Q1. What is Statistics?

Any raw Data, when collected and organized in the form of numerical or tables, is known as Statistics. Statistics is also the mathematical study of the probability of events occurring based on known quantitative Data or a Collection of Data.

Statistics attempts to infer the properties of a large Collection of Data from inspection of a sample of the Collection thereby allowing educated guesses to be made with a minimum of expense. There are generally 3 kinds of averages commonly used in Statistics. They are: (i) Mean, (ii) Median, and (iii) Mode.

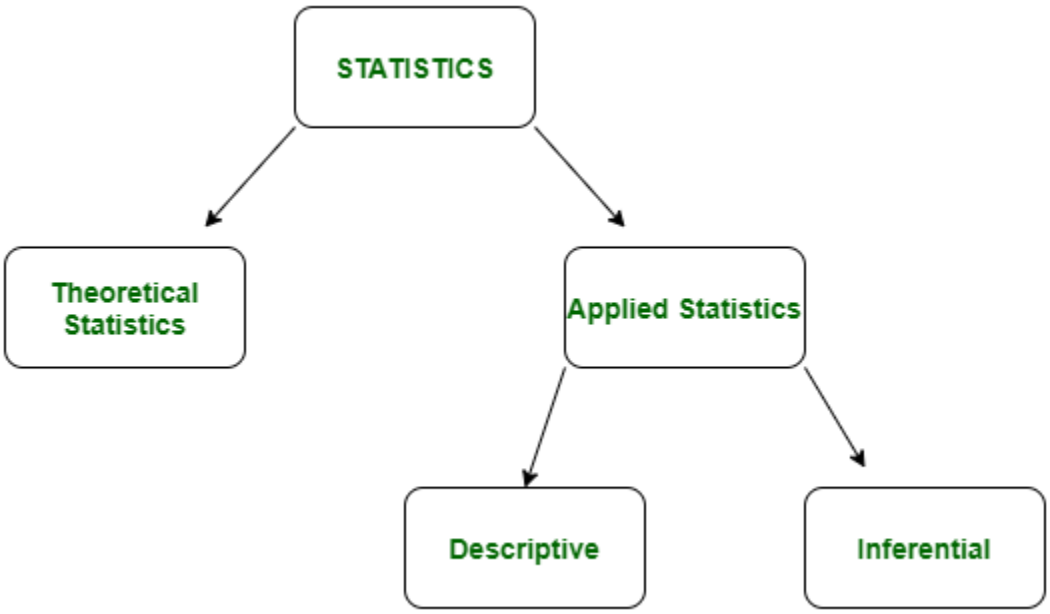
Statistics is the study of Data Collection, Analysis, Interpretation, Presentation, and organizing in a specific way. Mathematical methods used for different analytics include mathematical Analysis, linear algebra, stochastic Analysis, the theory of measure-theoretical probability, and differential equations. Collecting, classifying, organizing, and displaying numerical Data is associated with Statistics. This helps one to grasp different outcomes from it and foresee several possibilities of various events. Statistics discuss information, observations, and Data in the form of numerical Data. We can find different indicators of central tendencies and the divergence of various values from the centre with the help of Statistics.

The ability to analyse and interpret statistical Data is a vital skill for researchers and professionals from a wide variety of disciplines. You may need to make decisions based on statistical Data, interpret statistical Data in research papers, do your own research, and interpret the Data.

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data. It is basically a collection of quantitative data.

Types of Statistics:

- Theoretical Statistics
- Applied Statistics



Q2. Define the different types of statistics and give an example of when each type might be used.

There are two kinds of Statistics, which are divided into the following two categories.

Descriptive Statistics

In descriptive Statistics, the Data or Collection Data are described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

