment about the mean :

$= \mu_1 = 0.$ [by definition]

(vi) Second moment about the mean :

$$\mu_2 = \mu_2' - \mu_1'^2 = (m^2 + m) - (m)^2 = m.$$

i.e., Variance $= \mu_2 = m$

$\therefore$ S.D. $= \sqrt{m}$

Mean $= m = (\sqrt{m})^2 = (\text{S.D.})^2$.

(vii) Third moment about the mean :

$$\mu_3 = \mu_2' - 3\mu_2'\mu_1' + 2\mu_2'^3 = (m^3 + 3m^2 + m) - 3(m^2 + m)\,m + 2(m)^3$$

$$= m^3 + 3m^2 + m - 3m^3 - 3m^2 + 2m^3$$

$$= m.$$

(viii) Fourth moment about the mean :

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

$$= (m^4 + 6m^3 + 7m^2 + m) - 4(m^3 + 3m^2 + m)\,m + 6(m^2 + m)(m)^2 - 3(m)^4$$

$$= 3m^2 + m.$$

Pearson's Coefficients :

(ix) $\beta_1 = \mu_3^2/\mu_2^3 = m^2/m^3 = 1/m$

(x) $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3m^2 + m}{m^2} = 3 + \dfrac{1}{m}$.

(xi) $\gamma_1 = \sqrt{\beta_1} = 1/\sqrt{M}$

(xii) $\gamma_2 = \beta_2 - 3 = 1/m$.

(xiii) Expectation and variance : Let X be Poisson variate; then

$$E(X) = \sum_{r=0}^{\infty} e^{-m}\cdot\frac{m^r}{r!}.r = me^{-m}\sum_{r=1}^{\infty}\frac{m^{r-1}}{(r-1)!} = me^{-m}.e^{m} = m$$

$$E(X^2) = \sum_{r=0}^{\infty} e^{-m}\cdot\frac{m^r}{r!}.r^2 = \sum_{r=0}^{\infty} e^{-m}\cdot\frac{m^r}{r!}\{r(r-1)+r\}$$

$$= m^2e^{-m}\sum_{r=2}^{\infty}\frac{m^{r-2}}{(r-1)!} + m = m^2e^{-m}.e^{m} + m = m^2 + m.$$

$\therefore$ Var $(X) = E(X^2) - \{E(X)\}^2 = (m^2 + m) - m^2 = m.$

(xiv) Expected frequency : Let

$$P(X = x) = \frac{m^x e^{-m}}{x!},\quad x = 0, 1, 2, \ldots, \infty$$

and N = Total frequency.

Then expected frequency for $(X = x) = N\,.\,P(X = x)$.

§ 6.7. FITTING A POISSON DISTRIBUTION

In order to fit a Poisson distribution first of all calculate the mean of the obtained distribution. Then find $P(X = 0) = p(0) = e^{-m}$.

The other probabilities may be calculated by the recursion formula

$$p(x+1) = \frac{m}{x+1}\,.\,p(x).$$

Note. Proof of the above recursion formula. We have

$$p(x) = \frac{m^x e^{-m}}{x!} \quad \ldots(1)$$

$$\therefore \quad p(x+1) = \frac{m^{x+1} e^{-m}}{(x+1)!} = \frac{m}{x+1} \cdot \frac{m^x e^{-x}}{x!} = \frac{m}{x+1} \cdot p(x). \quad [\text{using (1)}]$$

ILLUSTRATIVE EXAMPLES

Example 1. ***For Poisson's distribution, prove that*** $M\sigma\gamma_1\gamma_2 = 1$, ***where symbols have their usual meanings.***

Solution. Here $M = m, \sigma = \sqrt{m}, \gamma_1 = 1/\sqrt{m}, \gamma_2 = 1/m$

$$M\sigma\gamma_1\gamma_2 = m\sqrt{m}(1/\sqrt{m})(1/m) = 1.$$ **Proved.**

Example 2. ***For Poisson's distribution, prove that***

$$\sqrt{\beta_1}\,(\beta_2 - 3)\, m\sigma = 1$$

where symbols have their usual meaning.

Solution. Here $\beta_1 = 1/m,\ \beta_2 = 3 + (1/m),\ m = m, \sigma = \sqrt{m}.$

$$\therefore \quad \sqrt{\beta_1}\,(\beta_2 - 3)\, m\sigma = \frac{1}{\sqrt{m}}\left(3 + \frac{1}{m} - 3\right) m\sqrt{m} = 1.$$ **Proved.**

Example 3. ***Criticise the following statement :***

'The mean of a Poisson's distribution is 5 while the standard deviation is 2.

Solution. For Poisson's distribution we know that

$$\text{mean} = (\text{S.D.})^2 \Rightarrow 5 = (2)^2$$

which is absurd. Hence the given statement is false.

Example 4. ***For Poisson's distribution, prove that as m tends to infinity,*** γ_1 ***and*** γ_2 ***both approach to zero.***

Solution. $\gamma_1 = \sqrt{\beta_1} = \sqrt{\left(\frac{\mu_3^2}{\mu_2^3}\right)} = \sqrt{\left(\frac{m^2}{m^3}\right)} = \frac{1}{\sqrt{m}} \to 0 \text{ as } m \to \infty$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{3m^2 + m}{m^2} - 3 = \frac{m}{m^2}$$

$$= 1m \to 0 \text{ as } m \to \infty.$$

Example 5. (a) ***Fit Poisson's distribution to the following and calculate theoretical frequencies*** $(e^{-0.5} = 0.61)$ **:**

Deaths	:	0	1	2	3	4
Frequency	:	122	60	15	2	1

Solution. Here total frequency

$$(N) = \Sigma f = 122 + 60 + 15 + 2 + 1 = 200$$

$$\text{Mean } (m) = \frac{\Sigma fx}{N} = \frac{122 \times 0 + 60 \times 1 + 15 \times 2 + 2 \times 3 + 1 \times 4}{122 + 60 + 15 + 2 + 1}$$

$$= \frac{60 + 30 + 6 + 4}{200} = \frac{1}{2} = 0.5.$$

Now $e^{-m} = e^{-0.5} = 1 - (0.5) + \frac{1}{2}(0.5)^2 - \frac{1}{6}(0.5)^3 + \ldots$

$$= 1 - 0.5 + 0.125 - 0.0208 + \ldots$$

$$= 0.61 \text{ (approx.)}$$

$\therefore$ Therefore required Poisson distribution (or theoretical frequencies of r deaths)

$$= Ne^{-m}\frac{m^r}{r!} = 200 \times (0.61) \times \frac{(0.5)^r}{r!}.$$

Computation of theoretical frequencies :

r	P(X = r)	N.P (X = r) Expected frequency
0	$e^{-m} = 0.61$	$200 \times 0.61 = 122$
1	$me^{-m} = 0.305$	$200 \times 0.305 = 61$
2	$\frac{m^2}{2!}e^{-m} = 0.0762$	$200 \times 0.0762 = 15$
3	$\frac{m^3}{3!}e^{-m} = 0.0127$	$200 \times 0.127 = 2$
4	$\frac{m^4}{4!}e^{-m} = 0.0016$	$200 \times 0.0016 = 0$
Total	$1.0055 = 1$	200

Thus it gives frequencies 122, 61, 15, 2 and 0 respectively for $r = 0, 1, 2, 3$ and 4.

Example 5. (b) ***Fit a Poisson distribution to the following data and calculate the theoretical frequencies*** **:**

x	0	1	2	3	4
f	192	100	24	3	1

Solution. $N = \Sigma f = 192 + 100 + 24 + 3 + 1 = 320$

$$\Sigma fx = 192 \times 0 + 100 \times 1 + 24 \times 2 + 3 \times 3 + 1 \times 4 = 161.$$

$\therefore$ Mean $(m) = \frac{\Sigma fx}{N} = \frac{161}{320} = 0.5031.$

$\therefore$ Required Poisson distribution

$$= N \cdot e^{-m} \cdot \frac{m^r}{r!} = 320 \times e^{-0.5031} \frac{(0.5031)^r}{r!}$$

$$= 320 \times 0.6047 \frac{(0.5031)^r}{r!} = 193.489 \frac{(0.5031)^r}{r!}$$

r	N . P(r)	Theoretical Frequencies
0	$193.489\,(0.5031)^0 / 0!$	$193.489 = 193$
1	$193.489\,(0.5031)^1 / 1!$	$97.344 = 97$
2	$193.489\,(0.5031)^2 / 2!$	$24.487 = 24$
3	$193.489\,(0.5031)^3 / 3!$	$4.106 = 4$
4	$193.489\,(0.5031)^4 / 4!$	$0.516 = 1$

Example 6. ***Find the probability that at most 5 defective fuses will be found in a box of 200 fuses, if experience shows that 2 per cent of such fuses are defective.***

Solution. Here $p = 2\% = 2/100 = 1/50$, $n = 200$.

$\therefore$ $np = 200/50 = 4$.

Maximum number of defective fuses = 5.

$\therefore$ $r \leq 5$.

Also $e^{-4} = 1 - 4 + \frac{1}{2} \cdot 4^2 - \frac{1}{3!} \cdot 4^3 + \frac{1}{4!} \cdot 4^4 - \ldots = 0.0183$

$\therefore$ Required probability ($r \leq 5$)

$$= \sum_{r=0}^{5} \frac{4^r \cdot e^{-4}}{r!} = e^{-4}\left[1 + \frac{4}{1!} + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{4^5}{5!}\right]$$

$$= 0.0183[1 + 4 + 8 + 10.6667 + 10.6667 + 8.5333]$$

$$= 0.0183 \times 42.8667 = 0.7845.$$ Ans.

Example 7. ***In 1,000 extensive sets of trials for an event of small probability the frequencies f of the number x of successes are found to be***

x :	0	1	2	3	4	5	6	7
f :	305	365	210	80	28	9	2	1

Assuming it to be a Poissonian distribution, calculate its mean, variance and expected frequencies for the Poissonian distribution with same mean.

Solution. Here mean $(m) = \frac{\Sigma fx}{\Sigma f} = \frac{1201}{1000} = 1.2$ (approx.)

$$\text{Variance} = \sigma^2 = \frac{2719}{1000} - (1.2)^2 = 2.719 - 1.44 = 1.279$$

$$e^{-m} = e^{-1.2} = 1 - 1.2 + \frac{(1.2)^2}{2!} - \frac{(1.2)^3}{3!} + \ldots = 0.3012$$

x	f	fx	fx^2
0	305	0	0
1	365	365	365
2	210	420	840
3	80	240	720
4	28	112	448
5	9	45	225
6	2	12	72
7	1	7	49
Total	1000	1201	2719

The expected frequencies for $x = 0, 1, 2, 3, 4, 5, 6, 7$ are respectively calculated below :

$$Ne^{-m} = 0.3012 \times 1000 = 301.2 \qquad [\because N = 1{,}000]$$

$$Ne^{-m} \cdot m = 301.2 \times 1.2 = 361.4$$

$$Ne^{-m} \frac{m^2}{2!} = 301.2 \times \frac{(1.2)^2}{2!} = 216.8$$

$$Ne^{-m} \frac{m^3}{3!} = 301.2 \times \frac{(1.2)^3}{3!} = 86.7$$

$$Ne^{-m} \frac{m^4}{4!} = 301.2 \times \frac{(1.2)^4}{4!} = 26$$

$$Ne^{-m} \frac{m^5}{5!} = 301.2 \times \frac{(1.2)^5}{5!} = 6.2$$

$$Ne^{-m} \frac{m^6}{6!} = 301.2 \times \frac{(1.2)^6}{6!} = 1.2$$

$$Ne^{-m} \frac{m^7}{7!} = 301.2 \times \frac{(1.2)^7}{7!} = 0.2$$

Example 8. ***In a certain factory turning razor blades, there is a small chance (1/500) for any blade to be defective. The blades are in packets of 10. Use Poisson's distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 10,000 packets.***

Solution. Here $p = 1/500$, $n = 10$, $N = 10{,}000$.

$\therefore$ $m = np = 10(1/500) = 0.02$

Now $e^{-m} = e^{-0.02} = 1 - 0.02 + \frac{1}{2!}(0.02)^2 - \ldots = 0.9802$ (approx.)

The respective frequencies (*i.e.*, number of packets) containing no defective, one defective and two defective blades are given respectively as follows :

$$Ne^{-m}, Ne^{-m} \cdot m, Ne^{-m} \cdot \frac{1}{2} m^2$$

i.e., $10000 \times 0.9802;\ 10000 \times 0.9802 \times 0.02;\ 10000 \times 0.9802 \times \frac{1}{2}(0.02)^2$

i.e., 9802; 196; 2. **Ans.**

Example 9. *In 1000 consecutive issues of the 'Utopian Seven Daily Chronicle' the deaths of centenarians were recorded, the number x having frequency f according to the table*

x :	0	1	2	3	4	5	6	7	8
f :	229	325	257	119	50	17	2	1	0

Show that the distribution is roughly Poissonian by calculating its mean, and then the frequencies in the Poissonian distribution with the same mean and the same total frequency of 1000. Also calculate variance of the given distribution and compare it with the mean (given $e^{-1.5} = 0.2231$ approx).

Solution.

x	f	fx	fx^2	x	f	fx	fx^2
0	229	0	0	5	17	85	415
1	325	325	325	6	2	12	72
2	257	514	1028	7	1	7	49
3	119	357	1071	8	0	0	0
4	50	200	800				
		Total	$\Sigma f = 1000$			$\Sigma fx = 1500$	$\Sigma fx^2 = 3770$

The mean (m) of the series $= \Sigma\, fx / \Sigma\, f = 1500/1000 = 1.5$.

$$\text{Variance} = \sigma^2 = \frac{3770}{1000} - (1.5)^2 = 3.77 - 2.25 = 1.52$$

$$\therefore \quad \sigma = 1.2, \quad m = 1.5.$$

But in the Poisson distribution,

$$\sigma = \sqrt{m} = \sqrt{1.5} = 1.2.$$

It shows that the distribution is roughly Poissonian. Consequently, the frequencies are given by $Ne^{-m}(m^r / r!)$.

Here $Ne^{-m} = 1000 \times e^{-1.5} = 1000 \times 0.2231 = 223.1$.

Now the corresponding frequencies for $r = x = 0, 1, \ldots, 8$ are as follows :

$$223.1 \times \frac{m^0}{0!}, \quad 223.1 \times \frac{1.5}{1!}, \quad 223.1 \times \frac{(1.5)^2}{2!},$$

$$223.1 \times \frac{(1.5)^3}{3!}, \quad 223.1 \times \frac{(1.5)^4}{4!}, \quad 223.1 \times \frac{(1.5)^5}{5!},$$

$$223.1 \times \frac{(1.5)^6}{6!}, \quad 223.1 \times \frac{(1.5)^7}{7!}, \quad 223.1 \times \frac{(1.5)^8}{8!},$$

i.e., $223.1, 334.7, 251.0, 125.5, 47.1, 14.1, 3.5, 0.8, 0.2$.

Example 10. *Find the mean and standard deviation for the table of deaths of women over 80 years old recorded in a three year period. No. of deaths recorded*

In a day :	0	1	2	3	4	5	6	7
No. of days :	364	376	218	89	33	13	2	1

Find the expected number of days with one death recorded for the Poisson series fitted to the data.

Solution.

x	f	fx	fx^2	x	f	fx	fx^2
0	364	0	0	4	33	132	528
1	376	376	376	5	13	65	325
2	218	436	872	6	2	12	72
3	89	267	801	7	1	7	49
		Total			1096	1295	3023

Now N (the total frequency) $= 1096$.

The mean (m) of the series $= \Sigma\, fx / \Sigma\, f = 1295/1096 = 1.18$

$$\sigma^2 = \frac{3023}{1096} - \left(\frac{1295}{1096}\right)^2 = 2.758 - 1.396 = 1.362$$

$$\therefore \quad \sigma = 1.17 \text{ (aprox.)}$$

Again $$e^{-m} = e^{-1.18} = 1 - 1.18 + \frac{(1.18)^2}{2!} - \frac{(1.18)^3}{3!} + \ldots$$

$$= 0.316 \text{ (on simplification).}$$

$\therefore$ The required frequency for one death $= Ne^{-m} \cdot m$

$$= 1096 \times (0.316) \times (1.18) = 397.$$

Ans.

Example 11. *A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused.*

(given $e^{-1.5} = 0.2231$)

Solution. Given $m = 1.5, e^{-m} = e^{-1.5} = 0.2231$.

The proportion of days when neither car is used $= e^{-m} \cdot \frac{m^0}{0!} = e^{-m} = 0.2231$.

The probability when no car, one car and two cars will be used

$$= e^{-m}\frac{m^0}{0!} + e^{-m} \cdot \frac{m}{1!} + e^{-m} \cdot \frac{m^2}{2!} = e^{-m}\left[1 + m + \frac{m^2}{2}\right]$$

$$= 0.2231\,[1 + 1.5 + 1.25] = 0.8087375.$$

$\therefore$ The proportion of days on which some demand is refused

$$= 1 - 0.8087375 = 0.1912625.$$

Ans.

Example 12. *For a Poisson distribution with mean m, show that*

$$\mu_{r+1} = mr \cdot \mu_{r-1} + m\frac{d\mu_r}{dm} \text{ where } \mu_r = \sum_{x=0}^{\infty} (x-m)^r \frac{e^{-m}m^x}{x!}.$$

Solution. $$\frac{d\mu_r}{dm} = \frac{d}{dm}\left[\sum_{x=0}^{\infty} (x-m)^r \frac{e^{-m}m^x}{x!}\right]$$

$$= -r \sum_{x=0}^{\infty} (x-m)^{r-1} \frac{e^{-m} m^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-m)^r}{x!} [xm^{x-1}e^{-m} - m^x e^{-m}]$$

$$= -r\mu_{r-1} + \sum_{x=0}^{\infty} \frac{(x-m)^r e^{-m} m^{x-1}(x-m)}{x!}$$

$$\therefore \quad m\frac{d\mu_r}{dm} + mr\mu_{r-1} = \sum_{x=0}^{\infty} \frac{(x-m)^{r+1} e^{-m} m^x}{x!} = \mu_{r+1}. \qquad \textbf{Proved.}$$

Exercise. If m is the mean and μ_r is the rth central moment of a Poisson distribution, then prove that

$$\mu_{r+1} = mr\mu_{r-1} + \frac{d\mu_r}{dm}.$$

Example 13. ***In a Poisson distribution with unity mean, show that the mean deviation from mean is 2/e times the standard deviation.***

Solution. We know that the probability of x successes in Poisson distribution

$$= e^{-m} m^x / x!$$

Here mean $= m = 1$.

$\therefore$ S.D. $= \sqrt{m} = \sqrt{1} = 1$.

$P(x)$ = Probability of x successes $= e^{-1}/x!$ $[\because m = 1]$

Now mean deviation (M.D.) from mean

$$= \sum_{x=0}^{\infty} |x-m| P(x), \text{ where } m = 1$$

$$= \sum_{x=0}^{\infty} \frac{|x-1| e^{-1}}{x!} = \frac{1}{e} \sum_{x=0}^{\infty} \frac{|x-1|}{x!}$$

$$= \frac{1}{e}\left[1 + \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \frac{4}{5!} + \ldots\right] \qquad \ldots(1)$$

General term of the series in (1)

$$= \frac{n}{(n+1)!} = \frac{(n+1)-1}{(n+1)!} = \frac{1}{n!} - \frac{1}{(n+1)!} \qquad \ldots(2)$$

Substituting $n = 1, 2, 3, \ldots$, etc. in (1) and using (2) for different terms, the M.D. from mean

$$= \frac{1}{e}\left[1 + \left(\frac{1}{1!} - \frac{1}{2!}\right) + \left(\frac{1}{2!} - \frac{1}{3!}\right) + \ldots\right]$$

$$= \frac{1}{e}[1+1] \qquad \text{[remaining terms cancel in pairs]}$$

$$= 2e = (2/e) \times 1 = (2/e).\text{(S.D.)}. \qquad \textbf{Proved.}$$

Example 14. ***A telephone switch handles 600 calls on the average during a rush hour. The board can make a maximum 20 connections per minute. Use Poisson distribution to estimate the probability that the board will be over-taxed during any given minute*** $(e^{-10} = 0.00004539)$.

