

UNIT 1:

Statistics

Measures of Central Tendency

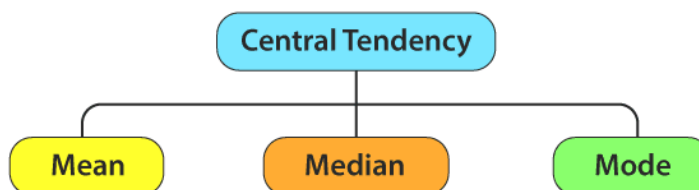
In statistics, the central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the centre of the data distribution. Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.

Definition

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

The central tendency of the dataset can be found out using the three important measures namely [mean, median and mode](#).

CENTRAL TENDENCY



Mean

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is

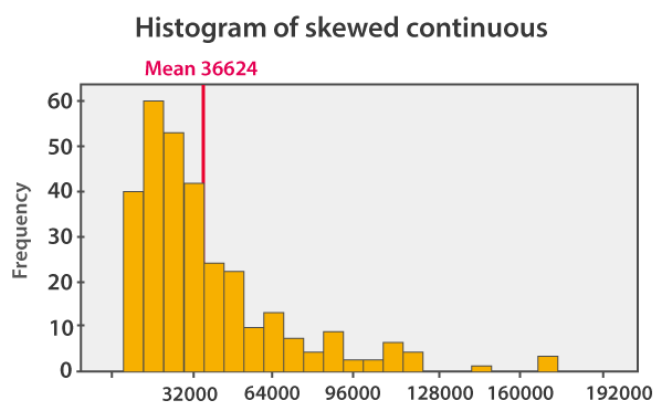
considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

- Geometric Mean
- Harmonic Mean
- Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy. The formula to calculate the mean value is given by:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.



In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

Median

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2

Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2

Here 12 is the middle or median number that has 6 values above it and 6 values below it.

Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17

Median even	
	40
	38
	35
	33
	32
	30
28	29
	27
	26
	24
	23
	22
	19
	17

When you look at the given dataset, the two middle values obtained are 27 and 29.

Now, find out the mean value for these two numbers.

i.e., $(27+29)/2 = 28$

Therefore, the median for the given data distribution is 28.

Mode

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5

Mode
5
5
5
4
4
3
2
2
1

Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

Based on the properties of the data, the measures of central tendency are selected.

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency.

Measures of Central Tendency and Dispersion

The central tendency measure is defined as the number used to represent the center or middle of a set of data values. The three commonly used measures of central tendency are the mean, median, and mode.

A statistic that tells us how the data values are dispersed or spread out is called the measure of dispersion. A simple measure of dispersion is the range. The range is equivalent to the difference between the highest and least data values. Another measure of dispersion is the [standard deviation](#), representing the expected difference (or deviation) among a data value and the mean.

Frequency Distribution Table – Data Collection

In our day to day life, recording information is very crucial. A piece of information or representation of facts or ideas which can be further processed is known as data. The weather forecast, maintenance of records, dates, time, and everything is related to data collection.

The collection, presentation, analysis, organization and interpretation of observations or data is known as statistics. We can make predictions about the nature of data based on the previous data using statistics. Statistics are helpful when a large amount of data is to be studied and observed.

The collected statistical data can be represented by various methods such as tables, bar graphs, pie charts, histograms, frequency polygons, etc.

What is Frequency Distribution Table in Statistics?

In statistics, a frequency distribution table is a comprehensive way of representing the organisation of raw data of a quantitative variable. This table shows how various values of a variable are distributed and their corresponding frequencies. However, we can make two frequency distribution tables:

- (i) Discrete frequency distribution
- (ii) Continuous frequency distribution (Grouped frequency distribution)

How to Make a Frequency distribution table?

Frequency distribution tables can be made using [tally marks](#) for both discrete and continuous data values. The way of preparing discrete frequency tables and continuous frequency distribution tables are different from each other.