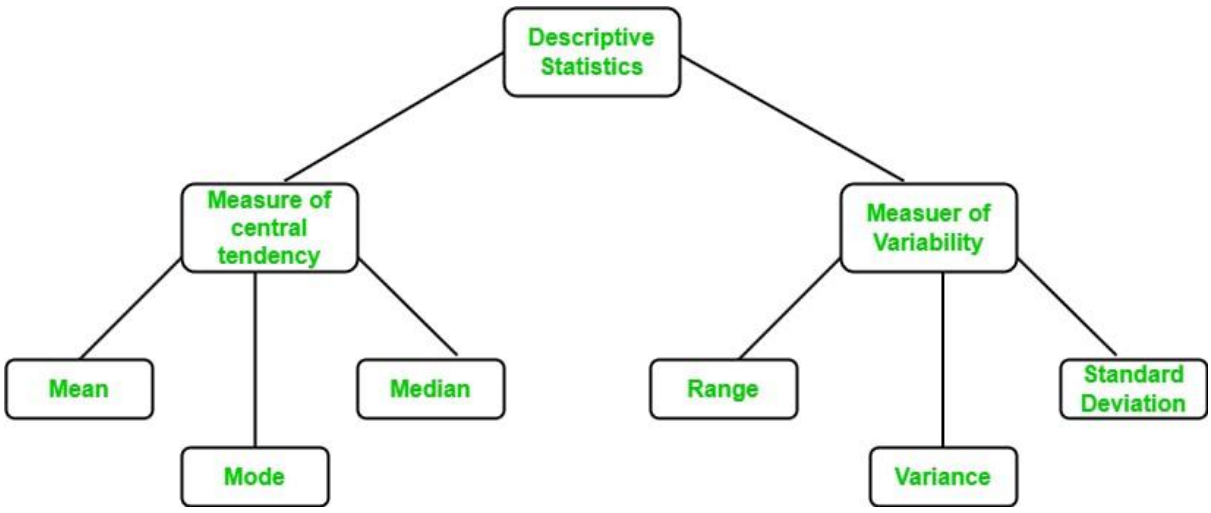
Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars, or scatter plots.

Descriptive Statistics are just descriptive and thus do not require normalization beyond the Data collected.

Descriptive statistics is a term given to the analysis of data that helps to describe, show, and summarize data in a meaningful way. It is a simple way to describe our data. Descriptive statistics is very important to present our raw data ineffective/meaningful way using numerical calculations or graphs or tables. This type of statistics is applied to already known data.

Types of Descriptive Statistics:

- Measure of Central Tendency
- Measure of Variability




Inferential Statistics

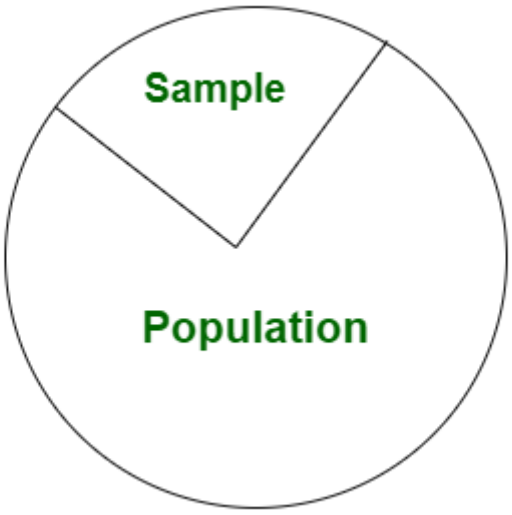
In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analysed, and summarised we use Inferential Statistics to describe the Meaning of the collected Data.

Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.

Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.

Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

2. Inferential Statistics: In inferential statistics, predictions are made by taking any group of data in which you are interested. It can be defined as a random sample of data taken from a population to describe and make inferences about the population. Any group of data that includes all the data you are interested in is known as population. It basically allows you to make predictions by taking a small sample instead of working on the whole population.




Example

In a class, the Data is the set of marks obtained by 50 students. Now when we take out the Data average, the result is the average of 50 students' marks. If the average marks obtained by 50 students are 88 out of 100, based on the outcome, we will draw a conclusion.

Mean, Median and Mode in Statistics

Mean: Mean is considered the arithmetic average of a Data set that is found by adding the numbers in a set and dividing by the number of observations in the Data set.

Median: The middle number in the Data set while listed in either ascending or descending order is the Median.

Mode: The number that occurs the most in a Data set and ranges between the highest and lowest value is the Mode.

For n number of observations, we have

$$\text{Mean} = \overline{x} = \frac{\sum x}{n} = \sum xn$$

$$\text{Median} = \frac{\left[\frac{n}{2} + 1\right]^{th} term}{2} \text{ if n is odd.}$$

$$\text{Median} = \frac{\left[\frac{n}{2}\right]^{th} term + \left[\frac{n}{2} + 1\right]^{th} term}{2} \text{ if n is even.}$$

Mode = The value which occurs most frequently

Measures of Dispersion in Statistics

The measures of central tendency do not suffice to describe the complete information about a given Data. Therefore, the variability is described by a value called the measure of dispersion.

The different measures of dispersion include:

1. The range in Statistics is calculated as the difference between the maximum value and the minimum value of the Data points.
2. The quartile deviation that measures the absolute measure of dispersion. The Data points are divided into 3 quarters. Find the Median of the Data points. The Median of the Data points to the left of this Median is said to be the upper quartile and the Median of the Data points to the right of this Median is said to be the lower quartile. Upper quartile - lower quartile is the interquartile range. Half of this is the quartile deviation.
3. The Mean deviation is the statistical measure to determine the average of the absolute difference between the items in a distribution and the Mean or Median of that series.
4. The standard deviation is the measure of the amount of variation of a set of values.

