3. STATISTICS OF LOCATION

Statistics of location are descriptive statistics serving to locate specific positions of the frequency distribution of a sample in the scale of scores of a variable. Some of them like mean, median and mode are known as the *measures for central values* or *central tendencies* because they indicate the location of some central points of the frequency distribution in the scale of the variable. Others like percentiles and quartiles serve to locate positions other than the central ones.

3.1 MEAN

Mean is the arithmetic average of the observed scores. The sample mean and the parametric mean are represented by the symbols $\bar{X}$ and μ respectively. Where X represents each individual score of a sample, ΣX is the sum of all the scoses, and N is the sample size or the total frequency of cases in the sample,

$$\bar{X} = \frac{\Sigma X}{N}.$$

As it follows that $\Sigma X = n\bar{X}$, mean may be defined as the score which each individual would have possessed if the total score of the sample (ΣX) were equally distributed among all the individuals. For a symmetrical distribution of scores, mean is the most reliable, stable and widely applicable central value. Its important properties are given below.

(a) The algebraic sum of the deviations of all the individual scores from the mean amounts to zero. Thus, where $(X - \bar{X})$ represents the deviation of each individual score from the sample mean,

$$\Sigma(X - \bar{X}) = 0.$$

This is because the sum of the positive deviations of some scores from the mean equals that of the negative deviations of the remaining scores from the latter.

(b) The *sum of squares* about the mean, i.e., $\Sigma(X - \bar{X})^2$, which is the sum of the squared deviations of scores from the mean, is the lowest of the sums of squares about the measures of central values.

(c) If the individual scores of a sample are all multiplied or divided by a constant number a, the mean also gets respectively multiplied and divided by the same number.

$$\frac{\Sigma Xa}{N} = \bar{X}a; \quad \frac{\Sigma X}{aN} = \frac{\bar{X}}{a}.$$

(d) If a constant number a is added to or subtracted from each individual score of a sample, the mean also gets respectively increased and decreased by the same number.

$$\frac{\Sigma(X+a)}{N} = \bar{X} + a; \quad \frac{\Sigma(X-a)}{N} = \bar{X} - a.$$

(e) If a score, having an extreme positive or negative deviation from the mean, is included in the sample, the mean is displaced towards that score unless it is counterbalanced by the simultaneous inclusion of another score with an equal but opposite deviation. This implies that the mean is unreliable as a measure of central value in an asymmetric distribution which has one of its tails longer than the other due to a few scores with large deviations in the longer tail.

(f) If the scores of a variable Y are the linear functions of the scores of another variable X, then the mean $\bar{Y}$ of the former is also a linear function of the mean $\bar{X}$ of the latter. Thus, if a is the vertical intercept and b the slope of the straight line formed by

plotting the Y scores against the X scores of the respective individuals in a sample,

$$\frac{Y = a + bX\ ;}{\therefore\ \overline{Y} = a + b\overline{X}.}$$

When there are k sets of scores, viz., $X_1, X_2, \ldots\ldots X_k$, of the same variable in as many samples, the grand mean $\overline{X}$ of all the samples is computed from the sample means, $\overline{X}_1, \overline{X}_2, \ldots \overline{X}_k$, and the respective sample sizes, viz., $n_1, n_2, \ldots n_k$. The sample sizes are the *weights* given to the respective sample means.

$$\Sigma X_1 = n_1\overline{X}_1\ ;\ \Sigma X_2 = n_2\overline{X}_2\ ;\ \Sigma X_k = n_k\overline{X}_k\ ;$$

$$\therefore\ \overline{X} = \frac{\Sigma X_1 + \Sigma X_2 + \cdots + \Sigma X_k}{n_1 + n_2 + \cdots + n_k}$$

$$= \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \cdots + n_k\overline{X}_k}{n_1 + n_2 + \cdots + n_k}$$

or, $$\overline{X} = \frac{\Sigma[(\text{sample size})(\text{sample mean})]}{\Sigma(\text{sample size})}$$

1. Computation from ungrouped data.

For the data not arranged into a frequency distribution, mean is computed by dividing the sum of the individual raw scores (X) by the total number (N) of individuals in the sample.

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

Example 3.1.1.

Compute the mean of the following body weights (kg) of a sample of men :

55, 60, 62, 58, 57, 61, 59, 60, 61, 62.