In this section, you will learn how to make a [discrete frequency distribution](#) table with the help of examples.

Examples

Suppose the runs scored by the 11 players of the Indian cricket team in a match are given as follows:

25,65,03,12,35,46,67,56,00,31,17

This type of data is in raw form and is known as raw data. The difference between the measure of highest and lowest value in a collection of data is known as the range. Here, the range is-

$|67 - 00| = 67$

When the number of observations increases, this type of representation is quite hectic, and the calculations could be quite complex. As statistics is about the presentation of data in an organized form, the data representation in tabular form is more convenient.

Considering another example: In a quiz, the marks obtained by 20 students out of 30 are given as:

12,15,15,29,30,21,30,30,15,17,19,15,20,20,16,21,23,24,23,21

This data can be represented in tabular form as follows:

Table 1: Frequency Distribution Table (Ungrouped)

Marks obtained in quiz	Number of students(Frequency)
12	1
15	4
16	1
17	1
19	1
20	2
21	3
23	2
24	1
29	1
30	3
Total	20

The number of times data occurs in a data set is known as the frequency of data. In the above example, frequency is the number of students who scored various marks as

tabulated. This type of tabular data collection is known as an ungrouped frequency table.

What happens if, instead of 20 students, 200 students took the same test. Would it have been easy to represent such data in the format of an ungrouped frequency distribution table? Well, obviously no. To represent a vast amount of information, the data is subdivided into groups of similar sizes known as class or class intervals, and the size of each class is known as class width or class size.

Frequency Distribution table for Grouped data

The frequency distribution table for grouped data is also known as the [continuous frequency distribution](#) table. This is also known as the grouped frequency distribution table. Here, we need to make the frequency distribution table by dividing the data values into a suitable number of classes and with the appropriate class height. Let's understand this with the help of the solved example given below:

Question:

The heights of 50 students, measured to the nearest centimetres, have been found to be as follows:

161, 150, 154, 165, 168, 161, 154, 162, 150, 151, 162, 164, 171, 165, 158, 154, 156, 172, 160, 170, 153, 159, 161, 170, 162, 165, 166, 168, 165, 164, 154, 152, 153, 156, 158, 162, 160, 161, 173, 166, 161, 159, 162, 167, 168, 159, 158, 153, 154, 159

(i) Represent the data given above by a grouped frequency distribution table, taking the class intervals as 160 – 165, 165 – 170, etc.

(ii) What can you conclude about their heights from the table?

Solution:

(i) Let us make the grouped frequency distribution table with classes:

150 – 155, 155 – 160, 160 – 165, 165 – 170, 170 – 175

Class intervals and the corresponding frequencies are tabulated as:

Class intervals	Frequency	Corresponding data values
150 – 155	12	150, 150, 151, 152, 153, 153, 153, 154, 154, 154, 154, 154
155 – 160	9	156, 156, 158, 158, 158, 159, 159, 159, 159
160 – 165	14	160, 160, 161, 161, 161, 161, 162, 162, 162, 162, 162, 164, 164
165 – 170	10	165, 165, 165, 165, 166, 166, 167, 168, 168, 168
170 – 175	5	170, 170, 171, 172, 173
Total	50	

(ii) From the given data and above table, we can observe that 35 students, i.e. more than 50% of the total students, are shorter than 165 cm.

Arithmetic, Geometric, and Harmonic means

Algebra is the branch of mathematics that deals with numbers and symbols and provides the rules to manipulate these symbols and numbers. An algebraic expression is formed of numbers, variables, and mathematical operators. Example: $6y+2$ is an algebraic expression. Comparing numerical values is an important aspect of mathematics, and algebra helps us deal with this. Mean is another entity that is used for finding relationships between Numbers/ Variables. There are three kinds of mean available: Arithmetic Mean(AM), Geometric Mean(GM), Harmonic Mean(HM) in algebraic mathematics which helps us in calculating Mean. Now, we will discuss the three means in mathematics: Arithmetic Mean(AM), Geometric Mean(GM), Harmonic Mean(HM), and compare three of them on the basis of magnitude.

