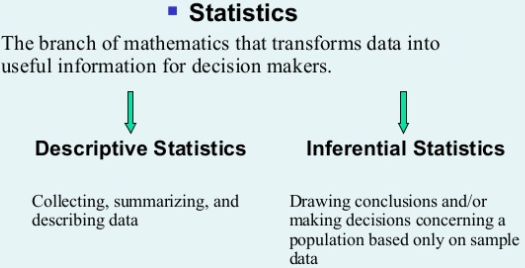
**Statistics**

A branch of mathematics taking and transforming numbers into useful information for decision makers. It is the process of **collecting data, analyzing data, interpreting data and presenting data.** It is a way to get information from data. Learning it is important as knowledge of statistics allows you to make a better sense of use of numbers.

**Types**:

Statistics can be classified into two types:

1. **Descriptive** Statistics
2. **Inferential** Statistics



**Descriptive statistics** present, summarize and organize characteristics of a data set(sample). A data set is a collection of responses or observations from a [sample or entire population](https://www.scribbr.com/methodology/population-vs-sample/).

In [quantitative research](https://www.scribbr.com/methodology/quantitative-research/), after [collecting data](https://www.scribbr.com/methodology/data-collection/), the first step of data analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

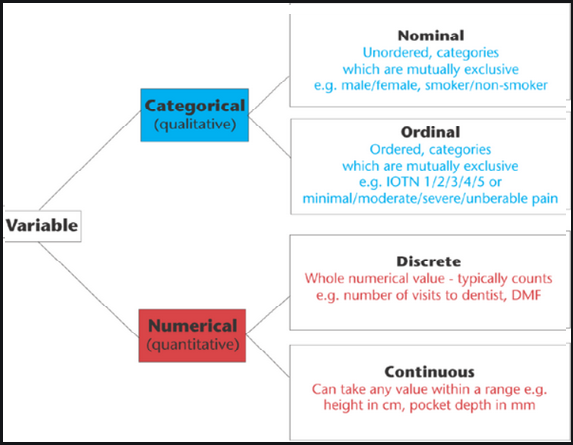
The next step is**inferential statistics**, which are tools that help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population. Drawing conclusions about a population based on data observed in a sample and use the data to learn about the [population](https://en.wikipedia.org/wiki/Statistical_population).

**Frequency Table:**

A frequency table is way of summarizing data by depicting the number of times a data value occurs so that it is easy to understand.

**Variables in Statistics:**

Statistical variables (data) can be classified in two ways, **quantitative** and **qualitative**

****

1. **Quantitative** **or** **Non-Numeric:**

Qualitative variables are variables that are **categorical** (string) in nature, or that don’t have any numerical representation. They are further classified as

* 1. **Ordinal:**

Ordinal variables are those that can be arranged in **orders** or during the arrangement where order matters.

* 1. **Nominal**

Nominal variables are those variables where order doesn’t matter at all.

1. **Quantitative** **or Numeric** **variables** :

**Quantitative**variables are variables that have some **numerical** representation and they contain some information numerically. Height, Age, Weight are the types that come under this category.They are further classfied as:

* 1. **Discrete**:

Variables whose values are **countable**(whole number)

e.g: Number of Students : 12

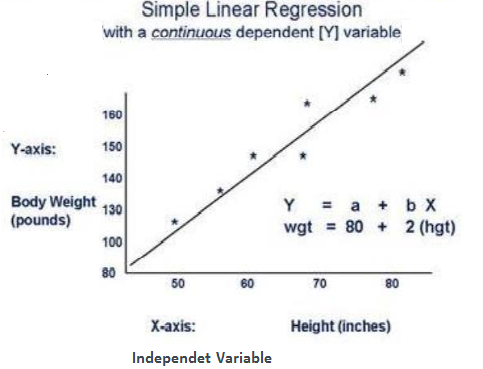
* 1. **Continuous**:

Variables whose values are **measurable**(continuous in nature - decimal)

e.g: Height : 175.5 cm.

Note: If age is 25(numerical) it is quantitative, if given as twenty five then it is qualitative.

**Variables  -  Dependent  &  Independent**

****

1. **Dependent  or** **Target** **or** **Class** **or** **Predeicted** **or** **output** **variable:**

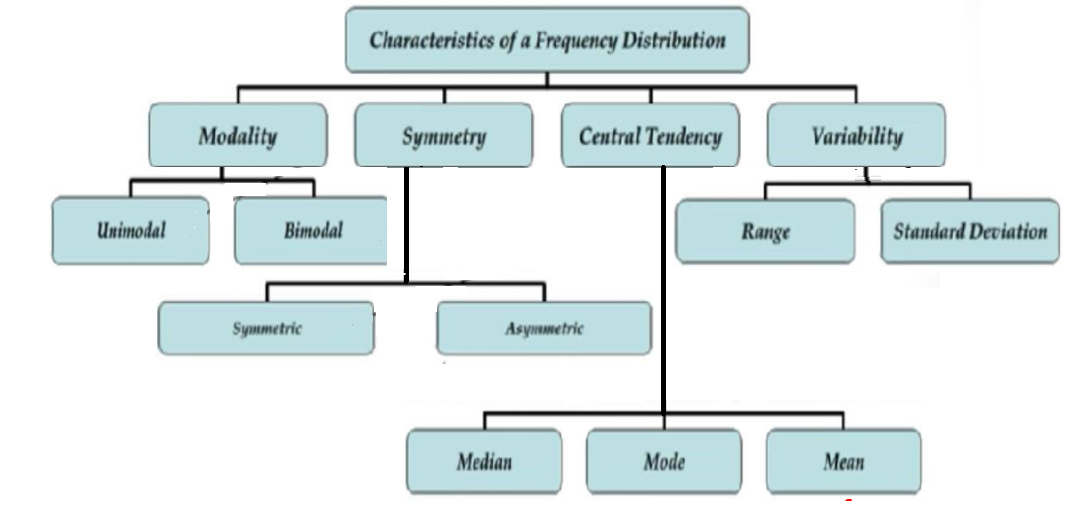
The dependent variables are the types of variables that are completely **dependent** on the independent variables (Feature/Column). The dependent variables are named as such because they are the values that are predicted or assumed by the predictor / independent variables.

1. **Independent  or** **Predictor variable:**

The independent variables are called as such because **independent** variables **predict** or forecast the values of the dependent variable in the model.

**Characteristics of frequency distribution  (Summarizing data):**

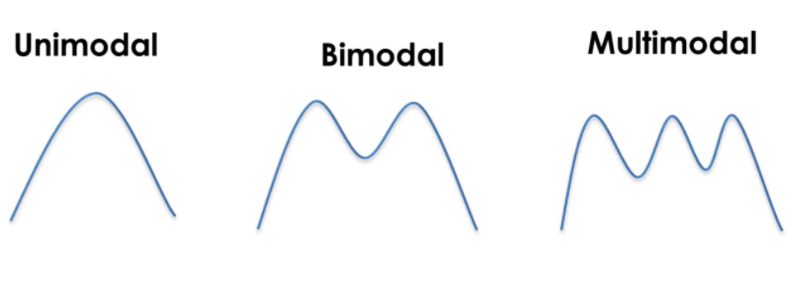
Frequency distribution in [statistics](https://www.toppr.com/guides/business-mathematics-and-statistics/statistical-description-of-data/introduction-to-statistics/) provides the information of the [number](https://www.toppr.com/guides/maths/can-you-see-the-pattern/pattern-in-figures-and-numbers/) of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list, table, or [graphical representation](https://www.toppr.com/guides/business-economics-cs/descriptive-statistics/graphic-presentation-of-data/)

****

1. **Modality:**

The **mode** or modal value is the number that appears **most frequently** in a data set. A set of numbers may have one mode, more than one mode, or no mode at all.

 Unimodal has single peak(mode=1),Bimodal has two peak(mode=2,there are two values that occured most frequently but it might have an issue)

****

1. **Symmetry:**

Symmetry can be classified as **symmetric** and **asymmetric.**

**Symmetric:**

Symmetrical distribution is a situation in which the values of variables occur at regular frequencies and the **mean, median, and mode** occur at **same** **points**.   Symmetric curve is **bell shaped curve**

**Asymmetric:**

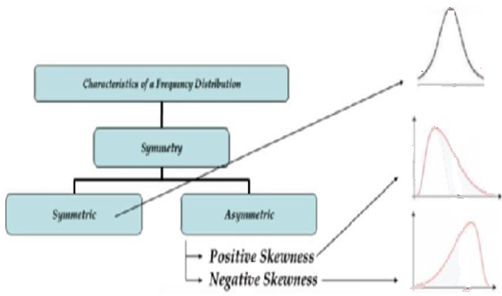
Asymmetrical distribution is a situation in which the values of variables occur at **irregular** frequencies and the **mean, median, and mode** occur at **different** points. An asymmetric distribution exhibits [**skewness**](https://www.investopedia.com/terms/s/skewness.asp).   Two types of skewness are :

1. **Positive skewness**or**right-skewed:**

Tail on the right side of the distribution is longer than left.

1. **Negative** **skewness**  or  **left-skewed:**

Tail on the left side of the distribution is longer than right.



1. **Central tendency:**

 A  **central tendency** (or **measure of central tendency**) is a central or typical value for a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution). It may also be called a **center** around which data is distributed.

The common measures of central tendency are **Mean**(average),   **mode**(most frequent value),   **Medain**(middle value in a sorted list of numbers in case of odd amount of numbers and average of middle values in a sorted list of even amount of numbers )

**Used for** : Missing value imputation.

1. **Variability**:

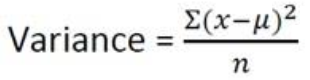
  When Central tendency fails(all mean,mode,median same),we check for consistency (dispersion measures) or reliabilty in data.The difference being exhibited by data points within a data set, as related to each other or as related to the mean. This can be expressed through the range, variance or standard deviation(tells how each value is distributed from mean) of a data set.

1. **Range**:

Difference between highest and lowest value.When we make use of time we use range.

1. **Standard deviation**(Deviation from the mean):

Measure of the amount of variation or dispersion of a set of values.  A low standard deviation indicates that the values tend to be close to the mean





**Percentile & Quartile:**

A percentile is a measure used in [statistics](https://en.wikipedia.org/wiki/Statistics) indicating the value below which a given [percentage](https://en.wikipedia.org/wiki/Percentage) of observations in a group of observations falls. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found. Equivalently, 80% of the observations are found above the 20th percentile.

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

****,

where P-The percentile you are interested.

n - Number of values

**Inter Quartile Range:**

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.It is a 5 number summary – Q1,Q2,Q3,min,max values.

Q1-First Quartile-25th percentile

Q2-First Quartile-50th percentile(Median)

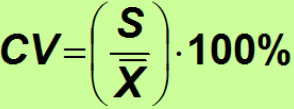
Q3-First Quartile-75th percentile

IQR(Inter Quartile Range)= **Q3-Q1**

**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. It gives variation in percentage. More variation means data will be more random.

Coefficient of variation = (Standard deviation/Mean)\*100%



**Univariate analysis** - Anaysis of one variable.

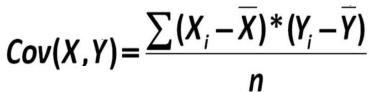
**Bivariate analysis** - It is the analysis of exactly two variables.

Note: Till now what we saw was Univariate analysis.

**Measures of assocaition between 2 variables (Bivariate analysis):**

1. Covariance
2. Correlation coefficient
3. **Covariance**:

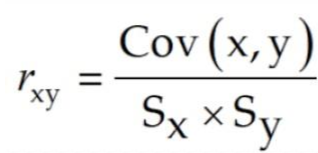
**Higher** the value **stronger** the relation between them. statistical measurements to know the relationship between the variance between the two variables (Variance between two columns).Magnitude is higher, covariance is stronger. Ranges from **-∞ to +∞**

****

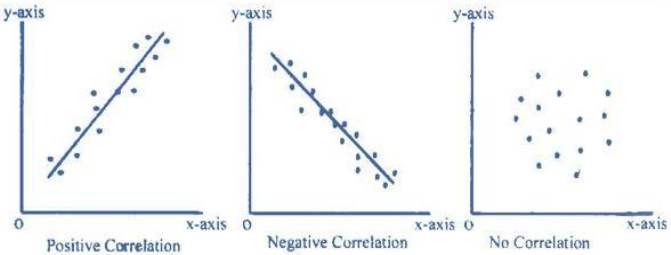
1. **Correlation coefficient:**

It is the statistical measure of the **strength of the relationship between the two variables**. Ranges from **-1 to +1**. A value of exactly **1.0** means there is a **perfect** **positive** relationship between the two variables. A value of **-1.0** means there is a **perfect negative** relationship between the two variables. . A value of **0** means there is a **no relationship** between the two variables(Independent). As covariance value is **unbound** we are not sure about the better results so we go for **correlation coefficient.** Value greater than +0.5 or -0.5, then has high correlation. If value is lesser than +0.5 or -0.5 then low correlation or neglible.

If there is a dependency between the independent variables, we try to resolve as it will affect the dependent variable

**. **

**Types of Correlation:**

****

**Population:**

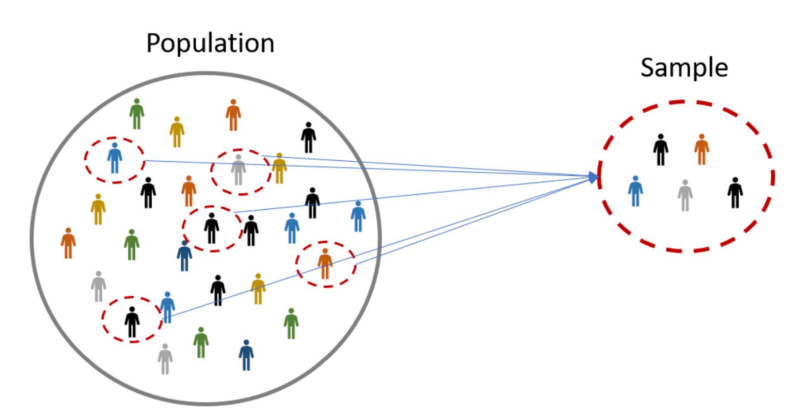
A **population** includes all of the [elements](https://stattrek.com/Help/Glossary.aspx?Target=element) from a set of data.The measurable characteristic of a population, such as a [mean](https://stattrek.com/Help/Glossary.aspx?Target=Mean) or [standard deviation](https://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation), is called a **parameter**

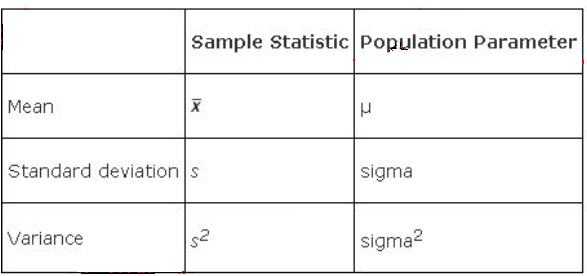
e.g: **Census**

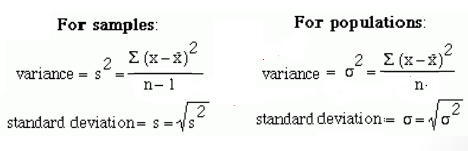
**Sample**

A **sample** consists one or more observations drawn from the population.The measurable characteristic of a population, such as a [mean](https://stattrek.com/Help/Glossary.aspx?Target=Mean) or [standard deviation](https://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation), is called a **statistic.**

**e.g**: Survey

****





**Degree of Freedom:**

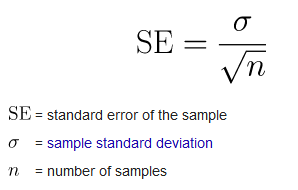
The number of independent values that can vary in an analysis that affect the measure. It is the number of values in the final calculation of a statistic that are free to vary

**Central Limit Theorem:**

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples(**n>=30**) from the population **with replacement**(replacing same value in another sample) , then the **distribution of the sample means will be approximately normally distributed irrespective of the distribution of the population** and the **mean of the sampling means**(Mean of the sample means) will be **approximately equal** to **population mean.**

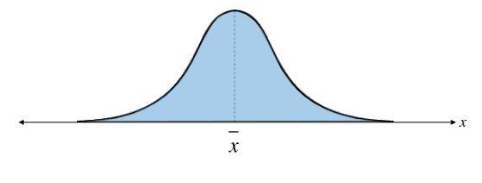
**Standard Error:**

It is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the parameter or the statistic is the mean, it is called the standard error of the mean also known as Standard deviation of sample mean.



**Properties of Normal Distribution:**

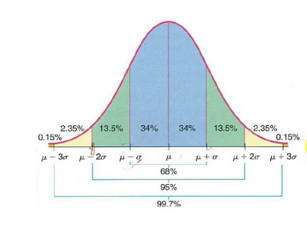
* The mean,mode and median are equal.
* The normal curve is bell-shaped and symmetric about the mean.
* Mostly We will convert data into normal distribution before feeding it into model.



**Empirical split or** [**Three-sigma rule**](https://www.investopedia.com/terms/t/three-sigma-limits.asp)**:**

The empirical rule, also referred to as the [**three-sigma rule**](https://www.investopedia.com/terms/t/three-sigma-limits.asp) or 68-95-99.7 rule, is a statistical rule which states that for a [**normal distribution**](https://www.investopedia.com/terms/n/normaldistribution.asp)**, almost all observed data will fall within three standard deviations (denoted by σ) of the mean or average (denoted by µ).**

In particular, the empirical rule predicts that 68% of observations falls within the first standard deviation (µ ± σ), 95% within the first two standard deviations (µ ± 2σ), and 99.7% within the first three standard deviations (µ ± 3σ).



**Z-Score:**

Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of [**standard deviations**](https://www.investopedia.com/terms/s/standarddeviation.asp) **from the mean**. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. To find in which region it lies in terms of S.D in the normal distribution.

**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.

**Variance Inflation Factor** **(VIF):**

It is a measure of the amount of [multicollinearity](https://www.investopedia.com/terms/m/multicollinearity.asp)(correlation high between independent variables) in a set of multiple [regression](https://www.investopedia.com/terms/r/regression.asp) variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model [variance](https://www.investopedia.com/terms/v/variance.asp) to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A **high VIF** indicates that the associated independent variable is **highly collinear** with the other variables in the model.

**Absolute deviation :**

The **standard** deviation is calculating the **square of the difference** whereas the mean absolute deviation is only looking at the absolute difference(without squaring we will take absolute of that value).

**Data Visualization:**

Data visualization is the **graphical representation** of information and data. By using visual elements **like charts, graphs**, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

**Matplotlib**:

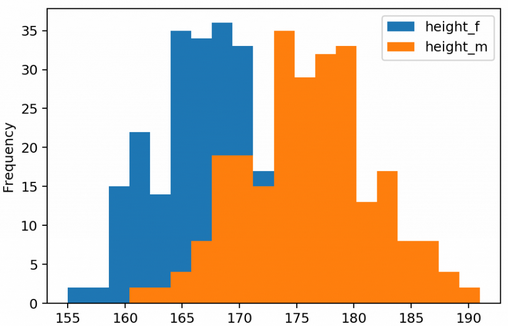
Matplotlib is plotting library for the Python programming language and its **numerical** **mathematics extension NumPy** library for creating static, animated, and interactive visualizations in Python.

**Seaborn:**

Seaborn is a Python visualization library **based on matplotlib**. It provides a high-level interface for drawing attractive statistical graphics

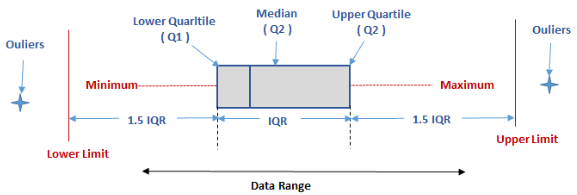
1. **Histogram:**

A **histogram** is a graphical display of data using bars of different heights. In a **histogram**, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A **histogram** displays the shape and spread of continuous sample data.



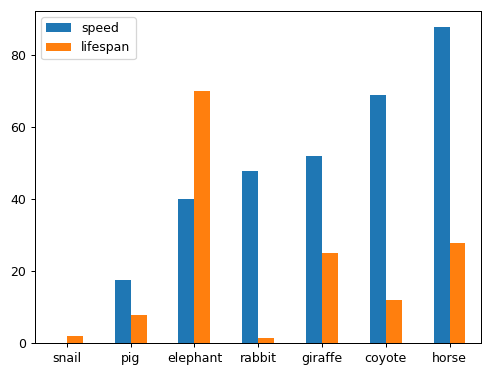
1. Box plot or Whisker plot:

A method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles. It is used to find outliers.

****

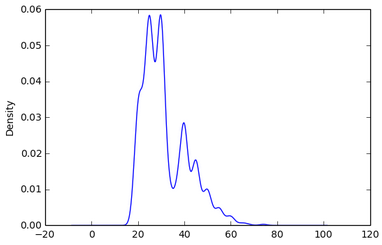
1. **Bar plot:**

Bar plots are used to compare variables. A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.

****

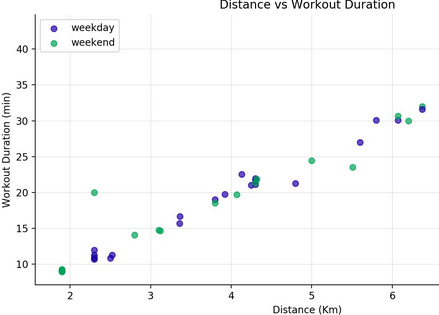
1. **Density plot:**

The density plot can be considered a continuous form of the histogram (for most purposes) that shows an estimate of the continuous distribution of a random variable

****

1. **Scatterplot:**

A **scatter plot** is a diagram drawn between two distributions of variables X and Y on a two dimensional plane. A scatter plot is used only as an **initial tool** in the process of finding any relationship between two variables.

****

**Statistical Simulation Link :**

<http://onlinestatbook.com/stat_sim/sampling_dist/index.html>