mode is not suitable.

4.13. MOMENTS

The word 'Moment' is derived from physics, where it is equal to the force multiplied by the perpendicular distance. In statistical terminology, 'Moment' is analogous to that in Mechanics. The frequency of each class corresponds to the force and the deviation of the variate from any measure of central tendency represents the distance. However, the term moment as used in Physics has nothing to do with the moment used in statistics, the only analogy being that in statistics, we talk of moment of random variable about some point. The moment in statistics are used to describe the various characteristics of frequency distribution like central tendency, variation, skewness and Kurtosis. The Greek Letter μ (read as mu) is generally used to denote the moments.

4.13.1. Moments About Mean ($\bar{x}$)

For Ungrouped Data. If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then r^{th} moment about mean $\bar{x}$ is denoted by μ_r and is defined as

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r; \quad r = 0,\ 1,\ 2,...$$

For Grouped Data. If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then r^{th} moment about mean $\bar{x}$ is denoted by μ_r and is defined as

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r; \quad r = 0,\ 1,\ 2,...$$

where $$N = \sum_{i=1}^{n} f_i$$

For $r = 0$, $$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^0 = \frac{1}{N}\sum_{i=1}^{n} f_i - 1 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{1}{N}\cdot N = 1$$

For $r = 1$, $$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - f_i\bar{x})$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \bar{x}\frac{1}{N}\sum_{i=1}^{n} f_i = \bar{x} - \bar{x}\cdot\frac{1}{N}\cdot N = \bar{x} - \bar{x} = 0$$

For $r = 2$, $$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = (\text{S.D.})^2 = \text{Variance and so on.}$$

$\therefore$ The second moment about the mean coincides with the variance of the distribution.

Similarly, $$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^3;$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^4$$

and so on.

4.13.2. Moments About Arbitrary Point

When the actual mean is in fractions, it is difficult to calculate moments by applying the above formula. In such a case, we can first compute moments about an arbitrary origin 'A' and then convert these moments into moments about the actual mean. Moments about arbitrary origin, also called *raw moments* are denoted by the symbol μ' to distinguish them from moments about mean which are denoted by μ.

For Ungrouped Data. If $x_2, x_2, ..., x_n$ are the values of the variable x under consideration, the rth moment about any point 'A' is denoted by μ'_r and is defined as

$$\mu'_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^r; \quad r = 0,1,2,...$$

For Grouped Data. If $x_2, x_2, ..., x_n$ are the values of a the variable x with the corresponding frequencies $f_1, f_2 \dots f_n$ respectively then the r^{th} moment about any point 'A' is denoted by μ'_r and is defined as

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)^r; \quad r = 0, 1, 2,...$$

where $$N = \sum_{i=1}^{n} f_i$$

For $r = 0$, $$\mu_0' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^0 = \frac{1}{N}\sum_{i=1}^{n} f_i \cdot 1 = \frac{1}{N}\cdot N = 1$$

For $r = 1$, $$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A) = \frac{1}{N}\sum_{i=1}^{n}(f_i x_i - f_i A)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i - a\frac{1}{N}\sum_{i=1}^{n} f_i = \overline{x} - a\cdot\frac{1}{N}\cdot N = \overline{x} - a$$

For $r = 2$, $$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2$$

For $r = 3$, $$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^3$$

For $r = 4$, $$\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^4$$

and so on....

4.13.3 Moments about the Origin

If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then r^{th} moment about origin in denoted by v_r and is defined as

$$v_r = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r;\ r = 0, 1, 2, ...$$

where $$N = \sum_{i=1}^{n} f_i$$

For $r = 0$, $$v_0 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^0 = \frac{1}{N}\cdot N = 1$$

For $r = 1$, $$v_1 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \overline{x}$$

For $r = 2$, $$v_2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2$$

For $r = 3$, $$v_3 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^3$$

For $r = 4$, $$v_4 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^4$$

and so on.....

4.13.4 Relation Between μ_r and μ'_r

We know that
$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_{i=1}^{n} f_i [(x_i - A) + (A - \bar{x})]^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \left[(x_i - A)^r + {}^rC_1 (x_i - A)^{r-1}(A - \bar{x}) + {}^rC_2 (x_i - A)^{r-2}(A - \bar{x})^2 + \ldots + (A - \bar{x})^r\right]$$

$\therefore$
$$\mu_r = \mu'_r + {}^rC_1 \mu'_{r-1}(A - \bar{x}) + {}^rC_2 \mu'_{r-2}(A - \bar{x})^2 + \ldots + (A - \bar{x})^r \quad \ldots(i)$$

Now
$$A - \bar{x} = -(\bar{x} - A) = -\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A) = -\mu'_1 \quad \ldots(ii)$$

From eqns. (*i*) and (*ii*), we obtain
$$\mu_r = \mu'_r - {}^rC_1 \mu'_{r-1}\mu'_1 + {}^rC_2 \mu'_{r-2}\mu'^2_1 + \ldots + (-1)^r \mu'^r_1 \quad \ldots(iii)$$

Putting $r = 2, 3, 4$ in (*iii*), we obtain

$$\boxed{\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \mu'_2 - \mu'^2 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3(\mu')^4 \end{aligned}}$$

4.13.5. Effect of Change of Origin and Scale on Moments

Let $u_i = \dfrac{x_i - A}{h}$ *i.e.*, $x_i = A + hu_i$

$\therefore$ $\bar{x} = A + h\bar{u}$, where bar denoted the mean of respective variable.

$\therefore$
$$x_i - \bar{x} = h(u_i - \bar{u})$$

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r = \frac{1}{N}\sum_{i=1}^{n} f_i h^r u_i^r = h^r \cdot \frac{1}{N}\sum_{i=1}^{n} f_i u_i^r$$

Also
$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_{i=1}^{n} f_i h^r (u_i - \bar{u})^r = h^r \cdot \frac{1}{N}\sum_{i=1}^{n} f_i (u_i - \bar{u})^r$$

Hence, the r^{th} moment of the variable x is the h^r times the corresponding moment of the variable u.

4.13.6 Relation Between μ_r and v_r

We know that
$$v_r = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r; \quad r = 0, 1, 2, \ldots$$

where
$$N = \sum_{i=1}^{n} f_i = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A + A)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i [(x_i - A)^r + {}^rC_1 (x_i - A)^{r-1} \ldots + A_r]$$

$$= \mu'_r + {}^rC_1 \mu_{r-1} A + \ldots + A^r$$

If we take $A = \bar{x}$ (for μ_r) then

$$\nu_r = \mu_r + {}^rC_1\mu_{r-1}\bar{x} + {}^rC_2\mu_{r-2}\bar{x}^2 + \ldots + \bar{x}^r \qquad \ldots(i)$$

Putting $r = 1, 2, 3, 4$ in (i), we obtain

$$\boxed{\begin{aligned}
\nu_1 &= \bar{x} = A + \mu_1' \\
\nu_2 &= \mu_2 + \bar{x}^2 = \mu_2 + \nu_1^2 \\
\nu_3 &= \mu_3 + 3\nu_2\bar{x} - 2\bar{x}^3 = \mu_3 + 3\nu_1\nu_2 - 2\nu_1^3 \\
\nu_4 &= \mu_4 + 4\nu_3\bar{x} - 6\bar{x}\nu_2\bar{x}^2 + 3\bar{x}^4 \\
&= \mu_4 + 4\nu_1\nu_3 - 6\nu_1^2\nu_2 + 3\nu_1^4
\end{aligned}}$$

4.13.4. Sheppard's Correction For Moments

For grouped data, we suppose that the observations are concentrated at the mid-point of the class-intervals, i.e., each member has an equal value of the variable in the class. But it is not so, so they need some adjustment. W.F. Sheppared suggested the following corrections for second and fourth moments:

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{12}h^2\mu_2 + \frac{7}{240}h^4$$

where h is the width of the class interval.

Note : μ_1 and μ_3 require no correction.

