TOPICS COVERED HERE:

- **Defnition of ststistics**
- **Why stats in data science**
- **Population and Sample**
- **Types of Statistics**
- **Central Tendency {Mean, Median, Mode}**
- **Skewness,Kurtosis**
- **Standard Deviation,Mean Deviation, Variance**
- **Range, Percentile, Quartiles**
- **Tables**
- **Graphs**

What is STATISTICS ?

- statistics is the science of using sample data to understand the population in the context of uncertainity
- It uses quantified models and representations for a given set of experimental data
- 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 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

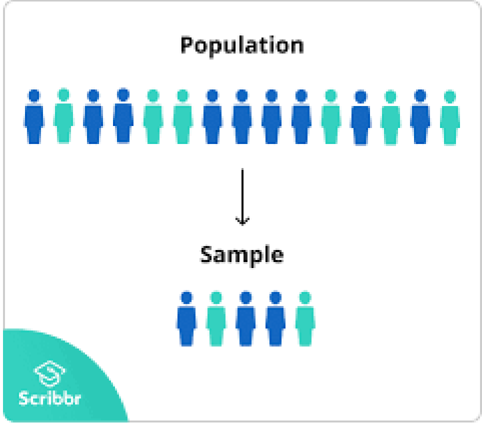
Some basic concepts in statistics

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



Types of STATISTICS

statistics can be divided into 2 types

- `Descriptive statistics`
- `Inferential staistics`

Descriptive statistics

Descriptive Analysis is the type of analysis of data that helps describe, show or summarize data points in a constructive way

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 3 ways:

1. **Statistical Measures**
2. **Tables**
3. **Graphs**

STATISTICAL MEASURE

Statistical Measures can be again classified into 3 types

1. **Central Tendency**
2. **Shape**
3. **Spread**

1. 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 = ODD

{ [{n/2} + ({n/2}+1)] ÷ 2 }th position of observation = EVEN

- 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 distribution

MODE

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

2.Shape

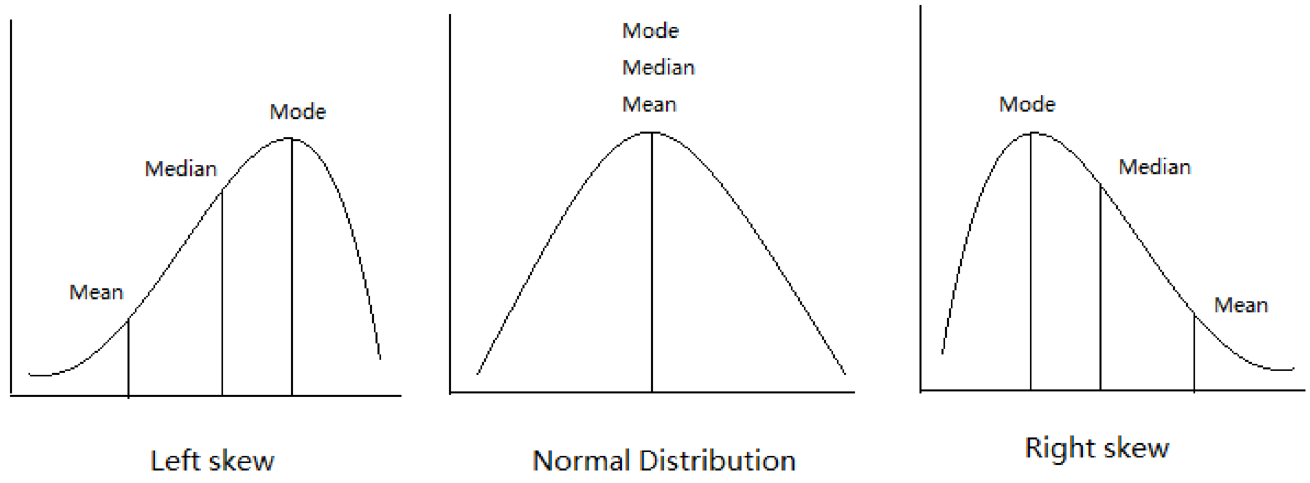
shapes can again be classified into two sub types

