Dr. SNS RAJALAKSHMI COLLEGE OF ARTS & SCIENCE (AUTONOMOUS)

Accredited by NAAC (Cycle III) with ‘A+’ Grade
Affiliated to Bharathiar University
Coimbatore-49

DEPARTMENT OF COMPUTER APPLICATIONS

II BCA

ELECTIVE TRACK 2 – HIGHER EDUCATION

21UCU807: DATA SCIENCE

UNIT – V

TIME SERIES ANALYSIS:

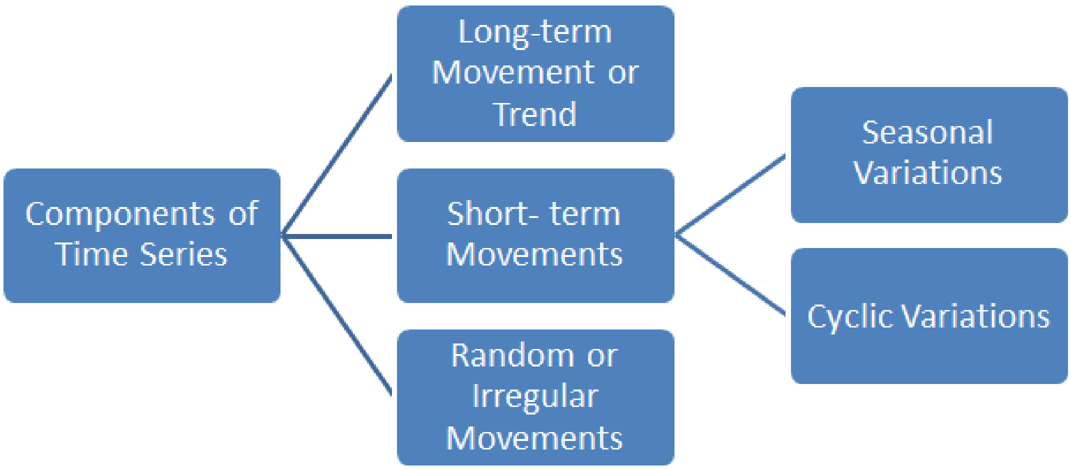
Time series analysis is a specific way of analysing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

What sets time series data apart from other data is that the analysis can show how variables change over time? In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data.

Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time. Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing. Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms. Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow's weather report to future years of climate change. Examples of time series analysis in action include:

- Weather data
- Rainfall measurements
- Temperature readings
- Heart rate monitoring (EKG)
- Brain monitoring (EEG)
- Quarterly sales
- Stock prices
- Automated stock trading
- Industry forecasts
- Interest rates



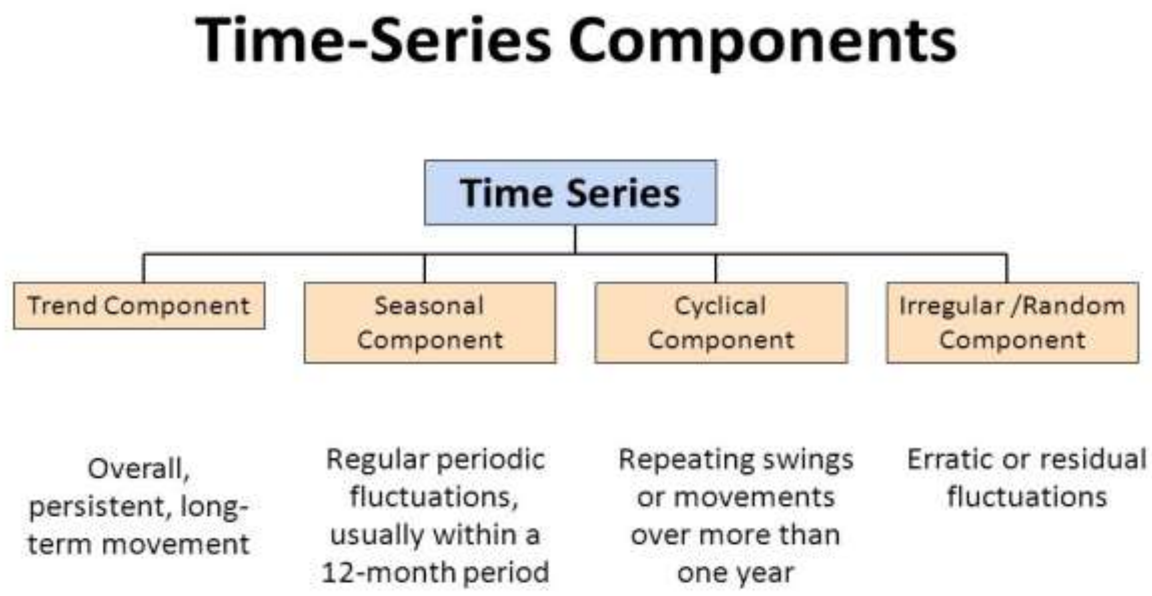
TIME SERIES ANALYSIS TYPES:

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can't account for all variances, and they can't generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to a lack of fit. Lack of fit or over fitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect.

Models of time series analysis include:

- **Classification:** Identifies and assigns categories to the data.
- **Curve fitting:** Plots the data along a curve to study the relationships of variables within the data.
- **Descriptive analysis:** Identifies patterns in time series data, like trends, cycles, or seasonal variation.
- **Explanative analysis:** Attempts to understand the data and the relationships within it, as well as cause and effect.
- **Exploratory analysis:** Highlights the main characteristics of the time series data, usually in a visual format.

- **Forecasting:** Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.
- **Intervention analysis:** Studies how an event can change the data.
- **Segmentation:** Splits the data into segments to show the underlying properties of the source information.



PROBABILITY:

Probability is simply how likely something is to happen. Whenever we're unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are. The analysis of events governed by probability is called statistics.

In science, the probability of an event is a number that indicates how likely the event is to occur. It is expressed as a number in the range from 0 and 1, or, using percentage notation, in the range from 0% to 100%. The more likely it is that the event will occur, the higher its probability.

It is based on the possible chances of something to happen. The theoretical probability is mainly based on the reasoning behind probability. For example, if a coin is tossed, the theoretical probability of getting a head will be ½.

