**Unit II**

**Correlation**

**Correlation** refers to a process for establishing the relationships between two variables. You learned a way to get a general idea about whether or not two variables are related, is to plot them on a “[scatter plot](https://byjus.com/maths/scatter-plot/)”. While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach.

**Correlation in Statistics**

This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.

A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.

For ordinal scales, the correlation coefficient can be calculated by using Spearman’s rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson’s r, ordinarily referred to as simply the correlation coefficient.

**What Does Correlation Measure?**

In statistics, Correlation studies and measures the direction and extent of relationship among variables, so the correlation measures co-variation, not causation. Therefore, we should never interpret correlation as implying cause and effect relation. For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change). Furthermore, if the correlation exists, it is linear, i.e. we can represent the relative movement of the two variables by drawing a straight line on graph paper.

## Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson’s r is presented. From the following examples, relatively small sample sizes are given. Later, data from larger samples are given.

## Scatter Diagram

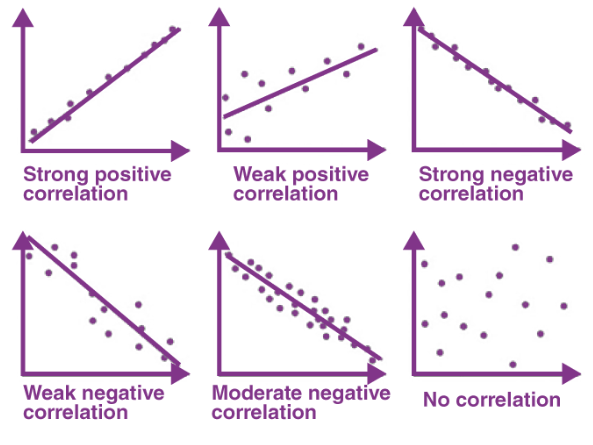
A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Later, when the regression model is used, one of the variables is defined as an independent variable, and the other is defined as a dependent variable. In regression, the independent variable X is considered to have some effect or influence on the dependent variable Y. Correlation methods are symmetric with respect to the two variables, with no indication of causation or direction of influence being part of the statistical consideration. A scatter diagram is given in the following example. The same example is later used to determine the correlation coefficient.

## Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

* Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
* Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
* No Correlation – when there is no linear dependence or no relation between the two variables.

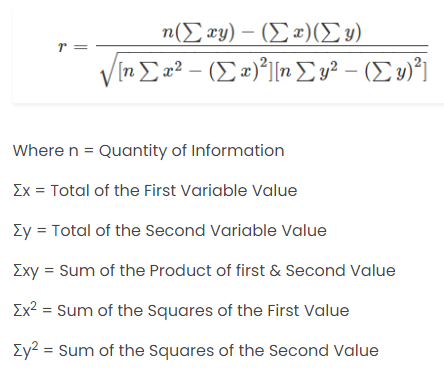


**Correlation Formula**

Correlation shows the relation between two variables. Correlation coefficient shows the measure of correlation. To compare two datasets, we use the correlation formulas.

### Pearson Correlation Coefficient Formula

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data sets. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.



**Regression**

Regression are used to predict continuous values, while classification categorizes data. Both are supervised learning tasks in machine learning.

## ****What is Regression?****

[Regression](https://www.geeksforgeeks.org/types-of-regression-techniques/) is a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The objective is to determine the most suitable function that characterizes the connection between these variables.

It seeks to find the best-fitting model, which can be utilized to make predictions or draw conclusions.

## Regression in Machine Learning

It is a supervised machine learning technique, used to predict the value of the dependent variable for new, unseen data. It models the relationship between the input features and the target variable, allowing for the estimation or prediction of numerical values.

Regression analysis problem works with if output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.

## Terminologies Related to the Regression Analysis in Machine Learning

Terminologies Related to Regression Analysis:

* **Response Variable**: The primary factor to predict or understand in regression, also known as the dependent variable or target variable.
* **Predictor Variable**: Factors influencing the response variable, used to predict its values; also called independent variables.
* **Outliers**: Observations with significantly low or high values compared to others, potentially impacting results and best avoided.
* **Multicollinearity**: High correlation among independent variables, which can complicate the ranking of influential variables.
* [**Underfitting and Overfitting:**](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/)Overfitting occurs when an algorithm performs well on training but poorly on testing, while underfitting indicates poor performance on both datasets.

### ****Regression Types****

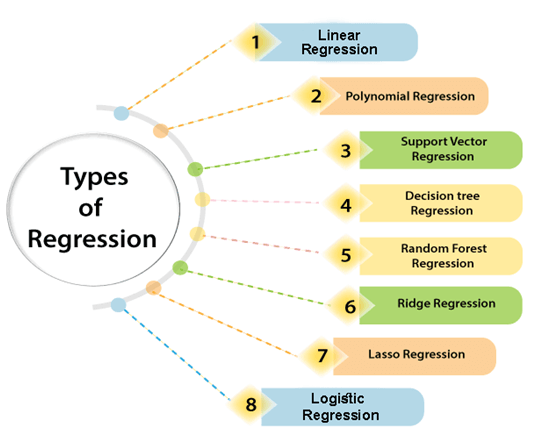
There are two main types of regression:

* **Simple Regression**
  + Used to predict a continuous dependent variable based on a single independent variable.
  + Simple linear regression should be used when there is only a single independent variable.
* **Multiple Regression**
  + Used to predict a continuous dependent variable based on multiple independent variables.
  + Multiple linear regression should be used when there are multiple independent variables.
* **NonLinear Regression**
  + Relationship between the dependent variable and independent variable(s) follows a nonlinear pattern.
  + Provides flexibility in modeling a wide range of functional forms.

## Types of Regression

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

* **Linear Regression**
* **Logistic Regression**
* **Polynomial Regression**
* **Support Vector Regression**
* **Decision Tree Regression**
* **Random Forest Regression**
* **Ridge Regression**
* **Lasso Regression:**



## Regression Algorithms

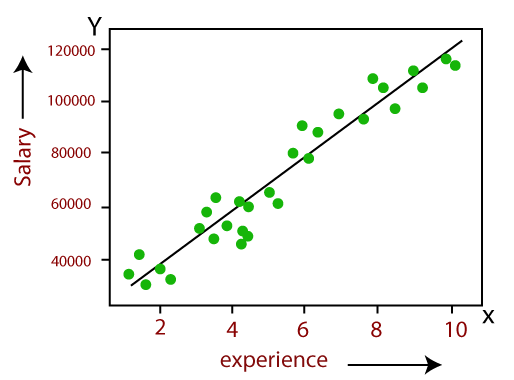
There are many different types of regression algorithms, but some of the most common include:

* **Linear Regression**

[Linear regression](https://www.geeksforgeeks.org/ml-linear-regression/) is one of the simplest and most widely used statistical models. This assumes that there is a linear relationship between the independent and dependent variables. This means that the change in the dependent variable is proportional to the change in the independent variables.

### Linear Regression:

* Linear regression is a statistical regression method which is used for predictive analysis.
* It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
* It is used for solving the regression problem in machine learning.
* Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
* If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
* The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



Below is the mathematical equation for Linear regression: Y= aX+b

**Here, Y = dependent variables (target variables),  
X= Independent variables (predictor variables),  
a and b are the linear coefficients**

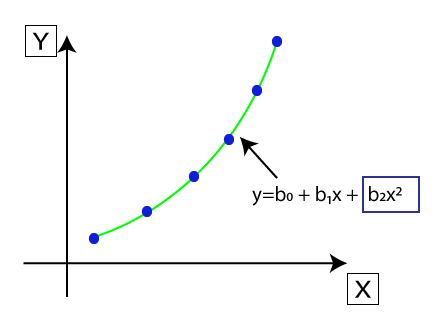
**Some popular applications of linear regression are:**

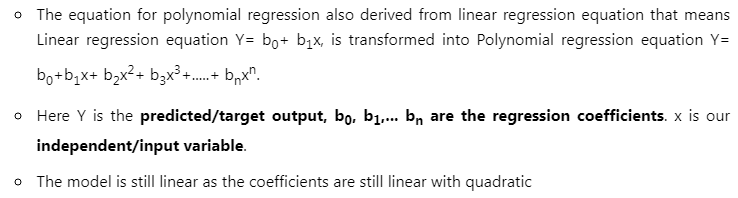
* Analyzing trends and sales estimates
* Salary forecasting
* Real estate prediction
* Arriving at ETAs in traffic.
* **Polynomial Regression**

[Polynomial regression](https://www.geeksforgeeks.org/polynomial-regression-from-scratch-using-python/)  is used to model nonlinear relationships between the dependent variable and the independent variables. It adds polynomial terms to the linear regression model to capture more complex relationships.

