CHAPTER 3

DESCRIPTIVE STATISTICS I: CENTRAL TENDENCY

3.1 INTRODUCTION

The tabular and graphical approaches of summarizing raw data are useful and effective in technical reports and documents and as visual aids when presentations must be made to an audience. Nevertheless, numerical descriptions in the form of indices or values often are preferred for summarizing a set of data. This is particularly true when the data set is a **sample** and is used as a basis for making statistical inferences about a **population**. From this decision-making point of view, use of numerical values or indices is an important and useful means of describing and summarizing data. These indices or values are all referred to as **descriptive statistics**.

When we are interested in describing the entire distribution of some observations or characteristics of individuals, there are two types of indices that are especially useful. These are the measures of central tendency and measures of variability. **Measures of central tendency** are numerical indices that attempt to answer the question: what is the typical value of the observations in this distribution? They are generally indicative of the 'centre' 'middle' 'representative' or the 'most typical' of a set. **Measures of variability** are numbers that attempt to answer questions like: how different from each other are the observations in the distribution? Or "how

do the observations in the distribution spread out around the typical value"? Measures of central tendency and measures of variability are both descriptive statistics because they are used to describe distributions of observations or characteristics of the individuals under study.

When we are interested in describing the position of a particular value relative to the other values in the distribution, we use numerical indices known as **measures of location**. Measures of location are also descriptive statistics, but they describe the position of one score relative to others rather than describing the whole set of data and hence they are not necessarily the central or middle values. In many texts, however, measures of 'location' and measures of 'central tendency' are considered as synonymous.

3.2 MEASURES OF CENTRAL TENDENCY

There are several different measures of central tendency. Each is an indicator of what a typical value is, but each employs a different definition of 'typical'. These measures are collectively called **statistical averages**. An average of a set of values of a variable is a single number, which is 'typical' of the whole set. Similarly, an average of a frequency distribution may be thought of a single value, which is assumed to be 'typical' of all values making up the frequency distribution. The purpose of an average is to represent the distribution and also to afford a basis of comparison with other distributions of similar nature. We describe some important averages in this section.

Among the several averages, the most commonly used averages are

(a) Mean
(b) Median and
(c) Mode

The mean can again be of three types:

- Arithmetic mean
- Geometric mean and
- Harmonic mean

In addition to these common measures, a few special measures are occasionally used. These, among others, are

- Quadratic mean

- Pooled mean
- Weighted mean
- Trimmed mean and
- Trimean

We present below a brief overview of the different types of averages referred to above

3.3 THE ARITHMETIC MEAN

The **arithmetic mean**, or simply the **mean**, is the most commonly used central value of a set of observations, more precisely of frequency distribution. A layman views this mean simply as an 'average'. It has all sorts of connotations in everyday language – we speak about a cricket player's batting average, we talk about the average income of a family, we describe a bride's or groom's appearance as average, we label your class performance as average and so forth. A statistician, however, views this as an **index** that provides a succinct description of the centrality of a set of data.

The arithmetic mean is calculated by totaling the results of all the observations and dividing this total by the number of observations when the data at hand are numerical. If the ages of six college-going boys are 16,18,17,15, 17 and 16 years, the mean age is

$$\frac{16+18+17+15+17+16}{6} = \frac{99}{6} = 16.5 \text{ years.}$$

It is obviously a typical value and centers around all the values of the set. Note that the average is, as it should be, a value that falls between the lowest value 15 and the highest value 18 and that it is not (and not necessarily should be) equal to any of the given values. This feature is true not only in this particular case, it is true in general.

We can now give a simple definition of arithmetic mean as follows:

***Definition* 3.1:** *The arithmetic mean is an average or central value of a set of observations obtained by summing these observations and then dividing this sum by the number of such observations.*

To give an algebraic formula for the mean, suppose let us suppose that we have n real numbers $x_1, x_2, \ldots x_n$. Then the arithmetic mean of these numbers denoted by $\bar{x}$ (read x bar), can be written in a symbolic form as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \quad \ldots (3.1)$$

Clearly, the mean $\bar{x}$ may be positive, negative or even zero depending on the nature of the variable included in any computation.

Introducing the symbol Σ (termed **sigma**, the Greek letter for S), we write the mean $\bar{x}$ as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \ldots (3.2)$$

Here Σ is the notation for summing the x_i values starting with i=1 (written beneath the summation) and ending with $i = n$ (written above the summation). In many instances, the limits i=1 and $i = n$ are omitted when all the values of n are to be summed unless otherwise required.

3.3.1 Sample Mean and Population Mean

If we regard the 6 students as a sample taken from all the students of the school (population), then $\bar{x}$ is the **sample mean**. This mean may be quite different from the one computed for all the students, which is referred to as the **population mean**. The formula for computing the population mean is similar to that of the sample mean, but we use different notation to indicate that we are dealing with the entire population. The number of items in the population is usually denoted by N, and the symbol for the population mean is a Greek letter μ (*mu*). Thus,

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \ldots (3.3)$$

It is important to note that the arithmetic mean can be calculated only for numerical (either interval or ratio) data. Categorical (nominal or ordinal) data such as sex, opinion, health status, do not permit calculation of arithmetic mean. For dichotomous variable coded as 0 and 1, the mean, however, has a special meaning. It is the proportion of cases 1 pertaining to the data.

3.3.2 Arithmetic Mean for Ungrouped Data

In computing sample mean $\bar{x}$ or population mean μ, we assumed in (3.2) and (3.3) that each value of the variable x occurs once in the

series $x_1, x_2, \ldots, x_k$. In the case when x_1 is repeated f_1 times, x_2 is repeated f_2 times, …, x_k is repeated f_k times, the formula for the arithmetic mean assumes the form

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots + f_kx_k}{f_1 + f_2 + \ldots\ldots + f_k} = \frac{\sum f_ix_i}{\sum f_i} \quad \ldots (3.4)$$

This is essentially a mean calculated from raw data arranged in the form of a simple frequency distribution. Here we have k distinct values of x, namely $x_1, x_2, \ldots, x_k$, each of the values occurring with frequencies $f_1, f_2, \ldots, f_k$ respectively. If $f_1 = f_2 = \ldots = f_k = 1$, the formula (3.4) is identical to formula (3.2). The following table illustrates the calculation of $\bar{x}$ using (3.4)

Value (x_i)	Frequency (f_i)	Product (f_ix_i)
x_1	f_1	f_1x_1
x_2	f_2	f_2x_2
..	..	..
x_k	f_k	f_kx_k
Total	Σf_i	Σf_ix_i

The mean is then calculated as a ratio of $\sum f_ix_i$ to $\sum f_i$

$$\bar{x} = \frac{\sum f_ix_i}{\sum f_i}$$

Example 3.1: A sample survey of Bangladesh Bureau of Statistics in a rural area collected the age at first marriage in years of 330 newly married women. The distribution was as follows:

Table 3.1: Age at first marriage (AFM) in years of 330 women

AFM	Women	AFM	Women
11	17	16	48
12	28	17	36
13	37	18	23
14	52	19	11
15	70	20	8

Calculate the mean age at first marriage for the women in the sample.

Solution: This is a frequency distribution for which the formula (3.4) is applicable. The computational steps involved in obtaining Σf_i and Σf_ix_i. Here the age at marriage, denoted by x, serves as the variable. This computation is shown below:

Age at marriage (x_i)	Number of women (f_i)	Product of col.1 & col, 2 (f_ix_i)
11	17	187
12	28	336
13	37	481
14	52	728
15	70	1050
16	48	768
17	36	612
18	23	414
19	11	209
20	8	160
Total	330	4945

Here k=10 and

$$\sum f_ix_i = 4945, \quad \sum f_i = 330$$

Hence employing (3.4), we arrive at the mean age as follows:

$$\bar{x} = \frac{\sum f_ix_i}{\sum f_i} = \frac{4945}{330} = 14.98 \text{ years}$$

Thus the mean age at marriage for the women under study is about 15 years. Note that this is a **sample mean**, since the 330 women were chosen from the community on a sample basis. We cannot compute the population mean μ unless we have the age at marriage of the entire population from which the sample was drawn. Nevertheless, if you can regard your sample as a representative of the population, your sample mean will serve as an **estimate** of the population mean.

3.3.3 Arithmetic Mean for Grouped Data

When the arithmetic mean is computed from grouped distribution, the mid-point of each class is taken as the representative value of that class. The various mid-values are multiplied by their respective class frequencies, the products are added, and the sum of the products is then divided by the total

number of observations to obtain the arithmetic mean. Symbolically, if y_1, y_2, ...y_n are the mid-values and f_1, f_2,...f_k the corresponding frequencies, where the subscript 'k' stands for the number of classes, then the mean $\bar{y}$ is

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i} \qquad \ldots (3.5)$$

The assumption underlying the choice of the mid-values y_1, y_2,...y_n in computing the mean is that the observations are uniformly or symmetrically distributed about the mid-values within the class intervals.

Example 3.2: Calculate arithmetic mean for the data in Table 2.6 (Chapter 2)

Solution: To compute the mean age, we reproduce the table under reference

Table 3.2: Computing arithmetic mean for data in Table 2.1

Age (in years)	Mid-value (y_i)	Frequency (f_i)	Product ($f_i y_i$)
24.5–29.5	27	3	81
29.5–34.5	32	9	288
34.5–39.5	37	15	555
39.5–44.5	42	12	504
44.5–49.5	47	7	329
49.5–54.5	52	4	208
Total	-	50	1965

Here

$$\sum f_i = 50 \text{ and } \sum f_i y_i = 1965$$

so that

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i} = \frac{1965}{50} = 39.3 \text{ years}$$

You can check that the arithmetic mean calculated from raw data in Table 2.1 is:

$$\bar{x} = \frac{25 + 29 + \ldots + 32 + 50}{50} = \frac{1960}{50} = 39.2 \text{ years.}$$

Note that the calculated mean from the grouped data differs by only a small margin of 0.1 from that calculated from raw data. In some instances,

this difference may even be larger. This is expected, since the items of each class are often not uniformly or systematically distributed throughout the intervals.

We can present the formula (3.5) as follows also:

$$\bar{y} = \sum\left(\frac{f_i}{\sum f_i}\right) y_i = \sum\left(\frac{f_i}{n}\right) y_i = \sum r_i y_i \qquad \ldots (3.6.)$$

where $r_i = f_i/n$ denotes the relative frequency of the ith class and $n = \Sigma f_i$. This shows that when you have a relative frequency distribution in hand, you can obtain the mean values of the data by simply multiplying the relative frequencies by the mid-points of the respective classes. This is also equally true for ungrouped data where x is used in place of y.

3.3.4 Short-cut Method of Computing Arithmetic Mean

The computation of mean is sometimes facilitated by reducing the magnitude of the given data through appropriately changing the origin and unit. Suppose we have a set of n values $x_1, x_2, \ldots, x_n$. We subtract a constant quantity 'a' say from each of the values of x and obtain a new set of values $d_1, d_2, \ldots, d_n$. Here 'a' is variously known as the **origin factor, assumed mean** or **provisional mean**. Then the mean for the ungrouped distribution can be calculated employing the formula

$$\bar{x} = a + \frac{\sum d_i}{n} = a + \bar{d} \qquad \ldots (3.7)$$

where $\bar{d}$ is analogous to $\bar{x}$, representing the arithmetic mean of the d values. The formula is proved as follows.

Since $d_1 = x_1 - a$, $d_2 = x_2 - a$, ..., $d_n = x_n - a$,

$$\bar{d} = \frac{d_1 + d_2 + \ldots + d_n}{n}$$

$$= \frac{(x_1 - a) + (x_2 - a) + \ldots + (x_n - a)}{n}$$

$$= \frac{(x_1 + x_2 + \ldots + x_n) - (a + a + \ldots + a)}{n}$$

$$= \frac{\sum x_i - na}{n} = \bar{x} - a$$

Rearranging the foregoing expression, we obtain

$$\bar{x} = a + \bar{d} \qquad \ldots (3.7a)$$

Let us illustrate the method by an example.

Example 3.3: Compute the arithmetic mean of the values 240, 211, 447, 380, 410, 190 by suitably changing the origin.

Solution: Let the origin be chosen arbitrarily at 300. This implies that a=300. We then construct a table of the following form to compute d_i values:

x_i	$d_i = x_i - a$
240	–60
211	–89
447	147
380	80
410	110
190	–110

Here $\sum d_i = 78$, so that the mean of the d_i values is

$$\bar{d} = \frac{\sum d_i}{n} = \frac{78}{6} = 13$$

Hence employing (3.7a)

$$\bar{x} = a + \bar{d} = 300 + 13 = 313$$

This mean is the same as the one computed from the data in column 1, as you can verify:

$$\bar{x} = \frac{240 + 211 + 447 + 380 + 410 + 190}{6} = \frac{1878}{6} = 313.$$

The formula can be more advantageously used when the frequency distributions have equal class-widths. Thus if h stands for a common scale factor for all classes, we can modify the formula (3.7) by redefining d_i as follows:

$$d_i = \frac{x_i - a}{h}$$

the x_i being the mid-value of the i-th class

The constant h is usually taken to be the class-width to make d values as small as possible to ease the computation. Having made this transformation in the variable x, the formula for computing the mean is as follows:

$$\bar{x} = a + h\left(\frac{\sum f_i d_i}{n}\right) = a + h\bar{d} \qquad \ldots (3.8a)$$

The method is sometimes called **coding method** or **short-cut method.**

The proof of (3.8a) is undertaken as a theorem later in this chapter (see Theorem 3.1).

Example 3.4: Calculate arithmetic mean of the age data in Table 3.2 by short-cut method.

Solution: For computational convenience, we reproduce the table below:

Age (in years)	Mid-value (x_i)	Frequency (f_i)	Transformed value $d_i = (x_i - a)/h$	Product ($f_i d_i$)
24.5–29.5	27	3	–2	–6
29.5–34.5	32	9	–1	–9
34.5–39.5	37	15	0	0
39.5–44.5	42	12	1	12
44.5–49.5	47	7	2	14
49.5–54.5	52	4	3	12
Total	-	50	–	23

For this example, we take

$$a = 37,\ h = 5 \text{ and } d = \frac{x - 37}{5}.$$

Thus

$$\bar{d} = \frac{\sum f_i d_i}{\sum f_i} = \frac{23}{50} = 0.46$$

so that

$$\bar{x} = a + h\bar{d} = 37 + 5(0.46) = 39.3$$

which exactly agrees with the one obtained earlier (see Table 3.2). Appreciable time is saved by this short-cut method, because we work with small whole numbers instead of mid-points, which may be very large

values with several digits. If the assumed mean can be chosen intelligently, and h is uniform throughout the distribution, the short-cut method will always be time-saving without any loss of accuracy.

The method is based on an important property of the mean namely, the algebraic sum of the observations from their mean is zero or symbolically, $\sum(x_i - \bar{x}) = 0$. This implies that the sum of the deviations from an arbitrarily chosen value other than the mean is not 'zero'. Thus, if the mid-point of a given class is taken as the **assumed** or **provisional mean** in computing the **exact mean**, it will require necessary correction. It may be worth to note here that although any mid-point (y) can be taken as the assumed mean, it is more convenient to choose the mid-point of the class nearer the centre of the distribution, which has relatively large frequency.

The computation of the arithmetic mean by the so called short-cut method is based on a simple theorem that we state and prove below:

***Theorem* 3.1:** *Arithmetic mean is dependent on both origin and scale of measurement.*

Proof: Let x be a quantitative variable taking on values $x_1, x_2, \ldots, x_k$, the corresponding class frequencies being $f_1, f_2, \ldots f_k$. Let d be a new variable taking on the values $d_1, d_2, \ldots, d_k$ such that

$$d_i = \frac{x_i - a}{h} \qquad \ldots (3.8b)$$

where a is an arbitrary value and h is a common class width.