Probability is the branch of mathematics concerning the occurrence of a random event, and four main types of probability exist:

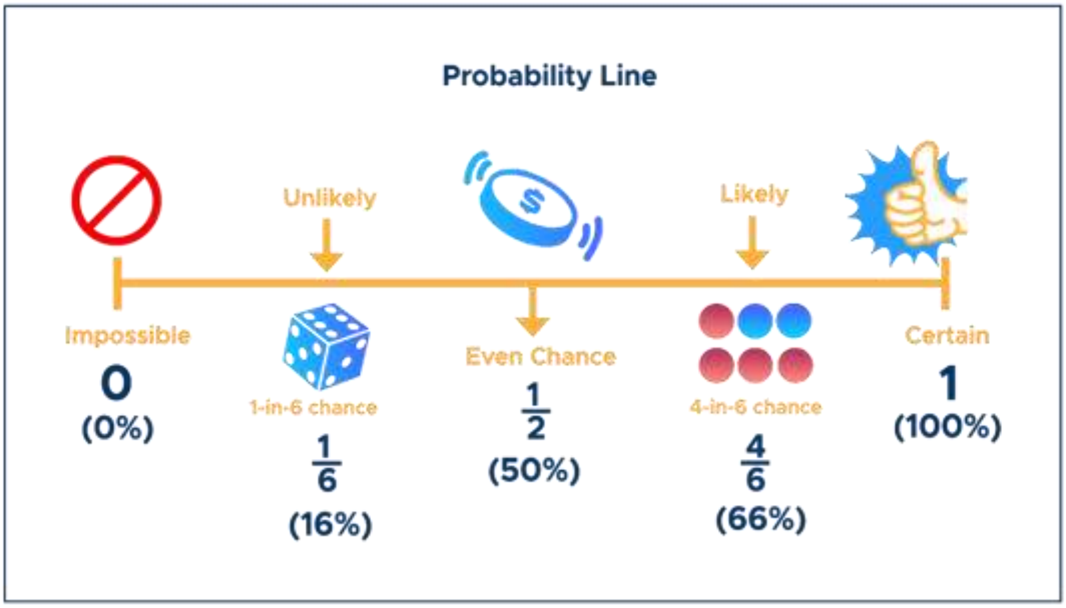
Classical,

Empirical,

Subjective

Axiomatic.

Probability is a number between 0 and 1 that describes the chance that a stated event will occur. An event is a specified set of outcomes of a random variable. Mutually exclusive events can occur only one at a time. Exhaustive events cover or contain all possible outcomes.



UNIFORM DISTRIBUTION:

In statistics, uniform distribution refers to a type of probability distribution in which all outcomes are equally likely. A deck of cards has within it uniform distributions because the likelihood of drawing a heart, a club, a diamond, or a spade is equally likely.

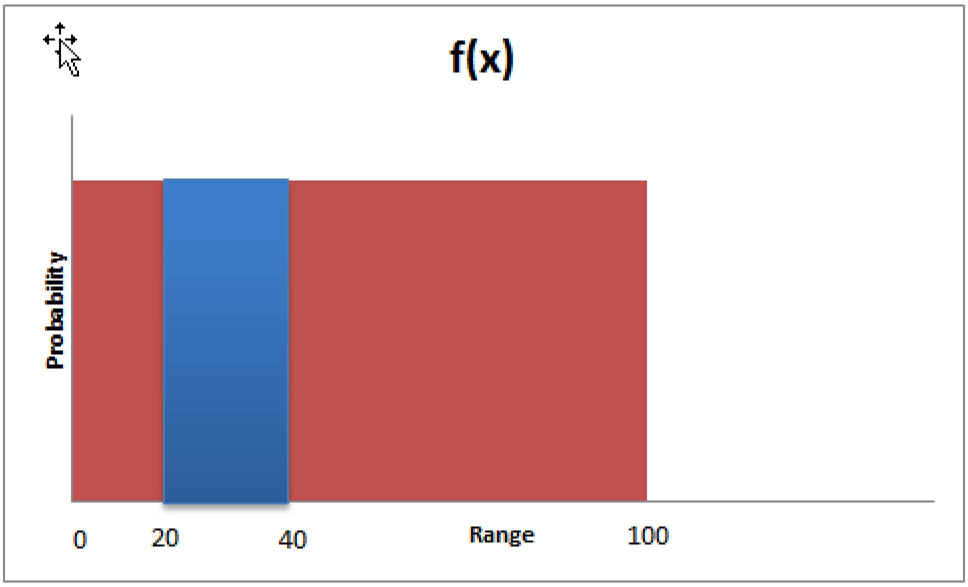
The uniform distribution gets its name from the fact that the probabilities for all outcomes are the same. Unlike a normal distribution with a hump in the middle or a chi-square distribution, a uniform distribution has no mode. Instead, every outcome is equally likely to occur.

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability. The probability density function and cumulative distribution function for a continuous uniform distribution on the interval

In uniform distribution, organisms are spread out in a fairly regular pattern. This occurs often where individuals must compete for a limiting resource, such as water or light. Desert shrubs and redwood trees grow in a uniform distribution—shrubs compete for water, while redwoods compete for light.

Uniform distribution is of two forms – discrete and continuous. This bifurcation depends on the type of outcomes with possibilities of occurrence. A discrete uniform distribution is the probability distribution where the researchers have a predefined number of equally likely outcomes.

One of the most important applications of the uniform distribution is in the generation of random numbers. That is, almost all random number generators generate random numbers on the (0,1) interval. For other distributions, some transformation is applied to the uniform random numbers.



PSUEDORANDOM NUMBERS:

A pseudorandom sequence of numbers is one that appears to be statistically random, despite having been produced by a completely deterministic and repeatable process.

The generation of random numbers has many uses, such as for random sampling, Monte Carlo methods, board games, or gambling. In physics, however, most processes, such as gravitational acceleration, are deterministic, meaning that they always produce the same outcome from the same starting point.

Some notable exceptions are radioactive decay and quantum measurement, which are both modelled as being truly random processes in the underlying physics. Since these processes are not practical sources of random numbers, people use pseudorandom numbers, which ideally have the unpredictability of a truly random sequence, despite being generated by a deterministic process.

In many applications, the deterministic process is a computer algorithm called a pseudorandom number generator, which must first be provided with a number called a random seed.

Since the same seed will yield the same sequence every time, it is important that the seed be well chosen and kept hidden, especially in security applications, where the pattern's unpredictability is a critical feature.

In some cases where it is important for the sequence to be demonstrably unpredictable, people have used physical sources of random numbers, such as radioactive decay, atmospheric electromagnetic noise harvested from a radio tuned between stations, or intermixed timings of people's keystrokes.

The time investment needed to obtain these numbers leads to a compromise: using some of these physics readings as a seed for a pseudorandom number generator.

NON DISCRETE VARIABLE:

The opposite of a discrete variable is a continuous variable, which can take on all possible values between the extremes. Thus this variable can vary in a continuous manner. For example, consider the length of a stretched rubber band**.**

NON CONTINUOUS RANDOM VARIABLE:

A discrete variable is a kind of statistics variable that can only take on discrete specific values. The variable is not continuous, which means there are infinitely many values between the maximum and minimum that just cannot be attained, no matter what.

NOTATION OF EXPECTATION:

Intuitively, the expectation is the average of a large number of independent realizations of the random variable. Let say X is our observation concerning the daily snow level, we want to estimate the average snow level. Therefore, we observed n data points on X and get an average — μ.

DEPENDENCE:

If two or more variables affect each other, we will talk of a statistical dependence or relationship. The strength of a dependence between variables can be measured simultaneously – for example, we could investigate to what extent the looks of a person influence his personal income and vice versa. The grade of such a relationship is referred to as correlation and can be measured by methods such as the Spearman's rank correlation coefficient.

