what is STATISTICS ?

1. statistics is the science of using sample data to understand the population in the context of uncertainity
2. It uses quantified models and representations for a given set of experimental data
3. It deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.

`OR`

Statistics are numbers that summarize raw facts and figures in some meaningful way..

Why STATISTICS in DATA SCIENCE ?

1. Understanding the fundamentals of statistics is a core capability for becoming a Data Scientist.
2. Statistics helps you make sense of unclear or imbalanced data into simple sets
3. These can be used in Machine Learning Algorithms for clean and unbiased prediction of data and eliminate the duplicate values

Some basic concepts in statistics

POPULATION

- The whole data set or world is called as a population

SAMPLE

- sample is the smaller form of the data set/population
- a pile of samples makes a population

SKEWNESS

Skewness means the supressed graph{gives the shape of distribution}

- If it is **right skewed** then mean is greater than median **{mean > median}**
- If it is **left skewed** then mean is lesser than median **{mean < median}**

Types of STATISTICS

statistics can be divided into 2 types

- `Descriptive statistics`

- `Inferential staistics`

Descriptive statistics

Descriptive Statistics can be useful for two reasons:

- To provide basic information about variables in a dataset
- To highlight potential relationship among variables

The Descriptive Statistics can be measured in two ways:

1. Measure of central Tendency
2. Measure of variability

MEASURE OF CENTRAL TENDENCY ¶

- In Statistics, the centarl tendency can be said as **central or typical value of distributon**
- It identifies the different central points in the data
- These are often referred colloquially as **Average**

Central tendency can be classified as :

- `MEAN`
- `MEDIAN`
- `MODE`

MEAN

1. It is the arthimetic average of data values
2. It is highly Susceptible to outliers
3. mean can never be larger or smaller than maximum (or) minimum values..., but..., can be either maximum (or) minimum value.

mean is represented by **μ**

The mathematical formula for mean is : **Mean = {Sum of Observation} ÷ {Total numbers of Observations}.**

- The only problem with mean in distribution is **MEAN always moves behind the OUTLIERS** i.e., if an outlier is present in a data set the mean value gets dragged with the outler value. [or] we can also say that only one outlier is enough to change the data,it attracts mean.

- If Outlier is Bigger, then the mean will also increase(mean is also bigger)
- If Outlier is Smaller, then the mean will also decrease(mean is also smaller)

MEDIAN

In an ordered Array, the median is the middle number

It is very important to set the data in an ordered format to calculate the median in given observation

The mathematical formula for median can be written as : **{(n + 1) ÷ 2}th position of observation**

- The Median is not dependent on outliers.It is totally based on number of observations in data set

`OR`

- The Median values are based on total distribuution

MODE

- Mode is the most occuring data point in the distribution
- There can alsobe more number of modes in a distribution.