A little rearrangement of the expression (3.8b) results in

$$x_i = a + hd_i \qquad \ldots (3.9)$$

Multiplying both sides of (3.9) by f_i and then summing

$$\sum f_i x_i = a\sum f_i + h\sum d_i \qquad \ldots (3.10)$$

Dividing (3.9) throughout by $n\left(=\sum f_i\right)$, we arrive at

$$\bar{x} = a + h\left(\frac{\sum f_i d_i}{n}\right) = a + h\bar{d} \qquad \ldots (3.10a)$$

Hence the proof.

Clearly, the arithmetic mean $\bar{x}$ does not remain free of a and h, meaning that $\bar{x}$ is dependent on these two factors. Note that when h=1, formula (3.10a) is identical to (3.7).

The computation of mean from grouped data by the so called short-cut method involves the following steps:

1. Determine the mid-values of each class of the distribution.
2. Decide on the provisional mean 'a'
3. Identify the size of the class. This is 'h'
4. Subtract 'a' from each mid-value determined in step (2).
5. Divide the resulting differences obtained in step (4) by h to obtain d.
6. Multiply each d by the corresponding f and add the resulting product. This is $\sum f_i d_i$.
7. Multiply $\sum f_i d_i$ by h, and divide the resulting value by n.
8. Add a to the obtained in step 7 above obtain $\bar{x}$. That is

$$\bar{x} = a + h\left(\frac{\sum f_i d_i}{n}\right)$$

It is important to note that the use of short-cut method is rewarding only when you can choose a and h intelligently and h is uniform throughout the distribution.

3.3.5 Pooled Mean

Pooled mean, also called **combined mean**, is obtained by combining two or more sub-group means into a single mean. Thus, if $\bar{x}_1$ is the mean of n_1 observations, and $\bar{x}_2$ is the mean of n_2 observations, then the pooled mean of $n_1 + n_2$ observations is calculated as

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad \ldots (3.11)$$

Example 3.5: A family consists of 10 members of which 6 are males and 4 are females. Their average ages were 48 and 37 years respectively. What is the average age of all the 10 members of the family?

Solution: Here $n_1 = 6$, and $n_2 = 4$, $\bar{x}_1 = 48$, $\bar{x}_2 = 37$. Hence

$$\bar{x}_c = \frac{6\times 48 + 4\times 37}{6+4} = \frac{436}{10} = 43.6 \text{ years.}$$

For k sub-groups, the mean is computed as

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum n_i\bar{x}_i}{\sum n_i} \quad \ldots (3.11a)$$

where $\bar{x}_i$ $(i = 1, 2, \ldots, k)$ is the mean of the i-th set or group.

Example 3.6: The average wages of the workers in Table 2.1 (Chapter 2) by their religion were as follows:

Religion	Number of workers	Average wages
Muslim	36	72.1
Hindu	9	73.0
Christian	5	77.6
Total	50	72.8

Compute the mean of all the 50 workers.

Solution: Here $n_1 = 36$, $n_2 = 9$, $n_3 = 5$, $\bar{x}_1 = 72.1$, $\bar{x}_2 = 73.0$ and $\bar{x}_3 = 77.6$ so that the combined mean on using (3.11a) is

$$\bar{x}_c = \frac{36 \times 72.1 + 9 \times 73.0 + 5 \times 77.6}{50} = 72.8$$

As you can verify, this mean is in complete agreement with the one computed directly from the wage data in Table 2.1:

$$\bar{x} = \frac{93 + 66 + \ldots\ldots + 71 + 57}{50} = \frac{3641}{50} = 72.8$$

Example 3.7: A department store reported an average sales of Tk. 20 lac in first 10 days for the month of April, 25 lac in the next 7 days of the month and 32 lac during the remaining 13 days of the month. Compute the sales per day during the whole month.

Solution: Given the values, the combined mean is

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{10 \times 20 + 7 \times 25 + 13 \times 32}{10 + 7 + 13} = 26.37 \text{ lac}$$

3.3.6 Weighted Arithmetic Mean

In an ordinary arithmetic mean discussed above, each item in the set is assumed to have equal importance. But there are situations where the relative importance of different items is not the same. When this is the

case, we compute **weighted arithmetic mean**. The term 'weight' stands for the relative importance of different items. This is the mean in which each value of the variable under study is weighted according to its importance in the group of items.

Definition **3.3:** *Weighted mean is an average of quantities to which have been attached a series of weights in order to make proper allowance for their relative importance.*

The weighted arithmetic mean of a set of observations $x_1, x_2, \ldots x_n$, with respective weights $w_1, w_2, \ldots w_n$ is given by

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad \ldots (3.12)$$

For an understanding of the use of weighted mean, note the following example.

Example 3.8: A rice hoarder anticipates that during the next monsoon, the price of rice will go up. This increase, as the hoarder guesses, however, will not be uniform for all varieties of rice. He speculates that variety C will show the highest increase, followed by B and variety A will almost be of equal demand as B. To maximize his profit, he therefore decides to purchase 270 maunds of variety C, 180 maunds of variety B and 150 maunds of variety A. The accompanying table shows the current price of the rice by variety and quantity to be purchased. The fourth column shows the hoarder's anticipated importance of varieties B and C relative to A.

Variety	Price per maund (in Tk.)	Quantity to be purchased (in maund)	Relative weight
A	425	150	1.0
B	350	180	1.2
C	400	270	1.8

Here $x_1 = 425$, $x_2 = 350$, $x_3 = 400$, and $w_1 = 150, w_2 = 180, w_3 = 270$, so that the weighted arithmetic mean based on (3.12) is

$$\bar{x}_w = \frac{150(425) + 180(350) + 270(400)}{150 + 180 + 270} = \frac{23,4750}{600} = 391.25$$

Note that the importance the hoarder assigns (here 1:1.2:1.8) to each variety are just his best guess in the face of uncertainty. Some other persons might assign a very different set of weights based on his own

experience. The weights are thus entirely subjective and may vary from situation to situation, or from person to person.

You will arrive at the same mean if instead of the absolute weights (here 150, 180, 270) you use the relative weights 1, 1.2, and 1.8. This can be seen from the following computation:

$$\bar{x}_w = \frac{1\times 425 + 1.2\times 350 + 1.8\times 400}{1+1.2+1.8} = \frac{1565}{4} = 391.25$$

as before.

Example 3.9: The accompanying table shows the unemployment rates and the civilian labor force in the United States in 2001 by regions. Compute the average unemployment rates for the entire United States.

Solution: To average the unemployment rates, we must weight them with their respective labor force (in millions). As before, we construct a table below to compute the overall average percentage.

Regions	Labor force (w)	Unemployment rate (%) (x)	Product ($w \times x$)
Northeast	26.9	4.1	110.29
South	50.6	4.7	237.82
Midwest	34.7	4.4	152.68
West	32.5	5.0	162.50
Total	144.7	$\bar{x}_w$	663.29

$$\sum w_i = 144.7 \text{ and } \sum w_i x_i = 663.29$$

Hence the weighted mean, applying (3.12) is

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{663.29}{144.7} = 4.58\%$$

Example 3.10: The accompanying table shows the dividend yield and market capitalization of 5 company groups registered with the Dhaka Stock Exchange during 1995–96. Compute the overall dividend yield for the company groups.

Company groups	Market capitalization (Taka in crores)	Dividend yield (%)
Bank and Investments	556.96	2.18
Engineering and Constructions	705.84	1.47
Food and Allied Products	823.72	3.15
Jute Industries	34.42	2.28
Textile Industries	641.43	1.66

Solution: The overall-mean yield is the weighted arithmetic mean of the individual dividend yield of the company groups. This average is worked out by weighting the individual company group's dividend yield by their corresponding aggregate market capitalization. Thus

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{556.96\times 2.18 + \ldots + 641.43\times 1.66}{556.96 + \ldots + 641.43} = \frac{5989.727}{2762.37} = 2.17$$

The choice of the weights did not pose any difficulties in either of the above examples, but there are situations in which the selection of the weights is far from obvious. For instance, if we wanted to compare figures on the cost of living in different cities or different years, it would have been difficult to account for the relative importance of such items as food, rent, medical care and so on. Personal judgment is sometimes applied in such cases to decide on the relative weights.

The weighted average is frequently used in index number construction. An index number is a statistical device used to study the changes in any event (such as price, consumption, birth, death etc) over time. If p_0 and p_n are the base year and current year prices respectively, we can compute a quotient p_n/p_o, known as **price relative** and weight this quotient by the value of the item obtained as a product of p_n and q_n (i.e. $p_n \times q_n$), where q_n is the current year quantity of the same item. The weighted arithmetic average of the price relatives I_{0n} is then defined as

$$I_{0n} = \frac{\sum p_n q_n \left(\frac{p_n}{p_0}\right)}{\sum p_n q_n} \times 100 \qquad \ldots (3.13)$$

where the sum extends over all commodities covered in the computation process. Note that the index number formula (3.13) is of the form (3.12).

Example 3.11: From the following data, calculate the weighted index number and comment on the changes in the cost of living of the consumers of the commodities listed in the table.

Commodities	Base year (2005) Price (p_0)	Base year (2005) Consumption (q_0)	Current year (2008) Price (p_n)	Current year (2008) Consumption (q_n)
Rice	22	8	38	8
Wheat	14	10	35	12
Sugar	28	5	35	4

Solution: To employ the formula above, we construct the following table

Commodities	$p_n q_n$	$\frac{p_n}{p_0}$	$p_n q_n \left(\frac{p_n}{p_0}\right)$
Rice	304	1.73	525.92
Wheat	420	2.50	1050.0
Sugar	140	1.25	175.0
Total	864	–	1750.92

The average increase is

$$\bar{x}_w = \frac{1750.92}{864} \times 100 = 202.6$$

Hence the cost of living of the consumers has increased by more than 102 percent (more than double) over a period of 3 years although quantity consumed in 2005 has not changed at all during this period.

To summarize, the pooled mean is used when we want to average several group or category means, while the weighted mean is used when we wan to average, rates, ratios, proportions and percentage values.

Properties of Arithmetic Mean

Arithmetic mean as a measure of central tendency has some appealing properties. These are discussed below:

***Property* 3.1:** *The algebraic sum of the deviations of the values $x_1, x_2, ..., x_n$ from their arithmetic mean $\bar{x}$ is zero. That is*

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad \ldots (3.14)$$

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0 \qquad \ldots (3.14a)$$

That is,

This is equally true for a frequency distribution. That is

$$\sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0$$

The proof is simple. The expression (3.14a) can be expanded and simplified as follows:

$$(x_1 + x_2 + \ldots + x_n) - \underbrace{(\bar{x} + \bar{x} + \ldots + \bar{x})}_{n \text{ terms}} = \sum x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

To numerically demonstrate this, let us consider 5 observations 2, 5, 6, 7, 10. The mean of these observations is 6. Thus subtracting 6 from each of the values and adding the resulting differences

$$(2–6)+(5–6)+(6–6)+(7–6)+(10–6)=0$$

It is important to note that no other value (except the mean) would satisfy (3.14a). Therefore, mean is the value that balances all of the values on either side of it. Thus, $\sum(x_i - k) = 0$ only when k is the arithmetic mean.

***Property* 3.2:** *The arithmetic mean is sensitive to extreme values.*

An arithmetic mean is very sensitive to the influence of a few large values in a collection of a predominantly small numbers; thus it sometimes fails to indicate the centre of a series of data. An example will make this point clear. Suppose a survey of 1000 families showed that each of 3 families has an income of Tk. 100,000 and each of the remaining 997 families has incomes of Tk. 3000. The arithmetic mean of these 1000 families is (3×100,000+997×3000)/1000=Tk.3219. But this average cannot be considered as a central value since 997 or 99.7 percent of all the families have incomes only slightly less than that amount.

To conceptualize this feature further, consider the following sets of observations:

Set A	Set B
2	2
3	3
5	5
7	7
8	33
Mean=5	Mean=10

Note that all the observations in both the sets are the same except that for the very large value of 33 in the second set. This one extreme value is sufficient to double the size of the mean. Clearly, 10 is typical of the five values entered into its computation. Thus when a distribution is markedly skewed, the mean provides a misleading estimate of the central tendency.

Property* 3.3:** *If a set consists of n_1 observations of the form $x_{11}, x_{12}, \ldots x_{1n_1}$ with mean $\bar{x}_1$ and a second set consists of n_2 observations of the form, $x_{21}, x_{22}, \ldots x_{2n_2}$ with mean $\bar{x}_2$. then the mean of all the n_1+n_2 observations called* ***combined *mean or* ***pooled mean,*** *is given by the expression*

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Proof: By definition, the mean $\bar{x}$ of the first set is

$$\bar{x}_1 = \frac{x_{11} + x_{12} + \ldots\ldots + x_{1n_1}}{n_1} = \frac{\sum x_{1i}}{n_1}$$

which gives

$$\sum x_{1i} = n_1\bar{x}_1 \qquad \ldots (a)$$

Similarly, the mean $\bar{x}_2$ of the second set is

$$\bar{x}_2 = \frac{x_{21} + x_{22} + \ldots\ldots + x_{2n_2}}{n_2} = \frac{\sum x_{2i}}{n_2}$$

from which

$$\sum x_{2i} = n_2\bar{x}_2 \qquad \ldots (b)$$

Adding (a) and (b)

$$\sum x_{1i} + \sum x_{2i} = n_1\bar{x}_1 + n_2\bar{x}_2 \qquad \ldots (c)$$

The combined set consists of $n_1 + n_2$ observations as below

$$x_{11}, x_{12}, \ldots x_{1n_1}, x_{21}, x_{22}, \ldots, x_{2n_2}$$

Thus the mean of the combined set is

$$\bar{x}_c = \frac{\left(x_{11} + x_{12} + \ldots + x_{1n_1}\right) + \left(x_{21} + x_{22} + \ldots + x_{2n_2}\right)}{n_1 + n_2}$$

$$= \frac{n_1\left(\frac{x_{11} + x_{12} + \ldots + x_{1n_1}}{n_1}\right) + n_2\left(\frac{x_{21} + x_{22} + \ldots + x_{2n_2}}{n_2}\right)}{n_1 + n_2}$$

$$= \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Hence the proof.

The result can be extended to k sets of observations, in which case the above expression assumes the form

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots\ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots\ldots + n_k} = \frac{\sum n_i\bar{x}_i}{\sum n_i} \qquad \ldots (3.15)$$

where $\bar{x}_i$ (i=1, 2, ..k) is the mean of the i-th set comprising n_i observations.

***Property* 3.4:** *The sum of squared deviation from the arithmetic mean is less than the sum of squared deviation from any other value.*

To illustrate this property, we examine the values in the accompanying table, where the deviations of the values have been taken from a number of values including the arithmetic mean.

x	$(x-2)^2$	$(x-3)^2$	$(x-\bar{x})^2$	$(x-5)^2$	$(x-6)^2$
2	0	1	4	9	16
3	1	0	1	4	9
4	4	1	0	1	4
5	9	4	1	0	1
6	16	9	4	1	0
Sum	30	15	10	15	30

* $\bar{x} = 4$

Note that the sum of squares is the smallest in column 4, when deviations are taken from the mean value 4. This property of the mean is known as the least-squares property of the mean, which is of considerable importance in statistics, particularly when it is applied to curve fitting.

