UNIT 4 MEASURES OF CENTRAL TENDENCY

Structure

4.0 OBJECTIVES

After going through this unit, you will be able to:

- compute numerical quantities that measure the central tendency of a set of data such as, mean, median, mode, geometric mean and harmonic mean, and
- use these measures.

4.1 INTRODUCTION

In the previous Unit we had discussed about condensation of raw data by grouping them into a few class intervals and presenting in the form of a table or diagram. Such tables or diagrams provide a rough idea of the distribution of observations. Often we need to compare between distributions. In such situations it is difficult to compare tables or diagrams simply by looking at them. It is much more convenient and useful for comparison if we could find out a single numerical value for describing the data.

Measures of Central Tendency (or Location) constitute one of the major statistics designed for this purpose. There are five main measures of central tendency. These are Arithmetic Mean, Geometric Mean, Harmonic Mean, Median and Mode. You will learn about each one of these measures below.

4.2 MEASURES OF CENTRAL TENDENCY

In frequency distributions of observations discussed in Unit 3 we notice that the observations tend to cluster around a central value. This phenomenon of clustering around a central value in a frequency distribution is called '*Central Tendency*'. Thus, it is of interest to locate such a value around which clustering of observations takes place. There are several measures of central tendency (or location) of a frequency distribution. These measures produce numbers that summarise a frequency distribution in terms of one of its properties, namely, central tendency.

4.2.1 Arithmetic Mean

The *average* or the *arithmetic mean*, or simply the *mean* when there is no ambiguity, is the most common measure of central tendency. It is defined as the sum total of all values in the sample divided by the number of observations. It is denoted by a bar above the symbol of the variable being averaged. Thus $\overline{X}$ stands for the mean of X-values in the sample. If in a sample a particular X-value, say X_i occurs with frequency f_i ($i = 1, 2, \ldots n$), its contribution to the total of X-values is $f_i X_i$. Thus, one can compute the mean of X-values by

$$\overline{X} = \frac{1}{N}(f_1X_1 + f_2X_2 + \cdots + f_nX_n) = \frac{\sum_{i=1}^{n} f_iX_i}{N}, \qquad \text{where } N = \sum_{i=1}^{n} f_i.$$

When observations are classified into class intervals, as for continuous variables, individual observations falling into a class interval are not separately identifiable and the contribution of the individual observation from a class interval to the total cannot be calculated. To avoid this difficulty, it is assumed that every observation falling into a class interval has a value equal to the *mid-point* into which these observations fall. Such a procedure will not give the exact mean had one computed it from raw data and may require what is called corrections for grouping.

Example 4.1: Compute the mean for discrete frequency distribution of Table 4.1.

Table 4.1
Frequency distribution of 100 households by size

Household Size (X_i)	Frequency (f_i)
1	3
2	16
3	25
4	33
5	12
6	7
7	2
8	2
Total	100

Let us compute the arithmetic mean of the data given in the above table.

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N} = \frac{1\times 3 + 2\times 16 + 3\times 25 + 4\times 33 + 5\times 12 + 6\times 7 + 7\times 2 + 8\times 2}{100} = \frac{374}{100} = 3.74.$$

Thus, mean household size based on 100 households is 3.74.

Example 4.2: Compute the mean for grouped frequency distribution of Table 4.2.

Table 4.2
Frequency distribution of 100 households by average monthly household expenditure on food

Expenditure class (Rs.)	Frequency
262.5 - 286.5	1
286.5 - 310.5	14
310.5 - 334.5	16
334.5 - 358.5	28
358.5 - 382.5	26
382.5 - 406.5	15
Total	**100**

For computation of the mean we have to construct table as given below.

