CHAPTER 25

Statistical Methods

1. Introduction. 2. Collection and classification of data. 3. Graphical representation. 4. Comparison of frequency distributions. 5. Measures of central tendency. 6. Measures of dispersion. 7. Coefficient of variation; Relations between measures of dispersion. 8. Standard deviation of the combination of two groups. 9. Moments. 10. Skewness. 11. Kurtosis. 12. Correlation. 13. Coefficient of correlation. 14. Lines of regression. 15. Standard error of estimate. 16. Rank correlation. 17. Objective Type of Questions.

25.1 INTRODUCTION

Statistics deals with the methods for collection, classification and analysis of numerial data for drawing valid conclusions and making reasonable decisions. It has meaningful applications in production engineering, in the analysis of experimental data, etc. The importance of statistical methods in engineering is on the increase. As such we shall now introduce the student to this interesting field.

25.2 (1) COLLECTION OF DATA

The collection of data constitutes the starting point of any statistical investigation. Data may be collected for each and every unit of the whole lot (*population*), for it would ensure greater accuracy. But complete enumeration is prohibitively expensive and time consuming. As such out of a very large number of items, a few of them (a *sample*) are selected and conclusions drawn on the basis of this sample are taken to hold for the population.

(2) **Classification of data.** The data collected in the course of an inquiry is not in an easily assimilable form. As such, its proper classification is necessary for making intelligent inferences. The classification is done by dividing the raw data into a convenient number of groups according to the values of the variable and finding the frequency of the variable in each group.

Let us, for example, consider the raw data relating to marks obtained in Mechanics by a group of 64 students :

79	88	75	60	93	71	59	85
84	75	82	68	90	62	88	76
65	75	87	74	62	95	78	63
78	82	75	91	77	69	74	68
67	73	81	72	63	76	75	85
80	73	57	88	78	62	76	53
62	67	97	78	85	76	65	71
78	89	61	75	95	60	79	83

This data can conveniently be grouped and shown in a tabular form as follows :

Class	*Frequency*	*Cumulative frequency*
50—54	1	1
55—59	2	3
60—64	9	12
65—69	7	19
70—74	8	27
75—79	17	44
80—84	6	50
85—89	8	58
90—94	3	61
95—99	3	64
	Total = 64	

It would be seen from the above table that there is one student getting marks between 50—54, two students getting marks between 55—59, nine students getting marks between 60—64 and so on. Thus the 64 figure have been put into only 10 groups, called the **classes**. The width of the class is called the **class interval** and the number in that interval is called the **frequency.** The mid-point or the mid-value of the class is called the **class mark.** The above table showing the classes and the corresponding frequencies is called a *frequency table. Thus a set of raw data summarised by distributing it into a number of classes alongwith their frequencies is known as a* **frequency distribution**.

While forming a frequency distribution, the number of classes should not ordinarily exceed 20, and should not, in general, be less than 10. As far as possible, the class intervals should be of equal width.

(3) **Cumulative frequency.** In some investigations, we require the number of items less than a certain value. We add up the frequencies of the classes upto that value and call this number as the *cumulative frequency*. In the above table, the third column shows the cumulative frequencies, *i.e.*, the number of students, getting less than 54 marks, less than 59 marks and so on.

25.3 GRAPHICAL REPRESENTATION

A convenient way of representing a sample frequency distribution is by means of graphs. It gives to the eyes the general run of the observations and at the same time makes the raw data readily intelligible. We give below the important types of graphs in use :

(1) **Histogram.** A histogram is drawn by erecting rectangles over the class intervals, such that the areas of the rectangles are proportional to the class frequencies. If the class intervals are of equal size, the height of the rectangles will be proportional to the class frequencies themselves (Fig. 25.1).

(2) **Frequency polygon.** A frequency polygon for an ungrouped data can be obtained by joining points plotted with the variable values as the abscissae and the frequencies as the ordinates. For a grouped distribution, the abscissae of the points will be the mid-values of the class intervals. In case the intervals are equal, the frequency polygon can be obtained by joining the middle points of the upper sides of the rectangles of the histogram by straight lines (shown by dotted lines in Fig. 25.1). If the class intervals become very very small, the frequency polygon takes the form of a smooth curve called the *frequency curve*.

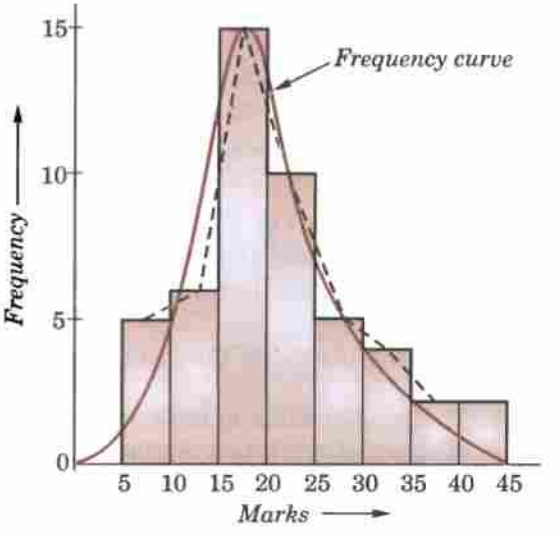


Fig. 25.1

(3) **Cumulative frequency curve-Ogive.** Very often, it is desired to show in a diagrammatic form, not the relative frequencies in the various intervals, but the cumulative frequencies above or below a given value. For example, we may wish to read off from a diagram the number or proportions of people whose income is not less than any given amount, or proportion of people whose height does not exceed any stated value. Diagrams of

this type are known as *cumulative frequency curves* or *ogives*. These are of two kinds '*more than*' or '*less than*' and typically they look somewhat like a long drawn *S* (Fig. 25.2).

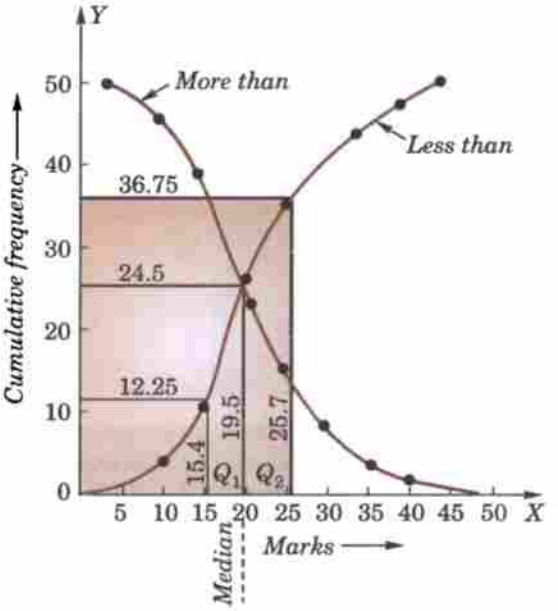


Fig. 25.2

Example 25.1. *Draw the histogram, frequency polygon, frequency curve and the ogive 'less than' and 'more than' from the following distribution of marks obtained by 49 students :*

Class (Marks group)	Frequency (No. of students)	Cumulative frequency	
		(Less than)	(More than)
5—10	5	5	49
10—15	6	11	44
15—20	15	26	38
20—25	10	36	23
25—30	5	41	13
30—35	4	45	8
35—40	2	47	4
40—45	2	49	2

Solution. In Fig. 25.1, the rectangles show the *histogram*; the dotted polygon represents the *frequency polygon* and the smooth curve is the *frequency curve*.

The *ogives* 'less than' and 'more than' are shown in Fig. 25.2.

25.4 COMPARISON OF FREQUENCY DISTRIBUTIONS

The condensation of data in the form of a frequency distribution is very useful as far as it brings a long series of observations into a compact form. But in practice, we are generally interested in comparing two or more series. The inherent inability of the human mind to grasp in its entirety even the data in the form of a frequency distribution compels us to seek for certain constants which could concisely give an insight into the important characteristics of the series. The chief constants which summarise the fundamental characteristics of the frequency distributions are (*i*) *Measures of central tendency*, (*ii*) *Measures of dispersion and* (*ii*) *Measures of skewness*.

25.5 MEASURES OF CENTRAL TENDENCY

A frequency distribution in general, shows clustering of the data around some central value. Finding of this central value or the average is of importance, as it gives a most representative value of the whole group.

Different methods give different averages which are known as the *measures of central tendency*. The commonly used measures of central value are *Mean, Median, Mode, Geometric mean and Harmonic mean.*

(1) Mean. If $x_1, x_2, x_3, ..., x_n$ are a set of n values of a variate, then the *arithmetic mean* (or simply *mean*) is given by

$$\bar{\mathbf{x}} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}, \text{ i.e. } \frac{\Sigma \mathbf{x}_i}{\mathbf{n}} \quad ...(1)$$

In a *frequency distribution*, if $x_1, x_2, ..., x_n$ be the mid-values of the class-intervals having frequencies $f_1, f_2, ..., f_n$ respectively, we have

$$\bar{\mathbf{x}} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{f_1 + f_2 + ... + f_n} = \frac{\Sigma \mathbf{f}_i\mathbf{x}_i}{\Sigma \mathbf{f}_i} \quad ...(2)$$

Calculation of mean. Direct method of computing especially when applied to grouped data involves heavy calculations and in order to avoid these, the following formulae are generally used :

I. Short-cut method $$\bar{\mathbf{x}} = \mathbf{A} + \frac{\Sigma \mathbf{f}_i d_i}{\Sigma \mathbf{f}_i} \quad ...(3)$$

II. Step-deviation method $$\bar{\mathbf{x}} = \mathbf{A} + \mathbf{h}\frac{\Sigma \mathbf{f}_i \mathbf{u}_i}{\Sigma \mathbf{f}_i} \quad ...(4)$$

where $d = x - A$ and $u = (x - A)/h$, A being an arbitrary origin and h the equal class interval.

Proof. If $x_1, x_2, ..., x_n$ are the mid-values of the classes with frequencies $f_1, f_2, ..., f_n$, we have

$$\Sigma f_i x_i = \Sigma f_i (A + d_i) = A\Sigma f_i + \Sigma f_i d_i$$

$$\therefore \quad \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

Further $u_i = d_i/h$ or $d_i = hu_i$. Substituting this value in (3), we get (4).