SOLVED EXAMPLES

Example 4.37. Calculate the first four moments about the mean of the following data :

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Solution. Let us first calculate moment about $x = 4$

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - 4)^r = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^r$$

where

$$d_i = x_i - 4$$

Calculation of First Four Moments

x	f	$d = x - 4$	fd	fd^2	fd^3	fd^4
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	$\Sigma f = 256$	$\Sigma d = 0$	$\Sigma fd = 0$	$\Sigma fd^2 = 512$	$\Sigma fd^3 = 0$	$\Sigma fd^4 = 2816$

$$\mu_1' = \frac{1}{N}\Sigma fd = 0$$

$$\mu_2' = \frac{1}{N}\Sigma fd^2 = \frac{512}{256} = 2$$

$$\mu_3' = \frac{1}{N}\Sigma fd^3 = 0$$

$$\mu_4' = \frac{1}{N}\Sigma fd^4 = \frac{2816}{256} = 11$$

Moments about mean are

$$\mu_1 = 0 \text{ (always)}$$
$$\mu_2 = \mu_2' - \mu_1'^2 = 2$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3(\mu_1')^4 = 11$$

Example 4.38. From the following data, calculate moments about :

(*i*) Assumed mean 25

(*ii*) Actual mean, and

(*iii*) Moments about zero.

Variable	0-10	10-20	20-30	30-40
Frequency	1	3	4	2

Solution.

Calculation of Moments about Arbitrary Origin

Variable	*Mid-point m*	*f*	$d = \frac{m-25}{10}$	*fd*	*fd*^{2}	*fd*^{3}	*fd*^{4}
0-10	5	1	–2	–2	4	–8	16
10-20	15	3	–1	–3	3	–3	3
20-30	25	4	0	0	0	0	0
30-40	35	2	1	2	2	2	2
				$\Sigma fd = -3$	$\Sigma fd^2 = 9$	$\Sigma fd^3 = -9$	$\Sigma fd^4 = 21$

(*i*) Moments about assumed mean

$$\mu_1' = \frac{1}{N}\Sigma\, fd \times h = \frac{-3}{10} \times 10 = -3$$

$$\mu_2' = \frac{1}{N}\Sigma\, fd^2 \times h^2 = \frac{9}{10} \times (10)^2 = \frac{9}{10} \times 100 = 90$$

$$\mu_3' = \frac{1}{N}\Sigma\, fd^3 \times h^3 = \frac{-9}{10} \times 1000 = -900$$

$$\mu_4' = \frac{1}{N}\Sigma\, fd^4 \times h^4 = \frac{21}{10} \times 10000 = 21000$$

(*ii*) Moments about mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 90 - (-3)^2 = 81$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -900 - 3\,[(-3)\,(90)] + 2(-3)^3$$
$$= -900 + 810 - 54 = -144$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3(\mu_1')^4$$
$$= 21000 - 4(-900)\,(-3) + 6(-3)^2 \times 90 - 3\,(-3)^4$$
$$= 21000 - 10800 + 4860 - 243 = 14817$$

(*iii*) Moments about zero

$$v_1 = A + \mu_1' \text{ or the mean} = 25 - 3 = 22$$

$$v_2 = \mu_2 + (v_1)^2 = 81 + (22)^2 = 565$$

$$v_3 = \mu_3 + 3v_1v_2 - 2v_1^3 = -144 + 3(22)\,(565) - 2(22)^3$$
$$= -144 + 37290 - 21296 = 15850$$

$$v_4 = \mu_4 + 4v_1\,v_3 - 6v_1^2\,v_2 + 3v_1^4$$
$$= 14817 + 4(22)\,(15850) - 6(22)^2\,(565) + 3(22)^4$$
$$= 14817 + 1394800 - 1640760 + 702768 = 471625.$$

Example 4.39. The first four moments of a distribution about $x = 4$ are 1, 4, 10 and 45. Determine the values of the mean, variance, μ_3 and μ_4.

Solution. We have $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 45$ and $A = 4$.

We know that,
$$\mu_1' = \bar{x} - A$$
$$\Rightarrow \quad 1 = \bar{x} - 4$$
$$\Rightarrow \quad \bar{x} = 5$$
i.e.,
$$\text{Mean} = 5$$
Again,
$$\mu_2 = \mu_2' - \mu_1'^2 = 4 - (1)^2 = 4 - 1 = 3$$
i.e.,
$$\text{Variance} = 3$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$= 10 - 3 \times 4 \times 1 + 2(1)^3 = 10 - 12 + 2 = 0$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1' - 3(\mu_1')^4$$
$$= 45 - 4 \times 10 \times 1 + 6 \times 4 \times (1)^2 - 3(1)^4$$
$$= 45 - 40 + 24 - 3 = 69 - 43 = 26.$$

Example 4.40. If the first four moments of a distribution about the value 5 are – 4, 22, – 117 and 560. Obtain the moments about: (*i*) mean and (*ii*) origin.

Solution. We are given moments about an arbitrary origin 5.

Thus $\mu_1' = -4$, $\mu_2' = 22$, $\mu_3' = -117$, $\mu_4 = 560$.

Moments about Mean. From these, we can find out moments about mean from the following relationships.
$$\mu_2 = \mu_2' - \mu_1'^2$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
Substituting the values,
$$\mu_2 = 22 - (-4)^2 = 22 - 16 = 6$$
$$\mu_3 = -117 - 3(-4)(22) + 2(-4)^2$$
$$= -117 + 260 - 128 = 19$$
$$\mu_4 = 560 - 4(-4)(117) + 6(22)(-4)^3 - 3(-4)^4$$
$$= 560 - 1872 + 2112 - 768 = 32$$

Thus, the moments about mean are $\mu_1 = 0$, $\mu_2 = 6$, $\mu_3 = 19$ and $\mu_4 = 32$.

Moments about origin. Let the moments about origin be denoted by v_1, v_2, v_3 etc.

We have the following formulae :
$$v_1 = \bar{x} = A + \mu_1^1 = 5 - 4 = 1$$
$$v_2 = \mu_2 + \bar{x}^2 = 6 + 1^2 = 7$$

$$v_3 = \mu_3 + 3v_1v_2 - 2v_1^3$$
$$= 19 + 3(1)(7) - 2(1)^3$$
$$= 19 + 21 - 2 = 38$$
$$v_4 = \mu_4 + 4v_1v_3 - 6v_1^2 v_2 + 3v_1^4$$
$$= 32 + 4(1)(38) - 6(1)(7) + 3(1)^4$$
$$= 32 + 152 - 42 + 3 = 145$$

Thus, $v_1 = 1$, $v_2 = 7$, $v_3 = 38$ and $v_4 = 145$

Example 4.41. The first four moments about the mean of a frequency distribution with class size 3 are $\mu_1 = 0$, $\mu_2 = 8.7579$, $\mu_3 = -2.924$, $\mu_4 = 281.453$. Use Sheppard's formulae to find the correct moments.

Solution. Given that $\mu_1 = 0$, $\mu_2 = 8.7579$, $\mu_3 = -2.924$, $\mu_4 = 281.453$.

Using Sheppard's correction, we have

$$\mu_1 \text{ (corrected)} = \mu_1 = 0$$

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} = 8.7579 - \frac{3^2}{12} = 8.0079$$

$$\mu_3 \text{ (corrected)} = \mu_3 = -2.924$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$$

$$= 281.453 - \frac{1}{2}(3)^2 \times 8.0079 + \frac{7}{240}(3)^4 = 244.40495.$$

EXERCISE 4.1

1. Define moments. What is the use of moments in the study of a frequency distribution?
2. Find the first four moments about the mean of the following series :
   1, 3, 7, 9, 10 [**Ans.** $\mu_1 = 0$, $\mu_2 = 12$, $\mu_3 = -12$, $\mu_4 = 208.8$]
3. Compute the first four moments about the mean from the following data :

Mid-value	5	10	15	20	25	30	35
Frequency	8	15	20	32	23	17	5

[**Ans.** $\mu_1 = 0$, $\mu_2 = 59.993$, $\mu_3 = -49.644$, $\mu_4 = 8351.08$]

4. The first three moments of a distribution about the value 2 of the variable are 1, 16, –40. Show that the mean is 3, variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24, 76.
5. The first four moments of a distribution about the value 4 of the variable are –15, 17, –30 and 108. Find the moments about mean. [**Ans.** $\mu_1 = 0$, $\mu_2 = 14.75$, $\mu_3 = 39.75$, $\mu_4 = 142.3125$]
6. Find the first four moments about mean from the following data :

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of students	8	12	20	30	15	10	5

[**Ans.** $\mu_1 = 0$, $\mu_2 = 236.76$, $\mu_3 = 264.336$, $\mu_4 = 141290.10$]

7. Express the third and fourth central moments in terms of the first four moments about the origin.
8. How we moments helpful in a study of the different aspects of the formation of a frequency distribution.