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:

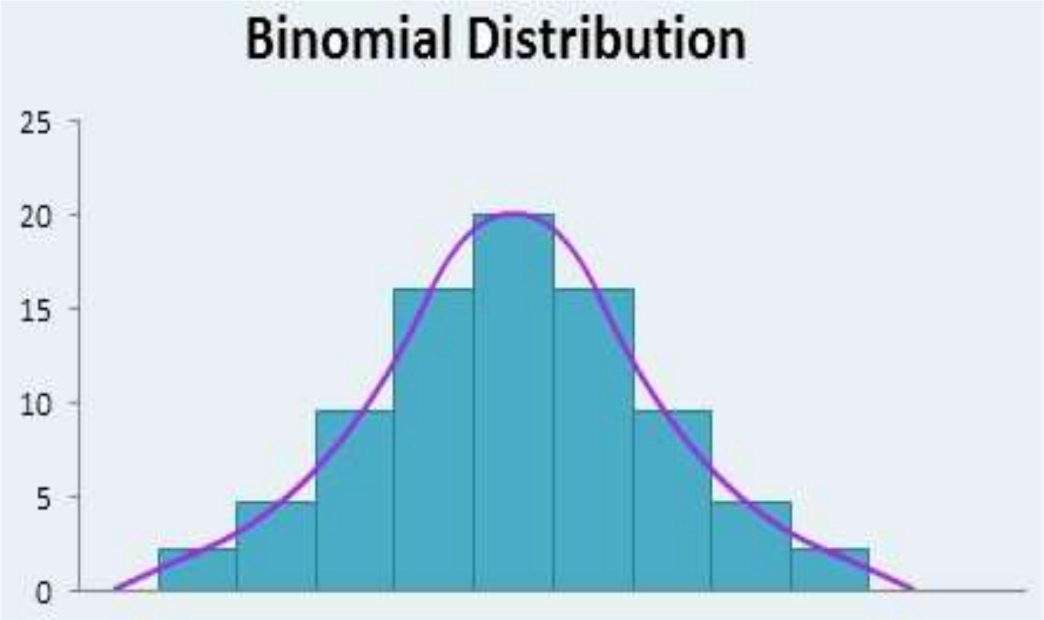
- P(A given B)
  The "*given*" is denoted using the pipe "|" operator; for example:
- P(A | B)
  The conditional probability for events *A* given event *B* is calculated as follows:
- P(A given B) = P(A and B) / P(B)
  This calculation assumes that the probability of event *B* is not zero, e.g. is not 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.

BINOMIAL DISTRIBUTION:

Binomial distribution is a probability distribution used in statistics that summarizes the likelihood that a value will take one of two independent values under a given set of parameters or assumptions.

The underlying assumptions of binomial distribution are that there is only one outcome for each trial, that each trial has the same probability of success, and that each trial is mutually exclusive, or independent of one another.



UNDERSTANDING BINOMIAL DISTRIBUTION:

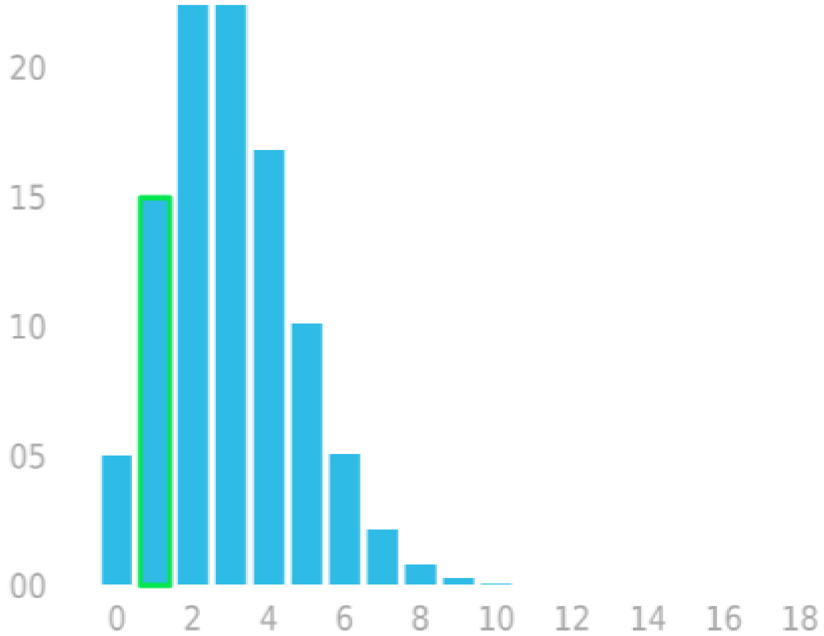
To start, the "binomial" in binomial distribution means two terms. We're interested not just in the number of successes, nor just the number of attempts, but in both. Each is useless to us without the other.

Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution. This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure) given a number of trials in the data. Binomial distribution thus represents the probability for x successes in n trials, given a success probability p for each trial.

Binomial distribution summarizes the number of trials, or observations when each trial has the same probability of attaining one particular value. Binomial distribution determines the probability of observing a specified number of successful outcomes in a specified number of trials.

POISSON DISTRIBUTION:

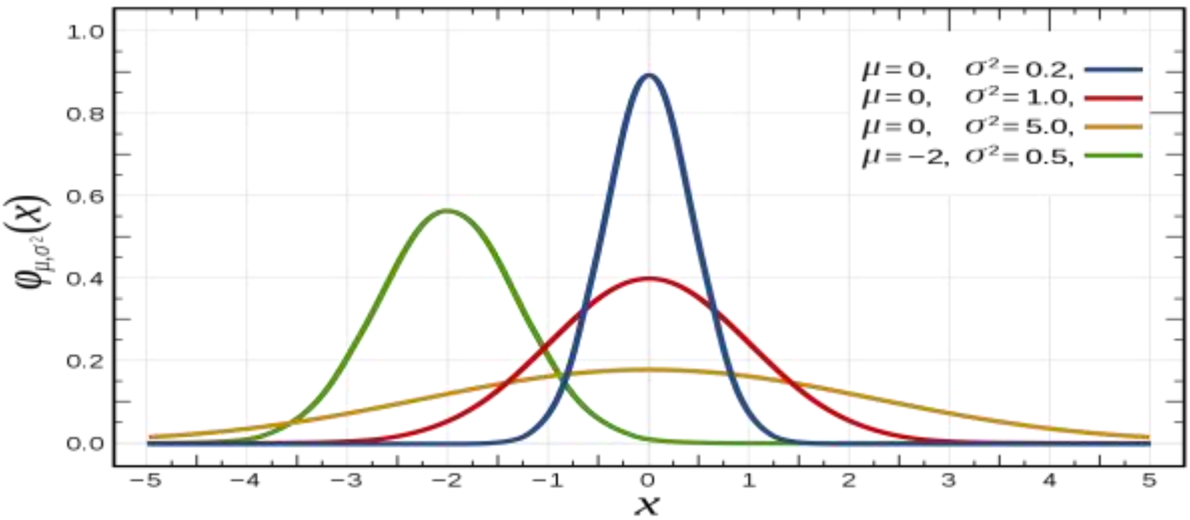
The Poisson distribution, named after the French mathematician Denis Simon Poisson, is a discrete distribution function describing the probability that an event will occur a certain number of times in a fixed time (or space) interval. It is used to model count-based data, like the number of emails arriving in your mailbox in one hour or the number of customers walking into a shop in one day, for instance.



NORMAL DISTRIBUTION:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

The normal distribution is an important probability distribution in math and statistics because many continuous data in nature and psychology display this bell-shaped curve when compiled and graphed.

