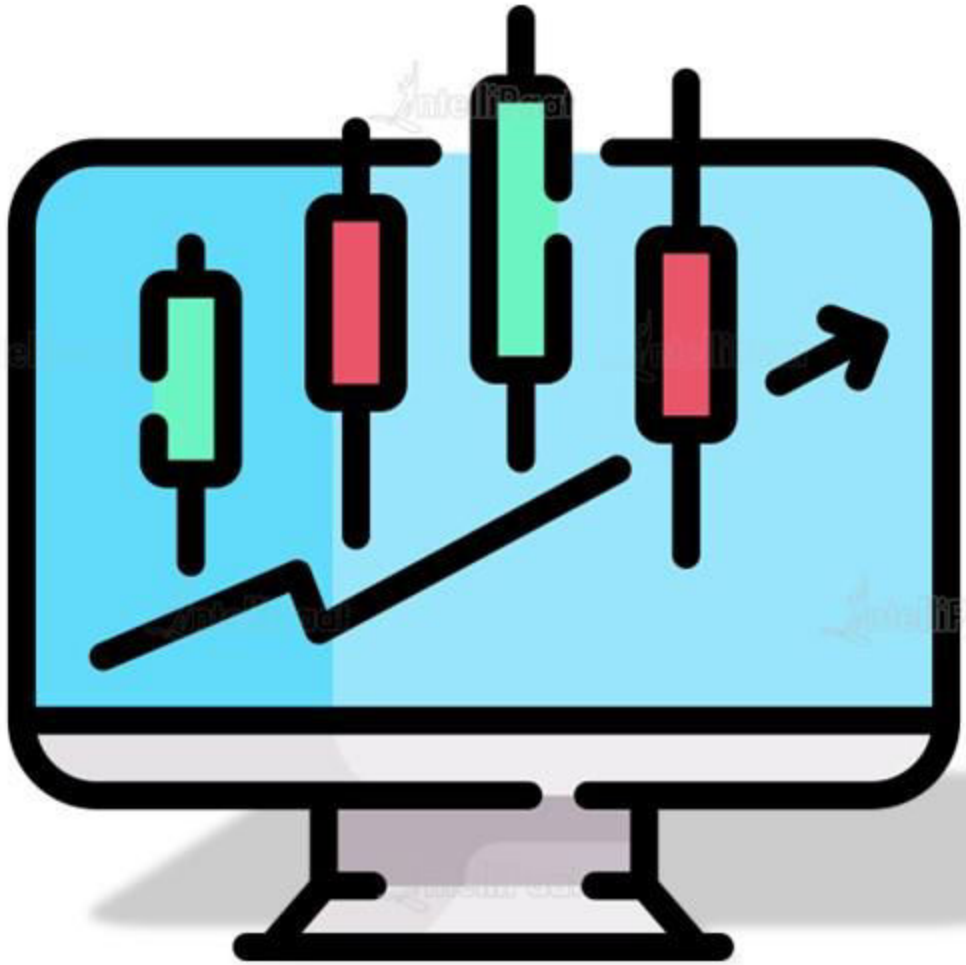
Introduction to Statistics



AGENDA

01 Intro to Statistics

02 Sampling

03 Central Tendencies

04 Variation

05 Correlation

06 Normal Distribution

07 Empirical Rule

08 Z Scores

09 Linear Regression

What is Statistics?

What is Statistics?

What?

Statistics is a branch of Mathematics that deals with collection, analyzing, and interpreting large amounts of data.

Why is Statistics important?

Why is Statistics important?

Statistics allows us to derive knowledge from large datasets and this knowledge can then be used to make predictions, decisions, classifications etc.

Where is Statistics used?

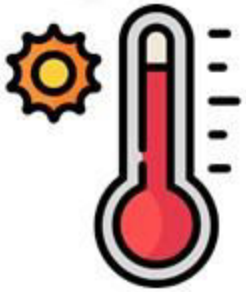
Where is Statistics used?

Statistics are used in various fields, some of them are:

Medical Research

Stock Market

Sales Projection

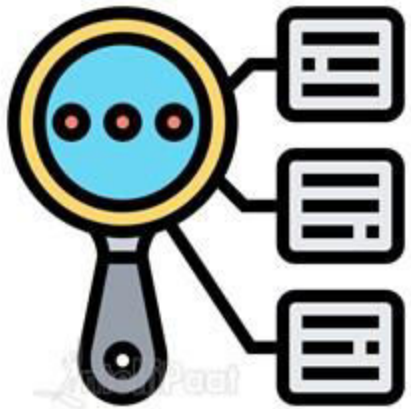


Weather Forecasting

Sampling

Sampling

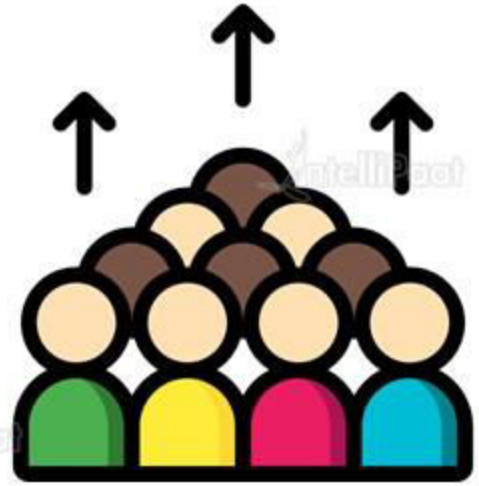
Sampling is the process of collecting data to perform analysis on

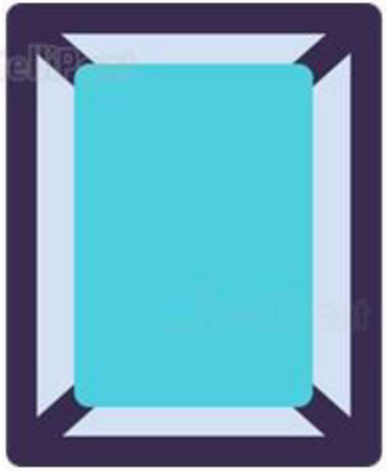


Sample vs Population

Sample vs Population

Population is the entire dataset such as the whole population of a country, **Sample** is subset of that population which is analyzed to make inferences

