

## MODULE 3

# Basic Statistical Concepts

You will learn about the 'Basic Statistical Concepts' in this module.

### Module Learning Objectives

At the end of this module, you will be able to:

- Explain the basic concepts of statistics along with its classification: descriptive and inferential statistics.
- Understand the measures of scales: nominal, ordinal, interval and ratio scales.
- Describe the measures of location: mean, median and mode.
- Enumerate the measures of variability: range, quartiles and interquartile range, variance and standard deviation.
- Explain the measures of shape: skewness and kurtosis.



### Module Topics

The following topics that will be covered in the module:

1. Introduction to Statistics
2. Scale of Measurements (Nominal, Ordinal, Ratio and Interval)
3. Measures of Location
4. Measures of Variability/Spread
5. Measures of Shape



## 1.2 Classification of Statistical Methods

The following illustration depicts the two major categories of statistical analysis.



We just learnt about the introduction to Statistics. Let's now learn about the classification of statistical methods. As depicted in the illustration above, there are two major categories of statistical methods. Each one is important for specific purposes. Some of the measures may be similar, but the goal and methodologies differ. Two major categories of statistical analysis are:

1. **Descriptive Statistics:** As the name suggests, the purpose of descriptive statistics is to collect, organize, summarize and describe data. It deals with a larger chunk of data and describes it with the help of summary charts and tables. The goal is to describe data and not derive conclusions out of it. Descriptive statistics is used to describe the characteristics of a sample set.
2. **Inferential Statistics:** Inferential statistics applies complex mathematical calculations to derive an inference and make decisions about a large population, based on the study of a smaller sample. Inferential statistics is used for making predictions.

## 1.3 Descriptive Statistics

### Descriptive Statistics:

- Is a set of methods to describe the collected data.
- Aims to describe the characteristics of a sample data set and understanding the specific set of observations.
- Summarizes and represents data using graphs, charts and tables.
- Involves following methods:
  - ↳ Measurement of central tendency
  - ↳ Measurement of dispersion/spread
  - ↳ Measurement of skewness and kurtosis
  - ↳ Exploration of relationships of paired data

Descriptive statistics is simple and straightforward. It describes a data set. For example, you will use descriptive statistics, if you want to study about a group of people and their shopping patterns. You will simply collect the details about their shopping behaviour, days on which they shop more, their items of interest, etc. You will summarize the data and present it using graphs and charts, which will

give a picture of the group's properties. The sample might be a representative of a larger population, but descriptive statistics do not intend to draw conclusions or make predictions about the larger population using the sample data.

Some more scenarios where descriptive statistics is applied: the portion of salary that is spent on vegetables each month; batting average of MS Dhoni in the last five years; the number of seats that a political party won during the last election.

There is no uncertainty associated with descriptive statistics, as we describe only the sample or population that we actually measure and there's no assumption. In this process, we actually take a considerably large number of data points and reduce them using summary values and graphs. The major advantage of this method over excel sheets and tables of data is that it allows one to gain more insights and visualize the data patterns.

Let's now have a look at the methods that are used in descriptive statistics.

- **Measurement of the central tendency:** Central tendency refers to the measure of finding out where most of the values in a given data set fall, i.e., the general trends that prevail among the values. Mean (average), Median (mid-value) and Mode (most commonly occurring value) are referred to as the measures of central tendency.
- **Measurement of dispersion/spread:** Measures of dispersion is the way of describing how the data is distributed and relate to each other. It helps us in finding out how far the data extends from the centre. If the dispersion is low, it indicates that the data is tightly clustered around the centre value and a higher dispersion value indicates scattered data. Using a graph, frequency distribution can be plotted. Common measures of dispersion/spread include:
  - Range - the entire range of values present in a data set.
  - Frequency distribution - frequency of occurrence of a particular value in a data set.
  - Quartiles - subgroups formed within a data set, after the values are divided into four equal parts across the range.
  - Variance - the amount of spread that exists among the data.
  - Standard deviation - the spread of data in accordance with the mean.
- **Measurement of skewness and kurtosis:** This measure tells about the symmetric nature of values in a given data set. If data is asymmetric, it will be skewed to the right or left of the central value. Similarly, Kurtosis tells us about the shape of the tails of the distribution on the far left or far right.
- **Exploration of relationships of paired data:** Relationships of data can be explored by measuring the correlation or by using scatterplots. Correlation refers to the change in one of the variables in a particular direction, in response to the change in the other variable. A correlation coefficient is the measure of the direction and strength of this tendency to vary together.