### Polynomial Regression:

* Polynomial Regression is a type of regression which models the **non-linear dataset** using a linear model.
* It is similar to multiple linear regression, but it fits a non-linear curve between the value of x and corresponding conditional values of y.
* Suppose there is a dataset which consists of datapoints which are present in a non-linear fashion, so for such case, linear regression will not best fit to those datapoints. To cover such datapoints, we need Polynomial regression.
* I**n Polynomial regression, the original features are transformed into polynomial features of given degree and then modeled using a linear model.** Which means the datapoints are best fitted using a polynomial line.





* **Support Vector Regression (SVR**)

[Support vector regression (SVR)](https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/) is a type of regression algorithm that is based on the support vector machine (SVM) algorithm. SVM is a type of algorithm that is used for classification tasks, but it can also be used for regression tasks. SVR works by finding a hyperplane that minimizes the sum of the squared residuals between the predicted and actual values.

* **Decision Tree Regression**

[Decision tree regression](https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/) is a type of regression algorithm that builds a decision tree to predict the target value. A decision tree is a tree-like structure that consists of nodes and branches. Each node represents a decision, and each branch represents the outcome of that decision. The goal of decision tree regression is to build a tree that can accurately predict the target value for new data points.

* **Random Forest Regression**

[Random forest regression](https://www.geeksforgeeks.org/random-forest-regression-in-python/) is an ensemble method that combines multiple decision trees to predict the target value. Ensemble methods are a type of machine learning algorithm that combines multiple models to improve the performance of the overall model. Random forest regression works by building a large number of decision trees, each of which is trained on a different subset of the training data. The final prediction is made by averaging the predictions of all of the trees.

### Regularized Linear Regression Techniques

**Ridge Regression**

[Ridge regression](https://www.geeksforgeeks.org/ml-ridge-regressor-using-sklearn/) is a type of linear regression that is used to prevent overfitting. Overfitting occurs when the model learns the training data too well and is unable to generalize to new data.

### Ridge Regression:

* Ridge regression is one of the most robust versions of linear regression in which a small amount of bias is introduced so that we can get better long term predictions.
* The amount of bias added to the model is known as **Ridge Regression penalty**. We can compute this penalty term by multiplying with the lambda to the squared weight of each individual features.
* The equation for ridge regression will be:



* A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
* Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
* It helps to solve the problems if we have more parameters than samples.

### Lasso Regression:

* Lasso regression is another regularization technique to reduce the complexity of the model.
* It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights.
* Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
* It is also called as **L1 regularization**. The equation for Lasso regression will be:



**Lasso regression**

[Lasso regression](https://www.geeksforgeeks.org/understanding-lars-lasso-regression/) is another type of linear regression that is used to prevent overfitting. It does this by adding a penalty term to the loss function that forces the model to use some weights and to set others to zero.

### ****Examples****

**Which of the following is a regression task?**

* Predicting age of a person
* Predicting nationality of a person
* Predicting whether stock price of a company will increase tomorrow
* Predicting whether a document is related to sighting of UFOs?

**Solution :**Predicting age of a person (because it is a real value, predicting nationality is categorical, whether stock price will increase is discrete-yes/no answer, predicting whether a document is related to UFO is again discrete- a yes/no answer).

### ****Applications of Regression****

* **Predicting prices:** For example, a regression model could be used to predict the price of a house based on its size, location, and other features.
* **Forecasting trends:** For example, a regression model could be used to forecast the sales of a product based on historical sales data and economic indicators.
* **Identifying risk factors:** For example, a regression model could be used to identify risk factors for heart disease based on patient data.
* **Making decisions:** For example, a regression model could be used to recommend which investment to buy based on market data.

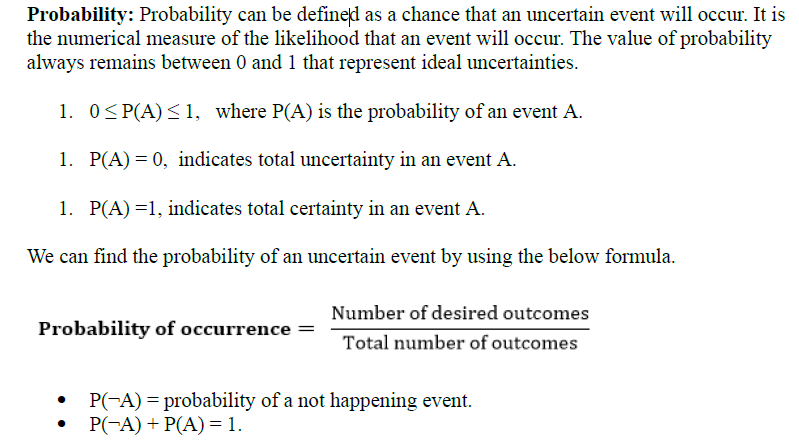
### ****Advantages of Regression****

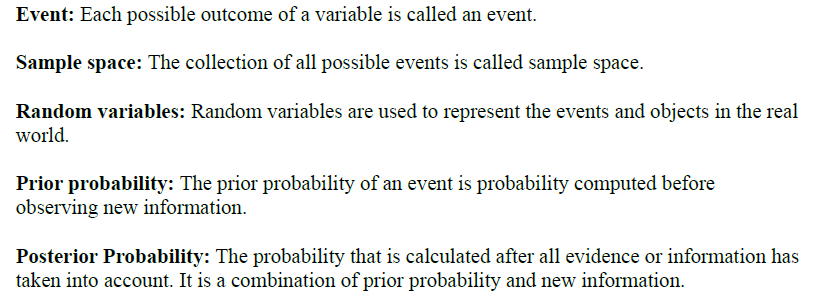
* Easy to understand and interpret
* Robust to outliers
* Can handle both linear and nonlinear relationships.

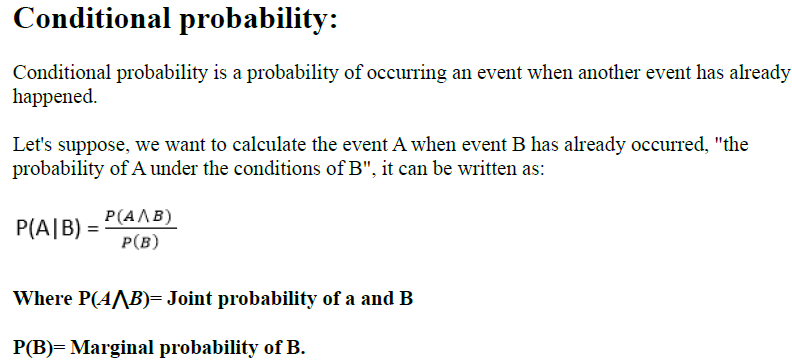
### ****Disadvantages of Regression****

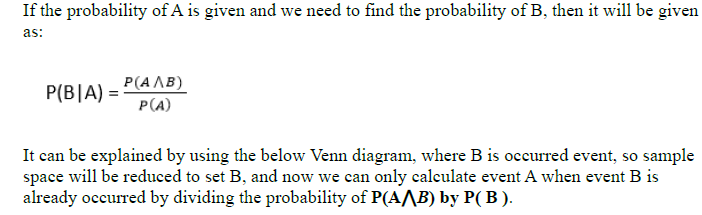
* Assumes linearity
* Sensitive to multi collinearity
* May not be suitable for highly complex relationships

