**What is Statistics?**

Statistics is a mathematical study and transforming numbers into useful information like maps/Graphs/tables for decision making.

Statistics is to do the following in the data.

* Analyzing
* Interpreting
* Collecting
* Organizing

Based on the above way a data scientist will do decisions on the data model.

Why does a Data Scientist need to learn Statistics?

Statistics is a way to get information from data , by following the simple formulas of Statistics a data scientist will make simple to complex decisions and build a data model and present the study.

Knowledge of Statistics allows you to make better sense and use of the numbers/ raw data.

**Statistics will help in:**

* Present and describe a data properly in a readable manner
* Draw conclusions from large sets and subsets of collected information/data
* Make a reliable decision /forecast about an activity involving business
* Improve the process

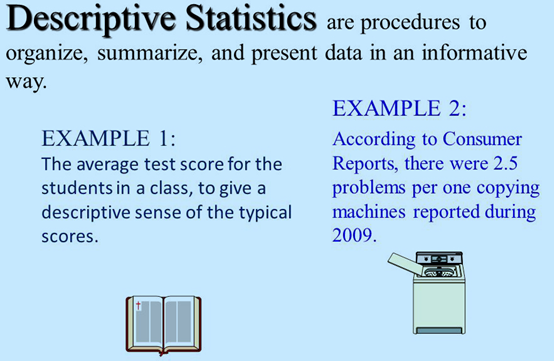
**Statistics Classification :**

Statistics is classified in to two

* Descriptive
* Inferential

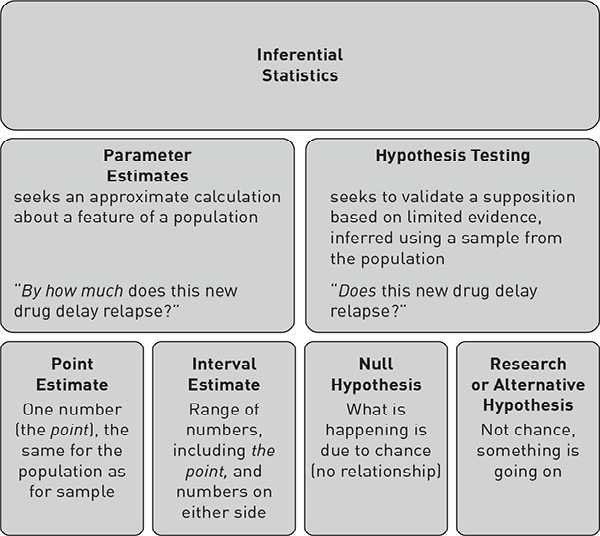
**Descriptive Statistics:**

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. ... Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows.

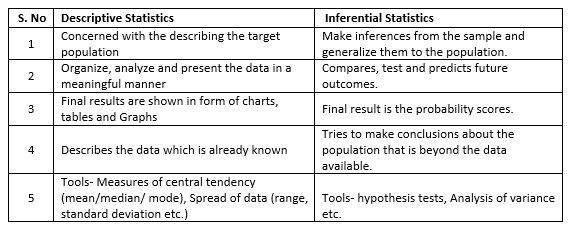


**Inferential Statistics:**

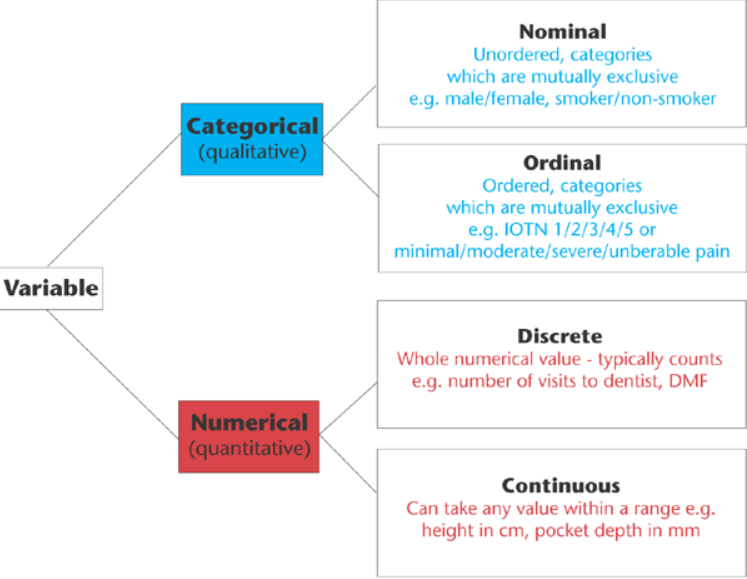
**Inferential statistics** are often **used to** compare the differences between the treatment groups. **Inferential statistics** use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.

