

# STATISTICS

## 1.BASIC STATISTICS

---

### 1.1WHAT IS STATISTICS?

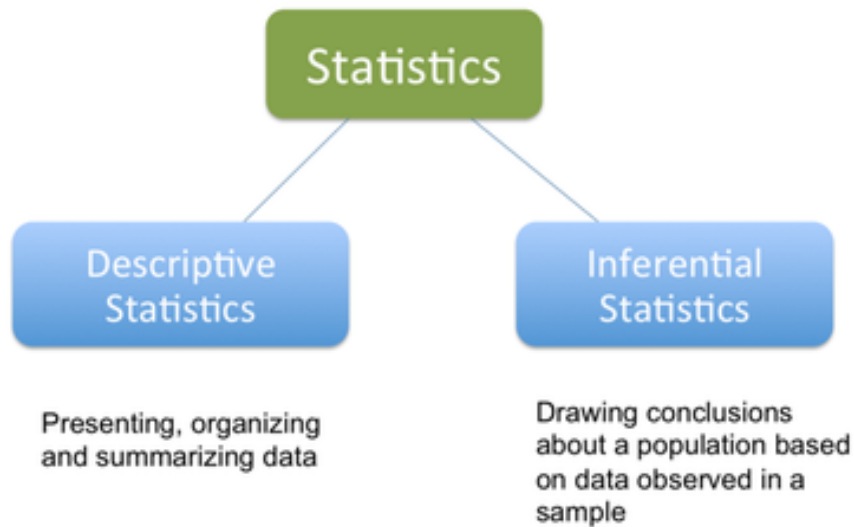


A branch of mathematics taking and transforming numbers into useful information for decision makers.

### 1.2WHY STATISTICS?

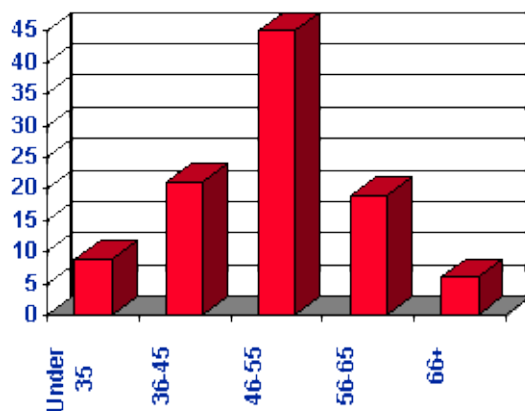
Make reliable forecasts about a business activity. Statistical methods and analyses are often used to communicate research findings and to support hypotheses and give credibility to research methodology and conclusions.

## 1.3 TYPES OF STATISTICS



### 1.3.1 DESCRIPTIVE STATISTICS

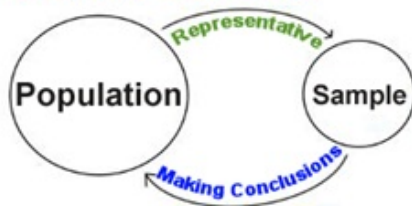
A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.



## 1.3.2 INFERENCE STATISTICS

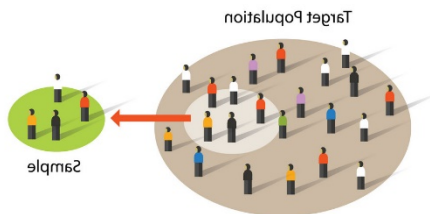
Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible.

### Inferential Statistics



## 1.3.3 POPULATION & SAMPLE

Population is the entire pool from which a statistical sample is drawn. In statistics, population may refer to people, objects, events, hospital visits, measurements, etc. A population can, therefore, be said to be an aggregate observation of subjects grouped together by a common feature. A sample refers to a set of observations drawn from a population. Often, it is necessary to use samples for research, because it is impractical to study the whole population.



### 1.4.1 PARAMETERS AND STATISTIC

A **parameter** is a characteristic of a population. A **statistic** is a characteristic of a sample

| Characteristic               | Mean              | Standard deviation                 |
|------------------------------|-------------------|------------------------------------|
| Sample Statistics            | $\bar{x}$         | $\delta$ or s.d.                   |
| Parameters of the population | $\mu$ (read : Mu) | $\sigma$ (read : sigma)<br>Or S.D. |

### 1.5 VARIABLE & DATA

A variable may also be called a **data item**. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye color and vehicle type are examples of variables. It is called a variable because the value may vary between data units in a population, and may change in value over time.

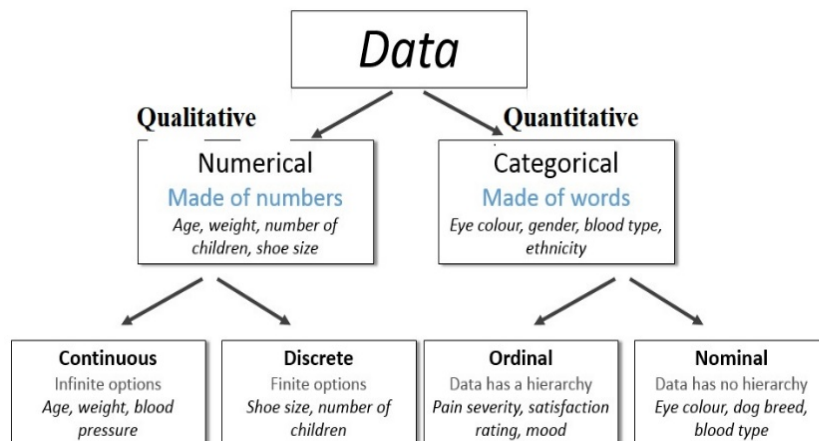
Data is facts or figures from which conclusions can be drawn. Data can take various forms, but are often numerical. As such, data can relate to an enormous variety of aspects, for example:

- the daily weight measurements of each individual in your classroom;
- the number of movie rentals per month for each household in your neighbourhood.

## 1.5.1 TYPES OF VARIABLES

| Independent Variables  | Dependent Variables   |
|--|---|
| Independent variables are variables that determine the value of the variables.                                       | Dependent variables are variables that get determined from independent variables.     |
| Independent variables are considered to be experiment controller these can be manipulated.                           | Dependent variables are experiment measure and are very difficult to manipulate.      |
| Independent variables takes the form of experiment stimulus having two attributes which is either present or absent. | Dependent variables have attributes which are direct, indirect or through constructs. |
| Independent variables are variables that can be termed as casual variables.  | Dependent variables are considered as the caused variables.                           |

