**Statistics**

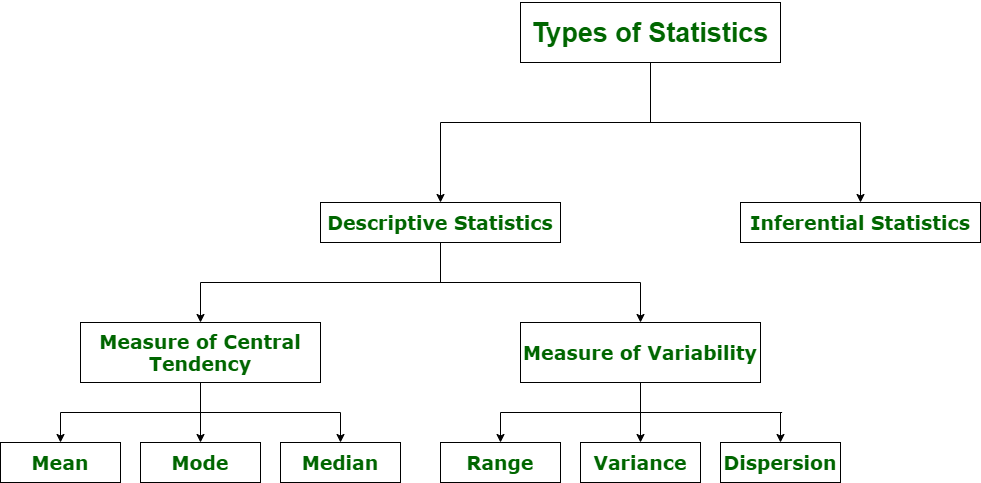
Statistics can be explained as the category of math which is used for summarizing, interpreting and analyzing the things we observe to bring meaning or make sense for things we observe. For instance, the family counselor might use statistics for describing the patient's behavior or the effect of his treatment.

In simple words, Statistics can be used to derive meaningful insights from data by performing mathematical computations on it.

**Importance of Statistics**

1. Statistics makes the work simple & provides a clear picture on the work we do on daily basis.
2. The statistical methods helps us to research on different streams such as medicine, economics, business, social science and so on.
3. Statistics provides us different types of organized data with the help of graphs, diagrams and charts.
4. Statistics comes handy while we do critical analysis.

**Types of Statistics :**



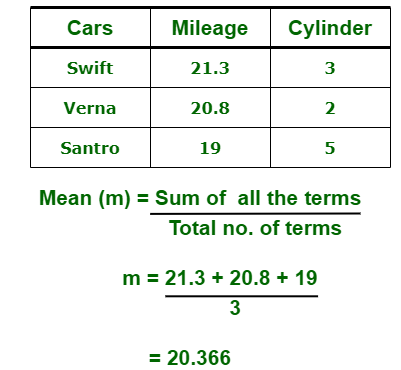
1. **Descriptive Statistics**

Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

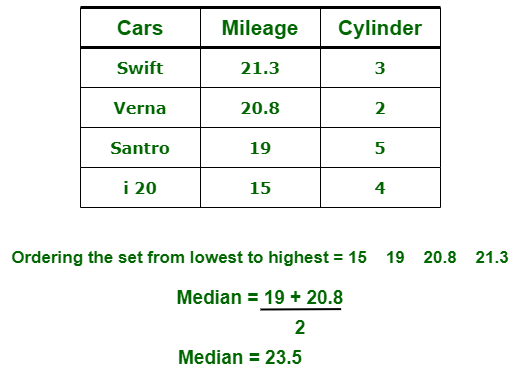
**(a). Measure of central tendency**

Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.  
In statistics, there are three common measures of central tendency as shown below:

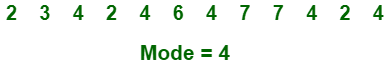
1. **Mean :**  
   It is measure of average of all value in a sample set.  
   For example,



1. **Median :**  
   It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.  
   For example,



1. **Mode :**  
   It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.  
   For example,



**(b). Measure of Variability or Measure of Dispersion.**  
Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below:

**(i) Range :**  
It is given measure of how to spread apart values in sample set or data set.

Range = Maximum value - Minimum value

**(ii) Variance :**  
It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

**S2=** ∑ni=1 [(xi - ͞x)2 ÷ n]

In these formula, **n** represent total data points, **͞x** represent mean of data points and **xi** represent individual data points.

