The second central moment gives the value of

$\therefore$ Variance $= \mu_2 = 16$

$\therefore$ Standard deviation $= \sqrt{\mu_2} = \sqrt{16} = 4$

Coefficient of skewness is given by,

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

Since μ_3 is negative, the distribution is negatively skewed. Coefficient of kurtosis is given by,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = 0.63$$

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

Ex. 14 : *The first four central moments of distribution are 0, 2.5, 0.7 and 18.75. Comment on the skewness and kurtosis of the distribution.* **(May 2009)**

Sol. : Testing of Skewness : $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Coefficient of skewness is given by,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.0314$$

Since, μ_3 is positive, the distribution is positively skewed slightly.

Testing of Kurtosis : Coefficient of kurtosis is given by

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

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3$$

Since, β_2 is exactly three, the distribution is mesokurtic.

EXERCISE 5.1

1. Find the Arithmetic Mean, Median and Standard deviation for the following frequency distribution.

x	5	9	12	15	20	24	30	35	42	49
f	3	6	8	8	9	10	8	7	6	2

Ans. $\bar{x} = 22.9851$, M = 20, $\sigma = 11.3538$

2. Age distribution of 150 life insurance policy-holders is as follows :

Age as on Nearest Birthday	Number
15 – 19.5	10
20 – 24.5	20
25 – 29.5	14
30 – 34.5	30
35 – 39.5	32
40 – 44.5	14
45 – 49.5	15
50 – 54.5	10
55 – 59.5	5

Calculate mean deviation from median age.

Ans. M.D. = 8.4284

3. The Mean and Standard deviation of 25 items is found to be 11 and 3 respectively. It was observed that one item 9 was incorrect. Calculate the Mean and Standard deviation if :

(i) The wrong item is omitted.

(ii) It is replaced by 13. **(May 2012)**

Ans. (i) $\bar{x} = 11.08, \sigma = 3.345$, (ii) $\bar{x} = 11.16, \sigma = 2.9915$

4. Following table gives the Marks obtained in a paper of statistics out of 50, by the students of two divisions :

C.I.	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
Div. A(f)	2	6	8	8	15	18	12	11	9	4
Div. B(f)	3	5	7	9	12	16	11	5	6	2

Find out which of the two divisions show greater variability. Also find the common mean and standard deviation.

Ans. B has greater variability, $\bar{x} = 26.1458$, $\sigma = 11.1267$

5. Calculate the first four moments about the mean of the following distribution. Find the coefficient of Skewness and Kurtosis.

x	1	2	3	4	5	6	7	8	9	10
f	6	15	23	42	62	60	40	24	13	5

Ans. $\mu_1 = 0$, $\mu_2 = 3.703$, $\mu_3 = 0.04256$, $\mu_4 = 37.5$, $\beta_1 = 0.00005572$, $\beta_2 = 2.8411$

6. The first four moments of a distribution about the mean value 4 are −1.5, 17, −30 and 108. Find the moments about the mean and β_1 and β_2.

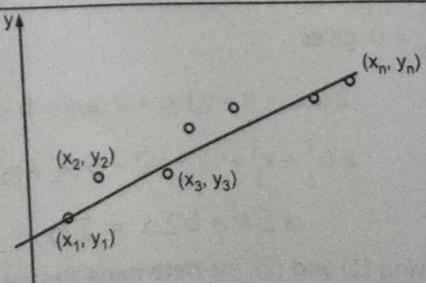
Ans. $\mu_1 = 0$, $\mu_2 = 14.75$, $\mu_3 = 39.75$, $\mu_4 = 142.31$; $\beta_1 = 0.4926$, $\beta_2 = 0.6543$.

5.7 CURVE FITTING

5.7.1 Least Square Approximation

As a result of certain experiment suppose the values of the variables (x_i, y_i) are recorded for i = 1, 2, 3, ... n.

