Calculation of Mean and Standard Deviation

Amt. of Scholarship (X)	No. of Children (f)	A = 50, d = X − 50	$d' = \frac{d}{10}$	fd'	fd'^2
30	10	−20	−2	−20	40
40	8	−10	−1	−8	8
50	7	0	0	0	0
60	3	+10	+1	3	3
70	2	+20	+2	4	8
	N = 30			Σfd' = −21	$\Sigma fd'^2 = 59$

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times i$$

$$= 50 - \frac{21}{30} \times 10 = 50 - \frac{210}{30} = 50 - 7 = 43$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$$

$$= \sqrt{\frac{59}{30} - \left(\frac{-21}{30}\right)^2} \times 10 = \sqrt{1.966 - 0.49} \times 10$$

$$= \sqrt{1.476} \times 10 = 1.214 \times 10 = 12.14$$

Variance

Variance is another measure of dispersion. The term variance was first used by R.A. Fisher in 1918. **Variance is the square of the standard deviation.** Symbolically,

$$\textbf{Variance} = (S.D.)^2 = \sigma^2$$

Calculation of Variance

(i) $\text{Variance} = \frac{\Sigma f(X - \bar{X})^2}{N}$ (Actual Mean Method)

(ii) $\text{Variance} = \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2$ (Assumed Mean Method)

(iii) $\text{Variance} = \left[\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2\right] \times i^2$ (Step Deviation Method)

Example 23. Calculate the mean and variance from the data given below:

Daily wages:	0—10	10—20	20—30	30—40	40—50
No. of workers:	2	7	10	5	3

Solution:

Calculation of Mean and Variance

Daily wages	f	M.V. (m)	A = 25, d = m − 25	$d' = \frac{d}{10}$	fd'	fd'^2
0—10	2	5	−20	−2	−4	8
10—20	7	15	−10	−1	−7	7
20—30	10	25 = A	0	0	0	0
30—40	5	35	+10	+1	+5	5
40—50	3	45	+20	+2	+6	12
	N = 27				Σfd' = 0	$\Sigma fd'^2 = 32$

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times i = 25 + \frac{0}{7} \times 10 = 25$$

$$\text{Variance} = \left[\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2\right] \times i^2 = \left[\frac{32}{27} - \left(\frac{0}{27}\right)^2\right] \times 10^2$$

$\Rightarrow$ $\sigma^2 = 1.185 \times 100 = 118.51$

$\therefore$ $\bar{X} = 25,\ \sigma^2 = 118.51$

EXERCISE 6.4

1. Calculate the standard deviation from the following data:

X:	63	67	64	59	61	67	68	66	63	61	68	61

[Ans. $\sigma = 3$]

2. Calculate the standard deviation from the following data using assumed mean method:

X:	48	75	54	60	63	69	72	51	57	56

[Ans. $\sigma = 8.62$]

3. Calculate the mean and standard deviation for the following data:

Size:	10	20	30	40	50	60	70
Frequency:	6	8	16	15	33	11	12

[Ans. $\bar{X} = 44.059,\ \sigma = 16.36$]

4. Calculate mean and standard deviation of the following series:

Daily wages:	0—10	10—20	20—30	30—40	40—50
No. of workers:	2	7	10	5	3

[Ans. $\bar{X} = 25,\ \sigma = 10.88$]

5. Calculate median and S.D. from the following data:

Variable:	21—25	26—30	31—35	36—40	41—45	46—50	51—55
Frequency:	5	15	28	42	15	12	3

[Ans. $M = 36.928,\ \sigma = 6.735$]

6. Calculate the mean and the standard deviation of the following series:

Marks (Above):	0	10	20	30	4-0	50	60	70
No. of students :	100	90	75	50	25	15	5	0

[Ans. $\overline{X} = 31, \sigma = 15.9$]

7. Calculate the mean and standard deviation from the following data:

Class Interval:	–40 to – 30	–30 to –20	–20 to –10	–10 to 0	0 to 10	10 to 20	20 to 30
Frequency:	10	28	30	42	65	180	10

[Ans. $\overline{X} = 4.29, \sigma = 14.75$]

8. The following table gives the marks obtained by a group of 80 students in an examination. Calculate the mean and variance.

Marks obtained	No. of students	Marks obtained	No. of students
10—14	2	34—38	10
14—18	4	38—42	8
18—22	4	42—46	4
22—26	8	46—50	6
26—30	12	50—54	2
30—34	16	54—58	4

[Ans. $\overline{X} = 33.5, \sigma^2 = 110.144$]

9. A charitable organisation decided to give old age pensions to people over 60 years of age. The scale of pension were fixed as follows:

Age group:	60—65	65—70	70—75	75—80	80—85
Pension per month (Rs.):	20	25	30	35	40

The ages of 25 persons who secured the pensions' rights are as given below:
74, 62, 84, 72, 61, 83, 72, 81, 64, 71, 63, 61, 60, 67, 74, 64, 79, 73, 75, 76, 69, 68, 78, 66, 67.
Calculate mean and S.D. of monthly pension. [Ans. $\overline{X} = 28.20, \sigma = 6.765$]

Combined Standard Deviation

Just as it is possible to calculate combined mean of two or more groups, similarly the combined standard deviation of two or more groups can be calculated. The combined standard deviation of two groups is denoted by σ_{12} and is computed as follows:

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

Where, σ_{12} = combined standard deviation;

σ_1 = standard deviation of the first group;

σ_2 = standard deviation of the second group;

$d_1 = \overline{X}_1 - \overline{X}_{12}$, $d_2 = \overline{X}_2 - \overline{X}_{12}$

The above formula can be extended to calculate the standard deviation of three or more groups. For example, combined S.D. of three groups is given by:

$$\sigma_{123} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_2 \sigma_3^2 + N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2}{N_1 + N_2 + N_3}}$$

Where, $d_1 = \overline{X}_1 - \overline{X}_{123}$; $d_2 = \overline{X}_2 - \overline{X}_{123}$; $d_3 = \overline{X}_3 - \overline{X}_{123}$

Example 24. Two samples of size 100 and 150 respectively have means 50 and 60 and standard deviations 5 and 6. Find the mean and standard of the combined sample of size 250.

Solution: Given, $N_1 = 100 \quad \overline{X}_1 = 50, \quad \sigma_1 = 5$

$N_2 = 150 \quad \overline{X}_2 = 60, \quad \sigma_2 = 6$

Now,

$$\overline{X}_{12} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2} = \frac{100 \times 50 + 150 \times 60}{100 + 150}$$

$$= \frac{100 \times 50 + 150 \times 60}{100 + 150}$$

$$= \frac{5000 + 9000}{250} = \frac{14000}{250} = 56$$

$$d_1 = \overline{X}_1 - \overline{X}_{12} = 50 - 56 = -6$$

$$d_2 = \overline{X}_2 - \overline{X}_{12} = 60 - 56 = +4$$

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times 25 + 150 \times 36 + 100(-6)^2 + 150(4)^2}{100 + 150}}$$

$$= \sqrt{\frac{2500 + 5400 + 3600 + 2400}{250}}$$

$$= \sqrt{\frac{13900}{250}} = 7.46$$

Hence, the combined mean is 56 and standard deviation is 7.46.

Example 25. For a group containing 100 observations, the arithmetic mean and standard deviation are 8 and $\sqrt{10.5}$. For 50 observations selected from the 100 observations, the mean and standard deviations are 10 and 2 respectively. Find the arithmetic mean and the standard deviations of the other half.

Solution: Given: $N = 100, \quad \overline{X}_{12} = 8 \quad \sigma_{12} = \sqrt{10.5}$

$N_1 = 50, \quad \overline{X}_1 = 10 \quad \sigma_1 = 2$

$N_2 = 100 - N_1 = 100 - 50 = 50$

We know that:

$$\bar{X}_{12}=\frac{N_1\bar{X}_1+N_2\bar{X}_2}{N_1+N_2}$$

$$8=\frac{50(10)+50(\bar{X}_2)}{100}$$

$$800=500+50\bar{X}_2$$

$$300=50\bar{X}_2$$

$$\bar{X}_2=\frac{300}{50}=6$$

$\therefore$

$$d_1=\bar{X}_1-\bar{X}_{12}=10-8=2 \Rightarrow d_1^2=4$$

$$d_2=\bar{X}_2-\bar{X}_{12}=6-8=-2 \Rightarrow d_2^2=4$$

$$\sigma_{12}=\sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_1d_1^2+N_2d_2^2}{N_1+N_2}}$$

Substituting the values, we get

$$\sqrt{10.5}=\sqrt{\frac{50\times4+50\sigma_2^2+50\times4+50\times4}{100}}$$

Squaring both sides,

$$10.5=\frac{200+50\sigma_2^2+200+200}{100}$$

$$10.5\times100=600+50\sigma_2^2$$

$$1050=600+50\sigma_2^2$$

$\therefore$ $50\sigma_2^2=450$

$\therefore$ $\sigma_2^2=\frac{450}{50}=9$

$\Rightarrow$ $\sigma_2=3$

Thus, $\bar{X}_2=6$, $\sigma_2=3$

Example 26. Find the missing information from the following:

	Group I	Group II	Group III	Combined
Number	50	60	90	200
Standard Deviation	6	7	8	7.746
Mean	113	—	115	116

Solution: We are given:

$$N=N_1+N_2+N_3=200 \quad N_1=50,\ N_3=90$$

$\therefore$ $N_2=N-(N_1+N_3)=200-140=60$

Now,

$$\bar{X}_{123}=\frac{N_1\bar{X}_1+N_2\bar{X}_2+N_3\bar{X}_3}{N_1+N_2+N_3}$$

We are given: $\bar{X}_1=113$, $\bar{X}_3=115$, $\bar{X}_{123}=116$

Substituting the values, we get

$$116=\frac{(50)(113)+(60)(\bar{X}_2)+(90)(115)}{200}$$

$$116\times200=50\times113+60\bar{X}_2+90\times115$$

$$23200=5650+60\bar{X}_2+10350$$

$\therefore$

$$60\bar{X}_2=23200-5650-10350=7200$$

$\therefore$

$$\bar{X}_2=\frac{7200}{60}=120$$

$$d_1=\bar{X}_1-\bar{X}_{123}=113-116=-3 \Rightarrow d_1^2=9$$

$$d_2=\bar{X}_2-\bar{X}_{123}=120-116=4 \Rightarrow d_2^2=16$$

$$d_3=\bar{X}_3-\bar{X}_{123}=115-116=-1 \Rightarrow d_3^2=1$$

$$\sigma_{123}=\sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_2\sigma_3^2+N_1d_1^2+N_2d_2^2+N_3d_3^2}{N_1+N_2+N_3}}$$

We are given: $\sigma_{123}=7.745$, $\sigma_1=6$, $\sigma_2=7$

Substituting the values, we get

$$7.746=\sqrt{\frac{50(36)+60(49)+90\sigma_3^2+50(9)+60(16)+90(1)}{50+60+90}}$$

$$7.746=\sqrt{\frac{1800+2940+90\sigma_3^2+450+960+90}{200}}$$

$$7.746=\sqrt{\frac{6{,}240+90\sigma_3^2}{200}}$$

Squaring both sides,

$$(7.746)^2=\frac{6{,}240+90\sigma_3^2}{200}$$

$$12000=6240+90\sigma_3^2$$

$\Rightarrow$ $90\sigma_3^2=12000-6240=5760$

$\Rightarrow$ $\sigma_3^2=\frac{5760}{90}=64$

$\Rightarrow$ $\sigma_3=\sqrt{64}=8$

Thus, $N_2=60$, $\bar{X}_2=120$, $\sigma_3=8$

IMPORTANT TYPICAL EXAMPLE

Example 27. The mean weight of 150 students is 60 kg. The mean weight of boys is 70 kg. with a standard deviation of 10 kg. For the girls, the mean weight is 55 kg and the standard deviation is 15 kg. Find the number of boys and girls and the combined standard deviation.

Solution: Given: $N = N_1 + N_2 = 150,\ \bar{X}_{12} = 60$

$$\bar{X}_1 = 70,\ \sigma_1 = 10,\ \bar{X}_2 = 55,\ \sigma_2 = 15$$

We have to determine the number of boys

$$\therefore \quad N_2 = 150 - N_1$$

Here, N_2 will be the number of girls and N_1 will be the number of boys.

We know, $$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Substituting the values, we get

$$60 = \frac{N_1(70) + (150 - N_1)(55)}{150}$$

$$60 \times 150 = 70\,N_1 + 8250 - 55\,N_1$$

$$9000 = 70\,N_1 + 8250 - 55\,N_1$$

$$9000 = 8250 + 15\,N_1$$

$$\Rightarrow \quad 15N_1 = 9000 - 8250 = 750$$

$$\Rightarrow \quad N_1 = \frac{750}{15} = 50$$

Hence, $N_2 = 150 - 50 = 100$

Thus, the number of boys and girls are 50 and 100 respectively.

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

Here, $N_1 = 50,\ \sigma_1 = 10,\ N_2 = 100,\ \sigma_2 = 15$

$$d_1 = \bar{X}_1 - \bar{X}_{12} = 70 - 60 = 10 \Rightarrow d_1^2 = 100$$

$$d_2 = \bar{X}_2 - \bar{X}_{12} = 55 - 60 = -5 \Rightarrow d_2^2 = 25$$

Substituting the values, we get

$$\sigma_{12} = \sqrt{\frac{50(100) + 100(225) + 50(100) + 100(25)}{50 + 100}}$$

$$\sigma_{12} = \sqrt{\frac{5000 + 22500 + 5000 + 2500}{150}} = \sqrt{\frac{35000}{150}} = \sqrt{233.33} = 15.28.$$

Thus, combined S.D. is 15.28.

EXERCISE 6.5

1. Two samples of size 40 and 60 respectively have means 20 and 25 and standard deviations 5 and 6 respectively. Find the combined mean and standard deviation of size 100.
   [**Ans.** $\bar{X}_{12} = 23,\ \sigma_{12} = 6.13$]
2. For two groups of observations the following results were available :

Group I	Group II
$\Sigma(X-5) = 8$	$\Sigma(X-8) = -10$
$\Sigma(X-5)^2 = 40$	$\Sigma(X-8)^2 = 70$
$N_1 = 20$	$N_2 = 25$

   Find mean and standard deviation of both the groups taken together.
   [**Hint:** See Example 45] [**Ans.** $\bar{X}_{12} = 6.62,\ \sigma_{12} = 1.864$]
3. The mean height of the students in a class is 152 cm. The mean height of boys is 158 cm with a standard deviation of 5 cm. And the mean height of girls is 148 cm with a standard deviation of 4 cm. Find the percentage of boys in the class and also the S.D of heights of all the students in the class. [**Ans.** Percentage of boys $= 40\%, \sigma_{12} = 6.603$]
4. The first of two subgroups has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$, find the mean and standard deviation of the second such group. [**Ans.** $\bar{X}_2 = 16,\ \sigma_2 = 4$]

● Correcting Incorrect Values of Mean and Standard Deviation

In certain cases, mean and standard deviation are calculated by using one or two incorrect values of the variable. Just as we can correct an incorrect mean, similarly, there is a procedure of correcting an incorrect standard deviation.

Steps: The various steps in the calculation of correct S.D. are as follows:

(*i*) Find out incorrected sum of the squared values of the variable, *i.e.*, find ΣX^2. This is to be found by using the following formula which involves incorrected $\bar{X}$ and σ.

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$$

or $$\sigma^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2$$

$$\therefore \text{ Incorrected } \Sigma X^2 = N[\sigma^2 + (\bar{X})^2]$$

(*ii*) Find corrected ΣX^2. To do so, we subtract the square of the incorrect item from incorrected ΣX^2 and add the square of correct item to incorrected ΣX^2. Thus,

$$\text{Corrected } \Sigma X^2 = \text{Incorrected } \Sigma X^2 - (\text{Incorrect value})^2 + (\text{Correct value})^2$$

(*iii*) Apply the following formula:

$$\text{Corrected } \sigma = \sqrt{\frac{\text{Corrected } \Sigma X^2}{N} - (\text{Corrected } \bar{X})^2}$$

Example 28. For a group of 100 observations, the mean and standard deviation were found to be 60 and 5 respectively. Later on it was discovered that a correct item 50 was wrongly copied as 30. Find the correct mean and standard deviation.

Solution: Given: $N = 100,\ \overline{X} = 60,\ \sigma = 5$

Calculation of Correct Mean

$$\overline{X} = \frac{\Sigma X}{N}$$

$$60 = \frac{\Sigma X}{100}$$

or $$\Sigma X = 6000$$

Incorrected $\Sigma X = 6000$

Corrected $\Sigma X = 6000 + \text{Correct item} - \text{Incorrect item}$

$$= 6000 + 50 - 30 = 6020$$

Hence, Corrected $\overline{X} = \frac{6020}{100} = 60.20$

Calculation of Correct S.D.

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\overline{X})^2}$$

Putting the values, we get

$$5 = \sqrt{\frac{\Sigma X^2}{100} - (60)^2}$$

Squaring both sides, we get

$$25 = \frac{\Sigma X^2}{100} - 3600$$

$$25 + 3600 = \frac{\Sigma X^2}{100}$$

$\therefore$ Corrected $\Sigma X^2 = 100[25 + 3600] = 362500$

Corrected $\Sigma X^2 = 362500 + (\text{Correct item})^2 - (\text{Incorrect item})^2$

$$= 362500 + (50)^2 - (30)^2$$

$$= 362500 + 2500 - 900$$

$\therefore$ Corrected $\Sigma X^2 = 364100$

Corrected $\sigma = \sqrt{\frac{364100}{100} - (60.20)^2}$

$$= \sqrt{3641.00 - 3624.04}$$

$$= \sqrt{16.96} \doteq 4.12$$

$\therefore$ Corrected $\overline{X} = 60.20$, Corrected $\sigma = 4.12$

IMPORTANT TYPICAL EXAMPLE

Example 29. The mean, standard deviation and range of a symmetrical distribution of weights of a group of 20 boys are 40 kgs, 5 kgs, and 6 kgs respectively. Find the mean and standard deviation of the group if the lightest and heaviest boys are excluded.

Solution: Since, the distribution is given to be symmetrical, the mean will lie at the middle of the range.

Therefore, the weight of the heaviest boy = 40 + 3 = 43 kgs and the weight of the lightest boy = 40 − 3 = 37 kgs.

We are given that $\overline{X} = 40, \sigma = 5$ and $N = 20$

$$\overline{X} = \frac{\Sigma X}{N} \quad \text{or} \quad 40 = \frac{\Sigma X}{20} \Rightarrow \Sigma X = 800$$

Corrected $\Sigma X = 800 - 43 - 37 = 720.$

$\therefore$ Corrected $\overline{X} = \frac{720}{18} = 40.$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\overline{X})^2}$$

$$5 = \sqrt{\frac{\Sigma X^2}{20} - (40)^2}$$

Squaring both sides,

$$(5)^2 = \frac{\Sigma X^2}{20} - (40)^2$$

$\therefore$ $\Sigma X^2 = 20[25 + 1{,}600] = 20 \times 1625 = 32{,}500$

Corrected $\Sigma X^2 = 32{,}500 - 43^2 - 37^2 = 29{,}282$

Corrected $\sigma = \sqrt{\frac{29{,}282}{18} - (40)^2} = \sqrt{26.777} = 5.17$

$\therefore$ Corrected $\sigma = 5.17$

EXERCISE 6.6

1. A student obtained the mean and standard deviation of 100 observations as 40 and 5 respectively. It was later found that one observation was wrongly copied as 50, the correct figure being 40. Find the correct mean and standard deviation. **[Ans.** $\overline{X} = 39.3, \sigma = 4.9$]
2. During nine days in a festival the highest sale of a shop was on Sunday and Rs 90 more than the average sale for other days. If the standard deviation of the sale during the festival is 33.33, find the standard deviation leaving that the highest sale.
[Hint: See Example 49] **[Ans.** $\sigma = 15.4$]

3. The mean age and standard deviation of a group of 200 persons (grouped in intervals 0—5, 5—10, ... etc.) were found to be 40 and 15. Later on it was discovered that the age 43 was misread as 53. Find the correct mean and standard deviation. [Ans. $\overline{X} = 39.95, \sigma = 14.97$]
[Hint: See Example 53]
4. The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking, it was found that one item 8 was incorrect. Calculate mean and standard deviation if: (*i*) the wrong item is omitted. (*ii*) it is replaced by 12.
[Ans. (*i*) $\overline{X} = 10.11, \sigma = 1.997$; (*ii*) $\overline{X} = 10.2, \sigma = 1.99$]

Determination of Missing Values

In certain situations, the values of one or more items may be missing from the given information. The method of computing missing values is explained with the help of the following examples:

Example 30. The mean of 5 observations is 4.4 and the variance is 8.24. If three of the observations are 4, 6 and 9, find the other two.

Solution: Let the missing observations be x_1 and x_2

X	X^2
4	16
6	36
9	81
x_1	x_1^2
x_2	x_2^2
$\Sigma X = x_1 + x_2 + 19$	$\Sigma X^2 = 133 + x_1^2 + x_2^2$

Here we are given $N = 5, \overline{X} = 4.4, \sigma^2 = 8.24$

As $\overline{X} = \frac{\Sigma X}{N}$

$\therefore$ $4.4 = \frac{x_1 + x_2 + 19}{5}$ or $x_1 + x_2 + 19 = 22$

$\therefore$ $x_1 + x_2 = 3 \Rightarrow x_2 = 3 - x_1$ ...(*i*)

Now, $(S.D.)^2 = Variance = \frac{\Sigma X^2}{N} - (\overline{X})^2$

$8.24 = \frac{133 + x_1^2 + x_2^2}{5} - (4.4)^2$

$\therefore$ $133 + x_1^2 + x_2^2 = 5[8.24 + 19.36]$

$\Rightarrow$ $133 + x_1^2 + x_2^2 = 138$

$\Rightarrow$ $x_1^2 + x_2^2 = 5$ ...(*ii*)

From (*i*) and (*ii*), $x_1^2 + (3 - x_1)^2 = 5$

$\Rightarrow x_1^2 + 9 + x_1^2 - 6x_1 = 5 \Rightarrow 2x_1^2 - 6x_1 + 4 = 0$

$\Rightarrow$ $x_1^2 - 3x_1 + 2 = 0$ $\Rightarrow$ $(x_1 - 1)(x_1 - 2) = 0$

$\therefore$ $x_1 = 1, 2$

If $x_1 = 1, x_2 = 2$ and if $x_1 = 2, x_2 = 1$.

Example 31. Mean and standard deviation of the following continuous series are 135.3 and 9.6 respectively. The distribution after taking step deviations is as follows:

d:	–4	–3	–2	–1	0	1	2	3
f:	2	5	8	18	22	13	8	4

Determine the actual class intervals.

Solution: Here, d is identical to d' which is referred as $d' = \frac{X - A}{i}$.

In order to ascertain the class intervals, we need two values—size of the class interval (*i*) and assumed mean (A). From the formula of S.D., we can determine the size of class interval (*i*) and from the formula of mean, we can determine the assumed mean (A).

Calculations for Determining *i* and *A*

d	f	fd	fd^2
–4	2	–8	32
–3	5	–15	45
–2	8	–16	32
–1	18	–18	18
0	22	0	0
1	13	13	13
2	8	16	32
3	4	12	36
	$N = 80$	$\Sigma fd = -16$	$\Sigma fd^2 = 208$

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

$$9.6 = \sqrt{\frac{208}{80} - \left(\frac{-16}{80}\right)^2} \times i$$

$\Rightarrow$ $9.6 = \sqrt{2.6 - 0.04} \times i$

$\Rightarrow$ $9.6 = \sqrt{2.56} \times i$

$9.6 = 1.6 \times i$

$i = \frac{9.6}{1.6} = 6$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

$$135.3 = A + \frac{(-16)}{80} \times 6 = A - 1.2$$

$$\therefore \quad A = 136.5$$

Using the values of A, i and d, we can write the mid-values as given below:

$$d = \frac{X - A}{i} \Rightarrow di = X - A$$

$$\Rightarrow \quad X = A + id$$

Here, $A = 136.5$, $i = 6$.

d:	–4	–3	–2	–1	0	1	2	3
M.V. $(X) = A + id$	136.5 + (6) (–4) = 112.5	118.5	124.5	130.3	136.5	142.5	148.5	154.5

The various class intervals shall be obtained by using the formula:

$$m \pm \frac{i}{2}$$

The various class intervals are:

$$112.5 \pm \frac{6}{2}, \quad 118.5 \pm \frac{6}{2}, \quad 124.5 \pm \frac{6}{2}, \quad 130.5 \pm \frac{6}{2},$$

$$136.5 \pm \frac{6}{3}, \quad 142.5 \pm \frac{6}{3}, \quad 148.5 \pm \frac{6}{3}, \quad 154.5 \pm \frac{6}{2}$$

i.e., 109.5 – 115.5, 115.5 – 121.5, 121.5 – 127.5, 127.5 – 133.5, 133.5 – 139.5, 139.5 – 145.5, 145.5 – 151.5, 151.5 – 157.5

Thus,

Class Intervals	f	Class Intervals	f
109.5—115.5	2	133.5—139.5	22
115.5—121.5	5	139.5—145.5	13
121.5—127.5	8	145.5—151.5	8
127.5—133.5	18	151.5—157.5	4

EXERCISE 6.7

1. The mean of 5 observations is 4.4 and variance is 8.24. If three of five observations are 1, 2, 6, find the other two.

   [Ans. If $x_1 = 9, x_2 = 4$ and if $x_1 = 4, x_2 = 9$]

2. Mean and S.D. of the following continuous series are 31 and 15.9. The distribution after taking step deviations is as follows:

d:	–3	–2	–1	0	1	2	3
f:	10	15	25	25	10	10	5

Determine the actual class intervals.

[Ans. 0—10, 10—20, 20—30, 30—40, 40—50, 50—60, 60—70]

Mathematical Properties of Standard Deviation

The important mathematical properties of standard deviation are as follows:

(1) The standard deviation of first n natural numbers can be found from the following formula:

$$\sigma = \sqrt{\frac{1}{12} \cdot (n^2 - 1)}$$

For example, the standard deviation of the first 5 natural numbers is given as:

$$\sigma = \sqrt{\frac{1}{12} \cdot (5^2 - 1)} = \sqrt{\frac{24}{12}} = \sqrt{2} = 1.414$$

(2) The combined S.D. of two or more groups can be found by using the following formula:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$ where, $d_1 = \bar{X}_1 - \bar{X}_{12}$, $d_2 = \bar{X}_2 - \bar{X}_{12}$

(3) The sum of the squares of the deviations of the items taken from arithmetic mean is least. That is why standard deviation is computed from the A.M.

(4) If a constant amount 'a' is added or subtracted from each item of a series, then S.D. remains unaffected, *i.e.*, S.D. is independent of the change of origin.

(5) If each item of a series is multiplied or divided by a constant 'a', then S.D. is affected by the same amount, *i.e.*, S.D. is not independent of the change of scale.

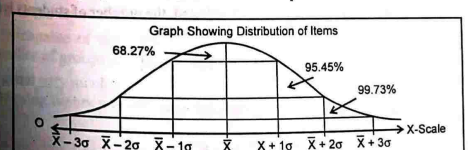
(6) The standard deviation has the following relation to the arithmetic mean in a symmetrical distribution:

$\bar{X} \pm 1\sigma$ includes 68.27% of the items.

$\bar{X} \pm 2\sigma$ includes 95.45% of the items.

$\bar{X} \pm 3\sigma$ includes 99.73% of the items.

The following figure illustrate the relationship:



(7) The standard deviation has the following relation to quartile deviation (Q.D.) and mean deviation (M.D.) in a symmetrical (or normal) distribution:

$$Q.D. = \frac{2}{3}\sigma, \quad M.D. = \frac{4}{5}\sigma, \quad Q.D : M.D : S.D :: 10 : 12 : 15$$

IMPORTANT TYPICAL EXAMPLES

Example 32. The following table gives the distribution of marks obtained by 90 students in an examination:

Marks:	0—10	10—20	20—30	30—40	40—50	50—60
No. of Students:	4	10	20	35	15	6

Calculate (*i*) Mean, (*ii*) Standard deviation and (*iii*) Percentage of students lying within the range (*a*) $\overline{X} \pm 1\sigma$ and (*b*) $\overline{X} \pm 2\sigma$.

Solution:

Calculation of $\overline{X}$ and σ

Classes	f	M.V. (m)	$d = m - A$	$d' = \frac{d}{i}$	fd'	fd'^2
0—10	4	5	–30	–3	–12	36
10—20	10	15	–20	–2	–20	40
20—30	20	25	–10	–1	–20	20
30—40	35	35 = A	0	0	0	0
40—50	15	45	+10	+1	15	15
50—60	6	55	+20	+2	12	24
	$N = 90$				$\Sigma fd' = -25$	$\Sigma fd'^2 = 135$

(*i*) $\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 35 + \frac{(-25)}{90} \times 10 = 32.22$

(*ii*) $\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{135}{90} - \left(\frac{-25}{90}\right)^2} \times 10 = 11.92$

(*iii*) $\overline{X} \pm 1\sigma = 32.22 \pm 11.92$

The limit of the range $\overline{X} \pm \sigma$ are 20.3 and 44.14. Under the assumption that observations in a class are uniformly distributed, the number of students lying within these limits are:

$$\frac{20}{10} \times (30 - 20.3) + 35 + \frac{15}{10}(44.14 - 40) = 60.61$$

$\therefore$ Percentage of students $= \frac{60.61}{90} \times 100 = 67.34\%$

$$\overline{X} \pm 2\sigma = 32.22 \pm 2 \times 11.92$$

Similarly, limits of the range $\overline{X} \pm 2\sigma$ are 8.38 and 56.06 and the number of students lying within these limits are:

$$\frac{4}{10} \times (10 - 8.38) + 10 + 20 + 35 + 15 + \frac{6}{10} \times (56.06 - 50) = 84.29$$

$\therefore$ Percentage of students $= \frac{84.29}{90} \times 100 = 93.66\%$

Example 33. You are the incharge of the rationing department of a state affected by food shortage. The following information is received from your local investigators :

Area	Mean Calories	Standard Deviation of Calories
X	2,500	500
Y	2,200	300

The estimated requirement of an adult is taken at 3,000 calories daily and absolute minimum at 1,250. Comment on the reported figures and determine which area needs more urgent action.

Solution: We shall compute the 3-sigma limits $\overline{X} \pm 3\sigma$ for each area, which will include approximately 99.73% of the population observation [assuming that the distribution is approximately normal].

	3-σ Limits = $\overline{X} \pm 3\sigma$
Area X	2500 ± 3 × 500 = 2500 ± 1500 = (1000, 4000)
Area Y	2200 ± 3 × 300 = 2200 ± 900 = (1300, 3100)

The absolute daily minimum calories requirement for a person is 1250. From the above figures we observe that almost all the persons in the area Y are getting more than the minimum calories requirement as the lower limit in this area is 1300. However, since in the area X, the lower 3-σ limit is 1000 which is less than 1250, quite a number of people in area X are not getting the minimum requirement of 1250 calories. Hence, as the incharge of the rationing department, it becomes my duty to take urgent action for the people of area X.

Merits and Demerits of Standard Deviation

Merits

(*i*) It is a rigidly defined.

(*ii*) It is based on all the observations.

(*iii*) It is capable of being treated mathematically. For example, if standard deviations of a number of groups are known, their combined standard deviation can be computed.

(*iv*) It is not very much affected by the fluctuations of sampling and, therefore, is widely used in sampling theory and test of significance.

Demerits

(*i*) As compared to the quartile deviation and range, etc., it is difficult to understand and difficult to calculate.

(*ii*) It gives more importance to extreme observations.

(*iii*) Since, it depends upon the units of measurement of the observations, it cannot be used to compare the dispersions of the distributions expressed in different units.

EXERCISE 6.8

1. Calculate the S.D. of the first 7 natural numbers. [Ans. $\sigma = 2$]
2. If mean and standard deviation of 75 observations is 40 and 8 respectively, find the new mean and standard deviation if :
   (*i*) Each observation is multiplied by 5.
   (*ii*) 7 is added to each observation.
   [Hint: See Example 50] [Ans. (*i*) New mean = 200, New S.D. = 40
   (*ii*) New mean = 47, New S.D. = 8]
3. 5 observations of a series are 4, 6, 8, 12 and 15. Their mean and standard deviation are 9 and 4 respectively. Make such alternations in the terms of the series that new standard deviation is 20 and mean is 50.
   [Hint: See Example 51] [Ans. 25, 35, 45, 65 and 80]
4. The following table gives the length of life of 300 persons:

Age (X):	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79	80—89
No. of persons:	6	15	33	39	45	27	18	10	7

   Calculate (*i*) Mean (*ii*) Standard deviation (*iii*) The percentage of persons whose length of life falls within $\overline{X} \pm 2\sigma$. [Ans. $\overline{X} = 41.85, \sigma = 18.5, 95\%$]
5. From the following figures determine the percentage of cases which lie outside the range:
   $\overline{X} \pm \sigma, \overline{X} \pm 2\sigma, \overline{X} \pm 3\sigma$.
   115, 117, 121, 125, 116, 120, 118, 117, 119, 116, 122, 124, 123, 118, 120, 118, 126, 127, 122, 123.
   [Ans. $\overline{X} = 120.35, \sigma = 3.45$, 3.5%, 0%, 0%]
6. A collar manufacturer is considering the production of a new style of collar to attract young men. The following statistics of neck circumference are available based on measurements of a typical group of college students. Compute the SD and use the criterion ($\overline{X} \pm 3\sigma$), to determine the largest and smallest sizes of collars, he should make in order to meet the needs of practically all his customers, bearing in mind, that collars are worn, on average $\frac{3}{4}$ inch larger than neck size.

Mid-points:	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
f:	4	19	30	63	66	29	18	1	1

   [Hints: See Example 52] [Ans. $\overline{X} = 14.232, \sigma = 0.719, \overline{X} \pm 3\sigma + \frac{3}{4} = 12.825$ to 17.139]

(5) COEFFICIENT OF VARIATION

Coefficient of variation is an important relative measure of dispersion. It was developed by Karl Pearson and is widely used in comparing the variability of two or more series. Coefficient of variation is denoted by C.V. and is given by:

$$\text{Coefficient of Variation (C.V.)} = \frac{\sigma}{\overline{X}} \times 100$$

Steps for Calculation

(*i*) First of all calculate $\overline{X}$.
(*ii*) Calculate σ.
(*iii*) Put the value of $\overline{X}$ and σ in the above formula.

Uses of Coefficient of Variation

Coefficient of variation is used to compare the variability, homogenity, stability, consistency and uniformity of two or more series. The series having less value of the coefficient of variation is considered more consistent in comparison to a series having a higher value of the coefficient of variation.

Example 34. From the prices of shares of X and Y given below, state which share is more stable in value:

X:	41	44	43	48	45	46	49	50	42	40
Y:	91	93	96	92	90	97	99	94	98	95

Solution: For finding out which share is more stable in value, we have to compare the coefficient of variation.

Calculation of C.V.

X	$A = 45$ dx	dx^2	Y	$A = 95$ dy	dy^2
41	−4	16	91	−4	16
44	−1	1	93	−2	4
43	−2	4	96	1	1
48	3	9	92	−3	9
45 = A	0	0	90	−5	25
46	1	1	97	2	4
49	4	16	99	4	16
50	5	25	94	−1	1
42	−3	9	98	3	9
40	−5	25	95 = A	0	0
$N = 10$ $\Sigma X = 448$	$\Sigma dx^2 = -2$	$\Sigma dx^2 = 106$	$\Sigma Y = 945$	$\Sigma dy = -5$	$\Sigma dy^2 = 85$

Share X: $\overline{X} = \frac{\Sigma X}{N} = \frac{448}{10} = 44.8$

$$\sigma_x = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2} = \sqrt{\frac{106}{10} - \left(\frac{-2}{10}\right)^2} = 3.25$$

$$\therefore C.V._X = \frac{3.25}{44.8} \times 100 = 7.25\%$$

Share Y: $Y = \frac{\Sigma Y}{N} = \frac{945}{10} = 94.5$

$$\sigma_y = \sqrt{\frac{\Sigma dy^2}{N} - \left(\frac{\Sigma dy}{N}\right)^2} = \sqrt{\frac{85}{10} - \left(\frac{-5}{10}\right)^2} = 2.87$$

$$\therefore C.V._Y = \frac{2.87}{94.5} \times 100 = 3.03\%$$

Since, the coefficient of variation is less for share Y, hence share Y is more stable in price.

Example 35. The scores of two batsmen A and B in ten innings during a certain match are:

A:	32	28	47	63	71	39	10	60	96	14
B:	19	31	48	53	67	90	10	62	40	80

Find out who is a better scorer and who is more consistent batsman.

Solution: For finding out which of the two batsman is a better scorer, we have to compare the arithmetic means and for finding out which batsman is more consistent, we have to compare the coefficient of variation.

Calculation of $\overline{X}$ and C.V.

X	$\overline{X}=46$ $x=X-\overline{X}$	x^2	Y	$\overline{Y}=50$ $y=Y-\overline{Y}$	y^2
32	– 14	196	19	– 31	961
28	– 18	324	31	– 19	361
47	+ 1	1	48	– 2	4
63	+ 17	289	53	+ 3	9
71	+ 25	625	67	+ 17	289
39	– 7	49	90	+ 40	1600
10	– 36	1296	10	– 40	1600
60	+ 14	196	62	+ 12	144
96	+ 50	2500	40	– 10	100
14	– 32	1024	80	+ 30	900
$N=10$ $\Sigma X=460$	$\Sigma x=0$	$\Sigma x^2=6500$	$\Sigma Y=500$	$\Sigma y=0$	$\Sigma y^2=5968$

Batsman A: $\overline{X}=\frac{\Sigma X}{N}=\frac{460}{10}=46,$ **Batsman B:** $\overline{Y}=\frac{\Sigma Y}{N}=\frac{500}{10}=50$

Since the arithmetic mean is higher for batsman B, hence batsman B is a better scorer.

$\sigma=\sqrt{\frac{\Sigma x^2}{N}}=\sqrt{\frac{6500}{10}}=25.495$ $\sigma=\sqrt{\frac{\Sigma y^2}{N}}=\sqrt{\frac{5968}{10}}=24.43$

$\therefore C.V.=\frac{\sigma}{\overline{X}}=\frac{25.49}{46}\times 100=55.41\%$ $\therefore C.V.=\frac{\sigma}{\overline{X}}=\frac{24.43}{50}\times 100=48.86\%$

Since, the coefficient variation is less for batsman B, hence batsman B is more consistent.

Example 36. Goals scored by two teams A and B in a football session were as follows:

No. of goals scored:	0	1	2	3	4
No. of matches by A:	27	9	8	5	4
No. of matches by B:	17	9	6	5	3

By calculating the coefficient of variation in each case, find which team may be considered more consistent.

Solution: **For team A:**

X (goals)	f (No. of matches)	$A=2$ $d=X-A$	fd	fd^2
0	27	–2	–54	108
1	9	–1	–9	9
2 = A	8	0	0	0
3	5	+1	+5	5
4	4	+2	+8	16
	$N=53$		$\Sigma fd=-50$	$\Sigma fd^2=138$

$$\overline{X}=A+\frac{\Sigma fd}{N}=2-\frac{50}{53}=2-0.94=1.06$$

$$\sigma=\sqrt{\frac{\Sigma fd^2}{N}-\left(\frac{\Sigma fd}{N}\right)^2}=\sqrt{\frac{138}{53}-\left(\frac{-50}{53}\right)^2}$$

$$=\sqrt{2.603-0.889}=\sqrt{1.74}=1.309$$

$$\text{C.V. for Team A}=\frac{\sigma}{\overline{X}}\times 100=\frac{1.309}{1.06}\times 100=123.49\%$$

For team B:

X (goals)	f (No. of matches)	$A=2$ $d=X-A$	fd	fd^2
0	17	–2	–34	68
1	9	–1	–9	9
2	6	0	0	0
3	5	+1	5	5
4	3	+2	6	12
	$N=40$		$\Sigma fd=-32$	$\Sigma fd^2=94$

$$\overline{X}=A+\frac{\Sigma fdx}{N}=2-\frac{32}{40}=2-0.8=1.2$$

$$\sigma=\sqrt{\frac{\Sigma fd^2}{N}-\left(\frac{\Sigma fd}{N}\right)^2}=\sqrt{\frac{94}{40}-\left(\frac{-32}{40}\right)^2}$$

$$=\sqrt{2.35-0.64}=\sqrt{1.14}=1.307$$

$$\text{C.V. for Team B}=\frac{\sigma}{\overline{X}}\times 100=\frac{1.307}{1.2}\times 100=108.9\%$$

Since, the coefficient of variation of Team B is less than team A, so team B is more consistent.

Note: Normally, the value of C.V. does not exceed 100% but in Z-shaped distribution, the value of C.V. exceeds 100%

Example 37. You are given below the daily wages paid to workers in two factories X and Y:

Daily wages	No. of workers	
	Factory X	Factory Y
12—13	15	25
13—14	30	40
14—15	44	60
15—16	60	35
16—17	30	12
17—18	14	15
18—19	7	5

Using appropriate measures, answer the following:

(*i*) Which factory pays higher average wages?

(*ii*) Which factory has a more consistent wage structure?

Solution: For finding out which factory pays higher wages, we have to compute the arithmetic means and for finding out which factory has a more consistent wage structure, we have to compare the coefficient of variation:

Calculation of $\overline{X}$ and C.V.

Wages	M.V. (m)	$d = m - A$	Factory X			Factory Y		
			f	fd	fd^2	f	fd	fd^2
12—13	12.5	−3	15	−45	135	25	−75	225
13—14	13.5	−2	30	−60	120	40	−80	160
14—15	14.5	−1	44	−44	44	60	−60	60
15—16	15.5	0	60	0	0	35	0	0
16—17	16.5	+1	30	+30	30	12	+12	12
17—18	17.5	+2	14	+28	56	15	+30	60
18—19	18.5	+3	7	+21	61	5	+15	45
			$N = 200$	$\Sigma fd = -70$	$\Sigma fd^2 = 448$	$N = 192$	$\Sigma fd = -158$	$\Sigma fd^2 = 562$

(*i*) **Factory X**

$$\overline{X} = A + \frac{\Sigma fd}{N} = 15.5 - \frac{70}{200} = 15.15$$

Factory Y

$$\overline{Y} = A + \frac{\Sigma fd}{N} = 15.5 - \frac{158}{192} = 14.67$$

Since, the arithmetic mean is higher for factory X, hence factory X pays higher average wage.

(*ii*) **Factory X**

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

$$= \sqrt{\frac{448}{200} - \left(\frac{-70}{200}\right)^2}$$

$$= \sqrt{2.24 - 0.1225} = 1.445$$

$$C.V._X = \frac{\sigma}{\overline{X}} \times 100$$

$$= \frac{1.455}{15.15} \times 100 = 9.60\%$$

Factory Y

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

$$= \sqrt{\frac{562}{192} - \left(\frac{-158}{192}\right)^2}$$

$$\sigma = \sqrt{2.93 - 0.677} \simeq 1.504$$

$$C.V._Y = \frac{\sigma}{\overline{Y}} \times 100$$

$$= \frac{1.504}{14.67} \times 100 = 10.25\%$$

Since, the coefficient of variation is less for factory X, hence factory X has more consistent wage structure.

Example 38. An analysis of the monthly wages paid to workers in firm A and B belonging to the same industry gives the following results:

	Firm A	Firm B
No. of workers:	500	600
Average monthly wage (Rs.):	186	175
Variance of distribution of wages (Rs.)	81	100

(*i*) Which firm pays a larger wage bill?

(*ii*) In which firm is there greater variability in individual wages?

(*iii*) Find the combined mean and standard deviation of wages of the two firms taken together.

Solution: (*i*) **Total wage bill of firm A**

$$\overline{X} = \frac{\Sigma X}{N}$$

$\therefore$ Total wage (ΣX) bill of firm A $= \overline{X} \times N = 186 \times 500 =$ Rs. 93,000.

Total wage bill of firm B

$$\overline{Y} = \frac{\Sigma Y}{N}$$

Total wage (ΣY) bill of firm B $= \overline{Y} \times N = 175 \times 600 =$ Rs. 1,05,000.

Hence, firm B pays larger wage bill.

(ii) To determine the firm in which there is greater variability in individual wages, we shall compare the coefficient of variation.

Firm A

$$C.V. = \frac{\sigma}{\overline{X}} \times 100$$

Given: $\sigma^2 = 81 \Rightarrow \sigma = \sqrt{81} = 9,\ \overline{X} = 186$

$$\therefore\ C.V._A = \frac{9}{186} \times 100 = 4.84\%$$

Firm B

$$C.V. = \frac{\sigma}{\overline{Y}} \times 100$$

Given: $\sigma^2 = 100 \Rightarrow \sigma = 10,\ \overline{Y} = 175$

$$\therefore\ C.V._B = \frac{10}{175} \times 100 = 5.71\%$$

Since, the coefficient of variation is greater in case of firm B, there is greater variability in individual wages of firm B.

***(iii)* Combined Mean and Standard Deviation.**

$$\overline{X}_{12} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2} = \frac{500 \times 186 + 600 \times 175}{500 + 600}$$

$$\overline{X}_{12} = \frac{93{,}000 + 1{,}05{,}000}{1{,}100} = \frac{1{,}98{,}000}{1{,}100} = \text{Rs. } 180$$

$$d_1 = \overline{X}_1 - \overline{X}_{12} = 186 - 180 = 6 \Rightarrow d_1^2 = 36$$

$$d_2 = \overline{X}_2 - \overline{X}_{12} = 175 - 180 = -5 \Rightarrow d_2^2 = 25$$

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{500 \times 81 + 600 \times 100 + 500 \times 36 + 600 \times 25}{500 + 600}}$$

$$= \sqrt{\frac{40500 + 60000 + 18000 + 15000}{1100}}$$

$$= \sqrt{\frac{133500}{1100}} = \sqrt{121.36} = 11.01$$

Example 39. Given: sum of squares of items = 2430, $\overline{X} = 7, N = 12$, find the coefficient of variation.

Solution: Given : $\Sigma X^2 = 2430, \overline{X} = 7, N = 12$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\overline{X})^2}$$

$$\sigma = \sqrt{\frac{2430}{12} - (7)^2} = \sqrt{153.5} = 12.38$$

$$\text{C.V.} = \frac{\sigma}{\overline{X}} \times 100$$

$$= \frac{12.38}{7} \times 100 = 176.85\%$$

EXERCISE 6.9

1. Batsmen X and Y scored following runs in different innings they played in a test series. Which of the two is a better scorer? Who is more consistent batsman?

X:	12	115	6	73	7	19	119	36	84	29
Y:	47	12	76	42	4	51	37	48	13	0

[**Ans.** $\overline{X} = 50, \sigma_X = 41.83, C.V._X = 83.66; \overline{Y} = 33, \sigma_Y = 23.37, C.V_Y = 70.81$
Batsman X is a better scorer; Batsman Y is a consistent batsman]

2. The following is the record number of bricks laid each day for 10 days by two layers A and B. Calculate the coefficient in each case and discuss the relative consistency of the two brick-layers.

A:	700	675	725	625	650	700	650	700	600	650
B:	550	600	575	550	650	600	550	525	625	600

If each of the values in respect of worker A is decreased by 10 and each of the values for worker B is increased by 50, how will it affect the results obtained earlier?

[Hint: See Example 43] [**Ans.** $\overline{X}_A = 667.5, \sigma_A = 37.15$, C.V.$_A$ = 5.56%
$\overline{X}_B = 582.5, \sigma_B = 37.15$, C.V.$_B$ = 6.38%]

3. Goals scored by two teams A and B in a footfall session were as follows:

No. of goals scored:	0	1	2	3	4	5
No. of matches by A:	15	10	7	5	3	2
No. of matches by B:	20	10	5	4	2	1

Find out which team is more consistent.

[**Ans.** C.V. for A = 102.06%, C.V. for B = 124.6%, Team A is more consistent]

4. A factory produces two types of electric bulbs A and B. In an experiment relating to their life, the following results were obtained:

Length of life (in hrs.)	No. of lamps (A)	No. of lamps (B)
500—700	5	4
700—900	11	30
900—1100	26	12
1100—1300	10	8
1300—1500	8	6

Which type of electric lamp do you prefer? Give reasons.

[**Ans.** C.V.(A) = 21.6, C.V.(B) = 23.4
As C.V. of A is less, so lamp A is preferred]

5. For two firms A and B belonging to the same industry, the following data is given:

	Firm A	Firm B
No. of wage earners:	586	648
Average monthly wage (Rs.):	52.5	47.5
Standard deviation:	10	11

(*i*) Which firm A or B pays larger amount as weekly wages?

(*ii*) Which firm shows greater variability in the wage rate?

(*iii*) Find the mean and S.D. of all workers in the two factories taken together.

[**Ans.** (*i*) Firm B, (*ii*) In firm B, there is greater variability, (*iii*) $\overline{X}_{12} = 49.87, \sigma_{12} = 10.83$]

6. From the following data, find out Range, Quartile Deviation, Mean Deviation and Coefficient of Variation when mean of the distribution is 37.4.

X:	0—10	10—20	20—30	30—40	40—50	50—60	60—70
f:	2	7	9	11	?	8	6

[**Ans.** Missing Frequency = 7, R = 70, Q_1 = 23.809, Q_3 = 59.88, Q.D. = 13.995, S.D. = 17.04, C.V. = 45.56%]

7. If 20 is substracted from every observation in a data set, then the coefficient of variation of the resulting set is 20%. If 40 is added to every observation of the same data, then the coefficient of variation of the resulting set of data is 10%. Find the $\overline{X}$ and σ of the original set of data.

[**Hint:** See Similar Example 57, $20 = \frac{\sigma \times 100}{\overline{X} - 20}$, $10 = \frac{\sigma \times 100}{\overline{X} + 40}$] [**Ans.** $\overline{X}$ = 80, σ = 12]

8. A fund manager is considering investment in the equity shares of one of two companies. The criterion for selecting the company for investment is consistency of return on net worth. The following data have been collected:

Financial Year	Return on Net worth (%)	
	Modern Industries Ltd. (MIL)	Pioneer Industries Ltd. (PIL)
2001—2002	19	20
2000—2001	20	24
1999—2000	16	16
1998—1999	13	15
1997—1998	12	10

You are required to identify the company in which the fund manager should invest.
[**Hint:** See Example 58]

[**Ans.** For MIL : $\overline{X}$ = 16%, σ = 3.16%, C.V. = 19.76%
For PIL : $\overline{X}$ = 17%, σ = 4.73%, C.V. = 27.83%
MIL is more consistent and investment be made in MIL]

9. The coefficient of variation of wages of male workers and female workers are 55 per cent and 70 per cent respectively, while the standard deviations are 22.0 and 15.4 respectively. Calculate the overall average wages of all workers given that 80 per cent of the workers are male.

