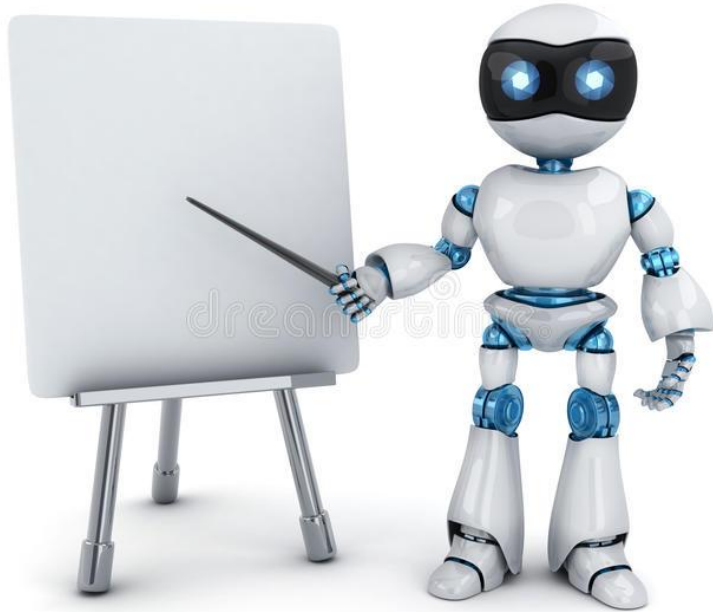


# Statistics Essentials



# Overview of Statistics

## Module - 1

The science of collecting, describing, and interpreting data is popularly known as Statistical leveraging in Data Science.



## Two areas of Statistics:

**Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way

**Inferential statistics** – The methods used to determine something about a population on the basis of a sample

**Descriptive statistics** are methods for organizing and summarizing data.

For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.

A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

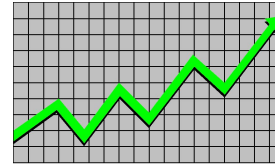
Collect data

e.g., Survey



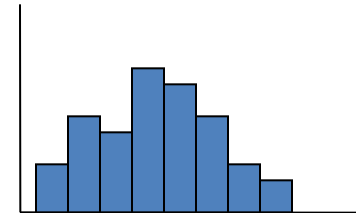
Present data

e.g., Tables and graphs



Summarize data

e.g., Sample mean = 
$$\frac{\sum X_i}{n}$$



- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.
- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg



**Inference** is the process of drawing conclusions or making decisions about a **population** based on **sample results**

**Population:** A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

**Sample:** A subset of the population.

**Variable:** A characteristic about each individual element of a population or sample.

**Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

# Definition to Basic terms

**Data (plural):** The set of values collected for the variable from each of the elements belonging to the sample.

**Random Variable:** Variable are placeholder where you can store anything. It can number, or string, sentences.

**Experiment:** A planned activity whose results yield a set of data.

**Parameter:** A numerical value summarizing all the data of an entire population.

**Statistic:** A numerical value summarizing the sample data.

Let's first understand the basic concepts of statistics.



## Statistical Population

A collection of all probable observations of a specific characteristic of interest

**Example:** All learners taking this course



## Sample

A subset of population

**Example:** A group of 20 learners selected for a quiz



## Variable

An item of interest that can acquire various numerical values

**Example:** The number of defective items manufactured in a factory

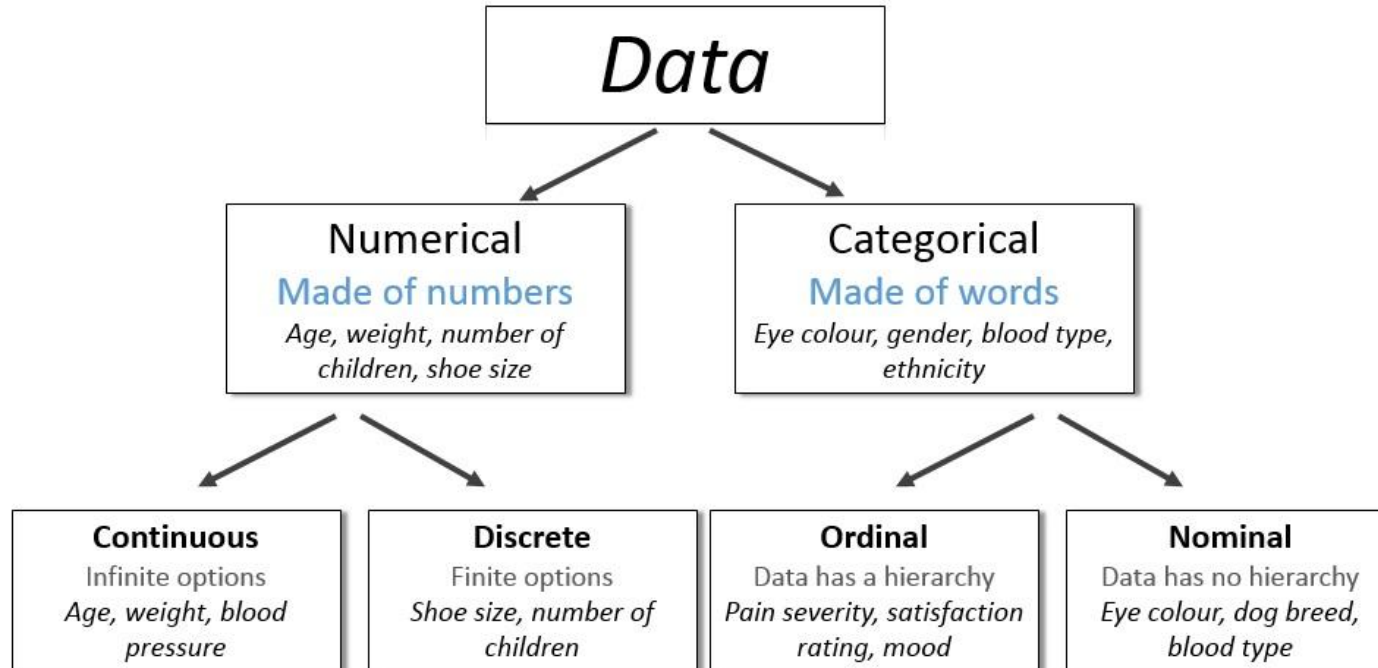


## Parameter

A population characteristic of interest

**Example:** The average income of a class of people





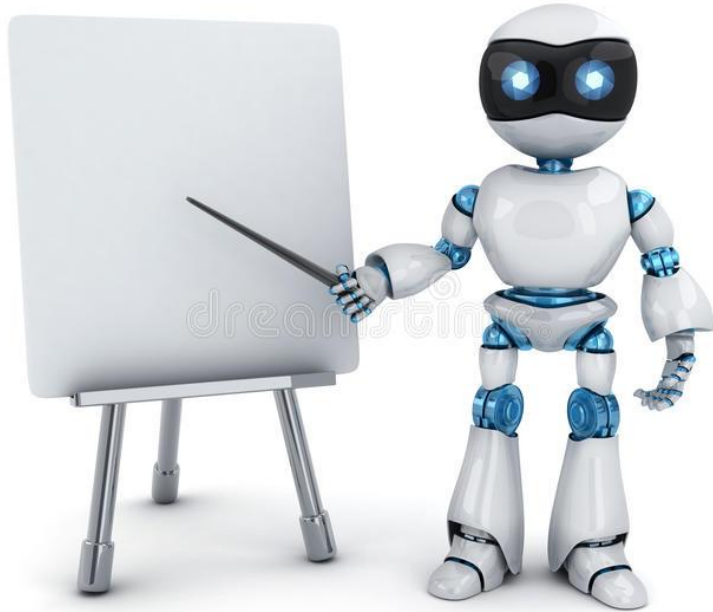
*Example:* Identify each of the following as examples of qualitative or numerical variables:

1. The temperature in Barrow, Alaska at 12:00 pm on any given day.
2. The model of automobile.
3. Whether or not a 6 volt lantern battery is defective.
4. The weight of a lead pencil.
5. The length of time billed for a long distance telephone call.
6. The brand of cereal children eat for breakfast.
7. The type of book taken out of the library by an adult.

*Example:* Identify each of the following as examples of

(1) nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

- The length of time until a pain reliever begins to work.
- The number of chocolate chips in a cookie.
- The number of colors used in a statistics textbook.
- The brand of refrigerator in a home.
- The overall satisfaction rating of a new car.
- The number of files on a computer's hard disk.
- The pH level of the water in a swimming pool.
- The number of staples in a stapler.



# Harnessing Data

## Module - 2

- Collecting the data
- Presenting the data-->Visualization using Matplotlib and Seaborn.
- Summarizing the data-->Module 3

A sample which is drawn from the population should have same characteristics as the population.

Sampling can be:

- **with replacement:** a member of the population may be chosen more than once
- **without replacement:** a member of the population may be chosen only once (lottery ticket)

**Step1:-** Define the object or aim of the experiment.

i.e Estimate the average life of electronic component

**Step2:-** Define the variable and population of interest.

i.e usage, power rating, battery life etc

**Step3:-** Defining the data collection scheme and data measuring scheme.

i.e sampling procedure, sample size, data measuring device.

**Step4:-** Defining the appropriate descriptive and inferential analysis techniques

**Experiment:** The investigator controls or modifies the environment and observes the effect on the variable under study.

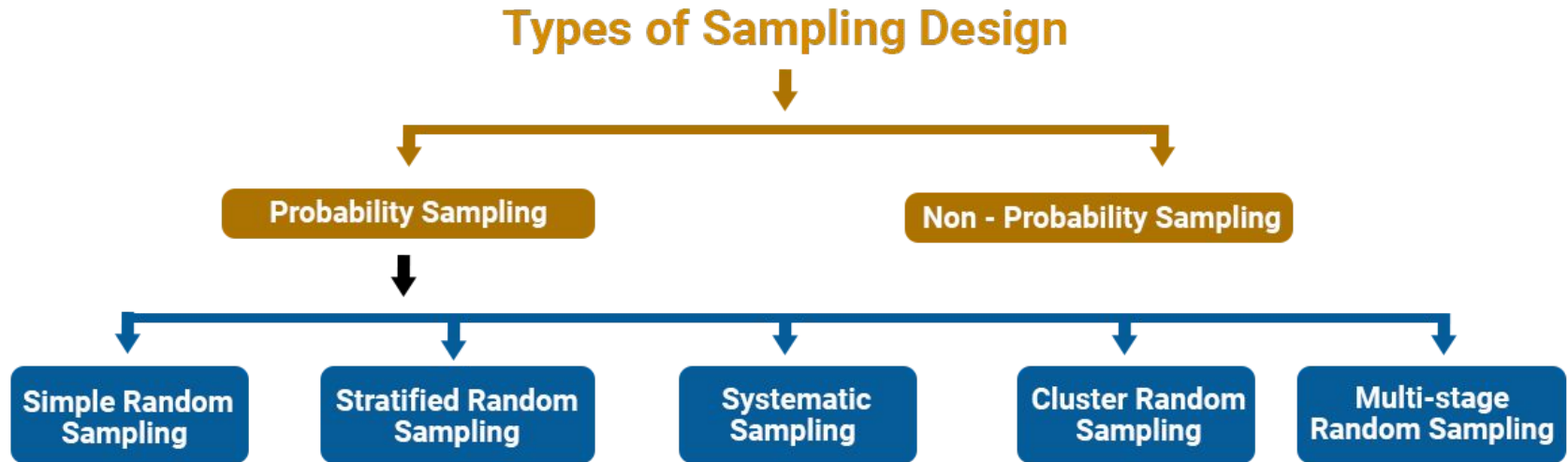
**Survey:** Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

**Census:** A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

**Judgment Samples:** It is a non-probability sampling technique in which the sample members are chosen only on the basis of the researcher's knowledge and judgment.

**Probability Samples:** Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.





- 1. Simple Random sampling** :-each sample of the same size has an equal chance of being selected.
- 2. Stratified Sampling** :-divide the population into groups called strata and then take a sample from each stratum.
- 3. Cluster sampling** :-divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.
- 4. Systematic sampling** :-randomly select a starting point and take every n-th piece of data from a listing of the population.
- 5. Multistage Random** :- divide the population into clusters and select some clusters at the first stage. At each subsequent stage, you further divide up those selected clusters into smaller clusters, and repeat the process until you get the desired sample size.

*Example:* An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

There are 2712 employees.

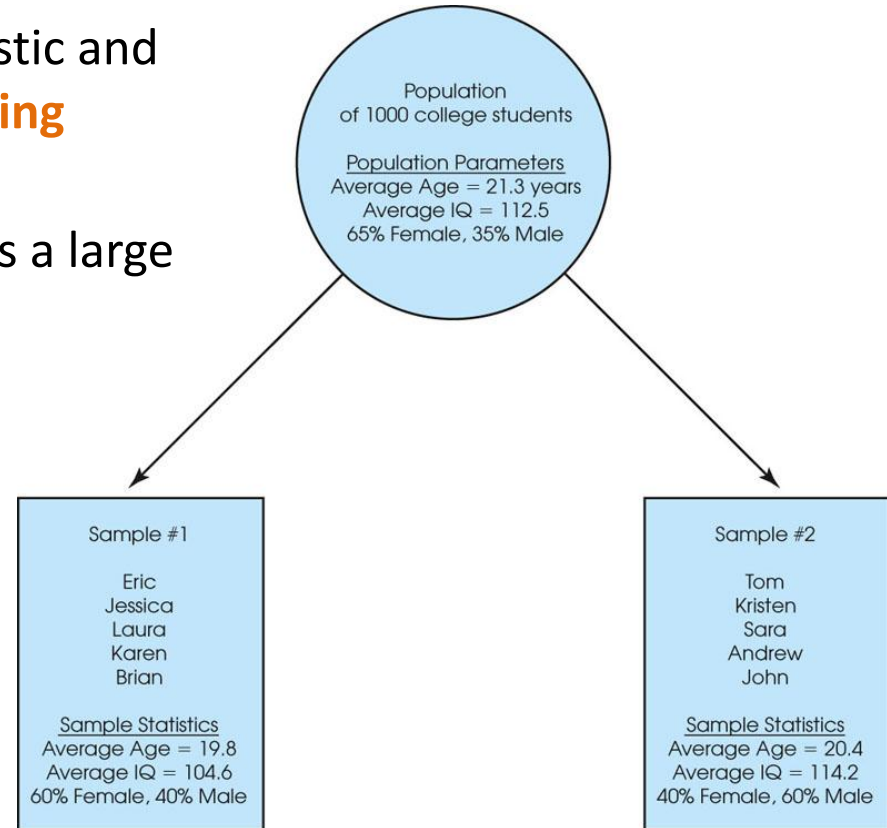
Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

