Institute for Advanced Computing And Software Development (IACSD)

Akurdi, Pune

Advanced Analytics using Statistics

Dr. D.Y. Patil Educational Complex, Sector 29, Behind Akurdi Railway Station,

Nigdi Pradhikaran, Akurdi, Pune - 411044.

Introduction to Analytics:

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.

The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision.

The key objective of data analysis :

The goal of data analysis is to organize the information in a logical manner. It aids in the analysis of data from various viewpoints and statistical approaches.

To extract information from data, the process of data analysis employs analytical and logical reasoning.

Data analysis' major goal is to identify meaning in data so that the information gained may be utilized to make better decisions.

Data analytics Life Cycle -The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

1. Discovery
2. Data preparation
3. Model planning
4. Model building implementation
5. Quality assurance
6. Documentation

1.Data Discovery and Formation

Everything begins with a defined goal. In this phase, you'll define your data's purpose and how to achieve it by the time you reach the end of the data analytics lifecycle.

The initial stage consists of mapping out the potential use and requirement of data, such aswhere the information is coming from what story you want your data to convey.

how your organization benefits from the incoming data Basically, as a data analysis expert, you'll need to focus on enterprise requirements related to data, rather than data itself.

Additionally, your work also includes assessing the tools and systems that are necessary to read, organize, and process all the incoming data.

Essential activities in this phase include structuring the business problem in the form of an analytics challenge and formulating the initial hypotheses (IHs) to test and start learning the data

The subsequent phases are then based on achieving the goal that is drawn in this stage.

2. Data Preparation and Processing

The data preparation and processing step involve collecting, processing, and cleansing the accumulated data One of the essential parts of this phase is to make sure that the data you need is actually available to you for processing . The earliest step of the data preparation phase is to collect valuable information and proceed with the data analytics lifecycle in a business ecosystem .

3.Design a Model

The model's building initiates with identifying the relation between data points to select the key variables and eventually find a suitable model .

This step also includes the teamwork to determine the methods, techniques, and workflow to build the model in the subsequent phase.

4. Model Building

This step of data analytics architecture comprises developing data sets for testing, training, and production purposes .The data analytics experts meticulously build and operate the model that they had designed in the previous step . They rely on tools and several techniques like decision trees, regression techniques and neural networks for building and executing the model.

5. Communicating results

The communication step starts with a collaboration with major stakeholders to determine if the project results are a success or failure . The project team is required to identify the key findings of the analysis, measure the business value associated with the result, and produce a narrative to summarise and convey the results to the stakeholders.

6. Documentation

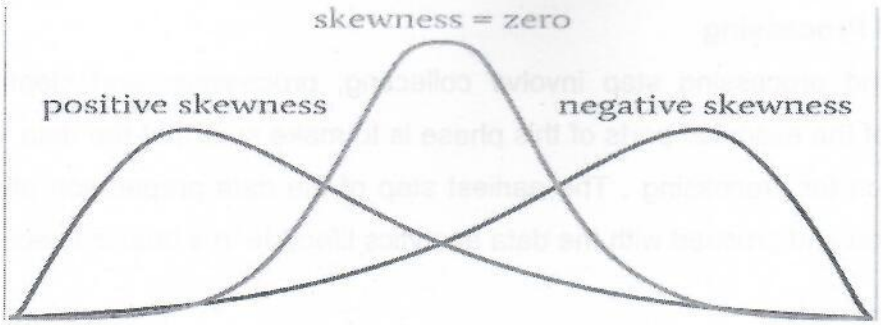
As your data analytics lifecycle draws to a conclusion, the final step is to provide a detailed report with key findings, coding, briefings, technical papers/ documents to the stakeholders Additionally, to measure the analysis's effectiveness, the data is moved to a live environment from the sandbox and monitored to observe if the results match the expected business goal.

Shape of Data -Shape of data is measured by

1. Skewness

Skewness measures the degree of asymmetry exhibited by the data

If skewness equals zero, the histogram is symmetric about the mean Kurtosis measures how peaked the histogram is.



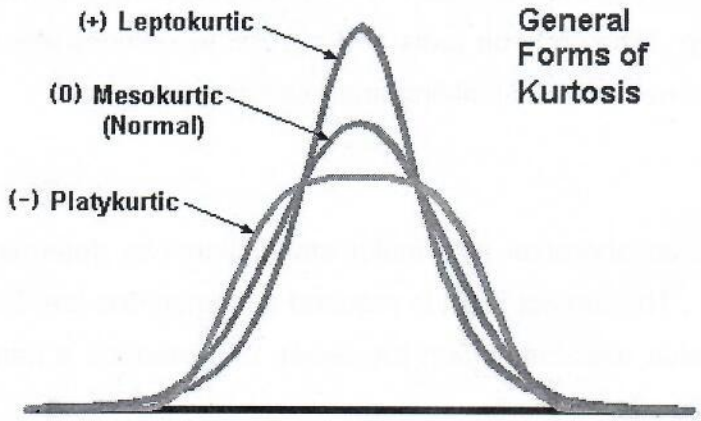
2. Kurtosis - The kurtosis of a normal distribution is 0

Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution .

Distributions with medium kurtosis (medium tails) are mesokurtic.

Distributions with low kurtosis (thin tails) are platykurtic.

Distributions with high kurtosis (fat tails) are leptokurtic.



Evolution of Analytics-

Types of Data Analytics

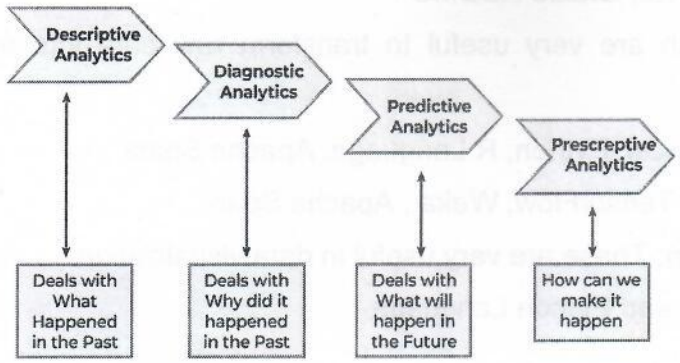
There are four major types of data analytics:

1.Descriptive (business intelligence and data mining) - Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data.

2.Diagnostic analytics - In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

3.Predictive (forecasting)- Historical patterns are used to predict specific Outcomes.

4.Prescriptive (optimization and simulation) -Advanced analytic techniques are applied to make recommendations based on outcomes.



Data analytics Steps-

1) Decide on the objectives: In your data analysis process, you must begin with the right set of questions. Some examples include: How can we reduce production costs without sacrificing quality? What are some ways to increase sales opportunities with our current resources? Do customers view our brand in a favorable way?

2) Set measurement priorities: This step involves processes to identify what to measure and how to measure? For example, should you perform your analysis weekly, monthly or annually? What is your unit of measure? What factors should be included in your analysis?

3)Data collection: Data Collection is the process of collecting information on targeted variables for your analysis. In this step, the emphasis is to ensure collecting accurate and honest data. Thus, before you start your hunt for collecting new data, find what information could be collected from existing data sources.

4)Data cleaning: Data Cleaning is the process of preventing and correcting errors in data. It includes steps such as structuring the data as required for the relevant analysis tools, removing duplicate values, identifying outliers in data, etc

5)Analysis of data: Various data analysis techniques can be used to understand, interpret, and derive conclusions based on the requirements. . Tools like Microsoft Excel, R, Python, Tableau, etc. can be used for performing analysis.

6)Interpreting the results: Does the analysis answer your objectives? Does the analysis help you defend against any objections?

If your interpretation of the data holds up, under all of these questions and considerations, then you likely have come to a productive conclusion. Create a report of your analysis in a format as required by stakeholders for business decisions. The feedback from stakeholders might result in additional analysis.

Data analytics tools:

1. **Database tools**: which are used to store the data

Database Tools: SQL, MongoDB, MySQL, Oracle RDBMS

2. **Data transformation tools**: which are very useful to transform raw data into meaningful information

Data transformation Tools: Microsoft Excel, Python, R Language, Apache Spark

3. Data Modeling Tools: Scikit-Learn, TensorFlow, Weka:, Apache Spark,

Data Modeling Tools in Visualization: These are very useful in data visualization.