[**Ans.** $\overline{X}_{12} = 36.4$]

10. The number of employees, wages per employee and variance of the wage per employee for two factories are given below:

	Factory A	Factory B
Number of employees	50	100
Average wage per employee per week (Rs.)	120	85
Variance of the wages per employee per week (Rs.)	9	16

(*i*) In which factory is there greater variation in the distribution of wages per employee?

(*ii*) Suppose in factory B, the wages of an employee were wrongly noted as Rs. 120 instead of Rs. 100. What would be the correct variance for factory B?

[**Ans.** (*i*) $C.V._A = 2.5$; $C.V._B = 4.71$ B is more variable (*ii*) $\sigma^2 = 5.96$]

(6) LORENZ CURVE

It is a graphical method of studying dispersion. Lorenz curve was given by famous statistician **Max O Lorenz.** Lorenz curve has great utility in the study of degree of inequality in the distribution of income and wealth between the countries. It is also useful for comparing the distribution of wages, profits, etc., over different business groups. Lorenz curve is a cumulative percentage curve in which the percentage of frequency (persons or workers) is combined with the percentage of other items such as income, profits, wages, etc.

Construction of a Lorenz Curve

Following steps are used while drawing a Lorenz Curve:

(*i*) The size of items (variable values) and frequencies are both cumulated. Taking grand total for each as 100, percentages are obtained for these various cumulative values.

(*ii*) Cumulative frequencies are plotted on X-axis while cumulative items are plotted on the Y-axis.

(*iii*) On both the axis, we start from 0 to 100 and both X and Y axis take the values from 0 to 100.

(*iv*) Draw a diagonal line **Y = X** joining the origin 0 (0.0) with the point P(100, 100). The line OP is called the **line of equal distribution.** Any point on this diagonal line shows the same per cent of X and Y.

(*v*) Plot the percentages of the cumulated values on the graph and a curve is obtained by joining different points. It is called **Lorenz curve.**

(*vi*) Closeness of the Lorenz curve to the line of equal distribution shows lesser variation in the distribution. Larger the gap between the line of equal distribution and the Lorenz curve, greater is the variation.

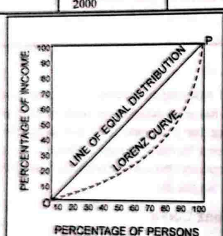
The following examples illustrate the procedure of drawing a Lorenz curve:

Example 40. Draw a Lorenz curve of the data given below:

Income:	100	200	400	500	800
No. of persons:	80	70	50	30	20

Solution:

Income	Cumulative Income	Cumulative percentage	No. of persons	Cumulative total	Cumulative percentage
100	100	$\frac{100}{2000}\times 100 = 5$	80	80	$\frac{80}{250}\times 100 = 32$
200	300	$\frac{300}{2000}\times 100 = 15$	70	150	$\frac{150}{250}\times 100 = 60$
400	700	$\frac{700}{2000}\times 100 = 35$	50	200	$\frac{200}{250}\times 100 = 80$
500	1,200	$\frac{1200}{2000}\times 100 = 60$	30	230	$\frac{230}{250}\times 100 = 92$
800	2,000	$\frac{2000}{2000}\times 100 = 100$	20	250	$\frac{250}{250}\times 100 = 100$



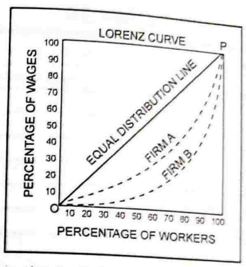
Example 41. Show inequality in wages in two different firms using Lorenz curve from following data:

Wages (Rs.):	50—70	70—90	90—110	110—130	130—150
No. of workers in firm A:	20	15	20	25	20
No. of workers in firm B:	150	100	90	110	50

Solution:

Wages (Rs.)	Mid-values (Rs.)	Cumulative Wages	Cumulative %	Firm A			Firm B		
				No. of workers	Cumulative total	Cumulative %	No. of workers	Cumulative total	Cumulative %
50—70	60	60	12	20	20	20	150	150	30
70—90	80	140	28	15	35	35	100	250	50
90—110	100	240	48	20	55	55	90	340	68
110—130	120	360	72	25	80	80	110	450	90
130—150	140	500	100	20	100	100	50	500	100

Note: The percentages are approximated to the nearest whole numbers.



It is obvious from the above figure that inequalities in the distribution of wages are more in Firm B than in Firm A.

EXERCISE 6.10

1. The following table shows number of firms in two different areas according to their annual profits. Present the data by way of Lorenz Curve.

Profit ('000 rupees):	6	25	60	84	105	150	170	400
Firms in Area A:	6	11	13	14	15	17	10	14
Firms in Area B:	2	38	52	28	38	26	12	4

2. The distribution of 9,400 Indian families according to income size is given below. Show inequality in the distribution of income by using Lorenz Curve.

Income:	0—1000	1000—5000	5000—10000	10000—20000	20000—40000
Families:	1,348	4,210	1,892	1,460	490

[**Hint:** Find out mid-value of class intervals]

3. Use Lorenz curve to compare the extant of inequalities in income distribution in two groups:

Monthly Income (Rs.):	1200—1400	1400—1600	1600—1800	1800—2000	2000—2200	2200—2400
No. of Persons in Gr.A:	800	960	1040	600	480	120
No. of Persons in Gr.B:	4800	6400	9600	3600	8000	4000

[Ans. Inequalities are more in Gr.A than Gr.B]

MISCELLANEOUS SOLVED EXAMPLES

Example 42. Calculate an appropriate measure of dispersion for the following data:

Wages per weak (in Rs.)	No. of wage earners
less than 35	14
35—37	62
38—40	99
41—43	18
over 43	7

Solution: The given distribution consists of open ended classes. One is less than 35 and the other is 'over 43'. The mid-values of these classes cannot be determined. Therefore, the appropriate measure of dispersion is Q.D. and coefficient of Q.D.

Calculation of Quartile Deviation

Wages (Rs.)	f	c.f.
Less 35	14	14
35—37	62	76
38—40	99	175
41—43	18	193
Over 43	7	200
	N = 200	

Location of Q_1

Size of Q_1 item $= \frac{N}{4} = \frac{200}{4}$, *i.e.*, 50th item which lies in 35—37 group which after adjustments becomes 34.5 or 37.5.

Now
$$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i$$
$$= 34.5 + \frac{50-14}{62} \times 2$$
$$= 34.5 + \frac{36}{62} \times 2 = 34.5 + \frac{72}{62} = 35.66$$

$\therefore \quad Q_1 = 35.66$

Location of Q_3

Size of Q_3 item $= \frac{3N}{4} = \frac{3 \times 200}{4} = 150$th item which lies in 38—40 group which after adjustment becomes 37.5 – 40.5.

Now,
$$Q_3 = l_1 + \frac{\frac{3N}{4} - c.f.}{f} \times i$$
$$= 37.5 + \frac{150-76}{99} \times 3 = 37.5 + \frac{74}{33} = 37.5 + 2.24 = 39.74$$

Calculation of Q.D. and its coefficient

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{39.74 - 35.66}{2} = \frac{4.08}{2} = 2.04$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{39.74 - 35.66}{39.74 + 35.66} = \frac{4.08}{75.4} = 0.54 \text{ approx.}$$

Example 43. The following is the record number of bricks laid each day for 10 days by two brick-layers A and B. Calculate the coefficient of variation in each case and discuss the relative consistency of the two brick-layers.

A:	700	675	725	625	650	700	650	700	600	650
B:	550	600	575	550	650	600	550	525	625	600

If the figures for A were in every case 10 more and those of B in every case 20 more than figure given above, how would the answer be affected?

Solution: (*i*) **Calculation for Mean and Standard Deviation**

Brick-layer A			Brick-layer B		
X	$dx' = \frac{X-700}{25}$	dx'^2	Y	$dy' = \frac{Y-625}{25}$	dy'^2
700	0	0	550	–3	9
675	–1	1	600	–1	1
725	1	1	575	–2	4
625	–3	9	550	–3	9
650	–2	4	650	1	1
700 = A	0	0	600	–1	1
650	–2	4	550	–3	9
700	0	0	525	–4	16
600	–4	16	625 = A	0	0
650	–2	4	600	–1	1
Total	–13	39		–17	51

Brick-layer A:

$$\overline{X} = A + \frac{\Sigma dx'}{N} \times i$$
$$= 700 - \frac{13}{10} \times 25 = 700 - 32.5 = 667.5 \text{ bricks per day}$$

$$\sigma_X = \sqrt{\frac{\Sigma dx'^2}{N} - \left(\frac{\Sigma dx'}{N}\right)^2} \times i$$

$$= \sqrt{\frac{39}{10} - \left(\frac{-13}{10}\right)^2} \times 25 = \sqrt{3.9 - 1.69} \times 25$$

$$= \sqrt{2.21} \times 25 = 1.486 \times 25 = 37.15$$

$$\text{C.V. (A)} = \frac{\sigma_X}{\overline{X}} \times 100 = \frac{37.15}{667.5} \times 100 = 5.56\%$$

Brick-layer B:

$$\overline{Y} = 625 - \frac{17}{10} \times 25 = 582.5 \text{ bricks per day}$$

$$\sigma_Y = \left[\sqrt{\frac{51}{10} - \left(\frac{-17}{10}\right)^2}\right] \times 25$$

$$= \sqrt{2.21} \times 25 = 37.15$$

$$\text{C.V. (B)} = \frac{37.15}{582.5} \times 100 = 6.38\%$$

As the coefficient of variation for brick-layer A is less than that of brick-layer B, brick-layer A is more consistent.

(*ii*) If the figures for A in every case were 10 more and that of B were 20 more, the arithmetic mean of A will increase by 10 and that of B by 20 but the standard deviations of both of them will remain unchanged.

[∵ S.D. is independent of the change of origin]

∴ A.M. of A will be 667.5 + 10 = 677.5 bricks per day

and A.M. of B will be 582.5 + 20 = 602.5 bricks per day

∴ Coefficient of variation of A will be $= \frac{37.15}{677.5} \times 100 = 5.48\%$

and coefficient of variation of B will be $= \frac{37.15}{602.5} \times 100 = 6.16\%$

After the change also brick-layer A remains more consistent than brick-layer B.

Example 44. Calculate arithmetic mean, median, mode and standard deviation for the following series:

Daily wages (Rs):	0—34.5	0—44.5	0—54.5	0—64.5	0—74.5	0—84.5
No. of workers:	4	24	62	86	96	100

Solution: The given data is in cumulative frequency form. It should first be converted into simple frequency data.

Calculation of $\overline{X}$, M, Z and σ

Daily wages	f	M.V. (m)	d	$d' = \frac{d}{10}$	fd'	fd'^2	$c.f.$
24.5—34.5	4	29.5	–20	–2	–8	16	4
34.5—44.5	20	39.5	–10	–1	–20	20	24
44.5—54.5	38	49.5 = A	0	0	0	0	62
54.5—64.5	24	59.5	+10	+1	+24	24	86
64.5—74.5	10	69.5	+20	+2	+20	40	96
74.5—84.5	4	79.5	+30	+3	+12	36	100
	$N = 100$				$\Sigma fd' = 28$	$\Sigma fd'^2 = 136$	

Arithmetic Mean

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 49.5 + \frac{28}{100} \times 10 = 52.3$$

Median: Size of median item $= \frac{N}{2} = \frac{100}{2} = 50$th item which lies in the class 44.5—54.5.

Now,
$$M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i$$

$$M = 44.5 + \frac{50 - 24}{38} \times 10 = 44.5 + \frac{26}{38} \times 10 = 44.5 + 6.84 = 51.34$$

Mode: By inspection, mode lies in the class 44.5 – 54.5

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$l_1 = 44.5$, $\Delta_1 = 38 - 20 = 18$, $\Delta_2 = 38 - 24 = 14$, $i = 10$

∴
$$Z = 44.5 + \frac{18}{18 + 14} \times 10 = 44.5 + \frac{18}{32} \times 10$$

$$= 44.5 + \frac{180}{32} = 44.5 + 5.625 = 50.125$$

S.D.:

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$$

$$= \sqrt{\frac{136}{100} - \left(\frac{28}{100}\right)^2} \times 10$$

$$= \sqrt{1.36 - 0.0784} \times 10 = \sqrt{1.2816} \times 10 = 11.32$$

Example 45. For two groups of observations the following results were available:

Group I	Group II
$\Sigma(X-5)=8$	$\Sigma(X-8)=-10$
$\Sigma(X-5)^2=40$	$\Sigma(X-8)^2=70$
$N_1=20$	$N_2=25$

Find the mean and standard deviation of both the groups taken together.

Solution: **Group I:**

Let $\Sigma d_1=\Sigma(X-5)=8$

$\Sigma d_1^2=\Sigma(X-5)^2=40$

$$\bar{X}_1=A+\frac{\Sigma d_1}{N}=5+\frac{8}{20}=5.40$$

$$\sigma_1=\sqrt{\frac{\Sigma d_1^2}{N}-\left(\frac{\Sigma d_1}{N}\right)^2}=\sqrt{\frac{40}{20}-\left(\frac{8}{20}\right)^2}=1.36$$

Group II:

Let $\Sigma d_2=\Sigma(X-8)=-10$

$\Sigma d_2^2=\Sigma(X-8)^2=70$

$$\bar{X}_2=A+\frac{\Sigma d_2}{N}=8+\frac{(-10)}{25}=7.6$$

$$\sigma_2=\sqrt{\frac{\Sigma d_2^2}{N}-\left(\frac{\Sigma d_2}{N}\right)^2}=\sqrt{\frac{70}{25}-\left(\frac{-10}{25}\right)^2}=1.62$$

$$\text{Combined Mean}(\bar{X}_{12})=\frac{N_1\bar{X}_1+N_2\bar{X}_2}{N_1+N_2}=\frac{20\times5.40+25\times7.6}{20+25}=6.62$$

$$\text{Combined S.D. }(\sigma_{12})=\sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_1d_1^2+N_2d_2^2}{N_1+N_2}}$$

$$d_1=\bar{X}_1-\bar{X}_{12}=5.40-6.62=-1.22$$

$$d_2=\bar{X}_2-\bar{X}_{12}=7.6-6.62=+0.98$$

$$\sigma_{12}=\sqrt{\frac{20(1.36)^2+25(1.62)^2+20(-1.22)^2+25(0.98)^2}{20+25}}$$

$$=\sqrt{3.476}=1.864$$

Example 46. **"After settlement the average weekly wage in a factory had increased from Rs. 8,000 to Rs. 12,000 and the standard deviation had increased from Rs. 100 to Rs. 150. After settlement the wage has become higher and more uniform." Do you agree?**

Solution: C.V. before settlement $=\frac{100}{8000}\times100=1.25\%$

C.V. after settlement $=\frac{150}{12000}\times100=1.25\%$

Since, there is no change in C.V., there is no improvement in uniformity.

Example 47. **Construct a continuous frequency distribution with class interval of 20 for the following table showing weight (in grams) of 50 applies:**

110	103	89	75	98	121	110	108	93	128
185	123	113	92	86	70	126	78	139	120
129	119	105	120	100	116	85	99	114	185
205	111	141	136	123	90	115	128	160	78
90	107	81	137	125	184	104	100	87	115

Also calculate the coefficient of variation of this distribution.

Solution: The lowest value is 70 and highest is 205. We have to take a class interval of 20. The various classes will be 70—90, 90—110, and so on up to 190—210.

Weight (in grams)	Tally Bars	Frequency
70—90	卌 卌 |	11
90—110	卌 卌 |	11
110—130	卌 卌 卌 ||||	19
130—170	||||	4
150—170	|	1
170—190	|||	3
190—210	|	1
		$N=50$

Calculation of Coefficient of Variation

Weight	f	M.V.	d	d'	fd'	fd'^2
70—90	11	80	− 60	− 3	− 33	99
90—110	11	100	− 40	− 2	− 22	44
110—130	19	120	− 20	− 1	− 19	19
130—170	4	140 = A	0	0	0	0
150—170	1	160	+20	+1	1	1
170—190	3	180	+40	+2	6	12
190—210	1	200	+60	+3	3	9
	$N=50$				$\Sigma fd'=-64$	$\Sigma fd'^2=184$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 140 - \frac{64}{50} \times 20 = 114.4$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{184}{50} - \left(\frac{-64}{50}\right)^2} \times 20$$

$$= \sqrt{368 - 1.6384} \times 20 = 28.5769$$

$$\text{Coefficient of variation} = \frac{\sigma}{\overline{X}} \times 100 = \frac{28.5769 \times 100}{114.4} = 24.97\%$$

Example 48. The coefficient of variation of a series is 58%. The standard deviation is 21.2. What is the arithmetic mean?

Solution:

$$\text{C.V.} = \frac{\sigma}{\overline{X}} \times 100$$

$$\Rightarrow \quad \overline{X} = \frac{\sigma}{C.V.} \times 100$$

$$\text{Mean } (\overline{X}) = \frac{21.2 \times 100}{58} = 36.6$$

Example 49. During nine days in a festival the highest sale of a shop was on Sunday and was Rs. 90 more than the average sale for other days. If the standard deviation of the sale during the festival is 33.33, find the standard deviation leaving that the highest sale.

Solution:

$$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}} \quad \text{(Formula of S.D.)}$$

$$33.33 = \sqrt{\frac{\Sigma(X - \overline{X})^2}{9}}$$

$$\Rightarrow \quad \Sigma x^2 = \frac{100}{3} \times \frac{100}{3} \times 9 = 10{,}000 \quad [\because x = X - \overline{X}]$$

Sunday value $= \overline{X} + 90$

$$x = X - \overline{X}$$

$$= (\overline{X} + 90) - \overline{X} = 90$$

Sum of the square of deviations excluding Sunday.

$$= 10000 - (90)^2 = 1900$$

$$\sigma \text{ for 8 days} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{8}}$$

$$= \sqrt{\frac{1900}{8}} = \sqrt{237.5} = 15.4$$

Example 50: If the mean and standard deviation of 75 observations is 40 and 8 respectively, find the new mean and standard deviation if

(*i*) Each observation is multiplied by 5.

(*ii*) 7 is added to each observation.

Solution: (*i*) New $\overline{X} = 200$, New $\sigma = 40$

The reason is that the change of scale affects the value of both $\overline{X}$ and σ.

(*ii*) New $\overline{X} = 47$, New $\sigma = 8$

The reason is that the mean is affected by change of origin but S.D. is not affected by change of origin.

Example 51. 5 observations of a series are 4, 6, 8, 12 and 15. Their mean and standard deviation are 9 and 4 respectively. Make such alternations in the terms of the series that new S.D. is 20 and new mean is 50.

Solution: Given $\overline{X} = 9, \sigma = 4$

The series with new S.D. is obtained when each observations is multiplied by 5. This operation will also increase the value of mean 5 times. In the given example, if each observation is multiplied by 5, the mean becomes 45 and S.D. becomes 20. The reason is that change of scale affects both mean and standard deviation.

Now, if 5 is added to each observation the mean becomes 50 while S.D. remains 20. The reason is that the change of origin affects only $\overline{X}$ and not σ. The transformation series corresponds to this can be written as $U = 5X + 5$. Thus, the changed observations will be: 25, 35, 45, 65 and 80.

Example 52. A collar manufacturer is considering the production of a new style of collar to attract youngmen. The following statistics of neck circumference are available based on measurements of a typical group of college students. Compute the SD and use the criterion $(\overline{X} \pm 3\sigma)$, to determine the largest and smallest sizes of collars, he should make in order to meet the needs of practically all his customers, bearing in mind, that collars are worn, on average $\frac{3}{4}$ inch larger than neck size.

Mid-points:	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
f:	4	19	30	63	66	29	18	1	1

Solution: Here, $i = 0.5$

Let, $A = 14.5 \quad \therefore d' = \left(\frac{X - A}{i}\right)$

Mid-points (X)	*f*	*d'*	d'^2	*fd'*	fd'^2
12.5	4	–4	16	–16	64
13.0	19	–3	9	–57	171
13.5	30	–2	4	–60	120
14.0	63	–1	1	–63	63
14.5	66	0	0	0	0
15.0	29	+1	1	29	29
15.5	18	+2	4	36	72
16.0	1	+3	9	3	9
16.5	1	+4	16	4	16
	$N = 231$			$\Sigma fd' = -124$	$\Sigma fd'^2 = 544$

Now, $\overline{X} = A + \frac{\Sigma fd'}{N} \times i$

$$= 14.5 + \frac{-124}{231} \times 0.5 = 14.5 - 0.268 = 14.232 \text{ inches}$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{544}{231} - \left(\frac{-124}{231}\right)^2} \times 0.5$$

$$= \sqrt{2.355 - 0.288} \times 0.5 = \sqrt{2.067} \times 0.5$$

$$= 1.437 \times 0.5 = 0.719 \text{ inches.}$$

Using the criterion $\overline{X} \pm 3\sigma$

Largest and smallest neck size

$$= \overline{X} \pm 3\sigma = 14.232 \pm 3(0.719)$$

$$= 14.232 \pm 2.157 = 12.075 \text{ and } 16.389$$

Since, the collars are worn on an average $\frac{3}{4}$ inch longer than the neck size, we should add 0.75 to these limits. Thus, the smallest and largest sizes of the collar should be: (12.075 + 0.75) and (16.389 + 0.75) = 12.825 and 17.139

Thus, the smallest size of the collar should be 12.825 inches long and largest 17.139 inches long.

Example 53. The mean age and standard deviation of a group of 200 persons (grouped in intervals 0—5, 5—10, ..., etc.) were found to be 40 and 15. Later on it was discovered that the age 43 was misread as 53. Find the correct mean and standard deviation.

Solution: $N = 200$, $\overline{X} = 40$ and $\sigma = 15$

As $\overline{X} = \frac{\Sigma fm}{N} \Rightarrow \Sigma fm = N\overline{X}$

Incorrected $\Sigma fm = N\overline{X} = 200 \times 40 = 8000$

Corrected age is 43 which falls in the group 40—45, the mid-value of which is 42.5.

Incorrected age is 53 which falls in the group 50—55, the mid-value of which is 52.5.

$$\text{Corrected } \overline{X} = \frac{8000 + 42.5 - 52.5}{200} = \frac{7990}{200} = 39.95$$

$$\text{Incorrected } \Sigma fm^2 = N(\sigma^2 + \overline{X}^2) = 200\,(15^2 + 40^2) = 3,65,000$$

$$\text{Corrected } \Sigma fm^2 = 3,65,000 - (52.5)^2 + (42.5)^2$$

$$= 3,65,000 - 2756.25 + 1806.25 = 3,64,050$$

$$\text{Corrected } \sigma = \sqrt{\frac{3,64,050}{200} - (39.95)^2}$$

$$= \sqrt{1820.25 - 1596.0025} = 14.97$$

Example 54. The monthly wages (in Rs.) of 100 workers are distributed as follows:

Wages (Rs.):	0—100	100—200	200—300	300—400	400—500	500—600
No. of workers:	12	x	27	y	17	6

If model wage is Rs. 256.25, find the missing frequencies and hence find % variation in the distribution.

Solution: Let the missing frequencies be x and y

Wages (Rs.)	f	c.f.
0—100	12	12
100—200	x f_0	$12 + x$
200—300	27 f_1	$39 + x$
300—400	y f_2	$39 + x + y$
400—500	17	$56 + x + y$
500—600	6	$62 + x + y$
	N = 100	

$$62 + x + y = 100$$

$$\Rightarrow \quad x + y = 100 - 62 = 38$$

As $Z = 256.25$, Mode lies in 200–300

Applying the formula,

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \quad \text{where, } \Delta_1 = |f_1 - f_0|;\ \Delta_2 = |f_2 - f_1|$$

$$256.25 = 200 + \frac{27 - x}{(27 - x) + (27 - y)} \times 100$$

$$256.25 - 200 = \frac{2700 - 100x}{27 - x + 27 - (38 - x)} \quad [\because y = 38 - x]$$

$$\Rightarrow \quad 56.25 = \frac{2700 - 100x}{27 - x + 27 - 38 + x}$$

$$\Rightarrow \quad 56.25 = \frac{2700 - 100x}{16}$$

$$\Rightarrow \quad 56.25 \times 16 = 2700 - 100x$$

$$\Rightarrow \quad 900 - 2700 = -100x$$

$$\Rightarrow \quad -1800 = -100x$$

$$\therefore \quad x = \frac{1800}{100} = 18$$

Now, $x + y = 38 \Rightarrow y = 38 - x = 38 - 18 = 20$

$\therefore$ $x = 18$ and $y = 20$

Now, in order to find out % variation in the distribution, we have to find out the co-efficient of variation.

Wages (Rs.)	f	M.V. (m)	d	d'	fd'	fd'^2
0—100	12	50	–200	–2	–24	48
100—200	18	150	–100	–1	–18	18
200—300	27	250 = A	0	0	0	0
300—400	20	350	100	1	20	20
400—500	17	450	200	2	34	68
500—600	6	550	300	3	18	54
	$N = 100$				$\Sigma fd' = 30$	$\Sigma fd'^2 = 208$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 250 + \frac{30}{100} \times 100 = 280$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{208}{100} - \left(\frac{30}{100}\right)^2} \times 100$$

$$= \sqrt{2.08 - 0.09} \times 100 = \sqrt{1.99} \times 100$$

$$= 1.4107 \times 100 = 141.07$$

$$C.V. = \frac{\sigma}{\overline{X}} \times 100 = \frac{141.07}{280} \times 100 = 50.38\%$$

Example 55. The following table shows the marks obtained by three students A, B and C in an examination:

Student ↓ Subject →	Maximum marks in each subject				
	800 S_1	700 S_2	900 S_3	600 S_4	1000 S_5
A	560	553	549	540	500
B	480	420	540	360	600
C	424	427	423	426	420

Determine which student has shown (*i*) most consistent performance and (*ii*) most inconsistent performance.

Solution: The given data has been shown below:

Student ↓ Subject →	Percentage marks					Average % marks	Standard deviation	C.V.
	S_1	S_2	S_3	S_4	S_5			
A	70	79	61	90	50	70	13.87	19.81
B	60	60	60	60	60	60	0	0
C	53	61	47	71	42	54.8	10.32	18.83

$$\text{A's average marks} = \frac{70+79+61+90+50}{5} = \frac{350}{5} = 70$$

$$\text{B's average marks} = \frac{60+60+60+60+60}{5} = \frac{300}{5} = 60$$

$$\text{C's average marks} = \frac{53+61+47+71+42}{5} = \frac{274}{5} = 54.8$$

A's standard deviation (σ_A)

$$= \sqrt{\frac{0+(9)^2+(-9)^2+(20)^2+(-20)^2}{5}} = \sqrt{\frac{962}{5}} = \sqrt{192.4} = 13.87$$

$$\text{B's standard deviation } (\sigma_B) = \sqrt{\frac{0}{5}} = 0$$

$$\text{C's standard deviation } (\sigma_C) = \sqrt{\frac{(1.8)^2+(6.2)^2+(-7.8)^2+(16.3)^2+(-12.8)^2}{5}}$$

$$= \sqrt{\frac{530.05}{5}} = \sqrt{106.41} = 10.32$$

$$\text{A's coefficient of variation (C.V.)} = \frac{\sigma}{\overline{X}} \times 100 = \frac{13.87}{70} \times 100 = 19.81\%$$

$$\text{B's coefficient of variation (C.V.)} = \frac{0}{60} \times 100 = 0\%$$

$$\text{C's coefficient of variation (C.V.)} = \frac{10.32}{54.8} \times 100 = 18.83\%$$

Hence, it is clear from the above data that

(*i*) B is most consistent as in his case, the coefficient of variation is the least.

(*ii*) A is most inconsistent as in his case, the coefficient of variation is the greatest.

Example 56. The salaries paid to the managers of a company had a mean of Rs. 20,000 with a standard deviation of Rs. 3,000. What will be the mean and standard deviation if all the salaries are increased by (*i*) 10%, (*ii*) 10% of the existing mean? Which policy would you recommend if the management does not want to have increased disparities in wages?

Solution: Increasing all the salaries by 10% ⇒ multiplying each salary each by 1.1. Hence, the mean is also multiplied by 1.1. Since, S.D. defends on change of scale, S.D. is also multiplied by 1.1.

∴ When all salaries are increased by 10%, the S.D. also increases by 10%.

If each salary is increased by 10%, the mean is also increased by 10%.

Increasing all the salaries by 10% of the existing mean ⇒ Adding a constant amount to each salary.

Since S.D. is independent of the change of origin, it will remain unchanged if each salary is increased by 10% of the mean. But mean will increase by 10% of the original mean.

If the management do not want to have increased disparities, it showed increase in the salary of the workers by 10% of the existing mean.

ample 57. If 10 is subtracted from every item in a data set then the coefficient of variation of the resulting set of data is 20%. If 20 is added to every item of the same data set then the coefficient of variation of the resulting set of data is 10%.

You are required to find out the coefficient of variation of the original set of data.

lution: Let the average and standard deviation of the original data set be $\overline{X}$ and s.

Average of all items '$X - 10$' $= \frac{\Sigma(X-10)}{N} = \frac{\Sigma X}{N} - \frac{N10}{N} = \overline{X} - 10$

Standard deviation of all items '$N - 10$' $= s$'

(This is because the value of standard deviation remains the same if each observation in a series is increased or decreased by the same quantity)

Given : $\frac{s}{\overline{X}-10} \times 100 = 20$

$\Rightarrow \quad 100s = 20\overline{X} - 200$

$\Rightarrow \quad 20\overline{X} - 100s = 200$ ...(*i*)

Average of all items '$X + 20$' $= \frac{\Sigma(X+20)}{N} = \frac{\Sigma X}{N} + \frac{N20}{N} = \overline{X} + 20$

Standard deviation of all items '$X + 20$' $= s$

(This is because the value of standard deviation remains the same if each observation in a series is increased or decreased by the same quantity)

Given : $\frac{s}{\overline{X}+20} \times 100 = 10$

$\Rightarrow \quad 100s = 10\overline{X} + 200$

$\Rightarrow \quad 10\overline{X} - 100s = 200$ ...(*ii*)

Subtracting equation (*ii*) from equation (*i*), we get

$(20\overline{X} - 100s) - (10\overline{X} - 100s) = 200 - (-200)$

$\Rightarrow \quad 10\overline{X} = 400$

$\Rightarrow \quad \overline{X} = \frac{400}{10} = 40$

Putting the value of $\overline{X}$ in equation (*ii*), we get

$10(40) - 100s = -200$

$\Rightarrow \quad 400 - 100s = -200$

$\Rightarrow \quad 100s = 600$

$\Rightarrow \quad s = \frac{600}{100} = 6$

$\therefore$ Coefficient of variation of the original data set:

$= \frac{s}{\overline{X}} \times 100 = \frac{6}{40} \times 100 = 15\%$

Example 58. A fund manager is considering investment in the equity shares of one of two companies. The criterion for selecting the company for investment is consistency of return on net worth. The following data have been collected:

Financial Year	Return on Net worth (%)	
	Modern Industries Ltd. (MIL)	Pioneer Industries Ltd. (PIL)
2001—2002	19	20
2000—2001	20	24
1999—2000	16	16
1998—1999	13	15
1997—1998	12	10

You are required to identify the company in which the fund manager should invest.

Solution: **Modern Industries Ltd. (MIL):**

Mean return on equity $(\overline{X}) = \frac{\Sigma X}{N} = \frac{19+20+16+13+12}{5} = \frac{80}{5} = 16\%$

Standard deviation (S.D.) $= \left[\frac{\Sigma(X-\overline{X})^2}{N}\right]^{\frac{1}{2}}$

$= \left[\frac{(19-16)^2+(20-16)^2+(16-16)^2+(13-16)^2+(12-16)^2}{5}\right]^{\frac{1}{2}}$

$= 3.16\%$ (approx.)

Coefficient of variation $= \frac{\text{S.D.}}{\overline{X}} \times 100 = \frac{3.16}{16} \times 100 = 19.76\%$ (approx.)

Pioneer Industries Ltd. (PIL):

Mean return on equity $(\overline{X}) = \frac{\Sigma X}{N} = \frac{20+24+16+15+10}{5} = \frac{85}{5} = 17\%$

Standard deviation (S.D.) $= \left[\frac{\Sigma(X-\overline{X})^2}{N}\right]^{\frac{1}{2}}$

$= \left[\frac{(20-17)^2+(24-17)^2+(16-17)^2+(15-17)^2+(10-17)^2}{5}\right]^{\frac{1}{2}}$

$= 4.73\%$ (approx.)

Coefficient of variation $= \frac{\text{S.D.}}{\overline{X}} \times 100 = \frac{4.73}{17} \times 100 = 27.83\%$ (approx.)

Since, the coefficient of variation for MIL is less than coefficient of variation for PIL, it can be inferred that profitability of MIL is more consistent than PIL. Hence, investment should be made in MIL.

IMPORTANT FORMULAE

1. Range

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

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

2. Quartile Deviation

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

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

3. Mean Deviation

For Individual Series:

$$\text{M.D.} = \frac{\Sigma\,|D|}{N}$$

For Discrete and Continuous Series:

$$\text{M.D.} = \frac{\Sigma f|D|}{N}$$

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

4. Standard Deviation

For Individual Series:

(i) $\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$: Actual Mean Method

(ii) $\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$: Assumed Mean Method

(iii) $\sigma = \sqrt{\frac{\Sigma d'^2}{N} - \left(\frac{\Sigma d'}{N}\right)^2} \times i$: Step Deviation Method

For Discrete / Continuous Series:

(i) $\sigma = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$: Actual Mean Method

(ii) $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$: Assumed Mean Method

(iii) $\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$: Step Deviation Method

5. Combined Standard Deviation

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

Where, $d_1 = \bar{X}_1 - \bar{X}_{12}$ and $d_2 = \bar{X}_2 - \bar{X}_{12}$

6. Coefficient of Variation

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

7. Variance

$$\text{Variance} = (S.D.)^2$$

or $\sigma = \sqrt{\text{variance}}$

QUESTIONS

1. What is meant by dispersion? What purpose does a measure of dispersion serve? .
2. What are the various measures of dispersion. Explain the relative merits and demerits of each.
3. (i) What are the properties of a good measure of variation?
   (ii) Why is standard deviation considered a better measure of dispersion.
4. What is coefficient of variation? What purpose does it serve? Also distinguish between variance and coefficient of variation.
5. Define range, interquartile range, quartile deviation, mean deviation and standard deviation. Describe their merits and demerits.
6. Define dispersion. Discuss the merits and demerits of different measures of dispersion.
7. What do you understand by standard deviation? Explain its important properties.
8. Explain the method of measuring inequalities of income by using Lorenz curve.
9. What do you understand by Lornez curve? Discuss the usefulness of Lornez curve.
10. Why is S.D. (σ) the most widely used measure of dispersion? Explain.
11. If a constant is subtracted from each score in a series, what will be its effect on $\bar{X}$ and σ?

7 Measures of Skewness

INTRODUCTION

In the preceeding two chapters, we have discussed the measures of central tendency and dispersion of frequency distributions for their summarisation and comparison with each other. These measures, however, do not adequately describe a frequency distribution in the sense that there could be two or more distributions with the same mean and standard deviation but still different from each other with regard to shape or pattern of distribution. This implies that there is need to develop some measures to further describe the distribution. These measures are known as measures of skewness.

MEANING OF SKEWNESS

The term skewness means lack of symmetry in a frequency distribution. Skewness denotes the degree of departure of a distribution from symmetry and reveals the direction of scatterness of the items. It gives us an idea about the shape of the frequency curve. When a distribution is not symmetrical, it is called a skewed distribution. Skewness tells us about the asymmetry of the frequency distribution.

DEFINITION OF SKEWNESS

Some important definitions of skewness are given below:

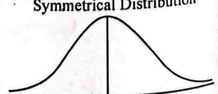
1. Skewness is the degree of asymmetry or departure from symmetry of a distribution.
   —**M.R. Speigal**
2. When a series is not symmetrical, it is said to be asymmetrical or skewed.
   —**Croxten and Cowden**
3. By skewness of a frequency distribution, we mean degree of its departure from symmetry.
   —**Simpson and Kafka**

SKEWNESS AND FREQUENCY DISTRIBUTION

The concept of skewness will be made more clear from the following diagrams showing a symmetrical distribution, a positively skewed distribution and a negatively skewed distribution.

(1) Symmetrical Distribution: In a symmetrical distribution or symmetrical curve, skewness is not present. The values of mean, median and mode coincide, *i.e.*, $\overline{X} = M = Z$. The spread of the frequencies is the same on both sides of the central point of the curve.

Symmetrical Distribution

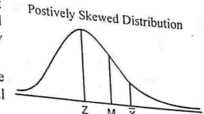


(2) Skewed Distribution: A distribution which is not symmetrical is called skewed distribution or asymmetrical distribution. A skewed distribution may be either positively skewed or negatively skewed.

***(a)* Postively Skewed Distribution:** If the longer tail of the frequency curve of distribution lies to the right of the central point, it is called a positively skewed distribution.

In the positively skewed distribution, the value of the mean will be greater than median and median be greater than mode, *i.e.*, $\overline{X} > M > Z$.

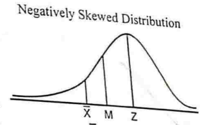
Postively Skewed Distribution



***(b)* Negatively Skewed Distribution:** If the longer tail of the frequency curve of the distribution lies to the left of the central point, it is called a negatively skewed distribution.

In the negatively skewed distribution, the value of the mean will be less then median and median be less than mode, *i.e.*, $\overline{X} < M < Z$.

Negatively Skewed Distribution



DIFFERENCE BETWEEN DISPERSION AND SKEWNESS

The main points of difference between dispersion and skewness are given as under:

(1) Dispersion is concerned with measuring the amount of variation in a series rather than with its direction. Skewness is concerned with direction of variation or the departure from symmetry.

(2) Dispersion tells us about the composition of the series whereas skewness tells us about the shape of the series.

(3) Measures of dispersion are based on averages of the first order such as $\overline{X}, M, Z$, etc., whereas measures of skewness are based on averages of first and second order such as $\overline{X}, M, Z, \sigma$, etc.

TESTS OF SKEWNESS

In order to find out whether a distribution is skewed or not, the following tests may be applied:

(1) Relationship between Averages: If in a distribution, the values of mean, median and mode are equal, *i.e.*, $\overline{X} = M = Z$, then skewness is absent in it. On the other hand, if the values of mean, median and mode are not identical, *i.e.*, $\overline{X} \neq M \neq Z$, then skewness is found present in the distribution.

(2) Distance of Quartiles from the Median: If in a distribution, the quartiles (Q_1 and Q_3) are equidistant from the median, *i.e.*, $Q_3 - M = M - Q_1$, then skewness is absent and if $Q_3 - M \neq M - Q_1$, then skewness is present in the distribution.

(3) Graph of the Data: When the data plotted on the graph paper gives us a bell shaped curve, skewness is absent. On the other hand, when the data plotted on the graph do not give the normal bell shaped curve, *i.e.*, the two values of the curve are not equal, then skewness is present in the distribution.

◘ MEASURES OF SKEWNESS

Measures of skewness help us to find out the direction and extent of asymmetry in a series. They may either be absolute or relative. The measures which expresses skewness in the units in which the values of the series are expressed are called **absolute measures of skewness.** The measures which expresses skewness in the form of ratios or percentage are called **relative measures of skewness.** Relative measures of skewness, also called **coefficient of skewness** are useful to compare the skewness of two or more series.

There are three important methods of measuring skewness, namely

(1) Karl Pearson's Method

(2) Bowley's Method

(3) Kelly's Method

◘ (1) Karl Pearson's Method

Karl Pearson's method is based on arithmetic mean ($\overline{X}$), mode (Z), median (M) and standard deviation (σ). Karl Pearson has given the following formulae for measuring skewness:

Absolute Measure of Skewness	Coefficient of Skewness
$S_K = \overline{X} - Z$	Coefficient of $S_K = \dfrac{\overline{X} - Z}{\sigma}$
When mode (Z) is ill defined, then $S_K = 3(\overline{X} - M)$	When mode (Z) is ill defined, then Coefficient of $S_K = \dfrac{3(\overline{X} - M)}{\sigma}$

The value of Karl Pearson's coefficient of skewness usually lies between ± 1. In case mode is ill defined, the value of coefficient of skewness lies between ± 3.

○ Steps for Calculation

(1) Calculate mean ($\overline{X}$) of the distribution.

(2) Calculate mode (Z) of the distribution.

(3) Calculate median (M) of the distribution.

(4) Calculate standard deviation (σ).

(5) Put these values in the formulae.

The following examples illustrate the procedure of calculating Pearson's coefficient of skewness:

○ Calculation of Coefficient of Skewness—Discrete Series

Example 1. From the following data, find out Karl Pearson's Coefficient of Skewness:

Height (in inches)	No. of persons
58	10
59	18
60	30
61	42
62	35
63	28

Solution:

Calculation of Karl Pearson's Coefficient of Skewness

Height (X)	f	$A = 60$ $(X - 60)$ d	fd	fd^2
58	10	−2	−20	40
59	18	−1	−18	18
60 A	30	0	0	0
61	42	+1	+42	42
62	35	+2	+70	140
63	28	+3	+84	252
	$N = 163$		$\Sigma fd = 158$	$\Sigma fd^2 = 492$

$$\overline{X} = A + \frac{\Sigma fd}{N} = 60 + \frac{158}{163} = 60 + 0.969 = 60.969$$

$$\sigma = \sqrt{\frac{fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{492}{163} - \left(\frac{158}{163}\right)^2}$$

$$= \sqrt{3.0184 - 0.9395} = \sqrt{2.0789} = 1.4418$$

By inspection, mode is 61 (as its frequency is maximum)

Thus, $\overline{X} = 60.969$, $\sigma = 1.4418$, $Z = 61$

$$\therefore \text{ Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{60.969 - 61}{1.4418}$$

$$= \frac{-0.031}{1.4418} = -0.0215$$

○ Calculation of Pearson's Coefficient of Skewness—Continuous Series

Example 2. Calculate Karl Pearson's Coefficient of Skewness from the following data:

Variable:	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
Frequency:	2	5	7	13	21	16	8	3

Solution:

Calculation of Karl Pearson's Coefficient of Skewness

Variable	f	M.V. (m)	d	$d' = \frac{d}{5}$	fd'	fd'^2
0 – 5	2	2.5	−20	−4	−8	32
5 – 10	5	7.5	−15	−3	−15	45
10 – 15	7	12.5	−10	−2	−14	28
15 – 20	13	17.5	−5	−1	−13	13
20 – 25	21	22.5 = A	0	0	0	0
25 – 30	16	27.5	+5	+1	16	16
30 – 35	8	32.5	+10	+2	16	32
35 – 40	3	37.5	+15	+3	9	27
	$N = 75$				$\Sigma fd' = -9$	$\Sigma fd'^2 = 193$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 22.5 + \frac{(-9)}{75} \times 5 = 22.5 - 0.6 = 21.9$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{fd'}{N}\right)^2} \times i = \sqrt{\frac{193}{75} - \left(\frac{-9}{75}\right)^2} \times 5$$

$$= \sqrt{2.5733 - 0.0144} \times 5 = \sqrt{2.5589} \times 5 = 1.599 \times 5 = 7.995$$

By inspection, modal class is 20—25

$$\therefore \quad Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 20 + \frac{21-13}{42-13-16} \times 5$$

$$= 20 + \frac{40}{13} = 20 + 3.08 = 23.08$$

Thus, $\overline{X} = 21.9$, $\sigma = 7.995$, $Z = 23.08$

$$\therefore \text{ Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{21.9 - 23.08}{7.995} = -0.15$$

Example 3. Calculate Karl Pearson's Coefficient of Skewness from the following data:

Wages:	300—400	400—500	500—600	600—700	700—800
No. of workers:	5	10	10	3	2

Solution: Since the given series is a bimodel series, the following formula for calculating skewness is used:

$$\text{Coeff. of } S_K = \frac{3(\overline{X} - M)}{\sigma}$$

Calculation of Coefficient of Skewness

Wages	f	M.V. (m)	d	d'	fd'	fd'^2	c.f.
300—400	5	350	−200	−2	−10	20	5
400—500	10	450	−100	−1	−10	10	15
500—600	10	550	0	0	0	0	25
600—700	3	650	+100	+1	3	3	28
700—800	2	750	+200	+2	4	8	30
	$N = 30$				$\Sigma fd' = -13$	$\Sigma fd'^2 = 41$	

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 550 - \frac{13}{30} \times 10$$

$$= 550 - 43.33 = 506.67$$

$$\text{Median} = \frac{N}{2} = \frac{30}{2} = 15\text{th item.}$$

Median lies in the class interval 400—500.

$$\therefore \quad M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 400 + \frac{15-5}{10} \times 100 = 400 + 100 = 500$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{41}{30} - \left(\frac{-13}{30}\right)^2} \times 100$$

$$= \sqrt{1.367 - 0.188} \times 100 = \sqrt{1.179} \times 100 = 1.086 \times 100 = 108.6$$

$\therefore \overline{X} = 506.67$, Median $= 500$, $\sigma = 108.6$

$$\text{Coefficient of } S_K = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(506.67 - 500)}{108.6} = \frac{20.01}{108.6} = +0.184$$

Example 4. Calculate the coefficient of skewness from the following data:

Mid-point:	15	20	25	30	35	40
Frequency:	12	18	25	24	20	21

Solution: As the mid-points of the different class intervals are given, we first find actual class intervals by using the formula $m \pm i/2$, where, m = mid-point and i = difference between two mid-points.

Calculation of Coefficient of Skewness

Classes	f	Mid-point m	d $(m-25)$	$d' = \frac{d}{5}$	fd'	fd'^2
12.5—17.5	12	15	−10	−10	−24	48
17.5—22.5	18	20	−5	−5	−18	18
22.5—27.5	25	25 = A	0	0	0	0
27.5—32.5	24	30	5	5	+24	24
32.5—37.5	20	35	10	+10	+40	80
37.5—42.5	21	40	15	+15	+63	189
	$N = 120$				$\Sigma fd' = 85$	$\Sigma fd'^2 = 359$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 25 + \frac{85}{120} \times 5 = 25 + 3.542 = 28.542$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{359}{120} - \left(\frac{85}{120}\right)^2} \times 5$$

$$= \sqrt{2.992 - 0.502} \times 5 = \sqrt{2.4902} \times 5 = 1.578 \times 5 = 7.89$$

Mode: The highest frequency is 25, mode lies corresponding to mid-point 25, *i.e.*, in the class 22.5—27.5.

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Here, $l_1 = 22{\cdot}5, \Delta_1 = 25 - 18 = 7, \Delta_2 = 25 - 24 = 1, i = 5$

$$Z = 22.5 + \frac{7}{7+1} \times 5 = 22.5 + 4.375 = 26.875$$

$$S_{K_P} = \frac{\overline{X} - Z}{\sigma} = \frac{28.542 - 26.875}{7.89} = \frac{1.667}{7.89} = 0.211$$

There is a low degree of positive skewness.

Example 5. Find the mean, mode, standard deviation and pearson's coefficient of skewness for the following data:

Year under:	10	20	30	40	50	60
No. of Persons:	15	32	51	78	97	109

Solution: Since it is a cumulative frequency series, it should first be converted into simple frequency series:

Years	f	M.V. (m)	d	d'	fd'	fd'^2
0—10	15	5	–30	–3	–45	135
10—20	17	15	–20	–2	–34	68
20—30	19	25	–10	–1	–19	19
30—40	27	35A	0	0	0	0
40—50	19	45	+10	+1	+19	19
50—60	12	55	+20	+2	+24	48
Total	$N = 109$				$\Sigma fd' = -55$	$\Sigma fd'^2 = 289$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 35 + \frac{(-55)}{109} \times 10 = 35 - 5.045 = 29.95$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{289}{109} - \left(\frac{-55}{109}\right)^2} \times 10$$

$$= \sqrt{2.6513 - 0.2546} \times 10 = \sqrt{2.3967} \times 10$$

$$= 1.548 \times 10 = 15.48$$

By inspection, modal class is 30 – 40

$$\therefore Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 30 + \frac{27-19}{54-19-19} \times 10 = 30 + \frac{8}{16} \times 10 = 35$$

$$\therefore \text{Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{2995 - 35}{15.48} = -0.32$$

Example 6. Calculate Pearson's coefficient of skewness from the following data:

Marks above:	10	20	30	40	50	60	70	80	90
No. of students:	100	97	90	70	40	25	15	8	3

Solution: Since it is a cumulative frequency series, it should first be converted into simple frequency series.

Marks	f	M.V. (m)	d	d'	fd'	fd'^2
10—20	3	15	–40	–4	–12	48
20—30	7	25	–30	–3	–21	63
30—40	20	35	–20	–2	–40	80
40—50	30	45	–10	–1	–30	30

50—60	15	55 = A	0	0	0	0
60—70	10	65	+10	+1	10	10
70—80	7	75	+20	+2	14	28
80—90	5	85	+30	+3	15	45
90—100	3	95	+40	+4	12	48
	$N = 100$				$\Sigma fd' = -52$	$\Sigma fd'^2 = 352$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 55 - \frac{52}{100} \times 10 = 55 - 5.2 = 49.8$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{352}{100} - \left(\frac{-52}{100}\right)^2} \times 10$$

$$= \sqrt{3.52 - 0.2704} \times 10 = \sqrt{3.2496} \times 10$$

$$= 1.8026 \times 10 = 18.02$$

By inspection, modal class is 40—50

$$\therefore Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 40 + \frac{30-20}{60-20-15} \times 10$$

$$= 40 + \frac{10 \times 10}{25} = 40 + 4 = 44$$

$$\therefore \text{Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{49.8 - 44}{18.02} = \frac{5.8}{18.02} = 0.32$$

Example 7. For a group of 20 items, $\Sigma X = 1452, \Sigma X^2 = 144280$ and mode = 63.7. Find Karl Pearson's co-efficient of Skewness.

Solution: Coefficient of skewness $= \frac{\overline{X} - Z}{\sigma}$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{1452}{20} = 72.6$$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (\overline{X})^2} = \sqrt{\frac{144280}{20} - (72.6)^2}$$

$$= \sqrt{7214 - 5270.76} = 44.082$$

$$\text{Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{72.6 - 63.7}{44.082} = 0.202$$

Example 8. In a certain distribution the following results were obtained:
C.V. = 40%, $\overline{X} = 25$, $Z = 20$
Find out coefficient of skewness.