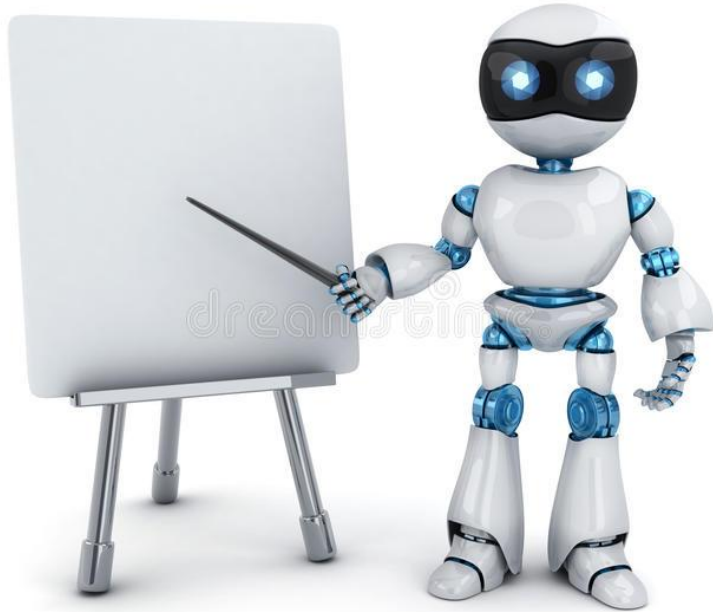
Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics

**Figure 1.2**

A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called **sampling error**.





# Exploratory Analysis

## Module - 3

# Measures of Central Tendencies

The property of data being concentrated in the centre.

- Mean
- Median
- Mode



The mean is the average of all numbers and is sometimes called the arithmetic mean.

The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median

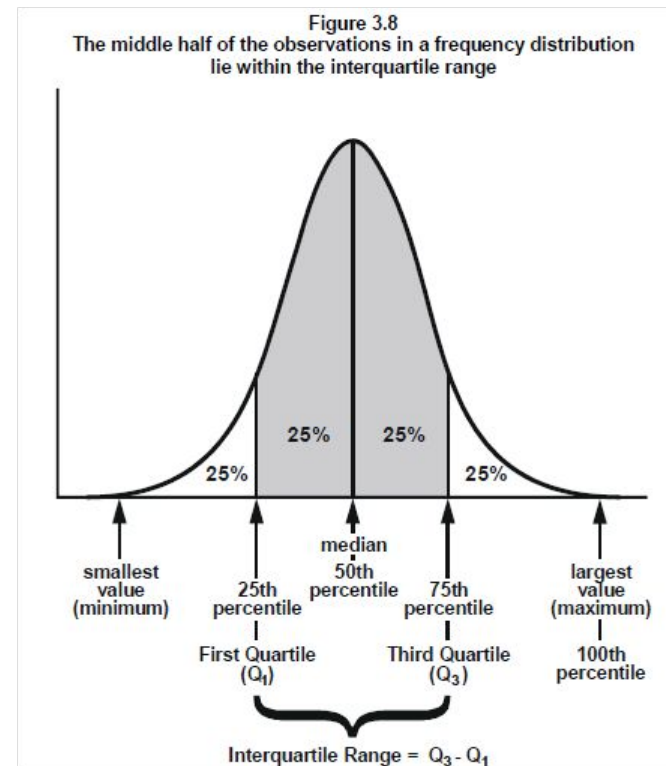
The mode is the number that occurs most often within a set of numbers.

# Measure of Spread / Data Variability

The range is the difference between the highest and lowest values within a set of numbers.

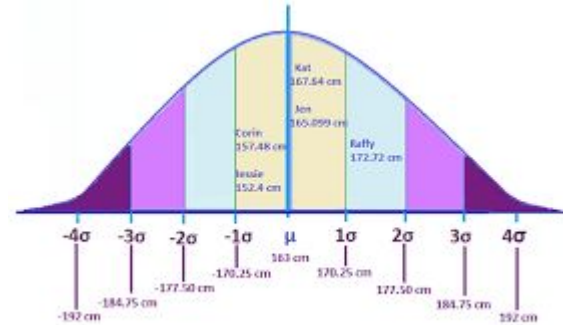
Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

The interquartile range is the middle half of the data.  
Mathematically the interquartile range includes the 50% of data points that fall between  $Q_1$  and  $Q_3$ .



# Standard Deviation ( $\sigma$ )

Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.



# Standard Deviation ( $\sigma$ )

## Standard Deviation Formula

Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$ <p>X - The Value in the data distribution <math>\mu</math> - The population Mean N - Total Number of Observations</p>	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$ <p>X - The Value in the data distribution <math>\bar{x}</math> - The Sample Mean n - Total Number of Observations</p>

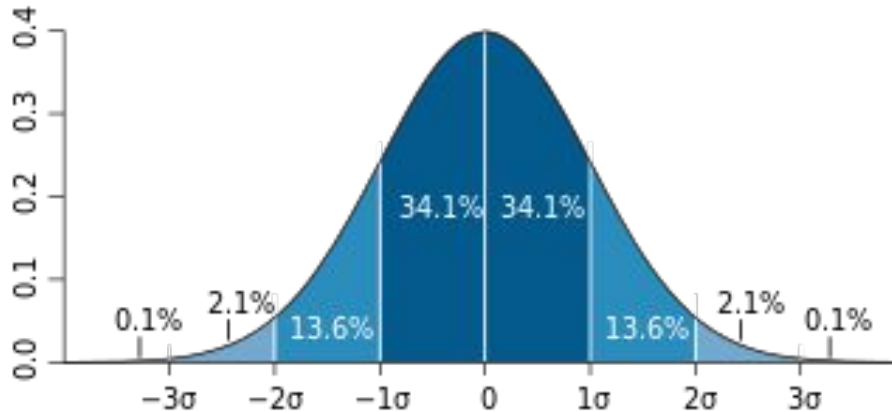


Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

- A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.

