CHAPTER 3

Prepared by Afreen Banu & Ashwin R G

Business Statistics

It involves **collecting, classifying, summarizing, organizing, analyzing, and interpreting data**. The main objective of Business Statistics is to make inferences about certain characteristics of a population in the business domain whether the population is people, objects, or collections of information.

Business Statistics is the science of Making Decisions on the bases of a lot of analysis in production, auditing, and econometrics. Every right manager makes company growth decisions on behalf of these statistics in times of uncertainty.

Business Statistics helps a business to: **Deal with uncertainties by forecasting seasonal, cyclic and general economic fluctuations**. Helps in Sound Decision making by providing accurate estimates about costs, demand, prices, sales etc. Helps in business planning on the basis of sound predictions and assumptions.

A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.

Business

Definition: Business is defined as an organised economic activity, wherein the exchange of goods and services takes place, for adequate consideration. It is nothing but a method of making money, from commercial transactions. It includes all those activities whose sole aim is to make available the desired goods and services to the society, in an effective manner.

It is a systematic attempt of the businesspersons to produce goods and services, and sell them at the market, to reap the reward, by way of profit.

Profit plays a pivotal role, as all the business activities are directed towards it, because it works an incentive to the entrepreneurs, for their efforts, and thus, necessary for every business.

Characteristics of Business

- Economic Activity: Business is an economic activity, as it is conducted with the primary objective of earning money, i.e. for an economic motive.
- Production/purchase of goods and services: Goods and services are produced or procured by business entities, so as to add value and sell them to the consumer. Goods are either manufactured by the company or procured from the supplier, with the aim of selling it further to the consumer, for profit.
- Selling of goods and services: Business must involve the transfer of goods to the customer for value, through selling, meaning that if the goods are acquired for personal consumption, then the transaction will not amount to business activity.
- Continuity in dealings: Every business requires regularity in transactions, i.e. an isolated transaction of exchange of goods or services will not be considered as business. So, to constitute business, the dealings must be carried out on a regular basis.
- Profit earning: The basic purpose of business is to make the profit from its activities. It is the spine of business, which keeps the business going, in the long term.
- Element of risk: Risk is the key element of every business, concerned with exposure to loss. Efforts are made to forecast future events and plan the business strategies accordingly. However, the factors that affect business are uncertain and so does the business opportunities, which can be a shift in demand, floods, fall in prices, strikes, lockout, money market fluctuation, etc.
- Uncertain return: In business, the return is never predictable and guaranteed, i.e. the amount of money which the business is going to reap is not certain. It may be possible that the business earns a huge profit or suffer heavy losses.
- Legal and Lawful: No matter, in which type of business the company is engaged, it should be legal in the eyes of the law, or else it will not be considered as business.
- Consumer satisfaction: The aim of business is to supply goods and services to consumers, so as to satisfy their wants, as when the consumer (final user) is satisfied, he/she will purchase the goods or services. But, if they are not, there are chances that they will look for substitutes.

What is Statistical Learning?

Statistical Learning is a set of tools for understanding data. These tools broadly come under two classes: supervised learning & unsupervised learning. Generally, supervised learning refers to predicting or estimating an output based on one or more inputs. Unsupervised learning, on the other hand, provides a relationship or finds a pattern within the given data without a supervised output.

Data presentations using Charts and Diagrams

What is data presentation?

Data presentation is defined as **the process of using various graphical formats to visually represent the relationship between two or more data sets so that an informed decision can be made based on them**.

What is the importance of data presentation?

Data Presentation helps the clients or the audience to not spend time grasping the concept and the future alternatives of the business and to convince them to invest in the company & turn it profitable both for the investors & the company.

What are the benefits of data presentation?

Data visualization **provides us with a quick, clear understanding of the information**. Thanks to graphic representations, we can visualize large volumes of data in an understandable and coherent way, which in turn helps us comprehend the information and draw conclusions and insights

What is the purpose of presenting data through charts and diagrams?

Charts and graphs **help to express complex data in a simple format**. They can add value to your presentations and meetings, improving the clarity and effectiveness of your message.

What is a diagram presentation?

Concept of Diagrammatic Presentation

It is **a technique of presenting numeric data through pictograms, cartograms, bar diagrams, and pie diagrams**. It is the most attractive and appealing way to represent statistical data. Diagrams help in visual comparison and they have a bird's eye view.

What is data presentation and analysis?

DEFINITION: The data presentation and analysis chapter **presents and analyses data collected from a research**. Some of the major issues discussed in this section include the response rate, the demographic profile of the respondents and the main research findings which are discussed as per objective.

What are the 5 methods of data presentation?

Some of the popular ways of presenting the data includes **Line graph, column chart, box pot, vertical bar, scatter plot**.

What is the importance of diagrams in statistics?

Diagrams play an important role in statistical data presentation. Diagrams are nothing but geometrical figures like lines, bars, circles, squares, etc. Diagrammatic data presentation **allows us to understand the data in an easier manner**.

Measure of Central Tendency, Variance

Central tendency and variation are two measures used in statistics to summarize data. **Measure of central tendency shows where the center or middle of the data set is located**, whereas measure of variation shows the dispersion among data values.

Central tendency focuses on the central distribution of data through a single value. Types of central tendency in real-life are average marks, rainfall, income, etc. Three commonly used measures of central tendency such as **arithmetic mean, median, and mode**.

How do you measure central tendency and variability?

Main Points

1. Measures of central tendency tell us what is common or typical about our variable.
2. Three measures of central tendency are the mode, the median and the mean.
3. The mode is used almost exclusively with nominal-level data, as it is the only measure of central tendency available for such variables.

Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

For more details

- Mode: the most frequent value.
- Median: the middle number in an ordered dataset.
- Mean: the sum of all values divided by the total number of values.

In addition to central tendency, the variability and distribution of your dataset is important to understand when performing descriptive statistics.

What Is Variance? | Definition, Examples & Formulas

The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean.

Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

Unlike range and interquartile range, variance is **a measure of dispersion that takes into account the spread of all data points in a data set**. It's the measure of dispersion the most often used, along with the standard deviation, which is simply the square root of the variance.

What is variance in simple terms?

Variance is **a measure of how data points differ from the mean**. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value. Variance means to find the expected difference of deviation from actual value.

What is variance and deviation?

The variance measures the average degree to which each point differs from the mean. While standard deviation is the square root of the variance, variance is the average of all data points within a group. The two concepts are useful and significant for traders, who use them to measure market volatility.

How do you calculate the variance?

The variance for a population is calculated by: **Finding the mean(the average). Subtracting the mean from each number in the data set and then squaring the result**. The results are squared to make the negatives positive.

Why do we calculate variance?

Investors use variance **to see how much risk an investment carries and whether it will be profitable**. Variance is also used in finance to compare the relative performance of each asset in a portfolio to achieve the best asset allocation. The square root of the variance is the standard deviation.

Understanding of Normal Distribution

What Is a 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".

KEY TAKEAWAYS

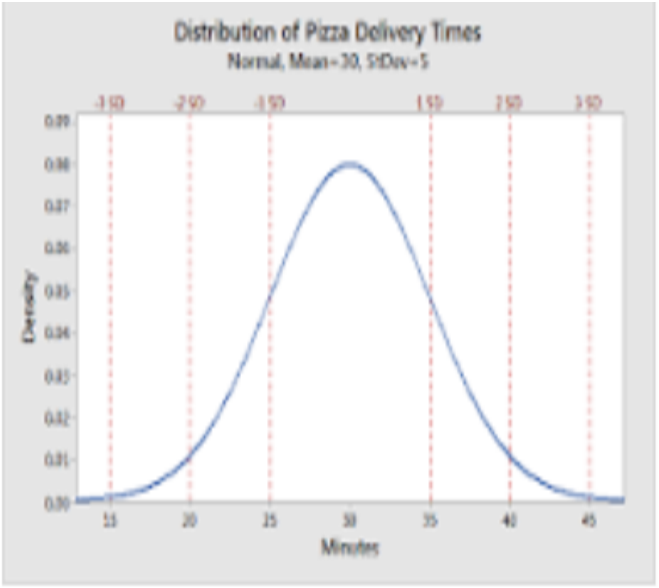
- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

What is a normal distribution in simple terms?

A normal distribution is **an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme**.

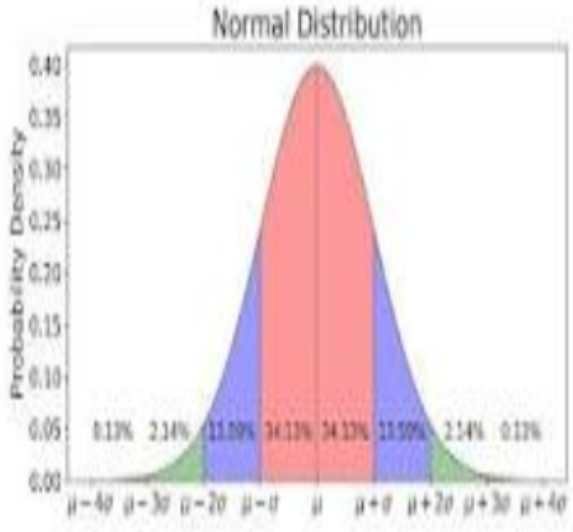
What is the importance of understanding normal distribution?

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because **it accurately describes the distribution of values for many natural phenomena**.



What is an example of normal distribution?

Height. **Height of the population** is the example of normal distribution. Most of the people in a specific population are of average height. The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short.