Q3. What are the different types of data and how do they differ from each other? Provide an example of each type of data?

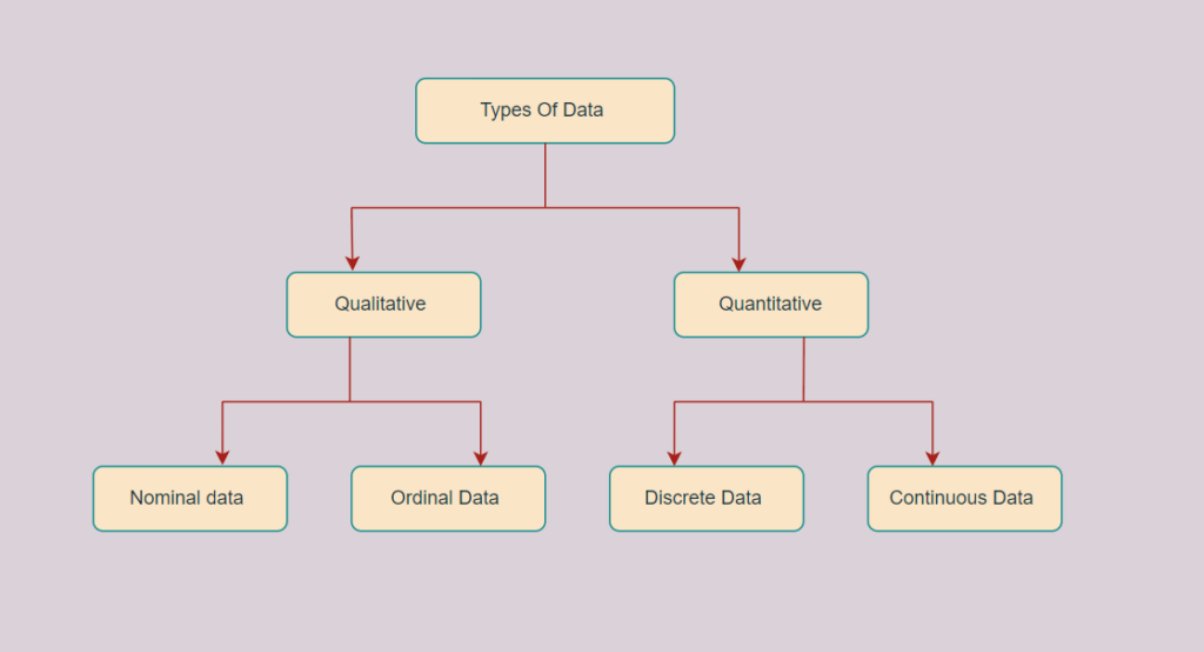
There are two types of data:

1. Qualitative Data.
2. Quantitative Data.

These data are further classified into four categories:

- Nominal data.
- Ordinal data.
- Discrete data.
- Continuous data.

Now business runs on data, and most companies use data for their insights to create and launch campaigns, design strategies, launch products and services or try out different things. According to a report, today, at least 2.5 quintillion bytes of data are produced per day.




**Qualitative or Categorical Data**

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.

## The other examples of qualitative data are:

- What language do you speak?
- Favourite holiday destination
- Opinion on something (agree, disagree, or neutral)
- Colours

## The Qualitative data are further classified into two parts:

## Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour.

The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

## Examples of Nominal Data:

- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

## Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered "in-between" qualitative and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

## Examples of Ordinal Data:

- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

### Quantitative Data

Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like "how much," "how many," and "how often." For example, the price of a phone, the computer's ram, the height or weight of a person, etc., falls under quantitative data.

Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.

### Examples of Quantitative Data:

- Height or weight of a person or object
- Room Temperature
- Scores and Marks (Ex: 59, 80, 60, etc.)
- Time

### The Quantitative data are further classified into two parts:

## Discrete Data

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

## Examples of Discrete Data :

- Total numbers of students present in a class.
- Cost of a cell phone
- Numbers of employees in a company
- The total number of players who participated in a competition.
- Days in a week

## Continuous Data

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

## Examples of Continuous Data:

- Height of a person
- Speed of a vehicle
- "Time-taken" to finish the work.
- Wi-Fi Frequency
- Market share price

