Statistics and Probability:

Statistics and Probability are the building blocks of the most revolutionary technologies in today's world. **From Artificial Intelligence to Machine Learning and Computer Vision, Statistics and Probability form the basic foundation to all such technologies.** In this article on Statistics and Probability, I intend to help you understand the math behind the most complex algorithms and technologies.

The following topics are covered in this Statistics and Probability

What Is Data?

Look around you, there is data everywhere. Each click on your phone generates more data than you know. This generated data provides insights for analysis and helps us make better business decisions. This is why data is so important.

What Is Data – Statistics and Probability

Data refers to facts and statistics collected together for reference or analysis.

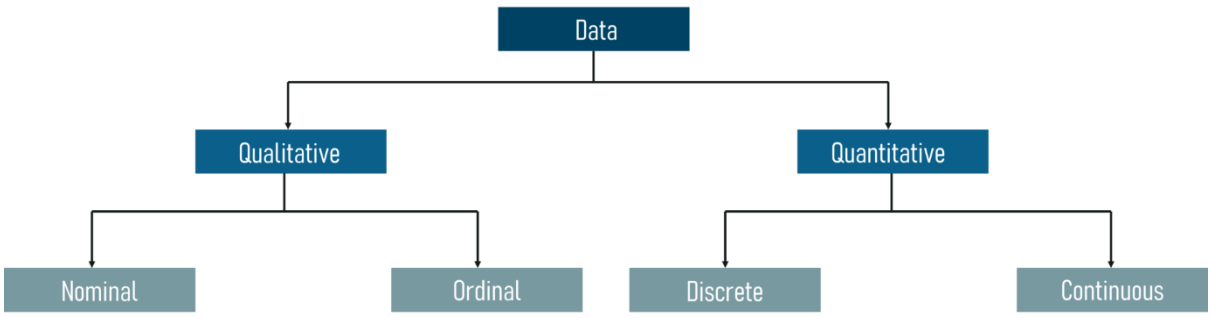
Data can be collected, measured and analyzed. It can also be visualized by using statistical models and graphs.

Categories Of Data

Data can be categorized into two sub-categories:

1. Qualitative Data
2. Quantitative Data

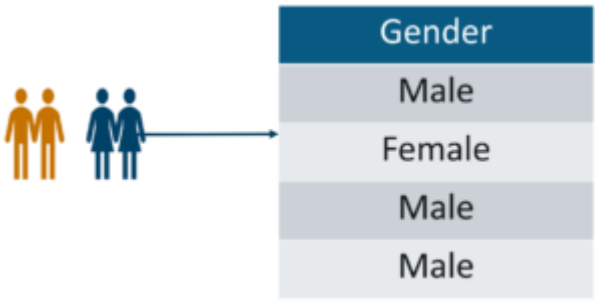
Refer the below figure to understand the different categories of data:



Categories Of Data – Statistics and Probability

Qualitative Data: *Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively. Qualitative data is further divided into two types of data:*

- *Nominal Data:* Data with no inherent order or ranking such as gender or race.



Gender
Male
Female
Male
Male

Nominal Data – Statistics and Probability

- *Ordinal Data:* Data with an ordered series of information is called ordinal data.

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

Ordinal Data – Statistics and Probability

Quantitative Data: *Quantitative data deals with numbers and things you can measure objectively. This is further divided into two:*

- *Discrete Data:* Also known as categorical data, it can hold a finite number of possible values.

Example: Number of students in a class.

- *Continuous Data:* Data that can hold an infinite number of possible values.

Example: Weight of a person.

So these were the different categories of data. The upcoming sections will focus on the basic Statistics concepts, so buckle up and get ready to do some math.

What Is Statistics?

Statistics is an area of applied mathematics concerned with data collection, analysis, interpretation, and presentation.