4.14. SKEWNESS

In a symmetrical distribution, the mean, median and mode coincide. When a frequency distribution is not symmetrical, it is said to be asymmetrical or skewed. According to Wessel and Willet, "Skewness is lack of symmetry". Skewness of a distribution can be defined as the tendency of a distribution to depart from symmetry. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a normal distribution.

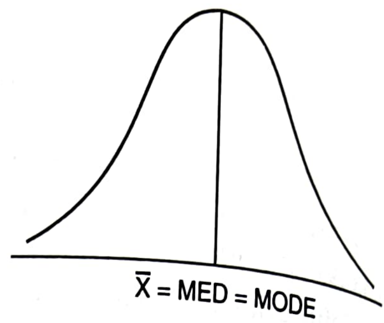

Fig. 4.2.

Skewness can be positive as well as negative. Skewness is positive if the longer tail of the distribution lies towards the right and negative if it lies towards the left.

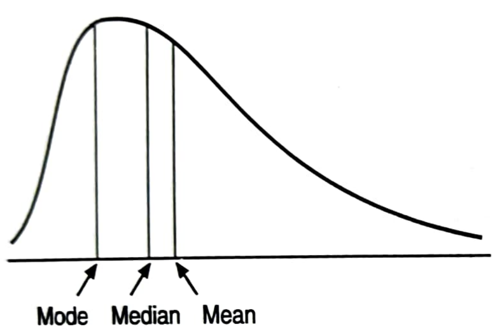

Fig. 4.3. Positively skewed distribution

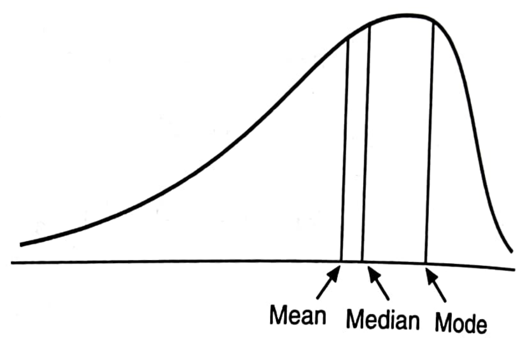

Fig. 4.4. Negatively skewed distribution

4.15. MEASURES

Measures of Skewness give us an idea about the extent of 'Lopsidedness' in a series. The various measures of skewness are :

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

1. Karl Pearson's Method. Karl Pearson method is based on the fact that in a symmetrical distribution, the value of arithmetic mean is equal to the value of mode. Karl Pearson's coefficient of skewness is defined as

$$S_K = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

If the mode is ill-defined, then using Mode = 3 Median – 2 Mean, we have

$$S_K = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Note : The value of Karl Pearson's coefficient of skewness lies between –3 and 3.

2. Bowley's Method. Bowley method is based on the fact that in a symmetrical distribution, the quartiles are equidistant from the median. Bowley's coefficient of skewness is defined as

$$S_K = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

Note : The value of Bowley's coefficient of skewness lies between –1 and 1.

3. Kelly's Method. Kelly's method is based on the fact that in a symmetrical distribution, the 10th percentile and 90th percentile are equidistant from the median.

$$\text{Kelly's coefficient of skewness} = \frac{P_{90} + P_{10} - 2\,\text{Median}}{P_{50} - P_{10}}$$

4. Method of Moments. A measure of skewness is obtained by making use of the third moment about the mean. When the method of moment is applied, β_1 is used as a relative measure of skewness. β_1 is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{or} \quad \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

This is also symbolised as γ_1. The value of β_1 shall be zero for a perfectly symmetrical distribution. The greater the value of β_1, the more skewed is the distribution.

SOLVED EXAMPLES

Example 4.42. Karl Pearson's coefficient of skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. Find the mode of the distribution.

Solution. We know that,

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

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

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

Example 4.43. Compute the coefficient of skewness from the following figures :

25, 15, 23, 40, 27, 25, 23, 25, 20

Solution.

Size of item (x)	*$d = x - 20$*	*$d^2 = (x - 20)^2$*
25	5	25
15	–5	25
23	3	9
40	20	400
27	7	49
25	5	25
23	3	9
25	5	25
20	0	0
	$\Sigma d = 43$	$\Sigma d^2 = 567$

Number of items $n = 9$

$$\text{Mean} = 20 + \frac{\Sigma d}{n} = 20 + \frac{43}{9} = 20 + 4.78 = 24.78$$

Since 25 occurs 3 times and no other item occur 3 or more than 3 times.

$\therefore$ Mode = 25

Standard deviation $\sigma = \sqrt{\frac{1}{n}\Sigma d^2 - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{567}{9} - \left(\frac{43}{9}\right)^2} = 6.33$

$\therefore$ Coefficient of skewness $= \dfrac{\text{Mean} - \text{Mode}}{\text{Standard devaition}} = \dfrac{24.78 - 25}{6.33} = -0.03$

Example. 4.44. In a frequency distribution, the coefficient of skewness based on the quartiles is 0.6. If the sum of the upper and the lower quartile is 100 and the median is 38, determine the values of the upper and the lower quartiles.

Solution. Coefficient of skewness based on quartiles $= \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$

Here $Q_3 + Q_1 = 100$ ...(*i*)

$Q_2 = 38$

$\therefore$ $0.6 = \dfrac{100 - 2 \times 38}{Q_3 - Q_1}$

$\Rightarrow$ $Q_3 - Q_1 = \dfrac{24}{0.6} = 40$ ...(*ii*)

From (*i*) and (*ii*), we get

$Q_3 = 70$ and $Q_1 = 30$

Example 4.45. For the following distribution, find the first four moments about the mean. Also find the value of β_1. Is it a symmetrical distribution?

x	2	3	4	5	6
f	1	3	7	3	1

Solution.

Calculation of moments

x	f	fx	$x - \bar{x}$	$f(x - \bar{x})$	$f(x - \bar{x})^2$	$f(x - \bar{x})^3$	$f(x - \bar{x})^4$
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
	$N = \Sigma f = 15$	$\Sigma fx = 60$		0	14	0	38

$$\bar{x} = \frac{\Sigma x}{N} = \frac{60}{15} = 4$$

Now,
$$\mu_1 = \frac{\Sigma f(x-\bar{x})}{N} = \frac{0}{15} = 0$$

$$\mu_2 = \frac{\Sigma f(x-\bar{x})^2}{N} = \frac{14}{15}$$

$$\mu_3 = \frac{\Sigma f(x-\bar{x})^3}{N} = \frac{0}{15} = 0$$

$$\mu_4 = \frac{\Sigma f(x-\bar{x})^4}{N} = \frac{38}{15}$$

Also,
$$\beta_1 = \frac{\mu_3^2}{\mu_2^2} = \frac{0}{\left(\frac{14}{15}\right)^2} = 0$$

In a symmetrical distribution, β_1 is zero. Hence, this distribution is symmetrical.