Solution :

$$\overline{X} = \frac{\Sigma X}{N} = \frac{55 + 60 + 62 + 58 + 57 + 61 + 59 + 60 + 61 + 62}{10}$$

$$= 59.5 \text{ kg.}$$

2. Computation from frequency tables

Frequency distributions of the discrete variables are often arranged in simple frequency tables in which the frequencies are entered against single distinct scores, each forming a class by itself (Table 2.3). In such cases, one or more individuals possess identical scores so that the scores are repeated in the data ; but the data cannot be classified continuously. The mean is computed from the frequencies (repetitions) of individual scores. Where $f_1, f_2, \ldots\ldots f_k$ are frequencies of the individual scores like $X_2, \ldots\ldots X_k$, and N is the total frequency sample size,

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + \cdots\cdots + f_kX_k}{N} = \frac{\Sigma fX_i}{N}.$$

Example 3.1.2.

Compute the mean of the data presented in Table 2.3.

Solution :

The relevant variable viz., number of children per family, is a discreet variable. From the data, arranged in a simple frequency table (Table 3.1),

$$\overline{X} = \frac{\Sigma fX_i}{N} = \frac{494}{197} = 2.5.$$

Table 3.1. Frequency table of number of children for computation of mean.

Number of children (X_i)	Number of families (f)	fX_i
0	7	0
1	35	35
2	67	134
3	43	129
4	32	128
5	10	50
6	3	18
Total	197 (N)	494 (ΣfX_i)

3. **Direct computation from grouped data**

This method is used for directly computing the mean of a continuous measurement variable whose scores have been arranged into regular frequency distributions.

The entire range of the observed scores should be divided into class intervals of *either equal or unequal lengths* and the scores should be arranged into a frequency distribution in terms of those intervals. Because the score of each observation in a class interval is assumed to be identical with the midpoint (X_c) of that interval, the sum of the scores of each interval is obtained by multiplying the frequency (f) of that interval by its midpoint (X_c) and may be represented by fX_c. Thus, the sum of the fX_c values of all the intervals gives the net sum of all the scores of the sample. Hence,

$$\bar{X} = \frac{\text{sum of all scores}}{\text{sample size}} = \frac{\Sigma fX_c}{N}$$

Minor discrepancies may arise in the mean computed in this way if the scores are grouped in a different set of class intervals. Moreover, the mean computed from midpoints of intervals may differ slightly from that computed directly from individual scores of ungrouped data.

The mean cannot be computed in this way if the data have been arranged in an *incomplete distribution* with some *open class intervals* (page 14) because the midpoint is not available for such an interval.

Example 3.1.3.

Compute the mean body weight from the following frequency distribution of body weights (kg) in a sample of humans.

Class intervals :	51-53	54-56	57-59	60-62	63-65	66-68	69-71
Frequencies :	4	7	12	25	13	6	3

Solution :

The data are arranged in Table 3.2.

(*i*) The midpoint (X_c) of each class interval is then computed as follows and entered against that interval in the table.

X_o = (lower score limit) + $\frac{1}{2}$[(upper score limit) − (lower score limit)]

For example, for the interval 54—56,

$X_o = 54 + \frac{1}{2}(56 - 54) = 55.$

Table 3.2. Frequency distribution for computing the mean body weight by the direct method.

Class intervals	X_o	f	x'	fx'	fX_o
51—53	52	4	−3	−12	208
54—56	55	7	−2	−14	385
57—59	58	12	−1	−12	696
60—62	61	25	0	−38	1525
63—65	64	13	+1	+13	832
66—68	67	6	+2	+12	402
69—71	70	3	+3	+9	210
				+34	
Total		70 (N)		$\Sigma fx' = -38 + 34 = -4$	4258

(*ii*) Each X_o is multiplied by the frequency (f) of cases in that interval to compute fX_o as the total of the scores in the latter. For example, the sum of the scores of all cases in the interval 57—59 is given by : $fX_o = 12 \times 58 = 696$.

(*iii*) The mean $\bar{X}$ is finally computed from the sum of the fX_o values of all the intervals and the total frequency (N) of the cases in the sample.

$$\bar{X} = \frac{\Sigma fX_o}{N} = \frac{4258}{70} = 60.8 \text{ kg.}$$

$M = AM + Ci$

$= 61 + \frac{-4}{70} \times 3$

$= 61 + (-\quad) \times 3$

$= 61 (-0.05) \times 3 = 60.85$

$= 61 - 0.15$

Example 3.1.4.

Compute the mean for the winglengths (mm) of houseflies given below :

3.9, 4.3, 4.8, 4.7, 4.6, 4.4, 3.7, 4.2, 4.1, 4.8, 5.3, 4.9, 4.6, 3.8, 4.0, 5.3, 5.7, 5.5, 3.9, 4.5, 3.8, 4.5, 5.0, 4.9, 4.8, 3.5, 4.3, 5.1, 3.9, 4.7, 5.6, 4.6, 4.4, 3.4, 5.1, 4.6, 3.9, 3.8, 4.8, 4.9.

