3

Measures of Central Tendency or Averages

3.1 INTRODUCTION

For practical purposes, the condensation of data set into a frequency distribution and the visual presentation are not enough, particularly, when two or more different data sets are to be compared. A data set can be summarized in a single value. Such a value, usually somewhere in the centre and representing the entire data set, is a value at which the data have a tendency to concentrate. The tendency of the observations to cluster in the central part of the data set is called *Central Tendency* and the summary value as a *measure of central tendency.* Since a measure of central tendency indicates the *location* or general position of the distribution or the data set in the range of observations, it is also known as a *measure of location or position*. The measures of central tendency or location are generally known as *Averages.* But in everyday language, 'the average' is often understood to refer to the arithmetic mean (a form of average to be discussed in section 3.4), it is for this reason that when anyone speaks of 'the average' (without qualification) of a set of observations, it may, as a rule, be assumed that the arithmetic mean is meant. The use of the term average has been traced to the time of Pythagoras (570-500 B.C.). Two points should be noted. First, a measure of central tendency should be somewhere within the range of the data, and secondly, it should remain unchanged by a rearrangement of the observations in a different order.

Since the late nineteenth century, the practice has been to make a distinction between a sample and a population from which the sample is drawn, by using Latin letters for numerical quantities describing the sample and Greek letters for corresponding quantities characterizing the population. It should be noted that population parameters are rarely calculated directly as all observations from the population are not usually available. The measures corresponding to population parameters are generally calculated from sample data and are regarded as the *estimates* of population parameters.

3.2 CRITERIA OF A SATISFACTORY AVERAGE

Several types of averages are defined to measure the representative or "typical" value of a set of data or distribution. It is therefore desirable that an average should be

(i) rigorously defined,
(ii) based on all the observations made,
(iii) simple to understand and easy to interpret,
(iv) quickly and easily calculated,
(v) amenable to mathematical treatment,
(vi) relatively stable in repeated sampling experiments, and
(vii) not unduly influenced by abnormally large or small observations.

An average that possesses all or most of the conditions stated above, is considered a *satisfactory* average.

3.3 TYPES OF AVERAGES

The most common types of averages are (i) the arithmetic mean or simply the mean, (ii) the geometric mean, (iii) the harmonic mean, (iv) the median and (v) the mode. The first three types are mathematical in character and give an indication of the magnitude of the observed values. The fourth type indicates the middle position while the last provides information about the most frequent value in the distribution or the data set.

3.4 THE ARITHMETIC MEAN

The *arithmetic mean* or simply the *mean* is the most familiar average. It is defined as a value obtained by dividing the sum of all the observations by their number, that is

$$\text{Mean} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

The mean may correspond to either a population or a sample from the population. If the given set of observations represents a population, the mean is called the *population mean* which is traditionally denoted by μ (the Greek letter *mu*). Thus the population mean of a set of N observations $x_1, x_2, \ldots, x_N$ is given as

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum x_i}{N}, \quad (i = 1, 2, \ldots, N)$$

where Σ, the Greek capital *sigma*, is a convenient symbol for summation.

If, instead, the given set of observations represents a sample, the mean is called the *sample mean*, usually denoted by placing a bar over the symbol used to represent the observations or the variable. Thus the mean of a set of n observations $x_1, x_2, \ldots, x_n$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}, \quad (i = 1, 2, \ldots, n)$$

where $\bar{x}$ is the mean of a sample of size n.

It is worthwhile to note that the population mean is a fixed quantity, whereas $\bar{x}$, the sample mean, is a variable because different samples from the same population tend to have different means.

In order to interpret the meaning of arithmetic mean, let x_i denote the marks obtained by the *ith* student in a class. Then $\sum x_i$ stands for the total marks obtained by all students and $\bar{x}$, the mean, represents the number of marks that *would have been obtained by each student if everyone in the class had obtained the same number of marks.* Geometrically the mean represents a point at which the distribution or the set of observations would balance.

Example 3.1 The marks obtained by 9 students are given below:

45, 32, 37, 46, 39, 36, 41, 48, 36.

Calculate the arithmetic mean.

The mean is given by

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{45 + 32 + 37 + 46 + 39 + 36 + 41 + 48 + 36}{9}$$

$$= \frac{360}{9} = 40 \text{ marks}$$

It is relevant to note that, if these marks represent the entire set of observations for the population, the above calculation gives the population mean, i.e. μ would equal to 40 marks.

3.4.1 The Weighted Arithmetic Mean. The multipliers or a set of numbers which express more or less adequately the relative importance of various observations in a set of data are technically called the *weights*. We assign weights $w_1, w_2, \ldots, w_n$ to the observations in a set of data according to their relative importance, when the observations are

not of equal importance. The *weighted mean*, denoted by $\bar{x}_w$, of a set of n values $x_1, x_2, \ldots, x_n$ with corresponding weights $w_1, w_2, \ldots, w_n$ is then defined as

$$\bar{x}_w = \frac{x_1w_1 + x_2w_2 + \ldots + x_nw_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\sum x_iw_i}{\sum w_i} \qquad (i = 1, 2, \ldots, n)$$

A weighted average is generally employed in the calculation of index numbers, birth and death rates, etc.

Example 3.2 Calculate the weighted mean from the following data:

Item	Expenditure (Rs.)	Weights
Food	290	7.5
Rent	54	2.0
Clothing	98	1.5
Fuel and Light	75	1.0
Other items	75	0.5

(P.U., B.A. (Optional), 1969, 94)

We calculate the weighted mean as below:

Item	Expenditure (x_i)	Weights (w_i)	x_iw_i
Food	290	7.5	2175.0
Rent	54	2.0	108.0
Clothing	98	1.5	147.0
Fuel and Light	75	1.0	75.0
Other items	75	0.5	37.5
Total	--	12.5	2542.5

Hence $\bar{x}_w = \frac{\sum x_iw_i}{\sum w_i} = \frac{2542.5}{12.5}$ = Rs. 203.4

3.4.2 Properties of the Arithmetic Mean. The arithmetic mean has the following four properties:

(i) For a set of data, the sum of the deviations of the observations x_i's from their mean, $\bar{x}$, taken with their proper signs, is equal to zero.

The sum of the deviations $= \sum_i (x_i - \bar{x}), \qquad (i = 1, 2, \ldots, n)$

$= \sum x_i - n\bar{x} \qquad (\because \bar{x} \text{ is constant})$

$= \sum x_i - \sum x_i = 0 \qquad (\because \bar{x} = \sum x/n)$

(ii) The sum of squared deviations of the x_i's from the mean, $\bar{x}$, is a minimum. In other words, $\sum(x_i - \bar{x})^2 \leq \sum(x_i - a)^2$, where a is an arbitrary value other than the mean.

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

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

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

$= \sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2 \quad [\because \sum(x_i - \bar{x}) = 0]$

It is obvious that $\sum (x_i - a)^2 > \sum(x_i - \bar{x})^2$ by $n(\bar{x} - a)^2$. The equality sign holds only when $\bar{x} = a$.

Hence $\sum(x_i - \bar{x})^2$ is always less than $\sum(x_i - a)^2$ if $a \neq \bar{x}$.

This property is usually called the *minimal* property of the mean.

(iii) If k subgroups of data consisting of $n_1, n_2, \ldots, n_k$, $(\sum n_i = n)$ observations have respective means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, then $\bar{x}$, the mean for all the data, is given by

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

$$= \frac{\sum n_i\bar{x}_i}{n}, \qquad (i = 1, 2, \ldots, k)$$

i.e. a weighted mean of all the subgroup means.

(iv) If $y_i = ax_i + b$ $(i = 1, 2, \ldots, n)$, where a and b are any two numbers and $a \neq 0$, then $\bar{y} = a\bar{x} + b$.

Now summing over all values of i, we obtain

$$\sum y_i = a \sum x_i + nb$$

Dividing both sides by n, we get

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

As the equation $y = ax + b$ represents a linear transformation from x to y, this property is usually called the *invariance* of the mean under a *linear transformation* and it provides the basis for so-called *coding* which refers to the operation of subtracting (or adding) a constant from each observation and then dividing (or multiplying) by another constant for computational convenience.