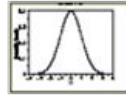


The graphical representation of all observations is known as distribution

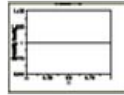


# Types Of Distributions

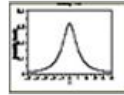
## Continuous Distribution



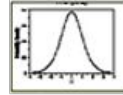
Normal Distribution



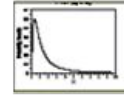
Uniform Distribution



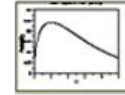
Cauchy Distribution



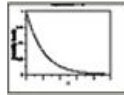
T Distribution



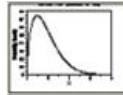
F Distribution



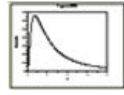
Chi-Square Distribution



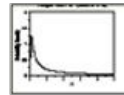
Exponential Distribution



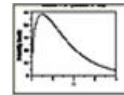
Weibull Distribution



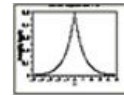
Lognormal Distribution



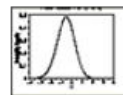
Bimbaum  
Saunders  
(Fatigue Life)  
Distribution



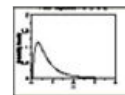
Gamma Distribution



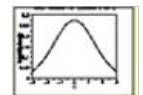
Double  
Exponential  
Distribution



Power Normal  
Distribution

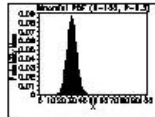


POwer  
Lognormal  
Distribution

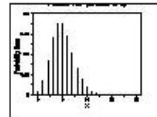


Tukey-Lamba  
Distribution

## Discrete Distribution



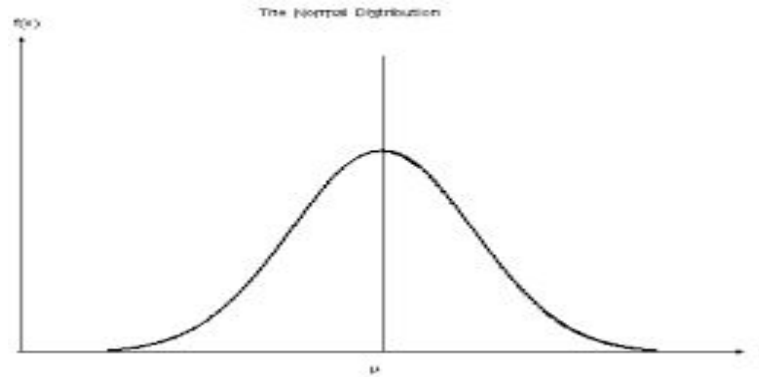
Binomial  
Distribution



Poisson  
Distribution

# Normal Distribution

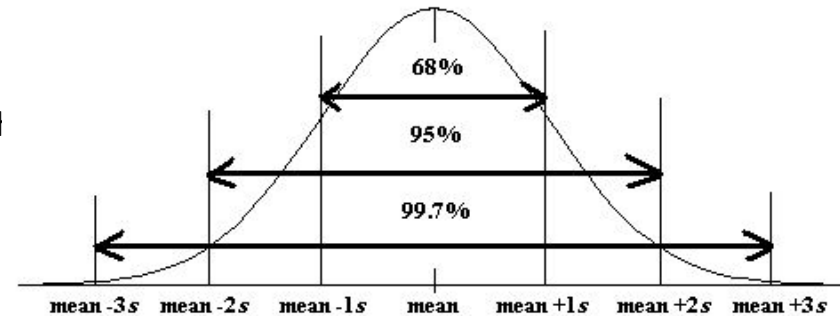
Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



1. Empirical Rule
2. Distortion in Normal Distribution
3. Central Limit Theorem
4. Standard Normal Distribution
5. Outliers
6. QQ plot
7. Log,Sqrt,Boxcox transformation

- The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- 68% of data falls within the first standard deviation from the mean
- 95% fall within two standard deviations. (2 Sigma)
- 99.7% fall within three standard deviations. (3 Sigma)
- Any points lying after 3 sigma are outliers.

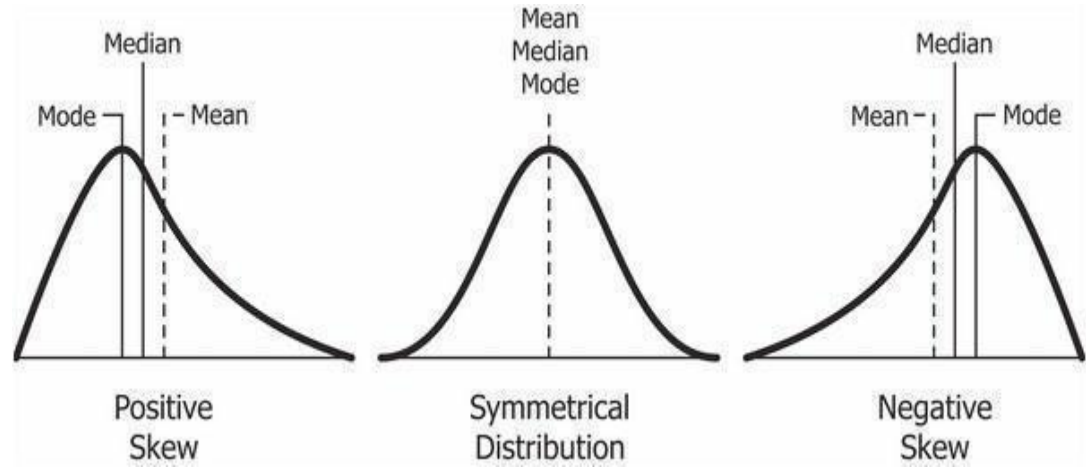


The distortion in normally distributed curves can be quantified in 2 ways

1. Skewness
2. Kurtosis



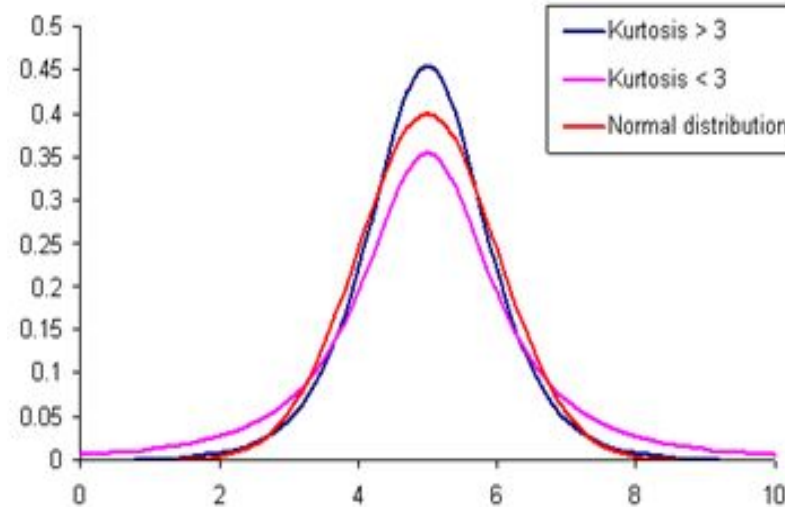
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution

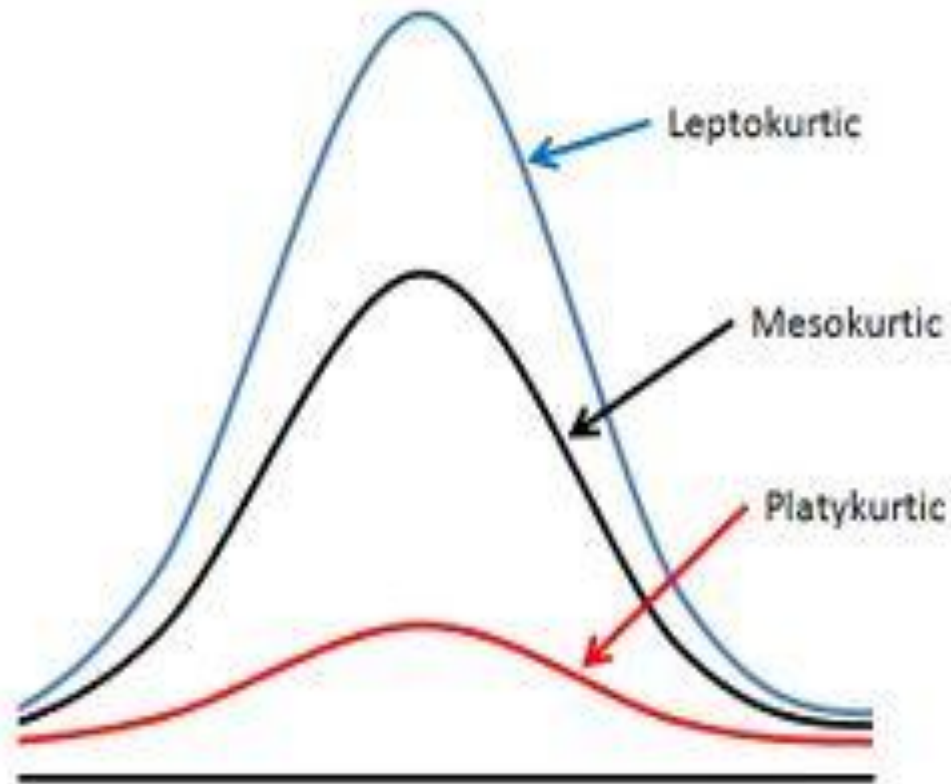


If the skewness is greater than 1 or less than -1, the data is highly skewed.

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and -0.5 or between 0.5 and 1, the data is moderately skewed.
- If the skewness is greater than 1 or less than -1, the data is highly skewed.
- A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis ( $>3$ ) can be visualized as a thin “bell” with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and “thickening” of the tails.

In probability theory and statistics, kurtosis is a measure of the “peakedness” of the probability distribution of a real-valued random variable.

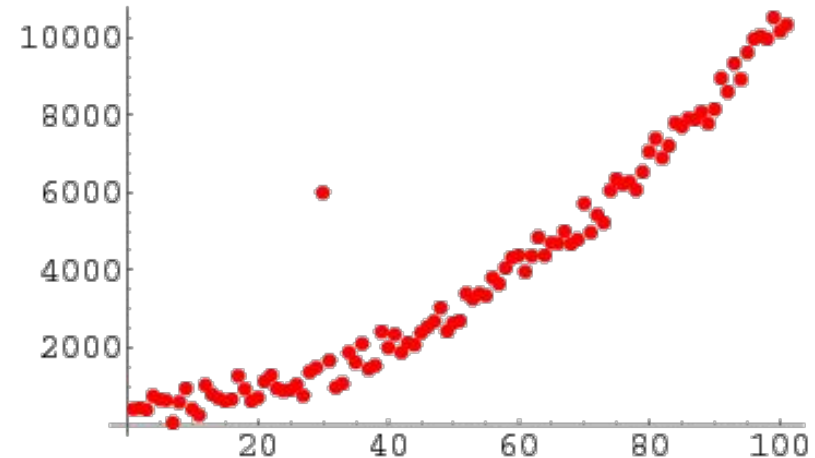




- When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.
- When you have categorical or discrete data, the median or mode is usually the best choice. For categorical data, you have to use the mode.

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

<https://tribe.datamites.com/posts/outliers>



# Central Limit Theorem

The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape.

**CLT in one sentence "Even if I'm not normal, the average is normal"**

When collecting means of the samples from any distribution, the no of samples taken for calculating the mean should be greater or equal to 30.



Probability density is the relationship between observations and their probability.

The overall shape of the probability density is referred to as a probability distribution, and the calculation of probabilities for specific outcomes of a random variable is performed by a probability density function, or PDF for short.

The probability density function for Normal distribution is given as

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

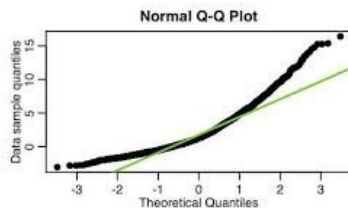
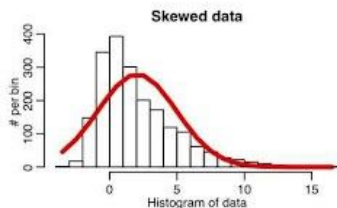
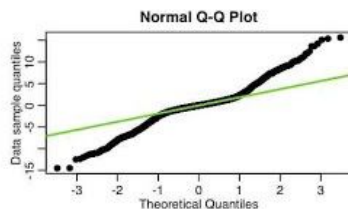
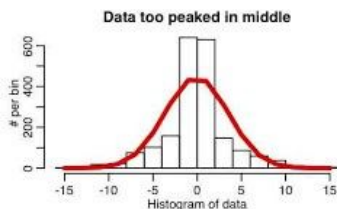
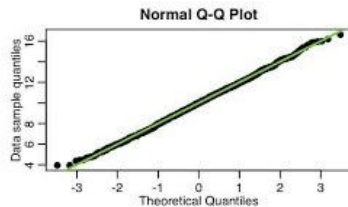
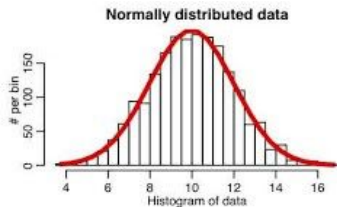
- $\mu$  is the **mean** or **expectation** of the distribution (and also its **median** and **mode**)
- $\sigma$  is the **standard deviation**
- $\sigma^2$  is the **variance**

CDF: It provides a shortcut for calculating many probabilities at once. We integrate the **pdf** function to get the cumulative probability.