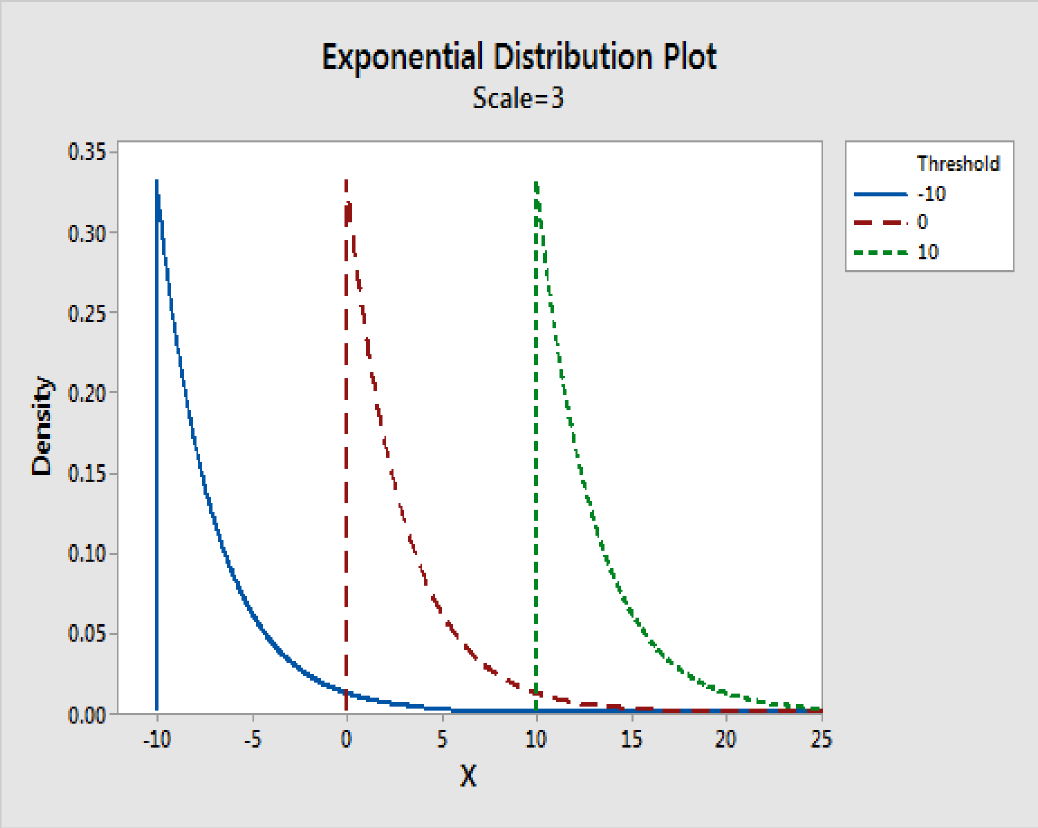


EXPONENTIAL DISTRIBUTION:

The exponential distribution is a continuous probability distribution that often concerns the amount of time until some specific event happens. It is a process in which events happen continuously and independently at a constant average rate. The exponential distribution has the key property of being memory less. The exponential random variable can be either more small values or fewer larger variables. For example, the amount of money spent by the customer on one trip to the supermarket follows an exponential distribution.

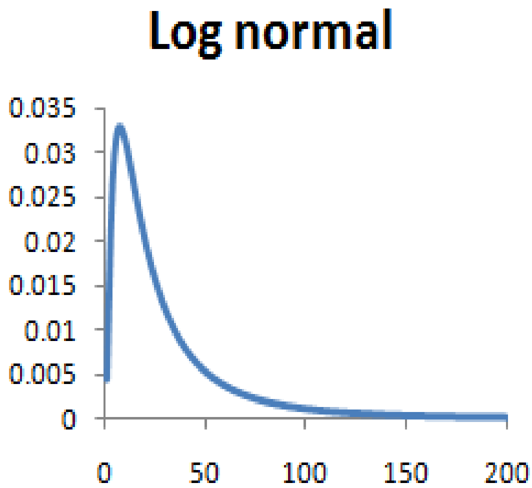
One of the widely used continuous distribution is the exponential distribution. It helps to determine the time elapsed between the events. It is used in a range of applications such as reliability theory, queuing theory, physics and so on. Some of the fields that are modelled by the exponential distribution are as follows:

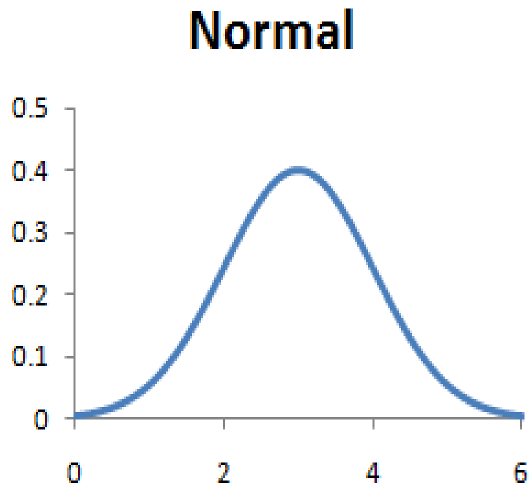
- Exponential distribution helps to find the distance between mutations on a DNA strand
- Calculating the time until the radioactive particle decays.
- Helps on finding the height of different molecules in a gas at the stable temperature and pressure in a uniform gravitational field
- Helps to compute the monthly and annual highest values of regular rainfall and river outflow volumes



LOG-NORMAL DISTRIBUTION:

A log-normal distribution is a continuous distribution of random variable y whose natural logarithm is normally distributed. For example, if random variable y = exp { y } has log-normal distribution then x = log (y) has normal distribution.





ENTROPY:

Entropy is one of the key aspects of Machine Learning. It is a must to know for anyone who wants to make a mark in Machine Learning and yet it perplexes many of us.

The focus of this article is to understand the working of entropy by exploring the underlying concept of probability theory, how the formula works, its significance, and why it is important for the Decision Tree algorithm

ORIGIN OF ENTROPY:

The term entropy was first coined by the German physicist and mathematician Rudolf Clausius and was used in the field of thermodynamics.

In 1948, Claude E. Shannon, mathematician, and electrical engineer, published a paper on A Mathematical Theory of Communication, in which he had addressed the issues of measure of information, choice, and uncertainty. Shannon was also known as the ‘father of information theory’ as he had invented the field of information theory.

“Information theory is a mathematical approach to the study of coding of information along with the quantification, storage, and communication of information.”

In his paper, he had set out to mathematically measure the statistical nature of “lost information” in phone-line signals. The work was aimed at the problem of how best to encode the information a sender wants to transmit. For this purpose, information entropy was developed as a way to estimate the information content in a message that is a measure of uncertainty reduced by the message.

So, we know that the primary measure in information theory is entropy. The English meaning of the word entropy is: it is a state of disorder, confusion, and disorganization. Let’s look at this concept in depth.

In simple words, we know that information is some facts learned about something or someone. Notionally, we can understand that information is something that can be stored in, transferred, or passed-on as variables, which can further take different values. In other words, a variable is nothing but a unit of storage. So, we get information from a variable by seeing its value, in the same manner as we get details (or information) from a message or letter by reading its content.