Solution:

$$\text{C.V.} = \frac{\sigma}{\overline{X}} \times 100 \Rightarrow 40 = \frac{\sigma}{25} \times 100 \Rightarrow \sigma = 10$$

$$\text{Coefficient of Skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{25 - 20}{10} = \frac{5}{10} = \frac{1}{2} = +0.5$$

EXERCISE 7.1

1. Calculate skewness and its coefficient from the following data: (Use Pearson's Formula).

Wages (Rs.) :	10	11	12	13	14	15	16
No . of workers :	4	7	9	15	8	5	2

[Ans. $S_K = -2.22$, coeff. of $S_K = -1.457$]

2. Calculate Pearson's Coefficient of Skewness from the following data:

Profits (Rs. lakhs):	70—80	60—70	50—60	40—50	30—40	20—30	10—20	0—10
No. of company:	11	22	30	35	21	11	6	5

[Ans. $\overline{X} = 46.84$, $Z = 47.86$, $\sigma = 17.08$, Coeff. of $S_K = -0.0304$]

3. Calculate Pearson's Coefficient of Skewness from the following:

Marks above:	0	10	20	30	40	50	60	70	80
No. of students:	150	140	100	80	80	70	30	14	0

[Hint: See Example 19] [Ans. $\overline{X} = 39.27, \sigma = 22.8, M = 45$; Coeff. of $S_K = -0.75$]

4. Calculate Pearson coefficient of skewness based on mean, median and standard deviation from the following data:

Age Groups:	0—10	10—20	20—30	30—40	40—50	50—60	60—70	70 and above
No. of workers:	18	16	15	12	10	5	2	1

[Ans. $\overline{X} = 26$ app., $M = 23.67$, $\sigma = 17.46$, Coeff. of $S_K = 0.4$]

5. The daily expenditure of 100 families is given below:

Daily expenditure:	0—20	20—40	40-60	60-80	80-100
No. of families:	13	?	27	?	16

If the mode of the distribution is 44, calculate Karl Pearson's coefficient of skewness.
[Hint: See Example 18] [Ans. Coeff. of $S_K = 0.237$]

6. Find Pearson's Coefficient of skewness from the following data:

Height (inches):	60—62	63—65	66—68	69—71	72—74
Frequency:	5	18	42	27	8

[Ans. Coeff. of $S_K = 0.034$]

7. From the marks secured by 120 students in Section A and 120 in Section B, the following measures are obtained :

Section A : $\overline{X} = 35.0$, $\sigma = 7$, $Z = 32$
Section B : $\overline{X} = 40.0$, $\sigma = 10$, $Z = 30$

Determine which distribution of marks is more skewed. [Ans. B is more skewed]

(2) Bowley's Method

Prof. Bowley has given another method of measuring skewness. It is based upon median (M), first quartile (Q_1) and third quartile (Q_3). It is also called **quartile method of measuring skewness.** Bowley has given the following formulae for measuring skewness:

Absolute Measure of Skewness	Bowley's Coefficient of Skewness
$Sk = Q_3 + Q_1 - 2M$	Coeff. of Sk $= \dfrac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

Steps for Calculation

(1) Calculate Q_1, *i.e.*, first quartile
(2) Calculate Q_3, *i.e.*, third quartile
(3) Calculate M, *i.e.*, median
(4) Substitute these values in the formulae.

The following examples illustrate the procedure of calculating Bowley's measure of skewness:

Calculation of Bowley's Coefficient of Skewness in Discrete Series

Example 9. Find Bowley's coefficient of skewness for the following frequency distribution:

No. of children per family:	0	1	2	3	4	5	6
No. of families:	7	10	16	25	18	11	8

Solution:

Calculation of Bowley's Coefficient of Skewness

No. of children (X)	No. of families (f)	c.f.
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95
	$N = 95$	

Bowley's coefficient of skewness is given by

$$\text{Coefficient of } S_K = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$Q_1 = \text{Size of} \left(\frac{N+1}{4}\right)^{th} \text{item} = \frac{95+1}{4} = 24^{th} \text{item}$$

Size of 24^{th} item is 2. Hence, $Q_1 = 2$

$$Q_3 = \text{Size of} \frac{3(N+1)}{4} \text{th item} = \frac{3(95+1)}{4} = 72 \text{ th item}$$

Size of 72^{th} item is 4. Hence, $Q_3 = 4$

$$M = \text{Size of } \frac{1}{2}(N+1)^{th} \text{ item} = \frac{95+1}{2} = 48^{th} \text{ item.}$$

Size of 48th item is 3. Hence, Median = 3

$$\text{Coeff. of skewness} = \frac{Q_3+Q_1-2M}{Q_3-Q_1}$$

$$= \frac{4+2-2\times3}{4-2} = \frac{0}{2} = 0$$

Calculation of Bowley's Coefficient of Skewness in Continuous Series

Example 10. Calculate coefficient of skewness based on quartiles and median from the following data:

Marks:	0—10	10—20	20—30	30—40	40—50	50—60	60—70	70—80
No. of students:	10	25	20	15	10	35	25	10

Solution: **Calculation of Bowley's Coefficient of Skewness**

Marks	f	c.f.
0—10	10	10
10—20	25	35
20—30	20	55
30—40	15	70
40—50	10	80
50—60	35	115
60—70	25	140
70—80	10	150
	N = 150	

$Q_1 = \frac{N}{4} = \frac{150}{4} = 37.5$th item. Q_1 lies in the class interval 20—30

$$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i = 20 + \frac{375-35}{20} \times 10 = 20 + 1.25 = 21.25$$

$Q_3 = \frac{3N}{4} = \frac{3}{4} \times 150 = 112.5$th item.

Q_3 lies in the class interval 50—60.

$$Q_3 = l_1 + \frac{\frac{3}{4}N - c.f.}{f} \times i = 50 + \frac{112.5-80}{35} \times 10 = 50 + 9.29 = 59.29$$

Median item $= \frac{N}{2} = \frac{150}{2} = 75$ th item.

Median lies in the class interval 40—50.

$$M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 40 + \frac{75-70}{10} \times 10 = 45$$

$\therefore$ $Q_1 = 21.25$, $Q_3 = 59.29$, $M = 45$

$$\text{Coeff. of } S_K = \frac{Q_3+Q_1-2M}{Q_3-Q_1}$$

$$= \frac{59.29+21.25-2\times45}{59.29-21.25}$$

$$= \frac{80.54-90}{38.04} = \frac{-9.46}{38.04} = -0.249$$

Example 11. Calculate Coefficient of Q.D. and Bowley's Coefficient of Skewness from the data given below:

Profits in lakhs (less than):	10	20	30	40	50	60	70
No. of Companies:	8	20	40	50	56	59	60

Solution: Since it is a cummulative frequency series, first we convert it into simple frequency series:

Size	f	c.f.
0—10	8	8
10—20	12	20
20—30	20	40
30—40	10	50
40—50	6	56
50—60	3	59
60—70	1	60
	N = 60	

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

Q_1 = Size of $\left(\frac{N}{4}\right)$th item $= \frac{60}{4} = 15$th item. Q_1 lies in the class 10—20.

$$Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i = 10 + \frac{15-8}{12} \times 10 = 10 + 5.833 = 15.833$$

Q_3 = Size of $\left(\frac{3}{4}N\right)$th item $= \frac{3}{4} \times 60 = 45$th item. Q_3 lies in the class 30—40.

$$Q_3 = l_1 + \frac{\frac{3}{4}N - c.f.}{f} \times i = 30 + \frac{45-40}{10} \times 10 = 35$$

$\therefore$ Coeff. of Q.D. $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1} = \dfrac{35 - 15.833}{35 + 15.833} = \dfrac{19.167}{50.833} = 0.377$

Median item = Size of $\dfrac{N}{2}$th item $= \dfrac{60}{2} = 30$th item. Median lies in the class 20—30.

$$\therefore \quad M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 20 + \frac{30 - 20}{20} \times 10 = 20 + 5 = 25$$

$$\text{Coeff. of } S_K = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{35 + 15.833 - 2(25)}{35 - 15.833} = \frac{0.833}{19.167} = 0.043$$

Example 12. Calculate Bowley's Coefficient of Skewness from the following data:

Mid-values:	1	2	3	4	5	6	7	8	9	10
Frequency:	2	9	11	14	20	24	20	16	5	2

Solution: As the mid-values of the different class intervals are given, we first find actual class intervals by using the formula $m \pm \frac{i}{2}$ where m = mid-value, i = difference between two mid-values.

Classes	Mid-values (m)	f	$c.f.$
0.5—1.5	1	2	2
1.5—2.5	2	9	11
2.5—3.5	3	11	22
3.5—4.5	4	14	36
4.5—5.5	5	20	56
5.5—6.5	6	24	80
6.5—7.5	7	20	100
7.5—8.5	8	16	116
8.5—9.5	9	5	121
9.5—10.5	10	2	123
		$N = 123$	

$$Q_1 = \text{size of } \frac{N}{4} \text{ th item} = \frac{123}{4} = 30.75$$

Q_1 lies in the class 3.5 – 4.5.

$$\therefore \quad Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i = 3.5 + \frac{30.75 - 22}{14} \times 1$$

$$= 3.5 + 0.625 = 4.125$$

$$Q_3 = \text{size of } \frac{3N}{4} \text{ th item} = \frac{3 \times 123}{4} = 92.25\text{th item}$$

Q_3 lies in the class 6.5 – 7.5

$$Q_3 = l_1 + \frac{\frac{3N}{4} - c.f.}{f} \times i = 6.5 + \frac{92.25 - 80}{20} \times 1$$

$$= 6.5 + 0.6125 = 7.1125$$

Median item = Size of $\left(\dfrac{N}{2}\right)$th item $= \dfrac{123}{2} = 61.5$th item.

Median lies in the class 5.5 – 6.5.

$$\therefore \quad M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 5.5 + \frac{61.5 - 56}{24} \times 1$$

$$= 5.5 + 0.2292 = 5.7292$$

$$\text{Coeff. of } S_K = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{7.1125 + 4.125 - 2(5.7292)}{7.1125 - 4.125}$$

$$= \frac{11.2375 - 11.4582}{2.9875} = \frac{-0.2207}{2.9875} = -0.073$$

Example 13. For a distribution, Bowley's coefficient of skewness is –0.56, $Q_1 = 16.4$ and Median = 24.2. Find Q_3 and coefficient of quartile deviation.

Solution: Given: Coeff. of $S_K = -0.56$, $Q_1 = 16.4$, $M = 24.2$

$$\text{Coeff. of } S_K = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$-0.56 = \frac{Q_3 + 16.4 - 2(24.2)}{Q_3 - 16.4}$$

$$\Rightarrow -0.56\,(Q_3 - 16.4) = Q_3 + 16.4 - 48.4$$

$$-0.56Q_3 + 9.184 = Q_3 + 16.4 - 48.4$$

$$-0.56Q_3 - Q_3 = 16.4 - 48.4 - 9.184$$

$$-1.56Q_3 = -41.184$$

$$Q_3 = 26.4$$

$$\therefore \quad Q.D. = \frac{Q_3 - Q_1}{2} = \frac{26.4 - 16.4}{2} = 5.$$

Example 14. Find coefficient of skewness from the following information:

Difference of two quartiles = 8

Mode = 11

Sum of two quartiles = 22

Mean = 8

Solution: We know,

$$Z=3M-2\overline{X}$$

$$3M=Z+2\overline{X}=11+2\times8=27$$

$$\Rightarrow \quad M=\frac{27}{3}=9$$

$$\Rightarrow \quad \text{Bowley's Coeff. of Skewness}=\frac{Q_3+Q_1-2M}{Q_3-Q_1}=\frac{22-2\times9}{8}=\frac{22-18}{8}=\frac{4}{8}=\frac{1}{2}=0.5$$

Example 15. For a distribution the distance of the median from the first quartile is five times of the third quartile from the median. Calculate Bowley's Coefficient of Skewness for the distribution.

Solution: We are given : $M-Q_1=5(Q_3-M)$

Bowley's coefficient of skewness is given by:

$$S_K(\text{Bowley})=\frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}=\frac{(Q_3-M)-5(Q_3-M)}{(Q_3-M)+5(Q_3-M)}$$

$$=\frac{-4(Q_3-M)}{6(Q_3-M)}=\frac{-2}{3}=-0.67$$

EXERCISE 7.2

1. Calculate coefficient of quartile deviation and Bowley's coefficient of skewness from the following data:

Size :	Below 10	10—20	20—30	30—40	40—50	Above 50
f:	5	12	20	16	5	2

[**Ans.** $Q_1=18.33, Q_3=35$, Coeff. of Q.D. = 0.313, Coeff. of $S_K=0.019$]

2. Find out Quartiles and Coefficient of Skewness from the following data:

X:	3	4	5	6	7	8	9	10
f:	2	5	7	11	10	8	5	3

[**Ans.** $Q_1=5, M=7, Q_3=8$, Coeff.of $S_K=-0.332$]

3. Calculate Bowley's Coefficient of skewness from the following data:

Mid-values:	75	100	125	150	175	200	225	250
Frequency:	35	40	48	100	125	80	50	22

[**Ans.** Coeff. of $S_K=-0.032$]

4. The mean, mode and Q.D. of a distribution are 42, 36 and 15 respectively. If its Bowley's coefficient of skewness is 1/3, find the values of two quartiles.

[**Hint:** Find $M=\overline{X}-\frac{1}{3}Z$] [**Ans.** $Q_1=20, Q_3=50$]

5. Calculate the quartile co-efficient of skewness for the following distribution:

Class:	1—5	6—10	11—15	16—20	21—25	26—30	31—35
f:	3	4	68	30	10	6	2

[**Ans.** Coeff. of $S_K=0.265$]

(3) Kelly's Method

The third method of measuring skewness is given by Prof. Kelly. It is based on percentiles and deciles. Kelly has given the following formulae for measuring skewness:

Absolute Measures of S_K	Coefficient of S_K
1. $S_K=P_{90}+P_{10}-2M$	1. Coeff. of $S_K=\frac{P_{90}+P_{10}-2M}{P_{90}-P_{10}}$
or	or
2. $S_K=D_9+D_1-2M$	2. Coeff. of $S_K=\frac{D_9+D_1-2M}{D_9-D_1}$

This method is not very popular in practice. It is suitable when the skewness is based on percentiles or deciles.

Steps for Calculation

(1) Calculate P_{90}, *i.e.*, Nineteenth Percentile

(2) Calculate P_{10}, *i.e.*, Tenth Percentile

(3) Calculate M, *i.e.*, Median.

(4) Substitute these in the formulae.

Example 16. From the data given below, find out Kelly's Coefficient of Skewness based on percentiles:

Marks:	0—10	10—20	20—30	30—40	40—50	50—60
No. of students:	4	6	20	10	7	3

Solution:

Marks	f	c.f.
0—10	4	4
10—20	6	10
20—30	20	30
30—40	10	40
40—50	7	47
50—60	3	50
	$N=50$	

$$P_{90}=\text{Size of }\frac{90N}{100}\text{ th item}=\frac{90\times50}{100}=45\text{th item.}$$

P_{90} lies in the class interval 40—50.

$$\therefore \quad P_{90}=l_1+\frac{\frac{90N}{100}-c.f.}{f}\times i=40+\frac{45-40}{7}\times10=47.14$$

$$P_{10}=\text{Size of }\frac{10N}{100}\text{ th item}=\frac{10\times50}{100}=5\text{th item.}$$

P_{10} lies in the class interval 10 — 20.

$$\therefore \quad P_{10} = l_1 + \frac{\frac{10N}{100} - c.f.}{f} \times i$$

$$= 10 + \frac{5-4}{6} \times 10 = 11.67$$

$$\text{Median item} = \text{Size of } \frac{N}{2}\text{th item} = \frac{50}{2} = 25\text{th item.}$$

M lies in the class interval 20–30.

$$\therefore \quad M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 20 + \frac{25-10}{20} \times 10$$

$$= 20 + \frac{15 \times 10}{20} = 27.5$$

$$\therefore \quad P_{90} = 47.17,\ P_{10} = 11.67,\ M = 27.5$$

$$\text{Kelly's Coeff. of Skew.} = \frac{P_{90} + P_{10} - 2M}{P_{90} - P_{10}} = \frac{47.14 + 11.67 - 2 \times 27.5}{47.14 - 11.67} = 0.11$$

EXERCISE 7.3

1. Calculate Kelly's coefficient of skewness from the data given below:

X:	110—115	115—120	120—125	125—130	130—135
f:	4	10	26	49	72
X:	135—140	140—145	145—150	150—155	155—160
f:	90	52	33	17	7

[Ans. Coeff. of $S_K = 0.013$]

2. Compute Kelly's coefficient of skewness based on percentiles from the following:

Marks:	15—20	20—25	25—30	30—35	35—40	40—45
No. of students:	1	2	15	22	7	3

[Ans. Coeff. of $S_K = 0.082$]

MISCELLANEOUS SOLVED EXAMPLES

Example 17: Calculate arithmetic mean, mode, standard deviation and coefficient of skewness for the following:

Marks (less than):	10	20	30	40	50	60
No. of students:	4	10	30	40	47	50

Solution: The above data are in cumulative form. Firstly these data will be converted into simple form:

Marks (X)	f	M.V. (m)	d	d'	fd'	fd'^2
0—10	4	5	−20	−2	−8	16
10—20	6	15	−10	−1	−6	6
20—30	20	25 = A	0	0	0	0
30—40	10	35	+10	+1	10	10
40—50	7	45	+20	+2	14	28
50—60	3	55	+30	+3	9	27
	$N = 50$				$\Sigma fd' = 19$	$\Sigma fd'^2 = 87$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 25 + \frac{19}{50} \times 10 = 28.8$$

By inspection, mode lies in the class 20—30

$$\therefore \quad Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 20 + \frac{20-6}{40-6-10} \times 10$$

$$= 20 + \frac{14}{24} \times 10 = 20 + 5.83 = 25.83$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{87}{50} - \left(\frac{19}{50}\right)^2} \times 10$$

$$= \sqrt{1.74 - 0.1444} \times 10 = 1.263 \times 10 = 12.63$$

$$\text{Coefficient of skewness} = \frac{\overline{X} - Z}{\sigma} = \frac{28.8 - 25.83}{12.63} = \frac{2.97}{12.63} = +0.235$$

Example 18. The daily expenditure of 100 families is given below:

Daily expenditure:	0—20	20—40	40—60	60—80	80—100
No. of families:	13	?	27	?	16

If the mode of the distribution is 44, calculate Karl Pearson Coefficient of skewness.

Solution: Let the missing frequency for the class 20—40 be X. The frequency for the class 60—80 shall be $100 - (56 + x) = 44 - x$

Expenditure	f	$c.f.$
0—20	13	13
20—40	x	13+x
40—60	27	40+x
60—80	$44 - x$	84
80—100	16	100
	$N = 100$	

The formula of mode is:

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Since the given modal value is 44, it lies in the class 40—60.

$$44 = 40 + \frac{27 - x}{54 - x - (44 - x)} \times 20$$

$$44 = 40 + \frac{27 - x}{10} \times 20$$

or $$44 - 40 = \frac{27 - x}{10} \times 20$$

or $$4 = \frac{27 - x}{10} \times 20$$

or $$40 = (27 - x)20$$

$$27 - x = 2$$

or $$x = 25$$

Thus, the frequency for the class 20—40 is 25 and the frequency of the class 60—80 is 44—25 = 19. Thus, the completed frequency distribution is:

0—20	20—40	40—60	60—80	80—100
13	25	27	19	16

Calculation of Coefficient of Skewness

Daily Expenditure	f	M.V. (m)	d	d'	fd'	fd'^2
0—20	13	10	− 40	− 2	− 26	52
20—40	25	30	− 20	− 1	− 25	25
40—60	27	50 = A	0	0	0	0
60—80	19	70	+ 20	+ 1	19	19
80—100	16	90	+ 40	+ 2	32	64
	$N = 100$				$\Sigma fd' = 0$	$\Sigma fd'^2 = 160$

Karl Pearson coefficient of skewness $= \frac{\overline{X} - Z}{\sigma}$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 50 + \frac{0}{100} \times 10 = 50$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{160}{100} - \left(\frac{0}{100}\right)^2} \times 20$$

$$= \sqrt{1.6 - 0} \times 20$$

$$= 1.265 \times 20 = 25.3$$

$Z = 44$ (given)

$$\therefore \quad \text{Coeff. of } S_K = \frac{\overline{X} - Z}{\sigma} = \frac{50 - 44}{25.3} = \frac{6}{25.3} = 0.237$$

Example 19. Calculate Karl Pearson's Coefficient of skewness from the following data:

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 above data are in cummulative frequency. Firstly we convert these data into simple form:

Marks	f	M.V. (m)	d	d'	fd'	fd'^2	$c.f.$
0—10	10	5	−40	−4	−40	160	10
10—20	40	15	−30	−3	−120	360	50
20—30	20	25	−20	−2	−40	80	70
30—40	0	35	−10	−1	0	0	70
40—50	10	45 = A	0	0	0	0	80
50—60	40	55	+10	+1	40	40	120
60—70	16	65	+20	+2	32	64	136
70—80	14	75	+30	+3	42	126	150
80—90	0	85	+40	+4	0	0	150
	$N = 150$				$\Sigma fd' = -86$	$\Sigma fd'^2 = 830$	

As this is a bimodal series (*i.e.*, there are two maximum frequencies), we will find coefficient of skewness by using the formula

$$\text{Coeffi. of } S_K = \frac{3(\overline{X} - M)}{\sigma}$$

$$\overline{X} = A + \frac{\Sigma fd'}{N} \times i = 45 + \frac{(-86)}{150} \times 10 = 45 - \frac{86}{15} = 39.27$$

Median item = Size of $\frac{N}{2} = \frac{150}{2}$ = 75th item. Median lies in class 40—50.

$$M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i$$

$$= 40 + \frac{75 - 70}{10} \times 10 = 40 + \frac{5}{10} \times 10 = 40 + 5 = 45$$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i = \sqrt{\frac{830}{150} - \left(\frac{-86}{150}\right)^2} \times 10$$

$$= \sqrt{5.53 - 0.33} \times 10 = \sqrt{5.20} \times 10 = 2.28 \times 10 = 22.8$$

$$\therefore \text{ Coeff. of skewness} = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(39.27 - 45)}{22.8} = \frac{3(-5.73)}{22.8} = \frac{-17.19}{22.8} = -0.75$$

Example 20. You are given the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of view of workers and that of management.

	Before	After
No. of workers	2,400	2,350
Mean wages (Rs.)	45.5	47.5
Median wages (Rs.)	48.0	45.0
Standard Deviation (Rs.)	12.0	10.0

Solution: The following comments can be made on the basis of the information given:

(*i*) By comparing the total wage bill we can comment on the increase or decrease in the level of wages.

Total wage bill *before* the settlement of dispute = 2,400 × 45.5 = Rs. 1,09,200

Total wage bill *after* the settlement of dispute = 2,350 × 47.5 = Rs. 1,11,625.

Hence the total wage bill has gone up after the settlement of dispute even though the number of workers has decreased from 2,400 to 2,350. This means that the average wage is now higher. This is definitely a gain to the workers.

Conversely, we cannot say that increased wage bill is necessarily a loss to management because if it results in greater efficiency of workers and, therefore, higher productivity, it would be a positive gain to management also.

(*ii*) Median before settlement of the dispute was 48 and after settlement it is 45. This means that formerly 50% of workers used to get above Rs. 48 and now they get only above Rs. 45.

(*iii*) By comparing the coefficient of variation before and after the settlement of dispute we can comment on the distribution of wages.

Coefficient of variation *before* the settlement of dispute

$$C.V. = \frac{\sigma}{\overline{X}} \times 100, \text{ where, } \sigma = 12, \overline{X} = 45.5$$

$$\therefore \quad C.V. = \frac{12}{45.5} \times 100 = 26.37$$

Coefficient of variation *after* the settlement of dispute $\sigma = 10, \overline{X} = 47.5$

$$\therefore \quad C.V. = \frac{10}{47.5} \times 100 = 21.05$$

Since the value of the coefficient of variation has decreased from 26.4 to 21.05 there is sufficient evidence to conclude that wages are more uniformly distributed after the settlement of dispute or, in other words, there is lesser inequality in the distribution of wages after the dispute is settled.

(*iv*) By comparing skewness, we can comment upon the nature of the distribution.

Coefficient of skewness *before* the settlement of dispute

$$S_{Kp} = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(45.5 - 48)}{12} = \frac{-7.5}{12} = -0.625$$

Coefficient of skewness *after* the settlement of dispute

$$S_{Kp} = \frac{3(47.5 - 45)}{10} = \frac{7.5}{10} = +0.75$$

Thus, the distribution is positively skewed after the settlement of dispute whereas it was negatively skewed before the settlement of dispute. This suggests that the number of workers getting low wages has increased considerably and that of workers getting high wages fallen, though the actual wage of workers has increased.

Example 21. Calculate Bowley's coefficient of skewness for the following data:

Size :	5—7	8—10	11—13	14—16	17—19
f:	14	24	38	20	4

Solution: Class intervals are in inclusive form. For finding median and quartiles, we convert the given distribution into exclusive form:

X	*f*	*c.f.*
4.5—7.5	14	14
7.5—10.5	24	38
10.5—13.5	38	76
13.5—16.5	20	96
16.5—19.5	4	100
	N = 100	

$$Q_1 = \text{Size of } \frac{N}{4} = \frac{100}{4} = 25\text{th item.}$$

Q_1 lies in the class interval 7.5—10.5.

$$\therefore \quad Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i = 7.5 + \frac{25 - 14}{24} \times 3 = 8.87$$

$$Q_3 = \text{Size of } \frac{3N}{4} = \frac{3(100)}{4} = 75\text{th item.}$$

Q_3 lies in class interval 10.5—13.5.

$$Q_3 = l_1 + \frac{\frac{3}{4}N - c.f.}{f} \times i = 10.5 + \frac{75 - 38}{38} \times 3 = 13.42$$

$$\text{Median} = \text{Size of } \frac{N}{2} = \frac{100}{2} = 50\text{th item.}$$

M lies in the class interval 10.5—13.5.

$$M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 10.5 + \frac{50-38}{38} \times 3 = 11.447$$

$\therefore$ Bowley's Coefficient of Skewness $= \dfrac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

$$= \frac{13.42 + 8.87 - 2 \times 11.447}{13.42 - 8.87} = -0.13$$

Example 22. In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of upper and lower quartiles is 100 and Median is 38, find the value of lower and upper quartiles.

Solution: Given: $Q_1 + Q_3 = 100, M = 38$ and $S_K = 0.6$

Bowley coeff. of skewness $= \dfrac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

Substituting the values in Bowley's formula, we get

$$0.6 = \frac{100 - 2(38)}{Q_3 - Q_1}$$

or $\quad 0.6(Q_3 - Q_1) = 100 - 76 = 24$

or $\quad Q_3 - Q_1 = \dfrac{24}{0.6} = 40$

Now $Q_3 + Q_1 = 100 \quad$ ...(*i*) (Given)

$Q_3 - Q_1 = 40 \quad$ ...(*ii*)

By adding (*i*) and (*ii*), we get

$2Q_3 = 140$

$\Rightarrow \quad Q_3 = 70$

Also $\quad Q_1 = 100 - 70 = 30$

$\therefore\ Q_1 = 30,\ Q_3 = 70$

Example 23. Pearson's coefficient of skewness for a distribution is 0.4 and coefficient of variance is 30%. Its mode is 88. Find the mean and median.

Solution: We are given $S_{Kp} = 0.4$, mode = 88 and coeff. of variance 30 %. We have to calculate mean and median.

$$S_K = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Coeff. of variance is 30% or $\dfrac{\sigma}{\overline{X}} = 0.3$

$$= \frac{1 - \frac{\text{Mode}}{\text{Mean}}}{\frac{\text{S.D.}}{\text{Mean}}} = \frac{1 - \frac{88}{\text{Mean}}}{0.3} = 0.4 = 1 - \frac{88}{\text{Mean}} = 0.12 = 1 - \frac{88}{\text{Mean}}$$

$$\Rightarrow \quad \frac{88}{\text{Mean}} = 1 - 0.12 = 0.88$$

0.88 Mean $= 88$ or Mean $= \dfrac{88}{0.88} = 100$

Mode $= 3$ Median $- 2\overline{X}$

$88 = 3$ Median $- 2(100)$

or $\quad 3$ Median $= 288$ or Median $= 96$

Hence, mean and median are 100 and 96 respectively.

Example 24. Consider the following distributions:

Items	Distribution A	Distribution B
Mean	100	90
Mode	90	80
Standard Deviation	10	10

(*i*) Distribution A has the same degree of the variation as distribution B.

(*ii*) Both distributions have the same degree of skewness. True/False? Comment, giving reasons.

Solution: (*i*) C.V. for distribution A $= 100 \times \dfrac{\sigma_A}{\overline{X}_A} = 100 \times \dfrac{10}{100} = 10$

C.V. for distribution A $= 100 \times \dfrac{\sigma_B}{\overline{X}_B} = 100 \times \dfrac{10}{90} = 11.11$

Since C.V.(B) > C.V.(A), the distribution B is more variable than the distribution A. Hence, the given statement that the distribution A has the same degree of variation as distribution B is wrong.

(*ii*) Karl Pearson's coefficient of skewness for the distributions A and B is given by:

$$S_K(\text{A}) = \frac{3(\overline{X} - M)}{\sigma} = \frac{3(100-90)}{10} = 3$$

and $\quad S_K(\text{B}) = \dfrac{3(90-80)}{10} = 3$

Since $S_K(\text{A}) = S_K(\text{B}) = 3$, the statement that both the distributions have the same degree of skewness is true.

IMPORTANT FORMULAE

MEASURES OF SKEWNESS

1. Pearson's Measures

Absolute skewness:

$$\text{Skewness} = \bar{X} - Z$$

when mode (Z) is ill defined, then

$$\text{Skewness} = 3(\bar{X} - M)$$

Relative measure of skewness:

(*i*) Coefficient of Skewness $= \dfrac{\bar{X} - Z}{\sigma}$

(*ii*) Coefficient of Skewness $= \dfrac{3(\bar{X} - M)}{\sigma}$ (when Z is ill defined).

2. Bowley's Measures

Absolute skewness:

$$\text{Skewness} = Q_3 + Q_1 - 2M$$

Relative measure of skewness:

$$\text{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

3. Kelly's Measures

Absolute skewness:

$$\text{Skewness} = P_{90} + P_{10} - 2M \quad \text{or} \quad D_9 + D_1 - 2M$$

Relative measure of skewness:

$$\text{Coefficient of skewness} = \frac{P_{90} + P_{10} - 2M}{P_{90} - P_{10}} \quad \text{or} \quad \frac{D_9 + D_1 - 2M}{D_9 - D_1}$$

QUESTIONS

1. What is skewness? How does it differ from dispersion? Describe the various measures of skewness.
2. Distinguish between dispersion and skewness and point out the various methods of measuring skewness.
3. What is skewness? What are tests of skewness? Draw rough sketches to indicate different types of skewness and locate rough the relative position of mean, median and mode in each case.
4. (*i*) Define skewness. How does it differ from dispersion?
   (*ii*) Explain different measures of skewness.

Moments and Measures of Kurtosis

8

INTRODUCTION

Like average, dispersion and skewness, kurtosis is the fourth characteristic of a frequency distribution which gives us an idea about the shape of a frequency distribution. Kurtosis indicates, whether a frequency distribution is more flat-topped or more peaked than the normal distribution. Before taking up a detailed study of Kurtosis, it is necessary to introduce the concept of moments which is essential for its study.

MOMENTS

Moments are the general statistical measures used to describe and analyse the characteristics of a frequency distribution. There are three basis for defining moments:

(1) Moments about the Mean

(2) Moments about Assumed Mean

(3) Moments about zero.

(1) Moments about the Mean

The moment about the mean are called **central moments**. They are denoted by Greek symbol μ (read as mu). If $X_1, X_2, X_3 \ldots X_n$ be the n values of a variable X and $\bar{X}$ be its actual mean, then the rth moment about the mean is defined and given by:

$$\mu_r = \frac{\Sigma(X - \bar{X})^r}{N} \quad \text{where } \mu_r = r\text{th moment about the mean, } r = 1, 2, 3, 4 \ldots$$

For a frequency distribution (or grouped data), the rth moment about mean is defined as:

$$\mu_r = \frac{\Sigma f(X - \bar{X})^r}{N} \quad \text{where } N = \Sigma f, r = 1, 2, 3, 4 \ldots$$

Putting $r = 1, 2, 3$ and 4, the various central moments are as follows:

Individual Series	Discrete/Continuous Series
μ_1 = First Central Moment $= \dfrac{\Sigma(X - \bar{X})^1}{N} = 0$	$\mu_1 = \dfrac{\Sigma f(X - \bar{X})^1}{N} = 0$
μ_2 = Second Central Moment $= \dfrac{\Sigma(X - \bar{X})^2}{N}$	$\mu_2 = \dfrac{\Sigma f(X - \bar{X})^2}{N}$
μ_3 = Third Central Moment $= \dfrac{\Sigma(X - \bar{X})^3}{N}$	$\mu_3 = \dfrac{\Sigma f(X - \bar{X})^3}{N}$
μ_4 = Fourth Central Moment $= \dfrac{\Sigma(X - \bar{X})^4}{N}$	$\mu_4 = \dfrac{\Sigma f(X - \bar{X})^4}{N}$

Moments are extended to higher powers but in practice the first four moments are obtair because of the difficulty of computation.

Note 1: The first central moment μ_1 is always zero, *i.e.*, $\mu_1 = 0$ because the sum of the deviations taken from the mean is zero ($\Sigma(X - \bar{X}) = 0$).

Note 2: The second central moment μ_2 is the square of the standard deviation, *i.e.*, $\mu_2 = (S.D.)^2$. It is equal to the variance of the distribution, *i.e.*, μ_2 = Variance = σ^2.

Example 1. Find the first four central moments of the following numbers: 1, 3, 7, 9, 10

Solution:

Calculation of Moments

X	$\bar{X}=6$ $(X-\bar{X})$	$(X-\bar{X})^2$	$(X-\bar{X})^3$	$(X-\bar{X})^4$
1	1 − 6 = −5	25	−125	625
3	3 − 6 = −3	9	−27	81
7	7 − 6 = 1	1	1	1
9	9 − 6 = 3	9	27	81
10	10 − 6 = 4	16	64	256
$\Sigma X = 30$ $N = 5$	$\Sigma(X-\bar{X}) = 0$	$\Sigma(X-\bar{X})^2 = 60$	$\Sigma(X-\bar{X})^3 = -60$	$\Sigma(X-\bar{X})^4 = 1044$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\mu_1 = \frac{\Sigma(X-\bar{X})}{N} = \frac{0}{5} = 0; \quad \mu_2 = \frac{\Sigma(X-\bar{X})^2}{N} = \frac{60}{5} = 12$$

$$\mu_3 = \frac{\Sigma(X-\bar{X})^3}{N} = \frac{-60}{5} = -12; \quad \mu_4 = \frac{\Sigma(X-\bar{X})^4}{N} = \frac{1044}{5} = 208.8.$$

Example 2. Calculate $\mu_1, \mu_2, \mu_3, \mu_4$ for the following frequency distribution:

Marks:	5—15	15—25	25—35	35—45	45—55	55—65
No. of students:	10	20	25	20	15	10

Solution:

Calculation of Moments

Marks	No. of students (f)	Mid-values X	fX	$\bar{X}=34$ $(X-\bar{X})$	$f(X-\bar{X})$	$f(X-\bar{X})^2$	$f(X-\bar{X})^3$	$f(X-\bar{X})^4$
5—15	10	10	100	−24	−240	5760	−138240	3317760
15—25	20	20	400	−14	−280	3920	−54880	768320
25—35	25	30	750	−4	−100	400	−1600	6400
35—45	20	40	800	6	120	720	4320	25920
45—55	15	50	750	16	240	3840	61440	983040
55—65	10	60	600	26	260	6760	175760	4569760
	$N = 100$		$\Sigma fX = 3400$		$\Sigma f(X-\bar{X}) = 0$	$\Sigma f(X-\bar{X})^2 = 21400$	$\Sigma f(X-\bar{X})^3 = 46800$	$\Sigma f(X-\bar{X})^4 = 29671200$

$$\text{Now, } \bar{X} = \frac{\Sigma fX}{N} = \frac{3400}{100} = 34; \qquad \mu_1 = \frac{\Sigma f(X-\bar{X})}{N} = \frac{0}{100} = 0$$

$$\mu_2 = \frac{\Sigma f(X-\bar{X})^2}{N} = \frac{21400}{100} = 214; \qquad \mu_3 = \frac{\Sigma f(X-\bar{X})^3}{N} = \frac{46800}{100} = 468;$$

$$\mu_4 = \frac{\Sigma f(X-\bar{X})^4}{N} = \frac{9671200}{100} = 96712.$$

(2) Moments about Assumed Mean

When the arithmetic mean is not in whole numbers but in fractions, the calculation of deviations from the mean would involve too many calculations and would take a lot of time. In such a case, the moments from the assumed mean are first calculated and then converted into central moments. The moments about assumed mean are called **non-central moments**. They are denoted by the Greek symbol μ' (pronounced as mu dash). If $X_1, X_2, X_3, ..., X_n$ be the n values of a variable X and A is its assumed mean, then the rth moment about assumed mean is defined and given by:

$$\mu'_r = \frac{\Sigma(X-A)^r}{N} \qquad \text{Where } \mu'_r = r\text{th moment about assumed mean A}$$

$$r = 1, 2, 3, 4.$$

For a frequency distribution, the rth moment about assumed mean (A) is defined as:

$$\mu'_r = \frac{\Sigma f(X-A)^r}{N}$$

Putting r = 1, 2, 3 and 4, the various non-central moments are as follows:

Individual Series	Discrete/Continuous Series
μ'_1 = First Moment about A = $\frac{\Sigma(X-A)^1}{N}$	$\mu'_1 = \frac{\Sigma f(X-A)^1}{N} = \frac{\Sigma fd^1}{N}$
μ'_2 = Second " " " = $\frac{\Sigma(X-A)^2}{N}$	$\mu'_2 = \frac{\Sigma f(X-A)^2}{N} = \frac{\Sigma fd^2}{N}$
μ'_3 = Third " " " = $\frac{\Sigma(X-A)^3}{N}$	$\mu'_3 = \frac{\Sigma f(X-A)^3}{N} = \frac{\Sigma fd^3}{N}$
μ'_4 = Fourth " " " = $\frac{\Sigma(X-A)^4}{N}$	$\mu'_4 = \frac{\Sigma f(X-A)^4}{N} = \frac{\Sigma fd^4}{N}$

Moments can be extended to higher powers in a similar fashion, but in practice, only the first four moments are computed because of the difficulty of computation.

Note: If there is some common factor in the X-column/Mid-value column (m) of a frequency distribution, the computation process of moments can further be simplified by dividing the deviations (d) taken from assumed mean by a common factor (i), and multiply the various moments by i, i^2, i^3 and i^4. Thus the four non-central moments μ'_1, μ'_2, μ'_3 and μ'_4 are calculated as follows:

$$\mu_1' = \frac{\Sigma fd'}{N} \times i; \qquad \mu_2' = \frac{\Sigma fd'^2}{N} \times i^2;$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3; \qquad \mu_4' = \frac{\Sigma fd'^4}{N} \times i^4.$$

Where, $d' = \frac{X - A}{i}$, i = common factor, X = values or mid-values of the X-column.

(3) Moments about Zero or Origin

The moments about zero or origin are denoted by Greek symbol ν(read as nu). If $X_1, X_2, \ldots X_n$ be the values of a variable X, then the rth moment about zero is defined and given by

$$\nu_r = \frac{\Sigma (X - 0)^r}{N} = \frac{\Sigma X^r}{N}$$

For a frequency distribution, rth moment about zero is defined by

$$\nu_r = \Sigma f(X - 0)^r = \frac{\Sigma fX^r}{N}$$

Putting r = 1, 2, 3 and 4, the various moment about zero are as follows:

Individual Series	Discrete/Continuous Series
$\nu_1 = \frac{\Sigma (X-0)^1}{N} = \frac{\Sigma X^1}{N} = \bar{X}$	$\nu_1 = \frac{\Sigma fX}{N} = \bar{X}$
$\nu_2 = \frac{\Sigma (X-0)^2}{N} = \frac{\Sigma X^2}{N}$	$\nu_2 = \frac{\Sigma fX^2}{N}$
$\nu_3 = \frac{\Sigma (X-0)^3}{N} = \frac{\Sigma X^3}{N}$	$\nu_3 = \frac{\Sigma fX^3}{N}$
$\nu_4 = \frac{\Sigma (X-0)^4}{N} = \frac{\Sigma X^4}{N}$	$\nu_4 = \frac{\Sigma fX^4}{N}$

Moments about zero can be extended to higher powers but in practice the first four moments are computed because of the difficulty of computation.

Conversion of Non-Central Moments including Zero into Central Moments

The central moments can be easily computed from the moments about the assumed mean including zero by using the following relations:

Using Moments about Assumed Mean	Using Moments about Origin
$\mu_1 = \mu_1' - \mu_1' = 0$ (Always)	$\mu_1 = 0$
$\mu_2 = \mu_2' - (\mu_1')^2$	$\mu_2 = \nu_2 - \nu_1^2$
$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1')^3$	$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$
$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$	$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2 \cdot \nu_1^2 - 3\nu_1^4$
$\bar{X} = A + \mu_1'$;	$\sigma^2 = \mu_2' - \mu_1'^2$

Conversion of Central Moments into Non-Central Moments including Zero

The moments about any value 'A' (including zero) can be easily computed from the central moments by using the following relations:

Moments about any Value 'A' from Central Moments	Moments about Zero from Central Moments
$\mu_1' = \bar{X} - A$	$\nu_1 = A + \mu_1'$ or $\bar{X}$
$\mu_2' = \mu_2 + (\mu_1')^2$	$\nu_2 = \mu_2 + \nu_1^2$
$\mu_3' = \mu_3 + 3\mu_2'.\mu_1' - 2(\mu_1')^3$	$\nu_3 = \mu_3 + 3\nu_2.\nu_1 - 2\nu_1^3$
$\mu_4' = \mu_4 + 4\mu_3'.\mu_1' - 6\mu_2'(\mu_1')^2 + 3(\mu_1')^4$	$\nu_4 = \mu_4 + 4\nu_3.\nu_1 - 6\nu_2\nu_1^2 + 3\nu_1^4$

Note: 1. The signs are reverse of what we had while converting moments about assumed mean into central moments.

2. It is necessary to find $\bar{X}$ for converting central moments into non-central moments.

UTILITY OF MOMENTS

Moments are useful in analysing the different aspects of frequency distribution. With the help of moments we can measure the central tendency of a set of observations, their variability, their asymmetry and the height of the peak their curves would make. The following is the summary of how moments help in analysing a frequency distribution:

	Moments	What it measures
1.	First moment about origin or zero (ν_1)	Mean
2.	Second moment about the mean (μ_2)	Variance
3.	Second and third moments about the mean (μ_2 and μ_3)	Skewness
4.	Second and fourth moments about the mean (μ_2 and μ_4)	Kurtosis

Example 3. Calculate the first four moments about mean from the following distribution:

X:	1	2	3	4	5	6	7
f:	2	9	25	35	20	8	1

Solution: We shall first determine moments about assumed mean, then calculate the central moments using the appropriate formula.

Calculations of Moments

X	f	$d = X - A$ $A = 4$	fd	fd^2	fd^3	fd^4
1	2	−3	−6	18	−54	162
2	9	−2	−18	36	−72	144
3	25	−1	−25	25	−25	25
4 A	35	0	0	0	0	0
5	20	+1	20	20	20	20
6	8	+2	16	32	64	128
7	1	+3	3	9	27	81
	$N = 100$		$\Sigma fd = -10$	$\Sigma fd^2 = 140$	$\Sigma fd^3 = -40$	$\Sigma fd^4 = 560$

$$\mu_1' = \frac{\Sigma fd^1}{N} = \frac{-10}{320} = -0.1 \qquad \mu_2' = \frac{\Sigma fd^2}{N} = \frac{140}{100} = 1.4$$

$$\mu_3' = \frac{\Sigma fd^3}{N} = \frac{-40}{100} = -0.4 \qquad \mu_4' = \frac{\Sigma fd^4}{N} = \frac{560}{100} = 5.6$$

Now, we convert moments about assumed mean into central moments by using the formula

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 1.4 - (-0.1)^2 = 1.4 - 0.01 = 1.39$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2\mu_1'^3 = (-0.4) - 3 \times 1.4 \times (-0.1) + 2(-0.1)^3$$

$$= -0.4 + 0.42 - 0.002 = 0.018$$

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

$$= 5.6 - 4 \times (-0.4) \times (-0.1) + 6 \times 1.4 \times (-0.1)^2 - 3(-0.1)^4$$

$$= 5.6 - 0.16 + 0.084 - 0.0003 = 5.5237$$

Example 4. Calculate the first four moments about mean for the following distribution:

X:	2.0	2.5	3.0	3.5	4.0	4.5	5.0
f:	5	38	65	92	70	40	10

Solution: We shall first determine moments about assumed mean and then calculate the central moments using the appropriate formula.

Calculation of Moments

X	*f*	$d = X - A$, $A = 3.5$	$d' = d/.5$	fd'	fd'^2	fd'^3	fd'^4
2.0	5	−1.5	−3	−15	45	−135	405
2.5	38	−1.0	−2	−76	152	−304	608
3.0	65	−0.5	−1	−65	65	−65	65
3.5A	92	0	0	0	0	0	0
4.0	70	+0.5	+1	70	70	70	70
4.5	40	+1	+2	80	160	320	640
5.0	10	+1.5	+3	30	90	270	810
	N = 320			$\Sigma fd' = 24$	$\Sigma fd'^2 = 582$	$\Sigma fd'^3 = 156$	$\Sigma fd'^4 = 2598$

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{24}{320} \times (0.5) = 0.0375$$

$$\mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{582}{320} \times (0.5)^2 = 0.4547$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3 = \frac{156}{320} \times (0.5)^3 = 0.0609$$

$$\mu_4' = \frac{\Sigma fd'^4}{N} \times i^4 = \frac{2598}{320} \times (0.5)^4 = 0.5074$$

Now we convert moments about assumed mean into central moments by using the formula

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 0.4547 - (0.0375)^2 = 0.4534$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2\mu_1'^3 = 0.0609 - 3(0.4547)(0.0375) + 2(0.0375)^3 = 0.0099$$

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

$$= 0.5074 - 4(0.0609)(0.0375) + 6(0.454)(0.0375)^2 - 3(.0375)^4 = 0.5021$$

Example 5. Calculate first four central mements from the following distribution:

Height (in inches):	60—62	63—65	66—68	69—71	72—74
Frequency:	5	18	42	27	8

Solution: We shall first determine moments about assumed mean and then calculate the central moments using the appropriate formula.

Calculation of Moments

Height (X)	*f*	M.V. (*m*)	*d*	$d' = d/3$	fd'	fd'^2	fd'^3	fd'^4
60—62	5	61	−6	−2	−10	20	−40	80
63—65	18	64	−3	−1	−18	18	−18	18
66—68	42	67 = A	0	0	0	0	0	0
69—71	27	70	+3	+1	+27	27	27	27
72—74	8	73	+6	+2	+16	32	64	128
	N=100				$\Sigma fd' = 15$	$\Sigma fd'^2 = 97$	$\Sigma fd'^3 = 33$	$\Sigma fd'^4 = 253$

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{15}{100} \times 3 = 0.45$$

$$\mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{97}{100} \times 9 = 8.73$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3 = \frac{33}{100} \times 27 = 8.91$$

$$\mu_4' = \frac{\Sigma fd'^4}{N} \times i^4 = \frac{253}{100} \times 81 = 204.93$$

Now, we convert moments about assumed mean into moments about mean by using the formula.

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 8.73 - (0.45)^2 = 8.73 - 0.20 = 8.53$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2(\mu_1')^3 = 8.91 - 3(8.73)(0.45) + 2(0.45)^3$$

$$= 8.91 - 11.79 + 0.18 = -2.70$$

$$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 3(\mu_1')^4$$
$$= 204.93 - 4\,(8.91)\,(0.45) + 6\,(8.73)\,(0.45)^2 - 3\,(0.45)^4$$
$$= 204.93 - 16.04 + 6\,(8.73)\,(0.25) - 3\,(0.04)$$
$$= 204.93 - 16.04 + 10.61 - 0.12 = 199.38$$

Example 6. The first four moments of a distribution about $x = 2$ are: 1, 2.5, 5.5 and 16. Calculate the four moments about $\overline{X}$ and about zero.

Solution: We are given $A = 2, \mu_1' = 1, \mu_2' = 2.5, \mu_3' = 5.5$ and $\mu_4' = 16$. From these moments about assumed mean (2), we can find out moments about mean with the help of the following formulae:

Moments about mean:

$$\mu_1 = 0$$
$$\mu_2 = \mu_2' - (\mu_1')^2 = 2.5 - (1)^2 = 1.5$$
$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2\mu_1'^3 = 5.5 - 3(2.5)(1) + 2(1)^3 = 5.5 - 7.5 + 2 = 0$$
$$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 3(\mu_1')^4$$
$$= 16 - 4(5.5)(1) + 6(2.5)(1)^2 - 3(1)^4 = 16 - 22 + 15 - 3 = 6$$

Thus, moments about mean are $\mu_1 = 0, \mu_2 = 1.5, \mu_3 = 0, \mu_4 = 6$

Moments about zero:

$$\overline{X} = A + \mu_1' = 2 + 1 = 3$$
$$v_1 = \overline{X} = 3$$
$$v_2 = \mu_2 + v_1^2 = 1.5 + (3)^2 = 10.5$$
$$v_3 = \mu_3 + 3v_2 . v_1 - 2v_1^3 = 0 + 3(10.5)(3) - 2(3)^3 = 40.5$$
$$v_4 = \mu_4 + 4v_3 . v_1 - 6v_2 . v_1^2 + 3v_1^4$$
$$= 6 + 4(40.5)\,(3) - 6(10.5)(3)^2 + 3(3)^4 = 6 + 486 - 567 + 243 = 168$$

Thus, moments about zero are $v_1 = 3, v_2 = 10.5, v_3 = 40.5; v_4 = 168$

Example 7. The arithmetic mean of a series is 22 and first four central moments are 0, 81, –144, and 14817. Find the first four moments (*i*) about the assumed mean '25' and (*ii*) about origin or zero.