Solution. Mean (m) = number of calls per minute

$= 600/60 = 10$

$\therefore$ The probability for using 0 to 20 calls per minute,

$$P(x \le 20) = \sum_{r=0}^{20} e^{-m} \frac{m^r}{r!} = e^{-10} \sum_{r=0}^{20} \frac{10^r}{r!} = 0.00004539 \sum_{r=0}^{20} \frac{10^r}{r!}.$$

Thus the probability that the board will be over-taxed during any given minute

= the probability when the calls are more than 20 connections per minute during any minute

$$= 1 - 0.00004539 \sum_{r=0}^{20} \frac{10^r}{r!}. \qquad \textbf{Ans.}$$

Example 15. ***In a book of 300 pages, a proof reader finds no error in 200 pages, in 75 pages one error on each page, in 20 pages two errors on each page and in 5 pages 3 errors on each page. Use Poisson distribution to these data and calculate theoretical frequency.*** $[e^{-0.43} = 0.6505]$.

Solution. The frequency distribution is

No. of errors x :	0	1	2	3	Total
No. of pages f :	200	75	20	5	300

$\therefore$ mean $m = \Sigma fx / \Sigma f$

$$= \frac{0 \times 200 + 1 \times 75 + 2 \times 20 + 3 \times 5}{300} = 0.43 \text{ (nearly)}.$$

$\therefore$ From Poisson's distribution,

$$P(x) = e^{-m}(m^x/x!) = e^{-0.43} \frac{(0.43)^x}{x!} = 0.6505 \frac{(0.43)^x}{x!}$$

$\therefore$ The theoretical frequency for x errors

$$= Ne^{-m} \frac{m^x}{x!} = 300 \times 0.6505 \frac{(0.43)^x}{x!} \qquad \ldots(1)$$

For $x = 0, 1, 2, 3, \ldots$ the theoretical frequencies are [putting for x in (1)] :

$$195.15, 83.91, 18.04, 2.90.$$

Example. 16. ***(a) In a Poisson distribution, the probability of assuming the value $x = 0$ is 10%. Find the mean for the distribution.*** $[\log_e 10 = 2.3026]$.

(b) If $P(x = 0) = P(x = 1) = a$ in the Poisson distribution, then show that $a = 1/e$.

(c) If $P(x = 2) = 9P(x = 4) + 90P(x = 6)$ in the Poisson distribution, then find $E(x)$.

Solution. Let the probability function of the Poisson distribution be as follows :

$$P(x) = e^{-m} \cdot (m^x/x!), \quad x = 0, 1, 2, \ldots$$

(a) $P(x = 0) = p(0) = e^{-m} = 10\%$

$\therefore$ $e^{-m} = (10/100) = 1/10$

$\Rightarrow$ $e^m = 10 \Rightarrow m = \log_e 10 = 2.3026$

(b) $P(x=0) = P(x=1) \Rightarrow e^{-m} = me^{-m} \Rightarrow m = 1$

$\therefore$ $P(x=0) = e^{-1} = a$ (given).

$\therefore$ $a = 1/e$

(c) $P(x=2) = 9P(x=4) + 90P(x=6)$

$\Rightarrow \frac{m^2}{2!} \cdot e^{-m} = 9\frac{m^4}{4!} \cdot e^{-m} + 90\frac{m^6}{6!} \cdot e^{-m} \Rightarrow m^4 + 3m^2 - 4 = 0$

$\Rightarrow (m^2+4)(m^2-1) = 0$

$\therefore$ $m = 1$, *i.e.*, mean $E(x) = 1$.

Example 17. ***Obtain Poisson distribution as a limiting form of binomial distribution. If the mean of Poisson variate X be m, then show that the mean of variate*** $\frac{X-m}{\sqrt{m}}$ ***is zero and the variance is one.***

Solution. For first part see § 6.5.

Second part. From § 6.6 (xiii), we know that if X be a Poisson variate, then

$$E(X) = m \quad \text{and} \quad E(X^2) = m^2 + m.$$

$\therefore$ The expectation of variate $\frac{X-m}{\sqrt{m}}$ is given by

$$E\left(\frac{X-m}{\sqrt{m}}\right) = E\left(\frac{X}{\sqrt{m}} - \sqrt{m}\right) = E\left(\frac{X}{\sqrt{m}}\right) - \sqrt{m}$$

$$= \frac{1}{\sqrt{m}} E(X) - \sqrt{m} = \frac{1}{\sqrt{m}} \cdot m - \sqrt{m} = 0.$$

Also the expectation of $\left(\frac{X-m}{\sqrt{m}}\right)^2$ is given by

$$E\left(\frac{X-m}{\sqrt{m}}\right)^2 = E\left(\frac{X^2 - 2mX + m^2}{m}\right) = E\left(\frac{X^2}{m} - 2X + m\right)$$

$$= \frac{1}{m} E(X^2) - 2E(X) + m = \frac{1}{m}(m^2 + m) - 2m + m = 1.$$

$\therefore$ $\text{var}\left(\frac{X-m}{\sqrt{m}}\right) = E\left(\frac{E-m}{\sqrt{m}}\right)^2 - \left\{E\left(\frac{X-m}{\sqrt{m}}\right)\right\}^2 = 1 - 0 = 1.$

Hence the mean of variate $\frac{X-m}{\sqrt{m}} = 0$ and variance $= 1$.

Example 18. ***If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals***

(i) exactly 3 ***(ii) more than 2 individuals***

(iii) none ***(iv) more than 1 individuals***

will suffer a bad reaction.

Solution. The probability of occurrence is very small, therefore it follows a Poisson distribution.

Mean $m = np = 2000 \times (0 \cdot 001) = 2$.

(i) Probability that exactly 3 suffer a bad reaction

$$= \frac{m^3 e^{-m}}{3!} = \frac{8e^{-2}}{6} = \frac{4}{3e^2} = 0 \cdot 180$$

(ii) Probability that more than 2 suffer a bad reaction

= 1 – [probability that no one suffers a bad reaction + probability that one suffers a bad reaction + probability that 2 suffer a bad reaction]

$$= 1 - \left[e^{-m} + \frac{m^1 e^{-m}}{1!} + \frac{m^2 e^{-m}}{2!}\right] = 1 - \left[\frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2}\right]$$

$$= 1 - \frac{5}{e^2} = 0 \cdot 323$$

(iii) Probability that none suffers a bad reaction $= e^{-m} = 1/e^2 = 0 \cdot 135$

(iv) Probability that more than 1 suffer a bad reaction

= 1 – [probability that no one suffers a bad reaction + probability that one suffers a bad reaction]

$$= 1 - \left[e^{-m} + \frac{m^1 e^{-m}}{1!}\right] = 1 - \left[\frac{1}{e^2} + \frac{2}{e^2}\right]$$

$$= 1 - \frac{3}{e^2} = 0 \cdot 594.$$

§ 6.8. MODE OF THE POISSON DISTRIBUTION

Definition. *That value of r which has a greater probability than any other, is called the mode of the Poisson distribution. Therefore,*

$$\frac{m^{r-1}e^{-m}}{(r-1)!} \le \frac{m^r e^{-m}}{r!} \le \frac{m^{r+1}e^{-m}}{(r+1)!}$$

$\Rightarrow$ $m \ge r \ge m-1 \Rightarrow m-1 \le r \le m.$

Clearly we have the following two cases :

Case I. If m is a +ve integer, then $m-1$ and m are two modes.

Case II. If m is not an integer then the integral part of m is the mode, *i.e.*, mode is the integral value between $m-1$ and m.

Example 19. ***A Poisson distribution has a double mode at x = 3 and x = 4; what is the probability that x will have one or the other of these two values ?***

Solution. We know that when m is a +ve integer, then there are two modes $m-1$ and m.

$\therefore$ $m - 1 = 3 \Rightarrow m = 4.$

$\therefore$ Using the formula

$$P(x) = e^{-m}(m^x / x!), \quad x = 0, 1, 2, \ldots \infty.$$

We have $P(x=3) = e^{-4} \cdot 4^3 / 3!$

and $P(x=4) = e^{-4} \cdot 4^4 / 4! = e^{-4} \cdot 4^3 / 3!$

$\therefore$ The required probability

$$P(x = 3 \text{ or } 4) = P(x = 3) + P(x = 4) = 2e^{-4} \cdot 4^3/3! = (64/3)e^{-4}.$$ **Ans.**

EXERCISE 6 (B)

1. Six coins are tossed 6400 times. Using Poisson's distribution find approximate probability of getting 6 heads x times.
2. Fit a Poisson's distribution for the following data :

x :	0	1	2	3	4	5	6	7	8	9	10
y :	103	143	98	42	8	4	2	0	0	0	0

$(e^{-1\cdot32} = 0\cdot2674)$

[Hint. Proceed exactly as Example 5 (a) of § 6.7]

3. Criticise the following statement :

'The mean of a Poisson distribution is 7, while the S.D. is 4.'

4. If X be a Poisson variate with parameter m and μ_r be the rth central moment, then show that

$$m[{}^rC_1\,\mu_{r-1} + {}^rC_2\,\mu_{r-2} + \ldots + {}^rC_r\,\mu_0] = \mu_{r+1}.$$

5. Let X be Poisson variate with parameter $m > 0$. If r is a non-zero integer and $\mu_r' = E(x^r)$ then show that

$$\mu_{r+1}' = m\left(\mu_r' + \frac{d\mu_r'}{dm}\right).$$

6. From records of 10 Prussian Army Corps kept over 20 years the following data were obtained showing the number of deaths caused by the kicks of a horse. Calculate the theoretical Poisson frequencies :

No. of details :	0	1	2	3	4	Total
Frequencies :	109	65	22	3	1	200

$[e^{-0\cdot61} = 0\cdot5436]$.

7. A skilled typist on routine work, kept a record of mistakes made per day during 300 working days. Compute the frequencies of the Poisson series, which has the same total frequency and mean as this distribution:

Mistakes per day :	0	1	2	3	4	5	6	Total
No. of days :	143	90	42	12	9	3	1	300

8. A book of 500 pages contains 500 misprints; estimate the probability that a given page contains at least three misprints.
9. A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality ?

$[e^{-5} = 0\cdot006738]$

10. Letters were received in an office on each of 100 days. Assuming the following data to form a random sample from a Poisson distribution, find the expected frequencies, correct to the nearest unit, taking $e^{-4} = 0\cdot0183$.

No. of letters :	0	1	2	3	4	5	6	7	8	9	10
Frequencies :	1	4	15	22	21	20	8	6	2	0	1

11. Suppose the number of telephone calls on an operator received from 900 to 905 follows a Poisson distribution with mean 3. Find the probability that
    (i) the operator will receive no call in that time interval tomorrow
    (ii) in the next three days the operator will receive a total of 1 call in that time interval.

    $[e^{-3} = 0\cdot04978]$.
12. An area of 144 square kilometres was selected for which the mean density of bombs appeared constant. To test the hypothesis that the bombs fell in clusters the area was divided into 576 squares of $\frac{1}{4}$ kilometres each and a count made of the number of squares containing 0, 1, 2, ..., etc. bombs of which there were 537 altogether. Calculate the theoretical frequencies.

Number of flying bombs per square :	0	1	2	3	4	5	and over	Total
Actual no. of square	143	211	93	35	7	1		576

13. If X is a Poisson variate and $P(X = 1) = P(X = 2)$, then find $P(X = 4)$.
14. If a Poisson distribution has a double mode at $x = 1$ and $x = 2$, show that $P(x = 1) = e^{-2}$.
15. If x is the number of occurrence of Poisson variate with mean m, show that $P(x \geq n) = P(x \geq n + 1) = P(x = n)$.
16. If X is a Poisson variate with mean m, show that $E(x^2) = mE(x + 1)$

    If $m = 1$, then show that $E(|x - 1|) = 2/e$.
17. If X is a Poisson variate with mean m, then find the mathematical expectation of e^{-kx} and kxe^{-kx} where k is a constant.
18. Discuss in brief the Poisson Distribution.
19. Compute mean, deviation and β_1 and β_2 for Poisson distribution.
20. Define Poisson distribution. Obtain Poisson distribution as a limiting form of binomial distribution. State the restrictions when binomial distribution tends to Poisson distribution.
21. Find $\mu_1, \mu_2, \mu_3, \mu_4, \beta_1, \beta_2, \gamma_1$ and γ_2 for Poisson distribution.
22. Show that $M\sigma\gamma_1\gamma_2 = 1$ for Poisson distribution, where symbols have their usual meanings.
23. For Poisson distribution, prove the following :
    (i) $\beta_1 = 1/m$ (ii) $\beta_2 = 3 + (1/m)$.
24. Define Poisson distribution. Show that if X and Y are two independent Poisson variates, then $X + Y$ is also a Poisson variate.
25. Show the following for a Poisson distribution : $\sqrt{(\beta_1)}\,(\beta_2 - 3)m\sigma = 1$.
26. Write a short note on Poisson distribution.

ANSWERS

1. $e^{-100}\{(100)^x/x!\}$ 2. $m = 1\cdot32, e^{-132} = 0\cdot2674$; 107, 141, 93, 41, 14, 4, 1, 0, 0, 0, 0.
3. False statement.

For Normal Distribution

Please refer unit II, article 2.7io

□■□

Chapter UNIT-IV

7

CORRELATION AND REGRESSION

§ 7.1. CORRELATION

Consider the annual rainfall in a certain region and the agricultural yield so as to find some sort of relation between the two-whether increase in rainfall results in greater agricultural yield.

We may find that a change in one variable results in change of second variable or does not have any effect on second variable.

Definition. *Whenever there exists a relationship between two variables such that a change in one variable results in a positive or negative change in the other and also greater change in one variable results in a corresponding greater change in the other, the relationship is called correlation and the two variables are called correlated.*

According to Croxton and Cowden :

"When the relationship is of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as **correlation.**"

According to Professor King. *"If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. This relationship is called* **correlation.**"

Positive and Negative Correlation :

Definition. Two variables are called positively correlated if corresponding to an increase (or decrease) in one variable results in an increase (or decrease) in the other.

Two variables are called negatively correlated if corresponding to an increase (or decrease) in one variable results in decrease (or increase) in the other.

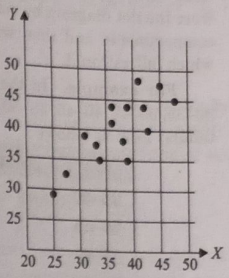
Examples. (i) Demand and price are positively correlated.

(ii) Supply and price are negatively correlated.

§ 7.2. SCATTER OR DOT-DIAGRAMS

Let x and y be two variables and suppose that there exists a correlation between them.

Let the points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$ correspond to the values of two variables. Plot these points on a graph paper referred to two perpendicular axes. Here the values of one of the variables (independent variable) are taken along x-axis and the other along y-axis. Such a graphical representation is said to be a **Scatter** or **Dot diagram**, in other words the diagram of dots so obtained is called a **scatter** or **dot diagram**.



For example. Following are data of marks of students in Analysis and Statistics at the B.C.A. examination, maximum marks of each subject being 50, to draw a dot diagram :

Roll No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Marks in Analysis	28	42	39	41	25	38	45	33	37	34	47	34	39	36	41
Marks in Statistics	32	40	35	47	39	43	47	39	38	35	45	36	43	41	43

Now plot the points (28, 32), (42, 40), (39, 35), ... on a graph paper as shown in the above figure. Here you are free to choose x and y. We have taken marks in Analysis as x's while marks in Statistics as y's.

By inspection of the graph, we find that there is tendency for small values of x to be related with small values of y and for large values of x to be related with large values of y.

Also the general trend of the dots is of a straight line.

§ 7.3. BIVARIATE DISTRIBUTION

Suppose we measure the heights and weight of a certain group of persons then here we have two variables – one variable relating to height and other relating to weight, such a distribution is called a bivariate distribution.

Definition. A universe, every member of which bears one of the values of each of two variates is called **bivariate distribution**.

§ 7.4. BIVARIATE FREQUENCY DISTRIBUTION

It may be possible that a particular value occurs more than once so that corresponding to one value (x_1, y_1) (say) there will be a number of dots. Suppose (x_1, y_1) occurs f_1 times then f_1 is called the frequency of (x_1, y_1). If these values are grouped according to the class-intervals then the frequency distribution so obtained is called a *bivariate frequency distribution.*

§ 7.5. CORRELATION TABLE

If the number of measurements is large, then it is convenient to choose some class intervals for the measurement of such variables. A table can be constructed from the dot diagram by sub-dividing the co-ordinate area into equal rectangular compartments and then writing within each compartment the number of dots which fall within it.

For example. In the example of § 8.2 above, if we take intervals 25–30, 30–35, 35–40, 40–45, 45–50 for the both x and y then the dots can be arranged in a tabular form as follows :

x / y	25–30	30–35	35–40	40–45	45–50
25–30	1				
30–35	1				
35–40		3	2		
40–45			3	2	
45–50				1	2

§ 7.6. (A) KARL PEARSON'S COEFFICIENT OF CORRELATION

It is the best mathematical method to find correlation since it is based on mean and standard deviation. By this method we can find not only the direction and magnitude of correlation but also its positive measure. Karl Pearson (1867–1936) developed a formula (in 1890) called correlation coefficient.

Correlation coefficient between two random variables X and Y, denoted by r, is a numerical measure of linear relationship between X and Y and is defined (by Karl Pearson) by

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

where $x = X - M_x$ = deviation of variable X measured from its mean M_x

$y = Y - M_y$ = deviation of variable Y measured from its mean M_y

σ_x = standard deviation of X-series

σ_y = standard deviation of Y-series

$p = \{\Sigma\,(xy)/\,n\}$

n = number of pairs of two variables

r is called the **product moment correlation coefficient.**

§ 7.6. (B) CHANGE OF ORIGIN AND SCALE

Consider two new variables defined by $u = \dfrac{X - X_0}{h}$, $v = \dfrac{Y - Y_0}{k}$, then

$$X = uh + X_0,\ Y = kv + Y_0.$$

If $\bar{u}$ and $\bar{v}$ denote the means of u series and v series respectively, then

Mean of X series $M_X = \bar{u}h + X_0$;

Mean of Y series $M_Y = \bar{v}k + Y_0$.

Also $\sigma_x^{\,2} = E(X - M_x)^2 = h^2\ E(u - \bar{u})^2 = h^2\sigma_u^{\,2}$.

Similarly, $\sigma_y^{\,2} = k^2\sigma_v^{\,2}$. Therefore, $\sigma_x = |h|\,\sigma_u$, $\sigma_y = |k|\,\sigma_v$ since σ_x and σ_y are always positive irrespective of the signs of h and k.

Now, $r_{XY} = \dfrac{\Sigma xy}{n\sigma_x\ \sigma_y} = \dfrac{\Sigma\,(X - M_X)\,(Y - M_Y)}{n\sigma_x\sigma_y}$ [See § 7.6 (A), take r_{XY} for r]

$$= \frac{hk\,\Sigma\,(u - \bar{u})\,(v - \bar{v})}{n\,|h|\,|k|\,\sigma_u\ \sigma_v} = \frac{hk}{|h|\,|k|}\,r_{uv}$$

$= r_{uv}$ [if h and k are of same sign, which is possible by changing the units of measurement only]

Hence coefficient of correlation is independent of X_0, Y_0 (change of origin) and h, k (change of scale).

Remark 1. If h and k are of opposite signs, then we have

$$r_{xy} = -r_{uv}.$$

Remark 2. Coefficient of correlation has no dimension and is a real number.

Remark 3. Since coefficient of correlation is independent of change of origin and scale, therefore, this property is useful in computation of r, as we can conveniently select any origin and scale.

Problem. Define coefficient of correlation. Show that the coefficient of correlation is independent of change of scale and origin of the variables.

§ 7.7. DEGREE OF CORRELATION

Since coefficient of correlation determines numerical measure of correlation, therefore they may be positive or negative.

Measure of Correlation at a Glance

S. No.	Measure of Correlation	Positive	Negative
1	Perfect	+ 1	− 1
2	High Degree	Between + 0·75 and + 1	Between − 0·75 and − 1
3	Mediate Degree	Between + 0·5 and + 0·75	Between − 0·5 and − 0·75
4	Low Degree	Between 0 and + 0·5	Between − 0 and − 0·5
5	No Correlation	0	0

ILLUSTRATIVE EXAMPLES

Example 1. ***The students got the following percentage of marks in principles of Economics and Statistics :***

Roll No. :	1	2	3	4	5	6	7	8	9	10
Marks in Economics :	78	36	98	25	75	82	90	62	65	39
Marks in Statistics :	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of Correlation.

Solution. Suppose the marks of two subjects are denoted by variables X and Y. Then the mean of X-series

$$= M_x = \Sigma x/\,n = 650/\,10 = 65$$

and the mean for Y-series $= M_y = \Sigma y/\,n = 660/\,10 = 66$.