Example 3.3 The mean heights and the number of students in three sections of a statistics class are given below:

Section	Number of boys	Mean height
A	40	62"
B	37	58"
C	43	61"

Find the overall mean height of 120 boys.

Here $n_1 = 40; n_2 = 37; n_3 = 43$, and

$\bar{x}_1 = 62", \bar{x}_2 = 58", \bar{x}_3 = 61"$

The mean height of the combined class is given as

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

$$= \frac{(40 \times 62) + (37 \times 58) + (43 \times 61)}{40 + 37 + 43} = \frac{7249}{120} = 60.4"$$

3.4.3 Mean From Grouped Data. When the number of observations is very large,, the data are organised into a frequency distribution, which is used to calculate the *approximate* values of descriptive measures as the identity of the observations is lost. To calculate the approximate value of the mean, the observations in each class are assumed to be identical with the class midpoint so that the product of the midpoint by the number of observations, *i.e.* frequency, would be approximately equal to the sum of observations for each class. Thus, if a frequency distribution has k classes with midpoints $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$), the mean is then given by the formula

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{f_1 + f_2 + \ldots + f_k}$$

$$= \frac{\sum f_ix_i}{n}, \quad (i = 1, 2, \ldots, k)$$

As a *weight* indicates the number of times an observation is to be counted, the mean calculated from a frequency distribution may also be regarded as the *weighted mean* where each class midpoint x_i, taken as the average value of the observations in that class, is weighted by the respective frequency f_i and the sum of the weighted products is divided by the sum of frequencies, i.e. weights.

Sometimes, there may be a slight difference in the values of $\bar{x}$ on account of errors caused by the assumption that all observations in any class may be treated as approximately the midpoint of that class, but experience tells us that this error is usually small and never serious. The following example illustrates this situation.

Example 3.4 Calculate the mean weight of apples from the data given in Example 2.2 directly from the observed values and from the data grouped into a frequency distribution.

The calculations are outlined below:

Weight (grams)	Sum of actual observations	f_i	Mean Weight of each class ($\bar{x}_i$)	$f_i\bar{x}_i$	Midpoints (x_i)	$f_i x_i$
65–84	695	9	77.2	694.8	74.5	670.5
85–104	947	10	94.7	947.0	94.5	945.0
105–124	1919	17	112.9	1919.3	114.5	1946.5
125–144	1325	10	132.5	1325.0	134.5	1345.0
145–164	766	5	153.2	766.0	154.5	772.5
165–184	716	4	179.0	716.0	174.5	698.0
185–204	956	5	191.2	956.0	194.5	972.5
Total	7324	60	--	7324.1	--	7350.0

Calculation based on Ungrouped data

We calculate the mean weight, $\bar{x}$, directly from all the observed values, which add to 7324. (The second column consists of subtotal of actual observations in any class).

$$\therefore \quad \bar{x} = \frac{\sum x_i}{n}, \quad (i = 1, 2, ..., 60)$$

$$= \frac{7324}{60} = 122.07 \text{ grams}$$

Next, we find the mean weight, $\bar{x}$, by multiplying the actual mean of the observations in any class by the corresponding frequency, adding the products and then dividing by n (column 5).

Thus $$\bar{x} = \frac{\sum f_i \bar{x}_i}{n}, \quad (i = 1, 2, ..., 7)$$

$$= \frac{7324.1}{60} = 122.07 \text{ grams}$$

Calculation based on Grouped data

Here we calculate the mean weight from grouped data, assuming that all observations in any class are identical with the midpoint of that class. The sixth column consists of class midpoints, x_i and the products are given in column 7.

Then $$\bar{x} = \frac{\sum f_i x_i}{n}, \quad (i = 1, 2, ..., 7)$$

$$= \frac{7350.0}{60} = 122.5 \text{ grams}$$

It should be noted that the numerical value of $\bar{x}$, calculated from the frequency distribution is slightly different from the value obtained directly from the ungrouped data.

3.4.4 Change of Origin and Scale. To reduce the computational labour and to save time, a change of the origin and scale can be made. If x_i denotes an observed value, a and b are two constants with $b \neq 0$, then the operations $x_i + a$, bx_i and $bx_i + a$ are known respectively as the *change of origin*, the *change of scale* and both *change of origin and scale.*

Let a be an arbitrary origin, sometimes called *assumed mean*, and let $x_i = a + hu_i$ where h denotes the unit of measurement. Then its corresponding *coded* value is $u_i = \frac{x_i - a}{h}$.

Now $$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum(a + hu_i)$$

$$= \frac{na}{n} + \frac{h\sum u_i}{n} = a + h\bar{u}$$

Thus the arithmetic mean can be calculated from any *origin* we may choose and using any *scale* we desire. This transformation is particularly useful for calculations based on grouped data, where h is the width of class interval and a is usually chosen the class midpoint lying in the region of the higher frequencies so that the larger frequencies may be multiplied by smaller values of u. This procedure gives us a *Short method* for hand calculations.

Example 3.5 Given the following frequency distribution of weights, calculate the mean weight by the Short Method.

Weight (grams)	65-84	85-104	105-124	125-144	145-164	165-184	185-204
f	9	10	17	10	5	4	5

To calculate the mean weight, let us take $u_i = \frac{x_i - 114.5}{20}$, where $a = 114.5$ is the midpoint corresponding to the largest frequency and $h = 20$ is the width of the uniform class interval. The necessary calculations are shown in the table below:

Weight (grams)	Midpoints (x_i)	f_i	u_i	$f_i u_i$
65 – 84	74.5	9	−2	−18
85 – 104	94.5	10	−1	−10
105 – 124	114.5	17	0	−28
125 – 144	134.5	10	1	10
145 – 164	154.5	5	2	10
165 – 184	174.5	4	3	12
185 – 204	194.5	5	4	20
Total		60	--	+52 / 24

Thus $\bar{x} = a + h\bar{u}$, where $\bar{u} = \frac{\sum f_i u_i}{n}$

$$= 114.5 + \frac{(24)\,(20)}{60} = 114.5 + 8.0 = 122.5 \text{ grams.}$$

Example 3.6 Compute the mean for the following frequency distribution of annual death rates:

Death Rate	Frequency
3.5 – 4.4	1
4.5 – 5.4	4
5.5 – 6.4	5
6.5 – 7.4	13
7.5 – 8.4	12
8.5 – 9.4	19
9.5 – 10.4	13
10.5 – 11.4	10
11.5 – 12.4	6
12.5 – 13.4	4
13.5 – 14.4	1
Total	88

The necessary calculations are given below:

Death Rate	Midpoints (x_i)	f_i	$u_i(=x_i-8.95)$	f_iu_i
3.5 – 4.4	3.95	1	–5	–5
4.5 – 5.4	4.95	4	–4	–16
5.5 – 6.4	5.95	5	–3	–15
6.5 – 7.4	6.95	13	–2	–26
7.5 – 8.4	7.95	12	–1	–12
8.5 – 9.4	8.95	19	0	–74
9.5 – 10.4	9.95	13	1	13
10.5 – 11.4	10.95	10	2	20
11.5 – 12.4	11.95	6	3	18
12.5 – 13.4	12.95	4	4	16
13.5 – 14.4	13.95	1	5	5
Total	--	88	--	+72 –2

Hence $\bar{x} = a + \dfrac{\sum_i f_i u_i}{n}$, where a is assumed mean and $h = 1$.

$$= 8.95 + \frac{(-2)}{88} = 8.95 - 0.02 = 8.93$$

3.5 THE GEOMETRIC MEAN

The *geometric mean*, G, of a set of n positive values $x_1, x_2, \ldots, x_n$ is defined as the positive nth root of their product, *i.e.*

$$G = \sqrt[n]{x_1 x_2 \ldots x_n}, \quad \text{where } x > 0$$

When n is large, the computation of the geometric mean becomes laborious, as we have to multiply all the values and then extract the nth root. The arithmetic is simplified by using logarithms to the base 10. Thus, taking logarithms, we get

$$\log G = \frac{1}{n}[\log x_1 + \log x_2 + \ldots + \log x_n]$$

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

Hence $$G = \text{antilog}\left[\frac{1}{n}\sum \log x_i\right]$$

It means the geometric mean is the anti-logarithm of the arithmetic mean of the logarithms of the values themselves.

