1. INTRODUCTION

Quantitative data in a mass exhibit certain general characteristics or they differ om each other in the following ways :

1. They show a tendency to concentrate at certain values, usually somewhere in the centre of the distribution. Measures of this tendency are called *measures of central tendency* or *averages*.
2. The data vary about a measure of central tendency and these measures of deviation are called *measures of variation or dispersion*.
3. The data in a frequency distribution may fall into symmetrical or asymmetrical patterns. The measures of the direction and degree of asymmetry are called *measures of skewness*.
4. Polygons of frequency distributions exhibit flatness or peakedness of the frequency curves. The measures of peakedness or flatness of the frequency curves are called *measures of kurtosis*.

2. FREQUENCY DISTRIBUTION

When observations, discrete or continuous, are available on a single characteristic f a large number of individuals, often it becomes necessary to condense the data as ar as possible without losing any information of interest. Let us consider the marks in tatistics obtained by 250 candidates selected at random from among those appearing n a certain examination.

TABLE 2·1 : MARKS IN STATISTICS OF 250 CANDIDATES

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46
41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38
31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	34	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as *array*. But this does not reduce the bulk of the data. A much better representation is given in *Table* 2·2 :

TABLE 2·2

Marks	No. of Students —Tally Marks	Total Frequency	Marks	No. of Students —Tally Marks	Total Frequency
15	II	= 2	40	𝍸 𝍸 I	= 11
17	III	= 3	41	𝍸 𝍸	= 10
18	II	= 2	42	𝍸 𝍸 III	= 13
19	II	= 2	43	𝍸 III	= 8
21	II	= 2	44	𝍸 𝍸 II	= 12
22	II	= 2	45	𝍸 II	= 7
23	III	= 3	46	𝍸 II	= 7
24	IIII	= 4	47	𝍸 III	= 8
25	I	= 1	48	𝍸 𝍸 II	= 12
26	III	= 3	49	III	= 3
27	I	= 1	50	𝍸 𝍸	= 10
28	III	= 3	51	IIII	= 4
29	II	= 2	52	𝍸	= 5
30	𝍸	= 5	53	IIII	= 4
31	𝍸 𝍸	= 10	54	III	= 3
32	𝍸 𝍸	= 10	55	II	= 2
33	𝍸 III	= 8	56	II	= 2
34	𝍸 𝍸 I	= 11	57	II	= 2
35	𝍸	= 5	58	II	= 2
36	𝍸	= 5	60	III	= 3
37	𝍸 𝍸 II	= 12	61	I	= 1
38	𝍸 𝍸 𝍸 II	= 17	62	I	= 1
39	𝍸 I	= 6	68	I	= 1

A bar (!) called *tally mark* is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (!) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

The representation of the data as above is known as *frequency distribution*. Marks are called the *variable* (x) and the '*number of students*' against the marks is known as the '*frequency*' (f) of the variable. The word '*frequency*' is derived from 'how frequently' variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks. This representation, though better than an 'array', does not condense the data much and it is quite cumbersome to go thorough this huge mass of data.

If the identity of the individuals about whom a particular information is taken is not relevant, nor the order in which the observations arise, then the first real step

condensation is to divide the observed range of variable into a suitable number of *class-intervals* and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in *Table* 2·3.

TABLE 2·3 : FREQUENCY TABLE

Marks (x)	*No. of students* (f)
15—19	9
20—24	11
25—29	10
30—34	44
35—39	45
40—44	54
45—49	37
50—54	26
55—59	6
60—64	5
65—69	1
Total	250

Such a table showing the distribution of the frequencies in the different classes is called a *frequency table* and the manner in which the class frequencies are distributed over the class intervals is called the *grouped frequency distribution* of the variable.

Remark. The classes of the type 15—19, 20—24, 25—29 etc., in which both the upper and lower limits are included are called *'inclusive classes'*. For example, the class 20— 24, includes all the values from 20 to 24, both inclusive and the classification is termed as *inclusive type classification.*

In spite of great importance of classification in statistical analysis, no hard and fast rules can be laid down for it. The following points may be kept in mind for classification :

1. The classes should be clearly defined and should not lead to any ambiguity.

2. The classes should be exhaustive, *i.e.,* each of the given values should be included in one of the classes.

3. The classes should be mutually exclusive and non-overlapping.

4. The classes should be of equal width. The principle, however, cannot be rigidly followed. If the classes are of varying width, the different class frequencies will not be comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The ratios thus obtained are called *'frequency densities'*.

5. Indeterminate classes, *e.g.,* the open-end classes like less than '*a* ' or greater than '*b*' should be avoided as far as possible since they create difficulty in analysis and interpretation.

6. The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However, the number of classes may be more than 15 depending upon the total frequency and the details required, but it is desirable that it is not less than 5 since in that case the classification may not reveal the essential characteristics of the population. The following formula due to Struges may be used to determine an approximate number k of classes :

$$k = 1 + 3{\cdot}322 \log_{10} N,$$ where N is the total frequency.

The Magnitude of the Class Interval. Having fixed the number of classes, divide the range (the difference between the greatest and the smallest observation) by it and the nearest integer to this value gives the magnitude of the class interval. Broad class intervals (*i.e.,* less number of classes) will yield only rough estimates while for high degree of accuracy small class intervals (*i.e.,* large number of classes) are desirable.

Class Limits. The class limits should be chosen in such a way that the mid-value of the class interval and actual average of the observations in that class interval are as near to each other as possible. If this is not the case then the classification gives a distorted picture of the characteristics of the data. If possible, class limits should be located at the points which are multiple of 0, 2, 5, 10, ... etc., so that the midpoints of the classes are the common figures, *viz.*, 0, 2, 5, 10, ... etc., the figures capable of easy and simple analysis.

2·2·1. Continuous Frequency Distribution. If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of above type.

Let us consider the distribution of age in years. If class intervals are 15—19, 20—24, etc. then the persons with ages between 19 and 20 years are not taken into consideration. In such a case, we form the class intervals as shown in the adjoining table.

Age (in years)
Below 5
5 or more but less than 10
10 or more but less than 15
15 or more but less than 20
20 or more but less than 25
and so on

Here all the persons with any fraction of age are included in one group or the other. For practical purposes we re-write the above classes as shown in the adjoining table.

0 — 5
5 — 10
10 — 15
15 — 20
20 — 25

The form of the frequency distribution with such classes is known as *continuous frequency distribution*. It should be clearly understood that in the above classes, the upper limits of each class are excluded from the respective classes. Such classes in which the upper limits are excluded from the respective classes and are included in the immediate next class are known as *'exclusive classes'* and the classification is termed as *'exclusive type classification'*.

2·3. GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION

It is often useful to represent a frequency distribution by means of a diagram which makes the unwieldy data intelligible and conveys to the eye the general run of the observations. Diagrammatic representation also facilitates the comparison of two or more frequency distributions. We consider below some important types of graphic representation.

2·3·1. Histogram. In drawing the histogram of a given continuous frequency distribution we first mark off along the x - axis all the class intervals on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If, however, the classes are of unequal width then the height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called *histogram*.

Remarks **1.** To draw the histogram for an ungrouped frequency distribution of a variable we shall have to assume that the frequency corresponding to the variate value x is spread over the interval $x - \frac{h}{2}$ to $x + \frac{h}{2}$, where h is the jump from one value to the next.

2. If the grouped frequency distribution is not continuous, first it is to be converted into continuous distribution and then the histogram is drawn.

3. Although the height of each rectangle is proportional to the frequency of the corresponding class, the height of a fraction of the rectangle is not proportional to the frequency of the corresponding fraction of the class, so that histogram cannot be directly used to read frequency over a fraction of a class interval.

4. The histogram of the distribution of marks of 250 students in *Table 2·3* (page 2·5) is obtained as follows :

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution with exclusive type classes as given in the following *Table 2·4 :*

TABLE 2·4

Marks	*No. of Students*
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1

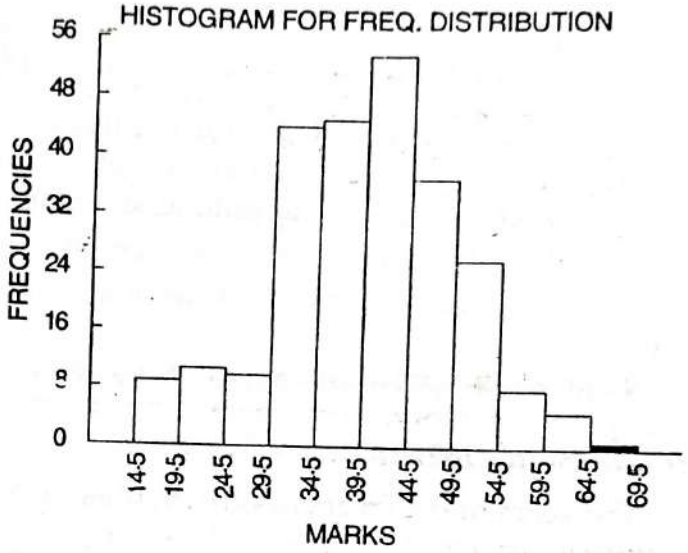


Fig. 2·1.

Note. The upper and lower class limits of the *new exclusive type classes* are known as *class boundaries.*

If d is the gap between the upper limit of any class and the lower limit of the succeeding class, the class boundaries for any class are then given by :

Upper class boundary = Upper class limit $+ \frac{d}{2}$; Lower class boundary = Lower class limit $- \frac{d}{2}$.

2·3·2. Frequency Polygon. For an ungrouped distribution, the frequency polygon is obtained by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points are mid-values of the classintervals. For equal classintervals the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width, the polygon can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

The frequency polygon so obtained should be extended to the base line (x—axis) at both the ends so that it meets the x-axis at the mid-points of two hypothetical classes, *viz.*, the class before the first class and the class after the last class, each assumed to have zero frequency.

2·4. AVERAGES (OR MEASURES OF CENTRAL TENDENCY)

According to Professor Bowley, averages are "*statistical constants which enable us to comprehend in a single effort the significance of the whole.*" They give us an idea about the

concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use :

(*i*) *Arithmetic Mean or Simple Mean*, (*ii*) *Median*, (*iii*) *Mode*,
(*iv*) *Geometric Mean, and* (*v*) *Harmonic Mean.*

2-4-1. Requisites for an Ideal Measure of Central Tendency. According to Professor Yule, the following are the characteristics to be satisfied by an ideal measure of central tendency :

(*i*) It should be rigidly defined.

(*ii*) It should be readily comprehensible and easy to calculate.

(*iii*) It should be based on all the observations.

(*iv*) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series, we should be able to calculate the average of the composite series obtained on combining the given series.

(*v*) It should be affected as little as possible by fluctuations of sampling.

In addition to the above criteria, we may add the following (which is not due to Prof. Yule) :

(*vi*) It should not be affected much by extreme values.

2-5. ARITHMETIC MEAN

Arithmetic mean of a set of observations is their sum divided by the number of observations, *e.g.*, the arithmetic mean $\bar{x}$ of n observations $x_1, x_2, \ldots, x_n$ is given by :

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots (2\cdot1)$$

In case of the frequency distribution $x_i | f_i$, $i = 1, 2, \ldots n$, where f_i is the frequency of the variable x_i,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N\right] \qquad \ldots (2\cdot1a)$$

In case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Remark. The symbol Σ is the letter capital sigma of the Greek alphabet and is used in mathematics to denote the sum of values.

Example 2·1. (*a*) *Find the arithmetic mean of the following frequency distribution :*

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6

(*b*) *Calculate the arithmetic mean of the marks from the following table :*

Marks :	0—10	10—20	20—30	30—40	40—50	50—60
No. of Students :	12	18	27	20	17	6

Solution.

(*a*) COMPUTATION OF MEAN

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	73	299

$$\therefore \quad \bar{x} = \frac{1}{N}\sum fx$$
$$= \frac{299}{73}$$
$$= 4{\cdot}09$$

(*b*) COMPUTATION OF MEAN

Marks	No. of Students (f)	Mid-point (x)	fx
0—10	12	5	60
10—20	18	15	270
20—30	27	25	675
30—40	20	35	700
40—50	17	45	765
50—60	6	55	330
Total	100		2,800

Arithmetic mean ($\bar{x}$)
$$= \frac{1}{N}\sum fx$$
$$= \frac{1}{100} \times 2{,}800$$
$$= 28$$

It may be noted that if the values of x or (and) f are large, the calculation of mean by formula (2·1*a*) is quite time-consuming and tedious. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A' as explained below :

Let $d_i = x_i - A$. Then $f_i d_i = f_i (x_i - A) = f_i x_i - A f_i$

Summing both sides over i from 1 to n, we get

$$\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} f_i x_i - A \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - A \,.\, N$$

$$\Rightarrow \quad \frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - A = \bar{x} - A,$$

where $\bar{x}$ is the arithmetic mean of the distribution.

$$\therefore \quad \bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i \qquad \ldots(2{\cdot}2)$$

This formula is much more convenient to apply than formula (2·1*a*).

Any number can serve the purpose of arbitrary point 'A' but, usually, the value of x corresponding to the middle part of the distribution will be much more convenient.

In case of grouped or continuous frequency distribution, the arithmetic is reduced to still greater extent by taking $d_i = \frac{x_i - A}{h}$, where A is an arbitrary point and h is the common magnitude of class interval. In this case, we have $hd_i = x_i - A$ and proceeding exactly similarly as above , we get

$$\bar{x} = A + \frac{h}{N} \sum_{i=1}^{n} f_i d_i \qquad \ldots (2\cdot3)$$

Example 2·2. *Calculate the mean for the following frequency distribution :*

Class interval :	0—8	8—16	16—24	24—32	32—40	40—48
Frequency :	8	7	16	24	15	7

Solution. Here we take $A = 28$ and $h = 8$.

COMPUTATION OF MEAN

Class interval	Mid-value (x)	Frequency (f)	$d = \frac{x-A}{h}$	fd
0—8	4	8	−3	−24
8—16	12	7	−2	−14
16—24	20	16	−1	−16
24—32	28	24	0	0
32—40	36	15	1	15
40—48	44	7	2	14
Total		77		−25

$$\bar{x} = A + \frac{h \sum fd}{N}$$

$$= 28 + \frac{8 \times (-25)}{77}$$

$$= 28 - \frac{200}{77}$$

$$= 25{\cdot}404$$

2·5·1. Properties of Arithmetic Mean

Property 1. *Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If $x_i \mid f_i$, $i = 1, 2, \ldots, n$ is the frequency distribution, then*

$$\sum_{i=1}^{n} f_i (x_i - \bar{x}) = 0, \ \bar{x} \text{ being the mean of distribution.}$$

Proof. $$\sum_i f_i (x_i - \bar{x}) = \sum_i f_i x_i - \bar{x} \sum_i f_i = \sum_i f_i x_i - \bar{x} \, . \, N, \qquad \ldots (*)$$

Also $$\bar{x} = \frac{\sum_i f_i x_i}{N} \quad \Rightarrow \quad \sum_i f_i x_i = N \bar{x}$$

Substituting in (*), we get : $$\sum_{i=1}^{n} f_i (x_i - \bar{x}) = N \, . \, \bar{x} - \bar{x} \, . \, N = 0$$

Property 2. *The sum of the squares of the deviations of a set of values is minimum when taken about mean.*

Proof. For frequency distribution $x_i \mid f_i$, $i = 1, 2, \ldots, n$, let $Z = \sum_{i=1}^{n} f_i (x_i - A)^2$, be the sum of the squares of the deviations of given values from any arbitrary point 'A'. We have to prove that Z is minimum when $A = \bar{x}$.

Applying the principle of maxima and minima from differential calculus, Z will be minimum for variations in A if $\frac{\partial Z}{\partial A} = 0$ and $\frac{\partial^2 Z}{\partial A^2} > 0$.

Now $$\frac{\partial Z}{\partial A} = -2 \sum_i f_i (x_i - A) = 0 \quad \Rightarrow \quad \sum_i f_i (x_i - A) = 0$$

$$\Rightarrow \qquad \sum f_i x_i - A \sum f = 0 \quad \text{or} \quad A = \frac{\sum f_i x_i}{N} = \bar{x}$$

$$\text{Again} \qquad \frac{\partial^2 Z}{\partial A^2} = -2 \sum_i f_i (-1) = 2 \sum_i f_i = 2N > 0$$

Hence Z is minimum at the point $A = \bar{x}$. This establishes the result.