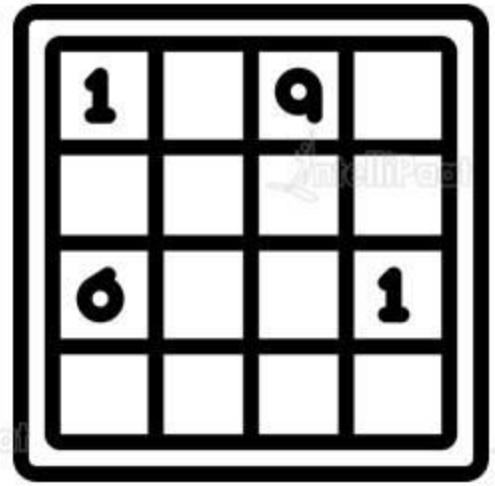




Sample Frame

Sample Frame

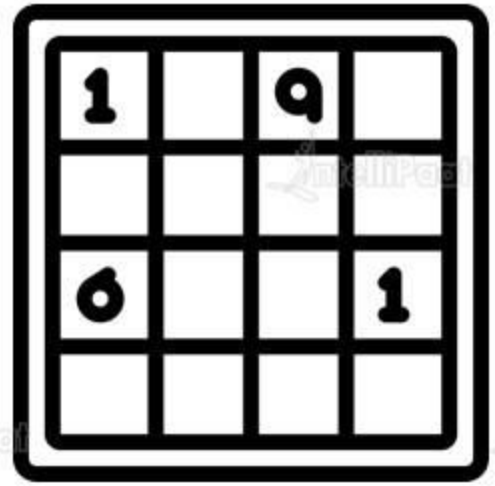
A **Sampling Frame** is a list from which sample is selected, such as a Citizen Register for a country or Employee List for a Company etc.



Sampling Error

Sampling Error

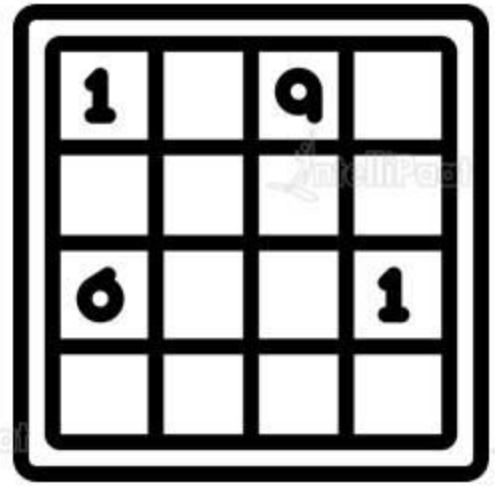
Sampling Error is an error that leads to our sample not accurately representing our population.



Non Sampling Error

Non Sampling Error

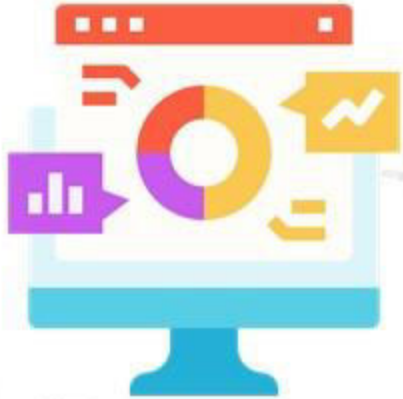
Non Sampling Error occurs dues to poor sample design, inaccurate measurements, bias in data collection etc.



Random Sampling

Random Sampling

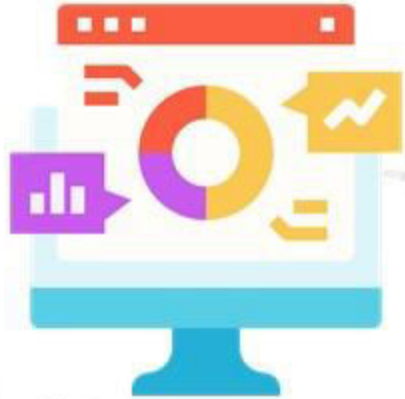
Random Sampling is the process of selecting a subset / sample from a population in such a way that every data point is equally likely to be included in the sample

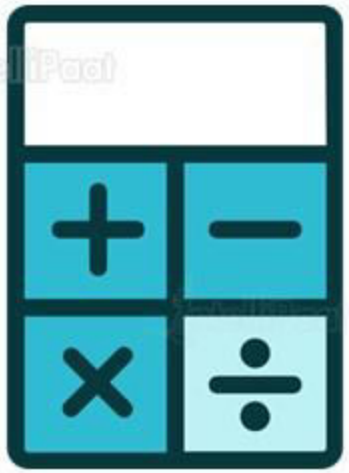


Stratified Sampling

Stratified Sampling

Stratified Sampling is the process of dividing your samples into layers or groups and then performing random sampling for each group

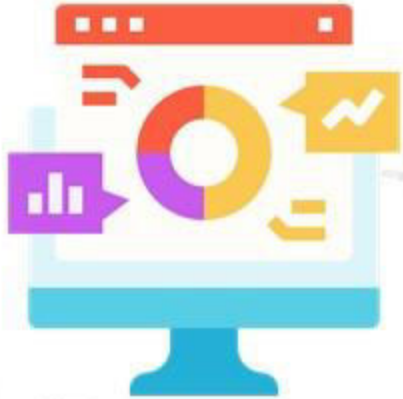


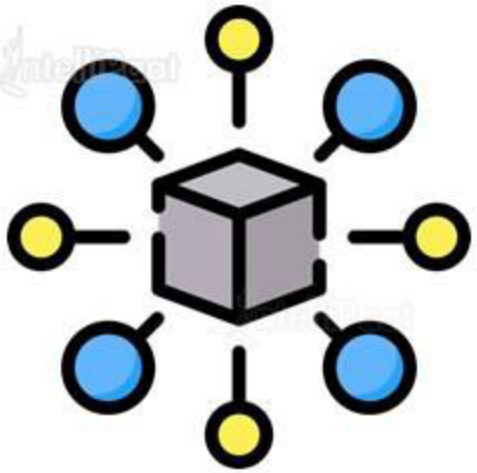


Systematic Sampling

Systematic Sampling

Systematic Sampling is the process of selecting your sample by picking every Kth element in your population. You don't need a list for this.