If the deviations of X's and Y's from their respective means be x and y, then the data may be arranged as shown in the table :

X	Y	$x = X - M_x$	$y = Y - M_y$	x^2	y^2	xy
78	84	13	18	169	324	234
36	51	−29	−15	841	225	435
98	91	33	25	1089	625	825
25	60	−40	−6	1600	36	240
75	68	10	2	100	4	20
82	62	17	−4	289	16	−68
90	86	25	20	625	400	500
62	58	−3	−8	9	64	24
65	53	0	−13	0	169	0
39	47	−26	−19	676	361	494
650	660	0	0	5398	2224	2704

Thus we have,

$$\Sigma x^2 = 5398,\ \Sigma y^2 = 2224,\ \Sigma xy = 2704.$$

$\therefore$ The coefficient of correlation

$$r = \frac{\Sigma xy}{\sqrt{\{(\Sigma x^2)(\Sigma y^2)\}}} = \frac{2704}{\sqrt{\{(5398 \times 2224)\}}}$$

$$= \frac{2704}{73{\cdot}4 \times 47{\cdot}1} = \frac{2704}{3457{\cdot}14} = \frac{270400}{345714} = 0{\cdot}78 \text{ (nearly).}$$

Example 2. ***Calculate the Karl Pearson's coefficient of correlation between X and Y series :***

X:	17	18	19	19	20	20	21	21	22	23
Y:	12	16	14	11	15	19	22	16	15	20

Solution. The data may be arranged in the following form :

X	Y	$x = X - M_x$	$y = Y - M_y$	x^2	y^2	xy
17	12	−3	−4	9	16	12
18	16	−2	0	4	0	0
19	14	−1	−2	1	4	2
19	11	−1	−5	1	25	5
20	15	0	−1	0	1	0
20	19	0	3	0	9	0
21	22	1	6	1	36	6
21	16	1	0	1	0	0
22	15	2	−1	4	1	−2
23	20	3	4	9	16	12
200	160	0	0	30	108	35

If the mean of X's and Y's are M_x and M_y respectively, then

$$M_x = \frac{\Sigma X}{n} = \frac{200}{10} = 20 \text{ and } M_y = \frac{\Sigma Y}{n} = \frac{160}{10} = 16.$$

If the standard deviations of X's and Y's are σ_x and σ_y then

$$\sigma_x = \sqrt{(\Sigma x^2 / n)} = \sqrt{(30/10)} = \sqrt{3} = 1{\cdot}73$$

$$\sigma_y = \sqrt{(\Sigma y^2 / n)} = \sqrt{(108/10)} = \sqrt{(10{\cdot}8)} = 3{\cdot}28.$$

$\therefore$ The coefficient of correlation r is given by

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

$$= \frac{35}{10 \times 1{\cdot}73 \times 3{\cdot}28} = \frac{3{\cdot}5}{1{\cdot}75 \times 3{\cdot}25} = 0{\cdot}616.$$

Example 3. ***A computer while calculating the correlation coefficient between two variates x and y from 25 pairs of observations obtain the following constants :***

$$n = 25,\ \Sigma x = 125,\ \Sigma x^2 = 650,\ \Sigma y = 100,\ \Sigma y^2 = 960,\ \Sigma xy = 508.$$

It was however, later discovered at the time of checking that he had copied down two pairs as $\begin{array}{c|c} x & y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ ***while the correct values were*** $\begin{array}{c|c} x & y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$. ***Obtain the correct value of the correlation coefficient.***

Solution. Since $8 + 6 = 6 + 8$ and $8^2 + 6^2 = 6^2 + 8^2, 14 + 6 = 12 + 8$ therefore, on account of mistake there is no change in Σx, Σy and Σx^2. But there will be change in Σy^2 and Σxy.

In Σy^2, instead of old value $14^2 + 6^2 = 232$, the new value $12^2 + 8^2 = 208$ is to be substituted.

In Σxy, instead of old value $(6 \times 14 + 8 \times 6 = 132)$, the new value $(8 \times 12 + 6 \times 8 = 144)$ is to be substituted.

Σy^2 (correct value) $= 960 - 232 + 208 = 960 - 24 = 936$.

Similarly Σxy (correct value) $= 508 - (132) + 144 = 508 + 12 = 520$.

Hence the coefficient of correlation r is given by

$$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} = \frac{520}{\sqrt{\{(650 \times 936)\}}} = \frac{520}{\sqrt{(608400)}} = \frac{520}{780} = \frac{2}{3} = 0{\cdot}666.$$

§ 7.8. SHORT CUT METHOD

In many cases it becomes easier if deviations are measured from assumed means. In this case a different formula is developed. The process is as follows :

Let M_x and M_y be the actual means of the X-series and Y-series and let A_x and A_y be their assumed means respectively. Also suppose that x and y are the deviations of X-series and Y-series from M_x and M_y respectively, and ξ, η the respective deviations from A_x and A_y. Then

$$\xi = X - A_x = X - M_x + M_x - A_x$$
$$= x + d_x \text{ where } d_x = M_x - A_x = \Sigma\xi / n$$
$$\eta = Y - A_y = Y - M_y + M_y - A_y$$
$$= y + d_y \text{ where } d_y = M_y - A_y = \Sigma\eta / n.$$
$$\therefore \quad \Sigma\xi\eta = \Sigma(x + d_x)(y + d_y) = \Sigma[xy + xd_y + yd_x + d_x\, d_y]$$
$$= \Sigma xy + d_y\, \Sigma x + d_x\, \Sigma y + \Sigma d_x\, d_y$$
$$= \Sigma xy + n\, d_x\, d_y \text{ since } \Sigma x = 0 = \Sigma y.$$
$$\therefore \quad \Sigma xy = \Sigma\xi\eta - n\, d_x\, d_y = \Sigma\xi\eta - \frac{\Sigma\xi\,\Sigma\eta}{n}$$
$$\sigma_x^2 = \frac{\Sigma\xi^2}{n} - \left(\frac{\Sigma\xi}{n}\right)^2, \quad \sigma_y^2 = \frac{\Sigma\eta^2}{n} - \left(\frac{\Sigma\eta}{n}\right)^2.$$

Substituting these values in the formula of Karl Pearson's coefficient of correlation (§ 7.6), we get

$$r = \frac{\Sigma xy}{n\,\sigma_x\,\sigma_y} = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\,\Sigma\eta}{n}}{n\sqrt{\left\{\dfrac{\Sigma\xi^2}{n} - \left(\dfrac{\Sigma\xi}{n}\right)^2\right\}}\sqrt{\left\{\dfrac{\Sigma\eta^2}{n} - \left(\dfrac{\Sigma\eta}{n}\right)^2\right\}}}$$

or
$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right\}\right]}}.$$

ILLUSTRATIVE EXAMPLES

Example 1. ***Find the coefficient of correlation between the value of X and Y.***

X:	1	3	5	7	8	10
Y:	8	12	15	17	18	20

Solution. Suppose that the assumed means for X and Y-series are 7 and 15 respectively. Now the data may be arranged in the following manner :

X	Y	$\xi = X - 7$	$\eta = Y - 15$	$\xi\eta$	ξ^2	η^2
1	8	−6	−7	42	36	49
3	12	−4	−3	12	16	9
5	15	−2	0	0	4	0
7	17	0	2	0	0	4
8	18	1	3	3	1	9
10	20	3	5	15	9	25
		−8	0	72	66	96

Here $n = 6$.

Substituting these values in Karl Pearson's formula namely

$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right\}\right]}}, \text{ we get}$$

$$r = \frac{72 - \dfrac{(-8)\times 0}{6}}{\sqrt{\left[\left\{66 - \dfrac{8^2}{6}\right\}\left\{96 - 0\right\}\right]}} = \frac{72}{\sqrt{55\cdot 333 \times 96}}$$

$$= \frac{72}{72\cdot 88} = 0\cdot 9879 = 1 \text{ (nearly)}.$$

Example 2. ***Calculate the value of r for the following values of X and Y:***

X:	2	5	7	9	19	17
Y:	25	27	26	29	34	35

Solution. Let the assumed means for X and Y series be 9 and 29 respectively.

X	Y	$\xi = X - 9$	$\eta = Y - 29$	$\xi\eta$	ξ^2	η^2
2	25	−7	−4	28	49	16
5	27	−4	−2	8	16	4
7	26	−2	−3	6	4	9
9	29	0	0	0	0	0
19	34	10	5	50	100	25
17	35	8	6	48	64	36
Total		5	2	140	233	90

Here $n = 6$.

∴ The required coefficient of correlation r is given by

$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right\}\right]}}$$

or
$$r = \frac{140 - \dfrac{5\times 2}{6}}{\sqrt{\left[\left\{233 - \dfrac{5^2}{6}\right\}\left\{90 - \dfrac{2^2}{6}\right\}\right]}}$$

$$= \frac{140 - 1\cdot 67}{\sqrt{[\{233 - 4\cdot 17\}\{90 - 0\cdot 67\}]}}$$

$$= \frac{138\cdot 33}{\sqrt{(228\cdot 83 \times 89\cdot 33)}} = \frac{138\cdot 33}{142\cdot 97} = 0\cdot 967.$$

Example 3. ***The ages of husbands and wives are given in the following table :***

Age of Husband x:	23	27	28	29	30
Age of Wife's y :	18	22	23	24	25

Calculate the coefficient of correlation between x and y from the above table.

Solution. Let the assumed means for x and y be $A_x = 28$ and $A_y = 23$ respectively, the

x	y	$\xi = x - 28$	$\eta = y - 23$	$\xi\eta$	ξ^2	η^2
23	18	−5	−5	25	25	25
27	22	−1	−1	1	1	1
28	23	0	0	0	0	0
29	24	1	1	1	1	1
30	25	2	2	4	4	4
Total		$\Sigma\xi = -3$	$\Sigma\eta = -3$	$\Sigma\xi\eta = 31$	$\Sigma\xi^2 = 31$	$\Sigma\eta^2 = 31$

Here $n = 5$.

Required coefficient of correlation is given by

$$r = \frac{\Sigma\xi\eta - \frac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \frac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \frac{(\Sigma\eta)^2}{n}\right\}\right]}}$$

$$= \frac{31 - \frac{(-3)(-3)}{5}}{\sqrt{\left\{\left(31 - \frac{9}{5}\right)\left(31 - \frac{9}{5}\right)\right\}}}$$

$$= \frac{5 \times 31 - 9}{\sqrt{(5 \times 31 - 9)(5 \times 31 - 9)}} = 1.$$

Thus there is a perfect positive correlation between the age of husband and wife.

Example 4. ***Calculate the coefficient of correlation for the following ages of husbands and wives :***

Husband's age (X):	24	27	28	28	29	30	32	33	35	35	40
Wife's age (Y) :	18	20	22	25	22	28	28	30	27	30	32

Solution. Let the assumed means for X and Y be $A_x = 30$ and $A_y = 28$ respectively, then the data may be arranged in the following manner :

X	Y	$\xi = X - 30$	$\eta = Y - 28$	$\xi\eta$	ξ^2	η^2
24	18	−6	−10	60	36	100
27	20	−3	−8	24	9	64
28	22	−2	−6	12	4	36
28	25	−2	−3	6	4	9
29	22	−1	−6	6	1	36
30	28	0	0	0	0	0
32	28	2	0	0	4	0
33	30	3	2	6	9	4
35	27	5	−1	−5	25	1
35	30	5	2	10	25	4
40	32	10	4	40	100	16
Total		11	−26	159	217	270

Here $n = 11$.

∴ Required coefficient of correlation is given by

$$r = \frac{\Sigma\xi\eta - \frac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \frac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \frac{(\Sigma\eta)^2}{n}\right\}\right]}} = \frac{159 - \frac{(11) \times (-26)}{11}}{\sqrt{\left[\left\{217 - \frac{11^2}{11}\right\}\left\{270 - \frac{(-26)^2}{11}\right\}\right]}}$$

$$= \frac{185}{\sqrt{206 \times 288 \cdot 55}} = \frac{185}{\sqrt{(4296 \cdot 13)}} = \frac{185}{207 \cdot 27} = 0 \cdot 89.$$

Example 5. ***The following marks have been obtained by a class of students in Statistics (out of 100).***

Paper I :	80	45	55	56	58	60	65	68	70	75	85
Paper II :	82	56	50	48	60	62	64	65	70	74	90

Compute the coefficient of correlation for the above data.

Solution. Let the assumed means for x and y be $A_x = 65$ and $A_y = 70$ respectively, thus we have the following table :

x	y	$\xi = x - A_x$ $(A_x = 65)$	$\eta = y - A_y$ $(A_y = 70)$	$\xi\eta$	ξ^2	η^2
80	82	15	12	180	225	144
45	56	−20	−14	280	400	196
55	50	−10	−20	200	100	400
56	48	−9	−22	198	81	484
58	60	−7	−10	70	49	100
60	62	−5	−8	40	25	64
65	64	0	−6	0	0	36
68	65	3	−5	−15	9	25
70	70	5	0	0	25	0
75	74	10	4	40	100	16
85	90	20	20	400	400	400
Total		2	−49	1393	1414	1865

Here $n = 11$.

$\therefore$ The required coefficient of correlation is obtained by putting the values in Pearson's formula, we have

$$r = \frac{\Sigma\xi\eta - \frac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \frac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \frac{(\Sigma\eta)^2}{n}\right\}\right]}}$$

$$= \frac{1393 - \frac{(2)\times(-49)}{11}}{\sqrt{\left[\left\{1414 - \frac{2^2}{11}\right\}\left\{1865 - \frac{(-49)^2}{11}\right\}\right]}}$$

$$= \frac{14\cdot92}{\sqrt{(1413\cdot64\times1646\cdot7)}} = \frac{1402}{1525\cdot7} = 0\cdot918.$$

Example 6. ***The following data regarding the heights (y) and weights (x) of 100 college students are given***

$\Sigma x = 15000$, $\Sigma x^2 = 2272500$, $\Sigma y = 6800$, $\Sigma y^2 = 463025$ *and* $\Sigma xy = 1022250$.

Find the correlation coefficient between height and weight.

Solution. Given $n = 100$.

Correlation coefficient r between x and y is given by

$$r = \frac{\Sigma xy - \frac{\Sigma x\,\Sigma y}{n}}{\sqrt{\left[\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}\right]}}$$

$$= \frac{1022250 - \frac{15000\times6800}{100}}{\sqrt{\left[\left\{2272500 - \frac{(15000)^2}{100}\right\}\left\{463025 - \frac{(6800)^2}{100}\right\}\right]}}$$

$$= \frac{1022250 - 1020000}{\sqrt{(2272500 - 2250000)(463025 - 462400)}}$$

$$= \frac{2250}{\sqrt{(22500\times625)}} = \frac{2250}{150\times25} = \frac{15}{25} = \frac{3}{5} = 0\cdot6.$$

Example 7. ***Establish the formula*** $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\,\sigma_y}$ ***where r is the correlation coefficient between x and y and*** $\sigma_x, \sigma_y, \sigma_{x-y}$ ***are concerned standard deviations. Hence evaluate r from the following data :***

x:	21	23	30	54	57	58	72	78	87	90
y:	60	71	72	83	110	84	100	92	113	135

Solution. To establish the formula

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\,\sigma_y}.$$

Let $\bar{x}$ and $\bar{y}$ be the means of x and y series respectively. Then

$$\sigma_x^2 = \frac{\Sigma(x-\bar{x})^2}{n}, \quad \sigma_y^2 = \frac{\Sigma(y-\bar{y})^2}{n}$$

where n is the number of terms in each of x and y series.

Now, mean of $(x-y)$ series $= \overline{(x-y)} = \bar{x} - \bar{y}$.

Also $\Sigma[(x-y)-(\bar{x}-\bar{y})]^2 = \Sigma[(x-\bar{x})-(y-\bar{y})]^2$

$$= \Sigma[(x-\bar{x})^2 + (y-\bar{y})^2 - 2(x-\bar{x})(y-\bar{y})]$$

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

$$\Rightarrow \quad \frac{\Sigma[(x-y)-(\bar{x}-\bar{y})]^2}{n} = \frac{\Sigma(x-\bar{x})^2}{n} + \frac{\Sigma(y-\bar{y})^2}{n} - \frac{2\Sigma\{(x-\bar{x})(y-\bar{y})\}}{n}$$

$$\Rightarrow \quad \sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\,\sigma_y \qquad \left[\because r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\,\sigma_y}, \text{ see } \S\ 8.6\right]$$

$$\Rightarrow \quad r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\,\sigma_y}. \qquad \ldots(1)$$

To evaluate r. Here $n = 10$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{570}{10} = 57, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{920}{10} = 92$$

Mean of $(x-y)$ series $= \overline{(x-y)} = \dfrac{\Sigma(x-y)}{n} = \dfrac{-350}{10} = -35$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-y)$	$(x-y)-\overline{(x-y)}$	$[(x-y)-\overline{(x-y)}]^2$
21	60	− 36	− 32	1296	1024	− 39	− 4	16
23	71	− 34	− 21	1156	441	− 48	− 13	169
30	72	− 27	− 20	729	400	− 42	− 7	49
54	83	− 3	− 9	9	81	− 29	6	36
57	110	0	18	0	324	− 53	− 18	324
58	84	1	− 8	1	64	− 26	9	81
72	100	15	8	225	64	− 28	7	49
78	92	21	0	441	0	− 14	21	441
87	113	30	21	900	441	− 26	9	81
90	135	33	43	1089	1849	− 45	− 10	100
$\Sigma x = 570$	$\Sigma y = 920$	—	—	$\Sigma(x-\bar{x})^2 = 5846$	$\Sigma(y-\bar{y})^2 = 4688$	$\Sigma(x-y) = -350$	—	$\Sigma[(x-y)-\overline{(x-y)}]^2 = 1346$

$$\sigma_x^2 = \frac{1}{n}\Sigma(x-\bar{x})^2 = \frac{5846}{10} = 584\cdot6$$

$$\sigma_y^2 = \frac{1}{n}\Sigma(y-\bar{y})^2 = \frac{4688}{10} = 468\cdot8$$

$$\sigma_{x-y}^2 = \frac{1}{n}\Sigma[(x-y)-\overline{(x-y)}]^2 = \frac{1346}{10} = 134\cdot6$$

Putting values in (1), we obtain

$$r = \frac{584\cdot6 + 468\cdot8 - 134\cdot6}{2\sqrt{584\cdot6}\,\sqrt{468\cdot8}} = \frac{918\cdot8}{1047\cdot016} = 0\cdot87754.$$

Example 8. ***Calculate the coefficient of correlation from the following results :***

$$n = 10,\ \Sigma x = 650,\ \Sigma y = 660,\ \Sigma (x-65)^2 = 5398,$$
$$\Sigma (y-66)^2 = 2224,\ \Sigma [(x-65)(y-66)] = 2704.$$

Solution. Here, we have

$$\bar{x} = \frac{\Sigma x}{n} = \frac{650}{10} = 65,\ \bar{y} = \frac{\Sigma y}{n} = \frac{660}{10} = 66.$$

$$\therefore \quad r = \frac{\Sigma [(x-\bar{x})(y-\bar{y})]}{\sqrt{[\Sigma (x-\bar{x})^2 . \Sigma (y-\bar{y})^2]}} = \frac{\Sigma [(x-65)(y-66)]}{\sqrt{[\Sigma (x-65)^2\ \Sigma (y-66)^2]}}$$

$$= \frac{2704}{\sqrt{(5398 \times 2224)}} = 0 \cdot 78704.$$

Example 9. ***Calculate the coefficient of correlation between x and y from the following results :***

$$n = 6,\ \Sigma x = 34,\ \Sigma y = 90,\ \Sigma (x-7)^2 = 66,$$
$$\Sigma (y-15)^2 = 96,\ \Sigma [(x-7)(x-15)] = 72.$$

Solution. Let $\bar{x} = \frac{\Sigma x}{n} = \frac{34}{6} = 5 \cdot 667 \neq 7.$

$\therefore$ Let $\xi = x - 7,\ \eta = y - 15$

$$\Sigma \xi = \Sigma (x-7) = \Sigma x - 6 \times 7 = 34 - 42 = -8$$
$$\Sigma \eta = \Sigma (y-15) = \Sigma y - 6 \times 15 = 90 - 90 = 0$$
$$\Sigma \xi \eta = \Sigma [(x-7)(y-15)] = 72$$
$$\Sigma \xi^2 = \Sigma (x-7)^2 = 66,\ \Sigma \eta^2 = \Sigma (y-15)^2 = 96.$$

$$\therefore \quad r = \frac{\Sigma \xi \eta - \frac{\Sigma \xi\ \Sigma \eta}{n}}{\sqrt{\left[\left\{\Sigma \xi^2 - \frac{(\Sigma \xi)^2}{n}\right\}\left\{\Sigma \eta^2 - \frac{(\Sigma \eta)^2}{n}\right\}\right]}}$$

$$\therefore \quad = \frac{72 - \frac{(-8) \times 0}{6}}{\sqrt{\left[66 - \frac{(-8)^2}{6}\right\}\{96 - 0\}}} = \frac{72}{\sqrt{(55 \cdot 333 \times 96)}} = \frac{72}{\sqrt{5311 \cdot 97}}$$

$$= \frac{72}{72 \cdot 88} = 0 \cdot 9879 = 1 \text{ (nearly)}.$$

Example 10. ***If x and y are two uncorrelated random variables, and if $u = x + y$, $v = x - y$, find the coefficient of correlation between u and v.***