Class interval (Rs.) (0)	Mid-point (X_i) (1)	Frequency (f_i) (2)	$f_i X_i$ (3)
262.5 - 286.5	274.5	1	274.5
286.5 - 310.5	298.5	14	4179.0
310.5 - 334.5	322.5	16	5160.0
334.5 - 358.5	346.5	28	9702.0
358.5 - 382.5	370.5	26	9633.0
382.5 - 406.5	394.5	15	5917.5
Total		**100**	**34866.0**

Thus, mean of monthly average household expenditure on food is

$$\bar{X} = \frac{34866}{100} = \text{Rs. } 348.66.$$

One may note from the above example that to find column (3) one needs to multiply the corresponding values of column (1) and (2), and often hand computations are long for each multiplication. These computations can be simplified, particularly when successive column (1) values are equidistant (but applicable otherwise also), by making the following simple transformation.

For $i = 1, 2, \ldots, n$

$$u_i = \frac{X_i - A}{h} \quad \text{i.e.,} \quad X_i = A + hu_i \quad \text{and so} \quad \bar{X} = A + h\bar{u}.$$

Often A is called the 'assumed mean' and $h\bar{u}$ as its correction to get $\bar{X}$. Choice of A and h are made so that computation of $\bar{u}$ becomes simple. Usually A is taken as that X value for which the frequency is largest. For equidistant successive X-values in column (1), h may be taken as the difference between two successive X-values. For equal length class intervals, the difference between successive mid-points is the same as the length of each class interval.

We will explain this method by re-computing the mean of the monthly average household food expenditure data given in Table 4.2. We construct Table 4.3 by using A and h as explained below.

We define A = Mid-point of the class with largest frequency = 346.5 and
h = Common length of each class interval = 24.

Thus, $u_i = \dfrac{X_i - 346.5}{24}$

Table 4.3
Computation of mean of data of Table 4.2

Class interval (Rs.)	Mid-point (X_i)	$u_i = \dfrac{X_i - 346.5}{24}$	Frequency (f_i)	$f_i u_i$
262.5 - 286.5	274.5	– 3	1	– 3
286.5 - 310.5	298.5	– 2	14	– 28
310.5 - 334.5	322.5	– 1	16	– 16
334.5 - 358.5	346.5	0	28	0
358.5 - 382.5	370.5	1	26	26
382.5 - 406.5	394.5	2	15	30
Total			100	9

We find out that

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{n} f_i u_i = \frac{1}{100} \times 9 = \frac{9}{100}$$

Thus, $\bar{X} = A + h \times \bar{u} = 346.5 + 24 \times \dfrac{9}{100}$ = Rs.348.66 as was computed earlier.

Properties of Arithmetic Mean

1) *The algebraic sum of deviations of a given set of observations is zero when taken from the arithmetic mean.*

   Let $X_1, X_2, \ldots . X_n$ be n observations with respective frequencies as f_1, $f_2, \ldots . f_n$. Mathematically, this property implies that $\sum_{i=1}^{n} f_i (X_i - \bar{X}) = 0$,

   where $X_i - \bar{X}$ is the deviation of i^{th} observation from mean.

   To prove the above property, we write

$$\sum_{i=1}^{n} f_i(X_i - \overline{X}) = \sum_{i=1}^{n} f_i X_i - \overline{X}\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i X_i - n.\overline{X} = 0.$$

Hence the result.

2) *The sum of squares of deviations of a given set of observations is minimum when taken from the arithmetic mean.*

Mathematically, this property implies that for any arbitrarily chosen origin, A,

$S = \sum_{i=1}^{n} f_i(X_i - A)^2$ is minimum when $A = \overline{X}$.

To prove this property, we note that the magnitude of S will depend upon the selected value of A. Thus, we can say that S is a function of A. We want to find that value of A for which S is minimum. Using calculus, this value is given by the equation $\frac{dS}{dA} = 0$ such that $\frac{d^2S}{dA^2} > 0$.

(Remember that the value of a function is minimum when first derivative is zero and second derivative is positive.)