Solution :

(*i*) The data are first arranged into a frequency distribution and entered in Table 3.3.

Highest score = 5.7. Lowest score = 3.4. Range = 3.4 − 5.7.

Sample size (N) = 40. Number of intervals chosen = 5.

Length of class intervals (i) = (5.7 − 3.4)/5 ≃ 0.5.

(*ii*) The midpoint X_o is computed for each class interval. For example, for the interval 3.9 − 4.3,

$X_o = 3.9 + \frac{1}{2}(4.3 - 3.9) = 4.1.$

(*iii*) Each X_o is multiplied by the frequency (f) of that interval to give fX_o which serves as a measure of the sum of all the scores in that interval. The sum of scores for the interval 3.9 − 4.3, for example, is given by :

$fX_o = 9 \times 4.1 = 36.9.$

(*iv*) The mean ($\bar{X}$) is finally computed from the sum of the fX_c values of all the intervals and the total frequeny (N) of the sample.

$$\bar{X} = \frac{\Sigma fX_c}{N} = \frac{180.5}{40} = 4.51 \text{ mm.}$$

Table 3.3. Frequency distribution for computing the mean winglength by direct method.

Class intervals	X_c	f	fX_c
3.4—3.8	3.6	6	21.6
3.9—4.3	4.1	9	36.9
4.4—4.8	4.6	14	64.4
4.9—5.3	5.1	8	40.8
5.4—5.8	5.6	3	16.8
Total		40	180.5

4. Computation by code method

This short method is applicable only to continuous frequency distributions having class intervals of *equal lengths*.

(*i*) The midpoint of an interval near the centre of the distribution is arbitrarily chosen as the *assumed mean* (A) and assigned a *code number* of 0 to show that this midpoint does not deviate from A (Table 3.4).

(*ii*) The midpoints of intervals, rising progressively higher than that of A, are assigned code numbers like +1, +2, etc., in an ascending order. Similarly, those of intervals, running progressively downwards from the interval of A, are given code numbers like −1, −2, etc., in a descending order. These code numbers (x') indicate the magnitudes of positive or negative deviations of the respective midpoints from A in terms of the interval units or code units which equal the length i of the intervals.

(*iii*) The code number (x') of each interval is multiplied by the frequency (f) of that interval and the algebraic sum ($\Sigma fx'$) of all these products is divided by the sample size (N) to get the *correction term* (c) in code units.

$$c = \frac{\Sigma fx'}{N}.$$

(*iv*) Next, c is reconverted to original units of the scores by multiplying it with the length i of the class interval.

(*v*) The algebraic sum of the assumed mean (A) and the correction term in original units (ci) gives the actual mean ($\bar{X}$).

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

Example 3.1.5.

Compute the mean by the code method for the following frequency distribution of memory test scores in a sample :

Class intervals :	18—20	21—23	24—26	27—29	30—32	33—35	36—38
Frequencies :	5	9	13	23	15	10	5
Midpoints :	19	22	25	28	31	34	37

Solution :

The frequency distribution is tabulated in Table 3.4.

(*i*) The midpoint (X_c) of the interval 27—29 is taken as the assumed mean (A): $A = 28$. This midpoint is then assigned the code number (x') of 0. The midpoints of intervals higher than 27—29 are then assigned positive code numbers ($+x'$) in an ascending order while those lower than 27—29 are given negative code number ($-x'$) in a descending order.

Table 3.4. Frequency distribution for computing the mean memory test score by the code method.

Class intervals	X_c	fx	f	cf	x'	fx'
36—38	37	185	5	80	+3	+15
33—35	34	340	10	75	+2	+20
30—32	31	465	15	65	+1	+15
27—29	28	644	23	50	0	0 +50
24—26	25	325	13	27	−1	−13
21—23	22	198	9	14	−2	−18
18—20	19	95	5	5	−3	−15
Total		$\Sigma fx = 2252$	80 (N)		$\Sigma fx' = +50$	−46
						+ 4 ($\Sigma fx'$)

Height ↑ length

$= +4$

(*ii*) The code number (x') of each interval is multiplied by the frequency (f) of that interval to give the fx' value of the latter. For the interval 24—26, for example, $fx' = 13 \times (-1) = -13$.

(*iii*) The correction term (c) is then computed from the the sum of the fx' values of all the intervals and the sample size (N).

$$c = \frac{\Sigma fx'}{N} = \frac{+4}{80} = 0.05.$$

$M = \frac{\Sigma fx}{N} = \frac{2252}{80} = 28.15$

(*iv*) The length i of the class intervals is used in computing the mean $\overline{X}$.

$$i = 3;$$
$$\therefore \quad \overline{X} = A + ci = 28 + 0.05 \times 3$$
$$= 28.15 \simeq 28.2.$$

5. Computation by the weighted mean method

The sample means ($\overline{X}_1$, $\overline{X}_2$, etc.) of a number of samples may be used to compute the grand mean ($\overline{X}$) of the full set, using the sample sizes (n_1, n_2, etc.) as the *weights* for the respective sample means.