Solution: Given $\overline{X} = 22$, $\mu_1 = 0, \mu_2 = 81, \mu_3 = -144$ and $\mu_4 = 14817$

(*i*) About Assumed mean '25' (A = 25)

$$\mu_1' = \overline{X} - A = 22 - 25 = -3$$
$$\mu_2' = \mu_2 + (\mu_1')^2 = 81 + (-3)^2 = 81 + 9 = 90$$
$$\mu_3' = \mu_3 + 3\mu_2' . \mu_1' - 2(\mu_1')^3$$
$$= -144 + 3(90)(-3) - 2(-3)^3$$
$$= -144 - 810 + 54 = -900$$

$$\mu_4' = \mu_4 + 4\mu_3' . \mu_1' - 6\mu_2' . \mu_1'^2 + 3(\mu_1')^4$$
$$= 14817 + 4(-900)\,(-3) - 6(90)\,(-3)^2 + 3(-3)^4$$
$$= 14817 + 10800 - 4860 + 243 = 21{,}000$$

(*ii*) About origin zero ($A = 0$)

$$v_1 = \overline{X} = 22$$
$$v_2 = \mu_2 + v_1^2 = 81 + (22)^2 = 81 + 484 = 565$$
$$v_3 = \mu_3 + 3v_2 . v_1 - 2v_1^3 = -144 + 3(565)\,(22) - 2(22)^3$$
$$= -144 + 37290 - 21296 = 15850$$
$$v_4 = \mu_4 + 4v_3 . v_1 - 6v_2 . v_1^2 + 3v_1^4$$
$$= 14817 + 4(15850)\,(22) - 6(565)\,(22)^2 + 3(22)^4$$
$$= 14817 + 1394800 - 1640760 + 702768 = 4{,}71{,}625$$

SHEPPARD CORRECTIONS FOR GROUPING ERRORS IN MOMENTS

In computing various moments in case of grouped data, it is assumed that the values of all items lying in a class are concentrated at the mid-point of the class. This assumption leads to grouping error in finding the values of the moments. This error is corrected by famous mathematician Sheppard, and therefore, called Sheppard's Corrections. According to Sheppard, first (μ_1) and third (μ_3) moments need no corrections. He has suggested the following formulae for correcting the second (μ_2) and the fourth (μ_4) moments which he regards as crude moments liable to be affected by the grouping error of a continuous series.

$$\mu_2\,(\text{corrected}) = \mu_2(\textit{uncorrected}) - \frac{i^2}{12}$$

$$\mu_4\,(\text{corrected}) = \mu_4(\textit{uncorrected}) - \frac{1}{2}i^2\ \mu_2\,(\textit{uncorrected}) + \frac{7}{240}i^4$$

Where, i = width of class interval.

The first and third moments need no correction.

The following conditions should be satisfied for the application of Sheppard's corrections:

(*i*) The correction should not be made unless the frequency is at least 1000 otherwise the moments will be more affected by sampling errors than by grouping errors.

(*ii*) The correction is not applicable to J-or U-shaped distributions or even to the skew form.

(*iii*) The observations should relate to a continuous variable.

(*iv*) The curve should approach the base line gradually and slowly at each end of the distribution.

Example 8. The first four central moments of a continuous series with class intervals of 3 are arrived at 0, 43.353, –9.774 and 5508.567. Find their corrected values using Sheppard's corrections.

Solution: According to Sheppard, the first and third moments about mean need no correction. Hence, the 2nd and 4th moments only are corrected as follows:

We are given, $\mu_2 = 43.353$ and $\mu_4 = 5508.567$ and $i = 3$

We have,

$$\mu_2(\text{corrected}) = \mu_2 - \frac{i^2}{12} = 43.353 - \frac{(3)^2}{12} = 43.353 - 0.75 = 42.603$$

$$\mu_4\,(\text{corrected}) = \mu_4 - \frac{1}{2}i^2 \cdot \mu_2 + \frac{7}{240}i^4 = 5508.567 - \frac{1}{2}(3)^2(43.353) + \frac{7}{240}(3)^4$$

$$= 5508.567 - 195.0885 + 2.3625 = 5315.841$$

◘ BETA AND GAMMA COEFFICIENTS (OR BETA AND GAMMA MEASURES) BASED ON MOMENTS

Karl Pearson has developed Beta and Gamma Coefficients (or Beta and Gamma Measures) based on the central moments which are given below:

Beta Coefficients or Beta Measures	Gamma Coefficients or Gamma Measures
$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$	$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}}$
$\sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}}$	$\gamma_2 = \beta_2 - 3$ or $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$
$\beta_2 = \frac{\mu_4}{\mu_2^2}$	

Note: These coefficients are used in the calculation of skewness and kurtosis. It needs to be mentioned here that the above coefficients are pure numbers and independent of the units of measurements.

Example 9. The first four central moments are: 0, 4, 8 and 144. Find β and γ coefficients.

Solution: We are given: $\mu_1 = 0, \mu_2 = 4, \mu_3 = 8, \mu_4 = 144$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(8)^2}{(4)^3} = \frac{64}{64} = 1 \qquad \therefore \quad \beta_1 = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{144}{(4)^2} = \frac{144}{16} = 9 \qquad \therefore \quad \beta_2 = 9$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{8}{\sqrt{(4)^3}} = \frac{8}{\sqrt{64}} = \frac{8}{8} = 1$$

$$\gamma_2 = \beta_2 - 3 = 9 - 3 = 6$$

◘ MEASURE OF SKEWNESS BASED ON CENTRAL MOMENTS

A mesure of skewness may be obtained by making use of the second and third central moments. Skewness is measured by β_1 coefficient (read as β_1 coefficient) which is defined and given by:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(\text{Third Central Moment})^2}{(\text{Second Central Moment})^3}$$

In a symmetrical distribution, β_1 shall be zero. The greater the values of β_1, the more skewed the distribution. But β_1 as a measure of skewness cannot tell us about the direction of skewness, *i.e.*, whether it is positive or negative. Therefore, instead of β_1, sometimes $\sqrt{\beta_1}$ is used as a measure of skewness. It is obtained as:

$$\sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} \qquad \text{where } \sqrt{\beta_1} = \text{Moment Coefficient of Skewness}$$

⚬ Interpretation

The value of $\sqrt{\beta_1}$ is interpreted as follows:

(a) If $\sqrt{\beta_1} = 0$, there is no skewness, *i.e.*, the distribution is symmetric.

(b) If $\sqrt{\beta_1} > 0$, there is positive skewness, *i.e.*, the distribution is positively skewed.

(c) If $\sqrt{\beta_1} < 0$, there is negative skewness, *i.e.*, the distribution is negatively skewed.

Example 10. The first three central moments of a distribution are: 0, 2.5, 0.7. Find the moment coefficient of skewness.

Solution: We are given: $\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7$

$$\text{Moment coefficient of skewness} = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0.7}{\sqrt{(2.5)^3}} = \frac{0.7}{\sqrt{15.625}} = \frac{0.7}{3.953} = 0.177$$

Example 11. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Calculate the moment coefficient of skewness.

Solution: We are given: $A = 5, \mu_1' = 2, \mu_2' = 20, \mu_3' = 40, \mu_4' = 50$

$$\mu_1 = 0,\ \mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2\mu_1'^3 = 40 - 3(2)(20) + 2(2)^3 = -64$$

$$\text{Moment coefficient of skewness} = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-64}{\sqrt{(16)^3}} = -1$$

EXERCISE 8.1

1. Calculate the first four moments about mean for the following data:

X:	3	6	8	10	18

[Ans. $\mu_1 = 0, \mu_2 = 25.6, \mu_3 = 97.2, \mu_4 = 1588$]

2. Calculate the first four moments about mean for the following data:

X :	2	3	4	5	6
f:	1	3	7	3	1

[Ans. $\mu_1 = 0, \mu_2 = 0.933, \mu_3 = 0, \mu_4 = 2.530$]

3. Calculate the first four central moments from the following data and also make Sheppards' corrections:

Variable :	0—10	10—20	20—30	30—40
Frequency :	1	3	4	2

[Ans. $\mu_1 = 0, \mu_2 = 81, \mu_3 = -144, \mu_4 = 14{,}817$, μ_2 (corrected) = 71, μ_4 (corrected) = 11058.67]

4. Calculate the first four moments about the mean from the following data and also find the value of β_1 and β_2:

Marks :	0—10	10—20	20—30	30—40	40—50	50—60	60—70
No. of students :	5	12	18	40	15	7	3

[Ans. $\mu_1 = 0, \mu_2 = 177.39, \mu_3 = 47.982, \mu_4 = 95009.364; \beta_1 = 0.0004, \beta_2 = 3.02$]

5. The first four moments of a distribution about the value A=5 are –2, 10, –25 and 50. Find the first four about $\bar{X}$ and about zero.
[Ans. $\mu_1 = 0, \mu_2 = 6, \mu_3 = 19, \mu_4 = 42; \nu_1 = 1, \nu_2 = 7, \nu_3 = 38, \nu_4 = 155$]

6. The arithmetic mean of a series is 5 and the first four central moments are 0, 3, 0 and 26. Find the four moments (*i*) based on assumed mean '4' and (*ii*) based on zero.
[Ans. (*i*) 1, 4, 10 and 45 (*ii*) 5, 28, 170, 1101]

7. Examine whether the following results for obtaining 2nd order central moments are consistent or not: $N = 50, \Sigma X = 100, \Sigma X^2 = 160$.
[Ans. Inconsistent]
[**Hint:** See Example 27]

8. If the first three moments about origin for distribution are 10, 225 and 0 respectively, calculate the first three moments about value '5' for the distribution.
[Ans. $\mu_1' = 5, \mu_2' = 150, \mu_3' = -2750$]

9. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness of the distribution.
[Ans. $\beta_1 = 0.31$, the distribution is slightly skewed]

10. The first four moments of a distribution about value 2 are 1, 2.5, 5.5 and 16 respectively. Calculate the four moments about mean and comment on the nature of distribution.
[Ans. $\mu_1 = 0, \mu_2 = 1.5, \mu_3 = 0, \mu_4 = 6, \beta_1 = 0$, symmetrical, $\beta_2 = 2.67$, platy-kurtic]

11. The first four central moments of a continous series with class intervals of 6 are arrived at 0, –60, 900 and –9500. Find their corrected values accending to Sheppard's corrections.
[Ans. – 63, –8382.2]

◘ KURTOSIS

Kurtosis is a Greek word meaning bulkiness. In statistics, it refers to the degree of flatness or peakedness of a frequency curve. The degree of kurtosis (or peakedness) of a distribution is measured relative to the peakedness of the normal curve. To quote **M.R. Speigal,** *"Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution"*. According to **Croxten and Cowden** *"A measure of kurtosis indicates the degree to which a frequency distribution is peaked or flat-topped". Thus, a measure of kurtosis tells us the extent to which a distribution is more peaked or flat-topped than the normal curve.*

○ Types of Kurtosis

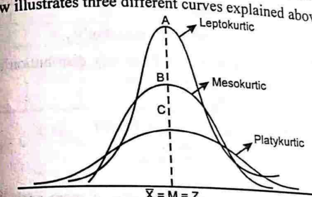
There are three types of kurtosis in a distribution:

(1) **Lepto-kurtic:** A curve having a high peak than the normal curve is called lepto-kurtic. In such a curve, there is too much concentration of the items near the centre.

(2) **Platy-kurtic:** A curve having a low peak (or flat topped) than the normal curve is called platy-kurtic. In such a curve, there is less concentration of items near the centre.

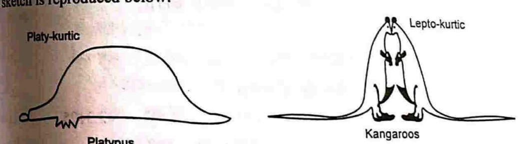
(3) **Meso-kurtic:** A curve having normal peak or the normal curve itself is called meso-kurtic. In such a curve, there is equal distribution of items around the central value.

The figure below illustrates three different curves explained above:



(A) Lepto-kurtic, (B) Meso-kurtic, (C) Platy-kurtic

A famous British statistician **William Gosset** (known as "Student") has very humorously described the nature of the curves in these words **"platy-kurtic curves are squat with short tails, like the platyplus, lepto-kurtic curves are high with long tails like the Kangaroos"**. Gosset's little sketch is reproduced below:



○ Measures of Kurtosis

Kurtosis is measured by β_2 (read as beta two) which is defined and given by:

$$\text{Measure of kurtosis } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\text{Fourth Central Moment}}{(\text{Second Central Moment})^2}$$

▶ **Interpretation**

The value of β_2 is interpreted as follows:

(*i*) If $\beta_2 > 3$, the curve is more peaked than the normal curve, *i.e.*, lepto-kurtic.

(*ii*) If $\beta_2 < 3$, the curve is less peaked than the normal curve, *i.e.*, platy-kurtic.

(*iii*) If $\beta_2 = 3$, the curve is having moderate peak, *i.e.*, meso-kurtic.

▶ Alternative Measure

Sometimes, the Kurtosis is measured by γ_2 (read as Gamma two) which is defined and given by:

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

▶ Interpretation

The value of γ_2 is interpreted as follows:

(*i*) If γ_2 or $\beta_2 - 3 = 0$, the curve is meso-kurtic

(*ii*) If γ_2 or $\beta_2 - 3 > 0$ the curve is lepto-kurtic

(*iii*) If γ_2 or $\beta_2 - 3 < 0$ the curve is platy-kurtic.

Note: It is easier to interpret kurtosis by calculating β_2 instead of γ_2.

Example 12. The first four moments about mean of a frequency distribution are 0, 100, –7 and 35,000. Discuss the kurtosis of the distribution.

$\mu_1 = 0, \mu_2 = 100, \mu_3 = -7, \mu_4 = 35{,}000$

Solution: We are given: $\mu_1 = 0, \mu_2 = 100, \mu_3 = -7, \mu_4 = 35{,}000$

Coefficient of Kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{35{,}000}{(100)^2} = 3.5 > 3.$

Since the value of β_2 is greater than 3, the curve is more peaked than the normal curve, *i.e.*, lepto-kurtic.

Example 13. The first four moments of a distribution about the value '4' of the variable are –1.5, 17, –30 and 108. Discuss the kurtosis of the distribution.

Solution: We are given: $A = 4, \mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 108$

For determining kurtosis, we need to determine μ_2 and μ_4.

$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$= 17 - (-1.5)^2 = 14.75$$
$$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 3(\mu_1')^4$$
$$= 108 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125$$

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

Since, $\beta_2 < 3$, the distribution is platy-kurtic.

Example 14. Calculate first four central moments and coefficient of kurtosis for the following distribution and comment on the result.:

Variable:	0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40
Frequency:	2	5	7	13	21	16	8	3

Solution: We shall first determine moments about assumed mean, and then calculate the central moments using the appropriate formulae:

Calculation of Moments

Variable	f	M.V. (m)	$d = X - A$	$d' = d/5$	fd'	fd'^2	fd'^3	fd'^4
0—5	2	2.5	–20	–4	–8	32	–128	512
5—10	5	7.5	–15	–3	–15	45	–135	405
10—15	7	12.5	–10	–2	–14	28	–56	112
15—20	13	17.5	–5	–1	–13	13	–13	13
20—25	21	22.5 = A	0	0	0	0	0	0
25—30	16	27.5	+5	+1	+16	16	+16	16
30—35	8	32.5	+10	+2	+16	32	+64	128
35—40	3	37.5	+15	+3	+9	27	+81	243
	$N = 75$				$\Sigma fd' = -9$	$\Sigma fd'^2 = 193$	$\Sigma fd'^3 = -171$	$\Sigma fd'^4 = 1429$

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{-9}{75} \times 5 = -0.6 \qquad \mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{193}{75} \times 25 = 64.33$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3 = \frac{-171}{75} \times 125 = -285$$

$$\mu_4' = \frac{\Sigma fd'^4}{N} \times i^4 = \frac{1429}{75} \times 625 = 11908.33$$

Using moments about assumed mean, central moments are calculated as:

$$\mu_1 = 0 \text{ (Always)}$$
$$\mu_2 = \mu_2' - \mu_1'^2 = 64.33 - (-0.6)^2 = 64.33 - 0.36 = 63.97$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -285 - 3(-0.6)(64.33) + 2(-0.6)^3$$
$$= -285 + 115.794 - 0.432 = -169.638$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2' \cdot \mu_1'^2 - 3\mu_1'^4$$
$$= 11908.33 - 4(-285)(-0.6) + 6(64.33)(-0.6)^2 + 3(-0.6)^4$$
$$= 11908.33 - 684 + 138.953 - 0.3888 = 11362.895$$

Coefficient of kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11362.895}{(63.97)^2} = \frac{11362.895}{4092.1609} = 2.776$

Since $\beta_2 < 3$, the distribution is platy-kurtic.

Example 15. The standard deviation of a symmetrical distribution is 3. What must be the value of the fourth moment about the mean in order that the distribution be meso-kurtic?

Solution: For a meso-kurtic distribution $\beta_2 = 3$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad \ldots(i)$$

We are given: $\sigma = 3$

$$\mu_2 = \sigma^2 = (3)^2 = 9$$

Thus, $\beta_2 = 3, \mu_2 = 9$

Putting the value in (*i*)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \Rightarrow 3 = \frac{\mu_4}{9^2} \text{ or } \mu_4 = 243$$

Thus, the fourth moment about the mean must be 243 in order that distribution be meso-kurtic.

Example 16. The following data are given to an economist for the purpose of economic analysis. The data refers to the length of a certain type of batteries:

$N = 100$, $\Sigma fd = 50$, $\Sigma fd^2 = 1970, \Sigma fd^3 = 2948$

and $\Sigma fd^4 = 86{,}752$ in which $d = X - 48$

Do you think that the distribution is platy-kurtic?

Solution: Given, $N = 100, \Sigma fd = 50, \Sigma fd^2 = 1970, \Sigma fd^3 = 2948$ and $\Sigma fd^4 = 86{,}752$.

We shall first determine moments about assumed mean, then calculate the central moments using the appropriate formula:

$$\mu_1' = \frac{\Sigma fd}{N} = \frac{50}{100} = 0.50; \qquad \mu_2' = \frac{\Sigma fd^2}{N} = \frac{1970}{100} = 19.70$$

$$\mu_3' = \frac{\Sigma fd^3}{N} = \frac{2948}{100} = 29.48; \qquad \mu_4' = \frac{\Sigma fd^4}{N} = \frac{86752}{100} = 867.52$$

Using moments about assumed mean, the central moments are:

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 19.70 - (0.50)^2 = 19.45$$

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

$$= 867.52 - 4(29.48)(0.50) + 6(19.70)(0.5)^2 - 3(0.5)^4$$

$$= 867.52 - 58.96 + 29.55 - 0.1875 = 837.9225$$

Coefficient of Kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{837.9225}{(19.45)^2} = \frac{837.9225}{378.3025} = 2.215 < 3$

Since $\beta_2 < 3$, the distribution is platy-kurtic.

COMBINED EXAMPLES ON SKEWNESS AND KURTOSIS

Example 17. Compute the coefficient of skewness (β_1) and coefficient of kurtosis (β_2) based on moments for the following data:

Age:	25—30	30—35	35—40	40—45	45—50	50—55	55—60	60—65
Frequency:	2	8	18	27	25	16	7	2

Solution:

Calculation of Moments

Age	f	M.V.	d	$d' = \frac{d}{5}$	fd'	fd'^2	fd'^3	fd'^4
25—30	2	27.5	−15	−3	−6	18	−54	162
30—35	8	32.5	−10	−2	−16	32	−64	128
35—40	18	37.5	−5	−1	−18	18	−18	18
40—45	27	42.5 = A	0	0	0	0	0	0
45—50	25	47.5	+5	+1	+25	25	+25	25
50—55	16	52.9	+10	+2	+32	64	+128	256
55—60	7	57.5	+15	+3	+21	63	+189	567
60—65	2	62.5	+20	+4	+8	32	+128	512
	N=105				$\Sigma fd'$ = 46	$\Sigma fd'^2$ =252	$\Sigma fd'^3$ =334	$\Sigma fd'^4$ =1668

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{46}{105} \times 5 = 2.19, \qquad \mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{252}{105} \times 5^2 = 60$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3 = \frac{334}{105} \times 5^3 = 397.625 \qquad \mu_4' = \frac{\Sigma fd'^4}{N} \times i^4 = \frac{1668}{105} \times 5^4 = 9892.85$$

Moments about Mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 60 - (2.19)^2 = 55.20$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2\mu_1'^3 = 397.625 - 3(60)(2.19) + 2(2.19)^3 = 24.43$$

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

$$= 9892.85 - 4\,(397.625)(2.19) + 6(60)(2.19)^2 - 3(2.19)^4$$

$$= 9892.85 - 3483.195 + 1726.596 - 69.007 = 8067.25$$

Now, Moment Coefficient of Skewness $(\beta_1) = \frac{\mu_3^2}{\mu_2^3} = \frac{(24.43)^2}{(55.20)^3} = 0.0035$

Moment Coefficient of Kurtosis $(\beta_2) = \frac{\mu_4}{\mu_2^2} = \frac{8067.25}{(55.20)^2} = 2.65 < 3$

Example 18. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution: Given $\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7$ and $\mu_4 = 18.75$.

Testing Skewness

Skewness is measured by the coefficient β_1 which is defined as:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since, $\beta_1 = 0.031$, the distribution is slightly skewed, *i.e.*, it is not perfectly symmetrical.

Testing Kurtosis

For testing kurtosis, we compute the value of β_2 which is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

Since β_2 is exactly 3, the distribution is meso-kurtic.

Example 19. The first four moments of a distribution about the value 4 are 1, 4, 10 and 45. Obtain a measure of skewness and kurtosis.

Solution: We are given $A = 4, \mu_1' = 1, \mu_2' = 4, \mu_3' = 10$ and $\mu_4' = 45$

Moments about Mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 4 - (1)^2 = 3$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2\mu_1'^3 = 10 - 3(4)(1) + 2(1)^3 = 10 - 12 + 2 = 0$$

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

$$= 45 - 4(10)(1) + 6(4)(1)^2 - 3(1)^4 = 45 - 40 + 24 - 3 = 26$$

Measure of Skewness

Skewness is measured by the coefficient β_1 which is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{(3)^3} = 0$$

Since $\beta_1 = 0$, the distribution is symmetrical.

Measures of Kurtosis

Kurtosis is measured by the coefficient β_2 which is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{26}{(3)^2} = 2.89$$

Since $\beta_2 < 3$, the distribution is platy-kurtic.

EXERCISE 8.2

1. The first four moments of a distribution about $x = 4$ are –1.5, 17, –30, 108. Discuss the kurtosis of the distribution. [Ans. $\beta_1 = 0.65$, platy-kurtic]
2. The first four central moments of a distribution are 0, 19.67, 29.26 and 866. Test the skewness and kurtosis of the distribution. [Ans. $\beta_1 = 1125$, $\beta_2 = 2.238$, platy-kurtic]
3. Find a measure of kurtosis for the following distribution:

Marks:	30—35	35—40	40—45	45—50	50—55	55—60	60—65	65—70
No. of students:	5	14	16	25	14	12	8	6

[Ans. $\beta_2 = 2.34$, platy-kurtic]

4. For a meso-kurtic distribution, the first moment about 7 is 23 and second moment about origin is 1000. Find coefficient of variation and fourth moment about mean.
[Hint: See Example 29] [Ans. C.V. = 33.33, μ_4 = 30,000]
5. Analyse the frequency distribution by the method of moments.

X:	2	3	4	5	6
f:	1	3	7	3	1

[Hint: See Example 24] [Ans. $\overline{X} = 4, \sigma = 0.966, \sqrt{\beta_1} = 0, \beta_2 = 2.91$]

6. For the following distribution, calculate the first four central moments and two beta coefficients:

Class-interval:	20—30	30—40	40—50	50—60	60—70	70—80	80—90
Frequency :	5	14	20	25	17	11	8

[Ans. $\mu_1 = 0, \mu_2 = 254, \mu_3 = 540, \mu_4 = 1,49,000, \beta_1 = 0.177945, \beta_2 = 2.31$]

7. For a distribution it has been found that the first four moments about 27 are 0, 256, – 2871 and 1,88,462 respectively. Obtain the first four moments about zero. Also calculate the values of β_1 and β_2 and comment.
[Hint: See Example 28] [Ans. $v_1 = 27, v_2 = 985, v_3 = 37548, v_4 = 15,29,579, \beta_1 = 0.49, \beta_2 = 2.875$]
8. For a distribution mean is 10, standard deviation is 4, $\sqrt{\beta_1} = 1$ and $\beta_2 = 4$. Obtain the first four moments about origin, *i.e.*, zero.
[Hint: See Example 21] [Ans. $v_1 = 10, v_2 = 116, v_3 = 1544, v_4 = 23,184$]

MISCELLANEOUS EXAMPLES

Example 20. Calculate first four central moments from the following and also find the value of β_1 and β_2:

Sales (Rs. crores):	40—50	50—60	60—70	70—80	80—90
No. of companies:	10	25	30	23	12

Solution:

Calculations for Moments

Sales (Rs. crores)	f	Mid values m	$d = m - A$	$d' = \frac{d}{10}$	fd'	fd'^2	fd'^3	fd'^4
40—50	10	45	– 20	–2	–20	40	–80	160
50—60	25	55	– 10	–1	–25	25	–25	25
60—70	30	65 = A	0	0	0	0	0	0
70—80	23	75	10	+1	+23	23	+23	23
80—90	12	85	20	+2	+24	48	+96	192
	$N = 100$				$\Sigma fd' = 2$	$\Sigma fd'^2 = 136$	$\Sigma fd'^3 = 14$	$\Sigma fd'^4 = 400$

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{2}{100} \times 10 = 0.2; \qquad \mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{136}{100} \times 100 = 136$$

$$\mu_3' = \frac{\Sigma f d'^3}{N} \times i^3 = \frac{14}{100} \times 1000 = 140;\ \mu_4' = \frac{\Sigma f d'^4}{N} \times i^4 = \frac{400}{100} \times 10{,}000 = 40{,}000$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 136 - (0.2)^2 = 135.96$$

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

$$= 140 - 3(136)(0.2) + 2(0.2)^3 = 140 - 81.6 + 0.016 = 58.416$$

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

$$= 40{,}000 - 4(140)(0.2) + 6(136)(0.2)^2 - 3(0.2)^4$$

$$= 40{,}000 - 112 = 32.64 - 0.0048 = 39{,}920.64$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(58.416)^2}{(135.96)^3} = 0.0014$$

β_1 is a measure of skewness. Since the value of β_1 is very close to zero, the distribution is more or less symmetrical.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{39{,}920.64}{(135.96)^2} = 2.16$$

β_2 is a measure of kurtosis. Since the value of β_2 is less than 3, the curve is platy-kurtic.

Example 21. For a distribution mean is 10, standard deviation is 4, $\sqrt{\beta_1} = 1$ and $\beta_2 = 4$. Obtain the first four moments about the origin.

Solution: Given: $\overline{X} = 10,\ \sigma = 4,\ \sqrt{\beta_1} = 1,\ \beta_2 = 4,$

$$\mu_2 = \sigma^2 = (4)^2 = 16$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \Rightarrow 1 = \frac{\mu_3^2}{(16)^3} \Rightarrow \mu_3^2 = 4096$$

$$\therefore \quad \mu_3 = \sqrt{4096} = 64$$

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

$\therefore$ $\mu_1 = 0$ (always), $\mu_2 = 16, \mu_3 = 64,\ \mu_4 = 1024$

Moments about Zero

$$v_1 = \overline{X} \text{ or } A + \mu_1' = 10$$

$$v_2 = \mu_2 + v_1^2 = 16 + (10)^2 = 116$$

$$v_3 = \mu_3 + 3v_2 . v_1 - 2v_1^3$$

$$= 64 + 3(116)(10) - 2(10)^3 = 64 + 3480 - 2000 = 1544$$

$$v_4 = \mu_4 + 4v_3 . v_1 - 6v_2 . v_1^2 + 3v_1^4$$

$$= 1024 + 4(1544)(10) - 6(116)(10)^2 + 3(10)^4$$

$$= 1024 + 61760 - 69600 + 30000 = 23{,}184$$

Example 22. The first four moments of a distribution about $X = 4$ are 1, 4, 10 and 45. Obtain the various characteristics of the distribution on the basis of the information given. Comment upon the nature of distribution.

Solution: We are given: $A = 4,\ \mu_1' = 1,\ \mu_2' = 4,\ \mu_3' = 10$ and $\mu_4' = 45$

According to the formulae on moments, the different possible characteristics of the distribution will be brought as under:

(*i*) Mean of distribution $\overline{X} = A + \mu_1' = 4 + 1 = 5$

(*ii*) S.D. of the distribution or $\sigma = \sqrt{\mu_2} = \sqrt{\mu_2' - (\mu_1')^2}$

$$= \sqrt{4 - (1)^2} = \sqrt{4 - 1} = \sqrt{3} = 1.732$$

(*iii*) Variance $= \sigma^2 = \mu_2 = \mu_2' - (\mu_1')^2 = 4 - 1 = 3$

(*iv*) Coefficient of variance or C.V. $= \frac{\sigma}{\overline{X}} \times 100$

$$= \frac{1.732}{5} \times 100 = 34.64\%$$

(*v*) Coefficient of skewness or $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

Where, $\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1')^3 = 10 - 3(4)(1) + 2(1)^3 = 10 - 12 + 2 = 0$

and $\mu_2 = 3$

Thus, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{(3)^2} = \frac{0}{27} = 0$

Comment: As $\beta_1 = 0$, the distribution is symmetric.

(*vi*) Coefficient of kurtosis or $\beta_2 = \frac{\mu_4}{\mu_2^2}$

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

$$= 45 - 4(10)(1) + 6(4)(1)^2 - 3(1)^4 = 45 - 40 + 24 - 3 = 26$$

and $\mu_2 = 3$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{26}{(3)^2} = 2.88$$

Comment: As $\beta_2 < 3$, the distribution is platy-kurtic.

Example 23. Compute the coefficient of skewness (β_1) and kurtosis (β_2) based on moments from the following data:

X:	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f:	1	5	12	22	17	9	4	3	1	1

Solution:

Calculation of Skewness and Kurtosis

X	f	$d' = \frac{X-44.5}{10}$	fd'	fd'^2	fd'^3	fd'^4
4.5	1	−4	−4	16	−64	256
14.5	5	−3	−15	45	−135	405
24.5	12	−2	−24	48	−96	192
34.5	22	−1	−22	22	−22	22
44.5	17	0	0	0	0	0
54.5	9	+1	+9	9	+9	9
64.5	4	+2	+8	16	+32	64
74.5	3	+3	+9	27	+81	243
84.5	1	+4	+4	16	+64	256
94.5	1	+5	+5	25	+125	625
	$N = 75$		$\Sigma fd' = -30$	$\Sigma fd'^2 = 224$	$\Sigma fd'^3 = -6$	$\Sigma fd'^4 = 2{,}072$

$$\mu_1' = \frac{\Sigma fd'}{N} \times i = \frac{-30}{75} \times 10 = -4; \qquad \mu_2' = \frac{\Sigma fd'^2}{N} \times i^2 = \frac{224}{75} \times 10^2 = 298.66$$

$$\mu_3' = \frac{\Sigma fd'^3}{N} \times i^3 = \frac{-6}{75} \times 10^3 = -80 \qquad \mu_4' = \frac{\Sigma fd'^4}{N} \times i^4 = \frac{2{,}072}{75} \times 10^4 = 27626.66$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 298.66 - (-4)^2 = 282.66$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2\mu_1'^3 = -80 - 3\,(298.66)(-4) + 2(-4)^3 = 3375.92$$

$$\mu_4 = \mu_4' - 4\,{}_1\mu_3'\mu' + 6\mu_2'\mu_1'^2 - 3(\mu_1')^4$$

$$= 27626.66 - 4\,(-80)(-4) + 6(298.66)(-4)^2 - 3(-4)^4$$

$$= 27626.66 - 1280 + 28671.36 - 768 = 302890.02$$

Skewness: $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(3375.92)^2}{(282.66)^3} = 0.504$

For kurtosis we have to compute the value of β_2

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{302890.02}{(282.66)^2} = 3.79 > 3$$

Since the value of β_2 is greater than 3, the curve is more peaked than the normal curve, *i.e.*, lepto-kurtic.

Example 24. Given the frequency distribution:

X:	2	3	4	5	6
f:	1	3	7	3	1

Show that the distribution is symmetric and platy-kurtic.

Solution: For determining the symmetricity and kurtosis of the distribution we are to assess the value of β_1 and β_2 and for this, we compute the first four moments about the mean, *i.e.*, μ_1, μ_2, μ_3 and μ_4.

Calculation for Central Moments

X	f	fX	$\bar{X} = 4$ $X - \bar{X}$ x	fx	fx^2	fx^3	fx^4
2	1	2	−2	−2	4	−8	16
3	3	9	−1	−3	3	−3	3
4	7	28	0	0	0	0	0
5	3	15	+1	+3	3	+3	3
6	1	6	+2	+2	4	+8	16
	$N = 15$	$\Sigma fX = 60$		$\Sigma fx = 0$	$\Sigma fx^2 = 14$	$\Sigma fx^3 = 0$	$\Sigma fx^4 = 38$

We have $\bar{X} = \frac{\Sigma fX}{N} = \frac{60}{15} = 4$

First four central moments are:

$$\mu_1 = \frac{\Sigma fx}{N} = \frac{0}{15} = 0; \qquad \mu_2 = \frac{\Sigma fx^2}{N} = \frac{14}{15} = 0.933$$

$$\mu_3 = \frac{\Sigma fx^3}{N} = \frac{0}{15} = 0; \qquad \mu_4 = \frac{\Sigma fx^4}{N} = \frac{38}{15} = 2.533$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(0.933)^2} = 0$$

Since, the value of β_1 is 0, the distribution is symmetric

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2.533}{(0.933)^2} = 2.91$$

Since, the value of β_2 is less than 3, the distribution is platy-kurtic.

Example 25. The first four central moments of a continuous series with class intervals of 10 are arrived at 0, 20, 40 and 50 respectively. Find the corrected values of the moments according to Sheppard's corrections.

Solution: According to Sheppard, the first and third moments about the mean need no correction. Hence, the 2nd and 4th moments only are corrected as follows:

We are given, $\mu_1' = 0, \mu_2' = 20, \mu_3' = 40$ and $\mu_4' = 50, i = 10$

We have

$$\mu_2(\text{corrected}) = \mu_2 - \frac{i^2}{12} = 20 - \frac{10^2}{12} = 20 - \frac{100}{12} = 20 - 8.33 = 11.67$$

and $\mu_4(\text{corrected}) = \mu_4 - \frac{1}{2} i^2 \cdot \mu_2 + \frac{7}{240}(i)^4 = 50 - \frac{1}{2} \cdot (10)^2(20) + \frac{7}{240}(10)^4$

$$= 50 - 1000 + \frac{7}{240}(10{,}000) = 50 - 1000 + 291.67 = -658.33$$

IMPORTANT TYPICAL EXAMPLES

Example 26. For a distribution, the mean is 10, standard deviation is 4, $\sqrt{\beta_1} = 1$, and $\beta_2 = 4$. Obtain the first four moments about '4'.

Solution: Given, $\overline{X} = 10, \sigma = 4, \sqrt{\beta_1} = 1$ and $\beta_2 = 4$

$$\mu_2 = \sigma^2 = (4)^2 = 16$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \Rightarrow 1 = \frac{\mu_3^2}{(16)^3} \Rightarrow \mu_3^2 = 4096$$

$$\therefore \quad \mu_3 = \sqrt{4096} = 64$$

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

$$\therefore \quad \mu_1 = 0, \mu_2 = 16, \mu_3 = 64, \mu_4 = 1024$$

Moments about 4

$$\mu_1' = \overline{X} - A = 10 - 4 = 6$$

$$\mu_2' = \mu_2 + (\mu_1')^2 = 16 + (6)^2 = 52$$

$$\mu_3' = \mu_3 + 3\mu_2' \cdot \mu_1' - 2(\mu_1')^3$$

$$= 64 + 3(52)(6) - 2(6)^3 = 64 + 936 - 432 = 568$$

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

$$= 1024 + 4(568)(6) - 6(52)(6)^2 + 3(6)^4$$

$$= 1024 + 13632 - 11232 + 3888 = 7312$$

Example 27. Examine whether the following results of a piece of computation for obtaining the second central moments are consistent or not:

$N = 120, \Sigma fX = -125, \Sigma fX^2 = 128$

Solution:

$$\mu_1 = \frac{-125}{120} = -1.042 \qquad \mu_2' = \frac{128}{120} = 1.066$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 1.066 - (1.042)^2 = 1.066 - 1.085 = -0.019$$

As the variance $\mu_2 = \sigma^2$ can never be negative, the data for obtaining μ_2 are not consistent.

Aliter:

$$\sigma^2 = \mu_2 = \frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2 = \frac{128}{120} - \left(\frac{-125}{120}\right)^2 = 1.066 - 1.085 = -0.019$$

Example 28. For a distribution it has been found that the first four moments about 27 are 0, 256, – 2871 and 1,88,462 respectively. Outline the first four moments about zero. Also calculate the values of β_1 and β_2 and comment.

Solution: Given, $A = 27, \mu_1' = 0, \mu_2' = 256, \mu_3' = -2871, \mu_4' = 1,88,462$

Moments about Mean:

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 256 - 0 = 256$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1'^3) = -2871 - 3(256)(0) + 2(0) = -2871$$

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

$$= 188462 - 4(-2871)(0) + 6(256)(0)^2 - 3(0)^4 = 188462$$

Moments about Zero:

$$v_1 = \overline{X} = A + \mu_1' = 27 + 0 = 27$$

$$v_2 = \mu_2 + v_1^2 = 256 + (27)^2 = 256 + 729 = 985$$

$$v_3 = \mu_3 + 3v_2 . v_1 - 2v_1^3$$

$$= -2871 + 3(985)(27) - 2(27)^3$$

$$= -2871 + 79785 - 39366 = 37548$$

$$v_4 = \mu_4 + 4v_3 . v_1 - 6v_2 . (v_1^2) + 3(v_1^4)$$

$$= 188462 + 4(37548)(27) - 6(985)(27)^2 + 3(27)^4$$

$$= 188462 + 4055184 - 4308390 + 1594323 = 1529579$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-2871)^2}{(256)^3} = \frac{8242641}{16777216} = 0.49$$

Comment : Since, $\beta_1 = 0.49$, the distribution is positively of skewed.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{188462}{(256)^2} = \frac{188462}{65536} = 2.875$$

Comment : As $\beta_2 < 3$, the distribution is platy-kurtic.

Example 29. For a meso-kurtic distribution, the first moment about 7 is 23 and the second moment about origin is 1000. Find the coefficient of variation and the fourth moment about the mean.

Solution: Since the distribution is given to be meso-kurtic, we have

$$\beta_2 = 3 \Rightarrow \frac{\mu_4}{\mu_2^2} = 3 \Rightarrow \mu_4 = 3\mu_2^2 \quad ...(i)$$

First moment about '7' is 23

i.e., μ_1' (about 7) = 23 (Given)

$$\therefore \quad \text{Mean} = 7 + \mu_1' = 7 + 23 = 30 \quad ...(ii)$$

But mean is the first moment about origin.

$\therefore$ μ_1' (about origin) = 30

Moments About Origin

$\mu'_1 = \text{Mean} = 30; \quad \mu'_2 = 1{,}000$ (Given)

$\therefore \quad \mu_2 = \mu'_2 - \mu'^2_1 = 1000 - (30)^2 = 100 \Rightarrow$ Variance $(\sigma^2) = 100 \Rightarrow$ S.D.$(\sigma) = 10$

$$\text{Coefficient of Variation (C.V.)} = \frac{100 \times \text{S.D.}}{\text{Mean}} = \frac{100 \times 10}{30} = 33.33$$

Substituting the value of $\mu_2 = 100$, in (*i*), the fourth moment about mean is given by:

$$\mu_4 = 3 \times (100)^2 = 30{,}000.$$

Example 30. If $\beta_1 = +1, \beta_2 = 4$ and variance $= 9$, find the values of μ_3 and μ_4 and comment upon the nature of the distribution.

Solution: We are given, $\beta_1 = +1, \beta_2 = 4$ and variance $= \mu_2 = 9$

$$\beta_1 = +1 \quad \Rightarrow \quad \frac{\mu_3^2}{\mu_2^3} = 1$$

$$\Rightarrow \quad \mu_3^2 = \mu_2^3 = 9 \times 9 \times 9 = (3 \times 3 \times 3)^2 = (27)^2 \Rightarrow \quad \mu_3 = \pm 27$$

$$\text{Also, } \beta_2 = 4 \quad \Rightarrow \quad \frac{\mu_4}{\mu_2^2} = 4 \quad \Rightarrow \quad \mu_4 = 4 \times 9 \times 9 = 324$$

$\therefore \quad \mu_3 = \pm 27$ and $\mu_4 = 324$.

Nature of the Distribution: Since $\beta_1 \neq 0$, but $\beta_1 = 1$, the distribution is moderately skewed. Further, since μ_3 $(= \pm 27)$ can be positive or negative, we cannot tell the direction of the skewness.

Also $\beta_2 = 4 > 3$. Hence, the given distribution is lepto-kurtic, *i.e.*, more peaked than the normal curve.

Example 31. The first three moments of the distribution about the value '2' of the variables are 1, 16 and – 40. Show that the mean is 3, variance is 15 and $\mu_3 = -86$.

Solution: We are given, $A = 2, \mu'_1 = 1, \mu'_2 = 16, \mu'_3 = -40$

$$\text{Mean } (\overline{X}) = A + \mu'_1 = 2 + 1 = 3$$

$$\text{Variance } (\sigma^2) = \mu_2 = \mu'_2 - (\mu'_1)^2 = 16 - (1)^2 = 16 - 1 = 15$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$
$$= -40 - 3(16)(1) + 2(1)^3$$
$$= -40 - 48 + 2 = -86$$

Example 32. The first four moments of a distribution about the value '3' of the variable are 1.2, 13, – 22 and 180. Find the value of β_2.

Solution: We are given, $\mu'_1 = 1.2, \mu'_2 = 13, \mu'_3 = -22, \mu'_4 = 180$

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

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 13 - (-1.2)^2 = 13 - 1.44 = 11.56$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$
$$= 180 - 4(-22)(-1.2) + 6(13)(-1.2)^2 - 3(-1.2)^4$$
$$= 180 - 105.6 + 112.32 - 6.2208$$
$$= 292.32 - 111.8208 = 180.4992$$

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

Here, $\mu_4 = 180.4992$, $\mu_2 = 11.56$

$$\beta_2 = \frac{180.4992}{(11.56)^2} = \frac{180.4992}{133.6336} = 1.35$$

Since β_2 is less than 3, so the curve is platy-kurtic.

IMPORTANT FORMULAE

- **Moments about Mean**

$$\mu_1 = \frac{\Sigma(X - \overline{X})^1}{N} = 0, \qquad \mu_2 = \frac{\Sigma(X - \overline{X})^2}{N}$$

$$\mu_3 = \frac{\Sigma(X - \overline{X})^3}{N}, \qquad \mu_4 = \frac{\Sigma(X - \overline{X})^4}{N}$$

- **For a Frequency Distribution**

$$\mu_1 = \frac{\Sigma f(X - \overline{X})^1}{N}, \qquad \mu_2 = \frac{\Sigma f(X - \overline{X})^2}{N} \text{ etc.}$$

- **Moments about Arbitrary Origin 'A'**

$$\mu'_1 = \frac{\Sigma(X - A)^1}{N}, \qquad \mu'_2 = \frac{\Sigma(X - A)^2}{N}$$

$$\mu'_3 = \frac{\Sigma(X - A)^3}{N}, \qquad \mu'_4 = \frac{\Sigma(X - A)^4}{N}$$

- **For a Frequency Distribution**

$$\mu'_1 = \frac{\Sigma f(X - A)^1}{N} \times i \quad \text{or} \quad \mu'_1 = \frac{\Sigma fd'}{N} \times i$$

$$\mu'_2 = \frac{\Sigma f(X - A)^2}{N} \times i^2 \quad \text{or} \quad \mu'_2 = \frac{\Sigma fd'^2}{N} \times i^2$$

$$\mu'_3 = \frac{\Sigma f(X - A)^3}{N} \times i^3 \quad \text{or} \quad \mu'_3 = \frac{\Sigma fd'^3}{N} \times i^3$$

- **Moments about Zero**

$$v_1 = \frac{\Sigma X^1}{N}, \qquad v_2 = \frac{\Sigma X^2}{N}$$

$$v_3 = \frac{\Sigma X^3}{N}, \qquad v_4 = \frac{\Sigma X^4}{N}$$

- **Relationship between Central and Non-central Moments**

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2\mu_1'^3$$

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

- **Relationship between Central Moments and Moments about Origin**

$$v_1 = \bar{X}, \qquad v_2 = \mu_2 + v_1^2$$

$$v_3 = \mu_3 + 3\mu_2 . v_1 + (v_1)^3, \qquad v_4 = \mu_4 + 4\mu_3 . v_1 + 6\mu_2 . (v_1)^2 + (v_1)^4$$

- **Skewness and Kurtosis**

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \qquad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \qquad \gamma_2 = \beta_2 - 3$$

QUESTIONS

1. What do you understand by skewness and kurtosis? Give formulae for measuring them.
2. Explain the term kurtosis. How does kurtosis differ from skewness?
3. What is kurtosis? How is it measured?
4. Define moments. How are skewness and kurtosis calculated from central moments?
5. Distinguish between skewness and kurtosis.
6. What is kurtosis? What purpose does it serve?
7. Define Moments. How are they useful in analysing the different aspects of a frequency distribution?
8. Discuss various measures of kurtosis?
9. How do you measure skewness and kurtosis by using moments?
10. Give measures of skewness and kurtosis.
11. What is kurtosis? Explain the methods to measure kurtosis.
12. What are Sheppard's corrections for grouping errors ? State the conditions under which Sheppard's corrections are applicable.

PART–II

Correlation

1

INTRODUCTION

In our day-to-day life, we find many examples when a mutual relationship exists between two variables, *i.e.*, with fall or rise in the value of one variable, the fall or rise may take place in the value of other variable. For example, price of a commodity rises as the demand for the commodity goes up. Upto a certain time-period, weight of a person increases with the increase in age. Similarly, the temperature rises with the rise in the sun light. These facts indicate that there is certainly some mutual relationship that exists between the demand for a commodity and its price, the age of a person and his weight, and the sunlight and temperature. The correlation refers to the statistical technique used in measuring the closeness of the relationship between the variables.

DEFINITION OF CORRELATION

Some important definitions of correlation are given below:

1. Correlation analysis deals with the association between two or more variables.
   —Simpson and Kafka
2. If two or more quantities vary in sympathy, so that movement in one tend to be accompanied by corresponding movements in the other, then they are said to be correlated.
   —Conner
3. Correlation analysis attempts to determine the degree of relationship between variables.
   —Ya-Lun Chou

Thus, correlation is a statistical technique which helps in analysing the relationship between two or more variables.

UTILITY OF CORRELATION

The study of correlation is of immense significance in statistical analysis and practical life, which is clear from the following points:

(1) Most of variables show same kind of relationship. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis, we can measure the degree of relationship in one figure between different variables like supply and price, income and expenditure, etc.

(2) Once we come to know that the two variables are mutually related, then we can estimate the value of one variable on the basis of the value of another. This function is performed by regression technique, which is based on correlation. In other words, the concept of regression is based on correlation.

(3) Correlation is also useful for economists. An economist specifies the relationship between different variables like demand and supply, money supply and price level by way of correlation.

(4) In business, a trader makes the estimation of costs, sales, prices, etc., with the help of correlation and makes appropriate plans.

Thus, in every field of practical life, correlation analysis is extremely useful in making a comparative study of two or more related phenomena and analyzing their mutual relationship.

◘ TYPES OF CORRELATION

Main types of correlation are given below:

(1) Positive and Negative Correlation: On the basis of direction of change of the variables, correlation can be classified into two types:

(i) **Positive Correlation:** If two variables X and Y move in the same direction, *i.e.*, if one rises, other rises too and vice versa, then it is a called as positive correlation. Examples of positive correlation are the relationship between price and supply, between money supply and prices, etc.

(ii) **Negative Correlation:** If two variables X and Y move in opposite direction, *i.e.*, if one rises, other falls, and if one falls, other rises, then it is called as negative correlation. Examples of negative correlation are the relationship between demand and price, investment and rate of interest, etc.

(2) Linear and Curvi-Linear Correlation: On the basis of change in proportion, correlation is of two types:

(i) **Linear Correlation:** If the ratio of change of two variables X and Y ($\Delta Y / \Delta X$) remains constant throughout, then they are said to be linearly correlated, like as when everytime supply of a commodity rises by 20% as often as its price rises by 10%, then such two variables have linear relationship. If values of these two variables are plotted on a graph, then all the points will lie on a straight line.

(ii) **Curvi-Linear Correlation:** If the ratio of change between the two variables is not constant but changing, correlation is said to be curvi-linear, like as when everytime price of a commodity rises by 10%, then sometimes its supply rises by 20%, sometimes by 10% and sometimes by 40%, then non-linear or curvi-linear correlation exists between them. In case of curvi-linear correlation, values of the variables plotted on a graph will give a curve.

(3) Simple Partial and Multiple Correlation: On the basis of number of variables studied, correlation may be classified into three types:

(i) **Simple Correlation:** When we study the relationship between two variables only, then it is called simple correlation. Relationship between price and demand, height and weight, income and consumption, etc., are all examples of simple correlation.

(ii) **Partial Correlation:** When three or more variables are taken but relationship between any two of the variables is studied, assuming other variables as constant, then it is called partial correlation. Suppose, under constant temperature, we study the relationship between amount of rainfall and wheat yield, then this will be called as partial correlation.

(iii) **Multiple Correlation:** When we study the relationship among three or more variables then it is called multiple correlation. For example, if we study the relationship between rainfall, temperature and yield of wheat, then it is called as multiple correlation.