## 1.5.2 TYPES OF DATA



## **1.6 FREQUENCY DISTRIBUTION & NORMAL DISTRIBUTION**

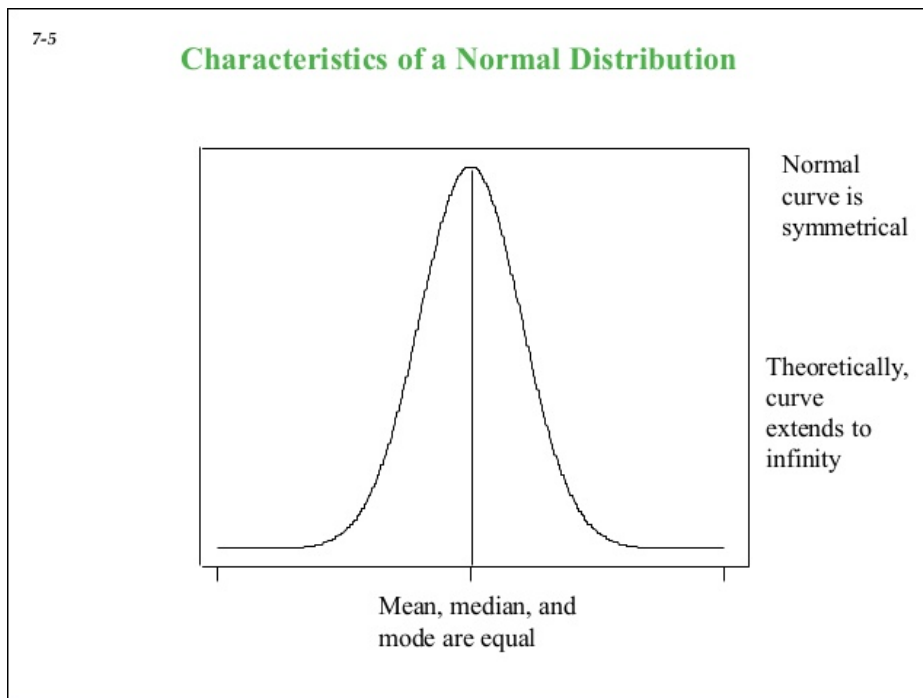
Frequency distribution, in statistics, a graph or data set organized to show the frequency of occurrence of each possible outcome of a repeatable event observed many times. Simple examples are election returns and test scores listed by percentile. A frequency distribution can be graphed as a histogram or pie chart. Frequency distribution, in statistics, a graph or data set organized to show the frequency of occurrence of each possible outcome of a repeatable event observed many times. Simple examples are election returns and test scores listed by percentile. A frequency distribution can be graphed as a histogram or pie chart.

### **NORMAL DISTRIBUTION**

The 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. In N.D mean, mode, median are all equal.

If a dataset follows a normal distribution, then about 68% of the observations will fall within  $\sigma$  of the mean  $\mu$ , which in this case is with the interval  $(-1,1)$ . About 95% of the observations

will fall within 2 standard deviations of the mean, which is the interval  $(-2,2)$  for the standard normal, and about 99.7% of the observations will fall within 3 standard deviations of the mean, which corresponds to the interval  $(-3,3)$  in this case. Although it may appear as if a normal distribution does not include any values beyond a certain interval, the density is actually positive for all values,  $(-\infty, \infty)$ .



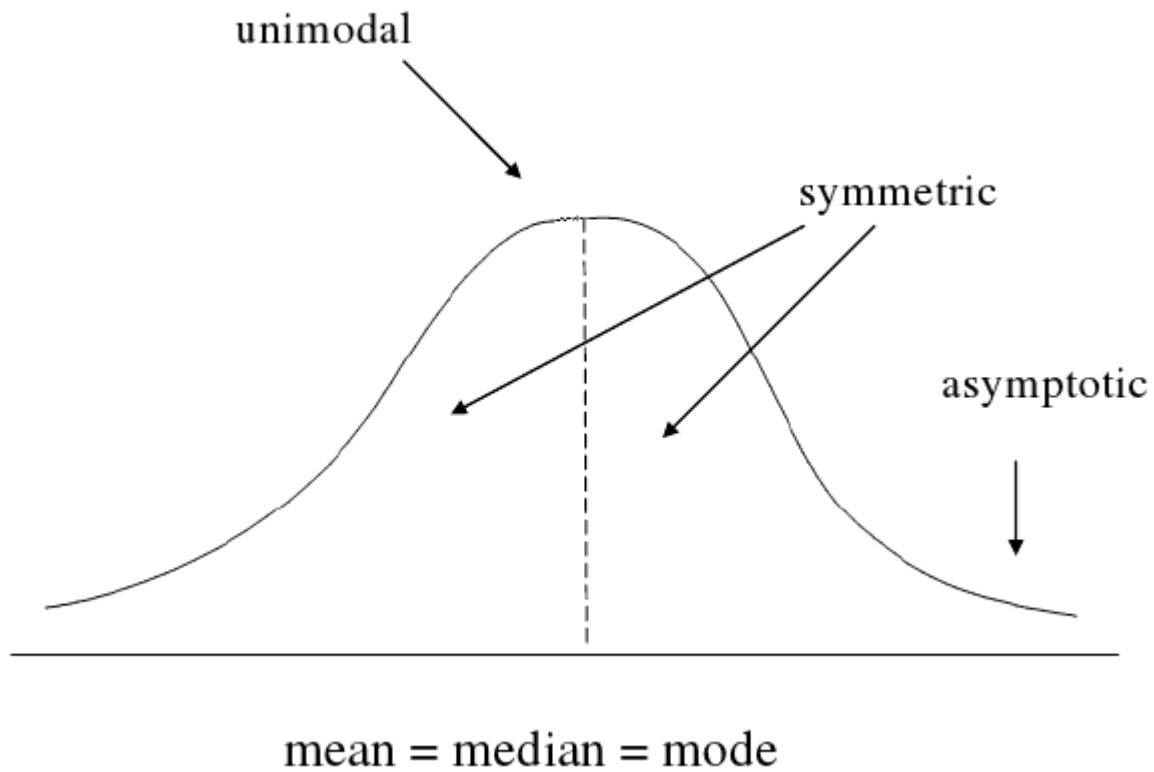
### **1.6.1 CHARACTERISTICS OF Frequency Distribution**

There are four important characteristics of frequency distribution.

They are as follows:

- Measures of central tendency and location (mean, median, mode)
- Measures of dispersion (range, variance, standard deviation)
- The extent of symmetry/asymmetry (skewness)
- The flatness or peakedness\modularity (kurtosis).