Property 3. *(Mean of the composite series). If $\bar{x}_i$ $(i = 1, 2, \ldots, k)$ are the means of k component series of sizes n_i, $(i = 1, 2, \ldots, k)$ respectively, then the mean $\bar{x}$ of the composite series obtained on combining the component series is given by the formula :*

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ldots + n_k \bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \sum_{i=1}^{k} n_i x_i \Big/ \sum_{i=1}^{k} n_i \qquad \ldots (2{\cdot}4)$$

Proof. Let $x_{11}, x_{12}, \ldots, x_{1n1}$ be n_1 members of the first series ; $x_{21}, x_{22}, \ldots, x_{2n2}$ be n_2 members of the second series, $\ldots, x_{k1}, x_{k2}, \ldots, x_{knk}$ be n_k members of the kth series.

Then, by *def.*,

$$\left.\begin{aligned} \bar{x}_1 &= \frac{1}{n_1}(x_{11} + x_{12} + \ldots + x_{1n_1}) \\ \bar{x}_2 &= \frac{1}{n_2}(x_{21} + x_{22} + \ldots + x_{2n_2}) \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ \bar{x}_k &= \frac{1}{n_k}(x_{k1} + x_{k2} + \ldots + x_{kn_k}) \end{aligned}\right\} \qquad \ldots (*)$$

The mean $\bar{x}$ of composite series of size $n_1 + n_2 + \ldots + n_k$ is given by

$$\bar{x} = \frac{(x_{11} + x_{12} + \ldots + x_{1n_1}) + (x_{21} + x_{22} + \ldots + x_{2n_2}) + \ldots + (x_{k1} + x_{k2} + \ldots + X_{kn_k})}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ldots + n_k \bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \sum_i n_i \bar{x}_i / \left(\sum_i n_i\right) \qquad \text{[From (*)]}$$

Example 2.3. *The average salary of male employees in a firm was Rs. 5200 and that of females was Rs. 4,200. The mean salary of all the employees was Rs. 5,000. Find the percentage of male and female employees.*

Solution. Let n_1 and n_2 denote respectively the number of male and female employees in the firm, and $\bar{x}_1$ and $\bar{x}_2$ denote respectively their average salary (in rupees). Let $\bar{x}$ denote the average salary of all the workers in the firm.

We are given that : $\bar{x}_1 = 5{,}200, \quad \bar{x}_2 = 4{,}200 \quad$ and $\quad \bar{x} = 5{,}000.$

Also we know $\bar{x} = \dfrac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \quad \Rightarrow \quad 5{,}000\,(n_1 + n_2) = 5{,}200\, n_1 + 4{,}200\, n_2$

$$\Rightarrow \quad (5{,}200 - 5{,}000)\, n_1 = (5{,}000 - 4{,}200)\, n_2 \quad \Rightarrow \quad 20 n_1 = 80 n_2 \quad \Rightarrow \quad \frac{n_1}{n_2} = \frac{4}{1}.$$

$\therefore$ The percentage of male employees in the firm $= \dfrac{4}{4+1} \times 100 = 80$

and the percentage of female employees in the firm $= \dfrac{1}{4+1} \times 100 = 20.$

Example 2·4. *The following is the age distribution of 1,000 persons working in a large industrial house :*

Age - group	*No. of persons*	*Age-group*	*No. of persons*
20—25	30	45—50	105
25—30	160	50—55	70
30—35	210	55—60	60
35—40	180	60—65	40
40—45	145		

Due to continuous heavy losses the management decides to bring down the strength to 30% of the present number according to the following scheme :

(i) To retrench the first 15% from the lower group.

(ii) To absorb the next 45% in other branches.

(iii) To make 10% from the highest age group retire permanently, if necessary.

Calculate the age limits of the persons retained and those to be transferred to other departments. Also find the average age of those retained.

Solution. Total number of persons in the industrial house is $N = 1{,}000$.

According to the conditions of the problem :

(*i*) The number of persons to be retrenched from the lower group

$$= 15\% \text{ of } N = \frac{15}{100} \times 1{,}000 = 150.$$

30 of these will be from the first group 20—25 and the remaining 150 – 30 = 120, from the next age group 25—30.

(*ii*) The number of persons (from the next groups) to be absorbed in other branches :

$$= 45\ \% \text{ of } N = \frac{45}{100} \times\ 1{,}000 = 450.$$

These will belong to the different age-groups as detailed below :

Age - group	*Number of persons*
25—30	160–120* = 40
30—35	210
35—40	180
40 —45	450 – (40 + 210 + 180) = 20

[* : Because 120 persons from this group have been retrenched in (*i*).]

(*iii*) The number of persons to retire (from the highest age-groups)

$$= 10\% \text{ of } N = \frac{10}{100} \times 1{,}000 = 100.$$

These 100 persons are from the highest age-groups as shown :

Age-group	*Number of persons*
55—60	60
60—65	40

Hence incorporating the steps in (*i*), (*ii*) and (*iii*), the frequency distribution of the number of persons retained in the industrial house is as shown in the adjoining table :

Age-group	*Number of persons*
40—45	145 – 20 = 125
45—50	105
50—55	70

CALCULATION FOR AVERAGE AGE OF THOSE RETAINED

Age-group	*Mid-value* (x)	f	$d = \frac{x - 47.5}{5}$	fd
40—45	42·5	125	− 1	− 125
45—50	47·5	105	0	0
50—55	52·5	70	1	70
Total	$N = 300$		$\sum fd = -55$	

The average age of those retained is given by :

$$\bar{X} = A + \frac{h \sum fd}{N} = 47{\cdot}5 + \frac{5 \times (-55)}{300}$$

$$= 46{\cdot}5833 \simeq 47$$

Hence, the average age of those retained in the industrial house is 47 years.

2·5·2. Merits and Demerits of Arithmetic Mean

Merits	*Demerits*
1. It is rigidly defined.	1. It cannot be determined by inspection nor it can be located graphically.
2. It is easy to understand and easy to calculate.	2. Arithmetic mean cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitively; such as, intelligence, honesty, beauty, etc.
3. It is based upon all the observations.	3. Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values.
4. It is amenable to algebraic treatment. The mean of the composite series in terms of the means and sizes of the component series is given by : $\bar{x} = \sum_{i=1}^{k} n_i \bar{x}_i / (\sum_{i=1}^{k} n_i)$	4. Arithmetic mean is affected very much by extreme values. In case of extreme items, arithmetic mean gives a distorted picture of the distribution and no longer remains representative of the distribution.
5. Of all the averages, arithmetic mean is affected least by fluctuations of sampling. This property is sometimes described by saying that arithmetic mean is a *stable* average. Thus, we see that arithmetic mean satisfies all the properties laid down by Prof. Yule for an ideal average.	5. Arithmetic mean may lead to wrong conclusions if the details of the data from which it is computed are not given. Let us consider the following marks obtained by two students A and B in three tests, *viz.*, terminal test, half-yearly examination and annual examination respectively.
	Marks — *I Test* — *II Test* — *III Test* — *Average* A — 50% — 60% — 70% — 60% B — 70% — 60% — 50% — 60%
	Thus average marks obtained by each of the two students at the end of the year are 60%. If we are given the average marks alone we conclude that the level of intelligence of both the students at the end of the year is same. This is a fallacious conclusion since we find from the data that student A has improved consistently while student B has deteriorated consistently.
	6. Arithmetic mean cannot be calculated if the extreme class is open, e.g., below 10 or above 90.
	7. In extremely asymmetrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location.

2·5·3. Weighted Mean. In calculating arithmetic mean we suppose that all the items in the distribution have equal importance. But in practice this may not be so. If some items in a distribution are more important than others, then this point must be borne in mind, in order that average computed is representative of the distribution. In such cases, proper weightage is to be given to various items; the weights attached to each item being proportional to the importance of the item in the distribution. For example, if we want to have an idea of the change in cost of living of a certain group of people, then the simple mean of the prices of the commodities consumed by them will not do, since all the commodities are not equally important, *e.g.*, wheat, rice and pulses are more important than cigarettes, tea, confectionery, etc.

Let w_i be the weight attached to the item x_i, $i = 1, 2, \ldots, n$. Then we define :

$$\text{Weighted arithmetic mean (or weighted mean)} = \sum_i w_i x_i \,/ \sum_i w_i \qquad \ldots (2\cdot5)$$

It may be observed that the formula for weighted mean is the same as the formula for simple mean with f_i, $(i = 1, 2, \ldots, n)$, *the frequencies replaced by* w_i, $(i = 1, 2, \ldots, n)$, *the weights.*

Weighted mean gives the result equal to the simple *mean* if the weights assigned to each of the variate values are equal. It results in higher value than the simple mean if smaller weights are given to smaller items and larger weights to larger items. If the weights attached to larger items are smaller and those attached to smaller items are larger, then the weighted mean results in smaller value than the simple mean.

Example 2·5. *Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.*

Solution. The first n natural numbers are : $1, 2, 3, \ldots, n$.

We know that

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

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

X	w	wX
1	1	1^2
2	2	2^2
3	3	3^2
⋮	⋮	⋮
n	n	n^2

$$\text{Simple A.M. } (\bar{X}) = \frac{\Sigma X}{n} = \frac{1 + 2 + 3 + \ldots + n}{n} = \frac{n+1}{2}$$

$$\text{Weighted A.M. } (\bar{X}_w) = \frac{\Sigma w X}{\Sigma w} = \frac{1^2 + 2^2 + \ldots + n^2}{1 + 2 + \ldots + n} = \frac{n(n+1)(2n+1)}{6} \cdot \frac{2}{n(n+1)} = \frac{2n+1}{3}$$

2·6. MEDIAN

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, *i.e.*, it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a *positional* average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 25, 20, 15, 35, 18, *i.e.*, 15, 18, 20, 25, 35 is 20 and the median of 8, 20, 50, 25, 15, 30, *i.e.*, of 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20 + 25) = 22{\cdot}5$.

Remarks. In case of even number of observations, in fact any value lying between the two middle values can be taken as median but conventionally we take it to be the mean of the middle terms.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below :

(*i*) Find $\frac{1}{2}N$, where $N = \sum f_i$.

(*ii*) See the (less than) cumulative frequency (*c.f.*) just greater than $\frac{1}{2}N$.

(*iii*) The corresponding value of x is median.

Example 2·6. *Obtain the median for the following frequency distribution :*

x :	1	2	3	4	5	6	7	8	9
f :	8	10	11	16	20	25	15	9	6

Solution.

COMPUTATION OF MEDIAN

x	f	*c.f.*
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
Total	$N = 120$	

Here $N = 120 \quad \Rightarrow \quad \frac{N}{2} = 60$

The cumulative frequency (*c.f.*) just greater than $\frac{1}{2}N$ is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

2·6·1. Median for Continuous Frequency Distribution. In the case of continuous frequency distribution, the class corresponding to the *c.f.* just greater than $\frac{1}{2}N$ is called the *median class* and the value of median is obtained by the following formula :

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right) \qquad \ldots (2\cdot6)$$

where l is the lower limit of the median class,

f is the frequency of the median class,

h is the magnitude of the median class,

'c' is the *c.f.* of the class *preceding* the median class,

and $N = \sum f$.

Derivation of the Median Formula (2·6). Let us consider the following continuous frequency distribution, $(x_1 < x_2 < \ldots < x_{n+1})$:

Class interval :	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$	...	$x_n - x_{n+1}$
Frequency :	f_1	f_2	...	f_k	...	f_n

The cumulative frequency distribution is given by :

Class interval :	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$	...	$x_n - x_{n+1}$
Frequency :	F_1	F_2	...	F_k	...	F_n

where $F_i = f_1 + f_2 + \ldots + f_i$. The class $x_k - x_{k+1}$ is the median class if and only if $F_{k-1} < \frac{1}{2}N < F_k$.

Now, if we assume that the variate values are uniformly distributed over the median-class which implies that the ogive is a straight line in the median-class, then we get from the *Fig.* 2·2.

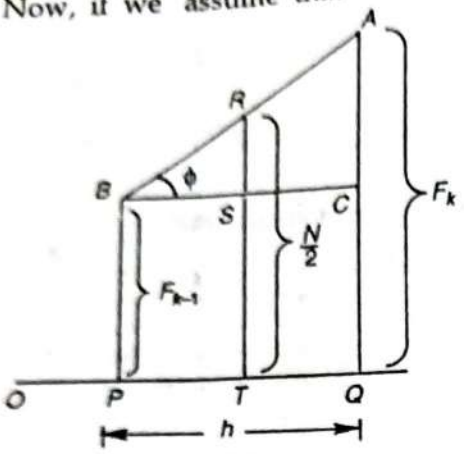


Fig. 2·2.

$$\tan\phi = \frac{RS}{BS} = \frac{AC}{BC}$$

i.e.,
$$\frac{RT - TS}{BS} = \frac{AQ - CQ}{BC}$$

or
$$\frac{RT - BP}{BS} = \frac{AQ - BP}{PQ}$$

or
$$\frac{\frac{1}{2}N - F_{k-1}}{BS} = \frac{F_k - F_{k-1}}{PQ} = \frac{f_k}{h},$$

where f_k is the frequency and h the magnitude of the median class.

$$\therefore \quad BS = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Hence

$$\text{Median} = OT = OP + PT = OP + BS = l + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right),$$ which is the required formula.

Remark. The median formula (2·6) can be used only for continuous classes without any gaps, *i.e.*, for *'exclusive type'* classification. If we are given a frequency distribution in which classes are of *'inclusive type'* with gaps, then it must be converted into a continuous *'exclusive type'* frequency distribution without any gaps before applying (2·6). This will affect the value of l in (2·6). As an illustration see *Example* 2·8.

Example 2·7. *Find the median wage of the following distribution :*

Wages (in Rs.) :	2,000—3,000	3,000—4,000	4,000—5,000	5,000—6,000	6,000—7,000
No. of workers :	3	5	20	10	5

Solution. COMPUTATION OF MEDIAN

Wages (in Rs.)	*No. of employees*	*c.f.*
2,000—3,000	3	3
3,000—4,000	5	8
4,000—5,000	20	28
5,000—6,000	10	38
6,000—7,000	5	43

Here $\frac{1}{2}N = \frac{1}{2}(43) = 21{\cdot}5$.

Cumulative frequency just greater than 21·5 is 28 and the corresponding class is 4,000—5,000. Thus median class is 4,000—5,000.

Hence using (2·6), $\text{Median} = 4{,}000 + \frac{1{,}000}{20}(21{\cdot}5 - 8) = 4{,}000 + 675 = 4{,}675.$

Thus median wage is Rs. 4,675.

Example 2·8. *In a factory employing 3,000 persons, in a day 5 per cent work less than 3 hours, 580 work from 3·01 to 4·50 hours, 30 per cent work from 4·51 to 6·00 hours, 500 work from 6·01 to 7·50 hours, 20 per cent work from 7·51 to 9·00 hours and the rest work 9·01 or more hours. What is the median hours of work ?*

Solution. The given information can be expressed in tabular from as follows :

CALCULATIONS FOR MEDIAN WAGES

Work hours	*No. of employees (f)*	*Less than c.f.*	*Class boundaries*
Less than 3	$\frac{5}{100} \times 3{,}000 = 150$	150	Below 3·005
3·01—4·50	580	730	3·005—4·505
4·51—6·00	$\frac{30}{100} \times 3{,}000 = 900$	1,630	4·505—6·005
6·01—7·50	500	2,130	6·005—7·505
7·51—9·00	$\frac{20}{100} \times 3{,}000 = 600$	2,730	7·505—9·005
9·01 and above	3,000 − 2,730 = 270	3,000 = N	9·005 and above

Here $N = 3{,}000 \Rightarrow \frac{1}{2}N = 1{,}500$.

The *c.f.* just greater than 1500 is 1,630. The corresponding class 4·51—6·00, whose class boundaries are 4·505—6·005, is the median class.

Using the median formula, we get

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right) = 4{\cdot}505 + \frac{1{\cdot}5}{900}(1{,}500 - 730) = 4{\cdot}505 + 1{\cdot}283 \approx 5{\cdot}79$$

Hence, the median hours of work are 5·79.

Example 2·9. *An incomplete frequency distribution is given as follows :*

Variable	*Frequency*	*Variable*	*Frequency*
10—20	*12*	*50—60*	*?*
20—30	*30*	*60—70*	*25*
30—40	*?*	*70—80*	*18*
40—50	*65*	*Total*	*229*

Given that the median value is 46, determine the missing frequencies, using the median formula.

Solution. Let the frequency of the class 30—40 be f_1 and that of 50—60 be f_2.

Then $f_1 + f_2 = 229 - (12 + 30 + 65 + 25 + 18) = 79.$

Since median is given to be 46, the class 40—50 is the median class.