In general, if $\bar{x}$ is the mean of a set of observations and a is any arbitrary constant, then the above property can be stated in general term as

$$\sum(x_i - a)^2 \geq \sum(x_i - \bar{x})^2$$

The above inequality is simple to prove:

$$\sum(x_i-a)^2=\sum(x_i-\bar{x}+\bar{x}-a)^2$$
$$=\sum(x_i-\bar{x})^2+n(\bar{x}-a)^2+2(\bar{x}-a)\sum(x_i-\bar{x})$$
$$=\sum(x_i-\bar{x})^2+n(\bar{x}-a)^2$$

showing that

$$\sum(x_i-a)^2\geq\sum(x_i-\bar{x})^2$$

***Property* 3.5**: *Arithmetic mean is the most stable measure of central tendency than any other measures.*

Suppose that we draw 30 random samples of 500 persons who are registered voters in a city election. If we compute the mean, the median and the modal age for each of the 30 samples, we would find that the frequency distribution constructed from the 30 sample means would show less variability than either the frequency distribution of the 30 sample modes or the sample medians.

***Property* 3.6:** *The arithmetic mean depends on both (i) origin and (ii) scale of measurement.*

This property has been illustrated numerically in examples (3.3 and 3.4) and proved as a theorem earlier (see Theorem 3.1).

***Property* 3.7:** *If a and b are constants such that $x=a\pm by$, where x and y are two variables assuming values $x_1, x_2,...x_n$ and $y_1, y_2,...,y_n$ respectively, then $\bar{x}=a\pm b\bar{y}$*

Proof: Since $x=a\pm by$,

$$\sum x_i=\sum(a\pm by_i)=(a\pm by_1)+(a\pm by_2)+...+(a\pm by_n)$$
$$=(a+a+...+a)\pm b(y_1+y_2+...+y_n)$$
$$=na\pm b\sum y_i$$

Dividing both sides by n, we have

$$\frac{\sum x_i}{n}=a\pm\frac{b\sum y_i}{n}$$

This shows that

$$\bar{x}=a\pm b\bar{y}$$

***Property* 3.8:** *The arithmetic mean of the sum of two or more variables is equal to the sum of their means. That is if $u=x\pm y$, then $\bar{u}=\bar{x}\pm\bar{y}$*

Proof: Since $u=x\pm y$,

$$\sum u_i=\sum(x_i\pm y_i)$$
$$=(x_1\pm y_1)+(x_2\pm y_2)+...+(x_n\pm y_n)$$
$$=(x_1+x_2+...+x_n)\pm(y_1+y_2+...+y_n)$$
$$=\sum x_i\pm\sum y_i$$

Dividing both sides by n,

$$\bar{u}=\bar{x}\pm\bar{y}$$

***Property* 3.9**: *The arithmetic mean of first n natural numbers is given by*

$$\bar{x}=\frac{n+1}{2}$$

The first n natural numbers are 1, 2, 3, …,n, the mean of which is

$$\bar{x}=\frac{1+2+...+n}{n}=\frac{\frac{n(n+1)}{2}}{n}=\frac{n+1}{2}$$

Example 3.12: Find the mean of the following sets:

(a) 1, 2, 3, …, 50, and (b) 510, 520, 530, …,610.

Solution:

(a) To sum the first 50 natural numbers 1, 2, ..,50 by property 9 is

$$\bar{x}=\frac{n+1}{2}=\frac{50+1}{2}=25.5$$

(b) Let x take on the values 510, 520, …,610. To convert the given set into a set of natural numbers, we subtract 500 from each of the given values and divide the resulting value by 10 and obtain a new set of values of the form 1, 2, 3, …,11. This is a set of 11 natural numbers so that a=500 and h=10.

$$\bar{d}=\frac{n+1}{2}=\frac{12}{2}=6$$

Hence

$$\bar{x}=a+h\bar{d}=500+10(6)=560$$

3.4 THE MEDIAN

We have already demonstrated that arithmetic mean is influenced by the presence of unusually very high or low values (see Property 2 of arithmetic mean). While such sensitivity of the mean is sometimes desirable in some instances, it is not always so. In the latter cases, the use of some other measures, such as the **median** will be more appropriate. Like mean, median is also a measure of central tendency. We define this as follows:

***Definition* 3.2:** *The median is the value in a set of ranked or ordered observations that divides the whole set of observations into two parts of equal size.*

The implication of the above definition is that a median is the middle- most value of the observations such that the number of observations above it is equal to the number of observations below it. The median is sometimes called **positional average** because it lies in the middle of the data set after the values in the set have been placed in an ordered array.

As with the mean, the method of obtaining median depends on whether the data are grouped or ungrouped.

3.4.1 Properties of the Median

(a) The median is unique; that is, like the mean, there is only one median for the whole set of data.

(b) It is unaffected by the outliers, the extremely large or small values, and is therefore, a valuable measure of central tendency when such values exist in the data set.

While discussing the property of arithmetic mean (see Property 2), we had two sets of data as we reproduce below:

Set A: 2, 3, 5, 7, 8
Set B: 2, 3, 5, 7, 33

The arithmetic means are respectively 5 and 10 for A and B. The significantly higher value of the mean for set B is entirely due to the presence of the extreme value 33 in the set. As you can note, the median in both the sets is 5, thus demonstrating that median is not affected by the presence of extreme values.

(c) The median can be computed from distributions with open-ends classes.

(d) Unlike the mean, the median can be obtained for all levels of data except the nominal.

A few disadvantages of the median are that

(a) An overall (pooled) median cannot be obtained from a set of medians, whereas an overall mean can be obtained as a weighted average of several means.

(b) Medians are less stable; that is, medians vary more in repeated sampling than do means. This is a serious disadvantage when it is desired to draw an inference concerning a population.

(c) Median cannot be calculated for nominal data, since raking of the observations is not possible.

3.4.2 Computing Median from Raw Data

Suppose we have a family of seven members whose ages in years are 12, 7, 2, 34, 17, 21 and 19. To compute the median of these numbers, we arrange them either in ascending order or descending order. In either orderings, the middle most value is the median. Arranging them in both orders, the series is

Ascending order: 2, 7, 12, **17**, 19, 21, 34.
Descending order: 34, 21, 19, **17**, 12, 7, 2.

The middle-most value in either orderings is the fourth value, which in this case is 17. Three observations in the first arrangement, *viz.* 2, 7, and 12 fall below this value and the other three values *viz.* 19, 21, and 34 fall above this value. Thus by definition, 17 is the median of these values.

The choice of the middle-most value does not pose any problem so long as the series consists of odd number of observations. How would we deal with the problem when the number of observations is even? Suppose a new member joins the family aged 26 years. The set of values now in ordered array is:

2, 7, 12, 17, 19, 21, 26, 34

In this instance, any value greater than 17 but less than 19 will be the median. It is a usual assumption that median lies at the centre of these two values and arithmetic mean of these two values is taken as the median of the set. If we let $\tilde{m}$ to stand for the median, then

$$\tilde{m} = \frac{17+19}{2} = 18$$

As a general rule, for n observations, the median either assumes the value of one of the observations or falls between two values. If n is divisible by 2 (i.e. $n/2$ is an integer), the median $\tilde{m}$ has the value half-way between the $n/2$-th value and $(n/2+1)$-th value:

$$\tilde{m} = \frac{\left(\frac{n}{2}\right)\text{th value} + \left(\frac{n}{2}+1\right)\text{th value}}{2}$$

If n is not exactly divisible by 2 (i.e. $n/2$ is not an integer), the median has the value of the next higher integer. That is

$$\tilde{m} = \left(\frac{n+1}{2}\right)\text{th value}$$

Example 3.13: The weights of 11 mothers in kg were recorded as follows:

47, 44, 42, 41, 58, 52, 55, 39, 40, 43 61

To obtain the median weight, we arrange the values in ascending order first. When we do so, the set appears as follows:

39, 40, 41, 42, 43, 44, 47, 52, 55, 58, 61

Since $n/2$=5.5 (which is not an integer), the next higher integer is 6. The 6th observation is 44 and hence 44 is the median. In other words

$$\tilde{m} = \left(\frac{11+1}{2}\right)\text{th value} = 6\text{th value} = 44$$

Example 3.14: If one woman with weight 52 kg (say) is removed from the data set in Example 3.13, the set now consists of the values

39, 40, 41, 42, 43, 44, 47, 55, 58, 61

Now we have n=10, which is divisible by 2. By definition, the median will be the average of the 5th and 6th observations.

$$\tilde{m} = \frac{\left(\frac{10}{2}\right)\text{th value} + \left(\frac{10}{2}+1\right)\text{th value}}{2}$$

$$= \frac{5\text{th value} + 6\text{th value}}{2} = \frac{43+44}{2} = 43.5$$

Example 3.15: Suppose in a survey, 5 individuals were ranked on the basis of their level of education. Out of five individuals, one was found to have 'no education', one 'primary', one 'secondary', one 'higher secondary'; and the remaining one as 'graduate'. The median ranking is 'secondary', because half of the ratings are above 'secondary', and the other half are below it.

3.4.3 Computing Median for Ungrouped Data

A discrete frequency distribution may appear in ungrouped form. We illustrate below how median can be computed in such cases.

Example 3.16: A survey was conducted among 100 families to know their family size. The result of this investigation gave the following distribution:

Family size:	1	2	3	4	5	6	7	8	9	Total
No. of families:	2	6	12	18	19	15	11	11	6	100

For this distribution, n (=100) is exactly divisible by 2. Thus the median lies in the mid-way between the 50th and 51st values. Note that this value cannot be readily located simply by counting as we did before. One way of locating this value is to form an explicit series from the distribution and count from the beginning until we reach the 50th and 51st positions and take the average of the values corresponding to these two positions. This can be done by writing the value 1 two times, 2 six times, 3 twelve times and so on. This process shows that both the 50th and 51st observations are 5. Hence the median is (5+5)/2 = 5.

The process just described is a cumbersome one. The process can be shortened intelligently if we form a cumulative frequency (CF) column and locate the observation between $\frac{1}{2}n$th and $(\frac{1}{2}n+1)$th i.e. between 50th and 51st positions of the distribution. This is done as follows:

Family size:	1	2	3	4	5	6	7	8	9	Total
Frequency :	2	6	12	18	19	15	11	11	6	100
CF :	2	8	20	38	57	72	83	94	100	-

From the cumulative frequency column, we note that the 50½th position corresponds to the family with 5 members. The median is thus 5. The computation of the median becomes easier if n were odd.

For example if n=101, we would have looked for the 1/2(n+1)th=1/2 (101+1)th =51st position, which is still 5. Note that addition of one value does not alter the median. But this is not true in general.

3.4.4 Computing Median for Grouped Data

When the data have been arranged in a grouped frequency distribution, the identities of the individual values have been lost. As a result, we cannot determine the exact value of the median. It can be estimated, however, by (i) locating the class in which the median lies and then (ii) interpolating within this class to arrive at the median. The conventional assumption made in such interpolation is that the individual values within a class are evenly distributed over the interval of that class. The class which contains this median is called the **median class**. The median class is the class whose cumulative frequency is greater than or equal to $n/2$. The formula for computing median is then

$$\tilde{m} = l_m + \frac{h}{f_m}\left(\frac{n}{2} - F_{(m)-1}\right) \quad \ldots (3.16)$$

The symbols used in the above equation have the following meanings:

l_m= Lower limit of the median class
n = Total frequency=Σf_i
f_m= Frequency of the median class
$F_{(m)-1}$= Cumulative frequency of the pre- median class
h = Width of the median class.

3.4.5 Developing a Formula for Median

To develop an algebraic expression for median, refer to the ogive as shown in Figure 3.2 (see section 3.4.6).

We first identify the **median class** as the class containing cumulative frequency $n/2$. This is best understood by referring to the ogive. The median is the abscissa of point P on the ogive whose ordinate is 50% of the total frequency, i.e. $n/2$. Note that for the two similar triangles PQR and RST as shown in the figure referred to above.

$$\frac{RQ}{RS} = \frac{PQ}{ST}$$

or

$$\frac{RQ}{h} = \frac{n/2 - F_{(m)-1}}{F_m - F_{(m)-1}} = \frac{n/2 - F_{(m)-1}}{f_m}$$

where

$F_{(m)-1}$=Cumulative frequency of the pre-median class,
F_m=Cumulative frequency of the median class,

f_m =Frequency of the median class
h=Width of the median class.

Hence

$$RQ = \frac{h}{f_m}\left(\frac{n}{2} - F_{(m)-1}\right)$$

where RQ is the fraction of the distance one must traverse through the median class to reach $n/2$. Letting l_m represent the lower class boundary of the median class, we reach to an algebraic expression for the median of a group frequency distribution as follows:

$$\tilde{m} = l_m + RQ = l_m + \frac{h}{f_m}\left(\frac{n}{2} - F_{(m)-1}\right)$$

The use of the above formula is facilitated through identifying first the median class. The median class is identified from the cumulated frequency column on the basis of the value of $n/2$. The following steps are involved in computing median from grouped data:

- Compute the less than type cumulative frequencies
- Determine $n/2$, one-half of the total number of cases
- Locate the median class for which the cumulative frequency is more than $n/2$.
- Determine the lower limit of the median class. This is l_m
- Sum the frequencies of all classes prior to median class. This is $F_{(m)-1}$
- Determine the frequency of the median class. This is f_m.
- Determine the width of the median class. This is h.

You have now all the quantities to compute median. Put them in formula (3.16) above and complete your calculation[1]. The method is illustrated below with an example.

Example 3.17: Use the age data in Table 3.2 to compute the median age of the workers.

[1] If the distribution of data is such that there is a class interval for which the cumulative frequency is exactly equal to $n/2$, then the upper limit of that class would be the median and no interpolation would be needed.

Solution: For computational convenience we reproduce below the table under reference with an additional column of cumulative frequency.

Table for computing median for the data in Table 3.2

Age (in years)	Frequency (f_i)	Cumulative frequency
24.5–29.5	3	3
29.5–34.5	9	12
34.5–39.5	15	27
39.5–44.5	12	39
44.5–49.5	7	46
49.5–54.5	4	50
Total	50	–

Here n=50. Looking at the cumulative frequency column, we find that half-way of 50 (i.e. 25) corresponds to the class 34.5–39.5. This is the so called median class. To apply the formula (3.16), we locate the following values from the table:

$$l_m = 34.5, h = 5, f_m = 15 \text{ and } F_{(m)-1} = 12$$

so that on using (3.16)

$$\tilde{m} = 34.5 + \frac{5}{15}(25 - 12) = 38.83$$

For a distribution with more than type cumulative frequencies, the formula to be used for computing median is as follows:

$$\tilde{m} = u_0 - \frac{h}{f_m}\left(\frac{n}{2} - F_{(m)+1}\right) \quad \ldots (3.17)$$

where

u_0 =Upper limit of the median class

F_{m+1} = Cumulative frequency of the post- median class.

It is important to note that median is almost always computed from the less than type cumulated frequency distribution because of its simplicity.

To illustrate use of the formula (3.17), we reconstruct Table 3.2 below with de-cumulative frequencies to represent more than type ogive and hence to compute the median from this distribution.

Table for computing median for more than type distribution

Age (in years)	Frequency (f_i)	Decumulative frequency
24.5–29.5	3	50
29.5–34.5	9	47
34.5–39.5	15	38
39.5–44.5	12	23
44.5–49.5	7	11
49.5–54.5	4	4
Total	50	–

Applying (3.17)

$$\tilde{m} = 39.5 - \frac{5}{15}\left(\frac{50}{2} - 23\right) = 38.83$$

as ought to be.

It is important to note that the question of odd or even value of n in the determination of the median from a grouped frequency distribution in most cases is trivial. Its effect is so small that no adjustment is needed for computing median from even or odd value of n.

3.4.6 Locating Median Graphically

In addition to computing median using formula just described, one can make use of the graphs and diagrams to locate the median. Two such approaches can be named for this purpose. These are described below.