Tableau, Power BI, Microsoft Excel, R, and Python Language

Some Key Terms and Definitions –

Statistics - It involves the process of converting raw data into a meaningful, organized, and informative form. The main purpose of using statistics is to plan the collected data in terms of experimental designs and statistical surveys. Statistics is considered a mathematical science that works with numerical data. In short, statistics is a crucial process which helps to make the decision based on the data.

Business Intelligence/Information Systems:

The process of gathering information in the field of business. It can be described as the process of enhancing data into information and then into knowledge. Business intelligence is carried out to gain sustainable competitive advantage, and is a valuable core competence in some instances.

Data Mining:

Data Mining is focused on better understanding characteristics and patterns among variables in large database using a variety of statistical and analytical tools.

Many standard statistical tools as well as more advanced ones are used extensively in data Mining Visualization: Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

Population: Populations is the collection of all possible observations of a specified characteristic of interest. All students taking the statistics course in a business school is an example of population.

Sample: sample is a subset of the population. Suppose you want to select a team of 20 students from 200 students of an MBA program for participating in a management quiz. The total number of students 200 is the population . 20 students selected for the quiz is the sample.

Sample space- A sample space is a collection or a set of possible outcomes of a random

experiment. The sample space is represented using the symbol, "S". The subset of possible outcomes of an experiment is called events. A sample space may contain a number of outcomes that depends on the experiment. If it contains a finite number of outcomes, then it is known as discrete or finite sample spaces. Consider a random experiment. The set of all the possible outcomes is called the sample space
of the experiment and is usually denoted by S. Any subset E of the sample space S is called an event. Here is an example

Tossing a coin. The sample space is S = {H, T }. E = {H} is an event.

Event An event is a subset of a sample space. Refer to Example "the sum of the dots is 6" is an event. It is expressible of a set of elements E = {(1, 5) (2, 4) (3, 3) (4, 2) (5, 1)}

Probability

The probability of an event is basically a number between 0 and 1, where, on an estimate, 0 designates the impossibility of the event, and 1 designates certainty.

Events: Events are actually a subset of possible outcomes of an experiment.

Probability can be-

Joint Probability: Probability of events *A* and *B*.

Marginal Probability: Probability of event X=*A* given variable *Y*.

Conditional Probability: Probability of event *A* given event *B*.

These types of probability form the basis of much of predictive modeling with problems such as classification and regression.

Joint Probability of Two Variables -

We may be interested in the probability of two simultaneous events, e.g. the outcomes of two different random variables.

The probability of two (or more) events is called the joint probability. The joint probability of two or more random variables is referred to as the joint probability distribution.

For example, the joint probability of event *A* and event *B* is written formally as:

P(A and B)

The "*and*" or conjunction is denoted using the upside down capital "*U*" operator "^" or sometimes a comma ",".

P(A ^ B)

P(A, B)

The joint probability for events *A* and *B* is calculated as the probability of event *A* given event *B* multiplied by the probability of event *B*.

This can be stated formally as follows:

P(A and B) = P(A given B) * P(B)

The calculation of the joint probability is sometimes called the fundamental rule of probability or the "*product rule*" of probability or the "chain rule" of probability.

Here, *P(A given B)* is the probability of event A given that event B has occurred, called the conditional probability, described below.

The joint probability is symmetrical, meaning that *P(A and B)* is the same as *P(B and A)*. The calculation using the conditional probability is also symmetrical, for example:

P(A and B) = P(A given B) * P(B) = P(B given A) * P(A)

Marginal Probability -

We may be interested in the probability of an event for one random variable, irrespective of the outcome of another random variable.

For example, the probability of *X=A* for all outcomes of *Y*.

The probability of one event in the presence of all (or a subset of) outcomes of the other random variable is called the marginal probability or the marginal distribution. The marginal probability of one random variable in the presence of additional random variables is referred to as the marginal probability distribution.

It is called the marginal probability because if all outcomes and probabilities for the two variables were laid out together in a table (*X* as columns, *Y* as rows), then the marginal probability of one variable (*X*) would be the sum of probabilities for the other variable (Y rows) on the margin of the table.

There is no special notation for the marginal probability; it is just the sum or union over all the probabilities of all events for the second variable for a given fixed event for the first variable.

P(X=A) = sum P(X=A, Y=yi) for all y

This is another important foundational rule in probability, referred to as the "*sum rule*."

The marginal probability is different from the conditional probability (described next) because it considers the union of all events for the second variable rather than the probability of a single event.

Conditional Probability -

We may be interested in the probability of an event given the occurrence of another event.

The probability of one event given the occurrence of another event is called the conditional probability. The conditional probability of one to one or more random variables is referred to as the conditional probability distribution.

For example, the conditional probability of event *A* given event *B* is written formally as:

P(A given B)

The "*given*" is denoted using the pipe "|" operator; for example:

P(A | B)

The conditional probability for events *A* given event *B* is calculated as follows:

P(A given B) = P(A and B) / P(B)

This calculation assumes that the probability of event *B* is not zero, e.g. is not impossible.

The notion of event *A* given event *B* does not mean that event *B* has occurred (e.g. is certain); instead, it is the probability of event *A* occurring after or in the presence of event *B* for a given trial.

Bayes' Theorem-

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

Bayes theorem states that the conditional probability of an event A, given the occurrence of another event B, is equal to the product of the likelihood of B, given A and the probability of A.

P(A|B)=P(B|A)P(A)/P(B)

Here, P(A) = how likely A happens(Prior knowledge)- The probability of a hypothesis is true before any evidence is present.

P(B) = how likely B happens(Marginalization)- The probability of observing the evidence.

P(A/B) = how likely A happens given that B has happened(Posterior)-The probability of a hypothesis is true given the evidence.

P(B/A) = how likely B happens given that A has happened(Likelihood)- The probability of seeing the evidence if the hypothesis is true.

Metrics and Data Classification

Nominal Scale - When numbers assigned to objects serve as labels for identification or categorization, then such numbers are in nominal scale. Such numbers have no quantitative meaning.

Ordinal Scale- When assigned numbers indicate relation between entities in terms of greater than, equal or less than but do not state how much greater than or less than, then the scale is called ordinal scale.

Interval Scale- When assigned numbers are such that difference in numbers is valid but not

ratios, then the scale is called interval scale.

Ratio Scale - When a scale contains absolute zero, it is called ratio scale

Scales	Nominal	Ordinal	Interval	Ratio
Description	Label	Label, order	Label, order, equal distance units	Label, order, equal distance units and absolute zero
Nature	Qualitative	Qualitative	Quantitative	Quantitative
Data	Discrete Data	Discrete Data	Continuous Data	Continuous Data
Test	Non- Parametric Test	Non- Parametric Test	Parametric Test	Parametric Test
Example	Gender can be male or female. Eye colour can be black, blue, green.	Height can be short, middle and tall. Customer satisfaction can be unhappy, neutral, happy.	What is temperature in your city? It can be. • below 0°C. • between 0°C - 20°C. • between 20°C - 40°C and • above 40°C.	What is your weight in kilograms? • Less than 50 KG. • 51- 100 KG. • 101- 150 KG. • More than 150 KG.

Random variable -

A random variable is a rule that assigns a numerical value to each outcome in a sample space. Random variables may be either discrete or continuous. A random variable is said to be discrete if it assumes only specified values in an interval. Otherwise, it is continuous. We generally denote the random variables with capital letters such as X and Y. When X takes values 1, 2, 3, ..., it is said to have a discrete random variable.

Types of Random Variable

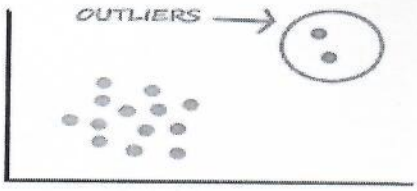
As discussed in the introduction, there are two random variables, such as:

Discrete Random Variable - A discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, ... and so on. The probability distribution of a random variable has a list of probabilities compared with each of its possible values known as probability mass function.