Obs. *The algebraic sum of the deviations of all the variables from their mean is zero, for*

$$\Sigma f_i (x_i - \bar{x}) = \Sigma f_i x_i - \bar{x}\,\Sigma f_i = \Sigma f_i x_i - \frac{\Sigma f_i x_i}{\Sigma f_i} \cdot \Sigma f_i = 0.$$

Cor. *If $\bar{x}_1, \bar{x}_2$ be the means of two samples of size n_1 and n_2, then the mean $\bar{x}$ of the combined sample of size $n_1 + n_2$ is given by*

$$\bar{\mathbf{x}} = \frac{\mathbf{n}_1\bar{\mathbf{x}}_1 + \mathbf{n}_2\bar{\mathbf{x}}_2}{\mathbf{n}_1 + \mathbf{n}_2}$$

For $n_1\bar{x}_1$ = sum of all observations of the first sample,

and $n_2\bar{x}_2$ = sum of all observations of the second sample.

$\therefore$ sum of the observations of the combined sample $= n_1\bar{x}_1 + n_2\bar{x}_2$.

Also number of the observations in the combined sample $= n_1 + n_2$.

$\therefore$ mean of the combined sample $= \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

Example 25.2. *The following is the frequency distribution of a random sample of weekly earnings of 509 employees :*

Weekly earnings :	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
No. of employees :	3	6	10	15	24	42	75	90	79	55	36	26	19	13	9	7

Calculate the average weekly earnings.

Solution. The calculations are arranged in the following table. The arbitrary origin is generally taken as the value corresponding to the maximum frequency.

By direct method, we have

$$\text{Mean } \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{13,315}{509} = 26.16$$

By step-deviation method, we have

$$\bar{x} = A + h\frac{\Sigma fu}{\Sigma f} = 25 + 2 \times \frac{295}{509}$$

$$= 25 + 1.16 = 26.16, \text{ which is same as found above.}$$

Weekly earnings	*Mid value* x	*No. of employees* f	*Step deviations* $f \times x$	$u = (x - 25)/2$	$f \times u$
10—12	11	3	33	−7	−21
12—14	13	6	78	−6	−36
14—16	15	10	150	−5	−50
16—18	17	15	255	−4	−60
18—20	19	24	456	−3	−72
20—22	21	42	882	−2	−84
22—24	23	75	1725	−1	−75
24—26	25	90	2250	0	−398
26—28	27	79	2133	1	79
28—30	29	55	1595	2	110
30—32	31	36	1116	3	108
32—34	33	26	858	4	104
34—36	35	19	665	5	95
36—38	37	13	481	6	78
38—40	39	9	351	7	63
40—42	41	7	287	8	56
					+693
		$\Sigma f = 509$	$\Sigma fx = 13{,}315$		$\Sigma fu = 295$

(2) Median. *If the values of a variable are arranged in the ascending order of magnitude, the median is the middle item if the number is odd and is the mean of the two middle items if the number is even.* Thus the median is equal to the mid-value, *i.e.*, the value which divides the total frequency into two equal parts.

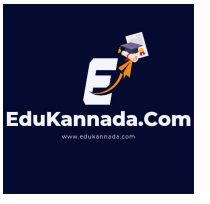
For the grouped data,

$$\textbf{Median} = \mathbf{L} + \frac{(\frac{1}{2}\mathbf{N} - \mathbf{C})}{\mathbf{f}} \times \mathbf{h}$$

where L = lower limit of the median class, N = total frequency,

f = frequency of the median class, h = width of the median class,

and C = cumulative frequency upto the class preceding the median class.



Quartiles. Quartiles are those values which divide the frequency into four equal parts, when the values are arranged in the ascending order of magnitude. The **lower quartile** (Q_1) is mid-way between the lower extreme and the median. The **upper quartile** (Q_3) is midway between the median the upper extreme.

For the grouped data, these are calculated by the formulae :

$$\mathbf{Q_1} = \mathbf{L} + \frac{(\frac{1}{4}\mathbf{N} - \mathbf{C})}{\mathbf{f}} \times \mathbf{h}$$

and

$$\mathbf{Q_3} = \mathbf{L} + \frac{(\frac{3}{4}\mathbf{N} - \mathbf{C})}{\mathbf{f}} \times \mathbf{h}$$

where L = lower limit of the class in which Q_1 or Q_3 lies, f = frequency of this class, h = width of the class

and C = cumulative frequency upto the class preceding the class in which Q_1 or Q_3 lies.

The difference between the upper and lower quartiles, i.e., Q_3—Q_1 is called the **inter-quartile range.**

Obs. The ogives give a ready method of marking on the curve the values of the median and the quartiles. The two ogives 'less than' and 'more than' cut each other at the median (Fig. 25.2).

(3) Mode. *The mode is defined as that value of the variable which occurs most frequently, i.e., the value of the maximum frequency.*

For a grouped distribution, it is given by the formula

$$\textbf{Mode} = \mathbf{L} + \frac{\Delta_1}{\Delta_1 + \Delta_2}\mathbf{h}$$

where L = lower limit of the class containing the mode,

Δ_1 = excess of modal frequency over frequency of preceding class,

Δ_2 = excess of modal frequency over following class,

and h = size of modal class.

For a frequency curve (Fig. 25.1), the abscissa of the highest ordinate determines the value of the mode. There may be one or more modes in a frequency curve. Curves having a single mode are termed as *unimodal*, those having two modes as *bi-modal* and those having more than two modes as *multi-modal*.

Obs. In a symmetrical distribution, the mean, median and mode coincide. For other distributions, however, they are different and are known to be connected by the empirical relationship :

Mean – Mode = 3(Mean – Median).

Example 25.3. *Calculate median and the lower and upper quartiles from the distribution of marks obtained by 49 students of example 25.1. Find also the semi-interquartile range and the mode.*

Solution. *Median* (or 49/2) falls in the class (15—20) and is given by

$$15 + \frac{(49/2) - 11}{15} \times 5 = 15 + \frac{13.5}{3} = 19.5 \text{ marks.}$$

Lower quartile Q_1 (or 49/4) = 12.25) also falls in the class 15—20.

$$\therefore \quad Q_1 = 15 + \frac{(49/4) - 11}{15} \times 5 = 15 + \frac{12.5}{3} = 15.4 \text{ marks}$$

Upper quartile (or $\frac{3}{4} \times 49 = 36.75$) falls in the class 25—30.

$$\therefore \quad Q_3 = 25 + \frac{36.75 - 36}{5} \times 5 = 25.75 \text{ marks.}$$

$$\textit{Semi-interquartile range} = \tfrac{1}{2}(Q_3 - Q_1) = \frac{25.75 - 15.4}{2} = \frac{10.35}{2} = 5.175.$$

Mode. It is seen that the mode value falls in the class 15—20. Employing the formula for the grouped distribution, we have

$$\textit{Mode} = 15 + \frac{15 - 6}{(15 - 6) + (15 - 10)} \times 5 = 18.2 \text{ marks.}$$

Obs. In Fig. 25.2, the ogives meet at a point whose abscissa is 19.5 which is the *median* of the distribution. The values for the lower and upper quartiles are similarly seen to be 15.4 (for frequency 12.25) and 25.7 (for frequency 36.75).

Example 25.4. *Given below are the marks obtained by a batch of 20 students in a certain class test in Physics and Chemistry.*

Roll No. of students	*Marks in Physics*	*Marks in Chemistry*	*Roll No. of students*	*Marks in Physics*	*Marks in Chemistry*
1	53	58	11	25	10
2	54	55	12	42	42
3	52	25	13	33	15
4	32	32	14	48	46
5	30	26	15	72	50
6	60	85	16	51	64
7	47	44	17	45	39
8	46	80	18	33	38
9	35	33	19	65	30
10	28	72	20	29	36

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

Solution. The subject for which the value of the median is higher will be the subject in which the level of knowledge of the students is higher. To find the median in each case, we arrange the marks in ascending order of magnitude :

Sr. No.	Marks in Physics	Marks in Chemistry	Sr. No.	Marks in Physics	Marks in Chemistry
1	25	10	11	46	42
2	28	15	12	47	44
3	29	25	13	48	46
4	30	26	14	51	50
5	32	30	15	52	55
6	33	32	16	53	58
7	33	33	17	54	64
8	35	36	18	60	72
9	42	38	19	65	80
10	45	39	20	72	85

Median marks in Physics = A.M. of marks of 10th and 11th terms

$$= \frac{45 + 46}{2} = 45.5$$

Median marks in Chemistry = A.M. of marks of 10th and 11th items.

$$= \frac{39 + 42}{2} = 40.5$$

Since the median marks in Physics is greater than the median marks in Chemistry; the level of knowledge in Physics is higher.

Example 25.5. *An incomplete frequency distribution is given as below :*

Variable :	*10—20*	*20—30*	*30—40*	*40—50*	*50—60*	*60—70*	*70—80*
Frequency :	*12*	*30*	*?*	*65*	*?*	*25*	*18*

Given that the total frequency is 229 and median is 46, find the missing frequencies.

Solution. Let f_1, f_2 be the missing frequencies of the classes 30—40 and 50—60 respectively.

Since the median lies in the class 40—50,

$$\therefore \quad 46 = 40 + \frac{229/2 - (12 + 30 + f_1)}{65} \times 10$$

which gives $f_1 = 33.5$ which can be taken as 34.

$$\therefore \quad f_2 = 229 - (12 + 30 + 34 + 65 + 25 + 18) = 45.$$

(4) Geometric mean. If $x_1, x_2, \ldots, x_n$ are a set of n observations, then the *geometric mean* is given by

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

or
$$\log \text{G.M.} = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) \qquad \ldots(1)$$

In a frequency distribution, let $x_1, x_2, \ldots, x_n$ be the central values with corresponding frequencies $f_1, f_2, \ldots, f_n$, we have

$$\text{G.M.} = \left[(x_1)^{f_1} \cdot (x_2)^{f_2} \ldots (x_n)^{f_n}\right]^{1/n} \quad \text{where } n = \Sigma f_i.$$

or
$$\log \text{G.M.} = \frac{1}{n}[f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n] \qquad \ldots(2)$$

Hence (1) and (2) show that logarithm of G.M. = A.M. of logarithms of the values.

(5) Harmonic mean. If $x_1, x_2, \ldots x_n$ be a set of n observations, then the *harmonic mean is defined as the reciprocal of the (arithmetic) mean of the reciprocals of the quantities.* Thus

$$\text{H.M.} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)}$$

In a frequency distribution, $\text{H.M.} = \dfrac{1}{\frac{1}{n}\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}\right)}$ where $n = \Sigma f_i$.