Arithmetic Mean(AM)

In mathematics, arithmetic Mean(AM) evaluates the total sum of the numbers upon the total count of numbers. Arithmetic Mean(AM) in simple terms is basically the average of all the given numbers whose Arithmetic Mean needs to be calculated. Consider two variables R and S then the arithmetic mean(AM) of these variables would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$AM = (R + S)/N$$

$$= (R + S)/2$$

Geometric Mean(GM)

In mathematics, geometric mean(GM) evaluates to the finding the nth root of the product of all the numbers whose geometric mean(GM) needs to be calculated. Consider two variables R and S then the geometric mean(GM) of these variables would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$GM = (R * S)^{1/N}$$
$$= (R * S)^{1/2}$$

Harmonic Mean(HM)

In mathematics, Harmonic Mean(HM) evaluates the total count of numbers upon the sum of reciprocal of the given numbers whose Harmonic Mean(HM) needs to be calculated. Consider two variables R and S then the Harmonic Mean(HM) of these variables would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$HM = N/(1/R + 1/S)$$
$$= 2/(1/R + 1/S)$$
$$= 2/(R + S/RS)$$
$$= 2RS/(R + S)$$

What is the comparison between the arithmetic, geometric, and harmonic means?

Solution:

Comparing Arithmetic Mean(AM), Geometric Mean(GM), Harmonic Mean(HM) on the basis of magnitude. So consider two numbers 4 and 5 replacing these variables in the above formulas.

Hence the arithmetic mean(AM) of these numbers would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$AM = (4 + 5)/2$$
$$= (4 + 5)/2$$
$$= 4.5$$

The Geometric Mean(GM) of these numbers would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$GM = (4 * 5)^{1/2}$$

$$= (4 * 5)^{1/2}$$

$$= (20)^{1/2}$$

$$= 4.47$$

The Harmonic Mean(HM) of these numbers would be given by the formula:

Let the total number of entities under consideration, here the two variables are entities so, assign entities(N) the value of 2.

$$HM = N / (1/4 + 1/5)$$

$$= 2 / (1/4 + 1/5)$$

$$= 2 / (4 + 5 / (4 * 5))$$

$$= (2 * 20) / 9$$

$$= 40/9$$

$$= 4.44$$

Comparing Arithmetic Mean(AM), Geometric Mean(GM), Harmonic Mean(HM) of the two numbers 4 and 5.

We have,

$$\text{Arithmetic Mean(AM)} = 4.5$$

$$\text{Geometric Mean(GM)} = 4.47$$

$$\text{Harmonic Mean(HM)} = 4.44$$

We see that arithmetic mean is the largest in magnitude, followed by Geometric Mean and then by Harmonic Mean.

So,

Arithmetic Mean > Geometric Mean > Harmonic Mean

or

Harmonic Mean < Geometric Mean < Arithmetic Mean

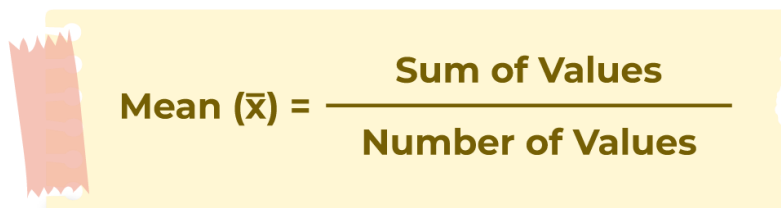
Mean, Median and Mode

Mean, Median, and Mode are measures of the **central tendency**. These values are used to define the various parameters of the given data set. The measure of central tendency (Mean, Median, and Mode) gives useful insights about the data studied, these are used to study any type of data such as the average salary of employees in an organization, the median age of any class, the number of people who plays cricket in a sports club, etc.

