CHAPTER THREE

3 MEASURES OF CENTRAL TENDENCY (AN AVERAGE)

Motivating Example

In the previous chapter, we deal on collection and summarizing the data using tables, diagrams and graphs.

Now suppose the following example,

- Students from two or more classes appeared in the examination and we wish to compare the performance of the classes in the examination or wish to compare the performance of the same class after some coaching over a period of time.
- When making such comparisons, it is not practicable to compare the full frequency distributions of marks. However compactly these may be presented. Therefore, for such statistical analysis, we need a single representative value that describes the entire mass of data given in the frequency distribution.
- This single representative value is called the central value, measure of location or an average around which individual values of a series cluster. This central value or an average enables us to get a gist of the entire mass of data, and its value lies somewhere in the middle of the two extremes of the given observations.
- For this reason, such a central value or an average is frequently called a measure of central tendency is needed. That is, a single value that describes the characteristics of the entire mass of data is called measures of central tendency or average.

3.1 Definition of Measure of Central Tendency

Numerical descriptive measures are commonly used to convey a mental image of pictures, objects, and other phenomena. There are two main reasons for this. First, graphical descriptive measures are inappropriate for statistical inference, because it is difficult to describe the similarity of a sample frequency histogram and the corresponding population frequency histogram. The second reason for using numerical descriptive measures is one of expediency—we never seem to carry the appropriate graphs or histograms with us, and so must resort to our powers of verbal communication to convey the appropriate picture. We seek several numbers, called numerical descriptive measures, which will create a mental picture of the frequency distribution for a set of measurements.

The two most common numerical descriptive measures are ***measures of central tendency*** and ***measures of variability;*** that is, we seek to describe the center of the distribution of measurements and also how the measurements vary about the center of the distribution.

More generally, when two or more different data sets are to be compared it is necessary to condense the data, but for comparison the condensation of data set into a frequency distribution and visual presentation are not enough. It is then necessary to summarize the data set in a single value. Such a value usually somewhere in the center and represent the entire data set and hence it is called **measure of central tendency or averages.** Since a measure of central tendency (i.e. averages) indicates the location or the general position of the distribution on the X-axis therefore it is also known as a measure of location or position.

3.2 Objectives of Measuring Central Tendency

A single value that describes the characteristics of the entire mass of data is called measures of central tendency or average.

Objectives of measuring central tendency are:

- To get a single value that represent(describe) characteristics of the entire data;
- To summarizing/reducing the volume of the data;
- To facilitating comparison within one group or between groups of data;
- To enable further statistical analysis;
- recognize and distinguish between the different types of averages;
- learn to compute different types of averages;
- draw meaningful conclusions from a set of data;
- develop an understanding of which type of average would be the most useful in a particular situation.

❖ **The Summation Notation (∑)**

Let a data set consists of a number of observations, represents by $x_1, x_2, ..., x_n$ where n (the last subscript) denotes the number of observations in the data and x_i is the i^{th} observation.

Then the sum $x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$. For instance, a data set consisting of six measurements 21, 13, 54, 46, 32 and 37 is represented by x_1, x_2, x_3, x_4, x_5 and x_6 where $x_1 = 21$, $x_2 = 13$, $x_3 = 54$, $x_4 = 46$, $x_5 = 32$ and $x_6 = 37$.

Their sum becomes $\sum_{i=1}^{6} x_i = 21+13+59+46+32+37=208$.

Similarly, $x_1^2 + x_2^2 + \ldots + x_n^2 = \sum_{i=1}^{n} xi^2$

❖ **Some Basic Properties of the Summation Notation**

1. $\sum_{i=1}^{n} c = n.c$ where c is a constant number.
2. $\sum_{i=1}^{n} b.x_i = b\sum_{i=1}^{n} x_i$ where b is a constant number.
3. $\sum_{i=1}^{n} (a + bx_i) = n.a + b\sum_{i=1}^{n} x_i$ where a and b are constant numbers
4. $\sum_{i=1}^{n} (x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$
5. $\sum_{i=1}^{n} x_i y_i \neq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$

Example:

Let $\sum_{i=1}^{12} x_i = 26$, $\sum_{i=1}^{12} y_i = 17$, $\sum_{i=1}^{12} x_i^2 = 484$, $\sum_{i=1}^{12} y_i^2 = 362$

Find *I)* $\sum_{i=1}^{12} (4x_i + 3y_i)$, *II)* $\sum_{i=1}^{12} 2x_i(x_i - 7)$

Solution: *I)* $\sum_{i=1}^{12} (4x_i + 3y_i) = 4\sum_{i=1}^{12} x_i + \sum_{i=1}^{12} Y_i = 4(26) + 3(17) = 155$

II) $\sum_{i=1}^{12} 2x_i(x_i - 7) = 2\sum_{i=1}^{12} x_i^2 - 14\sum_{i=1}^{12} x_i = 2(484) - 14(26) = 604$

❖ **Important Characteristics of a Good Measure of Central Tendency**

We say a measure of central tendency is best if it possesses most of the following characteristics. It should:

- Not be seriously affected by extreme observations,
- Exist and be unique,
- Based on all observations during computation,
- It should be easy to calculate and understand (interpret),
- It should be rigidly defined. (The definition should be clear and unambiguous so that it leads to one and only one interpretation by different persons. So that the personal biases of the investigator don't affect the value of its usefulness),
- It should be representative of the entire data, if it's from sample. (Then the sample should be random enough to be accurate representative of the population)**,**
- It should have sampling stability. (It shouldn't be affected by sampling fluctuations. This means that if we pick (take) two independent random samples of the same size from a given population and compute the average for each of these samples then the value obtained from different samples should not vary much from one another,
- Being capable for further statistical analysis and /or algebraic manipulation,

3.3 Types of Measures of Central Tendency

Several types of averages or measures of central tendency can be defined, the most commons are

- the mean
- the mode
- the median

3.3.1 The Mean

There are four types of means: Arithmetic mean, weighted arithmetic mean, Harmonic mean and Geometric mean.

3.3.1.1 Arithmetic mean

Arithmetic mean is defined as the *sum of the measurements of the items divided by the total number of items*.

- **For row data:** When the data is given by row**,** the arithmetic mean can be defined as:

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_k}{n} = \frac{\sum_{i=1}^{k} x_i}{n}$$

where, n= no of observations in the data set

Example 1: You measure the screen area of 10 desktop computer (in cm^2) and record the following measure:

185 96 94.2 112 150 76.5 201 80 102 99

Compute the mean area of a screen based on the collected data.

Soln.

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i}{n} = \frac{185 + 96 + \cdots + 99}{10} = \frac{1{,}195.7}{10} = 119.57$$

❖ **Arithmetic Mean for Ungrouped/ Discrete Frequency Distribution**

When the data are arranged or given on the form of ungrouped frequency distribution, then the formula for the mean is

$$\bar{X} = \frac{f_1x_1 + f_2x_2 + \cdots + f_kx_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_ix_i}{\sum_{i=1}^{k} f_i} \quad \text{Note that } \sum_{i=1}^{k} f_i = n$$

Example 2: Monthly incomes of fourth year regular students are given in the following frequency distribution.

Monthly income (birr)	54.5	64.5	74.5	84.5	94.5	104.5	114.5	$\sum_{i=1}^{k} f_ix_i$
Number of students	6	9	15	25	13	7	5	
f_ix_i	327	580.5	1117.5	2112.5	1228.5	731.5	572.5	6598

Compute the mean for these data.

❖ **Arithmetic Mean for Grouped/ Continuous Frequency Distribution**

If data are given in the form of continuous frequency distribution, the sample mean can be computed as

$$\bar{X} = \frac{f_1m_1 + f_2m_2 + \cdots + f_km_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_im_i}{\sum_{i=1}^{k} f_i} \quad \text{Note that } \sum_{i=1}^{k} f_i = n$$

Where m_i is the class mark of the i^{th} class; $i = 1, 2 \ldots k$

f_i = the frequency of the i^{th} class and k = the number of classes

Note that $\sum_{i=1}^{k} f_i = n$ = the total number of observations.

Example 3: A random sample of 30 states shows the number of low power FM radio stations for each state. Find the variance and standard deviation for the data.

Class limits	Frequency
1–9	5
10–18	7
19–27	10
28–36	3
37–45	3
46–54	2

Sopln.

Class limits	Frequency	Class Mark (m_i)	$f_i * m_i$
1–9	5	5	25
10–18	7	14	98
19–27	10	23	230
28–36	3	32	96
37–45	3	41	123
46–54	2	50	100
		$\sum_{i=1}^{k} fi * mi$	**672**

$$\Rightarrow \bar{X} = \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i} = \frac{672}{30} = 22.4$$

Exercise: Thirty automobiles were tested for fuel efficiency (in miles per gallon). The following frequency distribution was obtained. Find the mean fuel efficiency for the automobiles.