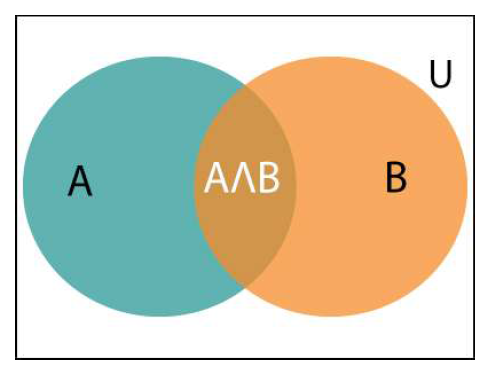
**Probability**

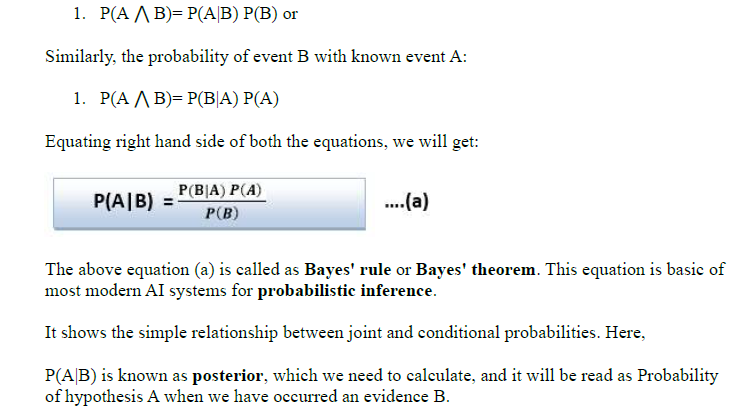
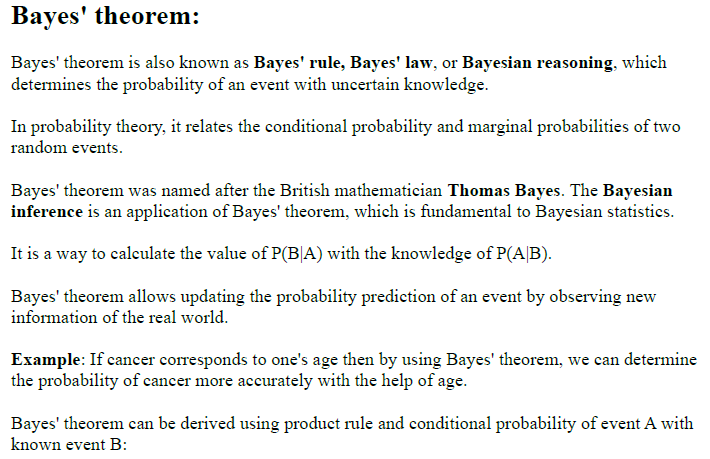
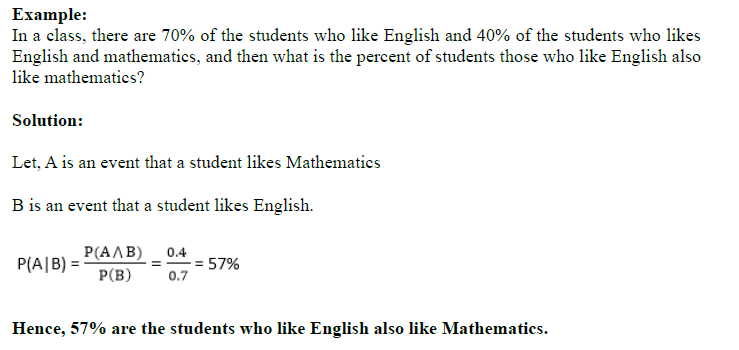


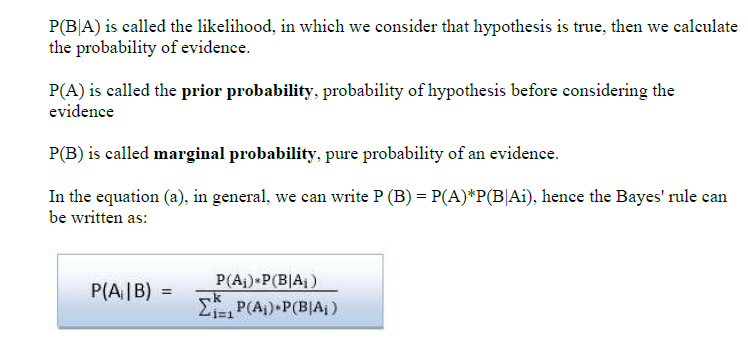
****

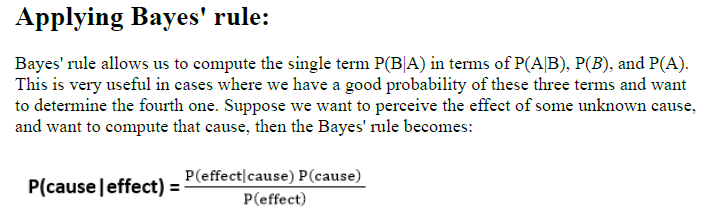
****

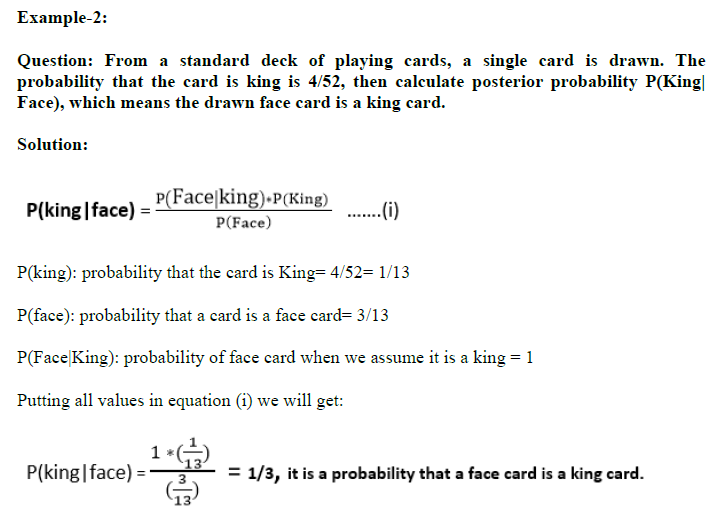
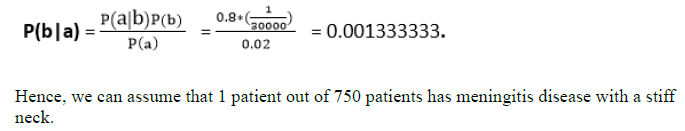
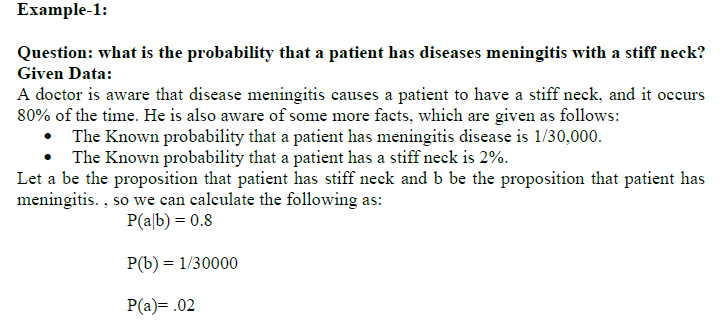