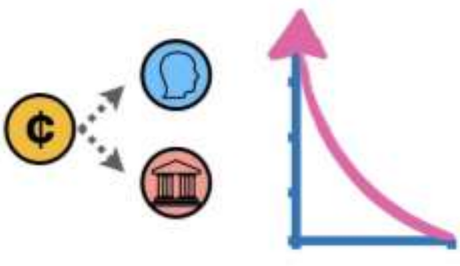
The entropy measures the “amount of information” present in a variable. Now, this amount is estimated not only based on the number of different values that are present in the variable but also by the amount of surprise that this value of the variable holds. Allow me to explain what I mean by the amount of surprise.

Let’s say, you have received a message, which is a repeat of an earlier text then this message is not at all informative. However, if the message discloses the results of the cliff-hanger US elections, then this is certainly highly informative. This tells us that the amount of information in a message or text is directly proportional to the amount of surprise available in the message.

Hence, one can intuitively understand that this storage and transmission of information is associated with the amount of information in that variable. Now, this can be extended to the outcome of a certain event as well. For instance, the event is tossing a fair coin that will have two equally likely outcomes. This will provide less information that is in other words, has less surprise as the result of the fair coin will either be heads or tails. Hence, the flipping of a fair coin has a lower entropy.

In information theory, the entropy of a random variable is the average level of “information“, “surprise”, or “uncertainty” inherent in the variable’s possible outcomes.

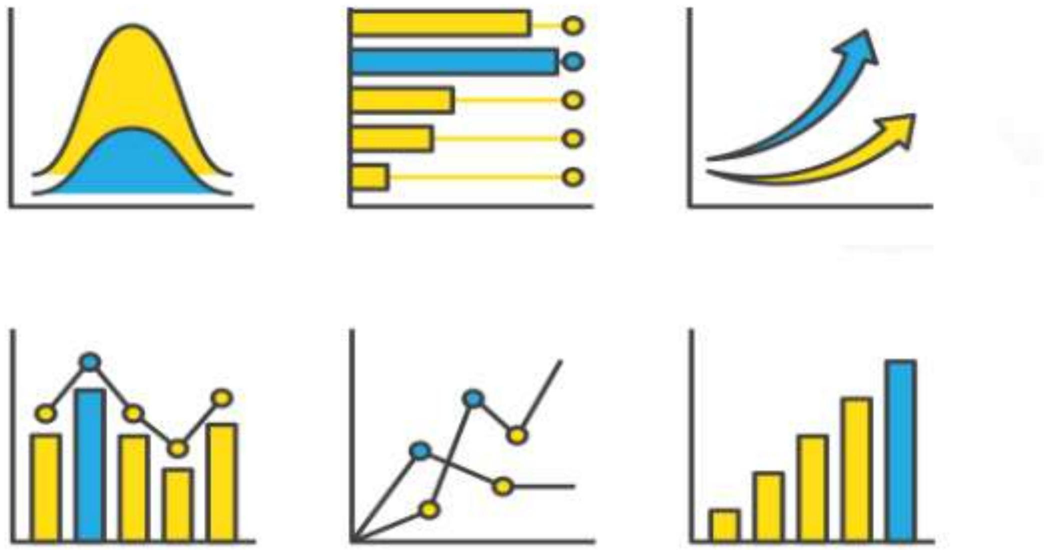
That is, the more certain or the more deterministic an event is, the less information it will contain. In a nutshell, the information is an increase in uncertainty or entropy.



$$\sum p(x)\log\left(\frac{1}{p(x)}\right)$$

STATISTICS:

In data science, statistics is at the core of sophisticated machine learning algorithms, capturing and translating data patterns into actionable evidence. Data scientists use statistics to gather, review, analyze, and draw conclusions from data, as well as apply quantified mathematical models to appropriate variables.

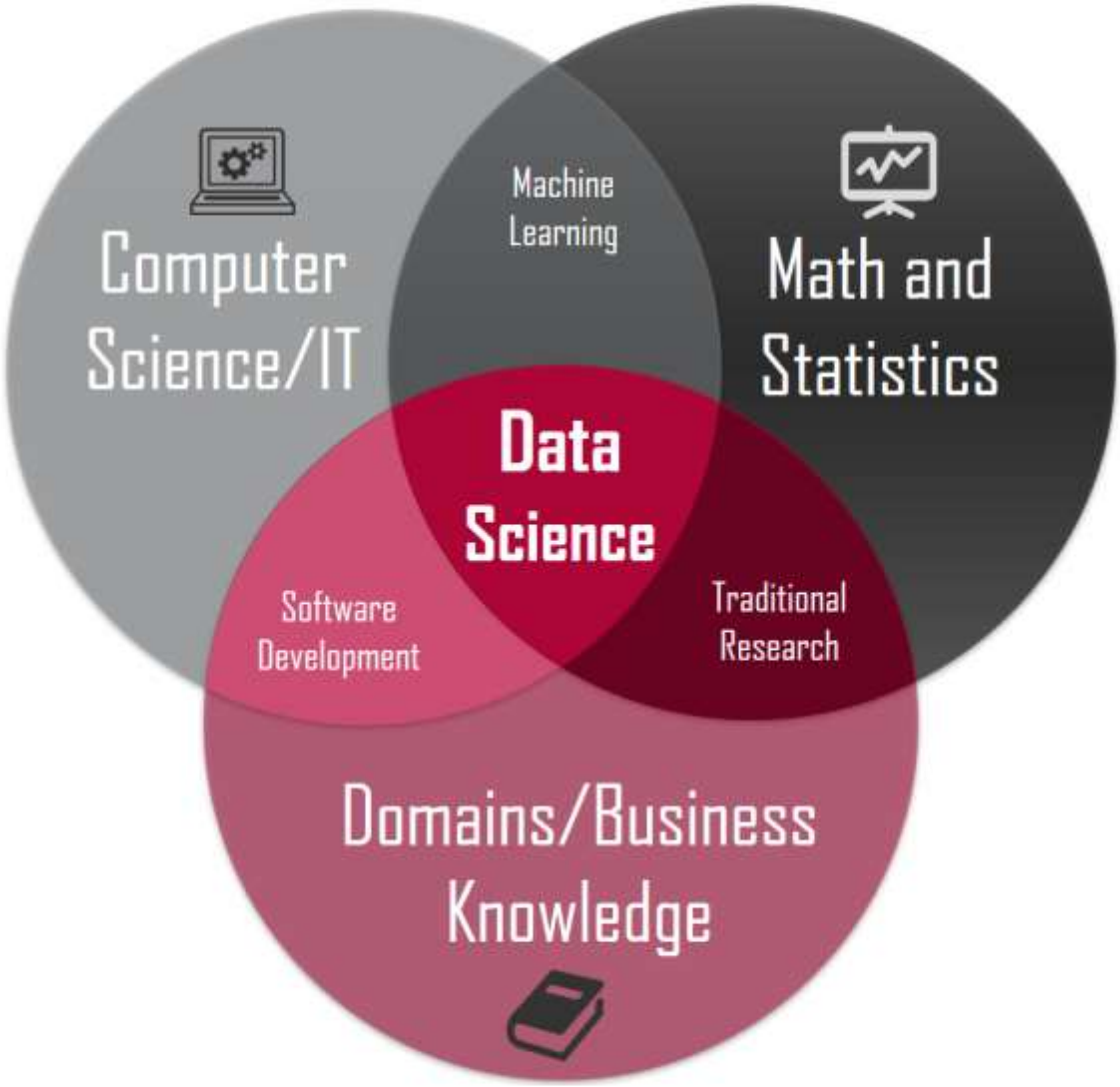


Data scientists use statistics to collect, evaluate, analyze, and draw conclusions from data, as well as to implement quantitative mathematical models for pertinent variables. Data science requires both technical skills, such as R and Python programming, and “soft skills,” such as communication and attention to detail.