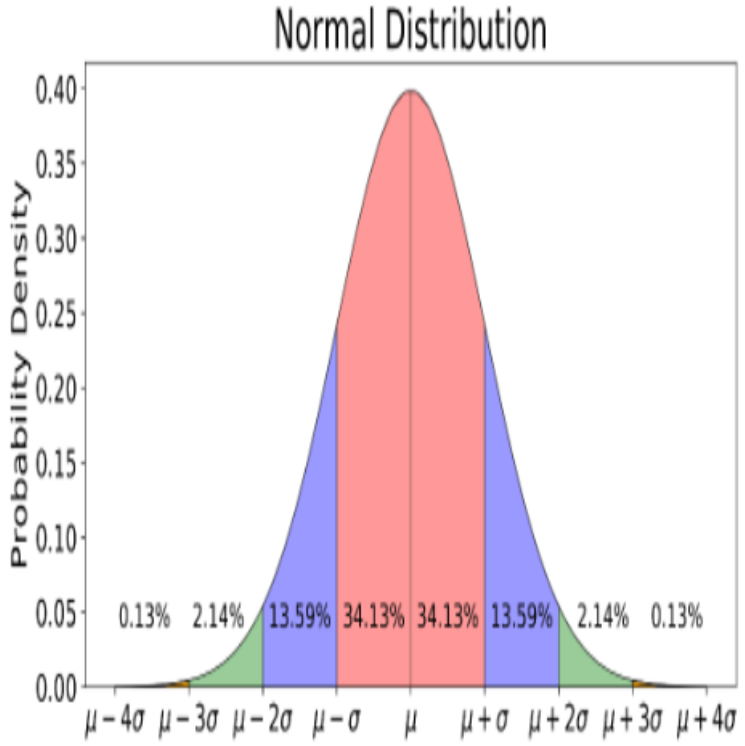


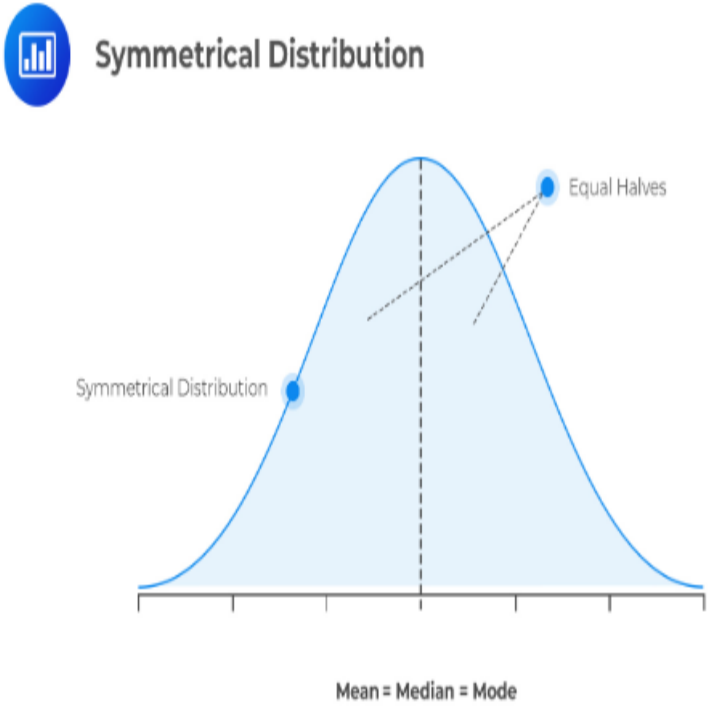
What is the benefit of normal distribution?

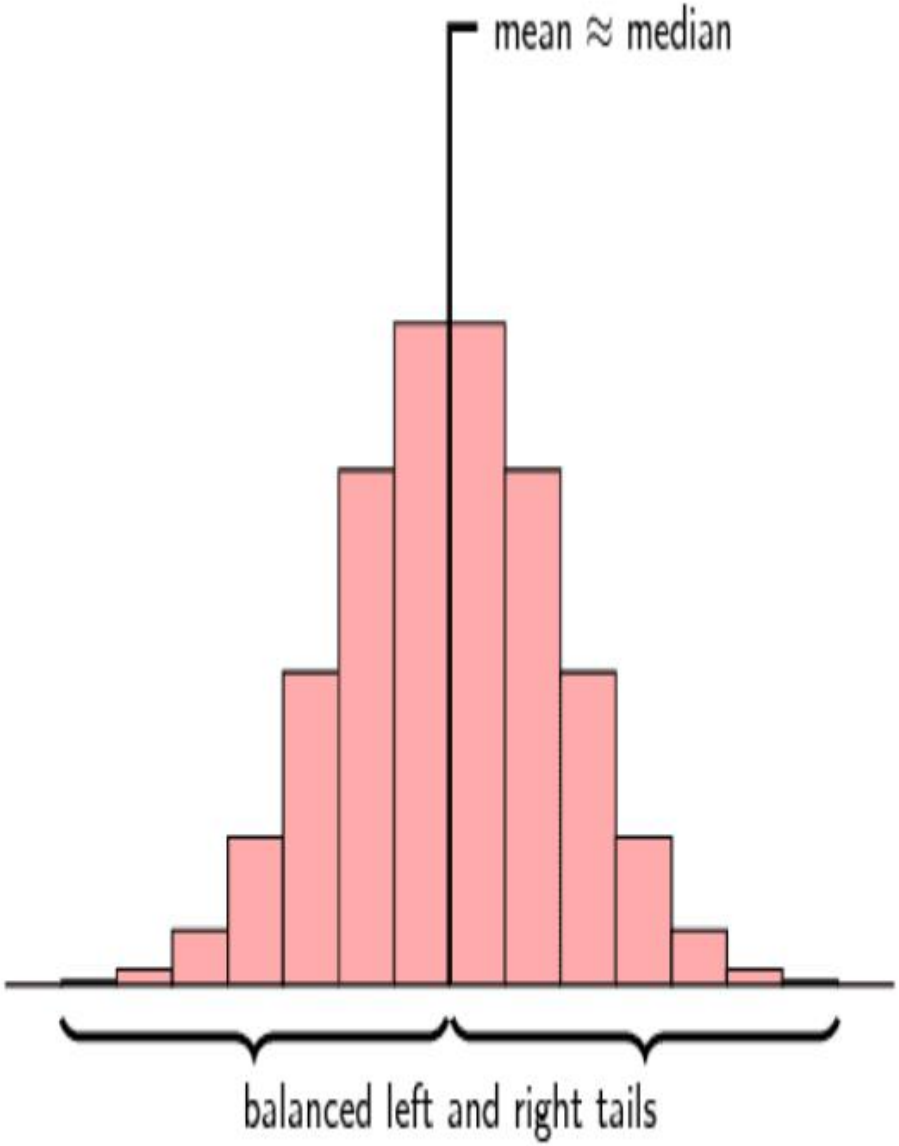
Answer. The first advantage of the normal distribution is that it is **symmetric and bell-shaped**. This shape is useful because it can be used to describe many populations, from classroom grades to heights and weights

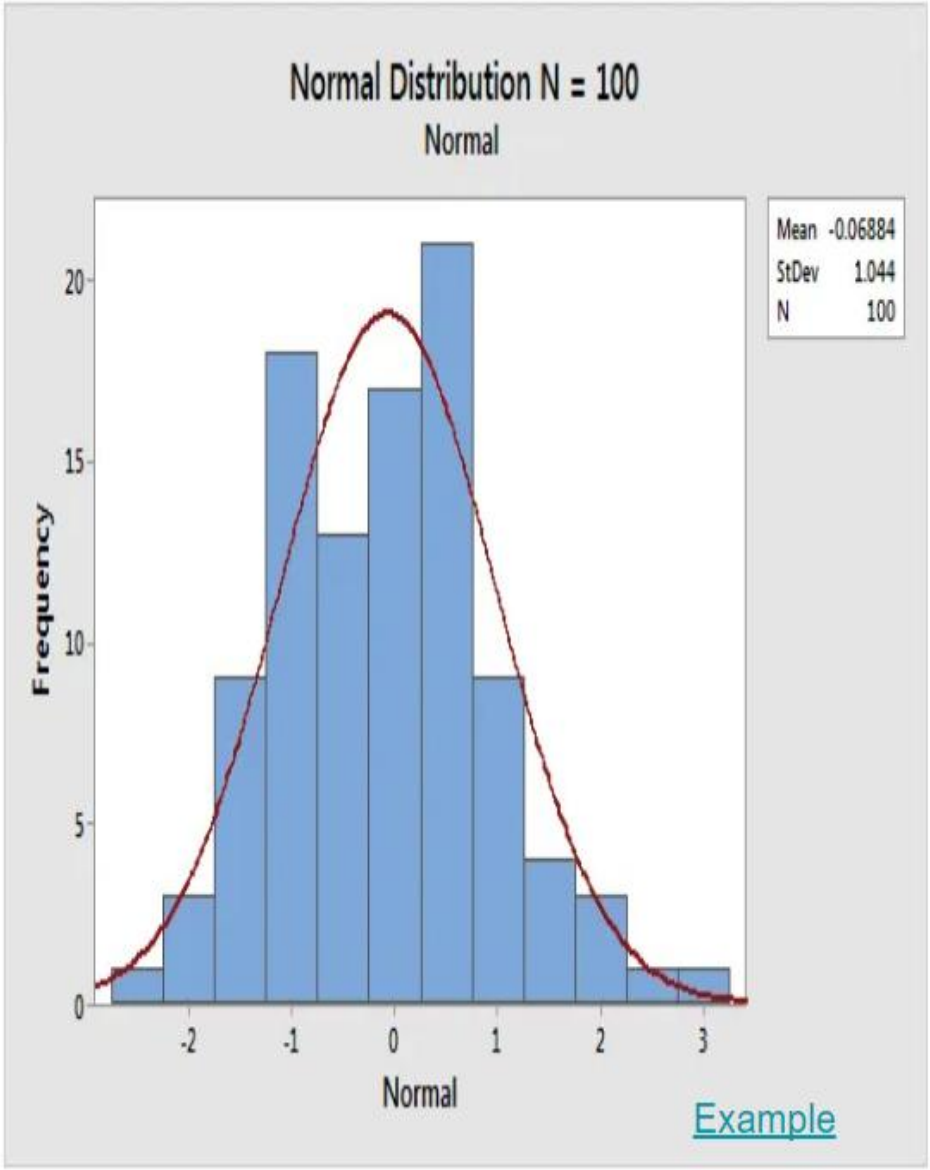
What is the difference between normal distribution and symmetric distribution?