In the forthcoming sections and modules, we'll see each of these in detail.



## 1.4 Inferential Statistics

### Inferential Statistics:

- Is a set of methods used to make a generalization, estimate, prediction or decision.
- Aims to test a hypothesis and derive conclusions about a population based on the sample.
- Analysis results are generalized from a sample to a larger population.
- Includes the following analysis tools:
  - ↳ Hypothesis tests
  - ↳ Confidence Intervals
  - ↳ Regression analysis

In inferential statistics, you draw a conclusion, i.e., an inference about a larger population by analysing a sample data set taken from the population. Once you derive a conclusion from the sample, you will then generalize it to the population. There needs to be the confidence that the sample accurately reflects the population.

Let's again take the example of people's shopping patterns. You have taken data of 100 people from a group of 1000. Around 50 of them buy chocolates a lot in April in a given year. With this data, you predict that 50% of all the population will buy more chocolates in April in the coming year. There's an uncertainty in this prediction, as people's buying behavior might change.

Let's look at the other examples given the previous slide as well: with the current month's spend on vegetables, you can estimate the amount that will be spent on vegetables the whole year; you can predict the result of a match in accordance with MS Dhoni's batting average; with previous election results you can try to predict the number of seats a party will win in the upcoming elections. Again, all these are associated with uncertainties.

In inferential statistics, random sampling is done, instead of picking up the group of choice, to have more confidence that the sample represents the population. When we have a small data set as the sample from a large population, the mean of the sample might not reflect the actual mean of the population. This difference between the sample statistics and the population statistics is referred to as 'sampling error'. Inferential statistics incorporates the estimates of these sampling errors into the results.

Most common tools used in inferential statistics are:

- **Hypothesis tests:** Consider a clinical trial where a drug is tested in a specific group of people and controls. Using Hypothesis tests, we can find out whether the drug will have the similar effect on the entire population. Hypothesis tests are thus used to make a claim about the population by analyzing the sample data set.

- **Confidence intervals:** Confidence intervals give the range of values for an unknown parameter, like mean or standard deviation, of the population by measuring the sample. This is expressed in terms of an interval and the degree of confidence that the parameter is, within the interval. Confidence intervals incorporate the sample errors and uncertainties to create the range within which the population values are likely to fall.
- **Regression analysis:** Regression analysis is used to describe a relationship between a set of independent variables and a dependent variable. Regression analysis also makes use of hypothesis tests to find if the relationship identified in the sample data exists in the population as well.

All these tools will be explained in detail in upcoming sections/modules. Some of the other techniques that statisticians use to explore the relationship among data, thereby to create inferential statistics are: linear regression analyses, logistic regression analyses, ANOVA, correlation analysis, structural equation modeling, and survival analysis. Common tests of significance include the chi-square and t-test. These methods tell the statistician if the results of the sample analysis represent the population.

We've now covered the introduction to statistics, where we learnt about the classification and the measures involved. We're now going to learn about the various measures of statistics.