According to Elite Data Science, a data science educational platform, data scientists need to understand the fundamental concepts of descriptive statistics and probability theory, open_in_new which include the key concepts of probability distribution, statistical significance, hypothesis testing and regression.

A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value.

Statistics is a mathematically-based field which seeks to collect and interpret quantitative data. In contrast, data science is a multidisciplinary field which uses scientific methods, processes, and systems to extract knowledge from data in a range of forms.



TYPES OF STATISTICS:

Statistics have majorly categorized into two types:

- Descriptive statistics
- Inferential statistics

DESCRIPTIVE STATISTICS:

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorized into four different categories:

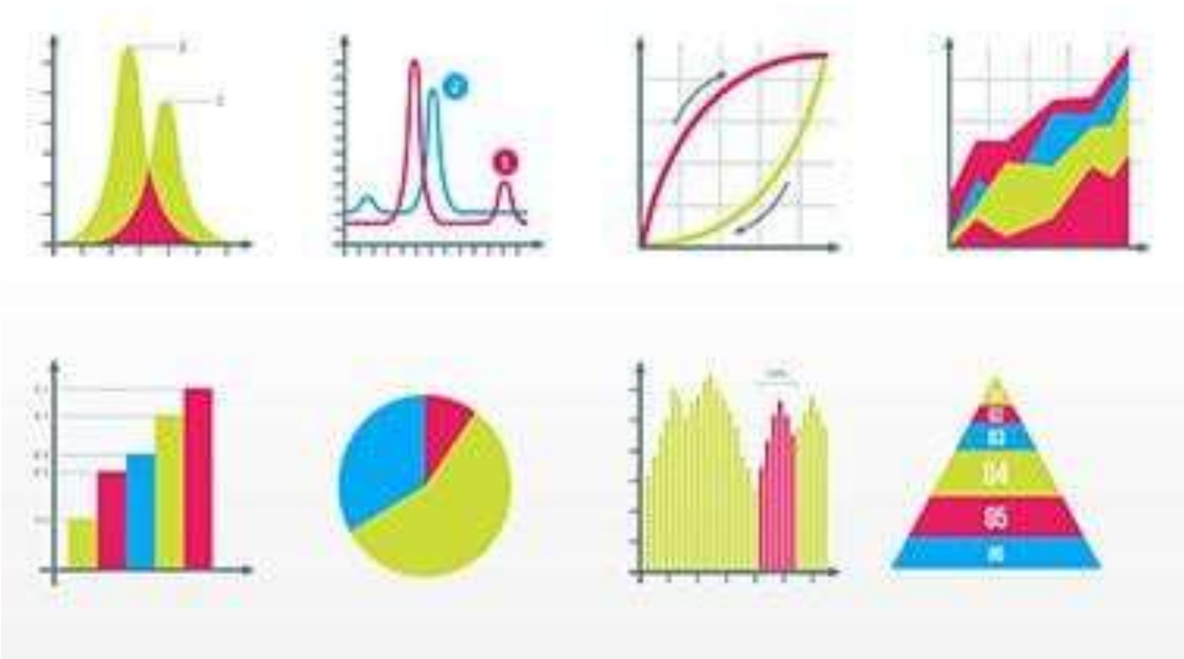
- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

INFERENTIAL STATISTICS:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.



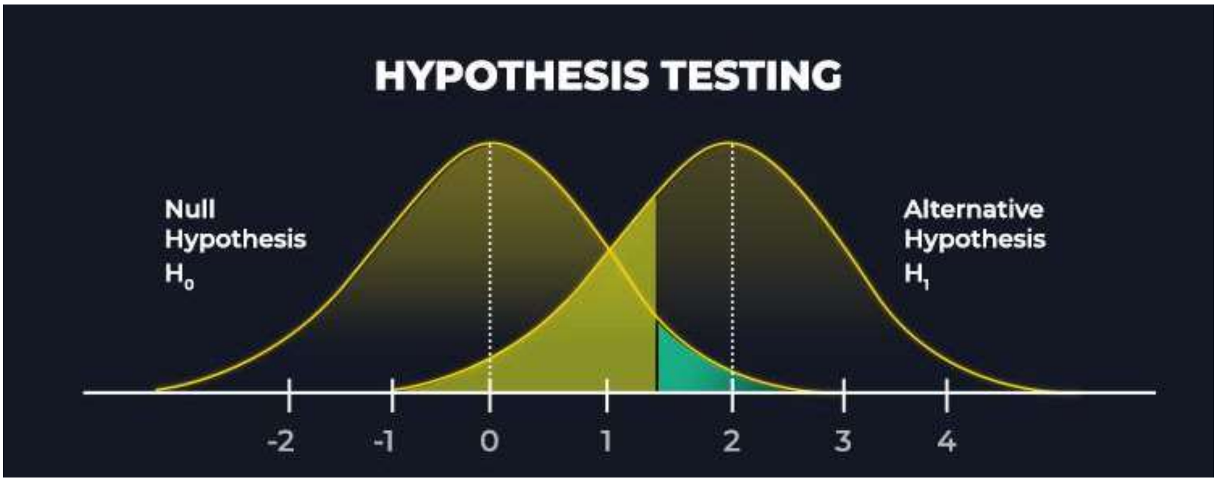
HYPOTHESIS TESTING:

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.

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.