(a) Median from histogram

The method of approximating the median from grouped frequency distribution can be best understood by referring to a histogram. We illustrate how to locate median from histogram from data as shown in Example 3.17 above. The histogram constructed from this table is as shown in Figure 3.1 below:

The total area under this histogram is equal to the sum of the areas of the individual rectangles. Since the width of each rectangle is 5,

Total area = 5(height of the first rectangle) + 5(height of the second rectangle) +....+ 5 (height of the sixth rectangle)

$$= 5(3 + 9 + 15 + 12 + 7 + 4) = 5(50) = 250$$

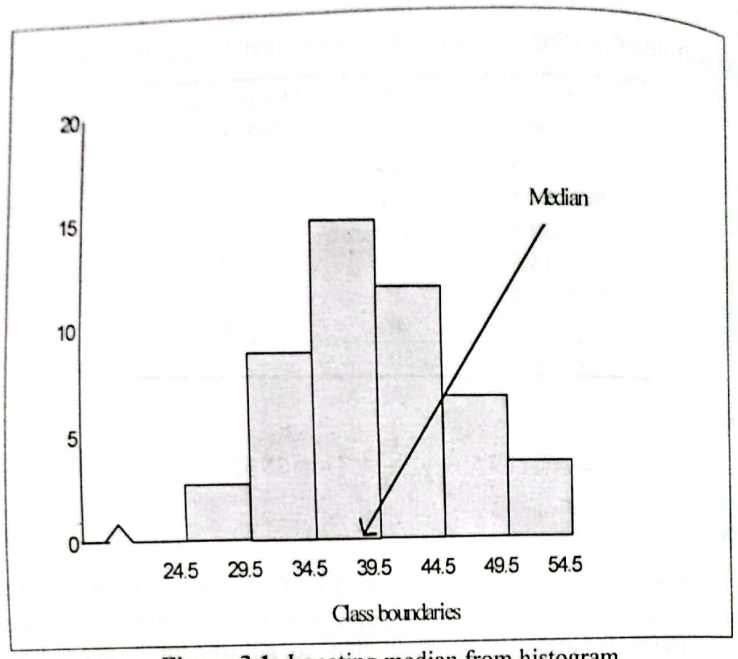


Figure 3.1: Locating median from histogram

[Precisely, a histogram should have been drawn with frequency densities rather than absolute frequencies so as to make the total area equal to the total frequency. Since heights of the rectangles in a histogram in that case are proportional to the frequencies, determination of median graphically with absolute frequency will not affect in any way the determination of median value.]

The median, by definition, is the value that divides the total area of the histogram into two equal parts. Since the total area is 250, the median will lie on the horizontal axis at a point where the area of the rectangles adds to (250)/2 = 125. Let us now locate the point where the area attains this value. The first two rectangles have a cumulated area of 5×3+5×9=60. This area corresponds to the lower limit (34.5) of the third rectangle. To arrive at an area of 125, we need an additional area of 65 (=125–60) to make up an area equal to 125. When the area of the third rectangle (5×15=75) is added, the cumulative area becomes 135, which exceeds the required area (125) by 10. This implies that the median lies somewhere in the range 34.5–39.5 of the third rectangle. Thus, if l_m is the lower limit of the third rectangle and x_0 is the point between the two limits 34.5 and 39.5 of the third rectangle, then the median $\tilde{M}$ is located at a distance of x_0 from

l_m. That is $\tilde{m} = l_m + x_0$. But $x_0 \times 15 = 65$, so that $x_0 = 4.33$. Hence $\tilde{m} = l_m + x_0 = 34.5 + 4.33 = 38.83$, which we obtained earlier directly using the usual formula.

(b) Median from ogive

To determine median graphically, draw a less than type ogive. Then draw a horizontal line through the point $n/2$ parallel to the x-axis. Draw another line perpendicular to the x-axis from the point of intersection of the line and the ogive. The point, at which the perpendicular cuts the x-axis, is approximately the median. We illustrate below in Figure 3.2 how the median can be located graphically from the age data in Table 3.2.

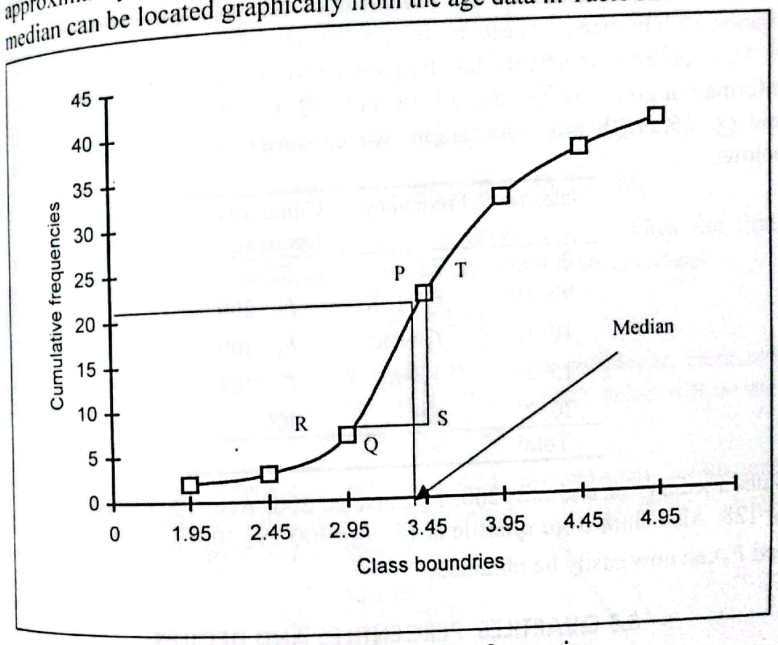


Figure 3.2: Locating median from ogive

As we can see from the graph, the median lies very close to the upper boundary of the interval 34.5–39.5, while the application of the formula (3.16) yields the median at 38.83, which falls in the interval indicated above.

When both the more than type and less than type ogives are drawn on the same scale and on the same graph paper, the median is the point at which

the perpendicular drawn from the intersection of the ogives, cuts the x-axis.

Example 3.18: The rate of sales tax as a percentage of the sales paid by 400 shopkeepers of a market during an assessment year ranged from 0 to 25%. The sales tax paid by 18% of them was not greater than 5%. The median rate of sales tax was 10% and the 3rd quartile rate of sales tax was 15%. If only 8% of the shopkeepers paid sales tax at a rate greater than 20% but not greater than 25%, summarize the information in the form of a frequency distribution taking intervals of 5%.

Solution: As described in the statement of the problem, there are five classes each of size 5. These are 0–5, 5–10, 10–15, 15–20 and 20–25. Let $f_1, f_2, \ldots, f_5$ be respectively the frequencies of these classes. The partial information given are $\sum f=400, f_1=.18\times400=72, f_5=.08\times400=32$, median=10 and $Q_3=15$. With this information, we construct the frequency table as below:

Sales tax	Frequency	Cumulative frequency
00–05	72	72
05–10	$f_2=128$	$F_2=200$
10–15	$f_3=100$	$F_3=300$
15–20	$f_4=68$	$F_4=368$
20–25	32	400
Total	400	–

Since median=10 and $n/2=200$, F_2 must be 200. Also $72+f_2=200$, so that $f_2=128$. Also since third quartile is 15, $F_3=300$. As $200+f_3=300, f_3=100$. f_4 and F_4 can now easily be obtained.

3.5 QUARTILES, PERCENTILES AND DECILES

A few other measures that are allied to the median include the **quartiles**, **deciles** and **percentiles.** These measures are based on their position in a series of observations. They are .not necessarily central values and hence they are referred to as **measures of location.** Collectively, they together are called **quantiles**, **fractiles** or **partition values**. We discuss these measures below.

3.5.1 The Quartiles

There are three quartiles in a data set, usually denoted by Q_1, Q_2 and Q_3, which divide the whole distribution into four equal parts. The second quartile Q_2, is identical with the median. The first quartile, Q_1, is the value at or below which one-fourth (25%) of all observations in the set fall; the third quartile, Q_3 is the value at or below which three-fourths (75%) of the observations lie.

For ungrouped data, a quartile, as does the median, either assumes the value of one of the items or falls between two values. If n is divisible by 4, the first quartile (Q_1) has the value half-way between the $n/4$th and $(n/4+1)$th observation. If n is not exactly divisible by 4 (i.e. $n/4$ is not an integer), the first quartile has the value of the next higher integer. To find the third quartile Q_3, we replace $n/4$ by $3n/4$.

Consider the following set of 12 values arranged in ascending order:

14, 17, 19, 23, 27, 32, 40, 49, 54, 59, 71, 80.

Here $n=12$, which is divisible by 4. The quotient is 3. Thus the first quartile will be the average value of the 3rd and the 4th observations:

$$Q_1=\frac{19+23}{2}=21$$

We add a new value 94 to the set so that $n/4$ is not an integer. Here $n/4$ =13/4= 3.25. The next higher integer is 4. Thus the 4th value will be the first quartile. Observe that the 4th value in the set is 23.

For $n=12$, the third quartile is the value mid-way between 3n/4th and $(3n/4+1)$th observation, since $3n/4=9$ is an integer. Thus Q_3 is the average of the 9th and 10th observations. That is

$$Q_3=\frac{54+59}{2}=56.5$$

If we add a new value 94 to the set, n becomes 13 and the third quartile now is the 10th observation, since $3n/4$ =39/4=9.75, which is not an integer. The next higher integer is 10. Thus Q_3 is the tenth value, which is equal to 59. That is $Q_3=59$.

For large data sets, the computation of quartile values simply by inspection of the data set is practically impossible. In such situations the quartile values can be calculated by first forming the cumulative frequency distribution and then locating the desired quartile from the given values

based on n values. The following is an example, which illustrates how to compute median from such ungrouped data.

Example 3.19: Distribution of 70 students according to their weight in kg is as follows:

Weight	No. of students	Cumulative frequencies
40	6	6
43	11	17
51	19	36
55	17	53
60	13	66
63	4	70
Total	70	-

To obtain Q_1 and Q_3, we cumulate the frequencies as shown above. Since $n/4=70/4=17.5$ is not an integer, the first quartile will be the 18th observation (next higher integer of the fraction 17.5), which in this case corresponds to 51. Since $3n/4 = 52.5$, the Q_3 is the 53rd value which equals 55.

With grouped data, the method of estimating the first and third quartile is similar to that of estimating the median. This can be accomplished through the following general formula:

$$Q_r = l_r + \frac{h}{f_r}\left(\frac{rn}{4} - F_{(r)-1}\right), \quad r = 1,2,3 \quad \ldots(3.18)$$

where

n = total number of observations in the distribution
h = Class width
$F_{(r-1)}$ = Cumulative frequency of the class prior to rth quartile class
f_r = Frequency of rth quartile class
l_r = Lower limit of the rth quartile class

Example 3.20: Compute the first quartile and third quartile from the data presented in Table 3.2.

Solution: For ready reference, we reproduce the table and construct an extra column (column 3) displaying the cumulative frequency to which the quartile values relate:

Age (in years)	Frequency (f_i)	Cumulative frequency
24.5–29.5	3	3
29.5–34.5	9	12
34.5–39.5	15	27
39.5–44.5	12	39
44.5–49.5	7	46
49.5–54.5	4	50
Total	50	–

Following (3.18) for r=1, the formula for computing Q_1 is

$$Q_1 = l_1 + \frac{h}{f_1}\left(\frac{n}{4} - F_{(1)-1}\right)$$

Here for r=1
$n/4=50/4=12.5, l_1 = 34.5,$, h=5, f_1=15 and $F_{(1)-1}$ =12

so that

$$Q_1 = 34.5 + \frac{5}{15}(12.5 - 12) = 34.67$$

To compute the third quartile Q_3, r=3 and the other values required are $3n/4 = 37.5$, $l_3 = 39.5$, $F_{(3)-1} = 27$, $f_3 = 12$, $h = 5$. Thus from (3.18)

$$Q_3 = l_3 + \frac{h}{f_3}\left(\frac{3n}{4} - F_{(3)-1}\right) = 39.5 + \frac{5}{12}(37.5 - 27) = 43.87$$

A value of 34.67 for Q_1 implies that 25 percent of the workers are below age 34.67. Similarly, there are 75 percent workers in the company who are below 43.87 years of age and only 25 percent of them are above this age as implied by the value of Q_3.

3.5.2 The Percentiles

Like quartiles, the statistical measure referred to as **percentile** offers a means for identifying the location of values in the data set that are not necessarily central values. Percentiles are the values, which divide the distribution into 100 equal parts. Thus there are 99 percentiles in a distribution, which are conventionally denoted by P_1, P_2, ...P_{99}.

Recall that in the discussion of the median, we found that the median divides the items arranged in order of magnitude into two equal parts. Thus in terms of percentiles, the median is the 50-th percentile. This means that $P_{50}=Q_2=\tilde{m}$. At times the 25-th percentile and/or the 75-th percentile may

be of particular interest. These two percentiles are in fact the first quartile (Q_1) and third quartile (Q_3) respectively, which we discussed earlier.

3.5.3 Percentiles for Ungrouped Data

With ungrouped data, the percentile either takes on the value half-way between the two observations or the value of one of the observations, depending on whether n is divisible by 100 or not. Consider the ordered observations

11, 14, 17, 23, 27, 32, 40, 49, 54, 59, 71, 80

To determine the 29th percentile, P_{29} say, we note that $(29\times12)/100 = 3.48$, which is not an integer. Thus the next higher integer 4 here will determine the 29th percentile value. On inspection, $P_{29} = 23$. Similarly P_{75} will be the average of the 9th and 10th observations, since $(75\times12)/100 = 9$, which is an integer. Thus the 75th percentile value is

$$P_{75} = \frac{1}{2}(9\text{th value} + 10\text{th value}) = \frac{1}{2}(54+59) = 56.5$$

If the percentile values are required for an ungrouped frequency distribution, the same procedure may be followed. Consider the distribution of Example 3.19. To compute 35th percentile, say, we obtain P_{35}. Here $rn/100 = (35\times\times70)/100 = 24.5$, which is not an integer. The next higher integer is 25, so that the value that corresponds to this integer is the 35th percentile. Looking at the cumulative frequency column, P_{35}=51. Based on this P-value, we can assert that 35% of the students scored 51 or less.

3.5.4 Percentile for Grouped Data

For a grouped frequency distribution, a formula similar to those determining the median and the quartiles may be used to determine the percentiles. In general, the ith percentile of a grouped distribution for n observations may be arrived at by using the following formula:

$$P_r = l_r + \frac{h}{f_r}\left(\frac{rn}{100} - F_{(r)-1}\right) \quad \ldots (3.19)$$

where

l_r = Lower limit of the rth percentile class

$F_{(r)-1}$ = Cumulative frequency of the pre-percentile class

f_r = Frequency of the rth percentile class

h= Width of the i-th percentile class interval

As an illustration, we compute the 30th percentile for the distribution presented in Table 3.2, which we reproduce here for computational convenience.

Age (in years)	Frequency (f_i)	Cumulative frequency
24.5–29.5	3	3
29.5–34.5	9	12
34.5–39.5	15	27
39.5–44.5	12	39
44.5–49.5	7	46
49.5–54.5	4	50
Total	50	–

The required percentile class is determined from $rn/100 = (30\times50)/100 = 15$. Looking at the cumulative frequency in the table, we find that this value falls in the range 34.5–39.5. The other required values are:

$$r = 30,\ l_{30} = 34.5,\ h = 5,\ f_{30} = 15 \text{ and } F_{(30)-1} = 12.$$

so that

$$P_{30} = 34.5 + \frac{5}{15}(15-12) = 35.5$$

This implies that of all the workers, 30% were under age 35.5 years.

3.5.5 Percentile Rank

The percentile rank of any score or observation is defined as the percentage of cases in a distribution that falls at or below that score. Percentile ranks are simple to calculate if the entire collection of raw scores is available. Consider the following collection of test scores as obtained by 20 applicants in a test arranged in ascending order.