Central Tendencies

Central Tendencies

Central Tendency is used to indicate where does the middle or center of the distribution of our data lies

Mean

Mode

Median

Mode

Mode

Mode is used to indicate the most frequent data point, in other words the one which occurs most number of times

Mean

Mode

Median

Median

Median

Median is the middle of the data. If the data is arranged in ascending order then the data element which occurs right at the center is the median

Mean

Mode

Median

Mean

Mean

Mean is the average of the data. In simpler terms it's the sum of values divided by total number of values. It's represented by Greek letter Sigma

Trimmed Mean

Weighted Mean

Trimmed Mean

Trimmed Mean

Trimmed Mean is used to deal with outliers by trimming or removing some data from both ends so as to get rid of outliers

Trimmed Mean

Weighted Mean

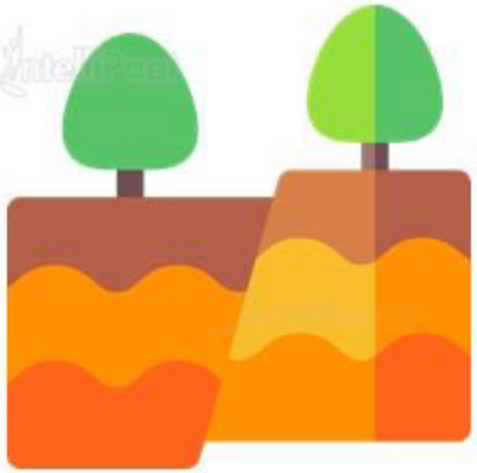
Weighted Mean

Weighted Mean

Weighted Mean is used when certain values are supposed to count more in some context. E.g.: Calculating average grade of a student based on their grade distribution

Trimmed Mean

Weighted Mean



Variation

Variation

Variation in statistics is used to show how data is dispersed, or spread out. Several measures of variation are used in statistics.

Range

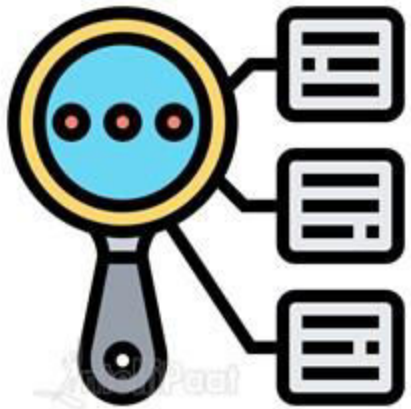
Quartiles

Variance

Range

Range

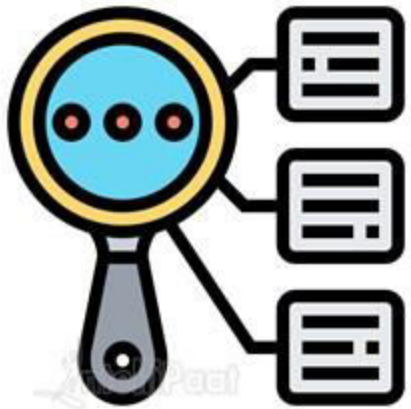
Range is the difference between the highest and the lowest values in our dataset. Range tells us the distance between the lowest and highest values in our data

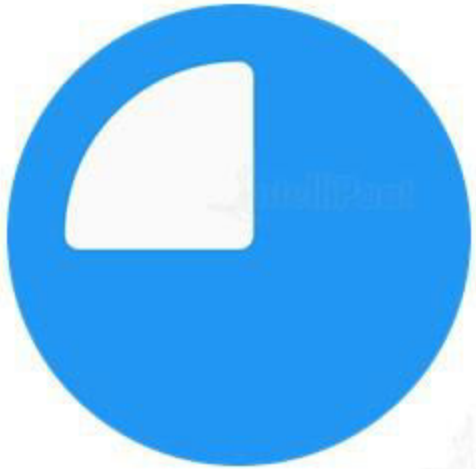


Percentiles

Percentiles

Percentiles are scores that are used to describe a value below which some Observations fall.
E.g.: If X is at 70th Percentile it mean 70% of other data points from our sample are below X

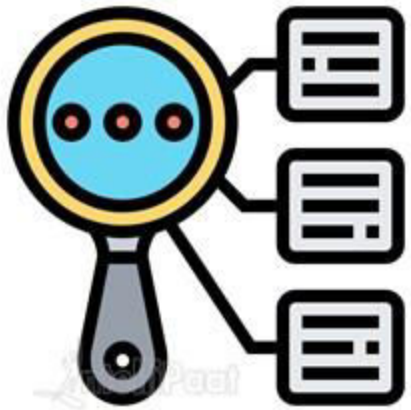




Quartiles

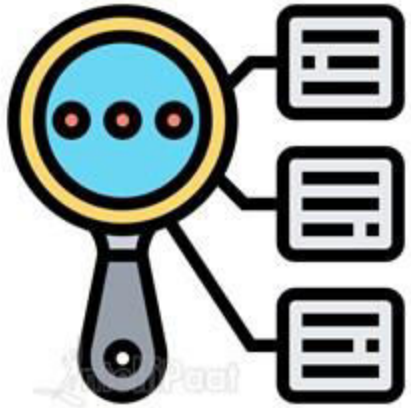
Quartiles

Quartiles are used to break the data into 4 parts so as to better find the spread of data in a way that is less influenced by outliers.