Differentiating S with respect to A and equating to zero, we get

$$\frac{dS}{dA} = -2\sum_{i=1}^{n} f_i(X_i - A) = 0$$

This implies that

$$\sum_{i=1}^{n} f_i X_i - A\sum_{i=1}^{n} f_i = 0 \quad \text{or} \quad A = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \overline{X}.$$

Further, it can be shown that $\frac{d^2S}{dA^2} > 0$ when $A = \overline{X}$.

4.2.2 Median

Median of a distribution locates a central point which divides a distribution into two equal halves, i.e., it is the middle most value among a set of observations. Let us start with examples in a discrete case. Consider a data set having 5 distinct observations: 2, 4, 9, 12, 19 (arranged in ascending order). Here 9 is the middle most value since an equal number of observations are to its left and to its right. Thus, 9 is the median of the above observations. Consider another data set having 6 distinct observations: 3, 8, 15, 25, 35, 43. Here any point between 15 and 25 has the property that equal number of observations are to its left and to its right. Any point in the interval 15 to 25 may be used as a median. Conventionally we take the middle point of such an interval to define median uniquely. Thus 20 is the median of 3, 8, 15, 25, 35, 43.

When a data set has non-distinct observations — a situation more common in practice — difficulties may arise. In such situations, it may not be always possible to locate the middle most value or the central point that divides the distribution into two equal halves, For example, in the case of the data set having

5 observations 2, 9, 9, 12, 19 the value 9 is repeated twice. Thus, a formal definition of median is needed to overcome such difficulties.

A median of a distribution is a point or a central value such that ***at least*** *50% of the observations are less than or equal to it and* ***at least*** *50% of the observations are greater than or equal to it.* With this definition of median and the convention of taking the middle point of a class in which each point i a median, median of a distribution can always be specified uniquely. Thus, median of observations 2, 9, 9, 12, 19 is 9 because 3 of the 5 observations (60%) are less than or equal to 9 and 4 of the 5 observations (80%) are greater than or equal to 9.

Let us find out the median household size from the frequency distribution in Table 4.1. We notice that 77 (out of 100) households have family size of less than or equal to 4 and 56 households have family size of more than or equal to 4. Thus median family size in this case is 4.

Median for a grouped frequency distribution of a continuous variable is easier to understand if one looks at the associated histogram with height of a rectangle equal to the frequency density, $\frac{f}{h}$, of the class. In such a histogram, the area of a rectangle gives the frequency of the corresponding class. The median, in this case, is a point in one of the classes such that the areas to its left and to its right are 50% each. First step is to locate the class, up to the right boundary of which the total area is at least 50% (called the *median class*). Then the median is computed by adding, to the lower boundary value of this class, the length of a part of this class interval in proportion to the frequency needed to achieve 50%. A convenient method of finding out the median class is to compute the cumulative frequency (discussed in Unit 2, Section 2.3.3) and identifying the class interval in which the $\frac{N}{2}$th observation lies.

This method of computing median is illustrated through the data on monthly average household expenditure on food given in Table 4.2.