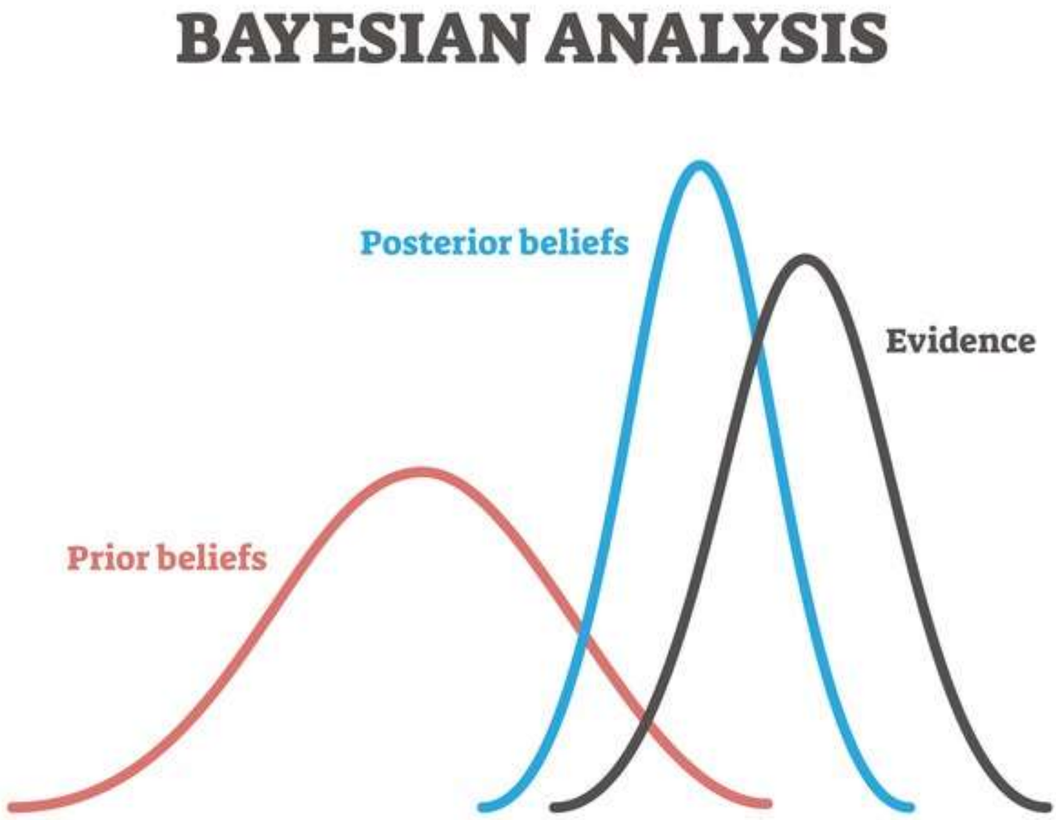


BAYESIAN STATISTICS:

The Bayesian technique is an approach in statistics used in data analysis and parameter estimation. This approach is based on the Bayes theorem. Bayesian Statistics follows a unique principle wherein it helps determine the joint probability distribution for observed and unobserved parameters using a statistical model.

Bayesian hypothesis testing enables us to quantify evidence and track its progression as new data come in. This is important because there is no need to know the intention with which the data were collected.

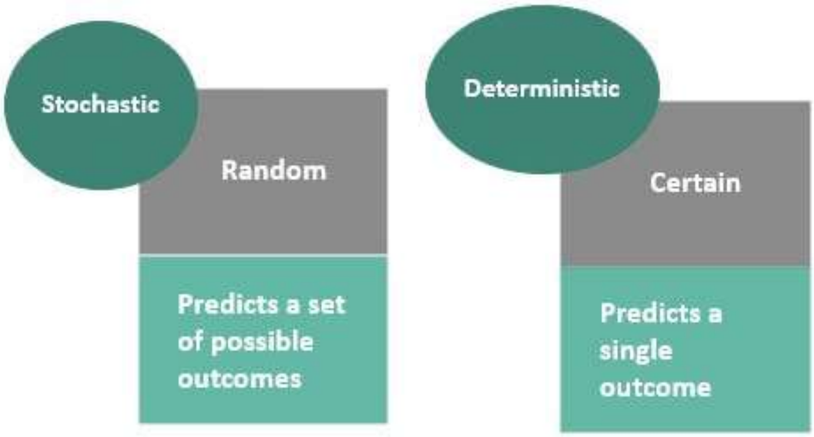
Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution, termed the prior and data distributions.



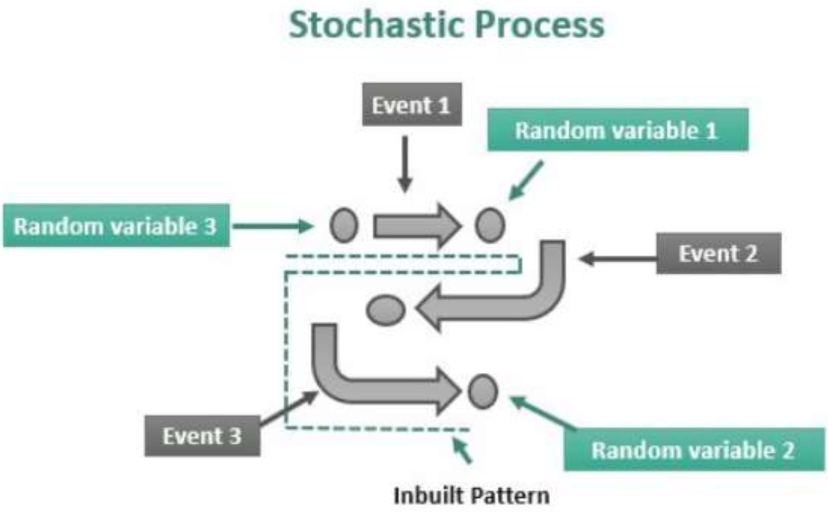
STOCHASTIC MODELING:

Stochastic modeling forecasts the probability of various outcomes under different conditions, using random variables. Stochastic modeling presents data and predicts outcomes that account for certain levels of unpredictability or randomness.

Stochastic modeling develops a mathematical or financial model to derive all possible outcomes of a given problem or scenarios using random input variables. It focuses on the probability distribution of possible outcomes. Examples are Monte Carlo Simulation, Regression Models, and Markov-Chain Models.



It has four main types – non-stationary stochastic processes, stationary stochastic processes, discrete-time stochastic processes, and continuous-time stochastic processes.

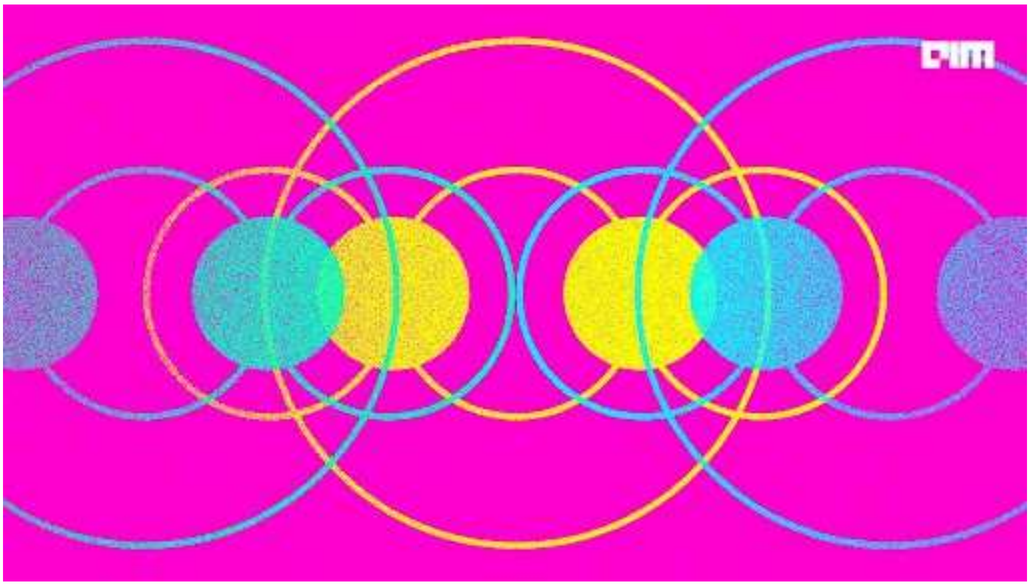


MARKOV CHAINS:

Markov Chains are devised referring to the memory less property of Stochastic Process which is the Conditional Probability Distribution of future states of any process depends only and only on the present state of those processes. Which are then used upon by Data Scientists to define predictions.