Example 4.46. Calculate Karl Pearson's coefficient of skewness from the data given below :

Income (₹)	**400-500**	**500-600**	**600-700**	**700-800**	**800-900**
Number of employees	**8**	**16**	**20**	**17**	**3**

Solution. We know that $S_K = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$

Calculation of Karl Pearson's coefficient of skewness

Income (₹)	***m.p.*** ***m***	***Number of employees*** ***f***	$\frac{m-650}{100}$ ***d***	***fd***	***fd²***
400-500	450	8	–2	– 16	32
500-600	550	16	–1	– 16	16
600-700	650	20	0	0	0
700-800	750`	17	1	17	17
800-900	850	3	2	6	12
		$N = 64$		$\Sigma fd = -9$	$\Sigma fd^2 = 77$

Mean :
$$\bar{x} = A + \frac{\Sigma fd}{N} \times i$$

$$A = 650, \Sigma fd = -9, N = 64, h = 100$$

$\therefore$
$$\bar{x} = 650 - \frac{9}{64} \times 100 = 650 - 14.06 = 635.94$$

Mode : By inspection mode lies in the class 600-700

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$L = 600, f_0 = 16, f_1 = 20, f_2 = 17, h = 100$

$$\therefore \qquad \text{Mode} = 600 + \frac{20-16}{40-16-17} \times 100$$

$$= 600 + 57.14 = 657.14$$

Standard Deviation :
$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$\Sigma fd^2 = 77, N = 64, \Sigma fd = -9, h = 100$$

$$\therefore \qquad \sigma = \sqrt{\frac{77}{64} - \left(\frac{-9}{64}\right)^2} \times 100 = \sqrt{1.203 - 0.0198} \times 100$$

$$= \sqrt{1.1832} \times 100 = 1.088 \times 100 = 108.8$$

$$\therefore \text{ Coefficient of skewness} = \frac{635.94 - 657.14}{108.8} = \frac{-21.2}{108.8} = -0.195.$$

Example 4.47. Calculate the quartile coefficient of skewness of the following distribution :

Variate (x)	1–5	6–10	11–15	16–20	21–25	26–30	31–35
Frequency (f)	3	4	68	30	10	6	2

Solution.

Calculation of Bowley's coefficient of skewness

Class	f	*c.f.*
1–5	3	3
6–10	4	7
11–15	68	75
16–20	30	105
21–25	10	115
26–30	6	121
31–35	2	123
	$N = 123$	

$\frac{N}{2} = 61.5$ Median class is 11–15

$$\therefore \quad \text{Median} = L+\frac{\left(\frac{N}{2}\right)\text{th}-c.f.}{f}\times h$$

$$= 11+\frac{61.5-7}{68}\times 5 = 11+\frac{272.5}{68} = 15$$

$$\frac{N}{4} = 30.75$$

$\therefore$ Class of lower quartile Q_1 is 11–15

$$Q_1 = L+\frac{\left(\frac{N}{4}\right)\text{th}-c.f.}{f}\times h$$

$$= 11+\frac{30.75-7}{68}\times 5 = 11+\frac{118.35}{68} = 12.75$$

$$\frac{3N}{4} = 92.25$$

$\therefore$ Class of upper quartile Q_3 is 16–20

$$Q_3 = L+\frac{\left(\frac{3N}{4}\right)\text{th}-c.f.}{f}\times h$$

$$= 16+\frac{92.25-75}{30}\times 5 = 16+\frac{16.50}{6} = 18.875$$

Quartile coefficient of skewness

$$= \frac{Q_3+Q_1-2\,\text{Median}}{Q_3-Q_1} = \frac{31.50-30}{6} = \frac{1.50}{6} = 0.25$$

EXERCISE 4.2

1. What do you understand by skewness? How is it measured?
2. Distinguish between positive and negative skewness with the help of diagrams.
3. For a moderately skewed data, the arithmetic mean is 100, the variance is 35 and Karl Pearson's coefficient of skewness is 0.2. Find its mode and median.

   [Ans. Mode = 98.81, Median = 99.61]
4. For a group of 20 items, $\Sigma x = 1452$, $\Sigma x^2 = 144280$ and mode = 63.7. Find Pearson's coefficient of skewness.

   [Ans. 0.202]
5. Compute the coefficient of skewness based on the moments for the following distribution :

x	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f	1	5	12	22	17	9	4	3	1	1

[Ans. $\beta_1 = 0.504$]

6. Calculate coefficient of skewness from the following data :

Marks above	0	10	20	30	40	50	60	70	80
Frequency	150	140	100	80	80	70	30	14	0

[**Ans.** – 0.8]

7. Find the coefficient of skewness for the following distribution :

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

[**Ans.** – 1]

8. Calculate measure of skewness based on quartiles and mean from the following data :

Variable	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency	358	2417	976	129	62	18	10

[**Ans.** 0.13]

9. In a certain distribution the following results were obtained, $\bar{x} = 45$, median = 48, coefficient of skewness = – 0.4. The person who gave you the data failed to give the value of standard deviation and you are required to estimate it with the help of the available information. [**Ans.** 22.5]

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

Data expenditure	0-20	20-40	40-60	60-80	80-100
Number of families	13	?	27	?	16

If the mode of the distribution is 44, calculate the Karl Pearson's coefficient of skewness.

[**Ans.** 0.237]

4.16. KURTOSIS

The 'peakedness' of the frequency distribution is another characteristics which might be measured. Kurtosis exhibits the extent to which the curve is more peaked or more flat-topped than the normal curve. It studies the concentration of items at the central part of a series. If the items concentrate too much in the centre, the curve becomes leptokurtic, and if the concentration in centre is comparatively little, the curve becomes platykurtic. If kurtosis is zero, the curve is mesokurtic *i.e.*, a normal curve.

The following diagram illustrates the shape of three different curves mentioned above.

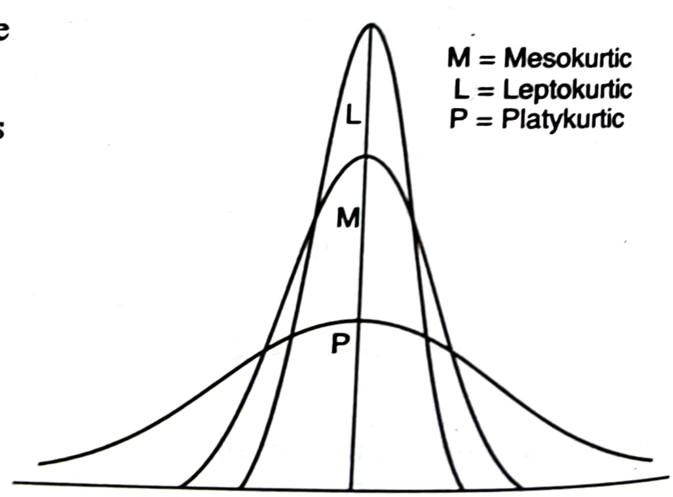

Fig. 8.5.

Kurtosis is measured by the coefficient β_2 which is given by

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

where μ_4 = 4th moment and μ_2 = 2nd moment

The greater the value of β_2, the more peaked the distribution.

Fisher's measure of kurtosis is expressed in terms of the Greek gamma (γ) i.e.,

$$\gamma_2 = \beta_2 - 3; \gamma_1 = +\sqrt{\beta_1}$$

If $\gamma_2 = 0$, i.e., $\beta_2 = 3$, the curve is normal or mesokurtic.

$\gamma_2 < 0$, i.e., $\beta_2 < 3$, the curve is platykurtic

$\gamma_2 < 0$, i.e., $\beta_2 > 3$, then curve is leptokurtic and if $\gamma_1 = 0$, the curve is symmetrical.

SOLVED EXAMPLES

Example 4.48. 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. Here

$$\mu'_1 = 1.5$$
$$\mu'_2 = 17$$
$$\mu'_3 = -30$$
$$\mu'_4 = 108$$

$$\therefore \quad \mu_2 = \mu'_2 - (\mu'_1)^2 = 17 - (-1.5)^2 = 14.75$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$
$$= 108 - 4(-1.5)(-30) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125$$

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

$\therefore$ The distribution is platykurtic.

Example 4.49. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Obtain as far as possible the various characteristics of this distribution on the basis of the information given.

Solution. We have, $\mu'_1 = 2$, $\mu'_2 = 20$, $\mu'_3 = 40$, $\mu'_4 = 55$.