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \cdots\cdots + n_k\overline{X}_k}{n_1 + n_2 + \cdots\cdots + n_k}.$$

Example 3.1.6.

The mean systolic blood pressure was found to be 129.4 and 133.6 mm Hg for two groups of 12 and 15 men, respectively. Find the mean systolic blood pressure of all the 27 men.

Solution :

$$\bar{X}_1 = 129.4 ; \quad n_1 = 12.$$
$$\bar{X}_2 = 133.6 ; \quad n_2 = 15.$$
$$\therefore \quad \bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$
$$= \frac{12 \times 129.4 + 15 \times 133.6}{12 + 15}$$
$$= 131.7 \text{ mm Hg.}$$

Example 3.1.7.

20% of a group of 80 men and 15% of a group of 120 women were found to be diabetic. Find the mean percentage of diabetics for both the groups combined.

Solution :

For men : $n_1 = 80$; percentage $(P_1) = 20$.
For women : $n_2 = 120$; $P_2 = 15$.

$$\therefore \quad \bar{X}_p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$
$$= \frac{80 \times 20 + 120 \times 15}{80 + 120} = 17\%$$

3.2 GEOMETRIC MEAN

Geometric mean (GM) is the nth root of the product of all the scores (Xs) of a sample when N is the total number of scores and none of the scores is negative or 0. It is thus the antilog of the mean of the logarithms of scores, provided all of the latter are higher than 0 in value. Where Π indicates the product of the terms following it,

$$GM = \sqrt[N]{X_1 . X_2 . X_3 . \ldots X_N} = (\Pi X)^{\frac{1}{N}}$$
$$= \text{Antilog} \left[\frac{\Sigma \log X}{N}\right].$$

Where all scores have positive values, $\bar{X}$ ordinarily either exceeds or equals GM. The latter is computed for frequency distributions with scores mostly concentrated in the low-value tail, for measurements in a logarithmic scale, as also for distributions of growth rates, rise in bacterial counts, reflex reaction times, and sensory responses such as differential perception of sound frequencies of matched intensities and optical potential changes at different intensities of illumination.

Example 3.2.1.

Compute the geometric mean of the following pH values of pancreatic juice collected from a sample of humans : 8.0, 8.3, 7.1, 8.2, 7.6, 7.7, 7.9, 8.0, 7.4, 8.2, 8.4, 8.1, 7.8, 7.9, 7.6.

Solution :

The logarithms of observed pH scores (X) are recorded against the respective scores and totalled in Table 3.5.

$$GM = \text{Antilog}\left[\frac{\Sigma \log X}{N}\right] = \text{Antilog}\left[\frac{13.441606}{15}\right]$$
$$= 7.87.$$

Table 3.5. Treatment of pH data for GM.

Serial No.	pH scores (X)	log X	Serial No.	pH scores (X)	log X
1	8.0	0.9030899	Total brought forward		7.1552611
2	8.3	0.919078	9	7.4	0.8692317
3	7.1	0.8512583	10	8.2	0.9138138
4	8.2	0.9138138	11	8.4	0.9242792
5	7.6	0.8808135	12	8.1	0.908485
6	7.7	0.8864907	13	7.8	0.8920946
7	7.9	0.897627	14	7.9	0.897627
8	8.0	0.9030899	15	7.6	0.8808135
Total carried forward		7.1552611	Total ($\Sigma \log X$)		13.441606

3.3 HARMONIC MEAN

Harmonic mean (H) is the reciprocal of the average of reciprocals of the observed scores (X) of a sample when each X score is higher than 0 in value. Where N is the sample size,

$$H = \frac{N}{\Sigma \frac{1}{X}}; \text{ or, } \frac{1}{H} = \frac{1}{N}\Sigma\frac{1}{X}.$$

H is ordinarily lower than $\bar{X}$ and either lower than or equal to GM. It is useful for measurements on a reciprocal scale. It is applicable to psychological tests involving measurements of time intervals for specific performances.

Example 3.3.1.

Compute the harmonic mean of the following O_2 percentages in human alveolar air: 13.9, 12.7, 12.9, 13.3, 13.6, 13.1, 12.5, 12.7, 13.3, 13.0.

Solution :

The reciprocal of O_2 percentages (X) are worked out, recorded against the respective X scores and totalled in Table 3.6.

$$H = \frac{N}{\Sigma \frac{1}{X}} = \frac{10}{0.765} = 13.07.$$

Table 3.6. Treatment of O_2 percentage data for harmonic mean.

