**Descriptive Statistics**

First of all, we should go over what statistics really is:

***Statistics is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.***

Within statistics, there are two main categories:

**1. Descriptive Statistics:**In Descriptive Statistics you are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables.

**2. Inferential statistics:**Inferential Statistics are produced by more complex mathematical calculations, and allow us to infer trends and make assumptions and predictions about a population based on a study of a sample taken from it.

**Descriptive Statistics**

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. Measures of Central Tendency
2. Measures of Dispersion (or Variability)

**Measures of Central Tendency**

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. This one number summary is of three types.

1. **Mean :** Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus, mean is a number around which the entire data set is spread.
2. **Median :** Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

* If the number of observations are odd, median is given by the middle observation in the sorted form.
* If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.

An important point to note that the order of the data (ascending or descending) does not effect the median.

**3.** **Mode :**Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.

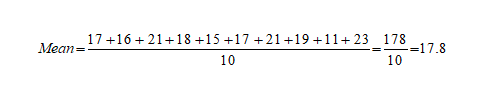
* If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.
* If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.
* If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal**.

*Example to compute the Measures of Central Tendency*

Consider the following data points.

**17, 16, 21, 18, 15, 17, 21, 19, 11, 23**

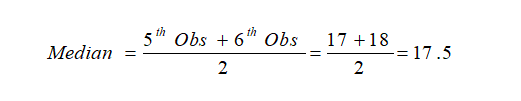
* Mean — Mean is calculated as



* Median — To calculate Median, lets arrange the data in ascending order.

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

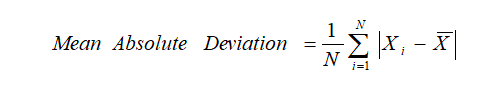


* Mode — Mode is given by the number that occurs maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

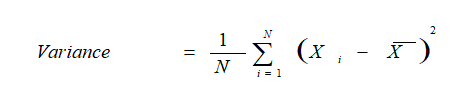
Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

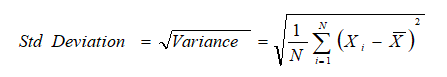
1. **Absolute Deviation from Mean** — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is calculated as



1. **Variance** — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as



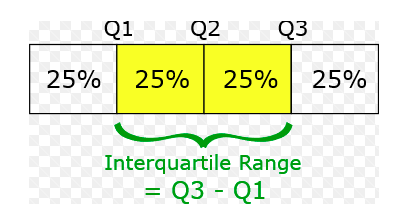
1. **Standard Deviation**— The square root of Variance is called the Standard Deviation. It is calculated as



1. **Range** — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as



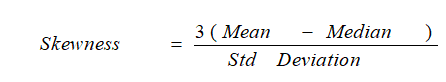
**5. Quartiles**— Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

* 25% of the data points lie below Q1 and 75% lie above it.
* 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
* 75% of the data points lie below Q3 and 25% lie above it.
* 

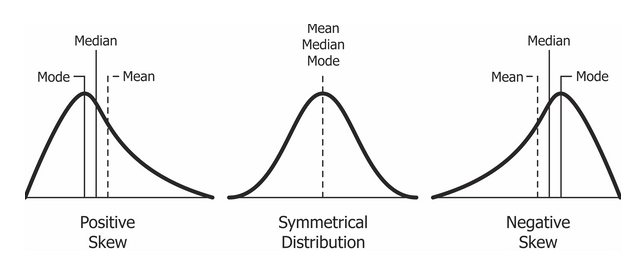
**6. Skewness**— The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

* Positive Skew — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.
* Negative Skew — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

The most commonly used method of calculating Skewness is



If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.



Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.