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

❖ **Properties and application of arithmetic mean**

1. It is easy to calculate and understand.
2. All observation involved in its calculation.
3. It is applicable to quantitative data only.
4. It cannot be computed for open end classes
5. It may not be the values which the variable actually takes and termed as a fictitious (unreal) average. E.g. The figure like on average 2.21 children per family, 3.4 accidents per day.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.
7. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
8. The mean is used in computing other statistics, such as the variance.
9. It is Unique: - a set of data has only one mean and not necessarily one of the data values.
10. If a constant k is added or subtracted from each value of a distribution, then the new mean for the new distribution will be the original mean plus or minus k, respectively.
11. The sum of the deviation of various values from their mean is zero i.e., $\sum (x_i - \bar{x}) = 0$
12. The sum of the squares of deviation of the given set of observations is minimum when taken from the arithmetic mean i.e., $\sum (x_i - A)^2 \rightarrow$ is minimum when taken from mean than any arbitrary value A from a set of observation.
13. It can be used for further statistical treatment, comparison of means, test of means.
14. When a set of observations is divided into k groups and $\bar{x}_1$ is the mean of n_1 observations of group 1, $\bar{x}_2$ is the mean of n_2 observations of group2, …, $\bar{x}_k$ is the mean of n_k observations of group k , then the combined mean ,denoted by $\bar{x}_c$, of all observations taken together is given by

$$\bar{X}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

15. If a wrong figure has been used in calculating the mean, we can correct if we know the correct figure that should have been used. Let

- X_{wr} denote the wrong figure used in calculating the mean
- X_c be the correct figure that should have been used
- $\bar{X}_{wr}$ be the wrong mean calculated using X_{wr} , then the correct mean, then $\bar{X}_{correct}$, is given by

$$\bar{X}_{correct} = \frac{n\bar{X}_{wr} + X_c - X_{wr}}{n}$$

Example 4: Last year there were three sections taking Stat 1044 course in ASTU. At the end of the semester, the three sections got average marks of 80, 83 and 76. If there were 28, 32 and 35 students in each section respectively. Find the mean mark for the entire students.

Solution: here we need to calculate combined mean as:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{28(80) + 32(83) + 35(76)}{28 + 32 + 35} = \frac{7556}{95} = \underline{79.54}$$

Example 5: An average weight of 10 students was calculated to be 65 kg, but latter, it was discovered that one measurement was misread as 40 kg instead of 80 kg. Calculate the corrected average weight.

Solution: $\bar{X}_{correct} = \frac{n\bar{X}_{wr} + X_c - X_{wr}}{n} = \frac{10(65)+80-40}{10} = 69$

Exercise:

The average score on the mid-term examination of 25 students was 75.8 out of 100. However, after the mid-term exam, a student whose score was 41 out of 100 dropped the course. What is the average/mean score among the 24 students?

3.3.1.2 Weighted Arithmetic Mean

In finding arithmetic mean and others type of mean for series of data, all items were assumed to be of equally importance (each value in the data set has equal weight). But, ***when the observations have different weight, we use weighted average.*** Weights are assigned to each item in proportion to its relative importance.

If $x_1, x_2, ..., x_k$ represent values of the items and $w_1, w_2, ..., w_k$ are the corresponding weights, then the weighted arithmetic mean, $(\bar{x}_w)$ is given by

$$\bar{X}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_kx_k}{w_1 + w_2 + \cdots + w_k} = \frac{\sum_{i=1}^{k} w_ix_i}{\sum_{i=1}^{k} w_i}$$

Example 6: A student's final mark in Mathematics, Physics, Chemistry and Biology are respectively 82, 80, 90 and 70. If the respective credits received for these courses are 3, 5, 3 and 1, determine the approximate average mark the student has got for one course.

Solution: We use a weighted arithmetic mean, weight associated with each course being taken as the number of credits received for the corresponding course.

x_i	82	80	90	70
w_i	3	5	3	1

Therefore, $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{(3\times 82) + (5\times 80) + (3\times 90) + (1\times 70)}{3+5+3+1} = 82.17$

3.3.1.3 Geometric Mean

The geometric mean is the n[th] root of the product of n positive values. If X1, $X_{2,\ldots,}X_n$ are n positive values, then their geometric mean is G.M $=(X_1X_{2\ldots}X_n)^{1/n}$. ***The geometric mean is usually used in average rates of change, ratios, percentage distribution, and logarithmical distribution.***

In case of number of observations is more than two it may be tedious taking out from square root, in that case calculation can be simplified by taking natural logarithm with base ten.

GM $= \sqrt[n]{\mathbf{x_1x_2x_3 \ldots x_n}} \Rightarrow$ GM $= (x_1x_2x_3 \ldots x_n)^{\frac{1}{n}}$ taking logarithm in both sides.

$\log(G.M) = \frac{1}{n} log(X_1.X_2.X_3 \ldots X_n) = \frac{1}{n} \sum_{i=1}^{n} \log x_i$

G. M = Antilog $\left[\frac{1}{n} \sum_1^n logX_i\right]$

Example 7: The ratios of prices in 1999 to those in 2000 for 4 commodities were 0.9, 1.25, 1.75 and 0.85. Find the average price ratio.

Solution G.M = antilog $\frac{\sum \log X_i}{n}$ = antilog $\frac{(\log 0.92 + \log 1.25 + \log 1.75 + \log 0.85)}{4}$

= antilog $\frac{(0.963 - 1 + 0.0969 + 0.2430 + 0.9294 - 1)}{4}$ = antilog0.5829 = 1.14

What is the arithmetic mean of the above values?

$$GM = \frac{0.92 + 1.25 + 1.75 + 0.85}{4} = 1.19$$

3.3.1.4 Harmonic Mean

Harmonic mean is a suitable measure of central tendency when the data pertains to speed, rate and time. $HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}}$

Example 8: A motorist travels 480km in 3 days. She travels for 10 hours at rate of 48km/hr on 1st day, for 12 hours at rate of 40km/hr on the 2nd day and for 15 hours at rate of 32km/hr on the 3rd day. What is her average speed?

- $HM = \frac{3}{\frac{1}{48} + \frac{1}{40} + \frac{1}{32}} = 39.92$

- **Relation Between AM GM HM**

Relation between AM GM HM is useful to better understand arithmetic mean (AM), geometric mean (GM), harmonic mean (HM). The product of arithmetic mean and harmonic mean is equal to the square of the geometric mean.

$$AM \times HM = GM^2.$$

The relation between AM GM HM can be understood from the statement that the value of AM is greater than the value of GM and HM. For the same given set of data points, the arithmetic mean is greater than geometric mean, and the geometric mean is greater than the harmonic mean. This relation between AM, GM HM can be presented as the following expression.

$$AM > GM > HM$$

For better understanding, let us first understand how to find AM, GM, HM. For any two numbers a & b the formula for the arithmetic mean (AM), geometric mean (GM), and harmonic mean (HP) is as follows. The arithmetic mean is also called the average of the given numbers, and for two numbers a, b, the arithmetic mean is equal to the sum of the two numbers, divided by 2.

$$AM = \frac{a + b}{2}$$

The geometric mean of two numbers is equal to the square roots of the product of the two numbers a, b. Further, if there are n numbers of data, then their geometric mean is equal to the nth root of the product of the n numbers.

$$GM = \sqrt{ab}$$

The harmonic mean of two numbers 1/a. 1/b is equal to the inverse of their arithmetic mean. The arithmetic mean of these two numbers 1/a, 1/b is equal to (a + b)/2ab, and the inverse of this results in the harmonic mean of the two numbers.

$$HM = \frac{2ab}{a+b}$$

The formula for the relation between AM, GM, HM is the product of arithmetic mean and harmonic mean is equal to the square of the geometric mean. This can be presented here in the form of this expression.

$$\text{AM} \times \text{HM} = \text{GM}^2$$

Let us try to understand this formula clearly, but deriving this formula.

$$AM * HM = \frac{\text{a} + \text{b}}{2} * \frac{2ab}{a+b} = ab$$

$$\Rightarrow ab = \sqrt{ab^2} = \left(\sqrt{ab}\right)^2 = GM^2$$

Thus, the square of the geometric mean is equal to the product of the arithmetic mean and harmonic mean.

Example: Find the value of geometric mean GM, if the arithmetic mean AM is 7, and harmonic mean HM is 48/7.

3.3.2 The Median

The median is the halfway point in a data set. Before you can find this point, the data must be arranged in ascending or increasing order. When the data set is ordered, it is called a data array. The median either will be the specific middle value or the arithmetic mean of the two middle values. We shall denote the median of $x_1, x_2, ..., x_n$ by $\tilde{x}$.

- **For ungrouped data the median is obtained by:**

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & if\ the\ number\ of\ items, n, is\ odd \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}) & if\ the\ number\ of\ items, n, is\ even \end{cases}$$

- **For grouped data** (continuous frequency distribution) **the median, obtained by interpolation method, is given by**

$$\tilde{X} = L_{med} + W\left(\frac{\frac{n}{2} - cf}{f_{med}}\right)$$

Where L_{med} = lower class boundary of the median class

cf = Sum of frequencies of all class lower than the median class (in other words it is the cumulative frequency immediately preceding the median class)

f_{med} = Frequency of the median class and W = is class width

☞ ***Note that: the median class is the class with the smallest cumulative frequency (less than type) greater than or equal to*** $n/2$ **.**

Examples 1: The birth weights in pounds of five babies born in a hospital on a certain day are 9.2, 6.4, 10.5, 8.1 and 7.8. Find the median weight of these five babies.

Solution: the median is 8.1.

Examples 2: The following table gives the distribution of the weekly wages of employees of a small firm.

Wages in birr	No. of employees
126 and below	3
127 – 135	5
136 – 144	9
145 – 153	12
154 – 162	5
163 – 171	4
172 and above	2

a) Find the median weekly wage.

b) Why is the median a more suitable measure of central tendency than the mean in this case?