Serial No.	X	$1/X$	Serial No.	X	$1/X$
1	13.9	0.072	Total brought forward		0.454
2	12.7	0.079	7	12.5	0.080
3	12.9	0.078	8	12.7	0.079
4	13.3	0.075	9	13.3	0.075
5	13.6	0.074	10	13.0	0.077
6	13.1	0.076			
otal carried forward		0.454	Total $\left(\Sigma \frac{1}{X}\right)$		0.765

4 PERCENTILES AND QUARTILES

Percentiles (P_p) are those points in a equency distribution, below which lie ecific percentages of the total number of ores of a sample. Thus, the 50th percentile $_{50}$) is that score below which should lie the west 50% of the scores; P_{50} is identical ith Mdn. Similarly, P_{25} and P_{75} are the cores below which lie respectively 25% and 5% of the total scores.

Quartiles (Q) are those points in a requency distribution, below which lie the pecific numbers of quarters of the total requency. Thus, Q_1, Q_2, Q_3 and Q_4 are the cores below which lie respectively one-fourth, alf, three-fourths and all of the total number f scores. $Q_2 = P_{50} = Mdn$; $Q_1 = P_{25}$; $Q_3 = P_{75}$; $Q_4 = P_{100} = N$.

1. Computation from cumulative frequencies

Percentiles and quartiles are computed from cumulative frequencies (cf).

$$P_p = X_l + i \times \frac{pN - cf_l}{f_o}$$

where P_p is the required percentile, cf_l is the cumulative frequency of all the intervals below the true lower limit X_l of the class interval containing P_p, f_o is the observed frequency in that class interval, i is the length of class intervals, and pN is the number of cases to be counted off in reaching P_p from the lowest score—pN *is* the product of the sample size N and the proportion of total cases below P_p.

Example 3.4.1.

Compute the 25th and 75th percentiles of the frequency distribution of body weights (kg) given in Table 2.8.

Solution :

The cumulative frequency distribution of Table 2.8 is reproduced in Table 3.7.

Table 3.7. Cumulative frequencies of body weights.

Class intervals	True limits		f	cf
	lower (X_l)	upper (X_u)		
51—53	50.5	53.5	5	5
54—56	53.5	56.5	7	12
57—59	56.5	59.5	14	26
60—62	59.5	62.5	28	54
63—65	62.5	65.5	15	69
66—68	65.5	68.5	8	77
69—71	68.5	71.5	3	80 (N)

(*i*) The length i of the intervals is obtained by subtracting X_u of any interval from that of the next higher one. Using the X_u scores of 65.5 and 62.5,

$$i = 65.5 - 62.5 = 3.$$

(*ii*) For computing P_{25} :

$$p = 0.25\ ; \quad N = 80\ ; \quad \therefore \quad pN = 0.25 \times 80 = 20.$$

So, 20 scores are counted off from the lowest class interval, thus reaching into the interval 57—59 in which P_{25} lies. The true lower limit X_l of this interval amounts of 56.5, the cumulative frequency cf_l upto 56.5 amounts to 12, and the frequency f_o in that interval amounts to 26.

$$X_l = 56.5\ ; \quad cf_l = 12\ ; \quad f_o = 14\ ; \quad pN = 20; \quad i = 3\ ;$$

$$P_{25} \text{ or } Q_1 = X_l + i \times \frac{pN - cf_l}{f_o} = 56.5 + 3 \times \frac{20 - 12}{14} = 58.2 \text{ kg.}$$

(*iii*) For computing P_{75} :

$$p = 0.75\ ; \quad pN = 0.75 \times 80 = 60.$$

In counting off 60 scores from the lowest interval, the interval 63—65 is reached in which lies P_{75}.

$$X_l = 62.5\ ; \quad cf_l = 54\ ; \quad f_o = 15\ ; \quad pN = 60\ ; \quad i = 3\ ;$$

$$P_{75} \text{ or } Q_3 = X_l + i \times \frac{pN - cf_l}{f_o} = 62.5 + 3 \times \frac{60 - 54}{15} = 63.7 \text{ kg.}$$

2. Graphical determination from ogive

Percentiles and quartiles may also be obtained graphically from ogives (Fig. 2.10). A line is drawn paralled to the X axis from that point on the Y axis which corresponds to the cP for the required percentile—the cP amounts to 25, 50 and 75 for respectively P_{25}, P_{50} and P_{75}. From the point of intersection of this line with the ogive, an ordinate is dropped to the X axis. The point of intersection of this ordinate with the X axis gives the required percentile or quartile.