In a symmetrical distribution the two sides of the distribution are a mirror image of each other. **A normal distribution is a true symmetric distribution of observed values**. When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape.






mean ≈ median
balanced left and right tails


Normal Distribution N = 100
Normal
Mean -0.06884
StDev 1.044
N 100
Frequency
0
5
10
15
20
-2
-1
0
1
2
3
Normal

Example

Data Pre-processing

What is data preprocessing?

Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process. More recently, data preprocessing techniques have been adapted for training machine learning models and AI models and for running inferences against them.

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

There are several different tools and methods used for preprocessing data, including the following:

- sampling, which selects a representative subset from a large population of data;
- transformation, which manipulates raw data to produce a single input;
- denoising, which removes noise from data;
- imputation, which synthesizes statistically relevant data for missing values;
- normalization, which organizes data for more efficient access; and
- feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

These tools and methods can be used on a variety of data sources, including data stored in files or databases and streaming data.

Feature Engineering (Feature Extraction and Normalization)

- What are the features of raw data?

  In computing, raw data may have the following attributes: it may possibly contain human, machine, or instrument errors, it may not be validated; it might be in different area (colloquial) formats; uncoded or unformatted; or some entries might be "suspect" (e.g., outliers), requiring confirmation or citation.

What is feature selection?
Feature Selection is **the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data**. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

Linear Algebra essentials

What is Linear Algebra?

Linear algebra can be defined as a branch of mathematics that deals with the study of linear functions in vector spaces. When information related to linear functions is presented in an organized form then it results in a matrix. Thus, linear algebra is concerned with vector spaces, vectors, linear functions, the system of linear equations, and matrices. These concepts are a prerequisite for sister topics such as geometry and functional analysis.

Linear Algebra Definition

The branch of mathematics that deals with vectors, matrics, finite or infinite dimensions as well as a linear mapping between such spaces is defined as linear algebra. It is used in both pure and applied mathematics along with different technical forms such as physics, engineering, natural sciences, etc.

Branches of Linear Algebra

Linear algebra can be categorized into three branches depending upon the level of difficulty and the kind of topics that are encompassed within each. These are elementary, advanced, and applied linear algebra. Each branch covers different aspects of matrices, vectors, and linear functions.

Elementary Linear Algebra

Elementary linear algebra introduces students to the basics of linear algebra. This includes simple matrix operations, various computations that can be done on a system of linear equations, and certain aspects of vectors. Some important terms associated with elementary linear algebra are given below:

Scalars - A scalar is a quantity that only has magnitude and not direction. It is an element that is used to define a vector space. In linear algebra, scalars are usually real numbers.

Vectors - A vector is an element in a vector space. It is a quantity that can describe both the direction and magnitude of an element.

Vector Space - The vector space consists of vectors that may be added together and multiplied by scalars.

Matrix - A matrix is a rectangular array wherein the information is organized in the form of rows and columns. Most linear algebra properties can be expressed in terms of a matrix.

Matrix Operations - These are simple arithmetic operations such as addition, subtraction, and multiplication that can be conducted on matrices.

Advanced Linear Algebra

Once the basics of linear algebra have been introduced to students the focus shifts on more advanced concepts related to linear equations, vectors, and matrices. Certain important terms that are used in advanced linear algebra are as follows:

Linear Transformations - The transformation of a function from one vector space to another by preserving the linear structure of each vector space.

Inverse of a Matrix - When an inverse of a matrix is multiplied with the given original matrix then the resultant will be the identity matrix. Thus, $A^{-1}A = I$.

Eigenvector - An eigenvector is a non-zero vector that changes by a scalar factor (eigenvalue) when a linear transformation is applied to it.

Linear Map - It is a type of mapping that preserves vector addition and vector multiplication.

Linear Algebra Topics

The topics that come under linear algebra can be classified into three broad categories. These are linear equations, matrices, and vectors. All these three categories are interlinked and need to be understood well in order to master linear algebra.

Linear Equations

A linear equation is an equation that has the standard form $a_1x_1+a_2x_2+...+a_nx_n$a1x1+a2x2+...+anxn. It is the fundamental component of linear algebra. The topics covered under linear equations are as follows:

- Linear Equations in One variable
- Linear Equations in Two Variables
- Simultaneous Linear Equations
- Solving Linear Equations
- Solutions of a Linear Equation
- Graphing Linear Equations
- Applications of Linear equations

Straight Line

Vectors

In linear algebra, there can be several operations that can be performed on vectors such as multiplication, addition, etc. Vectors can be used to describe quantities such as the velocity of moving objects. Some crucial topics encompassed under vectors are as follows:

- Types of Vectors
- Dot Product
- Cross Product
- Addition of Vectors

Matrices

A matrix is used to organize data in the form of a rectangular array. It can be represented as $A_{m \times n}$$A_{m \times n}$. Here, m represents the number of rows and n denotes the number of columns in the matrix. In linear algebra, a matrix can be used to express linear equations in a more compact manner. The topics that are covered under the scope of matrices are as follows:

- Matrix Operations
- Determinant
- Transpose of a Matrix
- Types of a Matrix

For More Details

Linear Algebra and its Applications

Linear algebra is used in almost every field. Simple algorithms also make use of linear algebra topics such as matrices. Some of the applications of linear algebra are given as follows:

- **Signal Processing** - Linear algebra is used in encoding and manipulating signals such as audio and video signals. Furthermore, it is required in the analysis of such signals.
- **Linear Programming** - It is an optimizing technique that is used to determine the best outcome of a linear function.
- **Computer Science** - Data scientists use several linear algebra algorithms to solve complicated problems.
- **Prediction Algorithms** - Prediction algorithms use linear models that are developed using concepts of linear algebra.

Linear Combination

What is Linear Combination?

A **linear equation** is an equation where the highest power of a variable is always 1.

This means we can use x and y as variables, but not x2ory3x2ory3.

A **linear equation with one variable** is an equation with only one variable, e.g., x+3=6.

This is very simple to solve.

A **linear equation with two variables** is something like x+2y=3, where there are two unknowns (x and y).

There is often a need to solve two of these equations simultaneously, e.g., 5x+y=17 and 3x+y=15.

There is more than one way of solving a system with two linear equations.

Many people know the way where one variable from one equation is expressed in terms of the other one, e.g., y=17-5x, and then substituted into the other equation.

This is not the linear combination method.

The **linear combination method** is a precise way of solving these kinds of equations.

Linear combination definition:

Using the linear combination method, a system of two linear equations is solved by *combining the two equations* to eliminate one of the variables.

How are linear combinations performed?

A linear equation with two variables is something like $x+2y=3$, where there are two unknowns (x and y).

There is often a need to solve two of these equations at the same time, e.g. $5x+y=17$ and $3x+y=15$.

The linear combination method solves a system of two linear equations by:

- Combining the two equations to eliminate one of the variables.
- Once one of the variables is eliminated, an equation with only one variable is left, and the value of the variable can be determined.
- This answer can then be substituted into one of the equations to give the value of the other variable.

Linear combinations of vectors

Most of the times, in linear algebra we deal with linear combinations of column vectors (or row vectors), that is, matrices that have only one column (or only one row).

Linear transformation and Matrices

A linear transformation can also be seen as a simple function. In functions, we usually have a scalar value as an input to our function. But rarely so far, we have experienced that input into a function can be a vector. So, **a linear transformation is actually a function that maps an input vector into an output vector.**

For a linear transformation, both input and output vectors are of the same length.

One of the most famous example of a linear transformation is the Discrete Fourier Transform. For instance, this transformation takes as an input a sequence of N signal samples and these samples are then mapped with the Fourier transform into a sequence of another N samples. These new samples are actually complex numbers in a Fourier domain. With complex numbers we can capture the amplitude in the frequency domain and phase (time-shift) of our original input signal.

Example

```
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt

%matplotlib inline
```

In [2]:

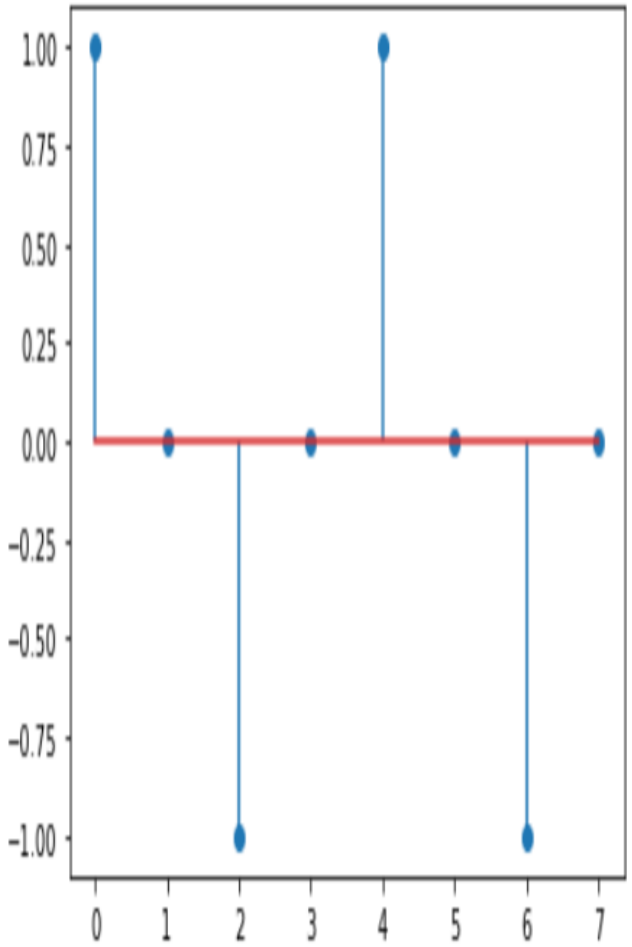
```
# A toy example shows how a sequence of samples from
# a time domain is mapped into a frequency domain using
# a Discrete Fourier Transformation.

k = np.arange(0,8)
x = np.cos(2*k*np.pi/4)

plt.stem(k, x, use_line_collection=True)
```