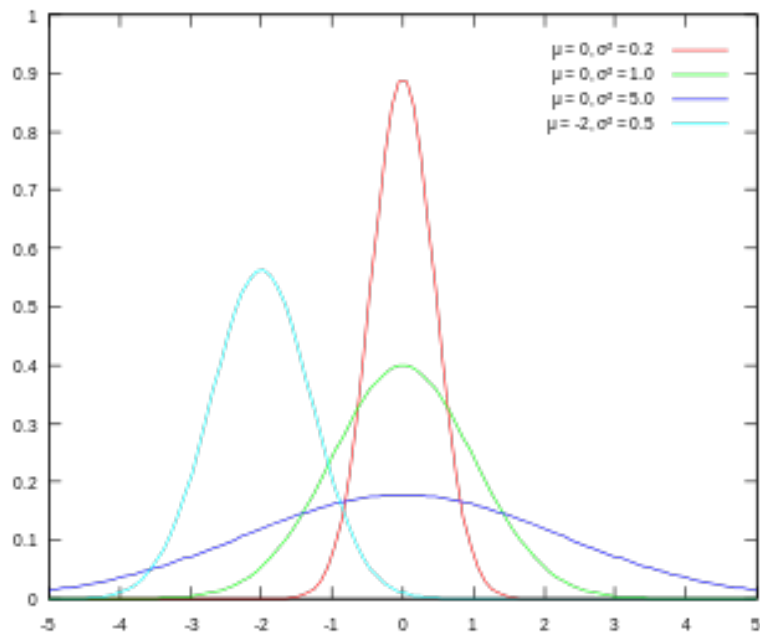




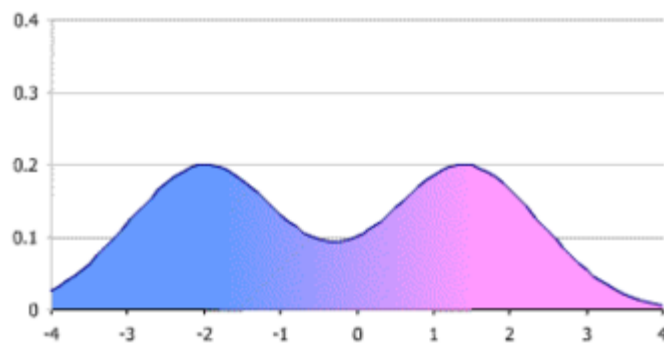
### 1.6.1.1 MODALITY

This depends on the mode of the data. If there is a single mode, the distribution function is called "unimodal". If it has more modes it is "bimodal" (2), "trimodal" (3), etc., or in general, "Multimodal".

## Unimodal



## Bi-model



### 1.6.1.2 SYMMETRY

The Symmetry is a character of the F.D which shows how the data is distributed. There are 2 types of symmetry:

#### 1.6.1.2.1 Symmetric:

**Symmetrical distribution** occurs when the values of variables occur at regular **frequencies** and the mean, median and mode occur at the same point. In graph form, **symmetrical distribution** often appears as a bell curve.

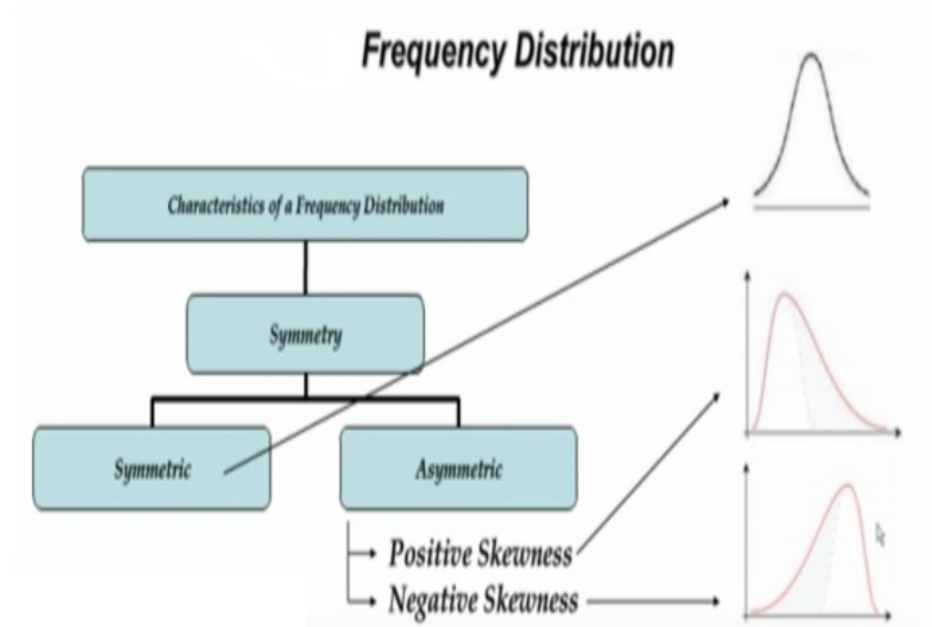
#### 1.6.1.2.2 Asymmetric:

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.

#### **Skewness:**

A distribution is asymmetric if it is not symmetric with zero skewness; in other words, it does not skew. An asymmetric distribution is either left-skewed or right-skewed. A left-skewed distribution, what is known as a negative distribution, has a

longer left tail. A right-skewed distribution, or a positively skewed distribution, has a longer right tail. Determining whether the mean is positive or negative is important when analyzing the skew of a data set because it affects data distribution analysis.



## **2. DESCRIPTIVE STATISTICS**

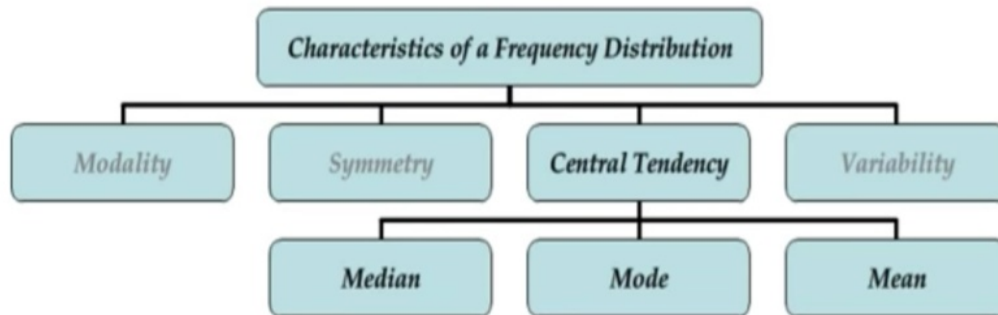
Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

In the characteristics on Frequency Distribution, two of the main characteristics falls under the boundaries of Descriptive Statistics.

### **2.1 CENTRAL TENDENCY:**

The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

1. Mean
2. Median
3. Mode



## **Mean:**

The Mean or **average** is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is  $167/8 = 20.875$ .

**Median:**

The Median is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order(Asc.order), and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median.

**Mode:**

The mode is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in

a bi-modal distribution there are two values that occur most frequently.

| Central Tendency Measures |                  |                           |
|---------------------------|------------------|---------------------------|
| Measure                   | Formula          | Description               |
| Mean                      | $\sum x/n$       | Balance Point             |
| Median                    | $n+1/2$ Position | Middle Value when ordered |
| Mode                      | None             | Most frequent             |

## Variability:

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution. There are some frequently used measures of variability: the range, coefficient of variance, and standard deviation.



**Range:**

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score.

**Coefficient of variance:**

The coefficient of variation (CV) is defined as the ratio of the standard deviation. It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a ratio scale, as these are the measurements that allow the division operation. The coefficient of variation may not have any meaning for data on an interval scale.

**Standard - Deviation:**

Standard deviation is a measure of dispersement in statistics. “Dispersement” tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the mean or average.