19, 22, 25, 30, 38, 39, 41, 43, 44, 47, 48, 49, 51, **54**, 56, 59, 61, 65, 67, 70

An applicant who received a score of 54 might ask himself "how does his score rank him among all the students who took part in the test?" The answer is that he scored the same as or better than 70% of the entire group, indicating that his **percentile rank** is 70%, or in other words, 70% of the students have scored 54 or below. Note that the score 54 ranks 14th from the bottom of the 20 scores so that the percentile rank of the score (π_r) 54 is

$$\pi_r = \frac{14}{20} \times 100 = 70\%$$

Thus his percentile rank is fourteenth (i.e.70%) out of 20.

For grouped distribution, the percentile rank is simply the solution of the equation (3.19) for π_r:

$$\pi_r = \frac{F_{(r)-1} + f_r\left(\frac{P_r - l_r}{h}\right)}{n} \times 100 \quad \ldots (3.20)$$

Let us obtain the percentile rank for an age of 35.5 for the data in Table 3.2. Here P_{30}=35.5, falling in the range 34.5–39.5, $F_{(30)-1} = 12$, $f_{30} = 15$, $l_{30} = 34.5$, h=5 and n =50. Using (3.20)

$$\pi_r = \frac{12 + 15\left(\frac{35.5 - 34.5}{5}\right)}{50} \times 100 = 30\%$$

Hence the percentile rank is 30%. This implies that 30% of the workers are aged 35.5 years or below.

3.5.6 The Deciles

When a distribution is divided into ten equal parts, each division is called a **decile**. Thus, there are 9 deciles in a distribution, which are denoted by D_1, $D_2, \ldots D_9$. Obviously $D_5 = \tilde{m} = P_{50}$.

The method of determining the deciles is similar to that for median, percentiles and quartiles. To compute 6th decile (D_6), for example, for the distribution in Example 3.19, determine $rn/10 = (6n)/10$. For n= 70, this quantity is 42. This being an integer, the average of 42nd and 43rd values will be the 6th decile. By inspection of the distribution, D_6=55.

For grouped data, the formula for the rth decile is

$$D_r = l_r + \frac{h}{f_r}\left(\frac{rn}{10} - F_{(r)-1}\right) \quad \ldots (3.21)$$

where

l_r = Lower limit of the rth decile class

$F_{(r)-1}$ = Cumulative frequency of the class prior to the pre- rth decile class

f_r = Frequency of the rth decile class

h = Width of the decile class

Example 3.21: Obtain D_4 for the distribution given in Table 3.2 and interpret your result.

Solution: Here n=50, r=4, $4n/10$=20, so that $l_4 = 34.5$ and $F_{(4)-1} = 12$, so that

$$D_4 = 34.5 + \frac{5}{15}\left(\frac{4 \times 50}{10} - 12\right) = 37.17$$

This value for the fourth decile implies that approximately 40% workers are under 37.17 years.

3.6 THE MODE

A third statistical measure, the **mode**, is sometimes used as a measure of central tendency of a distribution. The distribution we refer to here may be composed of both categorical and numeric data. The mode is interpreted as the value that occurs most frequently in a distribution.

Definition **3.3**: *Mode is the most frequently occurring value in a set of observations. In other words, mode is the value of a variable, which occurs with the highest frequency.*

For nominal data, such as sex (male, female), or marital status (married, single, widowed, divorced) of an individual, it does not make any sense to ask for the mean (or median) sex or mean (or median) marital status unless these variables can be assigned meaningful numerical values. It does however, make sense to ask, which **category** has the **most** people. The term **most** in this case means the largest number of individuals. This means that we are looking for a category of the variable, which has the highest frequency. Such a category, if exists, is called **modal category** and is used to locate the mode in a set of observations or in a distribution. If a population consists of 87 percent Muslims, 11 percent Hindus and the remaining 2 percent are of other religions; the modal category is the **Muslim**, which has the most people. These examples thus justify why we do sometimes need to introduce the concept of **mode** as a measure of central tendency in addition to arithmetic mean and median.

Example 3.22: The responses of 120 athletes on their preferred color of track suits were as follows:

Preferred Color	Number of customers
White	25
Green	21
Blue	54
Yellow	12
Orange	6
Red	2
Total	120

Here the most preferred color is blue. Thus blue may be regarded as modal choice

For numerical data, the mode is similarly obtained. For the observations 7, 2, 7, 7, 1, 7, 9, 7, the mode is 7, since it occurs with the highest frequency (here 5). The location of modal value becomes much easier if the observed values appear in the form of a frequency distribution.

Example 3.23: In 1996–97 Bangladesh Demographic and Health Survey asked 9127 ever-married women about the ideal number of children a couple should have. The responses were as follows:

Ideal Number	Percent of women
1	1.9
2	59.5
3	20.8
4	10.5
5+	1.5
Non-numeric response	5.8
Total	100.0

The table shows that most women (59.5%) reported a choice of 2 children as most desirable. So 2 is the mode.

3.6.1 Locating Mode for Grouped Data

To determine the value of the mode for a grouped frequency distribution, it is necessary to identify the modal class, in which the mode is located. In general, a **modal class** is one with the largest frequency. For the age data, presented in Table 3.2, the modal class is the class of 34.5–39.5, since it contains more items (15) than any other classes. Once the modal class is

identified, the next step is to locate the mode within the class. The value of the mode is usually estimated by the method of interpolation. The interpolation formula is as follows:

$$M_0 = l_0 + h\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \quad \ldots (3.22)$$

where

l_0 = lower limit of the modal class

h = size of the modal class

Δ_1 = differences in the frequencies of modal and pre - modal classes

Δ_2 = differences in the frequencies of modal and post - modal classes

If f_0 is the frequency of the modal class, f_{-1} that of the pre-modal class and f_1 that of the post-modal class, then the above formula can be presented in the following form:

$$M_0 = l_0 + h\left(\frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1}\right) \quad \ldots (3.23)$$

3.6.2 Developing Formula (3.22) to Compute Mode

To develop the interpolation formula shown above, consider the three adjacent rectangles of the histogram presented in Figure 3.3 displaying the age data in Example 2.19 (Chapter 2). The central rectangle RI_0I_1S corresponds to the modal class. The class widths are assumed to be equal for all intervals. The mode is the abscissa M_0 of the point of intersection P of the constructed lines QS and RT. Evidently, l_0 and l_1 are the lower and upper class boundaries of the modal class, and Δ_1 and Δ_2 represent respectively the excess of the modal class frequency over the class frequencies of the pre-modal and post-modal classes.

From the similar triangles PQR and PST, we have

$$\frac{EP}{RQ} = \frac{PF}{ST}$$

or

$$\frac{M_0 - l_0}{\Delta_1} = \frac{l_1 - M_0}{\Delta_2},$$

Solving for M_0

$$M_0 = \frac{\Delta_1 l_1 + \Delta_2 l_0}{\Delta_1 + \Delta_2}$$

Since $l_1 = l_0 + h$, h being the width of the class interval, the expression above becomes

$$M_0 = l_0 + h\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$$

Example 3.24: For the age data appearing in Table 3.2, compute the modal age.

Solution: We have 34.5–39.5 as the modal class since the highest frequency is contained in it (see table below).

Age (in years)	Frequency (f_i)
24.5–29.5	3
29.5–34.5	9
34.5–39.5	15
39.5–44.5	12
44.5–49.5	7
49.5–54.5	4

Examining the table, we find

$$l_0 = 34.5,\ \Delta_1 = 15-9 = 6,\ \Delta_2 = 15-12 = 3 \text{ and } h = 5$$

These values, when substituted in (3.22), yields

$$Mo = 34.5 + 5\left(\frac{6}{6+3}\right) = 37.83.$$

Example 3.25: Compare the mean, median and mode with the data on payment delays in Example 2.17 (Chapter 2) using both raw and grouped data.

Solution: Summing the payment values and dividing the resulting sum by the total number of values, we obtain the arithmetic mean:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{22+29+\ldots+16+21}{65} = \frac{1177}{65} = 18.11$$

The stem and leaf plot clearly identifies that 34th value, which is 17, is the median. The plot also shows that 16 is the value in the set which occurs

with the highest frequency (i.e. 9). Hence the mode of the given values is 16.

Note that the grouped distribution was formed as follows:

Payment delay	Frequency (f_i)	Mid-value (x_i)	Product ($f_i\ x_i$)	Cumulative frequency
09.5–12.5	3	11	33	3
12.5–15.5	14	14	196	17
15.5–18.5	23	17	391	40
18.5–21.5	12	20	240	52
21.5–24.5	8	23	184	60
24.5–27.5	4	26	104	64
27.5–30.5	1	29	29	65
Total	65	–	1177	–

The arithmetic mean is

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1177}{65} = 18.11.$$

To calculate the median, we employ the cumulative total column to locate the middle-most value and use the formula:

$$\tilde{m} = l_m + \frac{h}{f_m}\left(\frac{n}{2} - F_{(m)-1}\right).$$

Here $l_m = 15.5,\ h = 3,\ f_m = 23,\ F_{(m)-1} = 17$, so that the median is

$$\tilde{m} = 15.5 + \frac{3}{23}\left(\frac{65}{2} - 17\right) = 17.52.$$

The mode is straightforward to compute. We use the following formula to accomplish this:

$$M_0 = l_0 + h\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$$

From the table above, $l_0 = 15.5,\ h = 3,\ \Delta_1 = 23-14 = 9,\ \Delta_2 = 23-12 = 11$, so that

$$M_0 = 15.5 + 3\left(\frac{9}{9+11}\right) = 16.85.$$

Note that the mean, median and the mode calculated from the raw data are pleasingly consistent with those calculated from the grouped data. One of the explanations is that the groupings of the values have been made with judicial choice of the class boundaries.

3.6.3 Locating Mode from Histogram

A suitably constructed histogram may be used to provide a basis for locating mode. In a histogram, the class with the highest frequency represents the modal class. The two adjacent rectangles on both sides of the highest rectangle represent the pre-modal and post-modal classes. From the point of intersection of two connecting lines from the two top ends of the adjacent rectangles on both sides across the highest rectangle, a perpendicular is drawn to the X-axis. Then the mode is the abscissa of the point at which the perpendicular cuts the X-axis. The method is illustrated in Figure 3.3 below by the histogram drawn on the age data of Example 2.19 (Chapter 2).

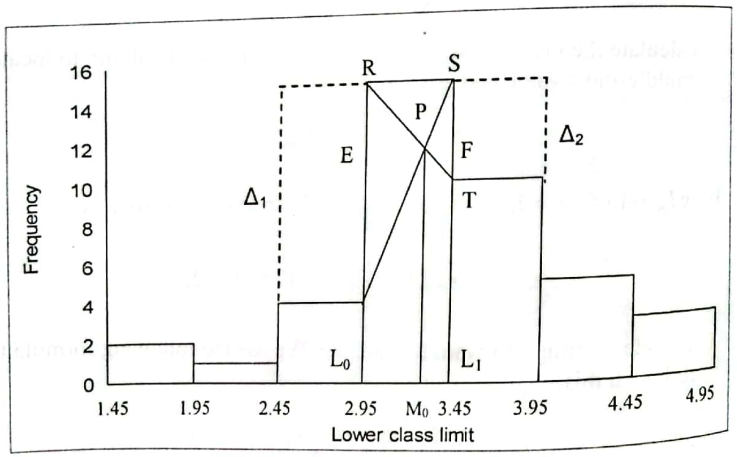


Figure 3.3: Locating mode from histogram

The perpendicular PM_0 cuts the X–axis at M_0. This is the point at which mode of the distribution is located. A perfect reading of this point will show that M_0=3.29. The accuracy of the estimate of mode graphically entirely depends on how accurately the histogram is drawn and how perfectly one can read the position of the mode from the graph.

3.6.4 An Empirical Relationship of Mode with Mean and Median

Unlike other measures of central tendency, a distribution may be **uni-modal** (having one mode), **bi-modal** (having two modes), **tri-modal** (having three modes)...or even **multi-modal** (having several modes). In instances, where a distribution possesses more than one mode, mode is said to be **ill-defined** and thus is difficult to interpret. In such cases, mode is not recommended as a measure of central tendency. However, one can attempt to compute the mode using the following formula, which is due to Karl Pearson and is based upon an **empirical relationship** between mean, median and mode of a moderately skewed distribution:

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean} \qquad \ldots (3.24)$$

When the distribution is symmetrical, the three measures of central tendency are of identical value i.e. for a symmetrical distribution

$$\text{Mean} = \text{Median} = \text{Mode} \qquad \ldots (3.25)$$

If the distribution is elongated to its right (positive) side (in which case the distribution is called **positively skewed**), we will find that

$$\text{Mode} < \text{Median} < \text{Mean} \qquad \ldots (3.26)$$

For a distribution, elongated to its left (in which case the distribution is called **negatively skewed**), mode is the largest of all the three measures, satisfying the relation

$$\text{Mode} > \text{Median} > \text{Mean} \qquad \ldots (3.27)$$

The relationships indicated in (3.25), (3.26) and (3.27) are illustrated in Figures (3.4a), (3.4b) and (3.4c) below:

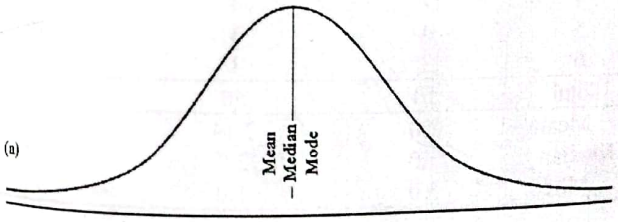


Figure 3.4a: Curve showing the symmetrical distribution

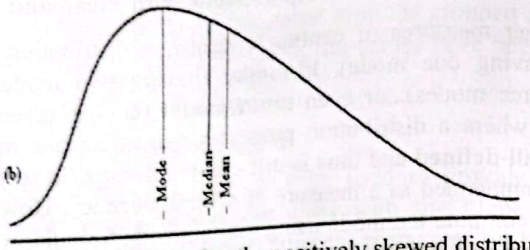


Figure 3.4b: Curve showing the positively skewed distribution

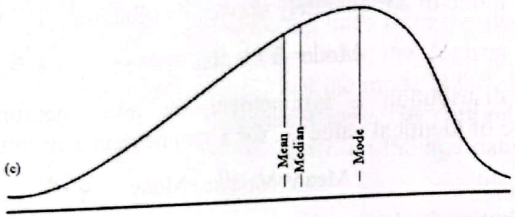
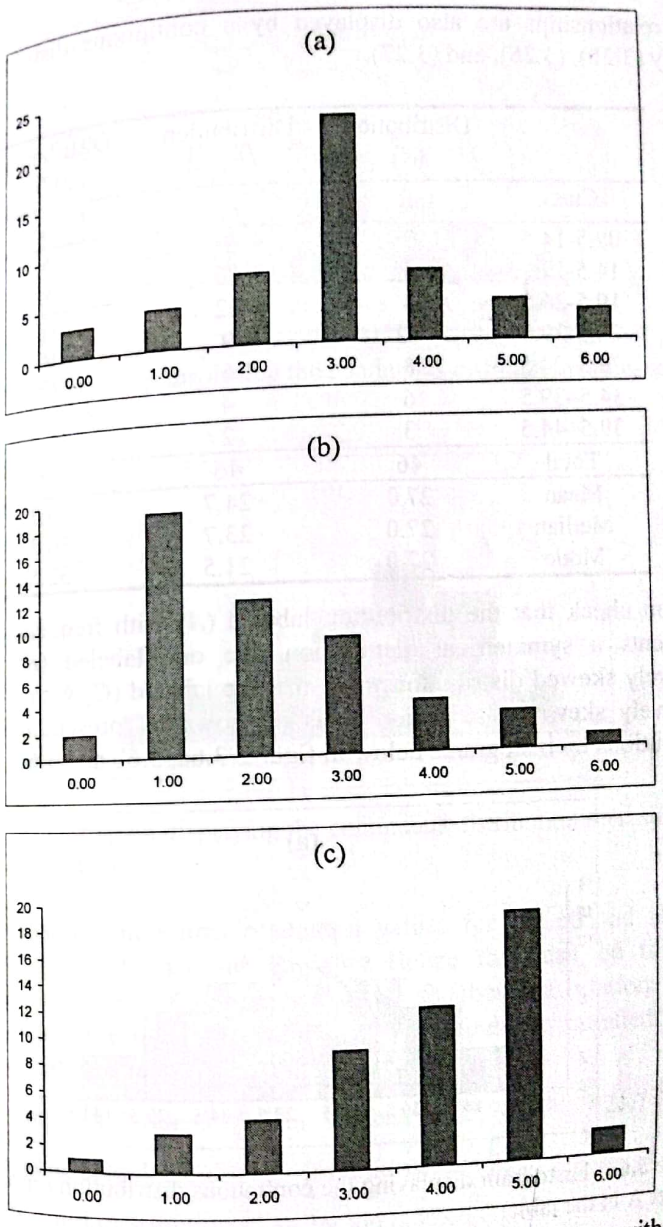


Figure 3.4c: Curve showing the negatively skewed distribution

These relationships are applicable for both discrete and continuous distributions. Let us illustrate this feature by an example with a discrete distribution first.