What Is Statistics – Statistics and Probability

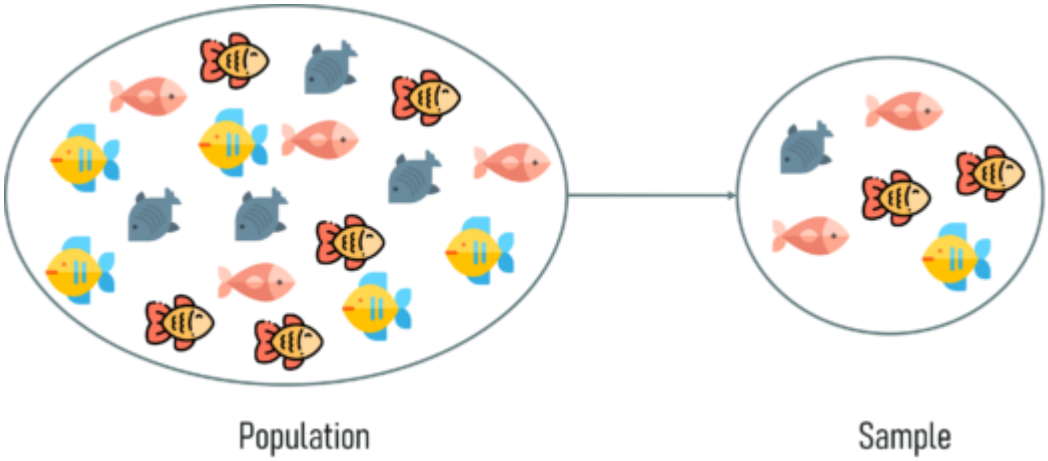
This area of mathematics deals with understanding how data can be used to solve complex problems. Here are a couple of example problems that can be solved by using statistics:

- Your company has created a new drug that may cure cancer. How would you conduct a test to confirm the drug's effectiveness?
- You and a friend are at a baseball game, and out of the blue, he offers you a bet that neither team will hit a home run in that game. Should you take the bet?
- The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?

These above-mentioned problems can be easily solved by using statistical techniques. In the upcoming sections, we will see how this can be done.

Basic Terminologies In Statistics

Before you dive deep into Statistics, it is important that you understand the basic terminologies used in Statistics. The two most important terminologies in statistics are population and sample.



Population and Sample – Statistics and Probability

Population: A collection or set of individuals or objects or events whose properties are to be analyzed

- **Sample:** A subset of the population is called 'Sample'. A well-chosen sample will contain most of the information about a particular population parameter.

Now you must be wondering how can one choose a sample that best represents the entire population.

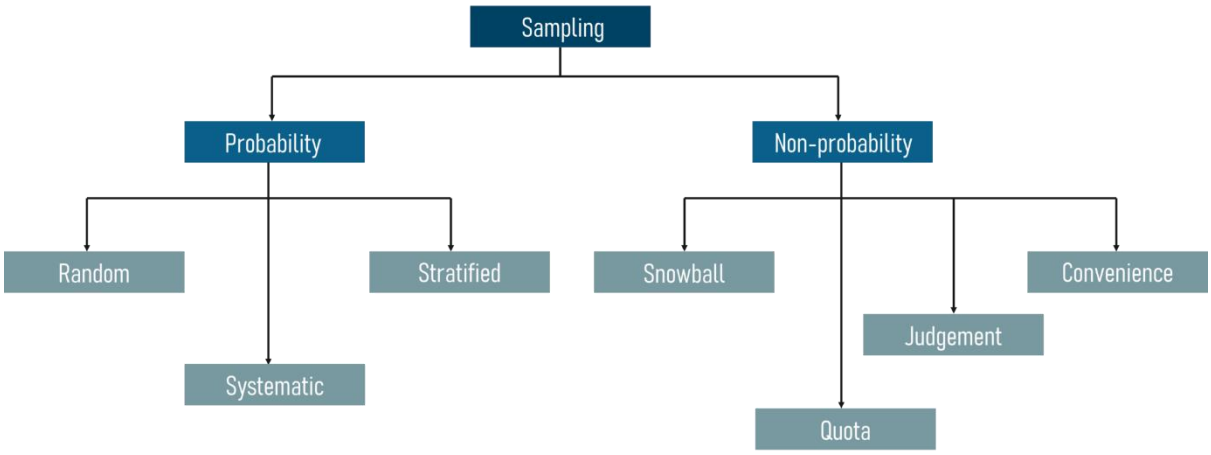
Sampling Techniques

Sampling is a statistical method that deals with the selection of individual observations within a population. It is performed to infer statistical knowledge about a population.

Consider a scenario wherein you're asked to perform a survey about the eating habits of teenagers in the US. There are over 42 million teens in the US at present and this number is growing as you read this blog. Is it possible to survey each of these 42 million individuals about their health? Obviously not! That's why sampling is used. It is a method wherein a sample of the population is studied in order to draw inference about the entire population.

There are two main types of Sampling techniques:

1. Probability Sampling
2. Non-Probability Sampling



Sampling Techniques – Statistics and Probability

In this blog, we'll be focusing only on probability sampling techniques because non-probability sampling is not within the scope of this blog.