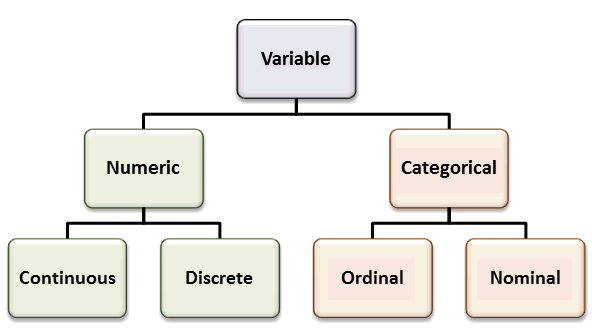
**(iii) Dispersion or Standard Deviation :**

It is a summary measure of the differences of each observation from the mean.

**σ=** √ (1÷n) ∑ni=1 (xi - μ)2

**2. Inferential Statistics :**  
Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion.

**Types Of Variables**



#### ****Categorical Variables:****

Both nominal and ordinal variables can be called categorical variables.

##### **1. Nominal Variable:**

A nominal variable is made up of various categories which has no order.

##### **Example:**

Gender of a patient may be Male or Female or State where they live in. Here each category differs from each other but there is no ranking order.

##### **2. Ordinal Variable:**

The second level of measurement is the ordinal level. There is not only a difference between the categories of a variable; there is also an order.  An example might be Highest paid, Average Paid and Lowest Paid employee.

#### ****Quantitative/ Numerical Variables:****

##### **1.  Continuous Variable:**

A variable is continuous if the possible values of the variable form an interval. An example is, again, the height of a patient. Someone can be 172 centimeters tall and 174 centimeters tall.  But also, for instance, 170.2461. We don’t have a set of separate numbers, but an infinite region of values.

##### **2. Discrete Variable:**

A variable is discrete if its possible categories form a set of separate numbers.

For the above breast cancer data **Uniformity of Cell Size: 1 – 10**is an example of discrete variable.