Solution. If r is the coefficient of correlation between u and v, then

$$r = \frac{\Sigma [(u-\bar{u})(v-\bar{v})]}{n \sigma_u\ \sigma_v} \qquad \ldots(1)$$

Now $u = x + y$ and $v = x - y.$

$\therefore$ $\bar{u} = \bar{x} + \bar{y}$ and $\bar{v} = \bar{x} - \bar{y}$

$$\Sigma [(u-\bar{u})(v-\bar{v})] = \Sigma [\{(x-\bar{x}) + (y-\bar{y})\}\{(x-\bar{x}) - (y-\bar{y})\}]$$
$$= \Sigma [(x-\bar{x})^2 - (y-\bar{y})^2] = \Sigma (x-\bar{x})^2 - \Sigma (y-\bar{y})^2$$
$$= n\sigma_x^2 - n\sigma_y^2$$

$$\sigma_u^2 = \frac{1}{n} \Sigma (u-\bar{u})^2 = \frac{1}{n} \Sigma [(x-\bar{x}) + (y-\bar{y})]^2$$
$$= \frac{1}{n} \Sigma [(x-\bar{x})^2 + (y-\bar{y})^2 + 2(x-\bar{x})(y-\bar{y})]$$
$$= \frac{1}{n} \Sigma (x-\bar{x})^2 + \frac{1}{n} \Sigma (y-\bar{y})^2 + \frac{2}{n} \Sigma [(x-\bar{x})(y-\bar{y})]$$
$$= \sigma_x^2 + \sigma_y^2$$

[Since x and y are uncorrelated, so $\Sigma [(x-\bar{x})(y-\bar{y})] = 0$]

Similarly $\sigma_v^2 = \sigma_x^2 + \sigma_y^2.$

Putting values in (1), $r = \dfrac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}.$

§ 7.9. STEP-DEVIATION METHOD

The calculation to find r becomes easier when we put $u = \xi / h$ and $v = \eta / h'$ where h and h' are scales for x and y series respectively, therefore putting $\xi = uh$ and $\eta = vh'$ in

$$r = \frac{\Sigma \xi \eta - \frac{\Sigma \xi\ \Sigma \eta}{n}}{\sqrt{\left[\left\{\Sigma \xi^2 - \frac{(\Sigma \xi)^2}{n}\right\}\left\{\Sigma \eta^2 - \frac{(\Sigma \eta)^2}{n}\right\}\right]}}$$, we get

$$r = \frac{\Sigma uv - \frac{\Sigma u\ \Sigma v}{n}}{\sqrt{\left[\left\{\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right\}\left\{\Sigma v^2 - \frac{(\Sigma v)^2}{n}\right\}\right]}}.$$

ILLUSTRATIVE EXAMPLE

Example 1. ***Find the coefficient of correlation for the following table :***

x:	10	14	18	22	26	30
y:	18	12	24	6	30	36

Solution. Let the assumed means of x and y be $A_x = 22$ and $A_y = 24$ respectively.

x	y	$\xi = x - A_x$	u	$\eta = y - A_y$	v	uv	u^2	v^2
10	18	−12	−3	−6	−1	3	9	1
14	12	−8	−2	−12	−2	4	4	4
18	24	−4	−1	0	0	0	1	0
22	6	0	0	−18	−3	0	0	9
26	30	4	1	6	1	1	1	1
30	36	8	2	12	2	4	4	4
Total			−3		−3	12	19	19

Here $h = 4$ (interval for x), $u = \xi / h = \xi / 4$,

$h' = 6$ (interval for y), $v = \eta / h' = \eta / 6$.

The coefficient of correlation

$$r = \frac{\Sigma uv - \frac{\Sigma u\,\Sigma v}{n}}{\sqrt{\left[\left\{\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right\}\left\{\Sigma v^2 - \frac{(\Sigma v)^2}{n}\right\}\right]}}$$

$$r = \frac{12 - \frac{(-3)(-3)}{6}}{\sqrt{\left[\left\{19 - \frac{(-3)^2}{6}\right\}\left\{19 - \frac{(-3)^2}{6}\right\}\right]}} = \frac{12 - 1{\cdot}5}{19 - 1{\cdot}5} = \frac{10{\cdot}5}{17{\cdot}5} = \frac{3}{5} = 0{\cdot}6.$$

EXERCISE 7 (A)

1. Calculate the coefficient of correlation between x and y for the values given below :

x :	-10	-5	0	5	10
y :	5	9	7	11	13

2. Find the coefficient of correlation between the values of x and y given below :

x :	4	8	10	12	16	22
y :	36	24	20	16	14	10

3. Calculate the coefficient of correlation between the values of x and y given below :

x :	28	80	97	69	59	79	68	61
y :	125	137	156	112	107	136	123	108

4. The 8 students got the following marks in Mathematics and English :

Mathematics :	76	90	98	69	54	82	67	52
English :	25	37	56	12	07	36	23	11

Calculate the coefficient of correlation.

5. Calculate the coefficient of correlation for the following ages of husband and wife :

(a) Husband's age :	23	27	28	29	30	31	33	35	36	39
Wife's age :	18	22	23	24	25	26	28	29	30	32
(b) Husband's age :	23	27	28	28	29	30	31	33	35	36
Wife's age :	18	20	22	27	21	29	27	29	28	29

6. Calculate the value of Karl Pearson's coefficient of correlation for the following series A and B :

A :	105	104	102	101	100	99	98	96	93	92
B :	101	103	100	98	95	96	104	92	97	94

7. Determine r from the following :

Family number :	1	2	3	4	5	6	7	8	9	10	11
Brother (x) :	71	68	66	67	70	71	70	73	72	65	66
Sister (y) :	69	64	65	63	65	72	65	64	66	69	62

8. Heights of fathers and sons are given as below; find r.

Height of father :	65	63	67	64	68	62	70	66	68	67	69	71
Height of son :	68	66	68	65	69	66	68	65	71	67	68	70

9. Ten students got the following percentage of marks in Economics and Statistics. Find the coefficient of correlation and interpret it.

Students :	1	2	3	4	5	6	7	8	9	10
Marks in Economics :	47	53	58	86	62	68	60	91	51	84
Marks in Statistics :	39	65	62	90	82	75	96	98	36	78

10. Find Karl Pearson's coefficient of correlation from the following index numbers and interpret it :

Wages :	100	101	103	102	100	99	97	98	96	95
Cost of living :	98	99	99	97	95	92	95	94	90	91

11. Find Karl Pearson's coefficient of correlation from the following table :

Marks :	55–58	58–61	61–64	64–67	67–70
Number of students :	12	17	23	18	11

ANSWERS

1. 0·9	2. −0·942	3. 0·957	4. 0·94
5. (a) 0·99	(b) 0·818	6. 0·6	7. −0·16
8. 0·7	9. 0·747	10. 0·85	11. −0·03.

§ 7.10. RANK CORRELATION

Professor Charles Spearman developed a method to find the correlation. This method is much easier as compared with Karl Pearson's method. This method is called **Spearman's Rank Difference Method**. In this method the knowledge of values of different terms of a series is not necessary but the method is applicable if we only know the ranks of different terms corresponding to their values. The top value is assigned rank 1, the second rank 2, the third rank 3 and so on.

It is given by $r = 1 - \frac{6\,\Sigma d^2}{n\,(n^2 - 1)}$ or $r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)}$

where d_i = difference in ranks of ith paired items in two series

n = number of individuals.

In the above formula 'r' is called **Spearman's Coefficient of rank correlation** or simply **rank correlation coefficient**.

Remark 1. If two individual values are equal *i.e.*, if ties occur, then they are assigned the average of the ranks they would have received if they had differed slightly. Therefore, if two items are tied for the 5th rank, each will be assigned the rank $\frac{1}{2}(5 + 6) = 5{\cdot}5$.

To find rank correlation coefficient in case of tied ranks.

If two or more ranks be equal then the formula

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)}$$

is corrected and the corrected formula is given by

$$r = 1 - \frac{6\left[\sum_{i=1}^{n} d_i^2 + \frac{1}{12}\sum_{j=1}^{k}(m_j^3 - m_j)\right]}{n(n^2-1)} \quad \ldots(1)$$

where m_j, $j = 1, 2, \ldots, k$ is the number of those terms whose ranks are equal.

The correction $\frac{1}{12}\Sigma(m_j^3 - m_j)$ is added to Σd_i^2.

Remark 2. Rank correlation coefficient lies between -1 and $+1$ including both the values.

ILLUSTRATIVE EXAMPLES

Example 1. ***Compute Spearman's rank correlation coefficient r for the following data :***

Person	I	II	III	IV	V	VI	VII	VIII	IX	X
Rank in Mathematics	10	5	4	6	2	3	1	9	7	8
Rank in Electronics	1	2	3	4	5	6	7	8	9	10

Solution.

Person	Rank in Mathematics X	Rank in Electronics Y	X – Y = d	d^2
I	10	1	9	81
II	5	2	3	9
III	4	3	1	1
IV	6	4	2	4
V	2	5	– 3	9
VI	3	6	– 3	9
VII	1	7	– 6	36
VIII	9	8	1	1
IX	7	9	– 2	4
X	8	10	– 2	4
			$\Sigma d = 0$	$\Sigma d^2 = 158$

Here $n = 10$.

$$\therefore \quad r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 158}{10(100-1)} = 1 - 0{\cdot}9576 = 0{\cdot}0424.$$

Example 2. ***Two Judges in a beauty contest rank the ten competitors in the following order :***

6	4	3	1	2	7	9	8	10	5
4	1	6	7	5	8	10	9	3	2

Do the two Judges appear agree in their standards ?

Solution. The given data may be arranged in the following manner :

First Judge's opinion X	Second Judge's opinion Y	Rank diff. X – Y = d	d^2
6	4	2	4
4	1	3	9
3	6	– 3	9
1	7	– 6	36
2	5	– 3	9
7	8	– 1	1
9	10	– 1	1
8	9	– 1	1
10	3	7	49
5	2	3	9
Totals		$\Sigma d = 0$	$\Sigma d^2 = 128$

Here $n = 10$.

$\therefore$ The required rank correlation coefficient is given by

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 128}{10(100-1)} = 1 - 0{\cdot}776 = 0{\cdot}224.$$

Hence judges agree, through lower rank.

Example 3. ***Calculate the rank correlation coefficient from the following table :***

X :	**78**	**89**	**97**	**69**	**59**	**79**	**68**	**57**
Y :	**125**	**137**	**156**	**112**	**107**	**136**	**123**	**108**

Solution.

X	Y	Rank in X	Rank in Y	X – Y = d	d^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	–1	1
59	107	7	8	–1	1
79	136	3	3	0	0
68	123	6	5	1	1
57	108	8	7	1	1
Total				$\Sigma d = 0$	$\Sigma d^2 = 4$

$\therefore$ The required rank correlation coefficient

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 4}{8(64-1)} = 1 - \frac{3}{63} = 60/63 = 20/21 = 0.95.$$

Example 4. ***The competitors in a beauty contest get marks by three judges in the following orders :***

First Judge	:	1	6	5	10	3	2	4	9	7	8
Second Judge	:	3	5	8	4	7	10	2	1	6	9
Third Judge	:	6	4	9	8	1	2	3	10	5	7

Use the rank correlation to discuss which pairs of judges have the nearest approach to common tastes in beauty.

Solution. Here we shall calculate the rank correlation for the following three judgements, and then we shall compare them

(i) First and second judges

(ii) Second and third judges

(iii) Third and first judges

Let the rank correlation coefficient be denoted by $r_{1,2}$, $r_{2,2}$ and $r_{3,1}$ for the above three cases respectively.

Rank by 1st Judge x	Rank by 2nd Judge y	Rank by 3rd Judge z	$d_1^2 = (x-y)^2$	$d_2^2 = (y-z)^2$	$d_3^2 = (z-x)^2$
1	3	6	$(-2)^2 = 4$	9	25
6	5	4	$1^2 = 1$	1	4
5	8	9	$(-3)^2 = 9$	1	16
10	4	8	$6^2 = 36$	16	4
3	7	1	$(-4)^2 = 16$	36	4
2	10	2	$(-8)^2 = 64$	64	0
4	2	3	$2^2 = 4$	1	1
9	1	10	$8^2 = 64$	81	1
7	6	5	$1^2 = 1$	1	4
8	9	7	$(-1)^2 = 1$	4	1
Total			$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 214$	$\Sigma d_3^2 = 60$

Here $n = 10$.

$\therefore$
$$r_{1,2} = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6\times 200}{10(100-1)} = -\frac{7}{33} = -0.212$$

$$r_{2,3} = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6\times 214}{10(100-1)} = -\frac{49}{165} = -0.297$$

and
$$r_{3,1} = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6\times 60}{10(100-1)} = \frac{7}{11} = -0.936.$$

From above calculations, we see that $r_{3,1}$ is maximum in value. Hence the judges third and first have the nearest approach to common tastes in beauty.

Example 5. ***Calculate the coefficient of correlation from the following data :***

Marks in Statistics	:	45	56	39	54	45	40	56	60	30	36
Marks in Maths	:	40	36	30	44	36	32	45	42	20	36

Solution.

Marks		Rank in	Rank in		
Stat. X	Maths. Y	X	Y	X − Y = d	d^2
45	40	5·5	4	1·5	2·5
56	36	2·5	6	−3·5	12·25
39	30	8	9	−1	1
54	44	4	2	2	4
45	36	5·5	6	−0·5	0·25
40	32	7	8	−1	1
56	45	2·5	1	1·5	2·25
60	42	1	3	−2	4
30	20	10	10	0	0
36	36	9	6	3	9
Total	—	—	—	$\Sigma d = 0$	$\Sigma d^2 = 36$

In the case of equal ranks, using the corrected formula

$$r = 1 - \frac{6[\Sigma d^2 + (1/12)\Sigma(m_j^3 - m_j)]}{n(n^2-1)}.$$

Here $m_1 = 2$, since the rank of mark 56 is A. M. of two numbers 2 and 3.

$m_2 = 2$, since the rank of marks 45 is A. M. of two numbers 5 and 6.

$m_3 = 3$, since the rank of marks 36 is A. M. of three numbers 5, 6 and 7.

$\therefore$
$$\frac{1}{12}\sum_{j=1}^{3}(m_j^3 - m_j) = \frac{1}{12}\{(2^3-2)+(2^3-2)+(3^3-3)\}$$

$$= (1/12)\{6+6+24\} = 36/12 = 3.$$

$\therefore$ Rank correlation coefficient is given by

$$r = 1 - \frac{6[\Sigma d_j^2 + (1/12)\Sigma(m_j^3 - m_j)]}{n(n^2-1)}$$

$$= 1 - \frac{6[36+3]}{10(10^2-1)} = 1 - \frac{6\times 39}{990} = 1 - 0.2364 = 0.7636.$$

Example 6. ***If $X_i + Y_i = n+1$, then show that $r = -1$.***

Solution. From rank difference, we have

$$X_i - Y_i = d_i \text{ and } X_i + Y_i = n+1 \text{ (given)}.$$

$\therefore$
$$d_i = 2X_i - (n+1).$$

$\therefore$
$$\Sigma d_i^2 = 4\Sigma X^2 - 4(n+1)\Sigma X + n(n+1)^2$$

$$= 4 \cdot \frac{n(n+1)(2n+1)}{6} - 4(n+1) \cdot \frac{n(n+1)}{2} + n(n+1)^2$$

$$= \frac{1}{3} n(n^2 - 1).$$

$$\therefore \quad r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \cdot \frac{1}{3} n(n^2-1)}{n(n^2-1)} = -1.$$

Example 7. ***If d_i stands for the difference in the ranks of the ith individual, then show that***

(i) the minimum value of Σd_i^2 is 0, and

(ii) the maximum value of Σd_i^2 is $\frac{1}{3}(n^3 - n)$.

Solution. We know that Spearman's rank correlation coefficient is given by

$$r = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}. \qquad ...(1)$$

(i) In case correlation is positive integral, $r = 1$.

Then equation (1) gives

$$1 = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \Rightarrow \Sigma d_i^2 = 0.$$

But $\Sigma d_i^2 = 0$ is possible only when each d_i is zero *i.e.*, $d_i = x_i - y_i = 0$.

$\Rightarrow x_i = y_i$. In other words $\Sigma d_i^2 = 0$, if the ranks of the ith individual in the two characteristics are equal.

Thus the minimum value of Σd_i^2 is zero and in this case the ranks of the ith individual (*i.e.*, ranks of every element) in the two characteristics are equal.

(ii) In case correlation in negatively integral, $r = -1$.

Then equation (1) gives

$$-1 = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \Rightarrow \frac{6\Sigma d_i^2}{n(n^2-1)} = 2$$

$$\Rightarrow \quad \Sigma d_i^2 = \frac{1}{3} n(n^2-1) = \frac{1}{3}(n^3 - n). \qquad ...(2)$$

Also, we know that $r \geq -1$ and hence (1) gives

$$1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \geq -1 \Rightarrow \frac{6\Sigma d_i^2}{n(n^2-1)} \leq 2$$

$$\Rightarrow \quad \Sigma d_i^2 \leq \frac{1}{3} n(n^2-1)$$

$$\Rightarrow \quad \Sigma d_i^2 \leq \frac{1}{3}(n^3 - n). \qquad ...(3)$$

Thus in view of (2) and (3), the maximum value of Σd_i^2 is $\frac{1}{3}(n^3 - n)$.

EXERCISE 7 (B)

1. Eight students got the following marks in Maths. and English :

Maths :	76	90	98	69	54	82	67	52
English :	25	37	56	12	7	36	23	11

Calculate rank correlation coefficient.

2. Ten students got the following percentage of marks in Economics and Statistics :

Students :	1	2	3	4	5	6	7	8	9	10
Marks in Economics :	8	36	96	25	75	82	90	62	65	39
Marks in Statistics :	84	51	91	60	68	62	86	58	53	47

Calculate rank correlation coefficient.

3. The ranking of 16 students in Mathematics and Physics are as follows :

Maths :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Physics :	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Calculate the rank correlation coefficient of proficiencies of this group in Mathematics and Physics.

4. The ranks of some 15 students in Mathematics and Statics were as follows. Two numbers within the brackets denote the ranks of the same student in the two subjects.
(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1), (9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13).
Calculate the rank correlation coefficient of proficiencies of this group in these two subjects.

5. Calculate the rank correlation coefficient from the following table :

x :	81	78	73	73	69	68	62	58
y :	10	12	18	18	18	22	20	24

6. Marks of twelve students in Linear Algebra and Analysis are given below :

Linear Algebra (x) :	60	34	40	50	45	40	22	43	42	66	64	46
Analysis (y) :	75	32	33	40	45	43	12	30	34	72	41	57

Calculate the rank correlation coefficient between x and y.

7. If d_i stands for the difference in ranks of the ith individual and if $d_i = 0$ for all values of i, prove that $r = 1$.

8. If d_i stands for the difference in ranks of the ith individual, show that the maximum values of Σd_i^2 is $\frac{1}{3}(n^3 - n)$.

Hence or otherwise show that the rank correlation coefficient lies between – 1 and + 1.

9. Show that in a ranked bivariate distribution in which no ties occur and in which the variables are independent :

(i) Σd_i^2 is always even,

(ii) there are not more than $\frac{1}{6}(n^3 - n) + 1$ possible values of r.

ANSWERS

1. 0·952 **2.** 0·39 **3.** 0·8 **4.** 0·51
5. – 0·9 **6.** 0·84.

§ 7.11. REGRESSION

Suppose there exists some relationship between two variables x and y, the dots of scatter diagram will be more or less cluster about a curve. This curve is called the *curve of regression*.

§ 7.12. LINE OF REGRESSION

Definition. If the curve of regression is a straight line, it is called a *line of regression* and the regression, in this case, is called **linear**.

§ 7.13. EQUATION TO THE LINE OF REGRESSION

Here we are to find the equation of a straight line which can be fitted to a set of points given on a scatter diagram. It can be done by the principle of least squares.

Suppose $Y = aX + b$ ...(1)

is the equation of the line of best fit of X.

Let M_x and M_y be the means of X-series and Y-series respectively. Now transfer the origin to the point (M_x, M_y) with the axes parallel to the original axes, the equation (1) of line with respect to the new origin becomes

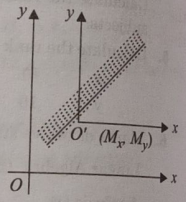
$$y = ax + b \qquad ...(2)$$

where $x = X - M_x$

and $y = Y - M_y$.

