#### Basic Statistics Concepts

A form of study that allows us to take data and define or summarize it using numerous techniques.

#### Basic terminologies

#### Population

It can said as the total number of individual humans, other organisms, or any other object that makes up a whole.

#### Parameter

Numerous mathematical calculations can be performed on the population such as finding the most common item or value occurring in the population or finding the average etc

#### Sample

A sample is a subset of the population (that ideally represents the population)

#### Categories in Statistics

Categorizations can help in getting a better insight into the statistics

* **Descriptive Statistics**

This form of statistics deals with the population as well as the sample. As the name suggests, descriptive statistics are used to describe the features and characteristics of the data.

* **Inferential Statistics**

This form of statistics is not only describing the data or stating the facts about it but draw some conclusions about the population based on the sample at hand.

#### Probability Distributions

In statistics, when we collect data, we can assess how it is spread or dispersed or simply distributed. This collection of values or distribution can be represented or visualized using graphs such as histograms

#### Univariate, Bivariate and Multivariate Statistics

Statistics can be performed using a single variable, two variables, and even multiple variables. When we use a single variable to perform statistics, where most of the descriptive statistics lie, then such statistics is known as univariate statistics. When we use two variables which is generally the case with inferential statistics where we are trying to assess the relationship between the two samples then this kind of statistic is commonly called a Bivariate Statistics. Finally when we have multiple variables where we simultaneously try to assess the relationship using multiple variables then this is known as multivariate statistics. For example, linear regression is a form of statistics where we try to assess the relationship of multiple independent features with a single dependent feature.

#### Central Limit Theorem

In a central limit theorem, the mean of the smaller sample is the same as that of the mean of the larger population. Therefore, the resulting standard deviation is equal to the standard deviation of the population. An important concept of the Central Limit Theorem is the estimation of the population mean. Margin error can be calculated by multiplying the standard error of the mean with the z-score of the percentage of confidence level.

## ****Measures in Statistics****

To statistically describe the data, we have to measure the data and there are four such common measures in statistics.

#### Measure of Frequency

Here we count how many times a value appears in the data. This is the most common and simplest measure

#### Measure of Central Tendency

The typical way in which we summarize the data is by finding its central point. This central point is calculated in numerous ways such as Mean, Median, and Mode which form the measure of the central tendency of the data.

#### Mean

It is the simple arithmetic average of the data

#### Median

It is that value that marks the 50th percentile i.e. the central value or middle value when we arrange the data in an ascending/descending order

#### Mode

It the most commonly occurring value or the value with the highest frequency

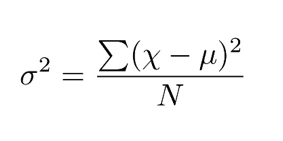
#### Measure of Dispersion

Consistence can be found by analyzing the dispersion of the values and this is where the measure of dispersion helps. The most common measures of dispersion include-

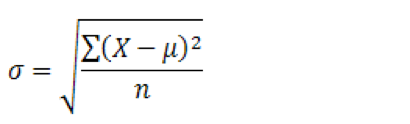
**Range**  
Finding the difference between the maximum and minimum value. This can be problematic as outliers can affect it adversely.

**Interquartile Range**  
Finding the difference between percentile 75th and 25th or Q3 and Q1. Here as we don’t use all the values, this method cannot be considered as a very reliable method for calculating the dispersion.

**Variance**  
Variance is calculated by finding the difference between the value and the mean aka deviation and then taking the square of it (as without squaring, the sum of deviation is always 0). We then sum all these squared deviations and divide it by the count of values.



**Standard Deviation**  
The problem with variance is that is changes the unit if the value. If the values are in km then the variance will be in km2 and this is something not very feasible. To correct this change, we take the square root of it and this makes the result to be called Standard Deviation.



**Measure of Shape**

Distribution of data can be found by plotting a histogram. This helps in revealing the shape of the distribution of the data. By analyzing the shape we can quickly gain good insights regarding the data and the kind of distribution it has. The shape can be divided into two parts- Symmetrical and Asymmetrical

**Symmetrical**  
This is where if we divide the data from the middle of the distribution then the remaining left side is the mirror image of the right side. The most common symmetrical distribution is Gaussian distribution where the mean is the same as the median which is the same as the mode. This allows for a symmetrical bell-shaped curve.

**Asymmetrical**  
When the distribution is skewed i.e. the distribution is tilted on either of the sides then the distribution is known as Asymmetrical. Such asymmetrical distribution can be found when the data is skewed. If the data is left-skewed (negative skewed) mode is greater than the mean and for right-skewed (positive skewed), the opposite is true.