Example 25.6. *Three cities A, B, C are equidistant from each other. A motorist travels from A to B at 30 km/hr, from B to C at 40 km/hr, from C to A at 50 km/hr. Determine the average speed.*

Solution. Let $AB = BC = CA = s$ km

Time taken to travel from A to $B = s/30$

Time taken to travel from B to $C = s/40$

Time taken to travel from C to $A = s/50$

$$\therefore \quad \text{average time taken} \quad = \frac{1}{3}\left(\frac{s}{30} + \frac{s}{40} + \frac{s}{50}\right)$$

$$\text{Thus the average speed} \quad = \frac{s}{\frac{1}{3}\left(\frac{s}{30} + \frac{s}{40} + \frac{s}{50}\right)}$$

In other words, the average speed is the harmonic mean of 30, 40, 50 km/hr.

$$\text{Hence the average speed} \quad = \frac{1}{\frac{1}{3}\left(\frac{1}{30} + \frac{1}{40} + \frac{1}{50}\right)} = 38.3 \text{ km/hr.}$$

Obs. Of the various measures of central tendency, *the mean* is the most important for it can be computed easily. *The median*, though more easily calculable, cannot be applied with ease to theoretical analysis. Median is of advantage when there are exceptionally large and small values at the ends of the distribution.

The mode, though most easily calculated, has the least significance. It is particularly misleading in distributions which are small in numbers or highly unsymmetrical.

The geometrical mean though difficult to compute, finds application in cases like populations where we are concerned with a quantity whose changes tend to be directly proportional to the quantity itself.

The harmonic mean is useful in limited situations where time and rate or prices are involved.

PROBLEMS 25.1

1. Draw the histogram and frequency polygon for the following distribution. Also calculate the arithmetic mean :

Class interval :	0—99	100—199	200—299	300—399	400—499	500—599	600—699	700—799
Frequency :	10	54	184	264	246	40	1	1

2. The following marks were given to a batch of candidates :

66	62	45	79	32	51	56	60	51	49
25	42	54	54	58	70	43	58	50	52
38	67	50	59	48	65	71	30	46	55
82	51	63	45	53	40	35	56	70	52
67	55	57	30	63	42	74	58	44	55

Draw a cumulative frequency curve.

Hence find the proportion of candidates securing more than 50 marks. Also mark off the median, the first and third quartiles.

3. Find the mean, median and mode for the following :

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

(Kerala, 1990)

4. Calculate mean, median and mode of the following data relating to weight of 120 articles :

Weight (in gm) :	0—10	10—20	20—30	30—40	40—50	50—60
No. of articles :	14	17	22	26	23	18

5. The population of a country was 300 million in 1971. It became 520 million in 1989. Calculate the percentage compound rate of growth per annum.

[**Hint.** Use $P_n = P_0(1 + r)^n$, r being the growth rate.]

6. The number of divorces per 1000 marriages in the United States increased from 84 in 1970 to 108 in 1990. Find the annual increase of the divorce rate for the period 1970 to 1990.

7. An aeroplane flies along the four sides of a square at speeds of 100, 200, 300 and 400 km/hr. respectively. What is the average speed of the plane in its flight around the square.

8. A man having to drive 90 km. wishes to achieve an average speed of 30 km/hr. For the first half of the journey, he averages only 20 km/hr. What must be his average speed for the second half of the journey if his overall average is to be 30 km/hr.

9. Following table gives the cumulative frequency of the age of a group of 199 teachers. Find the mean and median age of the group.

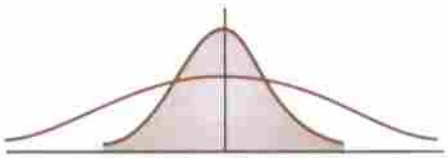
Age in years :	20—25	25—30	30—35	35—40	40—45	45—50	50—55	55—60	60—65	65—70
Cum. frequ. :	21	40	90	130	146	166	176	186	195	199

10. Recast the following cumulative table in the form of an ordinary frequency distribution and determine the median and the mode :

No. of days absent	*No. of students*	*No. of days absent*	*No. of students*
Less than 5	29	Less than 30	644
Less than 10	224	Less than 35	650
Less than 15	465	Less than 40	653
Less than 20	582	Less than 45	655
Less than 25	634		

25.6 MEASURES OF DISPERSION

Although measures of central tendency do exhibit one of the important characteristics of a distribution, yet they fail to give any idea as to how the individual values differ from the central value, *i.e.*, whether they are closely packed around the central value or widely scattered away from it. Two distributions may have the same mean and the same total frequency, yet they may differ in the extent to which the individual values may be spread about the average (*See* Fig. 25.3). The magnitude of such a variation is called *dispersion*. The important measures of dispersion are given below :



Same mean different dispersion

Fig. 25.3

(1) Range. *This is the simplest measure of dispersion and is given by the difference between the greatest and the least values in the distribution.* If the weekly wages of a group of labourers are

₹ 21 23 28 25 35 42 39 48

then *range* = Max. value – Min. value = 48 – 21 = ₹ 27.

(2) Quartile deviation or semi-interquartile range. One half of the interquartile range is called *quartile deviation,* or *semi-interquartile range.* If Q_1 and Q_3 are the first and third quartiles, the semi-interquartile range

$$Q = \tfrac{1}{2}(Q_3 - Q_1).$$

(3) Mean deviation. The mean deviation is the mean of the absolute differences of the values from the mean, median or mode. Thus *mean deviation* (*M.D.*)

$$= \frac{1}{n}\Sigma f_i \, |x_i - A|$$

where A is either the mean or the median or the mode. As the positive and negative differences have equal effects, only the absolute value of differences is taken into account.

(4) Standard deviation. The most important and the most powerful measure of dispersion is the *standard deviation* (*S.D.*) : generally denoted by σ. It is computed as the square root of the mean of the squares of the differences of the *variate* values from their mean.

Thus *standard deviation* (*S.D.*)

$$\sigma = \sqrt{\left[\frac{\Sigma \mathbf{f}_i(\mathbf{x}_i - \bar{\mathbf{x}})^2}{\mathbf{N}}\right]} \qquad ...(1)$$

where N is the total frequency Σf_i.

If however, the deviations are measured from any other value, say A, instead of $\bar{x}$, it is called the *root-mean-square deviation*.

The square of the standard deviation is known as the **variance.**

Calculation of S.D. The change of origin and the change of scale considerably reduces the labour in the calculation of standard deviation. The formulae for the computation of σ are as follows :

I. Short-cut method

$$\sigma = \sqrt{\left[\frac{\Sigma \mathbf{f}_i \mathbf{d}_i^2}{\Sigma \mathbf{f}_i} - \left(\frac{\Sigma \mathbf{f}_i \mathbf{d}_i}{\Sigma \mathbf{f}_i}\right)^2\right]} \quad \ldots(2)$$

II. Step-deviation method

$$\sigma = \mathbf{h}\sqrt{\left\{\frac{\Sigma \mathbf{f}_i \mathbf{d}_i'^2}{\Sigma \mathbf{f}_i} - \left(\frac{\Sigma \mathbf{f}_i \mathbf{d}_i'}{\Sigma \mathbf{f}_i}\right)^2\right\}} \quad \ldots(3)$$

where $d_i = x_i - A$ and $d'_i = (x_i - A)/h$, being the assumed mean and h the equal class interval.

Proof. We know that $x_i - \bar{x} = (x_i - A) - (\bar{x} - A)$

$$\therefore \quad \Sigma f_i (x_i - \bar{x})^2 = \Sigma f_i |d_i - (\bar{x} - A)|^2 = \Sigma f_i d_i^2 + (\bar{x} - A)^2 \Sigma f_i - 2(\bar{x} - A)\Sigma f_i d_i$$

$$= \Sigma f_i d_i^2 - \frac{(\Sigma f_i d_i)^2}{\Sigma f_i} \qquad \left[\because \ \bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}\right]$$

Hence $$\sigma^2 = \frac{\Sigma f_i (x - \bar{x})^2}{\Sigma f_i} = \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2$$

Further $d'_i = (x_i - A)/h = d_i/h$ or $d_i = hd'_i$, then substituting this value in (2), we get (3).

Obs. *The root mean square deviation is least when measured from the mean.*

The root mean square deviation is given by

$$s^2 = \frac{\Sigma f_i d_i^2}{\Sigma f_i} \quad \text{and} \quad \frac{\Sigma f_i d_i}{\Sigma f_i} = \left[A + \frac{\Sigma f_i d_i}{\Sigma f_i}\right] - A = \bar{x} - A$$

$\therefore$ from (2), we have $s^2 = \sigma^2 + (\bar{x} - A)^2$ ...(4)

This shows that s^2 is always $> \sigma^2$ and the least value of $s^2 = \sigma^2$. This occurs when $A = \bar{x}$.

25.7 (1) COEFFICIENT OF VARIATION

The ratio of the standard deviation to the mean, is known as the *coefficient of variation*. As this is a ratio having no dimension, it is used for comparing the variations between the two groups with different means. It is often expressed as a percentage.

$$\therefore \quad \textit{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$

(2) Relations between measures of dispersion

(*i*) *Quartile deviation = 2/3 (standard deviation)*

(*ii*) *Mean deviation = 4/5 (standard deviation)*

25.8 STANDARD DEVIATION OF THE COMBINATION OF TWO GROUPS

If m_1, σ_1 be the mean and S.D. of a sample of size n_1 and m_2, σ_2 be those for a sample of size n_2, then the S.D. σ of the combined sample of size $n_1 + n_2$ is given by

$$\mathbf{(n_1 + n_2)\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1D_1^2 + n_2D_2^2}$$

where $D_i = m_i - m$, m being the mean of combined sample.

From (4), we have $ns^2 = n\sigma^2 + n(\bar{x} - A)^2$ where n is the size of the sample.

i.e. sum of the squares of the deviations from $A = n\sigma^2 + n(\bar{x} - A)^2$.