Consider an arbitrary dot (x_r, y_r). The difference between this dot and the line (2) is

$$= y_r - ax_r - b.$$



Let U be the sum of the squares of these difference then

$$U = \Sigma (y - ax - b)^2, \ \forall \ r$$

$$= U(a, b) \text{ (say)}.$$

Now according to the principle of least squares, we have to choose a and b so that U is minimum, the conditions for which are

$$\partial U / \partial a = 0 = \partial U / \partial b$$

$$\partial U / \partial a = 0 \Rightarrow -2\Sigma x (y - ax - b) = 0$$

$$\Rightarrow \Sigma xy - a\Sigma x^2 - b\Sigma x = 0$$

$$\Rightarrow \Sigma xy - a\Sigma x^2 = 0 \qquad [\because \ \Sigma x = 0]$$

$$\therefore \quad a = \frac{\Sigma xy}{\Sigma x^2} = \frac{r\sigma_y}{\sigma_x}. \qquad ...(3)$$

$$\left[\because \ r \cdot \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2 \ \Sigma y^2)}} \times \frac{\sqrt{(\Sigma y^2 / n)}}{\sqrt{(\Sigma x^2 / n)}} = \frac{\Sigma xy}{\Sigma x^2} \right]$$

and $\partial U / \partial b = 0 \Rightarrow -2\Sigma (y - ax - b) = 0 \Rightarrow \Sigma y - a\Sigma x - nb = 0$

$$\Rightarrow -nb = 0 \qquad [\because \ \Sigma x = 0 = \Sigma y]$$

$$\therefore \quad b = 0. \qquad ...(4)$$

Therefore, the equation of line of best fit with respect to the new axes is

$$y = r(\sigma_y / \sigma_x) x. \qquad ...(5)$$

Hence the equation of line of best fit with respect to the original axes becomes

$$Y - M_y = r(\sigma_y / \sigma_x)(X - M_x). \qquad ...(6)$$

Equation (6) is called the **Regression line of Y on X**. Therefore if the straight line is selected in such a way that the sum of the squares of deviations parallel to y-axis be minimum, then this line is called the regression line of Y on X.

Similarly if X be taken as dependent variable, then the equation of another regression line, called the **regression line of X on Y**, is given by

$$X - M_x = r(\sigma_x / \sigma_y)(Y - M_y). \qquad ...(7)$$

§ 7.14. REGRESSION COEFFICIENTS

Regression coefficient of y on $x = r(\sigma_y / \sigma_x)$ and it is usually denoted by b_{yx}.

Similarly regression coefficient of x on $y = r(\sigma_x / \sigma_y)$ and it is usually denoted by b_{xy}.

§ 7.15. EXPLANATION OF REGRESSION LINES WHEN

(a) $r = 0$, (b) $r = \pm 1$.

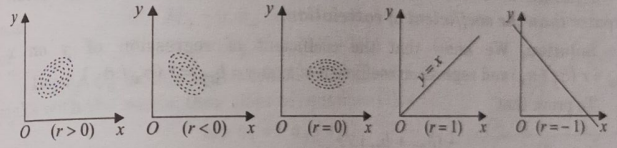
Explanation. The equation of regression line of Y on X is

$$Y - M_y = r(\sigma_y / \sigma_x)(X - M_x) \qquad ...(1)$$

and the equation of regression line of X on Y is

$$X - M_x = r(\sigma_x / \sigma_y)(Y - M_y). \qquad ...(2)$$

(a) If $r = 0$, then equations (1) and (2) become $Y = M_y$ and $X = M_x$, these are the equations of two straight lines parallel to X and Y-axes and passing through their means M_y and M_x.



Hence the regression lines are perpendicular to each other *i.e.*, means M_y and M_x do not change by change in Y and X *i.e.*, X and Y are independent.

(b) If $r = 1$, the equations of both regression lines are coincident and is given by

$$Y - M_y = (\sigma_y / \sigma_x)(X - M_x).$$

Similarly if $r = -1$, the equations of both the regression lines are coincident. In these cases the variates are called perfectly correlated. If $r = +1$, the variables are perfectly positively correlated *i.e.*, high values of one variate corresponds to high values of the other variate. If $r = -1$, the variables are perfectly negative correlated *i.e.*, high values of one variate correspond to low values of other variate.

ILLUSTRATIVE EXAMPLES

Example 1. ***Prove that the Karl Pearson's coefficient of correlation r lies between −1 and +1.***

Solution. We know that the sum of squares of deviations of dots from regression line of y on x is minimum :

$$U = \Sigma (y - ax - b)^2 = U(a, b), \text{(say)}.$$

Now from the condition of being U minimum, we have

$$\partial U / \partial a = 0 \Rightarrow -2\Sigma x\,(y - ax - b) = 0$$

$$\Rightarrow \Sigma xy - a\,\Sigma x^2 - b\,\Sigma x = 0 \Rightarrow \Sigma xy - a\,\Sigma x^2 = 0 \quad [\because\ \Sigma x = 0]$$

$$\therefore \quad a = \Sigma xy / \Sigma x^2$$

and $$\partial U / \partial b = 0 \Rightarrow -2\,\Sigma\,(y - ax - b) = 0$$

$$\Rightarrow \Sigma y - a\,\Sigma x - n\,b = 0 \Rightarrow -n\,b = 0 \qquad [\because\ \Sigma x = 0 = \Sigma y]$$

$$\therefore \quad b = 0.$$

$$\therefore \quad U = \Sigma\,(y - ax)^2 = \Sigma y^2 - 2a\,\Sigma xy + a^2\,\Sigma x^2 \qquad \ldots(2)$$

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

$$= \Sigma y^2 - \frac{2\,(\Sigma xy)^2}{\Sigma x^2} + \frac{(\Sigma xy)^2}{\Sigma x^2} = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$

$$= \Sigma y^2 \left\{1 - \frac{(\Sigma xy)^2}{\Sigma x^2\ \Sigma y^2}\right\} = (1 - r^2)\,\Sigma\, y^2 \qquad \left[\because\ r = \frac{\Sigma xy}{\sqrt{\Sigma x^2\ \Sigma y^2}}\right]$$

Since U is sum of squares, and it will not be negative.

Similarly, Σy^2 will not be negative. Thus

$$1 - r^2 \geq 0 \text{ or } r^2 \leq 1$$

or $$-1 \leq r \leq 1.$$

Example 2. ***Prove that the A. M. of the coefficient of regression is greater than the coefficient of correlation.***

Solution. We know that the coefficient of regression of y on x is $b_{yx} = r\,(\sigma_y / \sigma_x)$ and regression coefficient of x on $y = b_{xy} = r\,(\sigma_x / \sigma_y)$.

To prove that

$$\frac{1}{2}\,(b_{yx} + b_{xy}) > r$$

or $$\frac{1}{2}\left[\frac{r\sigma_y}{\sigma_x} + \frac{r\sigma_x}{\sigma_y}\right] > r \text{ or } \sigma_y^2 + \sigma_x^2 > 2\sigma_x\ \sigma_y$$

or $$(\sigma_y - \sigma_x)^2 \geq 0, \text{ which is true.}$$

Example 3. (a) ***Show that the coefficient of correlation is the G. M. of the coefficient of regression.***

Solution. We know that the coefficient of regression of y on x is $b_{yx} = \dfrac{r\sigma_y}{\sigma_x}$.

Similarly regression coefficient of x on $y = b_{xy} = \dfrac{r\sigma_x}{\sigma_y}$.

$$\therefore \quad b_{yx} \cdot b_{xy} = \frac{r\sigma_y}{\sigma_x} \cdot \frac{r\sigma_x}{\sigma_y} = r^2.$$

$\because$ G. M. of b_{yx} and $b_{xy} = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{r^2} = r.$

Example 3. (b) ***Show that the product of regression coefficients is less than or equal to 1.***

Solution. From example 3 (a) above, we have

$$b_{yx}\ b_{xy} = r^2.$$

But $$r^2 \leq 1$$

$$\therefore \quad b_{yx}\ b_{xy} \leq 1. \qquad \text{[See example 1]}$$

Example 3. (c) ***If $b_{yx} > 1$, then prove that $b_{xy} \leq 1$ provided $b_{yx}\ b_{xy} \leq 1$.***

Solution. Given $b_{yx} > 1$.

$$\therefore \quad \frac{1}{b_{yx}} < 1 \Rightarrow \frac{b_{xy}}{r^2} < 1 \qquad [\because\ b_{yx}\ b_{xy} = r^2]$$

$$\Rightarrow \quad b_{xy} < r^2 \Rightarrow b_{xy} \leq 1 \qquad [\because\ r^2 \leq 1]$$

Remark. Statements given in examples 1, 2, 3 (a), 3 (b) and 3 (c) are the **properties of regression coefficients.**

Example 4. ***If θ is the acute angle between the two regression lines in case of two variables x and y, show that $\tan\theta = \dfrac{1 - r^2}{r} \cdot \dfrac{\sigma_x\ \sigma_y}{\sigma_x^2 + \sigma_y^2}$ where r, σ_x, σ_y have their usual meanings.***

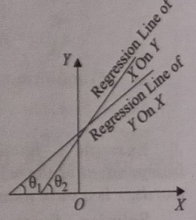
Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

Solution. We know that the equations of line of regression are as follows :

$$Y - M_y = (r\ \sigma_y / \sigma_x)\,(X - M_x) \qquad \ldots(1)$$

and $$X - M_x = (r\ \sigma_x / \sigma_y)\,(Y - M_y)$$

or $$Y - M_y = (\sigma_y / r\ \sigma_x)\,(X - M_x). \qquad \ldots(2)$$



Let θ_1, θ_2 be the angles which two regression lines make with the x-axis, then slope of equations (1) is

$$\tan\theta_1 = r\ \sigma_y / \sigma_x,$$

and slope of equation (2) is

$$\tan\theta_2 = \sigma_y / (r\ \sigma_x).$$

Let θ be the acute angle between the regression lines.

$$\therefore \quad \tan\theta = \tan(\theta_2 \sim \theta_1) = \frac{\tan\theta_2 \sim \tan\theta_1}{1 + \tan\theta_2\ \tan\theta_1}$$

$$\frac{\dfrac{\sigma_y}{r\,\sigma_x} \sim \dfrac{r\,\sigma_y}{\sigma_x}}{1 + \dfrac{\sigma_y}{r\,\sigma_x} \cdot \dfrac{r\,\sigma_y}{\sigma_x}} = \frac{1 - r^2}{r} \cdot \frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2}.$$

Case (i). When $r = 0$, then $\theta = \pi/2$ *i.e.*, two regression lines are perpendicular to each other. Thus the two estimated values of y are the same for all values of x or *vice-versa*.

Case (ii). When $r = \pm 1$, then $\theta = 0$. Hence $\theta = 0$ or π. Thus the lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point (M_x, M_y), they cannot be parallel and there is a perfect correlation between the two variables x and y. Hence two lines of regression coincide.

Example 5. ***For two random variables x and y with the same mean the two regression equations are*** $y = ax + b$ ***and*** $x = \alpha y + \beta$. ***Show that***

$$\frac{b}{\beta} = \frac{1-a}{1-\alpha}.$$

Find also the common mean.

Solution. Given equations of lines of regression are

$$y = ax + b \quad \ldots(1)$$

and $$x = \alpha y + \beta. \quad \ldots(2)$$

Let m be the same mean of two variables x and y, thus

$$m = M_x = M_y \text{ and } b_{yx} = a, b_{xy} = \alpha.$$

$\therefore$ Equation of line of regression of y on x is

$$y - m = a(x - m)$$

or $$y = ax + m(1 - a). \quad \ldots(3)$$

Similarly, equation of line of regression of x on y is

$$x - m = \alpha(y - m)$$

or $$x = \alpha y + m(1 - \alpha). \quad \ldots(4)$$

Now on comparing (1) and (3), (2) and (4), we get

$$b = m(1 - a) \quad \ldots(5)$$

$$\beta = m(1 - \alpha). \quad \ldots(6)$$

$\therefore$ Dividing (5) by (6), we have

$$\frac{b}{\beta} = \frac{m(1-a)}{m(1-\alpha)} = \frac{1-a}{1-\alpha}.$$

$\therefore$ (5) and (6) $\Rightarrow m = \dfrac{b}{1-a} = \dfrac{\beta}{1-\alpha}$

= common mean.

Example 6. ***Show that the value of the correlation coefficient is independent of the origin of reference and unit of measurement. In other words, r is a pure number.***

Solution. Suppose x and y are two given variables and consider two new variables U and V defined as

$$U = \frac{X-a}{h}, V = \frac{Y-b}{k} \Rightarrow X = a + hU, Y = b + kV$$

where a, b, h, k are constants.

Let M_x and M_y be the mean of X and Y series respectively and let M_u and M_v be the mean of U and V series respectively then we have

$$M_x = a + h\,M_u \text{ and } M_y = b + kM_v.$$

Let σ_u and σ_v be the standard deviation of U and V series, also we have :

$$\sigma_x = h\sigma_u \text{ and } \sigma_y = k\sigma_v$$

σ_u and σ_v are defined as

$$\sigma_u^2 = \frac{\Sigma U^2}{n} - \left(\frac{\Sigma U}{n}\right)^2, \quad \sigma_v^2 = \frac{\Sigma V^2}{n} - \left(\frac{\Sigma V}{n}\right)^2.$$

$\therefore$ $$x = X - M_x = X - a - h\,M_u = h(u - M_u) = hu$$

where u, is deviation of u from M_x.

Similarly $y = kv$.

$\therefore$ Coefficient of correlation $= r = \dfrac{\Sigma xy}{h\,\sigma_x\,\sigma_y} \Rightarrow r = \dfrac{\Sigma(hukv)}{n\,h\,\sigma_u\,k\sigma_v} \Rightarrow r = \dfrac{\Sigma\,u\,v}{n\,\sigma_u\,\sigma_v}.$

Example 7. ***The following marks have been obtained by a class of students in Statistics (out of 100) :***

Paper I	**80**	**45**	**55**	**56**	**58**	**60**	**65**	**68**	**70**	**75**	**85**
Paper II	**82**	**56**	**50**	**48**	**60**	**62**	**64**	**65**	**70**	**74**	**90**

Compute the coefficient of correlation for the above data. Find also the equation of the lines of regression.

Solution. The given data may be arranged in following manner :

	Paper I		Paper II			
x	$\xi = x - 65$	ξ^2	y	$\eta = y - 70$	η^2	$\xi\eta$
80	15	225	82	12	144	180
45	−20	400	56	−14	196	280
55	−10	100	50	−20	400	200
56	−9	81	48	−22	484	198
58	−7	49	60	−10	100	70
60	−5	25	62	−8	64	40
65	0	0	64	−6	36	0
68	3	9	65	−5	25	−15
70	5	25	70	0	0	0
75	10	100	74	4	16	40
85	20	400	90	20	400	400
	$\Sigma\xi = 2$	$\Sigma\xi^2 = 1414$		$\Sigma\eta = -49$	$\Sigma\eta^2 = 1865$	$\Sigma\xi\eta = 1393$

Here $n = 11$.

$\therefore$ Required correlation coefficient

$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right\}\right]}}$$

$$= \frac{1393 - \dfrac{2\times(-49)}{11}}{\sqrt{\left[\left\{1414 - \dfrac{(2)^2}{11}\right\}\left\{1865 - \dfrac{(-49)^2}{11}\right\}\right]}}$$

$$= \frac{1393 + 8\cdot 91}{\sqrt{[(1413\cdot 64)\times(1865 - 218\cdot 27)]}}$$

$$= \frac{1401\cdot 91}{\sqrt{(2327883\cdot 3)}} = \frac{1401\cdot 91}{1525\cdot 74} = 0\cdot 919$$

and $$\sigma_x = \sqrt{\left[\frac{\Sigma\xi^2}{n} - \left(\frac{\Sigma\xi}{n}\right)^2\right]} = \sqrt{\left[\frac{1414}{11} - \left(\frac{2}{11}\right)^2\right]}$$

$$= \sqrt{128 \cdot 545 - (0 \cdot 182)^2} = \sqrt{128 \cdot 545 - 0 \cdot 033} = \sqrt{128 \cdot 512} = 11 \cdot 336.$$

Similarly, $$\sigma_y = \sqrt{\left[\frac{\Sigma\eta^2}{n} - \left(\frac{\Sigma\eta}{n}\right)^2\right]} = \sqrt{\left[\frac{1865}{11} - \left(\frac{-49}{11}\right)^2\right]}$$

$$= \sqrt{169 \cdot 545 - (4 \cdot 455)^2} = \sqrt{169 \cdot 545 - 19 \cdot 847}$$

$$= \sqrt{149 \cdot 698} = 12 \cdot 235.$$

$\therefore$ Regression coeff. of y on x

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{(0 \cdot 919) \times (12 \cdot 235)}{(11 \cdot 336)} = \frac{21 \cdot 244}{11 \cdot 336} = 0 \cdot 992 \text{ (nearly)}$$

and regression coeff. of x on y

$$b_{xy} = \frac{r\sigma_x}{\sigma_y} = \frac{(0 \cdot 919) \times (11 \cdot 336)}{(12 \cdot 235)} = \frac{10 \cdot 418}{12 \cdot 235} = 0 \cdot 851 \text{ (nearly).}$$

Again $$M_x = \text{assumed mean} + \frac{\Sigma\xi}{n} = 65 + \frac{2}{11} = 65 \cdot 2 \text{ marks}$$

$$M_y = \text{assumed mean} + \frac{\Sigma\eta}{n} = 70 + \frac{-49}{11} = 65 \cdot 55 \text{ marks.}$$

Equation of line of regression of y on x is

$$y - M_y = r\frac{\sigma_y}{\sigma_x}(x - M_x) \quad \text{or} \quad y - 65 \cdot 55 = 0 \cdot 99\,(x - 65 \cdot 2).$$

Similarly, equation of line of regression of x on y is :

$$x - M_x = \frac{r\sigma_x}{\sigma_y}(y - M_y) \quad \text{or} \quad x - 65 \cdot 2 = 0 \cdot 85\,(y - 65 \cdot 55).$$

Example 8. ***Heights of father and sons are given in inches :***

Height of father :	**65**	**66**	**67**	**67**	**68**	**69**	**71**	**73**
Height of son :	**67**	**68**	**64**	**68**	**72**	**70**	**69**	**70**

(a) Find the coefficient of correlation for the above data.

(b) Form the two lines of regression and calculate the expected average height of son when the height of the father is 67 · 5 inches.

Solution. The calculation table is as follows :

x	y	$\xi = x - 69$	$\eta = y - 69$	ξ^2	η^2	$\xi\eta$
65	67	−4	−2	16	4	8
66	68	−3	−1	9	1	3
67	64	−2	−5	4	25	10
67	68	−2	−1	4	1	2
68	72	−1	3	1	9	−3
69	70	0	1	0	1	0
71	69	2	0	4	0	0
73	70	4	1	16	1	4
Total		$\Sigma\xi = -6$	$\Sigma\eta = -4$	$\Sigma\xi^2 = 54$	$\Sigma\eta^2 = 42$	$\Sigma\xi\eta = 24$

Here $n = 8$.

A. M. for x-series $= M_x =$ assumed mean $+ (\Sigma\xi / n)$

$= 69 + (-6)/8 = 68 \cdot 25.$

A. M. for y-series $= M_y =$ assumed mean $+ (\Sigma\eta / n)$

$= 69 + (-4)/8 = 68 \cdot 5$

$$\sigma_x = \sqrt{\left[\frac{\Sigma\xi^2}{n} - \left(\frac{\Sigma\xi}{n}\right)^2\right]} = \sqrt{\left[\frac{54}{8} - \left(\frac{-6}{8}\right)^2\right]} = \sqrt{6 \cdot 1875} = 2 \cdot 49$$

$$\sigma_y = \sqrt{\left[\frac{\Sigma\eta^2}{n} - \left(\frac{\Sigma\eta}{n}\right)^2\right]} = \sqrt{\left[\frac{42}{8} - \left(\frac{-4}{8}\right)^2\right]} = \sqrt{5} = 2 \cdot 23.$$

(a) The correlation coefficient is given by

$$r = \frac{\Sigma\xi\eta - \frac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left[\left\{\Sigma\xi^2 - \frac{(\Sigma\xi)^2}{n}\right\}\left\{\Sigma\eta^2 - \frac{(\Sigma\eta)^2}{n}\right\}\right]}}$$

$$= \frac{24 - \frac{(-6) \times (-4)}{8}}{\sqrt{\left[\left(54 - \frac{9}{2}\right)(42 - 2)\right]}} = \frac{21}{\sqrt{\left(\frac{99}{2} \times 40\right)}} = \frac{21}{\sqrt{(1980)}} = \frac{21}{44 \cdot 49} = \cdot 47.$$

(b) $\therefore$ Equation of line of regression of y on x is

$$y - M_y = r\frac{\sigma_y}{\sigma_x}(x - M_x)$$

or $$y - 68 \cdot 5 = 0 \cdot 47 \times \frac{2 \cdot 23}{2 \cdot 49}(x - 68 \cdot 25)$$

or $$y = 0 \cdot 421x + 39 \cdot 77. \qquad \ldots(1)$$

Similarly, equation of line of regression of x on y is

$$x - M_x = r\,(\sigma_x / \sigma_y)(y - M_y)$$

or $$x - 68 \cdot 25 = 0 \cdot 47 \times \frac{2 \cdot 49}{2 \cdot 23}(y - 68 \cdot 5)$$

$$x = 0 \cdot 52y + 32 \cdot 29. \qquad \ldots(2)$$

$\therefore$ From (1), corresponding to $x = 67 \cdot 5$, we have

$$y = 0 \cdot 421 \times 67 \cdot 5 + 39 \cdot 77 = 68 \cdot 19 \text{ inches.}$$

Thus the expected average height of the son is 68 · 19 inches corresponding to the given height of the father.