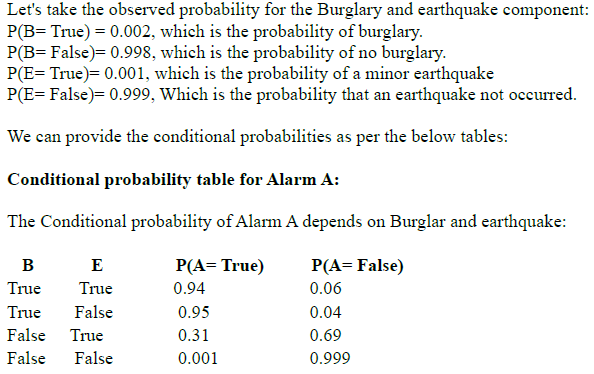
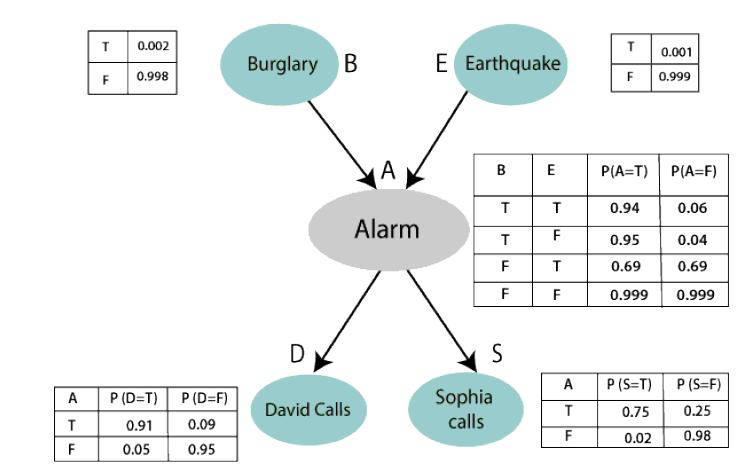
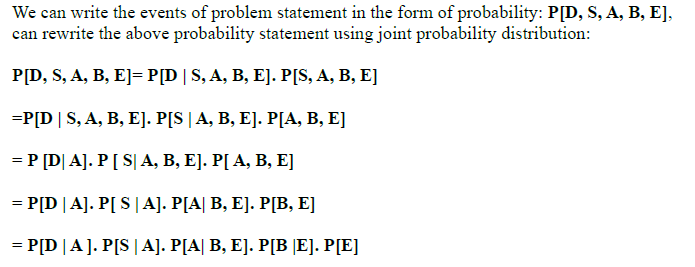
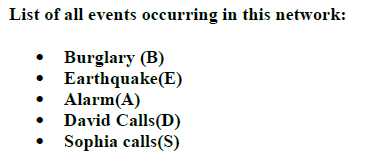
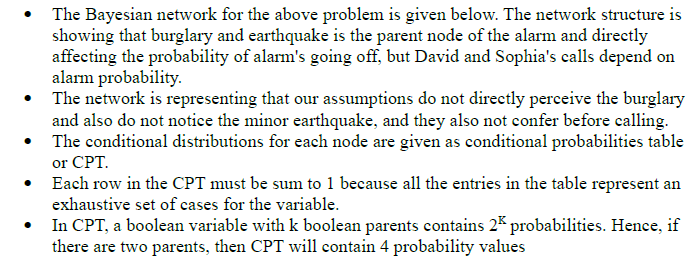
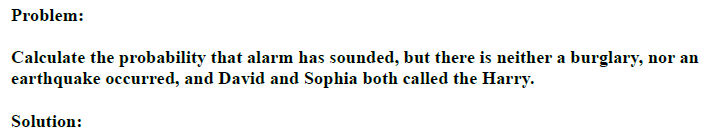
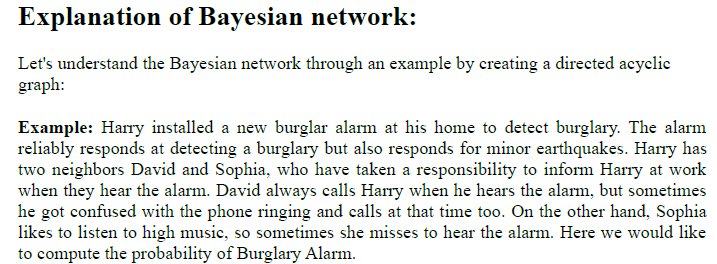
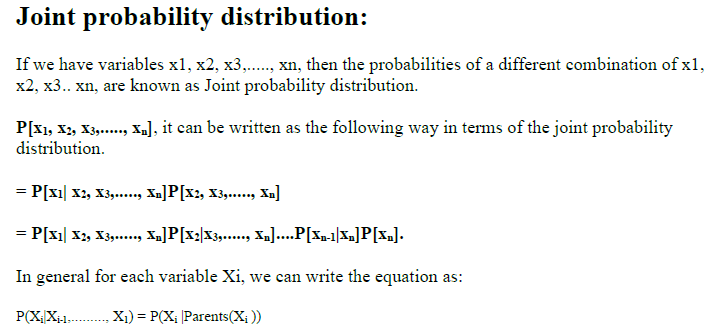
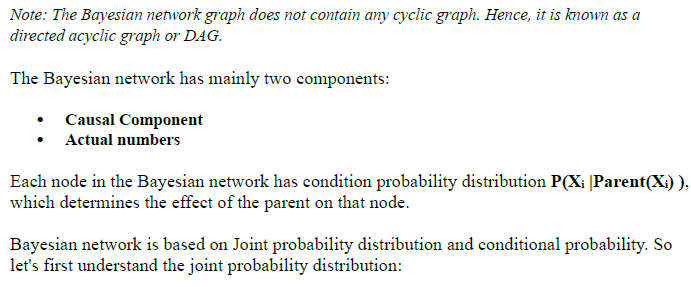
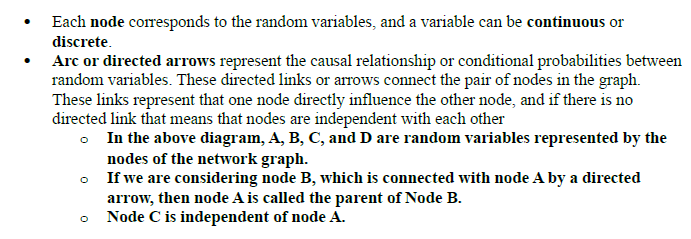
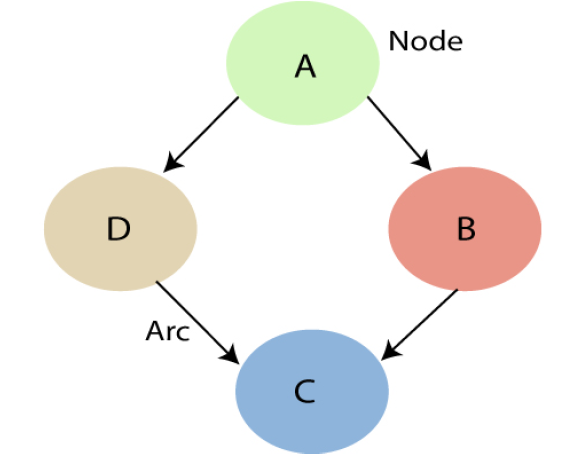
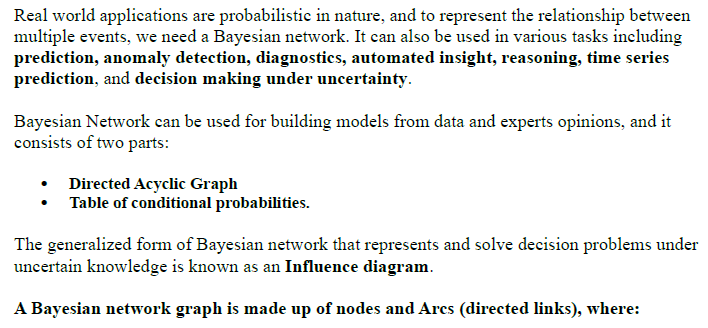
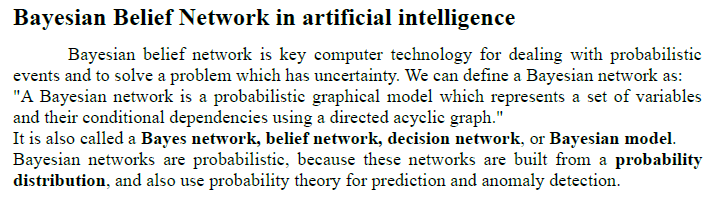
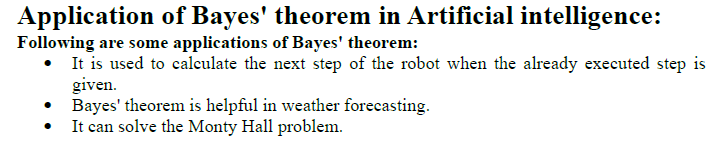