If these points are plotted, usually it is observed that a smooth curve passes through most of these points, while some the points are slightly away from this curve. The curve passing through these points may be a first degree curve i.e. a straight line say $y = ax + b$ or a second degree parabola such as $y = ax^2 + bx + c$ or in ge



Similarly, differentiating with respect to b

$$\frac{\partial D^2}{\partial b} = \Sigma 2 (ax^2 + by^2 - x) y^2 = 0$$

or $$\Sigma (ax^2y^2 + by^4 - xy^2) = 0$$

$\frac{\partial D^2}{\partial b} = 0$ gives,

$$a \Sigma x^2y^2 + b \Sigma y^4 - \Sigma xy^2 = 0 \quad \text{... (2)}$$

Here n = 5. Various summations are as follows :

x	y	x^4	x^2y^2	x^3	y^4	xy^2
1	3.35	1	11.22	1	125.94	11.22
2	5.92	16	140.19	8	1228.25	70.095
3	8.43	81	639.58	27	5050.22	213.193
4	10.93	256	1911.44	64	14271.86	477.86
5	13.45	625	4522.56	125	32725.72	904.512
	Σ	$\Sigma x^4 = 979$	$\Sigma x^2y^2 = 7224.99$	$\Sigma x^3 = 224$	$\Sigma y^4 = 53401.99$	$\Sigma xy^2 = 1676.88$

Using in equations (1) and (2), we obtain

$$979 (a) + 7224.99 (b) - 224 = 9$$

$$979a + 7224.99b = 224$$

$$a + 7.37997b = 0.2288 \quad \text{... (3)}$$

$$\text{and} 7224.99a + 53401.99b = 1676.88$$

$$a + 7.391289b = 0.232094 \quad \text{... (4)}$$

Solving (3) and (4)

$$0.011319b = 0.0032944$$

$$b = 0.29105$$

Using value of b in equation (3)

$$a + 7.37997 \times 0.29105 = 0.2288$$

$$a = -1.91914$$

$\therefore$ Equation of the required curve is, $-1.91914x^2 + 0.29105y^2 = x$

Exercise 5.2

1. Fit a straight line of the form y = mx + c to the following data, using least square criteria

x	0	1	2	3	4	5	6
y	–4	–1	2	5	8	11	14

Ans. $y = 3x - 4$

2. If a curve of the form $x = ay^2 + by + c$ satisfies the data :

x	–6	–8	–4	6	22	44	72
y	0	1	2	3	4	5	6

Find the best values of a, b, c.

Ans. $a = 3, b = -5, c = -6$.

3. Find the best values of a, b, c assuming that the following values of x, y are connected by the relation

$$y = ax^2 + bx + c$$

x	1	2	3	4	5
y	3.38	8.25	16.6	28.5	44

Ans. $a = 1.772, b = -0.383, c = 2.103$

4. Find the law of the form by = 10^{cx} where x, y are tabulated as

x	1	1.2	1.4	1.6	1.8
y	3.67	3.01	2.46	2.02	1.65

Ans. b = 0.1, c = – 0.4343

5. If x and y are connected by the relation $ax^2 + by^2 = x$, find the values of a and b by using least square criteria

x	1	2	3	4	5
y	3.35	5.92	8.43	10.93	13.45

Ans. a = – 1.25, b = 0.2

5.8 CORRELATION

We have already considered distributions involving one variable or what we call as univariate distributions. In many problems of practical nature, we are required to deal with two or more variables. If we consider the marks obtained by a group of students in two or more subjects, the distribution will involve two or more variables. Distributions using two variables are called *Bivariate distributions*. In such distributions, we are often interested in knowing whether there exists some kind of relationship between the two variables involved. In language of statistics, this means whether there is correlation or co-variance between the two variables. If the change in one variable affects the change in the other variable, the variables are said to be ***correlated***. For example, change in rainfall will affect the crop output and thus the variables 'Rainfall recorded' and 'crop output' are correlated. Similarly, for a group of workers, the variables 'income' and 'expenditure' would be correlated. If the increase (or decrease) in one variable causes corresponding increase (or decrease) in the other, the correlation is said to be ***positive*** or ***direct***. On the other hand, if increase in the value of one variable shows a corresponding decrease in the value of the other or vice versa, the correlation is called ***negative*** or ***inverse***. As the income of a worker increases, as a natural course his expenditure also increases, hence the correlation between income and expenditure is positive or direct. Correlation between heights and weights of a group of students will also be positive. If we consider the price and demand of a certain commodity then our experience tells us that as the price of a commodity rises, its demand falls and thus the correlation between these variables is negative or inverse. Several such examples can be given. Correlation can also be classified as linear and non-linear. It is based upon the constancy of the ratio of change between the two variables. As an example, consider the values assumed by variables x and y.