# Q-Q Plots

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution / Normal distribution or not. It plots the quantiles of your dataset against the quantiles of the theoretical distribution.



# Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.



- A z-score (aka, a standard score) indicates how many standard deviations an element is above or below from the mean. A z-score can be calculated from the following formula.
- $z = (X - \mu) / \sigma$

# Calculating Z-Score

Formula:

$$z\text{-score} = \frac{x_i - \bar{x}}{s}$$

---

$x_i$  = data point

$\bar{x}$  = mean

$s$  = standard deviation

---

Example:

$$z = \frac{231 - 130.1}{47.85} = 2.11$$

$$z = \frac{50 - 130.1}{47.85} = -1.67$$

- Z-scores generally range from -3.0 to +3.0.
- For bell shaped distributions, the empirical rule says 99.7% of all the data values have z-scores between -3.0 and +3.0.
- We consider any z-score that is either less than -3.0 or greater than +3.0 to be an **outlier**.

Transformation is nothing but taking a mathematical function and applying it to the data.

- Log Transformation [Each data point is replaced with  $\log(x)$  to obtain ND]
- Square-Root Transformation [Each data point is replaced by its square root]
- Reciprocal Transformation [It takes the inverse of  $x$  ie.,  $1/x$ ]
- BoxCox Transformation [flexible version of log or square root transformations — but instead of choosing manually, Box-Cox automatically finds the best value of  $\lambda$  (lambda) that makes your data as normal as possible.]

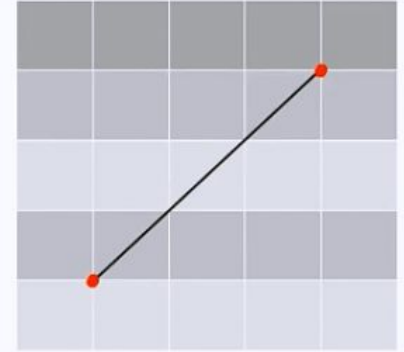
**Reason:** To transform the data to either reduce the skewness or to normalize the data or simply make the data easier to understand.

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance

# Euclidean Distance

It is a classical method to calculate the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space). This distance can be calculated by traveling along the line, connecting the points.

## Euclidean Distance



You can use the Pythagorean Theorem to compute this distance:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



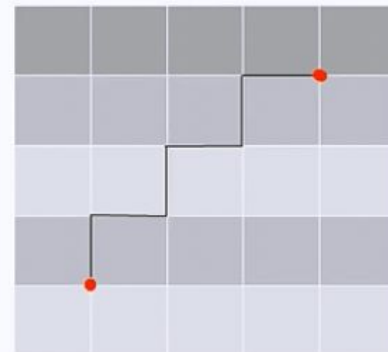
# Manhattan Distance

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

You can use the following formula to compute this distance:

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$

**Manhattan Distance**



It is a metric on the Euclidean space and can be considered as a generalization of both the Euclidean and Manhattan distances.

**You can use the following formula to compute this distance:**

$$D(x, y) = (|x_1 - y_1|^p + |x_2 - y_2|^p)^{\frac{1}{p}}$$

Where:

- $(x_1, x_2)$  and  $(y_1, y_2)$  are the coordinates of points  $x$  and  $y$ .
- $p$  is the order of the distance (like 1 for Manhattan, 2 for Euclidean, etc.).

For example:

- If  $p = 1$  (Manhattan distance):

$$D(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

- If  $p = 2$  (Euclidean distance):

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- If  $p = 3$ :

$$D(x, y) = (|x_1 - y_1|^3 + |x_2 - y_2|^3)^{\frac{1}{3}}$$

Statistical distances are used for several important reasons:

- Outlier Detection
- Clustering Analysis
- Feature Selection
- Data Cleaning
- Dimensionality Reduction

- It is the relationship between a pair of random variables where change in one variable causes change in another variable.
- It can take any value between  $-\infty$  to  $+\infty$ , where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

For Population:

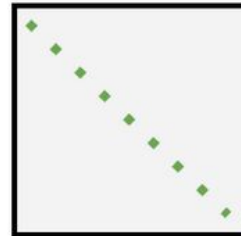
$$Covari(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{n}$$

For Sample

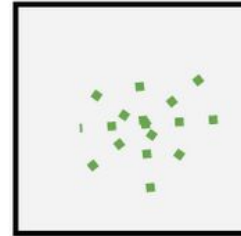
$$Covari(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{n - 1}$$

Here,  
 $x'$  and  $y'$  = mean of given sample set  
 $n$  = total no of sample  
 $x_i$  and  $y_i$  = individual sample of set

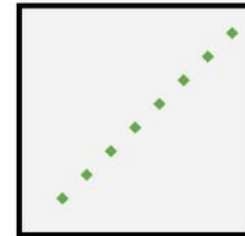
## COVARIANCE



Large Negative  
Covariance



Nearly Zero  
Covariance



Large Positive  
Covariance

- It is the scaled version of Covariance.
- Correlation is a step ahead of covariance as it quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change concerning each other (normalized covariance value)

Formula –

$$\text{Corr}(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{\sqrt{\sum_{i=1}^n (x_i - x')^2 \sum_{i=1}^n (y_i - y')^2}}$$

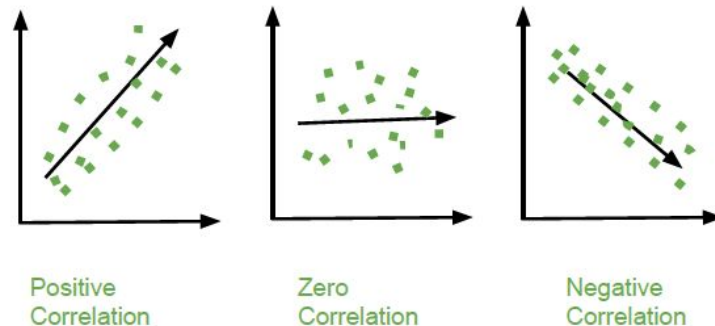
Here,

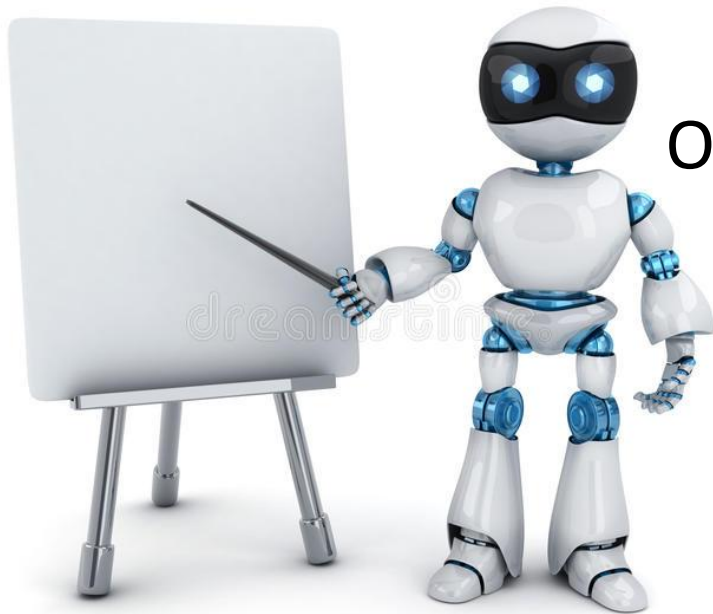
$x'$  and  $y'$  = mean of given sample set