Now let us apply this result to the first given sample taking A at m. Then, sum of the squares of the deviations of n_1 items from $m = n_1\sigma_1^2 + n_1(m_1 - m)^2$ ...(5)

Similarly for the second given sample taking A at m, sum of the squares of the deviations of n_2 items from $m = n_2\sigma_2^2 + n_2(m_2 - m)^2$ ...(6)

Adding (5) and (6), sum of the squares of the deviations of $n_1 + n_2$ items from m

$$= n_1\sigma_1^2 + n_2\sigma_2^2 + n_1(m_1 - m)^2 + n_2(m_2 - m)^2$$

$$\therefore \quad (n_1 + n_2)\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1D_1^2 + n_2D_2^2$$

This result can be extended to the combination of any number of samples, giving a result of the form

$$(\Sigma n_i)\sigma^2 = \Sigma(n_i\sigma_i^2) + \Sigma(n_iD_i^2).$$

Example. 25.7. *Calculate the mean and standard deviation for the following :*

Size of item :	6	7	8	9	10	11	12
Frequency :	3	6	9	13	8	5	4

(V.T.U., 2001)

Solution. The calculations are arranged as follows :

Size of item x	*Frequency f*	*Deviation d = x − 9*	*f × d*	*f × d²*
6	3	−3	−9	27
7	6	−2	−12	24
8	9	−1	−9	9
9	13	0	0	0
10	8	1	8	8
11	5	2	10	20
12	4	3	12	36
	$\Sigma f = 48$		$\Sigma fd = 0$	$\Sigma fd^2 = 124$

$$\therefore \qquad mean = 9 + \frac{\Sigma fd}{\Sigma f} = 9$$

$$Standard\ deviation = \sqrt{\left[\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}\right]} = \sqrt{\left(\frac{\Sigma fd^2}{\Sigma f}\right)} = \sqrt{\left(\frac{124}{48}\right)} = 1.607.$$

Example 24.8. *Calculate the mean and standard deviation of the following frequency distribution :*

Weekly wages is ₹	*No. of men*
4.5—12.5	4
12.5—20.5	24
20.5—28.5	21
28.5—36.5	18
36.5—44.5	5
44.5—52.5	3
52.5—60.5	5
60.5—68.5	8
68.5—76.5	2

Solution. The calculations are arranged in the table below :

Wages class ₹	*Mid value x*	*No. of men f*	*Step deviation* $d' = \frac{x - 32.5}{8}$	*fd'*	*fd'²*
4.5—12.5	8.5	4	−3	−12	36
12.5—20.5	16.5	24	−2	−48	96
20.5—28.5	24.5	21	−1	−21	21
28.5—36.5	32.5	18	0	0	0
36.5—44.5	40.5	5	1	5	5
44.5—52.5	48.5	3	2	6	12
52.5—60.5	56.5	5	3	15	45
60.5—68.5	64.5	8	4	32	128
68.5—76.5	72.5	2	5	10	50
		$\Sigma f = 90$		$\Sigma fd' = -13$	$\Sigma fd'^2 = 393$

$$\therefore \qquad mean\ wage = 32.5 + 8 \times \frac{\Sigma fd'}{\Sigma f} = 32.5 + 8\left(\frac{-13}{90}\right) = ₹\ 31.35$$

$$Standard\ deviation = 8\sqrt{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2} = 8\sqrt{\frac{393}{90} - \left(\frac{-13}{90}\right)^2} = ₹\ 16.64.$$

Example 25.9. *The following are scores of two batsmen A and B in a series of innings :*

A :	12	115	6	73	7	19	119	36	84	29
B :	47	12	16	42	4	51	37	48	13	0

Who is the better score getter and who is more consistent ? (V.T.U., 2004)

Solution. Let x denote score of A and y that of B.

Taking 51 as the origin, we prepare the following table :

x	$d(=x-51)$	d^2	y	$\delta(=y-51)$	δ^2
12	−39	1521	47	−4	16
115	64	4096	12	−39	1521
6	−45	2025	16	−35	1225
73	22	484	42	−9	81
7	−44	1936	4	−47	2209
19	−32	1024	51	0	0
119	68	4624	37	−14	196
36	−15	225	48	−3	9
84	33	1089	13	−38	1444
29	−22	484	0	−51	2601
Total	−10	17508		−240	9302

For A, $$\text{A.M. } \bar{x} = 51 + \frac{\Sigma d}{n} = 51 - \frac{10}{10} = 50$$

$$\text{S.D. } \sigma_1 = \sqrt{\left\{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2\right\}} = \sqrt{\{1750.8 - (-1)^2\}} = 41.8$$

$$\therefore \text{ coefficient of variation } = \frac{\sigma_1}{\bar{x}} \times 100 = \frac{41.8}{50} \times 100 = 83.6\%$$

For B, $$\text{A.M. } \bar{y} = 51 + \frac{\Sigma \delta}{n} = 51 - \frac{240}{10} = 27$$

$$\text{S.D. } \sigma_2 = \sqrt{\left\{\frac{\Sigma \delta^2}{n} - \left(\frac{\Sigma \delta}{n}\right)^2\right\}} = \sqrt{\{930.2 - (-24)^2\}} = 18.8$$

$$\therefore \text{ coefficient of variation } = \frac{\sigma^2}{\bar{y}} \times 100 = \frac{18.8}{27} \times 100 = 69.6\%$$

Since the A.M. of A > A.M. of B, it follows that A is a better score getter (*i.e.*, more efficient) than B.

Since the coefficient of variation of B < the coefficient of variation of A, it means that B is more consistent than A. Thus even though A is a better player, he is less consistent.

Example 25.10. *The numbers examined, the mean weight and S.D. in each group of examination by three medical examiners are given below. Find the mean weight and S.D. of the entire data when grouped together.*

Med. Exam.	*No. Examined*	*Mean Wt. (lbs.)*	*S.D. (lbs.)*
A	*50*	*113*	*6*
B	*60*	*120*	*7*
C	*90*	*115*	*8*

Solution. We have $n_1 = 50,\ \bar{x}_1 = 113,\ \sigma_1 = 6$

$n_2 = 60,\ \bar{x}_2 = 120,\ \sigma_2 = 7$

$n_3 = 90,\ \bar{x}_3 = 115,\ \sigma_3 = 8.$

If $\bar{x}$ is the mean of the entire data,

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{50\times113 + 60\times120 + 90\times115}{50 + 60 + 90} = \frac{23200}{200} = 116 \text{ lb.}$$

If σ is the S.D. of the entire data,

$$N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1D_1^2 + n_2D_2^2 + n_3D_3^2$$

where $N = n_1 + n_2 + n_3 = 200$, $D_1 = \bar{x}_1 - \bar{x} = -3$, $D_2 = \bar{x}_2 - \bar{x} = 4$ and $D_3 = \bar{x}_3 - \bar{x} = -1$.

$$\therefore \quad 200\sigma^2 = 50 \times 36 + 60 \times 49 + 90 \times 64 + 50 \times 9 + 60 \times 16 + 90 \times 1$$
$$= 1800 + 2940 + 5760 + 450 + 960 + 90$$
$$\sigma^2 = \frac{12000}{200} = 60. \quad \text{Hence} \quad \sigma = \sqrt{60} = 7.746 \text{ lb.}$$

PROBLEMS 25.2

1. The crushing strength of 8 cement concrete experimental blocks, in metric tonnes per sq. cm., was 4.8, 4.2, 5.1, 3.8, 4.4, 4.7, 4.1 and 4.5. Find the mean crushing strength and the standard deviation.
2. Show that the variance of the first n positive integers is $\frac{1}{12}(n^2 - 1)$. *(V.T.U., 2003)*
3. The mean of five items of an observation is 4 and the variance is 5.2. If three of the items are 1, 2 and 6, then find the other two. *(V.T.U., 2002)*
4. For the distribution

x:	5	6	7	8	9	10	11	12	13	14	15
f:	18	15	34	47	68	90	80	62	35	27	11

find the mean, median and lower and upper quartiles, variance and the standard deviation.

5. The following table shows the marks obtained by 100 candidates in an examination. Calculate the mean, median and standard deviation :

Marks obtained :	1—10	11—20	21—30	31—40	41—50	51—60
No. of candidates :	3	16	26	31	16	8

(Osmania, 2003 S ; V.T.U., 2003 S)

6. Compute the quartile deviation and standard deviation for the following :

x:	100—109	110—119	120—129	130—139	140—149	150—159	160—169	170—179
f:	15	44	133	150	125	82	35	16

7. Calculate (*i*) mean deviation about the mean, (*ii*) mean deviation about the median for the following distribution :

Class :	3—4.9	5—6.9	7—8.9	9—10.9	11—12.9	13—14.9	15—16.9
f :	5	8	30	82	45	24	6

(Madras, 2002)

8. Two observers bring the following two sets of data which represent measurements of the same quantity :

I.	105.1	103.4	104.2	104.7	104.8	105.0	104.9
II.	105.3	105.1	104.8	105.2	106.7	102.9	103.1

Calculate the standard deviation in each case. Which set of data is more reliable ? Can the same conclusion be reached by calculating the mean deviation ?

Obs. The smaller the coefficient of variation, the greater is the *reliability* or *consistency* in the data.

9. The heights and weights of the 10 armymen are given below. In which characteristics are they more variable ?

Height in cm.	170	172	168	177	179	171	173	178	173	179
Weight in kg.	75	74	75	76	77	73	76	75	74	75

10. The index number of prices of two articles A and B for six consecutive weeks are given below :

A:	314	326	336	368	404	412
B:	330	331	320	318	321	330

Find which has a more variable price ?

11. The scores of two golfers A and B in 12 rounds are given below. Who is the better player and who is the more consistent player ?

A:	74	75	78	72	78	77	79	81	79	76	72	71
B:	87	84	80	88	89	85	86	82	82	79	86	80

12. The scores obtained by two batsmen A and B in 10 matches are given below :

A:	30	44	66	62	60	34	80	46	20	38
B:	34	46	70	38	55	48	60	34	45	30

Calculating mean, S.D. and coefficient of variation for each batsman, determine who is more efficient and who is more consistent.

13. Find the mean and standard deviation of the following two samples put together :

Sample No.	*Size*	*Mean*	*S.D.*
1	50	158	5.1
2	60	164	4.6

14. A distribution consists of three components with frequencies 200, 250 and 300 having means 25, 10 and 15 and S.Ds. 3, 4 and 5 respectively. Show that the mean of the combined distribution is 16 and its S.D. is 7.2 approximately.

25.9 (1) MOMENTS

The rth moment about the mean $\bar{x}$ of a distribution is denoted by μ_r and is given by

$$\mu_r = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^r \qquad ...(1)$$

The corresponding moment about any point a is defined as

$$\mu_r' = \frac{1}{N}\Sigma f_i(x_i - a)^r \qquad ...(2)$$

In particular, we have $\mu_0 = \mu_0' = 1$ ...(3)

$$\mu_1 = \frac{1}{N}\Sigma f_i(x_i - \bar{x}) = 0;\ \mu_1' = \frac{1}{N}\Sigma f_i(x_i - a) = \bar{x} - a = d, \text{ say} \qquad ...(4)$$

$$\mu_2 = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^2 = \sigma^2. \qquad ...(5)$$

(2) Moments about the mean in terms of moments about any point.