Hence using median formula (2·6), we get

$$46 = 40 + \frac{114{\cdot}5 - (12 + 30 + f_1)}{65} \times 10$$

$$\Rightarrow \quad 46 - 40 = \frac{72{\cdot}5 - f_1}{65} \times 10 \quad \Rightarrow \quad 6 = \frac{72{\cdot}5 - f_1}{6{\cdot}5}$$

$$\Rightarrow \quad f_1 = 72{\cdot}5 - 39 = 33{\cdot}5 \approx 34 \qquad \text{(Since frequency is never fractional)}$$

$$\therefore \quad f_2 = 79 - 34 = 45 \qquad \text{(Since } f_1 + f_2 = 79\text{)}$$

Example 2·10. *A number of particular articles has been classified according to their weights. After drying for two weeks the same articles have again been weighted and similarly classified. It is known that the median weight in the first weighing was 20·83 gm. while in the second weighing it was 17·35 gm. Some frequencies a and b in the first weighing and x and y in the second are missing. It is known that* $a = \frac{1}{3}x$ *and* $b = \frac{1}{2}y$. *Find out the values of the missing frequencies.*

Class	Frequencies		Class	Frequencies	
	1st weighing	2nd weighing		1st weighing	2nd weighing
0—5	a	x	15—20	52	50
5—10	b	y	20—25	75	30
10—15	11	40	25—30	22	28

Solution. We are given : $a = \frac{1}{3}x$ and $b = \frac{1}{2}y \Rightarrow x = 3a$ and $y = 2b$.

First weighing	*Second Weighing*
N_1 = Total frequency $= a + b + 11 + 52 + 75 + 22$ $= 160 + a + b$	N_2 = Total Frequency $= x + y + 40 + 50 + 30 + 28$ $= 148 + x + y$ $= 148 + 3a + 2b$
Since median is given to be 20·83, the class 20—25 is the median class.	Since median is given to be 17·35, the class 15—20 is the median class.
Using the median formula, $Md = l + \frac{h}{f}\left(\frac{N_1}{2} - c\right)$	Using the median formula, we get
$\Rightarrow \quad 20{\cdot}83 = 20 + \frac{5}{75}\left[\frac{N_1}{2} - (63 + a + b)\right]$	$17{\cdot}35 = 15 + \frac{5}{50}\left[\frac{148 + 3a + 2b}{2} - (40 + x + y)\right]$
$\Rightarrow \quad 15\,(20{\cdot}83 - 20) = \frac{160 + a + b}{2} - (63 + a + b)$	$\Rightarrow \quad 10 \times 2{\cdot}35 = 74 + \frac{3a + 2b}{2} - 40 - 3a - 2b$
$\Rightarrow \quad 12{\cdot}45 = 17 - \frac{a + b}{2}$	$\Rightarrow \quad \frac{3a + 2b}{2} = 34 - 23{\cdot}5 = 10{\cdot}5$
$\therefore \quad a + b = 2\,(17 - 12{\cdot}45)$ $= 9{\cdot}10 \simeq 9 \quad \ldots (*)$	$\Rightarrow \quad 3a + 2b = 21 \quad \ldots (**)$

Multiplying (*) by 3, we get $3a + 3b = 27$ . . . (***)

Subtracting (**) from (***), we get $b = 6$.

Substituting in (*), we have $a = 9 - 6 = 3$.

$\therefore \quad a = 3, \; b = 6; \; x = 3a = 9, \quad y = 2b = 12.$

2·6·2. Merits and Demerits of Median

Merits	*Demerits*
1. It is rigidly defined.	1. In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.
2. It is easily understood and is easy to calculate. In some cases it can be located merely by inspection.	2. It is not based on all the observations. For example, the median of 10, 25, 50, 60 and 64 is 50. We can replace the observations 10 and 25 by any two values which are smaller than 50 and the observations 60 and 65 by any two values greater than 50, without affecting the value of median. This property is sometimes described by saying that median is *insensitive*.
3. It is not at all affected by extreme values.	3. It is not amenable to algebraic treatment.
4. It can be calculated for distributions with open-end classes.	4. As compared with mean, it is affected much by fluctuations of sampling.

Uses. (*i*) Median is the only average to be used while dealing with qualitative data which annot be measured quantitatively but still can be arranged in ascending or descending order of agnitude, *e.g.*, to find the average intelligence or average honesty among a group of people.

(*ii*) It is to be used for determining the typical value in problems concerning wages, istribution of wealth, etc.

7. MODE

Let us consider the following statements :

(*i*) The average height of an Indian (male) is 5'-6".

(*ii*) The average size of the shoes sold in a shop is 7.

(*iii*) An average student in a hostel spends Rs. 750 per month.

In all the above cases, the average referred to is mode. *Mode is the value which curs most frequently in a set of observations and around which the other items of the set uster densely. In other words, mode is the value of the variable which is predominant in the ries.* Thus, in the case of discrete frequency distribution, mode is the value of x rresponding to maximum frequency. For example, in the following frequency stribution :

x :	1	2	3	4	5	6	7	8
f :	4	9	16	25	22	15	7	3

lue of x corresponding to the maximum frequency, *viz.*, 25 is 4. Hence mode is 4.

But in any one (or more) of the following cases :

(*i*) if the maximum frequency is repeated,

(*ii*) if the maximum frequency occurs in the very beginning or at the end of the stribution, and

(*iii*) if there are irregularities in the distribution,

value of mode is determined by the *method of grouping*, which is illustrated below an example.

Example 2·11. *Find the mode of the following frequency distribution :*

Size (x)	:	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f)	:	3	8	15	23	35	40	32	28	20	45	14	6

Solution . Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below :

Size (x)	Frequency (i)	(ii)	(iii)	(iv)	(v)	(vi)
1	3					
		11				
2	8		23			
				26		
3	15				46	
		38				73
4	23		58			
				98		
5	35					
		75			107	
6	40		72			
						100
7	32					
		60		80		
8	28		48			
					93	
9	20					
		65				79
10	45		59			
				65		
11	14					
		20				
12	6					

The frequencies in column (i) are the original frequencies. Column (ii) is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two, we get column (iii). Combining the frequencies two by two after *leaving* the first two frequencies results in a repetition of column (ii). Hence, we proceed to combine the frequencies three by three, thus getting column (iv). The combination of frequencies three by three after leaving the first frequency results in column (v) and after leaving the first two frequencies results in column (vi).

The maximum frequency in each column is given in black type. To find mode we form the following table :

ANALYSIS TABLE

Column Number (1)	Maximum Frequency (2)	Value or combination of values of x giving max. frequency in (2) (3)
(i)	45	10
(ii)	75	5, 6
(iii)	72	6, 7
(iv)	98	4, 5, 6
(v)	107	5, 6, 7
(vi)	100	6, 7, 8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

2·7·1. Mode for Continuous Frequency Distribution. In case of continuous frequency distribution, mode is given by the formula :

$$Mode = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}, \quad \ldots (2\cdot7)$$

where l is the lower limit, h the magnitude and f_1 the frequency of the modal class, and f_0 and f_2 are frequencies of the classes preceding and succeeding the modal class respectively.

Derivation of the Mode Formula (2·7). Let us consider the continuous frequency distribution :

Class :	$x_1—x_2$	$x_2—x_3$	...	$x_k—x_{k+1}$	...	$x_n—x_{n+1}$
Frequency :	f_1	f_2	...	f_k	...	f_n.

where all the classes are of equal magnitude, say, h units.

If f_k is the maximum of all the frequencies, then the model class is $(x_k—x_{k+1})$.

Let us further consider a portion of the histogram, namely, the rectangles erected on the modal class and the two adjacent classes. The mode is the value of x for which the frequency curve has a maxima. Let the modal point be Q.

From the adjoining figure,

$$\tan\theta = \frac{LD}{LM} = \frac{NC}{MN}$$

and $$\tan\phi = \frac{LM}{AL} = \frac{MN}{NB}$$

$$\therefore \quad \frac{LM}{MN} = \frac{LD}{NC} = \frac{AL}{NB} = \frac{AL + LD}{NB + NC} = \frac{AD}{BC},$$

i.e., $$\frac{LM}{LN - LM} = \frac{PD - AP}{BR - CR}$$

or $$\frac{LM}{h - LM} = \frac{f_k - f_{k-1}}{f_k - f_{k+1}},$$

where 'h' is the magnitude of the modal class. Thus, solving for LM, we get

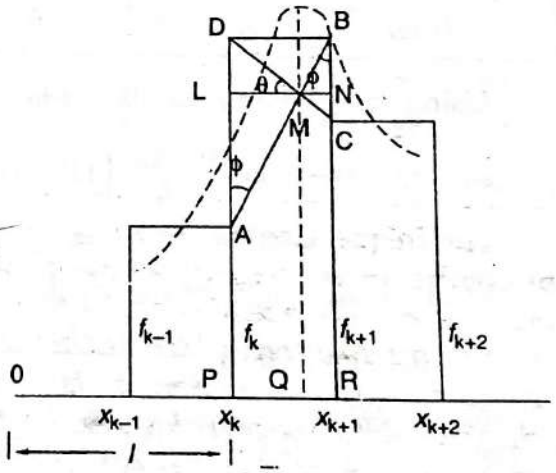


Fig. 2·3

$$LM = \frac{h(f_k - f_{k-1})}{(f_k - f_{k+1}) + (f_k - f_{k-1})} = \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Hence $$Mode = OQ = OP + PQ = OP + LM = l + \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}.$$

Example 2·12. *Find the mode for the following distribution :*

Class-interval :	0—10	10—20	20—30	30—40	40—50	50—60	60—70	70—80
Frequency :	5	8	7	12	28	20	10	10

Solution. Here maximum frequency is 28. Thus the class 40—50 is the modal class. Using mode formula (2·7), the value of mode is given by :

$$Mode = 40 + \frac{10(28 - 12)}{(2 \times 28 - 12 - 20)} = 40 + 6\cdot666 = 46\cdot67 \text{ (approx.)}$$

Example 2·13. *The median and mode of the following wage distribution are known to be Rs. 3,350 and Rs. 3,400 respectively. Find the values of f_3, f_4 and f_5 :*

Wages (in Rs.)	*No. of Employees*	*Wages (in Rs.)*	*No. of Employees*
0—1,000	4	4,000—5,000	f_5
1,000—2,000	16	5,000—6,000	6
2,000—3,000	f_3	6,000—7,000	4
3,000—4,000	f_4	*Total*	230

Solution .

CALCULATION FOR MEDIAN AND MODE

Wages (in Rs.)	*Frequency (f)*	*Less than c.f.*
0—1,000	4	4
1,000—2,000	16	20
2,000—3,000	f_3	$20 + f_3$
3,000—4,000	f_4	$20 + f_3 + f_4$
4,000—5,000	f_5	$20 + f_3 + f_4 + f_5$
5,000—6,000	6	$26 + f_3 + f_4 + f_5$
6,000—7,000	4	$N = 30 + f_3 + f_4 + f_5$
Total	$N = 230$	

From the adjoining table,

$N = \Sigma f = 30 + f_3 + f_4 + f_5 = 230$

$\Rightarrow f_3 + f_4 + f_5 = 230 - 30 = 200$ ... (*)

Since median is 3,350, which lies in the class 3,000—4,000, 3,000—4,000 is the median class.

Using the median formula : $Md = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$

$\Rightarrow \quad 3{,}350 = 3{,}000 + \frac{1{,}000}{f_4}\left[115 - (20 + f_3)\right]$ or $\frac{3{,}350 - 3{,}000}{1{,}000} = \frac{95 - f_3}{f_4}$

$\Rightarrow \quad 0{\cdot}35\, f_4 = 95 - f_3 \quad \Rightarrow \quad f_3 = 95 - 0{\cdot}35\, f_4$ ... (**)

Mode being 3,400, the modal class is also 3,000—4,000. Using mode formula :

$$3{,}400 = 3{,}000 + \frac{1{,}000\,(f_4 - f_3)}{2f_4 - f_3 - f_5} \Rightarrow \frac{3{,}400 - 3{,}000}{1{,}000} = \frac{f_4 + 0{\cdot}35\, f_4 - 95}{2f_4 - (200 - f_4)}$$ [Using (*) and (**)]

$$\Rightarrow \quad 0{\cdot}4 = \frac{1{\cdot}35\, f_4 - 95}{3f_4 - 200} \quad \Rightarrow \quad f_4 = \frac{95 - 80}{1{\cdot}35 - 1{\cdot}20} = \frac{15}{0{\cdot}15} = 100$$

Substituting in (**), we have $f_3 = 95 - 0{\cdot}35 \times 100 = 60$

Substituting the values of f_3 and f_4 in (*), we get $f_5 = 200 - f_3 - f_4 = 40$

Hence $f_3 = 60$, $f_4 = 100$ and $f_5 = 40$.

Remarks **1.** In case of irregularities in the distribution, or the maximum frequency being repeated or the maximum frequency occurring in the very beginning or at the end of the distribution, the modal class is determined by the method of grouping and the mode is obtained by using (2·7).

Sometimes, mode is estimated from the mean and the median. For a symmetric distribution (See § 2·16), mean, median and mode coincide. If the distribution is *moderately asymmetrical*, the mean, median and mode obey the following empirical relationship (due to Karl Pearson) :

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode}) \quad \Rightarrow \quad \text{Mode} = 3\,\text{Median} - 2\,\text{Mean} \quad \text{... (2}$$

2. If the method of grouping gives the modal class which does not correspond to the maximum frequency, *i.e.*, the frequency of modal class is not the maximum frequency, then in some situations we may get $2f_k - f_{k-1} - f_{k+1} = 0$. In such cases, the value of mode can be obtained by the formula :

$$\text{Mode} = l + \frac{h\,(f_k - f_{k-1})}{|f_k - f_{k-1}| + |f_k - f_{k+1}|}$$

2·7·2. Merits and Demerits of Mode

Merits	*Demerits*
1. Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection.	1. Mode is ill-defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called *bi-modal.* If a distribution has more than two modes, it is said to be *multimodal.*
2. Mode is not at all affected by extreme values.	2. It is not based upon all the observations.
3. Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open-end classes also do not pose any problem in the location of mode.	3. It is not capable of further mathematical treatment. 4. As compared with mean, mode is affected to a greater extent, by fluctuations of sampling.

Uses. Mode is the average to be used to find the ideal size, *e.g.*, in business forecasting, in the manufacture of ready made garments, shoes, etc.

2·8. GEOMETRIC MEAN

Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G, of n observations x_i; $i = 1, 2, ..., n$ is given by :

$$G = (x_1 . x_2 \ldots x_n)^{1/n} \qquad \ldots (2\cdot9)$$

The computation is facilitated by use of logarithms. Taking logarithm of both sides,

$$\log G = \frac{1}{n}\left(\log x_1 + \log x_2 + \ldots + \log x_n\right) = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$\therefore \quad G = \text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right) \qquad \ldots (2\cdot9a)$$

In case of frequency distribution $x_i \mid f_i$, $(i = 1, 2,..., n)$ geometric mean, G is :

$$G = \left(x_1^{f_1} . x_2^{f_2} \ldots x_n^{f_n}\right)^{\frac{1}{N}}, \quad \text{where } N = \sum_{i=1}^{n} f_i \qquad \ldots (2\cdot10)$$

Taking logarithms of both sides, we get

$$\log G = \frac{1}{N}\left(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n\right) = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i \qquad \ldots (2\cdot10a)$$

Thus we find that logarithm of geometric mean is the arithmetic mean of the logarithms of the given values. From (2·10*a*), we get

$$G = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x\right) \quad \ldots (2\cdot10b)$$

In the case of grouped or continuous frequency distribution, x is taken to be the value corresponding to the mid-point of the class intervals.