The first four moments about the mean are :

$$\mu_1 = 0$$
$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 20 - 4 = 16$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$
$$= 40 - 3(20)(2) + 2(2)^3 = -64$$
$$\mu_4 = \mu_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3(\mu'_1)^4$$
$$= 50 - 4(40)^2 + 6(20)(2)^2 - 3(2)^4$$
$$= 50 - 320 + 480 - 48 = 162$$

Mean of the distribution : $\bar{x} = A + \mu' = 5 + 2 = 7$

Standard deviation : $\sigma = \sqrt{\mu_2} = \sqrt{16} = 4$

Skewness : $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{(-64)^2}{(16)^3} = 1$

Kurtosis : $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{162}{(16)^2} = \dfrac{81}{128} = 0.63$

$$\gamma_1 = \sqrt{\beta_1} = 1, \gamma_2 = \beta_2 - 3 = 0.63 - 3 = -2.37$$

Example 4.50. The following data 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 $\Sigma fd^4 = 86752$, in which $d = x - 48$
Do you think that the distribution is platykurtic?

Solution. For finding out whether the distribution is playkurtic or not, we have to calculate the value of β_2.

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

$$\mu_1' = \frac{\Sigma fd}{N} = \frac{50}{100} = 0.5$$

$$\mu_2' = \frac{\Sigma fd^2}{N} = \frac{1970}{100} = 19.70$$

$$\mu_3' = \frac{\Sigma fd^3}{N} = \frac{2948}{100} = 29.48$$

$$\mu_4' = \frac{\Sigma fd^4}{N} = \frac{86752}{100} = 867.52$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 19.70 - (0.5)^2 = 19.70 - 0.25 = 19.45$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = 29.48 - 3(0.5)(29.48) + 2(0.5)^3$$

$$= 29.48 - 44.22 + 0.25 = -14.49$$

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

$$= 867.52 - 4(0.5)(29.48) + 6(19.70)(0.5)^2 - 3(0.5)^4$$

$$= 867.52 - 58.96 + 29.55 - 0.1875 = 837.9225$$

$$\therefore \quad \beta_2 = \frac{837.9225}{(19.45)^2} = \frac{837.9225}{378.3025} = 2.215$$

Since the value of β_2 is less than 3, the distribution is playkurtic.

Example 4.51. Find out the kurtosis of the data given below :

Class interval	0–10	10–20	20–30	30–40
Frequency	1	4	3	2

Solution. Calculation of kurtosis

Class interval	*mid-points* m	f	$\frac{m-15}{10}$ d	fd	fd^2	fd^3	fd^4
0–10	5	1	–1	–1	1	–1	1
10–20	15	4	0	0	0	0	0
20–30	25	3	1	+3	3	+3	3
30–40	35	2	2	+4	8	+16	32
	$N = \Sigma f = 10$			$\Sigma fd = 6$	$\Sigma fd^2 = 12$	$\Sigma fd^3 = 18$	$\Sigma fd^4 = 36$

$$\mu_1' = \frac{\Sigma fd}{N} \times h = \frac{6}{10} \times 10 = 6$$

$$\mu_2' = \frac{\Sigma fd^2}{N} \times h^2 = \frac{12}{10} \times 100 = 120$$

$$\mu_3' = \frac{\Sigma fd^3}{N} \times h^3 = \frac{18}{10} \times 1000 = 1800$$

$$\mu_4' = \frac{\Sigma fd^4}{N} \times h^4 = \frac{36}{10} \times 10000 = 36000$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 120 - (6)^2 = 120 - 36 = 84$$

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

$$= 36000 - 4(6)\,(1800) + 6(6)^2\,.\,120 - 3(6)^4$$

$$= 36000 - 43200 + 25920 - 3888 = 14832$$

Now $$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{14832}{(84)^2} = 2.102 < 3$$

which indicate that the frequency curve is less peaked at the top (*i.e.*, platykurtic).

EXERCISE 4.3

1. What do you understand by kurtosis ? How is it measured ?
2. The first four moments about mean of a frequency distribution are 0, 100, –7 and 35000. Discuss the kurtosis of the distribution. **[Ans.** Leptokurtic]
3. The first four moments of a distribution about $x = 4$ are 1, 4, 10 and 45. Find the various characteristics of the distribution on the information given. Comment upon the nature of the distribution. **[Ans.** $\beta_1 = 0$, $\beta_2 = 2.89$, Platykurtic]
4. The following data refers to the length of life (x) of a sample of good years tyres. Do you think that the distribution is playkurtic ?

Given that $$d = \frac{x - 35}{10} \quad N = 10, \Sigma fd = 50$$

$$\Sigma fd^2 = 1967.2 \ Sfd^3 = 2925.8, Sfd^4 = 86650.2$$

5. Discuss the kurtosis of the following distribution.

Marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Number of student	8	12	20	30	15	10	5

4.23. REGRESSION ANALYSIS

The regression analysis is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables. We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of on variable from that of the other. Here the variable whose value is to be predicted is called dependent or explained variable and the variable used for prediction is called *independent* or *explanatory variable*.

Remember: By regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables.

In a bivariate distribution, the analysis is restricted to only two variables only

Difference between Correlation and Regression Analysis

1. Correlation coefficient is quantitative measure of the extent or degree of linear relationship between the two variables. On the other hand regression means an average relationship between the two variables.
2. Correlation does not necessarily establish cause and effect relationship. However, in regression analysis there is a clear indication of cause and effect relationship. Here, the independent variable is the cause and dependent variable the effect.
3. Correlation may be nonsense or spurious while there is nothing like non sense regression.
4. In correlation analysis, r_{xy} measures the linear relationship between the variable x and y. Here $r_{xy} = r_{yx}$ *i.e.*, it is immaterial which of the two variables is taken as dependent or independent. However, in regression analysis, the identity of variable *i.e.*, which is dependent and which one is independent in important. As will be seen later, the two regression coefficients b_{xy} and b_{yx} are not symmetric like correlation coefficient *i.e.*, $b_{xy} \neq b_{yx}$.
5. Correlation coefficient r_{xy} is a relative measure of the linear relationship between x and y, while the regression coefficients, b_{yx} and b_{xy}, are absolute measures of change in the value of one variables corresponding to a unit change in value of another variable.
6. Whereas correlation analysis is confined only to the study of linear relationship between two variables, the regression analysis deals with linear and non-linear relationship between variables.

4.23.1. Regression Lines

For a simple regression analysis there are two regression lines called X on Y and Y on X, where X and Y are two variables. X on Y lines denotes that X is dependent and Y independent, *i.e.*, X depends on Y. Similarly Y on X denotes that Y depends on X.

Let $\bar{x}$ and $\bar{y}$ be the means for series X and series Y and σ_X the standard deviation for X series and σ_Y for Y-series. Let r be the correlation coefficient between them; then the regression line for

Y on X $\boxed{Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})}$, regression line for Y on X

X on Y, $\boxed{X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})}$, regression line for X on Y.

Here, the value of r lies between -1 to $+1$ and the terms $r\frac{\sigma_Y}{\sigma_X}$ and $r\frac{\sigma_X}{\sigma_Y}$ are called the *regression coefficients* for Y on X and X on Y series. These coefficients are denoted by b_{YX} and b_{XY}. Hence

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{\sum dx\, dy}{\sum x^2} \quad \ldots(i)$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{\sum dx\, dy}{\sum y^2} \quad \ldots(ii)$$

Multiplying (*i*) and (*ii*)

$$b_{XY} b_{YX} = r^2$$

$$\boxed{r = \sqrt{b_{XY} b_{YX}}}$$

4.23.2. Line of Regression of *y* on *x*

For getting the line of regression of y on x, we shall consider y as dependant variable and x as independent variable.

Let $y = a + bx$ be the equation of regression line of y on x.

The residual for i^{th} point is $E_i = y_i - a - bx_i$.

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \text{...}(i)$$

According to the *principle of Least squares*, the constants a and b are chosen in such a way that the sum of the squares of residuals is minimum.