We have

$$\mu_r = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^r = \frac{1}{N}\Sigma f[(x_i - a) - (\bar{x} - a)]^r$$

$$= \frac{1}{N}\Sigma f_i(X_i - d)^r \quad \text{where} \quad X_i = x_i - a,\ d = \bar{x} - a.$$

$$= \frac{1}{N}[\Sigma f_i X_i^r - {}^rC_1 d\Sigma f_i X_i^{r-1} + {}^rC_2 d^2 \Sigma f_i X_i^{r-2} - \ldots]$$

$$= \mu_r' - {}^rC_1\, d\mu_{r-1}' + {}^rC_2 d^2 \mu_{r-2}'' - \ldots\ldots \qquad ...(6)$$

In particular,

$$\mu_2 = \mu'_2 - \mu_1'^2 \qquad ...(7)$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 \qquad ...(8)$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4 \qquad ...(9)$$

These three results should be committed to memory. It should be noted that in each of these relations, the sum of the coefficients of the various terms on the right side is zero. Also each term on the right side is of the same dimension as the term on the left.

25.10 SKEWNESS

Skewness measures the degree of asymmetry or the departure from symmetry. If the frequency curve has a longer 'tail' to the right, *i.e.*, the mean is to the right of the mode [as in Fig. 25.4 (*a*)], then the distribution is said to have *positive skewness*. If the curve is more elongated to the left, then it is said to have *negative skewness* [Fig. 25.4 (*b*)].

The following three measures of skewness deserve mention :

(*i*) *Pearson's* coefficient of skewness* $= \dfrac{\text{mean} - \text{mode}}{\sigma}$

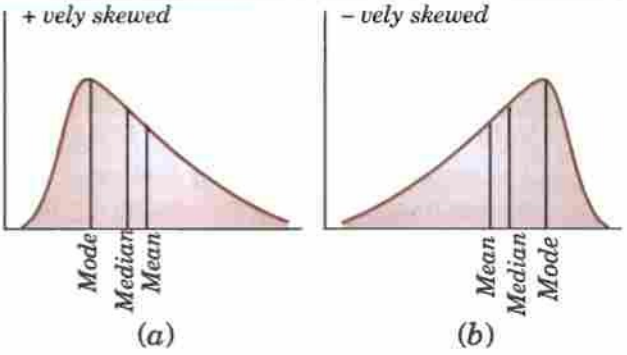


Fig. 25.4

* After the English statistician and biologist *Karl Pearson* (1857–1936) who did pioneering work and found the English school of statistics.

(*ii*) *Quartile coefficient of skewness* $= \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$

Its value always lies between – 1 and + 1.

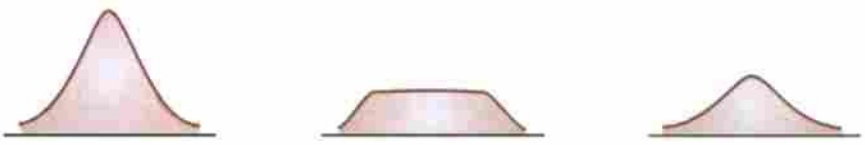
(*iii*) *Coefficient of skewness based on third moment* $\gamma_1 = \sqrt{\beta_1}$.

where $\beta_1 = \mu_3^2/\mu_2^3$

Thus $\gamma_1 = \sqrt{\beta_1}$ gives the simplest measure of *skewness.*

25.11 KURTOSIS

Kurtosis measures the degree of peakedness of a distribution and is given by $\beta_2 = \mu_4/\mu_2^2$.



(*a*) *Leptokurtic* (*b*) *Platykurtic* (*c*) *Mesokurtic*

Fig. 25.5

$\gamma_2 = \beta_2 - 3$ gives the excess of *Kurtosis*. The curves with $\beta_2 > 3$ are called *Leptokurtic* and those with $\beta_2 < 3$ as *Platykurtic*. The normal curve for which $\beta_2 = 3$, is called *Mesokurtic* [Fig. 25.5].

Example 25.11. *The first four moments about the working mean 28.5 of a distribution are 0.294, 7.144, 42.409 and 454.98. Calculate the moments about the mean. Also evaluate* β_1, β_2 *and comment upon the skewness and kurtosis of the distribution.* (V.T.U., 2...S)

Solution. The first four moments about the arbitrary origin 28.5 are $\mu'_1 = 0.294$, $\mu'_2 = 7.144$, $\mu'_3 = 42.409$, $\mu'_4 = 454.98$.

$$\therefore \quad \mu'_1 = \frac{1}{N}\sum f_i(x_i - 28.5) = \frac{1}{N}\sum f_i x_i - 28.5 = \bar{x} - 28.5 = 0.294 \text{ or } \bar{x} = 28.794$$

$$\mu_2 = \mu'_2 - \mu'^2_1 = 7.144 - (0.294)^2 = 7.058$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = 42.409 - 3(7.144)(0.294) + 2(0.294)^3 = 36.151.$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

$$= 454.98 - 4(42.409) \times (0.294) + 6(7.144)(0.294)^2 - 3(0.294)^4 = 408.738$$

Now $\beta_1 = \mu_3^2/\mu_2^3 = (36.151)^2/(7.058)^3 = 3.717$

$\beta_2 = \mu_4/\mu_2^2 = 408.738/(7.058)^2 = 8.205.$

$\therefore$ $\gamma_1 = \sqrt{\beta_1} = 1.928$, which indicates considerable skewness of the distribution.

$\gamma_2 = \beta_2 - 3 = 5.205$ which shows that the distribution is *leptokurtic.*

Example 25.12. *Calculate the median, quartiles and the quartile coefficient of skewness from the following data :*

Weight (lbs) :	*70–80*	*80–90*	*90–100*	*100–110*	*110–120*	*120–130*	*130–140*	*140–150*
No. of persons :	*12*	*18*	*35*	*42*	*50*	*45*	*20*	*8*

Solution. Here total frequency $N = \Sigma f_i = 230$.

The cummulative frequency table is

Weight (lbs) :	*70–80*	*80–90*	*90–100*	*100–110*	*110–120*	*120–130*	*130–140*	*140–150*
f :	12	18	35	42	50	45	20	8
cum. f. :	12	30	65	107	157	202	222	230

Now $N/2 = 230/2 = 115$th item which lies in 110–120 group.

$$\therefore \quad \text{median or } Q_2 = L + \frac{N/2 - C}{f} \times h = 110 + \frac{115 - 107}{50} \times 10 = 111.6$$

Also $N/4 = 230/4 = 57.5$ *i.e.* Q_1 is 57.5th or 58th item which lies in 90–100 group.

$$\therefore \quad Q_1 = L + \frac{N/4 - C}{f} \times h = 90 + \frac{57.5 - 30}{35} \times 10 = 97.85$$

Similarly, $3N/4 = 172.5$ *i.e.* Q_3 is 173rd item which lies in 120–130 group.

$$\therefore \quad Q_3 = L + \frac{3N/4 - C}{f} \times h = 120 + \frac{172.5 - 157}{45} \times 10 = 123.44$$

Hence quartile coefficient of skewness $= \dfrac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$

$$= \frac{97.85 + 123.44 - 2 \times 111.6}{123.44 - 97.85} = -0.07 \text{ (approx.)}.$$

PROBLEMS 25.3

1. Calculate the first four moments of the following distribution about the mean :

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

Also evaluate β_1 and β_2. *(V.T.U., 2004 ; Madras, 2003)*

2. The following table gives the monthly wages of 72 workers in a factory. Compute the standard deviation, quartile deviation, coefficients of variation and skewness. *(V.T.U., 2001)*

Monthly wages (in ₹)	*No. of workers*	*Monthly wages (in ₹)*	*No. of workers*
12.5—17.5	2	37.5—42.5	4
17.5—22.5	22	42.5—47.5	6
22.5—27.5	19	47.5—52.5	1
27.5—32.5	14	52.5—57.5	1
32.5—37.5	3		

3. Find Pearson's coefficient of skewness for the following data :

Class :	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
Frequency :	5	9	14	20	25	15	8	4

(V.T.U., 2000 S)

4. Compute the quartile coefficient of skewness for the following distribution :

x:	3–7	8–12	13–17	18–22	23–27	28–32	33–37	38–42
f:	2	108	580	175	80	32	18	5

(Madras, 2002 ; V.T.U., 2000)

Also compute the measure of skewness based on the third moment.

5. The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40. Show that the mean = 3, the variance = 15 and $\mu_3 = -86$. *(V.T.U., 2003 S)*
6. Compute skewness and kurtosis, if the first four moments of a frequency distribution $f(x)$ about the value $x = 4$ are respectively 1, 4, 10 and 45. *(Coimbatore, 1999)*
7. In a certain distribution, the first four moments about a point are – 1.5, 17, – 30 and 108. Calculate the moments about the mean, β_1 and β_2 ; and state whether the distribution is leptokurtic or platykurtic ?

25.12 CORRELATION

So far we have confined our attention to the analysis of observations on a single variable. There are, however, many phenomenae where the changes in one variable are related to the changes in the other variable. For instance, the yield of a crop varies with the amount of rainfall, the price of a commodity increases with the reduction in its supply and so on. Such a simultaneous variation, *i.e.* when the changes in one variable are associated or followed by changes in the other, is called *correlation.* Such a data connecting two variables is called *bivariate population.*

If an increase (or decrease) in the values of one variable corresponds to an increase (or decrease) in the other, the correlation is said to be *positive.* If the increase (or decrease) in one corresponds to the decrease (or increase) in the other, the correlation is said to be *negative.* If there is no relationship indicated between the variables, they are said to be *independent or uncorrelated.*

To obtain a measure of relationship between the two variables, we plot their corresponding values on the graph, taking one of the variables along the x-axis and the other along the y-axis. (Fig. 25.6).

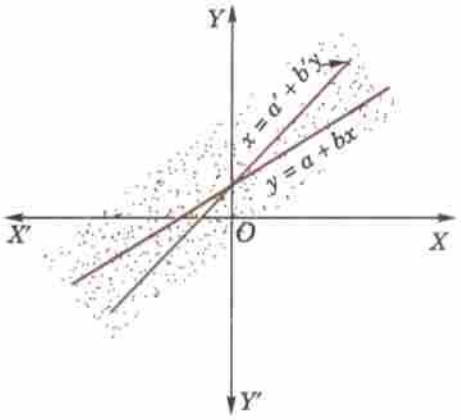


Fig. 25.6

Let the origin be shifted to $(\bar{x}, \bar{y})$, where $\bar{x}, \bar{y}$ are the means of x's and y's that the new co-ordinates are given by

$$X = x - \bar{x}, \qquad Y = y - \bar{y}.$$

Now the points (X, Y) are so distributed over the four quadrants of XY-plane that the product XY is positive in the first and third quadrants but negative in the second and fourth quadrants. The algebraic sum of the products can be taken as describing the trend of the dots in all the quadrants.