2·8·1. Geometric Mean of the Combined Group. If n_1 and n_2 are the sizes, G_1 and G_2 the geometric means of two series respectively, the geometric mean G, of the combined series is given by :

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \quad \ldots (2\cdot11)$$

Proof. Let $x_{1i}\,(i = 1, 2, \ldots, n_1)$ and $x_{2j}\,(j = 1, 2, \ldots, n_2)$ be n_1 and n_2 items of two series respectively. Then by *def.*,

$$G_1 = (x_{11} \cdot x_{12} \ldots x_{1n_1})^{1/n_1} \quad \Rightarrow \quad \log G_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \log x_{1i},$$

$$G_2 = (x_{21} \cdot x_{22} \ldots x_{2n_2})^{1/n_2} \quad \Rightarrow \quad \log G_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \log x_{2j}$$

The geometric mean G of the combined series is given by :

$$G = (x_{11} \cdot x_{12} \ldots x_{1n_1} \cdot x_{21} \cdot x_{22} \ldots x_{2n_2})^{1/(n_1+n_2)}$$

$$\Rightarrow \quad \log G = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1} \log x_{1i} + \sum_{j=1}^{n_2} \log x_{2j}\right) = \frac{1}{n_1 + n_2}\left(n_1 \log G_1 + n_2 \log G_2\right)$$

$$\Rightarrow \quad G = \text{Antilog}\left(\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right) \quad \ldots (2\cdot11a)$$

In general, if $G_1, G_2, \ldots, G_k$ are the geometric means of k groups having $n_1, n_2 \ldots, n_k$ observations respectively, the geometric mean G of the combined group consisting of $n_1 + n_2 + \ldots + n_k$ observations is given by :

$$G = \text{Antilog}\left(\frac{n_1 \log G_1 + n_2 \log G_2 + \ldots + n_k \log G_k}{n_1 + n_2 + \ldots + n_k}\right) \quad \ldots (2\cdot11b)$$

2·8·2. Merits and Demerits of Geometric Mean

Merits	*Demerits*
1. It is rigidly defined.	1. Because of its abstract mathematical character, geometric mean is not easy to understand and to calculate for a non-mathematics person.
2. It is based upon all the observations.	
3. It is suitable for further mathematical treatment.	
4. It is not affected much by fluctuations of sampling.	2. If any one of the observations is zero, geometric *mean* becomes zero and if any one of the observations is negative, geometric mean becomes imaginary regardless of the magnitude of the other items.
5. It gives comparatively more weight to small items.	

Uses. Geometric mean is used :

(*i*) To find the rate of population growth and the rate of interest.

(*ii*) In the construction of index numbers.

Example 2·14. *Show that in finding the arithmetic mean of a set of readings on a thermometer it does not matter whether we measure temperature in Centigrade or Fahrenheit, but that in finding the geometric mean it does matter which scale we use.*

Solution. Let $C_1, C_2, \ldots, C_n$ be the n readings on the Centigrade thermometer. Then their arithmetic mean is given by : $\overline{C} = \frac{1}{n}(C_1 + C_2 + \ldots + C_n)$

If F and C be the readings in Fahrenheit and Centigrade respectively, then we have the relation : $\frac{F-32}{180} = \frac{C}{100} \quad \Rightarrow \quad F = 32 + \frac{9}{5}C$.

Thus the Fahrenheit equivalents of $C_1, C_2, \ldots, C_n$ are :

$$32 + \frac{9}{5}C_1, 32 + \frac{9}{5}C_2, \ldots, 32 + \frac{9}{5}C_n, \text{ respectively.}$$

Hence the arithmetic mean of the readings in Fahrenheit is

$$\overline{F} = \frac{1}{n}\left\{(32 + \frac{9}{5}C_1) + (32 + \frac{9}{5}C_2) + \ldots + (32 + \frac{9}{5}C_n)\right\}$$

$$= \frac{1}{n}\left\{32n + \frac{9}{5}(C_1 + C_2 + \ldots + C_n)\right\} = 32 + \frac{9}{5}\left(\frac{C_1 + C_2 + \ldots + C_n}{n}\right)$$

$$= 32 + \frac{9}{5}\overline{C}, \text{ which is the Fahrenheit equivalent of } \overline{C}.$$

Hence in finding the arithmetic mean of a set of n readings on a thermometer, it is immaterial whether we measure temperature in Centigrade or Fahrenheit.

Geometric mean G, of n readings in Centigrade is : $G = (C_1 . C_2 \ldots C_n)^{1/n}$

Geometric mean G_1, (say), of Fahrenheit equivalents of $C_1, C_2, \ldots, C_n$ is :

$$G_1 = \left\{(32 + \frac{9}{5}C_1)(32 + \frac{9}{5}C_2) \ldots (32 + \frac{9}{5}C_n)\right\}^{1/n}$$

which is not equal to Fahrenheit equivalent of G, *viz.*, $\left\{\frac{9}{5}(C_1 . C_2 \ldots C_n)^{1/n} + 32\right\}$

Hence in finding the geometric mean of the n readings on a thermometer, the scale (Centigrade or Fahrenheit) is important.

Example 2·15. *The geometric mean of 10 observations on a certain variable was calculated as 16·2. It was later discovered that one of the observations was wrongly recorded as 12·9; in fact it was 21·9. Apply appropriate correction and calculate the correct geometric mean.*

Solution. The geometric mean G of n observations $x_1, x_2, \ldots, x_n$ is given by :

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

If x_1 is the observation copied wrongly instead of correct value x_1', then the *corrected geometric* mean G' is given by :

$$G' = (x_1' x_2 x_3 \ldots x_n)^{1/n} = \left(\frac{x_1'}{x_1} . x_1 x_2 \ldots x_n\right)^{1/n}$$

$$\Rightarrow \quad G' = (x_1 x_2 \ldots x_n)^{1/n}\left(\frac{x_1'}{x_1}\right)^{1/n} = G . \left(\frac{x_1'}{x_1}\right)^{1/n} \qquad \text{[From (*)]} \quad \ldots (**)$$

In the given problem, $G = 16{\cdot}2$, $n = 10$, $x_1 = 12{\cdot}9$, $x_1' = 21{\cdot}9$

$$\therefore \quad \text{Corrected } G.M. (G') = 16{\cdot}2 \times \left(\frac{21{\cdot}9}{12{\cdot}9}\right)^{1/10} \qquad \text{[From (**)]}$$

$$\Rightarrow \quad \log_{10} G' = \log_{10} 16{\cdot}2 + \frac{1}{10}(\log_{10} 21{\cdot}9 - \log_{10} 12{\cdot}9)$$

$$= 1{\cdot}2095 + \frac{1}{10}(1{\cdot}3404 - 1{\cdot}1106) = 1{\cdot}23248$$

$$\therefore \quad G' = \text{Antilog } (1{\cdot}2325) = 17{\cdot}08.$$

Example 2·16. *In a frequency table, the upper boundary of each class interval has a constant ratio to the lower boundary. Show that the geometric mean G may be expressed by the formula :* $\log G = x_0 + \frac{c}{N} \sum_i f_i (i-1)$,

where x_0 is the logarithm of the mid-value of the first interval and c is the logarithm of the ratio between upper and lower boundaries.

Solution. Let the ith class be denoted by $l_i - l_{i+1}$, with the corresponding frequency f_i ; $i = 1, 2, 3, \ldots$

Let the constant ratio between upper and lower limits of each class be λ, so that :

$$\frac{l_2}{l_1} = \frac{l_3}{l_2} = \ldots \frac{l_i}{l_{i-1}} = \ldots = \lambda, \text{ (say)} \Rightarrow l_2 = \lambda l_1, l_3 = \lambda l_2 = \lambda^2 l_1, \ldots, l_i = \lambda l_{i-1} = \lambda^{i-1} \cdot l_1 \quad \ldots (i)$$

If x_i is the mid-point of the ith class, then we have

$$x_1 = \frac{1}{2}(l_1 + l_2) = \frac{1}{2}(1 + \lambda) l_1 \quad \text{[From } (i)] \quad \ldots (ii)$$

and $$x_i = \frac{1}{2}(l_i + l_{i+1}) = \frac{1}{2}(\lambda^{i-1} l_1 + \lambda^i l_1)$$

$$= \frac{1}{2} l_1 (1 + \lambda) \lambda^{i-1} = x_1 \cdot \lambda^{i-1} \quad \text{[From } (ii)] \quad \ldots (iii)$$

Geometric mean G of the distribution is given by :

$$\log G = \frac{\sum f_i \log x_i}{\sum f_i} = \frac{1}{N} \sum_i f_i \log (x_1 \cdot \lambda^{i-1}) \quad \text{[From } (iii)]$$

$$= \frac{1}{N} \sum_i f_i \log x_1 + \frac{1}{N} \sum_i f_i (i-1) \log \lambda$$

$$= \log x_1 \cdot \frac{1}{N} \sum_i f_i + \log \lambda \cdot \frac{1}{N} \sum_i (i-1) f_i$$

$$= x_0 + c \sum_i (i-1) f_i \quad [\because x_0 = \log x_1 \text{ and } c = \log \lambda \text{ (Given)}]$$

2·9. HARMONIC MEAN

Harmonic mean of a number of observations, none of which is zero, is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus harmonic mean (H), of n observations x_i, $i = 1, 2, \ldots, n$ is given by :

$$H = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} (1/x_i)} \quad \ldots (2{\cdot}12$$

In case of frequency distribution $x_i \mid f_i$, $(i = 1, 2, \ldots, n)$,

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^{n} (f_i/x_i)}, \quad \left[N = \sum_{i=1}^{n} f_i\right] \quad \ldots (2{\cdot}12$$

If $x_1, x_2, \ldots, x_n$ are n observations with weights $w_1, w_2, \ldots, w_n$ respectively, the weighted harmonic mean is defined as : $H = \frac{\sum w_i}{\sum (w_i/x_i)}$. $\ldots (2{\cdot}12$

2·9·1. Merits and Demerits of Harmonic Mean

Merits	Demerits
Harmonic mean is rigidly defined, based upon all the observations and is suitable for further mathematical treatment. Like geometric mean, it is not affected much by fluctuations of sampling. It gives greater importance to small items and is useful only when small items have to be given a greater weightage.	Harmonic mean is not easily understood and is difficult to compute.

Example 2·17. *A cyclist pedals from his house to his college at a speed of 10 km. p.h. and back from the college to his house at 15 km. p.h. Find the average speed.*

Solution . Let the distance from the house to the college be x km. In going from house to college, the distance (x kilometre) is covered in $\frac{x}{10}$ hours, while in coming from college to house, the distance is covered in $\frac{x}{15}$ hours. Thus a total distance of $2x$ km. is covered in $\left(\frac{x}{10}+\frac{x}{15}\right)$ hours.

$$\text{Hence average speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}} = \frac{2x}{\left(\frac{x}{10}+\frac{x}{15}\right)} = 12 \text{ km. p.h}$$

In this case the average speed is given by the harmonic mean of 10 and 15 and not by the arithmetic mean.

Remarks 1. If equal distances are covered (travelled) per unit of time with speeds equal to $V_1, V_2, \ldots, V_n$, say, then the average speed is given by the harmonic mean of $V_1, V_2, \ldots, V_n$, *i.e.*,

$$\text{Average speed} = \frac{1}{\frac{1}{n}\left(\frac{1}{V_1}+\frac{1}{V_2}+\ldots+\frac{1}{V_n}\right)} = \frac{n}{\sum_i\left(\frac{1}{V_i}\right)} \quad \ldots (2\cdot12b)$$

2. Instead of fixed (constant) distance being travelled with varying speed, let us now suppose that different distances, say, $S_1, S_2, \ldots, S_n$ are travelled with different speeds, say, $V_1, V_2, \ldots, V_n$ respectively. In that case, the average speed is given by the weighted harmonic mean of the speeds, the weights being the corresponding distances travelled, *i.e.*,

$$\text{Average speed} = \frac{S_1+S_2+\ldots+S_n}{\left(\frac{S_1}{V_1}+\frac{S_2}{V_2}+\ldots+\frac{S_n}{V_n}\right)} = \frac{\sum_i S_i}{\sum_i\left(\frac{S_i}{V_i}\right)} \quad \ldots (2\cdot12c)$$

3. The harmonic mean, like arithmetic mean, is also used in averaging of ratios like price per unit, km. per hour, work done per hour, etc. under certain conditions. It may be noted here that a rate represents a ratio, *e.g.*, price $= \frac{\text{money}}{\text{quantity}}$, speed $= \frac{\text{distance}}{\text{time}}$, work done per hour $= \frac{\text{work done}}{\text{time taken}}$, *etc.* The average of a rate, defined by the ratio $\frac{p}{q}$, is given by the arithmetic mean of its values in different situations if the conditions are given in terms of q and by the harmonic mean if the conditions are given in terms of p.

Example 2·18. *(a) Milk is sold at the rates of 8, 10, 12 and 15 rupees per litre in four different months. Assuming that equal amounts are spent on milk by a family in the four months, find the average price in rupees per month.*

(*b*) *An individual purchases three qualities of pencils. The relevant data are given below :*

Quality	Price per pencil (Rs.)	Money spent (Rs.)
A	1·00	50
B	1·50	30
C	2·00	20

Calculate the average price per pencil.

Solution. (*a*). Since equal amounts of money are spent by the family for each of the four months, the average price of milk per month is given by the harmonic mean of 8, 10, 12 and 15.

$\therefore$ Average price of milk per month

$$= \text{Rs.}\ \frac{1}{\frac{1}{4}\left(\frac{1}{8}+\frac{1}{10}+\frac{1}{12}+\frac{1}{15}\right)} = \text{Rs.}\ \frac{4\times 120}{15+12+10+8} = \text{Rs.}\ \frac{4\times 120}{45} = \text{Rs. } 10{\cdot}67$$

(*b*) Here we are given : Total expenditure = Rs. (50 + 30 + 20) = Rs. 100

$$\text{Total number of pencils purchased} = \frac{50}{1}+\frac{30}{1{\cdot}50}+\frac{20}{2} = 80$$

$$\text{Average price per pencil} = \frac{\text{Total expenditure}}{\text{Total No. of pencils}} = \frac{100}{80} = \text{Rs. } 1{\cdot}25.$$

Remark. Average price of Rs. 1·25 can also be obtained by finding the weighted harmonic mean (*H.M.*) of 1, 1·5, and 2 with corresponding weights 50, 30 and 20 respectively.

Example 2·19. *You can take a trip which entails travelling 900 km. by train at an average speed of 60 km. per hour, 3,000 km. by boat at an average speed of 25 km. p.h., 400 km. by plane at 350 km. per hour and finally 15 km. by taxi at 25 km. per hour. What is your average speed for the entire distance ?*

Solution. Since different distances are covered with varying speeds, the required average speed for the entire distance is given by the weighted harmonic mean of the speeds (in km. p.h.), the weights being the corresponding distances covered (in km.).

COMPUTATION OF WEIGHTED H. M.

Speed (km./hr.) x	*Distance (in km.)* w	w/x
60	900	15·00
25	3,000	120·00
350	400	1·43
25	15	0·60
Total	4,315	137·03

Average speed

$$= \frac{\sum w_i}{\sum (w_i/x_i)} = \frac{4315}{137{\cdot}03} = 31{\cdot}489\ km.p.h$$

2·10. SELECTION OF AN AVERAGE

From the preceding discussion, it is evident that no single average is suitable for all practical purposes. Each one of the averages has its own merits and demerits and thus its own particular field of importance and utility. We cannot use the averages indiscriminately. A judicious selection of the average depending on the nature of the data and the purpose of the inquiry is essential for sound statistical analysis. Since

arithmetic mean satisfies all the properties of an ideal average as laid down by Prof. Yule ; is familiar to a layman and further has wide applications in statistical theory at large, it may be regarded as the best of all the averages.

2·11. PARTITION VALUES

These are the values which divide the series into a number of equal parts.

The three points which divide the series into four equal parts are called *quartiles*. The first, second and third points are known as the first, second and third quartiles respectively. The first quartile, Q_1, is the value which exceed 25% of the observations and is exceeded by 75% of the observations. The second quartile, Q_2, coincides with median. The third quartile, Q_3, is the point which has 75% observations before it and 25% observations after it.

The nine points which divide the series into ten equal parts are called *deciles* whereas *percentiles* are the ninety-nine points which divide the series into hundred equal parts. For example, D_7, the seventh decile, has 70% observations before it and P_{47}, the forty-seventh percentile, is the point which exceeds 47% of the observations. The methods of computing the partition values are the same as those of locating the median in the case of both discrete and continuous distributions.

Example 2·20. *Eight coins were tossed together and the number of heads resulting was noted. The operation was repeated 256 times and the frequencies (f) that were obtained for different values of x, the number of heads, are shown in the following table. Calculate median, quartiles, 4th decile and 27th percentile.*

x:	0	1	2	3	4	5	6	7	8
f:	1	9	26	59	72	52	29	7	1

Solution.

x:	0	1	2	3	4	5	6	7	8
f:	1	9	26	59	72	52	29	7	1
c.f. :	1	10	36	95	167	219	248	255	256

Median : Here $\frac{1}{2}N = \frac{1}{2} \times 256 = 128$. Cumulative frequency (*c.f.*) just greater than 128 is 167. Thus, median = 4.

Q_1 : Here $\frac{1}{4}N = 64$. and *c.f.* just greater than 64 is 95. Hence, $Q_1 = 3$.

Q_3 : Here $\frac{3}{4}N = 192$ and *c.f.* just greater than 192 is 219. Thus, $Q_3 = 5$.