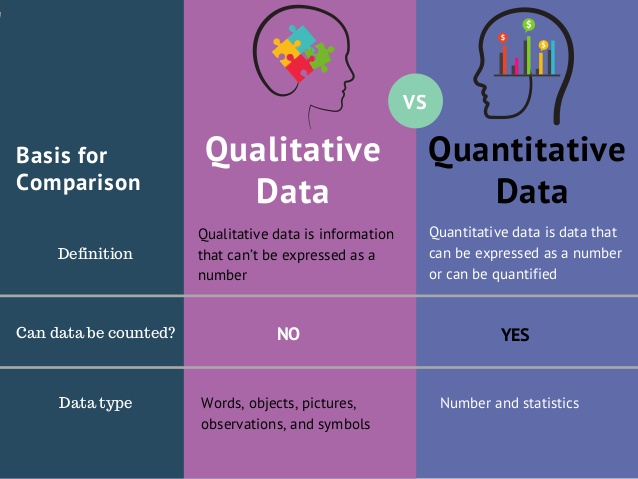


Comparison of Descriptive Vs Inferential

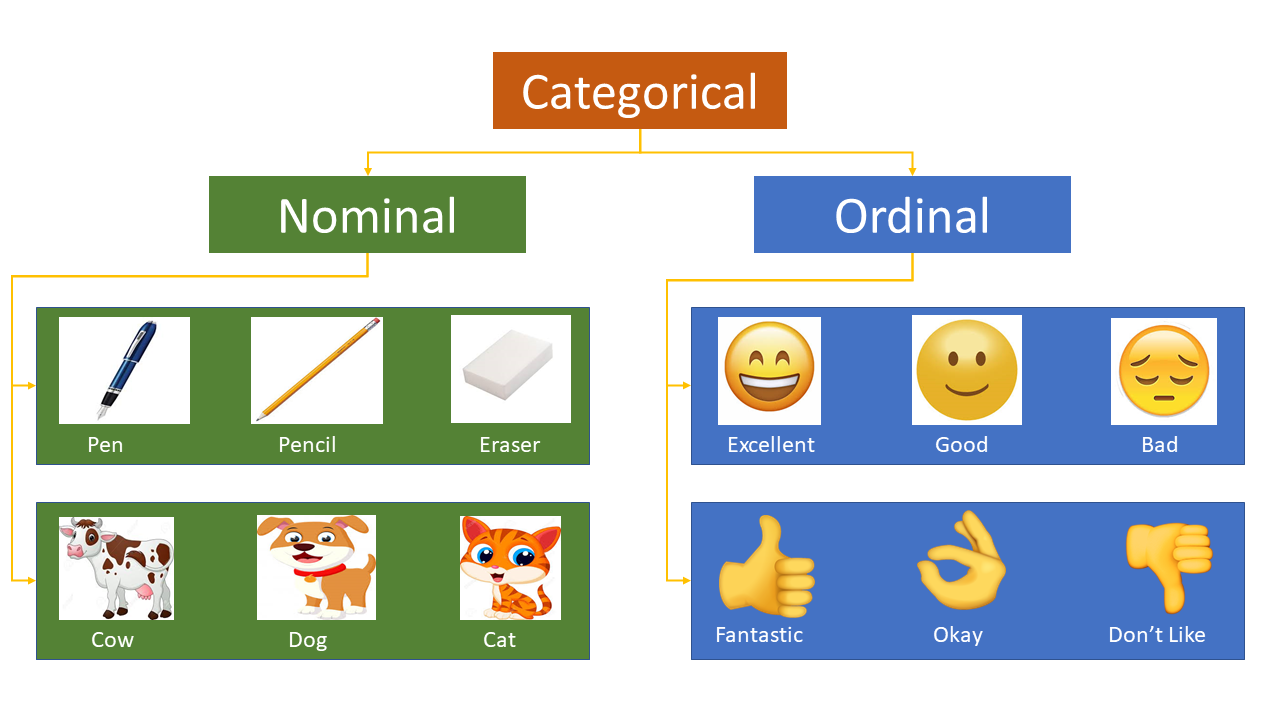


**Variables in Statistics:**

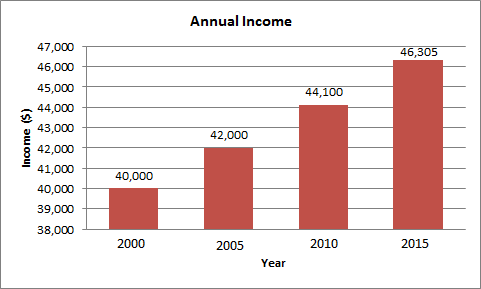




Examples of Categorical data:



Examples of Quantitative Data:



**Characteristics of Frequency Distribution:**

* Measures of central tendency and location i.e. mean, median, and mode.
* Measures of dispersion i.e. range, variance, and the standard deviation.