## What did You Grasp?



1. State True or False.  
Measurement of skewness is part of descriptive statistics.  
A) True  
B) False
2. Select all that apply.  
Which of the following options is/are true about inferential statistics?  
A) Inferential statistics involves random sampling  
B) Inferential statistics is used to describe a data set  
C) Inferential statistics is associated with uncertainties  
D) Measure of central tendency belongs to inferential statistics

## 2 Scale of Measurements (Nominal, Ordinal, Ratio and Interval)

### 2.1 Introduction to Measurement Scales

#### Scales of measurement:

- Ways in which variables or numbers are defined and categorized.
- Properties of measurement scales determine their appropriateness for use of analyses. These include:
  - ↳ Identity
  - ↳ Magnitude
  - ↳ Equal intervals
  - ↳ Absolute zero

Before doing any statistical procedure, we need to measure something. A scale gives the unit of measurement like categories, inches, seconds, etc. and the range of possible values. The scale affects the way in which data values and statistics using those data values are interpreted. It does not change the way that the data values can be analyzed statistically (although some of those statistical analyses would make no sense when you try to interpret them).

The unit can be either continuous (allows partial units, or decimal, values) or discrete (only whole units), but the value of any particular unit has the same meaning for all objects that are assigned that value on the scale. A scale can have an infinite range of possible values, i.e., a measure can take on any real-number value, or may be limited to as few as two possible values. If there is only one possible value, it is a constant, and no need for a scale.

Measurement scales can have the following properties.

- Identity - Each value on the measurement scale has a unique meaning.
- Magnitude - The ability to know if one score is greater than, less than, or equal to another score. Values on the measurement scale have an ordered relationship to one another.
- Equal Intervals refers to the magnitude that is represented by a unit of measurement that is the same no matter where it falls on the scale. Units along the scale are equal to one another.
- Absolute Zero refers to the scale that has a point equal to zero or a point where there is no score. Below absolute zero, no value exists.

## 2.2 Nominal Scales

### Nominal scales:

- The simplest scale of measurement, having only the identity property.
- Not a measure of quantity, but measures the quality, i.e., identity and difference.
- Indicates a category - A categorical variable, called nominal variable - for mutually exclusive categories.
- In a nominal scale, numbers identify and classify objects.

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category but have no inherent numerical value with respect to magnitude. A nominal scale uses categories that have no particular order. The units of a nominal scale are categories that do not overlap; an object cannot be in more than one category. Automobile make (Ford, Chevy, Honda, Toyota, etc.), teams (Cougars, Owls, Tigers, etc.), and gender (female, male) are nominal scale variables.

Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

## 2.3 Ordinal Scales

### Ordinal scale:

- Refers to order in measurement, indicating the direction, in addition to providing nominal information.
- Labels represent an order that indicates either preference or ranking.
- Numbers indicate the relative positions of objects, but not the magnitude of differences between them.
- Has both identity and magnitude properties.



The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale. An ordinal scale uses categories that have a particular order. The units of an ordinal scale are categories that do not overlap and defined such that a higher category has more of some property and characteristic than the preceding category.

An example of an ordinal scale in action would be the results of a horse race, reported as “win”, “place”, and “show”. We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a mile.

Class year (freshman, sophomore, junior, senior), game innings or quarters (1st, 2nd, 3rd, etc.), and meal (breakfast, brunch, lunch, hors-d’oeuvres, salad, soup, meat, dessert, etc.) are ordinal scale variables.

Nominal and ordinal scales involve discrete variables.

## 2.4 Interval Scales

### Interval scales:

- Have identity, magnitude and equal intervals.
- Magnitude between the consecutive intervals are equal.
- Scale that represents quantity and has equal units.
- Do not have a true zero, i.e., zero is just an additional point of measurement in the scale.

The interval scale of measurement has the properties of identity, magnitude, and equal intervals. The unit of measurement in an interval is equal across the whole range of possible values. There is no overlapping between the units and they relate an order to the property being measured. As mentioned above, an interval scale does not have a “true zero” value. The value of zero does not mean that there is no characteristic; zero is just a number used to describe a position on the scale. Negative values are also allowed by interval variables.

A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.



With an interval scale, you know not only whether different values are bigger or smaller, you also know how much bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Other examples of interval scales include attitude scales, calendar years.

## 2.5 Ratio Scales

### Ratio scale:

- The most sophisticated scale of measurement.
- Has all the four properties, identity, magnitude, equal intervals and absolute zero.
- Has a consistent unit of measurement.
- True zero value indicates absence of a characteristic.
- Measure of the magnitude of the data value using equal intervals between values and absolute zero.