Probability Sampling: This is a sampling technique in which samples from a large population are chosen using the theory of probability. There are three types of probability sampling:

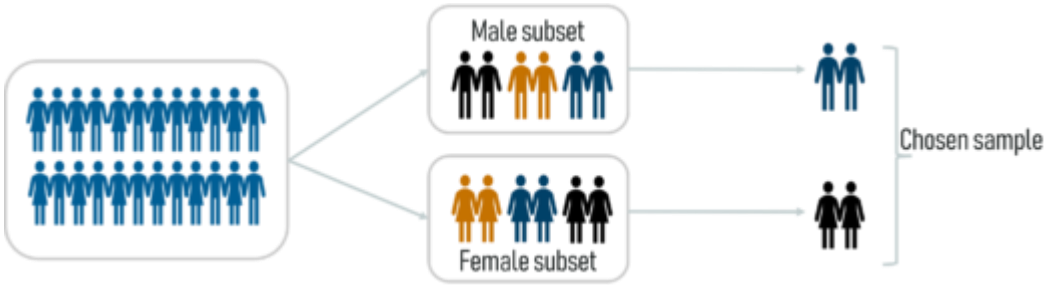
- **Random Sampling:** In this method, each member of the population has an equal chance of being selected in the sample.

Random Sampling – Statistics and Probability –

- **Systematic Sampling:** In Systematic sampling, every nth record is chosen from the population to be a part of the sample. Refer the below figure to better understand how Systematic sampling works.

Systematic Sampling – Statistics and Probability –

- **Stratified Sampling:** In Stratified sampling, a stratum is used to form samples from a large population. A stratum is a subset of the population that shares at least one common characteristic. After this, the random sampling method is used to select a sufficient number of subjects from each stratum.



Stratified Sampling – Statistics and Probability

Now that you know the basics of Statistics, let's move ahead and discuss the different types of statistics.

Types Of Statistics

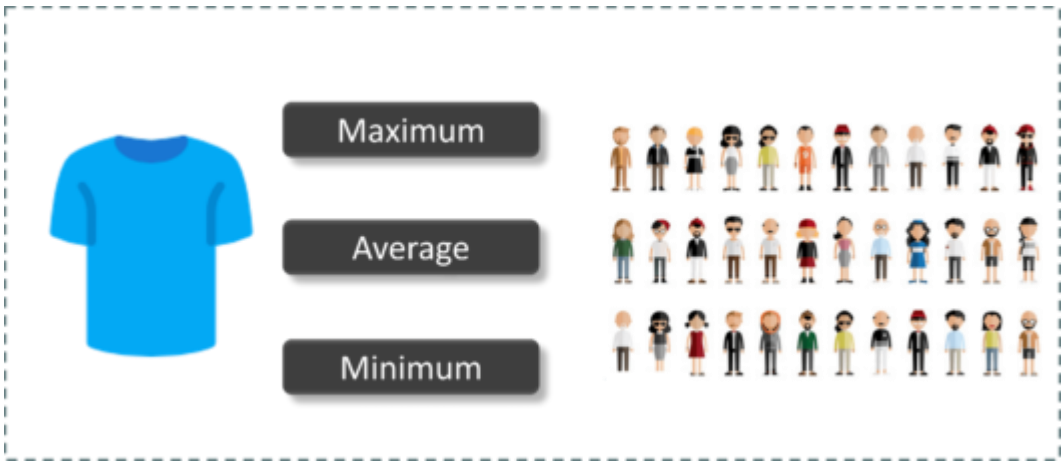
There are two well-defined types of statistics:

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive Statistics is mainly focused upon the main characteristics of data. It provides a graphical summary of the data.



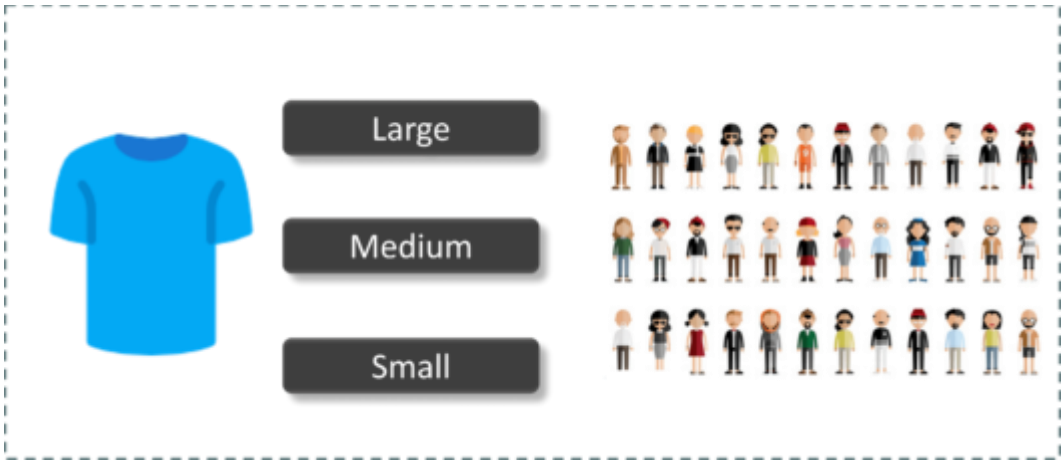
Descriptive Statistics – Statistics and Probability