## **VARIANCE:**

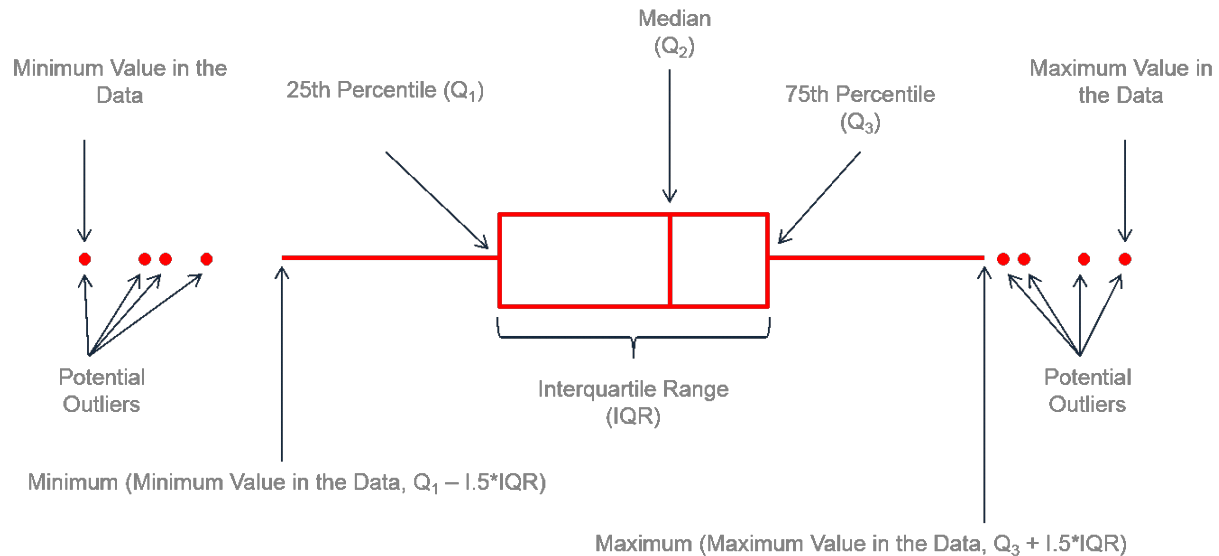
Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.

## **OUTLIERS:**

Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box plots, along with an analytic procedure for detecting outliers when the distribution is normal.

### *The BOX PLOT:*

The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is  $Q1$  and the upper quartile is  $Q3$ , then the difference ( $Q3 - Q1$ ) is called the interquartile range or IQR.



## Interquartile Range(IQR):

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  $Q_1$ ,  $Q_2$ , and  $Q_3$ , respectively.

## Covariance:

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, (i.e., the variables tend to show similar behavior), the covariance is positive. In the opposite case, when the greater values of one

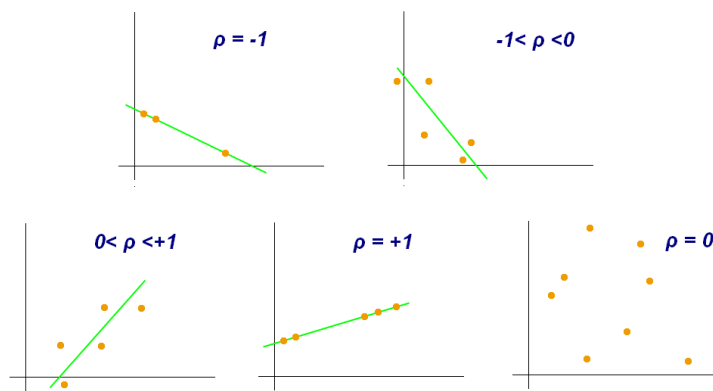
variable mainly correspond to the lesser values of the other, (i.e., the variables tend to show opposite behavior), the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

The formula:

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

### Correlation Coefficient:

It is a measure of the linear correlation between two variables X and Y. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

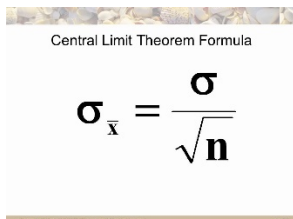


The Formula:

$$r = r_{xy} = \frac{\text{Cov}(x, y)}{S_x \times S_y}$$

## CENTRAL LIMIT THEOREM:

CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size. The formula is:

A graphic with a light blue background and a thin border. At the top, it says "Central Limit Theorem Formula". In the center, the formula is displayed: 
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
 At the bottom, there is a small copyright notice: "Copyright © 2010 by The McGraw-Hill Companies, All rights reserved."
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

## Empirical Rule:

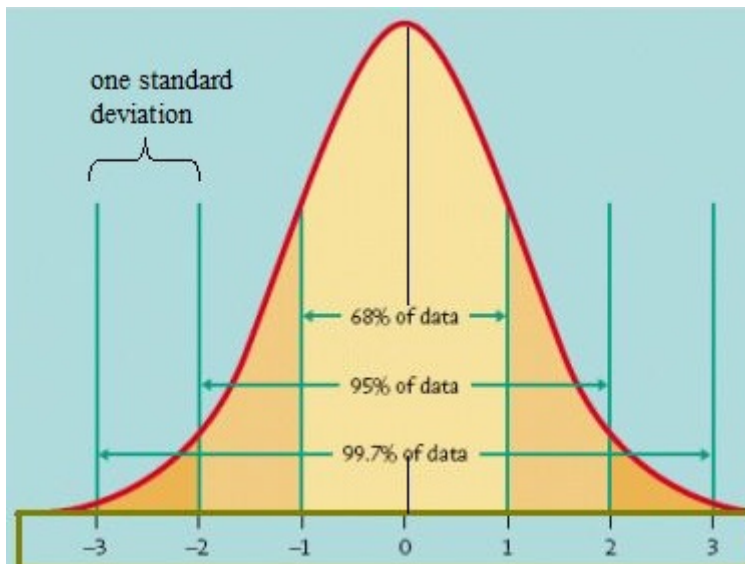
The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- 1) 68% of data falls within the first standard deviation from the mean.
- 2) 95% fall within two standard deviations.

3) 99.7% fall within three standard deviations.

The rule is also called the 68-95-99.7 Rule or the Three Sigma Rule.

The Empirical Rule is often used in statistics for forecasting, especially when obtaining the right data is difficult or impossible to get. The rule can give you a rough estimate of what your data collection might look like if you were able to survey the entire population.



### **Z-Score:**

It is a measure of how many standard deviations below or above the population mean a raw score is. A z-score is also known as a standard score and it can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of

the normal distribution curve). In order to use a z-score, you need to know the mean  $\mu$  and also the population standard deviation  $\sigma$ . The Formula is:

$$z = (x - \mu) / \sigma$$

## CONFIDENCE INTERVALS:

A confidence interval is how much uncertainty there is with any particular statistic. Confidence intervals are often used with a margin of error. It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population. Confidence intervals are intrinsically connected to confidence levels.

$$\bar{X} - z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$$

For a 90% confidence interval:  $z_{\alpha/2} = 1.65$

For a 95% confidence interval:  $z_{\alpha/2} = 1.96$

For a 99% confidence interval:  $z_{\alpha/2} = 2.58$

## Margin of errors:

The margin of error is the range of values below and above the sample statistic in a confidence interval. The confidence interval is a way to show what the uncertainty is with a certain statistic. A margin of error tells you how many percentage points your results will differ from the real population value. For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.

$$\text{MOE} = \pm z^* \left( \frac{\sigma}{\sqrt{n}} \right)$$



# INFERENTIAL STATISTICS

## HYPOTHESIS TESTING:

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.

### Steps in Hypothesis Testing:

#### Step 1: State the Null Hypothesis:

A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. The null hypothesis attempts to show that no variation exists between variables or that a single variable is no different than its mean. It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

## **Step 2: State the Alternative Hypothesis:**

The alternative hypothesis, denoted by  $H_1$  or  $H_a$ , is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

$$H_0: p = 0.5$$

$$H_a: p \neq 0.5$$

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. That is, we would conclude that the coin was probably not fair and balanced.