	Distribution *A*	Distribution *B*	Distribution *C*
Values	f_1	f_2	f_3
0	3	2	1
1	4	19	3
2	7	12	4
3	22	9	9
4	7	4	12
5	4	3	19
6	3	1	2
Total	50	50	50
Mean	3.0	2.14	3.86
Median	3.0	2.0	4.0
Mode	3.0	1.0	5.0

Note that the symmetrical (distribution A), the three measures of central average are the same and are equal to 3.0. The distribution B is positively skewed, thus satisfying (3.26), while the distribution C is negatively skewed and thus satisfies (3.27). These distributions are graphed in Figures below:



Figures: 3.5(a)–3.5(c): Different forms of discrete distributions with varying degrees of asymmetry.

The relationships are also displayed by a continuous distribution that satisfy (3.25), (3.26), and (3.27).

	Distribution (*A*)	Distribution (*B*)	Distribution (*C*)
Class	f_1	f_2	f_3
09.5-14.5	3	3	2
14.5-19.5	6	10	4
19.5-24.5	8	12	6
24.5-29.5	12	9	9
29.5–34.5	8	6	12
34.5–39.5	6	4	10
39.5–44.5	3	2	3
Total	46	46	46
Mean	27.0	24.7	29.3
Median	27.0	23.7	30.3
Mode	27.0	21.5	32.5

We can check that the distribution labeled (*A*) with frequency counts f_1 represents a symmetrical distribution, the one labeled (*B*) with f_2 a positively skewed distribution while the one labeled (*C*) with f_3 displays a negatively skewed distribution. This is shown by presenting the above distributions by histograms below in figures 3.6a, 3.6b and 3.6c.

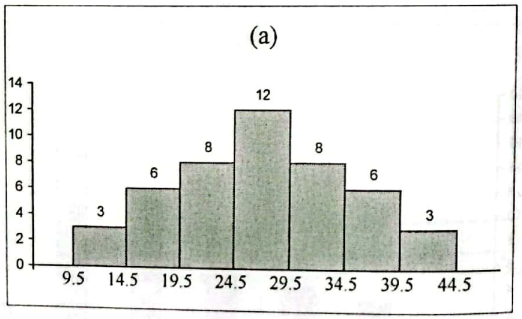


Figure 3.6a: Histogram displaying the continuous distribution representing the data set A in the table.

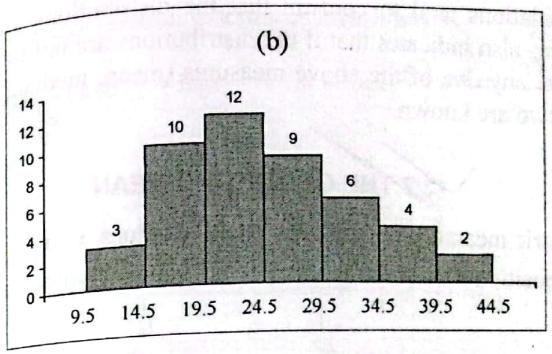


Figure 3.6b: Histogram displaying the continuous distribution representing the data set B in the table.

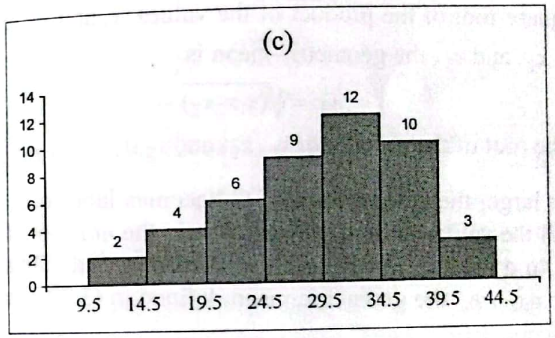


Figure 3.6c: Histogram displaying the continuous distribution representing the data set C in the table.

We note that the measures of central values for the second and third distribution do not vary substantially. Hence they can be treated as moderately skewed distributions. If this is so, then the relation between mean, median and mode as shown in (3.24) should be satisfied. For the distribution labeled *B*

$$\text{Empirical Mode}=3\,\text{Median}-2\,\text{Mean}=3\times23.7-2\times24.7=21.7$$

while the mode calculated directly from the data is 21.5.

And for the third distribution (distribution labeled C)

$$\text{Empirical Mode}=3\,\text{Median}-2\,\text{Mean}=3\times30.3-2\times29.3=32.3$$

while the mode calculated directly from C is 32.5.

These calculations tend to confirm that the distributions are moderately skewed. This also indicates that if the distributions are not too skewed, we can estimate any one of the above measures (mean, median, mode) if the remaining two are known.

3.7 THE GEOMETRIC MEAN

The geometric mean G of n non-zero positive values $x_1, x_2, ..., x_n$ is defined as the nth positive root of the product of the values. Symbolically,

$$G = \left(x_1 x_2 x_3 \ldots\ldots x_n\right)^{\frac{1}{n}} \quad \ldots(3.28)$$

For two values, say, x_1 and x_2, the geometric mean is

$$G = \sqrt{x_1 x_2}$$

i.e. the square root of the product of the values x_1 and x_2. Similarly for 3 values x_1, x_2, and x_3, the geometric mean is

$$G = \sqrt[3]{(x_1 x_2 x_3)}$$

i.e. the cube root of the product of x_1, x_2, and x_3.

When n is large, the computation of G becomes laborious, as we have to multiply all the values in the series and obtain the *nth* root. The task can be simplified to a great extent if the logarithms are used. Thus for n positive values $x_1, x_2, \ldots x_n$ the geometric mean defined in (3.28) can be expressed as follows:

$$\log G = \frac{1}{n}\log(x_1 x_2 \ldots x_n) = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) = \frac{1}{n}\sum \log x_i$$

Hence the logarithm of the geometric mean is the arithmetic mean of the logarithms of the values themselves.

To find G, we need to take anti-logarithm of log G. Thus

$$G = \text{Antilog}\,(\log G) = \text{Antilog}\left(\frac{\sum \log x_i}{n}\right) \quad \ldots(3.29)$$

For a frequency distribution, the geometric mean G can be expressed as follows:

$$G = \left(x_1^{f_1} x_2^{f_2} \ldots\ldots x_k^{f_k}\right)^{\frac{1}{n}} \quad \ldots(3.30)$$

Expressing the above relation in terms of logarithms

$$\log G = \frac{1}{n}\left(\log x_1^{f_1} + \log x_2^{f_2} + \ldots\ldots + \log x_k^{f_k}\right) = \frac{\sum f_i \log x_i}{n}$$

where $n = \sum f_i$.

As before

$$G = \text{Antilog}\left(\frac{\sum f_i \log x_i}{n}\right)$$

Example 3.26: Find the geometric mean of the values 47, 23, 71 and 81.

Solution: Using calculator

$$G^4 = 47 \times 23 \times 71 \times 81 = 6216831,$$

Hence

$$G = \sqrt[4]{6216831} = 49.93$$

Using logarithm

$$\log G = \tfrac{1}{4}(\log 6216831) = \tfrac{1}{4}(15.64277) = 3.9107$$

Taking antilog

$$G = 49.93$$

which is the same as obtained above.

You can also evaluate the geometric mean of the above values by adding the logarithmic values of each of the values. The computational procedure is as follows:

$$\text{Log } G = \tfrac{1}{4}(\log 47 + \log 23 + \log 71 + \log 81)$$

$$= \tfrac{1}{4}(3.8501 + 3.1355 + 4.2627 + 4.3944) = 3.9107$$

Taking antilog, G=49.92, which agrees exactly with the earlier results.

Geometric mean is used with numbers that tend to increase or decrease geometrically rather than arithmetically, that is, each number is the same multiple of the preceding number. It is typically used in averaging index numbers, rates of change, ratios, and other sets of values expressed in ratios or percentage forms. Suppose an economy with its initials size x_0 increases at a constant rate r. What would be the size of the economy at the end of fifth year? Let us examine the phenomenon below:

Let the size of the economy at the end of first year, second year, ..., fifth year be denoted by $x_1, x_2, \ldots, x_5$ respectively. Since the economy is growing

at a constant rate r, the absolute increase at the end of first year will be rx_0, so that its total size at the end of first year will be

$$x_1 = x_0 + rx_0 = x_0(1+r)$$

so that

$$1+r = \frac{x_1}{x_0} = Y_1 \text{ (say)}$$

At the end of second year, the absolute increase is rx_1 so that

$$x_2 = x_1 + rx_1 = x_1(1+r)$$

Hence

$$1+r = \frac{x_2}{x_0} = Y_2 \text{ (say)}$$

Similarly

$$1+r = \frac{x_3}{x_2} = Y_3, \quad 1+r = \frac{x_4}{x_3} = Y_4, \text{ and } 1+r = \frac{x_5}{x_0} = Y_5$$

Here $1+r$ is the growth factor with which the base year amount is accumulated to current year amount and r is the constant annual rate of changes in the economy. Multiplying the given ratios, we have

$$(1+r)^5 = Y_1 \times Y_2 \times Y_3 \times Y_4 \times Y_5$$

so that

$$G = \sqrt[5]{Y_1 \times Y_2 \times Y_3 \times Y_4 \times Y_5} \qquad \ldots (3.31)$$

This expression is the average of the ratios x_{i+1} / x_i, $i = 1, 2, \ldots, 5$ and is known as the **geometric mean** of the ratios. These ratios are assumed to be constant, a condition that must be satisfied for geometric mean to be applicable. But in reality, this condition is rarely satisfied. Nevertheless, in many applications, geometric mean seems to work well than arithmetic or other means.

If only two terminal values x_0 and x_5 are known and r is constant, a solution to $1+r$, henceforth we call it G (to stand for geometric mean) is

$$G = \sqrt[5]{\frac{x_5}{x_0}} \qquad \ldots (3.31a)$$

Thus to find the average change in rates for n-year period, we can use the following formula

$$G = \sqrt[n]{\frac{\text{Value at the end of period}}{\text{Value at the begining of period}}} = \sqrt[n]{\frac{x_n}{x_0}}$$

Since $G=1+r$,

$$r = \sqrt[5]{\frac{x_5}{x_0}} - 1 \qquad \ldots (3.31b)$$

and in general

$$r = \sqrt[n]{\frac{x_n}{x_0}} - 1 \qquad \ldots (3.31c)$$

This equation gives

$$x_n = x_0(1+r)^n \qquad \ldots (3.32)$$

If the increase is compounded at different rates $r_1, r_2, \ldots, r_n$, then the amount accumulated at the end of n-th year is

$$x_n = x_0(1+r_1)(1+r_2)\ldots(1+r_n) \qquad \ldots (3.33)$$

For the arithmetic change, we replace (3.31 c) by the following formula

$$A = \frac{x_n - x_0}{nx_0} = \frac{1}{n}\left(\frac{x_n}{x_0} - 1\right) \qquad \ldots (3.34)$$

Example 3.27: Suppose a piece of properties is purchased for Tk. 20 lac and sold 10 years later for Tk. 32 lac. What is the average annual return on the original investment?

Solution: Let r be the average annual rate of change. Employing the formula (3.32)

$$32 = 20(1+r)^{10}$$

from which

$$1+r = \sqrt[10]{\frac{32}{20}} = \sqrt[10]{1.6} = 1.048$$

so that

$$r = 0.048$$

or 4.8%. Thus the investment resulted in a mean return of 4.8 percent annually over the last 10 years. The corresponding arithmetic average change by (3.34) is

$$A=\frac{1}{10}\left(\frac{320000}{200000}-1\right)=0.06\ (\text{or } 6\%)$$

The geometric mean is unique and employs all the observations. It assigns less weight to extreme values. However it cannot be used when the data set contains 0 or one or more negative values. It is difficult to conceptualize also.

Example 3.28: An economy grows at the rate of 2% in the first year, 2.5% in the second year, 3% in the third year, 4% in the fourth year and 10% in the fifth year. What is the compound rate of growth of the economy for this five year period?

Solution: The accompanying table shows how with the given growth rates the original size (100) of the economy increases.

End of year (i)	Growth factor ($1+r_i$)	Size at the end (x_i)
1	1.02	$100(1.02)=102$
2	1.025	$102(1.025)=104.55$
3	1.03	$104.55(1.03)=107.69$
4	1.04	$107.69(1.04)=111.99$
5	1.10	$111.99(1.10)=123.19$

Here the accumulated size of the economy x_5, which grew at different rates, at the end of 5 years is

$$x_5=100(1+.02)(1+.025)(1+.03)(1+.04)(1+.10)=123.19$$

But

$$x_5=x_0(1+r)^5=100(1+r)^5$$

This gives

$$1+r=\sqrt[5]{\frac{123.19}{100}}=1.0426$$

Solving for r

$$r=1.0426-1=0.0426$$

The average annual rate of growth of the economy was thus 4.26% over the five year period.

Example 3.29: A car with the expected life of 3 years is depreciated at a uniform rate. The original cost of the car was Tk. 10 lac and it is estimated that at the end of 3 years, it will be worth 4 lac. Determine the constant rate of depreciation that reduces the value of the car to Tk. 4 lac.

Solution: Let r be the constant annual rate of depreciation so that by (3.31b)

$$10(1+r)^3=4$$

Solving

$$1+r=0.74$$

It follows that 26% of the preceding year-end value is the depreciation charge.

Example 3.30: The enumerated populations of Bangladesh at two census dates 1991 and 2001 were 106 and 124 millions. Find the annual rate of growth of this population if it is assumed that the population during this period increased geometrically, but not arithmetically. What would be the size of this population in 2011 if the population continues to grow at this rate?

Solution: We employ the formula (3.24)

$$r=\sqrt[n]{\frac{x_n}{x_0}}-1$$

Here $x_0=106, x_n=124, n=10,$ so that

$$r=\sqrt[10]{\frac{124}{106}}-1=0.015$$

or 1.5%. Therefore the population in the year 2011 with this rate of growth will be

$$x_{10}=x_0(1+r)^{10}=124(1+.015)^{10}=124(1.1605)=143.90 \text{ million.}$$

The arithmetic mean rate by (3.34) is

$$A=\frac{1}{10}\left(\frac{124}{106}-1\right)=0.017$$

so that

$$x_{10}=x_0(1+nA)=124(1+10\times.017)=145.08$$

3.8 THE HARMONIC MEAN

Another measure of central tendency, which is sometimes used, is the **harmonic mean**, commonly designated H. It is defined as the reciprocal of the arithmetic mean of the reciprocals of the individual values. For a set of n values $x_1, x_2, \ldots\ldots, x_n$ the harmonic mean H is obtained as follows:

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots+\frac{1}{x_n}} = \frac{n}{\sum(1/x_i)} \quad \ldots (3.35)$$

In a frequency distribution, if $x_1, x_2, \ldots, x_k$ represent k distinct values of a variable with corresponding frequencies $f_1, f_2, \ldots, f_k$, then

$$H = \frac{f_1+f_2+\ldots.+f_k}{f_1/x_1+f_2/x_2+\ldots+f_k/x_k} = \frac{\sum f_i}{\sum(f_i/x_i)}$$

To make the harmonic mean analogous to the arithmetic mean, the harmonic mean is sometimes expressed in its reciprocal form as follows:

$$\frac{1}{H} = \frac{\sum(f_i/x_i)}{n} \qquad (n=\sum f_i) \qquad \ldots (3.36)$$

which, when inverted, yields H.