❖ **Merits of median**

- The median is affected less than the mean by extremely high or extremely low values.
- The median is used to find the center or middle value of a data set.
- The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- Median can be calculated even in case of open-ended intervals.
- It can be computed for ratio, interval, and ordinal level of data.

- It is applicable to quantitative data only.

❖ **Demerits of median**

- It is not capable of further algebraic treatment.
- It is not a good representative of the data if the number of items (data) is small.
- The arrangement of items in order of magnitude is sometimes very tedious process if the number of items is very large.

3.3.3 The Mode

The mode is another measure of central tendency. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case and denoted by $\hat{x}$.A given set of data may have

- One mode – uni-model e.g. 3 ,3,7,6,2,1 $\hat{X}$=3
- Two modes – Bi-modal e.g. 10,10,9,9,6,3,2,1 $\hat{X}$= 10 and 9
- More than two modes- multi-modal. eg. 5,5,5,6,6,6,8,8,8,2,3,2 $\hat{X}$=5,6,8
- May not exist at all e.g. 1,3,2,4,5,6,7,8 ***no modal value***

For **discrete/ ungrouped** data, the mode or the modal value is the most frequently occurring score/observation in a series.

- **For grouped data** (continuous frequency distribution)**, the mode is found by the following formula:**

$$\hat{x} = L_{\text{mod}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)W$$

Where L_{mod} = lower class boundary of the modal class

Δ_1 = The difference between the frequency of the modal class and frequency of the class Immediately preceding the modal class

Δ_2 = The difference between the frequency of the modal class and frequency of the class Immediately follows the modal class

W = is the common class width

☞ ***Note that: the modal class is the class with the highest frequency in the distribution.***

Examples 1: The marks obtained by ten students in a semester exam in statistics are: 70, 65, 68, 70, 75, 73, 80, 70, 83 and 86. Find the mode of the students' marks.

Example 2: Find the mode for the frequency distribution of the birth weight (in kilogram) of 30 children given below.

Weight	1.9-2.3	2.3-2.7	2.7-3.1	3.1-3.5	3.5-3.9	3.9-4.3
No. of children	5	5	9	4	4	3

Solution: 2.7-3.1 is the modal class since it has the highest frequency

$$\Delta_1 = 9-5 = 4 \quad and \quad \Delta_2 = 9-4 = 5 \quad L_{\text{mod}} = 2.7$$

$$\hat{x} = 2.7 + \left(\frac{4}{4+5}\right) * 0.4 = 2.878$$

Merits of mode

- Mode is not affected by extreme values.
- The mode is used when the most typical case is desired.
- The mode is the easiest average to compute.
- Mode can be calculated even in the case of open-end intervals. And it is not necessary to know all observations.
- It can be computed for all level of data i.e., ratio, interval, ordinal or nominal.
- It is applicable for both qualitative and quantitative data.

Demerits of mode

- Mode may not exist in the series and even if it exists, it may not be a unique value.
- It does not fulfill most of the requirements of a good measure of central tendency

3.3.4 The Midrange

The midrange is a rough estimate of the middle. It is defined as the sum of the lowest and highest values in the data set, divided by 2. It is a very rough estimate of the average and can be affected by one extremely high or low value. The symbol MR is used for the midrange.

$$MR = \frac{lowest\ value\ +\ highest\ value}{2}$$

Examples 1: Of the 25 brightest stars, the distances from earth (in light-years) for those with distances less than 100 light-years are found below. Find the mean, median, mode, and midrange for the data.

8.6 36.7 42.2 16.8 33.7 77.5 87.9

4.4 25.3 11.4 65.1 25.1 51.5

Soln.

i) $Mean = \frac{\sum_{i=1}^{k} x_i}{n} = \frac{486.2}{13} = 37.4$

ii) Median, after an arrangement of the data as 4.4, 8.6, 11.4, 16.8, 25.1, 25.3, 33.7, 36.7, 42.2, 51.5, 65.1, 77.5, 87.9 since the number of observations is an odd (i.e., n = 13) the median $\tilde{x} = 33.7 = 7^{th} observation$

iii) Mode = no mode

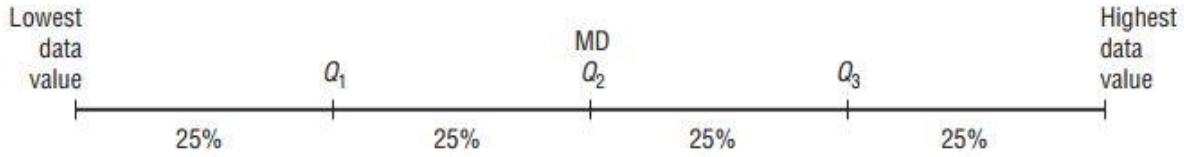
iv) $MR = \frac{lowest\ value + highest\ value}{2} = \frac{4.4+87.9}{2} = \frac{92.3}{2} = 46.15$

3.3.5 Quantiles

Quantiles are values which divides the data set arranged in order of magnitude in to certain equal parts. They are averages of position (non-central tendency). Some of these are quartiles, deciles and percentiles.

1. **Quartiles:** are values which divide the data set in to four equal parts, and denoted by Q_1, Q_2, and Q_3. The second quartile is equivalent to the median. The first quartile is also called the lower quartile and the third quartile is the upper quartile.

   Note that *Q*1 is the same as the 25th percentile; *Q*2 is the same as the 50th percentile, or the median; *Q*3 corresponds to the 75th percentile, as shown:



- ***For ungrouped data:***

  Let Q_j be the j^{th} quartile value for $j = 1,2,3$. Then $Q_j = \left(\frac{j}{4}(n+1)\right)^{th} item; \qquad j = 1,2,3.$

- ***For grouped data***

  We can apply the following formula: $Q_j = L_{Q_j} + \left(\frac{j \cdot n/4 - CF_{Q_j}}{f_{Q_j}}\right) W; \quad j = 1,2,3.$ were,

Q_j = the j^{th} quartile.

L_{Q_j} = Lower class boundary of the j^{th} quartile class.

CF_{Q_j} = Sum of frequencies of all classes lower than the j^{th} quartile class. (Cumulative frequency of the j^{th} quartile class)

f_{Q_j} = Frequency of the j^{th} quartile class and

W = Class width

- *The j^{th} quartile class is the class with the smallest cumulative frequency greater than or equal to $j \cdot n/4$.*

Example:

Find Q1, Q2, and Q3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Soln

Step 1 Arrange the data in order from lowest to highest.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50

↑

MD

$$MD = \frac{13 + 15}{2} = 14$$

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13

↑

Q_1

$$Q_1 = \frac{6 + 12}{2} = 9$$

So Q_1 is 9.

Step 4 Find the median of the data values greater than 14.

15, 18, 22, 50

↑

Q_3

$$Q_3 = \frac{18 + 22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

❖ **Other Method of Finding Quartiles**

To find the quartiles of a dataset or sample, follow the step-by-step guide below.

- ✓ Count the number of observations in the dataset (n).
- ✓ Sort the observations from smallest to largest.

➢ **Find the first quartile:**

Calculate n * (1 / 4).

- If n * (1 / 4) is an integer, then the first quartile is the mean of the numbers at positions n * (1 / 4) and n * (1 / 4) + 1.
- If n * (1 / 4) is not an integer, then round it up. The number at this position is the first quartile.

➢ **Find the second quartile:**

Calculate n * (2 / 4).

- If n * (2 / 4) is an integer, the second quartile is the mean of the numbers at positions n * (2 / 4) and n * (2 / 4) + 1.
- If n * (2 / 4) is not an integer, then round it up. The number at this position is the second quartile.

➢ **Find the third quartile:**

Calculate n * (3 / 4).

- If n * (3 / 4) is an integer, then the third quartile is the mean of the numbers at positions n * (3 / 4) and n * (3 / 4) + 1.
- If n * (3 / 4) is not an integer, then round it up. The number at this position is the third quartile.

Example:

Imagine you conducted a small study on language development in children 1–6 years old. You're writing a paper about the study and you want to report the quartiles of the children's ages.

Age (years)	1	2	3	4	5	6
Frequency	2	3	4	1	2	2

Step 1: Count the number of observations in the dataset

n = 2 + 3 + 4 + 1 + 2 + 2 = 14

Step 2: Sort the observations in increasing order

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Step 3: Find the first quartile

n * (1 / 4) = 14 * (1 / 4) = 3.5

3.5 is not an integer, so Q1 is the number at position 4.

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Q1 = 2 years

Step 4: Find the second quartile

n * (2 / 4) = 14 * (2 / 4) = 7

7 is an integer, so Q2 is the mean of the numbers at positions 7 and 8.

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Q2 = (3 + 3) / 2

Q2 = 3 years

Step 5: Find the third quartile

n * (3 / 4) = 14 * (3 / 4) = 10.5

10.5 is not an integer, so Q3 is the number at position 11.

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

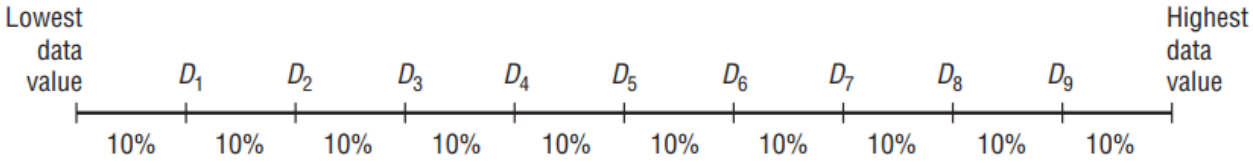
Q3 = 5 years

- In addition to dividing the data set into four groups, quartiles can be used as a rough measure of variability. This measure of variability which uses quartiles is called the ***interquartile range*** and is the range of the middle 50% of the data values.