$\therefore$ (*i*) If ΣXY is positive, the trend of the dots is through the first and third quadrants,

(*ii*) if ΣXY is negative the trend of the dots is in the second and fourth quadrants, and

(*iii*) if ΣXY is zero, the points indicate no trend *i.e.* the points are evenly distributed over the four quadrants.

The ΣXY or better still $\frac{1}{n}\Sigma XY$, *i.e.*, the average of n products may be taken as a measure of correlation. If we put X and Y in their units, *i.e.*, taking σ_x as the unit for x and σ_y for y, then

$$\frac{1}{n}\Sigma\frac{X}{\sigma_x}.\frac{Y}{\sigma_y}, \text{ i.e., } \frac{\Sigma XY}{n\sigma_x\sigma_y}$$

is the *measure of correlation.*

25.13 COEFFICIENT OF CORRELATION

The numerical measure of correlation is called the *coefficient of correlation* and is defined by the relation

$$r = \frac{\Sigma XY}{n\sigma_x\sigma_y}$$

where X = deviation from the mean $\bar{x} = x - \bar{x}$, Y = deviation from the mean $\bar{y} = y - \bar{y}$,

σ_x = S.D. of x-series, σ_y = S.D. of y-series and n = number of values of the two variables.

Methods of calculation :

(*a*) *Direct method.* Substituting the value of σ_x and σ_y in the above formula, we get

$$r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2 \Sigma Y^2)}} \qquad ...(1)$$

Another form of the formula (1) which is quite handy for calculation is

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{[\{n\Sigma x^2 - (\Sigma x)^2\} \times \{n\Sigma y^2 - (\Sigma y)^2\}]}} \qquad ...(2)$$

(*b*) *Step-deviation method.* The direct method becomes very lengthy and tedious if the means of the two series are not integers. In such cases, use is made of assumed means. If d_x and d_y are step-deviations from the assumed means, then

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{[\{n\Sigma d_x^2 - (\Sigma d_x)^2\} \times \{n\Sigma d_y^2 - (\Sigma d_y)^2\}]}} \qquad ...(3)$$

where $d_x = (x - a)/h$ and $d_y = (y - b)/k$.

Obs. The change of origin and units do not alter the value of the correlation coefficient since r is a pure number.

(*c*) *Co-efficient of correlation for grouped data.* When x and y series are both given as frequency distributions, these can be represented by a two-way table known as the *correlation-table.* It is double-entry table with one series along the horizontal and the other along the vertical as shown on page 848. The co-efficient of correlation for such a *bivariate frequency distribution* is calculated by the formula.

$$r = \frac{n(\Sigma f d_x d_y) - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{[\{n\Sigma f d_x^2 - (\Sigma f d_x)^2\} \times \{n\Sigma f d_y^2 - (\Sigma f d_y)^2\}]}} \quad ...(4)$$

where d_x = deviation of the central values from the assumed mean of x-series,
d_y = deviation of the central values from the assumed mean of y-series,
f is the frequency corresponding to the pair (x, y)
and $n(= \Sigma f)$ is the total number of frequencies.

Example 25.13. *Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R.) and engineering ratio (E.R.). Calculate the co-efficient of correlation.*

Student	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
I.R.	*105*	*104*	*102*	*101*	*100*	*99*	*98*	*96*	*93*	*92*
E.R.	*101*	*103*	*100*	*98*	*95*	*96*	*104*	*92*	*97*	*94*

(Andhra, 2000)

Solution. We construct the following table :

Student	*Intelligence ratio*		*Engineering ratio*				
	x	$x - \bar{x} = X$	y	$y - \bar{y} = Y$	X^2	Y^2	XY
A	105	6	101	3	36	9	18
B	104	5	103	5	25	25	25
C	102	3	100	2	9	4	6
D	101	2	98	0	4	0	0
E	100	1	95	−3	1	9	−3
F	99	0	96	−2	0	4	0
G	98	−1	104	6	1	36	−6
H	96	−3	92	−6	9	36	18
I	93	−6	97	−1	36	1	6
J	92	−7	94	−4	49	16	28
Total	990	0	980	0	170	140	92

From this table, mean of x, *i.e.*, $\bar{x} = 990/10 = 99$ and mean of y, *i.e.* $\bar{y} = 980/10 = 98$.

$\Sigma X^2 = 170$, $\Sigma Y^2 = 140$ and $\Sigma XY = 92$.

Substituting these values in the formula (1) p. 744, we have

$$r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2 \Sigma Y^2)}} = \frac{92}{\sqrt{(170 \times 140)}} = 92/154.3 = 0.59.$$

Example 25.14. *The correlation table given below shows that the ages of husband and wife of 53 married couples living together on the census night of 1991. Calculate the coefficient of correlation between the age of the husband and that of the wife.* *(J.N.T.U., 2003)*

Age of husband	*Age of wife*						*Total*
	15–25	*25–35*	*35–45*	*45–55*	*55–65*	*65–75*	
15–25	*1*	*1*	–	–	–	–	*2*
25–35	*2*	*12*	*1*	–	–	–	*15*
35–45	–	*4*	*10*	*1*	–	–	*15*
45–55	–	–	*3*	*6*	*1*	–	*10*
55–65	–	–	–	*2*	*4*	*2*	*8*
65–75	–	–	–	–	*1*	*2*	*3*
Total	*3*	*17*	*14*	*9*	*6*	*4*	*53*

Solution.

Age of husband				Age of wife x-series							Suppose $d_x = \frac{x-40}{10}$, $d_y = \frac{y-40}{10}$		
				15–25	25–35	35–45	45–55	55–65	65–75				
Years			Mid pt. x	20	30	40	50	60	70	Total f			
Age group	Mid pt. y		d_x / d_y	−20	−10	0	10	20	30		fd_y	fd_y^2	fd_xd_y
				−2	−1	0	1	2	3				
15–25	20	−20	−2	4 1	2 1					2	−4	8	6
25–35	30	−10	−1	4 2	12 12	0 1				15	−15	15	16
35–45	40	0	0		0 4	0 10	0 1			15	0	0	0
45–55	50	10	1			0 3	6 6	2 1		10	10	10	8
55–65	60	20	2				4 2	16 4	12 2	8	16	32	32
65–75	70	30	3					6 1	18 2	3	9	27	24
Total f				3	17	14	9	6	4	$53 = n$	16	92	86
fd_x				−6	−17	0	9	12	12	10	Thick figures in small sqs. stand for fd_xd_y		
fd_x^2				12	17	0	9	24	36	98			
fd_xd_y				8	14	0	10	24	30	86	**Check:** $\Sigma fd_xd_y = 86$ from both sides		

With the help of the above correlation table, we have

$$r = \frac{n(\Sigma fd_xd_y) - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{[n\Sigma fd_x^2 - (\Sigma fd_x)^2] \times [n\Sigma fd_y^2 - (\Sigma fd_y)^2]}}$$

$$= \frac{53 \times 86 - 10 \times 16}{\sqrt{[(53 \times 98 - 100) \times (53 \times 92 - 256)]}} = \frac{4398}{\sqrt{(5094 \times 4620)}} = \frac{4398}{4850} = 0.91 \text{ (approx.)}.$$

25.14 LINES OF REGRESSION

It frequently happens that the dots of the scatter diagram generally, tend to cluster along a well defined direction which suggests a linear relationship between the variables x and y. Such a line of best-fit for the given distribution of dots is called the *line of regression* (Fig. 25.6). In fact there are two such lines, one giving the best possible mean values of y for each specified value of x and the other giving the best possible mean values of x for given values of y. The former is known as the *line of regression of y on x* and the latter as the *line of regression of x on y*.

Consider first the line of regression of y on x. Let the straight line satisfying the general trend of n dots in a scatter diagram be

$$y = a + bx \qquad ...(1)$$

We have to determine the constants a and b so that (1) gives for each value of x, the best estimate for the average value of y in accordance with the *principle of least squares* (page 816), therefore, the normal equations for a and b are

$$\Sigma y = na + b\Sigma x \qquad ...(2)$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad ...(3)$$

(2) gives $$\frac{1}{n}\Sigma y = a + b \cdot \frac{1}{n}\Sigma x \text{ i.e., } \bar{y} = a + b\bar{x}.$$

This shows that $(\bar{x}, \bar{y})$, *i.e.*, the means of x and y, lie on (1).

Shifting the origin to $(\bar{x}, \bar{y})$, (3) takes the form

$$\Sigma(x-\bar{x})(y-\bar{y}) = a\Sigma(x-\bar{x}) + b\Sigma(x-\bar{x})^2, \text{ but } a\Sigma(x-\bar{x}) = 0,$$

$$\therefore \quad b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma XY}{\Sigma X^2} = \frac{\Sigma XY}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \qquad \left[\because r = \frac{\Sigma XY}{n\sigma_x\sigma_y}\right]$$

Thus the line of best fit becomes $$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \qquad ...(4)$$

which is the equation of *the line of regression of y on x*. Its slope is called the *regression coefficient of y on x*.

Interchanging x and y, we find that the line of regression of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \qquad ...(5)$$

Thus the *regression coefficient of y on x* $= r\sigma_y/\sigma_x$ ...(6)

and the *regression coefficient of x on y* $= r\sigma_x/\sigma_y$ ...(7)

Cor. *The correlation coefficient r is the geometric mean between the two regression co-efficients.*

For $$r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = r^2.$$

Example 25.15. *The two regression equations of the variables x and y are x = 19.13 – 0.87y and y = 11.64 – 0.50x. Find (i) mean of x's, (ii) mean of y's and (iii) the correlation coefficient between x and y.*

(V.T.U., 2004 ; Anna, 2003 ; Burdwan, 2003)

Solution. Since the mean of x's and the mean of y's lie on the two regression lines, we have

$$\bar{x} = 19.13 - 0.87\bar{y} \qquad ...(i)$$

$$\bar{y} = 11.64 - 0.50\bar{x} \qquad ...(ii)$$

Multiplying (ii) by 0.87 and subtracting from (i), we have

$$[1 - (0.87)(0.50)]\ \bar{x} = 19.13 - (11.64)(0.87) \text{ or } 0.57\ \bar{x} = 9.00 \text{ or } \bar{x} = 15.79$$

$$\therefore \quad \bar{y} = 11.64 - (0.50)(15.79) = 3.74$$

$\therefore$ regression coefficient of y on x is -0.50 and that of x on y is -0.87.