3.5 PERCENTILE RANKS

A percentile rank (PR) is the rank or graded position of a given score on a scale of 100 among all the scores of a sample. It is

estimated from the percentages of scores lying below it.

1. PR from cumulative frequencies

Where the given score X belongs to a class interval having the length i, the true lower limit X_l and the frequency f_o, and the cumulative frequency upto X_l amounts to cf_l,

$$PR = 100 \times \frac{cf_l + \frac{(X - X_l)f_o}{i}}{N}.$$

2. PR from ranked scores

When the given score has been assigned a numerical rank R in a descending order of magnitude in a sample of size N,

$$PR = 100 - \frac{100R - 50}{N};$$

but where the ranking is in the ascending order,

$$PR = \frac{100R - 50}{N}.$$

3. PR from ogive

An ordinate is raised on the X axis of an ogive at the given score X. A horizontal line is drawn from the point of intersection of the ordinate with the ogive. The point of intersection of this line with the Y axis gives the PR of the given score.

Example 3.5.1.

Find the PR of the score 64 of the data presented in Table 3.7 of Example 3.4.1.

Solution :

The score 64 belongs to the interval 63—65 with the true lower limit (X_l) of 62.5, a length (i) of 3, and a frequency (f_o) of 15 ; the cumulative frequency (cf_l) upto the X_l of 62.5 amounts to 54.

$$X = 64 ; \; X_l = 62.5 ; \; i = 3 ; \; f_o = 15 ; \; cf_l = 54 ; \; N = 80 ;$$

$$\therefore \quad PR = 100 \times \frac{cf_l + \frac{(X - X_l)f_o}{i}}{N}$$

$$= 100 \times \frac{54 + \frac{(64 - 62.5)15}{3}}{80} = 76.9.$$

Example 3.5.2.

Find the percentile ranks of students occupying the 3rd and 20th ranks in the descending order of merit in a Biology examination involving 80 students.

Solution :

(*i*) For the student with rank 3,

$$PR = 100 - \frac{100R - 50}{N} = 100 - \frac{100 \times 3 - 50}{80} = 96.9.$$

(*ii*) For the student with rank 20,

$$PR = 100 - \frac{100 \times 20 - 50}{80} = 75.6.$$

3.6 MEDIAN

Median (*Mdn*) is that point in the frequency distribution, above and below which lie equal numbers or 50% of scores or cases of the sample. *Mdn* is identical with P_{50} and Q_2. Some of the properties of *Mdn* are given below.

(a) The ordinate, drawn on the X axis at the Mdn, bisects the area of a frequency distribution into two equal halves.

(b) In symmetrical unimodal distributions, *Mdn* coincides with the mean and the mode, and the algebraic sum of deviations of the observed scores from the median amounts to 0.

(c) In asymmetric distributions, the median differs from the mean and the mode, the mean being located further than the median towards the longer tail of the distribution. In such distributions, the algebraic sum of deviations of the scores from the median differs from 0 in value and its positive or negative sign indicates a longer positive or negative tail, respectively, of the distribution. The degree of asymmetry (*skewness*) of a distribution is estimated from the difference between its median and either its mean or its mode—extreme observations predominate on that side of the median where the mean is situated, or on that side of the mode where the median is located.

(d) Because the median is less deflected than the mean by extreme deviations of a few observations, the former is more reliable and representative as a central value than the latter in an asymmetric distribution; e.g., distributions of blood pressures or body weights of aged humans, minimum effective doses of drugs, and bacterial counts in drinking water.

(e) Median can be computed even for a frequency distribution with *open class intervals* (page 14) which are unsuitable for computing the mean due to the lack of knowledge about their midpoints.

Median, however, finds very limited use in further statistical work.

1. Computation from ungrouped data

In an ungrouped set of data, the median is the $(N+1)/2$th score, counted from either the lowest or the highest score of the sample. (*i*) If there is an odd number of scores in the sample, i.e., N is an odd number, *Mdn* coincides with that observed score which belongs to the $(N+1)/2$th individual. (*ii*) If there is an even number of scores in the sample, the $(N+1)/2$th score falls midway between two observed scores and is given by the average of those two scores (Example 3.6.1). (*iii*) If the *Mdn* or $(N+1)/2$th score falls within a set of identical scores in the data, the value of the *Mdn* is found by extrapolation. For this, all the observed identical scores of that set are assumed to occupy one unit interval extending from 0.5 below the score to 0.5 above the latter. Each score of the set is assumed to cover that fraction of this unit interval as is given by the reciprocal of the number of scores in that set. *Mdn* is computed by adding to the lower limit of this unit interval as many of these fractions of the interval as the number of identical scores of the set covered on counting off $0.50N$ scores starting from the lowest score of the data (Example 3.6.2).

Example 3.6.1.