Out[2]:

<StemContainer object of 3 artists>



$$\begin{bmatrix}1\\2\\3\end{bmatrix} \quad f(\vec{v}) \quad \begin{bmatrix}2\\4\\6\end{bmatrix}$$

Input Vector **Output vector**

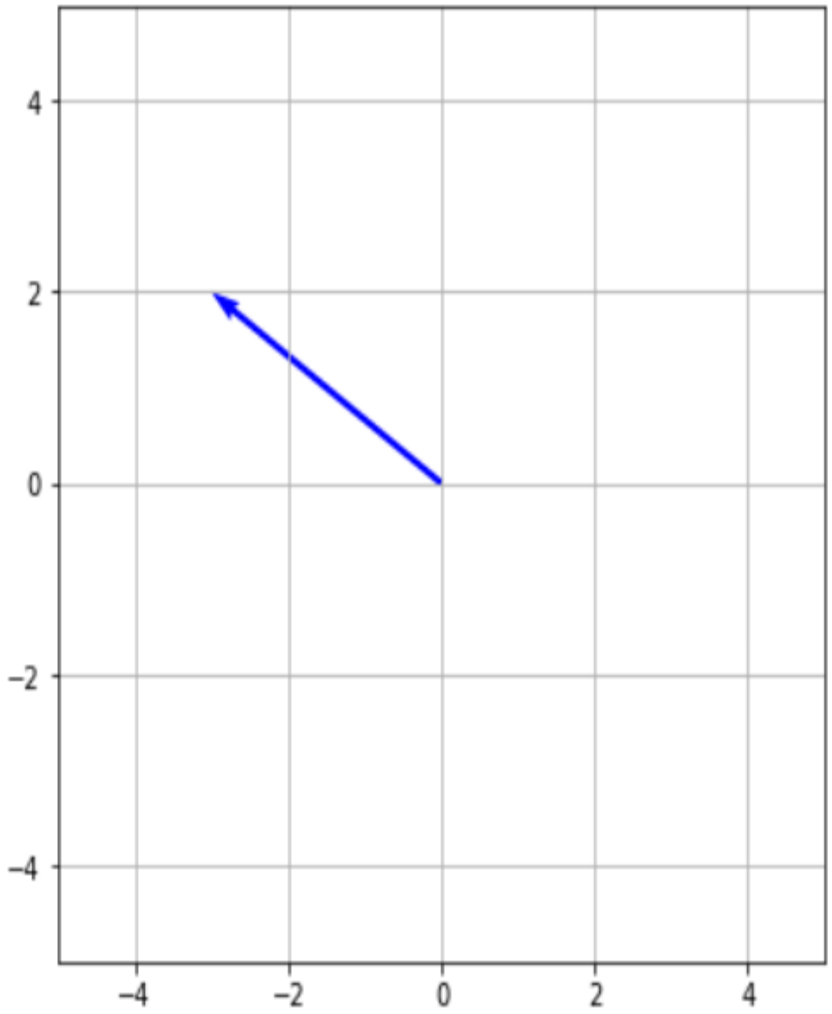
Ok, to make things simple again, we forget on complex numbers and we observe again our 2D coordinate system. We recall that a single vector in a 2D plane is represented with a pair (x, y). If we map this vector to another one, we say that this is actually a transformation. Recall that sometimes we refer to a vector as a movement. Then, with a linear transformation we are moving that vector again in our plane to get the output vector. Therefore, vectors can be seen as a displacement vectors and by transforming them we are actually moving them in some particular way.

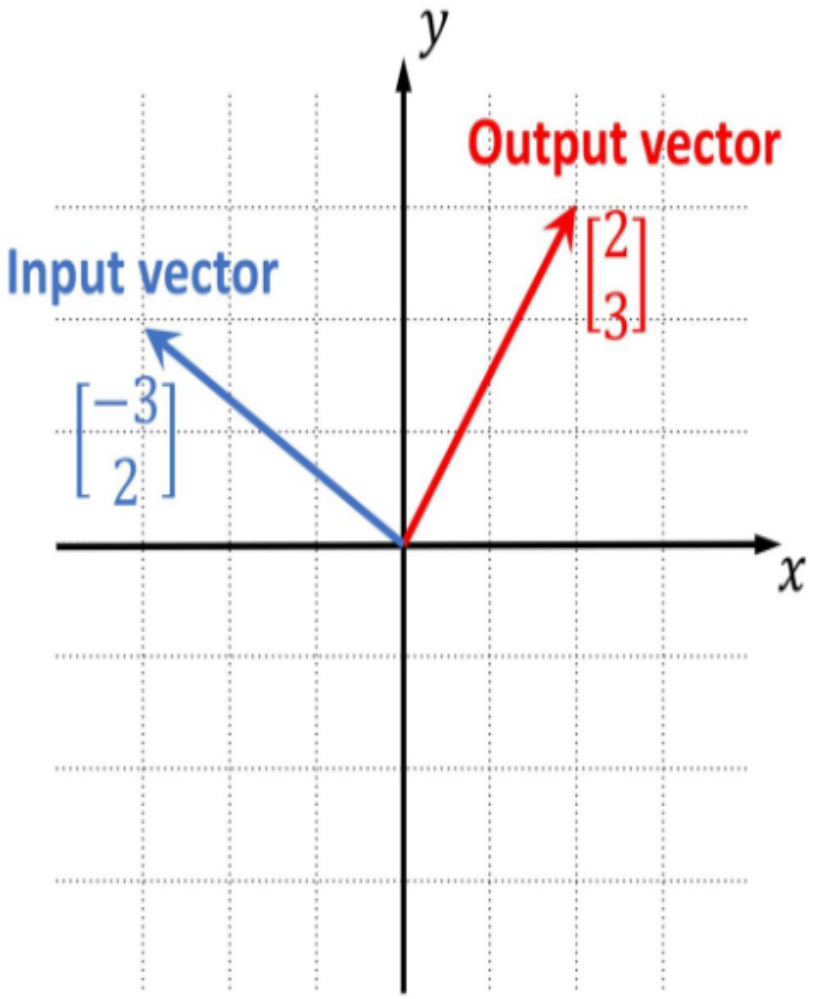
The word "transformation" suggests an association with the movement.

```
vec = np.array([[-3],[2]])
origin = np.zeros(vec.shape) # origin
point

plt.figure(figsize=(6,6))
plt.quiver(*origin, *vec, color=['b'],
scale=1, units='xy')

plt.grid()
plt.xlim(-5,5)
plt.ylim(-5,5)
plt.gca().set_aspect('equal')
plt.show()
```





Moreover, this same transform can be applied not only to a single vector, but can be actually applied on the whole set of vectors. So, basically, let's say that we want to transform the whole plane and to see where majority of the vectors from that plane will be mapped. One way to visualize this is to represent vectors not as displacement arrows, but as points (positions). Then, we can map each of these points and observe where they will land after the transformation. This will give us an idea how our transformation actually looks like.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$$

A 2×2 Matrix as a linear transformation

Now, it's interesting that actually the whole transformation is defined with two transformed basis vectors, and then, we can map the whole 2-D plane if we know the transformed basis vectors. Each of these vectors is specified with just two numbers: in this case [1−230]. Then, using these two vectors we can put them into a 2×2 matrix in a such a way that we stack these vectors along the columns and now this 2×2 matrix actually represent a useful matrix that we can use for further vector processing.

Actually, it's just the scaling two column vectors and then summing them and this is what we get as the resulting output. This can be more intuitive way to think about the matrix-vector multiplication.

Linear algebra in Neural Networks

A neural network is a powerful mathematical model combining linear algebra, biology and statistics to solve a problem in a unique way. The network takes a given amount of inputs and then calculates a specified number of outputs aimed at targeting the actual result.

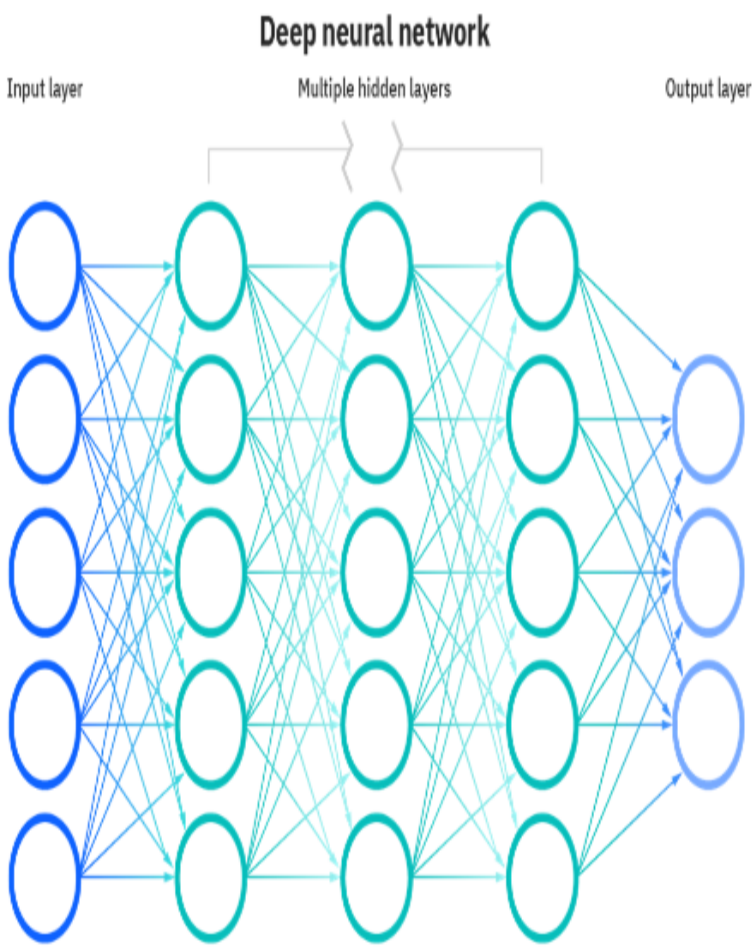
Neural Networks

Neural networks reflect the behavior of the human brain, allowing computer programs to recognize patterns and solve common problems in the fields of AI, machine learning, and deep learning.