The interquartile range (IQR) is the difference between the third and first quartiles.

IQR = Q3 - Q1

2. **Deciles:** are values dividing the data in to ten equal parts. Deciles are denoted by *D*1, *D*2, *D*3, . . ., *D*9, and they correspond to *P*10, *P*20, *P*30, . . ., *P*90. The fifth decile is equivalent to the median.



- **For ungrouped data**

Let D_j be the j^{th} percentile value for $j = 1, 2, ..., 9$. Then

$$D_j = \left(\frac{j}{10}(n+1)\right)^{th} item; \qquad j = 1, 2, ..., 9$$

- **For grouped data**

We can apply the following formula:

$$D_j = L_{D_j} + \left(\frac{j \cdot n/10 - CF_{D_j}}{f_{D_j}}\right) W; \quad j = 1, 2, ..., 9$$

Define the symbols similar way as we did in the case of quartiles.

The j^{th} decile class is the class with the smallest cumulative frequency greater than or equal to $j \cdot n/10$

3. **Percentiles:** Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group. They are values which divide the data in to one hundred equal parts and denoted by $P_1, P_2, ... P_{99}$. The fiftieth percentile is equivalent to the median.

- **For ungrouped data**

Let P_j be the percentile value for $j = 1,2,3,\ldots,99$. Then

$$P_j = \left(\frac{j}{100}(n+1)\right)^{th} item; \qquad j = 1,2,3,\ldots,99$$

- **For grouped data**

We can use the following formula:

$$P_j = L_{P_j} + \left(\frac{j \cdot {}^{n}\!/_{100} - F_{P_j}}{f_{P_j}}\right) W; \quad j = 1,2,3,\ldots,99$$

Define the symbols similar way as we did in the case of quartiles.

The j^{th} percentile class is the class with the smallest cumulative frequency greater than or equal to $j \cdot {}^{n}\!/_{100}$.

Interpretations

1. Q_j is the value below which $(j \times 25)$ *percent* of the observations in the series are found (where $j = 1,2,3$). For instance, Q_3 means the value below which 75 percent of observations in the given series are found.
2. D_j Is the value below which $(j \times 10)$ *percent* of the observations in the series are found (where $j = 1,2,\ldots,9$). For instance, D_4 is the value below which 40 percent of the values are found in the series.
3. P_j is the value below which j *percent* of the total observations are found (where $j = 1,2,3,\ldots,99$). For example, 73 percent of the observations in a given series are below P_{73}.

Exercise: The following table presents the male population of a certain region in Ethiopia.

Find a) all quartiles

b) The 9^{th} and 5^{th} decile and

c) 65^{th} and 75^{th} percentiles

Age groups (in years)	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
Male population	2580	3737	4620	5200	7250	620	297	355

CHAPTER FOUR

4. MEASURE OF VARIATION OR DISPURSION

4.1 Introduction

In the previous chapter, we have explained the measures of central tendency. It may be noted that these measures do not indicate the extent of dispersion or variability in a distribution. The dispersion or variability provides us one more step in increasing our understanding of the pattern of the data. Further, a high degree of uniformity (i.e. low degree of dispersion) is a desirable quality. If in a business there is a high degree of variability in the raw material, then it could not find mass production economical.

Suppose an investor is looking for a suitable equity share for investment. While examining the movement of share prices, he should avoid those shares that are highly fluctuating-having sometimes very high prices and at other times going very low. Such extreme fluctuations mean that there is a high risk in the investment in shares. The investor should, therefore, prefer those shares where risk is not so high.

4.2 Meaning and Definitions of Dispersion

The various measures of central value give us one single figure that represents the entire data. But the average alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of distribution.

Measures of dispersion help us in studying this important characteristic of a distribution.

- **Some important definitions of dispersion are given below:**

1. "Dispersion is the measure of the variation of the items." **-A.L. Bowley**
2. "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data." **-Spiegel**
3. Dispersion or spread is the degree of the scatter or variation of the variable about a central value." **-Brooks & Dick**
4. "The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion." **-Simpson & Kajka**

It is clear from above that dispersion (also known as scatter, spread or variation) measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the **second order**. An average is more meaningful when it is examined in the light of dispersion. For example, if the average wage of the workers of factory A is Rs. 3885 and that of factory B Rs. 3900, we cannot necessarily conclude that the workers of factory B are better off because in factory B there may be much greater dispersion in the distribution of wages.

The study of dispersion is of great significance in practice as could well be appreciated from the following example:

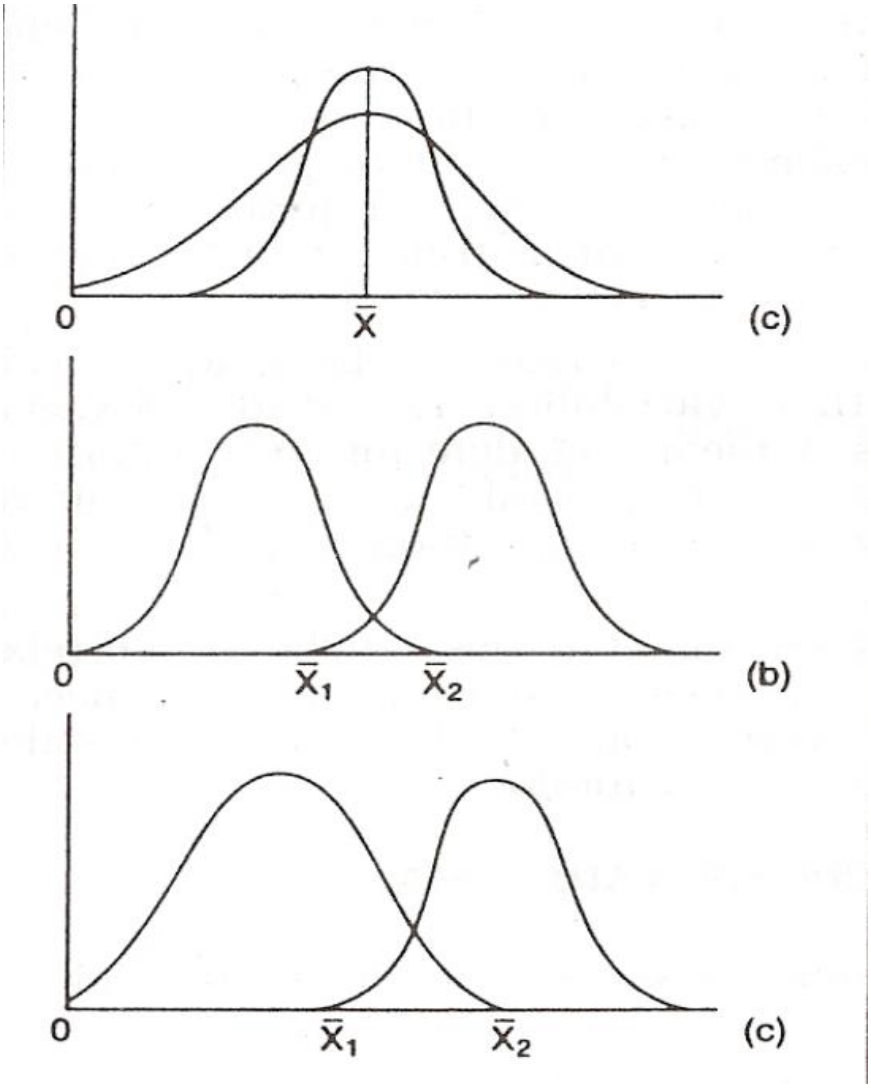
	Series A	**Series B**	**Series C**
Data	100	100	5
	100	110	20
	100	105	15
	100	107	390
	100	78	70
Total	**500**	**500**	**500**
Arithmetic Mean ($\bar{X}$)	100	100	100

Since arithmetic mean is the same in all three series, one is likely to conclude that these series are alike in nature. But a close examination shall reveal that distributions differ widely from one another.

In series A, in the table above, each and every item is perfectly represented by the arithmetic mean or in other words none of the items of series A deviates from the arithmetic mean and hence there is no dispersion. In series B, only one item is perfectly represented by the arithmetic mean and the other items vary but the variation is very small as compared to series C. In series C. not a single item is represented by the arithmetic mean and the items vary widely from one another. In series C, dispersion is much greater compared to series B.

The three figures given bellow represent frequency distributions with some of the characteristics. The two curves in diagram (a) represent two distractions with the same mean $\bar{X}$, but with different dispersions. The two curves in (b) represent two distributions with the same dispersion but with unequal means $\bar{X}_1$ and $\bar{X}_2$, (c) represents two distributions with unequal dispersion. The measures of

central tendency are, therefore insufficient. They must be supported and supplemented with other measures.



In the present chapter, we shall be especially concerned with the measures of variability or spread or dispersion. A measure of variation or dispersion is one that measures the extent to which there are differences between individual observation and some central or average value. In measuring variation we shall be interested in the amount of the variation or its degree but not in the direction. For example, a measure of 6 inches below the mean has just as much dispersion as a measure of six inches above the mean.