Ratio scales are the most sophisticated scales of measurement and possess all the four properties of real numbers, like identity, magnitude, equal intervals and absolute zero. Ratio scales have an origin that makes them different from the other three scales of measurement. Ratio scores have an infinite range and thus are continuous.

A ratio scale has a consistent unit of measurement. The true zero value indicates an absence of a characteristic. The units do not overlap and relate an order to the property being measured. Ratio scales also allow comparisons between values, like twice as high as, or one-half of, etc.

Some of the examples for ratio scales are mass, length, height, time, distance, energy, and many laboratory values. For example, the unit of energy (calories) has the same meaning at all parts of the scale. A zero value here indicates the absence of energy. Negative values do not have any meaning unless the definition of the measure is changed. For example, negative calories indicate an energy deficit rather than energy gain.

Interval and Ratio scales involve continuous variables.

## What did You Grasp?



1. Select all that apply.  
In a centigrade scale for temperature measurement, which of the following properties is/are satisfied?
  - A) Identity
  - B) Magnitude
  - C) Equal intervals
  - D) Absolute zero
2. State True or False.  
An ordinal scale exhibits only identity and magnitude.
  - B) True
  - C) False

## 3 Measures of Location

### 3.1 Measures of Location

#### Measures of Location:

- One of the tools covered under descriptive statistics.
- Measure of Location summarizes the data set by giving the range of the data values that describes its location relative to the entire data set arranged according to the magnitude.
- The measure of central tendency is the commonly used method to identify the centre of the data set.
- Used to describe a data set by identifying the central position within that data set.
- Measures of central tendency:
  - ↳ Mean
  - ↳ Median
  - ↳ Mode

Measure of location comes under descriptive statistics, which is also referred to as summary statistics. Location measures summarize the data and find one value that is typical for the data set, i.e., a single value that is a representative of the entire data set. Thus the measure of location is useful where it is not possible to list all the data, instead, a representative data value is used to describe the complete data set.

Most commonly applied measure of location is the measure of central tendency, i.e., identifying the centre of the data set, around which all other values are clustered. Measures of central tendency include: Mean, Median and Mode. We'll see each one of these in detail in the upcoming slides.

Other measures of location include:

- Minimum - the smallest value in the data set.
- Maximum - the largest value in a data set.

## 3.2 Mean

---

### Mean:

- The most popular measure of central tendency.
- Also referred to as the average value in the data set.
- Defined as the sum of all values in a data set, divided by the total number of values in a data set.
- Example: Let  $x_1, x_2, x_3, \dots, x_n$  be the values in the data set, and there are 'n' number of values in the data set:

Mean =

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad \text{or} \quad \frac{\sum x}{n}$$

You might have heard the term average countless number of times. What does it refer to? In statistics, we use the term mean in place of average. Mean is equal to the sum of all values in a data set divided by the total number of values in the data set. The Greek letter 'mu'  $\mu$  is used to represent mean.

The formula for calculating mean is given in the slide.

Let's take an example now: the temperature in Mumbai on five consecutive days is 35°C, 38°C, 40°C, 34°C and 38°C. The mean temperature of these five days is the sum of the temperatures on all five days divided by 5 i.e.,  $35+38+40+34+38/5$ ,  $187/5$ , which is equal to 37°C.

You might notice that the mean value is not one of the values in the data set. But mean is important in statistical analyses, as it minimises errors in the prediction of any value in the given data set. Also, mean is the only measure of central tendency, where the sum of the deviations of each value from the mean is always zero.

For mean, the distance between two numbers is defined to be the square of their difference. The sum of the squares of the differences between the data and the mean is smaller than the sum of squares

of the differences between the data and any other number. Similarly, the root mean square (rms) of the differences from the mean is smaller than the rms of the list of differences from any other number

Mean has a disadvantage, i.e., mean value is affected by the influence of outliers or out of range values. For example, if the distribution of data is not uniform and has extremely low or high numerical values compared to the mean. In this case, mean might not reflect the actual characteristic of the data set.