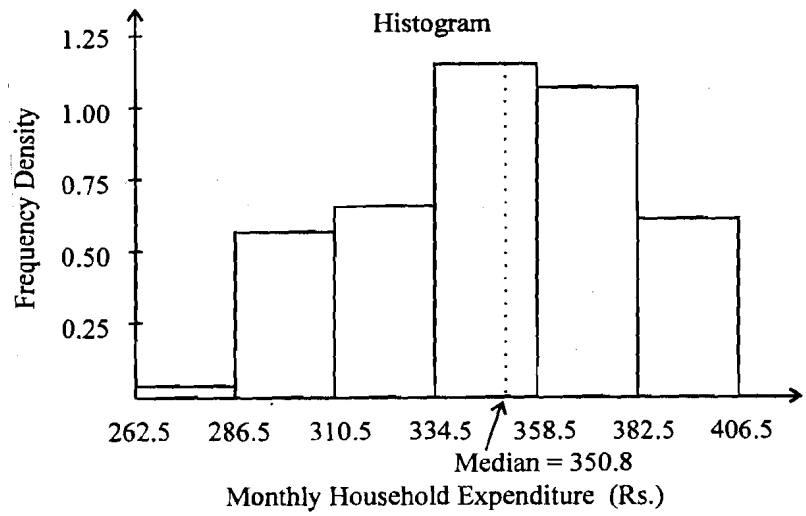

Fig. 4.1

Area up to the class boundary 334.5 is 31 and upto 358.5 is 59. Hence the median lies in the class 334.5 - 358.5. We now want to find a point in this class so that the area from 334.5 to the point is (50 - 31) = 19, where area up to 334.5 is 31. Since the rectangle over the interval 334.5 - 358.5 has an area of 28, and is of length 24, to get an area of 19 we need $\frac{19}{28}$th part of 24. This works out to be $\frac{19}{28} \times 24 = 16.3$. Thus the median is 334.5 + 16.3 = 350.8. Note also that the area in the class 350.8 to 358.5 is 28 - 19 = 9 and to the right of 350.8 is 9 + 41 = 50, as it should be.

Based on the above procedure, we can write a formula for the computation of median.

$$M_d = l_m + \frac{\frac{N}{2} - C}{f_m} \times h, \quad \text{where}$$

l_m is the lower limit of the median class, i.e., the class in which median lies,

N is the total frequency,

C is the cumulative frequency of classes preceding the median class (note that $C = 31$ in the above example),

f_m is the frequency of median class, and

h is the width of median class.

4.2.3 Mode

As has been pointed out earlier, often observations tend to cluster around a central value. A simple measure of this phenomenon is called mode.

Mode or modal value of a discrete variable is defined as that value of the variable for which frequency is maximum. Mode, however, is not the majority, i.e., it does not imply that most (50% or more) of the observations have the modal value.

From Table 4.1 we find that the mode or modal value of household size is 4 as this value occurs with largest frequency of 33 among 100 households.

There are, however, data sets when mode cannot be defined uniquely, i.e., the distribution has multiple mode. Raw data with 7 hypothetical observations with values 4, 3, 4, 1, 2, 5, 3, have two modes, 3 and 4. Distributions having two modes are called *bimodal distributions*, though the frequently encountered distributions have only one mode or are *unimodal*.

For observations on the continuous variable, like monthly household expenditure on food, no two observations are likely to have same value and so mode is not a meaningful measure of such raw data. However, central tendency comes out clearly when these raw data are grouped into various class intervals. For grouped data *modal class* is defined as the class having largest frequency. Since large class intervals are likely to include large number of observations and smaller class intervals are likely to have few observations, definition of modal class is meaningful only when class intervals have equal length.

For discrete data it is easier to find out the mode. But in the case of continuous data computation of the mode is done by the following formula:

$$M_o = l_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h, \text{ where}$$

l_m is the lower limit of the modal class, i.e., the class in which mode lies,

$\Delta_1 (= f_m - f_{m-1})$ is the difference of the frequencies of the modal class and its preceding class,

$\Delta_2 (= f_m - f_{m+1})$ is the difference of the frequencies of the modal class and its following class, and

h is the width of the modal class.

Let us look back to Table 4.2. Here modal class is 334.5 - 358.5 as it has the highest frequency, 28.

Thus, $l_m = 334.5$, $\Delta_1 = 28 - 16 = 12$, $\Delta_2 = 28 - 26 = 2$ and $h = 24$.

Hence $M_o = 334.5 + \frac{12}{12+2} \times 24 = 355.07$.

Mode is a useful measure of central tendency when a frequency distribution has a strong peak and it is particularly useless when a frequency distribution is almost flat.

Check Your Progress 1

1) The frequency distribution of a family size for 250 families in a ward of an industrial town is given below:

Family Size	Frequency
1	4
2	22
3	25
4	45
5	52
6	41
7	36
8	15
9	7
10	3
Total	**250**

Find the mean, median and mode.