◘ CORRELATION AND CAUSATION

Correlation is a numerical measure of direction and magnitude of the mutual relationship between the values of two or more variables. But the presence of correlation should not be taken as the belief that the two correlated variables necessarily have causal relationship as well. Correlation does not always arise from causal relationship but with the presence of causal relationship, correlation is certain to exist. Presence of high degree of correlation between different variables may be due to the following reasons:

(1) Mutual Dependence: The study of economic theory shows that it is not necessary that only one variable may affect other variable. It is possible that the two variables may affect each other mutually. In such situation, it is difficult to know which one is the cause and which one is the effect. For example, price of a commodity is affected by the forces of demand and supply. According to the law of demand, with the rise in price (other things remaining constant), demand for the commodity will fall. Here rise in price is the cause and fall in demand is the effect. On the other hand, with fall in demand, price of the commodity falls. Here fall in demand is the cause and fall in price is the effect. Thus there may be high degree of correlation between two variables due to mutual dependence, but it is difficult to know which one is the cause and which one is the effect.

(2) Due to Pure Chance: In a small sample it is possible that two variables are highly correlated but in universe, these variables are unlikely to be correlated, such correlation may be due to either the fluctuations of pure random sampling or due to the bias of investigator in selecting the sample. The following example makes the point clear:

Income (in Rs.)	5,000	6,000	7,000	8,000	9,000
Weight (in Kg.)	100	120	140	160	180

In the data as stated above, there is perfect positive correlation between income and weight, *i.e.*, weight increases with rise in income and the rate of change of the two variables is also the same. Still such kind of correlation cannot be said to be meaningful. Such relationship is said to be *spurious or non-sense*.

(3) Correlation Due to any Third Common Factor: Two variables may be correlated due to some common third factor rather than having direct correlation. For example, if there is high degree of positive correlation between per hectare field of tea and rice, then this does not imply that rice yield has risen due to the rich yield of tea. Another reason of the good yield of these two is the good rainfall well in time that affects both of these two.

◘ DEGREE OF CORRELATION

Degree of correlation can be known by coefficient of correlation (r). The following can be various types of the degree of correlation:

(1) Perfect Correlation

(2) High Degree of Correlation

(3) Moderate Degree of Correlation

(4) Low Degree of Correlation

(5) Absence of Correlation.

(1) Perfect Correlation: When two variables vary at constant ratio in the same direction, it is perfect positive correlation and when the direction of change is opposite, it is perfect negative correlation. In case of perfect positive correlation, correlation coefficient (r) is equal to +1, and in case of perfect negative correlation, correlation coefficient (r) is equal to –1.

(2) High Degree of Correlation: When correlation exists in very large magnitude, then it is called high degree of correlation. In such a case, correlation coefficient ranges between ±0.75 and ±1.

(3) Moderate Degree of Correlation: Correlation coefficient, on being within the limits ±0.25 and ±0.75 is termed as moderate degree of correlation.

(4) Low Degree of Correlation: When correlation exists in very small magnitude, then it is called as low degree of correlation. In such a case, correlation coefficient ranges between 0 and ± 0.25.

(5) Absence of Correlation: When there is no relationship between the variables, then correlation is found to be absent. In case of absence of correlation, the value of correlation coefficient is zero.

The degree of correlation on the basis of value of correlation coefficient can be summarised with the following table:

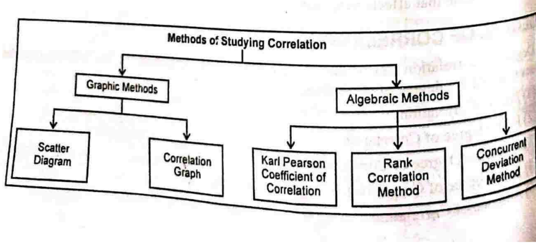
S.No.	Degree of Correlation	Positive	Negative
1.	Perfect Correlation	+1	–1
2.	High Degree of Correlation	Between +0.75 to +1	Between –0.75 to –1
3.	Moderate Degree of Correlation	Between +0.25 to +0.75	Between –0.25 to –0.75
4.	Low Degree of Correlation	Between 0 to + 0.25	Between 0 to –0.25
5.	Absence of Correlation	0	0

METHODS OF STUDYING CORRELATION

Correlation can be determined by the following methods:

(1) Graphic Methods
- (*i*) Scatter Diagram
- (*ii*) Correlation Graph

(2) Algebraic Methods
- (*i*) Karl Pearson's Coefficient of Correlation
- (*ii*) Spearman's Rank Correlation Method
- (*iii*) Concurrent Deviation Method



(1) GRAPHIC METHOD

(I) Scatter Diagram

Scatter diagram is a graphic method of finding out correlation between two variables. By this method, direction of correlation can be ascertained. For constructing a scatter diagram, X-variable is represented on X-axis and the Y-variable on Y-axis. Each pair of values of X and Y series is plotted in two-dimensional space of X—Y. Thus we get a scatter diagram by plotting all the pair of values. Different points may be scattered in various ways in the scatter diagram whose analysis gives us an idea about the direction and magnitude of correlation in the following ways:

(*i*) **Perfect Positive Correlation (r = +1):** If all points are plotted in the shape of a straight line, passing from the lower corner of left side to the upper corner at right side, then both series X and Y have perfect positive correlation, as is clear from the diagram (A) below.

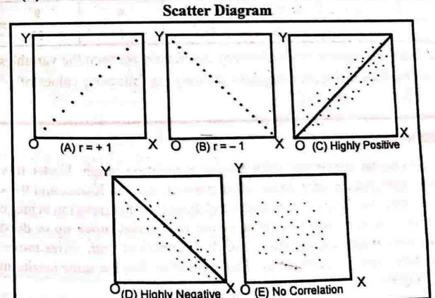
(*ii*) **Perfect Negative Correlation (r=–1):** When all points lie on a straight line from up to down, then X and Y have perfect negative correlation, as is clear from the diagram (B) below.

(*iii*) **High Degree of Positive Correlation:** When concentration of points moves from left to right upward and the points are close to each other, then X and Y have high degree of positive correlation, as is clear from the diagram (C) below.

(*iv*) **High Degree of Negative Correlation:** When points are concentrated from left to right downward, and the points are close to each other, then X and Y have high degree of negative correlation, as is clear from the diagram (D) below.

(*v*) **Zero Correlation (r = 0):** When all the points are scattered in four directions here and there and are lacking in any pattern, then there is absence of correlation, as is clear from the diagram (E) below.

Scatter Diagram



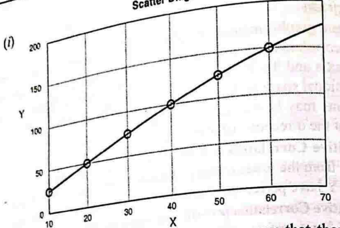
Example 1. Given the following pairs of values of the variable X and Y:

X :	10	20	30	40	50	60
Y :	25	50	75	100	125	150

(*i*) Make a Scatter Diagram.
(*ii*) Is there any correlation between the variables X and Y?

Solution: (*i*)

Scatter Diagram



(*ii*) By looking at the scatter diagram, we can say that there is perfect positive correlation between X and Y variables.

▶ **Merits and Demerits of Scatter Diagram**

Determining correlation by the method is easy because no mathematical computations are to be done. The major shortcoming of this method is that degree of correlation cannot be determined.

EXERCISE 1.1

1. Given the following pairs of values of the variables X and Y:

X:	2	3	5	6	8	9
Y:	6	5	7	8	12	11

(*a*) Make a scatter diagram. (*b*) Is there any correlation between the variables X and Y?

2. Draw three hypothetical scatter diagrams showing the following values of '*r*':
(*i*) r = –1 (*ii*) r = +1 (*iii*) r = 0

(ii) Correlation Graph

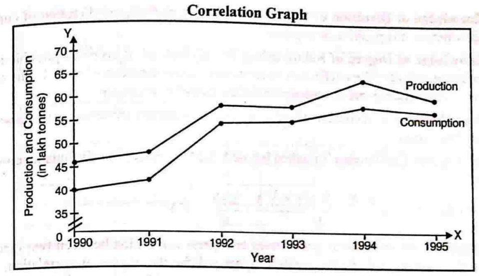
Correlation can also be determined with help of correlation graph. Under this method, two curves are drawn by marking the time, place, serial number, etc., on X-axis and the values of both correlated variables' series on Y-axis. The degree and direction of correlation is judged on the basis of these curves in the following ways: (*a*) If curves of both series move up or down in the same direction, then they have positive correlation, and (b) If curves of both series move in a opposite direction, then they have negative correlation. This method too has the same merits and demerits as those of a scatter diagram.

Example 2: Construct a correlation graph on the basis of the following data and comment on the relationship between production and consumption:

Year:	1990	1991	1992	1993	1994	1995
Production (in lakh tons):	46	48	58	58	64	60
Consumption (in lakh tons):	40	42	54	55	58	57

Solution:

Correlation Graph



In above shown graph, years are shown on OX axis and the production and consumption are shown on OY axis. This graph reveals that the two variables are closely related. Both curves are moving in one direction only. The distance between them also remains almost constant, therefore, there is high degree of positive correlation between them.

EXERCISE 1.2

1. From the following data, ascertain whether the income and expenditure of the 100 workers of a factory are correlated:

Year:	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Average income (in Rs.):	100	102	105	105	101	112	118	120	125	130
Average expenditure:	90	91	93	95	92	94	100	105	108	110

Use Correlation graph. **[Ans. Closely Related]**

2. From the following data, ascertain with the help of correlation graph, whether the demand and price of a commodity are correlated.

Year:	1986	1987	1988	1989	1990	1991	1992	1993
Demand in units:	50	55	62	70	75	78	80	82
Price in Rs.:	40	38	35	30	27	22	20	16

[Ans. Negatively correlated]

(2) ALGEBRAIC METHOD

(i) Karl Pearson's Coefficient of Correlation

It is quantitative method of measuring correlation. This method has been given by Karl Pearson and after his name, it is known as Pearson's coefficient of correlation. This is the best method of working out correlation coefficient. This method has the following main characteristics:

(1) **Knowledge of Direction of Correlation:** By this method, the direction of correlation is determined whether it is positive or negative.

(2) **Knowledge of Degree of Relationship:** By this method, it becomes possible to measure correlation quantitatively. The coefficient of correlation ranges between −1 and +1. The value of the coefficient of correlation gives knowledge about the size of relationship.

(3) **Ideal Measure:** It is considered to be an ideal measure of correlation as it is based on mean and standard deviation.

(4) **Covariance:** Karl Pearson's method is based on co-variance. The formula for co-variance is as follows:

$$\text{Cov}(X, Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N} = \frac{\Sigma XY}{N} - \bar{X}\bar{Y}$$

The magnitude of co-variance can be used to express correlation between two variables. As magnitude of co-variance becomes greater, higher will be the degree of correlation, otherwise lower. With positive sign of covariance, correlation will be positive. On the contrary, correlation will be negative if the sign of covariance is negative.

Calculation of Karl Pearson's Coefficient of Correlation

The calculation of Karl Pearson's coefficient of correlation can be divided into two parts:

(A) Calculation of Coefficient of Correlation in the case of Individual Series or Ungrouped Data.

(B) Calculation of Coefficient of Correlation in the case of Grouped Data.

(A) Calculation of Coefficient of Correlation in case of Individual Series or Ungrouped Data

The following are the main methods of calculating the coefficient of correlation in individual series:

(1) Actual Mean Method

This method is useful when arithmetic mean happens to be in whole numbers or integers. This method involves the following steps:

(1) First, we compute the arithmetic mean of X and Y series, *i.e.*, $\bar{X}$ and $\bar{Y}$ are worked out.

(2) Then from the arithmetic means of the two series, deviations of the individual values are taken. The deviations of *X*-series are denoted by *x* and of the *Y*-series by *y*, *i.e.*, $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

(3) Deviations of the two series are squared and added up to get Σx^2 and Σy^2.

(4) The corresponding deviations of the two series (*x* and *y*) are multiplied and summed up to get Σxy.

(5) Finally, correlation coefficient is found out by using the following formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \text{ or } \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

The correlation coefficient has the value always ranging between −1 and +1. The following examples clarify the computation procedure of this method:

Example 3. From the following data, calculate Karl Pearson's coefficient of correlation:

X:	2	3	4	5	6	7	8
Y:	4	7	8	9	10	14	18

Solution:

Calculation of Coefficient of Correlation

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
2	−3	9	4	−6	36	+18
3	−2	4	7	−3	9	+6
4	−1	1	8	−2	4	+2
5	0	0	9	−1	1	0
6	+1	1	10	0	0	0
7	+2	4	14	+4	16	+8
8	+3	9	18	+8	64	+24
$\Sigma X = 35$ N = 7	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 70$	$\Sigma y = 0$	$\Sigma y^2 = 130$	$\Sigma xy = 58$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{35}{7} = 5, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{70}{7} = 10$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{58}{\sqrt{28 \times 130}} = \frac{58}{\sqrt{3640}}$$

$$= \frac{58}{60.33} = +0.96$$

Thus, there is a high degree of positive correlation between the variables X and Y.

Example 4. From the following data, compute the coefficient of correlation between X and Y series.

	X-Series	Y-Series
Number of items:	15	15
Arithmetic mean:	25	18
Squares of deviations from mean:	136	138

Summation of product of deviations of X and Y series from their respective arithmetic means = 122.

Solution: We are given: $N = 15$, $\bar{X} = 25$, $\bar{Y} = 18$, $\Sigma x^2 = 136$, $\Sigma y^2 = 138$, $\Sigma xy = 122$

Applying the formula,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{122}{\sqrt{136 \times 138}}$$

$$= \frac{122}{\sqrt{18768}} = \frac{122}{136.996} = +0.89$$

IMPORTANT TYPICAL EXAMPLES

Example 5. From the following table, calculate the coefficient of correlation by Karl Pearson's method:

X:	6	2	10	4	8
Y:	9	11	–	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Solution: Let us first find the missing value of Y and let us denote it by a.

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{9+11+a+8+7}{5} = \frac{35+a}{5}$$

$\Rightarrow$ $$8 = \frac{35+a}{5}$$

$\therefore$ $$35 + a = 40 \quad \Rightarrow \quad a = 5$$

Thus, the complete series is:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Now we find the coefficient of correlation.

Calculation of Coefficient of Correlation

X	$\bar{X}=6$ x	x^2	Y	$\bar{Y}=8$ y	y^2	xy
6	0	0	9	1	1	0
2	–4	16	11	3	9	–12
10	4	16	5	–3	9	–12
4	–2	4	8	0	0	0
8	2	4	7	–1	1	–2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

Applying the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-26}{\sqrt{40 \times 20}}$$

$$= \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843}$$

$$= -0.9192$$

Example 6. From the data given below, find the number of items (N):
$r = 0.5$, $\Sigma xy = 120$, Standard Deviation of Y $(\sigma_y) = 8$, $\Sigma x^2 = 90$
Where, x and y are deviations from arithmetic means.

Solution: Given: $r = 0.5, \Sigma xy = 120, \Sigma x^2 = 90, \sigma_y = 8$

Now, $$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} \quad \text{when } y = Y - \bar{Y} \text{ [Formula of S.D.]}$$

$$8 = \sqrt{\frac{\Sigma y^2}{N}}$$, squaring both sides, we get

$$64 = \frac{\Sigma y^2}{N} \quad \Rightarrow \Sigma y^2 = 64N$$

Now, $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \quad \Rightarrow 0.5 = \frac{120}{\sqrt{90 \times 64N}}$$

Squaring both sides

$$0.25 = \frac{(120)^2}{90 \times 64N} \quad \Rightarrow 0.25 = \frac{14400}{5760N}$$

$\Rightarrow$ $(0.25)(5760)N = 14400$

$\Rightarrow$ $(1440)N = 14400$

$\therefore$ $$N = \frac{14400}{1440} = 10$$

EXERCISE 1.3

1. Calculate Karl Pearson's coefficient of correlation between the heights of fathers and sons from the following:

Height of fathers (in inches):	65	66	67	68	69	70	71
Height of sons (in inches):	67	68	66	69	72	72	69

[Ans. $r = 0.668$]

2. Calculate Pearson's coefficient of correlation between X and Y from the following data:

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	37	50	45	33	41	39

[Ans. $r = 0.947$]

3. Calculate the coefficient of correlation using Karl Pearson's formula based on actual mean value of the series given below:

X:	10	12	15	23	20
Y:	14	17	23	25	21

[Ans. $r = 0.864$]

4. From the following data, compute Karl Pearson coefficient of correlation:

	X-series	Y-series
Number of items:	7	7
Arithmetic mean:	4	8
Sum of squares of deviations from arithmetic mean:	28	76

Summation of products of deviations of X and Y series from their respective means is 46.
[Ans. $r = 0.997$]

5. If $r = 0.25$, $\Sigma xy = 45$, $\sigma_y = 3$, $\Sigma x^2 = 50$, where x and y denote deviations from their respective means, find the number of observations.
[Ans. $N = 72$]

6. Two variates X and Y when expressed as deviations from their respective means are given as follows:

x :	–4	–3	–1	–2	0	1	2	3	4
y :	3	–3	?	0	4	1	2	–2	–1

Find the coefficient of correlation between them. [Ans. $r = -0.06$]
[Hint: See Example 51]

7. Calculate Karl Pearson's coefficient of correlation, taking deviations from actual means 52 and 44 of the following data:

X:	44	46	46	48	52	54	?	56	60	60
Y:	36	40	42	40	?	44	46	48	50	52

[Ans. $r = +0.9504$]

8. Determine Pearson's coefficient of correlation from the following data:
$\Sigma X = 250, \Sigma Y = 300, N = 10, \Sigma(X-25)^2 = 480, \Sigma(Y-30)^2 = 600$ and
$\Sigma(X-25)(Y-30) = 150$ [Ans. $r = 0.28$]

(2) Assumed Mean Method

This method is useful when arithmetic mean is not in whole numbers but in fractions. In this method, deviations from assumed means of both the series (X and Y) are calculated. Correlation coefficient by this method can be determined in the following manner:

(1) Any values of X and Y are taken as their assumed mean, Ax and Ay.

(2) Deviations of the individual values of both the series (X and Y) are worked out from their assumed means. Deviations of X series $(X - Ax)$ are denoted by dx and of Y series $(Y - Ay)$ by dy.

(3) Deviations are summed up to get Σdx and Σdy.

(4) Then, squares of the deviations dx^2 and dy^2 are worked out and summed up to get Σdx^2 and Σdy^2 respectively.

(5) Each dx is multiplied by the corresponding dy and the products ($dxdy$) are added upto get $\Sigma dxdy$.

(6) Finally, correlation coefficient is obtained by using any one of following formula:

$$\text{or} \quad r = \frac{\Sigma dxdy - \frac{\Sigma dx \,.\, \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} \quad ...(i)$$

$$\text{or} \quad r = \frac{N \,.\, \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{\sqrt{N \,.\, \Sigma dx^2 - (\Sigma dx)^2} \,.\, \sqrt{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}} \quad ...(ii)$$

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N . \sigma_x . \sigma_y} \quad ...(iii)$$

Note: Unless otherwise specifically asked, formula (*ii*) should be used as it makes the computation work very easy.

The following examples clarify the computation process of this method:

Example 7. Find the coefficient of correlation from the following data:

X :	10	12	18	16	15	19	18	17
Y :	30	35	45	44	42	48	47	46

Solution:

Calculation of Coefficient of Correlation

X	A=16 dx	dx^2	Y	A=42 dy	dy^2	$dxdy$
10	–6	36	30	–12	144	72
12	–4	16	35	–7	49	28
18	+2	4	45	+3	9	6
16 = A	0	0	44	+2	4	0
15	–1	1	42 = A	0	0	0
19	+3	9	48	+6	36	18
18	+2	4	47	+5	25	10
17	+1	1	46	+4	16	4
ΣX = 125 N = 8	Σdx = –3	Σdx² = 71	ΣY = 337 N = 8	Σdy = 1	Σdy² = 283	Σdxdy = 138

$$\bar{X} = \frac{\Sigma X}{N} = \frac{125}{8} = 15.62 \quad , \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{337}{8} = 42.12$$

Since the actual means are not whole numbers, we take 16 as assumed mean for X and 42 as assumed mean for Y.

Applying the formula,

$$r = \frac{N \,.\, \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{\sqrt{N \,.\, \Sigma dx^2 - (\Sigma dx)^2} \,.\, \sqrt{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 138 - (-3)(1)}{\sqrt{8 \times 71 - (-3)^2}\ \sqrt{8 \times 283 - (1)^2}}$$

$$= \frac{1104+3}{\sqrt{568-9}\ \sqrt{2264-1}} = \frac{1107}{\sqrt{559}\ \sqrt{2263}}$$

$$= \frac{1107}{\sqrt{1265017}} = \frac{1107}{1124.72} = 0.98$$

Aliter: $\bar{X} = 15.62$, $\bar{Y} = 42.12$, $Ax = 16$, $Ay = 42$, $\Sigma dxdy = 138$

$$\sigma_x = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2} = \sqrt{\frac{71}{8} - \left(\frac{-3}{8}\right)^2} = \sqrt{\frac{71}{8} - \frac{9}{64}} = 2.95$$

$$\sigma_y = \sqrt{\frac{\Sigma dy^2}{N} - \left(\frac{\Sigma dy}{N}\right)^2} = \sqrt{\frac{283}{8} - \left(\frac{1}{8}\right)^2} = \sqrt{\frac{283}{8} - \frac{1}{64}} = 5.94$$

Applying the formula

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N.\sigma_x.\sigma_y}$$

$$= \frac{138 - 8(15.62-16)(42.12-42)}{8 \times 2.95 \times 5.94} = \frac{138 - 8(-0.38)(0.12)}{140.184}$$

$$= \frac{138 - 0.3648}{140.184} = \frac{137.6352}{140.184} = 0.98$$

Example 8. Calculate Karl Pearson's coefficient of correlation from the following data:

X:	24	27	28	28	29	30	32	33	35	35	40
Y:	18	20	22	25	22	28	28	30	27	30	22

(You may use 32 as working mean for X and 25 that for Y.)

Solution:

Calculation of Coefficient of Correlation

X	A=32 dx	dx^2	Y	A=25 dy	dy^2	$dxdy$
24	–8	64	18	–7	49	56
27	–5	25	20	–5	25	25
28	–4	16	22	–3	9	12
28	–4	16	25 = A	0	0	0
29	–3	9	22	–3	9	9
30	–2	4	28	+3	9	–6
32 = A	0	0	28	+3	9	0
33	1	1	30	+5	25	5
35	3	9	27	+2	4	6
35	3	9	30	+5	25	15
40	8	64	22	–3	9	–24
$N = 11$	$\Sigma dx = -11$	$\Sigma dx^2 = 217$		$\Sigma dy = -3$	$\Sigma dy^2 = 173$	$\Sigma dxdy = 98$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx . \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\ \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

$$= \frac{98 - \frac{(-11)(-3)}{11}}{\sqrt{217 - \frac{(-11)^2}{11}} \times \sqrt{173 - \frac{(-3)^2}{11}}} = \frac{98-3}{\sqrt{217-11}\ \sqrt{173-0.82}}$$

$$= \frac{95}{\sqrt{206 \times 172.18}} = \frac{95}{188.33} = 0.504$$

Aliter: r can be calculated by using the formula:

$$r = \frac{\Sigma dxdy \times N - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{98 \times 11 - (-11)(-3)}{\sqrt{11 \times 217 - (-11)^2} \times \sqrt{11 \times 173 - (-3)^2}}$$

$$= \frac{1078 - 33}{\sqrt{2387 - 121}\ \sqrt{1903 - 9}} = \frac{1045}{\sqrt{2266}\ \sqrt{1894}}$$

$$= \frac{1045}{\sqrt{4291804}} = \frac{1045}{2071.66} = 0.504$$

Example 9. Deviations of the items of two series X and Y from assumed mean are as under:

Deviations of X:	+5	–4	–2	+20	–10	0	+3	0	–15	–5
Deviations of Y:	+5	–12	–7	+25	–10	–3	0	+2	–9	–15

Calculate Karl Pearson's coefficient of correlation.

Solution:

dx	dx^2	dy	dy^2	$dxdy$
+5	25	+5	25	25
–4	16	–12	144	48
–2	4	–7	49	14
+20	400	+25	625	500
–10	100	–10	100	100
0	0	–3	9	0
+3	9	0	0	0
0	0	+2	4	0
–15	225	–9	81	135
–5	25	–15	225	75
$\Sigma dx = -8$	$\Sigma dx^2 = 804$	$\Sigma dy = -24$	$\Sigma dy^2 = 1262$	$\Sigma dxdy = 897$

$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{10\times 897-(-8)(-24)}{\sqrt{10\times 804-(-8)^2}\,\sqrt{10\times 1262-(-24)^2}}$$

$$=\frac{8970-192}{\sqrt{8040-64}\,\sqrt{12620-576}}=\frac{8778}{\sqrt{7976}\,\sqrt{12044}}$$

$$=\frac{8778}{\sqrt{96062944}}=\frac{8778}{9801.17}=0.895$$

Calculation of Coefficient of Correlation by taking a Common Factor

Common factor may be used to simplify the calculation of coefficient of correlation. It is important to note here that there will be no effects on the formula of coefficient of correlation if this common factor is used. The main reason is that the coefficient of correlation is independent of the change of origin and scale. If the origin is shifted or scale is changed, it will not affect the value of coefficient of correlation.

Example 10. Calculate coefficient of correlation from the following data:

X:	100	200	300	400	500	600
Y:	110	120	135	140	160	165

Solution: To simplify the calculation, let

$$dx=\frac{X-400}{100},\ dy=\frac{Y-140}{5}$$

Calculation of Coefficient of Correlation

X	dx	dx^2	Y	dy	dy^2	$dxdy$
100	–3	9	110	–6	36	18
200	–2	4	120	–4	16	8
300	–1	1	135	–1	1	1
400	0	0	140	0	0	0
500	+1	1	160	4	16	4
600	+2	4	165	5	25	10
$N=6$	$\Sigma dx=-3$	$\Sigma dx^2=19$		$\Sigma dy=-2$	$\Sigma dy^2=94$	$\Sigma dxdy=41$

$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{6\times 41-(-3)(-2)}{\sqrt{6\times 19-(-3)^2}\times\sqrt{6\times 94-(-2)^2}}$$

$$=\frac{246-6}{\sqrt{105}\,\sqrt{560}}=\frac{240}{\sqrt{58800}}=\frac{240}{242.487}=0.9897$$

IMPORTANT TYPICAL EXAMPLES

Example 11. From the following data, calculate the Karl Pearson's coefficient of correlation between age of students and their playing habits:

Age:	15	16	17	18	19	20
No. of students:	250	200	150	120	100	80
Regular players:	200	150	90	48	30	12

Solution: Since it is asked to find the correlation between age and playing habits, it is required to find the percentage of regular players which is obtained as follows:

No. of students	Regular players	% of Regular players
250	200	$\frac{200}{250}\times 100=80$
200	150	$\frac{150}{200}\times 100=75$
150	90	$\frac{90}{150}\times 100=60$
120	48	$\frac{48}{120}\times 100=40$
100	30	$\frac{30}{100}\times 100=30$
80	12	$\frac{12}{80}\times 100=15$

Now we calculate the correlation coefficient between age and percentage of regular players. Denoting the age by X and percentage of regular players by Y.

X	dx	dx^2	Y	dy	dy^2	$dxdy$
15	–2	4	80	+20	400	–40
16	–1	1	75	+15	225	–15
17 = A	0	0	60 = A	0	0	0
18	+1	1	40	–20	400	–20
19	+2	4	30	–30	900	–60
20	+3	9	15	–45	2025	–135
N = 6	$\Sigma dx=3$	$\Sigma dx^2=19$		$\Sigma dy=-60$	$\Sigma dy^2=3950$	$\Sigma dxdy=-270$

$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{6\times(-270)-(3)(-60)}{\sqrt{6\times 19-(3)^2}\times\sqrt{6\times 3950-(-60)^2}}$$

$$= \frac{-1620+180}{\sqrt{114-9}\sqrt{23700-3600}} = \frac{-1440}{\sqrt{2110500}} = \frac{-1440}{1452.75} - 0.99$$

There is a high degree of negative correlation between age and playing habits. It shows that as age increases, the tendency to play decreases.

Example 12. From the following data, calculate Karl Pearson's coefficient of correlation between age and blindness:

Age	No. of persons (in thousands)	Blinds
0—10	100	55
10—20	60	40
20—30	40	40
30—40	36	40
40—50	24	36
50—60	11	22
60—70	6	18
70—80	3	15

Solution: First, we shall find the number of blinds per lakh of population in each group as:

No. of persons ('000)	Blinds	No. of blinds (per lakh)
100	55	$\frac{55}{100000} \times 100000 = 55$
60	40	$\frac{40}{60000} \times 100000 = 67$
40	40	$\frac{40}{40000} \times 100000 = 100$
36	40	$\frac{40}{36000} \times 100000 = 111$
24	36	$\frac{36}{24000} \times 100000 = 150$
11	22	$\frac{22}{11000} \times 100000 = 200$
6	18	$\frac{18}{6000} \times 100000 = 300$
3	15	$\frac{15}{3000} \times 100000 = 500$

Denoting the Mid Value of Age by X and No. of Blinds per lakh by Y, we find coefficient of correlation.

Age	MV (X)	A = 35 $dx = \frac{X-35}{10}$	dx^2	Y	A=185 dy= Y–185	dy^2	$dxdy$
0—10	5	–3	9	55	–130	16900	390
10—20	15	–2	4	67	–118	13924	236
20—30	25	–1	1	100	–85	7225	85
30—40	35	0	0	111	–74	5476	0
40—50	45	+1	1	150	–35	1225	–35
50—60	55	+2	4	200	+15	225	30
60—70	65	+3	9	300	+115	13225	345
70—80	75	+4	16	500	+315	99225	1260
$N = 8$		$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 3$	$\Sigma dy^2 = 157425$	$\Sigma dxdy = 2311$

$$r = \frac{\Sigma\, dxdy - \frac{\Sigma\, dx\,.\Sigma\, dy}{N}}{\sqrt{\Sigma\, dx^2 - \frac{(\Sigma\, dx)^2}{N}}\sqrt{\Sigma\, dy^2 - \frac{(\Sigma\, dy)^2}{N}}} = \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}}$$

$$= \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2311 - 1.5}{\sqrt{42}\sqrt{157423.87}}$$

$$= \frac{2309.5}{\sqrt{42}\sqrt{157423.87}} = \frac{2309.5}{6.48 \times 396.76} = \frac{2309.5}{2571.004} = +0.898$$

Example 13. From the following data, calculate the coefficient of correlation between X-series and Y-series.

	X-series	Y-series
Mean	74.5	125.5
Assumed mean	69	112
Standard deviation (σ)	13.07	15.85

Sum of products of corresponding deviations of X and Y series from their assumed mean ($\Sigma\, dxdy$) = 2176 and no. of pairs of observations = 8.

Solution: Given:

$N = 8,\ \bar{X} = 74.5,\ A_x = 69,\ \sigma_x = 13.07,$

$\bar{Y} = 125.5,\ A_y = 112,\ \sigma_y = 15.85,\ \Sigma dxdy = 2176$

Applying the formula:

$$r = \frac{\Sigma dxdy - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N.\sigma_x.\sigma_y}$$

Substituting the values in the formula:

$$= \frac{2176 - 8(74.5 - 69)(125.5 - 112)}{8 \times 13.07 \times 15.85}$$

$$= \frac{2176 - 8(5.5)(13.5)}{8 \times 13.07 \times 15.85} = \frac{2176 - 594}{1657.276} = \frac{1582}{1657.276} = +0.9546$$

Example 14. From the following data, calculate the coefficient of correlation between 'age' and 'playing habits':

Age	No. of students	No. of regular players
15—16	200	150
16—17	270	162
17—18	340	170
18—19	360	180
19—20	400	180
20—21	300	120

Solution: First we shall find the percentage of regular players as follows:

No. of students	No. of regular players	% of regular players
200	150	$\frac{150}{200} \times 100 = 75$
270	162	$\frac{162}{270} \times 100 = 60$
340	170	$\frac{170}{340} \times 100 = 50$
360	180	$\frac{180}{360} \times 100 = 50$
400	180	$\frac{180}{400} \times 100 = 45$
300	120	$\frac{120}{300} \times 100 = 40$

Denoting Mid-Value of Age by X and Percentage of Regular Players by Y.

Calculation of Coefficient of Correlation

Age	M.V. (X)	A=17.5 dx	dx^2	% of Regular players (Y)	A= 50 dy	dy^2	$dxdy$
15-16	15.5	−2	4	75	+25	625	-50
16-17	16.5	−1	1	60	+10	100	-10
17-18	17.5 = A	0	0	50 = A	0	0	0
18-19	18.5	+1	1	50	0	0	0
19-20	19.5	+2	4	45	−5	25	-10
20-21	20.5	+3	9	40	−10	100	-30
$N = 6$		$\Sigma dx = 3$	$\Sigma dx^2 = 19$		$\Sigma dy = 20$	$\Sigma dy^2 = 850$	$\Sigma dxdy = -100$

Now,

$$r = \frac{N \times \Sigma dxdy - \Sigma dx \, \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2} \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times (-100) - (3)(20)}{\sqrt{6 \times 19 - (3)^2} \sqrt{6 \times 850 - (20)^2}}$$

$$= \frac{-660}{\sqrt{105}\sqrt{4700}} = \frac{-660}{702.4956} = -0.9395$$

It shows that there is high degree of negative correlation between age and playing habits.

Example 15. From the data given below, calculate the coefficient of correlation by Karl Pearson's method between density of population and death rate:

Cities	Area in sq. miles	Population (in '000)	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Solution: First we calculate density of population and death rate by using the formula and denote them by X and Y.

$$\text{Density of Population} = \frac{\text{Population}}{\text{Area}}$$

$$\text{Death Rate} = \frac{\text{No. of Deaths}}{\text{Population}} \times 1000$$

Cities	Area (in sq. mile)	Population ('000)	No. of deaths	Density (X)	Death rate (%) (Y)
A	150	30,000	300	$\frac{30,000}{150} = 200$	$\frac{300}{30,000} \times 1,000 = 10$
B	180	90,000	1440	$\frac{90,000}{180} = 500$	$\frac{1,440}{90,000} \times 1,000 = 16$
C	100	40,000	560	$\frac{40,000}{100} = 400$	$\frac{560}{40,000} \times 1,000 = 14$
D	60	42,000	840	$\frac{42,000}{60} = 700$	$\frac{840}{42,000} \times 1,000 = 20$
E	120	72,000	1224	$\frac{72,000}{120} = 600$	$\frac{1224}{72,000} \times 1,000 = 17$
F	80	24,000	312	$\frac{24,000}{80} = 300$	$\frac{312}{24,000} \times 1,000 = 13$

Calculation of Coefficient of Correlation

Cities	Density (X)	$\bar{X}=450$, $x=\frac{X-450}{50}$	x^2	Death Rate (Y)	$\bar{Y}=15$, $y=Y-15$	y^2	xy
A	200	−5	25	10	−5	25	25
B	500	+1	1	16	+1	1	1
C	400	−1	1	14	−1	1	1
D	700	+5	25	20	+5	25	25
E	600	+3	9	17	+2	4	6
F	300	−3	9	13	−2	4	6
$N=6$	$\Sigma X=2700$	$\Sigma x=0$	$\Sigma x^2=70$	$\Sigma Y=90$	$\Sigma y=0$	$\Sigma y^2=60$	$\Sigma xy=64$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{2700}{6}=450, \qquad \bar{Y}=\frac{\Sigma Y}{N}=\frac{90}{6}=15$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations:

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2}\cdot\sqrt{\Sigma y^2}}$$

$$=\frac{64}{\sqrt{70}\times\sqrt{60}}=\frac{64}{\sqrt{70\times 60}}$$

$$=\frac{64}{64.81}=+0.9875$$

There is a high degree of positive correlation between density of population and death rate.

EXERCISE 1.4

1. Calculate the Correlation Coefficient from the following data of marks obtained in Commerce (X) and Economics (Y):

X:	50	60	58	47	49	33	65	43	46	68
Y:	48	65	50	48	55	58	63	48	50	70

[Ans. $r=0.611$]

2. Seven students obtained the following percentage of marks in the college test (X) and in the final examination (Y). Find out the coefficient of correlation between these variables:

X:	50	62	72	25	20	60	60
Y:	48	65	74	33	25	55	66

[Ans. $r=0.974$]

3. Calculate Karl Pearson's coefficient of correlation between the values of X and Y for the following data:

X:	78	89	96	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

Assume 69 and 112 as the mean values for X and Y respectively. [Ans. $r=+0.954$]

4. From the following data, calculate the coefficient of correlation between X-series and Y-series:

	X-series	Y-series
Mean:	381.2	24.5
Assumed mean:	380	25
Standard deviation (σ):	16.79	2.97

Summation of products of corresponding deviations of X and Y series from their assumed means ($\Sigma dxdy$) = 390 and no. of pairs of observations = 10. [Ans. $r=0.794$]

5. The following table gives the distribution of items of production and also the relative defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size group:	15—16	16—17	17—18	18—19	19—20	20—21
No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114

[Hint: See Example 52] [Ans. $r=-0.95$]

6. Find out coefficient of correlation from the following data:

X:	300	350	400	450	500	550	600	650	700
Y:	800	900	1000	1100	1200	1300	1400	1500	1600

[Hint: Let $dx=\frac{X-500}{50}$, $dy=\frac{Y-1200}{100}$] [Ans. $r=+1$]

7. Calculate the coefficient of correlation between age group and mortality rate from the following data :

Age group :	0—20	20—40	40—60	60—80	80—100
Rate of mortality :	350	280	540	760	900

[Ans. $r=0.947$]

8. Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below:

Age:	16	17	18	19	20	21	22
No. of students:	350	320	280	240	180	120	50
Regular players:	315	256	182	132	63	18	4

[Ans. $r=-0.994$]

9. Following figures give the rainfall in inches and production in '00 tons for Rabi and Kharif crops for number of years. Find the coefficient of correlation between rainfall and total production:

Rainfall:	20	22	24	26	28	30	32
Rabi production:	15	18	20	32	40	39	40
Kharif production:	15	17	20	18	20	21	15

[Ans. $r = +0.917$]

10. With the following data in 4 cities, calculate the coefficient of correlation by Pearson's method between the density of population and the death rate:

Cities	Area in sq.km.	Population ('000)	No. of deaths
A	200	40	480
B	150	75	1200
C	120	72	1080
D	80	20	280

[Ans. $r = +0.821$]

11. Calculate 'r' from the following data:
$\Sigma X = 225, \Sigma Y = 189, N = 10, \Sigma(X-22)^2 = 85, \Sigma(Y-19)^2 = 25$ and $\Sigma(X-22)(Y-19) = 43$.
[Hint: See Example 53 Aliter] [Ans, $r = 0.96$]

(3) Method Based on the Use of Actual Data

This method is also known as **Product moment method**. When number of observations are few, correlation coefficient can also be calculated without taking deviations either from actual mean or from assumed mean *i.e.* from actual X and Y values. In this method, the correlation coefficient can be determined in the following way:

(1) First of all, values of the variables X and Y series are summed up to get ΣX and ΣY.

(2) The values of the variables of X and Y series are squared up and added to get ΣX^2 and ΣY^2.

(3) The values of X variable and Y variable are multiplied and the product is added up to get ΣXY.

(4) Finally, the following formula is used to get the correlation coefficient:

$$r = \frac{\Sigma XY - \frac{\Sigma X . \Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

or

$$r = \frac{N . \Sigma XY - \Sigma X . \Sigma Y}{\sqrt{N . \Sigma X^2 - (\Sigma X)^2} \sqrt{N . \Sigma Y^2 - (\Sigma Y)^2}}$$

Example 16. From the following data, find Karl Pearson coefficient of correlation:

X:	2	3	1	5	6	4
Y:	4	5	3	4	6	2

Solution:

Calculation of Coefficient of Correlation

X	X^2	Y	Y^2	XY
2	4	4	16	8
3	9	5	25	15
1	1	3	9	3
5	25	4	16	20
6	36	6	36	36
4	16	2	4	8
$N = 6, \Sigma X = 21$	$\Sigma X^2 = 91$	$\Sigma Y = 24$	$\Sigma Y^2 = 106$	$\Sigma XY = 90$

Applying the formula:

$$r = \frac{N . \Sigma XY - \Sigma X . \Sigma Y}{\sqrt{N . \Sigma X^2 - (\Sigma X)^2} \sqrt{N . \Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{6 \times 90 - (21)(24)}{\sqrt{6 \times 91 - (21)^2} \sqrt{6 \times 106 - (24)^2}} = \frac{540 - 504}{\sqrt{546 - 441} \sqrt{636 - 576}}$$

$$= \frac{36}{\sqrt{105} \sqrt{60}} = \frac{0.36}{\sqrt{6300}} = \frac{36}{79.37} = +0.453$$

Example 17. Calculate product moment correlation coefficient from the following data:

X:	–5	–10	–15	–20	–25	–30
Y:	50	40	30	20	10	5

Solution: In this question the mean of X and Y series may come in fractions or negative signs. It will pose a problem in computing deviations, so here method based on the use of actual values will be used.

Calculation of Coefficient of Correlation

X	X^2	Y	Y^2	XY
–5	25	50	2500	–250
–10	100	40	1600	–400
–15	225	30	900	–450
–20	400	20	400	–400
–25	625	10	100	–250
–30	900	5	25	–150
$\Sigma X = -105$, $N = 6$	$\Sigma X^2 = 2275$	$\Sigma Y = 155$	$\Sigma Y^2 = 5525$	$\Sigma XY = -1900$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{6\times(-1900)-(-105)(155)}{\sqrt{6\times 2275-(-105)^2}\sqrt{6\times 5525-(155)^2}}$$

$$= \frac{-11400+16275}{\sqrt{13650-11025}\sqrt{33150-24025}}$$

$$= \frac{4875}{\sqrt{2625}\sqrt{9125}} = \frac{4875}{\sqrt{23953125}} = \frac{4875}{4894.19} = 0.996$$

Example 18. Find the Coefficient of Correlation for the following data:

$N = 10,\ \bar{X} = 5.5,\ \bar{Y} = 4,\ \Sigma X^2 = 385,\ \Sigma Y^2 = 192,\ \Sigma(X+Y)^2 = 947$

Solution:

$\bar{X} = \frac{\Sigma X}{N} \Rightarrow 5.5 = \frac{\Sigma X}{10} \Rightarrow \Sigma X = 55$

$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow 4 = \frac{\Sigma Y}{10} \Rightarrow \Sigma Y = 40$

$\Sigma(X+Y)^2 = \Sigma X^2 + \Sigma Y^2 + 2\Sigma XY = 947$

$\Rightarrow 385 + 192 + 2\Sigma XY = 947 \Rightarrow 2\Sigma XY = 370$

$\Rightarrow \Sigma XY = 185$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

Putting the given values, we get

$$= \frac{10\times 185-(55)(40)}{\sqrt{10\times 385-(55)^2}\sqrt{10\times 192-(40)^2}}$$

$$= \frac{1850-2200}{\sqrt{3850-3025}\sqrt{1920-1600}} = \frac{-350}{\sqrt{825\times 320}}$$

$$= \frac{-350}{513.80} = -0.681$$

IMPORTANT TYPICAL EXAMPLES

Example 19. Calculate the coefficient of correlation from the following data and interpret the result:

$\Sigma XY = 8425, \bar{X} = 28.5, \bar{Y} = 28.0, \sigma_x = 10.5, \sigma_y = 5.6$ and $N = 10$

Solution: On the basis of informations given, we use direct method for the calculation of correlation coefficient:

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

For this formula, the value of ΣXY and N are known, the values of $\Sigma X, \Sigma Y, \Sigma X^2$ and ΣY^2 are to be calculated.

$\bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N\bar{X} = 10\times 28.5 = 285$ ...(*i*)

$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N\bar{Y} = 10\times 28.0 = 280$ ...(*ii*)

$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$ (Formula of S.D.)

$\Rightarrow \sigma_x^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2$

$\therefore \Sigma X^2 = N[\sigma^2_x + (\bar{X})^2] = 10[(10.5)^2 + (28.5)^2] = 9225$ ...(*iii*)

Similarly, $\Sigma Y^2 = N[\sigma_y^2 + (\bar{Y})^2] = 10[(5.6)^2 + (28.0)^2] = 8153.6$ ...(*iv*)

$\Sigma XY = 8425$ (given), $N = 10$

Now, $$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10\times 8425-(285)(280)}{\sqrt{9225\times 10-(285)^2}\sqrt{8153.6\times 10-(280)^2}}$$

$$= \frac{4450}{\sqrt{11025}\sqrt{3136}} = \frac{4450}{5880} = 0.756$$

Interpretation: There is a positive correlation between X and Y.

Aliter: r can be calculated as follows:

$$r = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}$$

$$Cov(X, Y) = \frac{1}{N}\Sigma(X-\bar{X})(Y-\bar{Y}) = \frac{1}{N}\Sigma XY - \bar{X}\,\bar{Y}$$

Substituting the values, we have

$$Cov(X, Y) = \frac{1}{10}(8425) - (28.5)(28.0)$$

$$= 842.5 - 798 = 44.5$$

Now, $$r = \frac{Cov.(X,Y)}{\sigma_x.\sigma_y} = \frac{44.5}{(10.5)(5.6)} = \frac{44.5}{58.8} = 0.756$$

From the value of $r = 0.756$, it appears that there is positive correlation between X and Y.

Example 20. The following are the nine pairs of values of variable X and Y:
$N=9,\ \Sigma X=45,\ \Sigma Y=135,\ \Sigma X^2=285,\ \Sigma Y^2=2085,\ \Sigma XY=731$
While checking it was found out that two pairs were copied as:

X	Y
8	10
6	8

instead of

12	6
10	7

Obtain the correlation coefficient for the corrected data.

Solution: $N=9,\ \Sigma X=45,\ \Sigma Y=135,\ \Sigma X^2=285,\ \Sigma Y^2=2085,\ \Sigma XY=731$

Replacing the wrong values by correct values, new values are

$\Sigma X=45-8-6+12+10=53$

$\Sigma Y=135-10-8+6+7=130$

$\Sigma X^2=285-64-36+144+100=429$

$\Sigma Y^2=2085-100-64+36+49=2006$

$\Sigma XY=731-80-48+72+70=745$

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\ \sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

$$=\frac{9\times745-(53)(130)}{\sqrt{9\times429-(53)^2}\ .\sqrt{9\times2006-(130)^2}}$$

$$=-0.153$$

Example 21. While calculating the coefficient of correlation between the variables X and Y, a computer obtained the following constants:

$N=20, r=0.3, \bar{X}=15, \bar{Y}=20, \sigma_x=4$ and $\sigma_y=5$

In the course of checking, however, it was detected that an item 27 has been wrongly taken as 17 in case of X series and 35 instead of 30 in case of Y series. Obtain the correct value of r.

Solution: Given $N=20, \bar{X}=15, \bar{Y}=20, \sigma_x=4, \sigma_y=5, r=0.3$

We have $\bar{X}=\frac{\Sigma X}{N} \Rightarrow \Sigma X=N\bar{X}=20\times15=300$

But this is not the correct value of ΣX due to mistakes

Corrected $\Sigma X=300-17+27=310$ ...(i)

$$\bar{Y}=\frac{\Sigma Y}{N} \Rightarrow \Sigma Y=N\bar{Y}=20\times20=400$$

But this is not the correct value of ΣY due to mistakes

Corrected $\Sigma Y=400-35+30=395$ ...(ii)

We know $\sigma_x=\sqrt{\frac{\Sigma X^2}{N}-(\bar{X})^2}$ (Formula of S.D.)

$\sigma_x^2=\frac{\Sigma X^2}{N}-(\bar{X})^2 \quad \therefore\ \Sigma X^2=N[\sigma_x^2+(\bar{X})^2]=20[16+225]=4820$

But this is not the correct value of ΣX^2 due to mistakes

Corrected $\Sigma X^2=4820-17^2+27^2=4820-289+729=5260$ ...(iii)

$\sigma_y=\sqrt{\frac{\Sigma Y^2}{N}-(\bar{Y})^2}$ (Formula of S.D.)

$\Rightarrow \sigma_y^2=\frac{\Sigma Y^2}{N}-(\bar{Y})^2 \quad \therefore\ \Sigma Y^2=N[\sigma_y^2+(\bar{Y})^2]=20[25+400]=8500$

But this is not the correct value of ΣY^2 due to mistakes

Corrected $\Sigma Y^2=8500-35^2+30^2=8500-1225+900=8175$ ...(iv)

Calculation of Corrected ΣXY

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\ \sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

$$0.3=\frac{20\times\Sigma XY-(300)(400)}{\sqrt{20\times4820-(300)^2}\ \sqrt{20\times8500-(400)^2}}$$

$$0.3=\frac{20\Sigma XY-1{,}20{,}000}{80\times100}$$

$0.3\times8000=20\Sigma XY-1{,}20{,}000$

$20\Sigma XY=1{,}22{,}400$

$\Rightarrow \Sigma XY=6120$

$\therefore$ Incorrected $\Sigma XY=6120$

But this is not the correct value of ΣXY due to mistakes

Corrected $\Sigma XY=6120+810-595=6335$ ...(v)

Now, the correct value of r would be calculated as:

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\ \sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

$$=\frac{20\times6335-(310)(395)}{\sqrt{20\times5260-(310)^2}\sqrt{20\times8175-(395)^2}}$$

$$=\frac{126700-122450}{\sqrt{105200-96100}\sqrt{163500-156025}}$$

$$=\frac{4250}{\sqrt{9100}\sqrt{7475}}=\frac{4250}{8247.57}=0.5153$$

EXERCISE 1.5

1. Find Karl Pearson's coefficient of correlation between X and Y from the following data:

X:	5	4	3	2	1
Y:	5	2	10	8	4

What will be the correlation coefficient between $2X+3$ and $5Y-4$.
[Hint: See Example 50] **[Ans.** $r=-0.1980$, No effect]

2. Calculate Karl Pearson's coefficient of correlation between the values of X and Y given below:

X:	−15	+18	−12	−10	+15	−20	−25	+15	+16	−14
Y:	+8	−10	+5	+12	−6	+4	+11	−9	−7	+13

[Ans. $r=-0.912$]

3. Calculate 'r' from the following data:
$\Sigma X=225, \Sigma Y=189, N=10, \Sigma(X-22)^2=85$
$\Sigma(Y-19)^2=25$ and $\Sigma(X-22)(Y-19)=43$
[**Hint:** See Example 53] **[Ans.** $r=0.9598$]

4. Following result were obtained from an analysis of 12 pairs of observations:
$n=12, \Sigma X=30, \Sigma Y=5, \Sigma X^2=670, \Sigma Y^2=285, \Sigma XY=334$
Later on it was discovered that one pair of values ($X=11, Y=4$) were copied wrongly, the correct values of the pair was ($X=10, Y=14$). Find the correct value of correlation coefficient. **[Ans.** $r=0.7746$]

5. Calculate the coefficient of correlation from the following data and interpret the result:
$N=10, \bar{X}=15, \bar{Y}=12, \Sigma XY=1500, \sigma_x=4, \sigma_y=9.0$ **[Ans.** $r=-0.833$]

6. Given the following:
$r=-1, \bar{X}=4.5, \bar{Y}=5.5, \sigma_x^2=5.25, \sigma_y^2=5.25, N=8$
One pair of observation ($X=9, Y=10$) omitted to be included and hence to be included, calculate the correct coefficient of correlation. **[Ans.** $r=-0.4$]

7. In two sets of variables X and Y with 50 items each, the following data were observed:
$\bar{X}=10, \sigma_x=3, \bar{Y}=6, \sigma_y=2, r=0.3$
However, on subsequent verification it was found that one value of X(=10) and one value of Y(=6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of correlation coefficient affected?
[**Hint:** See Example 54] **[Ans.** $r=0.3$, it is not affected]

(4) Variance-Covariance Method

This method of determining correlation coefficient is based on covariance. In this method, the following formula is used to obtain correlation coefficient:

$$r=\frac{Cov(X,Y)}{\sqrt{Var(X)}\ \sqrt{Var(Y)}}$$

Or

$$r=\frac{Cov(X,Y)}{\sigma_x.\sigma_y}=\frac{\frac{\Sigma XY}{N}-\overline{X}\overline{Y}}{\sigma_x.\sigma_y}$$

Where, $Cov(X,Y)=\frac{\Sigma xy}{N}=\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N}=\frac{\Sigma XY}{N}-\overline{X}\overline{Y}$

The formula can also be written as:

$$r=\frac{\Sigma xy}{N.\sigma_x.\sigma_y} \quad \text{where, } x=X-\bar{X},\ y=Y-\bar{Y}$$

Example 22. For two series X and Y, Cov $(X, Y)=15$, Var $(X)=36$, Var $(Y)=25$, calculate the coefficient of correlation.