* The extent of the symmetry or asymmetry i.e. skewness.
* The flatness or the peakedness i.e. kurtosis.

Can also be said as below

* Modality
* Symmetry
* Central tendency
* Variability

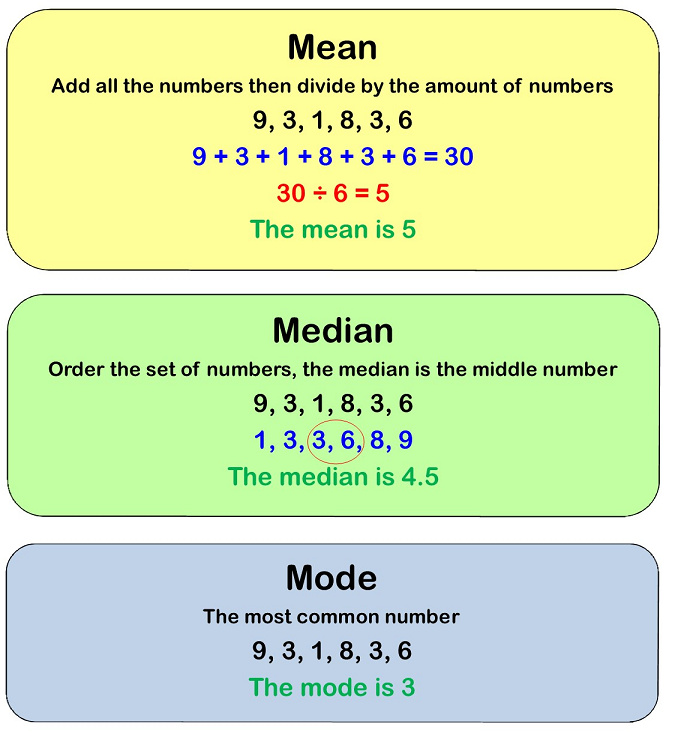
Modality is further divided into two unimodal and Bimodal

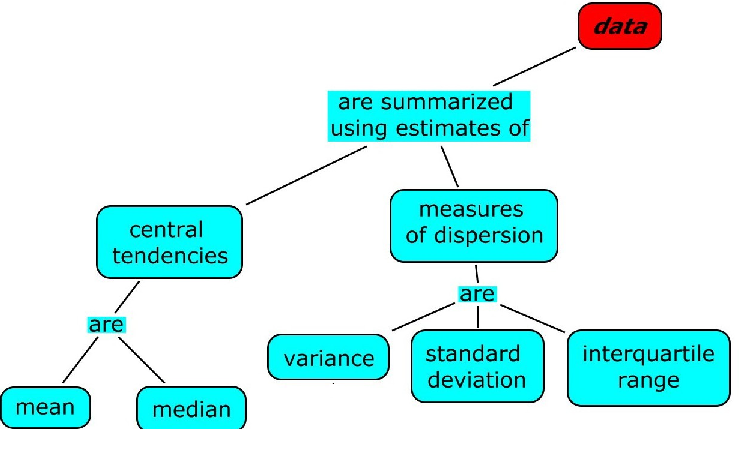
**Central Tendency**

In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages.