.....

.....

.....

.....

.....

2) Compute the mean, median and mode for the following frequency distribution.

Frequency Distribution of IQ for 309 Six-Year old Children

I.Q.	Frequency
160 - 169	2
150 - 159	3
140 - 149	7
130 - 139	19
120 - 129	37
110 - 119	79
100 - 109	69
90 - 99	65
80 - 89	17
70 - 79	5
60 - 69	3
50 - 59	2
40 - 49	1
Total	**309**

.....

.....

.....

.....

.....

.....

.....

.....

4.3 OTHER MEASURES OF CENTRAL TENDENCY

Besides the arithmetic mean, median and mode there are other averages which are relatively unimportant but may be appropriate in particular situations. These are Geometric Mean and Harmonic Mean. We will discuss these in Section 4.3.1.

Often we see that all the observations do not have equal importance. In such cases we need to give differential importance to different items. Here we use weighted means — arithmetic, geometric or harmonic — instead of simple means. This we will discuss in Section 4.3.2.

4.3.1 Geometric Mean and Harmonic Mean

Often we have to deal with data that are time dependent, i.e., time series data which are unlike one-time data of Tables 4.1 and 4.2. For time dependent data, it is often of interest to find the pattern of change over time. Consider the following two data sets.

Set I : 1000 1100 1200 1300 1400 1500 1600
Set II : 1100 1210 1331 1464 1611 1772 1949

First set looks like basic salary (in Rs.) of an employee for 7 years with annual increment of Rs. 100 per year.

Second set looks more like his gross salary (in Rs.). Annual increase in the two sets are given below.

Set I : 100 100 100 100 100 100
Set II : 110 121 133 147 161 177

Arithmetic mean of annual increase is 100 for Set I and 141.5 for Set II. On the basis of these average annual increases, if one works-out figures for the two sets, starting from the initial values, one would get the following.

Set I : 1000 1100 1200 1300 1400 1500 1600
Set II : 1100 1241.5 1383 1524.5 1666 1807.5 1949

That the use of arithmetic mean has worked well for Set I and not for Set II is because the progression of original numbers in the two sets are different. In Set I, increment has been a fixed quantum whereas in Set II, figures have increased at a fixed rate. Fixed quantum of increase is called arithmetic progression and arithmetic mean is appropriate to describe the increase. Fixed rate of increase is called geometric progression and geometric mean is most appropriate to describe the increase.

For n numbers $X_1, X_2, \ldots X_n$ geometric mean (GM) is defined as the nth root of the product of these n numbers, i.e.,

$$GM = (X_1 X_2 \cdots X_n)^{\frac{1}{n}} = \left[\prod_{i=1}^{n} X_i\right]^{\frac{1}{n}}$$

Clearly, GM is not defined unless all the n numbers are positive. By taking logarithm of GM, one has

$$\log GM = \left(\frac{1}{n}\right)(\log X_1 + \log X_2 + \cdots + \log X_n) = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

which shows now GM can be computed by using a log-table. Anti-logarithm of the arithmetic mean of log X values is GM. For the second data set, gross salary increased at the rate of 11% every year. In practice, however, increase/decrease will not be at a fixed rate over the years; and it is meaningful to talk about average rate because fixed rate situation is rare. In general, GM is more appropriate average for percentage (or proportionate) rates of change than arithmetic mean as in the case of rise in various price indices, cost of living indices, etc.