For data organised into a grouped frequency distribution, having k classes with classmarks $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$), the formula for the geometric mean is given by

$$G = [x_1^{f_1} \cdot x_2^{f_2} \ldots x_k^{f_k}]^{1/n},$$

In terms of logarithms, the formula becomes

$$\log G = \frac{1}{n}[f_1 \log x_1 + f_2 \log x_2 + \ldots + f_k \log x_k]$$

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

or $$G = \text{antilog}\left[\frac{1}{n}\sum f_i \log x_i\right]$$

The *weighted* geometric mean of the values $x_1, x_2, \ldots, x_k$ with corresponding weights $w_1, w_2, \ldots, w_k$ is given by

$$\log G_w = \frac{1}{\sum w_i}[\sum w_i \log x_i]$$

The geometric mean is appropriate to average ratios and rates of change.

Example 3.7 Find the geometric mean of 45, 32, 37, 46, 39, 36, 41, 48 and 36.

The geometric mean, G, is calculated as

$$G = \sqrt[9]{(45 \times 32 \times 37 \times 46 \times 39 \times 36 \times 41 \times 48 \times 36)}$$

Taking logs, we have

$$\log G = \frac{1}{9}\ [\log 45 + \log 32 + \log 37 + \log 46 + \log 39 + \log 36 + \log 41 + \log 48 + \log 36]$$

$$= \frac{1}{9}\ [1.65321 + 1.50515 + 1.56820 + 1.66276 + 1.59106 + 1.55630 + 1.61278 + 1.68124 + 1.55630]$$

$$\therefore \quad \log G = \frac{1}{9}[14.38700] = 1.59856$$

Hence $G = \text{anti-log}\ (1.59856) = 39.68$

Example 3.8 Given the following frequency distribution of weights, calculate the geometric mean.

Weight (grams)	65–84	85–104	105–124	125–144	145–164	165–184	185–204
f	9	10	17	10	5	4	5

The computation of the geometric mean is shown below:

Weight (grams)	x_i	f_i	$\log x_i$	$f_i \log x_i$
65 – 84	74.5	9	1.8722	16.8498
85 – 104	94.5	10	1.9754	19.7540
105 – 124	114.5	17	2.0589	35.0013
125 – 144	134.5	10	2.1287	21.2870
145 – 164	154.5	5	2.1889	10.9445
165 – 184	174.5	4	2.2418	8.9672
185 – 204	194.5	5	2.2889	11.4445
Σ	--	60	--	124.2483

$$\therefore \quad \log G = \frac{1}{n}\Sigma f_i \log x_i = \frac{124.2483}{60} = 2.0708$$

Hence G = Anti-log (2.0708) = 117.7 grams

3.6 THE HARMONIC MEAN

The *harmonic mean*, H, of a set of n values $x_1, x_2, \ldots, x_n$ is defined as the reciprocal of the arithmetic mean of the reciprocals of the values. In symbols,

$$H = \text{Reciprocal of } \frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}{n}$$

$$= \frac{n}{\Sigma \frac{1}{x_i}}, \quad \text{where } x \neq 0$$

The harmonic mean is an appropriate type to be used in averaging certain kinds of ratios or rates of change. To illustrate this formula, let us take an example. Suppose a car is running at the rate of 15 *km/hr* during the first 30 *km*; at 20 *km/hr* during the second 30 *km*; and at 25 *km/hr* during the third 30 *km*. The distance is constant but the times are

variable. Therefore, the harmonic mean is the correct average. In this case, the harmonic mean is

$$H = \text{Reciprocal of } \frac{\frac{1}{15} + \frac{1}{20} + \frac{1}{25}}{3}$$

$$= \frac{3}{0.06667 + 0.5000 + 0.04000}$$

$$= \frac{3}{0.15667} = 19.15 \; km/hr \text{ approximately.}$$

Care should be exercised to apply the harmonic mean. The following rule will help determine the application of the harmonic mean.

"When rates are expressed as x per y, and x is constant, the harmonic mean is required; but if y is constant, the arithmetic mean is required."

For data organised into a frequency distribution having k classes with classmarks $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$) the harmonic mean of the distribution is given by

$$H = \text{Reciprocal of } \frac{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_k}{x_k}}{f_1 + f_2 + \ldots + f_k}.$$

$$= \frac{n}{\sum f_i \frac{1}{x_i}}$$

Similarly, the weighted harmonic mean is defined as

$$H_w = \frac{w_1 + w_2 + \ldots + w_n}{w_1\left(\frac{1}{x_1}\right) + w_2\left(\frac{1}{x_2}\right) + \ldots + w_n\left(\frac{1}{x_n}\right)}$$

$$= \frac{\sum w_i}{\sum w_1\left(\frac{1}{x_i}\right)}$$

Example 3.9 Find the harmonic mean from the following frequency distribution of weights:

Weight (grams)	65–84	85–104	105–124	125–144	145–164	165–184	185–204
f	9	10	17	10	5	4	5

We calculate the harmonic mean as below:

Weight (grams)	x_i	f_i	$f_i\left(\frac{1}{x_i}\right)$
65 – 84	74.5	9	0.12081
85 – 104	94.5	10	0.10582
105 – 124	114.5	17	0.14847
125 – 144	134.5	10	0.07435
145 – 164	154.5	5	0.03236
165 – 184	174.5	4	0.02292
185 – 204	194.5	5	0.02571
Σ	--	60	0.53044

Hence $$H = \frac{n}{\Sigma f_i\left(\frac{1}{x_i}\right)} = \frac{60}{0.53044} = 113.11 \text{ grams}$$

Example 3.10 Compute the Geometric and Harmonic means for the following distribution of annual death rates:

x_i	3.95	4.95	5.95	6.95	7.95	8.95	9.95	10.95	11.95	12.95	13.95
f_i	1	4	5	13	12	19	13	10	6	4	1

(B.I.S.E. Lahore 1971)

We can construct the following table to compute the geometric and harmonic means:

x_i	f_i	$\log x_i$	$f_i \log x_i$	$\frac{1}{x_i}$	$f_i\left(\frac{1}{x_i}\right)$
3.95	1	0.59660	0.59660	0.25316	0.25316
4.95	4	0.69461	2.77844	0.20202	0.80808
5.95	5	0.77452	3.87260	0.16807	0.84035
6.95	13	0.84198	10.94574	0.14388	1.87044
7.95	12	0.90037	10.80444	0.12579	1.50948
8.95	19	0.95182	18.08458	0.11173	2.12287
9.95	13	0.99782	12.97166	0.10050	1.30650
10.95	10	1.03945	10.39450	0.09132	0.91320
11.95	6	1.07740	6.46440	0.08368	0.50208
12.95	4	1.11229	4.44916	0.07722	0.30888
13.95	1	1.14459	1.14459	0.07168	0.07168
Total	88	--	82.50671	--	10.50672

Now $\log G = \frac{1}{n}\sum f_i \log x_i = \frac{82.50671}{88} = 0.93758$

Hence G = anti-log (0.93758) = 8.66, and

$$\text{Harmonic mean} = \frac{n}{\sum f_i\left(\frac{1}{x_i}\right)} = \frac{88}{10.50672} = 8.38$$

3.7 THE MEDIAN

The *median* is defined as a value which divides a data set that have been ordered, into two equal parts, one part comprising of observations greater than and the other part smaller than it. Or more precisely, the median is a value at or below which 50% of the ordered data lie.