### 3.3 Median

#### Median:

- Middle value in a data set, after it is arranged in the order of magnitude.
- Values are first arranged from the smallest to the largest.
- Median is the middle value of the data set if the total number of values is an odd number; average of the middle two values, if the total number of values is an even number.
- Median =  $[(n+1)/2]$ th value, if there are odd number of values.
- Median = average of the  $[n/2]$ th and the  $[(n/2)+1]$ th values, if there are an even number of values.

Median is defined as the middle point of the data set after it is organized in the order of magnitude. It is estimated by first ordering the data from smallest to largest and then counting upwards for half the observations. The estimate of the median is either the value at the centre of the ordering in the case of an odd number of values, or the simple average of the middle two values if the total number of values is even.

To be specific, if there are an odd number of observations, median is the  $[(n+1)/2]$ th value, and if there are an even number of values, it is the average of the  $[n/2]$ th and the  $[(n/2)+1]$ th values.

#### Example:

Suppose we have the data below:

65, 55, 89, 56, 35, 14, 56, 55, 87, 45, 92

Rearrange that data into order of magnitude (smallest first):

14, 35, 45, 55, **56**, 56, 65, 87, 89, 92

Median is the middle value - in this case, 56 (highlighted in bold). It is the middle value because there are 5 values before it and 5 values after it.



Now, let's imagine if there are only 10 values? Take the middle two scores and average the result. So, if we look at the example below:

65, 55, 89, 56, 35, 14, 56, 55, 87, 45

Rearrange that data into order of magnitude (smallest first):

14, 35, 45, 55, 55, 56, 56, 65, 87, 89

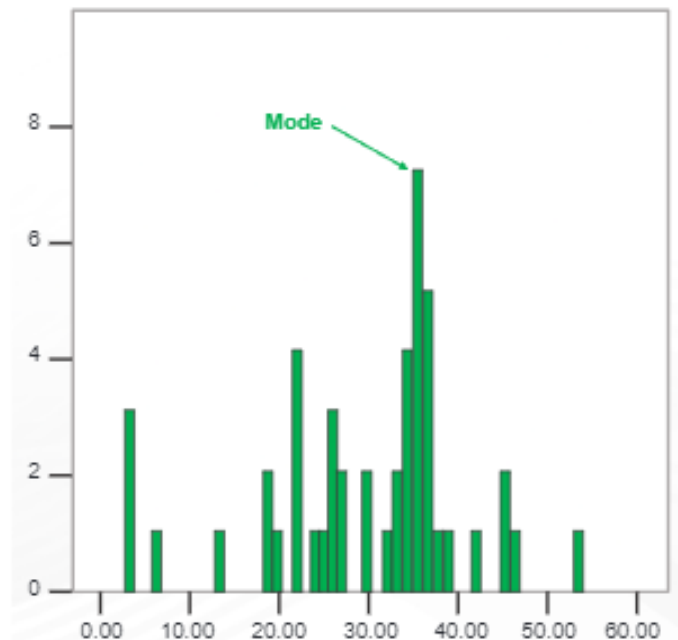
Only now we have to take the 5th and 6th score in our data set and then find the mean of the two (use the formula in the previous slide) to get a median of 55.5.

For median, the distance between two numbers is defined to be the absolute value of their difference. That is, the sum of the absolute values of the differences between a median and the data is no larger than the sum of the absolute values of the differences between any other number and the data.

### 3.4 Mode

#### Mode:

- Most frequently occurring value in the data set, i.e. value of the random sample that occurs with the greatest frequency.
- Used in a qualitative fashion.
- Applicable to all levels of data measurement, i.e., nominal, ordinal, interval and ratio scales.
- Bimodal - if values occur most frequently, two modes.
- Multimodal - Data sets with more than two modes.