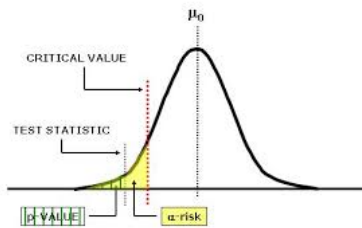
## **Step 3 Choose a test statistic:**

According to the data provided and the requirement choose the tests

## **Step 4 Interpret Critical Values:**

The critical value approach involves determining "likely" or "unlikely" by determining whether or not the observed test statistic is more extreme than would be expected if the null hypothesis were true. That is, it entails comparing the observed test statistic to some

cutoff value, called the "**critical value.**" If the test statistic is more extreme than the critical value, then the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic is not as extreme as the critical value, then the null hypothesis is not rejected.



## ONE VARIABLE TESTS:

### 1.Z-TEST

### 2.T-TEST

### 3. $\chi^2$ -TEST

### 4.F-TEST

## Z-TEST:

A z-test is used for testing the mean of a population versus a standard, or comparing the means of two populations, with large ( $n \geq 30$ ) samples whether you know the population standard deviation or not. It is also used for testing the proportion of some characteristic versus a standard proportion, or comparing the proportions of two populations.

### **When to use z-test:**

- 1) our sample size is greater than 30. Otherwise, use a t test.
- 2) Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.
- 3) Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.
- 4) Your data should be randomly selected from a population, where each item has an equal chance of being selected.

### **Running a Z test on your data requires five steps:**

1. State the null hypothesis and alternate hypothesis.
  2. Choose an alpha level.
  3. Find the critical value of z in a z table.
- $$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$
4. Calculate the z test statistic
  5. Compare the test statistic to the critical z value and decide if you should support or reject the null hypothesis.

### **T-TEST:**

A t-test is used for testing the mean of one population against a standard or comparing the means of two populations if you do not know the populations' standard deviation and when you have a limited sample ( $n < 30$ ). If you know the populations' standard deviation, you may use a z-test.

Example: Measuring the average diameter of shafts from a certain machine when you have a small sample.

### STEPS:

Step 1: Write your null hypothesis statement

Step 2: Write your alternate hypothesis. This is the one you're testing.

Step 3: Identify the following pieces of information you'll need to calculate the test statistic. The question should give you these items:

The sample mean( $\bar{x}$ ).

The population mean( $\mu$ ).

The sample standard deviation( $s$ )

Number of observations( $n$ )

Step 4: Insert the items from above into the t score formula.

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

This is your calculated t-value.

Step 5: Find the t-table value. You need two values to find this:

- ✓ The alpha level.
- ✓ The degrees of freedom.

## CHI-SQUARE TEST:

The **Chi Square** statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent. An example research question that could be answered using a Chi-Square analysis would be:

Is there a significant relationship between voter intent and political party membership?

The Chi Square Test for Normality can only be used if:

- ✓ Your expected value for the number of sample observations for each level is greater than 5
- ✓ Your data is randomly sampled
- ✓ The variable you are studying is categorical.

If your variable is continuous, you will need to bin the data before using the chi-square test for normality.

To apply the Chi-Square Test for Normality to any data set, let your null hypothesis be that your data is sampled from a normal distribution and apply the Chi-Square Goodness of Fit Test. Given your mean and standard deviation, you will need to calculate the expected values under the normal distribution for

every data point. Then use the formula chi square test for normality

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{(\text{expected})}$$

### **Goodness of Fit Test:**

The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution). In other words, it tells you if your sample data represents the data you would expect to find in the actual population.

### **F TEST:**

Any statistical test that uses F-distribution can be called an F-test. It is used when the sample size is small i.e.  $n < 30$ .

The F-test can be used to test the hypothesis that the population variances are equal.

### **STEPS:**

1. State the null hypothesis and the alternate hypothesis.
2. Calculate the F value. The F Value is calculated using the formula  $F = (SSE1 - SSE2 / m) / SSE2 / n - k$ , where SSE = residual sum of squares, m = number of restrictions and k = number of independent variables.

3. Find the F Statistic (the critical value for this test). The F statistic formula is:
4.  $F \text{ Statistic} = \text{variance of the group means} / \text{mean of the within group variances}$ .
5. You can find the F Statistic in the F-Table.
6. Support or Reject the Null Hypothesis.

### **F Test to Compare Two Variances**

A Statistical F Test uses an F Statistic to compare two variances,  $s_1$  and  $s_2$ , by dividing them. The result is always a positive number (because variances are always positive). The equation for comparing two variances with the f-test is:

$$F = s_1^2 / s_2^2$$

If the variances are equal, the ratio of the variances will equal

1. For example, if you had two data sets with a sample 1 (variance of 10) and a sample 2 (variance of 10), the ratio would be  $10/10 = 1$ .



You always test that the population variances are equal when running an F Test. In other words, you always assume that the variances are equal to 1. Therefore, your null hypothesis will always be that the variances are equal.

## **ANOVA:**

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits the aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, but the random factors do not. Analysts use the analysis of the variance test to determine the result that independent variables have on the dependent variable amid a regression study.

## **Types of ANOVA**

There are two types of analysis of variance: **one-way (or unidirectional) and two-way**. One-way or two-way refers to the number of independent variables in your Analysis of Variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, say salary and skill set. It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

The formula is:

$$\begin{aligned}SS_{total} &= \sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \\SS_{between} &= \sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2 \\SS_{within} &= \sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2\end{aligned}$$

© easycalculation.com

## PAIRED T TESTS:

The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations. Common applications of the paired sample t-test include case-control studies or repeated-measures designs

Like many statistical procedures, the paired sample t-test has two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis assumes that the true mean difference between the paired samples is zero. Under this model, all observable differences are explained by random variation. Conversely, the alternative hypothesis assumes that the true mean difference between the paired samples is not equal to zero. The alternative hypothesis can take one of several forms depending on the expected outcome. If the direction of the difference does not matter, a two-tailed hypothesis is used. Otherwise, an upper-tailed or lower-tailed hypothesis can be used to increase the power of the test. The null hypothesis remains the same for each type of alternative hypothesis.

- The null hypothesis ( $H_0$ ) assumes that the true mean difference ( $\mu_d$ ) is equal to zero.
- The two-tailed alternative hypothesis ( $H_1$ ) assumes that  $\mu_d$  is not equal to zero.
- The upper-tailed alternative hypothesis ( $H_1$ ) assumes that  $\mu_d$  is greater than zero.
- The lower-tailed alternative hypothesis ( $H_1$ ) assumes that  $\mu_d$  is less than zero.