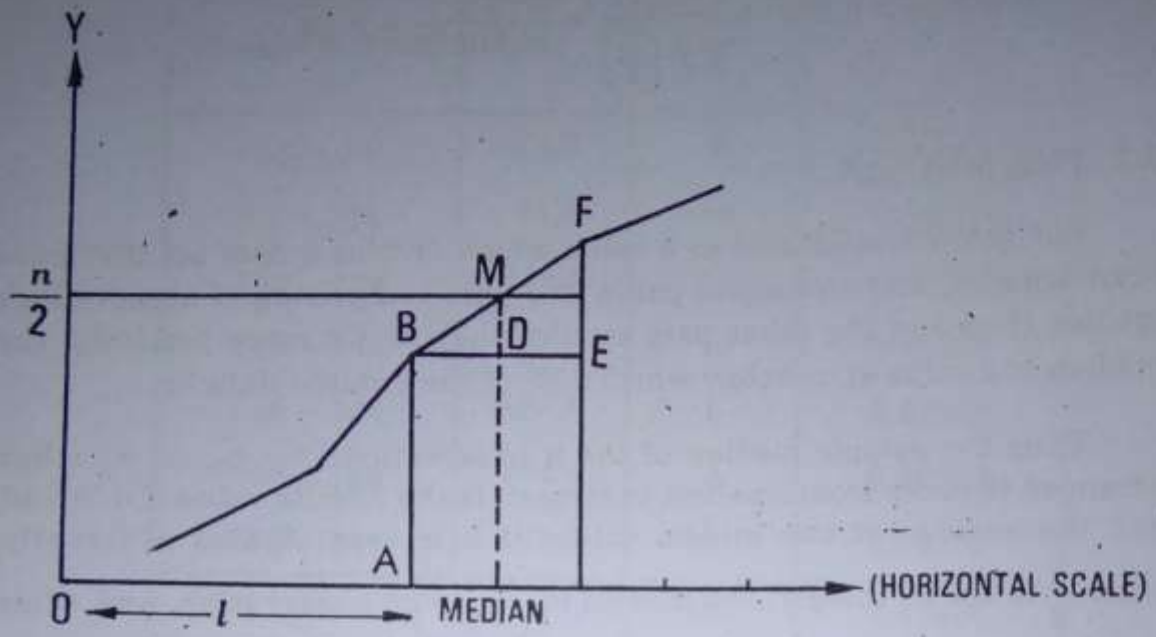
Thus the sample median of the n observations $x_1, x_2, \ldots, x_n$ when arranged in *order* from smallest to largest, is the middle value if n is odd, and the average of two middle values if n is even. Stated differently, when $\frac{n}{2}$ is not an integer, the median is $\left(\frac{n+1}{2}\right)th$ observation, and when $\frac{n}{2}$ is an integer, the median is the average of $\frac{n}{2}th$ and $\left(\frac{n}{2}+1\right)th$ observations.

The median in case of a discrete or ungrouped frequency distribution can be found as above by forming a cumulative frequency distribution.

For data grouped into a frequency distribution, the *median* is a value or a point on the horizontal scale through which a vertical line divides the histogram of the distribution into two parts of equal area. In other words, the median is that value on the horizontal scale which corresponds to a cumulative frequency $\frac{n}{2}$. This value would lie in a certain group, called the *median group*, but a single value for the median is often desirable. To obtain this single value, we *assume that the observations are evenly distributed within the median group.* Then it may be obtained as follows:

Let us consider a relevant portion of the cumulative frequency polygon as drawn on page 72. Then the median is the abscissa of the point M. That is

Median = OA + BD (see figure below).



Since BDM and BEF are similar triangles, therefore $\frac{BD}{BE} = \frac{DM}{EF}$

or $BD = BE \cdot \frac{DM}{EF}$

Now evidently BE is the width of the class-interval containing median and hence is equal to h.

$DM = \frac{n}{2} - AB$, where AB represents the cumulative frequency corresponding to the group preceding the median group. Let it be equal to C. Then $DM = \frac{n}{2} - C$.

EF is the difference between two cumulative frequencies, which is clearly the frequency corresponding to the median group and is denoted by f. If $OA = l$ (which is the lower boundary of median group), then substituting these values, we get the following formula

$$\text{Median} = l + \frac{h}{f}\left(\frac{n}{2} - C\right)$$

This process of determining the median is called *linear interpolation* and it does not require a uniform class interval. If this arithmetical process is not used, but the value of median (on the X-axis) corresponding to a cumulative frequency $\frac{n}{2}$ is read directly from the graph of Ogive curve, the process is called the *location of the median graphically*. The median is an average of position. It is also known as a partition value. The population median may be found in the same way from all the observations in the population.

3.7.1 Quantiles. When the number of observations is quite large, the principle according to which a distribution or an ordered data set is divided into two equal parts, may be extended to any number of divisions. The three values which divide the distribution into four equal parts, are called the *Quartiles*. These values are denoted by Q_1, Q_2 and Q_3 respectively. Q_1 is called the *first* or *lower quartile* and Q_3 is known as the *third or upper quartile*. In other words, the quartiles Q_1, Q_2 and Q_3 are the values at or below which lie respectively, the lowest 25, 50 and 75 per cent of the data. Similarly, the nine values which divide the distribution into ten equal parts, are called *Deciles* and are denoted by $D_1, D_2, \ldots, D_9$; while the ninety nine values dividing the data into one hundred equal parts, are called *Percentiles* and are denoted by $P_1, P_2, \ldots, P_{99}$. The second quartile or the fifth decile or the fiftieth percentile is obviously identical with the median.

Quartiles, deciles, percentiles and other values obtained by equal subdivision of the given set of data, are collectively called *Quantiles* or sometimes *Fractiles*. The Quantiles should be calculated when the number of observations is quite large.

It is interesting to note that all the Quantiles are percentiles. For example, the *3rd* quartile is the *75th* percentile and the *6th* decile corresponds to the *60th* percentile. We therefore use the following formula to compute P_j, the *jth* percentile from a set of n observations, arranged in *order* from smallest to largest.

(i) When $\frac{jn}{100}$ is not an integer, the *jth* percentile is given as

P_j = Observation with ordinal number $\left[\frac{jn}{100}\right] + 1$; and

(ii) when $\frac{jn}{100}$ is an integer, the *jth* percentile is

P_j = Average of two observation with ordinal numbers

$$\left(\frac{jn}{100}\right) \text{ and } \left(\frac{jn}{100}\right) + 1,$$

where $[x]$ stands for the largest integer in x.

In case of grouped data, Quantiles are calculated in the same way as the median.

Sir Francis Galton (1822–1911) is considered the originator of the terms *median* and *percentiles*, although it was Gauss (1777–1855) who actually suggested *median*.

Example 3.11 Given below are the marks obtained by 9 students:

45, 32, 37, 46, 39, 36, 41, 48 and 36.

Find the median and the quartiles.

To find the median and the quartiles, we first arrange the marks in *order* from lowest to highest. The ordered marks are;

32, 36, 36, 37, 39, 41, 45, 46, 48.

Here $n = 9$ and $\frac{n}{2}$, i.e. $\frac{9}{2}$ is not an integer, therefore

Median = Marks obtained by $\left(\left[\frac{n}{2}\right] + 1\right)th$ student

= Marks obtained by (4+1), *i.e* 5*th* student in ordered data,

= 39 marks

Now Q_1 = Marks obtained by $\left(\left[\frac{n}{4}\right] + 1\right)th$ student as $\frac{n}{4}$ is not an integer.

= Marks obtained by $\left(\left[\frac{9}{4}\right] + 1\right)th$, *i.e. 3rd* student.

= 36 marks.

Similarly,

Q_3 = Marks obtained by $\left(\left[\frac{3n}{4}\right] + 1\right)th$ student

= Marks obtained by $\left(\left[\frac{27}{4}\right] + 1\right)th$, *i.e. 7th* student.

= 45 marks.

Example 3.12 The following distribution relates to the number of assistants in 50 retail establishments.

No. of assistants	0	1	2	3	4	5	6	7	8	9
f	3	4	6	7	10	6	5	5	3	1

Find the median number of assistants. Also calculate the quartiles and the *7th* decile.

This is an example of ungrouped frequency distribution with unit class interval. To locate the median, the quartiles and the *7th* decile for such a distribution, we cumulate the frequencies as shown in the table.