The mode of a set of data is the most common value among the data. In other words, mode is the most frequently occurring value in a data set. If the data are grouped, the grouping with the highest frequency is the mode. It is rare that several data coincide exactly, unless the variable is discrete, or the measurements are reported with low precision.

Mode is not used much in statistical analysis since it depends on the accuracy with which the data are measured; although it may be useful for categorical data to describe the most frequent category. The expression 'bimodal' distribution is used to describe a distribution with two peaks in it and multimodal has more than two modes in it. This can be caused by mixing populations. For example, height might appear bimodal if one had men and women on the population.

Looking at the temperature example from Mean calculation: 35°C, 38°C, 40°C, 34°C and 38°C. Mode of this data set is 38°C since it occurs more number of times than the other values.

For mode, the distance between two numbers is defined to be zero if the numbers are equal, and one if they are not equal. That is, the number of data that differ from a mode is no larger than the number of data that differ from any other value. Equivalently, a mode is a number from which the fewest possible data differ: a “most common” value.

## What did You Grasp?



1. Calculate the mean of the following data set - 60, 56, 61, 68, 51, 53, 69, 54.  
A) 58  
B) 56  
C) 59  
D) 61
2. Find the median of the given data set - 6, 5, 4, 1, 7, 3  
A) 3  
B) 4.5  
C) 7.5  
D) 5

## 4 Measures of Variability/Spread

### 4.1 Measures of Variability/Spread

#### Measures of Spread:

- A measure of how the data is dispersed or spread around the mean, i.e. the average.
- Is used in quantitative data, as the variables can be arranged in a logical order, with a low and high value.
- Measures of spread:
  - ↳ Range
  - ↳ Quartiles and Interquartile range
  - ↳ Variance
  - ↳ Standard deviation

Measures of spread summarise the data in a manner that described the scatteredness of the values and how they differ from the mean value. Measures of variability need to have the following properties.

- The measure should be proportional to the spread of the data (small when the data are clustered together, and large when the data are widely spread).
- The measure should be independent of the number of values in the data set (otherwise, simply by taking more measurements the value would increase even if the scatter of the measurements was not increasing).
- The measure should be independent of the mean.

## 4.2 Range

### Range:

- The simplest measure of variability.
- The distance between the smallest value and the largest value in a dataset.
- Represents the width of the smallest interval that contains all the data.
- The range of a list is calculating the difference between the largest value and the smallest value.
- Affected by outliers, as variance may either be too low or too high because of outliers.

Range is defined as the distance between the smallest value to the largest value. Range of a data set is calculated by finding the difference between the largest and the smallest values in the data set.

For example, the range of the following data set: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8 is 4, the difference between the highest value (8 ) and the lowest value (4).

## 4.3 Quartiles and Interquartile Range

### Quartiles range:

- Divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters:
  - ↳ The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.
  - ↳ The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.
  - ↳ The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.
  - ↳ The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles and describes the middle 50% of values when ordered from lowest to highest.

The Interquartile Range (IQR) is the difference between the upper quartile (75th percentile) and the lower quartile (25th percentile). It is the width of the interval that contains the middle 50% of the data. It is insensitive to the most extreme values of the data (assuming that there are more than four data). The IQR is resistant: changing just one value has a limited effect on it. Neither the range nor the IQR is a range of numbers, despite their names—each is a single number.

A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts) or percentiles (hundred equal parts).

For example, consider the data set: 4, 5, 5, Q1 5, 6, 6 Q2 6, 6, 7 Q3 7, 7, 8. The quartiles are as follows:

$$Q1 = (5+5) / 2 = 5$$

$$Q2 = (6+6) / 2 = 6$$

$$Q3 = (7+7) / 2 = 7$$

The IQR for Dataset A is = 2

$$IQR = Q3 - Q1 = 2$$

## 4.4 Variance

### Variance:

- Variance is a measure of how the spread-out a data set is.
- Calculated as the average squared deviation of each number from the mean of a data set.



- The formula for calculating variance ( $S^2$ ) of a data set is:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Variance is a measure of how the spread-out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set.