**Percentile**

The value below which a percentage of data falls

### Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you

That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

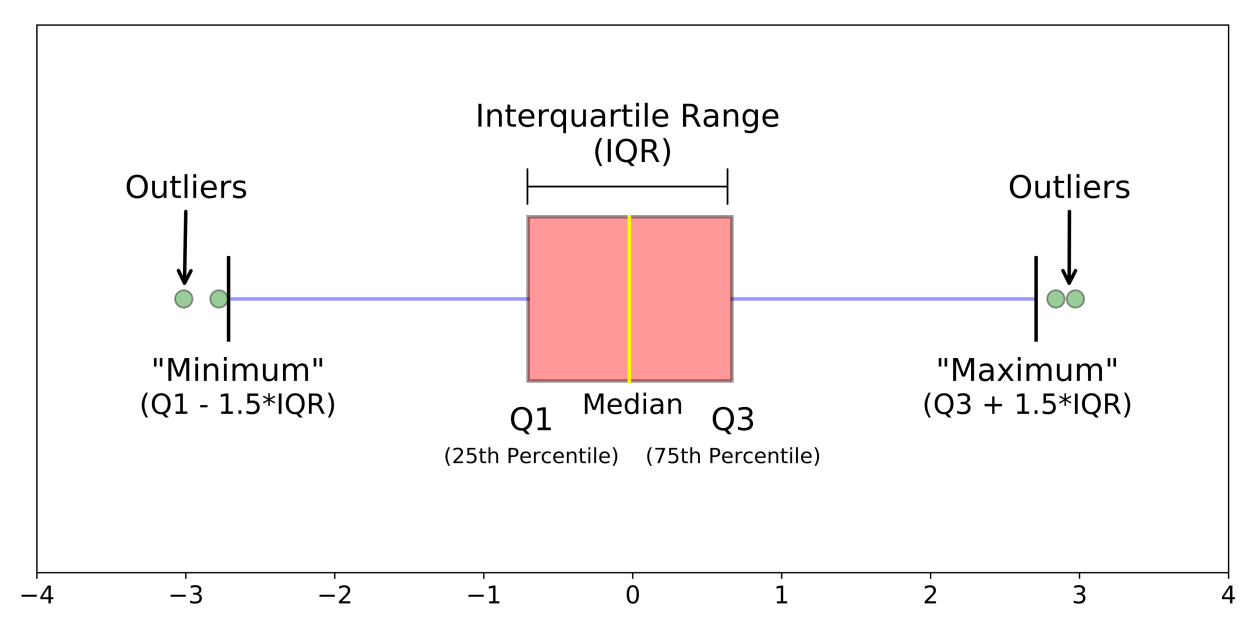
**Quartile**

Values that split the data set into quarters based on percentiles.

1. The first quartile is referred to as Q1, or the lower quartile. This value is the 25th percentile, in which the lower quarter of the values fall below the 25th percentile while three quarters are above it.
2. The second quartile, or Q2, is the value at the 50th percentile. This is the median of the data set.
3. Q3, the third quartile, is referred to as the 'upper quartile' and is the value of the 75th percentile, meaning only 25% of values in the set are above this value.

**Boxplot**

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.



\*\*IQR = Q3-Q1

**Coefficient of Variation**

 The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. It is generally expressed as a percentage.

### Formula

The **formula for the coefficient of variation**is:

Coefficient of Variation = (Standard Deviation / Mean) \* 100.  
In symbols: CV = (SD/[xbar](https://www.statisticshowto.com/wp-content/uploads/2009/09/xbar.bmp)) \* 100.

**Bi-variant Analysis**

**Covariance**

Covariance provides insight into how two variables are related to one another.

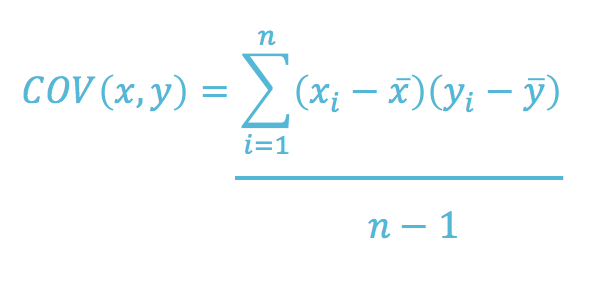
More precisely, covariance refers to the measure of how two random variables in a data set will change together.

A positive covariance means that the two variables at hand are positively related, and they move in the same direction.

A negative covariance means that the variables are inversely related, or that they move in opposite directions.

### How to Calculate Covariance

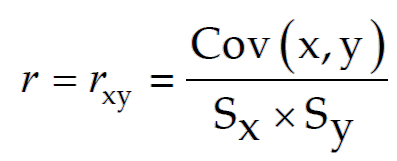
**The formula for covariance is as follows:**



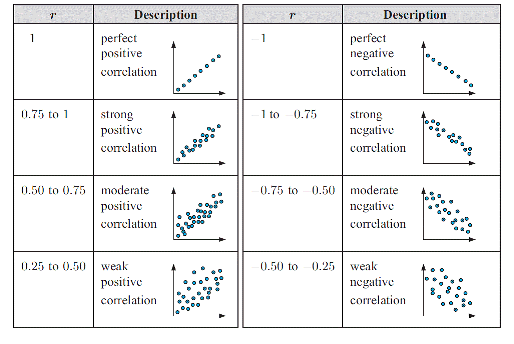
**Correlation coefficient**

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

* 1 indicates a strong positive relationship.
* -1 indicates a strong negative relationship.
* A result of zero indicates no relationship at all.



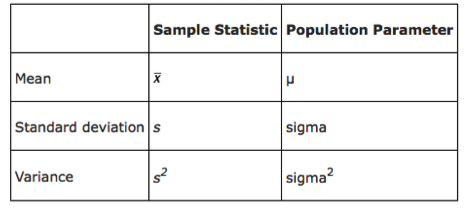
**Types of Correlation**



**Population and Sample**

Population : full set of Raw data / Business data usually available for any analysis.