No. of assistants (*x*)	0	1	2	3	4	5	6	7	8	9
Frequency	3	4	6	7	10	6	5	5	3	1
Cumulative Frequency	3	7	13	20	30	36	41	46	49	50

Since $\frac{n}{2}$, *i.e.* $\frac{50}{2}$ is an integer, therefore the median is the average number of assistants in $\left(\frac{n}{2}\right)th$ and $\left(\frac{n}{2}+1\right)th$, *i.e.*, *25th* and *26th* retail establishments. Looking at the cumulative frequency row, we find that these two values corresponding to the same value of x, *i.e.* 4.

Hence *median* number of assistants = 4.

For Q_1, we see that $\frac{n}{4}$, *i.e.* $\frac{50}{4}$ is not an integer, therefore

Q_1 = No. of assistants in $\left(\left[\frac{50}{4}\right]+1\right)th$ establishment.

= No. of assistants in (12 + 1), *i.e. 13th* establishment

= 2 assistants.

Similarly,

Q_3 = No. of assistants in $\left(\left[\frac{3\times 50}{4}\right]+1\right)th$ establishment as $\frac{3n}{4}$ is also not an integer.

= No. of assistants in 38*th* establishment.

= 6 assistants.

Again D_7 = Average number of assistants $\left(\frac{7\times 50}{10}\right)th$ and $\left(\frac{7\times 50}{10}+1\right)th$ establishment as $\frac{7n}{10}$ is an integer.

= Average number of assistants in 35*th* and 36*th* establishments

= 5 assistants (since both values correspond to 5)

Example 3.13 Find the median, the quartiles and the 8*th* decile for the distribution of examination marks given below:

Marks	30–39	40–49	50–59	60–69	70–79	80–89	90–99
Number of students	8	87	190	304	211	85	20

(P.U., B.A/B.Sc. 1970)

To find the median marks, quartiles, etc. by the process of *linear interpolation*, the data are assumed to be continuous. Thus we need the class boundaries as the cumulative frequencies correspond to upper class boundaries, *i.e.* to 39.5, 49.5, etc., and not to 39, 49, etc., the upper class

limits. Hence the class boundaries are shown in the first column and the last column of the table below consists of cumulative frequencies.

Class-boundaries (marks)	Midpoints (x_i)	Frequency (f_i)	Cumulative frequency (F)
29.5 – 39.5	34.5	8	8
39.5 – 49.5	44.5	87	95
49.5 – 59.5	54.5	190	285
59.5 – 69.5	64.5	304	589
69.5 – 79.5	74.5	211	800
79.5 – 89.5	84.5	85	885
89.5 – 99.5	94.5	20	905

Median = Marks obtained by $\left(\frac{n}{2}\right)th$ student

= Marks obtained by $\frac{905}{2}$, *i.e.* $452.5th$ student which corresponds to marks in the class 59.5 – 69.5. Therefore

Median = $l + \frac{h}{f}\left(\frac{n}{2} - C\right)$, where the letters have their usual significance.

$= 59.5 + \frac{10}{304}(452.5 - 285)$

$= 59.5 + \frac{1675}{304} = 59.5 + 5.5 = 65$ marks.

And Q_1 = Marks of $\left(\frac{n}{4}\right)th$ student

= Marks of $\frac{905}{4}$, *i.e.* $226.25th$ student which corresponds to a value in the class 49.5–59.5. Therefore

$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) = 49.5 + \frac{10}{190}(226.25 - 95)$

$= 49.5 + 6.9 = 56.4 = 56$ marks

Again Q_3 = Marks of $\left(\frac{3n}{4}\right)th$ student

= Marks of $\frac{3 \times 905}{4}$, *i.e.* $678.75th$ student which lies in the class 69.5 – 79.5

$$\therefore \quad Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C\right)$$

$$= 69.5 + \frac{10}{211}(678.75 - 589) = 69.5 + 4.2 = 73.7 = 74 \text{ marks}$$

And D_8 = Marks of $\left(\frac{8n}{10}\right)th$ student

= Marks of $\frac{8 \times 905}{10}$, *i.e.*$724th$ student which also lies in the class 69.5 – 79.5

Hence $D_8 = l + \frac{h}{f}\left(\frac{8n}{10} - C\right)$

$$= 69.5 + \frac{10}{211}(724 - 589) = 69.5 + 6.4 = 76 \text{ marks}$$

3.8 THE MODE

The French word *mode* meaning fashion, has been adopted to convey the idea of "most frequent". The *mode* is defined as a value which occurs most frequently in a set of data, that is it indicates the most common result. A set of data may have more than one mode or no mode at all when each observation occurs the same number of times.

In an ungrouped frequency distribution with classes consisting of single values, the mode can be immediately located by examining the distribution. For example, in the distribution relating to the number of assistants in 50 retail establishments (Example 3.12) the mode is 4, as the frequency for $x = 4$ is greater than for any other value of x.

When the data re organised into a grouped frequency distribution, the mode would lie in the class that carries the highest frequency. This class is called the *modal class*. For most practical purposes, it is sufficient to take the midpoint of the modal class as the mode but generally it is a poor approximation. It therefore becomes desirable to decide at what point of the modal class, the mode should be located. To meet this requirement a method based on three adjacent rectangles of the histogram, with the tallest in the middle, has been developed. The method is

$$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h,$$

where l = lower class boundary of the modal class,

f_m = frequency of the modal class,

f_1 = frequency associated with the class *preceding* the modal class,

f_2 = frequency associated with the class *following* the modal class, and

h = width of class interval.

The mode can also be calculated by the following formula:

$$\text{Mode} = l + \frac{f_2}{f_1 + f_2} \times h,$$

where the letters have their usual meaning. It should be noted that the first formula is more accurate and should be generally used in calculating the mode.

When a frequency distribution is displayed as a smooth curve, the mode is the abscissa of the highest ordinate. A distribution having a single mode, is called a *unimodal* distribution. A distribution having a single mode, is called a *unimodal* distribution, while a distribution with two or more modes, is called a *bimodal* or *multimodal* distribution. It has no meaning for flat-topped distributions. It should be remembered that, when a frequency distribution has classes of unequal widths, the modal class is the class with maximum frequency per unit. The mode should be calculated if the frequency distribution has a class that carries more frequencies than the others and this class should not be at the extremity of the distribution.

Example 3.14 Calculate the mode for the distribution of examination marks given in Example 3.13.

The class that carries the highest frequency is 59.5 – 69.5, which is thus the modal class.

Also $l = 59.5, f_1 = 190, f_2 = 211, f_m = 304$ and $h = 10$.

Hence $$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h,$$

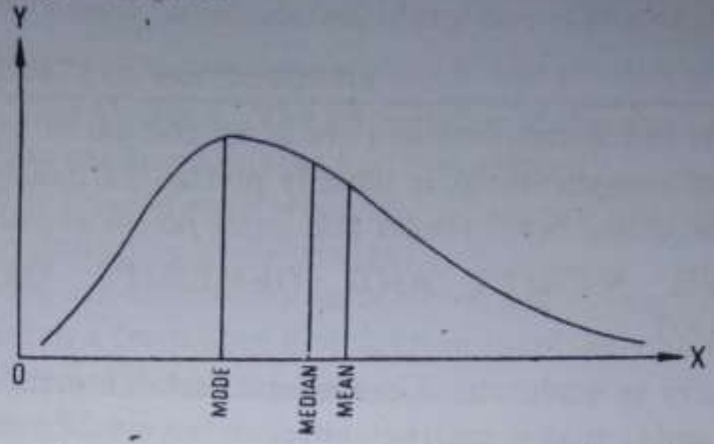
$$= 59.5 + \frac{304 - 190}{(304 - 190) + (304 - 211)} \times 10$$

$$= 59.5 + 5.8 = 65.3 = 65 \text{ marks.}$$

3.9 EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE

In a single-peaked frequency distribution, the values of the mean, median and mode coincide if the frequency distribution is absolutely

symmetrical. But if these values differ, the frequency distribution is said to be skewed or *asymmetrical.*



Experience tells us that in a unimodal curve of moderate skewness, the median is usually sandwiched between the mean and the mode and between them the following approximate relation holds good.