What are neural networks?

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

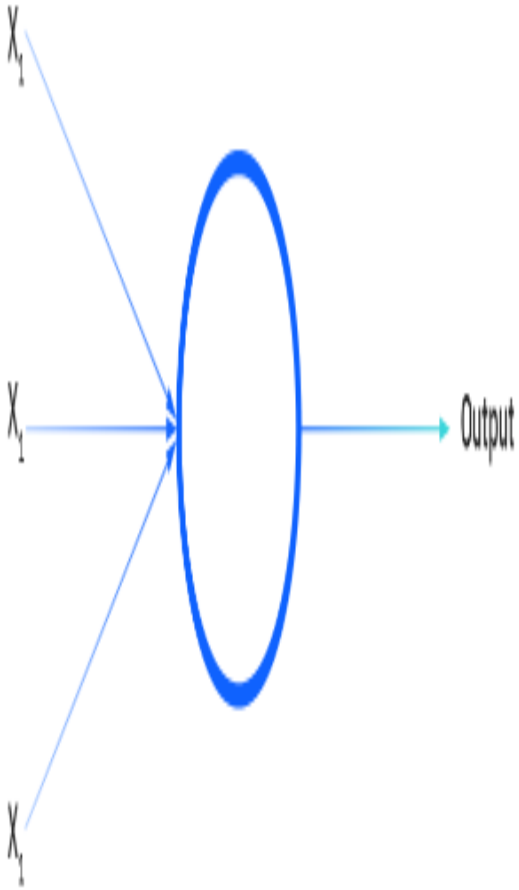


Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity.

Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm.

Types of neural networks

Neural networks can be classified into different types, which are used for different purposes. While this isn't a comprehensive list of types, the below would be representative of the most common types of neural networks that you'll come across for its common use cases:



The perceptron is the oldest neural network, created by Frank Rosenblatt in 1958. It has a single neuron and is the simplest form of a neural network:

Feedforward neural networks, or multi-layer perceptrons (MLPs), are comprised of an input layer, a hidden layer or layers, and an output layer. While these neural networks are also commonly referred to as MLPs, it's important to note that they are actually comprised of sigmoid neurons, not perceptrons, as most real-world problems are nonlinear. Data usually is fed into these models to train them, and they are the foundation for computer vision, natural language processing, and other neural networks.

Convolutional neural networks (CNNs) are similar to feedforward networks, but they're usually utilized for image recognition, pattern recognition, and/or computer vision. These networks harness principles from linear algebra, particularly matrix multiplication, to identify patterns within an image.

Recurrent neural networks (RNNs) are identified by their feedback loops. These learning algorithms are primarily leveraged when using time-series data to make predictions about future outcomes, such as stock market predictions or sales forecasting.

Neural networks vs. deep learning

Deep Learning and neural networks tend to be used interchangeably in conversation, which can be confusing. As a result, it's worth noting that the "deep" in deep learning is just referring to the depth of layers in a neural network. A neural network that consists of more than three layers—which would be inclusive of the inputs and the output—can be considered a deep learning algorithm. A neural network that only has two or three layers is just a basic neural network.

Probability

Probabilistic reasoning in Artificial intelligence

Uncertainty:

Till now, we have learned knowledge representation using first-order logic and propositional logic with certainty, which means we were sure about the predicates. With this knowledge representation, we might write A→B, which means if A is true then B is true, but consider a situation where we are not sure about whether A is true or not then we cannot express this statement, this situation is called uncertainty.

So to represent uncertain knowledge, where we are not sure about the predicates, we need uncertain reasoning or probabilistic reasoning.

Causes of uncertainty:

Following are some leading causes of uncertainty to occur in the real world.

1. Information occurred from unreliable sources.
2. Experimental Errors
3. Equipment fault
4. Temperature variation
5. Climate change.

Probabilistic reasoning:

Probabilistic reasoning is a way of knowledge representation where we apply the concept of probability to indicate the uncertainty in knowledge. In probabilistic reasoning, we combine probability theory with logic to handle the uncertainty.

We use probability in probabilistic reasoning because it provides a way to handle the uncertainty that is the result of someone's laziness and ignorance.

In the real world, there are lots of scenarios, where the certainty of something is not confirmed, such as "It will rain today," "behavior of someone for some situations," "A match between two teams or two players." These are probable sentences for which we can assume that it will happen but not sure about it, so here we use probabilistic reasoning.

Need of probabilistic reasoning in AI:

- When there are unpredictable outcomes.
- When specifications or possibilities of predicates becomes too large to handle.
- When an unknown error occurs during an experiment.

In probabilistic reasoning, there are two ways to solve problems with uncertain knowledge:

- **Bayes' rule**
- **Bayesian Statistics**

As probabilistic reasoning uses probability and related terms, so before understanding probabilistic reasoning, let's understand some common terms:

Probability: Probability can be defined as a chance that an uncertain event will occur. It is the numerical measure of the likelihood that an event will occur. The value of probability always remains between 0 and 1 that represent ideal uncertainties.

1. $0 \leq P(A) \leq 1$, where P(A) is the probability of an event A.
1. $P(A) = 0$, indicates total uncertainty in an event A.
1. $P(A) = 1$, indicates total certainty in an event A.

We can find the probability of an uncertain event by using the below formula.

$$\textbf{Probability of occurrence} = \frac{\text{Number of desired outcomes}}{\text{Total number of outcomes}}$$

- $P(\neg A)$ = probability of a not happening event.
- $P(\neg A) + P(A) = 1$.

Event: Each possible outcome of a variable is called an event.

Sample space: The collection of all possible events is called sample space.

Random variables: Random variables are used to represent the events and objects in the real world.

Prior probability: The prior probability of an event is probability computed before observing new information.

Posterior Probability: The probability that is calculated after all evidence or information has taken into account. It is a combination of prior probability and new information.

Conditional probability:

Conditional probability is a probability of occurring an event when another event has already happened.

Let's suppose, we want to calculate the event A when event B has already occurred, "the probability of A under the conditions of B", it can be written as:

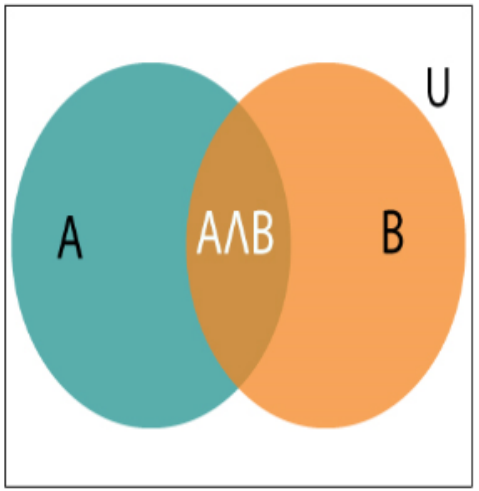
$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Where $P(A \wedge B)$= Joint probability of a and B

P(B)= Marginal probability of B.

If the probability of A is given and we need to find the probability of B, then it will be given as:

$$P(B \mid A) = \frac{P(A \wedge B)}{P(A)}$$



It can be explained by using the below Venn diagram, where B is occurred event, so sample space will be reduced to set B, and now we can only calculate event A when event B is already occurred by dividing the probability of **$P(A \wedge B)$ by P(B)**.

Example:

In a class, there are 70% of the students who like English and 40% of the students who likes English and mathematics, and then what is the percent of students those who like English also like mathematics?

Solution:

Let, A is an event that a student likes Mathematics

B is an event that a student likes English.

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{0.4}{0.7} = 57\%$$

Hence, 57% are the students who like English also like Mathematics.

Bayes' theorem in Artificial intelligence

Bayes' theorem:

Bayes' theorem is also known as **Bayes' rule, Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge.

In probability theory, it relates the conditional probability and marginal probabilities of two random events.

Bayes' theorem was named after the British mathematician **Thomas Bayes**. The **Bayesian inference** is an application of Bayes' theorem, which is fundamental to Bayesian statistics.

It is a way to calculate the value of P(B|A) with the knowledge of P(A|B).

Bayes' theorem allows updating the probability prediction of an event by observing new information of the real world.

Example: If cancer corresponds to one's age then by using Bayes' theorem, we can determine the probability of cancer more accurately with the help of age.

Bayes' theorem can be derived using product rule and conditional probability of event A with known event B:

As from product rule we can write:

$P(A \wedge B) = P(A|B)\, P(B)$ or

Similarly, the probability of event B with known event A:

$P(A \wedge B) = P(B|A)\, P(A)$

Equating right hand side of both the equations, we will get:

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)} \qquad \text{....(a)}$$

The above equation (a) is called as Bayes' rule or Bayes' theorem. This equation is basic of most modern AI systems for probabilistic inference.

I

It shows the simple relationship between joint and conditional probabilities. Here,

P(A|B) is known as **posterior**, which we need to calculate, and it will be read as Probability of hypothesis A when we have occurred an evidence B.

P(B|A) is called the likelihood, in which we consider that hypothesis is true, then we calculate the probability of evidence.