Now, the condition for U to be maximum or minimum is given by

$$\frac{\partial U}{\partial a} = 0 \text{ and } \frac{\partial U}{\partial b} = 0$$

From (*i*),

$$\frac{\partial U}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1)$$

$$\frac{\partial U}{\partial a} = 0 \text{ gives } 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \qquad \boxed{\sum y = na + b\sum x} \qquad \text{...}(ii)$$

Also from (*i*)

$$\frac{\partial U}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \qquad \boxed{\sum xy = a\sum x + b\sum x^2} \qquad \text{...}(iii)$$

(*ii*) and (*iii*) are called *normal equations.*

After solving (*ii*) and (*iii*), we can get the value 'a' and 'b',

$$b = \frac{\Sigma xy - \frac{1}{n}\Sigma x \Sigma y}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \qquad \text{...}(iv)$$

and

$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \bar{y} - b\bar{x} \qquad \text{...}(v)$$

(*v*) gives $\bar{y} = a + b\bar{x}$

Hence, $y = a + bx$ line passes through point $(\bar{x}, \bar{y})$.

Putting $a = \bar{y} - b\bar{x}$ in equation of line $y = a + bx$, we get

$$\boxed{y - \bar{y} = (x - \bar{x})} \qquad \ldots(vi)$$

(vi) is called *regression line* of y on x and 'b' is called the regression co-efficient of y on x and is usually denoted by b_{yx}.

Hence, eqn. (vi) can be rewritten as

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})}$$

where $\bar{x}$ and $\bar{y}$ are mean values while

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

In (iii), shifting the origin to $(\bar{x}, \bar{y})$, we get

$$\sum(x-\bar{x})(y-\bar{y}) = a\sum(x-\bar{x}) + b\sum(x-\bar{x})^2$$

$$\Rightarrow \quad nr\,\sigma_x\sigma_y = a(0) + bn\sigma_x^2$$

$$\Rightarrow \quad b = r\frac{\sigma_y}{\sigma_x}$$

$$\left\{\begin{array}{l} \because \sum(x-\bar{x}) = 0 \\ \dfrac{1}{n}\sum(x-\bar{x})^2 = \sigma_x^2 \\ \text{and } \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\sigma_y} = r \end{array}\right.$$

Hence, regression coefficient b_{yx} can also be defined as

$$\boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}}$$

where r is the coefficient of correlation while σ_x and σ_y are the standard deviation of x and y series respectively.

4.23.3. Line of Regression of *x* on *y*

Proceeding in the same, we can derive the regression line of x on y as

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where b_{xy} is the regression coefficient of x on y and is given by

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} \quad \text{or} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where the terms have their usual meanings.

4.23.4. Properties of Regression Co-efficients

Property I. Correlation co-efficient is the geometric mean between the regression co-efficients.

The co-efficient of regression are $\dfrac{r\sigma_y}{\sigma_x}$ and $\dfrac{r\sigma_x}{\sigma_y}$.

Geometric mean between them $= \sqrt{\dfrac{r\sigma_y}{\sigma_x} \times \dfrac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r =$ co-efficient of correlation.

Property II. If on of the regression co-efficients is greater than unity, the other must be less than unity.

The two regression co-efficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Let $\qquad b_{yx} > 1$, then $\frac{1}{b_{yx}} < 1 \qquad$ ...(*i*)

Since $\qquad b_{xy} \cdot b_{xy} = r^2 \leq 1 \qquad (\because -1 \leq r \leq 1)$

$\therefore \qquad b_{xy} \leq \frac{1}{b_{yx}} < 1. \qquad$ [using (*i*)]

Similarly, if $b_{xy} > 1$, then $b_{yx} < 1$.

Property III. Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.

i.e., $$\frac{b_{yx} + b_{xy}}{2} > r$$

or $$r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} > 2r$$

or $$\sigma_x^2 + \sigma_y^2 > 2\sigma_x\sigma_y$$

or $$(\sigma_x - \sigma_y)^2 > 0 \text{ which is true.}$$

Property IV. Regression co-efficients are independent of the origin but not of scale.

Proof. Let $\qquad u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$, where a, b, h and k are constants

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r \cdot \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h} b_{vu}$$

Similarly, $$b_{xy} = \frac{h}{k} b_{uv}.$$

Thus b_{yx} and b_{xy} are both independent of a and b but not of h and k.

Property V. The correlation co-efficient and the two regression co-efficients have same sign.

Proof. Regression co-efficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression co-efficient of x on $y = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive; b_{yx}, b_{xy} and r must have be same sign.

4.23.5. Angle between Two Lines of Regression

If θ is the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2},$$ where r, σ_x, σ_y have their usual meanings.

Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

Proof. Equations to the lines of regression of y on x and x on y are

$$y-\bar{y} = \frac{r\sigma_y}{\sigma_x}(x-\bar{x}) \text{ and } (x-\bar{x}) = \frac{r\sigma_x}{\sigma_y}(y-\bar{y}) \quad ...(i)$$

Their slopes are
$$m_1 = \frac{r\sigma_y}{\sigma_x} \text{ and } m_2 = \frac{\sigma_y}{r\sigma_x} \quad ...(ii)$$

$$\therefore \quad \tan\theta = \pm\frac{m_2-m_1}{1+m_2m_1} = \pm\frac{\frac{\sigma_y}{r\sigma_x}-\frac{r\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}} \quad \{\text{using } (i) \text{ and } (ii)\}$$

$$= \pm\frac{1-r^2}{r}\cdot\frac{\sigma_y}{\sigma_x}\cdot\frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2} = \pm\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Since $r^2 \le 1$ and σ_x, σ_y are positive

$\therefore$ +ve sign gives the acute angle between the lines.

Hence,
$$\tan\theta = \frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

when $r = 0$, $\theta = \frac{\pi}{2}$ $\therefore$ The two lines of regression are perpendicular to each other.

Hence, the estimated values of y is the same for all values of x and vice-versa.

When $r = \pm 1$, $\tan\theta = 0$ so that $\theta = 0$ or π.

Hence, the lines of regression coincide and there is perfect correlation between the two variates x and y.

SOLVED EXAMPLES

Example 4.95. Obtain the line of regression of y on x for the data given below:

x	1.53	1.78	2.60	2.95	3.42
y	33.50	36.30	40.00	45.80	53.50

Solution. The line of regression of y on x is given by

$$y-\bar{y} = b_{yx}(x-\bar{x}) \quad ...(i)$$

where b_{yx} is the coefficient of regression given by

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2-(\Sigma x)^2} \quad ...(ii)$$

Now we form the table as,

x	y	x^2	xy
1.53	33.50	2.3409	51.255
1.78	36.30	2.1684	64.614
2.60	40.00	6.76	104
2.95	45.80	8.7025	135.11
3.42	53.50	11.6964	182.97
$\Sigma x = 12.28$	$\Sigma y = 209.1$	$\Sigma x^2 = 32.6682$	$\Sigma xy = 537.949$

For $x = 5$

$$b_{yx} = \frac{(5\times 537.949)-(12.28\times 209.1)}{(5\times 32.6682)-(12.28)^2} = \frac{121.997}{12.543} = 9.726$$

Also, mean $\bar{x} = \frac{\Sigma x}{n} = \frac{12.28}{5} = 2.456$

and $$y = \frac{\Sigma y}{n} = \frac{20.91}{5} = 41.82$$

$\therefore$ From (i), we get

$$y - 41.82 = 9.726(x - 2.456) = 9.726x - 23.887$$

$$y = 19.932 + 9.726x$$

which is the required line of regression of y on x.

Example 4.96. The following data regarding the heights (y) and weights (x) of 100 college student are given:

$$\Sigma x = 15{,}000,\ \Sigma x^2 = 22{,}72{,}500,\ \Sigma y = 6{,}800,\ \Sigma y^2 = 4{,}63{,}025 \text{ and } \Sigma xy = 10{,}222{,}50.$$

Find the equation of regression line of height on weight.

Solution. $$\bar{x} = \frac{\Sigma x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{6800}{100} = 68$$

Regression coefficient of y on x,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(100\times 1022250)-(15000\times 6800)}{(100\times 2272500)-(15000)^2}$$

$$= \frac{102225000 - 102000000}{227250000 - 225000000} = \frac{225000}{2250000} = 0.1$$

Regression line of height (y) on weight (x) is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow \quad y - 68 = 0.1\,(x - 150)$$

$$\Rightarrow \quad y = 0.1x - 15 + 68$$

$$\Rightarrow \quad y = 0.1x + 53.$$

Example 4.97. If the regression coefficients are 0.8 and 0.2, what would be the value of co-efficient of correlation?

Solution. We know that $r^2 = b_{yx} \cdot b_{xy} = 0.8 \times 0.2 = 0.16$ $\quad [\because\ b_{yx} = 0.8;\ b_{xy} = 0.2]$

Since r has the same sign as both the regression coefficients b_{yx} and b_{xy}.

Hence, $\quad r = \sqrt{0.16} = 0.4$

Example 4.98. Calculate linear regression co-efficients from the following:

x	1	2	3	4	5	6	7	8
y	3	7	10	12	14	17	20	24

Solution. Linear regression coefficients are given by

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad \text{and} \quad b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

x	y	x^2	y^2	xy
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\sum x = 36$	$\sum y = 107$	$\sum x^2 = 204$	$\sum y^2 = 1763$	$\sum xy = 599$

Here $n = 8$

$$\therefore \quad b_{yx} = \frac{(8\times 599) - (36\times 107)}{(8\times 204) - (36)^2} = \frac{4792 - 3852}{1632 - 1296} = 2.7976$$

and

$$b_{xy} = \frac{(8\times 599) - (36\times 107)}{(8\times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

Example 4.99. From the given data obtain the two regression equations using the method of least squares.

X	2	4	6	8	10
Y	5	7	9	8	11

Solution. Computation of regression equation

X	Y	XY	X^2	Y^2
2	5	10	4	25
4	7	28	16	49
6	9	54	36	81
8	8	64	64	64
10	11	110	100	121
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 266$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

For Y on X, normal equations

$$\Sigma Y = na + b\Sigma X \qquad [\because\ n = 5]$$

$$\Sigma XY = a\Sigma X + b\Sigma Y^2$$

$\Rightarrow$
$$40 = 5a + 30b$$

and
$$266 = 30a + 220b$$

Solving these two equations, we have

$$a = 4.1,\ b = 0.65$$

So the required equation $\quad Y = a + bX$

$$Y = 4.1 + 0.65X$$

For X on Y, normal equations

$$\Sigma X = na + b\Sigma Y$$

and
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

$\Rightarrow$
$$30 = 5a + 40b$$

$$266 = 40a + 340b$$

On solving, $a = -4.4$, $b = 1.3$

$\therefore$ The required equation is

$$X = a + bY$$

$$X = -4.4 + 1.3Y$$

Example 4.100. 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.

Solution.

x	y	xy	x^2
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here,
$$n = 5$$
$$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5$$
$$\bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$\therefore$
$$b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2} = 0.95$$

Regression line of y on x is given by
$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$\Rightarrow \quad y - 19.8 = 0.95(x - 5)$$
$$\Rightarrow \quad y = 0.95x + 15.05$$

When $x = 4$ years, $y = (0.95 \times 4) + 15.05 = 18.85$ hundred rupees = ₹ 1885.

Example 4.101. In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are eligible:

Variance of $x = 9$

Regression equations: $8x - 10y + 66 = 0$, $40x - 18y = 214$.

What were: (*a*) the mean values of x and y, (*b*) the co-efficient of correlation between x and y (*c*) the standard deviation of y and .

Solution. (*a*) Since both the lines of regression pass through the point $(\bar{x}, \bar{y})$ therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad ...(i)$$
$$40\bar{x} - 18\bar{y} - 214 = 0 \quad ...(ii)$$

Multiplying (*i*) by 5,
$$40\bar{x} - 50\bar{y} + 330 = 0 \quad ...(iii)$$

Subtracting (*iii*) from (*ii*),
$$32\bar{y} - 544 = 0 \qquad \therefore \quad \bar{y} = 17$$

$\therefore$ From eqn. (*i*),
$$8\bar{x} - 170 + 66 = 0 \text{ or } 8\bar{x} = 104 \quad \therefore \quad \bar{x} = 13$$

(b) Variance of $x = \sigma_x^2 = 9$

$$\sigma_x = 3.$$

$\therefore$ The equations of lines of regression can be written as

$$y = 0.8x + 6.6 \quad \text{and} \quad x = 0.45y + 5.35$$

$\therefore$ The regression coefficient of y on x is $\dfrac{r\sigma_y}{\sigma_x} = 0.8$ ...(iv)

The regression coefficient of x on y is $\dfrac{r\sigma_x}{\sigma_y} = 0.45$ ...(v)

Multiplying (iv) and (v), $r^2 = 0.8 \times 0.45 = 0.36 \therefore \quad r = 0.6$

(+ve sign with square root is taken because regression coefficients are +ve)

(c) From (iv), we have

$$\frac{r\sigma_y}{\sigma_x} = 0.8$$

$$\Rightarrow \qquad \sigma_y = \frac{0.8\sigma_x}{r} = \frac{0.8 \times 3}{0.6} = 4.$$

Example 4.102. A study of prices of a certain commodity at Hapur and Kanpur yields the following data:

	Hapur (₹)	*Kanpur* (₹)
Average price/kilo	2.463	2.797
Standard deviation	0.326	0.207

Correlation coefficient between prices at Hapur and Kanpur is 0.774.

Estimate, from the above data, the most likely price at hapur corresponding to the price of ₹ 3.052 per kilo at Kanpur.

Solution. Here $\bar{x} = 2.463,\ \bar{y} = 2.797,\ \sigma_x = 0.326,\ \sigma_y = 0.207$

and $$r = 0.774$$

Regression co-efficient $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = (0.774)\left(\frac{0.326}{0.207}\right) = 1.218956522$$

Regression line of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\Rightarrow \qquad x - 2.463 = 1.218956522\,(y - 2.797)$$

$$x = 1.218956522y - 0.9464213$$

when $y = 3.052$, $$x = 1.218956522\,(3.052) - 0.9464213 = 2.7738$$

Example 4.103. The data for advertising and sales are given below:

	Adv. Exp (X) (₹ *lakhs*)	*Sales* (Y) (₹ *lakhs*)
Mean	**10**	**90**
S.D.	**3**	**12**

Correlation coefficient = 0.8.

(*i*) Calculate the two regression lines.

(*ii*) Find the likely sales when advertisement expenditure is ₹ 15 lakhs.

(*iii*) What should be the advertisement expenditure if the company wants to attain a sales target of ₹ 120 lakhs?

Solution. (*i*) Regression line for X on Y

$$(X - \bar{X}) = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$$X - 10 = 0.8\,\frac{3}{12}\,(Y - 90)$$

$$X - 10 = \frac{1}{5}\,(Y - 90)$$

$$X - 10 = 0.2Y - 18$$

or

$$X = 0.2Y - 8$$

Regression line for Y on X

$$Y - \bar{Y} = r\,\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$Y - 90 = 0.8\,\frac{12}{3}\,(X - 10)$$

$$Y - 90 = 3.2\,(X - 10)$$

or

$$Y = 3.2\,X + 58$$

(*ii*) Given $X = 15$, find Y,

Using Y on X line

$$Y = 3.2\,(15) + 58$$

$$Y = 106 \text{ lakhs.}$$

(*iii*) Given $Y = 120$, find X

Using X on Y line,

$$X = 0.2(120) - 8$$

$$X = 16 \text{ lakhs.}$$

Example 4.104. The lines of regression of y on x and x on y are respectively

$$y = x + 5 \text{ and } 16x - 9y = 94$$

find the variance of x if the variance of y is 16.

Solution. Regression equation of y on x is

$$y = x + 5$$

$\Rightarrow$

$$b_{yx} = 1$$

Regression equation of x on y is

$$16x - 9y = 94$$

i.e.,

$$x = \frac{9}{16}y - \frac{94}{16}$$

$\Rightarrow$

$$b_{xy} = \frac{9}{16}$$

We know that,

$$r = \sqrt{b_{xy}\, b_{yx}}$$

$$r = \sqrt{\frac{9}{16} \times 1}$$

$\Rightarrow$

$$r = \frac{3}{4}$$

Example 4.105. The regression lines of y on x and x on y are respectively $y = ax + b$, $x = cy + d$. Show that

$$\frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}},\ \bar{x} = \frac{bc + d}{1 - ac} \text{ and } \bar{y} = \frac{ad + b}{1 - ac}.$$

Solution. The regression line of y on x is

$$y = ax + b \qquad \text{...}(i)$$

$\therefore$

$$b_{yx} = a$$

The regression line of x on y is

$$x = cy + d \qquad \text{...}(ii)$$

$\therefore$

$$b_{xy} = c$$

We know that,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \qquad \text{...}(iii)$$

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \qquad \text{...}(iv)$$

Dividing (*iii*) by (*iv*), we get

$$\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{a}{c} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}$$

Since both the regression lines pass through the point $(\bar{x}, \bar{y})$ therefore,

$$\bar{y} = a\bar{x} + b \text{ and } \bar{x} = c\bar{y} + d$$

$\Rightarrow$

$$a\bar{x} - \bar{y} = -b \qquad \text{...}$$

$$\bar{x} - c\bar{y} = d \qquad \text{...(}$$

Multiplying equation (*vi*) by a and then subtracting from (c), we get

$$(ac - 1)\,\bar{y} = -ad - b \Rightarrow \bar{y} = \frac{ad + b}{1 - ac}$$

Similarly, we get

$$\bar{x} = \frac{bc + d}{1 - ac}.$$

Example 4.106. A panel of two judges, A and B, graded seven T.V. serial performances by awarding marks independently as shown in the following table:

Performance	1	2	3	4	5	6	7
Marks by A	46	42	44	40	43	41	45
Marks By B	40	38	36	35	39	37	41

The eighth T.V. performance which judge B could not attend, was awarded 37 marks by judge A. If the judge B had also been present, how many marks would be expected to have been awarded by him to the eighth T.V. performance?

Use regression analysis to answer this question.

Solution. Let the marks awarded by judge A be denoted by x and the marks awarded by judge B be denoted by y.

Here, $n = 7$;

$$\bar{x} = \frac{\sum x}{n} = \frac{46+42+44+40+43+41+45}{7} = 43$$

$$\bar{y} = \frac{\sum y}{n} = \frac{40+38+36+35+39+37+41}{7} = 38$$

Let us form the table as

z	y	xy	x^2
46	40	1840	2116
42	38	1596	1764
44	36	1584	1936
40	35	1400	1600
43	39	1677	1849
41	37	1517	1681
45	41	1845	2025
$\Sigma x = 301$	$\Sigma y = 266$	$\Sigma xy = 11459$	$\Sigma x^2 = 12971$

Regression coefficient,

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{(7\times 11459)-(301\times 266)}{(7\times 12971)-(301)^2}$$

$$= \frac{80213-80066}{90797-90601} = \frac{147}{196} = 0.75$$

Regression line of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 38 = 0.75(x-43)$$

$\Rightarrow$

$$y = 0.75x + 5.75$$

When $x = 37$,

$$y = 0.75(37) + 5.75 = 33.5 \text{ marks}$$

Hence, if judge B had also been present, 33.5 marks would be expected to have been awarded to the eighth T.V. performance.

EXERCISE 4.8

1. Find the regression coefficient y on x for the following data. Also find the value of y when $x = 11$

$$\Sigma x = 15 = \Sigma y, \Sigma x^2 = 49 = \Sigma y^2, \Sigma xy = 44, n = 5.$$

$$\left[\textbf{Ans.}\ b_{xy} = \frac{1}{4}, y = 1\right]$$

2. If $b_{xy} = -0.4$ and $b_{xy} = -0.9$, find r_{xy}. **[Ans.** 0.6]
3. Find the regression coefficient b_{xy} between x and y for the following data
$\Sigma x = 27, \Sigma y = 44, \Sigma xy = 306, \Sigma x^2 = 164,$
$\Sigma y^2 = 574, n = 4.$ **[Ans.** 2.1]
4. Find b_{yx} from the following data:

x	1	2	3	4	5
y	6	8	7	6	8

[Ans. 0.2]

5. Find the line of regression of y on x for the following data:

x	10	9	8	7	6	4	3
y	8	12	7	10	8	9	6

[Ans. $y = 8.582 + 1.094x$]

6. If the equations of the regression lines are

$$3x + 4y - 7 = 0 \text{ and } 4x + y - 5 = 0.$$

Find the correlation between x and y. **[Ans.** – 0.433]

7. For the following observations, find the regression coefficient b_{yx} and b_{xy} and hence find the correlation between x and y:

$$[(x, y)] = \{(4, 2), (2, 3), (3, 2), (4, 4), (2, 4)\}$$

$$\left[\textbf{Ans.}\ b_{yx} = \frac{1}{4}, bxy = -\frac{1}{4} \text{ and } r = -\frac{1}{4}\right]$$

8. If the regression coefficient be given by $b_{yx} = 1.6$ and $b_{xy} = 0.4$, and θ be the angle between two regression lines, find the value of $\tan\theta$. **[Ans.** 0.18]
9. Find two regression coefficient b_{yx} and b_{xy}, when the lines of regression are

$$13x - 10y + 11 = 0$$

$$2x - y - 1 = 0$$

$$\left[\textbf{Ans.}\ b_{yx} = \frac{13}{10}, b_{xy} = \frac{1}{2}\right]$$

10. $4x - 2y = 3$ and $2x - 3y = 5$ are two lines of regression of x on y and y on x respectively. Find:
(i) regression coefficients b_{xy}, b_{yx}.
(ii) correlation coefficient between x and y
(iii) estimate y when $x = 3$.

$$\left[\textbf{Ans.}\ (i)\ b_{xy} = \frac{1}{2}, b_{yx} = \frac{2}{3}; (ii)\ \rho(X, Y) = 0.577; (iii)\ y = \frac{1}{3}\right]$$

11. The equations of two regression lines in a correlation analysis are as follows:

$$3x + 2y = 26$$

and $$6x + y = 31$$

A student obtains the mean values $\bar{x} = 7$, $\bar{y} = 4$ and the value of the correlation coefficient $r = 0.5$. Do you gree with him? If not, suggest your results.[**Ans.** No. $\bar{x} = 4$, $\bar{y} = 7$, and $r = 0.5$]

12. The correlation coefficient between x and y is 0.60. If the variance of $x = 2.25$, the variance of $y = 4.00$, mean of $x = 10$ and mean of $y = 20$, find the equations of the regression lines of: (*i*) y on x (*ii*) x on y. [**Ans.** $y = 0.8x + 12$ and $x = 0.45y + 1$]

13. For observations of pairs (x, y) of the variables x and y, the following results are obtained, $\Sigma x = 110$, $\Sigma y = 70$, $\Sigma x^2 = 2500$, $\Sigma y^2 = 2000$, $\Sigma xy = 100$; and $n = 20$.
Find equation of the line of regression of x on y. Estimate the value of x if $y = 4$.

$$\left[\textbf{Ans. } x - 5.5 = -\frac{19}{117}(y - 3.5),\ x = 5.419\right]$$

14. Given $x = 4y + 5$ and $y = \dfrac{x}{17} + 4$ are the lines regression of x on y and y on x respectively.
Find the mean values of x and y and the correlation coefficient between x and y.

[**Ans.** $\bar{x} = 28$, $\bar{y} = 5.75$, $r = +0.5$]

15. The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men:

	x	y
Mean	53	142
Variance	130	165

and $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Find the approximate regression equation and use it to estimate the blood pressure of a man whose age is 45. [**And.** $y = 0.935x + 92.445$, 134.52]

16. Two variables x and y have zero means, the same variance σ^2 and zero correlation, show that $u = x \cos\alpha + y \sin\alpha$ and $v = x \sin\alpha - y \cos\alpha$ have the same variance σ^2 and zero correlation.