Example 9. ***(a) The following data regarding the heights (y) and weights (x) of 100 college students are given :***

$$\Sigma x = 15000,\ \Sigma x^2 = 2272500,\ \Sigma y = 6800,\ \Sigma y^2 = 463025,\ \Sigma xy = 1022250.$$

Find the correlation coefficient between height and weight and equation of regression line of height on weight.

Solution. Given $n = 100$. We have

$$\sigma_x = \sqrt{\left[\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2\right]} = \sqrt{\left[\frac{2272500}{100} - \left(\frac{15000}{100}\right)^2\right]}$$

$$= \sqrt{(22725 - 22500)} = \sqrt{225} = 15$$

$$\sigma_y = \sqrt{\left[\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2\right]} = \sqrt{\left[\frac{463025}{100} - \left(\frac{6800}{100}\right)^2\right]}$$

$$= \sqrt{(4630 \cdot 25 - 4624)} = \sqrt{6 \cdot 25} = 2 \cdot 5.$$

Now
$$r = \frac{\Sigma xy - \frac{\Sigma x\, \Sigma y}{n}}{\sqrt{\left[\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}\right]}}$$

$$= \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x\, \Sigma y}{n^2}}{\sqrt{\left[\left\{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2\right\}\left\{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2\right\}\right]}}$$

$$= \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x\, \Sigma y}{n^2}}{\sigma_x\, \sigma_y} = \frac{\frac{1022250}{100} - \frac{15000 \times 6800}{100 \times 100}}{15 \times 2 \cdot 5}$$

$$= \frac{10222 \cdot 5 - 10200}{15 \times 2 \cdot 5} = \frac{22 \cdot 5}{37 \cdot 5} = 0 \cdot 6.$$

Again $\bar{x} = \frac{\Sigma x}{n} = \frac{15000}{100} = 150, \ \bar{y} = \frac{\Sigma y}{n} = \frac{6800}{100} = 68.$

Equation of regression line of height (y) on weight (x) is

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x}),$$

or $$y - 68 = 0 \cdot 6 \times \frac{2 \cdot 5}{15}(x - 150) \text{ or } y - 68 = \frac{1}{10}(x - 150)$$

or $$10y = x + 530.$$

Example 9. (*b*) *For 10 observations on price (x) and supply (y), the following data were obtained (in appropriate units),*

$\Sigma x = 130, \Sigma y = 220, \Sigma x^2 = 2288, \Sigma y^2 = 5506$ ***and*** $\Sigma xy = 3467.$

Obtain the two lines of regression and estimate the supply when the price is 16 units.

Solution. $\bar{x} = \frac{\Sigma x}{n} = \frac{130}{10} = 13, \ \bar{y} = \frac{\Sigma y}{n} = \frac{220}{10} = 22$ [$\because n = 10$]

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{2288}{10} - \left(\frac{130}{10}\right)^2} = \sqrt{228 \cdot 8 - 169} = 7 \cdot 7330$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2} = \sqrt{\frac{5506}{10} - \left(\frac{220}{10}\right)^2} = \sqrt{550 \cdot 6 - 484} = 8 \cdot 1609$$

$$r = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x\, \Sigma y}{n^2}}{\sigma_x\, \sigma_y} = \frac{\frac{3467}{10} - \frac{130 \times 220}{100}}{7 \cdot 7330 \times 8 \cdot 1609}$$

$$= \frac{346 \cdot 7 - 286}{63 \cdot 1082} = \frac{60 \cdot 7}{63 \cdot 1082} = 0 \cdot 9618.$$

$\therefore$ Regression line of y on x is

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

or $$y - 22 = 0 \cdot 9618 \times \frac{8 \cdot 1609}{7 \cdot 733}(x - 13)$$

or $$y = 22 + 1 \cdot 015\,(x - 13)$$

or $$y = 1 \cdot 015x + 8 \cdot 805. \qquad \ldots(1)$$

Regression line of x on y is

$$x - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \text{ or } x - 13 = 0 \cdot 9618 \times \frac{7 \cdot 733}{8 \cdot 1609}(y - 22)$$

or $$x = 13 + 0 \cdot 9114\,(y - 22) \text{ or } x = 0 \cdot 9114y - 7 \cdot 0508. \qquad \ldots(2)$$

Now given price (*i.e.*, x) = 16. Putting $x = 16$ in (1), we get

$$y = 1 \cdot 015 \times 16 + 8 \cdot 805 \Rightarrow y = 25 \cdot 045 \text{ units.}$$

Hence supply = 25 · 045.

Example 9. (*c*) *Find the line of regressive of y on x for the data given below* :

x	1·53	1·78	2·60	2·95	3·42
y	33·5	36·3	40·0	45·8	53·5

Solution. Here we are giving another method [*i.e.*, by the principle of least squares] to find line of regression of y on x.

S.N.	x	y	xy	x^2
1	1·53	33·5	51·255	2·3409
2	1·78	36·3	64·614	3·1684
3	2·60	40·0	104·000	6·7600
4	2·95	45·8	135·110	8·7025
5	3·42	53·5	182·97	11·6964
$n = 5$	$\Sigma x = 12 \cdot 28$	$\Sigma y = 209 \cdot 1$	$\Sigma xy = 537 \cdot 949$	$\Sigma x^2 = 32 \cdot 6682$

Let the equation of line of regression of y on x be

$$y = a + bx \qquad \ldots(1)$$

where a and b are obtained by the following normal equations

$$\Sigma y = na + b\,\Sigma x \Rightarrow 209 \cdot 1 = 5a + 12 \cdot 28b \qquad \ldots(2)$$

$$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2 \Rightarrow 537 \cdot 949 = 12 \cdot 28\,a + 32 \cdot 6682\,b. \qquad \ldots(3)$$

Solving (2) and (3), $a = 17 \cdot 931442, \ b = 9 \cdot 7266117.$

Putting values in (1), the required line of regression is given by **Ans.**

$$y = 17 \cdot 931442 + 9 \cdot 7266117\,x.$$

Example 9. ***(d) The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men :***

	x	y
Mean	53	142
Variance	130	165

and $\Sigma(x-\bar{x})(y-\bar{y}) = 1220.$

Find the approximate regression equation and use it to estimate the blood pressure of a man whose age is 45.

Solution. From given records, we have

$$\bar{x} = 53, \quad \bar{y} = 142, \; n = 10$$

$$\sigma_x^{\,2} = 130, \; \sigma_y^{\,2} = 165.$$

$$\therefore \quad r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\,\sigma_y} = \frac{1220}{10\times\sqrt{130}\,\sqrt{165}} = 0\cdot 8330.$$

Regression line of blood pressure (y) on age (x) is

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$\Rightarrow \quad y - 142 = 0\cdot 833 \times \frac{\sqrt{165}}{\sqrt{130}}(x-53)$$

$$\Rightarrow \quad y = 142 + 0\cdot 9385\,(x-53)$$

$$\Rightarrow \quad y = 0\cdot 9385x + 92\cdot 26.$$

$\therefore$ When $x = 45$,

$$y = 0\cdot 9385\times 45 + 92\cdot 26 = 134\cdot 49. \qquad \textbf{Ans.}$$

Example 10. ***In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible : variance of x = 9.***

Regression line equations $8x - 10y + 66 = 0, \; 40x - 18y = 214$

what were (a) the mean value of x and y,

(b) standard deviation of y,

(c) the coefficient of correlations between x and y.

Solution. (a) We know that both the lines of regression pass through the point (M_x, M_y) where M_x, M_y are the mean of x and y respectively.

$\therefore$ From the given equations of regression lines, we have

$$4M_x - 5M_y + 33 = 0 \quad \text{and} \quad 20M_x - 9M_y = 107.$$

Solving, we get $M_x = 13, M_y = 17.$

(c) Suppose the lines of regression of y on x and of x on y are $4x - 5y + 33 = 0$ and $20x - 9y = 107$ respectively.

These equations can be written in the form :

$$y = 0\cdot 8x + 6\cdot 6 \quad \text{and} \quad x = 0\cdot 45y + 5\cdot 35$$

$\therefore$ Regression coefficient of y on $x = b_{yx} = r\,\sigma_y / \sigma_x = 0\cdot 8,$

and Regression coefficient of x on $y = b_{xy} = r\,\sigma_x / \sigma_y = 0\cdot 45.$

Hence $r^2 = b_{yx}\cdot b_{xy} = 0\cdot 8\times 0\cdot 45 = 0\cdot 36$

$\Rightarrow \quad r = 0\cdot 6.$

(b) Since variance of $x = \sigma_x^2 = 9.$

$$\therefore \quad \sigma_x = 3.$$

(a) We have, $b_{yx} = r\cdot\dfrac{\sigma_y}{\sigma_x} \Rightarrow 0\cdot 8 = 6\times\dfrac{\sigma_y}{3}$

$$\therefore \quad \sigma_y = 4.$$

Example 11. ***(a) Given x = 4y + 5, y = kx + 4 are the regression lines of x on y and y on x respectively. Show that*** $0 \le 4k \le 1$. ***If k = 1/16, find the means of the two variables and the coefficient of correlation between them.***

Solution. Here the equations of lines of regression are as follows :

$$y = kx + 4 \quad \text{and} \quad x = 4y + 5.$$

$\therefore$ Regression coefficient of y on $x = b_{yx} = k$

and Regression coefficient of x on $y = b_{xy} = 4.$

Now, $\sqrt{(b_{xy}\cdot b_{yx})} = r$

$\Rightarrow \quad b_{xy}\cdot b_{yx} = r^2 \Rightarrow 0 \le b_{xy}\cdot b_{yx} \le 1 \qquad [\because r^2 \le 1]$

$\Rightarrow \quad 0 \le 4k \le 1.$

If $k = 1/16$, then $b_{xy}\cdot b_{yx} = 1/16$ and $b_{xy} = 4.$

$\therefore \quad r^2 = b_{xy}\cdot b_{yx} = 4\times k = 4\times(1/16) = 1/4.$

$\therefore \quad r = 1/2.$

If $\bar{x}, \bar{y}$ are the means of the variables x, y and since lines of regression pass through $(\bar{x}, \bar{y})$, therefore

$$\bar{y} = (1/16)\,\bar{x} + 4 \quad \text{and} \quad \bar{x} = 4\bar{y} + 5.$$

Solving, we get

$$\bar{x} = 28 \quad \text{and} \quad \bar{y} = (23/4).$$

Example 11. ***(b) Two lines of regression are given by x + 2y − 5 = 0 and*** $2x + 3y - 8 = 0$ ***and*** $\sigma_x^{\,2} = 12$. ***Calculate the mean values of x and y, variance of y and the coefficient of correlation between x and y.***

Solution. Here the equations of lines of regression are

$$x + 2y - 5 = 0 \qquad \ldots(1)$$

and $$2x + 3y - 8 = 0. \qquad \ldots(2)$$

Let $\bar{x}$ and $\bar{y}$ be the means of x and y respectively, then the above lines of regression will pass through the point $(\bar{x}, \bar{y})$ and so

$$\bar{x} + 2\bar{y} - 5 = 0 \quad \text{and} \quad 2\bar{x} + 3\bar{y} - 8 = 0.$$

Solving, $\bar{x} = 1, \bar{y} = 2.$

The lines of regression (1) and (2) may be rewritten as

$$y = -\frac{1}{2}x + \frac{5}{2} \qquad \ldots(3)$$

and $$x = -\frac{3}{2}y + 4. \qquad \ldots(4)$$

$\therefore$ Regression coefficient of y on x is [from (3)] $\qquad \ldots(5)$

$$r\,\frac{\sigma_y}{\sigma_x} = -\frac{1}{2}$$

and regression coefficient of x on y is [from (4)]

$$r\frac{\sigma_x}{\sigma_y} = -\frac{3}{2}. \qquad ...(6)$$

Multiplying (5) and (6), $r^2 = \frac{3}{4} \Rightarrow r = \pm\frac{\sqrt{3}}{2}$.

Since the coefficients of x and y in the given equations (1) and (2) are of the same sign, therefore,

$$r = -\frac{\sqrt{3}}{2} = -\frac{1 \cdot 732}{2} = -0 \cdot 866.$$

Given $\sigma_x^2 = 12 \Rightarrow \sigma_x = 2\sqrt{3}$.

Substituting values of σ_x and r in (5), we get

$$\left(-\frac{\sqrt{3}}{2}\right)\cdot\left(\frac{\sigma_y}{2\sqrt{3}}\right) = -\frac{1}{2} \Rightarrow \sigma_y = 2.$$

$\therefore$ variance $(y) = \sigma_y^2 = 4$

i.e., var $(y) = 4$.

Example 12. ***Given the following information :***

	Mean	Standard Deviation
x	40	10
y	6	1·5

Correlation coefficient (between x and y), r = 0·9. Find :

(i) The value of y for x = 60

and (ii) The value of x for y = 10.

Solution. Here $M_x = 40$, $\sigma_x = 10$, $M_y = 6$, $\sigma_y = 1 \cdot 5$, $r = 0 \cdot 9$ where notations have their usual meaning.

(i) The equation of regression line of y on x is

$$y - M_y = r(\sigma_y / \sigma_x)(x - M_x)$$

i.e., $$y - 6 = \frac{0 \cdot 9 \times 1 \cdot 5}{10}(x - 40)$$

$\Rightarrow$ $$y = 6 + 0 \cdot 135(x - 40)$$

$\Rightarrow$ $$y = 6 + 0 \cdot 135x - 5 \cdot 4$$

$\Rightarrow$ $$y = 0 \cdot 135x + 0 \cdot 6.$$

When $x = 60$, we have

$$y = 0 \cdot 135 \times 60 + 0 \cdot 6 \Rightarrow y = 8 \cdot 7.$$

(ii) The equation of regression line of x on y is

$$x - M_x = r(\sigma_x / \sigma_y)(y - M_y)$$

i.e., $$x - 40 = \frac{0 \cdot 9 \times 10}{1 \cdot 5}(y - 6)$$

$\Rightarrow$ $$x = 40 + 6(y - 6) \Rightarrow x = 6y + 4.$$

When $y = 10$, we have

$$x = 6 \times 10 + 4 \Rightarrow x = 64.$$

Example 13. ***Given the following information :***

	Mean	Standard Deviation
yield of wheat (kilogram per unit area)	10	8
Rainfall (cm.)	8	2

Correlation coefficient between production (yield) and rainfall r = 0·5. Estimate the yield when rainfall is 9 cm.

Solution. Let us denote x = yield, y = rainfall.

Then $M_x = 10$, $\sigma_x = 8$, $M_y = 8$, $\sigma_y = 2$, $r = 0 \cdot 5$.

The equation of regression line of x on y is

$$x - M_x = r(\sigma_x / \sigma_y)(y - M_y)$$

i.e., $$x - 10 = \frac{0 \cdot 5 \times 8}{2}(y - 8)$$

$\Rightarrow$ $$x = 2y - 6.$$

When $x = 9$ cm, then $x = 2 \times 9 - 6 \Rightarrow x = 12$.

$\therefore$ The required estimated yield is 12 kg. per unit area.

Example 14. ***Two variables x and y are related to each other by the equation ax + by + c = 0. Prove that the coefficient of correlation between them is − 1 if a and b are of same sign and is + 1 if they are of opposite sign.***

Solution. The given equation is

$$ax + by + c = 0. \qquad ...(1)$$

Writing the equation (1) as the regression line of x on y as follows :

$$x = -(b/a)y - (c/a).$$

$\therefore$ $$b_{xy} = -(b/a). \qquad ...(2)$$

Again rewriting (1) as the regression line of y on x as follows :

$$y = -(a/b)x - (c/a) \qquad ...(3)$$

$\therefore$ $$b_{yx} = -(a/b).$$

Multiplying (2) and (3), we get

$$r^2 = b_{xy} \cdot b_{yx} = \left(-\frac{b}{a}\right)\left(-\frac{a}{b}\right) = 1 \Rightarrow r = \pm 1.$$

(i) If a and b both are of same sign *i.e.*, either $a > 0, b > 0$ or $a < 0, b < 0$, then the regression coefficients b_{xy} and b_{yx} both are negative. Hence $r = -1$.

(ii) If a and b are of opposite signs *i.e.*, either $a > 0, b < 0$ or $a < 0, b > 0$, then the regression coefficients b_{xy} and b_{yx} both are positive. Hence $r = +1$.

Example 15. ***The regression lines of y on x and of x on y are respectively y = ax + b and x = cy + d. Show that***

(i) Means are $\bar{x} = \dfrac{bc + d}{1 - ac}$, $\bar{y} = \dfrac{ad + b}{1 - ac}$.

(ii) If means of x and y are equal, then $\dfrac{b}{d} = \dfrac{1 - a}{1 - c}$.

(iii) Correlation coefficient between x and y is $\pm\sqrt{ac}$.

(iv) The ratio of the standard deviation of x and y is $\sqrt{c/a}$.

Solution. (i) The given regression lines are

$$y = ax + b \quad ...(1)$$

and

$$x = cy + d. \quad ...(2)$$

Let $\bar{x}$ and $\bar{y}$ be the means of x and y respectively, then the above lines will pass through the point $(\bar{x}, \bar{y})$ and therefore,

$$\bar{y} = a\bar{x} + b \text{ and } \bar{x} = c\bar{y} + d$$

i.e., $$a\bar{x} - \bar{y} + b = 0 \quad ...(3)$$

$$\bar{x} - c\bar{y} - d = 0 \quad ...(4)$$

Solving $$\frac{\bar{x}}{d + bc} = \frac{\bar{y}}{b + ad} = \frac{1}{-ac + 1}$$

$$\therefore \quad \bar{x} = \frac{bc + d}{1 - ac}, \ \bar{y} = \frac{ad + b}{1 - ac}.$$

(ii) If $\bar{x} = \bar{y}$ then, we have

$$\frac{bc + d}{1 - ac} = \frac{ad + b}{1 - ac}$$

$$\Rightarrow \quad bc + d = ad + b \Rightarrow b(1 - c) = d(1 - a)$$

$$\Rightarrow \quad \frac{b}{d} = \frac{1 - a}{1 - c}.$$

(iii) From equation (1), the regression coefficient (b_{yx}) of y on x is given by

$$b_{yx} = r(\sigma_y / \sigma_x) = a. \quad ...(5)$$

Similarly from equation (2),

$$b_{xy} = r(\sigma_x / \sigma_y) = c. \quad ...(6)$$

Multiplying (5) and (6), we have

$$r^2 = ac \Rightarrow r = \pm\sqrt{ac}.$$

Also from relations (5) and (6) it follows that r is positive if $a > 0, c > 0$ and r is negative if $a < 0, c < 0$.

(iv) Dividing (6) and (5), we have

$$r\frac{\sigma_x}{\sigma_y} \cdot \frac{\sigma_x}{r\sigma_y} = \frac{c}{a} \Rightarrow \frac{{\sigma_x}^2}{{\sigma_y}^2} = \frac{c}{a} \Rightarrow \frac{\sigma_x}{\sigma_y} = \sqrt{\left(\frac{c}{a}\right)}.$$

§ 7.16. STANDARD ERROR OF PREDICTION

Standard error of prediction is defined as the deviation of the predicted value from the observed value.

Standard error of prediction (or estimate) of y on x is denoted by E_{yx} and is given by

$$E_{yx} = \sqrt{\frac{\Sigma(y - y_r)^2}{n}}$$

where y is the observed (or actual) value and y_r is the predicted value (as given by the regression line of y on x).