The mathematical representations of the null and alternative hypotheses are :

H0:  $\mu_d = 0$

H1:  $\mu_d \neq 0$  (two-tailed)

H1:  $\mu_d > 0$  (upper-tailed)

H1:  $\mu_d < 0$  (lower-tailed)

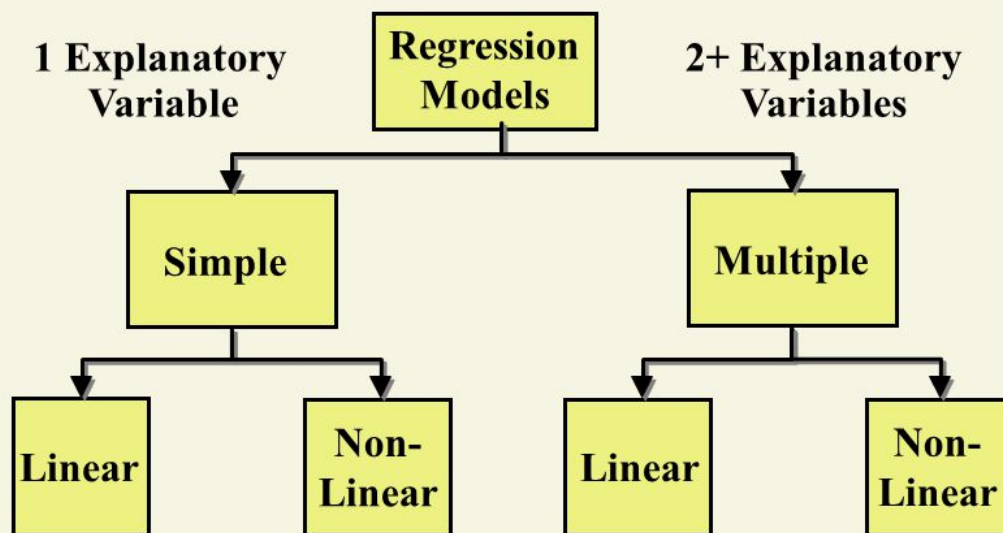
The Formula :

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{(N-1)(N)}}}$$

## REGRESSION:

Regression is a statistical measure used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

### Types of Regression Models



## **SIMPLE LINEAR REGRESSION:**

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

Simple linear regression plots one independent variable  $X$  against one dependent variable  $Y$ . Technically, in regression analysis, the independent variable is usually called the predictor variable and the dependent variable is called the criterion variable. However, many people just call them the independent and dependent variables. Regression analysis can result in linear or nonlinear graphs. A linear regression is where the relationships between your variables can be described with a straight line. Non-linear regressions produce curved lines

## Linear regression Assumptions:

The regression has five key assumptions:

- ✓ Linear relationship
- ✓ Multivariate normality
- ✓ No or little multicollinearity
- ✓ No auto-correlation
- ✓ Homoscedasticity

### The formula:

The diagram shows the linear regression formula  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$  with the following labels and arrows:

- Dependent Variable** points to  $Y_i$ .
- Population Y intercept** points to  $\beta_0$ .
- Population Slope Coefficient** points to  $\beta_1$ .
- Independent Variable** points to  $X_i$ .
- Random Error term** points to  $\epsilon_i$ .

Below the formula, two blue curly braces indicate components:

- A brace under  $\beta_0 + \beta_1 X_i$  is labeled **Linear component**.
- A brace under  $\epsilon_i$  is labeled **Random Error component**.

### Least squares criterion:

Sum of squares of error  $SSE = \sum e^2$

$$e^2 = (y - \hat{y})^2$$

## **Calculating SSR**

The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

**The Total Sum of Squares (SST) is equal to SSR + SSE.**

**Mathematically,**

$$\text{SSR} = \sum (y - \bar{y})^2 \quad (\text{measure of explained variation})$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 \quad (\text{measure of unexplained variation})$$

$$\text{SST} = \text{SSR} + \text{SSE} = \sum (y - \bar{y})^2 \quad (\text{measure of total variation in } y)$$

## **The Coefficient of Determination**

The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as  $R^2$ .

$$R^2 = \text{SSR}/\text{SST} = \text{SSR}/(\text{SSR}+\text{SSE})$$

The value of  $R^2$  can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.



## Standard Error of Regression:

The Standard Error of a regression is a measure of its variability. It can be used in a similar manner to standard deviation, allowing for prediction intervals.

$y \pm 2$  standard errors will provide approximately 95% accuracy, and 3 standard errors will provide a 99% confidence interval.

Standard Error is calculated by taking the square root of the average prediction error.

$$\text{Standard Error} = \sqrt{\text{SSE}/N-k}$$

Where  $n$  is the number of observations in the sample and  $k$  is the total number of variables in the model.

The output of a simple regression is the coefficient  $\beta$  and the constant  $A$ . The equation is then:

$$\Delta y / \Delta x$$

$$y = A + \beta * x + \varepsilon$$

where  $\varepsilon$  is the residual error.

$\beta$  is the per unit change in the dependent variable for each unit change in the independent variable. Mathematically:

$$\beta = \Delta y / \Delta x$$

## Gradient Descent:

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. Gradient descent is also known as steepest descent.

The Mean-Square-Error :

$MSE = SSE/n$ , Where MSE stands for mean square error.

## Inference about slope

$$SE = sb1 = \frac{\sqrt{\sum (y_i - \hat{y}_i)^2 / (n - 2)}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

where  $y_i$  is the value of the dependent variable for observation  $i$ ,  $\hat{y}_i$  is estimated value of the dependent variable for observation  $i$ ,  $x_i$  is the observed value of the independent variable for observation  $i$ ,  $\bar{x}$  is the mean of the independent variable, and  $n$  is the number of observations.

**Degrees of freedom.** For simple linear regression (one independent and one dependent variable), the degrees of freedom (DF) is equal to:

$$DF = n - 2$$

where  $n$  is the number of observations in the sample.

Test statistic. The test statistic is a  $t$  statistic ( $t$ ) defined by the following equation.

$$t = b_1 - \beta_1 / SE$$

where  $b_1$  is the slope of the sample regression line, and SE is the standard error of the slope.

### Confidence Interval:

**A  $100(1 - \alpha)\%$  Confidence Interval for the Mean Value of  $y$  for  $x = x_p$**

$$\hat{y} \pm t_{\alpha/2} (\text{Estimated standard deviation of } \hat{y})$$

or

$$\hat{y} \pm (t_{\alpha/2})s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where  $t_{\alpha/2}$  is based on  $(n - 2)$  df

**A  $100(1 - \alpha)\%$  Prediction Interval for an Individual  $y$  for  $x = x_p$**

$$\hat{y} \pm t_{\alpha/2} [\text{Estimated standard deviation of } (y - \hat{y})]$$

or

$$\hat{y} \pm (t_{\alpha/2})s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where  $t_{\alpha/2}$  is based on  $(n - 2)$  df

### POLYNOMIAL REGRESSION:

More than one independent variable can be used to explain variance in the dependent variable, as long as they are not linearly related.

**A multiple regression takes the form:**

$$y = A + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

where  $k$  is the number of variables, or parameters.

Some general guidelines to keep in mind when estimating a polynomial regression model are:

The fitted model is more reliable when it is built on a larger sample size  $n$ .

Do not extrapolate beyond the limits of your observed values, particularly when the polynomial function has a pronounced curve such that an extrapolation produces meaningless results beyond the scope of the model.

Consider how large the size of the predictor(s) will be when incorporating higher degree terms as this may cause numerical overflow for the statistical software being used.