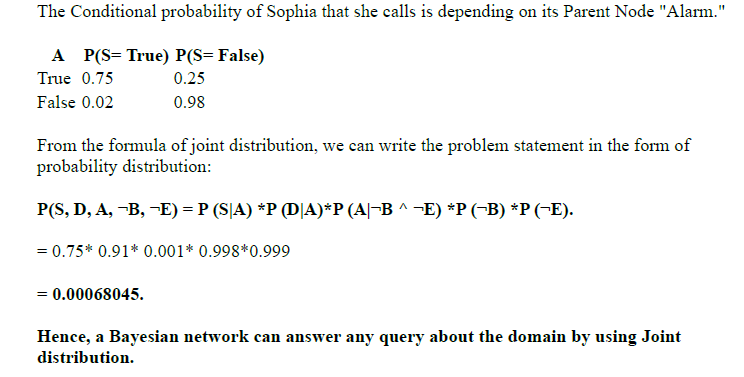
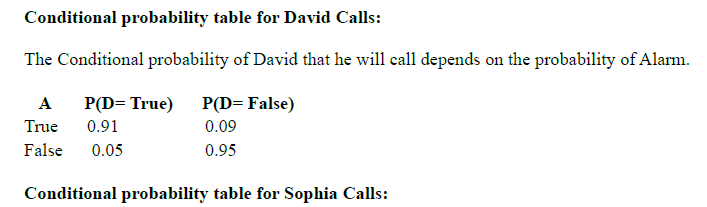


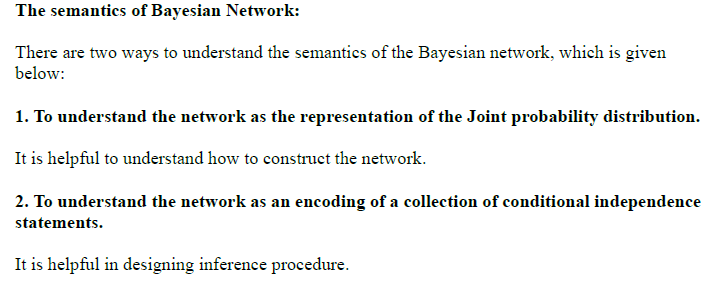












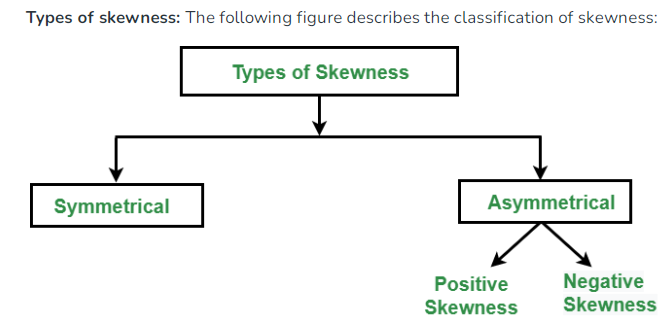
**Skewness**

**Skewness** is an important statistical technique that helps to determine asymmetrical behavior than of the frequency distribution, or more precisely, the lack of symmetry of tails both left and right of the frequency curve. A distribution or dataset is symmetric if it looks the same to the left and right of the center point.

“Skewness essentially is a commonly used measure in descriptive statistics that characterizes the asymmetry of a data distribution, while kurtosis determines the heaviness of the distribution tails.”

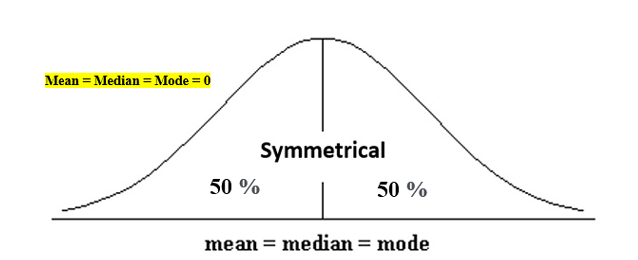
## What Is Skewness?

Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side. Skewness helps in understanding the shape and outliers in a dataset.



**1. Symmetric Skewness:**A perfect symmetric distribution is one in which frequency distribution is the same on the sides of the center point of the frequency curve. In this, Mean = Median = Mode. There is no skewness in a perfectly symmetrical distribution.

The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.



When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness  
comes into the picture.

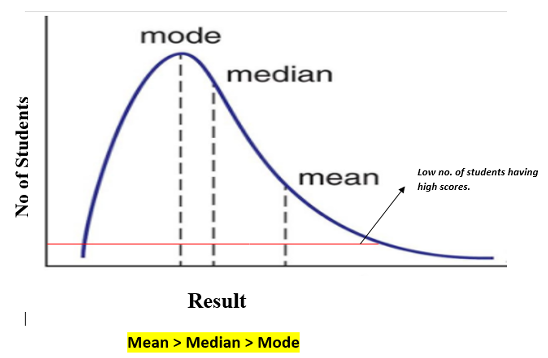
**2. Asymmetric Skewness:**A asymmetrical or skewed distribution is one in which the spread of the frequencies is different on both the sides of the center point or the frequency curve is more stretched towards one side or value of Mean. Median and Mode falls at different points.

* **Positive Skewness:**In this, the concentration of frequencies is more towards higher values of the variable i.e. the right tail is longer than the left tail. **Positive skewness** indicates a longer tail on the right side of the distribution
* **Negative Skewness:**In this, the concentration of frequencies is more towards the lower values of the variable i.e. the left tail is longer than the right tail. **Negative skewness** indicates a longer tail on the left side.

## Types of Skewness

#### Positive Skewed or Right-Skewed  (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



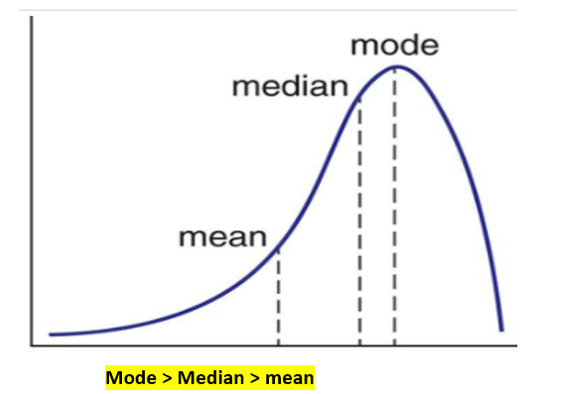
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.

Extreme positive skewness is not desirable for a distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

#### Negative Skewed or Left-Skewed (Negative Skewness)

A negatively skewed or left-skewed distribution has a long left tail; it is the complete opposite of a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.