$n$  = total no of sample

$x_i$  and  $y_i$  = individual sample of set

## CORRELATION





# Hypothesis Testing & Other computational Techniques

## Module - 4

- Hypothesis is a statement, assumption or claim about the value of the parameter (mean, variance, median etc.).
- A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.
- Hypothesis Testing is a formal procedure to test whether an assumption (**hypothesis**) about a population is true or not.

Ex:-if we make a statement that “Dhoni is the best Indian Captain ever.” This is an assumption that we are making based on the average wins and losses team had under his captaincy. We can test this statement based on all the match data.

- When a hypothesis specifies an exact value of parameter, it is simple hypothesis. For eg., A motorcycle company claiming that a certain model gives an average mileage of 100 km per litre, this is a case of simple hypothesis.
- If a hypothesis specifies a range of values then it is called a composite hypothesis. For eg., Average age of students in a class is greater than 20. This statement is a composite hypothesis.



- The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true.
- Represents the default assumption.
- The concept of the null is similar to innocent until proven guilty.

- Represents the opposite of the null hypothesis such that both the alternate and null hypotheses together cover all possible values of the population parameter.
- It is the hypothesis that researchers aim to support or prove based on evidence from sample data.
- The process in hypothesis testing involves gathering evidence against the null hypothesis to support the alternative hypothesis.

## **Consider a court of law:**

The null hypothesis always begins with the assumption that the defendant is innocent.

- We require evidence to reject the null hypothesis (convict the defendant).
- When we collect evidence and try to reject null hypothesis, there are 2 errors that could potentially occur: Type 1 or Type 2 error.

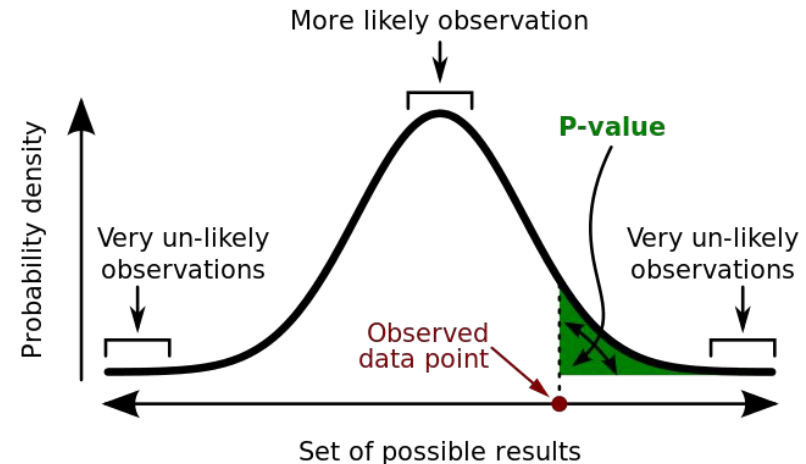
# Type I and Type II Error

H0: Person is innocent

H1: Person is not innocent

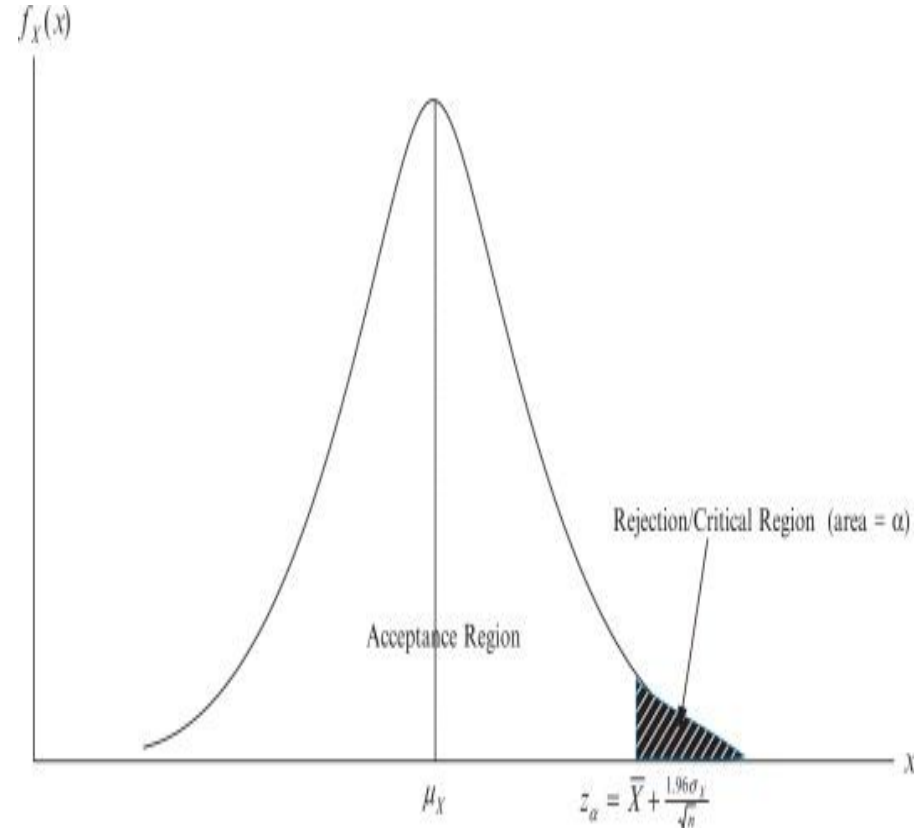
Decision	H0 True	H0 False
Reject H0	Type I error	Correct Decision
Do not reject H0	Correct Decision	Type II error

- Since we know that passing value of a test can be 1% or 5%. But we must also know what are the test score and this test score is known as p- value.
- Technically p-value (probability value) is the smallest level of significance at which a null hypothesis can be rejected.
- If p-value is greater than alpha, we do not reject the null hypothesis.
- If p-value is smaller than alpha, we reject the null hypothesis.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

- The critical region is that region in the sample space in which if the calculated value lies then we reject the null hypothesis.
- The critical region lies in one tail or two tails on the probability distribution curve according to the alternative hypothesis.
- The value of critical region is denoted by  $\alpha$ .
- It is known as level of significance. i.e what is passing criteria of test.