Now since the coefficient of correlation is the geometric mean between the two regression coefficients.

$$\therefore \quad r = \sqrt{[(-0.50)(-0.87)]} = \sqrt{(0.43)} = -0.66.$$

[– ve sign is taken since both the regression coefficients are – ve]

Example 25.16. *In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales :*

Salesmen	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Test scores	*40*	*70*	*50*	*60*	*80*	*50*	*90*	*40*	*60*	*60*
Sales (000)	*2.5*	*6.0*	*4.5*	*5.0*	*4.5*	*2.0*	*5.5*	*3.0*	*4.5*	*3.0*

Calculate the regression line of sales on test scores and estimate the most probable weekly sales volume if a salesman makes a score of 70.

Solution. With the help of the table below, we have

$\bar{x}$ = mean of x (test scores) = 60 + 0/10 = 60

$\bar{y}$ = mean of y (sales) = 4.5 + (– 4.5)/10 = 4.05.

Regression line of sales (y) on scores (x) is given by

$$y - \bar{y} = r(\sigma_y / \sigma_x)(x - \bar{x})$$

where
$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{\sigma_x \sigma_y} \times \frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{(\sigma_x)^2} = \left[\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{n}\right] \Big/ \left[\Sigma d_x^2 - (\Sigma d_x)^2 / n\right]$$

$$= \frac{140 - \dfrac{0 \times (-4.5)}{10}}{2400 - 0^2/10} = \frac{140}{2400} = 0.06$$

$\therefore$ the required regression line is

$$y - 4.05 = 0.06(x - 60) \quad \text{or} \quad y = 0.06x + 0.45.$$

For $x = 70, y = 0.06 \times 70 + 0.45 = 4.65$.

Thus the most probable weekly sales volume for a score of 70 is 4.65.

Test scores x	*Sales* y	*Deviation of x from assumed mean (= 60)* d_x	*Deviation of y from assumed average (= 4.5)* d_y	$d_x \times d_y$	d_x^2	d_y^2
40	2.5	– 20	– 2	40	400	4
70	6.0	10	1.5	15	100	2.25
50	4.5	– 10	0	0	100	0
60	5.0	0	0.5	0	0	2.25
80	4.5	20	0	0	400	0
50	2.0	– 10	– 2.5	25	100	6.25
90	5.5	30	1	30	900	1.00
40	3.0	– 20	– 1.5	30	400	2.25
60	4.5	0	0	0	0	0
60	3.0	0	– 1.5	0	0	2.25
		$\Sigma d_x = 0$	$\Sigma d_y = -4.5$	$\Sigma d_x d_y$ = 140	Σd_x^2 = 2400	Σd_y^2 =18.25

Example 25.17. *If θ is the angle between the two regression lines, show that*

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

Explain the significance when $r = 0$ and $r = \pm 1$. *(U.P.T.U., 2007 ; V.T.U., 2007)*

Solution. The equations to the line of regression of y on x and x on y are

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

$\therefore$ their slopes are $m_1 = r\sigma_y/\sigma_x$ and $m_2 = \sigma_y/r\sigma_x$

Thus
$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2} = \frac{\sigma_y / r\sigma_x - r\sigma_y / \sigma_x}{1 + \sigma_y^2 / \sigma_x^2} = \frac{1 - r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

When $r = 0$, $\tan\theta \to \infty$ or $\theta = \pi/2$ i.e. when the variables are independent, the two lines of regression are perpendicular to each other.

When $r = \pm 1$, $\tan\theta = 0$ i.e., $\theta = 0$ or π. Thus the lines of regression coincide i.e., there is perfect correlation between the two variables.

Example 25.18. *In a partially destroyed laboratory record, only the lines of regression of y on x and x on y are available as $4x - 5y + 33 = 0$ and $20x - 9y = 107$ respectively. Calculate $\bar{x}, \bar{y}$ and the coefficient of correlation between x and y.* *(S.V.T.U., 2009 ; U.P.T.U., 2009 ; V.T.U., 2005)*

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

$$4\bar{x} - 5\bar{y} + 33 = 0, \quad 20\bar{x} - 9\bar{y} = 107.$$

Solving these equations, we get $\bar{x} = 13$, $\bar{y} = 17$.

Rewriting the line of regression of y on x as $y = \frac{4}{5}x + \frac{33}{5}$, we get

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{4}{5} \qquad ...(i)$$

Rewriting the line of regression of x on y as $x = \frac{9}{20}y + \frac{107}{9}$, we get

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{9}{20} \qquad ...(ii)$$

Multiplying (i) and (ii), we get

$$r^2 = \frac{4}{5} \times \frac{9}{20} = 0.36 \qquad \therefore \qquad r = 0.6$$

Hence $r = 0.6$, the positive sign being taken as b_{yx} and b_{xy} both are positive.

Example 25.19. *Establish the formula* $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

Hence calculate r from the following data :

x :	*21*	*23*	*30*	*54*	*57*	*58*	*72*	*78*	*87*	*90*
y :	*60*	*71*	*72*	*83*	*110*	*84*	*100*	*92*	*113*	*135*

(U.P.T.U., 2002)

Solution. (*a*) Let $z = x - y$ so that $\bar{z} = \bar{x} - \bar{y}$.

$\therefore$ $$z - \bar{z} = (x - \bar{x}) - (y - \bar{y})$$

or $$(z - \bar{z})^2 = (x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})$$

Summing up for n terms, we have

$$\Sigma(z - \bar{z})^2 = \Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2 - 2\Sigma(x - \bar{x})(y - \bar{y})$$

or $$\frac{\Sigma(z - \bar{z})^2}{n} = \frac{\Sigma(x - \bar{x})^2}{n} + \frac{\Sigma(y - \bar{y})^2}{n} - 2\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n}$$

i.e., $$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad \left[\because\ r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}\right.$$

which is the required result.

(*b*) To fine r, we have to calculate σ_x, σ_y and σ_{x-y}. We make the following table :

x	$X = x - 54$	X^2	y	$Y = y - 100$	Y^2	$y - x$	$(x - y)^2$
21	− 33	1089	60	− 40	1600	39	1521
23	− 31	961	71	− 29	841	48	2304
30	− 24	576	72	− 28	784	42	1764
54	0	0	83	− 17	289	29	841
57	3	9	110	10	100	53	2809
58	4	16	84	− 16	256	26	676
72	18	324	100	0	0	28	784
78	24	576	92	− 8	64	14	196
87	33	1089	113	13	169	26	676
90	36	1296	135	35	1225	45	2025
Total	30	5936		− 80	5328	350	13596

$$\therefore \quad \sigma_x^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{5636}{10} - \left(\frac{30}{10}\right)^2 = 593.6 - 9 = 584.6$$

$$\sigma_y^2 = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2 = \frac{5328}{10} - \left(\frac{-80}{10}\right)^2 = 532.8 - 64 = 468.8$$

$$\sigma^2_{x-y} = \frac{\Sigma(x-y)^2}{N} - \left\{\frac{\Sigma(x-y)}{N}\right\}^2 = 1359.6 - 1225 = 134.6$$

From the above formula,

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma^2_{x-y}}{2\sigma_x\sigma_y} = \frac{584.6 + 468.8 - 134.6}{2 \times 24.18 \times 23.85} = 0.876.$$

Example 25.20. *While calculating correlation coefficient between two variables x and y from 25 pairs of observations, the following results were obtained : $n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$. Later it was discovered at the time of checking that the pairs of values*

x	y
8	12
6	8

were copied down as

x	y
6	14
8	6

Obtain the correct value of correlation coefficient. *(V.T.U., 2011 S ; S.V.T.U., 2009)*

Solution. To get the correct results, we subtract the incorrect values and add the corresponding correct values.

$\therefore$ The correct results would be

$$\Sigma n = 25, \Sigma x = 125 - 6 - 8 + 8 + 6 = 125, \Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$
$$\Sigma y = 100 - 14 - 6 + 12 + 8 = 100, \Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$
$$\Sigma xy = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y^2)\}]}} = \frac{25 \times 520 - 125 \times 100}{\sqrt{[\{25 \times 650 - (125)^2\}\{25 \times 436 - (100)^2\}]}}$$

$$= \frac{20}{\sqrt{(25 \times 36)}} = \frac{2}{3}.$$

25.15 STANDARD ERROR OF ESTIMATE

The sum of the squares of the deviations of the points from the line of regression of y on x is

$$\Sigma(y - a - bx)^2 = \Sigma(Y - bX)^2, \text{ where } X = x - \bar{x}, Y = y - \bar{y}$$

$$= \sum \left(Y - r\frac{\sigma_y}{\sigma_x}X\right)^2 = \Sigma Y^2 - 2r(\sigma_y/\sigma_x)\,\Sigma XY + r^2(\sigma_y^2/\sigma_x^2)\,\Sigma X^2$$

$$= n\sigma_y^2 - 2r(\sigma_y/\sigma_x)\,r \,.\, n\sigma_x\sigma_y + r^2(\sigma_y^2/\sigma_x^2) \,.\, n\sigma_x^2 = n\sigma_y^2(1 - r^2).$$

Denoting this sum of squares by nS_y^2, we have $S_y = \sigma_y\sqrt{(1 - r^2)}$ ...(1)

Since S_y is the root mean square deviation of the points from the regression line of y on x, it is called the *standard error of estimate* of y. Similarly the standard error of estimate of x is given by

$$S_x = \sigma_x\sqrt{(1 - r^2)} \qquad ...(2)$$

Since the sum of the squares of deviations cannot be negative, it follows that

$$r^2 \le 1 \quad \text{or} \quad -1 \le r \le 1.$$

i.e., correlation coefficient lies between – 1 and 1. *(J.N.T.U., 2006)*

If $r = 1$ or -1, the sum of the squares of deviations from either line of regression is zero. Consequently each deviation is zero and all the points lie on both the lines of regression. These two lines coincide and we say that the correlation between the variables is *perfect*. The nearer r^2 is to unity the closer are the points to the lines of

regression. Thus *the departure of r^2 from unity is a measure of departure from linearity of the relationship between the variables.*

25.16 RANK CORRELATION

A group of n individuals may be arranged in order to merit with respect to some characteristic. The same group would give different orders for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the *rank correlation* in the characteristics A and B for that group of individuals.