Find the median for the following strengths (in degrees of arc) of reflex knee jerk movements observed in a sample of athletes : 19, 21, 22, 26, 28, 30, 31, 35, 35, 37.

Solution :

$$N = 10. \quad Mdn = \frac{N+1}{2}\text{th score} = \frac{10+1}{2} \text{ or } 5.5\text{th score.}$$

Thus, the *Mdn* lies between the 5th and 6th scores counted from either the lowest or the highest score. Counting from the lowest score of 19, the 5th and 6th scores amount to 28 and 30 respectively.

$$\therefore \quad Mdn = 5.5\text{th score} = \frac{28+30}{2} = 29^{\circ}.$$

Example 3.6.2.

Find the median for the following body weight (kg) data obtained from a sample of men : 55, 57, 58, 59, 61, 61, 61, 63, 67, 68, 70.

Solution :

$$N = 11. \quad \therefore \quad Mdn = \frac{N+1}{2}\text{th score} = \frac{11+1}{2} \text{ or 6th score.}$$

Thus, five scores lie below the *Mdn* and the remaining five above it. But in counting off five scores from the lowest one, the first of three identical scores, viz., the first 61, gets included in those five scores. The three identical scores forming the set of 61, may be assumed to occupy one unit interval extending from 60.5 to 61.5, each score occupying $\frac{1}{3}$ or 0.33 of this interval. On counting off only one of these three identical scores for arriving at the *Mdn*, the upper limit of that score is reached at 60.5 + 0.33 or 60.83. This, therefore, is the *Mdn*. $\therefore$ *Mdn* = 60.8.

2. Computation from grouped data

For a continuous frequency distribution, *Mdn* is computed as P_{50} or Q_2 from cumulative frequencies.

$$Mdn = X_l + i \times \frac{0.50\,N - cf_l}{f_o}$$

where cf_l is the cumulative frequency of all class intervals below the true lower limit X_l of the interval containing the *Mdn*, f_o is the observed frequency in that interval, i is the length of class intervals, and $0.50N$ gives the proportion of the total frequency N to be counted off from one end of the distribution to reach the *Mdn*.

Such computations sometimes pose problems. (*a*) The *Mdn* may fall between two intervals, each with its own f_o ; in such cases, the true upper limit X_u of the lower of those two intervals is taken as the *Mdn* (Example 3.6.4). (*b*) The *Mdn* may fall in a vacant interval containing no case ; the mid-point X_o of this interval is then taken as the *Mdn* (Example 3.6.5).

Example 3.6.3.

Compute the median of the following frequency distribution of body weights (kg).

Class intervals :	51-53	54-56	57-59	60-62	63-65	66-68	69-71
Frequencies :	5	7	14	28	15	8	3

Solution :

(*i*) The data are arranged in Table 3.8. The true lower and upper limits (X_l and X_u) and the cumulative frequency (*cf*) of each interval are computed (pages 13-15, 26-27).

Table 3.8. Cumulative frequencies of body weight data.

Class intervals	True limits		*f*	*cf*
	lower (X_l)	upper (X_u)		
51—53	50.5	53.5	5	5
54—56	53.5	56.5	7	12
57—59	56.5	59.5	14	26
60—62	59.5	62.5	28	54
63—65	62.5	65.5	15	69
66—68	65.5	68.5	8	77
69—71	68.5	71.5	3	80 (*N*)

(*ii*) The length *i* of the class intervals is obtained by subtracting the X_l of any interval from the X_l of the next higher one. Thus, $i = 59.5 - 56.5 = 3$.

(*iii*) The number of scores, to be counted off from one end of the distribution to reach the *Mdn*, is given by $0.50N$.

$$0.50N = 0.50 \times 80 = 40.$$

(*iv*) The counting off of 40 scores with effect from the lowest score leads into the interval 60 – 62 in which the *Mdn* lies. The true lower limit (X_l) of this interval is 59.5, the interval has a frequency (f_o) of 28 cases, and the cumulative frequency (cf_l) upto the X_l of this interval amounts to 26.

$$X_l = 59.5 \; ; \; cf_l = 26 \; ; \; f_o = 28.$$

(*v*) The *Mdn* is then computed as follows :

$$Mdn = X_l + i \times \frac{0.50N - cf_l}{f_o}$$

$$= 59.5 + 3 \times \frac{40 - 26}{28} = 61.0 \text{ kg.}$$

Example 3.6.4.

Calculate the median for the following frequency distribution of achievement test scores in a group of students.

Class intervals :	67-76	77-86	87-96	97-106	107-116	117-126
Frequencies :	8	13	18	19	15	5

Solution :

The data are arranged in Table 3.9.