## Difference between Discrete and Continuous Data

| Discrete Data | Continuous Data |
|---|---|
| Discrete data are countable and finite; they are whole numbers or integers | Continuous data are measurable; they are in the form of fractions or decimal |

| Discrete Data | Continuous Data |
|---|---|
| Discrete data are represented mainly by bar graphs | Continuous data are represented in the form of a histogram |
| The values cannot be divided into subdivisions into smaller pieces | The values can be divided into subdivisions into smaller pieces |
| Discrete data have spaces between the values | Continuous data are in the form of a continuous sequence |
| **Examples:** Total students in a class, number of days in a week, size of a shoe, etc | **Example:** Temperature of room, the weight of a person, length of an object, etc |

Q4. Categorise the following datasets with respect to quantitative and qualitative data types:

(i) Grading in exam: A+, A, B+, B, C+, C, D, E: Ordinal
(ii) Colour of mangoes: yellow, green, orange, red: Nominal
(iii) Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,] - Continuous
(iv) Number of mangoes exported by a farm: [500, 600, 478, 672, …]- Discrete.

Q5. Explain the concept of levels of measurement and give an example of a variable for each level.

**Levels of measurement,** also called scales of measurement. In scientific research, a variable is anything that can take on different values across your data set (e.g., height or test scores).

There are 4 levels of measurement:

- Nominal: the data can only be categorized.
- Ordinal: the data can be categorized and ranked.
- Interval: the data can be categorized, ranked, and evenly spaced.
- Ratio: the data can be categorized, ranked, evenly spaced, and has a natural zero.

Depending on the level of measurement of the variable, what you can do to analyse your data may be limited. There is a hierarchy in the complexity and precision of the level of measurement, from low (nominal) to high (ratio).

**Nominal, ordinal, interval, and ratio data**

Going from lowest to highest, the 4 levels of measurement are cumulative. This means that they each take on the properties of lower levels and add new properties.

| Nominal level | Examples of nominal scales |
|---|---|

| | |
|---|---|
| You can categorize your data by labelling them in mutually exclusive groups, but there is no order between the categories. | • City of birth<br>• Gender<br>• Ethnicity<br>• Car brands<br>• Marital status |
| **Ordinal level** | **Examples of ordinal scales** |
| You can categorize and rank your data in an order, but you cannot say anything about the intervals between the rankings.<br>Although you can rank the top 5 Olympic medallists, this scale does not tell you how close or far apart they are in number of wins. | • Top 5 Olympic medallists<br>• Language ability (e.g., beginner, intermediate, fluent)<br>• Likert-type questions (e.g., very dissatisfied to very satisfied) |
| **Interval level** | **Examples of interval scales** |
| You can categorize, rank, and infer equal intervals between neighbouring data points, but there is no true zero point.<br>The difference between any two adjacent temperatures is the same: one degree. But zero degrees is defined differently depending on the scale – it doesn't mean an absolute absence of temperature.<br>The same is true for test scores and personality inventories. A zero on a test is arbitrary; it does not mean that the test-taker has an absolute lack of the trait being measured. | • Test scores (e.g., IQ or exams)<br>• Personality inventories<br>• Temperature in Fahrenheit or Celsius |
| **Ratio level** | **Examples of ratio scales** |
| You can categorize, rank, and infer equal intervals between neighbouring data points, and there is a true zero point.<br>A true zero means there is an absence of the variable of interest. In ratio scales, zero does mean an absolute lack of the variable.<br><br>For example, in the Kelvin temperature scale, there are no negative degrees of temperature – zero means an absolute lack of thermal energy. | • Height<br>• Age<br>• Weight<br>• Temperature in Kelvin |

Q6. Why is it important to understand the level of measurement when analyzing data? Provide an example to illustrate your answer.

The level at which you measure a variable determines how you can analyse your data. The different levels limit which descriptive statistics you can use to get an overall summary of your data, and which type of inferential statistics you can perform on your data to support or refute your hypothesis.

In many cases, your variables can be measured at different levels, so you must choose the level of measurement you will use before data collection begins.

Example of a variable at 2 levels of measurement

You can measure the variable of income at an ordinal or ratio level.

Ordinal level: You create brackets of income ranges: $0–$19,999, $20,000–$39,999, and $40,000–$59,999. You ask participants to select the bracket that represents their annual income. The brackets are coded with numbers from 1–3.

Ratio level: You collect data on the exact annual incomes of your participants.

| Participant | Income (ordinal level) | Income (ratio level) |
|---|---|---|
| A | Bracket 1 | $12,550 |
| B | Bracket 2 | $39,700 |
| C | Bracket 3 | $40,300 |

At a ratio level, you can see that the difference between A and B's incomes is far greater than the difference between B and C's incomes.