Quartiles

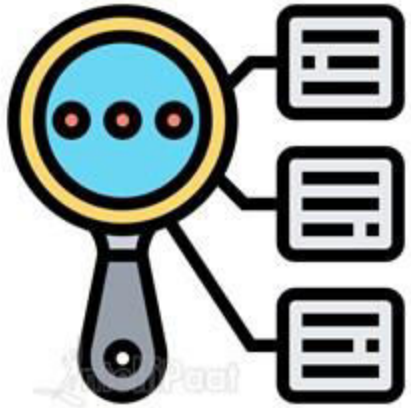
Quartiles are expressed in percentiles. 1st Quartile is 25th Percentile, 2nd Quartile is 50th Percentile (Median) and 3rd Quartile is 75th Percentile



Interquartile Range (IQR)

Interquartile Range (IQR)

Interquartile Range (IQR) is the difference between the lower and upper quartile. This gives us a better idea of the range of data.



Standard Variance and Standard Deviation

Standard Variance and Standard Deviation

Standard Variance measures how far a set of numbers are spread out from their average value.

Standard Deviation is used to express the magnitude by which the members of a group differ from the mean value for the group

Sample Variance

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

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Standard Variance and Standard Deviation

Standard Deviation is the square root of **Standard Variance.**

Sample Variance

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

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Correlation

Correlation

Correlation is a term that is a measure of the strength of a **linear relationship** between **two quantitative variables**

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Positive Correlation

Positive Correlation

Positive Correlation is a term that is used to describe a **positive linear relationship** between **two quantitative variables**

Positive Correlation

No Correlation

Negative Correlation

No Correlation

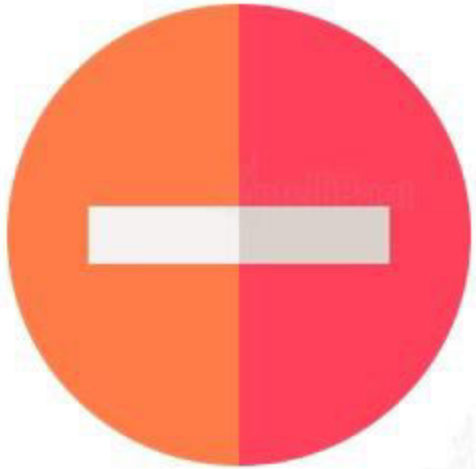
No Correlation

No Correlation is a term used to describe no **linear relationship** between **two quantitative variables**

Positive Correlation

No Correlation

Negative Correlation



Negative Correlation

Negative Correlation

Negative Correlation is a term that is used to describe the strength of a **Negative linear relationship** between **two quantitative variables**