P(A) is called the **prior probability**, probability of hypothesis before considering the evidence

P(B) is called **marginal probability**, pure probability of an evidence.

In the equation (a), in general, we can write P (B) = P(A)*P(B|Ai), hence the Bayes' rule can be written as:

$$P(A_i \mid B) = \frac{P(A_i) * P(B \mid A_i)}{\sum_{i=1}^{k} P(A_i) * P(B \mid A_i)}$$

Where A_1, A_2, A_3,....., A_n is a set of mutually exclusive and exhaustive events.

Applying Bayes' rule:

Bayes' rule allows us to compute the single term P(B|A) in terms of P(A|B), P(*B*), and P(A). This is very useful in cases where we have a good probability of these three terms and want to determine the fourth one. Suppose we want to perceive the effect of some unknown cause, and want to compute that cause, then the Bayes' rule becomes:

$$\mathbf{P(cause|effect)} = \frac{\mathbf{P(effect|cause)\,P(cause)}}{\mathbf{P(effect)}}$$

Example-1: Question: what is the probability that a patient has diseases meningitis with a stiff neck?

Given Data:

A doctor is aware that disease meningitis causes a patient to have a stiff neck, and it occurs 80% of the time. He is also aware of some more facts, which are given as follows:

- The Known probability that a patient has meningitis disease is 1/30,000.
- The Known probability that a patient has a stiff neck is 2%.

Let a be the proposition that patient has stiff neck and b be the proposition that patient has meningitis. , so we can calculate the following as:

P(a|b) = 0.8

P(b) = 1/30000

P(a)= .02

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} = \frac{0.8*(\frac{1}{30000})}{0.02} = 0.001333333.$$

Hence, we can assume that 1 patient out of 750 patients has meningitis disease with a stiff neck.

Example-2: Question: From a standard deck of playing cards, a single card is drawn. The probability that the card is king is 4/52, then calculate posterior probability P(King|Face), which means the drawn face card is a king card.

Solution:

$$P(king|face) = \frac{P(Face|king) * P(King)}{P(Face)} \quad .....(i)$$

P(king): probability that the card is King= 4/52= 1/13

P(face): probability that a card is a face card= 3/13

P(Face|King): probability of face card when we assume it is a king = 1

Putting all values in equation (i) we will get:

$$P(king|face) = \frac{1 * (\frac{1}{13})}{(\frac{3}{13})} = 1/3,$$ **it is a probability that a face card is a king card.**

Application of Bayes' theorem in Artificial intelligence:

Following are some applications of Bayes' theorem:

- It is used to calculate the next step of the robot when the already executed step is given.
- Bayes' theorem is helpful in weather forecasting.
- It can solve the Monty Hall problem.

Bayesian Belief Network in artificial intelligence

Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertainty. We can define a Bayesian network as:

"A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."

It is also called a **Bayes network, belief network, decision network**, or **Bayesian model**.

Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.

Joint and Marginal Probabilities

Probability of One Random Variable

Probability quantifies the likelihood of an event.

Specifically, it quantifies how likely a specific outcome is for a random variable, such as the flip of a coin, the roll of a dice, or drawing a playing card from a deck.

For a random variable *x*, *P(x)* is a function that assigns a probability to all values of *x*.

- Probability Density of x = P(x)

The probability of a specific event *A* for a random variable x is denoted as *P(x=A)*, or simply as *P(A)*.

- Probability of Event A = P(A)

Probability is calculated as the number of desired outcomes divided by the total possible outcomes, in the case where all outcomes are equally likely.

- Probability = (number of desired outcomes) / (total number of possible outcomes)

This is intuitive if we think about a discrete random variable such as the roll of a die. For example, the probability of a die rolling a 5 is calculated as one outcome of rolling a 5 (1) divided by the total number of discrete outcomes (6) or 1/6 or about 0.1666 or about 16.666%.

The sum of the probabilities of all outcomes must equal one. If not, we do not have valid probabilities.

- Sum of the Probabilities for All Outcomes = 1.0.

The probability of an impossible outcome is zero. For example, it is impossible to roll a 7 with a standard six-sided die.

- Probability of Impossible Outcome = 0.0

The probability of a certain outcome is one. For example, it is certain that a value between 1 and 6 will occur when rolling a six-sided die.

- Probability of Certain Outcome = 1.0

The probability of an event not occurring, called the complement.

This can be calculated by one minus the probability of the event, or *1 – P(A)*. For example, the probability of not rolling a 5 would be 1 – P(5) or 1 – 0.166 or about 0.833 or about 83.333%.

- Probability of Not Event A = 1 – P(A)

Now that we are familiar with the probability of one random variable, let's consider probability for multiple random variables.

Probability of Multiple Random Variables

In machine learning, we are likely to work with many random variables.

For example, given a table of data, such as in excel, each row represents a separate observation or event, and each column represents a separate random variable.

Variables may be either discrete, meaning that they take on a finite set of values, or continuous, meaning they take on a real or numerical value.

As such, we are interested in the probability across two or more random variables.

This is complicated as there are many ways that random variables can interact, which, in turn, impacts their probabilities.

This can be simplified by reducing the discussion to just two random variables (X, Y), although the principles generalize to multiple variables.

And further, to discuss the probability of just two events, one for each variable ($X=A$, $Y=B$), although we could just as easily be discussing groups of events for each variable.

Therefore, we will introduce the probability of multiple random variables as the probability of event A and event B, which in shorthand is $X=A$ and $Y=B$.

We assume that the two variables are related or dependent in some way.

There are three main types of probability we might want to consider; they are:

- **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. For example:

- The probability of a row of data is the joint probability across each input variable.
- The probability of a specific value of one input variable is the marginal probability across the values of the other input variables.
- The predictive model itself is an estimate of the conditional probability of an output given an input example.

Joint, marginal, and conditional probability are foundational in machine learning.

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.

Probability of Independence and Exclusivity

When considering multiple random variables, it is possible that they do not interact.

We may know or assume that two variables are not dependent upon each other instead are independent.

Alternately, the variables may interact but their events may not occur simultaneously, referred to as exclusivity.

Independence

If one variable is not dependent on a second variable, this is called independence or statistical independence.

This has an impact on calculating the probabilities of the two variables.

For example, we may be interested in the joint probability of independent events *A* and *B*, which is the same as the probability of *A* and the probability of *B*.

Probabilities are combined using multiplication, therefore the joint probability of independent events is calculated as the probability of event *A* multiplied by the probability of event *B*.

This can be stated formally as follows:

- **Joint Probability**: P(A and B) = P(A) * P(B)

As we might intuit, the marginal probability for an event for an independent random variable is simply the probability of the event.

It is the idea of probability of a single random variable that are familiar with:

- **Marginal Probability**: P(A)

We refer to the marginal probability of an independent probability as simply the probability.

Similarly, the conditional probability of *A* given *B* when the variables are independent is simply the probability of *A* as the probability of *B* has no effect. For example:

- **Conditional Probability**: P(A given B) = P(A)

We may be familiar with the notion of statistical independence from sampling. This assumes that one sample is unaffected by prior samples and does not affect future samples.

Many machine learning algorithms assume that samples from a domain are independent to each other and come from the same probability distribution, referred to as independent and identically distributed, or i.i.d. for short.

Exclusivity

If the occurrence of one event excludes the occurrence of other events, then the events are said to be mutually exclusive.

The probability of the events are said to be disjoint, meaning that they cannot interact, are strictly independent.

If the probability of event *A* is mutually exclusive with event *B*, then the joint probability of event *A* and event *B* is zero.

- P(A and B) = 0.0

Instead, the probability of an outcome can be described as event *A* or event *B*, stated formally as follows:

- P(A or B) = P(A) + P(B)

The "or" is also called a union and is denoted as a capital "*U*" letter; for example:

- P(A or B) = P(A U B)

If the events are not mutually exclusive, we may be interested in the outcome of either event.

The probability of non-mutually exclusive events is calculated as the probability of event *A* and the probability of event *B* minus the probability of both events occurring simultaneously.

This can be stated formally as follows:

- P(A or B) = P(A) + P(B) – P(A and B)

Probability Distribution

n order to understand probability distribution, let us first understand what probability is. Probability is the measure of the likelihood of an event occurring in an experiment. In simple terms, it tells us how likely is it that the event will occur. The value of the probability of an event occurring ranges from 0 (being least probable) to 1 (being most probable).

The probability distribution is a function that provides the probabilities of different outcomes for experimentation. It shows the possible values that a random variable can take and how often do these values occur.

In probability distribution, the sum of all these probabilities always aggregates to 1. In the data science domain, one of the usages of the probability distribution is for calculating confidence intervals and for calculating the critical regions in the hypothesis tests.

Continuous and Discrete Distributions

The type of probability distribution to be used depends upon whether the variable contains discrete values or continuous values. A discrete distribution can only take a limited set of values whereas continuous distributions can take in any value within the specified range.

The continuous distributions are represented in terms of probability density as there can be infinite values in a certain range and the probability of each value will be zero. In the case of discrete distribution, we can obtain a probability for each value as the number of values is limited.

Types of Distributions – Discrete Distribution

Binomial Distribution

It is a type of distribution where the number of outcomes in a single trial is only two. Each trial is independent of another trial; that is, the outcome of each trial does not have an impact on the outcome of other trials. The trials that are conducted in this experiment are identical to each other.

Thus, the probability of success and failure would be the same for each trial. For example, if the probability of success for a trial is 0.8 (which means the probability of failure would be 0.2), then it will be the same for the rest of the trials as well.

Multi nominal Distribution

This is the generalized version of binomial distribution where the number of outcomes can be greater than two. The other properties of this distribution are similar to that of the binomial distribution. For example, consider when a fair die is rolled, the probability of each outcome is going to be the same for all trials as these trials are independent of each other.