What is Mean?

Mean is the sum of all the values in the data set divided by the number of values in the data set. It is also called the Arithmetic Average. Mean is denoted as \bar{x} and is read as **x bar**.

The formula to calculate the mean is,


$$\text{Mean } (\bar{x}) = \frac{\text{Sum of Values}}{\text{Number of Values}}$$

Mean Formula

The formula to calculate the mean is,

$$\text{Mean } (\bar{x}) = \text{Sum of Values} / \text{Number of Values}$$

If $x_1, x_2, x_3, \dots, x_n$ are the values of a data set then the mean is calculated as:

$$\bar{x} = (x_1 + x_2 + x_3 + \dots + x_n) / n$$

Example: Find the mean of data sets 10, 30, 40, 20, and 50

Solution:

Mean of the data 10, 30, 40, 20, 50 is

Mean = (sum of all values) / (number of values)

Mean = (10 + 30 + 40 + 20 + 50) / 5

= 30

Mean of Grouped Data

Mean for the grouped data can be calculated by using various methods. The most common methods used are discussed in the table below,

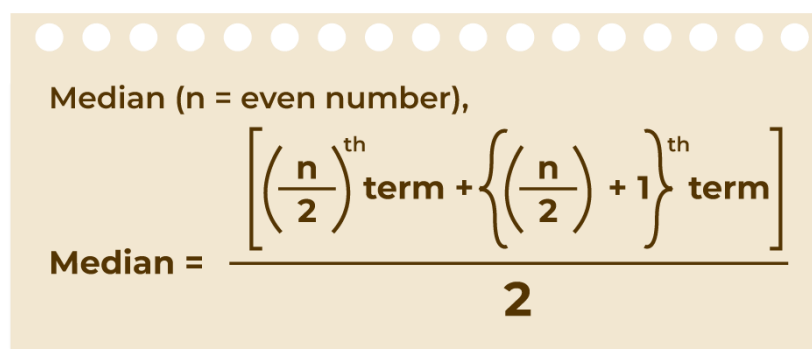
Direct Method	Assumed Mean Method	Step Deviation Method
<p>Mean</p> $\bar{x} = \sum f_i x_i / \sum f_i$ <p>where, $\sum f_i$ is the sum of all frequencies</p>	<p>Mean</p> $\bar{x} = a + \sum f_i x_i / \sum f_i$ <p>where, a is Assumed mean d_i is equal to $x_i - a$ $\sum f_i$ the sum of all frequencies</p>	<p>Mean</p> $\bar{x} = a + h \sum f_i x_i / \sum f_i$ <p>where, a is Assumed mean u_i = $(x_i - a)/h$ h is Class size $\sum f_i$ the sum of all frequencies</p>

Also, Check Mean

What is Median?

A Median is a middle value for sorted data. The sorting of the data can be done either in ascending order or descending order. A median divides the data into two equal halves.

The formula to calculate the median of the number of terms if the number of terms is even is shown in the image below,



Median (n = even number),

$$\text{Median} = \frac{\left[\left(\frac{n}{2} \right)^{\text{th}} \text{ term} + \left\{ \left(\frac{n}{2} \right) + 1 \right\}^{\text{th}} \text{ term} \right]}{2}$$

The formula to calculate the median of the number of terms if the number of terms is odd is shown in the image below,

Median (n = odd number),

$$\text{Median} = \left[\frac{(n+1)}{2} \right]^{\text{th}} \text{ term}$$

Median Formula

The formula for the median is,

If the number of values (n value) in the data set is odd then the formula to calculate the median is,

$$\text{Median} = \left[\frac{(n + 1)}{2} \right]^{\text{th}} \text{ term}$$

If the number of values (n value) in the data set is even then the formula to calculate the median is:

$$\text{Median} = \left[\frac{(n/2)^{\text{th}} \text{ term} + \{(n/2) + 1\}^{\text{th}} \text{ term}}{2} \right]$$

Example: Find the median of given data set 30, 40, 10, 20, and 50

Solution:

Median of the data 30, 40, 10, 20, 50 is,

Step 1: Order the given data in ascending order as:

10, 20, 30, 40, 50

Step 2: Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.