Finally, we discuss about another measure of location called harmonic mean (HM). This mean comes naturally in many situations as in the following illustration. A stockist stocks Rs. 5000 worth of an item at the beginning of every month. Unit rate (in Rs.) of the item for five successive months had been 10.75, 11.80, 14.00, 11.45

and 12.00. The stockist wants to find average rate per unit of the item he has stocked for five months. Computation is presented below :

Month	Amount Spent (Rs.)	Unit Rate (Rs.)
1	5000	10.75
2	5000	11.80
3	5000	14.00
4	5000	11.45
5	5000	12.00
Total	**25000**	

$$\text{Average price (in Rs.) of his entire stock} = \frac{\text{Total Money Spent}}{\text{Total Quantity Purchased}}$$

$$= \frac{5 \times 5000}{\frac{5000}{10.75} + \frac{5000}{11.80} + \frac{5000}{14.00} + \frac{5000}{11.45} + \frac{5000}{12.00}}$$

$$= \frac{5}{\frac{1}{10.75} + \frac{1}{11.80} + \frac{1}{14.00} + \frac{1}{11.45} + \frac{1}{12.00}}$$

$$= \frac{1}{\frac{1}{5}\left(\frac{1}{10.75} + \frac{1}{11.80} + \frac{1}{14.00} + \frac{1}{11.45} + \frac{1}{12.00}\right)} = 11.91.$$

The last expression is the reciprocal of the arithmetic mean of reciprocals and is called harmonic mean (HM). For a set of n values $X_1, X_2, \ldots\ldots X_n$, HM is defined as

$$\text{HM} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Note that HM is not defined when any observation is zero.

If the stockist, instead of stocking Rs. 5000 worth of items, stocks 3000 items at the beginning of every month at the given prices, the appropriate average would be arithmetic mean. To verify this, we can write

$$\text{Average Price} = \frac{\text{Total Money Spent}}{\text{Total Quantity Purchased}}$$

$$= \frac{3000 \times 10.75 + 3000 \times 11.80 + 3000 \times 14.00 + 3000 \times 11.45 + 3000 \times 12.00}{3000 \times 5}$$

$$= \frac{10.75 + 11.80 + 14.00 + 11.45 + 12.00}{5} = \text{AM of the given prices.}$$

4.3.2 Weighted Means

For many practical applications weighted means (arithmetic, geometric or harmonic) reflect phenomenon more clearly than unweighted or simple means that have been

computed so far. For computation of, say, consumer price index, not all commodities are equally important. Increase in fuel cost may affect consumer price index more than an increase in agricultural prices. For stock market, stock of some key companies may be a trend setter. Weighted means are more appropriate in such situations. To find weighted mean, a weight w_i is attached to each X_i and the means are computed as if w_i's are, symbolically, frequencies of the corresponding X_i's. The computational formulae are as given below:

$$\text{Weighted AM} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

$$\text{Weighted GM} = \left(\prod_{i=1}^{n} X_i^{w_i}\right)^{\frac{1}{\sum w_i}} \quad \text{and}$$

$$\text{Weighted HM} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{X_i}}$$

Weighted mean becomes equal to unweighted mean when each w_i is same or equal to unity.

4.3.3 Pooled Mean

Often we come across situations when means have been computed for different sources or samples. In such situations we become interested to find an overall mean if it is meaningful. This is done by computing what is called a *pooled mean.* The procedure of computing a pooled mean is given below.

Let $m_1, m_2, \ldots m_r$ be r arithmetic (or geometric or harmonic) means, computed on the basis of $n_1, n_2, \ldots n_r$ observations respectively. Then

$$\text{Pooled arithmetic mean} = \frac{1}{n}\sum_{i=1}^{r} m_i n_i, \text{ where } n = \sum_{i=1}^{r} n_i$$

$$\text{Pooled geometric mean} = \left(\prod_{i=1}^{r} m_i^{n_i}\right)^{\frac{1}{n}} \quad \text{and}$$

$$\text{Pooled harmonic mean} = \frac{n}{\sum_{i=1}^{r} \frac{n_i}{m_i}}$$

where $n = n_1 + n_2 + .. + n_r$

Note that the above expressions are similar to the expressions for weighted means.