Positive Correlation

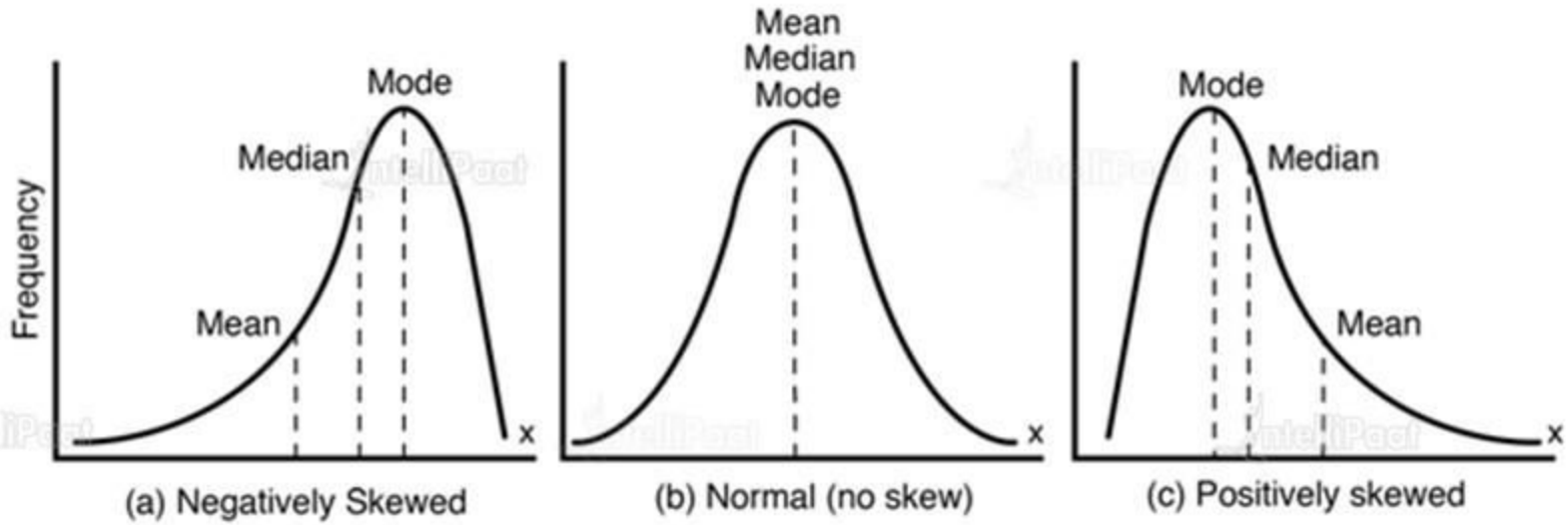
No Correlation

Negative Correlation

Normal Distribution

Normal Distribution

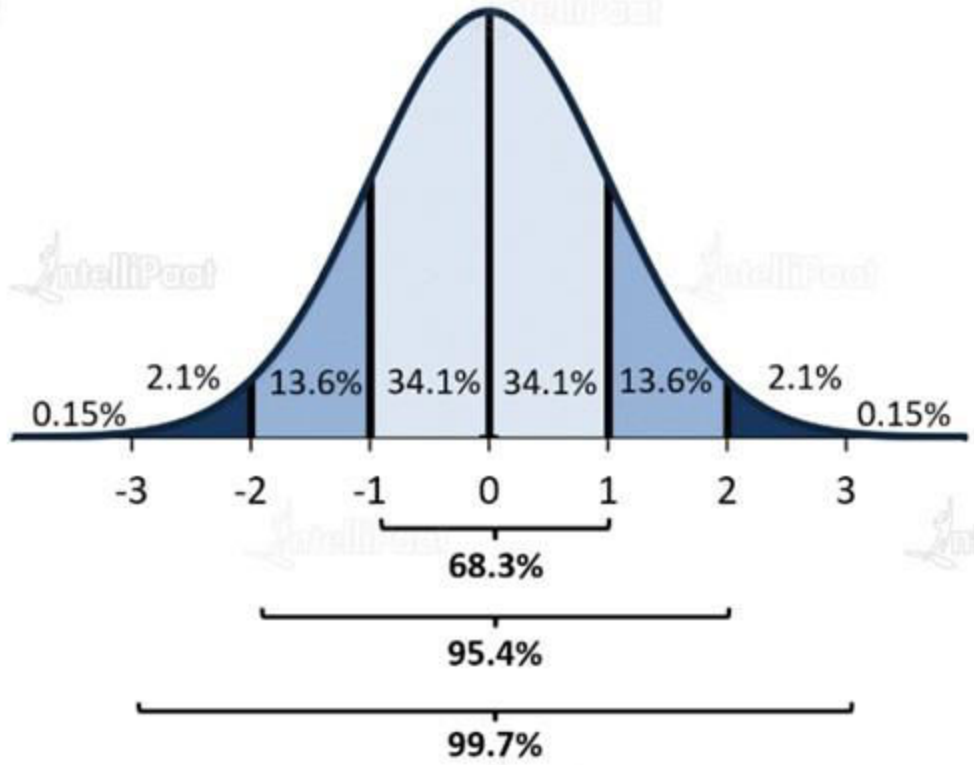
Normal Distribution is a term that is used to describe a distribution which when plotted gives us a shape of bell curve. It has mean of zero and standard deviation of 1



Empirical Rule

Empirical Rule

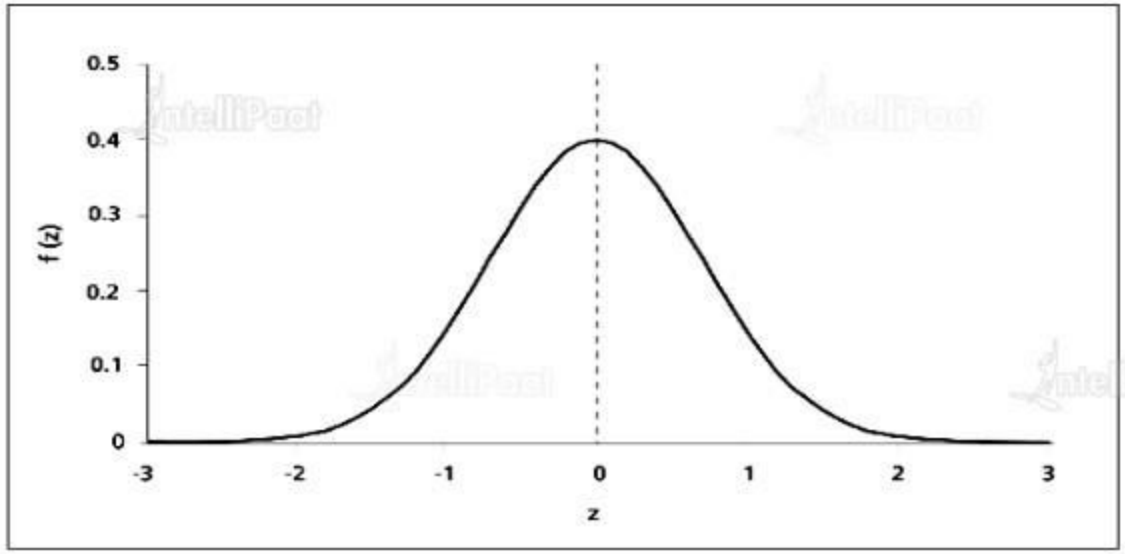
Empirical Rule, is used to remember the **percentage of values** that lie within a band around the mean in a **normal distribution** with a width of **two, four and six standard deviations**



Z Scores

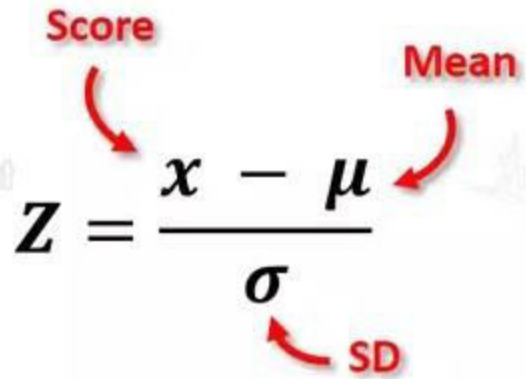
Z Scores

Z-Score is a measure of how many **standard deviations** below or above **the population mean** a raw score. Z-score can be placed on a **normal distribution curve.**



Z Scores

Z-Score is a measure of how many **standard deviations** below or above **the population mean** a raw score. Z-score can be placed on a **normal distribution curve.**