Step 3: Here, n = 5 (odd)

$$\text{Median} = \left[\frac{(n + 1)}{2} \right]^{\text{th}} \text{ term}$$

$$\begin{aligned} \text{Median} &= \left[\frac{(5 + 1)}{2} \right]^{\text{th}} \text{ term} \\ &= 30 \end{aligned}$$

Median of Grouped Data

The median of the grouped data median is calculated using the formula,

$$\text{Median} = l + \left[\frac{(n/2 - cf)}{f} \right] \times h$$

where

l is lower limit of median class

n is number of observations

f is frequency of median class

h is class size

cf is cumulative frequency of class preceding the median class.

What is Mode?

A mode is the most frequent value or item of the data set. A data set can generally have one or more than one mode value. If the data set has one mode then it is called "Uni-modal". Similarly, If the data set contains 2 modes then it is called "Bimodal" and if the data set contains 3 modes then it is known as "Trimodal". If the data set consists of more than one mode then it is known as "multi-modal" (can be bimodal or trimodal). There is no mode for a data set if every number appears only once.

The formula to calculate the mode is shown in the image below,



Mode Formula

Mode = Highest Frequency Term

Example: Find the mode of the given data set 1, 2, 2, 3, 3, 4, 5

Solution:

Given set is {1, 2, 2, 2, 3, 3, 4, 5}

As the above data set is arranged in ascending order.

By observing the above data set we can say that,

Mode = 2

As, it has highest frequency (3)

Mode of Grouped Data

The mode of the grouped data is calculated using the formula,

$$\text{Mode} = l + [(f_1 + f_0) / (2f_1 - f_0 - f_2)] \times h$$

where,

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class

h is the size of class intervals

l is the lower limit of modal class

Also, Check Mode

Relation between Mean Median Mode

For any group of data, the relation between the three central tendencies mean, median, and mode is shown in the image below,


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

Mean Median Mode Formula

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

Range

It is the difference between the highest value and the lowest value. It is a way to understand how the numbers are spread in a data set. The range of any data set is easily calculated by using the formula given in the image below,


$$\text{Range} = \text{Highest Value} - \text{Lowest Value}$$

Range Formula

The formula to find the Range is:

$$\text{Range} = \text{Highest value} - \text{Lowest Value}$$

Example: Find the range of the given data set 12, 19, 6, 2, 15, 4

Solution:

Given set is {12, 19, 6, 2, 15, 4}

Here,

Lowest Value = 2

Highest Value = 19

*Range = 19 - 2
= 17*

Measures of Dispersion

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, [standard deviation](#), quartile deviation, etc.

The types of absolute measures of dispersion are:

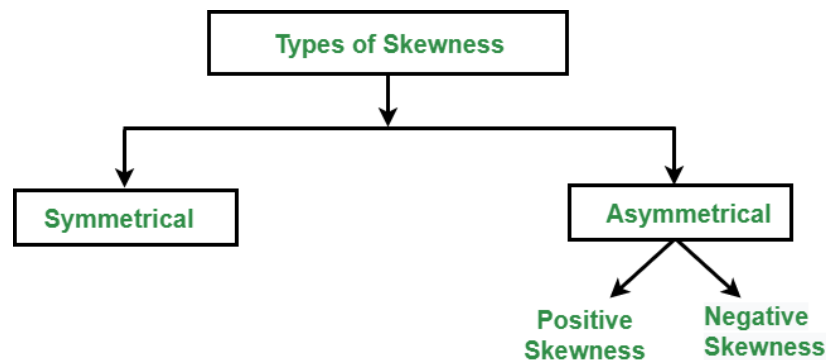
1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3, 5, 6, 7 => Range = 7 - 1 = 6
2. **Variance:** Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance.
$$\text{Variance } (\sigma^2) = \sum (X - \mu)^2 / N$$
3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e.
$$\text{S.D.} = \sqrt{\sigma}$$

4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Skewness and Kurtosis

Skewness is an important statistical technique that helps to determine asymmetrical behavior than of the frequency distribution, or more precisely, the lack of symmetry of tails both left and right of the frequency curve. A distribution or dataset is symmetric if it looks the same to the left and right of the center point.