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

or $$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}.$$

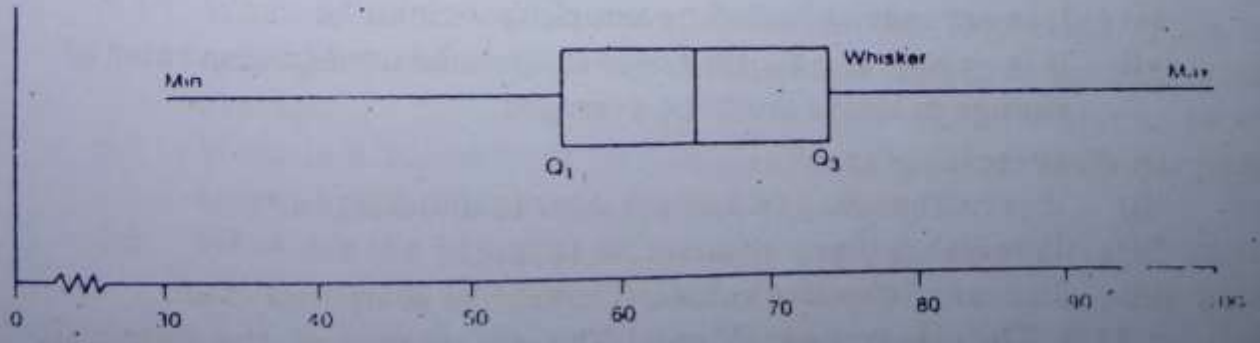
This empirical relation does not hold in case of a J-shaped or an extremely skewed distribution.

3.10 THE BOX PLOTS

The Box Plots, which are graphically very simple, are based on the Median, a measure of location and the Interquartile Range (*IQR*), a measure of data's variability. They are informative and effective for comparing two or more data sets / distributions.

A box plot is constructed by drawing a rectangle (the *box*) with the ends (called the *hinges*) drawn at the lower and upper quartiles (Q_1 and Q_3). The median of the data is shown in the box usually by a " + " sign. The straight lines (called the *whiskers*) are drawn from each hinge to the most extreme observations. The entire graph is called a *Box and Whiskers plot*. If one whisker is longer, the distribution of data is skewed in the direction of the longer whisker. The box plot given below represents the distribution of examination marks given in Example 3.13.

BOX PLOT FOR DATA IN EXAMPLE 3.13



When two or more distributions are to be compared by drawing box plots, the scale of measurement is usually plotted vertically. Sometimes, two sets of limits, called *inner fences* and *outer fences* are also used.

3.11 RELATIVE MERITS AND DEMERITS OF VARIOUS AVERAGES

It is necessary to understand the merits and demerits of each one of the averages in order that it may be appropriately employed.

3.11.1 The Arithmetic Mean. The advantages of the mean are:

(i) It is rigorously defined by a mathematical formula.
(ii) It is based on all the observations in the data.
(iii) It is easy to calculate and simple to comprehend.
(iv) It is determined for almost every kind of data.
(v) It is a relatively stable statistic with the fluctuations of sampling. That is why it is universally used.
(vi) It is amenable to mathematical treatment.

The disadvantages of the mean are:

(i) It is greatly affected by extreme values in the data.
(ii) It gives sometimes fallacious conclusions.
(iii) In a highly skewed distribution, the mean is not an appropriate measure of average.
(iv) If the grouped data have "open-end" classes, mean cannot be calculated without assuming the limits.

3.11.2 The Geometric Mean. The advantages of the geometric mean are:

(i) It is rigorously defined by a mathematical formula.
(ii) It is based on all observed values.
(iii) It is amenable to mathematical treatment in certain cases.
(iv) It gives equal weightage to all the observations.
(v) It is not much affected by sampling variability.
(vi) It is an appropriate type of average to be used in case rates of change or ratios are to be averaged.

Its disadvantages are:

(i) It is neither easy to calculate nor to understand.
(ii) It vanishes if any observation is zero.
(iii) In case of negative values, it cannot be computed at all.

3.11.3 The Harmonic Mean. The advantages of the harmonic mean are:

(i) It is rigorously defined by a mathematical formula.
(ii) It is based on all the observations in the data.
(iii) It is amenable to mathematical treatment.
(iv) It is not much affected by sampling variability.

(v) It is an appropriate type for averaging rates and ratios.

The disadvantages of the harmonic mean are:

(i) It is not readily understood.

(ii) It cannot be calculated, if any one of the observations is zero.

(iii) It gives too much weightage to the smaller observations.

3.11.4 The Median. The advantages of the median are:

(i) It is easily calculated and understood.

(ii) It is located even when the values are not capable of quantitative measurement.

(iii) It is not affected by extreme values. It can be computed even when a frequency distribution involves "open-end" classes like those of income and prices.

(iv) In a highly skewed distribution, median is an appropriate average to use.

The median has the following disadvantages:

(i) It is not rigorously defined.

(ii) It is not capable of lending itself to further statistical treatment.

(iii) It necessitates the arrangement of data into an array which can be tedious and time consuming for a large body of data.

3.11.5 The Mode. The advantages of the mode are:

(i) It is simply defined and easily calculated. In many cases, it is extremely easy to locate the mode.

(ii) It is not affected by abnormally large or small observations.

(iii) It can be determined for both the quantitative and the qualitative data.

The disadvantages of the mode are:

(i) It is not rigorously defined.

(ii) It is often indeterminate and indefinite.

(iii) It is not based on all the observations made.

(iv) It is not capable of lending itself to further statistical treatment.

(v) When the distribution consists of a small number of values, the mode may not exist.

EXERCISES

3.1 What is a statistical average? Name the important types of averages. Discuss the advantages and disadvantages of each average. (P.U., B.A. (Hons.) 1960)

3.2 What is a measure of "central tendency"? What is the purpose served by it? What are its desirable qualities?

3.3 What are the principal criteria for a satisfactory average? State giving reasons the circumstances in which it would be preferable

to use (i) the mean, (ii) the median (iii) the mode, (iv) geometric mean and (v) harmonic mean.

3.4 What criteria do you apply to judge the merits of an average? Discuss the merits and demerits of the different averages in common use with special reference to these criteria.

3.5 In what circumstances would you consider the Arithmetic mean, the Geometric mean and the Harmonic mean respectively, the most suitable statistic to describe the central tendency of distributions? (P.U., B.A./B.Sc. 1989)

3.6 What are the different measures of central tendency? Describe the manner of computation of any *three* of them with suitable illustrations. (P.U., M.A. Econ. 1967)

3.7 Define weighted average and explain how it differs from simple mean. Give the method of its computation and discuss the use of weighted mean in Statistics. (P.U., M.A. Econ. 1974)

3.8 What is the median? What are its advantages and disadvantages? Give reasons why the statistician usually prefers the arithmetic mean to the median. (P.U., M.A. Econ. 1981)

3.9 Define, and explain how to compute, the following quantities for a grouped distribution: $\bar{x}$, Q_1, Q_3, D_7 and Mode.

3.10 Define the arithmetic mean, the mode and the median. Discuss the relationship of these three measures of location in a skewed distribution. State the chief advantages of the arithmetic mean as a form of average.

3.11 Define Mean, Median and Mode. What are their advantages and limitations in the analysis of data? Give various methods of calculating Arithmetic mean, with illustrations.
(P.U., B.A./B.Sc. 1958, 1960)

3.12 (a) Define Mean, Median and Mode. Give an empirical relation between them. Does this relation give correct value for the mode?

(b) Criticise the following statements:

(i) An average does not reveal all the information about the data.

(ii) The median is described as *the value of the average* rather than the *average value*.

3.13 Comment on the following statements:

(i) The depth of a river at four different points is 2, 7, 5, 6 feet respectively. The average depth is 5 feet. Therefore all the people with heights above 5 feet can cross it.

(ii) The average marks of one class of students are 30. Therefore every student is hopeless.

(iii) The average income of a king and his household servants is £20,000 per month, therefore all the household servants must be fabulously paid.

(iv) On an average, the number of accidents occurring in the middle of the road are 5 per thousand. The number of deaths at other places is 30 per thousand. Therefore, it is safer to walk in the middle of the road.