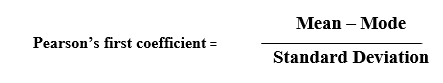


Median is the middle value, and mode is the most frequent value. Due to an unbalanced distribution, the median will be higher than the mean.

## How to Calculate the Skewness Coefficient?

Skewness can be calculated using various methods, whereas the most commonly used method is Pearson’s coefficient.

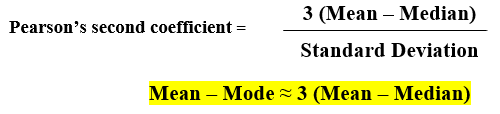
**Pearson’s first coefficient of skewness**  
To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.



As Pearson’s correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of **-1 to +1.** That accurately shows the range of the correlation values.

Pearson’s first coefficient of skewness is helping if the data present high mode. But, if the data have low mode or various modes, Pearson’s first coefficient is not preferred, and Pearson’s second coefficient may be superior, as it does not rely on the mode.

**Pearson’s second coefficient of skewness**  
subtract the median from the mean, multiply the difference by 3, and divide the product by the standard deviation.



Rule of thumb : **If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.**  
**If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.**  
**If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.**

## Kurtosis

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.

Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset. The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.

Peakedness in a data distribution is **the degree to which data values are concentrated around the mean**. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

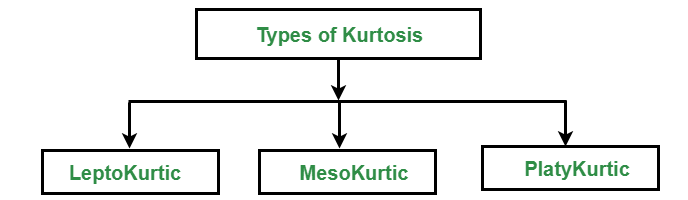
Example:

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

### Kurtosis:

It is also a characteristic of the frequency distribution. It gives an idea about the shape of a frequency distribution. Basically, the measure of kurtosis is the extent to which a frequency distribution is peaked in comparison with a normal curve. It is the degree of peakedness of a distribution.

**Types of kurtosis:**The following figure describes the classification of kurtosis:



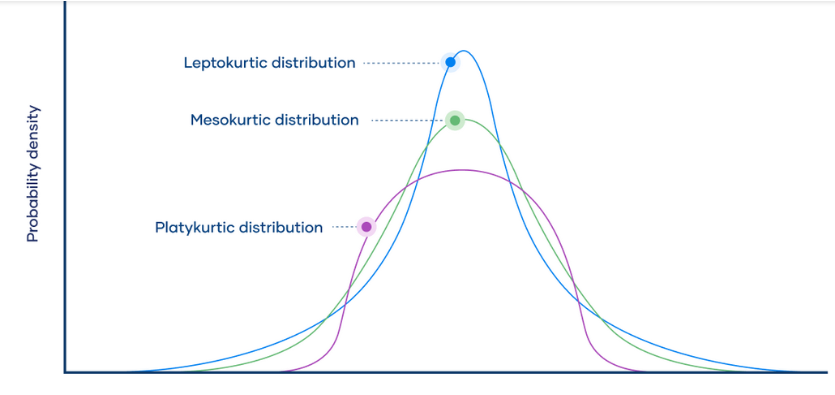
1. **Leptokurtic:**Leptokurtic is a curve having a high peak than the normal distribution. In this curve, there is too much concentration of items near the central value.
2. **Mesokurtic:**Mesokurtic is a curve having a normal peak than the normal curve. In this curve, there is equal distribution of items around the central value.
3. **Platykurtic:**Platykurtic is a curve having a low peak than the normal curve is called platykurtic. In this curve, there is less concentration of items around the central value.

## What Is Excess Kurtosis?

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution).

## Types of Excess Kurtosis

1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
2. Mesokurtic (kurtosis same as the normal distribution).
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution).

**Leptokurtic**

Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

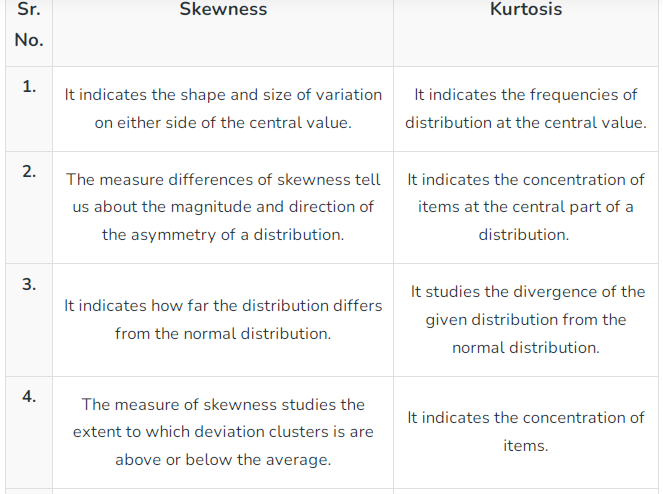
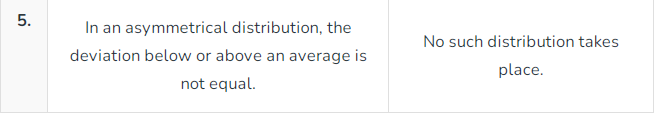
#### Platykurtic

Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

#### Mesokurtic

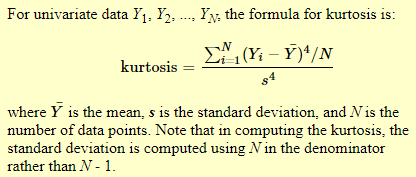
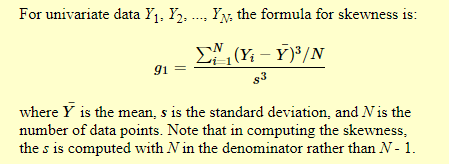
Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

**Comparison of Skewness and Kurtosis**

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

The [histogram](https://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm) is an effective graphical technique for showing both the skewness and kurtosis of data set.

**Key Points**

* Skewness is a statistical measure of the asymmetry of a probability distribution. It characterizes the extent to which the distribution of a set of values deviates from a normal distribution.
* Skewness between -0.5 and 0.5 is symmetrical.
* Kurtosis measures whether data is heavily heavy-tailed or light-tailed.
* Data sets with high kurtosis have heavy tails and more outliers, while data sets with low kurtosis tend to have light tails and fewer outliers.
* Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution).

**Multivariate ANOVA (MANOVA)**

Multivariate ANOVA (MANOVA) extends the capabilities of analysis of variance (ANOVA) by assessing multiple dependent variables simultaneously. ANOVA statistically tests the differences between three or more group means. For example, if you have three different teaching methods and you want to evaluate the average scores for these groups, you can use ANOVA. However, ANOVA does have a drawback. It can assess only one dependent variable at a time. This limitation can be an enormous problem in certain circumstances because it can prevent you from detecting effects that actually exist.

ANOVA hereby helps to compare two means at the same time, but can only include one dependent variable in the analysis. On the other hand, MANOVA can determine the relationship between multiple variables concurrently.

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

MANOVA provides a solution for some studies. This statistical procedure tests multiple dependent variables at the same time. By doing so, MANOVA can offer several advantages over ANOVA.

Multivariate analysis of covariance (MANCOVA) is a statistical technique that is the extension of analysis of covariance (ANCOVA). Basically, it is the multivariate analysis of variance (MANOVA) with a covariate(s).).

## ANOVA Restrictions

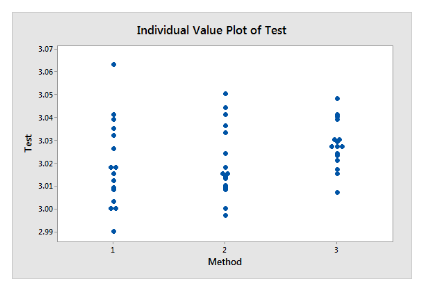
## Regular ANOVA tests can assess only one dependent variable at a time in your model. Even when you fit a general linear model with multiple independent variables, the model only considers one dependent variable. The problem is that these models can’t identify patterns in multiple dependent variables.