Do not go strictly by low  $p$ -values to incorporate a higher degree term, but rather just use these to support your model only if the resulting residual plots looks reasonable. This is an example of a situation where you need to determine "practical significance" versus "statistical significance".

In general, as is standard practice throughout regression modeling, your models should adhere to the hierarchy principle, which says that if your model includes  $X_h X_h$  and  $X_h X_h$  is shown to be a statistically significant predictor of  $Y$ , then your model should also include each  $X_j X_j$  for all  $j < h$ , whether or not the coefficients for these lower-order terms are significant.

### **Under-fitting:**

Under-fitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Intuitively, under-fitting occurs when the model or the algorithm does not fit the data well enough.

Specifically, under-fitting occurs if the model or algorithm shows low variance but high bias. Under-fitting is often a result of an excessively simple model has **high bias and low variance**.

### **Over-fitting:**

Over-fitting occurs when a statistical model or machine learning algorithm captures the noise of the data. Intuitively, over-fitting occurs when the model or the algorithm fits the data too well. Specifically, over-fitting occurs if the model or algorithm shows low bias but high variance. Over-fitting is often a result of an excessively complicated model, and it can be prevented by fitting multiple models and using validation or cross-validation to compare their predictive accuracy on test data. Has **high variance and low bias**.

## Regularization:

This is a form of regression, that constrains/regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

A simple relation for linear regression looks like this. Here  $Y$  represents the learned relation and  $\beta$  represents the coefficient estimates for different variables or predictors( $X$ ).

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 .$$

Now, this will adjust the coefficients based on your training data. If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

## Ridge Regression:

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

Above image shows ridge regression, where the RSS is modified by adding the shrinkage quantity. Now, the coefficients are estimated by minimizing this function. Here,  $\lambda$  is the tuning parameter that decides how much we want to penalize the flexibility of our model. The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept  $\beta_0$ . This intercept is a measure of the mean value of the response when  $x_{i1} = x_{i2} = \dots = x_{ip} = 0$ .

When  $\lambda = 0$ , the penalty term has no effect, and the estimates produced by ridge regression will be equal to least squares. However, as  $\lambda \rightarrow \infty$ , the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. As can be seen, selecting a good value of  $\lambda$  is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are also known as the L2 norm.

The coefficients that are produced by the standard least squares method are scale equivariant, i.e. if we multiply each input by  $c$

then the corresponding coefficients are scaled by a factor of  $1/c$ . Therefore, regardless of how the predictor is scaled, the multiplication of predictor and coefficient ( $X_j\beta_j$ ) remains the same. However, this is not the case with ridge regression, and therefore, we need to standardize the predictors or bring the predictors to the same scale before performing ridge regression. The formula used to do this is given below.

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

### **Lasso Regression:**

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|.$$

**Lasso** is another variation, in which the above function is minimized. It's clear that this variation differs from ridge regression only in penalizing the high coefficients. It uses  $|\beta_j|$  (modulus) instead of squares of  $\beta$ , as its penalty. In statistics, this is known as the L1 norm.

Lets take a look at above methods with a different perspective. The ridge regression can be thought of as solving an equation, where summation of squares of coefficients is less than or equal to  $s$ . And the Lasso can be thought of as an equation where summation of modulus of coefficients is less than or equal to  $s$ . Here,  $s$  is a constant that exists for each value of shrinkage factor  $\lambda$ . These equations are also referred to as constraint functions.



Consider there are 2 parameters in a given problem. Then according to above formulation, the ridge regression is expressed by  $\beta_1^2 + \beta_2^2 \leq s$ . This implies that ridge regression coefficients have the smallest RSS(loss function) for all points that lie within the circle given by  $\beta_1^2 + \beta_2^2 \leq s$ .

Similarly, for lasso, the equation becomes,  $|\beta_1| + |\beta_2| \leq s$ . This implies that lasso coefficients have the smallest RSS(loss function) for all points that lie within the diamond given by  $|\beta_1| + |\beta_2| \leq s$ .

### Elastic Net:

In statistics in particular, the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

The elastic net technique solves this regularization problem. For an  $\alpha$  strictly between 0 and 1, and a non-negative  $\lambda$ , elastic net solves the problem

$$\min_{\beta_0, \beta} \left( \frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta) \right),$$

where

$$P_\alpha(\beta) = (1-\alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1 = \sum_{j=1}^p ((1-\alpha) \beta_j^2 + \alpha |\beta_j|).$$

Elastic net is the same as lasso when  $\alpha = 1$ . As  $\alpha$  shrinks toward 0, elastic net approaches ridge regression. For other values of  $\alpha$ , the penalty term  $P_\alpha(\beta)$  interpolates between the L1 norm of  $\beta$  and the squared L2 norm of  $\beta$ .



## **LOGISTIC REGRESSION:**

The logistic model (or logit model) is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables. The two possible dependent variable values are often labeled as "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. The binary logistic regression model can be generalized to more than two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model.

### **The Odds:**

The odds is not the same as the probability. The odds is the number of "successes" (deaths) per "failure" (continue to live), while the probability is the proportion of "successes". I find it instructive to compare how one would estimate these two: An estimate of the odds would be the ratio of the number of successes over the number of failures, while an estimate of the probability would be the ratio of the number of success over the total number of observations.

Odds and probabilities are both ways of quantifying how likely an event is, so it is not surprising that there is a one to one relation between the two. You can turn a probability ( $p$ ) into an odds ( $o$ ) using the following formula:  $o = p / (1 - p)$ . You can turn an odds into a probability like so:  $p = o / (1 + o)$

## **LOG ODDS:**

Log odds are an alternate way of expressing probabilities, which simplifies the process of updating them with new evidence. Unfortunately, it is difficult to convert between probability and log odds. The log odds is the log of the odds ratio. Thus, the log odds of A are

$$= \log(P(A)/P(\neg A)).$$

## **Logistic regression assumptions:**

- It has no outliers.
- No or little multicollinearity.
- Linear relationship between logical outcomes and each predictor variable.

## **Transforming a proportion:**

- The value should be between 0 and 1.
- Odds are always positive.  
$$\text{odds} = (p / (1 - p)) \rightarrow [0, \infty)$$
- Log odds is continuous.  
$$\text{Log odds} = \ln(p / (1 - p)) \rightarrow (-\infty, \infty)$$