$D_4 : \frac{4}{10}N = 4 \times 25{\cdot}6 = 102{\cdot}4$ and *c.f.* just greater than 102·4 is 167. Hence, $D_4 = 4$.

$P_{27} : \frac{27}{100}N = 27 \times 2{\cdot}56 = 69{\cdot}12$ and *c.f.* just greater than 69·12 is 95. Hence $P_{27} = 3$.

Example 2·21. *Following is the distribution of marks obtained by 500 candidates in Statistics paper of a civil services examination :*

Marks more than :	0	10	20	30	40	50
Number of Candidates :	500	460	400	200	100	30

Calculate the lower quartile marks . If 70% of the candidates pass in the paper, find the minimum marks obtained by a pass candidate.

Solution.

COMPUTATION OF LOWER QUARTILE AND THIRD DECILE

Marks more than	*Cumulative frequency*	*Class intervals*	*Frequency (f)*	*Less than c.f.*
0	500 = N	0—10	500 – 460 = 40	40
10	460	10—20	460 – 400 = 60	100
20	400	20—30	400 – 200 = 200	300
30	200	30—40	200 – 100 = 100	400
40	100	40—50	100 – 30 = 70	470
50	30	50 and above	30	500

$\frac{1}{4}N = \frac{1}{4}(500) = 125.$

The less than *c.f.* just greater than 125 is 300. Hence the corresponding class 20—30 contains Q_1.

$$\therefore \quad Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - C\right) = 20 + \frac{10}{200}(125 - 100) = 21{\cdot}25$$

Since out of 500 students, 70% pass the test, *i.e.*, 30% fail in the test, the minimum marks obtained by a pass candidate are given by D_3.

$\frac{3N}{10} = \frac{3 \times 500}{10} = 150$. The *c.f.* just greater than 150 is 300. Therefore, D_3 lies in the corresponding class 20—30, and is given by :

$$D_3 = l + \frac{h}{f}\left(\frac{3N}{10} - C\right) = 20 + \frac{10}{200}(150 - 100) = 22{\cdot}5$$

Hence minimum marks obtained by the pass candidate are 22·5.

Example 2·22. *For a group of 5,000 shopkeepers the daily earnings vary from Rs. 200 to Rs. 800. The earnings of 4 per cent of the shopkeepers are under Rs. 250 and those of 10 per cent are under Rs. 300 ; 15 per cent of the shopkeepers earn Rs. 600 and over, and 5 per cent of them earn Rs. 700 and over. The quartile earnings are Rs. 400 and Rs. 540, and the sixth decile is Rs. 500. Put this information in the form of a frequency table.*

Solution. We are given $N = 5{,}000$.

(*i*) 4% of 5,000 = (4 × 5,000/100) = 200 shopkeepers earn under Rs. 250.
(*ii*) 10% of 5,000 = 500 shopkeepers earn under Rs. 300.
(*iii*) 15% of 5,000 = 750 shopkeepers earn Rs. 600 and over, *i.e.*, ≥ Rs. 600.
(*iv*) 5% of 5,000 = 250 shopkeepers earn ≥ Rs. 700.
(*v*) Q_1 = Rs. 400 ⇒ 25% of 5,000 = 1, 250 shopkeepers earn less than Rs. 400.
(*vi*) Q_3 = Rs. 540 ⇒ 75% of 5,000 = 3,750 shopkeepers earn less than Rs. 540.
(*vii*) D_6 = Rs. 500 ⇒ 60% of 5,000 = 3,000 shopkeepers earn less than Rs. 500.

Since the earnings vary from Rs. 200 to Rs. 800, the first class will start with 200 and the last class will end with 800.

The above information can be expressed in the following Table :

FREQUENCY DISTRIBUTION OF EARNINGS OF SHOPKEEPERS

Value of the Variable (daily earnings in Rs.)	*No. of Shopkeepers*	*Daily Earnings (Rs.)*	*No. of Shopkeepers*
Less than 250	200	200—250	200
Less than 300	500	250—300	500 – 200 = 300
Less than 400	1,250	300—400	1,250 – 500 = 750
Less than 500	3,000	400—500	3,000 – 1,250 = 1,750
Less than 540	3,750	500—540	3,750 – 3,000 = 750
----------		540—600	5,000 – (3,750 + 750) = 500
More than 600	750	600—700	750 – 250 = 500
More than 700	250	700—800	250

FREQUENCY DISTRIBUTION OF EARNINGS OF SHOPKEEPERS

Daily Earnings (Rs.)	*No. of Shopkeepers (f)*
200 and under 300	200 + 300 = 500
300 and under 400	750
400 and under 500	1,750
500 and under 600	750 + 500 = 1,250
600 and under 700	500
700 and under 800	250

In the above table, the various classes are of unequal widths. Rearranging and combining them to have classes with equal magnitude of 100 each, the final frequency distribution of wages of 5,000 workers is as shown in the adjoining table.

2·11·1. Graphical Location of the Partition Values. The partition values, *viz.*, quartiles, deciles and percentiles, can be conveniently located with the help of a curve called the *'cumulative frequency curve'* or *'Ogive'*. The procedure is illustrated below :

First, form the 'less than cumulative frequency' table. Take the class intervals (or the variate values) along the x-axis and plot the corresponding cumulative frequencies along the y-axis against the *upper limit* of the class interval (or against the variate value in the case of discrete frequency distribution). The curve obtained on joining the points so obtained by means of free hand drawing is called the *less than cumulative frequency curve* or *less than ogive*. Similarly, by plotting the more than *c.f.* against the lower limit of the corresponding class and joining the points so obtained by a smooth free hand curve, we obtain *'more than ogive'*. The graphical location of partition values from this curve is explained below by means of an example.

Example 2·23. *Draw the cumulative frequency curve for the following distribution showing the number of marks of 59 students in Statistics.*

Marks-group :	0—10	10—20	20—30	30—40	40—50	50—60	60—70
No. of Students :	4	8	11	15	12	6	3

Solution.

Marks-group	*No. of Students*	*Less than c.f.*	*More than c.f.*
0—10	4	4	59
10—20	8	12	55
20—30	11	23	47
30—40	15	38	36
40—50	12	50	21
50—60	6	56	9
60—70	3	59	3

The smooth curve obtained on joining these points is called *'less than'* ogive.

If we plot the 'more than' cumulative frequencies, *viz.*, 59, 55, ..., 3 against the lower limits of the corresponding classes, *viz.*, 0, 10, ..., 60 and join the points by a smooth curve, we get 'more than cumulative frequency curve' or *'more than'* ogive.

Taking the marks-group along x-axis and *c.f.* along y-axis, we plot the less than cumulative frequencies, *viz.*, 4, 12, 23, ... , 59 against the upper limits of the corresponding classes, *viz.*, 10, 20, 30, ..., 70 respectively.

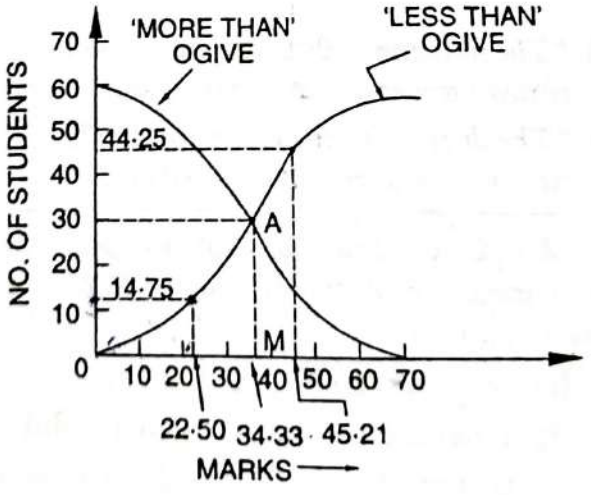


Fig. 2·4

To locate graphically the value of median, mark a point corresponding to $\frac{1}{2}N$ along y-axis. At this point draw a line parallel to x-axis meeting the ogive at the point 'A' (say). From 'A' draw a *line* perpendicular to x-axis meeting it in 'M' (say). Then abscissa of 'M' *viz.*, OM gives the value of median.

To locate the values of Q_1 (or Q_3), we mark the points along y-axis corresponding to $\frac{1}{4}N$ (or $\frac{3}{4}N$) and proceed exactly similarly.

In the above example, we get Median = 34·33, Q_1 = 22·50 and Q_3 = 45·21

Remarks 1. The median can also be located as follows :

From the point of intersection of 'less than' ogive and 'more than' ogive, draw perpendicular to OX. The abscissa of the point so obtained gives median.

2. Other partition values, *viz.*, deciles and percentiles, can be similarly located from 'ogive'.

2·12. DISPERSION

Averages (or the measures of central tendency) give us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone, we cannot form a complete idea about the distribution as will be clear from the following example.

Consider the series (*i*) 7, 8, 9, 10, 11, (*ii*) 3, 6, 9, 12, 15, and (*iii*) 1, 5, 9, 13, 17. In all these cases we see that n, the number of observations, is 5 and the mean is 9. If we are given that the mean of 5 observations is 9, we cannot form an idea as to whether it is the average of first series or second series or third series or of any other series of 5 observations whose sum is 45. Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures. One such measure is *Dispersion.*

Literal meaning of dispersion is '*scatteredness*'. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. In the above case we say that series (*i*) is more homogeneous (less dispersed) than the series (*ii*) or (*iii*) or we say that series (*iii*) is more heterogeneous (more scattered) than the series (*i*) or (*ii*).

Some important definitions of dispersion are given below :

(*i*) *"Dispersion is the measure of extent to which individual items vary."*
—L.R. Connor

(*ii*) *"The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion."* —Simpson and Kafka

(*iii*) *"The degree to which numerical data tend to spread about an average value is called variation or dispersion of the data."* —Spiegel

2·12·1. Characteristics for an Ideal Measure of Dispersion. The desiderata for an ideal measure of dispersion are the same as those for an ideal measure of central tendency, *viz.*,

(*i*) It should be rigidly defined.

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

(*iii*) It should be based on all the observations.

(*iv*) It should be amenable to further mathematical treatment.

(*v*) It should be affected as little as possible by fluctuations of sampling.

2·13. MEASURES OF DISPERSION

Various measures of dispersion can be classified into two broad categories :

(*a*) The measures which express the spread of observations in terms of distance between the values of selected observations. These are also termed as *distance measures*, *e.g.*, range and interquartile range (or quartile deviation).

(*b*) The measures which express the spread of observations in terms of the average of deviations of observations from some central value, *e.g.*, mean deviation and standard deviation.

2·13·1. Range. The range is the difference between two extreme observations of the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is given by :

$$\text{Range} = X_{max} - X_{min} = A - B \qquad \ldots (2\cdot13)$$

Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

2·13·2. Quartile Deviation. Quartile deviation or semi-interquartile range Q is given by :

$$Q = \frac{1}{2}(Q_3 - Q_1), \qquad \ldots (2\cdot13a)$$

where Q_1 and Q_3 are the first and third quartiles of the distribution respectively.

Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

2·13·3. Mean Deviation. If $x_i \mid f_i$, $i = 1, 2, \ldots, n$ is the frequency distribution, then mean deviation from the average A (usually mean, median or mode) is given by :

$$\text{Mean deviation from average } A = \frac{1}{N}\sum_{i=1}^{n} f_i \mid x_i - A \mid, \qquad \sum_i f_i = N \qquad \ldots (2\cdot14)$$

where $\mid x_i - A \mid$ represents modulus or the absolute value of the deviation $(x_i - A)$, where the negative sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

Remark. It may be pointed out here that mean deviation is least when taken from median. (The proof is given for continuous variable in chapter 5. Example 5·29)

Example 2·24. *Calculate : (i) Quartile deviation (Q.D.), and (ii) Mean Deviation (M.D.) from mean, for the following data :*

Marks.	:	*0—10*	*10—20*	*20—30*	*30—40*	*40—50*	*50—60*	*60—70*
No. of Students	:	*6*	*5*	*8*	*15*	*7*	*6*	*8*

Solution.

CALCULATION FOR Q.D. AND M.D. FROM MEAN

Marks	Mid-value (x)	No. of Students (f)	$d = \frac{x-35}{10}$	fd	$\lvert x-\bar{x}\rvert = \lvert x-33{\cdot}4\rvert$	$f\lvert x-\bar{x}\rvert$	Less than c.f.
0—10	5	6	−3	−18	28·4	170·4	6
10—20	15	5	−2	−10	18·4	92·0	11
20—30	25	8	−1	−8	8·4	67·2	19
30—40	35	15	0	0	1·6	24·0	34
40—50	45	7	1	7	11·6	81·2	41
50—60	55	6	2	12	21·6	129·6	47
60—70	65	3	3	9	31·6	94·8	50
Total		50		−8		659·2	

(*i*) Here $N = 50$; $\frac{1}{4}N = 12{\cdot}75$; $\frac{3}{4}N = 37{\cdot}25$

The *c.f.* just greater than 12·75 is 19. Hence, the corresponding class 20—30 contains Q_1.

$\therefore \quad Q_1 = 20 + \frac{10}{8}(12{\cdot}75 - 11) = 22{\cdot}19$

The *c.f.* just greater than 37·25 is 41. Hence, the corresponding class 40—50 contains Q_3.

$\therefore \quad Q_3 = 40 + \frac{10}{7}(37{\cdot}25 - 34) = 44{\cdot}64$

Hence, $Q.D. = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(44{\cdot}64 - 22{\cdot}19) = 11{\cdot}23$

(*ii*) Mean, $(\bar{x}) = A + \frac{h\sum fd}{N} = 35 + \frac{10\times(-8)}{50} = 33{\cdot}4$ marks.

$\therefore \quad$ M.D. (from mean) $= \frac{1}{N}\sum f\lvert x-\bar{x}\rvert = \frac{659{\cdot}2}{50} = 13{\cdot}184.$

Example 2·25. *Prove that the mean deviation about the mean $\bar{x}$ of the variable x, the frequency of whose ith size x_i is f_i is given by :*

$$\frac{2}{N}\left(\bar{x}\sum_{x_i<\bar{x}} f_i - \sum_{x_i<\bar{x}} f_i x_i\right), \quad N = \sum_i f_i$$

Solution. Mean deviation (M.D.) about mean $= \frac{1}{N}\sum_i f_i \lvert x_i - \bar{x}\rvert$

$$= \frac{1}{N}\left(\sum_{x_i<\bar{x}} f_i\lvert x_i-\bar{x}\rvert + \sum_{x_i>\bar{x}} f_i\lvert x_i-\bar{x}\rvert\right) = \frac{1}{N}\left[\sum_{x_i<\bar{x}} f_i(\bar{x}-x_i) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right]$$

$$= \frac{1}{N}\left[-\sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right] \quad \ldots(*)$$

Since, the algebraic sum of deviations about mean is zero, we have

$$\sum_i f_i(x_i-\bar{x}) = 0 \Rightarrow \sum_{x_i>\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) = 0 \quad \Rightarrow \quad \sum_{x_i>\bar{x}} f_i(x_i-\bar{x}) = -\sum_{x_i<\bar{x}} f_i(x_i$$

Substituting in (*), we get

$$\text{M.D.} = \frac{1}{N}\left[-2\sum_{x_i<\bar{x}} f_i(x_i-\bar{x})\right] = \frac{2}{N}\left[-\sum_{x_i<\bar{x}} f_i x_i + \bar{x}\sum_{x_i<\bar{x}} f_i\right] = \frac{2}{N}\left[\bar{x}\sum_{x_i<\bar{x}} f_i - \sum_{x_i<\bar{x}} f_i x_i\right]$$

2·13·4. Standard Deviation and Root Mean Square Deviation. Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i \mid f_i ; i = 1, 2, \ldots, n$,

$$\sigma = \sqrt{\frac{1}{N}\sum_i f_i (x_i - \bar{x})^2}, \qquad \ldots (2\cdot15)$$

where $\bar{x}$ is the arithmetic mean of the distribution and $\sum_i f_i = N$.

The step of squaring the deviations $(x_i - \bar{x})$ overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment [see equation (2·21), page 2·37]. Moreover, of all the measures, standard deviation is affected least by fluctuations of sampling.

Thus, we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion except for the general nature of extracting the square root which is not readily comprehensible for a non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favour with economists or businessmen who are not interested in the results of the modal class. Taking into consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and the most powerful measure of dispersion.