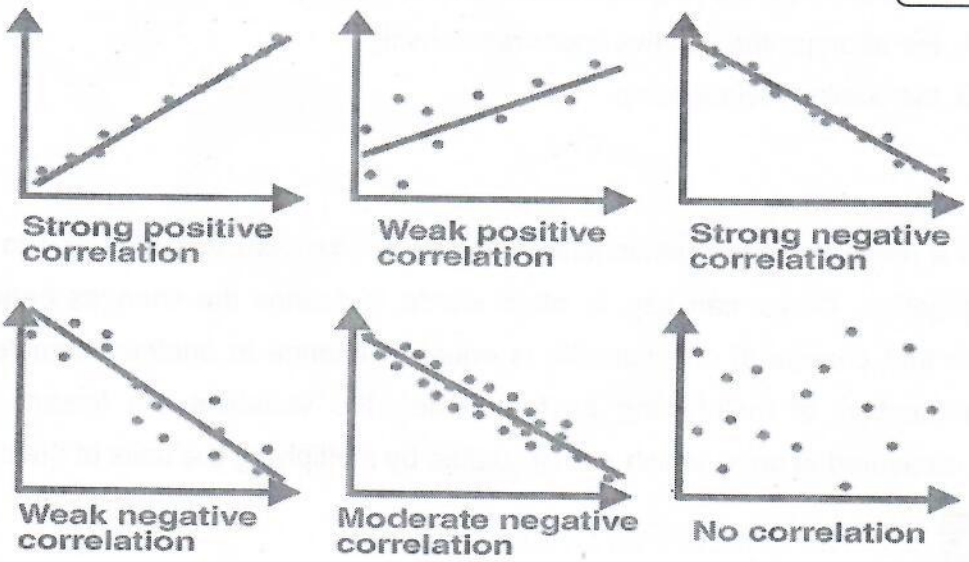
Continuous Random Variable - A numerically valued variable is said to be continuous if, in any unit of measurement, whenever it can take on the values a and b. If the random variable X can assume an infinite and uncountable set of values, it is said to be a continuous random variable.

Outlier.

In a dataset, Outliers are values that differ significantly from the mean of characteristic features of a dataset. With the help of an outlier, we can determine either variability in the measurement or an experimental error.



Correlation - Correlation refers to the statistical relationship between the two entities. It measures the extent to which two variables are linearly related. For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people.



Correlation Analysis-

When we try to find the relationship between two variables, we use correlation analysis.

Correlation Analysis provides two pieces of information

The strength of relationship

The direction of relationship

Karl Pearson's Coefficient of Correlation(r)

The strength and direction of relationship is calculated by Karl Pearson's coefficient of correlation, denoted by 'r' and defined by the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Features of Correlation Coefficient, r –

Unit free

Ranges between –1 to 1

If r= 1, Perfect Positive Linear Relationship.

If r= -1, Perfect Negative Linear Relationship

The closer to –1, the stronger the negative linear relationship

The closer to 1, the stronger the positive linear relationship

The closer to 0, the weaker Relationship.

Covariance

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Covariance Formula

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

Measures of central tendency

Measures of central tendency yield information about "particular places or locations in a group of numbers."

Common Measures -

Mean

Median

mode

Percentiles

Quartiles

1. mean - mean represents the average of the given collection of data. It is applicable for both continuous and discrete data.It is equal to the sum of all the values in the collection of data divided by the total number of values.

Suppose we have *n* values in a set of data namely as x_1, x_2, x_3, …, x_n, then the mean of data is given by:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots \ldots + x_n}{n}$$

2.Median- Given that the data collection is arranged in ascending or descending order, the following method is applied:

If number of values or observations in the given data is odd, then the median is given by $[(n+1)/2]^{th}$ observation.

If in the given data set, the number of values or observations is even, then the median is given by the average of $(n/2)^{th}$ and $[(n/2)+1]^{th}$ observation.

3.mode -The most frequently occurring value in a data set.

4.Quartiles- Measures of central tendency that divide a group of data into four subgroups. In statistics, **Quartiles** are the set of values which has three points dividing the data set into four identical parts

Q_1: 25% of the data set is below the first quartile

Q_2: 50% of the data set is below the second quartile

Q_3: 75% of the data set is below the third quartile

Measures of Dispersion -

The averages and measures of central tendency give one single figure which is a representative of the entire data. However , it does not indicate the variations or dispersions in the values of the data.

Thus , if we have two or more sets of observations with the same arithmetic mean but with different variations in the values and average are inadequate to describe a distribution and the knowledge of the dispersion which measures the extent to which the items of the data vary from the average is necessary.

The measures of dispersions are:

- **Range**

- **Quartile Deviation**
- **Mean Deviation**
- **Standard Deviation**

Range- The difference between the largest and the smallest values in a set of data.

If given data is – 2,4,5,7,9,10

Range =10-1=9

Quartile Deviation-

The values (three in numbers) which divide the given data into four equal parts are knows as the quartiles Q1,Q2 and Q3. The first quartile (Q1) is the value of the (N/4)th observation(i.e. it has 25% of the observations below it and 75% of the observations above it and the third quartile (Q3) is the value of the (3N/4)th observation. Q2 divides the data into two equal parts and hence coincides with the median.

Q1 = Value of (N+1)/4 observation.

And Q3= Value of [3(N+1/4)] observation.

Quartile Deviation = Q3-Q1/2

Mean Deviation –

It is observed by taking the arithmetic mean of the absolute deviations of all the values from the average. Absolute deviation is obtained by ignoring the sign (+ or -) of the deviation, thus treating it is always as positive. Mean deviation is also known as Mean(absolute) deviation or average deviation.

$$\textbf{Mean Deviation} = \frac{\sum_1^n |x_i - \mu|}{n}$$

$$\text{where the mean is } \mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Standard Deviation - It is defined as the positive square root of the arithmetic mean of the squares of the deviations of all the observations from their arithmetic mean.

Population Standard Deviation Formula $\sigma = \sqrt{\frac{\sum(X-\mu)^2}{n}}$

Sample Standard Deviation Formula $s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$

Common Data Types

Before we jump on to the explanation of distributions, let's see what kind of data we can encounter. The data can be discrete or continuous.

Discrete Data, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5, or 6, not 1.5 or 2.45. (Discrete Probability

Distribution)

Continuous Data can take any value within a given range. The range may be finite or infinite. For example, a girl's weight or height, the length of the road. The weight of a girl can be any value – 54 kgs, 54.5 kgs, or 54.5436kgs. (Continuous Probability Distribution)

Probability distribution-

Probability is *the systematic consideration of the outcomes of a random experiment.* For example, when we do the coin toss, there are two possible outcomes – heads or tails. Each of these options has the same probability of the number of successes occurring during each flip. The probability of either heads or tails on a single coin flip is ½, which is symmetric distribution in probability.

Types of Distributions

Bernoulli Distribution -

Let's start with the easiest distribution, which is Bernoulli Distribution.

All you cricket junkies out there! At the beginning of any cricket match, how do you decide who will bat or ball? A toss! It all depends on whether you win or lose the toss, right? Let's say if the toss results in a head, you win. Else, you lose. There's no midway.

A **Bernoulli distribution** has only two bernoulli trials or possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X with a Bernoulli distribution can take the value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

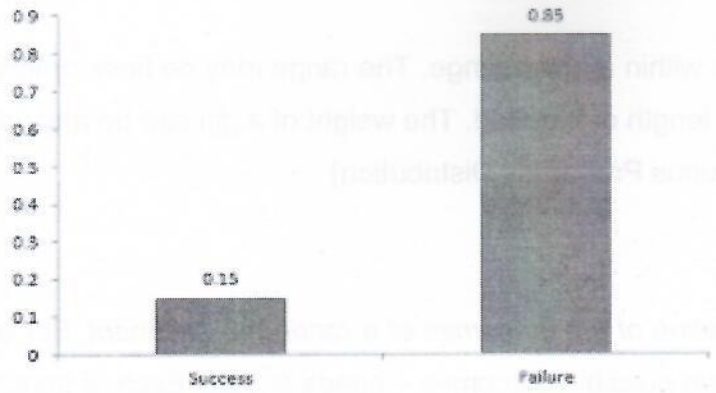
Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure. Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is given by: px(1-p)1-x where x € (0, 1). It can also be written as

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The probabilities of success and failure need not be equally likely, like the result of a fight between Undertaker and me. He is pretty much certain to win. So, in this case probability of my success is 0.15, while my failure is 0.85

Here, the probability of success(p) is not the same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.



Here, the probability of success = 0.15, and the probability of failure = 0.85. The expected value is exactly what it sounds like. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows:

$E(X) = 1*p + 0*(1-p) = p$

The variance of a random variable from a bernoulli distribution is:

$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p)$

There are many examples of Bernoulli distribution, such as whether it will rain tomorrow or not, where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

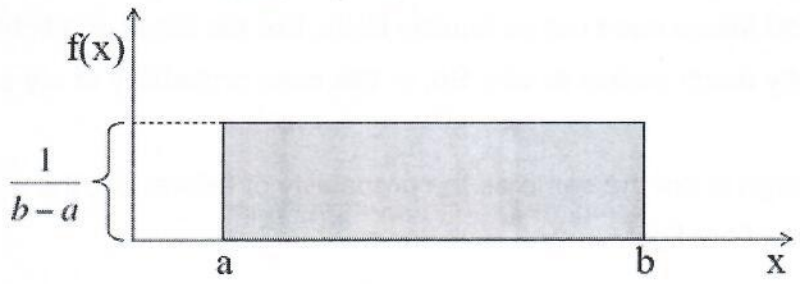
Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely, which is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$

The graph of a uniform distribution curve looks like



You can see that the shape of the Uniform distribution curve is rectangular, the reason why

Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

The number of bouquets sold daily at a flower shop is uniformly distributed, with a maximum of 40 and a minimum of 10.

Let's try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution are:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = $(b-a)^2/12$

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Binomial Distribution

Let's get back to cricket. Suppose you won the toss today, indicating a successful event. You toss again, but you lose this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let's assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, the probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is the same for all the trials is called a Binomial Distribution.

Based on the above explanation, the properties of a Binomial Distribution are:

Each trial is independent.

There are only two possible outcomes in a trial – success or failure.

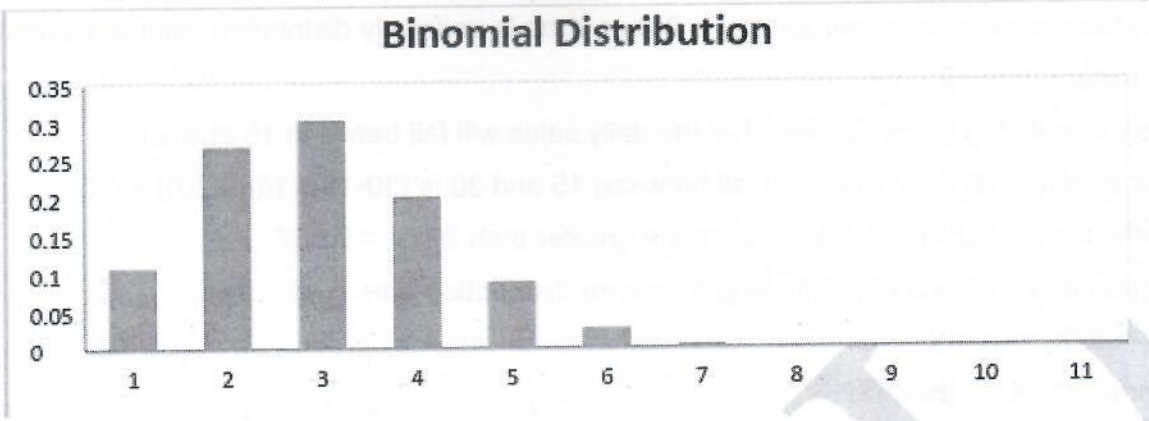
A total number of n identical trials are conducted.

The probability of success and failure is the same for all trials. (Trials are identical.)

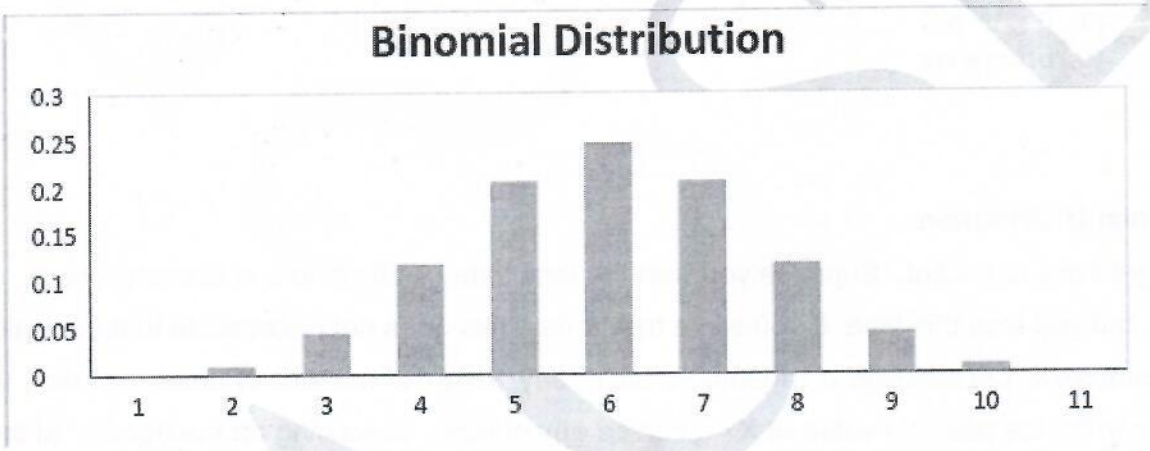
The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like this.



Now, when the probability of success = probability of failure, in such a situation, the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean -> $\mu = n*p$

Variance -> $Var(X) = n*p*q$

Normal Distribution or Gaussian Distribution

The **normal distribution** represents the behavior of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

The mean, median, and mode of the distribution coincide.

The curve of the distribution is bell-shaped and symmetrical about the line $x=\mu$.

The total area under the curve is 1.

Exactly half of the values are to the left of the center, and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials

approaches infinity, then the shapes will be quite similar.

The PDF of a random variable X, following a normal distribution, is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

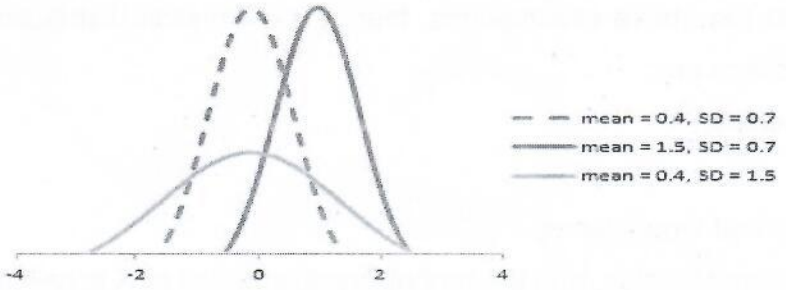
The mean and variance of a random variable X, which is said to be normally distributed, is given by:

Mean -> E(X) = μ

Variance -> Var(X) = σ^2

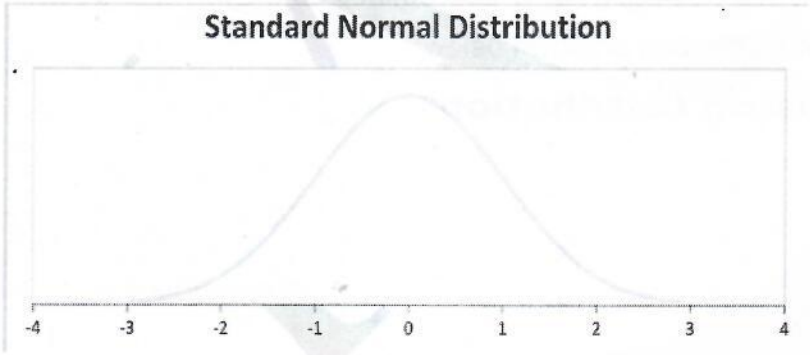
Here, μ (mean) and σ (standard deviation) are the parameters.

The graph of a random variable X ~ N (μ, σ) is shown below.



A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1. For such a case, the PDF becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad for -\infty < x < \infty$$



Poisson Distribution

Suppose you work at a call center; approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are:

The number of emergency calls recorded at a hospital in a day.

The number of thefts reported in an area in a day.

The number of customers arriving at a salon in an hour.

The number of suicides reported in a particular city.

The number of printing errors on each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called a **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal its probability over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions, then it is a Poisson distribution. Some notations used in Poisson distribution are:

λ is the rate at which an event occurs,

t is the length of a time interval,

And X is the number of events in that time interval.

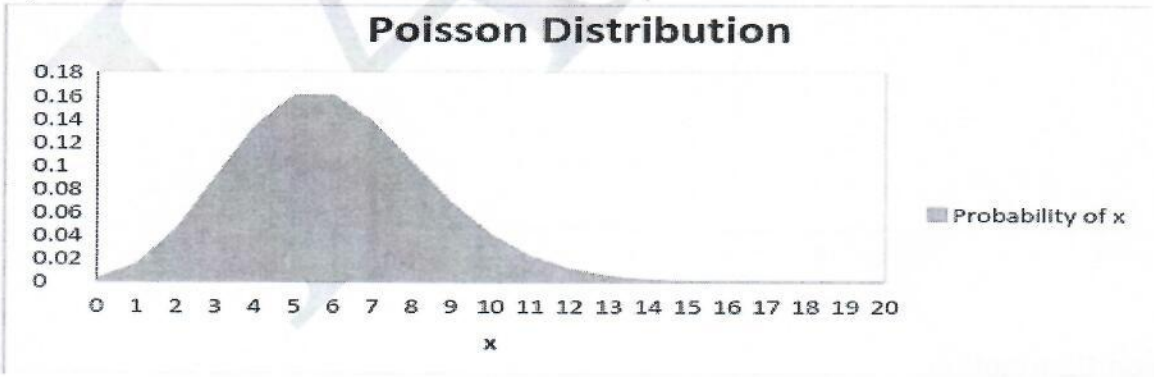
Here, X is called a Poisson Random Variable, and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

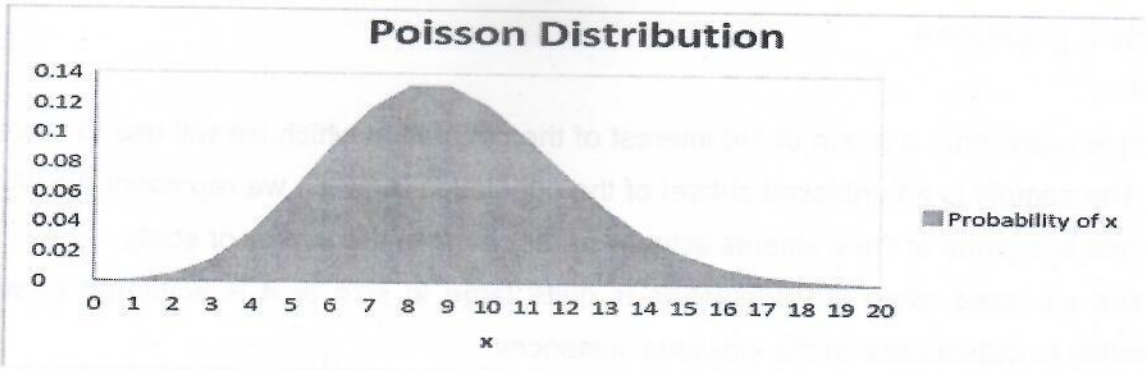
The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots..$$

The mean μ is the parameter of this distribution. μ is also defined as the λ times the length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to the increase in the mean.



It is perceptible that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution:

Mean -> $E(X) = \mu$

Variance -> $Var(X) = \mu$

Exponential Distribution -

Let's consider the call center example one more time. What about the interval of time between the calls? Here, the exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

The exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an **exponential distribution** with PDF:

$f(x) = \{ \lambda e^{-\lambda x}, \ x \geq 0$

And parameter **$\lambda > 0$,** which is also called the rate.

For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

Mean -> $E(X) = 1/\lambda$

Variance -> $Var(X) = (1/\lambda)^2$

Also, the greater the rate, the faster the curve drops, and the lower the rate, the flatter the curve. This is explained better with the graph shown below.

To ease the computation, there are some formulas given below.

$P\{X \leq x\} = 1 - e^{-\lambda x}$ corresponds to the area under the density curve to the left of x.

$P\{X > x\} = e^{-\lambda x}$ corresponds to the area under the density curve to the right of x.

$P\{x1 < X \leq x2\} = e^{-\lambda x1} - e^{-\lambda x2}$, corresponds to the area under the density curve between x1 and x2.

Sample & population

Sample:

A sample represents a group of the interest of the population which we will use to represent the data. The sample is an unbiased subset of the population in which we represent the whole data. A sample is a group of the elements actually participating in the survey or study.

Samples are used when the population is quite large in size or it is scattered or when it is impossible to collect data on the individual instances.

example: Let us assume the population of India is 10 million, and recent elections were conducted in India between two parties 'party A ' and 'party B' now researchers want to find which party is winning so here we will create a group of few people lets say 10,000 from different regions and age groups so that sample is not biased. Then ask them who they voted we can get the exit poll.

Population:

A complete collection of the objects or measurements is called the population or else everything in the group we want to learn about will be termed as population. or else In statistics population is the entire set of items from which data is drawn in the statistical study. It can be a group of individuals or a set of items.

For example: let us assume that there are 50 employees in my company, so 50 people is a complete set hence it will represent the population of my company.

Sample

A subset from a larger data set.

Population

The larger data set or idea of a data set.

N (n)

The size of the population (sample).

Random sampling

Drawing elements into a sample at random.

Stratified sampling

Dividing the population into strata and randomly sampling from each strata.

Stratum (pl., strata)

A homogeneous subgroup of a population with common characteristics.

Simple random sample

The sample that results from random sampling without stratifying the population.

Bias

Systematic error.

Sample bias

A sample that misrepresents the population.

SAMPLING METHODS –

1. PROBABILITY SAMPLING - Each element of the population has a known and equal probability of selection in the sample.

It relies on a random selection of elements

It is used in case of 'Finite Population'

TYPES OF PROBABILITY SAMPLING –

- **Simple Random Sampling**
- **Systematic Sampling**
- **Stratified Sampling**
- **Cluster Sampling**
- **Area Sampling**

Simple Random Sampling - A method of sampling that relies on a random or chance selection method so that every element of the sampling frame has a known probability of being selected.

Systematic Random Sampling: - Uses an ordered list.

E.g. selecting every 10^{th} name from the membership register.

Example – 300 Students(Population)

We want to select 15.

Stratified Random Sampling

A method of sampling in which sample elements are selected separately from population strata that are identified in advance by the researcher. The strata will be formed on the basis of Homogeneity .(Homogenous group)

Cluster Sampling

Sampling in which elements are selected in two or more stages, with the first stage being the random selection of naturally occurring clusters and the last stage being the random selection of elements within clusters.

Area Sampling

Clusters are heterogeneous in nature.

When the clusters are selected on the basis of geographical area , then it is also called as Area Sampling

NON-PROBABILITY SAMPLING - It is not possible to specify , for each element of the population, the relative likelihood that it will be included in the sample.

It is used in case of 'Infinite Population'

Random selection of elements is not necessary.

It relies on personal judgment of the researcher.

The researcher can arbitrarily or consciously decide what elements to include in the sample.

TYPES OF PROBABILITY SAMPLING –

- Accidental/Convenience
- Quota Sampling
- Purposive/Judgmental
- Snowball Sampling

1.Convenience sampling/ Accidental Sampling: Samples are drawn at the convenience of the interviewer.

E.g. in mall intercept interviews, the interviewer selects people who are accessible and willing to participate.

2. Judgment sampling/ Purposive Sampling: The researcher believes that the sample of key respondents possesses the attributes valuable to the researcher.

E.g. selecting sarpanch of a gramsabha in a rural study.

3. Quota Sampling- Quota sampling insures inclusion of diverse elements of the population in the sample and make sure that these diverse elements take account of the proportions in which they occur in the population.

- For example, we take a sample from a population with equal number of boys and girls, and that there is a difference between the two groups in the characteristic we wish to study and we fail to interview any girls, the results of the study would almost certainly be extremely misleading generalizations about the population.

4. Snowball sampling: Also called referral sampling: one sampling unit, or subject refers another, who refers another, where the characteristics is dispersed thinly in the population and so on. E.g. selecting a sample of mountaineers, Selection of specialized doctors.

Central limit theorem

Central limit theorem is a statistical theory which states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard

deviation $\sigma / \sqrt{n}$.

As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean. If the sample size is small, the actual distribution of the data may or may not be normal, but as the sample size gets bigger, it can be approximated by a normal distribution. This statistical theory is useful in simplifying analysis while dealing with stock indexes and many more.

Statistical Inference Terminology

Hypothesis Tests

Hypothesis testing can be defined as a statistical tool that is used to identify if the results of an experiment are meaningful or not. It involves setting up a null hypothesis and an alternative hypothesis. These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa. An example of hypothesis testing is setting up a test to check if a new medicine works on a disease in a more efficient manner.

Hypothesis tests, also called *significance tests*, are ubiquitous in the traditional statistical analysis.. Their purpose is to help you learn whether random chance might be responsible for an observed effect.

Key Terms for Hypothesis Tests

Null hypothesis

Alternative hypothesis

An analyst performs hypothesis testing on a statistical sample to present evidence of the plausibility of the null hypothesis. Measurements and analyses are conducted on a random sample of the population to test a theory. Analysts use a random population sample to test two hypotheses: the null and alternative hypotheses.

The null hypothesis is typically an equality hypothesis between population parameters; for example, a null hypothesis may claim that the population means return equals zero. The alternate hypothesis is essentially the inverse of the null hypothesis (e.g., the population means the return is not equal to zero). As a result, they are mutually exclusive, and only one can be correct. One of the two possibilities, however, will always be correct.

Null Hypothesis and Alternate Hypothesis

The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

H0 is the symbol for it, and it is pronounced H-naught.

The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.

Let's understand this with an example.

A sanitizer manufacturer claims that its product kills 95 percent of germs on average.

To put this company's claim to the test, create a null and alternate hypothesis.

H0 (Null Hypothesis): Average = 95%.

Alternative Hypothesis (H1): The average is less than 95%.

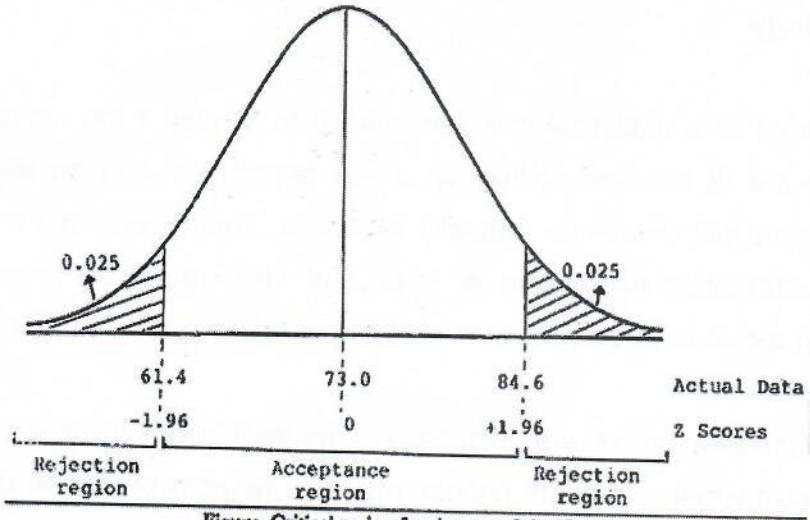


Figure. Critical region for the test of significance.

Steps of Hypothesis Testing

Step 1: Specify Your Null and Alternate Hypotheses

It is critical to rephrase your original research hypothesis (the prediction that you wish to study) as a null (Ho) and alternative hypothesis so that you can test it quantitatively. Your first hypothesis, which predicts a link between variables, is generally your alternate hypothesis. The null hypothesis predicts no link between the variables of interest.

Step 2: Gather Data

For a statistical test to be legitimate, sampling and data collection must be done in a way that is meant to test your hypothesis. You cannot draw statistical conclusions about the population you are interested in if your data is not representative.

Step 3: Conduct a Statistical Test

Other statistical tests are available, but they all compare within-group variance (how to spread out the data inside a category) against between-group variance (how different the categories are from one another). If the between-group variation is big enough that there is little or no overlap between groups, your statistical test will display a low p-value to represent this. This suggests that the disparities between these groups are unlikely to have occurred by accident. Alternatively, if there is a large within-group variance and a low between-group variance, your statistical test will show a

high p-value. Any difference you find across groups is most likely attributable to chance. The variety of variables and the level of measurement of your obtained data will influence your statistical test selection.

Step 4: Determine Rejection Of Your Null Hypothesis

Your statistical test results must determine whether your null hypothesis should be rejected or not. In most circumstances, you will base your judgment on the p-value provided by the statistical test. In most circumstances, your preset level of significance for rejecting the null hypothesis will be 0.05 - that is, when there is less than a 5% likelihood that these data would be seen if the null hypothesis were true. In other circumstances, researchers use a lower level of significance, such as 0.01 (1%). This reduces the possibility of wrongly rejecting the null hypothesis.

Step 5: Present Your Results

The findings of hypothesis testing will be discussed in the results and discussion portions of your research paper, dissertation, or thesis. You should include a concise overview of the data and a summary of the findings of your statistical test in the results section. You can talk about whether your results confirmed your initial hypothesis or not in the conversation. Rejecting or failing to reject the null hypothesis is a formal term used in hypothesis testing. This is likely a must for your statistics assignments.

One-Tailed and Two-Tailed Hypothesis Testing

The One-Tailed test, also called a directional test, considers a critical region of data that would result in the null hypothesis being rejected if the test sample falls into it, inevitably meaning the acceptance of the alternate hypothesis.

In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.

In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.

If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

Right Tailed Hypothesis Testing

If the larger than (>) sign appears in your hypothesis statement, you are using a right-tailed test, also known as an upper test. Or, to put it another way, the disparity is to the right. For instance, you can contrast the battery life before and after a change in production. Your hypothesis statements can be the following if you want to know if the battery life is longer than the original (let's say 90 hours):

The null hypothesis is (H0 <= 90) or less change.

A possibility is that battery life has risen (H1) > 90.

The crucial point in this situation is that the alternate hypothesis (H1), not the null hypothesis, decides whether you get a right-tailed test.

Left Tailed Hypothesis Testing

Alternative hypotheses that assert the true value of a parameter is lower than the null hypothesis are tested with a left-tailed test; they are indicated by the asterisk "<".

Example:

Suppose H0: mean = 50 and H1: mean not equal to 50

According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.

In a similar manner, if H0: mean >=50, then H1: mean <50

Here the mean is less than 50. It is called a One-tailed test.

Type 1 and Type 2 Error

A hypothesis test can result in two types of errors.

Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.

Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Example:

Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.

H0: Student has passed

H1: Student has failed

Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks [H0 was true].

Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing marks [H1 is true].

Level of Significance

The alpha value is a criterion for determining whether a test statistic is statistically significant. In a statistical test, Alpha represents an acceptable probability of a Type I error. Because alpha is a probability, it can be anywhere between 0 and 1. In practice, the most commonly used alpha values are 0.01, 0.05, and 0.1, which represent a 1%, 5%, and 10% chance of a Type I error, respectively (i.e. rejecting the null hypothesis when it is in fact correct).

A confidence interval shows the probability that a parameter will fall between a pair of values

around the mean. Confidence intervals show the degree of uncertainty or certainty in a sampling method.

P-Value

A p-value is a metric that expresses the likelihood that an observed difference could have occurred by chance. As the p-value decreases the statistical significance of the observed difference increases. If the p-value is too low, you reject the null hypothesis.

Here you have taken an example in which you are trying to test whether the new advertising campaign has increased the product's sales. The p-value is the likelihood that the null hypothesis, which states that there is no change in the sales due to the new advertising campaign, is true. If the p-value is .30, then there is a 30% chance that there is no increase or decrease in the product's sales. If the p-value is 0.03, then there is a 3% probability that there is no increase or decrease in the sales value due to the new advertising campaign. As you can see, the lower the p-value, the chances of the alternate hypothesis being true increases, which means that the new advertising campaign causes an increase or decrease in sales.

Parametric Tests: ANOVA, t-test

The one-way analysis of variance is used to test the claim that three or more population means are equal

This is an extension of the two independent samples t-test

The *response* variable is the variable you're comparing

The *factor* variable is the categorical variable being used to define the groups

We will assume *k* samples (groups)

The *one-way* is because each value is classified in exactly one way

Examples include comparisons by gender, race, political party, color, etc.

As an analyst, you might use Analysis of Variance (ANOVA) to test a particular hypothesis. You'd use ANOVA to figure out how your various groups react, with the null hypothesis being that the means of the various groups are equal. If the difference between the two populations is statistically significant, then the two populations are unequal.

One-Way ANOVA

The most common method of performing an ANOVA test is one-way ANOVA. The one-way ANOVA means that the analysis of variance has one independent variable.

You can use the one-way ANOVA to see if there are any significant differences between the means of your independent variables. When you know how each independent variable's mean

differs from the others, you can figure out which of them is linked to your dependent variable and start to figure out what's driving that behaviour.

Population normality : Data is numerical data representing samples from normally distributed populations.

ANOVA test is a parametric test which assumes -

Homogeneity of Variance: the variances of the groups are "similar", P-VALUE >0.05.

The sizes of the groups are "similar"

The groups should be independent.

The residuals are normally distributed.

Two way ANOVA

The two-way analysis of variance is a variation of the one-way analysis. There are two independent variables in this equation (hence the name two-way). Factors are the two independent variables in a two-way ANOVA. The concept is that the dependent variable is influenced by two variables, or factors.

Assumptions of Two way ANOVA –

Population normality : Data is numerical data representing samples from normally distributed populations.

Homogeneity of Variance: the variances of the groups are "similar"

The sizes of the groups are "similar"

The groups should be independent.

The residuals are normally distributed

What Is a T-Test?

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

Understanding the T-Test

A t-test compares the average values of two data sets and determines if they came from the same population. In the above examples, a sample of students from class A and a sample of students from class B would not likely have the same mean and standard deviation. Similarly, samples

taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and standard deviation.

Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement. It assumes a null hypothesis that the two means are equal.

Using the formulas, values are calculated and compared against the standard values. The assumed null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance.

one sample t-test

In one sample t –test, we have only 1 group; want to test the mean value of the group against a hypothetical mean.

Assumptions of one sample t-test-

The data should be continuous (metric scale).

The data follow the normal probability distribution.

The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Independent samples t-test

Independent samples t-test compares the mean value of 2 groups. Groups are independent to each other and people randomly assigned to a single group.

Assumptions of Independent Samples t-test

The data of response variable should be continuous (metric).

The data follow the normal probability distribution.

The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Paired t-test

Paired t-test has two means. Either same people in both groups before and after, or people are related, e.g., husband-wife.

Paired t-test Assumptions

The data should be continuous (metric-Interval and Ratio).

The data, i.e., the differences for the matched-pairs, follow a normal probability distribution.

The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Non-parametric Tests- chi-Square, U-Test

1. chi-Square Test -

Chi-square test is used for finding significant relations. It is used to determine if categorical data shows dependency or the two classifications are independent.

There are two popular types of Chi-Square tests.

Chi-square test for goodness of fit- analysis of single categorical variable. The chi-square is used to find the bias of respondents regarding various related factors.

Chi-square test for independence or relatedness –analysis of relationship between two categorical variables

Assumptions of chi-Square Test

Assumption 1: Two variables should be measured at an ordinal or nominal level (i.e., categorical data).

Assumption 2: Two variable should consist of two or more categorical, independent groups.

Assumption 3: The data consists of entire populations or be randomly sampled from the population.

Assumption 4: 80% of the expected frequencies should be 5 or more. However the observed frequencies can be any value, including Zero.

Assumption 5: Preferably no data point should be zero.

2. U Test -

The Mann-Whitney U test is the nonparametric equivalent of the two sample t-test. While the t-test makes an assumption about the distribution of a population (i.e. that the sample came from a t-distributed population), the Mann Whitney U Test makes no such assumption.

Null Hypothesis for the Test

The test compares two populations. The null hypothesis for the test is that *the probability is 50% that a randomly drawn member of the first population will exceed a member of the second population.*

Another option for the null hypothesis is that the two samples come from the same population (i.e. that they both have the same median).

The result of performing a Mann Whitney U Test is a U Statistic. For small samples, use the direct method to find the U statistic; For larger samples, a formula is necessary. Or, you can use technology like SPSS to run the test.

When to use which Statistical Test –

Tests	Factor Variable/Independent Variable /predictor	Response Variable/Dependent Variable/ Criterion	Sample(s)	Application
One sample t- test	Metric (Test Value)	Metric (Interval and Ratio)	One Sample	Compare Means
Independent Sample t-test	Non-Metric	Metric	Two Independent Samples	Compare Means
Paired t-test	Metric	Metric	Related sample (Before and After)	Compare Means
Chi-Square	Non-Metric	Non-Metric	One sample, Two or More sample	Test of Independence & Goodness of fit

Predictive Modelling-

A model is a simplified representation of reality created to serve a purpose. It is simplified based on some assumptions about what is and is not important for the specific purpose, or sometimes based on constraints on information or tractability. For example, a map is a model of the physical world. It abstracts away a tremendous amount of information that the mapmaker deemed irrelevant for its purpose. Prediction - In data science, prediction more generally means to estimate an unknown value . This value could be something in the future (in common usage, true prediction), but it could also be something in the present or in the past. Indeed, since data mining usually deals with historical data, models very often are built and tested using events from the past.

Supervised Segmentation

Predictive model focuses on estimating the value of some particular target variable of interest. An intuitive way of thinking about extracting patterns from data in a supervised manner is to try to segment the population into subgroups that have different values for the target variable (and within the subgroup the instances have similar values for the target variable). if the segmentation is done using values of variables that will be known when the target is not, then these segments can be used to predict the value of the target variable . Moreover, the segmentation may at the same time provide a human-understandable set of segmentation patterns. One such segment expressed in English might be: "Middle-aged professionals who reside in New York City on average have a

churn rate of 5%."Specifically, the term "middle-aged professionals who reside in New York City" is the definition of the segment (which references some particular attributes) and "a churn rate of 5%" describes the predicted value of the target variable for the segment.

Often we are interested in applying data mining when we have many attributes, and are not sure exactly what the segments should be. In our churn-prediction problem, who is to say what are the best segments for predicting the propensity to churn? If there exist in the data segments with significantly different (average) values for the target variable, we would like to be able to extract them automatically.

This brings us to our fundamental concept: how can we judge whether a variable contains important information about the target variable? How much? We would like automatically to get a selection of the more informative variables with respect to the particular task at hand (namely, predicting the value of the target variable). Even better, we might like to rank the variables by how good they are at predicting the value of the target.

Selecting Informative Attributes

Given a large set of examples, how do we select an attribute to partition them in an informative way? Let's consider a binary (two class) classification problem, and think about what we would like to get out of it There are two types of heads: square and circular; and two types of bodies: rectangular and oval; and two of

the people have gray bodies while the rest are white. These are the attributes we will use to describe the people. Above each person is the binary target label, Yes or No , indicating (for example) whether the person becomes a loan write-off. We could describe the data on these people as:

- Attributes:
  - — head-shape: square, circular
  - — body-shape: rectangular, oval
  - — body-color: gray, white
- Target variable:
  - — write-off: Yes, No

So let's ask ourselves: which of the attributes would be best to segment these people into groups, in a way that will distinguish write-offs from non-write-offs? Technically, we

would like the resulting groups to be as pure as possible. By pure we mean homogeneous with respect to the target variable . If every member of a group has the same value for the target, then the group is pure. If there is at least one member of the group that has a different value for the target variable than the rest of the group, then the group is impure.

Unfortunately, in real data we seldom expect to find a variable that will make the segments pure.

However, if we can reduce the impurity substantially, then we can both learn something about the data (and the corresponding population), and importantly for this chapter, we can use the attribute in a predictive model—in our example, predicting that members of one segment will have higher or lower write-off rates than those
in another segment. If we can do that, then we can for example offer credit to those with the lower predicted write-off rates, or can offer different credit terms based on the different predicted write-off rates.

Supervised Segmentation with Tree-Structured Models

If we select the single variable that gives the most information gain, we create a very simple segmentation. If we select multiple attributes each giving some information gain, it's not clear how to put them together. Recall from earlier that we would like to create segments that use multiple attributes, such as "Middle-aged professionals who reside in New York City on average have a churn rate of 5%."