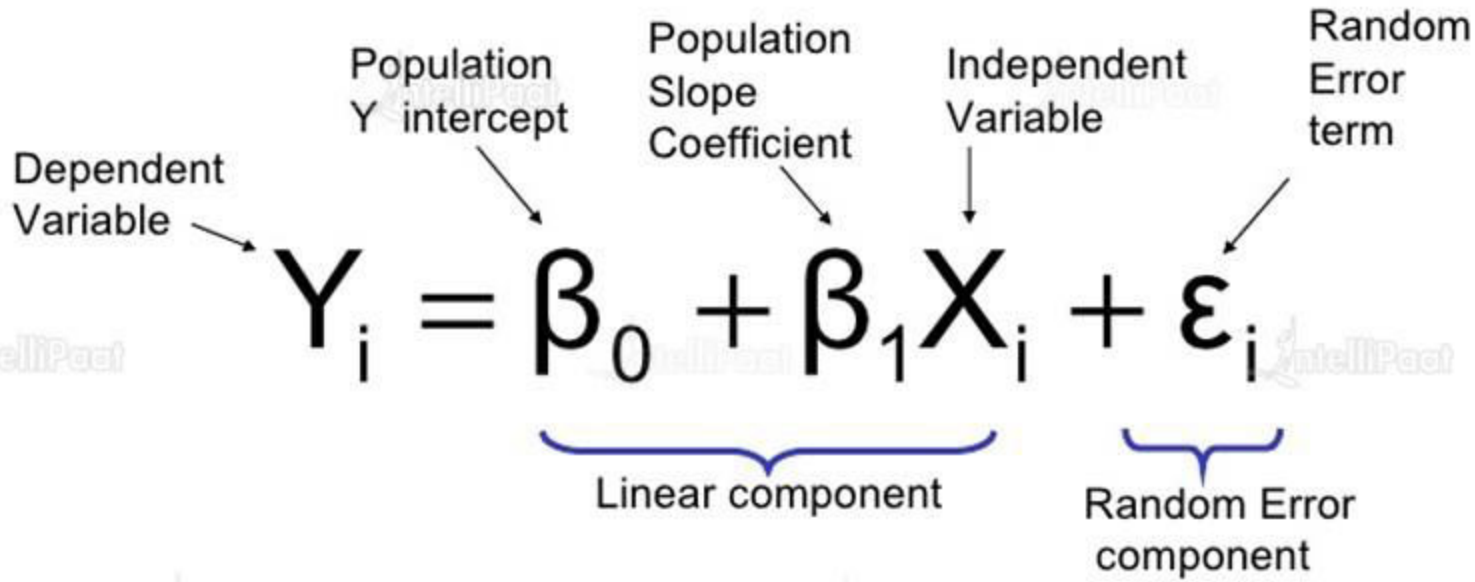


Linear Regression

Linear Regression

Linear regression is a basic and commonly used type of **predictive analysis.**

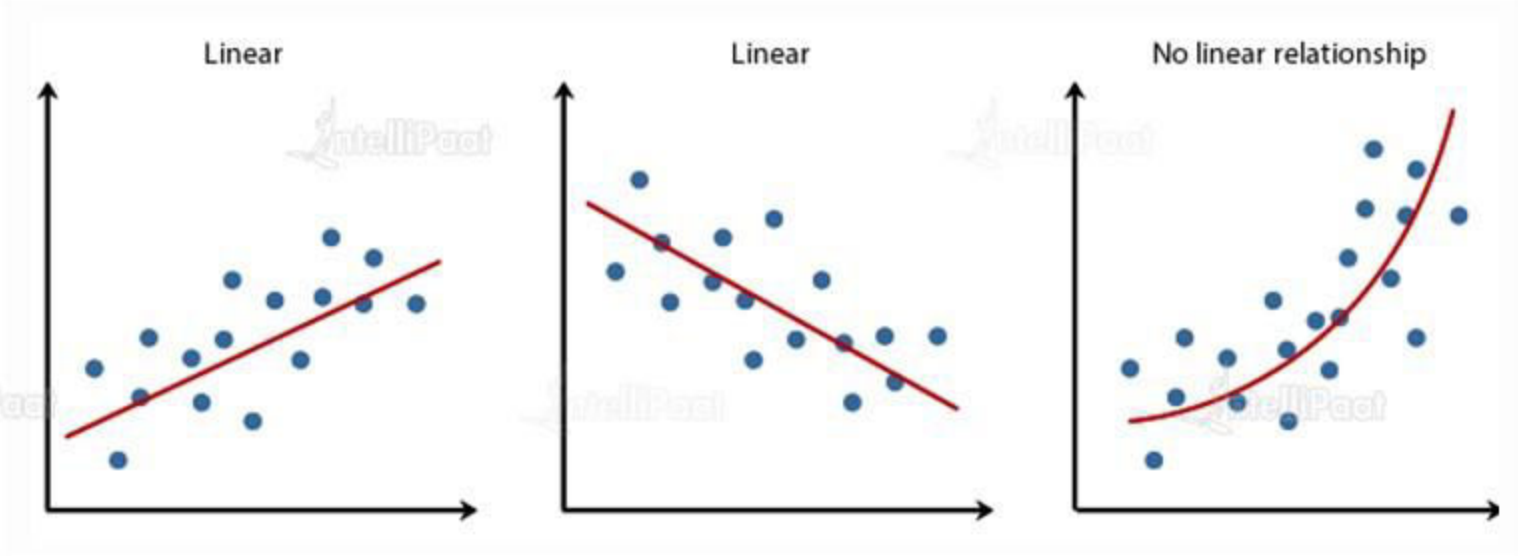
It is used to create a model that can predict a **dependent variable** using an **independent variable**



Linear Regression

Simple linear regression is useful for finding relationship between two continuous variables

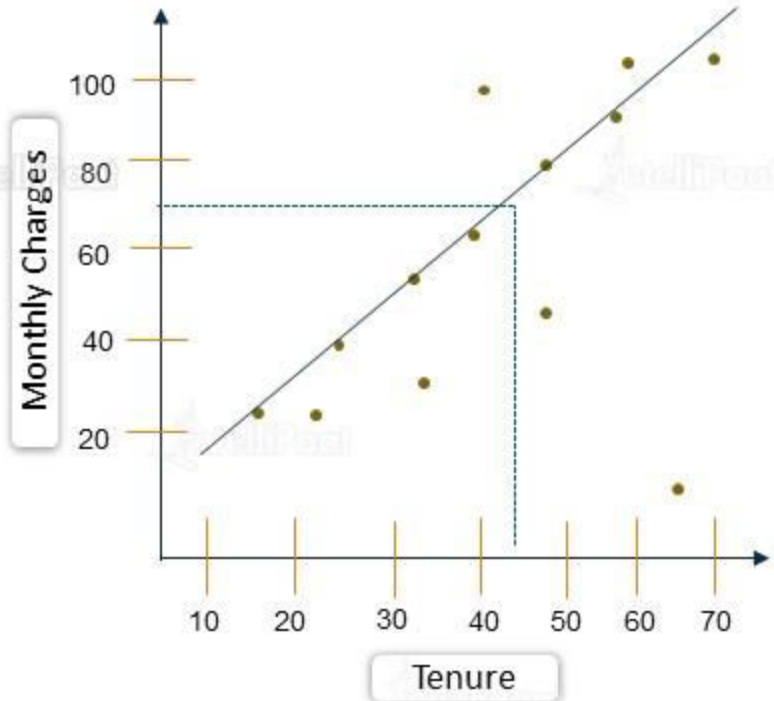
One is predictor or independent variable and the other is the response or dependent variable