This shows that the reciprocal of the harmonic mean is the arithmetic mean of the reciprocals of the values $x_1, x_2, \ldots\ldots, x_k$.

When $f_1, f_2, \ldots, f_k$ are replaced respectively by a set of weights $w_1, w_2, \ldots, w_k$, then we arrive at what we call **weighted harmonic mean** H_w:

$$H_w = \frac{\sum w_i}{\sum(w_i/x_i)} \qquad \ldots (3.37)$$

In general, the harmonic mean is used when rates are expressed as 'x' per y, and x is constant, such as miles per hour, income per household, production per acre etc. If y is constant, the arithmetic mean is used. For example, if the average cost per unit of a product is required, but the data are expressed as 'so many units' of product for certain amount of cost, the harmonic mean should be used. The rule is illustrated below.

Suppose we have n ratios, which we want to average. Define the ith ratio r_i as follows:

$$r_i = \frac{x_i}{y_i}, \quad i=1, 2, \ldots\ldots\ldots, n$$

Average of the ratio (R) is

$$R = \frac{\text{Total amount of } x}{\text{Total amount of } y} = \frac{\sum x_i}{\sum y_i}$$

If the units x_1 are fixed and equal to k (say),

$$r_i = \frac{k}{y_i}$$

and hence

$$y_i = \frac{k}{r_i}$$

Thus

$$R = \frac{\sum x_i}{\sum y_i} = \frac{\sum k}{\sum y_i} = \frac{nk}{k\sum(1/r_i)} = \frac{n}{\sum(1/r_i)}$$

which is the harmonic mean of the ratios r's.

Now we assume that the y_i's are constant and are equal to c. Then

$$r_i = \frac{x_i}{c} \quad \text{or} \quad x_i = r_i c$$

and hence

$$R = \frac{\sum x_i}{\sum y_i} = \frac{\sum r_i c}{nc} = \frac{\sum r_i}{n}$$

which is the simple arithmetic mean of the same ratios.

Example 3.31: An automobile driver travels from plain to hill station 100-km distance at a speed of 20 km per hour and then a makes his return journey at a speed of 8 km per hour. What is the average speed with which he traveled the entire distance of 200 km?

Solution: Here harmonic mean is appropriate because the distance traveled is constant (100 km, 100 km). Hence the required average is

$$H = \frac{2}{\frac{1}{20}+\frac{1}{8}} = 11.43$$

That this is the correct mean is verified as follows:

$$\text{Average speed} = \frac{\text{Total distance traversed}}{\text{Total time taken}}$$

The distance from plain to hill station is 100 km and the distance from hill station is also 100 km so that total distance covered is 100+100=200 km. Since the driver traveled from plain to hill station at a speed of 20 km, the time required is 100/20 hours. In his return journey, he took 100/8 hours so that the total time required for the entire journey is

$$\left(\frac{100}{20}+\frac{100}{8}\right) \text{hours.}$$

Hence

$$\text{Average speed} = \frac{200}{\frac{100}{20}+\frac{100}{8}} = \frac{2}{\frac{1}{20}+\frac{1}{8}} = H$$

Example 3.32: Suppose that a person spent Tk. 100 in each of the three fruit shops. In the first shop, he bought oranges at 4 taka a piece; in the second shop, he bought oranges at 5 taka a piece; and in the third shop, he bought orange at 10 taka each. What is the average price he paid per piece of orange?

Solution: The data are expressed as 'so many pieces of orange in 100 taka' while we wish to know the amount of money paid per piece of orange. If we compute the arithmetic mean

$$A = \frac{4+5+10}{3} = 6.33$$

we will be in error because more oranges were bought at 4 taka than at 5 taka and more were bought at 5 taka than at 10 taka. In other words, a weighted average is required. Since the person could buy 25 oranges from the first shop, 20 from the second shop and 10 from the third shop given the amount of money to spend, the weighted average is

$$WA = \frac{25\times4+20\times5+10\times10}{25+20+10} = 5.45$$

We check that the harmonic mean provides the same value:

$$\frac{1}{H} = \frac{1}{3}\left(\frac{1}{4}+\frac{1}{5}+\frac{1}{10}\right) = \frac{11}{60}$$

so that

$$H = \frac{60}{11} = 5.45$$

Example 3.33: Three families A, B, and C have equal monthly expenditure. The per capita expenditure (expenditure per person) of these families are Tk. 4,000, Tk. 5,000 and Tk. 10,000 respectively. Calculate the per capita expenditure of these 3 families together.

Solution: This is a problem that fits the harmonic mean, since the 3 families have constant expenditure and we are interested in computing per head expenditure. Thus

$$\frac{1}{H} = \frac{1}{3}\left(\frac{1}{4,000}+\frac{1}{5,000}+\frac{1}{10,000}\right) = \frac{0.00055}{3}$$

which gives

$$H = 5454$$

The arithmetic mean is

$$\text{AM} = \frac{1}{3}(4,000+5,000+10,000) = 6333$$

To verify that the harmonic mean is the correct average, let $\bar{X}$ be the average per capita expenditure. Then

$$\bar{X} = \frac{\text{Total expenditures}}{\text{Total number of persons}}$$

Let us arbitrarily choose that each family spends taka 20,000. If w_1, w_2, w_3 are the number of persons in A, B, and C respectively, then

$$w_1 = \frac{20000}{4000} = 5,\ w_2 = \frac{20000}{5000} = 4,\ w = \frac{20000}{10000} = 2$$

Hence

$$\bar{X} = \frac{20000+20000+20000}{5+4+2} = \frac{60000}{11} = 5454$$

This verifies that harmonic mean is the appropriate average in this instance.

Example 3.34: Find the arithmetic mean and harmonic mean of the first n natural numbers if the numbers are weighted by their respective sizes.

According to the problem statement, we require to compute the mean of the following frequency distribution:

x	f	fx
1	1	1^2
2	2	2^2
..	..	.
.	.	.
		.
n	n	n^2

Hence the arithmetic mean is

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1^2 + 2^2 + \ldots + n^2}{1 + 2 + \ldots + n} = \frac{\frac{n(n+1)(2n+1)}{6}}{\frac{n(n+1)}{2}} = \frac{2n+1}{3}$$

The harmonic mean is

$$H = \frac{\sum f}{\sum \frac{f}{x}} = \frac{1 + 2 + \ldots + n}{\frac{1}{1} + \frac{2}{2} + \ldots + \frac{n}{n}} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$$

3.9 OTHER MEASURES OF AVERAGE

There are several other measures of central values of data, which are rarely used. Among them, are quadratic mean, trimean and trimmed mean. We provide below a brief account of these measures.

3.9.1 Quadratic Mean

The quadratic mean, also called **root mean square** of a set of n values $x_1, x_2, \ldots, x_n$ is defined as

$$QM = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}} = \sqrt{\frac{\sum x_i^2}{n}} \qquad \ldots (3.38)$$

For a set of 5 numbers 1, 3, 4, 5 and 7,

$$QM = \sqrt{\frac{1^2 + 3^2 + 4^2 + 5^2 + 7^2}{5}} = \sqrt{\frac{100}{5}} = 4.47$$

Check that for these values, the arithmetic mean, geometric mean and the harmonic mean are respectively 4, 3.3, and 2.6. This mean is seen to have applications in physical and engineering sciences.

3.9.2 Trimmed Mean

As the name suggests, the trimmed mean is computed by 'trimming away' a certain percent of both the largest and smallest set of values. By the very nature of the measure, a trimmed mean is not a unique one for a particular set of data. Thus we can compute a 5% or a10% trimmed mean for the same set of data. A 10% trimmed mean is obtained by eliminating the largest 10% and smallest 10% values and computing the average of the remaining values. This helps to alleviate the distortion caused by extreme values, from which ordinary arithmetic mean suffers.

For a set of 10 observations, 32, 37, 47, 43, 36, 42, 38, 43, 86, and 26 the 10% trimmed mean will be based on 8 observations eliminating the largest value 86 and the smallest value 26. The resulting trimmed mean TM is

$$TM = \frac{32 + 37 + 47 + 43 + 36 + 42 + 38 + 43}{8} = \frac{318}{8} = 39.75$$

The arithmetic mean and the median of the set are respectively, 39.9 and 39.5 which are close to the trimmed mean. The trimmed mean approach is, of course, more insensitive to **outliers** (too large or too small values in the data set), than the mean but not as insensitive as the median. On the other hand, the trimmed mean approach makes use of more information in the data set. The trimmed mean approach is best understood in relation to exploratory data analysis (EDA) to be discussed in the following chapter.

3.9.3 Trimean

The partition values (quartiles, deciles and percentiles) discussed earlier, are some useful measures of location of a distribution. The three quartiles, Q_1, Q_2 and Q_3 in particular, when combined, give rise to a measure of location known as **trimean**. This is defined as follows:

$$\text{Trimean} = \frac{Q_1 + 2Q_2 + Q_3}{4} \qquad \ldots (3.39)$$

The concept of trimean is closely related to the EDA and the first and third quartiles are the approximations of the two hinges (see section 4.13, Chapter 4, for definition of hinges).

Let us check with the data in Example 3.2 how does the trimean compare with the arithmetic mean. For the data, the arithmetic mean, Q_1, Q_2 and Q_3 are respectively 39.3, 34.67, 38.83, and 43.87. With these values, the trimean works out to 39.05:

$$\text{Trimean} = \frac{34.67 + 2(38.83) + 43.87}{4} = 39.05 .$$

This closely agrees with the arithmetic mean 39.3.

3.10 COMPARING THE AVERAGES

The arithmetic mean, median and mode are regarded as the most important averages because of their simplicity and usefulness. However, as discussed before, the three averages are not equally applicable to any conditions. The choice will depend on the type of data being considered and on what is desired of a representative value. In making our choice of an appropriate average, we should, however, adhere to the following characteristics:

- An average should be easy to compute
- An average should be easily understandable.
- An average should be amenable to further algebraic manipulation.
- An average should be based on all the observations.
- An average should be so defined that it has one and only one interpretation.
- An average should not have much sampling variability.
- An average should remain unaffected by the presence of extreme values.

We compare below the averages in the light of the above criteria

Mean: Mean is the simplest of all measures of central tendency. It is easy to compute but applicable to only quantitative (interval level) data. It is based on all the observations. Algebraic treatment of arithmetic mean is possible in the sense that when several group means are available, a combined mean can be computed by applying simple arithmetic.

As arithmetic mean is a reflection of the total quantity represented by the values, it is highly influenced by extreme observations (too small or too big). When the distribution has open ends[1], mean cannot be computed meaningfully. For nominal data, the mean can never be computed.

[1] An open-end distribution is one that has one or both-end classes unspecified. Income less than Tk.250 (<250) or income more than Tk.5000 are examples of open-end classes and distributions with such classifications are open-end distributions.

Median: In contrast to mean, median is not influenced by the presence of extreme values in the data set. One of the major advantages of the median over the mean is that the median can be easily determined for an open-end distribution.

Like mean, median cannot be computed for nominal data, since ranking or ordering of the observations is not possible. Since it is a positional measure, its value is not based on all the observations. It is not also possible to compute a pooled or combined mean when median of several groups of data are desired.

Mode: When speed is the most important consideration, the mode would be the preferred measure of central tendency. Most importantly, mode can be used with any of the level of measurements and is the only measure of central tendency, which can be used meaningfully with nominal data.

The use of mode, however, is not recommended for ordinal level data, since it ignores much of the available information. Like median, computation of mode does not pause any problem for open-end distribution, while it is not possible in the case of mean.

Despite its several advantages, the mode estimated from a highly skewed distribution is too close to one end of the distribution, and thus fails to be a good representative of the data set. It is not also possible to calculate combined mode when two or more group modes are available.

Apart from the above drawbacks, the location of the modal class, and consequently the value of the mode depends on the ways in which the data are classified. The position of the mode varies with the selected length of the class interval. Thus, while the mode may be a useful measure of central tendency in some instances, it remains to be the most volatile among all the averages.

Perhaps the most important distinction between the median and the mean is that in the problems of inference (estimation, prediction, etc.), the median is generally less reliable than the mean. In other words, the median is generally subject to greater enhanced fluctuation than the mean; that is, it is apt to vary more from sample to sample.

Geometric mean and harmonic mean: Other measures of central tendency, less frequently used, include the **geometric mean** and the **harmonic mean**. Geometric mean is useful in averaging ratios and percentages. It gives less weight to large values and more to small values

than does the arithmetic mean. It is because of this reason it never exceeds the arithmetic mean. Geometric mean cannot be calculated when one or more of the values are zero or negative.

Harmonic mean is also rarely used. In problem related to time and rates, harmonic mean provides better results than any other measure of average. It gives the largest weight to the smallest observation and hence is not suitable for economic data. Harmonic mean cannot be calculated when one or more of the observations are zero.

Generally, the average we should use depends on the pattern of the distribution of the data and the purpose for which it is intended. If the distribution is symmetric, or nearly so, it does not make much difference which measure is used. If it is skewed in either direction, it may be more appropriate to use the mode or the median since the mean is not a typical value under the circumstance.

We will now state and prove one important relationship that connects arithmetic mean with geometric mean and harmonic mean.

Theorem 3.2: *For a set of n non-zero positive values* $x_1, x_2, \ldots\ldots, x_n$, *prove that* $A \geq G \geq H$, *where A, G, and H are respectively the arithmetic mean, geometric mean, and harmonic mean.*

Proof: By definition

$$A = \frac{\sum x_i}{n}, \quad G = (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

Taking log of G

$$\log G = \frac{1}{n}\sum \log x_i$$

$$= \frac{1}{n}\sum \log\left(\frac{\sum x_i}{n} - \frac{\sum x_i}{n} + x_i\right)$$

$$= \frac{1}{n}\sum \log(A - A + x_i)$$

$$= \frac{1}{n}\sum \log A\left(1 + \frac{x_i - A}{A}\right)$$

$$= \frac{1}{n}\sum \log A + \frac{1}{n}\sum \log\left(1 + \frac{x_i - A}{A}\right)$$

$$= \log A + \frac{1}{n}\sum\left\{\frac{x_i - A}{A} - \frac{1}{2}\left(\frac{x_i - A}{A}\right)^2 + \frac{1}{3}\left(\frac{x_i - A}{A}\right)^3\right\} + \ldots.$$

The expansion is valid only when $\left|\frac{x_i - A}{A}\right| \leq 1$. Under this condition, all third and higher order values of this expansion will be negligible and consequently

$$\log G = \log A + \frac{1}{n}\sum\left(\frac{x_i - A}{A}\right) - \frac{1}{2n}\sum\left(\frac{x_i - A}{A}\right)^2$$

The second term on the right hand side of the equation being zero,

$$\log G = \log A - \frac{1}{2n}\sum\left(\frac{x_i - A}{A}\right)^2$$

$$= \log A - \text{a positive quantity less than } \log A$$

Hence

$$A \geq G \qquad \ldots (a)$$

Using the relation $A \geq G$, we write

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \geq (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

Replacing $x_1, x_2, \ldots, x_n$ by $1/x_1, 1/x_2, \ldots, 1/x_n$ respectively

$$\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right) \geq \left(\frac{1}{x_1} \times \frac{1}{x_2} \times \ldots \times \frac{1}{x_n}\right)^{\frac{1}{n}}$$

That is

$$\frac{1}{H} \geq \frac{1}{G}$$

or

$$G \geq H \qquad \ldots (b)$$

Combining (a) and (b), $A \geq G \geq H$.