x	5	8	11	15	17	19	20
	16	22	30	34	38	40	

$$= \frac{1}{n}\left[\sum (y_i - \bar{y})^2 + b_{yx}^2 \sum (x_i - \bar{x})^2 - 2\, b_{yx} \sum (x_i - \bar{x})(y_i - \bar{y})\right]$$

$$= \sigma_y^2 + b_{yx}^2\, \sigma_x^2 - 2\, b_{yx} \operatorname{cov}(x, y) \qquad \left(\because b_{yx} = r\frac{\sigma_y}{\sigma_x},\ \operatorname{cov}(x, y) = r\,\sigma_x\,\sigma_y\right)$$

$$= \sigma_y^2 + r^2\,\sigma_y^2 - 2r^2\,\sigma_y^2 = \sigma_y^2\,(1 - r^2)$$

Hence the standard error of regression estimate of y on x is

$$\boxed{S_y = \sigma_y \sqrt{1 - r^2}}$$

Note that larger the value of r^2, smaller is the error. Hence the regression estimates are close to the actual values of y_i for large r^2. If $r = \pm 1$, the correlation is perfect and the standard error is zero, which means observed values and estimated values of y agree.

The standard error of regression estimate of x on y is given by,

$$\boxed{S_x = \sigma_x \sqrt{1 - r^2}}$$

Note : The above discussion leads to conclusion that rather than r we should consider r^2 for ***testing reliability of regression estimates***. Therefore, regression analysis claims validity if r^2 is sufficiently large. The quantity r^2 is called as the ***coefficient of determination***.

EXERCISE 5.3

1. Find Karl Pearson's coefficient of correlation for the following data and determine the probable error.

x	20	22	23	25	25	28	29	30	30	34
y	18	20	22	24	21	26	26	25	27	29

[**Hint :** Probable error $= 0.6745\left(\frac{1 - r^2}{\sqrt{N}}\right)$ where, r is the coefficient correlation and N the number of pairs of observations.]

Ans. 0.952, 0.02.

2. Find the coefficient of correlation for the following table : **(Dec. 2006, 2014; May 2017) Ans.** r = 0.6013

x	10	14	18	22	26	30
y	18	12	24	6	30	36

3. The following marks have been obtained by a group of students in Engineering Mathematics.

Paper I	80	45	55	56	58	60	65	68	70	75	85
Paper II	82	56	50	48	60	62	64	65	70	74	90

Calculate the coefficient of correlation. **Ans.** 9277

4. For the following tabulated data, find the coefficient of correlation.

x / y	18	19	20	21	Total
10 – 20	4	2	2	–	8
20 – 30	5	4	6	4	19
30 – 40	6	8	10	11	35
40 – 50	4	4	6	8	22
50 – 60	–	2	4	4	10
60 – 70	–	2	3	1	6
Total	**19**	**22**	**31**	**28**	**100**

Ans. 0.25

5. Coefficient of correlation between two variables X and Y is 0.8. Their covariance is 20. The variance of X is 16. Find the standard deviation of Y series.

Ans. 1.5625.

6. Determine the equations of regression lines for the following data :

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

and obtain an estimate of y for x = 4.5. **(May 2007)**

Ans. 0.95 x + 7.25, x = 0.957 – 6.4 = 11.525.

7. Two examiners A and B independently award marks to seven students.

R. No.	1	2	3	4	5	6	7
Marks by A	40	44	28	30	44	38	31
Marks by B	32	39	26	30	38	34	28

Obtain the equations of regression lines. If examiner A awards 36 marks to Roll No. 8, what would be the marks expected to be awarded by examiner B to the same candidate ?

Ans. y = 11.885 + 0.587 x, 33.017.

8. The two regression equations of the variables x and y are

$$x = 19.13 - 0.87\,y \qquad y = 11.64 - 0.50\,x$$

Find (i) $\bar{x}$, $\bar{y}$, (ii) The correlation coefficient between x and y. **(Dec. 2006)**

Ans. $\bar{x} = 15.935$, $\bar{y} = 3.673$, $r = 0.6595$

9. Determine the reliability of estimates for the data :

x	10	14	19	26	30	34	39
y	12	16	18	26	29	35	38

Ans. $r^2 = 0.988$ high.

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

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

❐❐❐