Similarly, the standard error of prediction (or estimate) of x on y is denoted by E_{xy} and is given by

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

where x is the observed (or actual) value and x_r is the predicted value (as given by the regression line of x on y).

Theorem. *To show that*

(i) $E_{yx} = \sigma_y \cdot \sqrt{1 - r^2}$, *(ii)* $E_{xy} = \sigma_x \sqrt{1 - r^2}$.

Proof. To equation of regression line of y on x is given by

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$\Rightarrow \quad y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

$$\therefore \quad y_r = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}). \quad ...(1)$$

We know that

$$E_{yx} = \sqrt{\frac{\Sigma(y - y_r)^2}{n}}$$

$$\Rightarrow \quad E_{yx}^2 = \frac{\Sigma(y - y_r)^2}{n} = \Sigma\left[\frac{1}{n}\left\{(y - \bar{y}) - \frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right\}^2\right] \quad \text{[using (1)]}$$

$$= \frac{1}{n}\Sigma\left\{(y - \bar{y})^2 + r^2\frac{{\sigma_y}^2}{{\sigma_x}^2}(x - \bar{x})^2 - 2r\frac{\sigma_y}{\sigma_x}(x - \bar{x})(y - \bar{y})\right\}$$

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

$$= {\sigma_y}^2 + r^2\frac{{\sigma_y}^2}{{\sigma_x}^2}{\sigma_x}^2 - 2r\frac{\sigma_y}{\sigma_x} \cdot r\sigma_x\sigma_y$$

$$= (1 + r^2 - 2r^2){\sigma_y}^2 = (1 - r^2){\sigma_y}^2$$

$$\Rightarrow \quad E_{yx} = \sigma_y\sqrt{1 - r^2}.$$

(ii) Proceed as above in (i).

Example. ***Find the standard error of estimate of y on x for the data given below :***

x:	1	2	3	4	5
y:	2	5	3	8	7

Solution.

S. No.	x	y	xy	x^2
1	1	2	2	1
2	2	5	10	4
3	3	3	9	9
4	4	8	32	16
5	5	7	35	25
$n = 5$	$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma xy = 88$	$\Sigma x^2 = 55$

Let the equation of line of regression of y on x be

$$y = ax + b \qquad ...(1)$$

where a and b are obtained by the following normal equations :

$$\Sigma y = a\,\Sigma x + nb \Rightarrow 25 = 15a + 5b \qquad ...(2)$$

$[\because\ n = 5]$

$$\Sigma xy = a\,\Sigma x^2 + b\,\Sigma x \Rightarrow 88 = 55a + 15b \qquad ...(3)$$

Solving (2) and (3), $a = \frac{13}{10}$, $b = \frac{11}{10}$.

Putting values of a and b in (1), the required line of regression of y on x is $y = \frac{13}{10}x + \frac{11}{10}$.

To find standard error of estimate, we form the following table :

x	y	$y_r = \frac{13}{10}x + \frac{11}{10}$	$y - y_r$	$(y - y_r)^2$
1	2	$\frac{13}{10} + \frac{11}{10} = \frac{24}{10}$	$-\frac{4}{10}$	$\frac{16}{100}$
2	5	$\frac{26}{10} + \frac{11}{10} = \frac{37}{10}$	$\frac{13}{10}$	$\frac{169}{100}$
3	3	$\frac{39}{10} + \frac{11}{10} = 5$	-2	4
4	8	$\frac{52}{10} + \frac{11}{10} = \frac{63}{10}$	$\frac{17}{10}$	$\frac{289}{100}$
5	7	$\frac{65}{10} + \frac{11}{10} = \frac{66}{10}$	$\frac{4}{10}$	$\frac{16}{100}$
				$\Sigma(y - y_r)^2 = 89/10$

$\therefore$ $$E_{yx} = \sqrt{\frac{(y - y_r)^2}{n}} = \sqrt{\frac{89}{10 \times 5}} \qquad [\because\ n = 5]$$

$$= \sqrt{1 \cdot 78} = 1 \cdot 3342$$

Ans.

EXERCISE 7 (C)

1. If $r = 0$, then show that two lines of regression are parallel to the axes.
2. Is the following statement correct ? Give reason.
   (a) The regression coefficient of x on y is 3·2 and that of y on x is 0·8.
   (b) $b_{yx} = 0 \cdot 8$ and $b_{xy} = 2 \cdot 4$.
   (c) $b_{xy} = 0 \cdot 8$, $b_{yx} = 0 \cdot 2$ and $r = -0 \cdot 4$.
3. If two regression coefficients are 0·8 and 0·2, then what would be the values of coefficient of correlation ?
4. (a) Calculate the regression equation from the following data :

Age of husband :	18	19	20	21	22	23	24	25	25	26
Age of wife :	17	17	18	18	18	19	19	20	21	22

   (b) Find the equations of regression in its simplest form of the following ages of Husbands and Wives at marriage :

Husbands' age :	23	27	28	28	29	30	31	33	35	36
Wife's age :	18	20	22	27	21	29	27	29	28	29

5. Write down the two regression equations that may be associated with the following pairs of values :

x :	152	114	138	154	144	153	141	117	126	154
y :	193	300	414	594	676	549	320	483	481	659

6. The different values of two variates are given by the following table :

x :	42	44	48	55	89	98	66
y :	56	49	53	58	65	76	58

   Calculate the regression coefficients and establish the regression equation which may be related with the given values of the variates.

7. Ten students got the following marks in Statistics (x) and Mathematics (y) :

x :	56	55	58	58	57	56	60	54	59	56
y :	68	67	67	70	65	68	70	66	68	66

   (a) Compute the correlation coefficient.
   (b) Estimate the marks in Maths. of a student who received 62 marks in Statistics.
   (c) Estimate the marks in Statistics of a student who received 69 marks in Maths.
8. You are given the following results for the heights (x) and weights (y) of 100 Policemen
   $M_x = 68$ inches, $M_y = 150$ lbs, $\sigma_x = 2 \cdot 5$ inches, $\sigma_y = 20$ lbs., $r = 0 \cdot 6$.
   Estimate from the above data :
   (a) The height of a particular policeman whose weight is 200 lbs.
   (b) The weight of a particular policeman who is 5 feet tall.
9. You are given the following measures for the heights (x) and weight (y) of 120 workers of a factory :
   $M_x = 67''$, $M_y = 14$ kg., $\sigma_x = 2 \cdot 4''$, $\sigma_y = 10 \cdot 5$ kg., $r = 0 \cdot 7$.
   Estimate from the above data :
   (a) The height of a particular worker whose weight is 72 kg.
   (b) The weight of a particular worker whose height is 62".
10. Calculate the coefficient of correlation and obtain the lines of regression for the following data :

x :	1	2	3	4	5	6	7	8	9
y :	9	8	10	12	11	13	14	16	15

    Obtain the estimate of y which would correspond on the average to $x = 6 \cdot 2$.

11. In the following table are recorded data showing the test scores made by 10 salesmen on intelligence test and their weekly sales :

Test Scores :	40	70	50	60	80	50	90	40	60	60
Sales :	2·5	6·0	4·5	5·0	4·5	2·0	5·5	3·0	4·5	3·0

Calculate the regression line of sale on rest score, and estimate the probable weekly sales volume if salesman makes a score of 70.

12. Two random variables have the least square regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation coefficient between x and y.

13. The following regression equations have been obtained from a correlation analysis :

$$8y = 18x + 30, 50x = 12y + 4.$$

Find (a) the mean of x and y, and

(b) the correlation coefficient between x and y.

14. The regression lines of y on x and of x on y are respectively $y = ax + b$ and $x = cy + d$, show that

(a) Means are $\bar{x} = (bc + d)/(1 - ac)$ and $\bar{y} = (ad + b)/(1 - ac)$.

(b) Correlation coefficient between x and y is $\pm\sqrt{ac}$.

(c) The ratio of the standard deviations of x and y is $\sqrt{c/a}$.

15. If the lines of regression of y on x and of x on y are respectively

$$a_1 x + b_1 y + c_1 = 0 \text{ and } a_2 x + b_2 y + c_2 = 0,$$

prove that $a_1 b_2 \leq a_2 b_1$.

16. Find the two lines of regression and coefficient of correlation for the bivariate data :

$$n = 18, \Sigma x^2 = 60, \Sigma y^2 = 96, \Sigma x = 12, \Sigma y = 18, \Sigma xy = 48.$$

ANSWERS

2. (a) No, $r = 1{\cdot}671 > 1$ which is impossible

(b) No, $r = 1{\cdot}92 > 1$ which is impossible

(c) No, $r = 0{\cdot}4$ since r, b_{yx}, b_{xy} should have same signs but given $r = -0{\cdot}4$.

3. 0·4

4. (a) $y - 18{\cdot}2 = 1{\cdot}42\,(x - 22{\cdot}5)$, $x - 22{\cdot}5 = 1{\cdot}74\,(y - 18{\cdot}9)$

(b) $x - 141 = 0{\cdot}034\,(y - 144{\cdot}45)$, $y = 3{\cdot}575x - 34{\cdot}67$.

5. $x = 0{\cdot}03\,y + 126{\cdot}693$, $y = 3{\cdot}595\,x - 34{\cdot}67$.

6. $b_{yx} = 0{\cdot}372$, $b_{xy} = 2{\cdot}197$; $y = 0{\cdot}372x + 35{\cdot}266$, $x = 2{\cdot}197y - 65{\cdot}68$

7. (a) 0·55 (b) 70 (c) 58·8

8. (a) 71·75 inches (b) 111·6 lbs.

9. (a) 76·28″ (b) − 1·3125 kg.

10. $r = 0{\cdot}95$, $y = 0{\cdot}95x + 27{\cdot}25$, $x = 0{\cdot}95y - 6{\cdot}4$; 33·14

11. $y = 0{\cdot}0583x + 0{\cdot}5070$; 4·5880

12. $\bar{x} = 4$, $\bar{y} = 7$, $r = -0{\cdot}5$

13.(a) $\bar{x} = 2{\cdot}56$, $\bar{y} = 9{\cdot}51$ (b) $r = 0{\cdot}73$

16. $y = -0{\cdot}18x + 1{\cdot}12$, $x = -0{\cdot}12\,y + 0{\cdot}89$, $r = -0{\cdot}147$.

§ 7.17. MULTIPLE CORRELATION

The theory of correlation involving three or more than three variables comes under the heading *multiple correlation*. In the previous chapter, we have seen that simple correlation deals with the degree of relationship between two variables such as ages of husbands and wives; height and weight etc. *In multiple correlation, we find the degree of inter-relationship among three or more than three variables.* Following are some of the examples of multiple correlated variables :

(i) Yield of crop per acre (x_1) depends upon rainfall (x_2) and quantity of manure (x_3). Here three variables x_1, x_2, x_3 are related, where x_1 is dependent variable and x_2, x_3 are independent variables.

(ii) Results of houses (x_1), depends upon tax rates (x_2) and building costs (x_3). Here x_1 is dependent variable while x_2, x_3 are independent variables.

(iii) Success of a student in an examination (x_1), depends upon the books available with him (x_2) and mental ability (x_3).

(iv) The yield of crop in a year (x_1), depends upon rainfall (x_2), manure (x_3), the average temperature (x_4) and average humidity during the period between sowing and harvesting of the crop (x_5). Here five variables are related out of which x_1 is dependent and others independent variables.

In the above examples we have seen that one variable (dependent variable) is influenced by two or more variables (independent variables). Hence it becomes necessary to find correlation between three or more than three variates.

In multiple correlation we study that how for the dependent variable is influenced by independent variables.

If x_1 be the dependent variable and $x_2, x_3, x_4, \ldots, x_n$ the independent variables, then multiple correlation between x_1 and $x_2, x_3, \ldots, x_n$ is denoted by $R_{1(2\,3\,\ldots\,n)}$ or by $R_{1.\,2\,3\,\ldots\,n}$.

§ 7.18. PARTIAL CORRELATION

Definition. Partial correlation between two variables x_1 and x_2 is the simple correlation between x_1 and x_2. When the influence of other variables $x_3, \ldots, x_n$ in them *i.e.*, in (x_1 and x_2 both) has been eliminated. It is denoted by $r_{12.\,34\,\ldots\,n}$.

In other words, the simple correlation between two variables when other variables are kept constant is said to be *partial correlation*.

For three variables x_1, x_2 and x_3. The simple correlation between x_1 and x_2 when the linear effect of x_3 in x_1 and x_2 both has been eliminated is called *partial correlation* between x_1 and x_2 and is denoted by $r_{12\cdot3}$ or by $r_{12\,(3)}$.

For example 1. The correlation between statures of mother and sons, when the stature of the father has a particular value, say 70 inches, is an example of partial correlation between the statures of mother sons, when the stature of father has a constant value 70 inches.

For example 2. In example (i) of § 7.1 above, if the quantity of manure (x_3) is fixed then the correlation between x_1 and x_2 is a partial correlation between them.

§ 7.19. MULTIPLE CORRELATION COEFFICIENT

The regression equation of x_1 on x_2 and x_3 is

$$x_1 = b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3. \qquad ...(1)$$

$\therefore$ Residual, $x_{1\cdot23} = x_1 - b_{12\cdot3}\, x_2 - b_{13\cdot2}\, x_3$

$\Rightarrow$ $b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3 = x_1 - x_{1\cdot23}.$ ...(2)

Let the estimated (expected) value of x_1 (observed value of the variable) be X_1 [as determined by (1)], then

$$X_1 = b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3 = x_1 - x_{1\cdot23} \qquad [\text{using (2)}]$$

The multiple correlation coefficient of x_1 on x_2 and x_3 is *defined as the simple correlation coefficient between* x_1 *and* X_1 *and is denoted by* $R_{1\,(23)}$ or $R_{1\cdot23}$. Thus

$$R_{1\,(23)} = \frac{\sum x_1 X_1}{\sqrt{[(\sum x_1^2)(\sum X_1^2)]}} \qquad ...(3)$$

Now $\sum x_1 X_1 = \sum x_1 (b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3)$

$= \sum x_1 (x_1 - x_{1\cdot23}) = \sum x_1^2 - \sum x_1 \cdot x_{1\cdot23}$

$= \sum x_1^2 - \sum x_{1\cdot23}^2$

$= n(\sigma_1^2 - \sigma_{1\cdot23}^2),$

$\sum x_1^2 = n\sigma_1^2$

and $\sum X_1^2 = \sum (x_1 - x_{1\cdot23})^2 = \sum (x_1^2 - 2x_1 x_{1\cdot23} + x_{1\cdot23}^2)$

$= \sum x_1^2 - \sum x_{1\cdot23}^2 \qquad [\because \sum x_1 x_{1\cdot23} = \sum x_{1\cdot23}^2]$

$= n(\sigma_1^2 - \sigma_{1\cdot23}^2).$

Substituting values in (3), we have

$$R_{1\,(23)} = \frac{n(\sigma_1^2 - \sigma_{1\cdot23}^2)}{n\sigma_1 (\sigma_1^2 - \sigma_{1\cdot23}^2)^{1/2}} = \frac{(\sigma_1^2 - \sigma_{1\cdot23}^2)^{1/2}}{\sigma_1} = \left(1 - \frac{\sigma_{1\cdot23}^2}{\sigma_1^2}\right)^{1/2} \qquad ...(4)$$

$\therefore$ The required coefficient of multiple correlation is given by (4).

Other Forms of $R_{1\,(23)}$. Squaring both sides of (4), we have

$$R_{1\,(23)}^2 = 1 - \frac{\sigma_{1\cdot23}^2}{\sigma_1^2} \qquad ...(5)$$

or $$R_{1\,(23)}^2 = 1 - \frac{\Delta}{\Delta_{11}} \qquad ...(5')$$

This is one another form for $R_{1\,(23)}$.

Again we know that

$$\sigma_{1\cdot23}^2 = \frac{\sigma_1^2}{(1 - r_{23}^2)}[1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}]$$

$$\Rightarrow \quad 1 - \frac{\sigma_{1\cdot23}^2}{\sigma_1^2} = 1 - \frac{1}{1 - r_{23}^2}[1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}]$$

$$\Rightarrow \quad R_{1\,(23)}^2 = \frac{1}{1 - r_{23}^2}[r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}]. \qquad ...(6)$$

This is again one another form for $R_{1\,(23)}$.

Particular Case. In case $r_{12} = r_{13} = r_{23} = r$, (say). Then (6) gives

$$R_{1\,(23)}^2 = \frac{2r^2 - 2r^3}{1 - r^2} = \frac{2r^2(1 - r)}{(1 + r)(1 - r)} = \frac{2r^2}{1 + r}, \qquad [\text{provided } r \neq 1]$$

$$\Rightarrow \quad R_{1\,(23)} = \frac{\sqrt{2}\, r}{\sqrt{(1 + r)}}.$$

Remark 1. Since $\sigma_1^2 \geq \sigma_{1\cdot23}^2$, so (5) $\Rightarrow R_{1\,(23)} \geq 0$.

Remark 2. In case $R_{1\,(23)} = 1$, then (5) implies that $\sigma_{1\cdot23}^2 = 0$ which implies that all the residuals $x_{1\cdot23}$ vanish, which means that observed and estimated value of x_1 coincide and hence *observed value of* x_1 *becomes a linear function of* x_2 *and* x_3.

§ 7.20. PARTIAL CORRELATION COEFFICIENT

As defined (see § 7.18), in trivariate distribution the partial correlation between x_1 and x_2 is the simple correlation between x_1 and x_2 when the linear effect of x_3 is eliminated from x_1 and x_2 both. Therefore, the partial correlation coefficient between x_1 and x_2, denoted by $r_{12\cdot3}$, is the simple correlation coefficient between $x_{1\cdot3}$ and $x_{2\cdot3}$ after the linear effect of x_3 has been eliminated from both.

Therefore, we shall subtract from x_1 of each point that part of x_1 which is due to the linear effect of x_1 which is due to the linear effect of x_3, as given by the regression of x_1 on x_3 and let the residual be denoted by $x_{1\cdot3}$, so

$$x_{1\cdot3} = x_1 - b_{13\cdot2}x_3 \;\Rightarrow\; x_{1\cdot3} = x_1 - \frac{\sigma_1}{\sigma_3} r_{13}\, x_3$$

Similarly, $x_{2\cdot3} = x_2 - \dfrac{\sigma_2}{\sigma_3} r_{23} x_3.$

$\therefore$ $r_{12\cdot3}$ = correlation coefficient between $x_{1\cdot3}$ and $x_{2\cdot3}$

$$= \frac{\sum x_{1\cdot3}x_{2\cdot3}}{\sqrt{[\sum (x_{1\cdot3}^2 \sum x_{2\cdot3}^2]}}. \qquad ...(1)$$

Now $\sum x_{1\cdot3}x_{2\cdot3} = \sum \left[\left(x_1 - \dfrac{\sigma_1}{\sigma_3} r_{13}x_3\right)\left(x_2 - \dfrac{\sigma_2}{\sigma_3} r_{23}x_3\right)\right]$

$= \sum x_1x_2 - r_{23}\dfrac{\sigma_2}{\sigma_3}\sum x_1x_3 - r_{13}\dfrac{\sigma_1}{\sigma_3}\sum x_2x_3 + r_{13}r_{23}\dfrac{\sigma_1\sigma_2}{\sigma_3^2}\sum x_3^2$

$= nr_{12}\,\sigma_1\sigma_2 - r_{23}\dfrac{\sigma_2}{\sigma_3}.nr_{13}\,\sigma_1\sigma_3 - r_{13}\dfrac{\sigma_1}{\sigma_3}.nr_{23}\,\sigma_2\sigma_3 + r_{13}r_{23}\dfrac{\sigma_1\sigma_2}{\sigma_3^2}n\sigma_3^2$

$= n\sigma_1\sigma_2\,(r_{12} - r_{12}r_{23})$ [since last two terms cancel] ...(2)

$$\sum x_{1\cdot 3}^2 = \sum \left(x_1 - \frac{\sigma_1}{\sigma_3} r_{13}\, x_3 \right)^2$$

$$= \sum x_1^2 - 2r_{13} \frac{\sigma_1}{\sigma_3} \sum x_1 x_3 + r_{13}^2 \frac{\sigma_1^2}{\sigma_3^2} \sum x_3^2$$

$$= n\sigma_1^2 - 2r_{13} \frac{\sigma_1}{\sigma_3} \cdot nr_{13}\, \sigma_1\sigma_3 + r_{13}^2 \frac{\sigma_1^2}{\sigma_3^2} \cdot n\sigma_3^2$$

$$= n\sigma_1^2 (1 - r_{13}^2) \qquad \ldots(3)$$

and similarly $\sum x_{2\cdot 3}^2 = n\sigma_2^2 (1 - r_{23}^2)$. $\qquad \ldots(4)$

Substituting values from (2), (3), (4) in (1), we get

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = -\frac{\Delta_{12}}{(\Delta_{11}\, \Delta_{22})^{1/2}}. \qquad \ldots(5)$$

Similarly values of $r_{23\cdot 1}$ and $r_{13\cdot 2}$ can be written.