Markov chains have many applications, ranging from modeling communication networks to analyzing stock prices. They are particularly useful in modeling systems that have a finite number of states and transitions between those states and can be used to analyze and predict the long-term behavior of such systems.

Markov chains are used to calculate the probability of an event occurring by considering it as a state transitioning to another state or a state transitioning to the same state as before.



A Markov chain is a mathematical process that transitions from one state to another within a finite number of possible states. It is a collection of different states and probabilities of a variable, where its future condition or state is substantially dependent on its immediate previous state.

VITERBI ALGORITHM:

Viterbi Algorithm is dynamic programming and computationally very efficient. We will start with the formal definition of the Decoding Problem, then go through the solution and finally implement it.

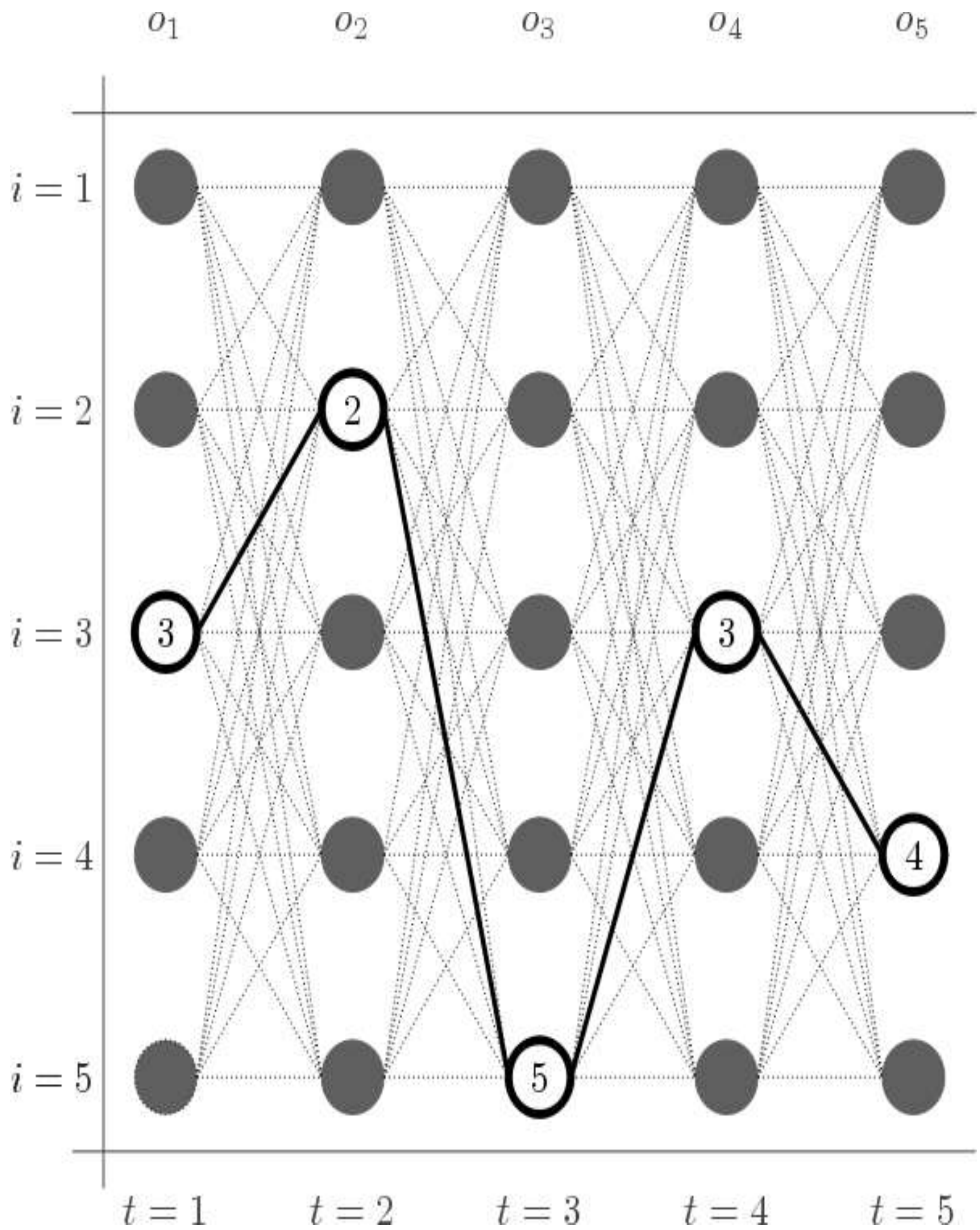
The Viterbi algorithm is a dynamic programming algorithm for obtaining the maximum a posteriori probability estimate of the most likely sequence of hidden states Called the Viterbi path that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models (HMM).

The algorithm has found universal application in decoding the convolutional codes used in both CDMA and GSM digital cellular, dial-up modems, satellite, deep-space communications, and 802.11 wireless LANs. It is now also commonly used in speech recognition, speech synthesis, divarication keyword spotting, computational linguistics, and bioinformatics. For example, in speech-to-text (speech recognition), the acoustic signal is treated as the observed sequence of events, and a string of text is considered to be the "hidden cause" of the acoustic signal. The Viterbi algorithm finds the most likely string of text given the acoustic signal.

HISTORY:

The Viterbi algorithm is named after Andrew Viterbi, who proposed it in 1967 as a decoding algorithm for convolutional codes over noisy digital communication links. It has, however, a history of multiple invention, with at least seven independent discoveries, including those by Viterbi, Needleman and Wunsch, and Wagner and Fischer. It was introduced to Natural Language Processing as a method of part-of-speech tagging as early as 1987.

Viterbi path and Viterbi algorithm have become standard terms for the application of dynamic programming algorithms to maximization problems involving probabilities. For example, in statistical parsing a dynamic programming algorithm can be used to discover the single most likely context-free derivation (parse) of a string, which is commonly called the "Viterbi parse". Another application is in target tracking, where the track is computed that assigns a maximum likelihood to a sequence of observations.



CONTINUOUS-TIME MARKOV PROCESSES:

A continuous-time Markov chain (CTMC) is a continuous stochastic process in which, for each state, the process will change state according to an exponential random variable and then move to a different state as specified by the probabilities of a stochastic matrix.

A gas station has a single pump and no space for vehicles to wait (if a vehicle arrives and the pump is not available, it leaves). Vehicles arrive to the gas station following a Poisson process with a rate of $\lambda=3/20$ vehicles per minute, of which 75% are cars and 25% are motorcycles.

The refueling time can be modelled with an exponential random variable with mean 8 minutes for cars and 3 minutes for motorcycles, that is, the services rates are $\mu c=1/8$cars and $\mu m=1/3$ motorcycles per minute respectively.

The Poisson distribution, named after the French mathematician Denis Simon Poisson, is a discrete distribution function describing the probability that an event will occur a certain number of times in a fixed time (or space) interval.

The Poisson distribution, named after the French mathematician Denis Simon Poisson, is a discrete distribution function describing the probability that an event will occur a certain number of times in a fixed time (or space) interval. It is used to model count-based data, like the number of emails arriving in your mailbox in one hour or the number of customers walking into a shop in one day, for instance.