(v) In a country, 2,000 vaccinated persons died. Therefore vaccination is useless.

3.14 Define the mean, the median and the mode of a frequency distribution.

It is commonly true that the median lies between the other two measures and is approximately twice as far from the mode as from the mean. State with reasons, whether you expect this relationship to hold, and which of the three statistics is likely to be the most useful single statistic for each of the following distributions:

(i) The annual earnings of employed males in Pakistan.

(ii) The percentage of sky, to the nearest 10 per cent, covered by cloud at Karachi at mid-day.

(iii) The exact length of rods cut to a standard size by machine.

(P.C.S., 1971)

3.15 (a) Define arithmetic mean and describe its properties.

(b) If the arithmetic mean of n numbers $x_1, x_2, \ldots, x_n$ is M and A is any arbitrary number, then show that

$$\sum(x_i - A)^2 = \sum (x_i - M)^2 + n(M - A)^2$$ (P.U., B.A./B.Sc. 1977, 82)

(c) A distribution consists of three components with frequencies 3, 4 and 5, and having means 2, 5.5 and 10. Find the mean of the combined distribution. (P.U., B.A./B.Sc. 1977)

3.16 (a) State the properties of the arithmetic mean.

(b) Show that $\sum(x_i - a)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - a)^2$. In other words, show that the sum of squares about $a = \bar{x}$ is smallest.

(c) A distribution consists of 3 components with frequencies 45, 40 and 65, having their means 2, 2.5 and 2 respectively. Show that the mean of the combined distribution is 2.13 approximately.

3.17 (a) The number of cars crossing a certain bridge in a big city in 10 intervals of five minutes each were recorded as follows:

25, 15, 18, 30, 20, 20, 12, 9, 16, 15

Calculate (i) the arithmetic mean, (ii) the median and (iii) the geometric mean.

(b) Explain why the mean calculated for a set of ungrouped data might differ from the mean if the same data were grouped into a frequency distribution.

3.18 (a) Define Arithmetic mean, Geometric mean and Harmonic mean; and prove that for any two positive numbers a and b,

$$A.M. \geq G.M. \geq H.M.$$ (P.U. B.A./B.Sc. 1991)

(b) The monthly incomes of ten families in rupees in a certain locality are given below:

Family: A B C D E F G H I J

Income (Rs.): 85, 70, 10, 75, 500, 8, 42, 250, 40, 36.

Calculate the arithmetic mean, the geometric mean and the harmonic mean of the above incomes. Which one of the above three averages represents the above figures best?

3.19 Calculate the arithmetic mean, the geometric mean and the harmonic mean of the annual incomes of fifteen families as given below:

Rs. 60, 80, 90, 96, 120, 150, 200, 360, 480, 520, 1060, 1200, 1450, 2500, 7200.

3.20 (a) In a company having 80 employees, 60 earn Rs. 3.00 per hour and 20 earn Rs. 2.00 per hour. (i) Determine the mean earnings per hour. (ii) Do you consider this mean hourly wage to be typical? (P.U., B.A./B.Sc. 1980-S)

(b) An examination candidate's percentages are: English, 73; French, 82; Mathematics, 57; Science, 62; History, 60. Find the candidate's weighted mean if weights of 4, 3, 3, 1, 1 respectively are allotted to the subjects.

3.21 Find (i) the simple average of prices in column 2 and (ii) the weighted average, using the quantities in column 3 as weights, and explain the difference between the two results.

(1) Piece goods	(2) Price per metre (Rs.)	(3) Quantity (millions metres)
Unbleached	8.37	286
Bleached	9.50	255
Printed flags	9.16	64
Other sorts	9.84	172
Dyed in piece	13.65	165
Of dyed yarn	11.95	80

3.22 The following are the monthly salaries in rupees of 30 employees of a firm:

139	126	114	100	88	62	77	99	103	108
144	129	148	63	69	148	132	118	142	116
123	104	95	80	85	106	123	133	140	134

The firm gave bonuses of Rs. 10, 15, 20, 25, 30 and 35 for individuals in the respective salary groups: exceeding 60 but not exceeding 75, exceeding 75 but not exceeding 90 and so on upto exceeding 135 but not exceeding 150. Find the average bonus paid per employee. (P.U., M.A. Econ. 1974; B.Z.U. M.A., Econ. 1991)

3.23 The following table shows the age distribution of 1,143 horses.

Age (years)	Number of horses (f_i)	Average age ($\bar{x}_i$)
1 – 4	12	2.7
5 – 9	223	7.6
10 – 14	435	12.0
15 – 19	272	16.3
20 – 24	161	20.8
25 – 29	34	25.8
30 and over	6	31.8

Compute the average age of these horses (a) from the first two columns of the table by the usual short method, (b) from the last two columns by weighting the group averages by the number of horses in the groups. Compare the two results. Which one is more nearly the real average age?

3.24 Find the arithmetic and geometric means of the series 1, 2, 4, 8, 16, ..., 2^n. Find also the harmonic mean. (P.U. D. St. 1960)

3.25 Find (i) arithmetic man, (ii) geometric mean, and (iii) harmonic mean of the series 1, 3, 9, 27, 81, .., 3^n. (P.U., B.A/B.Sc. 1973, 82)

3.26 (a) Define Geometric mean and describe its advantages and disadvantage.

(b) Given two sets, each of n positive values, $x_{11}, x_{12}, ..., x_{1n}; x_{21}, x_{22}, ..., x_{2n}$; prove that the geometric mean of the ratios of corresponding values in the two sets is equal to the ratio of the geometric mans of the two sets. (P.U. B.A/B.Sc. 1987)

Hint. Let a ratio be defined as $X = \frac{X_1}{X_2}$.

Then $\log X = \log X_1 - \log X_2$

Sum for all pairs of X_1's and X_2's.

Hence $G = \frac{G_1}{G_2}$.

3.27 A man gets a rise of 10% in salary at the end of his fist year of service, and further rises of 20% and 25% at the end of the second and third years respectively, the rise in each case being calculated on his salary at the beginning of the year. To what annual percentage increase is this equivalent?

3.28 (a) Define Harmonic mean. How does it differ from arithmetic mean? What are its advantages and disadvantages?

(b) A man travels from A to B at average speed of 30 miles per hour and returns from B to A along the same route at an average speed of 60 miles per hour. Find the average speed of the entire journey. (P.U., B.A./B.Sc. 1972)

(c) Find out the average speed of person who rides the first mile at the rate of 8 miles an hour, the next mile at the rate of 7.5 miles an hour and the third mile at the rate of 5.5 miles an hour. (P.U., B.A./B.Sc. 1981)

3.29 (a) A bus travelling 200 kilometres has 10 stages at equal intervals. The speed of the bus in the various stages was observed to be 10, 15, 20, 25, 20, 30, 40, 50, 30, 40 kilometres per hour. Find the average speed at which the bus travels.

(b) Find out the average rate of (i) motion in the case of a person who rides the first mile at the rate of 10 miles an hour, the next mile at the rate of 8 miles per hour, and the third mile at the rate of 6 miles per hour; (ii) increase in population, which in the first decade has increased 20%, in the next 25% and in the third 44%. (P.U., B.A. (Part I) 1962-S)

3.30 Find the geometric mean and the harmonic mean of the following frequency distribution:

Weekly Income (Rs.)	35–39	40–44	45–49	50–54	55–59	60–64	65–69
No. of workers	15	13	17	29	11	10	5

(P.U., B.A. (Hons. in Econ.) 1966)

3.31 Calculate the geometric and the harmonic means for the distribution given below:

Variable	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	2	5	7	13	21	16	8	3

(P.U., M.A. Econ. 1970-S)

3.32 Find the mean or median, whichever you think more suitable, in each of the following:

(i) Salaries of 5 men in an industrial concern:

Rs. 950, Rs. 2100, Rs. 1500, Rs. 100, Rs. 10,000.

(ii) Heights of 6 boys: 64", 65", 65", 66", 66", 67".

(iii) Handicaps of four golfers: 4, 18, 18, 20.