Linear Regression

Let us discuss linear regression with an example. We want to know how do the monthly charges of a customer vary with respect to the tenure

Estimating the value of monthly charges with the tenure of the customer



Linear Regression

In general, the data doesn't fall exactly on a line, so the regression equation should include an implicit error term

The fitted values (predicted values) are typically denoted by $\hat{Y}$ (Y-hat)

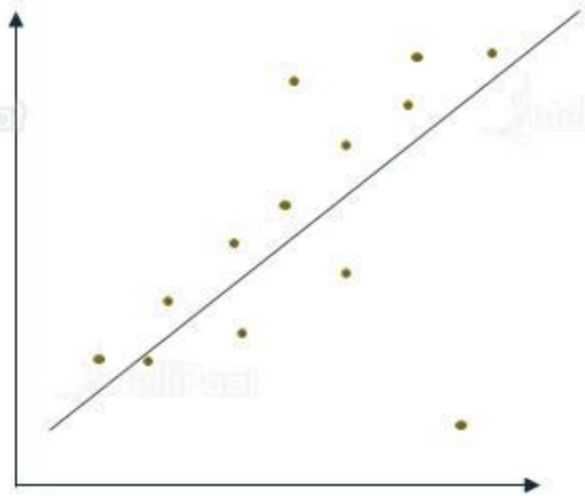
$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

Linear Regression

If b1 > 0, then x(predictor) and y(target) have a positive relationship, i.e., an increase in x will increase y

b1 > 0 ⟹ Positive Relationship

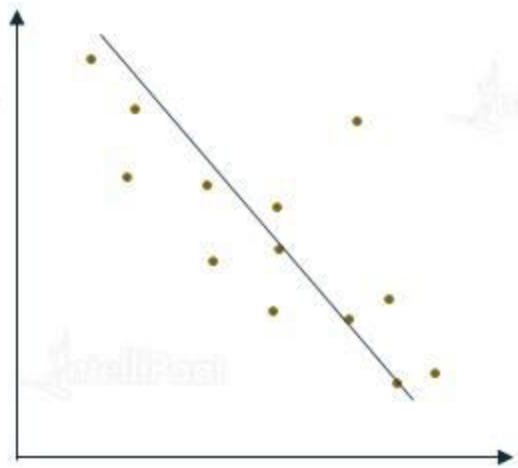


$$y = b_0 + b_1 x$$

Linear Regression

If b1 < 0, then x(predictor) and y(target) have a negative relationship, i.e., an increase in x will decrease y