(*i*) The true lower and upper limits (X_l and X_u) and the cumulative frequency (*cf*) of each interval are computed (pages 13-15, 26-27).

(*ii*) The length i of the class intervals is obtained by subtrating the X_l of any interval from the X_l of the next higher one. Thus, $i = 86.5 - 76.5 = 10$.

Table 3.9. Cumulative frequencies of achievement test scores.

Class intervals	True limits		f	x'	fx'	cf	fx'^2
	lower (X_l)	upper (X_u)					
117—126	116.5	126.5	5	+2	10	78 (N)	20
107—116	106.5	116.5	15	+1	15	73	15
97—106	96.5	106.5	19	0	0	58	0
87—96	86.5	96.5	18	−1	−18	39	18
77—86	76.5	86.5	13	−2	−26	21	52
67—76	66.5	76.5	8	−3	−24	8	72
					$\Sigma fx' = -43$		$\Sigma fx'^2 = 177$

(*iii*) The number of scores, to be counted off from one end of the distribution to reach the *Mdn*, is given by $0.50N$.

$$0.50N = 0.50 \times 78 = 39.$$

(*iv*) The counting off of 39 scores, starting from the lowest interval, leads exactly upto the X_u of the interval 87-96 which has a *cf* of 39. The *Mdn*, therefore, falls between the intervals 87-96 and 97-106. So, the X_u 96·5 of the last interval counted off, viz., 87—96, is taken as the median. Thus, $Mdn = 96.5$.

[The same result is also obtained on applying the formula used in the last example.

$$Mdn = X_l + i \times \frac{0.50N - cf_l}{f_0}$$

$$= 96.5 + 10 \times \frac{39 - 39}{19} = 96.5.]$$

Example 3.6.5.

Calculate the median of the following frequency distribution of serum iron ($\mu g\ dl^{-1}$) in 32 humans.

Class intervals :	97—106	107—116	117—126	127—136	137—146	147—156
Frequencies :	3	5	8	0	11	5

Solution :

The data are arrangcd in Table 3.10.

(*i*) The true lower and upper limits (X_l and X_u) and the cumulative frequency (*cf*) of each interval are computed (pages 13-15, 26-27).

(*ii*) The length i of the class intervals is obtained by subtracting the X_l of any interval from the X_l of the next higher interval ; thus, $i = 116.5 - 106.5 = 10$.

Table 3.10. Cumulative frequencies of serum iron data.

Class intervals	True limits		f	cf
	lower (X_l)	upper (X_u)		
97—106	96.5	106.5	3	3
107—116	106.5	116.5	5	8
117—126	116.5	126.5	8	16
127—136	126.5	136.5	0	16
137—146	136.5	146.5	11	27
147—156	146.5	156.5	5	32 (N)

(*iii*) The number of scores, to be counted off from one end of the distribution to reach the *Mdn*, is given by $0.50N$.

$$0.50N = 0.50 \times 32 = 16.$$

(*iv*) The counting off of 16 scores starting from the lowest interval brings us exactly upto the X_u of the interval 117-126 which has the *cf* of 16. The next higher interval 127-136, in which the *Mdn* should fall, is a vacant interval with no case. So, the midpoint (X_c) of that interval is taken as the *Mdn*. For this interval,

$$X_c = X_l + \tfrac{1}{2}(X_u - X_l)$$
$$= 126.5 + \tfrac{1}{2}(136.5 - 126.5) = 131.5.$$
$$\therefore \quad Mdn = 131.5.$$

3.7 MODE

The mode (M_o) is that score of the variable which belongs to the largest number of individuals in a sample. It is, therefore, the most frequent score in the sample and coincides with that point on the X axis of a frequency distribution which corresponds to the peak of the latter. A distribution is called *unimodal*, *bimodal* or *multimodal* according to the presence of one or more peaks and as many M_o values. In a perfectly symmetrical unimodal distribution, M_o, $\overline{X}$ and *Mdn* are identical. In an asymmetric distribution, *Mdn* lies between M_o and $\overline{X}$ while M_o lies on that side of the *Mdn* which leads to the shorter tail of the distribution.

The amount and algebraic sign of the deviation of the mean or the median from the mode indicate respectively the degree and direction of asymmetry of the distribution.

In a symmetrical or slightly asymmetric distribution, $M_o = 3Mdn - 2\overline{X}$. For example, the mode of a frequency distribution, having a mean of 73.12 and a median of 73.0, is given by:

$$M_o = 3Mdn - 2\overline{X}$$
$$= 3 \times 73 - 2 \times 73.12 = 72.76.$$

In grouped data, M_o may be approximately given by the midpoint X_c of the class interval having the highest frequency of cases. In ungrouped data, the most frequent score is an approximate measure of the M_o.