Literally meaning of dispersion is 'scatteredness'. Average or the measures of central tendency gives us an idea of the concentration of the observations about the central part of the distribution. If we

know the average alone, we cannot form a complete idea about the distribution. But with the help of dispersion, we have an idea about homogeneity or heterogeneity of the distribution.

4.3 Significance and Properties of Measuring Variation

Measures of variation are needed for four basic purposes:

1. Measures of variation point out as to how far an average is representative of the mass. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is a good estimate of the average in the corresponding universe. On the other hand, when dispersion is large, the average is not so typical, and unless the sample is very large, the average may be quite unreliable.
2. Another purpose of measuring dispersion is to determine nature and cause of variation in order to control the variation itself. In matters of health variations in body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production efficient operation requires control of quality variation the causes of which are sought through inspection is basic to the control of causes of variation. In social sciences a special problem requiring the measurement of variability is the measurement of "inequality" of the distribution of income or wealth etc.
3. Measures of dispersion enable a comparison to be made of two or more series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity of consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean great uniformity or consistency.
4. Many powerful analytical tools in statistics such as correlation analysis, the testing of hypothesis, analysis of variance, the statistical quality control, and regression analysis is based on measures of variation of one kind or another.

A good measure of dispersion should possess the following properties

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each and every item of the distribution.
5. It should be amenable to further algebraic treatment.
6. It should have sampling stability.
7. Extreme items should not unduly affect it.

4.4 Absolute and Relative Measures of Dispersion

Measures of dispersion/variation may be either absolute or relative.

What are Absolute Dispersion and Relative Dispersion?

Absolute & relative dispersion are two different ways to measure the spread of a data set. They are used extensively in biological statistics, as biological phenomena almost always show some variation and spread.

I. Absolute measures of dispersion

Absolute measures of dispersion are expressed in the same unit of measurement in which the original data are given. These values may be used to compare the variation in two distributions provided that the variables are in the same units and of the same average size. The easiest way to differentiate absolute dispersion with relative dispersion is to check whether your statistic involves units. Absolute measures always have units, while relative measures do not.

Absolute measures of dispersion include:

- The range,
- The quartile deviation,
- The mean deviation,
- The standard deviation and variance.

Absolute measures of dispersion use the original units of data, and are most useful for understanding the dispersion within the context of your experiment and measurements.

In case the two sets of data are expressed in different units, however, such as quintals of sugar versus tons of sugarcane or if the average sizes are very different such as manager's salary versus worker's salary, the absolute measures of dispersion are not comparable. In such cases measures of relative dispersion should be used.

II. Relative Measures of Dispersion

A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate measure of central tendency. It is sometimes called coefficient of dispersion because the word "coefficient" represents a pure number (that is independent of any unit of measurement). It should be noted that while computing the relative dispersion, the average (the measure of central tendency) used as a base should be the same one from which the absolute deviations were measured.

Note also that the value of a relative dispersion is unit less quantity.

Relative measures of dispersion are calculated as ratios or percentages; for example, one relative measure of dispersion is the ratio of the standard deviation to the mean. Relative measures of dispersion are always dimensionless, and they are particularly useful for making comparisons between separate data sets or different experiments that might use different units. One of the best examples of relative measure of dispersion is **Coefficient of Variation.**

Some Commonly Used Measures of Relative Dispersion / Absolute Dispersion. The simplest measure of absolute dispersion is the range. This is just the largest data point minus the smallest. We can write this as R = H – L.

For example, if a data set consisted of the points 2, 4, 5, 8, and 18, the range would be 18 – 2 = 16. The analogous relative measure of dispersion is the coefficient of range. This is given by (H – L)/ (H + L). For our example data set, it would be the ratio (18 – 2)/ (18 + 2), so (16/20) or 4/5.

The standard deviation is a more complicated measure of absolute dispersion, you could calculate it by squaring the difference between each data point and the mean, summing those squares, dividing by a number that is one less than the number of your data points, and then taking the square root of that. Since your values are squared and in the end the square root is taken again, the standard deviation is given in the original units of measure.

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

- Co-efficient of Range
- Co-efficient of Variation
- Co-efficient of Standard Deviation
- Co-efficient of Quartile Deviation
- Co-efficient of Mean Deviation

4.5 Types of Measures of Dispersion

4.5.1 The Range and Relative Range

Range (R) is defined as the difference between the largest and the smallest observation in a given set of data. That is, $R = x_{\max} - x_{\min}$ where x_{max} and x_{min} are the largest and the smallest observations in the series respectively.

In case grouped data, range is found by taking the difference between the class mark of the last class and that of the first class. That is, $R = M_{last} - M_{first}$ where M_{last} and M_{first} are the class marks of the last class and that of the first class respectively.

A relative range (RR), also known as coefficient of range, is given by

$$RR = \frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{R}{x_{\max} + x_{\min}} \ldots\ldots\ldots for\ ungrouped\ data$$

$$RR = \frac{M_{last} - M_{first}}{M_{last} + M_{first}} = \frac{R}{M_{last} + M_{first}} \ldots\ldots\ldots for\ grouped\ data$$

❖ **Properties of Range and Relative Range**

- Range and relative range are easy to calculate and simple to understand.
- Both cannot be computed for grouped data with open ended classes.
- They do not tell us anything about the distribution of values in the series.
- It is not based on all observation of the serious.
- It is affected by sampling fluctuation.
- It is affected by extreme values in the series.

Example 1: Find the range and relative range for the cost (in dollar) of ten laptops having different brand in a certain company.

462 480 534 624 498 552 606 588 516 570

Solution:

$$x_{\max} = 624\,dollar \qquad x_{\min} = 462\,dollar$$

$$R = x_{\max} - x_{\min} = 624\,dollar - 462\,dollar = 162\,dollar$$

$$RR = \frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{624\,dollar - 462\,dollar}{624\,dollar - 462\,dollar} = \frac{162\,dollar}{1086\,dollar} = 0.149$$

Example 2: Find the values of the range and relative range for the following frequency distribution: which shows the distribution of the maximum loads supported by a certain number of cables.

Maximum load (in kilo-Newton)	Number of cables
93 – 97	2
98 – 102	5
103 – 107	12
108 – 112	17
113 – 117	14
118 – 122	6
123 – 127	3
128 – 132	1

Solution:

$M_{first} = 95kN \quad M_{last} = 130kN$

$R = M_{last} - M_{first} = 130kN - 95kN = 35kN$

$$RR = \frac{M_{last} - M_{first}}{M_{last} + M_{first}} = \frac{130kN - 95kN}{130kN + 95kN} = \frac{35kN}{225kN} = 0.156$$

Inter Quartile Range: Is the difference between 3rd and 1st quartile and it is a good indicator of the absolute variability than range. IQR $= Q_3 - Q_1$

4.5.2 Quartile Deviation (semi – inter quartile Range) is a half of inter quartile range

$$QD = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$$

Coefficient of quartile Deviation The relative measure of quartile deviation also called the *coefficient of quartile deviation* is defined as: $Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

❖ ***Properties of Quartile Deviations***

i) The size of quartile deviation gives an indication about the uniformity. If QD is small, it denotes large uniformity. Thus, a coefficient of quartile deviation is used for comparing uniformity or variation in different distribution.

ii) Quartile deviation is not a measure of dispersion in the sense that it doesn't show the scatter around an average but only a distance on scale. As result it is regarded as a measure of partition.

iii) It can be computed when the distribution has an open-ended class. it is quite suitable in the case of open – ended distribution

iv) As compared to range, it is considered a superior measure of dispersion.

v) Since it not influenced by the extreme values in a distribution, it is particularly suitable in highly skewed or irregular distribution.

Examples 3. For the following frequency distribution find

a) Inter– quartile range.
b) Quartile deviation
c) CQD

Class limit	Frequency
21 – 22	10
23 – 24	22
25 – 26	20
27 – 28	14
29 – 30	14
Total	80

$\Rightarrow$ n/4 = 80/4 = 20, $20^{\text{th ordered}}$ observation

$\Rightarrow$ The 1^{st} quartile class is 23 -24

$$Q_1 = LCb + \frac{\left(n/4 - cf\right)w}{f} = 22.5 + \frac{(20 - 10)2}{22} = 23.4$$

$$Q_2 = 2\left(\frac{n}{4}\right) = 2\left(\frac{80}{4}\right) = 40, \quad Q_2 \; is 40^{th} \; obeservation$$

$\Rightarrow$ The class interval containing Q_2 is 25 – 26.

$$Q_2 = LCb_{Q2} + \frac{\left(2\left(n/4\right) - cf\right)w}{f} = 24.5 + \frac{(40 - 3)\ x2}{20} = \underline{\underline{25.3}}$$

And $Q_3 = 3\left(\frac{n}{4}\right) = 60$, Q_3 is 60^{th} position observation.

$\Rightarrow$ The class limits containing Q_3 is 27 – 28

$$Q_3 = LCb_{Q3} + \frac{\left(3\left(n/4\right) - cf\right)w}{f} = 26.5 + \frac{(60 - 52)}{14} = 27.84$$

a) Inter quartile range $= Q_3 - Q_1 = 27.64 - 23.44 = 4.23$

b) $Q.D = 1/2\ (Q_3 - Q_1) = 4.23/2 = 2.115$

c) CQD =4.23/51.24

The quartile deviation is more stable than the range as it defenses on two intermediate values. This is not affected by extreme values since the extreme values are already removed. However, quartile deviation also fails to take the values of all deviations.