b1 < 0 ⟹ Negative Relationship

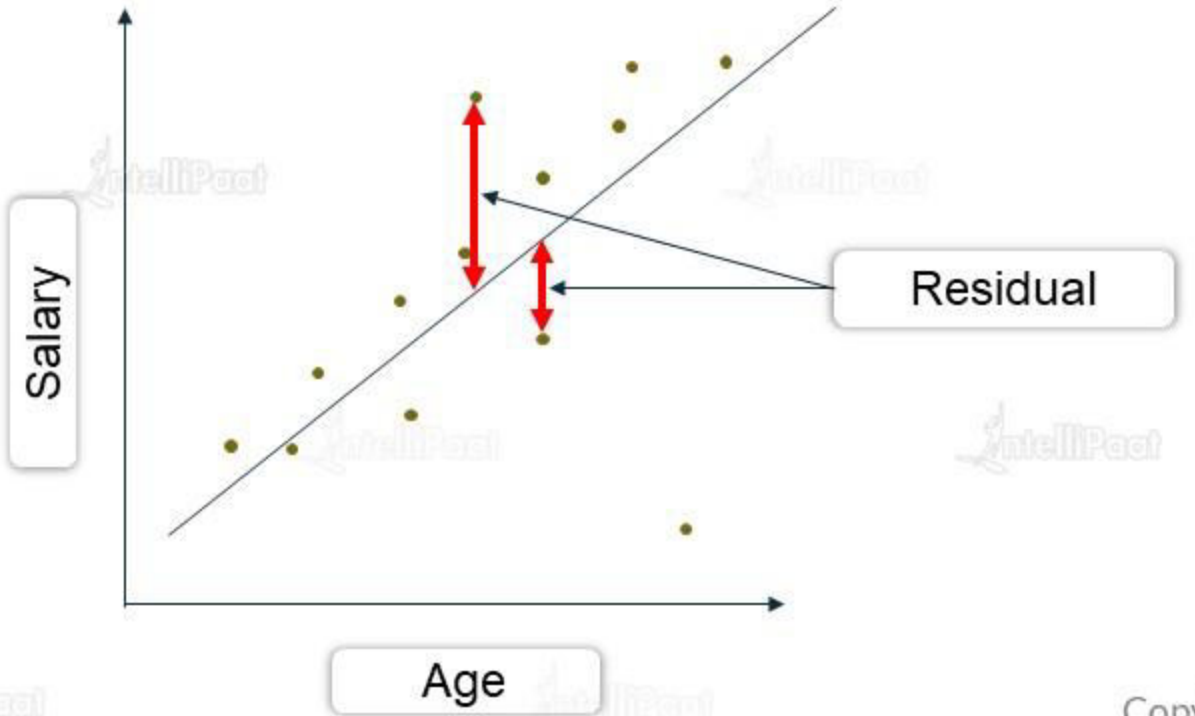


$$y = b_0 + b_1 x$$

Linear Regression

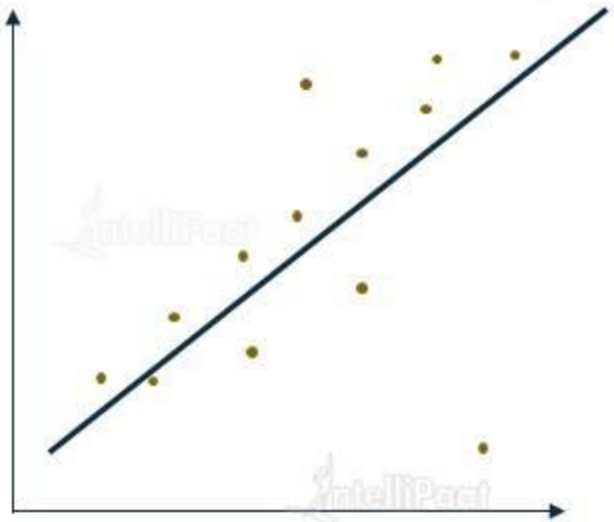
What are residuals?

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called residual. Each data point has one residual

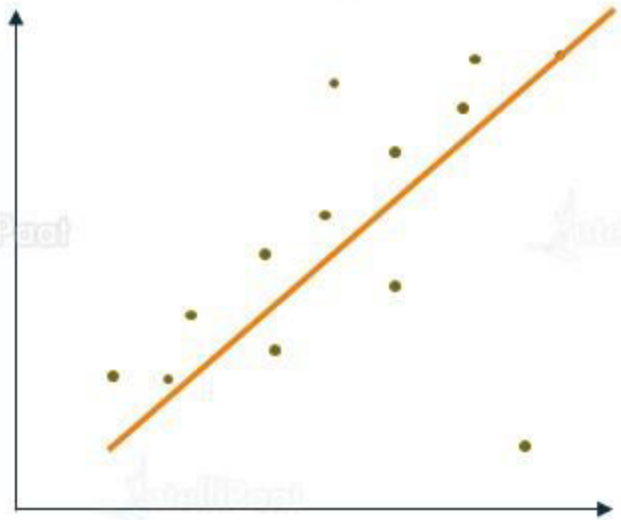


Linear Regression

There could be multiple fit lines passing through the points, so how shall we choose the line of best fit?



Line Fit 1



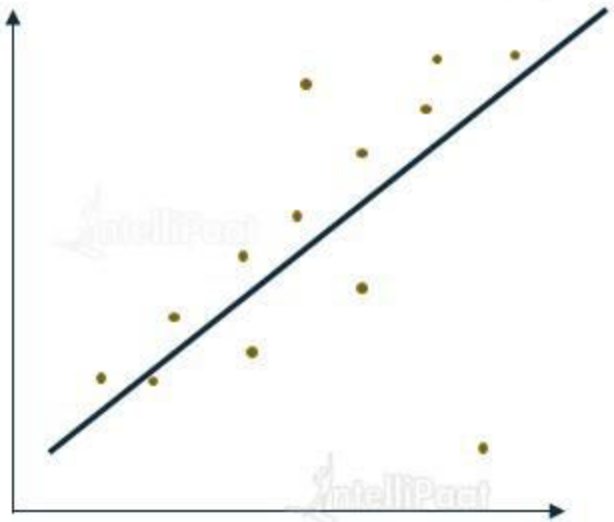
Line Fit 2

Line Fit 3

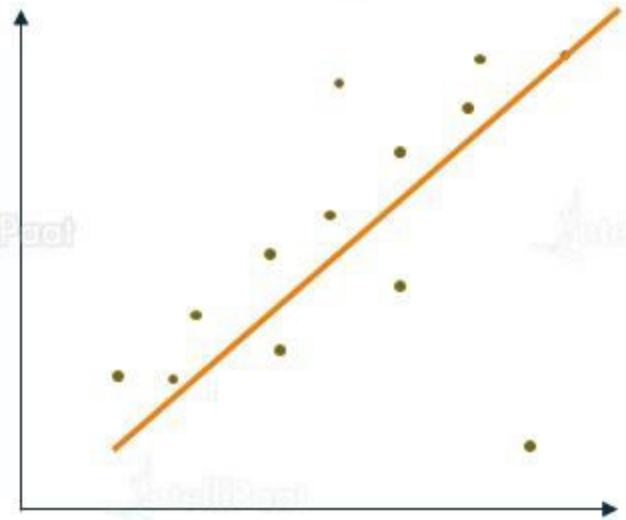
Linear Regression

The line with the lowest value of the residual sum of squares would be the best fit line

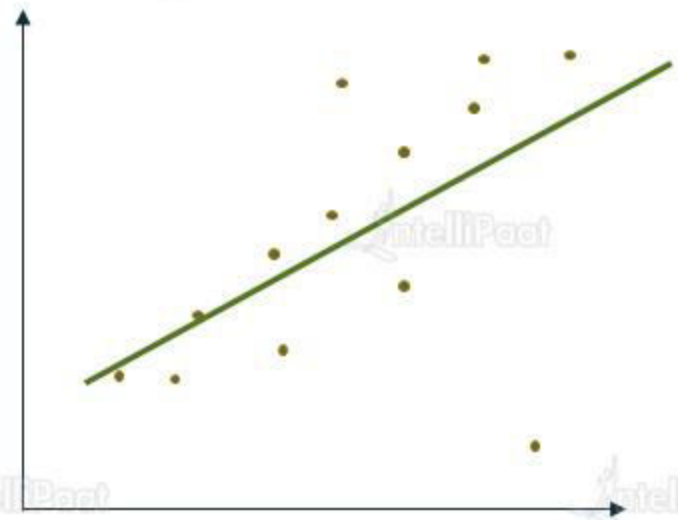
$$\text{RSS} = \sum_{k=1}^{n} (Actual - Predicted)^2$$



RSS = 120



RSS = 80



RSS = 132

India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com

24/7 Chat with Our Course Advisor