3.33 The following data relate to sizes of shoes sold at a store during a given week. Find the median of the shoes. Also calculate the quartiles, the 7th decile and the 64th percentile.

Size of Shoes	5	$5\frac{1}{2}$	6	$6\frac{1}{2}$	7	$7\frac{1}{2}$	8	$8\frac{1}{2}$	9	$9\frac{1}{2}$
No. of Pairs	2	5	15	30	60	40	23	11	4	1

(P.U., B.A. (Hons.) 1962)

3.34 Calculate the Mean, Median and Modal numbers of persons per house from the data:

No. of persons per house	1	2	3	4	5	6	7	8	9	10
No. of houses	26	113	120	95	60	42	21	14	5	4

(P.U., B.A. (Hons.) 1969)

3.35 Draw (i) a Histogram and (ii) an Ogive from the following data:

Daily wages (Rs.)	4–6	6–8	8–10	10–12	12–14	14–16
No. of employees	13	111	152	105	19	7

Find approximate value of the median from the Ogive and check your answer by calculation. (B.I.S.E. Sargodha, 1969-S)

3.36 Estimate graphically and by formula the median and quartile ages of head of household from the following distribution:

Age of head (yrs)	Number of households
under 25	44
25 and under 30	79
30 and under 40	152
40 and under 50	122
50 and under 60	141
60 and under 65	100
65 and under 70	58
70 and under 75	32
75 and under 85	28

3.37 Compute the median and quartiles of the following distribution of heights and check the results on a graph.

Heights (inches)	57.5–,	60.0–,	62.5–,	65.0–,	67.5–,	70.0–,	72.5–
Number	6	26	190	281	412	127	38

(P.U., M.A. Econ., 1969)

3.38 Explain when median is more representative than mean. Calculate the median of the following distribution.

Class	Number	Class	Number	Class	Number
100–104	4	125–129	298	150–154	260
105–109	14	130–134	380	155–159	128
110–114	60	135–139	450	160–164	66
115–119	138	140–144	500	165–169	28
120–124	236	145–149	430	170–174	12

(P.U., B.A./B.Sc. 1960)

3.39 The frequency distribution of a group of persons according to age is given below:

Age in years	<1	1–4	5–9	10–19	20–29	30–39	40–59	60–79
No. of persons	5	10	11	12	22	18	8	7

Calculate the Mean and the Median ages of the distribution.

3.40 (a) Describe the merits and demerits of mean and median.

(b) Calculate the median, the upper and lower quartiles from the following data: Also draw a box plot.

Class-Interval	Frequency
under 25	222
25 – 29	405
30 – 34	508
35 – 39	520
40 – 44	525
45 – 49	490
50 – 54	457
55 – 59	416
60 and over	166

(P.U., B.A./B.Sc., 1968)

3.41 The following distribution shows Kilowatt-Hours of Electricity used in one month by 75 residential consumers in a certain locality of Lahore.

Consumption in kilowatt hours	5–24	25–44	45–64	65–84	85–104	105–124	125–144	145–164
No. of consumers	4	6	14	22	14	5	7	3

Estimate the mean, the median and the two quartiles.

3.42 The yields of grain (x lb) from 500 small plots are grouped in classes with a common class-interval (0.2 lb.) in the table below, the values of x given being the midvalues of the classes. Show that the mean of the distribution is 3.95 lb.; the median is 3.95 lb.; and quartiles are 3.63 lb. and 4.28 lb.

x	f	x	f
2.8	4	4.2	69
3.0	15	4.4	59
3.2	20	4.6	35
3.4	47	4.8	10
3.6	63	5.0	8
3.8	78	5.2	4
4.0	88	Total	500

3.43 The weights in milligrams of 2538 seeds of the long leef pine were as follows:

Weight (milligrams)	Number of Seeds	Weight (milligrams	Number of Seeds
10–24.9	16	85–99.9	655
25–39.9	68	100–114.9	803
40–54.9	204	115–129.9	294
55–69.9	233	130–144.9	21
70–84.9	240	145–159.9	4

(a) Find the average weight, the median weight and the most common weight (mode) of the seeds.

(b) Find the first and third quartiles. Find the third decile and the 45*th* percentile.

(c) Explain your answers as you would to a person who had never studied statistics.

3.44 In a group of 500 wage-earners, the weekly wages of 4% were under Rs. 60 and those of 15% were under Rs. 62.50. 15% of the workers earned Rs. 95 and over, and 5% of them got Rs. 100 and over.

The median and quartile wages were Rs. 82.25, Rs. 72.75 and Rs. 90.50; the fourth and sixth decile wages were Rs. 78.75 and Rs. 85.25 respectively.

Put the above information in the form of a frequency distribution and estimate the mean wage of the 500 wage-earners therefrom.

Hint. First put the information in the form of a cumulative frequency table.

3.45 (a) Describe the advantages and disadvantages of the mean, the median and the mode. Explain the empirical relation between them. (P.U. B.A/B.Sc. 1971)

(b) The weight of the 40 male students at a university are given in the following frequency table:

Weight	118–126	127–135	136–144	145–153	154–162	163–171	172–180
Frequency	3	5	9	12	5	4	2

Calculate the mean, median and the mode.

(P.U., B.A./B.Sc. 1969)

3.46 The following table shows the distribution of the maximum loads in short tons supported by certain cables produced by a company.

Max. loads (Short tons)	9.8–10.2	10.3–10.7	10.8–11.2	11.3–11.7	11.8–12.2	12.3–12.7
No. of cables	7	12	17	14	6	4

Determine the mean, the median and the mode.

3.47 The following is the distribution of wages per thousand employees in a certain factory.

Daily wages (Rs.)	22	24	26	28	30	32	34	36	38	40	42	44
Number of employees	3	13	43	102	175	220	204	139	69	25	6	1

Calculate the Modal and Median wages and explain why there is a difference between the two.

3.48 (a) Define the mode of a frequency distribution. How does it compare with other types of averages?

(b) Write down the empirical relation between mean, median, and mode for unimodal distributions of moderate asymmetry. Illustrate graphically the relative positions of the mean, median and mode for frequency curves which are skewed to the right and to the left. (P.U., B.A/B.Sc. 1972,80-S)

(c) For a certain frequency distribution, the mean was 40.5 and median 36. Find the mode approximately using the formula connecting the three. (P.U. B.A./B.Sc. Optional, 1971-S)

3.49 (a) What types of averages would be suitable for the following cases? Give reasons.

(i) Size of agricultural holdings.

(ii) Heights of students.

(iii) Marks obtained in any examination.

(iv) Income of workers in a factory.

(v) Per capita income in Pakistan.

(vi) Comparison of intelligence.

(vii) Volumes of sales of ready-made shirts, shoes and collars.

(viii) Number of petals of flowers.

(b) What measures of central tendency would you recommend for the following cases. Give reasons in support of your answer.

(i) Symmetrical Distribution.

(ii) A J-shaped Distribution.

(iii) Distribution having "open-end" classes at the end of the classes.

(iv) Frequency distribution of a quantitative variable.

(P.U. M.A Econ., 1977)

3.50 (a) A distribution $x_1, x_2, \ldots, x_r, \ldots, x_k$ with frequencies $f_1, f_2, \ldots, f_r, \ldots, f_k$ is transformed into the distribution $X_1, X_2, \ldots, X_k$, with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constant. Show that the mean, mode and median of the new distribution are given in terms of those of the first distribution by the same transformation.

(b) A distribution has values of the variable $x_1, x_2, \ldots, x_k$ with corresponding frequencies $f_1, f_2, \ldots, f_k$. A new distribution with the same frequencies is formed by taking $X_r = 2x_r - 3$ for values of r ($r = 1, 2, \ldots, k$). If the values of the mean, median and mode of the original distribution are a, b and c respectively, what are these values of the new distribution?

(P.U., B.A/B.Sc. 1980)

❖❖❖❖❖❖❖❖❖

$L = x_i - \frac{h}{2}$

$h = x_i - x_{i-1}$

$L = x_i - \frac{H}{2}$

$H = x_i - x_{i-1}$