To establish the above inequality for two, three or four values, we proceed as follows:

When n=2

Let the values be x_1 and x_2 so that

$$A=\frac{x_1+x_2}{2} \text{ and } G=\sqrt{x_1x_2}$$

Taking the difference of these two means

$$A-G=\frac{x_1+x_2}{2}-\sqrt{x_1x_2}$$
$$=\frac{1}{2}\left(x_1+x_2-2\sqrt{x_1x_2}\right)$$
$$=\frac{1}{2}\left(\sqrt{x_1}-\sqrt{x_2}\right)^2$$

which is always positive so that $A-G\geq 0$.

Hence

$$A\geq G \qquad \ldots (c)$$

Again considering the difference of G and H, we have

$$G-H=\sqrt{x_1x_2}-\frac{2x_1x_2}{x_1+x_2}$$
$$=\frac{\sqrt{x_1x_2}}{x_1+x_2}\left\{(x_1+x_2)-2\sqrt{x_1x_2}\right\}$$
$$=\frac{\sqrt{x_1x_2}}{x_1+x_2}\left(\sqrt{x_1}-\sqrt{x_2}\right)^2, \text{ which is always}$$

positive. That is $G-H\geq 0$ .Hence

$$G\geq H. \qquad \ldots (d)$$

Hence

$$A\geq G\geq H.$$

When n=3

For three values x_1, x_2 and x_3, we consider the difference A–G as before.

$$A-G=\frac{x_1+x_2+x_3}{3}-\sqrt[3]{x_1x_2x_3}=\frac{1}{3}\left(x_1+x_2+x_3-3\sqrt[3]{x_1x_2x_3}\right)$$
$$=\frac{1}{3}\left\{\left(\sqrt[3]{x_1}\right)^3+\left(\sqrt[3]{x_2}\right)^3+\left(\sqrt[3]{x_3}\right)^3-3\sqrt[3]{x_1}\sqrt[3]{x_2}\sqrt[3]{x_2}\right\}$$

$$=\frac{1}{3}\left(\sqrt[3]{x_1}+\sqrt[3]{x_2}+\sqrt[3]{x_3}\right)$$
$$\times\frac{1}{2}\left\{\left(\sqrt[3]{x_1}-\sqrt[3]{x_2}\right)^2+\left(\sqrt[3]{x_2}-\sqrt[3]{x_3}\right)^2+\left(\sqrt[3]{x_3}-\sqrt[3]{x_1}\right)^2\right\}$$

which is always positive. That is $A-G\geq 0$.

Hence

$$A\geq G$$

Again

$$G-H=\sqrt[3]{x_1x_2x_3}-\frac{3x_1x_2x_3}{x_1x_2+x_2x_3+x_3x_1}$$
$$=\frac{\sqrt[3]{x_1x_2x_3}}{x_1x_2+x_2x_3+x_3x_1}\left\{x_1x_2+x_2x_3+x_3x_1\right\}$$
$$-3\sqrt[3]{(x_1x_2)(x_2x_3)(x_3x_1)}$$
$$=\frac{\sqrt[3]{x_1x_2x_3}}{x_1x_2+x_2x_3+x_3x_1}$$
$$\times\left\{\left(\sqrt[3]{x_1x_2}\right)^3+\left(\sqrt[3]{x_2x_3}\right)^3+\left(\sqrt[3]{x_3x_1}\right)^3-3\sqrt{x_1^2x_2^2x_3^2}\right\}$$
$$=\frac{\sqrt[3]{x_1x_2x_3}}{x_1x_2+x_2x_3+x_3x_1}$$
$$\times\frac{1}{2}\left\{\left(\sqrt[3]{x_1x_2}-\sqrt[3]{x_2x_3}\right)^2+\left(\sqrt[3]{x_2x_3}-\sqrt[3]{x_3x_1}\right)^2+\left(\sqrt[3]{x_3x_1}-\sqrt[3]{x_1x_2}\right)^2\right\}$$
$$\times\left(\sqrt[3]{x_1x_2}+\sqrt[3]{x_2x_3}+\sqrt[3]{x_3x_1}\right), \text{ which is positive.}$$

Thus gives $G-H\geq 0$ and hence

$$G\geq H.$$

Hence

$$A\geq G\geq H.$$

When n=4

To establish the result for 4 numbers x_1, x_2, x_3 and x_4, let us define

$$a=\frac{x_1+x_2}{2}, \quad b=\frac{x_3+x_4}{2}$$

Using the case for n=2,

$$\frac{a+b}{2} \geq \sqrt{ab}$$

or

$$\frac{x_1+x_2+x_3+x_4}{4} \geq \left(\frac{x_1+x_2}{2}\right)^{\frac{1}{2}}\left(\frac{x_3+x_4}{2}\right)^{\frac{1}{2}} \quad \ldots \text{(e)}$$

Using the inequality A≥G for two values

$$\left(\frac{x_1+x_2}{2}\right)^{\frac{1}{2}} \geq (x_1x_2)^{\frac{1}{4}} \quad \ldots\text{(f)}$$

and

$$\left(\frac{x_3+x_4}{2}\right)^{\frac{1}{2}} \geq (x_3\,x_4)^{\frac{1}{4}} \quad \ldots \text{(g)}$$

Multiplying (f) and (g) and combining the result with (e)

$$\frac{x_1+x_2+x_3+x_4}{4} \geq (x_1\,x_2\,x_3\,x_4)^{\frac{1}{4}} \quad \ldots \text{(h)}$$

or

$$A \geq G \quad \ldots \text{(i)}$$

Now putting $x_1 = \frac{1}{y_1},\ x_2 = \frac{1}{y_2},\ x_3 = \frac{1}{y_3},\ x_4 = \frac{1}{y_4}$ in (h)

$$\frac{1}{4}\left(\frac{1}{y_1}+\frac{1}{y_2}+\frac{1}{y_3}+\frac{1}{y_4}\right) \geq \left(\frac{1}{y_1y_2y_3y_4}\right)^{\frac{1}{4}}$$

from which $\frac{1}{H} \geq \frac{1}{G}$ and hence $G \geq H$

It follows that

$$A \geq G \geq H.$$

***Theorem* 3.3:** *If $G_1, G_2, \ldots, G_k$ are the geometric means of k sets of observations, then prove that the geometric mean G of the combined set is given by*

$$\log G = \frac{\sum n_i \log G_i}{\sum n_i}$$

where $n_1, n_2, \ldots n_k$ are the number of observations in set 1, set 2,....set k respectively.

Proof: Let the product of the n_1 observations in the first set be denoted by x_1, that of the n_2 observations in the second set be denoted by x_2,.....and that of the n_k observations of the kth set be denoted by x_k. Then by definition

$$G_1 = (x_1)^{\frac{1}{n_1}},\ G_2 = (x_2)^{\frac{1}{n_2}}, \ldots\ldots, G_k = (x_k)^{\frac{1}{n_k}}.$$

Obviously

$$x_1 = G_1^{n_1},\ x_2 = G_2^{n_2}, \ldots\ldots, x_k = G_k^{n_k}$$

The combined set now consists of the observations $x_1, x_2, \ldots, x_k$, the number of observations being $n_1+n_2+\ldots+n_k = \sum n_i = N$. Hence the geometric mean G of the combined set is

$$G = (x_1x_2\ldots\ldots\ldots x_k)^{\frac{1}{N}}$$

Taking logarithm

$$\log G = \frac{1}{N}\log(x_1x_2\ldots x_k)$$
$$= \frac{1}{N}(\log x_1 + \log x_2 + \ldots + \log x_k)$$
$$= \frac{1}{N}\left(\log G_1^{n_1} + \log G_2^{n_2} + \ldots + \log G_k^{n_k}\right)$$
$$= \frac{\sum n_i \log G_i}{\sum n_i}$$

For two values, the above result is expressed as follows:

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

If $n_1 = n_2 = \ldots = n_k = n$, then the expression above turns out to be

$$\log G = \frac{\sum \log G_i}{k}$$

or that

$$k \log G = \sum \log G_i = \log(G_1G_2\ldots G_k)$$

Hence $$G = (G_1 G_2 \ldots . G_k)^{\frac{1}{k}}$$

showing that the geometric mean of the combined set with equal number of observations in each set, is equal to the geometric mean of their individual geometric mean.

***Corollary* 3.1:** *If the variable x takes on n values $x_1, x_2, \ldots, x_n$ and $y=cx$, then the geometric mean of y is $G_y = cG_x$, where G_x and G_y are respectively the geometric mean of x and geometric mean of y. We prove this result as follows:*

The geometric mean of $x_1, x_2, \ldots, x_n$ is

$$G_x = (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

We have further $y_1 = cx_1,\ y_2 = cx_2,\ \ldots .\ y_n = cx_n$, so that

$$G_y = (y_1 y_2 \ldots . y_n)^{\frac{1}{n}} = (cx_1 \times cx_2 \times \ldots \times cx_n)^{\frac{1}{n}}$$

$$= c(x_1 x_2 \ldots . x_n)^{\frac{1}{n}} = cG_x$$

***Corollary* 3.2:** *If X assumes values $x_1, x_2, \ldots . x_n$ and y assumes values $y_1, y_2, \ldots . y_n$, then the geometric mean of the ratio of x to y is equal to the ratio of their geometric means.*

To prove this, let

$$r_i = \frac{x_i}{y_i}, \qquad i = 1, 2, \ldots, n$$

Geometric mean of $x_1, x_2, \ldots, x_n$ is:

$$G_x = (x_1 x_2 \ldots . x_n)^{\frac{1}{n}}$$

Geometric mean of $y_1, y_2, \ldots, y_n$ is:

$$G_y = (y_1 y_2 \ldots . y_n)^{\frac{1}{n}}$$

Geometric mean of $r_1, r_2, \ldots, r_n$ is:

$$G_r = (r_1 r_2 \ldots . r_n)^{\frac{1}{n}}$$

$$= \left(\frac{x_1}{y_1} \times \frac{x_2}{y_2} \times \ldots \times \frac{x_n}{y_n}\right)^{\frac{1}{n}}$$

$$= \frac{(x_1 x_2 \ldots . x_n)^{\frac{1}{n}}}{(y_1 y_2 \ldots . y_n)^{\frac{1}{n}}} = \frac{G_x}{G_y}$$

ollary **3.3:** *If x assumes values $x_1, x_2, \ldots x_n$ and y assumes values y_1, y_2, then the geometric mean of the product of x and y is equal to the duct of their geometric means. For, if $p = xy$,*

$$G_p = \{(x_1 y_1) \times (x_2 y_2) \ldots . \times (x_n y_n)\}^{\frac{1}{n}}$$

$$= (x_1 x_2 \ldots . x_n)^{\frac{1}{n}} \times (y_1 y_2 \ldots . y_n)^{\frac{1}{n}}$$

$$= G_x G_y$$

is completes the proof.

eorem **3.4:** *A variable X takes on n values, which are in geometric gression. Show that $A \times H = G^2$, where A, G and H are respectively ithmetic mean, geometric mean, and harmonic mean.*

roof: Let the variable assume values $a, ar, ar^2, \ldots, ar^{n-1}$, which are in ometric progression Then by definition

$$A = \frac{a + ar + ar^2 + \ldots . + ar^{n-1}}{n} = \frac{a(1-r^n)}{n(1-r)}, \qquad r < 1$$

$$G = \left(a.ar.ar^2 \ldots . ar^{n-1}\right)^{\frac{1}{n}}$$

$$= \left\{a^n r^{(1+2+\ldots . +n-1)}\right\}^{\frac{1}{n}} = a\left\{r^{\frac{n(n-1)}{2}}\right\}^{\frac{1}{n}} = a\left\{r^{\frac{n-1}{2}}\right\}$$

nce,

$$G^2 = a^2 r^{n-1}$$

ain

$$H = \frac{n}{\frac{1}{a} + \frac{1}{ar} + \frac{1}{ar^2} + \ldots . + \frac{1}{ar^{n-1}}}$$

$$= \frac{na}{1 + r^{-1} + r^{-2} + \ldots + r^{-(n-1)}}$$

$$= \frac{na}{\frac{\left(r^{-1}\right)^n - 1}{\left(r^{-1}\right) - 1}} = \frac{na\,r^{n-1}(1-r)}{1-r^n}$$

Now
$$A \times H = \frac{a(1-r^n)}{n(1-r)} \times \frac{nar^{n-1}(1-r)}{1-r^n} = a^2 r^{n-1} = G^2$$

This proves the theorem.

EXERCISES 3

1. Under what condition(s) would it make no difference whether the mean, the median, or the mode is used as a measure of central tendency? Explain
2. Describe a situation in which it is more appropriate to use each of the following as a measure of central tendency: (a) median (b) mode (c) mean, (d) geometric mean, (e) harmonic mean. Stating the conditions involved, write down the empirical relationships between mean, median and the mode.
3. What is an average? Why is it necessary? How do common people view this measure? Give a broad outline of the statistical averages you are familiar with.
4. What is meant by central tendency of data? How does this concept differ from location? What are the various measures of central tendency and location? Explain them with examples.
5. What are the criteria by which you can judge the adequacy of an average? Compare the different averages on the basis of these criteria.
6. What are appropriate measures of central tendency for nominal and ordinal data? Which measure of central tendency is applicable at all levels of measurement? Why?
7. What is a median? How do you compute median for grouped and ungrouped data? Describe how you would determine median from (i) a histogram and (ii) an ogive.
8. What criteria do you apply to judge the merits of an average? Discuss the merits and demerits of different averages in common use with special reference to these criteria.
9. What is a median? How do you compute median for grouped and ungrouped data? Describe how you would determine median from (i) a histogram and (ii) an ogive.

What is a weighted average? When is it appropriate to use a weighted average rather than a simple average? When they become identical? Discuss the method of computing weighted average from appropriate data.

Discuss the importance of mean, median and mode in statistical research. Describe in brief the advantages and disadvantages of one over the others. Discuss how median and mode can be located from graphs.

The following values represent the results of several measurements on the concentration of calcium in a stream [in milligrams per decimeter (mg/dm)]

14 8 9 8 13 14 8 8 12 11 10 7 10 12 12

(a) Calculate the mean concentration [**Ans.**: Mean =10.4]

(b) Calculate the median and the mode [**Ans.**: Median=10 and Mode=8.]

In order to establish a base time required to read and type symbols in a recognition study, a psychologist ran several trials on a number of occasions with each subject and recorded the time in milliseconds (msec). The following values represent the results from ten occasions for one of the subjects:

Occasion	Number of Trials	Mean response time (msec)
1	25	950.08
2	20	894.90
3	50	978.26
4	50	831.52
5	30	982.10
6	30	868.30
7	20	901.80
8	50	796.72
9	10	1013.60
10	25	966.12

Calculate the mean response time for these subjects. [**Ans.**: 902.62}

Six samples of students of Bangladesh Agricultural University (BAU), of sizes 50, 55, 40, 48, 35 and 60 were taken from six faculties, giving mean heights 4.3, 4.7, 5.0, 5.1, 4.8, and 5.7 ft. respectively. Use an appropriate measure of central tendency to estimate the mean height of the students of BAU

What are quartiles, deciles and percentiles? What role do they play in statistics as measures of central tendency? Define percentile rank and distinguish it from percentile.

The following values are the total times in minutes that 16 battery packs operated before requiring recharge: