STATISTICS FOR DATA SCIENCE

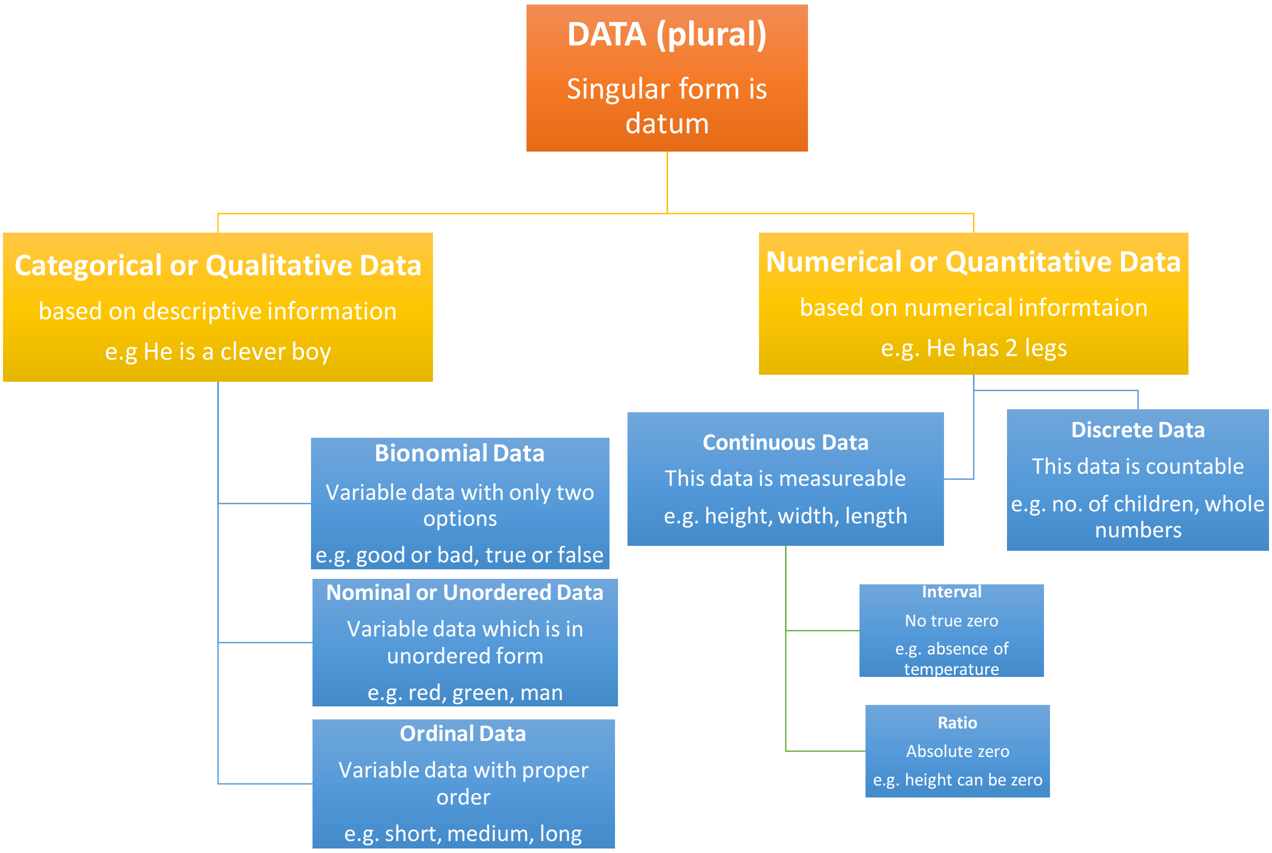
A. DESCRIPTIVE STATISTICS:

Before going to discuss about descriptive statistics, first we recall the basic concept of data and its types again here before starting descriptive statistics …..

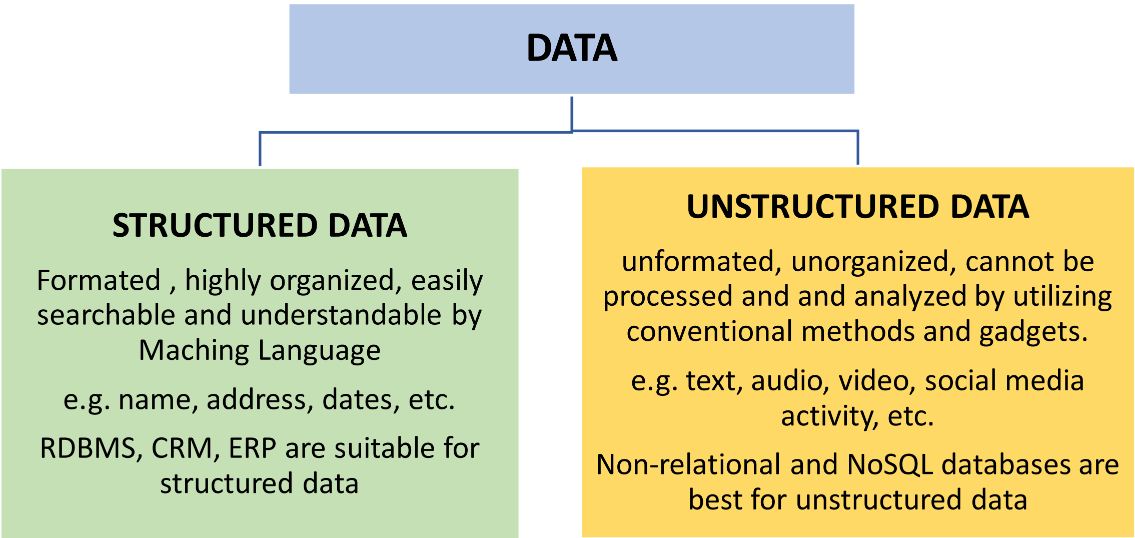
Data:

Data is a collection of factual information based on numbers, words, observations, measurements which can be utilized for calculation, discussion and reasoning.

TYPES OF DATA:


DATA (plural)
Singular form is datum
Categorical or Qualitative Data
based on descriptive information
e.g He is a clever boy
Numerical or Quantitative Data
based on numerical informtaion
e.g. He has 2 legs
Bionomial Data
Variable data with only two options
e.g. good or bad, true or false
Nominal or Unordered Data
Variable data which is in unordered form
e.g. red, green, man
Ordinal Data
Variable data with proper order
e.g. short, medium, long
Continuous Data
This data is measureable
e.g. height, width, length
Discrete Data
This data is countable
e.g. no. of children, whole numbers
Interval
No true zero
e.g. absence of temperature
Ratio
Absolute zero
e.g. height can be zero

The crude dataset is the basic foundation of data science and it may be of different kinds like Structured Data (Tabular structure), Unstructured Data (pictures, recordings, messages, PDF documents and so forth.) and Semi Structured.



DATA

STRUCTURED DATA

Formated , highly organized, easily searchable and understandable by Maching Language

e.g. name, address, dates, etc.

RDBMS, CRM, ERP are suitable for structured data

UNSTRUCTURED DATA

unformated, unorganized, cannot be processed and and analyzed by utilizing conventional methods and gadgets.

e.g. text, audio, video, social media activity, etc.

Non-relational and NoSQL databases are best for unstructured data

Furthermore, there are two kinds of data i.e. population data and sample data.

- **Population Data:**

Population data is the collection of all items of interest which is denoted by ‘N’ and the numbers we obtained when using population are called parameters.

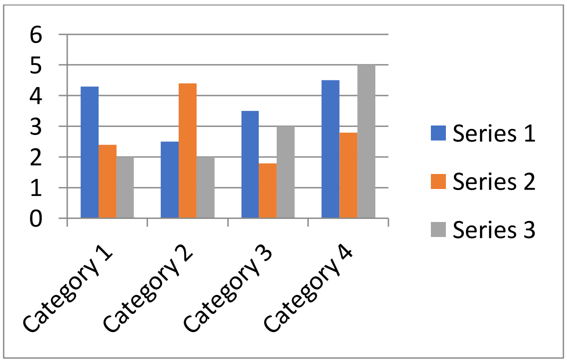
- **Sample Data:**

Sample data is a subset of the population which is denoted by ‘n’ and the numbers we obtained when using sample are called statistics.

Graphical Representation of variables in form of Graph & Tables:

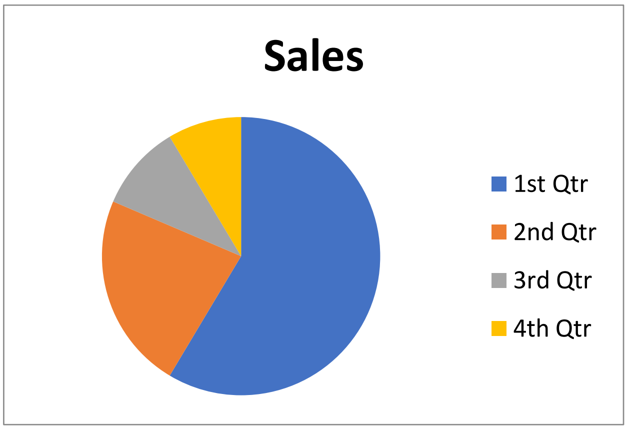
i. Bar Chart:

Bar charts are frequently being used to display data. In bar chart, each bar represents a category and y-axis shows the frequency as shown in figure



ii. Pie chart:

Pie Charts are frequently being used to display market share. If we want to see the share of any item as a part of the total then we utilized pie chart, as shown in figure below:



iii. Frequency Distribution Table:

Frequency distribution table shows the category and its corresponding absolute frequency as shown in figure

Category	Frequency
Black	12
Brown	5
Blond	3
Red	7

Relative frequency = frequency / total frequency

Measures of central tendency:

It is a single value that explains a set of data by identifying the central positing within that set of data. Measure of central tendency is also called measure of central location. The measures of central tendency are:

i. Mean
ii. Median
iii. Mode

i. Mean:

It is most popular to measures the central tendency. It is used with both discrete and continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. Therefore, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{x}$ is given by,

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

$$\bar{x} = \frac{\sum x}{n}$$

If we intend to calculate the population means instead of sample mean then we use the greet letter μ as

$$\mu = \frac{\sum x}{n}$$

ii. Median:

It is the mid score of a dataset that has been arranged in order of magnitude. In order to calculate the median, suppose we have the following dataset:

10	20	30	15	20	30	15	20	30	15	20

First of all, we re-arrange this data into order of magnitude from smaller to larger

10	15	15	15	20	**20**	20	20	30	30	30

Therefore, in this case bold figure 20 is our median. It is the middle mark, as there are 5 scores before it and 5 scores after it. However, if we have an odd number of scores like this one,

10	20	30	15	20	15	20	30	15	20

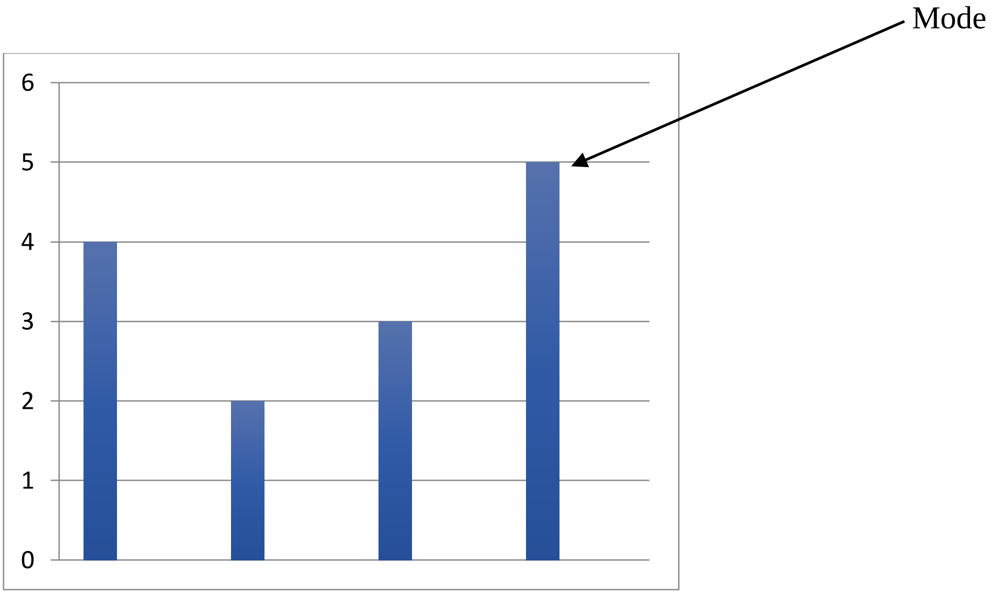
Then, we re-arrange this data into order of magnitude and we obtain

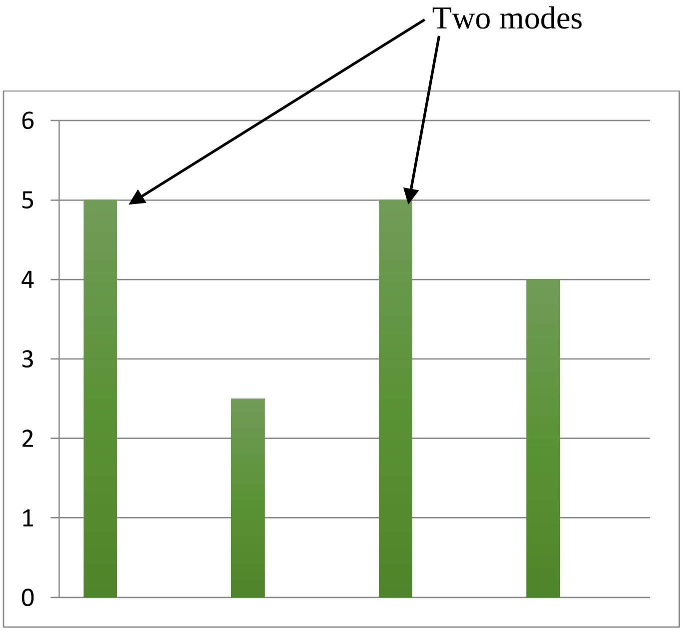
10	15	15	15	**20**	**20**	20	20	30	30

In this case, we have to take two values i.e. 20, 20 and average them to get a median i.e. 20.

iii. Mode:

It is a value that most often score in our dataset. A dataset can have no mode, one mode or multiple modes. It can be calculated by finding the value with the maximum frequency. For instance,





Measure of Asymmetry:

Skewness:

It is the measure of asymmetry that shows whether the observations in a dataset are focused on one side. Skewness can be calculated by the following formula

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i-}\,\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i-}\,\bar{x})^2}^{\,3}}$$

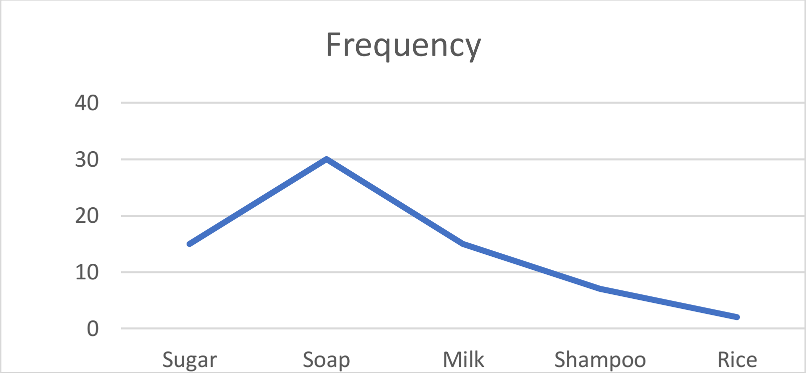
There are two types of skewness,

i. Right or Positive Skewness

ii. Left or Negative Skewness

i. Right or Positive Skewness:

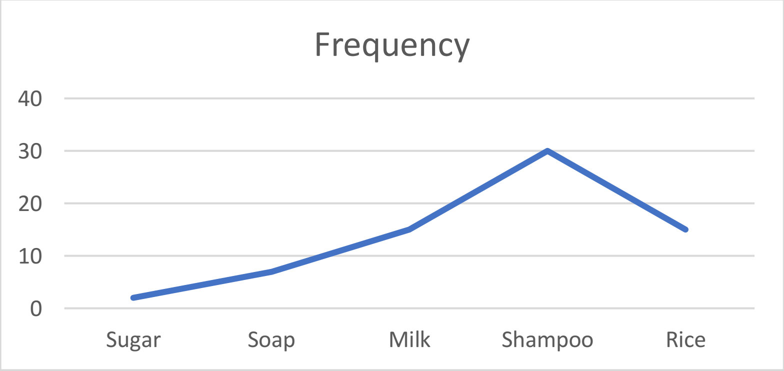
Right or positive skewness means that the outliers are to the right (long tail to the right) as shown in figure



Mean > median

ii. Left or Negative Skewness:

Left or negative skewness means that the outliers are to the left (long tail to the left) as shown in figure



Mean < median

However, if mean = median = mode then no skew and therefore, distribution will be symmetrical.

Variance and Standard Deviation:

Variance and Standard Deviation measure the dispersion of a set of data points around its means value.

Sample Variance formula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_{i-}\bar{x})^2}{n-1}$$

Population Variance formula:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_{i-}\mu)^2}{N}$$

Sample Standard Deviation formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_{i-}\bar{x})^2}{n-1}}$$

Population Standard Deviation formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_{i-}\mu)^2}{N}}$$

Coefficient of Variation:

There is no unit of measurement for Coefficient of variation. Coefficient of variation is perfect for comparison and universal across datasets. Formula of coefficient of variation is given below: -

$$CV = \frac{S}{\bar{x}}$$

Covariance and correlation:

Covariance	Correlation
Covariance is a statistical measure which is defined as a systematic relationship between a pair of random variables wherein change in one variable responded by an equivalent change in another variable	Correlation is a statistical measure which is defined as a systematic relationship between a pair of random variables wherein movement in one variable responded by an equivalent movement in another variable
The value of covariance lies between -∞ and +∞	The value of correlation lies between -1 and +1
A covariance of 0 means that the two variables	A correlation of 0 means that the two variables

are independent.	are independent
A positive covariance means that two variables move together	A correlation of 1 means perfect positive correlation
A negative covariance means that the two variables move in opposite directions	A correlation of -1 means perfect negative correlation.
Sample Covariance formula: $S_{xy} = \frac{\sum_{i=1}^{n}(x_{i-}\bar{x}) * (y_{i-}\bar{y})}{n-1}$ Population Covariance formula: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_{i-}\mu_x) * (y_{i-}\mu_y)}{N}$	Sample Correlation formula: $r = \frac{s_{xy}}{s_x s_y}$ Population Correlation formula: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

B. INFERENTIAL STATISTICS:

Probability distribution:

It is a statistical function that explains all the possible values and likelihoods that a random variable can take within a given range. This range will be bounded between the least and the highest possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors like distribution's mean, standard deviation, skewness, and kurtosis. Few examples of distributions are,

- Normal distribution
- Binominal distribution
- Student's T distribution
- Uniform distribution
- Poission distribution

Mostly, there is a confusion that distribution is a graph but in fact, it is the rule that help us in determining how the values are positioned in relation to each other.

i. Normal Distribution:

It is also known as Gaussian distribution or Bell Curve. It is mostly used in regression analysis. A lot of things closely follow this distribution:

- heights of people

- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test
- stock market information

When data is normal distributed then distribution is symmetric and

Mean = median = mode

$$N \sim (\mu, \sigma^2)$$

Where, N for normal, ~ for distribution, μ is mean, and σ^2 is the variance

ii. Standard Normal Distribution:

It is a normal distribution with a mean of 0 and a standard deviation of 1. Every normal distribution can be standardized using the following formula

$$a = \frac{x - \mu}{\sigma}$$

$$N \sim (0, 1)$$

Standardization permit to compare different normally distributed datasets, test hypothesis, detect outliers and normality, create confidence intervals and perform regression analysis.

iii. Central Limit Theorem:

This theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are the same in size), regardless of population distribution shape. If the sample sizes= or >30 are considered enough for the Central Limit Theorem to hold. The main aspect of this theorem is that the average of the sample means and standard deviations will equal the population mean and standard deviation. Furthermore, an adequately large sample size can forecast the characteristics of a population accurately. In Central Limit Theorem,

- No matter the distribution
- The more samples, closer to Normal (k ->∞)
- The bigger the samples, the closer to Normal (n -> ∞)

Estimators and Estimates:

Estimators:

It is a mathematical function of the sample that tell us that how to calculate an estimate of a parameter from a sample. Smaller the variance, most efficient the estimator. Hence, we required to find what are the "good" estimators. Few vital criteria for goodness of an estimator are based on these properties: -

- Bias
- Variance
- Mean Square Error

Examples of estimators and equivalent parameters are given in below table.

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	R	ρ

Estimates:

An estimate is the output value that you can get from an estimator. There are following types of estimates:

i. Point Estimates – a single value, e.g. 1, 6, 12.34, 0.123

ii. Confidence Interval Estimates – an interval, e.g. (1,4), (43, 45), (3.22, 5.33), (-0.24, 0.26). We mostly used confidence interval estimates when making inferences because it is more precise as compare to point estimates.

Confidence Interval:

It is an interval within which we are assured with certain %age of confidence, the population parameter will fall.

Margin of Error:

A margin of error explains how many percentage points your results will differ from the real population value. It can be calculated by the following two ways:

i. Margin of error = Critical value x Standard deviation
ii. Margin of error = Critical value x Standard error of the statistic

Student's T Distribution:

It is mostly used to estimate population parameters when the sample size is small and/or population variance are not known. It is pertinent to mention here that it is very useful in such cases where we have not enough information or too much cost is involve to acquire the requisite information. It has fatter tails as compare to normal distribution and lower peak. Following formula can be used to get the student's T distribution for a variable with a normally distributed population:

$$t_{v,\alpha} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \text{ where v are the degree of freedom}$$

C. HYPOTHESIS TESTING:

Scientific Method:

The scientific method is a process for gathering data and processing information. It was first sketched by Sir Francis Bacon (1561-1626) to provide logical, rational problem solving across many scientific fields. The main principle of scientific method is systematic observation, predictability, verifiability and amendment of hypothesis. The basic steps of the scientific method are:

- Make an observation that explains the issue
- Make an hypothesis or potential solution to the problem
- Test the hypothesis
- If the hypothesis is true then find further evidence or against-evidence
- If the hypothesis is false then create a new one or try again
- Draw conclusions and purify the hypothesis

What is hypothesis?

A hypothesis is an assumption based on inadequate evidence that requires further testing and experimentation. After further testing, a hypothesis can generally be confirmed true or false.

Null Hypothesis (H_0):

A null hypothesis is a hypothesis which is required to be tested. It is the hypothesis that the investigator is trying to show to be false. It is a status-quo. The concept of null is similar to someone remain innocent until enough evidence to prove guilty. For instance, someone say, data engineer normal salary is Rs.1,25,000/- but in our opinion he may be wrong, so, we make statistical testing to reject this hypothesis, it is called null hypothesis.

Alternative Hypothesis (H_1 or H_A):

An alternative hypothesis is inverses of the null hypothesis which is usually based on our own opinion. For instance, someone say, data engineer normal salary is Rs.1,25,000/- but in our opinion, data engineer cannot earn this value (less salary), it is called alternative hypothesis.

DECISIONS:

After testing, there will be two possibility of decisions i.e. accept the null hypothesis or reject the null hypothesis. Accept the null hypothesis means there is insufficient data to support the alteration or novelty brought by the unconventional. Reject the null hypothesis means there is sufficient statistical evidence that show this null hypothesis is false.

Level of Significance:

It is the probability of rejecting a null hypothesis by the test when it is really true. It is denoted by α (Alpha).

Confidence Level:

It is a possibility of a parameter that lies within a specified range of values. It is denoted as C. Level of significance is connected with the confidence level and the relationship between them is denoted by $c = 1 - \alpha$. The common level of significance and the corresponding confidence level are given below:-

- ❖ The level of significance 0.10 is related to the 90% confidence level.
- ❖ The level of significance 0.05 is related to the 95% confidence level.
- ❖ The level of significance 0.01 is related to the 99% confidence level.

The rejection rule is given below:-

- ✓ If p-value ≤ level of significance, then reject the null hypothesis.
- ✓ If p-value > level of significance, then do not reject the null hypothesis.

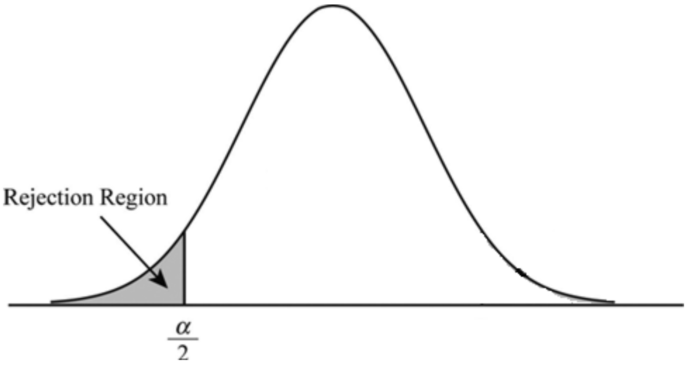
Rejection region:

The rejection region is the values of test statistic for which the null hypothesis is rejected.

Non rejection region:

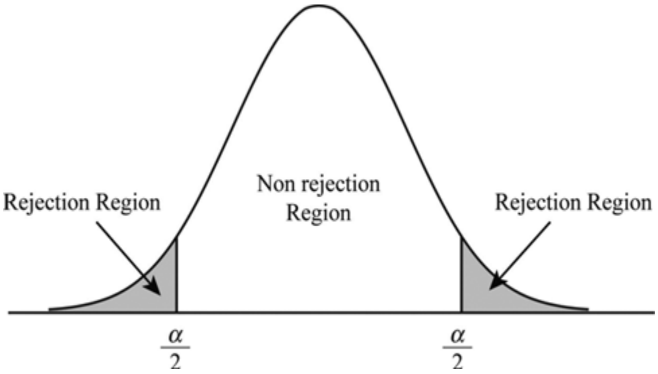
The set of all possible values for which the null hypothesis is not rejected is called the rejection region.

One sided (one-tailed) test is used when the null does not contain equality or inequality sign (<, >, ≤, ≥). The rejection region for one-sided (one-tailed) test is shown in figure:



- In the left-tailed test, the rejection region is shaded in left side (as shown in above figure).
- In the right-tailed test, the rejection region is shaded in right side.

Two sided (two-tailed) test is used when the null contains equality (=) or inequality (≠) sign. The rejection region for two-sided (two-tailed) test is shown in figure:-



Statistical Errors:

There are two types of statistical errors:

i. Type I Error (False Positive)

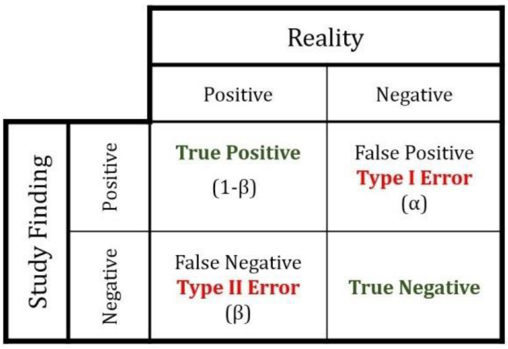
ii. Type II Error (False Negative)

Type-I Error (False Positive):

Type-I error occurs when we **reject** a null hypothesis that is actually **true.** The probability of committing type-I error is denoted by α (alpha).

Type-I Error (False Negative):

Type-II error occurs when we **accept** a null hypothesis that is actually **false.** The probability of committing type-II error is denoted by β (Beta).



		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

P-value:

The p-value is the smallest level of marginal significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in support of the alternative hypothesis. Usually, *p*-value is found with 3 digits after the dot (x.xxx).

The *p*-value is a number between 0 and 1 and can be interpreted as:

- A small *p*-value (typically ≤ 0.05) represents strong evidence against the null hypothesis, so, we reject the null hypothesis.
- A large *p*-value (> 0.05) represents weak evidence against the null hypothesis, so, we fail to reject the null hypothesis. 0.05 is often the *"cut-off-line"*.