4.5.3 The Mean Deviation and Coefficient of Mean Deviation

The *mean deviation (MD)* measures the average deviation of a set of observations about their central value, generally the mean or the median, ignoring the plus/minus sign of the deviations.

The mean deviation of a sample of n observations $x_1, x_2, ..., x_n$ is given as

$$MD = \frac{\sum |x_i - A|}{n}$$ Where A is a central measure (the mean or the median)

In case of grouped data, the formula for MD becomes

$$MD = \frac{\sum f_i |m_i - A|}{n}$$ Where m_i is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

- The mean deviation about the arithmetic mean is, therefore, given by

$$MD = \frac{\sum |x_i - \bar{x}|}{n} \text{.... for ungrouped data}$$

$$MD = \frac{\sum f_i |m_i - \bar{x}|}{n} \text{.... for grouped frequency distribution; where } m_i \text{ is the class mark of the } i^{th}$$

class, f_i is the frequency of the i^{th} class and $n = \sum f_i$

- The mean deviation about the median is also given by

$$MD = \frac{\sum |x_i - \tilde{x}|}{n} \text{.... for ungrouped data}$$

$$MD = \frac{\sum f_i |m_i - \tilde{x}|}{n} \text{.... for grouped frequency distribution; where } m_i \text{ is the class mark of the } i^{th}$$

class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

The *coefficient of mean deviation (CMD)* is the ratio of the mean deviation of the observations to their appropriate measure of central tendency: the arithmetic mean or the median.

In general, $CMD = \frac{MD}{A}$ where A is a measure of central tendency: the arithmetic mean or the median.

That is, CMD about the arithmetic mean is given by $CMD = \frac{MD}{\bar{x}}$ where MD is the mean deviation calculated about the arithmetic mean. On the other hand, CMD about the median is given by $CMD = \frac{MD}{\tilde{x}}$ in which case MD is calculated about the median of the observations.

❖ **Properties of Mean Deviation and coefficient of mean deviation**

- It is easy to understand and compute.
- It is based on all observations.
- It is not affected very much by the values of extreme value(s).
- It is not capable of further mathematical treatments and it is not a very accurate measure of dispersion.
- It can be calculated by using any average.

Example 4. Find the mean deviation about the median for the data in example 2 above and also the coefficient of mean deviation

4.5.4 The Variance, the Standard Deviation and Coefficient of Variation

4.5.4.1 The Variance

Variance is the arithmetic mean of the square of the deviation of observations from their arithmetic mean.

▪ Population Variance (σ^2)

➢ **For ungrouped data**

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \ldots = \frac{1}{N}\left(\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{N}\right)$$ Where μ is the population arithmetic mean and N is the total number of observations in the population.

➢ **For grouped data**

$$\sigma^2 = \frac{\sum f_i (m_i - \mu)^2}{N} = \frac{1}{N}\left(\sum f_i m_i^{\,2} - \frac{\left(\sum f_i m_i\right)^2}{N}\right)$$ Where μ is the population arithmetic mean, m_i is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $N = \sum f_i$.

▪ Sample Variance (S^2)

➢ **For ungrouped data**

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = alternatively = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$$ where, $\bar{x}$ is the sample arithmetic mean and *n* is the total number of observations in the sample.

For grouped data

$$S^2 = \frac{\sum f_i (m_i - \bar{x})^2}{n-1} = \ldots = \frac{1}{n-1}\left(\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{n}\right)$$ Where $\bar{x}$ is the sample arithmetic mean, m_i is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

4.5.4.2 The Standard Deviation

Standard deviation is the positive square root of the variance.

▪ Population Standard Deviation (σ)

$\sigma = \sqrt{\sigma^2}$ where σ^2 is the population variance.

▪ Sample Standard Deviation (S)

$S = \sqrt{S^2}$ where S^2 is the sample standard deviation.

Example: compute the variance and standard deviation for the following data

value	3	6	9	12	15	Total
frequency	1	4	10	3	2	20
$f_i x_i$	3	24	90	36	30	183
$x_i - \mu$	-6.15	-3.15	-0.15	2.85	5.85	
$(x_i - \mu)^2$	37.8225	9.9225	0.0225	8.1225	34.2225	
$f_i (x_i - \mu)^2$	37.8225	39.69	0.225	24.3675	68.445	170.55

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{183}{20} = 9.15, \quad where\ n = \sum_{i=1}^{5} f_i = 20$$

and $$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{170.55}{19} = 8.976$$

$$\Rightarrow s = \sqrt{s^2} = \sqrt{8.976} = 2.99$$

Example: The mean and standard deviation of 20 observations are found to be 10 and 2 respectively. On rechecking it was found that an observation 8 was incorrect. Calculate the correct mean and standard deviation in each of the following cases:

i. If wrong item is omitted

ii. If it is replaced by 12

Soln.

i. Given

- Number of observation n = 20
- Incorrect mean = 10
- Incorrect standard deviation = 2

From the given data we can have:

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{n}$$

$\Rightarrow\ 10 = \frac{1}{20}\sum_{i=}^{20} x_i$

$\Rightarrow\ \sum_{i=}^{20} x_i = 200$

So, the incorrect sum of observations = 200

⇒ Correct sum of observations = 200 – 8 = 192

Therefore, the correct mean and standard deviation will be:

$$\text{Correct mean} = \frac{correct\ sum}{19} = \frac{192}{19} = \mathbf{10.1} \quad and$$

$$Standard\ deviation\ (\delta) = \sqrt{\frac{1}{n}\sum_{i=1}^{20} {x_i}^2 - \frac{1}{n^2}(\sum_{i=1}^{20} x_i)^2}$$

$$\Rightarrow \quad 2 = \sqrt{\frac{1}{n} incorrect \sum_{i=1}^{20} {x_i}^2 - (incorrect\ mean)^2} = \sqrt{\frac{1}{20} incorrect \sum_{i=1}^{20} {x_i}^2 - 10^2}$$

$$\Rightarrow\ 4 = \frac{1}{20} incorrect \sum_{i=1}^{20} {x_i}^2 - 10^2 = \frac{1}{20} incorrect \sum_{i=1}^{20} {x_i}^2 - 100$$

$$\Rightarrow\ incorrect \sum_{i=1}^{20} {x_i}^2 = 2080$$

$$\text{So,} \quad correct \sum_{i=1}^{20} {x_i}^2 = \ incorrect \sum_{i=1}^{20} {x_i}^2 - (incorrect\ observation)^2$$

$$= 2080 - 8^2 = 2080 - 64 = 2016$$

$$\Rightarrow \text{Standard deviation } (\delta)_{correct} = \sqrt{\frac{1}{19} correct \sum_{i=1}^{20} x_i{}^2 - (correct\ mean)^2}$$

$$\delta_{correct} = \sqrt{\frac{1}{19}[2016] - (10.1)^2} = \sqrt{106.1 - 102.01} = \sqrt{4.09} = \mathbf{2.02}$$

ii. Exercise!!!

❖ **Properties of the Variance and the Standard Deviation**

Variance

- It cannot be negative.
- It removes most of the demerits or drawbacks of the measures of dispersion discussed so far.
- It is only used to measure spread or dispersion around the mean of a data set.
- Its unit is the square of the unit of measurement of values. For example, if the variable is measured in *kg*, the unit of variance is $kg^{2.}$
- It is calculated based on all the observations/data in the series.
- It gives more weight to extreme values and less to those which are near to the mean.

Standard Deviation

- It cannot be negative.
- It is considered to be the best measure of dispersion.
- It is only used to measure spread or dispersion around the mean of a data set.
- For data with almost the same mean, the greater the spread, the greater the standard deviation.
- [Demerits] If the values of two series have different unit of measurement, then we cannot compare their variability just by comparing the values of their respective standard deviations.
- It is calculated based on all the observations/data in the series. Standard deviation is capable of further algebraic treatment.
- Standard deviation can be used in conjunction with the mean in order to calculate data intervals when analyzing normally distributed data.
- It is sensitive to outliers. A single outlier can raise σ and in turn, distort the picture of spread.
- Similar to the variance, standard deviation gives more weight to extreme values and less to those which are near to the mean.

- If a constant c is added to each value of a population function, then

- ☞ The new variance is the same as that of the old variance.
- ☞ The new standard deviation is also the same as that of the old standard deviation.

Proof

Consider the data items: $x_1, x_2, x_3, \ldots x_n$ having mean $\bar{x}$. Add a constant c to each data item: $x_1 + c$, $x_2 + c$, $x_3 + c$, . . . $x_n + c$ and the new mean $\bar{x} + c$.

The new variance δ^2_{new} will be:

$$= \frac{(x_1 + c - (\bar{x} + c))^2 + (x_2 + c - (\bar{x} + c))^2 + (x_3 + c - (\bar{x} + c))^2 + \cdots + (x_n + c - (\bar{x} + c))^2}{n}$$

$$= \frac{(x_1 + c - \bar{x} - c)^2 + (x_2 + c - \bar{x} - c)^2 + (x_3 + c - \bar{x} - c)^2 + \cdots + (x_n + c - \bar{x} - c)^2}{n}$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$$= \textbf{old } \boldsymbol{\delta^2}$$

NB: If all the values of a population are increased by a constant c then, the mean is also increased by c while the standard deviation remains unchanged.