# Cases Of Critical Region

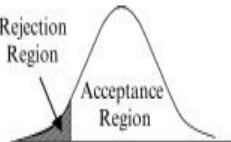


- If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in null hypothesis, it is called Two-tailed test.

E.g.  $H_0$ : mean = 100  $\rightarrow$   $H_1$ : mean not equal to 100

Here according to  $H_1$ , mean can be greater than or less than 100.

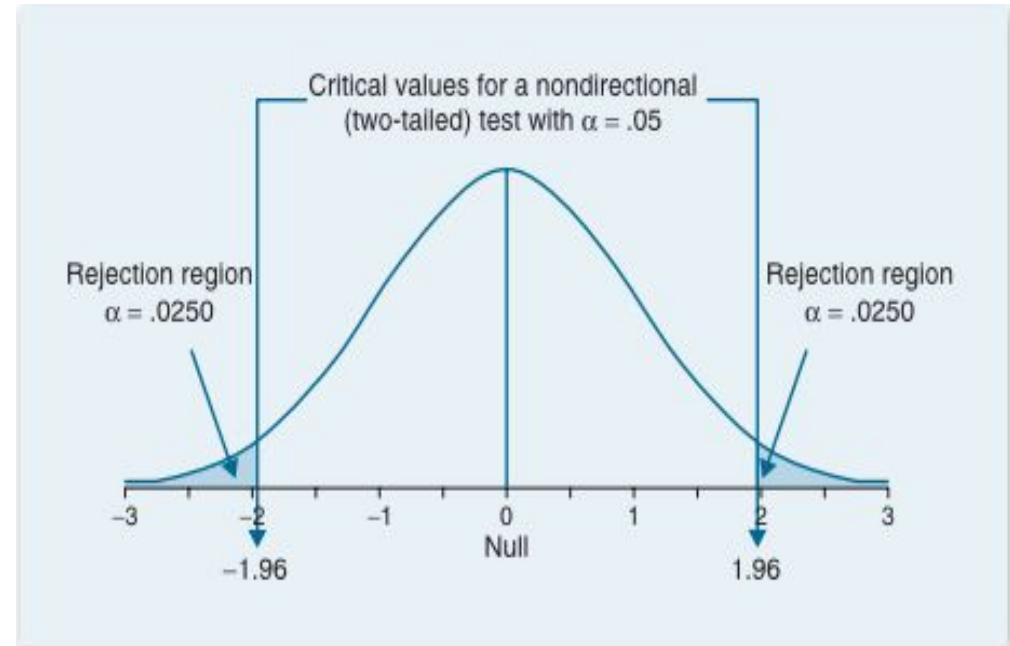
- If the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in null hypothesis, it is called One-tailed test.

E.g.  $H_0$ : mean  $\geq$  100  $\rightarrow$   $H_1$ : mean < 100

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

**Three cases of Critical Region arise.**

1. Set up Null hypothesis and Alternate hypothesis.
2. Decide the level of significance (1% or 5%).
3. Select the test as per requirement.
4. Calculate the p-value.
5. If p-value less than level of significance, reject the null hypothesis.
6. If p-value more than level of significance, accept the null hypothesis.





- **Parametric Tests:-**Those test which considers the shape of distribution of sample.
- **Non – Parametric Tests:-**Those test which do not considers the shape of distribution of sample.

- Feature Selection
- Model Evaluation & Selection
- Hyper Parameter Tuning
- Anomaly Detection

1. Z test
2. T/Student's T test
3. Paired t Test
4. One Way ANOVA

1. Chi Square Test
2. Mann-Whitney Test
3. Wilcoxon Signed-Rank Test
4. Kruskal-Wallis Test
5. Friedman's ANOVA

- A t-test is an analysis of two populations means through the use of statistical examination; a t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known.
- This helps in finding the association between Categorical and Continuous features.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

$\bar{X}$  -> Mean of sample set

$\mu$  -> Mean of Population

$s$  -> Standard deviation of sample

$N$  -> Sample size

S.No	Experience	Gender	Role	Salary
1	4.5	Female	Sr.Executive	10,00,000
2	9	Male	Manager	12,58,000
3	3.6	Male	Sr.Executive	9,08,000
4	8	Male	Sr.Executive	8,40,000
5	1	Female	Executive	5,00,000
6	12	Female	Sr.Manager	15,00,000
7	4	Male	Sr.Executive	8,00,000
8	5.5	Female	Manager	12,50,000
9	2.3	Male	Executive	5,50,000
10	1.9	Male	Executive	6,30,000
11	9.5	Female	Manager	11,00,000
12	14	Male	Sr.Manager	16,30,000
13	3	Male	Executive	6,00,000

Gender	Salary	Female	Male
Female	10,00,000	5,00,000	5,50,000
Male	12,58,000	10,00,000	6,00,000
Male	9,08,000	11,00,000	6,30,000
Male	8,40,000	12,50,000	8,00,000
Female	5,00,000	15,00,000	8,40,000
Female	15,00,000		9,08,000
Male	8,00,000		12,58,000
Female	12,50,000		16,30,000
Male	5,50,000		
Male	6,30,000	<b>10,70,000</b>	<b>9,02,000</b>
Female	11,00,000		
Male	16,30,000		
Male	6,00,000		

The T-Test helps to determine whether there is a statistically significant difference in the average salaries between these two groups.

H0: There is no significant difference between two groups.

$$T = \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The paired t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures.

For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.

Blood Pressure	Before	140	135	150	155	160	145	148	165	152	158
	After	130	125	135	140	145	135	140	150	138	142



# One-Way ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. Eg.,

	Executive	Sr.Executive	Manager	Sr.Manager
	5,00,000	8,00,000	11,00,000	15,00,000
	5,50,000	8,40,000	12,50,000	16,30,000
	6,00,000	9,08,000	12,58,000	
	6,30,000	10,00,000		
Mean	5,70,000	8,87,000	12,02,667	15,65,000

Within Group Variance - Values within the group are close to each other

Between Group Variance - Values between groups are not close to each other

- If the within-group variation is less and between-group variation is high, then it means this feature impacts the target variable. Hence it is an important feature.

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population. Chi- square score is given by

$$\chi^2 = \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

Observed frequency = No. of observations of class

Expected frequency = No. of expected observations of class if there was no relationship between the feature and the target. Expected Frequency = (Row Total \* Column Total)/N

Gender	Role
Female	Sr.Executive
Male	Manager
Male	Sr.Executive
Male	Sr.Executive
Female	Executive
Female	Sr.Manager
Male	Sr.Executive
Female	Manager
Male	Executive
Male	Executive
Female	Manager
Male	Sr.Manager
Male	Executive

### Contingency Table:

	Executive	Sr.Executive	Manager	Sr.Manager
Female	1	1	2	1
Male	3	3	1	1

The chi-squared value for the given contingency table is 0.257

Significance level is 0.05

The observed p-value is 0.967902

Since  $p\text{-value} > 0.05$ , we accept the Null hypothesis.

# Thank You