Trees as Sets of Rules

Before moving on from the interpretation of classification trees, we should mention their interpretation as logical statements .

You classify a new unseen instance by starting at the root node and following the attribute tests downward until you reach a leaf node, which specifies the instance's predicted class. If we trace down a single path from the root node to a leaf, collecting the conditions as we go, we generate a rule. Each rule consists of the attribute tests along the path connected with AND. Starting at the root node and choosing the left branches of the tree

The classification tree is equivalent to this rule set. If these rules look repetitive, that's because they are: the tree gathers common rule prefixes together toward the top of the tree. Every classification tree can be expressed as a set of rules this way. Whether the tree or the rule set is more intelligible is a matter of opinion; in this simple example, both are fairly easy to understand. As the model becomes larger, some people will prefer the tree or the rule set.

Probability Estimation

There is another, even more insidious problem with models that give simple classifications, rather than estimates of class membership probability. Consider the problem of estimating credit default. Under normal circumstances, for just about any segment of the population to which we would be considering giving credit, the probability of write- off will be very small—far less than 0.5. In this case, when we build a model to estimate the classification (write-off or not), we'd have to say that for each segment, the members are likely not to default—and they will all get the same

classification (not write-off). For example, in a naively built tree model every leaf will be labeled "not write-off." This turns out to be a frustrating experience for new data miners: after all that work, the model really just says that no one is likely to default? This does not mean that the model is useless. It may be that the different segments indeed have very different probabilities of write-off, they just all are less than 0.5. If instead we use these probabilities for assigning credit, we may be able reduce our risk substantially.

For completeness, let's quickly discuss one easy way to address this problem of small samples for tree-based class probability estimation. Instead of simply computing the frequency, we would often use a "smoothed" version of the frequency-based estimate, known as the Laplace correction, the purpose of which is to moderate the influence of leaves with only a few instances.

where n is the number of examples in the leaf belonging to class c , and m is the number of examples not belonging to class c .

Optimization -Optimization is the process of selecting values of decision variables that *minimize* or *maximize* some quantity of interest and is the most important tool for prescriptive analytics. Optimization models have been used extensively in operations and supply chains, finance, marketing, and other disciplines for more than 50 years to help managers allocate resources more efficiently and make lower-cost or more-profitable decisions.

Building Linear Optimization Models

Developing any optimization model consists of four basic steps:

1. Identify the decision variables.
2. Identify the objective function.
3. Identify all appropriate constraints.
4. Write the objective function and constraints as mathematical expressions.

Decision variables are the unknown values that the model seeks to determine. Depending on the application, decision variables might be the quantities of different products to produce, amount of money spent on R&D projects, the amount to ship from a warehouse to a customer, the amount of shelf space to devote to a product, and so on. The quantity we seek to minimize or maximize is called the objective function; for example, we might wish to maximize profit or revenue, or minimize cost or some measure of risk.

Constraints are limitations, requirements, or other restrictions that are imposed on any solution, either from practical or technological considerations or by management policy.

The presence of constraints along with a large number of variables usually makes identifying an optimal solution considerably more difficult and necessitates the use of powerful software tools. The essence of building an optimization model is to first identify these model components, and

then translate the objective function and constraints into mathematical expressions.

Identifying Elements for an Optimization Model -

Managers can generally describe the decisions they have to make, the performance measures they use to evaluate the success of their decisions, and the limitations and requirements they face or must ensure rather easily in plain language. The task of the analyst is to take this information and extract the key elements that form the basis for developing a model. Here is a simple scenario.

Decision Analytics: Analytic models and analyses provide decision makers with a wealth of information; however, people make the final decision. Good decisions don't simply implement the results of analytic models; they require an assessment of intangible factors and risk attitudes. Decision making is the study of how people make decisions, particularly when faced with imperfect or uncertain information, as well as a collection of techniques to support decision choices. Decision analysis differs from other modeling approaches by explicitly considering individual's preferences and attitudes toward risk, and modeling the decision process itself.

Decisions involving uncertainty and risk have been studied for many years. A large body of knowledge has been developed that helps to explain the philosophy associated with making decisions and also provide techniques for incorporating uncertainty and risk in making decisions.

Decision Trees

A useful approach to structuring a decision problem involving uncertainty is to use a graphical model called a decision tree. Decision trees consist of a set of nodes and branches. Nodes are points in time at which events take place. The event can be a selection of a decision from among several alternatives, represented by a decision node, or an outcome over which the decision maker has no control, an event node.

Event nodes are conventionally depicted by circles, and decision nodes are expressed by squares. Branches are associated with decisions and events. Many decision makers find decision trees useful because sequences of decisions and outcomes over time can be modeled easily.

Simulation -

Simulation is the imitation of the operation of a real-world process or system over time. The act of simulating something first requires that a model be developed; this model represents the key characteristics or behaviors of the selected physical or abstract system or process. The model represents the system itself, whereas the simulation represents the operation of the system over time.

MONTE –CARLO SIMULATION -It is used to study probabilistic simulations where the given process has a random or chance component. Using Monte Carlo Simulation, a given problem is solved by simulating the original data with random number generators.

STEPS OF MONTE –CARLO METHOD - Identify the input variables, collect data on them and write down the probability distribution for them. This represents the simulation model.

Rewrite the cumulative probability distribution.

Identify the random number intervals corresponding to these cumulative probabilities figures. If Probabilities are expressed in two identical digits use random numbers from 00 to 99 and if they are in three decimal digits use random numbers from 000 to 999.

Prepare a table for the random numbers (to be generated to simulate the model) and the expected (simulated) values of the inputs variables etc.

Generate ' n' random numbers(if not given) from the random umber tables.Process the simulated information and then summarize the data and interpret the results.

Decision analysis (DA) is a systematic, quantitative, and visual approach to addressing and evaluating the important choices that businesses sometimes face. Decision analysis uses a variety of tools to evaluate all relevant information to aid in the decision-making process and incorporates aspects of psychology, management techniques, training, and economics. It is often used to assess decisions that are made in the context of multiple variables and that have many possible outcomes or objectives. The process can be used by individuals or groups attempting to make a decision related to risk management, capital investments, and strategic business decisions.A graphical representation of alternatives and possible solutions, as well as challenges and uncertainties, can be created on a decision tree or influence diagram. More sophisticated computer models have also been developed to aid in the decision-analysis process.

Factor analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors.

It is often used to identify a small number of factors that explain most of the variance embedded in a larger number of variables.

Thus, factor analysis is about data reduction.

Assumptions in Factor Analysis -

Interval or Ratio Data: Interval or ratio data are assumed.

Sampling Adequacy : (KMO >0.5) or A sample of 100 subjects is acceptable (or No. of variables* 5)

Overall significance of correlation matrix should be checked with Bartlett test sphericity (Sig value<

0.05)

Multicollinearity but No perfect multicollinearity: Factor analysis is an interdependency technique. There should not be perfect multicollinearity between the variables.

No Outliers, Normality and Linearity

PCA-

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the **Principal Components**. It is one of the popular tools that is used for exploratory data analysis and predictive modeling. It is a technique to draw strong patterns from the given dataset by reducing the variances.

PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are ***image processing, movie recommendation system, optimizing the power allocation in various communication channels.*** It is a feature extraction technique, so it contains the important variables and drops the least important variable.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Some common terms used in PCA algorithm:

- **Dimensionality:** It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.
- **Correlation:** It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.
- **Orthogonal:** It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.
- **Eigenvectors:** If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
- **Covariance Matrix:** A matrix containing the covariance between the pair of variables is called the Covariance Matrix.

Principal Component

Cluster analysis

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

1. Hierarchical Cluster Analysis

2. Nonhierarchical Cluster Analysis (K-Means Clustering)

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

Agglomerative Approach: The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

Divisive Approach: The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping

data objects into microclusters, macro clustering is performed on the microcluster.

K-Means Clustering -

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

Advantages of Cluster Analysis:

It can help identify patterns and relationships within a dataset that may not be immediately obvious.

It can be used for exploratory data analysis and can help with feature selection.

It can be used to reduce the dimensionality of the data.

It can be used for anomaly detection and outlier identification.

It can be used for market segmentation and customer profiling.