The square of standard deviation is called the *variance* and is given by :

$$\sigma^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 \qquad \ldots (2\cdot15a)$$

Root mean square deviation, denoted by 's', is given by :

$$s = \sqrt{\frac{1}{N}\sum_i f_i (x_i - A)^2} \qquad \ldots (2\cdot16)$$

where A is any arbitrary number. s^2 is called *mean square deviation.*

Remark. *Relation between* σ *and* s. By def., we have

$$s^2 = \frac{1}{N}\sum_i f_i (x_i - A)^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x} + \bar{x} - A)^2$$

$$= \frac{1}{N}\sum_i f_i \{(x_i - \bar{x})^2 + (\bar{x} - A)^2 + 2(\bar{x} - A)(x_i - \bar{x})\}$$

$$= \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 + (\bar{x} - A)^2 \frac{1}{N}\sum_i f_i + 2(\bar{x} - A)\frac{1}{N}\sum_i f_i (x_i - \bar{x}),$$

$(\bar{x} - A)$, being constant is taken outside the summation sign. But $\sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of the deviations of the given values from their mean. Thus

$$s^2 = \sigma^2 + (\bar{x} - A)^2 = \sigma^2 + d^2, \text{ where } d = \bar{x} - A$$

Obviously, s^2 will be least when $d = 0$, *i.e.*, $\bar{x} = A$. Hence mean square deviation and consequently root mean square deviation is least when the deviations are taken from $A = \bar{x}$, *i.e.*, *standard deviation is the least value of root mean square deviation.*

(*iv*) $\beta_2 = \frac{\mu_4}{\mu_2^2} \Rightarrow \mu_4 = \beta_2 \cdot \mu_2^2 = 3 \times 9^2 = 243$ [From (*)]

Uncorrected $\{\Sigma(x-\bar{x})^4\} = N\mu_4 = 250 \times 243 = 60{,}750$

$\therefore$ Corrected $\{\Sigma(x-\bar{x})^4\}$

$= 60{,}750 - \{(64-54)^4 + (50-54)^4\} + \{(62-54)^4 + (52-54)^4\} = 54{,}606$

$\Rightarrow$ Corrected $\mu_4 = \frac{54{,}606}{250} = 218{\cdot}42$

Hence, Corrected $\beta_2 = \frac{\text{Corrected } \mu_4}{\text{Corrected } \mu_2^2} = \frac{218{\cdot}42}{(8{\cdot}81)^2} = 2{\cdot}81.$

Example 2·41. *Show that if a range of six times the standard deviation covers at least 18 class intervals, Sheppard's correction will make a difference of less than 0·5 per cent in the uncorrected value of the standard deviation.*

Solution. Let σ be the calculated standard deviation and σ_1 be the corrected standard deviation after applying Sheppard's correction for grouping. If h is the width of the class intervals, then we are given :

$$6\sigma \geq 18h \quad \Rightarrow \quad h \leq \frac{\sigma}{3} \qquad \ldots (*)$$

After applying Sheppard's correction for μ_2, we get

$$\sigma_1^2 = \sigma^2 - \frac{h^2}{12} \quad \Rightarrow \quad \sigma^2 = \sigma_1^2 + \frac{h^2}{12} \leq \sigma_1^2 + \frac{\sigma^2}{12 \times 9} \qquad \text{[From (*)]}$$

$$\therefore \quad \sigma^2\left(1 - \frac{1}{108}\right) \leq \sigma_1^2$$

$$\Rightarrow \quad \sigma \leq \sigma_1\left(1 - \frac{1}{108}\right)^{-1/2} = \sigma_1\left\{1 + \left(-\frac{1}{2}\right)\left(-\frac{1}{108}\right) + \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{2!}\left(-\frac{1}{108}\right)^2 + \ldots\right\}$$

$$\Rightarrow \quad \sigma \leq \sigma_1\left(1 + \frac{1}{216}\right), \text{ approx.}$$

$$\therefore \quad \sigma - \sigma_1 \leq \frac{\sigma_1}{216} < \frac{\sigma_1}{200} = 0{\cdot}005\,\sigma_1 = 0{\cdot}5\% \text{ of } \sigma_1 .$$

Hence, the result.

2·16. SKEWNESS

Literally, skewness means *'lack of symmetry'*. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if —

(*i*) Mean, median and mode fall at different points, *i.e.*, Mean ≠ Median ≠ Mode;

(*ii*) Quartiles are not equidistant from median; and

(*iii*) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

2·16·1. Measures of Skewness. Various measures of skewness are :

(1) $S_k = M - M_d$, (2) $S_k = M - M_0$,

where M is the mean, M_d, the median and M_0, the mode of the distribution.

(3) $S_k = (Q_3 - M_d) - (M_d - Q_1)$.

These are the *absolute measures* of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the *relative measures* called the *coefficients of skewness* which are pure numbers independent of units of measurement. The following are the coefficients of skewness.

I. *Prof. Karl Pearson's Coefficient of Skewness.*

$$S_k = \frac{(M - M_0)}{\sigma}, \text{ where } \sigma \text{ is the standard deviation of the distribution.} \quad \ldots (2{\cdot}40)$$

If mode is ill-defined, then using the empirical relation, $M_0 = 3M_d - 2M$, for a moderately asymmetrical distribution, we get

$$S_k = \frac{3\,(M - M_d)}{\sigma} \quad \ldots (2{\cdot}40\,a)$$

From (2·40) and 2·40 (*a*), we observe that $S_k = 0$ if $M = M_0 = Md$.

Hence for a symmetrical distribution, mean, median and mode concide

Skewness is positive if $M > M_0$ or $M > M_d$ and negative if $M < M_0$ or $M < M_d$.

Remark. *Limits for Karl Pearson's Coefficient of Skewness :* $S_k = \frac{3(M - Md)}{\sigma}$

$$|M - Md| = \left|\frac{1}{n}\sum_{i=1}^{n} x_i - Md\right| = \left|\frac{1}{n}\sum_{i=1}^{n}(x_i - Md)\right| \le \frac{1}{n}\sum_{i=1}^{n}|x_i - Md| \le \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad \ldots (*)$$

(∵ The sum of the absolute deviations is minimum when taken about median.)

$$|S_k|^2 = \left|\frac{3\,(M - Md)}{\sigma}\right|^2 \le \frac{\left[3\cdot\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|\right]^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\left[3\sum_{i=1}^{n}|x_i - \bar{x}|\right]^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \{\text{From } (*)\} \quad \ldots (**)$$

∴ Using Cauchy-Schwartz inequality : {*c.f.* Chapter 7}

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right), \quad \text{with } b_i = 1;\ i = 1, 2, \ldots, n, \text{ we get}$$

$$\left(\sum_{i=1}^{n} a_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} 1\right) \quad \Rightarrow \quad \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{n\sum_{i=1}^{n} a_i^2} \le 1 \quad \ldots (***)$$

∴ $|S_K|^2 \le 3^2$ [From (**) and (***)] $\Rightarrow$ $|S_k| \le 3$ or $-3 \le S_k \le 3$.

Hence, the limits for Karl Pearson's coefficient of skewness are ± 3. However, in practice, these limits are rarely attained.

II. *Prof. Bowley's Coefficient of Skewness.* Based on quartiles,

$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \quad \ldots (2{\cdot}41)$$

Remarks **1.** Bowley's coefficient of skewness is also known as *Quartile Coefficient of Skewness* and is especially useful in situations where quartiles and median are used, *viz.*,

(*i*) When the mode is ill-defined and extreme observations are present in the data.

(*ii*) When the distribution has open end classes or unequal class intervals.

In these situations Pearson's coefficient of skewness cannot be used.

2. From (2·41), we observe that : $S_k = 0$, if $Q_3 - Md = Md - Q_1$.
This implies that for a symmetrical distribution ($S_k = 0$), median is equidistant from the upper and lower quartiles. Moreover skewness is positive if :

$$Q_3 - Md > Md - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 > 2Md$$

and skewness is negative if

$$Q_3 - Md < Md - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 < 2Md$$

3. *Limits for Bowley's Coefficient of Skewness.* We know that for two real positive numbers a and b (*i.e.*, $a > 0$ and $b > 0$), the modulus value of the difference $(a - b)$ is always less than or equal to the modulus value of the sum $(a + b)$, *i.e.*,

$$|a - b| \leq |a + b| \quad \Rightarrow \quad \left|\frac{a-b}{a+b}\right| \leq 1 \qquad \ldots (*)$$

We also know that $(Q_3 - Md)$ and $(Md - Q_1)$ are both non-negative.
Thus taking $a = Q_3 - Md$ and $b = Md - Q_1$ in (*), we get

$$\left|\frac{(Q_3 - Md) - (Md - Q_1)}{(Q_3 - Md) + (Md - Q_1)}\right| \leq 1 \quad \Rightarrow \quad |S_k \text{ (Bowley)}| \leq 1 \text{ or } -1 \leq S_k \text{ (Bowley)} \leq 1.$$

Thus, Bowley's coefficient of skewness ranges from – 1 to 1.
Further, we note from (2·41) that :

$S_k = +1$, if $Md - Q_1 = 0$, *i.e.*, if $Md = Q_1$
$S_k = -1$, if $Q_3 - Md = 0$, *i.e.*, if $Q_3 = Md$

4. It should be clearly understood that the values of the coefficients of skewness obtained by Bowley's formula and Pearson's formula are not comparable, although in each case, $S_k = 0$ implies the absence of skewness, *i.e.*, the distribution is symmetrical. It may even happen that one of them gives positive skewness while the other gives negative skewness.

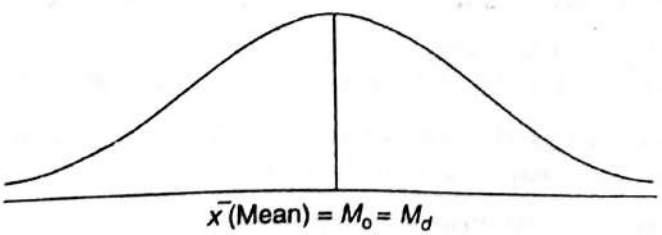
5. In Bowley's coefficient of skewness, the disturbing factor of variations is eliminated by dividing the absolute measure of skewness, *viz.*, $(Q_3 - Md) - (Md - Q_1)$ by the measure of dispersion $(Q_3 - Q_1)$, *i.e.*, quartile range.

6. The only and perhaps quite serious limitations of this coefficient is that it is based only on the central 50% of the data and ignores the remaining 50% of the data towards the extremes.

III. *Based upon moments, coefficient of skewness is :*

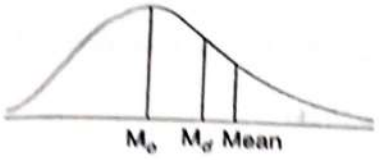
$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\,\beta_2 - 6\,\beta_1 - 9)} \qquad \ldots (2\cdot42)$$

where symbols have their usual meaning. Thus $S_k = 0$ if either $\beta_1 = 0$ or $\beta_2 = -3$. But since $\beta_2 = \mu_4/\mu_2^2$, cannot be negative, $S_k = 0$ if and only if $\beta_1 = 0$. Thus for a symmetrical distribution $\beta_1 = 0$. In this respect β_1 is taken to be a measure of skewness. The coefficient in (2·42) is to be regarded as without sign.

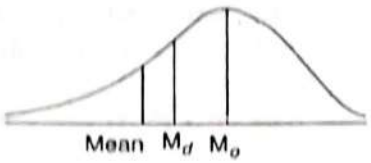


(Symmetric Distribution)
Fig. 2·5.

We observe in (2·40) and (2·41) that skewness can be positive as well as negative. The skewness is positive if the larger tail of the distribution lies towards the higher values of the variate (the right), *i.e.*, if the curve drawn with the help of the given data is stretched more to the right than to the left and is negative in the contrary case.



(Positively Skewed Distribution)
Fig. 2·6(a)



(Negatively Skewed Distribution)
Fig. 2·6(b)

2·17. KURTOSIS

If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the following figure in which all the three curves A, B and C are symmetrical about the mean 'm' and have the same range.

In addition to these measures, we should know one more measure which Prof. Karl Pearson calls as the *'Convexity of the Frequency Curve'* or *Kurtosis*. Kurtosis enables us to have an idea about the *'flatness or peakedness'* of the frequency curve. It is measured by the coefficient β_2 or its derivation γ_2 given by :

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \qquad \gamma_2 = \beta_2 - 3 \qquad \ldots (2\cdot42)$$

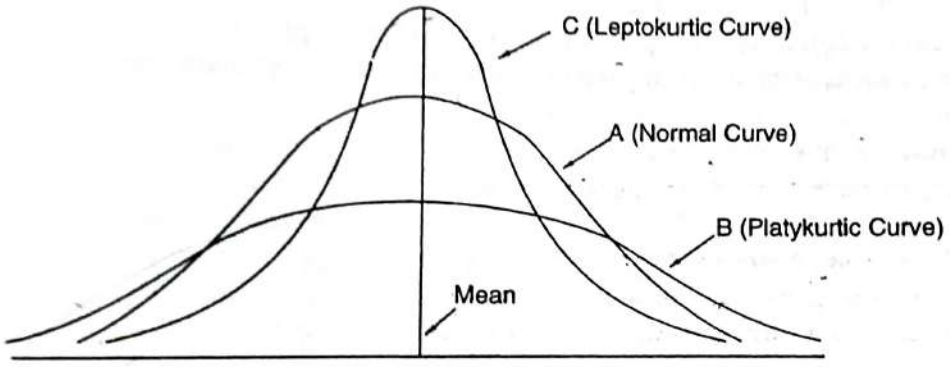


Fig. 2·7.

Curve of the type 'A' which is neither flat nor peaked is called the *normal curve* or *mesokurtic curve* and for such a curve $\beta_2 = 3$, *i.e.*, $\gamma_2 = 0$. Curve of the type 'B' which is flatter than the normal curve is known as *platykurtic* and for such a curve $\beta_2 < 3$, *i.e.*, $\gamma_2 < 0$. Curve of the type 'C' which is more peaked than the normal curve is called *leptokurtic* and for such a curve $\beta_2 > 3$, *i.e.*, $\gamma_2 > 0$.

Example 2·42. *For a distribution, the mean is 10, variance is 16, γ_1 is + 1 and β_2 is 4. Obtain the first four moments about the origin, i.e., zero. Comment upon the nature of distribution.*

Solution. We are given : Mean = 10, $\mu_2 = 16 \Rightarrow \sigma^2 = 16$ or $\sigma = 4$, $\gamma_1 = +1$, $\beta_1 = 4$

First four moments about origin $(\mu_1', \mu_2', \mu_3', \mu_4')$.

μ_1' = First moment about origin = Mean = 10

$\mu_2 = \mu_2' - \mu_1'^2 \quad \Rightarrow \quad \mu_2' = \mu_2 + \mu_1'^2 = 16 + 10^2 = 116$

We have $\gamma_1 = +1 \quad \Rightarrow \quad \dfrac{\mu_3}{\mu_2^{3/2}} = \dfrac{\mu_3}{\sigma^3} = 1$ or $\mu_3 = \sigma^3 = 4^3 = 64$

$$\therefore \quad \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\Rightarrow \quad \mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3 = 64 + 3 \times 116 \times 10 - 2 \times 1{,}000 = 1{,}544$$

Now $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 4 \quad \Rightarrow \quad \mu_4 = 4 \times 16^2 = 1024$

and $\mu_4 = \mu'_4 - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

$$\Rightarrow \quad \mu_4 = 1{,}024 + 4 \times 1{,}544 \times 10 - 6 \times 116 \times 100 + 3 \times 10{,}000 = 23{,}184.$$

Comments on nature of the distribution. Since $\gamma_1 = +1$, the distribution is moderately positively skewed, *i.e.*, if we draw the curve of the given distribution, it will have longer tail towards the right. Further since $\beta_2 = 4 > 3$, the distribution is leptokurtic, *i.e.*, it will be slightly more peaked then the normal curve.

Example 2·43. *For the frequency distribution of scores in mathematics of 50 candidates selected at random from among those appearing at a certain examination, compute the first four moments about the mean of the distribution.*

Scores :	50—60	60—70	70—80	80—90	90—100	100—110	110—120	120—130
Frequency :	1	0	0	1	1	2	1	0
Scores :	130—140	140—150	150—160	160—170	170—180	180—190	190—200	200—210
Frequency :	4	4	2	5	10	11	4	1
Scores :	210—220		220—230					
Frequency :	1		2					

Find also the corrected values of the moments after Sheppard's corrections are applied. Also obtain moment coefficients of skewness and kurtosis and comment on the nature of the distribution.

Solution.