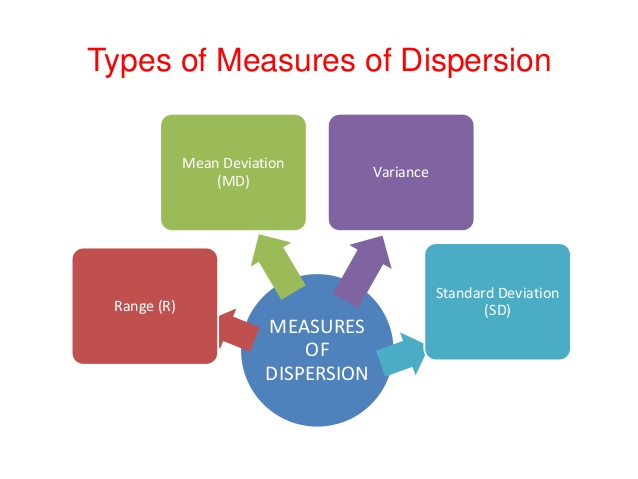
Measures of Central tendency are :

* Mean
* Median
* Mode





**Dispersion Measures**



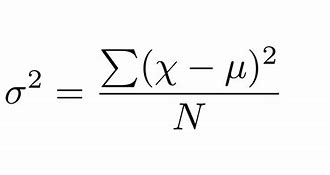
Range:

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X max and X minare the two extreme observations then Range = X max – X min

Quartile Division:

The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number. Quartile deviation or semi-inter-quartile deviation is Q = ½ × (Q3– Q1)

**variance** is the expectation of the squared deviation of a random variable from its mean, and it informally measures how far a set of (random) numbers are spread out from their mean



Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation.

Mean Deviation:

Practically speaking, the Range and the Quartile deviation separately cannot provide us the actual measurement of the variability of the values of a variable from their mean because they cannot ideally express the central value and the extent of scatteredness of those values around their average value. Moreover, these measures are not prepared on the basis of all the observations given for the variable. In order to avoid such limitations, we use another better method (as it is claimed) of dispersion known as the ‘Mean Deviation’.

**Outliers:**

Outliers are data values that differ greatly from the majority of a set of data. These values fall outside of an overall trend that is present in the data. A careful examination of a set of data to look for outliers causes some difficulty. Although it is easy to see, possibly by use of a stemplot, that some values differ from the rest of the data, how much different does the value have to be to be considered an outlier? We will look at a specific measurement that will give us an objective standard of what constitutes an outlier.

Interquartile Range

The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244) is what we can use to determine if an extreme value is indeed an outlier. The interquartile range is based upon part of the [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237) of a data set, namely the [first quartile and the third quartile](https://www.thoughtco.com/what-are-first-and-third-quartiles-3126235). The calculation of the interquartile range involves a single arithmetic operation. All that we have to do to find the interquartile range is to subtract the first quartile from the third quartile. The resulting difference tells us how spread out the middle half of our data is.

**Using Boxplot to identify Outliers**

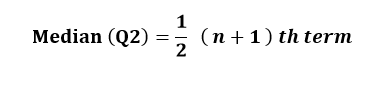
Box Plot Diagram

Box plot diagram also termed as **Whisker’s plot** is a graphical method typically depicted by quartiles and inter quartiles that helps in defining the upper limit and lower limit beyond which any data lying will be considered as outliers. ***The very purpose of this diagram is to identify outliers and discard it from the data series*** before making any further observation so that the conclusion made from the study gives more accurate results not influenced by any extremes or abnormal values.

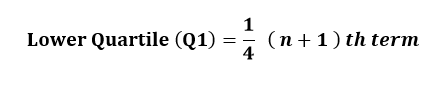
Identifying Outliers

Let **n** be the number of data values in the data set.