## Gradient descent:

$$J(\theta) = -1/m [ \sum y \log h_0(x) + (1-y) \log (1-h_0(x)) ]$$

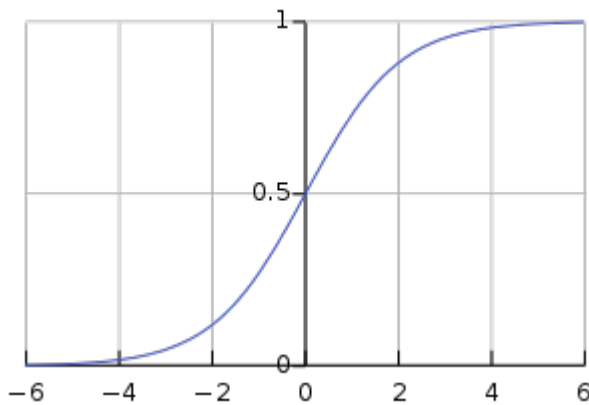
To find minimum  $J(\theta)$

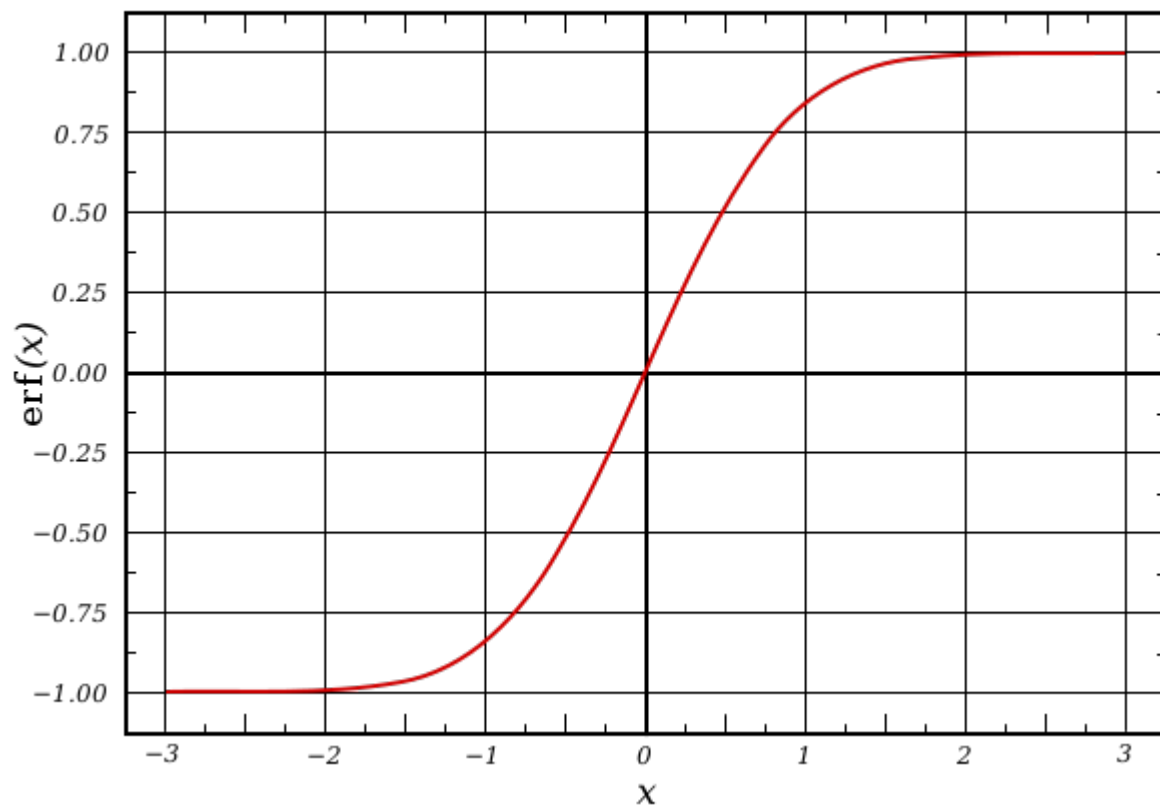
Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \}$$

## SIGMOID Function:

A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve. Often, sigmoid function refers to the special case of the logistic function.





$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}} \rightarrow [0, 1]$$

### Cost Function:

The cost function for the linear regression as follows:

$$\begin{aligned} \text{cost}(h_0(x), y) &= -\log(h_0(x)) \text{ if } y=1 \\ &= -\log(1-h_0(x)) \text{ if } y=0 \end{aligned}$$

### Metrics for Performance Evaluation

#### Confusion Matrix:

A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one

(in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).[\[2\]](#) The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

|              |          | Predicted class      |                      |
|--------------|----------|----------------------|----------------------|
|              |          | <i>P</i>             | <i>N</i>             |
| Actual Class | <i>P</i> | True Positives (TP)  | False Negatives (FN) |
|              | <i>N</i> | False Positives (FP) | True Negatives (TN)  |

## ROC(Receiver Operating Characteristics):

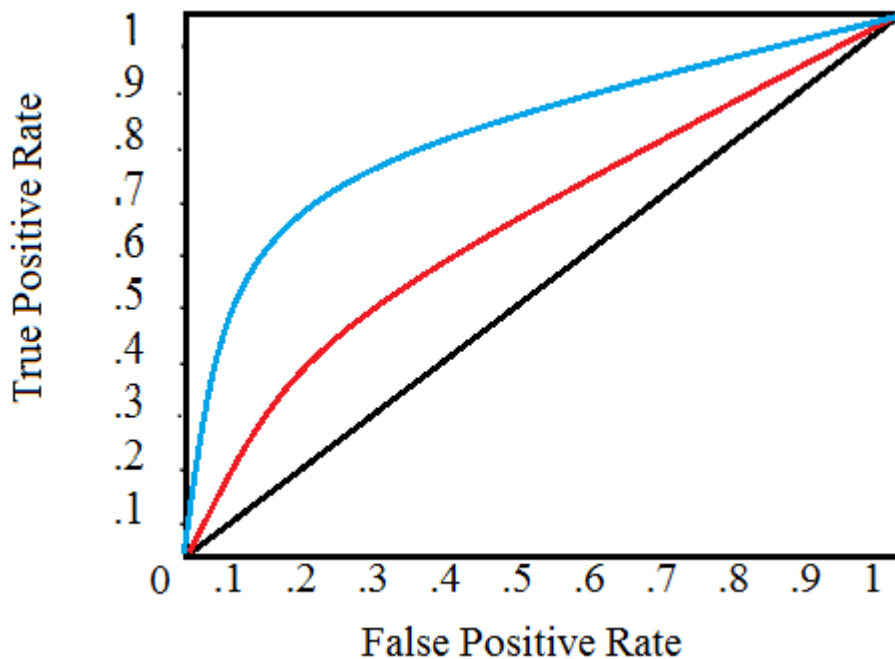
**Operating Characteristics curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity,

recall or probability of detection[1] in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm[1] and can be calculated as  $(1 - \text{specificity})$ . It can also be thought of as a plot of the Power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.





## AUC:

Area Under Curve provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions.

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions.

## Root Mean Squared Error (RMSE):

RMSE is the most popular evaluation metric used in regression problems. It follows an assumption that error is unbiased and follow a normal distribution.

Here are the key points to consider on RMSE:

The power of **‘square root’** empowers this metric to show large number deviations.

The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values. In other words, this metric aptly displays the plausible magnitude of error term.

It avoids the use of absolute error values which is highly undesirable in mathematical calculations.

When we have more samples, reconstructing the error distribution using RMSE is considered to be more reliable.

RMSE is highly affected by outlier values. Hence, make sure you’ve removed outliers from your data set prior to using this metric. As compared to mean absolute error, RMSE gives higher weightage and punishes large errors.

RMSE metric is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

where, N is Total Number of Observations.

### **MAPE:**

The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. Where  $A_t$  is the actual value and  $F_t$  is the forecast value, this is given by:

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

The mean absolute percentage error (MAPE) is the most common measure used to forecast error, and works best if there are no extremes to the data (and no zeros).