Example: Given the population function, 1,3,5,7, find the variance and standard deviation. Then add a constant 2 to each data item and find the new variance and standard deviation. Compare the old and new variance and standard deviation:

- If each data item of a population function is multiplied by a constant k,
  - ☞ The new variance is k^2 times the old variance. And
  - ☞ The new standard deviation is |k| times the old standard deviation.

Proof

Consider the population function $x_1, x_2, x_3, \ldots x_n$ whose mean is $\bar{x}$ and variance $\square^2$. Multiply each data item by a constant k. New data items: $kx_1, kx_2, kx_3, \ldots cx_n$ having new mean = $k\bar{x}$. The new variance δ^2_{new} will be:

$$= \frac{(kx_1 - (k\bar{x}))^2 + (kx_2 - (k\bar{x}))^2 + (kx_3 - (k\bar{x}))^2 + \cdots + (kx_n - (k\bar{x}))^2}{n}$$

New variance δ^2_{new}

$$= \frac{k^2[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}{n}$$

$$= k^2\delta^2$$

$$\therefore\ \delta^2{}_{new} = k^2 \text{ the old variance}$$

$$\Rightarrow New\ standard\ deviation = \sqrt{k^2\delta^2} = |k|\delta$$

$$= |k|\ Old\ standard\ deviation$$

N.B: If all the values of a population are multiplied by a constant k then,

i) The new mean is k * the old mean
ii) The new standard deviation is |k| * the old standard deviation
iii) The variance also will be k^2 * old variance

Example: Given the population function, 2, 1, 4, 5, find the mean, variance and standard deviation. Then multiply each data item by constant k=3 and find the new mean, variance and standard deviation. Compare the old and new mean, variance and standard deviation:

Solution: Old data items: 2, 1, 4, and 5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+1+4+5}{4} = \frac{12}{4} = 3$$

$$\delta^2 = \frac{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$$= \frac{(2-3)^2 + (1-3)^2 + (4-3)^2 + (5-3)^2}{4}$$

$$= \frac{(-1)^2 + (-2)^2 + (1)^2 + (2)^2}{4}$$

$$= \frac{1+4+1+4}{4} = \frac{10}{4} = \frac{5}{2} = \mathbf{2.5}$$

$$\Rightarrow\ \delta = \sqrt{2.5} = \mathbf{1.58}$$

New data items: 6, 3, 12, and 15

The new mean will be:

$$\bar{x}_{new} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6+3+12+15}{4} = \frac{36}{4} = 9 \text{ and}$$

$$\delta^2{}_{new} = \frac{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$$= \frac{(6-9)^2 + (3-9)^2 + (12-9)^2 + (15-9)^2}{4}$$

$$= \frac{(-3)^2 + (-6)^2 + (3)^2 + (6)^2}{4}$$

$$= \frac{9 + 36 + 9 + 36}{4} = \frac{90}{4} = \frac{45}{2} = \mathbf{22.5}$$

$$\Rightarrow \delta^2{}_{new} = 3^2 old\ variance = 9 * old\ variance = 9 * 2.5 = 22.5\ and$$

$$\delta_{\text{new}} = \sqrt{\delta^2{}_{new}} = \sqrt{22.5} = \mathbf{4.74}$$

$$\Rightarrow \delta_{\text{new}} = |3|\delta_{\text{old}} = 3 * 1.58 = 4.74$$

- If the standard deviation of $x_1, x_2, x_3, \ldots x_n$ is S then the standard deviation of

  $a + kx_1, a + kx_2, a + kx_3, \ldots a + kx_n$, would be **|k|s.**

4.5.4.3 Coefficient of Variation

The standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation (CV).*

Coefficient of variation is used in such problems where we want to compare the variability of two or more different series. Coefficient of variation is the ratio of the standard deviation to the arithmetic mean, usually expressed in percent.

$CV = \frac{S}{\bar{x}} \times 100$. Where, S is the standard deviation of the observations.

A distribution having less coefficient of variation is said to be less variable or more consistent or more uniform or more homogeneous.

Example: Last semester, the students of Biology and Chemistry Departments took *Stat 273* course. At the end of the semester, the following information was recorded.

Department	**Biology**	**Chemistry**
Mean score	79	64
Standard deviation	23	11

Compare the relative dispersions of the two departments' scores using the appropriate way.

Solution:

Biology Department	Chemistry Department
$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{23}{79} \times 100 = 29.11\%$	$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{11}{64} \times 100 = 17.19\%$

- **Interpretation**: Since the CV of Biology Department students is greater than that of Chemistry Department students, we can say that there is more dispersion relative to the mean in the distribution of Biology students' scores compared with that of Chemistry students.

Example: The following table illustrates the frequency distribution of masses of 100 male students in Gander University.

Mass (kg)	60-62	63-65	66-68	69-71	72-74
No. of students	5	18	42	27	8

Find: a) the variance b) the standard deviation c) the coefficient of variation

d) Calculate mean deviation?

Solution:

Mass (kg)	60-62	63-65	66-68	69-71	72-74	total
No. of students(f_i)	5	18	42	27	8	100
class mark(m_i)	61	64	67	70	73	
$f_i m_i$	305	1152	2814	1890	584	6745
$f_i m_i^2$	18605	73728	188538	132300	42632	455803
$\lvert m_i - \bar{x} \rvert$	6.45	3.45	0.45	2.55	5.55	
$f_i \lvert m_i - \bar{x} \rvert$	32.25	62.1	18.9	68.85	44.4	226.5

$$\sum_{i=1}^{5} f_i m_i = 6745, \quad \sum_{i=1}^{5} m_i^2 = 455803, \quad n = \sum_{i=1}^{5} f_i = 100$$

and $$\bar{x} = \frac{\sum_{i=1}^{5} f_i m_i}{n} = \frac{6745}{100} = 67.45$$

a) $$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{5} f_i m_i^2 - \frac{(\Sigma f_i m_i)^2}{n}\right) = \frac{1}{99}\left(455803 - \frac{(6745)^2}{100}\right) = 8.61$$

b) $S = \sqrt{S^2} = \sqrt{8.61} = 2.93$

c) $CV = \frac{S}{\bar{x}} * 100 = \frac{2.93}{67.45} * 100 = 4.344$

d) $MD = \frac{\sum f_i |m_i - \bar{x}|}{n} = \frac{226.5}{100} = 2.265$

4.5.5 The Standard Scores (Z-Scores)

A standard score is a measure that describes the relative position of a single score in the entire distribution of scores in terms of the mean and standard deviation. It also gives us the number of standard deviations a particular observation lie above or below the mean.

Population standard score: $Z = \frac{x - \mu}{\sigma}$ where x is the value of the observation, μ and σ are the mean and standard deviation of the population respectively.

Sample standard score: $Z = \frac{x - \bar{x}}{S}$ where x is the value of the observation, $\bar{x}$ and S are the mean and standard deviation of the sample respectively.

Interpretation:

$$If\ Z\ is \begin{cases} positive, & the\ observation\ lies\ above\ the\ mean \\ negative, & the\ observation\ lies\ below\ the\ mean \\ zero, & the\ observation\ equals\ to\ the\ mean \end{cases}$$

Example: Two sections were given an exam in a course. The average score was 72 with standard deviation of 6 for *section 1* and 85 with standard deviation of 5 for *section 2*. Student A from *section 1* scored 84 and student B from *section 2* scored 90. Who performed better relative to his/her group?

Solution: Section 1: $\bar{x}$ = 72, S = 6 and score of student A from Section 1; x_A = 84

Section 2: $\bar{x}$ = 85, S = 5 and score of student B from Section 2; x_B = 90

Z-score of student A: $Z = \frac{x_A - \bar{x}_1}{S_1} = \frac{84 - 72}{6} = 2.00$

Z-score of student B: $Z = \frac{x_B - \bar{x}_2}{S_2} = \frac{90 - 85}{5} = 1.00$

From these two standard scores, we can conclude that student A has performed better relative to his/her section students because his/her score is two standard deviations above the mean score of *selection 1* while the score of student B is only one standard deviation above the mean score of *section 2* students.

4.6. Measure of shape

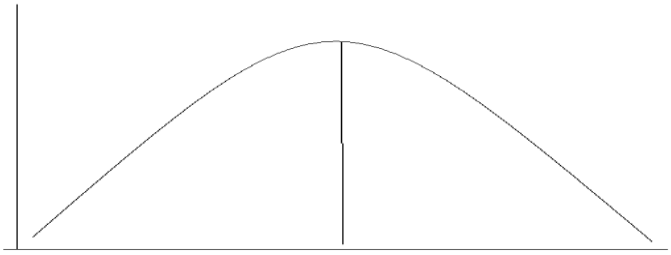
We have seen that averages and measure of dispersion can help in describing the frequency distribution. However, they are not sufficient to describe the nature of the distribution. For this purpose, we use the other concepts known as Skewness and Kurtosis.

4.6.1 Skewness

Skewness means lack of symmetry. When the values are uniformly distributed around the mean a distribution is said to be symmetrical. For example, the following distribution is symmetrical about its mean 3.