4.3.4 Choosing a Measure of Central Tendency

It has already been discussed when a particular mean, AM or GM or HM, is more appropriate than the other two. However, when one has grouped data in which either of the end classes are open ended, i.e., of the type 'upto c_1' and / or 'c_{k-1} and above', mid-points of such classes cannot be computed. Consequently, no mean can be computed. There is, however, no problem in computing median or mode in such cases. On the other hand, a pooled median or mode cannot be computed, like the case for mean, unless all the sets of data are made available in their entirety. These problems are related to computational difficulties and not to appropriateness of measure.

Since graphical representation of data is more appealing, median or mode are more useful in such a situation because their crude values can be obtained easily without having to go through any computations. Also, median and mode are simple concepts for communication and comparison between graphs. It has, however, been observed that median is less stable than arithmetic mean in repeated sampling and one needs to be careful when comparing graphs.

For data that has a distribution close in shape to what is called normal with one peak and going down symmetrically on either side, one may use one of mean, median or mode because for a normal shape distribution, these measures have the same value.

It should be clearly understood that choosing an appropriate measure of central tendency is not an end to data analysis, and much still remains. For example, by saying that household average monthly expenditure on food is Rs. 348.66, it does not say whether a large number of households have very low monthly average expenditure on food or a few households have a very good menu. Next set of analysis aims at answering such questions.

4.4 PERCENTILES

Concept of percentiles will be explained by using mainly Table 4.2 data on average monthly household expenditure. Percentiles are used in two directions, depending on the question to be answered. Direction of a question may be, what per cent of households have monthly average food expenditure upto Rs. 350.80? Or it may be, what is the maximum monthly average food expenditure of the lower 50% of the households? Note, from our earlier computation of median of Table 4.2 distribution, that the answer to one question is the figure in the other, i.e., 50% of the households have Rs. 350.80 as maximum average monthly food expenditure. Depending on interest, percentage below a cut-off point may be called for : when a poverty line is decided, it is of interest to know the percentage below the poverty line. In the other direction, it may also be of interest to find the status of lower 10% or upper 5% of the population. These are answered by using what are called percentiles.

4.4.1 Percentile: Definition and Computation

For any given percentage v, vth percentile is P_v, a value of the variable being studied, so that at least v percent of the observations are less than or equal to P_v and at least (100 - v) percent of the observations are greater than or equal to P_v.

For example, for Table 4.1, distribution of household size, $P_v = 5$ for any v from 78 to 89.

For grouped data, percentiles are more clearly understood when one looks at the cumulative distribution function. Let $F(X)$ be the proportion of observations less than or equal to X. Any given value X_0 is then the 100 $F(X_0)$th percentile. For Table 4.2, class boundaries, one has $F(286.5) = 0.01$, $F(310.5) = 0.15$, $F(334.5) = 0.31$, $F(358.5) = 0.59$ and $F(382.5) = 0.85$, and consequently Rs. 286.5 = P_{10}, Rs.310.5 = P_{15}, Rs. 334.5 = P_{31}, Rs. 358.5 = P_{59}, and Rs. 382.5 = P_{85}. Note that any amount less than Rs. 262.5 (lower boundary of first class interval) is zero-th percentile and any amount more than Rs.406.5 (upper boundary of last class interval) is 100th percentile.

4.4.2 Quartiles and Deciles

Depending on its use, some specific percentiles go by different names. Every 25th percentile is called a quartile, and every 10th percentile is called a decile. For example,

25th percentile = $P_{25} = Q_1$ = first quartile
50th percentile = $P_{50} = Q_2$ = second quartile
75th percentile = $P_{75} = Q_3$ = third quartile
10th percentile = $P_{10} = d_1$ = first decile
20th percentile = $P_{20} = d_2$ = second decile, etc., and
$P_{50} = Q_2 = d_5$ = median.