Types of skewness: The following figure describes the classification of skewness:



Types of Skewness

1. Symmetric Skewness: A perfect symmetric distribution is one in which frequency distribution is the same on the sides of the center point of the frequency curve. In this, Mean = Median = Mode. There is no skewness in a perfectly symmetrical distribution.

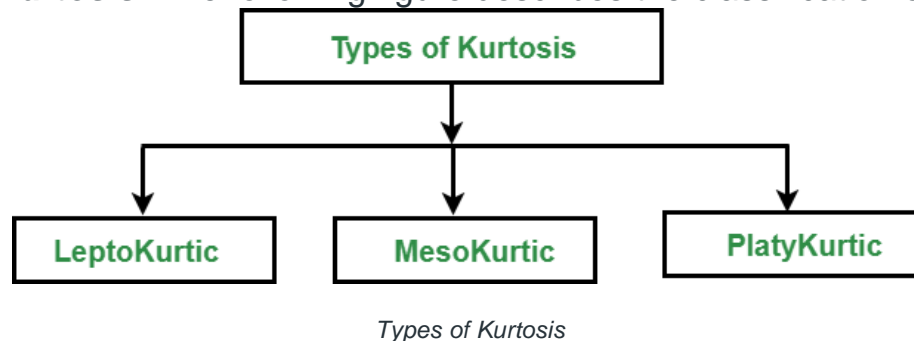
2. Asymmetric Skewness: A asymmetrical or skewed distribution is one in which the spread of the frequencies is different on both the sides of the center point or the frequency curve is more stretched towards one side or value of Mean. Median and Mode falls at different points.

- **Positive Skewness:** In this, the concentration of frequencies is more towards higher values of the variable i.e. the right tail is longer than the left tail.
- **Negative Skewness:** In this, the concentration of frequencies is more towards the lower values of the variable i.e. the left tail is longer than the right tail.

Kurtosis:

It is also a characteristic of the frequency distribution. It gives an idea about the shape of a frequency distribution. Basically, the measure of kurtosis is the extent to which a frequency distribution is peaked in comparison with a normal curve. It is the degree of peakedness of a distribution.

Types of kurtosis: The following figure describes the classification of kurtosis:



1. **Leptokurtic:** Leptokurtic is a curve having a high peak than the normal distribution. In this curve, there is too much concentration of items near the central value.
2. **Mesokurtic:** Mesokurtic is a curve having a normal peak than the normal curve. In this curve, there is equal distribution of items around the central value.
3. **Platykurtic:** Platykurtic is a curve having a low peak than the normal curve is called platykurtic. In this curve, there is less concentration of items around the central value.

Difference Between Skewness and Kurtosis

Sr. No.	Skewness	Kurtosis
1.	It indicates the shape and size of variation on either side of the central value.	It indicates the frequencies of distribution at the central value.
2.	The measure differences of skewness tell us about the magnitude and direction of the asymmetry of a distribution.	It indicates the concentration of items at the central part of a distribution.

Sr. No.	Skewness	Kurtosis
3.	It indicates how far the distribution differs from the normal distribution.	It studies the divergence of the given distribution from the normal distribution.
4.	The measure of skewness studies the extent to which deviation clusters is are above or below the average.	It indicates the concentration of items.
5.	In an asymmetrical distribution, the deviation below or above an average is not equal.	No such distribution takes place.