Bernoulli's Distribution

This is another variant of Binomial distribution. It is a special case of Binomial distribution where the number of trials conducted in an experiment is 1 (n = 1). As there is only one trial, it can be defined using only one parameter (p) which is generally the probability of success.

Some examples of such events are as follows: **a team will win a championship or not, a student will pass or fail an exam, and a rolled dice will either show a 6 or any other number**.

A Bernoulli trial is **an experiment with two possible outcomes: Success or Failure**. "Success" in one of these trials means that you're getting the result you're measuring. For example: If you flip a coin 100 times to see how many heads you get, then the Success is getting heads and a Failure is getting tails.

Negative Binomial Distribution

The following conditions in a negative binomial distribution differ from the binomial distribution:

- The number of trials conducted in an experiment is not fixed.
- The random variable indicates the number of trials required to attain a desired number of successes.

For binomial distribution, the random variable is the number of successes required i.e. We focus only on the number of successes no matter how many trails fail. But in the case of negative binomial, it focuses on how many trials will be required for achieving the number of successes i.e. The number of failures (negatives) is also brought into consideration which is why it is called a negative binomial distribution.

The process is continued only till the desired number of successes have been attained. This causes the number of trials for an experiment to be arbitrary. It is also called Pascal Distribution.

Example: Take a standard deck of cards, shuffle them, and choose a card. Replace the card and repeat until you have drawn two aces. Y is the number of draws needed to draw two aces. As the number of trials isn't fixed (i.e. you stop when you draw the second ace), this makes it a negative binomial distribution

Poisson Distribution

Poisson Distribution provides the probability of a discrete number of events occurring in a specific period of time, provided we know the average number of events that occurred during the same period. These events occur independently and have no effect over other events. For implementing this distribution, it assumes that the rate of occurrence remains constant over the time period.

Examples of Poisson distributions

- A death by horse kick is an "event."
- The time interval is one year.
- The mean number of events per time interval, λ, is 0.61.
- The number of deaths by horse kick in a specific year is k.

Discrete Uniform Distribution

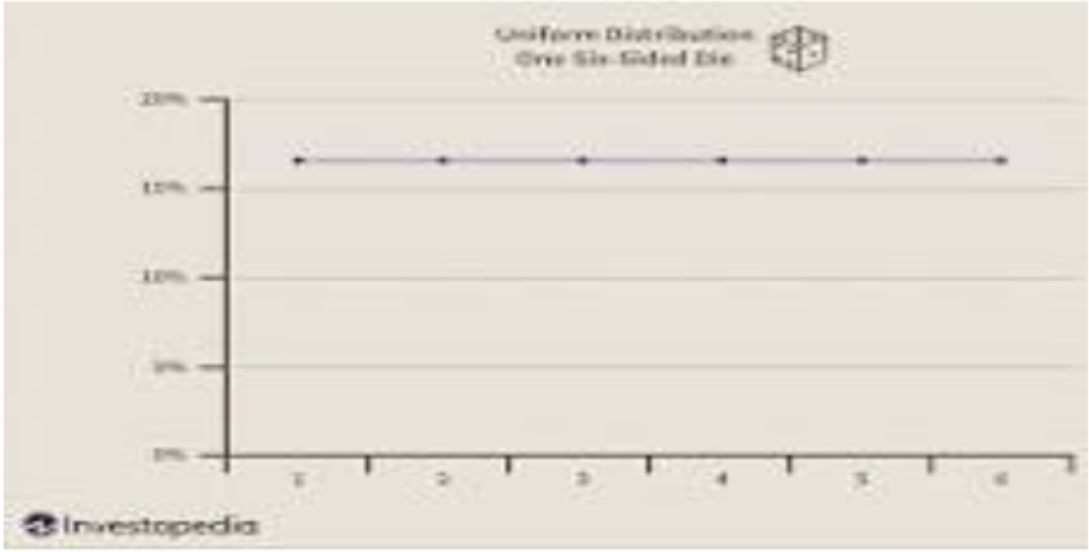
In uniform distribution, the probabilities of all the outcomes are equal. For example, consider when a fair die is rolled, the probability of any outcome ranging from 1 to 6 is going to be equal. The probability mass function of this distribution is 1/n where n is the total number of discrete values.

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. A coin also has a uniform distribution because the probability of getting either heads or tails in a coin toss is the same.

Types of Distributions – Continuous Distribution

Continuous Uniform Distribution

The uniformity in the distribution can be applied to continuous values as well. It indicates that the probability distribution is uniform between the specified range. It is also called a rectangular distribution due to the shape it takes when plotted on a graph.



Normal Distribution

A normal distribution (also known as a bell curve) is a type of continuous distribution that is symmetrical from both the ends of the mean. It generally indicates the one-half of the samples lie on the left side of the mean, while the other half lies on the right side. For a normal distribution, the mean, the mode, and the median are equal.

Normally distributed data generally follow the empirical rule. The empirical rule shows the spread of the data in terms of standard deviation and mean as follows: –

- 68% probability that the random variable falls within 1 standard deviation of the mean.
- 95% probability that the random variable falls within 2 standard deviations of the mean.
- 99.7% probability that the random variable falls within 3 standard deviations of the mean.

T – Distribution

It is similar to a normal distribution, but it has a higher probability towards the extreme values of the data. This makes it more liable to take values that are farther from the mean. When plotted on a graph, the curve seems shorter and fatter than the normal distribution curve.

It is preferred when the number of samples is smaller in size. With the increase in the size of samples, the t-distribution curve starts to appear like a normal distribution curve. As the formulae for normal distribution and t- distribution are very complex and time-consuming to calculate, we instead compute the values of Z-score and T-score respectively.

Chi – Square Distribution

Chi-square distribution is the distribution of the summation of the square of the random variables taken from a normal distribution. The degrees of freedom used in this distribution is equal to the number of variables taken from the normal distribution. The mean of a chi-square distribution is equa to the number of degrees of freedom.

This distribution is widely used in calculating the confidence intervals and in hypothesis testing. It is specific case of gamma distribution. It is also used in the chi-square test which is the goodness of fit test for observed distribution which helps in indicating if the sample data is a good representation of the entire population.

Degree of Freedom

Degrees of freedom refers to **the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample**. Degrees of freedom are commonly discussed in relation to various forms of hypothesis testing in statistics, such as a chi-square.

To calculate degrees of freedom, **subtract the number of relations from the number of observations**. For determining the degrees of freedom for a sample mean or average, you need to subtract one (1) from the number of observations, n.

Degrees of freedom are important **for finding critical cutoff values for inferential statistical tests**. Depending on the type of the analysis you run, degrees of freedom typically (but not always) relate the size of the sample.

Random Variables

What is a Random Variable?

The term "Random Variable" is used extremely often within the realm of probability and statistics, but what does it mean? Well, a random variable is defined as a variable whose possible values are outcomes of a random phenomenon. In specific terms, it is a function that maps the outcomes of an unpredictable process in numerical terms, often represented as a real number. So, a random variable may represent the outcome of an experiment that has yet to be performed, or a value that is currently uncertain. However, there are a couple of properties that are required of random variables.

A random variable conveys the results of an objectively random process, like rolling a die, or a subjectively random process, like an individual who is uncertain of an outcome due to incomplete information. A random variable must be measurable, which allows for the assignment of probabilities to the potential outcome. Furthermore, its outcome sometimes depends on environmental factors, like wind during a coin toss, however these additional factors are often excluded.

Random variables have a domain defined by the set of all possible outcomes of an event. Additionally, they have a probability distribution. This distribution can be either continuous, measuring numerical values in an interval, or discrete, as specified by a list of countable values. Within probability theory, random variables are used as functions defined by a sample space whose outcomes are numerical values.

Applications of Random Variables

As mentioned above, random variables are very common within almost any facet of mathematics and/or the scientific method and are often used in computer science. Random variables can be either discrete or continuous, as defined by the context of their application.

Discrete - Coin Toss

Imagine a coin toss where, depending on the side of the coin landing face up, a bet of a dollar has been placed. The possibility of winning a dollar corresponding to the outcome of a coin toss before tossing the coin defines the random variable. The outcome of the coin toss is either heads, or tails, creating an equal probability of either outcome. Because the value of the random variable is defined as a real-valued dollar, the probability distribution is discrete.

Continuous - Height

Any random variable that is defined through measuring, rather than counting, is continuous. In this case, imagine wanting to study the effects of caffeine intake on height. One's height would be the continuous random variable as it is unknown before the completion of the experiment, and its value is taken from measuring within a range.

Random Variables in Machine Learning

Random variables are an invaluable tool within applications of machine learning. As a neural network makes decisions using machine learning, it creates functions for understanding possible outcomes. These possible outcomes are often defined by random variables.

Theory of Estimation, Estimation Process, Statistical Inference

The estimation is the process of providing numerical values of the unknown parameter to the population. There are mainly two types of estimation process Point estimation and Interval estimation and confidence interval is the part of the interval estimation. We will also discuss about the elements of the estimation like parameter, statistic and estimator.

The characteristics of estimators are – (i) **Unbiasedness** – This is desirable property of a good estimator. (ii) **Consistency** – An estimator is said to be consistent if increasing the sample size produces an estimate with smaller standard error. (iii) **Efficiency** – An estimator should be an efficient estimator. (iv) **Sufficiency** – An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

The methods of estimation are– (i) **Method of maximum likelihood**, (ii) **Method of least square**, (iii) **Method of minimum variance**, (iv) **Method of moments**.

Element For Estimation

Parameter

Parameter is an unknown numerical factor of the population. The primary interest of any survey lies in knowing the values of different measures of the population distribution of a variable of interest. The measures of population distribution involves its mean, standard deviation etc. which is calculated on the basis of the population values of the variable. In other words, the parameter is a functional form of all the population unit.

Statistic

Any statistical measure calculated on the basis of sample observations is called Statistic. Like sample mean, sample standard deviation, etc. Sample statistic are always known to us.