Particular Case. In case $r_{12} = r_{13} = r_{23} = r$, say, then from (5), we obtain

$$r_{12\cdot 3} = \frac{r}{1 + r}.$$

Also in this case $r_{23\cdot 1} = r_{13\cdot 2} = \dfrac{r}{1 + r}$.

Remark. $r_{12\cdot 3}$ can also be defined as

$$r_{12\cdot 3} = \sqrt{(b_{12\cdot 3} \times b_{21\cdot 3})}.$$

By substituting the values of $b_{12\cdot 3}$ and $b_{21\cdot 3}$, we can easily obtain formula (5).

Similarly, we define

$$r_{23\cdot 1} = \sqrt{(b_{23\cdot 1} \times b_{32\cdot 1})}$$

and

$$r_{31\cdot 2} = \sqrt{(b_{31\cdot 2} \times b_{13\cdot 2})}.$$

ILLUSTRATIVE EXAMPLES

Example 1. *If $r_{12} = 0\cdot 80$, $r_{13} = -0\cdot 40$, $r_{23} = -0\cdot 56$, then find $r_{12\cdot 3}$, $r_{13\cdot 2}$, $r_{23\cdot 1}$ and $R_{1\cdot 23}$, the symbols have their usual meaning.*

Solution. We have

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23})^2}}$$

$$= \frac{0\cdot 80 - (-0\cdot 40)(-0\cdot 56)}{\sqrt{\{1 - (-0\cdot 40)^2\}\{1 - (-0\cdot 56)^2\}}}$$

$$= \frac{0\cdot 576}{\sqrt{(0\cdot 84)(0\cdot 6864)}} = \frac{0\cdot 576}{0\cdot 7593} = 0\cdot 7586$$

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{-0\cdot 40 - 0\cdot 80(-0\cdot 56)}{\sqrt{\{1 - (0\cdot 80)^2\}\{1 - (-0\cdot 56)^2\}}}$$

$$= \frac{0\cdot 0480}{\sqrt{(0\cdot 36)(0\cdot 6864)}} = 0\cdot 0966 \qquad [\because\ r_{32} = r_{23}]$$

$$r_{23\cdot 1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{-0\cdot 56 - (0\cdot 80)(-0\cdot 40)}{\sqrt{(1 - 0\cdot 8^2)\{1 - (-0\cdot 40)^2\}}}$$

$$= \frac{-0\cdot 2400}{\sqrt{(0\cdot 36)(0\cdot 84)}} = -0\cdot 4364$$

$$R_{1\cdot 23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \qquad [\text{see (6), § 7.19}]$$

$$= \frac{(0\cdot 8)^2 + (-0\cdot 40)^2 - 2(0\cdot 8)(-0\cdot 4)(-0\cdot 56)}{1 - (-0\cdot 56)^2}$$

$$= \frac{0\cdot 4416}{0\cdot 6864} = 0\cdot 6434.$$

Therefore, $R_{1\cdot 23} = \sqrt{0\cdot 6434} = 0\cdot 8021$.

Example 2. ***On the basis of observations made on 35 cotton plants the total correlations of yield of cotton (x_1), number of balls i.e., seed vessels (x_2), and height (x_3) are found to be $r_{12} = 0\cdot 863$, $r_{13} = 0\cdot 648$ and $r_{23} = 0\cdot 709$.***

Determine the multiple correlation $R_{1\cdot 23}$ and the partial correlations $r_{12\cdot 3}$ and $r_{13\cdot 2}$ and interpret your results.

Solution. Proceeding as Example 1 above, we have

$$R_{1\cdot 23} = 0\cdot 365,\ r_{12\cdot 3} = 0\cdot 751,\ r_{13\cdot 2} = 0\cdot 101.$$

Interpretation. Here $R_{1\cdot 23}$ has a very large value, therefore, it implies that x_2 and x_3 both have considerable influence on x_1 *i.e.*, the regression equation of x_1 on x_2 and x_3 will be excellent.

Again the value of the total correlation r_{12} is sufficiently large this implies that for predicting x_1, we should make x_2 as an independent variable.

In addition to it $r_{13\cdot 2}$ being $0\cdot 101$ (a small quantity), we should also take x_3 as an independent variable in addition to x_2. This would considerably increase the accuracy of the prediction.

Example 3. ***If $R_{1\,(23)} = 0$, then show that $R_{2\,(13)}$ is not necessarily zero.***

Solution. We have

$$R_{1\,(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} = 0 \qquad [\because\ R_{1\,(23)} = 0, \text{ given}]$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0 \Rightarrow 2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2 \qquad \ldots(1)$$

Now $$R_{2\,(13)}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{21}r_{23}r_{13}}{1 - r_{13}^2}$$

$$\Rightarrow \quad R_{2\,(13)} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} \qquad \text{[using (1) and } r_{21} = r_{12}\text{]}$$

$$= \sqrt{\frac{r_{23}^2 - r_{13}^2}{1 - r_{13}^2}}$$

which is not necessarily zero.

Example 4. *If* $R_{1\,(23)} = 1$, *show that* $R_{2\,(13)} = 1$, $R_{3\,(12)} = 1$.

Solution. $R_{1\,(23)} = \sqrt{\dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{(1 - r_{23}^2)}} = 1$ [$\because$ $R_{1\,(23)} = 1$, given]

Squaring $r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$

$\Rightarrow \quad r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{13}^2$. ...(1)

Now $R_{2\,(13)} = \sqrt{\dfrac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} = \sqrt{\dfrac{1 - r_{13}^2}{1 - r_{13}^2}} = 1$ [using (1)]

$$R_{3\,(12)} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} = 1 \qquad \text{[by using (1)]}$$

Example 5. *Show that* $b_{12\cdot3}\, b_{23\cdot1}\, b_{31\cdot2} = r_{12\cdot3}\, r_{23\cdot1}\, r_{31\cdot2}$.

Solution. We have

$$b_{12\cdot3} = \frac{\sum x_{1\cdot3} x_{2\cdot3}}{\sum x_{2\cdot3}^2} = \frac{r_{12\cdot3}\,\sigma_{1\cdot3}}{\sigma_{2\cdot3}}$$

$$b_{23\cdot1} = \frac{\sum x_{2\cdot1} x_{3\cdot1}}{\sum x_{3\cdot1}^2} = \frac{r_{23\cdot1}\,\sigma_{2\cdot1}}{\sigma_{3\cdot1}}$$

and

$$b_{31\cdot2} = \frac{\sum x_{3\cdot2} x_{1\cdot2}}{\sum x_{1\cdot2}^2} = \frac{r_{31\cdot2}\,\sigma_{3\cdot2}}{\sigma_{1\cdot2}}.$$

Multiply above expressions and get the result.

Example 6. *In a distribution* $\sigma_1 = 2$, $\sigma_2 = \sigma_3 = 3$, $r_{12} = 0{\cdot}7$, $r_{13} = 0{\cdot}5$, $r_{23} = 0{\cdot}5$.

Find (*i*) $b_{12\cdot3}$, $b_{13\cdot2}$ (*ii*) $\sigma_{1\cdot23}$ (*iii*) $R_{1\cdot23}$.

Solution. (i) We have

$$b_{12\cdot3} = \frac{\sigma_{1\cdot3}}{\sigma_{2\cdot3}}\, r_{12\cdot3} \text{ and } b_{13\cdot2} = \frac{\sigma_{1\cdot2}}{\sigma_{3\cdot2}}\, r_{13\cdot2} \qquad ...(1)$$

Now, $\sigma_{1\cdot3} = \sigma_1\sqrt{1 - r_{13}^2} = 2\sqrt{1 - (0{\cdot}5)^2} = 1{\cdot}7321$

$\sigma_{1\cdot2} = \sigma_1\sqrt{1 - r_{12}^2} = 2\sqrt{1 - (0{\cdot}7)^2} = 1{\cdot}4283$

$\sigma_{2\cdot3} = \sigma_2\sqrt{1 - r_{23}^2} = 3\sqrt{1 - (0{\cdot}5)^2} = 2{\cdot}5981$

$\sigma_{3\cdot2} = \sigma_3\sqrt{1 - r_{32}^2} = 3\sqrt{1 - (0{\cdot}5)^2} = 2{\cdot}5981$.

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0{\cdot}7 - (0{\cdot}5)(0{\cdot}5)}{\sqrt{(1 - 0{\cdot}25)(1 - 0{\cdot}25)}} = 0{\cdot}6000$$

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{0{\cdot}5 - (0{\cdot}7)(0{\cdot}5)}{\sqrt{(1 - 0{\cdot}49)(1 - 0{\cdot}25)}} = 0{\cdot}2425.$$

Putting values in (1), we get

$$b_{12\cdot3} = \frac{1{\cdot}7321 \times 0{\cdot}6}{2{\cdot}5981} = 0{\cdot}40$$

and

$$b_{13\cdot2} = \frac{1{\cdot}4283 \times 0{\cdot}2425}{2{\cdot}5981} = 0{\cdot}1333.$$

(ii) $\sigma_{1\cdot23} = \sigma_1\sqrt{\dfrac{\Delta}{\Delta_{11}}}$

where

$$\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$= 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} = 0{\cdot}36$$

Δ_{11} = co-factor of r_{11} in $\Delta = 1 - r_{23}^2 = 0{\cdot}75$.

$$\therefore \quad \sigma_{1\cdot23} = 2\sqrt{\frac{0{\cdot}36}{0{\cdot}75}} = 1{\cdot}3856.$$

(iii) $R_{1\cdot23}^2 = \dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} = 0{\cdot}52$ [putting values and simplifying]

$$\therefore \quad R_{1\cdot23} = \pm\, 0{\cdot}721.$$

Example 7. *If* $r_{12\cdot3} = 0$, *then prove that* $r_{13\cdot2} = r_{13}\sqrt{\left(\dfrac{1 - r_{23}^2}{1 - r_{12}^2}\right)}$.

Solution. We have

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{\{(1 - r_{13}^2)(1 - r_{23}^2)\}}} = 0. \qquad ...(1)$$

$$\therefore \quad r_{12} = r_{13}\, r_{23}.$$

$$\therefore \quad r_{13\cdot2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{\{(1 - r_{12}^2)(1 - r_{23}^2)\}}} = \frac{r_{13} - (r_{13}\, r_{23})\, r_{23}}{\sqrt{\{(1 - r_{12}^2)(1 - r_{23}^2)\}}} \qquad \text{[using (1)]}$$

$$= \frac{r_{13}(1 - r_{23}^2)}{\sqrt{\{(1 - r_{12}^2)(1 - r_{23}^2)\}}}$$

$$= r_{13}\sqrt{\left(\frac{1 - r_{23}^2}{1 - r_{12}^2}\right)}.$$

Proved.

Example 8. *Prove that* $b_{123} = \dfrac{b_{12} - b_{13}\, b_{32}}{1 - b_{23}\, b_{32}}$.

Solution. Since variates are measured from their respective means, therefore,

$$b_{123} = \frac{\Sigma x_{13}\, x_{23}}{\Sigma x_{23}^2} = \frac{\Sigma (x_1 - b_{13}\, x_3)(x_2 - b_{23}\, x_3)}{\Sigma (x_2 - b_{23}\, x_3)^2}$$

$$= \frac{\Sigma x_1 x_2 - b_{13}\, \Sigma x_2 x_3 - b_{23}\, \Sigma x_1 x_3 + b_{13}\, b_{23}\, \Sigma x_3^2}{\Sigma x_2^2 - 2b_{23}\, \Sigma x_2 x_3 + b_{23}^2\, \Sigma x_3^2}$$

where $b_{12} = \dfrac{\Sigma x_1 x_2}{\Sigma x_2^2} \Rightarrow \Sigma x_1 x_2 = N \sigma_2^2\, b_{12}$.

Similarly, $\Sigma x_2 x_3 = N\sigma_3^2\, b_{23}$, $\Sigma x_1 x_3 = N\sigma_3^2\, b_{13}$.

$$\therefore \quad b_{123} = \frac{N[\sigma_2^2\, b_{12} - b_{13}\, b_{23}\, \sigma_3^2 - b_{23}\, b_{13}\, \sigma_3^2 + b_{13} b_{23}\, \sigma_2^2]}{N[\sigma_2^2 - 2b_{13}\, b_{23}\, \sigma_3^2 + b_{23}^2\, \sigma_3^2]}$$

$$= \frac{b_{12} - b_{13}\, b_{32}}{1 - 2b_{32}\, b_{23} + b_{23}\, b_{32}} \qquad [\because\ b_{32}\, \sigma_2^2 = b_{23}\, \sigma_3^2]$$

$$= \frac{b_{12} - b_{13}\, b_{32}}{1 - b_{23}\, b_{32}}.$$

EXERCISE 7 (D)

1. If $r_{12} = 0{\cdot}86$, $r_{13} = 0{\cdot}65$, $r_{23} = 0{\cdot}72$, find r_{123}.
2. For a trivariate distribution, prove that $1 - R_{1(23)}^2 = (1 - r_{12}^2)(1 - r_{132}^2)$.

   Hence deduce that $R_{1(23)} \geq r_{12}$.

   [**Hint.** $\sigma_{123}^2 = \dfrac{1}{n}\sum x_{123}^2 = \dfrac{1}{n}\sum x_{12}\, x_{132}$

   $= \dfrac{1}{n}\sum x_{12}(x_1 - b_{123}\, x_2 - b_{132}\, x_3)$

   $= \dfrac{1}{n}(\sum x_{12}^2 - b_{132} \sum x_{12}\, x_{32})$

   $= \sigma_{12}^2 (1 - b_{132}\, b_{312})$

   $= \sigma_1^2 (1 - r_{12}^2)(1 - r_{132}^2)$ $[\because\ \sigma_{12}^2 = \sigma_1^2 (1 - r_{12}^2)]$ ...(1)

   Also $\sigma_{123}^2 = \sigma_1^2 (1 - R_{1(23)}^2)$. ...(2)

   $\therefore$ (1) and (2) $\Rightarrow 1 - R_{1(23)}^2 = (1 - r_{12}^2)(1 - r_{132}^2)$. ...(3)

   Also (3) $\Rightarrow 1 - R_{1(23)}^2 \leq 1 - r_{12}^2$

   $\Rightarrow R_{1(23)}^2 \geq r_{12}^2 \Rightarrow R_{1(23)} \geq r_{12}$.]
3. If $r_{123} = 0$, prove that $r_{132} = r_{13}\sqrt{\dfrac{1 - r_{23}^2}{1 - r_{13}^2}}$.
4. If $r_{23} = 1$, show that $r_{12}^2 = r_{13}^2$; $\sigma_{123}^2 = \sigma_1^2 (1 - r_{12}^2)$.

5. Prove that $b_{12} = \dfrac{b_{123} + b_{132}\, b_{321}}{1 - b_{132}\, b_{312}}$.
6. $r_{123}\dfrac{\sigma_{13}}{\sigma_{23}} = \dfrac{\sigma_1}{\sigma_2}\left(\dfrac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)$.
7. $\dfrac{\sigma_{123}\, \sigma_{231}}{r_{123}} = -\dfrac{\Delta\, \sigma_1\, \sigma_2}{\Delta_{12}}$.
8. If $r_{23} = 1$, prove that
   (a) $r_{12}^2 = r_{13}^2$, (b) $\sigma_{123}^2 = \sigma_1^2 (1 - r_{12}^2)$.
9. If $r_{23} = 0$, prove that
   (a) $R_{1(23)}{}^2 = r_{12}^2 + r_{13}^2$,
   (b) $\sigma_{123}{}^2 = \sigma_1^2 (1 - r_{12}^2 - r_{13}^2)$.
10. Explain partial and multiple correlation.

    If $\gamma_{12} = +0{\cdot}80$, $\gamma_{13} = -0{\cdot}40$, $\gamma_{23} = -0{\cdot}56$, then find partial correlation.

ANSWERS

1. 0·747.
10. $\gamma_{123} = 0{\cdot}759$, $\gamma_{132} = 0{\cdot}097$, $\gamma_{23\cdot1} = -0{\cdot}436$

EXERCISE 7 (E)

Objective Type Questions

1. For a bivariate distribution, which of the following relation is true :
   (a) $R_{1(23)} < 0$ (b) $R_{1(23)} > 0$
   (c) $R_{1(23)} < r_{12}$ (d) None of these.
2. For a bivariate distribution, which of the following statements is true :
   (a) $\sigma_{12}^2 > \sigma_1^2$ (b) $\sigma_{13}^2 > \sigma_1^2$
   (c) $\sigma_{132}^2 > \sigma_1^2$ (d) $\sigma_{123}^2 < \sigma_1^2$.
3. If $R_{1(23)} = 1$ for a trivariate distribution, then which of the following relations is true :
   (a) $\sigma_1^2 = 0$ (b) $\sigma_{12}^2 = 0$
   (c) $\sigma_{123}^2 = 0$ (d) None of these.
4. The formula for the measurement of regression coefficient b_{123} is :

   (a) $\dfrac{\sigma_1}{\sigma_2} \cdot \dfrac{\begin{vmatrix} r_{13} & r_{23} \\ r_{12} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}$ (b) $\dfrac{\sigma_1}{\sigma_2} \cdot \dfrac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & r_1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}$

   (c) $\dfrac{\sigma_2}{\sigma_1} \cdot \dfrac{\begin{vmatrix} r_{21} & r_{13} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix}}$ (d) $\dfrac{\sigma_3}{\sigma_1} \cdot \dfrac{\begin{vmatrix} r_{13} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} r_1 & r_{13} \\ r_{13} & 1 \end{vmatrix}}$.

5. The partial correlation coefficient of x_1 and x_2, when x_3 is constant, is given by :

(a) $r_{123} = \dfrac{b_{123}}{b_{213}}$ (b) $r_{123} = \dfrac{b_{213}}{b_{123}}$

(c) $r_{123} = b_{123} \times b_{213}$ (d) $r_{123} = \sqrt{(b_{123} \times b_{213})}$.

6. For a trivariate distribution, which of the following relations is true :

(a) $\Sigma\, x_1\, x_{2\cdot 13} = 0$ (b) $\Sigma\, x_{23}\, x_{123} = 0$

(c) $\Sigma\, x_{123}\, x_{12} = \Sigma\, x_{123}\, x_1$ (d) $\Sigma\, x_1\, x_{123} = 0$.

7. If $R_{123} = 0$, then which of the following is true :

(a) $R_{2\cdot 13} = 1, R_{3\cdot 12} = 0$ (b) $R_{2\cdot 13} = 0, R_{3\cdot 12} = 1$

(c) $R_{2\cdot 13} = 1, R_{3\cdot 12} = 1$ (d) $R_{2\cdot 13} = 0, R_{3\cdot 12} = 0$.

8. The correlation coefficient between the residuals $x_{1\cdot 23}$ and $x_{2\cdot 13}$ is :

(a) Twice (b) Thrice

(c) Equal

(d) Equal but of opposite sign of the correlation coefficient between x_{13} and x_{23}.

9. If $R_{123} = 0$, then which of the following relation is true for $R_{2\cdot 13}$ and $R_{3\cdot 12}$:

(a) Both are necessarily zero (b) One of the two is necessarily zero

(c) Both are not necessarily zero (d) One of the two is not necessarily zero.

10. If $r_{12} = k, r_{23} = -k$, then which of the following range is true for r_{13} :

(a) Between $1 - 2k^2$ and 1 (b) Between -1 and $1 - 2k^2$

(c) Between $1 + 2k^2$ and $1 - k^2$ (d) Between $1 - k^2$ and $1 + k^2$

11. Which of the following relations is true for a trivariate distribution :

(a) $b_{123}\, b_{23\cdot 1}\, b_{312} = r_{13}\, r_{23}\, r_{3\cdot 1}$ (b) $b_{123}\, b_{23\cdot 1}\, b_{312} = r_{12}^2\, r_{23}^2\, r_{3\cdot 1}^2$

(c) $b_{123}\, b_{23\cdot 1}\, b_{312} = r_{123}\, r_{23\cdot 1}\, r_{312}$ (d) $b_{123}\, b_{23\cdot 1}\, b_{312} = r_{123}^2\, r_{23\cdot 1}^2\, r_{312}^2$.

ANSWERS

1. (b) **2.** (d) **3.** (c) **4.** (b) **5.** (d) **6.** (d) **7.** (c) **8.** (d)
9. (c) **10.** (b) **11.** (c).