Solution: Given $Cov(X,Y)=15$, $Var(X)=36$, $Var(Y)=25$

$$r=\frac{Cov(X,Y)}{\sqrt{Var(X)}\ \sqrt{Var(Y)}}=\frac{15}{\sqrt{36}\ \sqrt{25}}$$

$$=\frac{15}{\sqrt{900}}=\frac{15}{30}=+0.50$$

Example 23. From the following data, compute the coefficient of correlation between X and Y:

X-series	*Y-series*
$N=30$	$N=30$
$\bar{X}=40$	$\bar{Y}=50$
$\sigma_x=6$	$\sigma_y=7$
$\Sigma xy=360$	

(Where, x and y are deviations from their respective means)

Solution: We are given $N=30, \bar{X}=40, \bar{Y}=50, \sigma_x=6, \sigma_y=7, \Sigma xy=360$

Karl Pearson's coefficient of correlation is given by:

$$r=\frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

$$=\frac{360}{30\times6\times7}=\frac{360}{1260}=\frac{2}{7}=+0.286$$

Aliter:

$$r = \frac{Cov(X,Y)}{\sigma_x . \sigma_y}, \quad Cov(X,Y) = \frac{\Sigma xy}{N} = \frac{360}{30} = 12 \text{ where, } x = X - \bar{X}, y = Y - \bar{Y}$$

$$r = \frac{12}{6 \times 7} = \frac{12}{42} = +0.286$$

IMPORTANT TYPICAL EXAMPLE

Example 24. From two series X and Y, Cov (X,Y) = 25, r = 0.6, variance of X=36. Calculate standard deviation of y.

Solution: Given, $Cov(X,Y) = 25, r = 0.6, var(X) = 36 \Rightarrow \sigma_x = \sqrt{36} = 6.$ $[\because \sigma = \sqrt{variance}]$

$$r = \frac{Cov(X,Y)}{\sigma_x . \sigma_y}$$

$$+0.6 = \frac{25}{6 \times \sigma_y}$$

$$(0.6)(6 \times \sigma_y) = 25$$

$$(3.6)(\sigma_y) = 25$$

$$\sigma_y = \frac{25}{3.6} = 6.94$$

EXERCISE 1.6

1. The following results are obtained regarding two series. Compute coefficient of correlation:

	X-series	Y-series
No. of items:	15	15
Arithmetic mean:	25	18
Standard deviation:	3.01	3.03

Sum of products of deviations of X and Y series from their means = 122. **[Ans. $r = 0.89$]**

2. Calculate the coefficient of correlation where Cov (X,Y) = 488; Variance of X = 824 and Variance of Y = 325. **[Ans. $r = +0.943$]**

3. If covariance between X and Y is 10 and the variance of X and Y are 16 and 9 respectively, find the coefficient of correlation. **[Ans. $r = +0.83$]**

4. Karl Pearson's coefficient of correlation between two variables X and Y is 0.64, their covariance is 16. If the variance of X is 9, find the standard deviation of Y-series. **[Ans. $\sigma_y = 8.33$]**

5. The coefficient of correlation between two variables X and Y is 0.48 and their covariance is 36. If the variance of X-series is 16, find the second moment about mean of Y-series [*i.e.*, variance of Y-series]. **[Ans. $\sigma_y^2 = \mu_2 = 351.5625$]**

(B) Calculation of Coefficient of Correlation in Grouped Data/Bivariate Distribution

When number of items in two series is very large, then we present them by means of a two-way frequency table. This table gives the frequency distribution of two variables X and Y. The class intervals for Y-variables are presented in column heading (captions) and class intervals for X-variables are presented in row headings (stubs). Frequencies of the each cell of the table are counted by means of using tally bars.

Correlation coefficient in case of grouped data is computed by using the following formula:

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx . \Sigma fdy}{N}}{\sqrt{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}} \sqrt{\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}}}$$

Or

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2} \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

Steps

(1) Step deviations of X-variables are worked out and these are denoted by '*dx*'. Similarly, step deviations of Y-variables are calculated and these are denoted by '*dy*'.

(2) Step deviations of X-variables are multiplied by the corresponding frequencies and added up to get Σfdx. Similarly Σfdy is obtained.

(3) By multiplying the squared deviations of X-variables with the corresponding frequencies or multiplying Σfdx by *dx* and adding up, we get Σfdx^2. Similarly Σfdy^2 are obtained.

(4) Multiplying *dx* and *dy* and further multiplying them with their corresponding cell frequencies yields *fdxdy*. This product is written in the cell down at the right side/corner. Adding together all the cornered values vertically and horizontally gives $\Sigma fdxdy$.

(5) Putting the values of Σfdx, Σfdx^2, Σfdy^2 and $\Sigma fdxdy$ in the above formula to obtain correlation coefficient.

The following examples make clear the computation of correlation in grouped data:

Example 25. 30 pairs of X and Y are given below:

X:	14	20	33	25	41	18	24	29	38	45
Y:	147	242	296	312	518	196	214	340	492	568
X:	23	32	37	19	28	34	38	29	44	40
Y:	382	400	288	292	431	440	500	512	415	514
X:	22	39	43	44	12	27	39	38	17	26
Y:	382	481	516	598	122	200	451	387	245	413

Prepare a correlation table taking class interval of X as 10 to 20, 20 to 30, etc. and that of Y as 100 to 200, 200 to 300, etc. and find Karl Pearson's coefficient of correlation.

Solution:

Preparation of Bivariate Frequency Distribution

Y / X →	10—20	20—30	30—40	40—50	Total
100—200	|||(3)				3
200—300	|| (2)	||| (3)	|| (2)		7
300—400		||||(4)	| (1)		5
400—500		|| (2)	卌 (5)	| (1)	8
500—600		| (1)	|(1)	卌 (5)	7
Total	5	10	9	6	N=30

(Lanscape Table Given at Page 35)

Applying the formula,

$$r = \frac{N \times \Sigma fdxdy - \Sigma fdx . \Sigma fdy}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2} \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{30 \times 35 - (16)(9)}{\sqrt{30 \times 38 - (16)^2} \sqrt{30 \times 55 - (9)^2}}$$

$$= \frac{1050 - 144}{\sqrt{1140 - 256} \sqrt{1650 - 81}} = \frac{906}{\sqrt{884} \sqrt{1569}}$$

$$= \frac{906}{29.73 \times 39.61} = \frac{906}{1177.60} = 0.76$$

Example 26. Calculate Karl Pearson's coefficient of correlation from the following data:

X/Y	10—25	25—40	40—55
0-20	10	4	6
20-40	5	40	9
40-60	3	8	15

Solution: **(Landscape Table Given at Page 36)**

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2} \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{100 \times 16 - (6)(12)}{\sqrt{100 \times 46 - (6)^2} \sqrt{100 \times 48 - (12)^2}}$$

$$= \frac{1600 - 72}{\sqrt{4564} \sqrt{4656}}$$

$$= \frac{1528}{4609.77} = 0.33$$

Correlation Table of Solution 25

Let $dx = \frac{X - 25}{10}$, $dy = \frac{Y - 350}{100}$

Y ↓ M.V.			X → M.V. / dy \ dx	10—20	20—30	30—40	40—50				
			M.V.	15	25	35	45				
				−10	0	+10	+20				
	M.V.		dx / dy	−1	0	+1	+2	f	fdy	fdy^2	fdxdy
100-200	150	−200	−2	2, 3, 6	—	—	—	3	−6	12	6
200-300	250	−100	−1	1, 2, 2	0, 3, 0	−1, 2, −2	—	7	−7	7	0
300-400	350	0	0	—	0, 4, 0	0, 1, 0	—	5	0	0	0
400-500	450	+100	+1	—	0, 2, 0	1, 5, 5	2, 1, 2	8	8	8	7
500-600	550	+200	+2	—	0, 1, 0	2, 1, 2	4, 5, 20	7	14	28	22
			f	5	10	9	6	N =30	Σ fdy =9	Σ fdy^2 =55	Σ fdxdy =35
			fdx	−5	0	9	12	Σ fdx =16			
			fdx^2	5	0	9	24	Σ fdx^2 =38			
			fdxdy	8	0	5	22	Σ fdxdy =35			

Correlation Table of Solution 26

Let $dx = \frac{X-30}{20}, dy = \frac{Y-32.50}{15}$

(In each cell: product $dx \cdot dy$ in the corner, frequency, and $fdxdy$.)

X ↓	M.V.		dx	Y → 10—25	25—40	40—55	f	fdx	fdx^2	fdxdy
			M.V.	17.5	32.50	47.5				
				−15	0	+15				
			dy	−1	0	+1				
0-20	10	−20	−1	(1) 10, 10	(0) 4, 0	(−1) 6, −6	20	−20	20	4
20-40	30	0	0	(0) 5, 0	(0) 40, 0	(0) 9, 0	54	0	0	0
40-60	50	+20	+1	(−1) 3, −3	(0) 8, 0	(1) 15, 15	26	26	26	12
			f	18	52	30	N = 100	Σ fdx =6	Σ fdx^2 =46	Σ fdxdy =16
			fdy	−18	0	30	Σ fdy =12			
			fdy^2	18	0	30	Σ fdy^2 =48			
			fdxdy	7	0	9	Σ fdxdy =16			

Correlation Table of Solution 27

Let $dx = X - 20, dy = \frac{Y-12.5}{5}$

Y ↓	M.V.		dy	X → M.V. 18	19	20	21	22	f	fdy	fdy^2	fdxdy
			dx	−2	−1	0	+1	+2				
0-5	2.5	−10	−2	—	—	—	(−2) 3, −6	(−4) 1, −4	4	−8	16	−10
5-10	7.5	−5	−1	—	—	—	(−1) 3, −3	(−2) 2, −4	5	−5	5	−7
10-15	12.5	0	0	—	—	(0) 7, 0	(0) 10, 0	—	17	0	0	0
15-20	17.5	+5	+1	—	(−1) 5, −5	(0) 4, 0	—	—	9	9	9	−5
20-25	22.5	+10	+2	(−4) 3, −12	(−2) 2, −4	—	—	—	5	10	20	−16
			f	3	7	11	16	3	N = 40	Σ fdy = 6	Σ fdy^2 = 50	Σ fdxdy = −38
			fdx	−6	−7	0	16	6	Σ fdx = 9			
			fdx^2	12	7	0	16	12	Σ fdx^2 = 47			
			fdxdy	−12	−9	0	−9	−8	Σ fdxdy = −38			

Example 27. Calculate Karl Pearson's coefficient of correlation from the following data:

Y / X	18	19	20	21	22	Total
0—5	—	—	—	3	1	4
5—10	—	—	—	3	2	5
10—15	—	—	7	10	—	17
15—20	—	5	4	—	—	9
20—25	3	2	—	—	—	5
Total	3	7	11	16	3	40

Solution: **(Landscape Table Given at Page 37)**

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2}\ \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{40(-38) - (6)(9)}{\sqrt{40(47) - (9)^2}\ \sqrt{40(50) - (6)^2}}$$

$$= \frac{-1574}{\sqrt{1799}\ \sqrt{1964}} = \frac{-1574}{42.41 \times 44.32} = \frac{-1574}{1879.61}$$

$$= -0.837'0 \approx -0.84.$$

It shows a high degree of negative correlation between X and Y.

EXERCISE 1.7

1. Calculate Karl Pearson's coefficient of correlation for the following distribution:

Y \ X	200—300	300—400	400—500	500—600	600—700
10—15	-	-	-	3	7
15—20	-	4	9	4	3
20—25	7	6	12	5	-
25—30	3	10	19	8	-

Also calculate its probable error. **[Ans.** $r = -0.438$, PE $= 0.0544$**]**

2. Calculate the coefficient of correlation between marks and age from the following data:

Marks \ Age	18	19	20	21
200—250	4	4	2	1
250—300	3	5	4	2
300—350	2	6	8	5
350—400	1	4	6	10

Can we conclude that increase in age causes increase in marks? **[Ans.** $r = 0.415$**]**

3. 24 pairs of X and Y are given below:

X	15	0	1	3	16	2	18	5
Y	13	1	2	7	8	9	12	9
X	4	17	6	19	14	9	8	13
Y	17	16	6	18	11	3	5	4
X	10	13	11	11	12	18	9	7
Y	10	11	14	7	18	15	15	3

Prepare a correlation table taking the magnitude of each class interval as four and the first interval as equal to 0 and less than 4. Calculate Karl Pearson's coefficient between X and Y.

[Ans. $r = 0.578$**]**

4. The frequency distribution of marks obtained in Physics and Chemistry by 100 students are given in the following table. Determine:

(*i*) Percentage of students passed in Physics and Chemistry, while for passing minimum 60% is required.

(*ii*) Coefficient of correlation.

Physics \ Chemistry	40—49	50—59	60—69	70—79	80—89	90—99	Total
90—99	—	—	—	2	4	4	10
80—89			1	4	6	5	16
70—79	—	—	5	10	8	1	24
60—69	1	4	9	5	2	—	21
50—59	3	6	6	2	—	—	17
40—49	3	5	4	—	—	—	12
Total	7	15	25	23	20	10	100

[Ans. (*i*) % of students in Physics = 71%,
% of students passed in Chemistry = 78%,
(*ii*) $r = 0.8056$**]**

Assumptions of Karl Pearson' Coefficient of Correlation

Karl Pearson's coefficient of correlation is based on the following assumptions:

(1) Affected by a Large Number of Independent Causes: Series or variables which are correlated, are affected by a large number of factors that result in a normal distribution.

(2) Cause and Effect Relation: There is a cause and effect relationship between the forces affecting the distribution of the items in the two series.

(3) Linear Relationship: Two variables are linearly related. Plotting the values of the variables in a scatter diagram yields a straight line.

Properties of the Coefficient of Correlation

The following are the important properties of the correlation coefficient (r):

(1) **Limits of Coefficient of Correlation:** Karl Pearson's coefficient of correlation lies between −1 and +1. Symbolically

$$-1 \le r \le +1$$

This implies r can never exceed +1 and never becomes less than −1. It always lies between −1 and +1.

(2) **Change of Origin and Scale:** Shifting the origin or scale does not affect in any way the value of correlation coefficient. coefficient of correlation is independent of the change of origin and scale. If the scale of a series is changed or the origin is shifted, then correlation coefficient remains unchanged.

(3) **Geometric Mean of Regression Coefficients:** Correlation coefficient is the geometric mean of the regression coefficients *byx* and *bxy*. Symbolically:

$$r = \sqrt{bxy \, . \, byx}$$

(4) If X and Y are independent variables, then coefficient of correlation is zero but the converse is not necessarily true. [For proof, See Example 55].

(5) **Pure Number:** 'r' is a pure number and is independent of the units of measurements. This implies that even if the two variables are expressed in two different units of measurements viz.; rainfall in inches, and yields of crops in quintals, the value of correlation coefficient comes out with a pure number. Thus, it does not require that the units of both the variables should be the same.

(6) **Symmetric:** The coefficient of correlation between the two variables x and y is symmetric i.e., $r_{xy} = r_{yx}$. It means that either we compute the value of correlation coefficient between x and y or between y and x, the coefficient of correlation remains the same.

Interpreting the Coefficient of Correlation

Coefficient of correlation measures the degree of relationship between two variables. It is denoted by 'r'. The value of correlation coefficient lies between −1 and +1. The value of correlation coefficient can be interpreted in the following ways:

(*i*) If $r = +1$, then there is perfect positive correlation.

(*ii*) If $r = 0$, then there is absence of linear correlation.

(*iii*) If $r = +0.25$, then there will be low degree of positive correlation.

(*iv*) If $r = +0.50$, then there is moderate degree of correlation.

(*v*) If $r = +0.75$, then there is high degree of positive correlation.

Similarly, negative values of r can be interpreted.

Probable Error and Karl Pearson's Coefficient of Correlation

To test the reliability of Karl Pearson's correlation coefficient, probable error is used. The following formula is used to determine probable error:

$$\textbf{Probable Error (P.E.)} = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

Where, r is the coefficient of correlation and N, the number of pairs of observations.

If the constant 0.6745 is omitted from the above formula of probable error, we get the **standard error of the coefficient of correlation.** Thus,

$$SE_r = \frac{1 - r^2}{\sqrt{N}}$$

Utility of Probable Error: (1) Probable error is used to interpret the value of the correlation coefficient. Interpretation of r with the help of probable error is made clear by the following points:

(*i*) If $|r| > 6$ P.E., then coefficient of correlation (r) is taken to be significant.

(*ii*) If $|r| < 6$ P.E., then coefficient of correlation (r) is taken to be insignificant. This means that, there is no evidence of the existence of correlation in both the series.

(2) Probable error also determines the upper and lower limits within which the correlation of a randomly selected sample from the same universe will fall. Symbolically,

Upper Limit $= r +$ P.E.$_r$, Lower Limit $= r -$ P.E.

Example 28. Find the Karl Pearson's coefficient of correlation from the following data:

X:	9	28	45	60	70	50
Y:	100	60	50	40	33	57

Also calculate probable error and point out whether the coefficient of correlation is significant or not.

Solution:

Calculation of Coefficient of Correlation

X	dx	dx^2	Y	dy	dy^2	$dxdy$
9	−36	1296	100	50	2500	−1800
28	−17	289	60	10	100	−170
45 = A	0	0	50 = A	0	0	0
60	15	225	40	−10	100	−150
70	25	625	33	−17	289	−425
50	5	25	57	7	49	35
$N = 6$	$\Sigma dx = -8$	$\Sigma dx^2 = 2460$		$\Sigma dy = 40$	$\Sigma dy^2 = 3038$	$\Sigma dxdy = -2510$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx \, . \, \Sigma dy}{\sqrt{N \times \Sigma dx^2 - (\Sigma dx)^2} \ \sqrt{N \times \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times (-2510) - (-8)(40)}{\sqrt{6 \times 2460 - (-8)^2} \ \sqrt{6 \times 3038 - (40)^2}}$$

$$= \frac{-15060+320}{\sqrt{14760-64}\ \sqrt{18228-1600}} = \frac{-14740}{\sqrt{14696}\ \sqrt{16628}}$$

$$= \frac{-14740}{121.227 \times 128.95} = \frac{-14740}{15632.221} = -0.94$$

Calculation of P.E.

$$\text{P.E.} = 0.6745 \times \frac{1-r^2}{\sqrt{N}} = 0.6745 \times \frac{1-(-0.94)^2}{\sqrt{6}}$$

$$= 0.6745 \times \frac{0.1164}{2.449} = 0.03205$$

Significance of r

$$\frac{|r|}{P.E.} = \frac{0.94}{0.03205} = 29.32$$

$\Rightarrow$ $|r| = 29.32$ P.E.

Since, $|r|$ is more than 6 times the P.E., so, correlation coefficient is highly significant.

Example 29. A student calculates the value of r as 0.7 when the value of n is 5 and concludes that r is highly significant. Is he correct?

Solution: We know that if the value of $r > 6$ P.E., then it is considered to be significant.

$$\text{P.E.} = 0.6745 = \frac{1-r^2}{\sqrt{N}}$$

$$= 0.6745 \times \frac{1-(0.7)^2}{\sqrt{5}} = 0.15$$

Now, $\frac{r}{P.E.} = \frac{0.7}{0.15} = 4.67 \Rightarrow r = 4.67\ P.E.$

Since r is less than six times the P.E., r is insignificant and the student is wrong in his calculation.

Example 30. Show by calculation which 'r' is more significant: (*i*) $r = 0.90$, P.E. $= 0.03$ (*ii*) $r = 0.70$, P.E. $= 0.02$.

Solution: r is most significant in that case in which it is the highest number of times the P.E. It is compared as below:

(*i*) $\frac{r}{P.E.} = \frac{0.90}{0.03} = 30$, so r is 30 times of P.E.

(*ii*) $\frac{r}{P.E.} = \frac{0.70}{0.02} = 35$, so r is 35 times of P.E.

It is clear from the above that coefficient of correlation is the most significant in case (*ii*).

EXERCISE 1.8

1. Find Karl Pearson's Coefficient of correlation from the following series of marks secured by 10 students in a class test in Mathematics and Statistics.

Maths (X):	45	70	65	30	90	40	50	75	85	60
Statistics (Y):	35	90	70	40	95	40	60	80	80	50

Also calculate probable error. Is the value of r significant or not?

[Ans. $r = 0.903$, P.E. $= 0.039$, Highly significant**]**

2. Calculate the coefficient of correlation between the heights of fathers and sons from the following:

Height of Fathers (inches):	65	66	67	68	69	70	71
Height of Sons (inches):	67	68	66	69	72	72	69

Also calculate its probable error. Is the value of r significant or not?

[Ans. $r = 0.668$, P.E. $= 0.141$, Not significant**]**

3. (a) Find r if $N = 100$, P.E. $= 0.05$ (b) Find N if P.E. $= 0.025$, $r = .80$ **[Ans.** (*a*) $r = 0.5086$ (b) $N = 94$**]**

4. Comment on the significance of r in the following situations:

   (*i*) $N = 25$, $r = 0.8$

   (*ii*) $N = 100$, P.E. $= 0.04$ **[Ans.** (*i*) P.E. $= 0.049$, significant (*ii*) $r = 0.63$, significant**]**

5. The correlation coefficient of a sample of 100 pairs of items was 0.92. Within what limits does it hold good for another sample taken from the same universe?

[Ans. PE $= 0.0103$, 0.92 ± 0.0103**]**

(ii) Spearman's Rank Correlation Method

This method of determining correlation was propounded by Prof. Spearman in 1904. By this method, correlation between qualitative data namely beauty, honesty, intelligence, etc., can be computed. Such types of variables can be assigned ranks but their quantitative measurement is not possible. Thus, rank correlation method is used in such cases. The following is the formula for the computation of rank correlation coefficient:

$$R = 1 - \frac{6\Sigma D^2}{N(N^2-1)} \quad \text{or} \quad 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Where, R = Rank coefficient of correlation, D = Difference between two ranks ($R_1 - R_2$),

N = Number of pair of observations.

The value of rank correlation coefficient always lies between -1 and $+1$.

Note: 1. The value of rank correlation coefficient will be equal to the value of Pearson's Coefficient of Correlation for the two characteristics taking the ranks as values of the variables, provided no rank value is repeated *i.e.* the rank values of all the variables are different.

2. The sum total of rank difference (*i.e.*, ΣD) is always equal to zero, *i.e.*, $\Sigma D = \Sigma(R_1 - R_2) = 0$. This serves as check on the calculation work.

This method can be studied in the following three different situations:

(1) When ranks are given

(2) When ranks are not given

(3) When equal or tied ranks.

► **(1) When ranks are given**

When ranks are given, the following procedure is adopted to find the rank correlation coefficient:

(*i*) Ranks difference is found out by deducting the ranks of Y series from the corresponding ranks of X series. This is denoted by D, *i.e.*, $D = R_1 - R_2$.

(*ii*) Squaring the rank differences and summing them up, we get ΣD^2.

(*iii*) Finally, the following formula is used:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

The following examples make the above said method clear:

Example 31. In a fancy-dress competition, two judges accorded the following ranks to eight participants:

Judge X:	8	7	6	3	2	1	5	4
Judge Y:	7	5	4	1	3	2	6	8

Calculate coefficient of rank correlation.

Solution:

Calculation of Rank Correlation Coefficient

Judge X R_1	Judge Y R_2	$D = R_1 - R_2$	D^2
8	7	+1	1
7	5	+2	4
6	4	+2	4
3	1	+2	4
2	3	−1	1
1	2	−1	1
5	6	−1	1
4	8	−4	16
$N = 8$		$\Sigma D = 0$	$\Sigma D^2 = 32$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 32}{8^3 - 8} = 1 - \frac{192}{504}$$

$$= 1 - 0.381 = 0.619$$

There is, thus, moderate degree of positive relationship between the two judgements.

Example 32. Two ladies were asked to rank 10 different types of lipsticks. The ranks given by them are given below:

Lipsticks:	A	B	C	D	E	F	G	H	I	J
Neelu:	1	6	3	9	5	2	7	10	8	4
Neena:	6	8	3	7	2	1	5	9	4	10

Calculate Spearman's rank correlation coefficient.

Solution:

Calculation of Rank Correlation Coefficient

R_1	R_2	$D = R_1 - R_2$	D^2
1	6	−5	25
6	8	−2	4
3	3	0	0
9	7	+2	4
5	2	+3	9
2	1	+1	1
7	5	+2	4
10	9	+1	1
8	4	+4	16
4	10	−6	36
$N = 10$		$\Sigma D = 0$	$\Sigma D^2 = 100$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 100}{10^3 - 10} = 1 - \frac{600}{990}$$

$$= 1 - \frac{60}{99} = 1 - 0.606 = 0.394$$

Example 33. Ten competitors in a beauty contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution: In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare the rank correlation coefficient between the judgements of

(*i*) 1st Judge and 2nd Judge

(*ii*) 2nd Judge and 3rd Judge

(*iii*) 1st Judge and 3rd Judge.

Calculation of Rank Correlation Coefficient

Rank by 1st Judge (R_1)	Rank by 2nd Judge (R_2)	Rank by 3rd Judge (R_3)	$(R_1 - R_2)^2$ D_{12}^2	$(R_2 - R_3)^2$ D_{23}^2	$(R_1 - R_3)^2$ D_{13}^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
$N = 10$	$N = 10$	$N = 10$	$\Sigma D_{12}^2 = 200$	$\Sigma D_{23}^2 = 214$	$\Sigma D_{13}^2 = 60$

Applying the formula,

$$R_{12} = 1 - \frac{6\Sigma D_{12}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6\Sigma D_{23}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = 1 - \frac{6\Sigma D_{13}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = +0.636$$

Since the coefficient of rank correlation is positive and maximum in the judgement of the first and third judges, we conclude that they have the nearest approach to common tastes in beauty.

(2) When ranks are not given

When we are given the actual data and not the ranks, the following procedure is adopted to find out rank correlation coefficient:

(i) First of all, ranks are assigned to the items of X and Y series on the basis of their size. The largest value is assigned rank first, second largest second rank and similarly other values

are ranked. Sometimes, the smallest value is assigned the highest rank *i.e.* in descending order of the values. However, the same order (*i.e.* ascending order or descending order) of assigning the ranks must be maintained in both the series.

(ii) Rank difference of both the series ($D = R_1 - R_2$) is found and squared up. The squared rank difference, thus obtained is summed upto get ΣD^2.

(iii) Finally, the following formula is used to obtain rank correlation coefficient:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

The following example gives clarity to the above said method and its procedure.

Example 34. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	15	17	14	13	11	12	16	18	10	9
Y:	18	12	4	6	7	9	3	10	2	5

Solution:

Calculation of Rank Correlation Coefficient

X	Rank R_1	Y	Rank R_2	$D = R_1 - R_2$	D^2
15	4	18	1	+3	9
17	2	12	2	0	0
14	5	4	8	–3	9
13	6	6	6	0	0
11	8	7	5	+3	9
12	7	9	4	+3	9
16	3	3	9	–6	36
18	1	10	3	–2	4
10	9	2	10	–1	1
9	10	5	7	+3	9
$N = 10$				$\Sigma D = 0$	$\Sigma D^2 = 86$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, $N = 10$, $\Sigma D^2 = 86$

$$R = 1 - \frac{6 \times 86}{10^3 - 10}$$

$$= 1 - \frac{516}{990} = 1 - 0.52 = 0.48$$

Thus, there is positive correlation between X and Y.

(3) When equal or tied ranks

When two or more items have equal values in a series, then in such case, items of equal values are assigned common ranks, which is average of the ranks. For example, when item 10 appears twice in a series and their rank turns out to be 7 and 8 respectively, then they should be assigned $\frac{7+8}{2} = 7.5$ rank. In such case, some modification has to be made in the formula. Here, the following formula is used to determine rank correlation coefficient:

$$R = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots]}{N^3 - N}$$

Here, m = Number of items of equal ranks.

The correction factor of $\frac{1}{12}(m^3 - m)$ is added to ΣD^2 for such number of times as the cases of equal ranks in the question.

Example 35. Calculate coefficient of rank correlation from the following data:

X	15	10	20	28	12	10	16	18
Y	16	14	10	12	11	15	18	12

Make corrections for tied ranks.

Solution:

Calculation of Coefficient of Rank Correlation

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
15	5	16	2	3	9.00
10	7.5	14	4	3.5	12.25
20	2	10	8	–6	36.00
28	1	12	5.5	–4.5	20.25
12	6	11	7	–1	1.00
10	7.5	15	3	4.5	20.25
16	4	18	1	3	9.00
18	3	12	5.5	–2.5	6.25
$N = 8$				$\Sigma D = 0$	$\Sigma D^2 = 114$

In this question, the cases of equal rank are two, one for X series and other for Y series. Hence $\frac{1}{12}(m^3 - m)$ would be added for two times in ΣD^2.

Here, number 10 is repeated twice in series X and number 12 is repeated twice in series Y. Therefore, in both X and Y, $m = 2$.

$$R = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N^3 - N}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{8^3 - 8}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(6) + \frac{1}{12}(6)]}{512 - 8} = 1 - \frac{6[114 + 0.5 + 0.5]}{504}$$

$$= 1 - \frac{6[115]}{504} = 1 - \frac{690}{504} = 1 - 1.369 = -0.369$$

Example 36. Calculate coefficient of correlation by means of ranking method from the following data:

X:	40	50	60	60	80	50	70	60
Y:	80	120	160	170	130	200	210	130

Solution:

Calculation of Rank Coefficient of Correlation

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
40	8	80	8	0	0
50	6.5	120	7	–0.5	0.25
60	4	160	4	0	0
60	4	170	3	1	1
80	1	130	5.5	–4.5	20.25
50	6.5	200	2	4.5	20.25
70	2	210	1	1	1
60	4	130	5.5	–1.5	2.25
$N = 8$				$\Sigma D = 0$	$\Sigma D^2 = 45.00$

In this question in X series, the values 60 and 50 are repeated thrice and twice. The average rank for the value 60 is 4 (3 + 4 + 5 ÷ 3) while for the value 50 it is 6.5 (6 + 7 ÷ 2). In both the cases, the correlation factor will be $\frac{1}{12}(3^3 - 3)$ and $\frac{1}{12}(2^3 - 2)$. In series Y, the 130 is repeated twice. The average rank for the value 130 is 5.5 (5 + 6 ÷ 2). In this case, correction factor will be $\frac{1}{12}(2^3 - 2)$.

Applying the formula,

$$R = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)]}{(N^3 - N)}$$

$\Sigma D^2 = 45$, $m_1 = 3$, $m_2 = 2$, $m_3 = 2$, $N = 8$

By subsituting values in the above formula, we get

$$R = 1 - \frac{6\left[45 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{8(8^2 - 1)}$$

$$= 1 - \frac{6(45 + 2 + 0.5 + 0.5)}{8(63)} = 1 - \frac{6(48)}{504} = 1 - \frac{288}{504}$$

$$= 1 - 0.571 = 0.429$$

IMPORTANT TYPICAL EXAMPLES

Example 37. The ranks of the same 8 students in tests in Mathematics and Statistics were as follows, the two numbers within brackets denoting the ranks of the same students in Mathematics and Statistics respectively:

(1, 4), (2, 2), (3, 1), (4, 6), (5, 8), (6, 3), (7, 5), (8, 7)

(*i*) Calculate the rank correlation for proficiencies of this group in Math's and Statistics.

(*ii*) What does the value of the coefficient obtained indicates?

(*iii*) If you have found out Karl Pearson's simple coefficient of correlation between the ranks of these 16 students. Would your results have been the same as obtained in (*i*) or any different?

Solution: (*i*) **Calculation of Rank Correlation Coefficient**

Ranks in Maths (R_1)	Ranks in Statistics (R_2)	$D = R_1 - R_2$	D^2
1	4	−3	9
2	2	0	0
3	1	+2	4
4	6	−2	4
5	8	−3	9
6	3	+3	9
7	5	+2	4
8	7	+1	1
$N = 8$			$\Sigma D^2 = 40$

Applying the formula

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 40}{8^3 - 8} = 1 - \frac{240}{504}$$

$$= \frac{504 - 240}{504} = \frac{264}{504} = +0.523$$

(*ii*) The value of rank correlation coefficient indicates that there is moderate degree of positive correlation.

(*iii*) **Calculation of Karl Pearson's Coefficient of Correlation**

Ranks in Maths (X)	A = 4 dx	dx^2	Ranks in Statistics (Y)	A = 4 dy	dy^2	dx dy
1	−3	9	4	0	0	0
2	−2	4	2	−2	4	4
3	−1	1	1	−3	9	3
4 = A	0	0	6	+2	4	0
5	+1	1	8	+4	16	4
6	+2	4	3	−1	1	−2
7	+3	9	5	+1	1	3
8	+4	16	7	+3	9	12
$N = 8$	$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 4$	$\Sigma dy^2 = 44$	$\Sigma dxdy = 24$

Applying the formula

$$r = \frac{N.\Sigma dxdy - \Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 24 - (4)(4)}{\sqrt{8 \times 44 - (4)^2}\sqrt{8 \times 44 - (4)^2}}$$

$$= \frac{192 - 16}{\sqrt{336}\sqrt{336}} = \frac{176}{336} = 0.523$$

It is evident that the value of correlation coefficient computed by using Karl Pearson is the same as obtained by rank correlation method. The reason is that when the ranks of the students are not repeated, then the two methods give the same answer.

Example 38. Calculate rank correlation coefficient from the following data:

Serial No.:	1	2	3	4	5	6	7	8	9	10
Rank Difference:	−2	?	−1	+3	+2	0	−4	+3	+3	−2

Solution: The total of rank differences (ΣD) is always equal to zero and on this base the missing rank difference will be calculated. Let the missing item be 'a'.

As $\Sigma D = 0 \Rightarrow -2 + a - 1 + 3 + 2 + 0 - 4 + 3 + 3 - 2 = 0$

$\therefore\ a = -2$

Calculation of Coefficient of Rank Correlation

Sr. No.	Rank Difference D	D^2
1	−2	4
2	−2	4
3	−1	1
4	+3	9
5	+2	4
6	0	0
7	−4	16
8	+3	9
9	+3	9
10	−2	4
$N = 10$	$\Sigma D = 0$	$\Sigma D^2 = 60$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 60}{10^3 - 10}$$

$$= 1 - \frac{360}{990} = 1 - 0.364 = +0.636$$

Example 39. The coefficient of rank correlation of marks obtained by 10 students in English and Mathematics was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution: Given, R = 0.5, $N = 10$, Incorrect difference of ranks (D) = 3

Correct difference of ranks (D) = 7

We know that:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.5 = 1 - \frac{6\Sigma D^2}{10^3 - 10}$$

$$0.5 = 1 - \frac{6\Sigma D^2}{990}$$

$$\frac{6\Sigma D^2}{990} = 1 - 0.5 = 0.5$$

$\Rightarrow$ Incorrected $\Sigma D^2 = 82.5$

Corrected $\Sigma D^2 = 82.5 - (\text{Incorrect value})^2 + (\text{Correct value})^2$

$= 82.5 - 3^2 + 7^2 = 122.5$

Corrected Coefficient of Rank Correlation (R) $= 1 - \frac{6\Sigma D^2}{N^3 - N}$

$$= 1 - \frac{6 \times 122.5}{10^3 - 10} = 1 - \frac{735}{990} = 0.258$$

Thus, the correct value of rank correlation coefficient is 0.258.

Example 40. The rank correlation coefficient between marks obtained by some students in 'Statistics' and 'Accountancy' is found to be 0.8. If the total of squares of rank differences is 33, find the number of students.

Solution: Given, R = 0.8, $\Sigma D^2 = 33$

Now, $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.8 = 1 - \frac{6 \times 33}{N^3 - N}$$

$$\frac{198}{N^3 - N} = 1 - 0.8 = 0.2$$

$\Rightarrow$ $$N^3 - N = \frac{198}{0.2} = 990$$

$N(N^2 - 1) = 990$ $\quad [\because a^2 - b^2 = (a+b)(a-b)]$

$N(N+1)(N-1) = 990$

$(N-1)(N)(N+1) = 9 \times 10 \times 11$ $\quad$ Comparing both sides, we get:

$\therefore$ $N - 1 = 9$

$\Rightarrow$ $N = 10$

Example 41. The rank correlation coefficient between marks obtained by 10 students in Mathematics and Economics was found to be 0.5. Find the sum of squares of differences of ranks.

Solution: Given, $R = 0.5$, $N = 10$

Now, $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$\frac{6\Sigma D^2}{N^3 - N} = 1 - R$$

$$\frac{6\Sigma D^2}{10^3 - 10} = 1 - 0.5 = 0.5$$

$$6\Sigma D^2 = 0.5 \times 990$$

$\Rightarrow$ $$\Sigma D^2 = \frac{0.5 \times 990}{6} = 82.5$$

▶ Merits and Demerits of Rank Correlation Method

Merits

(1) This method is simple to understand and easy to apply as compared to Karl Pearson's method.

(2) When the data are of qualitative nature like beauty, honesty, intelligence, etc., this is the only method to be employed.

(3) When we are given the ranks and not the actual data, this method can be usefully employed.

Demerits

(1) This method cannot be used for finding correlation in a grouped frequency distribution.

(2) When the number of items exceed 30, the calculations become quite tedious and require a lot of time.

EXERCISE 1.9

1. Ten commerce graduates appeared before a selection board consisting of two members X and Y for the post of probationary officer in a certain bank. If the rank order of each of two members is given below, find out the coefficient of rank correlation:

Rank order by X:	1	6	5	10	3	2	4	9	7	8
Rank order by Y:	3	5	8	4	7	10	2	1	6	9

[Ans. $R = -0.212$]

2. Ten competitors in an intelligence test are ranked by three judges in the following order:

Judge I:	9	3	7	5	1	6	2	4	10	8
Judge II:	9	1	10	4	3	8	5	2	7	6
Judge III:	6	3	8	7	2	4	1	5	9	10

Use the rank correlation coefficient to determine:

(*i*) Which pair of judges agree the most?

(*ii*) Which pair of judges disagree the most?

[Ans. $R_{12} = 0.71$, $R_{23} = 0.467$, $R_{13} = 0.842$
(*i*) Ist and IIIrd (*ii*) IInd and IIIrd]

3. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	75	88	95	70	60	80	81	50
Y:	120	130	130	115	110	140	142	100

[Ans. R = 0.7976]

4. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	46	56	39	45	54	58	36	40
Y:	30	60	40	50	70	70	30	50

[Ans. R = 0.75]

5. Find the rank correlation coefficient from the following marks awarded by the examiners in statistics:

R. Nos.:	1	2	3	4	5	6	7	8	9	10	11
Marks Awarded by Examiner A:	24	29	19	14	30	19	27	30	20	28	11
Marks Awarded by Examiner B:	37	35	16	26	23	27	19	20	16	11	21
Marks Awarded by Examiner C:	30	28	20	25	25	30	20	24	22	29	15

[Ans. $R_{AB} = -0.027$, $R_{BC} = 0.5272$, $R_{AC} = 0.26136$]

6. From the following data, calculate Spearman's coefficient of correlation:

X:	80	78	75	75	68	67	60	59
Y:	12	13	14	14	14	16	15	17

[Ans. R = – 0.928]

7. The ranks of the same 16 students in tests in Mathematics and Statistics were as follows, the two numbers within brackets denoting the ranks of the same students in Mathematics and Statistics respectively:
(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8), (10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

(*i*) Calculate the rank correlation for proficiencies of this group of Math's and Statistics.

(*ii*) What does the value of the coefficient obtained indicates?

(*iii*) If you have found out Karl Pearson's simple coefficient of correlation between the ranks of these 16 students would your results have been the same as obtained in (a) or any difference? [Ans. R = 0.8, r = 0.8]

8. From the following data, calculate Spearman's coefficient of correlation:

Sr. No.:	1	2	3	4	5	6	7	8	9	10
Rank differences:	–2	–4	–1	+3	+2	0	?	+3	+3	–2

[Ans. R = +0.636]

9. The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of squares of the difference in ranks is given to be 48, find the value of N. [Ans. N = 7]

(iii) CONCURRENT DEVIATION METHOD

Concurrent deviation method of determining the correlation is extremely simple method. In this method, correlation is determined on the basis of direction of the deviations. Under this method, taking into consideration the direction of deviations, they are assigned (+) or (–) or (0) signs. The following steps are taken to find out correlation in this method:

(1) Under this method, whatever the series X and Y are to be studied for correlation, each item of the series is compared with its preceding item. If the value is more than its preceding value, then its deviation is assigned (+) sign, if less than preceding value then (–) sign and if equal to the

preceeding value then (0) sign is assigned. After this, third item is compared with the second, fourth item is compared with the third and this process goes on till the deviations of all items in a series are worked out.

(2) The deviations of X and Y series (dx) and (dy) are multiplied to get dxdy. Product of similar signs will be positive (+) and opposite signs will be negative (–) like:

(+) (+) = +,
(–) (–) = +,
(0) (0) = +,
(–) (+) = –,
(+) (–) = –,
(0) (–) = –,
(–) (0) = –,
(0) (+) = –

(3) Summing the positive dxdy signs, their number is counted. This is known as the number of concurrent deviations. It is denoted by the sign 'C'. The deviations with minus signs are excluded from the computation. They are ignored. If all the deviations in a series have minus signs, then number of concurrent deviations will be zero *i.e.* C=0.

(4) Finally, the following formula is used for determining coefficient of concurrent deviations

$$r_c = \pm\sqrt{\pm\left(\frac{2C-n}{n}\right)}$$

Here, r_c =Coefficient of concurrent deviations;

C = Number of concurrent deviations or Number of positive signs obtained after multiplying *dx* with *dy*;

*n = Number of pairs of observations minus one = $N-1$.

Note: In this formula ± sign is used both inside and outside the radical sign. If the value of (2C–*n*) is positive, then (+) sign will be used both inside and outside the radical sign because in such case correlation will be positive. On the contrary, if (2C–*n*) has negative sign, then minus sign will be used both inside and outside the radical sign because correlation will be negative.

The value of coefficient of concurrent deviation always lies between –1 and +1.

The following examples make the procedure of concurrent deviation method clear.

Example 42. Find coefficient of concunent deviation from the following data:

X:	85	91	56	72	95	76	89	51	59	90
Y:	18.3	20.8	16.9	15.7	19.2	18.1	17.5	14.9	18.9	15.4

** Since there is no sign for the first value of X and Y, n is always taken to be one less than the actual number of observations.*

Solution:

X	Deviation signs (*dx*)	Y	Deviation signs (*dy*)	*dxdy*
85		18.3		
91	+	20.8	+	+
56	–	16.9	–	+
72	+	15.7	–	–
95	+	19.2	+	+
76	–	18.1	–	+
89	+	17.5	–	–
51	–	14.9	–	+
59	+	18.9	+	+
90	+	15.4	–	–
$n = (10-1) = 9$		$n = (10-1) = 9$		C = 6

Here, 2C– *n*, *i.e.*, 2 × 6 – 9 = 3 is positive, therefore we use positive (+) sign in the formula. Thus,

$$r_c = \pm\sqrt{\pm\frac{(2C-n)}{n}}$$

$$r_c = \pm\sqrt{\pm\frac{(2\times 6-9)}{9}} = +\sqrt{+\frac{3}{9}} = 0.577$$

Thus, there is positive correlation between X and Y.

Example 43. Compute the coefficient of correlation for the following data by the concurrent deviation method:

Year:	1971	1972	1973	1974	1975	1976	1977
Demand:	150	154	160	172	160	165	180
Price:	200	180	170	160	190	180	172

Solution: Denoting Demand and Prices by X and Y.

Year	Demand X	Deviation signs (*dx*)	Price Y	Deviation signs (*dy*)	*dxdy*
1971	150		200		
1972	154	+	180	–	–
1973	160	+	170	–	–
1074	172	+	160	–	–
1075	160	–	190	+	–
1976	165	+	180	–	–
1977	180	+	172	–	–
$n = (7-1) = 6$					C = 0

Here, $2C-n$, *i.e.*, $2\times 0-6=-6$ is negative, therefore we use negative (–) sign in the formula. Thus,

$$r_c=\pm\sqrt{\pm\frac{(2C-n)}{n}}$$

$$=\pm\sqrt{\pm\frac{(2\times 0-6)}{6}}=-\sqrt{-(-1)}=-1$$

There is perfect negative correlation between price and demand.

Example 44. Calculate coefficient of correlation by concurrent deviation method from the following data:

X	112	125	126	118	118	121	125	125	131	135
Y	106	102	102	104	98	96	97	97	95	90

Solution:

Calculation of Coefficient of Concurrent Deviation

X	Deviation signs (*dx*)	*Y*	Deviation signs (*dy*)	*dxdy*
112		106		
125	+	102	–	–
126	+	102	0	–
118	–	104	+	–
118	0	98	–	–
121	+	96	–	–
125	+	97	+	+
125	0	97	0	+
131	+	95	–	–
135	+	90	–	–
$n=10-1=9$				C=2

Here, $2C-n$, *i.e.*, $2\times 2-9=-5$ is negative, therefore we use negative (–) sign in the formula. Thus,

$$r_c=\pm\sqrt{\pm\frac{(2C-n)}{n}}\ ;\ C=2,\ n=9$$

$$=\pm\sqrt{\pm\frac{(2\times 2-9)}{9}}=-\sqrt{-\left(\frac{-5}{9}\right)}$$

$$=-\sqrt{0.5556}=-0.75$$

Thus, there is high degree of negative correlation between X and Y.

IMPORTANT TYPICAL EXAMPLE

Example 45. During the first 9 months of the financial year 1999-2000, the following changes in the price index of shares A and B were recorded as below. Calculate the coefficient of correlation by a suitable method:

Changes over the previous month

Month:	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Share A:	–4	–3	–4	0	+3	+4	+2	–3	+3
Share B:	+3	–3	–2	–4	–3	–4	0	–2	–3

Solution: In this question changes are given in comparison to preceding month and in such a case only concurrent deviation method may be used. The value of 'C' will be calculated on the basis of multiplication of signs only (values will be ignored)

Calculation of Coefficient of Correlation

Months	Share A	Deviation signs (*dx*)	Share B	Deviation signs (*dy*)	*dxdy*
April	–4	–	+3	+	–
May	–3	–	–3	–	+
June	–4	–	–2	–	+
July	0	0	–4	–	–
August	+3	+	–3	–	–
September	+4	+	–4	–	–
October	+2	+	0	0	–
November	–3	–	–2	–	+
December	+3	+	–3	–	–
	$n=9$				$C=3$

Applying the formula,

$$r_c=\pm\sqrt{\pm\frac{(2C-n)}{n}}$$

Here, $C=3$, $n=9$

$\therefore$

$$r_c=\pm\sqrt{\pm\frac{(2\times 3-9)}{9}}$$

$$=\pm\sqrt{\pm\frac{(-3)}{9}}=-\sqrt{-\frac{(-3)}{9}}=-\sqrt{0.33}=-0.574$$

Note: Generally, the value of '*n*' is written on the basis of N–1, but in the above example, it will not be applicable because deviation sign of first item is also known.

▶ Merits and Demerits of Concurrent Deviation Method

Merits

(1) This method is simple to understand.
(2) Its computations involve less time.
(3) When the number of items is very large, we can use this method to have a quick idea about the correlation.
(4) This method is useful in studying short term fluctuations.

Demerits

(1) By applying this method, we can get an idea only about the direction of correlation.
(2) This method is not useful for finding correlation of long term changes.
(3) This method is less accurate than Karl Pearson's method.