Estimator An estimator is a measure computed on the basis of sample values. It is a functional from of all sample observe prorating a representative value of the collected sample. **Relation Between Parameter And Statistic**

Parameter is a fixed measure describing the whole population (population being a group of people, things, animals, phenomena that share common characteristics.) A statistic is a characteristic of a sample, a portion of the target population. A parameter is fixed, unknown numerical value, while the statistic is a known number and a variable which depends on the portion of the population. Sample statistic and population parameters have different statistical notations: In population parameter, population proportion is represented by P, mean is represented by μ (Greek letter mu), σ2 represents variance, N represents population size, σ (Greek letter sigma) represents standard deviation, $\sigma_{\bar{x}}$ represents Standard error of mean, σ/μ represents Coefficient of variation, (X-μ)/σ represents standardized variate (z), and σp represents standard error of proportion.

In sample statistics, mean is represented by $\bar{x}$(x-bar), sample proportion is represented by $\hat{p}$(p-hat), s represents standard deviation, s2 represents variance, sample size is represented by n, $s_{\bar{x}}$ represents Standard error of mean, sp represents standard error of proportion, s/($\bar{x}$) represents Coefficient of variation, and (x-x)/s¯represents standardized variate (z).

What Is Estimation?

Estimation refers to the process by which one makes an idea about a population, based on information obtained from a sample.

Suppose we have a random sample $x1, x2, \ldots, xn$ on a variable x, whose distribution in the population involves an unknown parameter θ. It is required to find an estimate of θ on the basis of sample values. The estimation is done in two different ways: (i) Point Estimation, and (ii) Interval Estimation.

In point estimation, the estimated value is given by a single quantity, which is a function of sample observations. This function is called the 'estimator' of the parameter, and the value of the estimator in a particular sample is called an 'estimate'.

Interval estimation, an interval within which the parameter is expected to lie in given by using two quantities based on sample values. This is known as Confidence interval, and the two quantities which are used to specify the interval, are known as Confidence Limits.

Point Estimation

Many functions of sample observations may be proposed as estimators of the same parameter. For example, either the mean or median or mode of the sample values may be used to estimate the parameter μ of the normal distribution with probability density function

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Interval Estimation

In statistical analysis it is not always possible to find out an exact point estimate to form an idea about the population parameters. An approximately true picture can be formed if the sample estimations satisfy some important property such as unbiasedness consistency, sufficiency, efficiency & so on. So, a more general concept of estimation would be to find out an interval based on sample values which is expected to include the unknown parameter with a specified probability. This is known as the theory of interval estimator.

Confidence Interval

Let $x1, x2, \ldots, xn$ be a random sample from a population involve an unknown parameter θ. Our job is to find out two functions $t1$ & $t2$ of the sample values. Such that the probability of θ being included in the random interval $t1,t2$ has a given value say $1-\alpha$. So,

$$P(t_1 \leq \theta \leq t_2) = 1 - \alpha$$

Here the interval $[t1,t2]$ is called a $100 \times (\,) 1-\alpha$ % confidence interval for the parameter θ. The quantities $t1$ & $t2$ which serve as the lower & upper limits of the interval are known as confidence limits. $1-\alpha$ is called the confidence coefficient. The is a sort of measures of the trust or confidence that one may place in the interval for actually including θ.

Statistical Inference

Statistical inference is **the process of drawing conclusions about populations or scientific truths from data**. There are many modes of performing inference including statistical modeling, data oriented strategies and explicit use of designs and randomization in analyses.

There are two broad areas of statistical inference: statistical estimation and statistical hypothesis testing.

Test of Hypothesis, Decision Errors, One Level of Significance

Hypothesis testing is **an act in statistics whereby an analyst tests an assumption regarding a population parameter**. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

Decisions Errors refer to **the probability of making a wrong conclusion when doing hypothesis testing**. When a researcher sets out to do a study, she typically has a hypothesis, or a prediction of what she thinks the results will be.

The level of significance is defined as **the fixed probability of wrong elimination of null hypothesis when in fact, it is true**. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

The significance level is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

Regression Models

A regression model is **a statistical model that estimates the relationship between one dependent variable and one or more independent variables using a line** (or a plane in the case of two or more independent variables).

A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable. For example, the relationship between height and weight may be described by a **linear regression model**.

There are **two kinds of Linear Regression Model**:-

Simple Linear Regression: A linear regression model with one independent and one dependent variable. Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable.

Coefficient of Determination, R-square, Adjusted R-square

The coefficient of determination (R^2) is **a number between 0 and 1 that measures how well a statistical model predicts an outcome**. You can interpret the R^2 as the proportion of variation in the dependent variable that is predicted by the statistical model.

What Is R-Squared? R-squared (R2) is **a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model**.

Adjusted R squared is **calculated by dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R2 is always less than or equal to R2.

Forecasting & Time Series Analysis

An AI-based forecasting solution **uses an ensemble of machine learning algorithms to optimize forecasts**. The system then selects a model that's uniquely suited for the particular business metric that you're forecasting

Time series forecasting is one of the most important topics in data science. Almost every business needs to predict the future in order to make better decisions and allocate resources more effectively.

This repository provides examples and best practice guidelines for building forecasting solutions. The goal of this repository is to build a comprehensive set of tools and examples that leverage recent advances in forecasting algorithms to build solutions and operationalize them. Rather than creating implementations from scratch, we draw from existing state-of-the-art libraries and build additional utilities around processing and featurizing the data, optimizing and evaluating models, and scaling up to the cloud.

The examples and best practices are provided as Python Jupyter notebooks and R markdown files and a library of utility functions. We hope that these examples and utilities can significantly reduce the “time to market” by simplifying the experience from defining the business problem to the development of solutions by orders of magnitude. In addition, the example notebooks would serve as guidelines and showcase best practices and usage of the tools in a wide variety of languages.

What is time series forecasting?

Time series forecasting is the process of analyzing time series data using statistics and modeling to make predictions and inform strategic decision-making. It's not always an exact prediction, and likelihood of forecasts can vary wildly—especially when dealing with the commonly fluctuating variables in time series data as well as factors outside our control.

However, forecasting insight about which outcomes are more likely—or less likely—to occur than other potential outcomes. Often, the more comprehensive the data we have, the more accurate the forecasts can be. While forecasting and "prediction" generally mean the same thing, there is a notable distinction.

In some industries, forecasting might refer to data at a specific future point in time, while prediction refers to future data in general. Series forecasting is often used in conjunction with time series analysis. Time series analysis involves developing models to gain an understanding of the data to understand the underlying causes. Analysis can provide the "why" behind the outcomes you are seeing. Forecasting then takes the next step of what to do with that knowledge and the predictable extrapolations of what might happen in the future.

Def

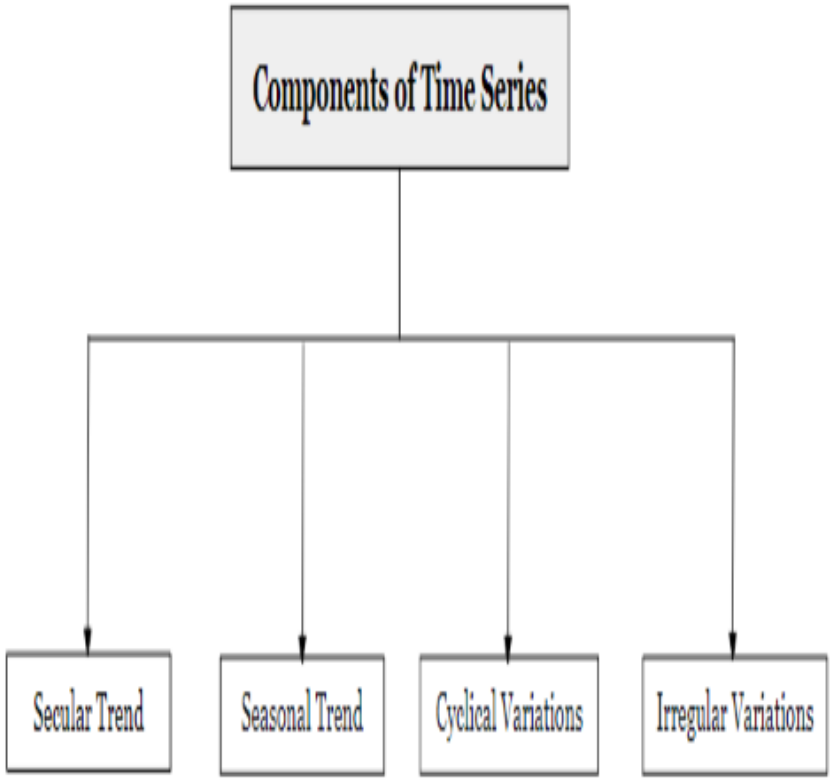
Time series components

An observed time series can be decomposed into three components: **the trend (long term direction), the seasonal (systematic, calendar related movements) and the irregular (unsystematic, short term fluctuations)**.

What are the 4 components of a time series?

These four components are:

- Secular trend, which describe the movement along the term;
- Seasonal variations, which represent seasonal changes;
- Cyclical fluctuations, which correspond to periodical but not seasonal variations;
- Irregular variations, which are other nonrandom sources of variations of series.



Various Forecasting Techniques

What Is Forecasting? Forecasting is **a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends**. Businesses utilize forecasting to determine how to allocate their budgets or plan for anticipated expenses for an upcoming period of time.

A forecast is based on past data, as opposed to a prediction, which is more subjective and based on instinct, gut feel, or guess. For example, the evening news gives the weather "forecast" not the weather "prediction." Regardless, the terms forecast and prediction are often used inter-changeably.

The Classification Problem

A classification problem is **when the output variable is a category**, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

In machine learning, classification refers to **a predictive modeling problem where a class label is predicted for a given example of input data**. Examples of classification problems include: Given an example, classify if it is spam or not. Given a handwritten character, classify it as one of the known characters.