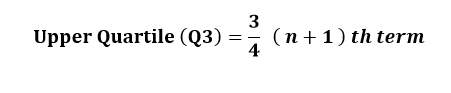
The **Median (Q2**) is the middle value of the data set.



The **Lower quartile (Q1)** is the median of the lower half of the data set



The **Upper quartile (Q3)** is the median of the upper half of the data set.



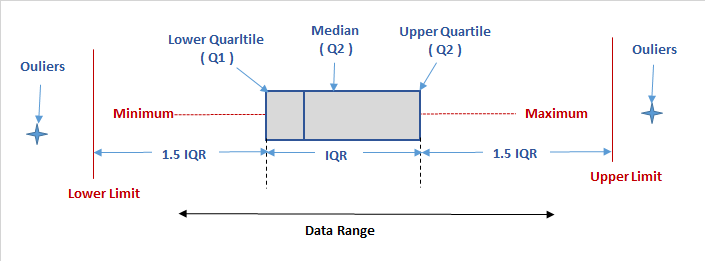
The **Interquartile range (IQR)** is the spread of the middle 50% of the data values.

Interquartile Range (IQR) = Upper Quartile (Q3) – Lower Quartile (Q1)

IQR = Q3 – Q1

**Lower Limit = Q1 – 1.5 IQR.**

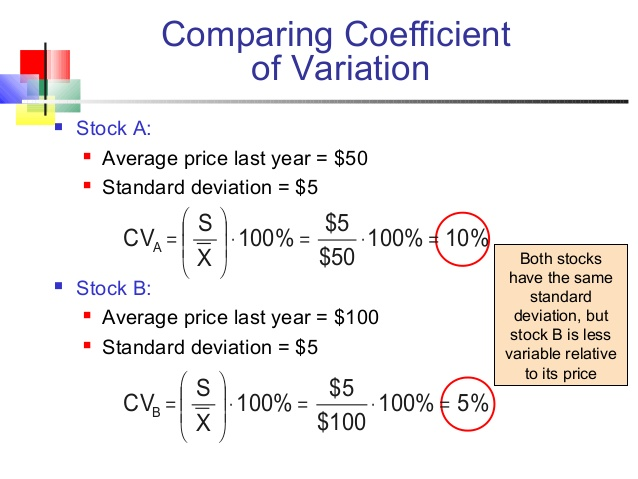
**Upper Limit = Q3 + 1.5 IQR**



**Coefficient of Variation:**

The coefficient of variation (relative standard deviation) is a **statistical measure of the dispersion of data points around the mean**. The metric is commonly used to compare the data dispersion between distinct series of data.

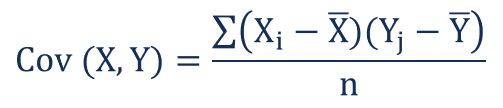
CV=SD/Mean \*100 %



**Measure of association between two variables**

* Covariance
* Correlation Coefficient

**covariance** is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, i.e., the variables tend to show similar behavior, the covariance is positive.



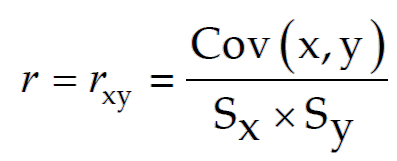
Higher the Value stronger the relation between X and Y

**Correlation Coefficient**

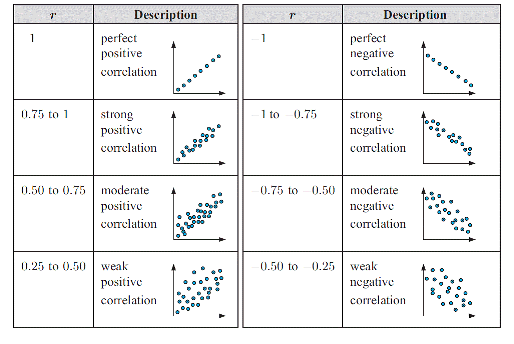
A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution

A measure of relationship not affected by the units of measurement

Ranges from =1 to +1.