EXERCISE 1.10

1. Calculate the coefficent of correlation by the method of concurrent deviation from the following data:

X:	65	50	35	55	60	25	45	80	85
Y:	45	35	55	40	70	30	40	65	80

[Ans. $r_c = 0.707$**]**

2. Calculate coefficient of concurrent deviation from the following data:

X:	65	40	35	75	63	80	35	20	80	60	50
Y:	60	55	50	56	30	70	40	35	80	75	80

[Ans. $r_c = +0.89$**]**

3. Find coefficient of correlation by concurrent deviation method of the following data:

Students:	A	B	C	D	E	F	G	H
Marks in Economics:	70	45	40	80	68	85	40	25
Marks in Statistics:	65	60	55	61	35	75	45	40

[Ans. $r_c = +1$**]**

4. Obtain a suitable measure of correlation from the following data regarding changes in price index of two shares A and B during the year:

Changes over the Previous Month

	J	F	M	A	M	J	J	A	S	O	N	D
Shares A:	+4	+3	+2	–1	–3	+4	–5	+1	+2	–7	+2	–3
Shares B:	–2	+5	+3	–2	–1	–3	+4	–1	+3	+6	+4	+2

[Ans. $r_c = -0.408$**]**

5. Find out the coefficient of correlation between X and Y by the method of concurrent deviation:

X:	26	30	30	24	29	25	25	32	32	38
Y:	62	58	55	68	67	64	64	75	81	78

[Ans. $r_c = -0.577$**]**

◘ COEFFICIENT OF DETERMINATION

The concept of coefficient of determination is used for the interpretation of coefficient of correlation and comparing the two or more correlation coefficients. **The coefficient of determination is defined as the square of the coefficient of correlation.** It is denoted by r^2. The coefficient of determination explains the percentage variation in the dependent variable Y that can be explained in terms of the independent variable X. If correlation coefficient (r) is 0.9 then coefficient of determination (r^2) will be 0.81 which implies that 81% of the total variations in the dependent variable (Y) occurs due to the independent variable (X). The remaining 19% variation occurs due to outside or external factors. Thus, the coefficient of determination is defined as the ratio of the explained variance to the total variance. In terms of formula:

$$\text{Coefficient of Determination } (r^2) = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

Coefficient of Non-Determination: By dividing the unexplained variation by the total variation, the coefficient of non-determination can be determined. Assuming the total of variation as 1, then the coefficient of determination can be determined by subtracting the coefficient of determination from 1. It is denoted by K^2. In terms of formula,

Coefficient of non-determination $(K^2) = 1 - r^2$

In the above example $r^2 = 0.81$, then the coefficient of non-determination will be 0.19 (1 – 0.81). It indicates that 19% of the variations are due to other factors.

Coefficient of Alientation $= \sqrt{1 - r^2}$

Generally, the coefficient of determination $(r)^2$ is widely used in practice.

Example 46. The coefficient of correlation (r) between consumption expenditure (C) and disposable income (Y) in a study was found to be +0.8. What percentage of variation in C are explained by variation in Y?

Solution: Here, $r = 0.8 \Rightarrow r^2 = (0.8)^2 = 0.64$. It means that 0.64 or 64% of the variation in consumption expenditure are explained by variation in income.

Example 47. Is it true that a correlation coefficient (r) = 0.8 indicates a relationship twice as close as $r = 0.4$?

Solution: The statement can be verified by using coefficient of determination, *i.e.*, r^2.

Now, **Ist case:** $r^2 = (0.8)^2 = 0.64$

2nd case: $r^2 = (0.4)^2 = 0.16$

This shows that 64% of the variation is explained in the first case and 16% of the variation is explained in the second case. Hence $r = 0.8$ does not indicate a relationship twice as close as $r = 0.4$.

Example 48. A correlation coefficient of 0.5 implies that 50% of the data are explained. Comment.

Solution: Coefficient of determination (r^2) show the percentage of variation in Y which are explained by the variation in X.

Now,

$$r^2 = (0.5)^2 = 0.25$$

Thus, the coefficient of correlation of 0.5 shows that 25% of the data are explained by X. In other words, 25% of the variation in Y is due to X and the remaining variation is due to other factors.

Example 49. The data relating to import price (X) and import quantity (Y) in respect of a given commodity are as under:

Year:	1975	'76	'77	'78	'79	'80	'81	'82	'83	'84
Import price :	2	3	6	5	4	3	5	7	8	7
Quantity imported:	6	5	4	5	7	10	9	7	8	9

(*i*) Calculate Karl Pearson's coefficient of correlation.

(*ii*) Find the percentage of variation in quantity imported that is explained by the variation in the import price.

Solution: (*i*)

Calculation of Coefficient of Correlation

X	$\bar{X}=5$ $X-\bar{X}$ x	x^2	Y	$\bar{Y}=7$ $Y-\bar{Y}$ y	y^2	xy
2	–3	9	6	–1	1	3
3	–2	4	5	–2	4	4
6	+1	1	4	–3	9	–3
5	0	0	5	–2	4	0
4	–1	1	7	0	0	0
3	–2	4	10	+3	9	–6
5	0	0	9	+2	4	0
7	+2	4	7	0	0	0
8	+3	9	8	+1	1	3
7	+2	4	9	+2	4	4
$N = 10$ $\Sigma X = 50$	$\Sigma x = 0$	$\Sigma x^2 = 36$	$\Sigma Y = 70$	$\Sigma y = 0$	$\Sigma y^2 = 36$	$\Sigma xy = 5$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{50}{10} = 5 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{70}{10} = 7$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{5}{\sqrt{36}\sqrt{36}} = \frac{5}{36} = 0.1389$$

(*ii*) Here, $r = 0.1389$

$\Rightarrow$ r^2 = coefficient of determination = $(0.1389)^2 = 0.0192$ or 1.92%

It means that 1.92% of the variations in quantity imported are explained by the variations in the import price.

EXERCISE 1.11

1. The relationship between consumption (C) and disposable income (Y) is expressed by $C = a + by$. In this context, explain what the value of r^2 measures.
2. "A correlation coefficient of 0.3 implies that 30% of the data are explained." Comment.
3. A correlation coefficient of 0.6 indicates a relationship twice as close to as where $r = 0.3$. Comment.

Quantity (Y) :	69	76	52	56	57	77	58	55	67	63	72	64
Price (X) :	9	12	6	10	9	10	7	8	12	6	11	8

(*i*) Calculate the Karl Pearson's coefficient of correlation between price and quantity.

(*ii*) Find the percentage of variation in quantity demanded that is explained by variation in the price of the commodity. **[Ans.** (*i*) $r = 0.645$, (*ii*) 42%**]**

4. Calculate from the given information:

X :	45	70	65	30	90	40	50	75	75	85	60
Y :	35	90	70	40	95	40	60	80	80	80	50

(*i*) Karl Pearson's coefficient of correlation.

(*ii*) Probable Error and show whether 'r' is significant or not?

(*iii*) Coefficient of non-determination and coefficient of alientation.

[Ans. (*i*) $r = 0.904$, (*ii*) P.E. = 0.0390, r is significant, (*iii*) $1 - r^2 = 0.183$, 0.4277**]**

MISCELLANEOUS SOLVED EXAMPLES

Example 50. (*i*) Find out the coefficient of correlation between X and Y from the following data:

X:	2	2	4	5	5
Y:	6	3	2	6	4

(*ii*) Multiply each X value by 2 and add 3. Multiply each value of Y by 5 and subtract 4. Find the correlation coefficient between two new sets of values. Explain why do or do not obtain the same result as in (*i*).

Solution: (*i*)

Calculation of Karl Pearson's Coefficient of Correlation

X	X^2	Y	Y^2	XY
2	4	6	36	12
2	4	3	9	6
4	16	2	4	8
5	25	6	36	30
5	25	4	16	20
$\Sigma X = 18$ $N = 5$	$\Sigma X^2 = 74$	$\Sigma Y = 21$	$\Sigma Y^2 = 101$	$\Sigma XY = 76$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5 \times 76 - 18 \times 21}{\sqrt{5 \times 74 - (18)^2}\sqrt{5 \times 101 - (21)^2}} = 0.036$$

(*ii*) **Let us define new variables U and V as follows:**

$$U = 2X + 3 \text{ and } V = 5Y - 4$$

We now calculate the coefficient of correlation between two new sets of values U and V as given

X	Y	$U = 2X + 3$	$V = 5Y - 4$	U^2	V^2	UV
2	6	7	26	49	676	182
2	3	7	11	49	121	77
4	2	11	6	121	36	66
5	6	13	26	169	676	338
5	4	13	16	169	256	208
		$\Sigma U = 51$	$\Sigma V = 85$	$\Sigma V^2 = 557$	$\Sigma V^2 = 1765$	$\Sigma UV = 871$

$$r = \frac{N.\Sigma UV - \Sigma U.\Sigma V}{\sqrt{N.\Sigma U^2 - (\Sigma U)^2}\ \sqrt{N.\Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{5 \times 871 - (51)(85)}{\sqrt{5 \times 557 - (51)^2}\sqrt{5 \times 1765 - (85)^2}}$$

$$= \frac{20}{\sqrt{184}\sqrt{1600}}$$

$$= 0.036$$

The value of r_{uv} is the same as that of r_{xy}. This is so because the correlation coefficient is independent of the change of origin and scale and U and V are obtained from X and Y by change of origin and scale so that we have r_{xy} and r_{uv}.

Example 51. Two variates X and Y when expressed as deviations from their respective means are given as follows:

x:	0	–4	4	–2	2
y:	1	3	?	0	–1

Find the Karl Pearson Coefficient of correlation between them.

Solution: In this question, one deviation in y series is missing. Let us denote the missing item by *a*. We know that the sum of deviations taken from mean is always zero.

So, $\Sigma y = 0$

$\therefore\ (1) + (3) + a + (0) + (-1) = 0$

$3 + a = 0$

$\therefore\ a = -3$

Thus the complete series is:

x:	0	–4	4	–2	2
y:	1	3	–3	0	–1

Now, we find the coefficient of correlation.

Calculation of Coefficient of Correlation

x	x^2	y	y^2	xy
0	0	1	1	0
–4	16	3	9	–12
4	16	–3	9	–12
–2	4	0	0	0
2	4	–1	1	–2
$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

Applying the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-26}{\sqrt{40 \times 20}}$$

$$= \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843} = -0.9192$$

Example 52. The following table gives the distribution of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality and its probable error. Is the value of 'r' significant or not?

Size group:	15—16	16—17	17—18	18—19	19—20	20—21
No. of items:	200	270	340	360	100	300
No. of defective items:	150	162	170	180	170	114

Solution: In this question, as correlation has to be found between size and defect in quality, hence defect in quality has to be first determined as % of the defective items.

Calculation of % of Defective Items

No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114
% of defective items:	$\frac{150}{200}\times100$ =75	$\frac{162}{270}\times100$ =60	$\frac{170}{340}\times100$ =50	$\frac{180}{360}\times100$ =50	$\frac{180}{400}\times100$ =45	$\frac{114}{300}\times100$ =38

Let us denote the mid value of the size group by X and % of defective items as Y

X (MV)	A=18.5 dx	dx^2	Y	A=50 dy	dy^2	dxdy
15.5	–3	9	75	25	625	–75
16.5	–2	4	60	10	100	–20
17.5	–1	1	50 = A	0	0	0
18.5 = A	0	0	50	0	0	0
19.5	+1	1	45	–5	25	–5
20.5	+2	4	38	–12	144	–24
$N=6$	$\Sigma dx=-3$	$\Sigma dx^2=19$		$\Sigma dy=18$	$\Sigma dy^2=894$	$\Sigma dxdy=-124$

Applying the formula,

$$r=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{6\times(-124)-(-3)(18)}{\sqrt{6\times19-(-3)^2}\sqrt{6\times894-(18)^2}}$$

$$=\frac{-744+54}{\sqrt{114-9}\sqrt{5364-324}}=\frac{-690}{\sqrt{105}\sqrt{5040}}=\frac{-690}{727.46}$$

$$=-0.948\approx-0.95$$

Probable Error (P.E.)

$$PE=0.6745\times\frac{1-r^2}{\sqrt{N}}$$

$$=0.6745\times\left(\frac{1-(-0.95)^2}{\sqrt{6}}\right)$$

$$=0.6745\times\frac{0.0975}{2.45}$$

$$=0.027$$

Significance of '*r*'

$$\frac{|r|}{P.E.}=\frac{0.95}{0.027}=35.18$$

$\therefore\ |r|=35.18\ P.E.$

As the value of $|r|$ is more than 6 times the P.E., so '*r*' is highly significant.

Example 53. Calculate correlation coefficient from the following results:

$N=10,\ \Sigma X=140,\ \Sigma Y=150$

$\Sigma(X-10)^2=180,\ \Sigma(Y-15)^2=215$

$\Sigma(X-10)(Y-15)=60$

Solution: For calculating correlation coefficient we need the values of $\Sigma X^2, \Sigma Y^2, \Sigma XY$ which we can determine from the values given:

$\Sigma(X-10)^2=\Sigma(X^2+100-20X)=\Sigma X^2+\Sigma100-20\Sigma X$

$=\Sigma X^2+N\times100-20\Sigma X$ [$\because\ \Sigma a=Na$]

$=\Sigma X^2+1000-20\times140$

$=\Sigma X^2+1000-2800=\Sigma X^2-1800$

$\Rightarrow\ \Sigma X^2-1800=180$ [$\because\ \Sigma(X-10)^2=180$]

$\therefore\ \Sigma X^2=1980$

$\Sigma(Y-15)^2=\Sigma(Y^2+225-30Y)=\Sigma Y^2+\Sigma225-30\Sigma Y$

$=\Sigma Y^2+N\times225-30\Sigma Y$ [$\because\ \Sigma a=Na$]

$=\Sigma Y^2+2250-30\times150$

$=\Sigma Y^2+2250-4500=\Sigma Y^2-2250$

$\Rightarrow\ \Sigma Y^2-2250=215$ [$\because\ \Sigma(Y-15)^2=215$]

$\therefore\ \Sigma Y^2=2465$

$\Sigma(X-10)(Y-15)=\Sigma(XY-15X-10Y+150)$

$=\Sigma XY-15\Sigma X-10\Sigma Y+\Sigma150$

$=\Sigma XY-15\Sigma X-10\Sigma Y+N\times150$

$=\Sigma XY-15\times140-10\times150+10\times150$

$=\Sigma XY-2100-1500+1500$

$=\Sigma XY-2100$

$\Rightarrow\ \Sigma XY-2100=60$ [$\because\ \Sigma(X-10)(Y-15)=60$]

$\therefore\ \Sigma XY=2160$

Applying the formula,

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

$$=\frac{10\times2160-140\times150}{\sqrt{10\times1980-(140)^2}\sqrt{10\times2465-(150)^2}}$$

$$=\frac{21600-21000}{\sqrt{19800-19600}\sqrt{24650-22500}}$$

$$=\frac{600}{\sqrt{200\times2150}}=\frac{600}{655.74}=+0.915$$

Aliter: r can also be calculated in the following manner:

$$\bar{X}=\frac{\Sigma X}{N}=\frac{140}{10}=14,\ \bar{Y}=\frac{\Sigma Y}{N}=\frac{150}{10}=15$$

Thus, deviations $\Sigma(X-10)$ and $\Sigma(Y-15)$ are from assumed means (A_x=10 and A_y=15)

Let, $\Sigma dx=\Sigma(X-10)=\Sigma X-\Sigma 10=\Sigma X-N.10=140-10\times 10=40$

$\Sigma dy=\Sigma(Y-15)=\Sigma Y-\Sigma 15=\Sigma Y-N.15=150-15\times 10=0$

$\Sigma\, dxdy=\Sigma(X-10)(Y-15)=60,\ \Sigma dx^2=180,\ \Sigma dy^2=215$ (Given)

Applying the formula,

$$r=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{10\times 60-40\times 0}{\sqrt{10\times 180-(40)^2}\sqrt{10\times 215-(0)^2}}$$

$$=\frac{600}{\sqrt{200\times 2150}}=\frac{600}{655.744}=0.915$$

Example 54. In two sets of variables X and Y with 50 items each, the following data were observed:

$\bar{X}=10,\ \sigma_x=3,\ \bar{Y}=6,\ \sigma_y=2,\ r=0.3,\ N=50$

However, on subsequent verification it was found that one value of X(=10) and one value of Y(= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of correlation coefficient affected?

Solution: Given: $N=50,\ \bar{X}=10,\ \bar{Y}=6,\ \sigma_x=3,\ \sigma_y=2,\ r=0.3$

$\therefore\quad \bar{X}=\frac{\Sigma X}{N}\Rightarrow \Sigma X=N\bar{X}$

$\therefore\quad$ Incorrected $\Sigma X=N\bar{X}=50\times 10=500$ ...(i)

$\therefore\quad \bar{Y}=\frac{\Sigma Y}{N}\Rightarrow \Sigma Y=N\bar{Y}$

$\therefore\quad$ Incorrected $\Sigma Y=N\bar{Y}=50\times 6=300$ ...(ii)

$\sigma_x^2=\frac{\Sigma X^2}{N}-(\bar{X})^2$

$\Rightarrow\quad \Sigma X^2=N(\sigma_x^2+\bar{X}^2)$ [Formula of variance of X]

Incorrected $\Sigma X^2=N(\sigma_x^2+\bar{X}^2)=50(9+100)=5450$ ...(iii)

$\sigma_y^2=\frac{\Sigma Y^2}{N}-(\bar{Y})^2$

$\Rightarrow\quad \Sigma Y^2=N(\sigma_y^2+\bar{Y}^2)$ [Formula of variance of Y]

Incorrected $\Sigma Y^2=N(\sigma_y^2+\bar{Y}^2)=50(4+36)=2000$ ...(iv)

We know,

$$r=\frac{Cov.(X,Y)}{\sigma_x.\sigma_y}$$

$\Rightarrow\quad r.\sigma_x.\sigma_y=Cov.(x,y)$

$\because\quad Cov.(X,Y)=\frac{1}{N}.\Sigma(X-\bar{X})(Y-\bar{Y})=\frac{1}{N}\Sigma XY-\bar{X}\bar{Y}$

$\therefore\quad r.\sigma_x.\sigma_y=\frac{1}{N}.\Sigma XY-\bar{X}.\bar{Y}$

$\therefore\quad \Sigma XY=N[r.\sigma_x.\sigma_y+\bar{X}\bar{Y}]$

$\Sigma XY=50[0.3\times 3\times 2+10\times 6]$

$=50\,[1.8+60]$

$=50\,[61.8]=3090$

Incorrected $\Sigma XY=3090$

Thus, we have the following incorrect values:

$\Sigma X=500,\ \Sigma Y=300,\ \Sigma X^2=5450,\ \Sigma Y^2=2000\ \Sigma XY=3090$

After dropping out the incorrect values, the corrected values for the remaining 49 pairs of items are given as:

Corrected values:

Corrected ΣX = 500 – 10 = 490

Corrected ΣY = 300 – 6 = 294

Corrected $\Sigma X^2=5450-10^2=5350$

Corrected $\Sigma Y^2=2000-6^2=1964$

Corrected $\Sigma XY=3090-10\times 6=3030$

$N=49$

Using these corrected values, we get

$$r=\frac{\Sigma XY-\frac{\Sigma X.\Sigma Y}{N}}{\sqrt{\Sigma X^2-\frac{(\Sigma X)^2}{N}}.\sqrt{\Sigma Y^2-\frac{(\Sigma Y)^2}{N}}}$$

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

$$=\frac{3030-2940}{\sqrt{450}\sqrt{200}}=\frac{90}{300}=+0.3$$

Hence the correlation coefficient is unaffected in this case.

Example 55. "If two variables are independent, the correlation between them is zero, but the converse is not always true." Comment.

Solution: If X and Y are two independent variables, then the covariance between them *i.e.* $Cov(X,Y)=0$ and hence $r_{xy}=\dfrac{Cov(X,Y)}{\sigma_x.\sigma_y}=0$. Thus, if X and Y are independent, they are uncorrelated.

The converse of this property implies that if $r_{xy}=0$, then X and Y may not necessarily be independent. To prove this property, let the two variables X and Y are connected by the relation $Y=X^2$ and consider the following data:

X	−3	−2	−1	0	1	2	3	ΣX = 0
Y	9	4	1	0	1	4	9	ΣY = 28
XY	−27	−8	−1	0	1	8	27	ΣXY = 0

Here, $\Sigma X=0, \Sigma Y=28$ and $\Sigma XY=0$

$$\therefore \quad Cov(X,Y)=\frac{1}{N}\Sigma XY-\frac{\Sigma X}{N}\cdot\frac{\Sigma Y}{N}=\frac{1}{7}\cdot(0)-\frac{0}{7}\cdot\frac{28}{7}=0$$

Thus,
$$r_{xy}=\frac{Cov(X,Y)}{\sigma_x.\sigma_y}=0$$

A close examination of the data would reveal that although $r_{xy}=0$ but X and Y are not independent. In fact, the variables are related by the equation $Y=X^2$, *i.e.*, there is a quadratic relation (*i.e.*, non-linear relationship) between the variables. This property implies that r_{xy} is only a measure of the linear relationship between X and Y. If the relationship is non-linear, the computed value of r_{xy} is no longer a measure of the degree of relationship between the two variables.

IMPORTANT FORMULAE

A. INDIVIDUAL SERIES

1. Karl Pearson's Coefficient of Correlation (When deviations are taken from actual mean):

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2\times\Sigma y^2}} \quad or \quad \frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

Where, $x=(X-\bar{X})$ $y=(Y-\bar{Y})$

2. When deviations are taken from assumed mean:

$$r=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

Where, $dx=(X-A)$ and $dy=(Y-A)$

3. When we use actual values of X and Y:

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

4. When we are given Variance and Covariance of X and Y:

$$r=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

where, $Cov(X,Y)=\frac{1}{N}.\Sigma(X-\bar{X})(Y-\bar{Y})=\frac{1}{N}.\Sigma XY-\bar{X}.\bar{Y}$

B. GROUPED SERIES

5. In a Bivariate or Grouped Frequency Distribution:

$$r=\frac{N\Sigma fdxdy-\Sigma fdx\Sigma fdy}{\sqrt{N.\Sigma fdx^2-(\Sigma fdx)^2}\sqrt{N.(\Sigma fdy)^2-(\Sigma fdy)^2}}$$

6. Spearmen's Rank Correlation Coefficient:

(*i*) When actual ranks are given:

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

(*ii*) When ranks are not repeated

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

(*iii*) When ranks are repeated

$$R=1-\frac{6\left[\Sigma D^2+\frac{1}{12}(m^3-m)+\frac{1}{12}(m^3-m)+...\right]}{N^3-N}$$

7. Concurrent Deviation Method

$$r_c=\pm\sqrt{\pm\left(\frac{2C-n}{n}\right)}$$

8. Probable Error and Standard Error

$$P.E.=0.6745\times\frac{1-r^2}{\sqrt{N}} \qquad S.E._r=\frac{1-r^2}{\sqrt{N}}$$

9. Coefficient of Determination

$$r^2=\frac{\text{Explained Variance}}{\text{Total Variance}}$$

QUESTIONS

1. Define correlation. Explain the various methods of studying correlation. What is the significance of studying correlation?
2. What is correlation? Explain various types of correlation. Does it always signify cause and effect relationship between the two variables?
3. Define Pearsons' coefficient of correlation. Interpret r when $r = 1, -1$ and 0.
4. Define rank correlation coefficient. How is it measured? When is it preferred to Karl Pearson's coefficient of correlation?
5. What is meant by coefficient of concurrent deviation? How is it measured?
6. What is scatter diagram and how is it useful in the study of correlation?
7. Explain the followings:
   (*i*) Probable Error (*ii*) Coefficient of Determination.
8. Explain the properties of correlation coefficient.

Linear Regression Analysis 2

INTRODUCTION

The study of regression has special importance in statistical analysis. We know that the mutual relationship between two series is measured with the help of correlation. Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.

MEANING AND DEFINITION

According to **Oxford English Dictionary**, the word 'regression' means "Stepping back" or "Returning to average value". The term was first of all used by a famous Biological Scientist in 19th century, **Sir Francis Galton** relating to a study of hereditary characteristics. He found out an interesting result by making a study of the height of about one thousand fathers and sons. His conclusion was that (*i*) Sons of tall fathers tend to be tall and sons of short fathers tend to be short in height (*ii*) But mean height of the tall fathers is greater than the mean height of the sons, whereas mean height of the short sons is greater than the mean height of the short fathers. The tendency of the entire mankind to **twin** back to average height, was termed by **Galton** 'Regression towards Mediocricity' and the line that shows such type of trend was named as 'Regression Line'.

In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, $D = f(P)$. Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.

DEFINITION OF REGRESSION

Some important definitions of regression are as follows:

1. Regression is the measure of the average relationship between two or more variables. **—M.M. Blair**
2. Regression analysis measures the nature and extent of the relation between two or more variables, thus enables us to make predictions. **—Hirsch**

In brief, regression is a statistical method of studying the nature of relationship between two variables and to make prediction.

UTILITY OF REGRESSION

The study of regression is very useful and important in statistical analysis, which is clear by the following points:

(1) Nature of Relationship: Regression analysis explains the nature of relationship between two variables.

(2) Estimation of Relationship: The mutual relationship between two or more variables can be measured easily by regression analysis.

(3) Prediction: By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.

(4) Useful in Economic and Business Research: Regression analysis is very useful in business and economic research. With the help of regression, business and economic policies can be formulated.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION

The main difference between correlation and regression is as follows:

(1) Degree and Nature of Relationship: Correlation is a measure of degree of relationship between X and Y whereas regression studies the nature of relationship between the variables so that one may be able to predict the value of one variable on the basis of another.

(2) Cause and Effect Relationship: Correlation does not always assume cause and effect relationship between two variables. Though two variables may be highly correlated, yet it does not necessary follow that one variable is the cause and another variable is the effect. But regression clearly expresses the cause and effect relationship between two variables. One variable is considered independent in regression, for which the value is given and other variable is dependent, which is estimated. The independent variable is the cause and the dependent variable is effect.

(3) Prediction: Correlation does not help in making prediction whereas regression enable us to make prediction. With the help of regression line of Y on X, the probable values of Y can be predicted on the basis of the values of X.

(4) Symmetric: In correlation analysis, correlation coefficient (r_{xy}) is the measure of direction and degree of linear relationship between the two variables X and Y. r_{xy} and r_{yx} are symmetrical, *i.e.*, $r_{xy} = r_{yx}$. This implies that it is immaterial which of X and Y is dependent variable and which is independent. In regression analysis, the regression coefficients b_{xy} and b_{yx} are not symmetric, *i.e.*, $b_{xy} \neq b_{yx}$. Thus, correlation coefficients r_{xy} and r_{yx} are symmetric whereas regression coefficients b_{yx} and b_{xy} are not symmetric.

(5) Non-sense Correlation: Sometimes, there may exist spurious or non-sense correlation between two variables by chance, like the correlation, if any between rise in income and rise in weight is a non-sense correlation but in regression analysis, there is nothing like non-sense regression.

(6) Origin and Scale: Correlation coefficient is independent of the change of origin and scale whereas regression coefficient is independent of change of origin but not of scale. This implies that if some common factor is taken out from X and Y variable, then no adjustment in correlation formula has to be made, whereas in case of regression, we have to make an adjustment for it in our formula.

TYPES OF REGRESSION ANALYSIS

The main types of regression analysis are as follows:

(1) Simple and Multiple Regression: In simple regression analysis, we study only two variables at a time, in which one variable is dependent and another is independent. The functional relationship between income and expenditure is an example of simple regression. On the contrary, we study more than two variables at a time in multiple regression analysis (*i.e.*, at least three variables) in which one is dependent variable and others are independent variable. The study of effect of rain and irrigation on yield of wheat is an example of multiple regression.

(2) Linear and Non-linear Regression: When one variable changes with other variable in some fixed ratio, this is called as linear regression. Such type of relationship is depicted on a graph by means of a straight line or a first degree equation. On the contrary, when one variable varies with other variable in a changing ratio, then it is referred to as curvi-linear/non-linear regression. This relationship, expressed on a graph paper takes the form of a curve. This is presented by way of 2nd or 3rd degree equation.

(3) Partial and Total Regression: When two or more variables are studied for functional relationship but at a time, relationship between only two variables is studied and other variables are held constant, then it is known as partial regression. On the other hand, in total regression all variables are studied simultaneously for the relationship among them.

SIMPLE LINEAR REGRESSION

In practice, simple linear regression is often used and under this, **Regression Lines, Regression Equations** and **Regression Coefficients** concepts are very important to be studied, which are as follows:

Regression Lines

The regression line shows the average relationship between two variables. This is also known as the **Line of Best Fit.** On the basis of regression line, we can predict the value of a dependent variable on the basis of the given value of the independent variable. If two variables X and Y are given, then there are two regression lines related to them which are as follows:

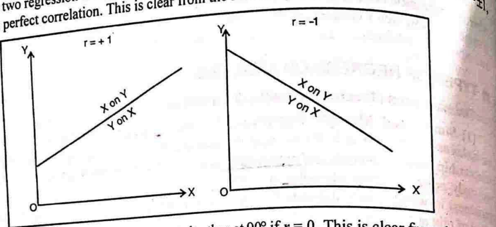
(1) Regression Line of X on Y: The regression line of X on Y gives the best estimate for the value of X for any given value of Y.

(2) Regression Line of Y on X: The regression line of Y on X gives the best estimate for the value of Y for any given value of X.

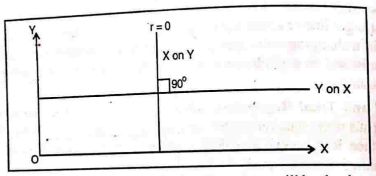
Nature of Regression Lines (or Relation between Correlation and Regression)

With the help of the direction and magnitude of correlation, the nature of regression lines can be known. The main points regarding the relationship among them are as follows:

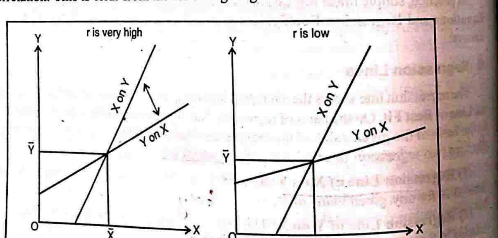
(1) The two regression lines are coincident or there will be only one regression line if $r = \pm 1$, *i.e.*, there is perfect correlation. This is clear from the following diagrams:



(2) The two regression lines intersect each other at 90° if $r = 0$. This is clear from the diagram given below:

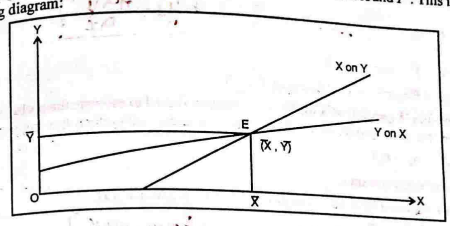


(3) The nearer the regression lines are to each other, the greater will be the degree of correlation. On the contrary, the greater the distance between the two regression lines, the lesser will be the degree of correlation. This is clear from the following diagrams:



(4) If regression lines rise from left to right upward, then correlation is positive. On the other side, if these line move from right to left, then correlation is negative.

(5) The regression lines cut each other at the point of intersection of $\bar{X}$ and $\bar{Y}$. This is clear from the following diagram:



Methods of Obtaining Regression Lines

(1) Scatter Diagram Method,

(2) Least Square Method.

(1) Scatter Diagram Method

This is the simplest method of constructing regression lines. In this method, values of the related variables are plotted on a graph. A straight line is drawn passing through the plotted points. The straight line is drawn with freehand. This shape of regression line can be linear or non-linear also. This depends upon the location of plotted points. This method is very rarely used in practice because in this method, the decision of the person who draws the regression lines very much affects the result.

(2) Least Square Method

Regression lines are also constructed by least square method. Under this method, a regression line is fitted through different points in such a way that the sum of squares of the deviations of the observed values from the fitted line shall be least. The line drawn by this method is called as the **Line of Best Fit.** In other words, under this method, the two regression lines, are drawn in such a way that sum of the squared deviations becomes minimum. The regression line of Y on X is so drawn such that vertically, the sum of squared deviations becomes minimum relating to the different points and the regression line on X on Y is so drawn such that horizontally, squared deviations of different points add up to the minimum.

Regression Equations

Regression equations are the algebraic formulation of regression lines. Regression equations represent regression lines. Just as there are two regression lines, similarly there are two regression equations, which are as follows:

(1) Regression Equation of Y on X: This equation is used to estimate the probable values of Y on the basis of the given values of X. This equation is expressed in the following way:

$$Y = a + bX$$

Here, a and b are constants.

Regression equation of Y on X can also be presented in another way as:

$$Y - \bar{Y} = r \cdot \frac{\sigma_y}{\sigma_x} \cdot (X - \bar{X})$$

or $Y - \bar{Y} = byx\,(X - \bar{X})$

Here, byx = Regression coefficient of Y on X.

(2) Regression Equation of X on Y: This equation is used to estimate the probable values of X on the basis of the given values of Y. This equation is expressed in the following way:

$$X = a_0 + b_0 Y$$

Here, a_0 and b_0 are constants.

Regression equation of X on Y can also be written in another way:

$$X - \bar{X} = r \cdot \frac{\sigma_x}{\sigma_y}\,(Y - \bar{Y})$$

or $X - \bar{X} = bxy\,(Y - \bar{Y})$

Here, bxy = Regression coefficient of X on Y.

Regression Coefficients

Just as there are two regression equations, similarly there are two regression coefficients. Regression coefficient measures the average change in the value of one variable for a unit change in the value of another variable. Regression coefficient, in fact, represents the slope of a regression line. For two variables X and Y, there are two regression coefficients, which are given as follows:

(1) Regression Coefficient of Y on X: This coefficient shows that with a unit change in the value of X variable, what will be the average change in the value of Y variable. This is represented by *byx*. Its formula is as follows:

$$byx = r \cdot \frac{\sigma_y}{\sigma_x}$$

The value of *byx* can also be determined by other formulae.

(2) Regression Coefficient of X on Y: This coefficient shows that with a unit change in the value of Y variable, what will be the average change in the value of X-variable. It is represented by *bxy*. Its formula is as follows:

$$bxy = r \cdot \frac{\sigma_x}{\sigma_y}$$

The value of *bxy* can also be found out by other formulae.

Properties of Regression Coefficients

The main properties of the regression coefficients are as follows:

(1) Coefficient of correlation is the geometric mean of the regression coefficients, *i.e.*,

$$r = \sqrt{bxy \times byx}$$

This property can be proved in the following manner:

Regression coefficient of X on Y $(bxy) = r \cdot \frac{\sigma_x}{\sigma_y}$ ...(*i*)

Regression coefficient of Y on X $(byx) = r \cdot \frac{\sigma_y}{\sigma_x}$ ...(*ii*)

Multiplying (*i*) and (*ii*)

$$bxy \cdot byx = r \cdot \frac{\sigma_x}{\sigma_y} \cdot r \cdot \frac{\sigma_y}{\sigma_x}$$

or $r^2 = bxy \cdot byx$

Hence, $r = \pm\sqrt{bxy \cdot byx}$

(2) Both the regression coefficients must have the same algebraic signs. The means either both regression coefficients will be either positive or negative. In other words, when one regression coefficient is negative, the other would be also negative. It is never possible that one regression coefficient is negative while the other is positive.

(3) The coefficient of correlation will have the same sign as that of regression coefficients. If both regression coefficient are negative, then the correlation coefficient would be negative. And if *byx* and *bxy* have positive signs, then *r* will also take plus sign.

(4) Both the regression coefficients cannot be greater than unity: If one regression coefficient of *y* on *x* is greater than unity, then the regression coefficient of *x* on *y* must be less than unity. This is because

$$r = \sqrt{byx \cdot bxy} = \pm 1$$

and never greater than one. If both the regression coefficients happen to be more than 1 then their geometric mean will exceed 1 which will not give the correlation coefficients whose value never exceeds 1.

(5) Arithmetic mean of two regression coefficients is either equal to or greater than the correlation coefficient. In terms of the formula:

$$\frac{byx + bxy}{2} \geq r$$

(6) Shift of origin does not affect regression coefficients but shift in scale does affect regression coefficients. Regression coefficients are independent of the change of origin but not of scale. This means if some common factor is taken out from the items of the series, then in that case, we will have to make adjustment in the regression coefficient formula which is shown below:

$$byx = bvu \cdot \frac{i_y}{i_x} \quad \text{and} \quad bxy = buv \cdot \frac{i_x}{i_y}$$

Where, $u = \frac{X - a}{h}$ and $v = \frac{Y - b}{k}$ and

i_y and i_x are common factors of *Y* and *X* series respectively.

To Obtain Regression Equations

The computation of regression equations can be divided into two parts:

(A) Regression Equations in case of Individual Series.

(B) Regression Equations in case of Grouped Data.

▶ **(A) Methods to Obtain Regression Equations in case of Individual Series**

In individual series, regression equations can be worked out by two methods, which are as follows:

(1) Regression Equations using Normal Equations.

(2) Regression Equations using Regression Coefficients.

▶ **(1) Regression Equations using Normal Equations**

This method is also called as **Least Square Method.** Under this method, computation of regression equations is done by solving out two normal equations. This method becomes clear by the following:

Regression Equation of Y on X

Regression Equation of Y on X is expressed as follows:

$$Y = a + bX$$

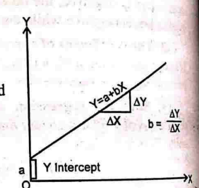
Where, Y = Dependent variable, X = Independent variable,

a = Y-intercept, b = Slope of the line.

Under least square method, the values of a and b are obtained by using the following two normal equations:

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$



Solving these equations, we get the following value of a and b.

$$b_{yx} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$$

$$a = \overline{Y} - b\overline{X}$$

Finally, the calculated value of a and b is put in the equation $Y = a + bX$. The regression equation of Y and X will be used to estimate the value of Y when the value of X is given.

Note: a is the Y-intercept, which indicates the minimum value of Y for $X = 0$ and b **is the slope of the line or called regression coefficient of Y and X,** which indicates the absolute increase in Y for a unit increase in X.

Regression Equation of X on Y

Regression Equation of X on Y is expressed as follows:

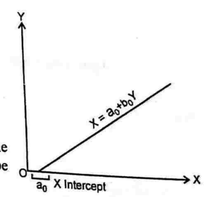
$$X = a_0 + b_0 Y$$

Where, X = Dependent variable, Y = Independent variable, a_0 = X-intercept, b_0 = Slope of the line.

Under least square method, the values of a_0 and b_0 are obtained by using the following two normal equations:

$$\Sigma X = Na_0 + b_0\Sigma Y$$

$$\Sigma XY = a_0\Sigma Y + b_0\Sigma Y^2$$



Solving these equations, we get the following value of a_0 and b_0.

$$b_0 = b_{xy} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2}$$

$$a_0 = \overline{X} - b_0\overline{Y}$$

Finally, the calculated value of a_0 and b_0 are put in the equation $X = a_0 + b_0 Y$. The regression equation of X on Y will be used to estimate the value of X when the value of Y is given.

Note: a_0 is the X-intercept, which indicates the minimum value of X for Y = 0 and b_0 **is the slope of the line or called regression coefficient of X on Y.**

The following examples makes the above said method more clear:

Example 1. Calculate the regression equation of X on Y from the following data by the method of least square:

X:	1	2	3	4	5
Y:	2	5	3	8	7

Solution:

Calculation of Regression Equation

X	X^2	Y	Y^2	XY
1	1	2	4	2
2	4	5	25	10
3	9	3	9	9
4	16	8	64	32
5	25	7	49	35
N = 5, ΣX = 15	ΣX^2 = 55	ΣY = 25	ΣY^2 = 151	ΣXY = 88

Regression Equation of X on Y is

$$X = a + bY$$

The two normal equations are:

$$\Sigma X = Na + b\,\Sigma Y$$

$$\Sigma XY = a\,\Sigma Y + b\,\Sigma Y^2$$

Substituting the values, we get

$$15 = 5a + 25b \quad ...(i)$$

$$88 = 25a + 151b \quad ...(ii)$$

Multiplying (*i*) by 5 and subtracting it from (*ii*)

$$88 = 25a + 151b$$

$$75 = 25a + 125b$$

$$13 = 26b$$

$$\therefore \quad b = \frac{13}{26} = 0.5$$

Putting the value of b in equation (*i*)

$$15 = 5a + 25 \times 0.5$$
$$15 = 5a + 12.5$$
$$5a = 2.5$$
$$a = 0.50$$

$$\therefore X = 0.5 + 0.5Y$$

Aliter: The value of a and b can also be obtained by using the following formula:

$$b_{xy} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \qquad a = \overline{X} - b\overline{Y}$$

Substituting the values, we get

$$b_{xy} = \frac{5 \times 88 - (15)(25)}{5 \times 151 - (25)^2} = \frac{440 - 375}{755 - 625} = \frac{65}{130} = \frac{1}{2} = 0.5$$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3, \overline{Y} = \frac{\Sigma Y}{N} = \frac{25}{5} = 5$$

$$\therefore \quad a = \overline{X} - b\overline{Y} = 3 - \frac{1}{2} \times 5 = 3 - 2.5 = 0.5$$

$$\therefore \quad X = 0.5 + 0.5Y$$

Example 2. Obtain the regression equation of Y on X by the least square method for the following data:

X:	1	2	3	4	5
Y:	9	9	10	12	11

Also estimate the value of Y when X = 10.

Solution:

Calculation of Regression Equation of Y on X

X	Y	XY	X^2
1	9	9	1
2	9	18	4
3	10	30	9
4	12	48	16
5	11	55	25
$N = 5$, $\Sigma X = 15$	$\Sigma Y = 51$	$\Sigma XY = 160$	$\Sigma X^2 = 55$

Regression Equation of Y on X is

$$Y = a + bX$$

The two normal equations are

$$\Sigma Y = Na + b\,\Sigma X$$
$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2$$

Substituting the values, we get

$$51 = 5a + 15b \quad ...(i)$$
$$160 = 15a + 55b \quad ...(ii)$$

Multiplying (*i*) by 3 and subtracting it from (*ii*)

$$160 = 15a + 55b$$
$$153 = 15a + 45b$$
$$- \quad - \quad -$$
$$7 = 10b$$

$$\therefore \quad b = \frac{7}{10} = 0.7$$

Putting the value of b in equation (*i*)

$$51 = 5a + 15(0.7) = 5a + 10.5$$
$$5a = 40.5$$
$$a = 8.1$$

Hence, the required regression equation of Y on X is given by

$$Y = 8.1 + 0.7X$$

Estimation for Y

For $X = 10$, $Y = 8.1 + 0.7(10) = 15.1$

Example 3. Given the following data:

$N = 8, \Sigma X = 21, \Sigma X^2 = 99, \Sigma Y = 4, \Sigma Y^2 = 68, \Sigma XY = 36$

Using the values, find

(*i*) Regression equation of Y on X.

(*ii*) Regression equation of X on Y.

(*iii*) Most approximate value of Y for X = 10

(*iv*) Most approximate value of X for Y = 2.5

Solution: **(*i*) Regression Equation of Y on X**

$$Y = a + bX$$

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{8 \times 36 - (21)(4)}{8 \times 99 - (21)^2} = 0.581$$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{21}{8} = 2.625, \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{4}{8} = 0.5$$

$$\therefore \quad a = \overline{Y} - b\overline{X} = 0.5 - (0.581)(2.625) = -1.025$$

$$\therefore \quad Y = -1.025 + 0.581X$$

(ii) Regression Equation of X on Y

$$X = a_0 + b_0 Y$$

$$b_0 = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{8 \times 36 - (21)(4)}{8 \times 68 - (4)^2} = 0.386$$

$$a_0 = \bar{X} - b_0\bar{Y} = 2.625 - (0.386)(0.5) = 2.432$$

$\therefore$ $X = 2.432 + 0.386Y$

(iii) Prediction for Y

When $X = 10, Y = -1.025 + 0.581(10) = 4.785$

(iv) Prediction for X

When $Y = 2.5, X = 2.432 + 0.386(2.5) = 3.397$

EXERCISE 2.1

1. Obtain the line of regression of Y on X by least square method for the following data:

X:	1	2	3	4	5
Y:	2	3	5	4	6

Also obtain an estimate of Y when $X = 2$. **[Ans.** $Y = 1.3 + 0.9X; 3.1$]

2. Find the regression of Y on X and X on Y by the least square method for the following data:

X:	1	2	3
Y:	2	4	5

Also find coefficient of correlation. **[Ans.** Y = 0.667 + 1.5X; X = –0.357 + 0.643Y; $r = 0.982$]

3. Compute the appropriate regression for the following data:

X (Independent variable):	1	3	4	8	9	11	14
Y (Dependent variable):	1	2	4	5	7	8	9

[Ans. $Y = 0.63X + 0.64$]

4. Obtain the two lines of regression from the following data:
$N = 3, \Sigma X = 6, \Sigma X^2 = 14, \Sigma Y = 15, \Sigma Y^2 = 77, \Sigma XY = 31$

[Ans. $Y = 0.5X + 4,\ X = 0.5Y - 0.5$]

5. Given: $\Sigma X = 15, \Sigma Y = 110, \Sigma XY = 400, \Sigma X^2 = 250, \Sigma Y^2 = 3200, N = 10$
Find the following:
(i) Regression coefficient of Y on X and the Y-intercept.
(ii) X-intercept, and the regression coefficient of X on Y.
(iii) Most approximate value of Y for X = 5.
(iv) Most approximate value of X for Y = 25.

[Ans. (i) $b = 1.033, a = 9.451$, (ii) $a = 0.201, b = 0.118$, (iii) $Y = 14.616, X = 3.151$]

(2) Regression Equations using Regression Coefficients

Regression equations can also be computed with the help of regression coefficients. For this, we will have to find out $\bar{X}, \bar{Y}$, byx and bxy from the given data. Regression equations can be computed from the regression coefficients by any of the following methods:

(1) Using the actual values of X and Y series.
(2) Using deviations from Actual Means.
(3) Using deviations from Assumed Means.
(4) Using r, σ_x, σ_y and $\bar{X}, \bar{Y}$.

(1) Using the Actual Values of X and Y Series

In this method, actual values of X and Y are used to determine regression equations. With regard to regression coefficients, regression equations are put in the following way:

Regression Equation of Y on X

$$Y - \bar{Y} = byx(X - \bar{X})$$

or $$Y = \bar{Y} + byx(X - \bar{X})$$

Here, $\bar{X}$ = Arithmetic mean of X series $= \frac{\Sigma X}{N}$

$\bar{Y}$ = Arithmetic mean of Y series $= \frac{\Sigma Y}{N}$

byx = Regression coefficient of Y on X

Using actual values, the value of byx can be calculated as:

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} \quad \text{or} \quad byx = \frac{\Sigma XY / N - \bar{X}.\bar{Y}}{\sigma_x^2}$$

Note: This formula is based on the normal equations, yet its use avoids the solution of normal equations.

Regression Equation of X on Y

$$X - \bar{X} = bxy(Y - \bar{Y})$$

or $$X = \bar{X} + bxy(Y - \bar{Y})$$

Where bxy = Regression coefficient of X on Y.

Using actual values, the value of bxy can be calculated as:

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \quad \text{or} \quad bxy = \frac{\Sigma XY / N - \bar{X}.\bar{Y}}{\sigma_y^2} = \frac{Cov(X_1Y)}{\sigma_y^2}$$

The following examples make this method more clear:

Example 4. Calculate the regression equations of X on Y and Y on X from the following data:

X:	1	2	3	4	5
Y:	2	5	3	8	7

Solution:

Calculation of Regression Equations

X	X^2	Y	Y^2	XY
1	1	2	4	2
2	4	5	25	10
3	9	3	9	9
4	16	8	64	32
5	25	7	49	35
N=5, ΣX=15	ΣX^2=55	ΣY=25	ΣY^2=151	ΣXY=88

$$\overline{X}=\frac{\Sigma X}{N}=\frac{15}{5}=3,\quad \overline{Y}=\frac{\Sigma Y}{N}=\frac{25}{5}=5$$

Regression Coefficient of Y on X (byx):

$$byx=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma X^2-(\Sigma X)^2}$$

$$=\frac{5\times 88-(15)(25)}{5\times 55-(15)^2}=\frac{440-375}{275-225}=\frac{65}{50}=1.3$$

Regression Coefficient of X on Y (bxy):

$$bxy=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma Y^2-(\Sigma Y)^2}$$

$$=\frac{5\times 88-(15)(25)}{5\times 151-(25)^2}=\frac{440-375}{755-625}=\frac{65}{130}=+0.5$$

Regression Equation of Y on X

$$Y-\overline{Y}=byx(X-\overline{X})$$

Substituting the values,

$$Y-5=1.3(X-3)$$
$$Y-5=1.3X-3.9$$
$$Y=1.3X-3.9+5$$
$$Y=1.3X+1.1$$

Regression Equation of X on Y

$$X-\overline{X}=bxy(Y-\overline{Y})$$
$$X-3=+0.5(Y-5)$$
$$X-3=0.5Y-2.5$$
$$X=0.5Y+0.5$$

Example 5. Calculate the two regression equations from the following data:

$$\Sigma X=30, \Sigma Y=23, \Sigma XY=168, \Sigma X^2=224, \Sigma Y^2=175, N=7$$

Hence or otherwise find Karl Pearson's coefficient of correlation.

Solution:

$$\overline{X}=\frac{\Sigma X}{N}=\frac{30}{7}=4.286$$

$$\overline{Y}=\frac{\Sigma Y}{N}=\frac{23}{7}=3.286$$

$$byx=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma X^2-(\Sigma X)^2}=\frac{7\times 168-(30)(23)}{7\times 224-(30)^2}=0.728$$

$$bxy=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma Y^2-(\Sigma Y)^2}=\frac{7\times 168-(30)(23)}{7\times 175-(23)^2}=0.698$$

Regression Equation of Y on X

$$Y-\overline{Y}=byx(X-\overline{X})$$
$$Y-3.286=0.728(X-4.286)$$
$$Y-3.286=0.728X-3.120$$
$$\therefore\quad Y=0.728X+0.166$$

Regression Equation of X on Y

$$X-\overline{X}=bxy(Y-\overline{Y})$$
$$X-4.286=0.698(Y-3.286)$$
$$X-4.286=0.698Y-2.294$$
$$\therefore\quad X=0.698Y+1.992$$

Karl Pearson's Coefficient of Correlation

$$r=\sqrt{byx\cdot bxy}$$
$$r=\sqrt{0.728\times 0.698}=0.712$$

IMPORTANT TYPICAL EXAMPLES

Example 6. In order to find the correlation coefficient between the two variables X and Y from 12 pairs of observations, the following calculations were made:

$$\Sigma X=30, \Sigma X^2=670, \Sigma Y=5, \Sigma Y^2=285, \Sigma XY=344$$

On subsequent verifications, it was discovered that the pair (X = 11, Y = 4) was copied wrongly, the correct values being (X = 10, Y = 14). After making necessary corrections, find:

(*i*) the two regression coefficients.

(*ii*) the two regresssion equations.

(*iii*) the correlation coefficient.

