Descriptive Statistics

Statistics:

Statistics is a branch of mathematics taking and transforming numbers into useful information for decision makers. The word statistics also refers to numbers that are used to describe data or relationships.

It is way to get information from data. It deals with data collection, organization, analysis, interpretation and presentation.

Why statistics: Knowledge of statistics helps us understand and make decision better based on data given.

Statistics helps in the study of many other fields, such as science, medicine, economics, psychology, politics and marketing.

Statistics is classified as

- **Descriptive statistics**
- **Inferential statistics**

Descriptive Statistics:

Numbers that are used to summarize and describe data are called descriptive statistics. Presenting, organizing and summarizing data.

Inferential Statistics:

Once the results have been summarized and described, it can be used for prediction. Numbers that make predictions about that you can't see are called inferential statistics.

Descriptive Statistics:

Types of Descriptive statistics:

- **Measure of Central Tendency**
- **Measure of Spread/Dispersion**

Measure of Central Tendency:

Mean: Sum of all the values and divide it by the number of values. The problem with the mean is it doesn't tell anything about how the value is distributed. The value that are very large or very small change a mean a lot.

Mean = Sum of x / Number of values

Median: Middle item of the data. Sort the data in ascending order and choose the number in the middle. If there is an even number of data, choose the two middle ones and calculate their mean.

Median = (n+1)/2

Mode: The **Mode** is the most frequent item of data.

Measure of Spread/Dispersion:

Standard deviation:

Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

$$\text{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x - \mu)^2}$$

Variance :

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

Variance = Square of standard deviation

Range:

Range is the difference between lowest and highest value.

Percentile:

It shows the value below which a given percentage of observations falls

Nth percentile states that there are at least N% of values less than or equal to this value and (100-N) values are greater or equal to this value.

I = (N/100)*n

N – The Percentile you are interested in

N = number of values.

- If I decimal, round off the next value. If it is integer then average of i and i+1.

For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

The 35th percentile, for instance, is the value of all observations below 35% of the datapool.

Quartile:

The 25th, 50th and 75th percentiles are called **quartiles**. The 50th percentile is also called the median**.**

Inter Quartile Range: - Q3 – Q1

Coefficient of Variation:-

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean.

Coff of variation = (Std.dev/Mean) * 100

Covariance:

It is the relationship between a pair of random variables where change in one variable causes change in another variable.

Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). If the correlation value is 0 then it means there is no Linear Relationship between variables

In the study of covariance only sign matters. A positive value shows that both variables vary in the same direction and negative value shows that they vary in the opposite direction.

Covariance between two variables x and y can be calculated as follows:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Central Limit Theorem:

If we calculate the mean of a sample, it will be an estimate of the mean of the population distribution. But, like any estimate, it will be wrong and will contain some error. If we draw multiple independent samples, and calculate their means, the distribution of those means will form a Gaussian distribution.

- Standard deviation of sample mean = population standard deviation/ square root(n)
- Mean of sample means distribution = population mean.