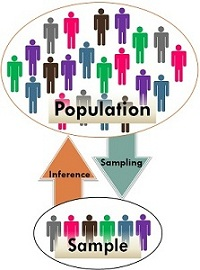
Types of correlation.



**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

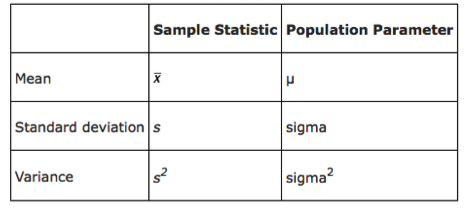


**Statistics and Parameter**

* Parameter is a numerical value summarizing all the data of an entire population . Ex Mean, Variance.
* Statistic is a numerical value summarizing the sample data ex Sample Mean , Sample Variance.

Examples:

* Average income of all employees in a company is a Parameter.
* Average income of a department in a company is a Statistic



**Central Limit Theorem.**

The central limit theorem (CLT) states that the **distribution of sample means approximates a normal distribution as the sample size gets larger**. Sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold.

Otherwise CLT states when you take a sample size of n of more than 30 samples in a population then your data will be always a Normal Distribution.

This is also called as Standard Error

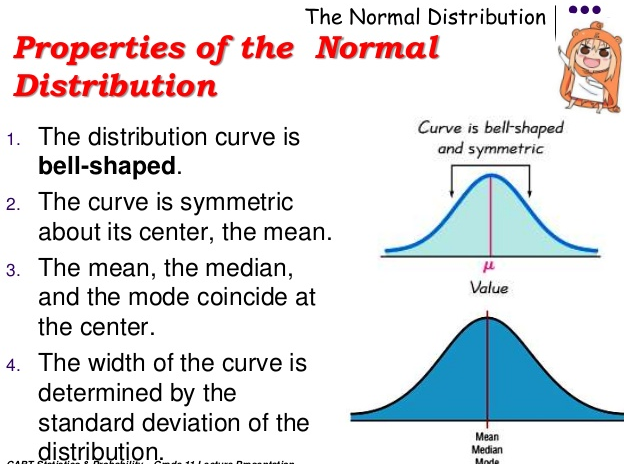
Standard Deviation of sample mean = (Population Standard deviation /

Square root of (No of samples))