Sample , is a set size taken from the population to derive at the inference / forecast of the data to take on business



**Central Limit Theorem (CLT)**

The Central Limit Theorem is the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the data distribution. An essential component of the Central Limit Theorem is the average of sample means will be the population mean.

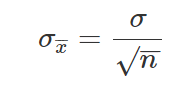
Similarly, if you find the average of all of the standard deviations in your sample, you will find the actual standard deviation for your population.

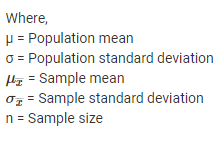
* Mean of sample is same as mean of the population.
* Standard deviation of the sample is equal to standard deviation of the population divided by square root of sample size.

Central limit theorem is applicable for a sufficiently large sample sizes (n≥30). The formula for central limit theorem can be stated as follows:

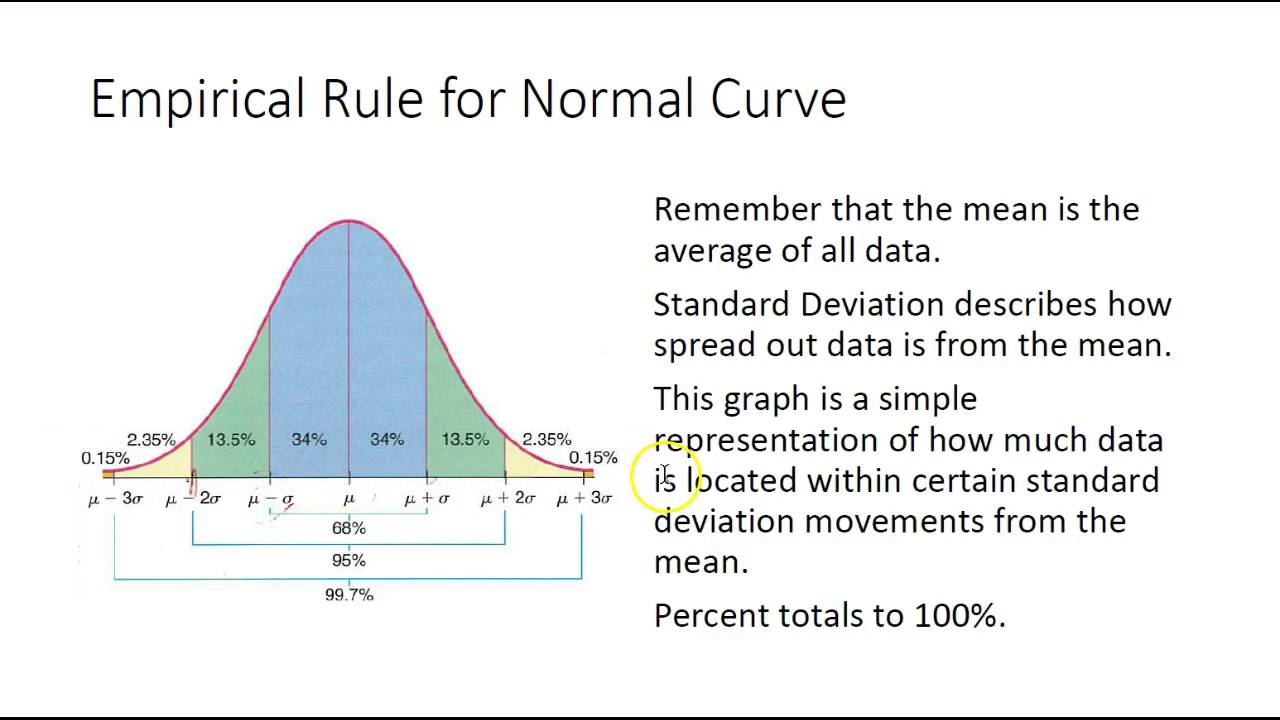


and



****

**Empirical Split in Normal distribution:**



**Z-Score**

The basic z score formula for a sample is: **z = (x – μ) / σ**

**The Z score tells you where is your score in the Normal distribution in the empirical split**