- **Skewness**
- **Kurtosis**

Skewness

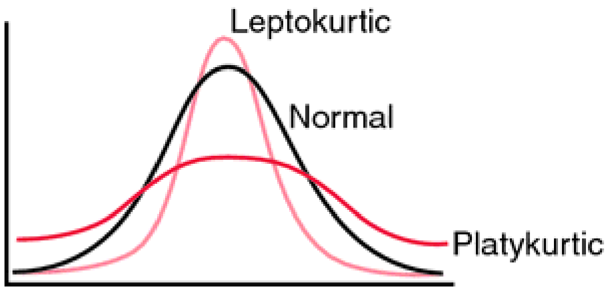
Skewness means the supressed graph{gives the shape of distribution}

- 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}**
- If the distribution is correct then **{Mean = Median = Mode}**



Kurtosis

kurtosis means the spread of tails in a graph



3.Spread

Spread defines the spread of data between minimum and maximum values in a plot

The spread can be subdivide again into:

- **standard deviation**
- **mean deviation**
- **variance**
- **range**
- **percentile**
- **quartiles**

Standard Deviation

Standard Deviation is the measurement of the average distance between each quantity and mean.i.e.,how the data is spread out from the mean.

Low S.D indicates the data points tend to be close to 'Mean' , where as high S.D shows that the data points are spread out over a wide range.

The formula for Standard Deviation can be given as:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ : Population standard deviation
x : Datapoint value
μ : Population mean
N : Population size

Mean Deviation

It is an average of absolute differences between each value in a set of values, and the average of all values of that set.

$$\textbf{M.D.} = \frac{1}{n}\sum_{i=0}^{n}|x_i - \bar{x}|$$

Variance

Variance is a square of average distance between each quantity and mean. That is it is square of standard deviation.

[OR]

It could also be said as spread of the mean.

variance can be defined as the square of standard deviation

$$\textbf{Variance} = (S.D.)^2$$

Range

It is the difference between the lowest and highest value.

Percentile

Percentile is a way to represent position of a values in data set. To calculate percentile, values in data set should always be in ascending order.

Quartiles

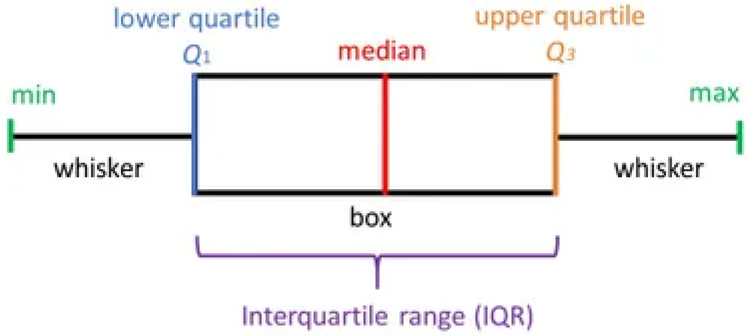
Quartiles of a set of data is a similar processto find the median

There are mainly Q1,Q3 Ranges in box plot which define the minimum and maximum values in the range. Here the 50% of total distribution lies in between Q1 and Q3.

Here the first 25% of the total data lies between "minimum" to Q1 and the last 25% of total data lies between Q3 to "maximum"

If any value ranging after Q3 and before Q1 can be termed as **Outliers**

Here **Q1** is said as **Lower Quartile Range** and **Q3** is said as **Upper Quartile Range** and **Median** as **IQR**



In general IQR can be considered as Range

IQR= Q3-Q1

TABLES

In general these are not much used. These show the **Tabular format** of the data

GRAPHS

Graphs show the **pictorial representation** of the data like **Pie chart,Histogarm,etc..** to show the spread of data or to display the relationships between data in categories.