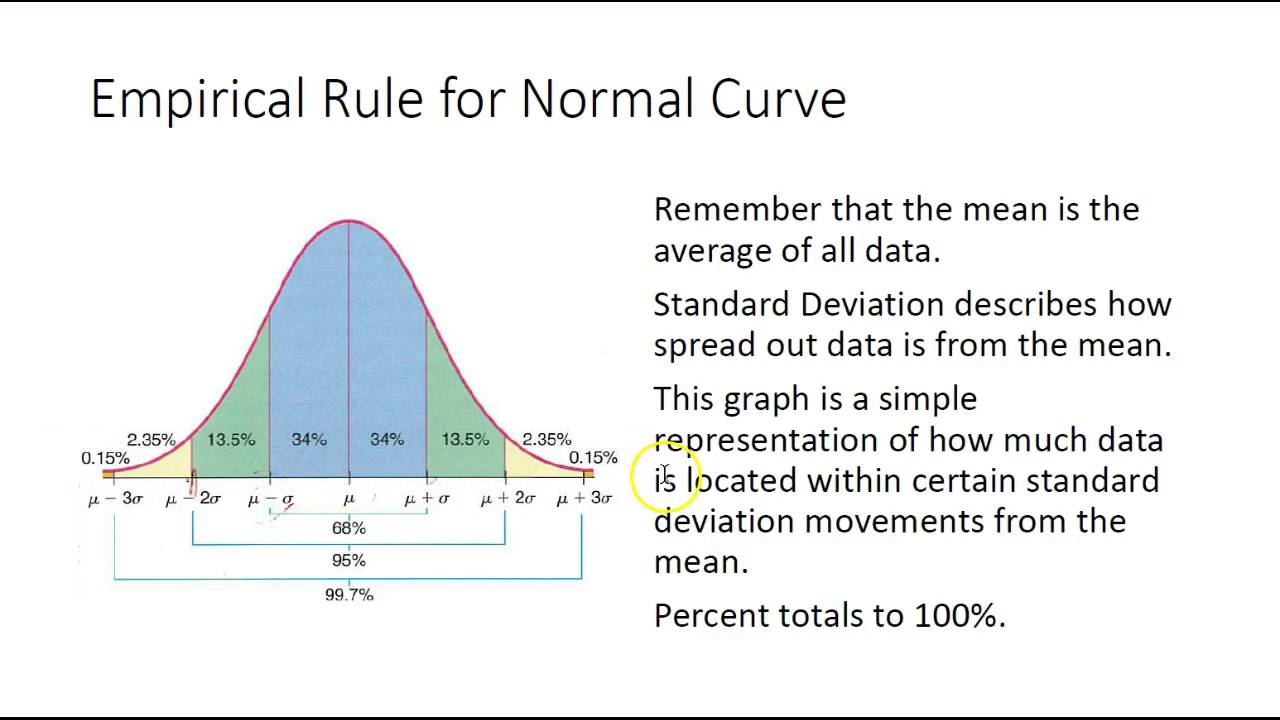
Mean of Sample mean = Population Mean

As Sample size increases Standard Error Decreases.

**Normal Distribution and its properties**



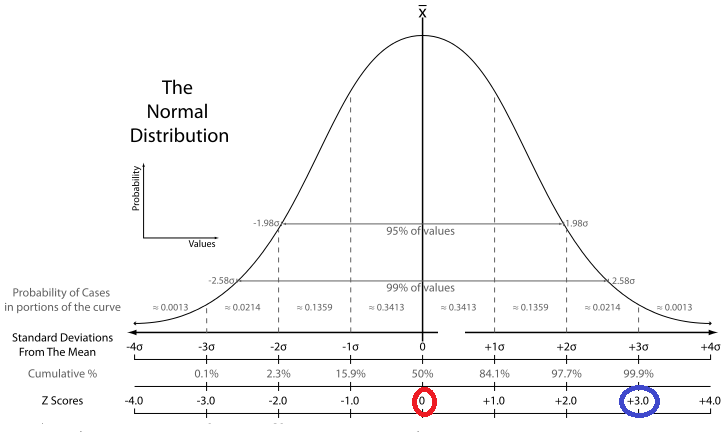
**Empirical Split in Normal distribution:**



**Z-Score**

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

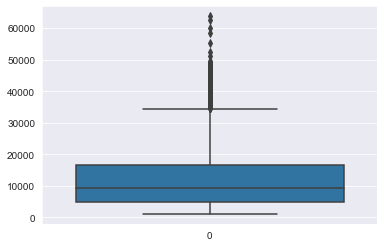
**The Z score tells you where is your score in the Normal distribution in the empirical split**

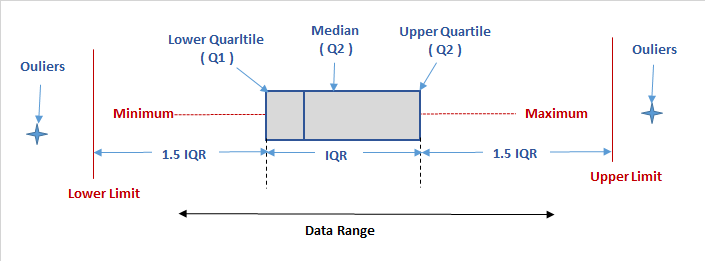


**Data Visualization Plots**

**Box Plot.**

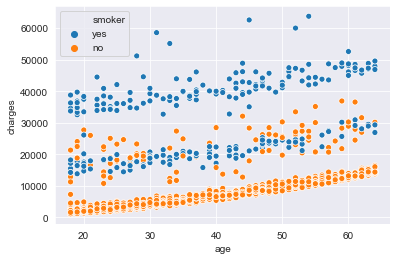
This shows the data spread for Individual columns.





**Scatter Plot :**

This shows the relationship between two columns



**Density Plot :**

This shows the Distribution of the data.