Use this step-by-step approach to find the variance of a data set.

- Calculate the mean.
- Subtract the mean from each value.
- Square each of the resulting values.
- Add these squared results together.
- Divide this total by the number of values (variance,  $S^2$ ).

## 4.5 Standard Deviation (SD)

---

**SD:**

- The mean distance of all values from the overall mean.
- The square root of variance.
- SD(s) of a sample data set can be calculated using the formula:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$


(where  $x_i$  is the individual value,  $\bar{x}$  is the mean of the sample set,  $n$  is the number of values in the distribution.)

In datasets with a small spread, all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.

The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency.

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation. In that sense, the standard deviation is a good indicator of the presence of outliers. This makes standard deviation a very useful measure of spread for symmetrical distributions with no outliers. The units of the SD are the same as the original units of measurement. For example, if the list is comprised of measurements of heights in inches, the SD has units of inches.

## What did You Grasp?



1. State True or False.

The range of a data set is not affected by outliers.

A) True  
B) False

## 5 Measures of Shape

### 5.1 Measures of Shape

#### Measures of shape:

- Describe the pattern of distribution of values within a data set.
- Data distribution may be:
  - Symmetric] - two sides of the distribution are a mirror image of each other. e.g.. Normal distribution
  - Skewed - asymmetrical distribution where the values tend to be more frequent around the high or low ends of the x-axis.



Measures of shape help us in identifying the pattern of data distribution within a data set. Distribution of values in a data set may be symmetrical or asymmetrical. Common examples of symmetry and asymmetry are the 'normal distribution' and the 'skewed distribution', respectively.

The distribution shape of quantitative data can be described as there is a logical order to the values, and the 'low' and 'high' end values on the x-axis of the histogram are able to be identified. The distribution shape of a qualitative data cannot be described as the data are not numeric.

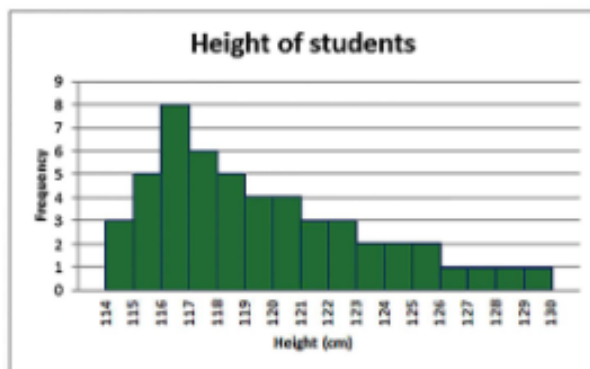
As mentioned in the slide, in a symmetrical distribution, the two sides of a distribution are mirror images of each other. Example for a truly symmetrical distribution is a normal distribution. When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.

In an asymmetrical distribution, the two sides will not be mirror images of each other. A skewed distribution can be positively or negatively skewed. We'll see about skewed distributions in the next slide.

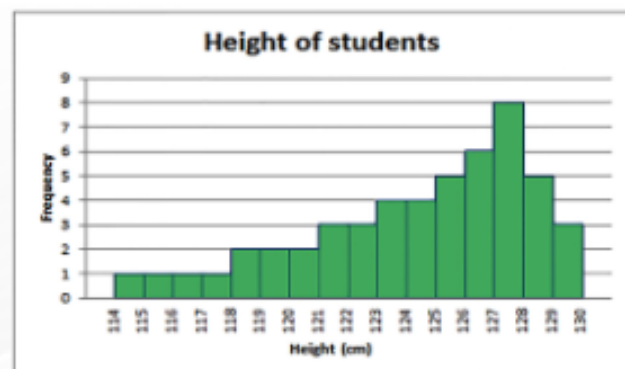
## 5.2 Skewness

### Skewness:

- The amount and direction of skew, i.e. deviation from the horizontal symmetry.
- Two types of skewed distribution:



Positively skewed distribution



Negatively skewed distribution

Skewness refers to the tendency of the values in a data set to be clustered around the high or low ends of the axis. By constructing a histogram and looking at the shape of the distribution, skewness can be identified.