Mid-value (x)	f	$d = \dfrac{x-135}{10}$	fd	fd^2	fd^3	fd^4
55	1	−8	−8	64	−512	4096
65	0	−7	0	0	0	0
75	0	−6	0	0	0	0
85	1	−5	−5	25	−125	625
95	1	−4	−4	16	−64	256
105	2	−3	−6	18	−54	162
115	1	−2	−2	4	−8	16
125	0	−1	0	0	0	0
135	4	0	0	0	0	0
145	4	1	4	4	4	4
155	2	2	4	8	16	32
165	5	3	15	45	135	405
175	10	4	40	160	640	2,560
185	11	5	55	275	1,375	6,875
195	4	6	24	144	864	5,184
205	1	7	7	49	343	2,401
215	1	8	8	64	512	4,096
225	2	9	18	162	1458	13,122
Total	50		150	1,038	4,584	39,834

The raw moments of variable d (about origin) are computed as :

$$\mu_1' = \frac{\Sigma fd}{\Sigma f} = \frac{150}{50} = 3{\cdot}00, \qquad \mu_2' = \frac{\Sigma fd^2}{\Sigma f} = \frac{1{,}038}{50} = 20{\cdot}76$$

$$\mu_3' = \frac{\Sigma fd^3}{\Sigma f} = \frac{4{,}584}{50} = 91{\cdot}68, \qquad \mu_4' = \frac{\Sigma fd^4}{\Sigma f} = \frac{39{,}834}{50} = 796{\cdot}68$$

The central moments of variable X are then computed as shown below :

$$\mu_2 = (\mu_2' - \mu_1'^2) \times h^2 = (20{\cdot}76 - 9{\cdot}00) \times 100 = 1{,}176$$

$$\mu_3 = (\mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3) \times h^3$$

$$= \{91{\cdot}68 - 3 \times 20{\cdot}76 \times 3 + 2\,(27)\} \times 1{,}000 = -41{,}160$$

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

$$= \{796{\cdot}68 - 4\,(91{\cdot}68)\,(3) + 6(20{\cdot}76)\,(9) - 3(81)\} \times 10{,}000 = 57{,}45{,}600.$$

Sheppard's Corrections for Moments

$$\bar{\mu}_2 = \mu_2 - \frac{h^2}{12} = 1{,}176 - \frac{100}{12} = 1{,}167{\cdot}67, \qquad \bar{\mu}_3 = \mu_3 = -41{,}160$$

$$\bar{\mu}_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4 = 57{,}45{,}600 - 58{,}800 + 291{\cdot}67 = 56{,}87{,}091{\cdot}67$$

Moment coefficient of skewness is given by :

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{-41{,}160}{1{,}176\sqrt{1{,}176}} = -\frac{41{,}160}{40{,}328{\cdot}40} = -1{\cdot}02$$

$$\text{Moment coefficient of kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{57{,}45{,}600}{(1{,}176)^2} = 4{\cdot}15.$$

Interpretation : Since $\gamma_1 = -1{\cdot}02$, *i.e.*, (negative), the distribution is moderately negatively skewed, *i.e.*, the frequency curve of the given distribution has a longer tail towards the left .

Further, since $\beta_2 = 4{\cdot}15 > 3$, the given distribution is *leptokurtic*, *i.e*, the frequency curve is more peaked than the normal curve.

CHAPTER CONCEPTS QUIZ

1. Which measure of location will be suitable to compare :
   (*i*) heights of students in two classes ; (*ii*) size of agricultural holdings ;
   (*iii*) average sales for various years ; (*iv*) intelligence of students ;
   (*v*) per capita income in several countries ;
   (*vi*) marks obtained 10, 8, 12, 4, 7, 11, and X ($X < 5$).
2. Which of the following are true for all sets of data ?
   (*i*) Arithmetic mean ≤ median ≤ mode, (*ii*) Arithmetic mean ≥ median ≥ mode,
   (*iii*) Arithmetic mean = median = mode (*iv*) None of these
3. Which of the following are true in respect of any distribution ?
   (*i*) The percentile points are in the ascending order.
   (*ii*) The percentile points are equispaced.
   (*iii*) A unique median value exists for each and every distribution.

12. What is standard deviation ? Explain its superiority over other measures of dispersion.
13. (*a*) Define the raw and central moments of a frequency distribution. Obtain the relation between the central moments of order r in terms of the raw moments. What are Sheppard's corrections to the central moments ?
    (*b*) Define moments. Establish the relationship between the moments about mean, *i.e.*, Central Moments in terms of moments about any arbitrary point and *vice versa*.
14. Define 'Moment'. What is its use ? Express first four central moments in terms of moments about the origin. What is the effect of change of origin and scale on moments ?
15. What is Sheppard's correction ? What will be the corrections for the first four moments ?
16. What do you understand by skewness ? How is it measured ? Distinguish clearly, by giving figures, between positive and negative skewness. Also show the relative positions of mean, median and mode in the figures, for positively and negatively skewed distributions.
17. Explain the methods of measuring skewness and kurtosis of a frequency distribution.
18. Show that for any frequency distribution :
    (*i*) Kurtosis is *greater* than unity.
    (*ii*) Bowley's coefficient of skewness is *less* than unity numerically.
    (*iii*) Karl Pearson's coefficient of skewness lies between – 3 and 3.
19. (*a*) Why do we calculate in general only the first four moments about mean of a distribution and not the higher moments ?
    (*b*) "Measures of central tendency, dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distribution." Explain.
20. (*a*) Define Pearsonian coefficients β_1 and β_2, and discuss discuss their utility in Statistics.
    (*b*) What do you mean by skewness and kurtosis of a distribution ? Show that the Pearson's Beta coefficients satisfy the inequality $\beta_2 - \beta_1 - 1 \geq 0$. Also deduce the $\beta_2 \geq 1$.
    (*c*) Define Karl Pearson's coefficients γ_1 and γ_2 and discuss their utility in Statistics.

ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

2·1. The following numbers give the weights of 55 students of a class. Prepare a suitable frequency table :

42	74	40	60	82	115	41	61	75	83	63
53	110	76	84	50	67	65	78	77	56	95
68	69	104	80	79	79	54	73	59	81	100
66	49	77	90	84	76	42	64	69	70	80
72	50	79	52	103	96	51	86	78	94	71

(*i*) Draw the histogram and frequency polygon of the above data. From the histogram, obtian an approximate value of mode.

(*ii*) For the above weights, prepare a cumulative frequency table and draw the less than ogive. Hence, obtain an approximate value of median.

2·2. (*a*) What are the points to be borne in mind in the formation of a frequency table ?

Choosing appropriate classintervals, form a frequency table for the following data :

10·2	0·5	5·2	6·1	3·1	6·7	8·9	7·2	8·9
5·4	3·6	9·2	6·1	7·3	2·0	1·3	6·4	8·0

4·3	4·7	12·4	8·6	13·1	3·2	9·5	7·6	4·0
5·1	8·1	1·1	11·5	3·1	6·8	7·0	8·2	2·0
3·1	6·5	11·2	12·0	5·1	10·9	11·2	8·5	2·3
3·4	5·2	10·7	4·9	6·2				

(*b*) What are the considerations one has to bear in mind while forming a frequency distribution ?

A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 integers and to begin the first class at 199·5, find the class limits and class marks of the seven classes.

(*c*) The class marks in a frequency table (of whole numbers) are given to be 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. Find out the following :

(*i*) the true classes, (*ii*) the true class limits, (*iii*) the true upper class limits.

2·3. (*a*) The following table shows the distribution of the number of students per teacher in 750 colleges :

Students :	1	4	7	10	13	16	19	22	25	28
Frequency :	7	46	165	195	189	89	28	19	9	3

Draw the histogram for the data and superimpose on it the frequency polygon.

(*b*) Draw the histogram and frequency curve for the following data :

Monthly wages in '000 Rs. :	11—13	13—15	15—17	17—19	19—21	21—23	23—25
No. of workers :	6	53	85	56	21	16	8

(*c*) Draw a histogram for the following data :

Age (in years) :	2—5	5—11	11—12	12—14	14—15	15—16
No. of boys :	6	6	2	5	1	3

2·4. (*a*) Three persons *A*, B and C were given the job of finding the average of 5,000 numbers. Each one did his own simplification. *A*'s method : Divide the sets into sets of 1,000 each, calculate the average in each set and then calculate the average of these averages. *B*'s method : Divide the set into 2,000 and 3,000 numbers, take average in each set and then take the average of the averages. *C*'s method : 500 numbers were unities. He averaged all other numbers and then added one. Are these methods correct ?

(*b*) The total sale (in '000 rupees) of a particular item in shop, on 10 consecutive days, is reported by a clerk as, 35·00, 29·60, 38·00, 30·00, 40·00, 41·00, 42·00, 45·00, 3·60, 3·80. Calculate the average. Later it was found that there was a number 10·00 in the machine and the reports of 4th to 8th days were 10·00 more than the true values and in the last 2 days he put a decimal in the wrong place thus for example 3·60 was really 36·0. Calculate the true mean value.

2·5. (*a*) Given below is the distribution of 140 candidates obtaining marks *X* or higher in a certain examination (all marks are given in whole numbers) :

X :	10	20	30	40	50	60	70	80	90	100
More than *c.f.* :	140	133	118	100	75	45	25	9	2	0

Calculate the mean, median and mode of the distribution.

(*b*) The four parts of a distribution are as follows :

Part	*Frequency*	*Mean*
1	50	61
2	100	70
3	120	80
4	30	83

Find the mean of the distribution.

2·6. (*a*) Define a 'weighted mean'. If several sets of observations are combined into a single set, show that the mean of the combined set is the weighted mean of several sets.

(*b*) The weighted geometric mean of three numbers 229, 275 and 125 is 203. The weights for the first and second numbers are 2 and 4 respectively. Find the weight of third.

2·7. The following table shows some data collected for the regions of a country :

Region	*Number of inhabitants (million)*	*Percentage of literates*	*Average annual income per person (Rs.)*
A	10	52	850
B	5	68	620
C	18	39	730

Obtain the overall figures for the three regions taken together. Prove the formulae you use.

2·8. For the two frequency distributions given in the adjoining table, the mean calculated from the first was 25·4 and that from the second was 32·5. Find the values of x and y.

Class	*Distribution I Frequency*	*Distribution II Frequency*
10—20	20	4
20—30	15	8
30—40	10	4
40—50	x	$2x$
50—60	y	y

2·9. Draw the ogives and hence estimate the median.

Class	: 0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79
Frequency	: 8	32	142	216	240	206	143	13

2·10. The adjoining data relate to the ages of a group of workers in a factory. Draw the cumulative frequency curve and find from the graph the number of workers between the ages 28—48.

Ages	*No. of workers*	*Ages*	*No. of workers*
20—25	35	40—45	90
25—30	45	45—50	74
30—35	70	50—55	51
35—40	105	55—60	30

2·11. (*a*) The mean of marks obtained in an examination by a group of 100 students was found to be 49·96. The mean of the marks obtained in the same examination by another group of 200 students was 52·32. Find the mean of the marks obtained by both the groups of students taken together.

(*b*) A distribution consists of three components with frequencies 300, 200 and 600, having their means 16, 8 and 4 respectively. Find the mean of the combined distribution.

(*c*) The mean marks obtained by 300 students in the subject of Statistics are 45. The mean of the top 100 of them was found to be 70 and the mean of the last 100 was known to be 20. What is the mean of the remaining 100 students ?

(*d*) The mean weight of 150 students in a certain class is 60 kilograms. The mean weight of boys in the class is 70 kilograms and that of the girls is 55 kilograms.

Find the number of boys and number of girls in the class.

2·12. From the following table showing the wage distribution in a certain factory determine :

(*a*) the mean wages, (*b*) the median wages, (*c*) the modal wages,

(*d*) the wage limits for the middle 50% of the wage earners,

(*e*) the percentages of workers who earned between Rs. 75 and Rs. 125,

(*f*) the percentage of workers who earned more than Rs. 150 per week, and

(*g*) the percentage of workers who earned less than Rs. 100 per week.

Weekly wages (Rs.)	*No. of employees*	*Weekly wages (Rs.)*	*No. of employees*
20—40	8	120—140	35
40—60	12	140—160	18
60—80	20	160—180	7
80—100	30	180—200	5
100—120	40		

2·13. (*a*) The following table gives the frequency distribution of marks in a class of 65 students :

Marks	*No. of Students*	*Marks*	*No. of students*
0—4	10	14—18	5
4—8	12	18—20	3
8—12	18	20—25	4
12—14	7	25 and over	6
Total			65

Calculate : (*i*) Upper and lower quartiles.

(*ii*) Number of students who secured marks more than 17.

(*iii*) Number of students who secured marks between 10 and 15.

(*b*) The following table shows the age distribution of heads of families in a certain country during the year 2001. Find the median, the third quartile and the second decile of the distribution. Check your results by the graphical method.

Age of head of family (years)	Under 25	25—29	30—34	35—44	45—54	55—64	65—74	above 74
Number (million)	2·3	4·1	5·3	10·6	9·7	6·8	4·4	1·8

2·14. The following data represent travel expenses (other than transportation) for 7 trips made during November by a salesman for a small firm :

Trip	*Days*	*Expenses (Rs.)*	*Expenses per day (Rs.)*
1	0·5	13·50	27
2	2·0	12·00	6
3	3·5	17·50	5
4	1·0	9·00	9
5	9·0	27·00	3
6	0·5	9·00	18
7	8·5	17·00	2
Total	25·0	105·00	70

An auditor criticised these expenses as excessive, asserting that the average expenses per ay is Rs. 10 (Rs. 70 divided by 7). The salesman replied that the average is only Rs. 4·20 Rs. 105 divided by 25) and that in any event the median is the appropriate measure and is only

Rs. 3. The auditor rejoined that the arithmetic mean is the appropriate measure, but that the median is Rs. 6.

You are required to :

(*a*) Explain the proper interpretation of each of the four averages mentioned.

(*b*) Which average seems appropriate to you ?

2·15. (*a*) An assessee depreciated the machinery of his factory by 10% each in the first two years and by 40% in the third year and thereby claimed 21% average depreciation relief from taxation department, but the I.T.O. objected and allowed only 20%. Show which of the two is right.

(*b*) A given machine is assumed to depreciate 40% in value in the first year, 25% in the second year and 10% per annum for the next three years, each percentage being calculated on the diminishing value. What is the average percentage depreciation, reckoned on the diminishing value for the five years ?

(*c*) An economy grows at the rate of 2% in the first year, 2·5% in the second, 3% in the third, 4% in the fourth, ... and 10% in the tenth year. What is the average rate of growth of the economy ?

(*d*) The export of a commodity increased by 30% in 1998, decreased by 22% in 1999 and then increased by 45% in the following year. The increase/decrease in each year, being measured in comparison to its previous year. Calculate the average rate of change of the exports per annum.

2·16. (*a*) You take a tip which entails travelling 900 miles by train at an average speed of 60 m.p.h.; 300 miles by boat at an average of 25 m.p.h.; 400 miles by plane at 350 m.p.h. and finally 15 miles by taxi at 25 m.p.h. What is your speed for the entire distance ?

(*b*) A train runs 25 miles at a speed of 30 m.p.h.; another 50 miles at a speed of 40 m.p.h.; then due to repairs of the track travels for 6 minutes at a speed of 100 m.p.h and finally covers the remaining distance of 24 miles at a speed of 24 m.p.h. What is the average speed in m.p.h ?

(*c*) A man motors from *A* to *B*. A large part of the distance is uphill and he gets a mileage of only 10 per gallon of gasoline. On the return trip he makes 15 miles per gallon. Find the harmonic mean of his mileage. Verify the fact that this is the proper average to be used by assuming that the distance from *A* to *B* is 60 miles.

(*d*) Calculate the average speed of a car running at the rate of 15 km.p.h. during the first 30 km.; at 20 km.p.h. during the second 30 km. and at 25 km.p.h during the third 30 km.

2·17. (*a*) The numbers 3·2, 5·8, 7·9 and 4·5 have frequencies x, $(x + 2)$, $(x - 3)$ and $(x + 6)$ respectively. If their arithmetic mean is 4·876, find the value of x.

(*b*) If $M_{x.x}$ is the geometric mean of Nx's and $M_{x.y}$ is the geometric mean of Ny's , then the geometric mean M_g of the $2N$ values is given by $M_g^2 = M_{x.x}\, M_{x.y}$.

2·18. If $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{x}_2 = \frac{1}{n}\sum_{i=2}^{n+1} x_i$ and $\bar{x}_3 = \frac{1}{n}\sum_{i=3}^{n+2} x_i$, then show that

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

2·19. A distribution $x_1, x_2, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_n$ is transformed into the distribution $X_1, X_2, \ldots, X_n$ with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constants. Show that the mean, median and mode of the new distribution are given in terms of those of the first distribution by the same transformation.