## Comparison of MANOVA to ANOVA Using an Example

MANOVA can detect patterns between multiple dependent variables. But, what does that mean exactly? It sounds complex, but graphs make it easy to understand. Let’s work through an example that compares ANOVA to MANOVA.

Suppose we are studying three different teaching methods for a course. This variable is our independent variable. We also have student satisfaction scores and test scores. These variables are our dependent variables. We want to determine whether the mean scores for satisfaction and tests differ between the three teaching methods. Here is the CSV file for the [MANOVA\_example](https://statisticsbyjim.com/wp-content/uploads/2017/03/MANOVA_example.csv).

The graphs below display the scores by teaching method. One chart shows the test scores and the other shows the satisfaction scores. These plots represent how one-way ANOVA tests the data—one dependent variable at a time.

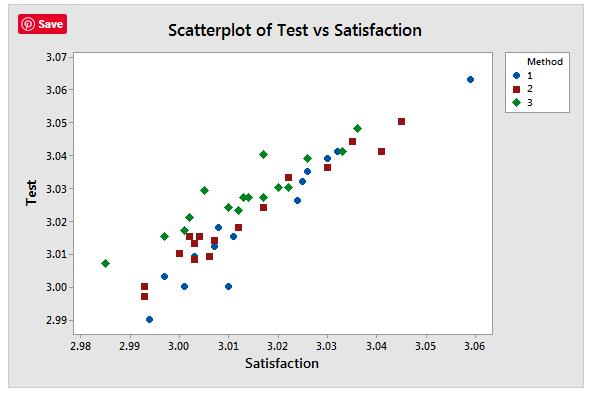




Both of these graphs appear to show that there is no association between teaching method and either test scores or satisfaction scores. The groups seem to be approximately equal. Consequently, it’s no surprise that the one-way ANOVA P-values for both test and satisfaction scores are insignificant (0.923 and 0.254). The teaching method isn’t related to either satisfaction or test scores.

## How MANOVA Assesses the Data

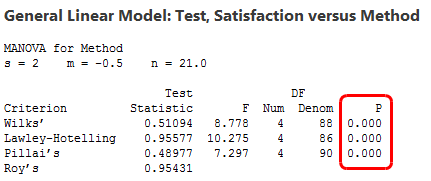
Let’s see what patterns we can find between the dependent variables and how they are related to teaching method. I’ll graph the test and satisfaction scores on the scatterplot and use teaching method as the grouping variable. This multivariate approach represents how MANOVA tests the data. These are the same data, but sometimes how you look at them makes all the difference.



The graph displays a positive [correlation](https://statisticsbyjim.com/glossary/correlation/) between Test scores and Satisfaction. As student satisfaction increases, test scores tend to increase as well. Moreover, for any given satisfaction score, teaching method 3 tends to have higher test scores than methods 1 and 2. In other words, students who are equally satisfied with the course tend to have higher scores with method 3. MANOVA can test this pattern statistically to help ensure that it’s not present by chance.

In your preferred statistical software, fit the MANOVA model so that Method is the independent variable and Satisfaction and Test are the dependent variables.

The MANOVA results are below.



.**When MANOVA Provides Benefits**

Use multivariate ANOVA when your dependent variables are correlated. The correlation structure between the dependent variables provides additional information to the model which gives MANOVA the following enhanced capabilities:

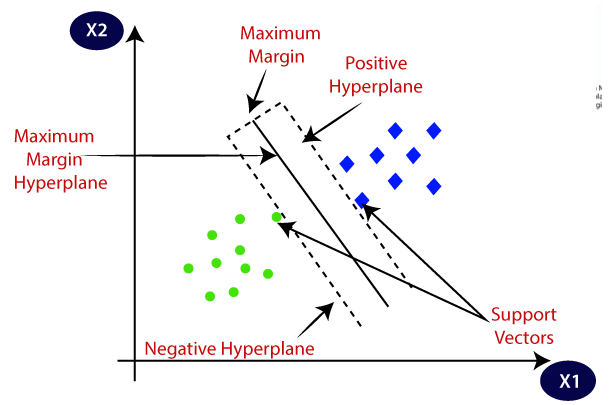
* **Greater statistical power**: When the dependent variables are correlated, MANOVA can identify effects that are smaller than those that regular ANOVA can find.
* **Assess patterns between multiple dependent variables**: The [factors](https://statisticsbyjim.com/glossary/factors/) in the model can affect the relationship between dependent variables instead of influencing a single dependent variable. As the example in this post shows, ANOVA tests with a single dependent variable can fail completely to detect these patterns.
* **Limits the joint error rate**: When you perform a series of ANOVA tests because you have multiple dependent variables, the joint probability of rejecting a true null hypothesis increases with each additional test. Instead, if you perform one MANOVA test, the error rate equals the [significance level](https://statisticsbyjim.com/glossary/significance-level/).

# Support Vector Machine Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

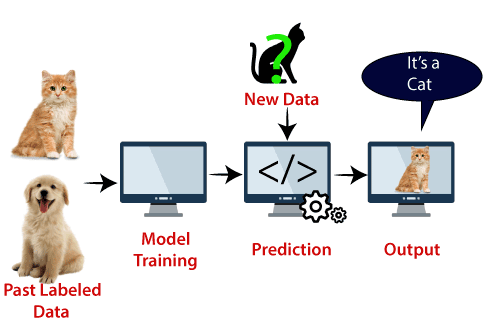
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

****

**Example:**

SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

****

SVM algorithm can be used for **Face detection, image classification, text categorization,** etc.

## ****Types of SVM****

**SVM can be of two types:**

* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

## ****Hyperplane and Support Vectors in the SVM algorithm:****

**Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

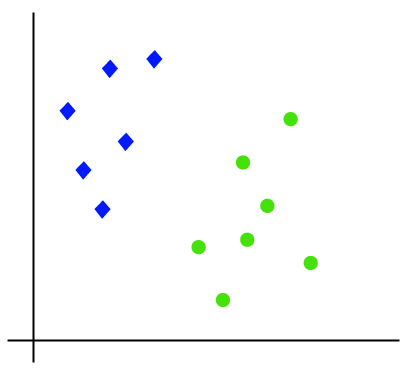
**Support Vectors:**

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

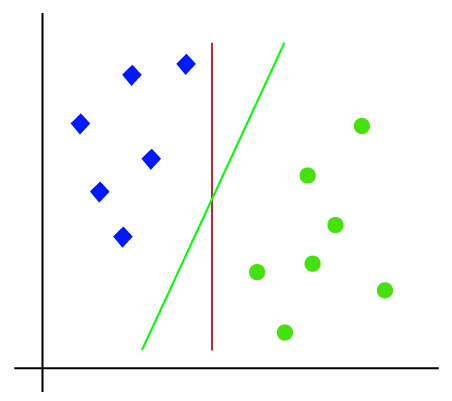
## ****How does SVM works?****

**Linear SVM:**

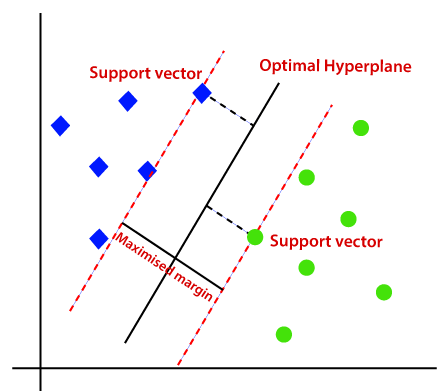
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:

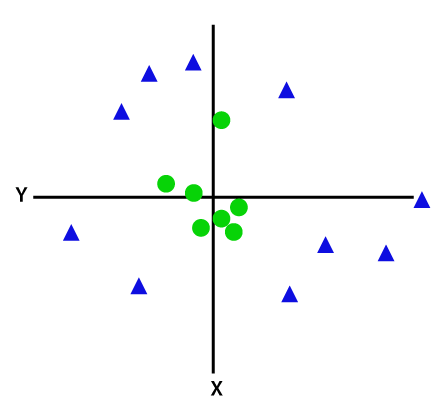


Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



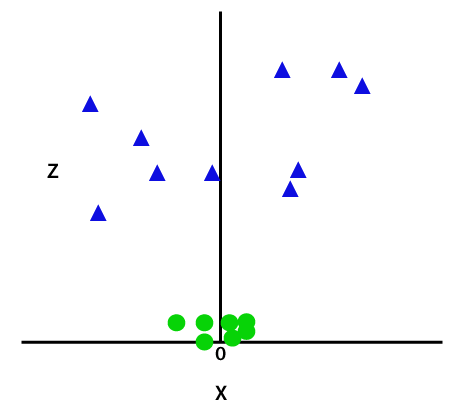
**Non-Linear SVM:**

If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:

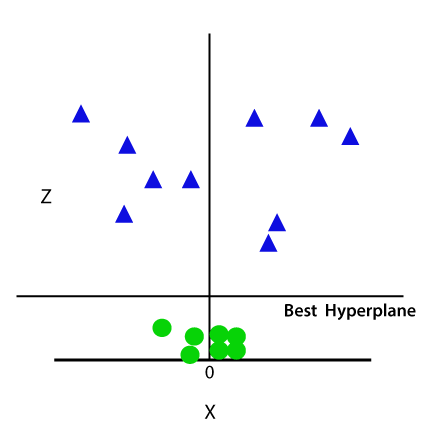


So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as: z=x2 +y2

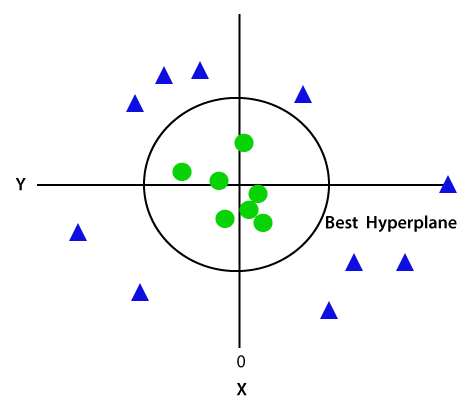
By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



Hence we get a circumference of radius 1 in case of non-linear data.

**Time Series Data**

A time series is a collection of data points gathered over a period of time and ordered chronologically. The primary characteristic of a time series is that it’s indexed or listed in time order, which is a critical distinction from other types of data sets. If you were to plot the points of time series data on a graph, and one of your axes would always be time.

Time series metrics refer to a piece of data that is tracked at an increment in time. For instance, a metric could refer to how much inventory was sold in a store from one day to the next.

Time series data is everywhere, since time is a constituent of everything that is observable. As our world gets increasingly instrumented, sensors and systems are constantly emitting a relentless stream of [time series data](https://www.influxdata.com/time-series-database/). Such data has numerous applications across various industries. Let’s put this in context through some examples.

Examples of time series analysis:

* Electrical activity in the brain
* Rainfall measurements
* Stock prices
* Number of sunspots
* Annual retail sales
* Monthly subscribers
* Heartbeats per minute

## Time series examples

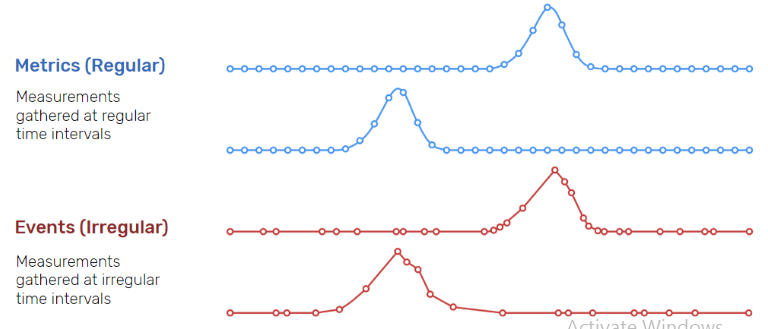
Weather records, economic indicators and patient health evolution metrics—all are time series data. Time series data could also be server metrics, application performance monitoring, network data, sensor data, events, clicks and many other types of analytics data.

## 

**Types of time series data**

Time series data can be classified into two types:

1. Measurements gathered at regular time intervals (metrics)
2. Measurements gathered at irregular time intervals (events)



Because they happen at irregular intervals, events are unpredictable and cannot be modeled or forecasted since forecasting assumes that whatever happened in the past is a good indicator of what will happen in the future.

A time series data example can be any information sequence that was taken at specific time intervals (whether regular or irregular). Common data examples could be anything from heart rate to the unit price of store goods.

### Linear vs. nonlinear time series data

A linear time series is one where, for each data point Xt, that data point can be viewed as a linear combination of past or future values or differences. Nonlinear time series are generated by nonlinear dynamic equations. They have features that cannot be modelled by linear processes: time-changing variance, asymmetric cycles, higher-moment structures, thresholds and breaks. Here are some important considerations when working with linear and nonlinear time series data:

* If a regression equation doesn’t follow the rules for a linear model, then it must be a nonlinear model.
* Nonlinear regression can fit an enormous variety of curves.
* The defining characteristic for both types of models are the functional forms.

## What is time series analysis?

Time series analysis is the collection of data at specific intervals over a period to identify trends, seasonality, and residuals to aid in forecasting a future event. Time series analysis involves inferring what has happened to a series of data points in the past and attempting to predict future values. Analyzing time series data allows for extracting meaningful statistics and other data characteristics. As the name suggests, time series data is a collection of observations created by repeating measurements over time. Once you have that information, you can plot it on a graph and learn more about precisely what you’re tracking.

A very straightforward time series analysis example might be the rise and fall of the temperature over the course of a day. By tracking the specific temperature outside at hourly intervals for 24 hours, you have a complete picture of the rise and fall of the temperature in your area. Then, suppose you know that the next day will be relatively similar in terms of things like precipitation and humidity. In that case, you can make a more educated guess about the temperature at specific times. This analysis is an oversimplified example, yes—but the underlying structure is the same regardless of what it is that you’re talking about.

## What is a time series graph?

Time series graphs are simply plots of time series data on one axis (typically Y) against time on the other axis (typically X). Graphs of time series data points can often illustrate trends or patterns in a more accessible, intuitive way.

## What are time plot statistics?

A time series plot is a graph in which the x-axis represents some measure of time. In fact, the x-axis is labeled as the time-axis. The y-axis represents the variable being measured. Data points are displayed and connected with straight lines in most cases, allowing for interpretation of the resulting graph.

The term ‘time series patterns’ describes long-term changes in the series. Whether measured as a trend, seasonal, or cyclic pattern, the correlation can be calculated in a number of ways (linear, exponential, etc.), and the direction may change at any given time.

Time series data is used in time series analysis (historical or real-time) and time series forecasting to detect and predict patterns — essentially looking at change over time. Following is a brief overview of each.

## Time series analysis methods

Time series analysis is a method of analyzing a series of data points collected over a period of time. In time series analysis, data points are recorded at regular intervals over a set period of time, rather than intermittently or at random.

Time series analysis is the use of statistical methods to analyze time series data and extract meaningful statistics and characteristics about the data. TSA helps identify trends, cycles, and seasonal variances to aid in the forecasting of a future event. Factors relevant to TSA include stationarity, seasonality and autocorrelation.

Time series analysis can be useful to see how a given variable changes over time (while time itself, in time series data, is often the independent variable). Time series analysis can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.