A distribution is said to be positively skewed when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).

A distribution is said to be negatively skewed when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values), with increasingly lesser values on the left side of the x-axis (i.e. the smaller values).

Key features of the skewed distribution:

- asymmetrical shape
- mean and median have different values and do not lie at the centre of the curve
- there can be more than one mode
- the distribution of the data tends towards the high or low end of the dataset

## 5.3 Kurtosis

---

### Kurtosis:

- Measure of how tall and sharp the central peak is, relative to a standard bell curve.
- Also called as the measure of tailedness of a data distribution.
- Outliers impact the kurtosis of a data distribution.
- Formula to calculate kurtosis:

$$\frac{\sum_{i=1}^N \frac{(X_i - \bar{X})^2}{N}}{s^4}$$

Where  $X_i$  is the individual value,  $\bar{X}$  is the mean,  $N$  is the total number of values and  $s$  is the standard deviation

Kurtosis is defined as the measure of flatness or peakedness of the given distribution for the random variable along its tail. The outliers in the given data have more effect on this measure. Kurtosis does not have any unit.

Based on the value of kurtosis, the distribution can be classified into three categories.

- The distribution with kurtosis equal to 3 is known as mesokurtic. A random variable which follows normal distribution has kurtosis 3.
- If the kurtosis is less than three, the distribution is called as platykurtic. Here, the distribution has shorter and thinner tails than the normal distribution. Moreover, the peak is lower and also broader when compared to the normal distribution.
- If the kurtosis is greater than three, the distribution is called as leptokurtic. Here, the distribution has longer and fatter tails than the normal distribution. Moreover, the peak is higher and also sharper when compared to the normal distribution.



## What did You Grasp?



1. Which of the following options is/are true about a skewed distribution?
  - A) Skewed distribution is a symmetrical distribution
  - B) Skewed distribution can have more than one mode.
  - C) A positively skewed distribution has its right tail longer than the left side
  - D) Data distribution tends towards the centre of the data set.

## In a nutshell, we learnt:



1. An introduction to statistics along with its classification: descriptive and inferential statistics.
2. The four measures of scales: nominal, ordinal, interval and ratio scales.
3. Measures of location: mean, median and mode.
4. Measures of variability: range, quartiles and interquartile range, variance and standard deviation.
5. Measures of shape: skewness and kurtosis.

# 1 Introduction to Statistics

## 1.1 Introduction to Statistics

### Statistics:

- The branch of mathematics that deals with the collection, analysis, interpretation and presentation of masses of numerical data.
- The discipline that utilizes data samples to support claims about populations (a larger data set).
- Utilizes models, representations and synopses for a set of experimental data and applies the results to a larger population.



Think about the following questions:

- What is the population of India?
- What is the literacy rate in India?
- Which state in India excels in healthcare?
- Which state in India has the largest number of automobile manufacturing units?
- Which Indian state recorded most number of accidents in a given year?

We come across a lot of questions like this on a day to day basis. There exist answers to all these questions. And all this is statistics, but it is not limited to data like this. In most of the cases like this, data from the entire population is not necessary. A portion of the population, that is, a sample is enough to give insights about the population. Imagine you want to find the average salary of an entry-level software developer in India. Will you go and meet each one of them in person to get the data? You will send a survey to a set of people and analyse it to get an average number. You will then apply it to the entire population. So, what is statistics?

As given in the slide, "Statistics is that branch of mathematics, which deals with the collection, analysis, interpretation and presentation of masses of numerical data." Statistics takes data from samples, perform interpretations and applies the results to populations, i.e., a larger data set. The sample is thus a representative of a population. In a nutshell, statistics is a process that is used to characterize a data set. It involves the process of gathering and evaluating data and summarizing the data into a mathematical form.

Statistics has applications in almost every discipline like education, healthcare, manufacturing, government sectors IT, Physical and Social sciences, Psychology, Humanities and so forth.