Use the method indicated above to find the mean of the following distribution :

x (duration of telephone conversation in seconds)	49·5	149·5	249·5	349·5	449·5	549·5	649·5	749·5	849·5	949·5
f (respective frequency)	6	28	88	180	247	260	132	48	11	5

2·20. If $\bar{x}_w$ is the weighted mean of x_i's with weights w_i, prove that

$$\left(\sum_{i=1}^{n} w_i\right)\left(\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2\right) = \sum_{i=1}^{n}\sum_{j>1} w_i w_j (x_i - x_j)^2, \text{ where } \sum_{i=1}^{n} w_i \neq 0.$$

2·21. (*a*) Compute quartile deviation graphically for the following data :

Marks	:	20—30	30—40	40—50	50—60	60—70	70 & over
Number of students	:	5	20	14	10	8	5

(*b*) Compute a suitable measure of dispersion for the adjoining frequency distribution, giving reasons :

Classes	*Frequency*
Less than 20	30
20—30	20
30—40	15
40—50	10
50—60	5

2·22. Age distribution of hundred life insurance policy holders is as follows :

Age as on nearest birthday	*Number*	*Age as on nearest birthday*	*Number*
17—19·5	9	41—50·5	14
20—25·5	16	51—55·5	12
26—35·5	12	56—60·5	6
36—40·5	26	61—70·5	5

Calculate mean deviation from median age.

2·23. Calculate the mean and standard deviation of the following distribution :

x :	2·5—7·5	7·5—12·5	12·5—17·5	17·5—22·5
f :	12	28	65	121
x :	22·5—27·5	27·5—32·5	32·5—37·5	37·5—42·5
f :	175	198	176	120
x :	42·5—47·5	47·5—52·5	52·5—57·5	57·5—62·5
f :	66	27	9	3

2·24. Explain clearly the ideas implied in using arbitrary working origin, and scale for the calculation of the arithmetic mean and standard deviation of a frequency distribution. The values of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the analysis in the above manner are 40·604 lb. and 7·92 lb. respectively.

x :	−3	−2	−1	0	1	2	3	4	Total
f :	3	15	45	57	50	36	25	9	240

Determine the actual class intervals.

2·25. (*a*) The arithmetic mean and variance of a set of 10 figures are known to be 17 and 33 respectively. Of the 10 figures, one figure (*i.e.*, 26) was subsequently found inaccurate, and was weeded out. What is the resulting (*a*) arithmetic mean, and (*b*) standard deviation.

(*b*) The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item 8 was incorrect. Calculate the mean and standard deviation if (*i*) the wrong items is omitted, and (*ii*) it is replaced by 12.

(*c*) For a frequency distribution of marks in Statistics of 200 candidates (grouped in intervals 0—5, 5—10, ..., etc.), the mean and standard deviation were found to be 40 and 15 respectively. Later it was discovered that the score 43 was misread as 53 in obtaining the frequency distribution. Find the corrected mean and standard deviation corresponding to the corrected frequency distribution.

2·26 (*a*) Complete a table showing the frequencies with which words of different numbers of letters occur in the extract reproduced below (omitting punctuation marks) treating as the variable the number of letters in each word, and obtain the mean, median and coefficient of variation of the distribution :

"Her eyes were blue : blue as autumn distance - blue as the blue we see, between the retreating mouldings of hills and woody slopes on a sunny September morning : a misty and shady blue, that had no beginning or surface, and was looked into rather than at."

(*b*) Treating the number of letters in each word in the following passage as the variable *x*, prepare the frequency distribution table and obtain its mean, median, mode and variance.

"The reliability of data must always be examined before any attempt is made to base conclusions upon them. This is true of all data, but particularly so of numerical data, which do not carry their quality written large on them. It is a waste of time to apply the refined theoretical methods of Statistics to data which are suspect from the beginning."

2·27. The mean of 5 observations is 4·4 and variance is 8·24. If three of the five observations are 1, 2 and 6, find the other two.

2·28. Scores of two golfers for 24 rounds were as follows :

Golfer *A* : 74, 75, 78, 72, 77, 79, 78, 81, 76, 72, 72, 77, 74, 70, 78, 79, 80, 81, 74, 80, 75, 71, 73.

Golfer *B* :86, 84, 80, 88, 89, 85, 86, 82, 82, 79, 86, 80, 82, 76, 86, 89, 87, 83, 80, 88, 86, 81, 81, 87.

Find which golfer may be considered to be more consistent player ?

2·29. The sum and sum of squares corresponding to length *X* (in cms.) and weight *Y* (in gms.) of 50 tapioca tubers are given below :

$\Sigma X = 212$, $\Sigma X^2 = 902{\cdot}8$, $\Sigma Y = 261$, $\Sigma Y^2 = 1457{\cdot}6$.

Which is more varying, the length or wieght ?

2·30. Lives of two models of refrigerators turned in for new models in a recent survey are given in the adjoining table. What is the average life of each model of these refrigerators ? Which model shows more uniformity ?

Life (No. of years)	*Model A*	*Model B*
0—2	5	2
2—4	16	7
4—6	13	12
6—8	7	19
8—10	5	9
10—12	4	1

2·31. Goals scored by two teams *A* and *B* in a football season were as shown in the adjoining table.

Find out which team is more consistent ?

No. of goals scored in a match	*No. of matches*	
	Team A	*Team B*
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

2·32. An analysis of monthly wages paid to the workers in two firms, *A* and *B* belonging the same industry, gave the following results :

	Firm A	*Firm B*
Number of wage-earners	986	548
Average hourly wages	Rs. 52·5	Rs. 47·5
Variance of distribution of wages (Rs.^{2})	100	121

(*i*) Which firm, *A* or *B*, pays out larger amount as hourly wages ?

(*ii*) In which firm *A* or *B*, is there grater variability in individual wages ?

(*iii*) What are the measures of average hourly wages and the variability in individual wages, of all the workers in the two firms, *A* and *B* taken together ?

2·33. The following data give the arithmetic averages and standard deviations of three sub-groups. Calculate the arithmetic average and standard deviation (s.d.) of the whole group :

Sub-group	*No. of men*	*Average wages (Rs.)*	*s.d. of wages (Rs.)*
A	20	61·0	8·0
B	100	70·0	9·0
C	120	80·5	10·0

2·34. Find the missing information from the following data :

	Group I	*Group II*	*Group III*	*Combined*
Number	50	?	90	200
Standard Deviation	6	7	?	7·746
Mean	113	?	115	116

2·35. A collar manufacturer is considering the production of a new style collar to attract young men. The following statistics of neck circumference are available based on the measurement of a typical group of students :

Mid-value (in inches) :	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
No. of students :	4	19	30	63	66	29	18	1

Compute the mean and standard deviation and use the criterion $\bar{x} \pm 3\sigma$, to obtain the largest and smallest size of collar he should make in order to meet needs of practically all his customers, bearing in mind that the collars are worn on an average 3/4 inch larger than neck size.

2·36. A frequency distribution is divided into two parts. The mean and standard deviation of the first part are m_1 and s_1 and those of the second part are m_2 and s_2 respectively. Obtain the mean and standard deviation for the combined distribution.

2·37. (*a*) The means of two samples of size 50 and 100 respectively are 54·1 and 50·3 and the standard deviations are 8 and 7. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples.

(*b*) A distribution consists of three components with frequencies 200, 250 and 300 having means 25, 10 and 15 and standard deviations 3, 4 and 5 respectively. Show that the mean of the combined group is 16 and its standard deviation is 7·2 approximately.

2·38. In a certain test for which the pass marks is 30, the distribution of marks of passing candidates classified by sex (boys and girls)were as given below :

Marks	*Frequency*	
	Boys	*Girls*
30—34	5	15
35—39	10	20
40—44	15	30
45—49	30	20
50—54	5	5
55—59	5	—
Total	70	90

The overall mean and standard deviation of marks for boys including the 30 failed were 38 and 10 respectively. The corresponding figures for girls including the 10 failed were 35 and 9.

(*i*) Find the mean and standard deviation of marks obtained by the 30 boys who failed in the test.

(*ii*) The moderation committee argued that percentage of passes among girls is higher because the girls are very studious and if the intention is to pass those who are really intelligent, a higher pass marks should be used for girls. Without questioning the propriety of this argument, suggest what the pass marks should be which would allow only 70% of the girls to pass.

(*iii*) The prize committee decided to award prizes to the best 40 candidates (irrespective of sex) judged on the basis of marks obtained in the test. Estimate the number of girls who would receive prizes.

2·39. Find the mean and variance of first n-natural numbers.

2·40. If the mean and standard deviation of a variable x are m and σ respectively, obtain the mean and standard deviation of $(ax + b)/c$, where a, b and c are constants.

2·41. In a series of measurements we obtain m_1 values of magnitude x_1, m_2 values of magnitude x_2, and so on. If $\bar{x}$ is the mean value of all the measurements, prove that the standard deviation is : $\sqrt{\frac{\sum m_r (k - x_r)^2}{\sum m_r} - \delta^2}$, where $\bar{x} = k + \delta$ and k is any constant.

2·42. (*a*) Show that in a discrete series if deviations are small compared with mean M so that $(x/M)^2$ and higher powers of (x/M) are neglected, prove that

(*i*) $MH = G^2$ (*ii*) $M - 2G + H = 0$,

where G is geometric mean and H is harmonic mean.

(*b*) The mean and standard deviation of a variable x are m and σ respectively. If the deviations are small compared with the value of the mean, show that

(*i*) Mean $(\sqrt{x}) = \sqrt{m}\left(1 - \frac{\sigma^2}{8m^2}\right)$ (*ii*) Mean $\left(\frac{1}{\sqrt{x}}\right) = \frac{1}{\sqrt{m}}\left(1 + \frac{3\sigma^2}{8m^2}\right)$ approximately.

(*c*) If the deviation $X_i = x_i - M$ is very small in comparison with mean M and (X_i/M) are neglected, prove that $V = \sqrt{\frac{2(M - G)}{M}}$.

where G is the geometric mean of the values $x_1, x_2, \ldots, x_n$ and V is the coefficient of dispersion $(\sigma/\bar{x})$.

2·43. Show that, if the variable takes the value 0, 1, 2, ..., n with frequencies proportional to the binomial coefficients ${}^nC_0, {}^nC_1, {}^nC_2, \ldots, {}^nC_n$ respectively then the mean of the distribution is $(n/2)$, the mean square deviation about $x = 0$ is $n(n+1)/4$ and the variance is $n/4$.

2·44. (*a*) Let r be the range and s be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$; then prove by general reasoning or otherwise that $s \leq r$.

(*b*) Let r be the range and $S = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}$, be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$, then prove that $S \leq r\left(\frac{n}{n-1}\right)^{\frac{1}{2}}$.

2·45. (*a*) The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40. Show that the mean is 3, the variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

2·46. The first four moments of distribution about the value 5 of the variable are 2, 20, 40 and 50. Obtain, as far as possible, the various characteristics of the distribution on the basis of the information given.

2·47. (*a*) If the first four moments of a distribution about the value 5 are equal to – 4, 22, – 117 and 560, determine the corresponding moments (*i*) about the mean, and (*ii*) about zero.

(*b*) The first four moments of a distribution about $x = 4$ are 1, 4, 10 and 45. Show that the mean is 5 and the variance is 3 and μ_3 and μ_4 are 0 and 26 respectively.

2·48. For the following data, calculate (*i*) Mean, (*ii*) Median, (*iii*) Semi-inter quartile range, (*iv*) Coefficient of variation, and (*v*) β_1 and β_2 coefficients.

Wages in Rupees :	170—180	180—190	190—200	200—210	210—220	220—230	230—240	240—250
Number of Persons :	52	68	85	92	100	95	70	28

2·49. Find the second, third and fourth central moments of the frequency distribution given below. Hence find and measure of skewness (γ_1) and measure of kurtosis (γ_2).

Class Limits	*Frequency*
110·0—114·9	5
115·0—119·9	15
120·0—124·9	20
125·0—129·9	35
130·0—134·9	10
135·0—139·9	10
140·0—144·9	5

Also apply Sheppard's corrections for moments.

2·50. The standard deviation of a symmetrical distribution is 5. What must be the value of the fourth moment about the mean in order that the distribution be (*i*) leptokurtic, (*ii*) mesokurtic, and (*iii*) platykurtic.

2·51. (*a*) Obtain Karl Pearson's measure of skewness for the following data :

Values	*Frequency*	*Values*	*Frequency*
5—10	6	25— 30	15
10—15	8	30—35	11
15—20	17	35—40	2
20—25	21		

(*b*) Assume that a firm has selected a random sample of 100 from its production line and has obtain the data shown in the table below :

Class interval	*Frequency*	*Class interval*	*Frequency*
130 —134	3	150—154	19
135—139	12	155—159	12
140—144	21	160—164	5
145—149	28		

Compute the following :

(*i*) The arithmetic mean, (*ii*) The standard deviation and

(*iii*) Karl Pearson's coefficient of skewness.

2·52. For the frequency distribution given below, calculate the coefficient of skewness based on quartiles :

Monthly Sales (Rs. lakh)	*No. of Firms*	*Monthly Sales* (Rs. lakh)	*No. of Firms*
Less than 20	30	Less than 70	644
Less than 30	225	Less than 80	650

Less than 40	465	Less than 90	665
Less than 50	580	Less than 100	680
Less than 60	634		

2·53. (*a*) (*i*) Karl Pearson's coefficient of skewness of a distribution is 0·32, its s.d. is 6·5 and mean is 29·6. Find the mode of the distribution.

(*ii*) If the mode of the above distribution 24·8, what will be the s.d. ?

(*b*) In a frequency distribution, the coefficient of skewness based upon the quartiles is 0·6. If the sum of the upper and lower quartiles is 100 and median is 38, find the value of the upper and lower quartiles.

2·54. (*a*) A frequency distribution gave the following results :

(*i*) C.V. = 5, (*ii*) Karl Pearson's coefficient of skewness = 0·5, (*iii*) $\sigma = 2$.

Find the mean and mode of the distribution.

(*b*) Find the C.V. of a frequency distribution given that its mean is 120, mode is 123 and Karl Pearson's coefficient of skewness is – 0·3.

(*c*) The first three moments of distribution about the value 2 are 1, 16 and – 40 respectively. Examine the skewness of the distribution.

2·55. (*a*) The scores in Economics of 250 candidates appearing at an examination have :

Mean $= \bar{x} = 39{\cdot}72$; Variance $= \sigma^2 = 97{\cdot}80$;

Third Central moment $= \mu_3 = -114{\cdot}18$; Fourth central moment $= \mu_4 = 28{,}396{\cdot}14$

It was later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moments.

(*b*) In calculating the moments of a frequency distribution based on 100 observations, the following results are obtained :

Mean = 9, Variance = 19, $\beta_1 = 0{\cdot}7$ (μ_3 + ive), $\beta_2 = 4$

But later on its was found that one observation 12 was read as 21. Obtain the correct values of the first four central moments, and β_1 and β_2.

2·56. The first three moments about the origin are given by :

$$\mu_1' = \frac{n+1}{2}, \quad \mu_2' = \frac{(n+1)(2n+1)}{6} \quad \text{and} \quad \mu_3' = \frac{n(n+1)^2}{4}.$$

Examine the skewness of the data.

2·57. Show that, if the class interval of a grouped distribution is less than one third of the calculated standard deviation, Sheppard's adjustment makes a difference of less than $\frac{1}{2}$ % in the estimate of the standard deviation.

2·58. Show that if a range of six times the standard deviation covers at least 18 class intervals, Sheppard's correction will make a difference of less than 0·5 per cent in the correct value of the standard deviation.

2·59. If ∂_r is the *rth* absolute moment about zero, use the mean value of

$$\{u + x|^{(r-1)/2} + v \mid x|^{(r+1)/2}\}^2 \quad \text{to show that} \quad \partial_r^{2r} \leq \partial_{r-1}^r \partial_{r+1}^r.$$

From this derive the following inequalities : (*i*) $\partial_r^{r+1} \leq \partial_{r+1}{}^r$, (*ii*) $\partial_r^{1/r} \leq \partial_{r+1}{}^{1/(r+1)}$.

2·60. For a random variable X moments of all orders exist. Denoting by μ_j and δ_j, the central moment and *j*th absolute moment respectively, show that

(*i*) $(\mu_{2j+1})^2 \leq \mu_{2j}\mu_{2j+2}$, (*ii*) $(\partial_j)^{1/i} \leq (\partial_{j+1})^{1/(j+1)}$.