The formulae for Q_1 and Q_3 are similar to the formula for the median. These can be directly written as given below.

$$Q_1 = l_{Q_1} + \frac{\frac{N}{4} - C}{f_{Q_1}} \times h, \text{ and}$$

$$Q_3 = l_{Q_3} + \frac{\frac{3N}{4} - C}{f_{Q_3}} \times h,$$

where C denotes the cumulative frequency of classes preceding the first (or third) quartile class and h is the corresponding class width.

Using similar notations, it is possible to write the formula for any partition value. For example, the formula for 40th percentile can be written as

$$P_{40} = l_{P_{40}} + \frac{\frac{40N}{100} - C}{f_{P_{40}}} \times h$$

Percentiles also go by the name of fractiles when proportions, instead of percentages, are used. For example, P_{30} is 0.3 fractile.

Just as one does not get a complete picture of a distribution by looking at a measure of location, too many percentiles may be needed to describe the spread or dispersion of a distribution. It is felt that there should be some simple measures of dispersion. This is the topic of discussion of the next unit.

Check Your Progress 2

1) Given below are the prices in ratios for five commodities with the corresponding weights. Calculate the Weighted Arithmetic Mean and Geometric Mean.

Commodity	Price Ratio	Weight
1	2.20	30
2	1.85	25
3	1.80	22
4	2.05	13
5	1.75	10

2) The earnings of five nationalised banks, in crores of rupees, is given below.

217.40 330.50 682.55 1263.59 2249.63

Find the Geometric Mean of the earnings.

.....

.....

.....

.....

.....

.....

3) The distribution of age of males at the time of marriage was as follows :

Age (years)	No. of Males
18 - 20	5
20 - 22	18
22 - 24	28
24 - 26	37
26 - 28	24
28 - 30	22

Find at the time of marriage (i) the average age, (ii) modal age, (iii) the median age, (iv) third quartile, (v) sixth decile, (vi) nineteenth percentile.

.....

.....

.....

.....

.....

.....

.....

4) In a factory, a mechanic takes 15 days to fabricate a machine, the second mechanic takes 18 days, the third mechanic takes 30 days and the fourth mechanic takes 90 days. Find the average number of days taken by the workers to fabricate the machine. Which average would you use, and why?

.....

.....

.....

.....

.....

.....

5) The amount of interest paid on each of the three different sums of money yielding 10%, 12% and 15% simple interest per annum are equal. What is the average yield percent on the total sum invested?

.....

.....

.....

.....

.....

.....

4.5 LET US SUM UP

In this unit you have learned to compute various measures of central tendency. These measures of central tendency can be divided into two broad categories, namely mathematical averages and positional averages. Positional averages are mode, median, quartiles, percentiles, etc., while arithmetic mean, geometric mean and harmonic mean are mathematical averages. Geometric Mean is most suitable for averaging ratio and proportional rates of growth while Arithmetic mean or Harmonic mean can be used to find average rates like price, speed, etc. depending upon the nature of the given condition.

4.6 KEY WORDS

Arithmetic Mean : Sum of observed values of a set divided by the number of observations in the set is called a mean or an average.

Frequency Distribution : The arrangement of data in the form of frequency distribution that describes the basic pattern which the data assumes in the mass.

Geometric Mean : It is the mean of n values of a variable computed as the nth root of their product.

Harmonic Mean : It is the inverse of the arithmetic mean of the reciprocals of the observations of a set.

Median : In a set of observations, it is the value of the middlemost item when they are arranged in order of magnitude.

Mode : In a set of observations, it is the value which occurs with maximum frequency.

4.7 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G U. and M. G Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

4.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) 5.1, 5, 5
2) 108.48 ; 108.41 ; 111.42

Check Your Progress 2

1) Rs. 1.96 ; Rs. 1.95
2) Rs. 674.31 crores
3) (i) 25.83 years (ii) 24.82 years (iii) 24.86 years (iv) 27.30 years (v) 25.59 years (vi) 28.79 years
4) Arithmetic Mean, 38.25 days
5) Harmonic Mean, 12%.