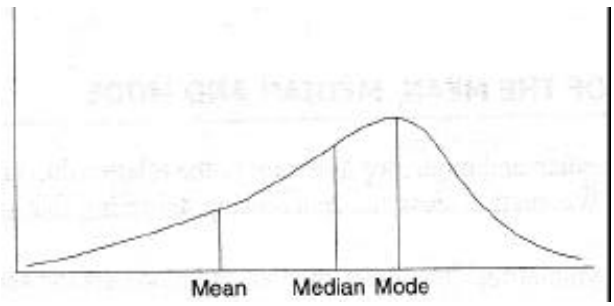
X_i :	1	2	3	4	5
f_i :	5	9	12	9	5

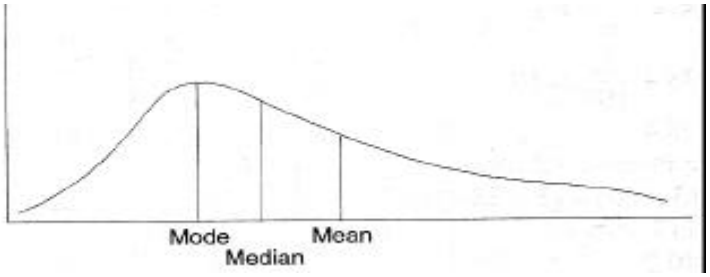
In a symmetrical distribution the mean, median and mode coincide, that is, $\overline{X} = \tilde{X} = \hat{X}$.



$\overline{X} = \tilde{X} = \hat{X}$ *Symmetrical distribution*

When a distribution is skewed to the right; mean > median > mode. If we take income distribution for different number of families; Income distribution is skewed to the right mean that a large number of families have relatively have low





Left (-) skewed Distribution

Right (+) skewed Distribution

income and a small number of families have extremely high income. In such a case, the mean is pulled up by the extreme high incomes and the relation among these three measures is as shown in figure. Here, we find that mean > median > mode. When a distribution is skewed to the left, then mode > median > mean. This is because here mean is pulled down below the median by extremely low values.

Karl Pearson's Measure of Skewness

In case the distribution is symmetric we will have Arithmetic mean. = Median = Mode; unless they will not be equal if the distribution is skewed.

Therefore, the distance between the A.M. and the Mode (A.M – Mode) can also be used as a measure of skewness. However, since the measure of skewness should be a pure number we define as $Sk = \frac{A.M - Mode}{\delta}$, Where δ is the standard deviation of the distribution.

For distribution which are bell shaped and are moderately skewed, we have an approximate relationship between the A.M, Median and mode.

A. M – Mode = 3 (A. M – Median)

Accordingly, we may define skewness as follows $Sk = \frac{3(A.M - Median)}{\delta}$

I. Bowley's Formula for Measure of Skewness

Bowley gave a measure of skewness on the assumption that in a asymmetric distribution, the second quartile is not equidistance from the first and the third quartile. Thus, Bowley's formula for the measure of skewness will be:

$$\alpha_3 = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Example: The following data is the amount of income per day for 60 households.

Income (in birr)	Number of Household
0-10	2
10-20	3
20-30	12
30-40	8
40-50	10
50-60	17
60-70	4
70-80	3
80-90	1

Find the coefficient of skewness using the Bowley's formula.

Skewness based on Central Moment

The central moment denoted by M_r is defined as:

$$M_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n-1}$$

The other measure uses the β (beta) coefficient which is given by

$$\beta_1 = \frac{M_3^2}{M_2^3}$$

where, M_2 & M_3 are, the second and the third central moments respectively.

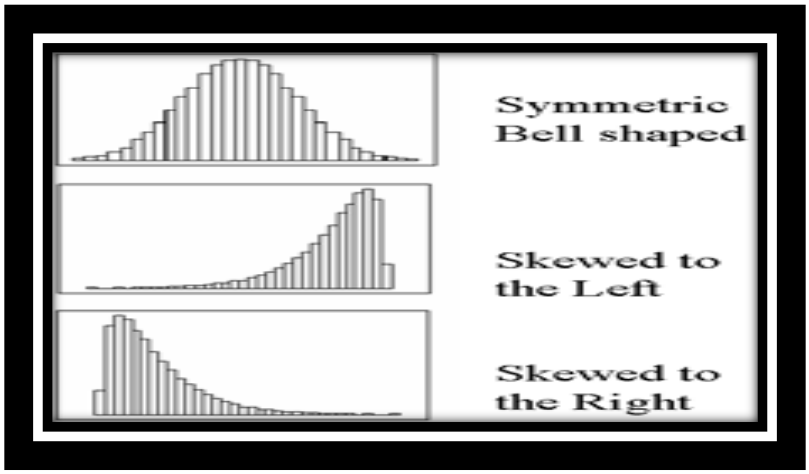
The second central moment is equivalent to the variance. The sample estimate of this coefficient is $b_1 = m_3^2/m_2^3$ where m_2& m_3 are sample central moments given by,

$$m_2 = \frac{\sum(X-\overline{X})^2}{n-1} \text{ or } \frac{\sum f(X-\overline{X})^2}{n-1}, \quad m3 = \frac{\sum(X-\overline{X})^3}{n-1} \text{ or } \frac{\sum f(X-\overline{X})^3}{n-1}$$

For a symmetrical distribution b_1 is zero. And also Skewness is positive or negative depend upon whether m_3 is positive or negative.

Example 4.12 The first four moments about mean of the distribution are 0, 2.5, 0.7, and 18.75. Test the Skewness of distribution.

For a **symmetrical distribution** Sk = 0. If the distribution **negatively skewed**, then the value of S_k is negative, and if it is **positively skewed** then S_k is positive. The range for values of S_k is from -3 to 3.



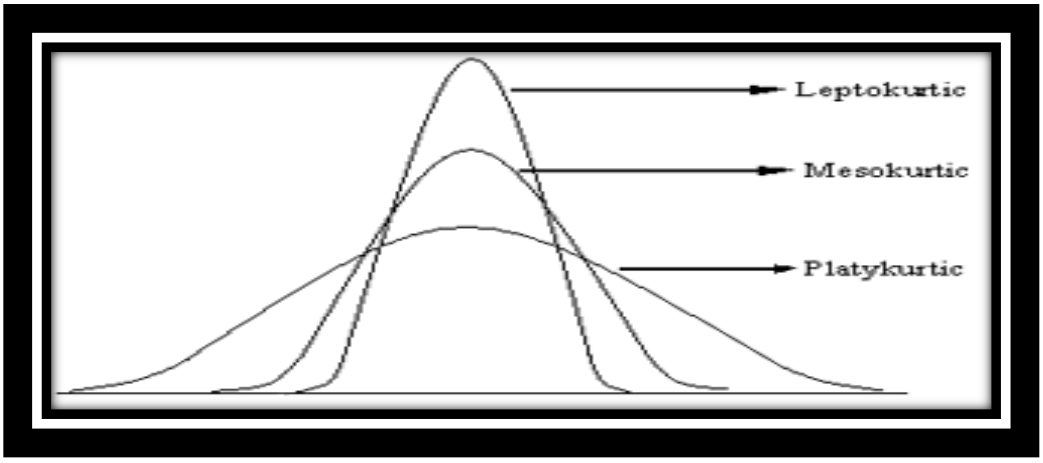
4.6.2 Kurtosis

A measure of the peakedness or convexity of a curve is known as Kurtosis. All the three curves are symmetrical about the mean. Still they are not of the same type. One has different peak as compared to that of others. Curve (1) is known as mesokurtic (normal curve); curve (2) is known as leptokurtic (leaping curve) and curve (3) is known as platykurtic (flat curve). Kurtosis is measured by Pearson's coefficient, β_2. It is given by

$$\beta_2 = \frac{M_4}{M_2^2} = \frac{M_4}{\delta^{2^2}}$$

The sample estimate of this coefficient is $b_2 = m_4/m_2^2$, where m_4 is the 4th central moment given by $m_4 = \frac{\sum (X - \bar{X})^4}{n-1}$.

The distribution is called **mesokurtic** if the value of $b_2 = 3$. When b_2 is more than 3 the distribution is said to be **leptokurtic**. And also, if b_2 is less than 3 the distribution is said to be **platykurtic.**



Example 4.13 The measure of skewness and kurtosis are given below for data in table.

Value(xi)	3	4	5	6	7	8	9	10
Frequency(f)	4	6	10	26	24	15	10	5

Value(xi)	Frequency(f)	$d=X-\overline{X}$	$f*d^2$	$f*d^3$
3	4	-3.7	54.76	-202.612
4	6	-2.7	43.74	-118.098
5	10	-1.7	28.90	-49.130
6	26	-0.7	12.74	-8.918
7	24	0.3	2.16	0.648
8	15	1.3	25.35	32.955
9	10	2.3	52.90	121.670
10	5	3.3	54.45	179.685

$\mathbf{m_2}=s^2=\frac{\sum f_i(X-\overline{X})^2}{n-1}=\frac{275}{99}=2.7777$

$\mathbf{m_4}=\frac{\sum f_i(X-\overline{X})^4}{n-1}=\frac{2074.13}{99}=20.9508,$ $\mathbf{m_3}=\frac{\sum f_i(X-\overline{X})^3}{n-1}=\frac{-43.8}{99}=-0.4424$

$\mathbf{b_1}=\frac{m_3^{\,2}}{m_2^{\,3}}=\frac{(-0.4424)^2}{(2.7777)^3}=-0.0091$, $\mathbf{b_2}=\frac{m_4}{m_2^{\,2}}=\frac{20.9508}{(2.7777)^2}=2.7153$

It is negatively skewed since m_3 is negative. The value of b_2 is 2.7153 which is less than 3. Hence the distribution is also a platykurtic.