At an ordinal level, however, you only know the income bracket for each participant, not their exact income. Since you cannot say exactly how much each income differs from the others in your data set, you can only order the income levels and group the participants.

Q7. How nominal data type is different from ordinal data type.

## Difference between Nominal and Ordinal Data

| **Nominal Data** | **Ordinal Data** |
|---|---|
| Nominal data can't be quantified, neither they have any intrinsic ordering | Ordinal data gives some kind of sequential order by their position on the scale |
| Nominal data is qualitative data or categorical data | Ordinal data is said to be "in-between" qualitative data and quantitative data |
| They don't provide any quantitative value, neither can we perform any arithmetical operation | They provide sequence and can assign numbers to ordinal data but cannot perform the arithmetical operation |
| Nominal data cannot be used to compare with one another | Ordinal data can help to compare one item with another by ranking or ordering |
| **Examples:** Eye color, housing style, gender, hair color, religion, marital status, ethnicity, etc | **Examples:** Economic status, customer satisfaction, education level, letter grades, etc |

Q8. Which type of plot can be used to display data in terms of range?

Histogram. If the groups depicted in a bar chart are continuous numeric ranges, we can push the bars together to generate a histogram. Bar lengths in histograms typically correspond to counts of data points, and their patterns demonstrate the distribution of variables in your data.

Q9. Describe the difference between descriptive and inferential statistics. Give an example of each type of statistics and explain how they are used.

## Difference between Descriptive and Inferential statistics:

| S.No. | Descriptive Statistics | Inferential Statistics |
|---|---|---|
| 1. | It gives information about raw data which describes the data in some manner. | It makes inferences about the population using data drawn from the population. |
| 2. | It helps in organizing, analyzing, and to present data in a meaningful manner. | It allows us to compare data, and make hypotheses and predictions. |
| 3. | It is used to describe a situation. | It is used to explain the chance of occurrence of an event. |
| 4. | It explains already known data and is limited to a sample or population having a small size. | It attempts to reach the conclusion about the population. |
| 5. | It can be achieved with the help of charts, graphs, tables, etc. | It can be achieved by probability. |

Q10. What are some common measures of central tendency and variability used in statistics? Explain how each measure can be used to describe a dataset.

**Measures of Central Tendency and Variability:**

**Central tendency**

**Definition**: the tendency of quantitative data to cluster around some central value. The closeness with which the values surround the central value is commonly quantified using the standard deviation. They are classified as summary statistics.

**Measures of Central Tendency**

**Mean**: The sum of all measurements divided by the number of observations. Can be used with discrete and continuous data. It is value that is most common.

**Median:** The middle value that separates the higher half from the lower half. Mean and median can be compared with each other to determine if the population is of normal distribution or not. Numbers are arranged in either ascending or descending order. The middle number is than taken.

**Mode:** The most frequent value. It shows most popular option and is the highest bar in histogram. Example of use: to determine the most common blood group.

**Geometric mean** - the nth root of the product of the data values.

**Harmonic mean** - the reciprocal of the arithmetic mean of the reciprocals of the data values.

**Weighted mean** - an arithmetic mean that makes use of weighting to certain data elements.

**Truncated mean** - the arithmetic mean of data values that do not include the whole set of values, such as ignoring values after a certain number or discarding a fixed proportion of the highest and lower values.

**Midrange** - the arithmetic mean of the maximum and minimum values of a data set

**Variability (dispersion)**

**Definition:** dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions. It is the variability or spread in a variable or a probability distribution i.e. They tell us how much observations in a data set vary. They allow us to summarise our data set with a single value hence giving a more accurate picture of our data set.

**Measures of variability**

**Variance:** A measure of how far a set of numbers is spread out from each other. It describes how far the numbers lie from the mean (expected value). It is the square of standard deviation.

**Standard deviation (SD): it** is only used for data that are “normally distributed”. SD indicates how much a set of values is spread around the average. SD is determined by the variance (SD=the root of the variance).

**Interquartile range (IQR):** the interquartile range (IQR), is also known as the 'midstream' or 'middle fifty', is a measure of statistical dispersion, being equal to the difference between the third and first quartiles. IQR = Q3 − Q1. Unlike (total) range, the interquartile range is a more commonly used statistic, since it excludes the lower 25% and upper 25%, therefore reflecting more accurately valid values and excluding the outliers.

**Range:** it is the length of the smallest interval which contains all the data and is calculated by subtracting the smallest observation (sample minimum) from the greatest (sample maximum) and provides an indication of statistical dispersion. It bears the same units as the data used for calculating it. Because of its dependence on just two observations, it tends to be a poor and weak measure of dispersion, with the only exception being when the sample size is large.