Solution:

Corrected ΣX = 30 + Correct value – Incorrect value

= 30 + 10 – 11 = 29

Corrected ΣY = 5 + 14 – 4 = 15

Corrected $\Sigma X^2 = 670 +$ (Correct value)^2–(Incorrect value)^2

$= 670 + 10^2 - 11^2 = 649$

Corrected $\Sigma Y^2 = 285 + 14^2 - 4^2 = 465$

Corrected $\Sigma XY = 344 + (10)(14) - (11)(4) = 440$

$$\bar{X} = \frac{\text{Corrected } \Sigma X}{N} = \frac{29}{12} = 2.416$$

$$\bar{Y} = \frac{\text{Corrected } \Sigma Y}{N} = \frac{15}{12} = 1.25$$

(i) Regression Coefficients

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{12 \times 440 - 29 \times 15}{12 \times 649 - (29)^2}$$

$$= \frac{5280 - 435}{7788 - 841} = \frac{4845}{6947} = +0.697$$

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{12 \times 440 - 29 \times 15}{12 \times 465 - (15)^2}$$

$$= \frac{5280 - 435}{5580 - 225} = \frac{4845}{5355} = 0.904$$

(ii) Two Regression Equations

X on Y	**Y on X**
$X - \bar{X} = bxy(Y - \bar{Y})$	$Y - \bar{Y} = byx(X - \bar{X})$
$X - 2.416 = 0.904(Y - 1.25)$	$Y - 1.25 = 0.697(X - 2.416)$
$X - 2.416 = 0.904Y - 1.13$	$Y - 1.25 = 0.697X - 1.683$
$X = 0.904Y + 1.286$	$Y = 0.697X - 0.433$

(iii) Correlation coefficient

$$r = \sqrt{byx.bxy} = \sqrt{(0.697)(0.904)} = +0.793$$

Example 7. Given that

$\bar{X} = 15, \bar{Y} = 12, \Sigma XY = 1500, \sigma_x = 6.4, \sigma_y = 9.0, N = 10$

Compute: (a) Two regression Coefficients

(b) Correlation coefficient between X and Y.

Solution: Given: $\bar{X} = 15, \bar{Y} = 12, \Sigma XY = 1500, \sigma_x = 6.4, \sigma_y = 9.0, N = 10$

Regression Coefficient of Y on X

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$$

The values of N and ΣXY are given and the values of $\Sigma X^2, \Sigma Y^2, \Sigma X$ and ΣY are to be calculated as follows:

$$\bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N.\bar{X} = 10 \times 15 = 150$$

$$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N.\bar{Y} = 10 \times 12 = 120$$

$$\Sigma X^2 = N[\sigma_x^2 + (\bar{X})^2] = 10[6.4^2 + 15^2] = 2659.6$$

$$\Sigma Y^2 = N[\sigma_y^2 + (\bar{Y})^2] = 10[9^2 + 12^2] = 2250$$

Now,

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{10 \times 1500 - (150)(120)}{10 \times 2659.6 - (150)^2}$$

$$= \frac{15000 - 18000}{26596 - 22500} = \frac{-3000}{4096} = -0.73$$

Aliter: byx can also be calculated as follows:

$$byx = \frac{\frac{\Sigma XY}{N} - \bar{X}.\bar{Y}}{\sigma_x^2} = \frac{\frac{1500}{10} - (15)(12)}{(6.4)^2}$$

$$= \frac{150 - 180}{40.96} = \frac{-30}{40.96} = -0.73$$

Regression Coefficient of X on Y

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{10 \times 1500 - (150)(120)}{10 \times 2250 - (120)^2}$$

$$= \frac{15000 - 18000}{22500 - 14400} = \frac{-3000}{8100} = -0.37$$

Aliter: bxy can also be calculated as follows:

$$bxy = \frac{\frac{\Sigma XY}{N} - \bar{X}.\bar{Y}}{\sigma_y^2} = \frac{\frac{1500}{10} - (15)(12)}{(9)^2}$$

$$= \frac{150 - 180}{81} = \frac{-30}{81} = -0.37$$

Coefficient of Correlation

$$r = \pm\sqrt{byx.bxy} = -\sqrt{(-0.73) \times (-0.37)} = -0.519$$

Example 8. Find out the regression coefficients of Y on X and X on Y from the following data:

$\Sigma X = 50, \bar{X} = 5, \Sigma Y = 60, \bar{Y} = 6, \Sigma XY = 350$, Variance of X = 4, Variance of Y = 9.

Solution: We know that: $\overline{X} = \frac{\Sigma X}{N} \Rightarrow 5 = \frac{50}{N} \Rightarrow N = 10$

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} \text{ or } \frac{\Sigma XY / N - \overline{X}.\overline{Y}}{\sigma_x^2} \text{ or } \frac{Cov(X,Y)}{\sigma_x^2}$$

$$\therefore \quad byx = \frac{\frac{350}{10} - (5)(6)}{4} = \frac{35-30}{4} = \frac{5}{4}$$

$$= 1.25$$

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \text{ or } \frac{\Sigma XY / N - \overline{X}.\overline{Y}}{\sigma_y^2} \text{ or } = \frac{Cov(X,Y)}{\sigma_y^2}$$

$$\therefore \quad bxy = \frac{\frac{350}{10} - (5)(6)}{9} = \frac{35-30}{9} = \frac{5}{9}$$

EXERCISE 2.2

1. Given the following bivariate data:

X:	–1	5	3	2	1	1	7	3
Y:	–6	1	0	0	1	2	1	5

   (*i*) Fit a regression line of Y on X and predict Y if X=10.

   (*ii*) Fit a regression line of X on Y and predict X if Y=2.5

   [**Ans.** $Y = -1.025 + 0.581X$; $X = 2.432 + 0.386Y$; $Y_{10} = 4.785$; $X_{2.5} = 3.397$]

2. By using the following data, find the regression equation of Y on X and compute the value of Y when X = 10.

   $\overline{X} = 5.5$, $\overline{Y} = 4.0$, $\Sigma X^2 = 385$, $\Sigma Y^2 = 192$, $\Sigma(X+Y)^2 = 947$ and $N = 10$

   [**Ans.** $Y = -0.42X + 6.31$, $Y_{10} = 2.11$]

3. Given that:

   $\Sigma X = 250, \Sigma Y = 300, \sigma_x = 5, \sigma_y = 10, \Sigma XY = 7900, N = 10$

   Compute : (*i*) Two regression coefficients,

   (*ii*) Correlation coefficient between X and Y.

   (*iii*) Most approximate value of Y when X = 55 and X when Y = 40.

   [**Ans.** $byx = 1.6$, $bxy = 0.4$, $r = 0.8$, $Y_{30} = 78$, $X_{40} = 29$]

4. By using the following data, find correlation coefficient and regression equation of Y on X and estimated value of Y when X = 20

   $N = 10, \Sigma X = 140, \Sigma Y = 150, \Sigma(X-10)^2 = 180, \Sigma(Y-15)^2 = 215, \Sigma(X-10)(Y-15) = 60$

   [**Hint:** See Example 53 on Correlation]

   [**Ans.** $r = 0.915$, $Y = 3X - 27$, $Y_{20} = 33$]

5. Following information was computed through a computer:

   $\Sigma X = 125, \Sigma Y = 100, \Sigma X^2 = 650, \Sigma Y^2 = 460, \Sigma XY = 508, N = 25$

   Later on it was discovered that two pairs of X and Y were miscopied as (6, 14) and (8, 6) instead of (8, 12) and (6, 8). Determine (*i*) the correct regression equations (*ii*) correct coefficient of correlation.

   [**Ans.** (*i*) $X = 0.556Y + 2.776$, $Y = 0.8X$, (*ii*) $r = 0.67$]

6. On each of 30 sets, two measurements are made. The following summaries are given :

   $\Sigma X = 15, \Sigma Y = -6, \Sigma XY = 56, \Sigma X^2 = 61$ and $\Sigma Y^2 = 90$

   Calculate the product moment correlation coefficient and the slope of regression line of Y on X.

   [**Hint:** See Example 52]

   [**Ans.** $r = 0.856$, $byx = 1.10$]

(2) Using Deviations taken from Actual Means

When the size of the values of X and Y is very large, then the method using actual values becomes very difficult to use. In such case, in place of actual values, deviations taken from arithmetic means ($\overline{X}$, $\overline{Y}$) are used to simplify the computation process. In such a case, regression equations are expressed as follows:

Regression Equation of Y on X

$$Y - \overline{Y} = byx\,(X - \overline{X})$$

or

$$Y = \overline{Y} + byx\,(X - \overline{X})$$

Here, $\overline{X}$ = Arithmetic mean of X

$\overline{Y}$ = Arithmetic mean of Y

byx = Regression coefficient of Y on X

Using deviations from actual means, the value of *byx* can be calculated as:

$$byx = \frac{\Sigma xy}{\Sigma x^2}$$

Where, $x = X - \overline{X}$; $y = Y - \overline{Y}$

Regression Equation of X on Y

$$X - \overline{X} = bxy\,(Y - \overline{Y})$$

or

$$X = \overline{X} + bxy(Y - \overline{Y})$$

Where, bxy = Regression coefficient of X on Y.

Using deviations from actual means, the value of *bxy* can be calculated as:

$$bxy = \frac{\Sigma xy}{\Sigma y^2}$$

Where, $x = X - \overline{X}$; $y = Y - \overline{Y}$

The following examples make this method more clear.

Example 9. Obtain the two regression equations from the following data:

X:	2	4	6	8	10	12
Y:	4	2	5	10	3	6

Solution:

X	$\bar{X}=7$ $(X-\bar{X})$ x	x^2	Y	$\bar{Y}=5$ $(Y-\bar{Y})$ y	y^2	xy
2	–5	25	4	–1	1	+5
4	–3	9	2	–3	9	+9
6	–1	1	5	0	0	0
8	+1	1	10	+5	25	+5
10	+3	9	3	–2	4	–6
12	+5	25	6	+1	1	+5
$\Sigma X=42$ N = 6	$\Sigma x=0$	$\Sigma x^2=70$	$\Sigma Y=30$	$\Sigma y=0$	$\Sigma y^2=40$	$\Sigma xy=18$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{42}{6}=7;\quad \bar{Y}=\frac{\Sigma Y}{N}=\frac{30}{6}=5$$

Since, the actual means of X and Y are whole numbers, we should taken deviations from $\bar{X}$ and $\bar{Y}$ to simplify calculations:

$$byx=\frac{\Sigma xy}{\Sigma x^2}=\frac{18}{70}=0.257$$

$$bxy=\frac{\Sigma xy}{\Sigma y^2}=\frac{18}{40}=0.45$$

Regression Equation of Y on X

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-5=0.257\,(X-7)$$
$$Y-5=0.257X-1.799$$
$$\therefore \quad Y=0.257X+3.201$$

Regression Equation of X on Y

$$X-\bar{X}=bxy\,(Y-\bar{Y})$$
$$X-7=0.45(Y-5)$$
$$X-7=0.45Y-2.25$$
$$X=0.45Y-2.25+7$$
$$\therefore \quad X=0.45Y+4.75$$

Example 10. The following are the intermediate results of the two series X and Y:
$\bar{X}=90, \bar{Y}=70, N=10, \Sigma x^2=6360, \Sigma y^2=2860, \Sigma xy=3900$
(Where x and y are deviations from the respective means)
Find two regression equations.

Solution:

Regression Coefficient of Y on X

$$byx=\frac{\Sigma xy}{\Sigma x^2}=\frac{3900}{6360}=0.613$$

Regression Coefficient of X on Y

$$bxy=\frac{\Sigma xy}{\Sigma y^2}=\frac{3900}{2860}=1.363$$

Regression Equation of X on Y

$$X-\bar{X}=bxy\,(Y-\bar{Y})$$
$$X-90=1.363\,(Y-70)$$
$$X-90=1.363Y-95.41$$
$$\therefore \quad X=1.363Y-5.41$$

Regression Equation of Y on X

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-70=0.613\,(X-90)$$
$$Y-70=0.613X-55.17$$
$$\therefore \quad Y=0.613X+14.83$$

Example 11. The following table gives the aptitude test scores and productivity indices of 10 workers at random:

Aptitude score	Productivity index
60	68
62	60
65	62
70	80
72	85
48	40
53	52
73	62
65	60
82	81

Estimate:

(*i*) the test score of a worker whose productivity index is 75.

(*ii*) the productivity index of a worker whose test score is 92.

Solution:

Calculation of Regression Equations

Aptitude Score X	($\bar{X}=65$) x	x^2	Productivity index Y	($\bar{Y}=65$) y	y^2	xy
60	–5	25	68	+ 3	9	– 15
62	–3	9	60	–5	25	+ 15
65	0	0	62	–3	9	0
70	+ 5	25	80	+ 15	225	+ 75
72	+ 7	49	85	+ 20	400	+ 140
48	–17	289	40	–25	625	+ 425
53	–12	144	52	–13	169	+ 156
73	+ 8	64	62	–3	9	– 24
65	0	0	60	–5	25	0
82	+ 17	289	81	+ 16	256	+ 272
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 894$	$\Sigma Y = 650$	$\Sigma y = 0$	$\Sigma y^2 = 1752$	$\Sigma xy = 1044$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65 : \bar{Y} = \frac{\Sigma Y}{N} = \frac{650}{10} = 65$$

Regression Equation of X on Y: $X - \bar{X} = bxy(Y - \bar{Y})$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{1044}{1752} = +0.596$$

$$X - 65 = 0.596\,(Y - 65)$$

$$X - 65 = 0.596\,Y - 38.74$$

or $$X = 26.26 + 0.596\,Y$$

For finding out the test score (X) of a person whose productivity index (Y) is 75, put Y = 75 in the above equation:

$$X_{75} = 26.26 + 0.596(75) = 26.26 + 44.7 = 70.96.$$

Regression Equation of Y on X: $Y - \bar{Y} = byx\,(X - \bar{X})$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{1044}{894} = +1.168$$

$$Y - 65 = 1.168\,(X - 65)$$

$$Y - 65 = 1.168\,X - 75.92 \text{ or } Y = -10.92 + 1.168\,X$$

For finding out the productivity index (Y) of a worker whose test score (X) is 92, we put X = 92 in the above equation.

$$Y_{92} = -10.92 + 1.168(92)$$

$$= -10.92 + 107.456 = 96.536$$

IMPORTANT TYPICAL EXAMPLES

Example 12. The following table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period:

Year:	1	2	3	4	5
Motor registration:	600	630	720	750	800
No. of tyres sold:	1,250	1,100	1,300	1,350	1,500

Find the regression equation to estimate the sale of tyres when motor registration is known. Estimate the sale of tyres when registration is 850.

Solution: Let X denotes number of motor registrations and Y denotes the number of tyres sold by a firm.

To simplify the calculation, let

$$x = \frac{X - \bar{X}}{i_x} \qquad y = \frac{Y - \bar{Y}}{i_y}$$

X	$x = \frac{X - \bar{X}}{10}$	x^2	Y	$y = \frac{Y - \bar{Y}}{50}$	y^2	xy
600	– 10	100	1,250	– 1	1	+ 10
630	– 7	49	1,100	– 4	16	+ 28
720	2	4	1,300	0	0	0
750	5	25	1,350	+ 1	1	+ 5
800	10	100	1,500	+ 4	16	+ 40
$\Sigma X = 3500$ $N = 5$	$\Sigma x = 0$	$\Sigma x^2 = 278$	$\Sigma Y = 6500$ $N = 5$	$\Sigma y = 0$	$\Sigma y^2 = 34$	$\Sigma xy = 83$

$$\bar{X} = \frac{3500}{5} = 700, \quad \bar{Y} = \frac{6500}{5} = 1300$$

Here, we have the regression of Y on X.

$$byx = \frac{\Sigma xy}{\Sigma x^2} \times \frac{i_y}{i_x} = \frac{83}{278} \times \frac{50}{10} = 1.4928$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 1300 = 1.4928\,(X - 700)$$

$$Y - 1300 = 1.4928X - 1044.96$$

$$Y = 1.4928X + 255.04$$

The estimate of sale of tyres (Y) when registration $X = 850$ is given by

$$Y = 1.4928 \times 850 + 255.04$$

$$= 1268.88 + 255.04 = 1523.92 \approx 1524$$

since the number of tyres cannot be fractional.

Example 13. Calculate the correlation coefficient from the following results:

$N = 10, \Sigma X = 350, \Sigma Y = 310$

$\Sigma(X-35)^2 = 162, \Sigma(Y-31)^2 = 222, \Sigma(X-35)(Y-31) = 92$

Also find the regression line of Y on X.

Solution:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{350}{10} = 35$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{310}{10} = 31$$

Thus, the given deviations $(X-35)$ and $(Y-31)$ are from actual means $(\bar{X} = 35, \bar{Y} = 31)$.

Thus, $\Sigma(X-35)^2 = 162$ or $\Sigma x^2 = 162$ where, $x = X - \bar{X}$

$\Sigma(Y-31)^2 = 222$, or $\Sigma y^2 = 222$ $\quad y = Y - \bar{Y}$

$\Sigma(X-35)(Y-31) = 92$ *or* $\Sigma xy = 92$

Coefficient of Correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{92}{\sqrt{162}\sqrt{222}} = +0.485$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx(X - \bar{X})$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{92}{162} = 0.568$$

$\therefore$

$$Y - 31 = 0.568(X - 35)$$

$$Y - 31 = 0.568X - 19.88$$

$$Y = 0.568X - 19.88 + 31$$

$$Y = 0.568X + 11.12$$

Graphing Regression Lines

It is quite easy to graph the regression lines once they have been computed. The procedure adopted is as follows:

(*i*) **Regression line of X on Y.** The regression line of X on Y can be drawn with the help of regression equation of X on Y, *i.e.,*

$$X = a + bY$$

If we put the respective values of Y in the above regression equation, we will find the estimated values of X. If we plot estimated values of X with the actual values of Y on the graph, we can draw regression line of X on Y.

(*ii*) **Regression line of Y on X.** The regression line of Y on X can be drawn with the help of regression equation of Y on X, *i.e.,*

$$Y = a + bX$$

If we put the respective values of X in the above equation, we will find the estimated values of Y. If we plot estimated values of Y with the actual values of X on the graph, we can draw regression line of Y on X.

The following example illustrate the graphing of regression lines.

Example 14. From the following data:

(*i*) Obtain the two regression equations.

(*ii*) Draw up the two regression lines on the graph paper with the help of two regression equations.

X:	1	2	3
Y:	5	4	6

Solution:

Calculation of Regression Equation

X	$\bar{X}=2$ x	x^2	Y	$\bar{Y}=5$ y	y^2	xy
1	–1	1	5	0	0	0
2	0	0	4	–1	1	0
3	+1	1	6	+1	1	+1
$\Sigma X = 6$ $N = 3$	$\Sigma x = 0$	$\Sigma x^2 = 2$	$\Sigma Y = 15$ $N = 3$	$\Sigma y = 0$	$\Sigma y^2 = 2$	$\Sigma xy = +1$

$$\therefore \quad \bar{X} = \frac{\Sigma X}{N} = \frac{6}{3} = 2; \qquad \bar{Y} = \frac{15}{3} = 5$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{1}{2} \qquad bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{1}{2}$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx(X - \bar{X})$$

$$Y - 5 = \frac{1}{2}(X - 2)$$

$$Y - 5 = \frac{1}{2}X - 1$$

$$\therefore \quad Y = \frac{1}{2}X + 4$$

Regression Equation of X on Y

$$X - \bar{X} = bxy(Y - \bar{Y})$$

$$X - 2 = \frac{1}{2}(Y - 5)$$

$$X - 2 = \frac{1}{2}Y - \frac{5}{2}$$

$$\therefore \quad X = \frac{1}{2}Y - \frac{1}{2}$$

(*ii*) **Regression Lines:** In order to draw up the two regression lines on the graph, we shall have to plot the given values of X and the computed values of Y and the given values of Y and the computed values of X

Computed Values of Y

Regression equation of Y on X

$$Y = \frac{1}{2}X + 4$$

Computed Values of X

Regression equation of X on Y

$$X = \frac{1}{2}Y - \frac{1}{2}$$

when $X=1,\ Y=\frac{1}{2}(1)+4=4.5$

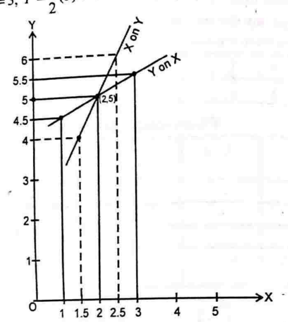
when $X=2,\ Y=\frac{1}{2}(2)+4=5.0$

when $X=3,\ Y=\frac{1}{2}(3)+4=5.5$

when $Y=5,\ X=\frac{1}{2}(5)-\frac{1}{2}=2$

when $Y=4,\ X=\frac{1}{2}(4)-\frac{1}{2}=1.5$

when $Y=6,\ X=\frac{1}{2}(6)-\frac{1}{2}=2.5$



Example 15. Compute the appropriate regression equation for the following data:

X (Independent variable)	Y (Dependent variable)
2	18
4	12
5	10
6	8
8	7
11	5

Solution: The appropriate regression equation will be Y on X

X	$\bar{X}=6$ x	x^2	Y	$\bar{Y}=10$ y	y^2	xy
2	−4	16	18	8	64	−32
4	−2	4	12	2	4	−4
5	−1	1	10	0	0	0
6	0	0	8	−2	4	0
8	+2	4	7	−3	9	−6
11	+5	25	5	−5	25	−25
$\Sigma X=36$	$\Sigma x=0$	$\Sigma x^2=50$	$\Sigma Y=60$	$\Sigma y=0$	$\Sigma y^2=106$	$\Sigma xy=-67$

$$\bar{X}=\frac{36}{6}=6;\qquad \bar{Y}=\frac{60}{6}=10$$

$$byx=\frac{\Sigma xy}{\Sigma x^2}$$

$$=\frac{-67}{50}=-1.34$$

Regression Equation of Y on X

$$Y-\bar{Y}=byx(X-\bar{X})$$

$$Y-10=-1.34(X-6)$$

$$\therefore\quad Y-10=-1.34X+8.04$$

$$Y=-1.34X+18.04$$

EXERCISE 2.3

1. For the following data, set up regression equation and estimate sales for an advertisement expenditure of Rs. 75 lakh.

Sales (Rs. crore):	14	16	18	20	24	30	32
Adv. expenditure (Rs. lakh):	52	62	65	70	76	80	78

[**Hint:** Let X denote sales) **[Ans. X = 0.621Y − 20.85, $X_{75,}$ = 25.725]**

2. Find the correlation coefficient and the equations of regression lines for the following values of X and Y:

X:	11	7	2	5	8	6	10
Y:	7	5	3	2	6	4	8

[Ans. r = 0.884, X = 0.75 + 1.25Y, Y = 0.625 + 0.625X]

3. The following data relate to marketing expenditure and the corresponding sales:

Expenditure (X) (Rs. lac):	10	12	15	20	23
Sales (Y) (Rs. crore) :	14	17	23	21	35

Estimate the marketing expenditure to obtain a sales target of Rs. 40 crore.

[Ans. $X=0.59Y+3.02;\ X_{40}=26.62$]

4. The following are the intermediate results of the two series X and Y
$\bar{X}=65, \bar{Y}=65, N=10, \Sigma x^2=894, \Sigma y^2=1752, \Sigma xy=1044$
(Where x and y are deviations from the respective means)
Find two regression equations. Also estimate Y when $X=92$ and X when $Y=75$.

[Ans. $Y=1.168X-10.92,\ Y_{92}=96.536;\ X=0.596Y+26.26,\ X_{75}=70.96$]

5. An investigation in to the demand for Television sets in 7 towns has resulted in the following data:

Population ('000) (X):	11	14	14	17	17	21	25
No. of T.V. sets demanded (Y):	15	27	27	30	34	38	46

Calculate the regression equation of Y on X and estimate the demand for T.V. sets for a town with a population of 30 thousand. **[Ans.** $Y = -3 + 2X$; $Y_{30} = 57$**]**

6. A departmental store gives in-service training to its salesmen which is followed by a test. It is considering whether it should terminate the service of any salesman who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period:

Test scores:	14	19	24	21	26	22	15	20	19
Sales ('00 Rs.):	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low test scores is justified ? If the firm wants a minimum sales volume of Rs. 3,000, what is the minimum test score that will ensure continuation of service ? Also estimate the most probable sales volume of a salesman making a score of 28.
[Hint : See Example 57] **[Ans.** $r = 0.9471$, justified, $X = 14.422 \approx 14$, $Y = 5286.64$**]**

7. The following table gives the marks in Economics and Statistics of 10 students selected at random:

Marks in Economics:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics:	43	46	49	41	36	32	31	30	33	39

Find (*i*) The two regression equations.
(*ii*) The coefficient of correlation between marks in Economics and Statistics.
(*iii*) The most likely marks in statistics when marks in economics are 30.
[Ans. (*i*) $X = -0.2337Y + 40.8806$, $Y = -0.6643X + 59.2576$
(*ii*) $r = -0.394$, (*iii*) 39.3286, or 39 marks**]**

8. The profits (Y) of a company in the Xth year of its life were as follows:

Years of life (X):	1	2	3	4	5
Profits (Y) (in lakh of Rs.):	1250	1400	1650	1950	2300

Estimate the profit of a company in the 6th year. **[Ans.** $Y = 265X + 915$, $Y_6 =$ Rs.2505 lakh**]**

9. From the following data:
(*i*) Obtain the two regression equations.
(*ii*) Draw up two regression lines on the graph paper.

X:	65	66	67	68	69	70	71
Y:	67	68	64	70	70	69	68

[Ans. $X = 0.462Y + 36.72$, $Y = 0.353X + 44$**]**

(3) Using Deviations taken from Assumed Means

When actual means turn out to be in fractions rather than the whole numbers like 24.69, 25.12 etc., then it becomes difficult to take deviations from actual means and squaring them up. To avoid such difficulty, deviations from assumed means rather than actual means are used. In such case, regression equations are expressed as follows:

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

Here, byx= Regression coefficient of Y on X.

Using deviations from assumed means, the value of byx can be calculated as:

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

or

$$byx = \frac{\Sigma dxdy - \dfrac{\Sigma dx \,.\, \Sigma dy}{N}}{\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}}$$

Where, $dx = X - A_x$, $dy = Y - A_y$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

Where, bxy= Regression coefficient of X on Y.

Using deviations from assumed means, the value of bxy can be calculated as:

$$bxy = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}$$

or

$$bxy = \frac{\Sigma dxdy - \dfrac{\Sigma dx \,.\, \Sigma dy}{N}}{\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}}$$

Where, $dx = X - A_x$, $dy = Y - A_y$

The following examples will clarify this method.

Example 16. Obtain the two regression equations for the following data:

X:	43	44	46	40	44	42	45	42	38	40	52	57
Y:	29	31	19	18	19	27	27	29	41	30	26	10

Also find the value of X when $Y = 49$ and Y when $X = 50$. Hence or otherwise find 'r'.

Solution:

Calculation of Regression Equations

X	A = 42 dx	dx^2	Y	A = 27 dy	dy^2	$dxdy$
43	1	1	29	2	4	2
44	2	4	31	4	16	8
46	4	16	19	–8	64	–32
40	–2	4	18	–9	81	18
44	2	4	19	–8	64	–16
42	0	0	27 = A	0	0	0
45	3	9	27	0	0	0
42 = A	0	0	29	2	4	0
38	–4	16	41	14	196	–56
40	–2	4	30	3	9	–6
52	10	100	26	–1	1	–10
57	15	225	10	–17	289	–255
$N = 12$ $\Sigma X = 533$	$\Sigma dx = 29$	$\Sigma dx^2 = 383$	$\Sigma Y = 306$	$\Sigma dy = -18$	$\Sigma dy^2 = 728$	$\Sigma dxdy = -347$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{533}{12} = 44.42, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{306}{12} = 25.5$$

Since the actual means of X and Y are in fractions, we should take deviations from assumed mean to simplify the calculations.

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{12 \times (-347) - (29)(-18)}{12 \times 383 - (29)^2} = \frac{-4164 + 522}{4596 - 841} = \frac{-3642}{3755}$$

$$= -0.969 = -0.97$$

$$bxy = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}$$

$$= \frac{12 \times (-347) - (29)(-18)}{12 \times 728 - (-18)^2} = \frac{-4164 + 522}{8736 - 324} = \frac{-3642}{8412}$$

$$= -0.432 = -0.43$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 44.42 = -0.43(Y - 25.5)$$
$$X - 44.42 = -0.43Y + 10.965$$
$$X = -0.43Y + 55.385$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 25.5 = -0.97\,(X - 44.42)$$
$$Y - 25.5 = -0.97X + 43.0874$$
$$Y = -0.97X + 68.5874$$

When Y = 49,

$$X = -0.43Y + 55.385$$
$$= -0.43\,(49) + 55.385$$
$$= -21.07 + 55.385$$
$$\therefore X_{49} = 34.315$$

When X = 50,

$$Y = -0.97(50) + 68.5874$$
$$= -48.5 + 68.5874$$
$$= 20.0874$$
$$\therefore Y_{50} = 20.0874$$

Coefficient of Correlation

$$r = \sqrt{byx.bxy}$$
$$= -\sqrt{(-0.97) \times (-0.43)} = -0.645$$

Example 17. Obtain the regression equation of Y on X from the following data:

X:	78	89	97	69	59	79	68	61
Y:	125	137	156	112	107	136	124	108

Solution:

Calculation of Regression Equations

X	A = 69 dx	dx^2	Y	A = 112 dy	dy^2	$dxdy$
78	+9	81	125	+13	169	+117
89	+20	400	137	+25	625	+500
97	+28	784	156	+44	1936	+1232
69 = A	0	0	112 = A	0	0	0
59	–10	100	107	–5	25	+50
79	+10	100	136	+24	576	+240
68	–1	1	124	+12	144	–12
61	–8	64	108	–4	16	+32
$N = 8$ $\Sigma X = 600$	$\Sigma dx = 48$	$\Sigma dx^2 = 1530$	$\Sigma Y = 1005$	$\Sigma dy = 109$	$\Sigma dy^2 = 3491$	$\Sigma dxdy = 2159$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{600}{8} = 75, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1005}{8} = 125.625$$

$$byx = \frac{N.\Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{8 \times 2159 - (48)(109)}{8 \times 1530 - (48)^2} = \frac{17272 - 5232}{12240 - 2304} = \frac{12040}{9936} = 1.212$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 125.625 = 1.212(X - 75)$$
$$Y - 125.625 = 1.212X - 90.9$$
$$Y = 1.212X + 34.725$$

IMPORTANT TYPICAL EXAMPLES

Example 18. A panel of judges A and B graded seven independently and awarded the following marks:

Debator:	1	2	3	4	5	6	7
Marks by A:	40	34	28	30	44	38	31
Marks by B:	32	39	26	30	38	34	28

An eight debator was awarded 36 marks by Judge A while Judge B was not present. If the Judge B was also present, how many marks would you expect him to award to eighth debator assuming degree of relationship exists in judgement?

Solution: Let marks awarded by Judge A be denoted by X and marks awarded by judge B be denoted by Y. The marks expected to be awarded by Judge B can be determined by fitting regression equations of Y on X.

Calculation of Regression Equations

X	A = 30 dx	dx^2	Y	A = 30 dy	dy^2	$dxdy$
40	+10	100	32	2	4	20
34	+4	16	39	9	81	36
28	−2	4	26	−4	16	8
30 = A	0	0	30 = A	0	0	0
44	14	196	38	8	64	112
38	8	64	34	4	16	32
31	1	1	28	−2	4	−2
$N = 7$ $\Sigma X = 245$	$\Sigma dx = 35$	$\Sigma dx^2 = 381$	$\Sigma Y = 227$	$\Sigma dy = 17$	$\Sigma dy^2 = 185$	$\Sigma dxdy = 206$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{245}{7} = 35, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{227}{7} = 32.43$$

$$byx = \frac{N.\Sigma dxdy - \Sigma dx.\Sigma dy}{N.\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{7 \times 206 - (35)(17)}{7 \times 381 - (35)^2}$$

$$= \frac{1442 - 595}{2667 - 1225} = \frac{847}{1442} = 0.587$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx(X - \bar{X})$$

$$Y - 32.43 = 0.587(X - 35)$$

$$Y - 32.43 = 0.587X - 20.545$$

$$Y = 0.587X + 11.885$$

For $X = 36$, Y shall be

$$Y = 0.587(36) + 11.885 = 21.132 + 11.885 = 33.017 \text{ or } 33 \text{ approx.}$$

Thus, if the Judge B was also present, he would have awarded 33 marks to the eighth debator.

Example 19. Simple observations obtained to study the relation between the measure of the waist and the length of the trousers are shown as under :

Measure of the Waist (in cm) :	70	72.5	75	77.5	80	82.5	85	87.5	90	92.5
Length of Trousers (in cm):	100	102	100	95	105	110	95	98	100	105

Obtain the line of best fit (regression) of length of trousers on measurement of the waist. Calculate the coefficient of determination.

Solution: Let X = measure of waist and Y = length of trousers.

Here, $N = 10$, $\Sigma X = 812.5$, $\Sigma Y = 1010$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{812.5}{10} = 81.25 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1010}{10} = 101$$

Since, $\bar{X}$ is not an integer, we will take the deviation of X from assumed value. Taking $dx = X - 80$ and $dy = Y - 101$.

The calculations are:

X	dx	dx^2	Y	dy	dy^2	$dxdy$
70	−10	100	100	−1	1	10
72.5	−7.5	56.25	102	+1	1	−7.5
75	−5	25	100	−1	1	+5
77.5	−2.5	6.25	95	−6	36	+15
80 = A	0	0	105	+4	16	0
82.5	25	6.25	110	+9	81	+22.5
85	5	25	95	−6	36	−30
87.5	7.5	56.25	98	−3	9	−22.5
90	10	100	100	−1	1	−10
92.5	12.5	156.25	105	+4	16	+50
$\Sigma X = 812.5$	$\Sigma dx = 12.5$	$\Sigma dx^2 = 531.25$	$\Sigma Y = 1010$	$\Sigma dy = 0$	$\Sigma dy^2 = 198$	$\Sigma dxdy = 32.5$

Regression Coefficient of Y on X:

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx.\Sigma dy}{N.\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{(10 \times 32.5) - (12.5 \times 0)}{(10 \times 531.25) - (12.5)^2} = \frac{325}{5312.5 - 156.25}$$

$$= \frac{325}{5156.25} = 0.06$$

Line of regression of length of trousers on the measurement of the waist, *i.e.*, the line of regression of Y on X is

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-101=0.06\,(X-81.25)$$
$$Y-101=0.06X-4.875$$
$$Y=0.06X+96.125$$

Coefficient of Determination:

$$r^2=\left[\frac{N\,\Sigma dxdy-(\Sigma dx).(\Sigma dy)}{\sqrt{N\,\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N\,\Sigma dy^2-(\Sigma dy)^2}}\right]^2$$

$$=\left[\frac{(10\times 32.5)-(125\times 0)}{\sqrt{10\times 531.25-(12.5)^2}\,\sqrt{10\times 198-0}}\right]^2$$

$$=\left[\frac{325}{\sqrt{5312.5-156.25}\times\sqrt{1980}}\right]^2=\left[\frac{325}{\sqrt{5156.25}\times\sqrt{1980}}\right]^2$$

$$=\frac{(325)^2}{5156.25\times 1980}=\frac{105625}{10209375}=0.01$$

Example 20. For a bivariate data, you are given the following information:

$\Sigma(X-58)=46 \quad \Sigma(X-58)^2=3086$

$\Sigma(Y-58)=9 \quad \Sigma(Y-58)^2=483$

$\Sigma(X-58)(Y-58)=1095$

$N=7$

(Assumed means of X and Y series are both 58)

You are required to determine (*i*) the two regression equations and (*ii*) the coefficient of correlation between X and Y series.

Solution: Since the assumed means of X and Y series are both 58, we have,

$\Sigma dx=46, \quad \Sigma dx^2=3086$

$\Sigma dy=9 \quad \Sigma dy^2=483$

$\Sigma dxdy=1095 \quad N=7$

$$byx=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{N.\Sigma dx^2-(\Sigma dx)^2}$$

$$=\frac{7\times 1095-(46)(9)}{7\times 3086-(46)^2}$$

$$=\frac{7665-414}{21602-2116}=\frac{7251}{19486}=0.37$$

$$bxy=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{N.\Sigma dy^2-(\Sigma dy)^2}$$

$$=\frac{7\times 1095-(46)(9)}{7\times 483-(9)^2}=\frac{7251}{3300}=2.20$$

Further, $\bar{X}=A+\frac{\Sigma dx}{N}=58+\frac{46}{7}=64.57$

$$\bar{Y}=A+\frac{\Sigma dy}{N}=58+\frac{9}{7}=59.29$$

Regression Equation of X on Y:

$$X-\bar{X}=bxy\,(Y-\bar{Y})$$
$$X-64.57=2.20\,(Y-59.29)$$
$$X-64.57=2.20Y-130.44$$
$$X=2.20Y-130.44+64.57$$
$$X=2.20Y-65.87$$

Regression Equation of Y on X:

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-59.29=0.37(X-64.57)$$
$$Y-59.29=0.37X-23.891$$
$$Y=0.37X-23.891+59.29$$
$$Y=0.37X+35.399$$
$$r=\sqrt{bxy\cdot byx}$$
$$r=\sqrt{2.20\times 0.37}=0.902$$

EXERCISE 2.4

1. Obtain the two regression equations for the following data:

X:	8	6	4	7	5	3
Y:	9	8	5	6	2	6

Also find the coefficient of correlation from the regression coefficients.

[Ans. $X=3.1+0.4Y;\ Y=2.23+0.685X;\ r=0.523$**]**

2. Obtain the two regression equations from the following data:

Age of husband (X):	18	19	20	21	22	23	24	25	26	27
Age of wife (Y):	17	17	18	18	18	19	19	20	21	21

Also find the coefficient of correlation from the regression coefficients.

[Ans. $Y=0.47X+8.225,\ X=1.99Y-14.9,\ r=+0.967$**]**

3. The height of fathers and sons in inches are:

Height of Fathers:	65	66	68	69	71	73	67	68	70	72	69
Height of Sons:	67	68	64	72	70	69	70	68	68	73	65

Estimate (*i*) the height of son if the height of the father is 64 inches, and (*ii*) the height of father if the height of son is 71.
Also calculate the value of Spearman's coefficient of correlation between them.
[Ans. (*i*) 66.18, (*ii*) 69.2, (*iii*) R= 0.4636]

4. The age and blood pressure of 10 university teachers are:

Age :	56	42	36	47	49	42	60	72	63	55
Blood Pressure:	147	125	118	128	145	140	155	160	149	150

(*i*) Find the correlation coefficient between age and blood pressure.
(*ii*) Determine the least square regression equation of blood pressure on age.
(*iii*) Estimate the blood pressure of a teacher whose age is 45 years.
[Hint: See Example 51] **[Ans.** $r = 0.89$, $Y = 1.11X + 83.758$, $Y_{45} = 133.708 \approx 134$]

5. The following table gives age (X) in years of cars and annual maintenance cost (Y) in hundred rupees:

X:	1	3	5	7	9
Y:	15	18	21	23	22

Estimate the maintenance cost for a 4 year old car after finding the regression equation.
[Ans. $Y = 0.95X + 15.05$; $Y_4 = 18.85$]

6. Obtain the two regression equations from the following data:

X:	4	5	6	8	11
Y:	12	10	8	7	5

Verify that the coefficient of correlation is the geometric mean of the two regression coefficients.
[Hint: See Example-50]
[Ans. $X = 15.024 - 0.979Y$; $Y = -0.929X + 14.717$; $r = -0.954$]

7. Calculate from the following data:
   (*i*) Two regression equations
   (*ii*) Coefficient of correlation
   (*iii*) Most likely value of X when Y = 10.

X:	45	55	56	58	60	65	68	70	75	80	85
Y:	56	50	48	60	62	64	65	70	74	82	90

[Ans. $X = 9.403 + 0.8514Y$, $Y = 0.884 + 0.992X$, $r = 0.9187$, $X_{10} = 17.913$]

(4) To Obtain Regression Equations from Coefficient of Correlation, Standard Deviations and Arithmetic Means of X and Y:

When the values of $\overline{X}$ and $\overline{Y}$, σ_x and σ_y and r of X and Y series are given, then regression equations are expressed in the following manner:

(1) Regression Equation of Y on X

$$Y - \overline{Y} = byx\,(X - \overline{X}) \qquad \text{where, } byx = r\,.\frac{\sigma_y}{\sigma_x}$$

or $$Y - \overline{Y} = r\,.\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

(2) Regression Equation of X on Y

$$X - \overline{X} = bxy\,(Y - \overline{Y}) \qquad \text{where, } bxy = r\,.\frac{\sigma_x}{\sigma_y}$$

or $$X - \overline{X} = r\,.\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

Note: The above said form, of the regression equation is used only when the values of $\overline{X}$ and $\overline{Y}$, σ_x and σ_y and r are given.

The following examples makes the above said method more clear.

Example 21. You are given the following information:

	X	Y
Arithmetic mean:	5	12
Standard deviation:	2.6	3.6
Correlation coefficient:	$r = 0.7$	

(*i*) Obtain two regression equations.
(*ii*) Estimate Y when X = 9.
(*iii*) Estimate X when Y = 12.

Solution: Given, $\overline{X} = 5, \overline{Y} = 12, \sigma_x = 2.6, \sigma_y = 3.6, r = 0.7$.

(*i*) Regression Equation of X on Y

$$X - \overline{X} = = r\,.\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

Putting the values in the equation, we get

$$X - 5 = 0.7 \times \frac{2.6}{3.6}(Y - 12)$$

$$X - 5 = 0.51\,(Y - 12)$$

$$X - 5 = 0.51Y - 6.12$$

$$X = 0.51Y - 1.12$$

or $$X = -1.12 + 0.51Y$$

Regression Equation of Y on X

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

Putting the values in the equation, we get

$$Y-12=0.7\times\frac{3.6}{2.6}(X-5)$$

$$Y-12=0.97(X-5)$$

$$Y-12=0.97X-0.97\times 5$$

$$Y-12=0.97X-4.85$$

$$Y=0.97X-4.85+12$$

$$Y=0.97X+7.15$$

or $$Y=7.15+0.97X$$

(ii) **Most likely value of Y when X = 9**

For this purpose, we use regression of Y on X

$$Y=7.15+0.97X$$

Putting X = 9 in the equation, we get

$$Y=7.15+0.97(9)=7.15+8.73=15.88$$

(iii) **Most likely value of X when Y = 12**

For this purpose, we use regression of X on Y

$$X=-1.12+0.51Y$$

Putting Y = 12 in the equation, we get

$$X=-1.12+0.51(12)$$

$$X=-1.12+6.12=5$$

Example 22. You are given below the following information about advertisement and sales:

	Adv. Expenditure (Rs. crore)	Sales (Rs. crore)
Mean	20	120
S.D.	5	25
Correlation coefficient, r_{xy} = + 0.8		

(i) Calculate the two regression equations.

(ii) What should be the advertisement budget if the company wants to attain sales target of Rs. 150 crore?

(iii) Find the most likely sales when advertisement expenditure is Rs. 25 crore.

Solution: Let X = Adv. Expenditure and Y = Sales

Thus, we have $\bar{X}=20, \bar{Y}=120, \sigma_x=5, \sigma_y=25, r_{xy}=0.8$

(i) *(a)* **Regression Equation of X on Y**

$$X-\bar{X}=r.\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

$$X-20=0.8\times\frac{5}{25}(Y-120)$$

$$X-20=0.16(Y-120)$$

$$X-20=0.16Y-19.2$$

$$X=0.16Y-19.2+20$$

$$\therefore \quad X=0.16Y+0.8$$

(b) **Regression Equation of Y on X**

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$Y-120=0.8\times\frac{25}{5}(X-20)$$

$$Y-120=4(X-20)$$

$$Y-120=4X-80$$

$$Y=40+4X$$

(ii) When sales target (Y) is Rs. 150 crore, then the advertisement expenditure (X) is

$$X=0.8+0.16Y$$

Put $Y=150$, $X=0.8+0.16(150)$

$$=0.8+24=24.8 \text{ crore.}$$

(iii) When advertisement expenditure (X) is Rs. 25 crore, the sales (Y) is

$$Y=40+4X$$

Put $X=25$, $Y=40+4(25)$

$$=40+100=140 \text{ crore.}$$

Example 23. Find the regression equations when you know:

$$\bar{X}=68.2,\ \bar{Y}=9.9,\ \frac{\sigma_y}{\sigma_x}=0.44, r=0.76$$

Solution: Given, $\bar{X}=68.2$, $\bar{Y}=9.9$, $\frac{\sigma_y}{\sigma_x}=0.44, r=0.76$

(i) **Regression Equation of Y on X**

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

Putting the values in the equation, we get

$$Y-9.9=0.76\times0.44(X-68.2)$$
$$Y-9.9=0.3344(X-68.2)$$
$$Y-9.9=0.3344X-22.81$$
$$Y=0.3344X-22.81+9.9$$
$$Y=0.3344X-12.91$$

or

(ii) Regression Equation of X on Y:

When $\frac{\sigma_y}{\sigma_x}=0.44$ or $\frac{44}{100}$

Then $\frac{\sigma_x}{\sigma_y}=\frac{100}{44}$ or 2.27

$$X-\bar{X}=r.\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

$$X-68.2=0.76\times2.27(Y-9.9)$$
$$X-68.2=1.725(Y-9.9)$$
$$X-68.2=1.725Y-17.08$$
$$X=1.725Y-17.08+68.2$$
$$X=1.725Y+51.12$$

Example 24. Find the expected price in Mumbai when price in Calcutta is Rs. 70 using the following data:

Average Price in Calcutta	:	Rs. 65
Average Price in Mumbai	:	Rs. 67
S.D. of Price in Calcutta	:	2.5
S.D. of Price in Mumbai	:	3.5
Correlation coefficient between price of Mumbai and Calcutta	:	0.8

Solution: Let X = Price in Calcutta, Y = Price in Mumbai

Given: $\bar{X}=65, \bar{Y}=67, \sigma_x=2.5, \sigma_y=3.5, r=0.8$

Expected Price in Mumbai (Y) when price in Calcutta (X) = 70 can be found from regression equation of Y on X.

Regression Equation of Y on X

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

Putting the values, we get

$$Y-67=0.8\times\frac{3.5}{2.5}(X-65)$$

$$Y-67=1.12(X-65)$$

$$Y-67=1.12X-72.8$$
$$Y=1.12X-72.8+67$$
$$Y=1.12X-5.8$$

When X = 70, $Y=1.12\,(70)-5.8=78.4-5.8$
$$=72.6$$

Thus, the expected price in Mumbai is Rs. 72.6 corresponding to Rs. 70 at Calcutta.

Example 25. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8, the average of husband age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations:

(i) the expected age of husband when wife's age is 20 years and

(ii) the expected age of wife when husband's age is 33 years.

Solution: Let age of wife be denoted by Y and age of husband by X. We are given:

$$\bar{X}=25, \bar{Y}=22, \sigma_x=4, \sigma_y=5, r=0.8$$

(i) For estimating age of husband when wife's age is 20 years, we use regression of X on Y as follows:

$$X-\bar{X}=r.\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

$$X-25=0.8\times\frac{4}{5}(Y-22)$$
$$X-25=0.64(Y-22)$$
$$X-25=0.64Y-14.08$$
$$X=0.64Y+10.92$$

When Y = 20, $X=0.64\,(20)+10.92=12.8+10.92=23.72$

Thus, the expected age of husband when wife's age is 20 years shall be 23.72 years.

(ii) For estimating age of wife when husband's age is 33 years, we use regression equation of Y on X as follows:

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$Y-22=0.8\times\frac{5}{4}(X-25)$$
$$Y-22=1(X-25)$$
$$Y-22=X-25 \Rightarrow Y=X-3$$

When X = 33, Y = 33 − 3 = 30

Thus, the expected age of wife when husband's age is 33 is 30 years.

IMPORTANT TYPICAL EXAMPLES

Example 26. The following data based on 450 students are given for marks in Statistics and Economics at a certain Examination:

Mean Marks in Statistics	:	40
Mean Marks in Economics	:	48
S.D. of Marks in Statistics	:	12
The variance of marks in Economics	:	256

Sum of the products of deviations of marks from their respective means is 42075.

(i) Obtain the equations of two lines of regression.

(ii) Estimate the average marks in Economics of candidates who obtained 50 marks in Statistics.

Solution: (i) Let X denote marks in Statistics and Y denote marks in Economics. We are given:

$\bar{X}=40,\quad \bar{Y}=48$

$\sigma_x=12,\quad \sigma_y^2=256 \Rightarrow \sigma_y=16$

$\Sigma xy=42075$

Before we obtain the regression equations, we compute the coefficient of correlation (r) by using the formula:

$$r=\frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

$$=\frac{42075}{450\times12\times16}=\frac{42075}{86400}$$

$=+0.49$ approx.

Regression Equation of X on Y

$$X-\bar{X}=r.\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

$$X-40=0.49\times\frac{12}{16}(Y-48)$$

$$X-40=\frac{5.88}{16}(Y-48)$$

$$X-40=0.3675(Y-48)$$

$$X-40=0.3675Y-17.64$$

$$X=0.3675Y-17.64+40$$

$$X=0.3675Y+22.36$$

Regression Equation of Y on X

$$Y-\bar{Y}=r.\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$Y-48=0.49\times\frac{16}{12}(X-40)$$

$$Y-48=\frac{7.84}{12}(X-40)$$

$$Y-48=0.653(X-40)$$

$$Y-48=0.653X-26.12$$

$$Y=0.653X-26.12+48$$

$$Y=0.653X+21.88$$

(ii) To estimate the marks in Economics when 50 marks in Statistics is given, we use regression of Y on X.

$$Y=0.653X+21.88$$

When X = 50,

$$Y=0.653\,(50)+21.88$$

$$=32.65+21.88$$

$$=54.53 \text{ or } 55 \text{ marks}$$

Thus, the expected marks in Economics is 55.

Example 27. If $\bar{X}=25,\ \bar{Y}=120, bxy=2$

Estimate the value of X when Y = 130.

Solution: Given, $\bar{X}=25,\ \bar{Y}=120, bxy=2$

For estimating X when Y = 130, we use regression equation of X on Y as follows:

$$X-\bar{X}=bxy(Y-\bar{Y})$$

or

$$X=\bar{X}+bxy(Y-\bar{Y})$$

$$X=25+2(130-120)$$

$$X=25+2(10)=45$$

Thus, the value of X is 45 when Y = 130.

Example 28. If $\sigma_x^2=9, \sigma_y^2=1600, r_{xy}=0.5$, obtain bxy.

Solution: Given, $\sigma_x^2=9(or\ \sigma_x=3), \sigma_y^2=1600(or\ \sigma_y=40), r_{xy}=0.5,$

We know, $bxy=r.\frac{\sigma_x}{\sigma_y}$

$$\therefore\quad bxy=0.5\times\frac{3}{40}=\frac{1.5}{40}$$

$$=0.0375$$