Let x_i, y_i be the ranks of the ith individuals in A and B respectively. Assuming that no two individuals are bracketed equal in either case, each of the variables taking the values 1, 2, 3, ..., n, we have

$$\bar{x} = \bar{y} = \frac{1+2+3+\ldots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

If X, Y be the deviations of x, y from their means, then

$$\Sigma X_i^2 = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 + n(\bar{x})^2 - 2\bar{x}\,\Sigma x_i = \Sigma n^2 + \frac{n(n+1)^2}{4} - 2\frac{n+1}{2} \cdot \Sigma n$$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)^2}{4} - \frac{n(n+1)^2}{2} = \frac{1}{12}(n^3 - n)$$

Similarly $\Sigma Y_i^2 = \frac{1}{12}(n^3 - n)$

Now let $d_i = x_i - y_i$ so that $d_i = (x_i - \bar{x}) - (y_i - \bar{y}) = X_i - Y_i$

$\therefore$ $\Sigma d_i^2 = \Sigma X_i^2 + \Sigma Y_i^2 - 2\Sigma X_i Y_i$

or $$\Sigma X_i Y_i = \tfrac{1}{2}(\Sigma X_i^2 + \Sigma Y_i^2 - \Sigma d_i^2) = \frac{1}{12}(n^3 - n) - \frac{1}{2}\Sigma d_i^2.$$

Hence the correlation coefficient between these variables is

$$r = \frac{\Sigma X_i Y_i}{\sqrt{(\Sigma X_i^2 \Sigma Y_i^2)}} = \frac{\frac{1}{12}(n^3 - n) - \frac{1}{2}\Sigma d_i^2}{\frac{1}{12}(n^3 - n)} = 1 - \frac{6\,\Sigma d_i^2}{n^3 - n}$$

This is called the *rank correlation coefficient* and is denoted by ρ.

Example 25.21. *Ten participants in a contest are ranked by two judges as follows :*

x :	1	6	5	10	3	2	4	9	7	8
y :	6	4	9	8	1	2	3	10	5	7

Calculate the rank correlation coefficient ρ. *(V.T.U., 2002)*

Solution. If $d_i = x_i - y_i$, then $d_i = -5, 2, -4, 2, 2, 0, 1, -1, 2, 1$

$\therefore$ $\Sigma d_i^2 = 25 + 4 + 16 + 4 + 4 + 0 + 1 + 1 + 4 + 1 = 60$

Hence $$\rho = 1 - \frac{6\Sigma d_i^2}{n^3 - n} = 1 - \frac{6 \times 60}{990} = 0.6 \text{ nearly.}$$

Example 25.22. *Three judges, A, B, C, give the following ranks. Find which pair of judges has common approach*

A :	1	6	5	10	3	2	4	9	7	8
B :	3	5	8	4	7	10	2	1	6	9
C :	6	4	9	8	1	2	3	10	5	7

(J.N.T.U., 2003)

Solution. Here $n = 10$.

$A\ (=x)$	*Ranks by* $B\ (=y)$	$C\ (=z)$	d_1 $x-y$	d_2 $y-z$	d_3 $z-x$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−3	5	4	9	25
6	5	4	1	1	−2	1	1	4
5	8	9	−3	−1	4	9	1	16
10	4	8	6	−4	−2	36	16	4
3	7	1	−4	6	−2	16	36	4
2	10	2	−8	8	0	64	64	0
4	2	3	2	−1	−1	4	1	1
9	1	10	8	−9	1	64	81	1
7	6	5	1	1	−2	1	1	4
8	9	7	−1	2	−1	1	4	1
Total			0	0	0	200	214	60

$$\therefore \quad \rho(x, y) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = -0.2$$

$$\rho(y, z) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -0.3$$

$$\rho(z, x) = 1 - \frac{6\Sigma d_3^2}{n(n-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 0.6$$

Since $\rho(z, x)$ is maximum, the pair of judges A and C have the nearest common approach.

PROBLEMS 25.4

1. Find the correlation co-efficient and the regression lines of y and x and x on y for the following data :

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

(V.T.U., 2010)

2. Find the correlation coefficient between x and y from the given data :

x :	78	89	97	69	59	79	68	57
y :	125	137	156	112	107	138	123	108

(J.N.T.U., 2005)

3. Find the co-efficient of correlation between industrial production and export using the following data and comment on the result.

Production (in crore tons) :	55	56	58	59	60	60	62
Exports (in crore tons) :	35	38	38	39	44	43	45

(Madras, 2000)

4. Ten people of various heights as under, were requested to read the letters on a car at 25 yards distance. The number of letters correctly read is given below :

Height (in feet) :	5.1	5.3	5.6	5.7	5.8	5.9	5.10	5.11	6.0	6.1
No. of letters :	11	17	19	14	8	15	20	6	8	12

Is there any correlation between heights and visual power ?

5. Using the formula $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$, find r from the following data :

x :	92	89	87	86	83	77	71	63	53	50
y :	86	88	91	77	68	85	52	82	37	57

6. Find the correlation between x (marks in Mathematics) and y (marks in Engineering Drawing) given in the following data :

y \ x	10—40	40—70	70—100	Total
0—30	5	20	—	25
30—60	—	28	2	30
60—90	—	32	13	45
Total	5	80	15	100

7. Find two lines of regression and coefficient of correlation for the data given below :
$n = 18, \Sigma x = 12, \Sigma y = 18, \Sigma x^2 = 60, \Sigma y^2 = 96, \Sigma xy = 48.$ *(U.P.T.U., MCA, 2009)*

8. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}(3/8)$, show that $\sigma_x = \frac{1}{2}\sigma_y$. *(V.T.U., 2004)*

9. For two random variables x and y with the same mean, the two regression lines are $y = ax + b$ and $x = \alpha y + \beta$. Show that $\frac{b}{\beta} = \frac{1-a}{1-\alpha}$. Find also the common mean. *(U.P.T.U., 2010)*

10. Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation coefficient between x and y. *(Madras, 2002)*

11. The regression equations of two variables x and y are $x = 0.7y + 5.2$, $y = 0.3x + 2.8$. Find the means of the variables and the coefficient of correlation between them. *(Osmania, 2002)*

12. In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively, $7x - 16y + 9 = 0$, $5y - 4x - 3 = 0$. Calculate the co-efficient of correlation, $\bar{x}$ and $\bar{y}$.

13. The following results were obtained from records of age (x) and 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 appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45.

14. Compute the standard error of estimate S_x for the respective heights of the following 12 couples :

Height x of husband (inches) :	68	66	68	65	69	66	68	65	71	67	68	70
Height y of wife (inches) :	65	63	67	64	68	62	70	66	68	67	69	71

15. Calculate the rank correlation coefficient from the following data showing ranks of 10 students in two subjects :

Maths :	3	8	9	2	7	10	4	6	1	5
Physics :	5	9	10	1	8	7	3	4	2	6

16. Find the rank correlation for the following data :

x :	56	42	72	36	63	47	55	49	38	42	68	60
y :	147	125	160	118	149	128	150	145	115	140	152	155

(S.V.T.U., 2009 ; J.N.T.U., 2003)

25.17 OBJECTIVE TYPE OF QUESTIONS

PROBLEMS 25.5

Select the correct answer or fill up the blanks in each of the following questions :

1. The median of the numbers 11, 10, 12, 13, 9 is
(a) 12.5 (b) 12 (c) 10.5 (d) 11.

2. The mode of the numbers 7, 7, 7, 9, 10, 11, 11, 11, 12 is
(a) 11 (b) 12 (c) 7 (d) 7 and 11.

3. S.D. is defined as

(a) $\sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}$ (b) $\frac{\Sigma f(x-\bar{x})}{\Sigma f}$ (c) $\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}$.

4. Coefficient of variation is

(a) $\frac{\sigma}{\bar{x}} \times 100$ (b) $\frac{\sigma}{x}$ (c) $\sqrt{\frac{\sigma^2}{x}} \times 100$.

5. Average scores of three batsman A, B, C are respectively 40, 45 and 55 and their S.D.s are respectively 9, 11, 16. Which batsman is more consistent ?

(a) A (b) B (c) C.

6. The equations of regression lines are $y = 0.5x + a$ and $x = 0.4y + b$. The correlation coefficient is

(a) $\sqrt{0.2}$ (b) 0.45 (c) $-\sqrt{0.2}$.

7. If the correlation coefficient is 0, the two regression lines are

(a) parallel (b) perpendicular (c) coincident (d) inclined at 45° to each other.

8. If r_1 and r_2 are two regression coefficients, then signs of r_1 and r_2 depend on .....

9. Regression coefficient of y on x is 0.7 and that of x on y is 3.2. Is the correlation coefficient r consistent ?

10. The standard deviation of the numbers 24, 48, 64, 36, 53 is .....

11. If $y = x + 1$ and $x = 3y - 7$ are the two lines of regression then $\bar{x}$ = ..., $\bar{y}$ = ... and r = ...

12. If the two regression lines are perpendicular to each other, then their coefficient of correlation is .....

13. Quartile deviation is defined as .....

14. The minimum value of correlation coefficient is .....

15. Prediction error of Y is defined as .....

16. If X and Y are independent, then the correlation coefficient between X and Y is .....

17. The point of intersection of the two regression lines is .....

18. The smaller the coefficient of variation, the greater is the ..... in the data.

19. The moment coefficient of skewness is given by .....

20. Kurtosis measures the ..... of a distribution.

21. The equation of the line of regression of y on x is .....

22. Coefficient of variation = ...

23. The angle between two regression lines is given by .....

24. A frequency curve is said to be Mesokurtic when β_2 is .....

25. Correlation coefficient is the geometrical mean between .....

26. When the variables are independent, the two lines of regression are .....

27. Arithmetic mean of the coefficients of regression is ..... than the coefficient of correlation.

28. If two regression lines coincide then the coefficient of correlation is .....

29. The rank coefficient is given by .....

30. The ratio of the standard deviation to the mean is known as .....

31. The value of $\Sigma f(x - \bar{x})$ = ...

32. The value of coefficient of correlation lies between ..... and .....

33. If the two regression coefficients are – 0.4 and – 0.9, then the correlation coefficient is .....

34. A distribution with the following constants is positively skew : $Q_1 = 25.8$, median = 49.0, $Q_3 = 64.2$.

(True or False)

35. Quartile coefficient of skewness is $\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$. (True or False)

36. Skewness indicates peakedness of the frequency distribution. (True or False)