Suppose you want to gift all your classmate's t-shirts. To study the average shirt size of students in a classroom, in descriptive statistics you would record the shirt size of all students in the class and then you would find out the maximum, minimum and average shirt size of the class.

Inferential Statistics

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

Inferential statistics generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using sample data.



Inferential Statistics – Statistics and Probability

So, if we consider the same example of finding the average shirt size of students in a class, in Inferential Statistics, you will take a sample set of the class, which is basically a few people from the entire class. You already have had grouped the

class into large, medium and small. In this method, you basically build a statistical model and expand it for the entire population in the class.

So that was a brief understanding of Descriptive and Inferential Statistics. In the further sections, you'll see how Descriptive and Inferential statistics works in depth.

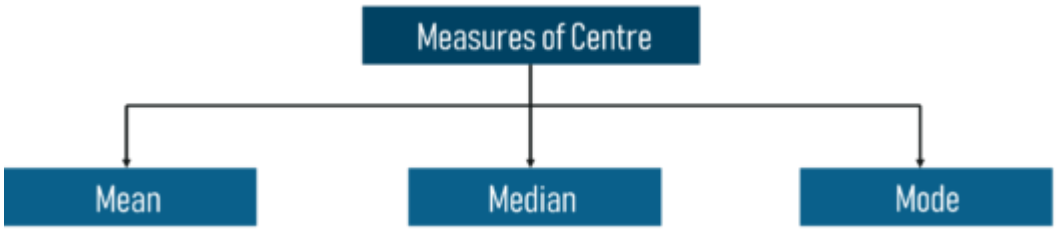
Understanding Descriptive Statistics

Descriptive Statistics is broken down into two categories:

1. Measures of Central Tendency
2. Measures of Variability (spread)

Measures Of Centre

Measures of the center are statistical measures that represent the summary of a dataset. There are three main measures of center:



Measures Of Centre – Statistics and Probability –

Mean: Measure of the average of all the values in a sample is called Mean.

1. **Median:** Measure of the central value of the sample set is called Median.
2. **Mode:** The value most recurrent in the sample set is known as Mode.

To better understand the Measures of central tendency let's look at an example. The below cars dataset contains the following variables:

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

DataSet – Statistics and Probability

- Cars
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Using descriptive Analysis, you can analyze each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

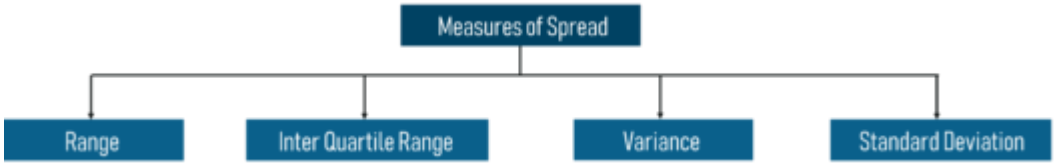
If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values. In this case, we'll take the sum of the Horse Power of each car, divided by the total number of cars:
Mean = (110+110+93+96+90+110+110+110)/8 = 103.625

If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.
The mpg for 8 cars: 21,21,21.3,22.8,23,23,23,23
Median = (22.8+23)/2 = 22.9

If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated the most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set, you can see that the most recurring value is 6. Hence 6 is our Mode.

Measures Of The Spread

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.



Measures Of Spread – Statistics and Probability

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

- Range: It is the given measure of how spread apart the values in a data set are. The range can be calculated as:

$$Range = Max(x\_i) - Min(x\_i)$$

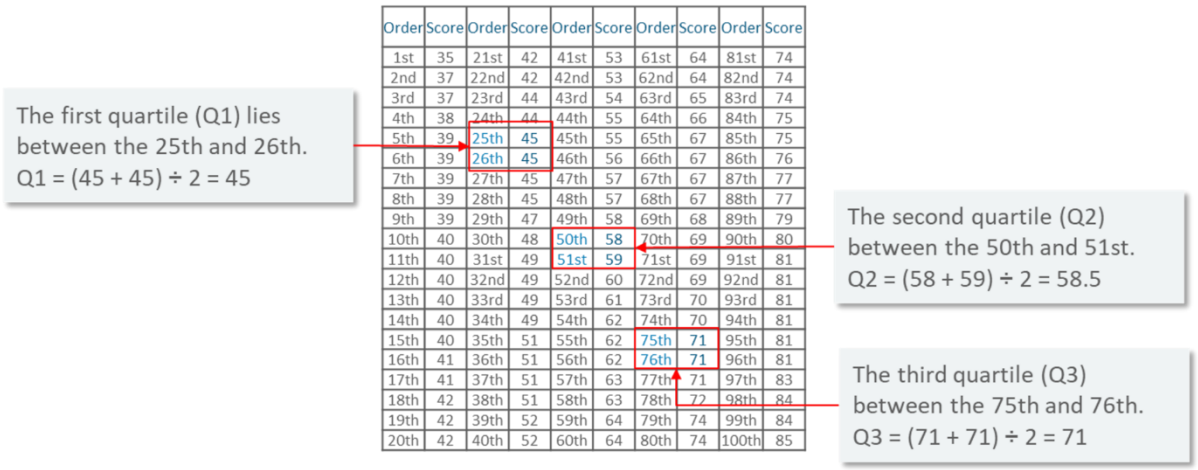
Here,

Max(x_i): Maximum value of x

Min(x_i): Minimum value of x

- Quartile: Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

To better understand how quartile and the IQR are calculated, let's look at an example.



Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

Measures Of Spread example – Statistics and Probability –

The above image shows marks of 100 students ordered from lowest to highest scores. The quartiles lie in the following ranges:

1. The first quartile (Q1) lies between the 25th and 26th observation.
2. The second quartile (Q2) lies between the 50th and 51st observation.
3. The third quartile (Q3) lies between the 75th and 76th observation.

- Inter Quartile Range (IQR): It is the measure of variability, based on dividing a data set into quartiles. The interquartile range is equal to Q3 minus Q1, i.e. IQR = Q3 – Q1

- Variance: It describes how much a random variable differs from its expected value. It entails computing squares of deviations. Variance can be calculated by using the below formula:

$$s^2 = \frac{\sum_{n=1}^{n} (x_i - \bar{x})^2}{n}$$

Measures Of Spread Variance – Statistics and Probability

Here,

x: Individual data points
n: Total number of data points
$\bar{x}$: Mean of data points

- Deviation is the difference between each element from the mean. It can be calculated by using the below formula:

Deviation = (x_i – μ)

- Population Variance is the average of squared deviations. It can be calculated by using the below formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} = (x_i - \mu)^2$$

Measures Of Spread Population Variance – Statistics and Probability –

- Sample Variance is the average of squared differences from the mean. It can be calculated by using the below formula:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{N} = (x_i - \bar{x})^2$$

Measures Of Spread Sample Variance – Statistics and Probability

Standard Deviation: It is the measure of the dispersion of a set of data from its mean. It can be calculated by using the below formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Measures Of Spread Standard Deviation – Statistics and Probability

To better understand how the Measures of spread are calculated, let's look at a use case.

Problem statement: *Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.*

Let's look at the solution step by step:

Step 1: Find out the mean for your sample set.

The mean is = 9+2+5+4+12+7+8+11+9+3

Then work out the mean of those squared differences.

+7+4+12+5+4+10+9+6+9+4 / 20
μ=7

Step 2: Then for each number, subtract the Mean and square the result.

$(x\_i - \mu)^2$

$(9-7)^2 = 2^2 = 4$
$(2-7)^2 = (-5)^2 = 25$
$(5-7)^2 = (-2)^2 = 4$
And so on...

We get the following results:
4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Step 3: Then work out the mean of those squared differences.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9 / 20
$\therefore \sigma^2 = 8.9$

Step 4: Take the square root of σ^2.

$\sigma = 2.983$

To better understand the measures of spread and center, let's execute a short demo by using the R language.

Descriptive Statistics In R

R is a statistical programming language, that is mainly used for Data Science, Machine Learning and so on. If you wish to learn more about R, give this *R Tutorial – A Beginner's Guide to Learn R Programming* blog a read.

Now let's move ahead and implement Descriptive Statistics in R.

In this demo, we'll see how to calculate the Mean, Median, Mode, Variance, Standard Deviation and how to study the variables by plotting a histogram. This is quite a simple demo but it also forms the foundation that every Machine Learning algorithm is built upon.

Step 1: Import data for computation

```
set.seed(1)

#Generate random numbers and store it in a variable called data
>data = runif(20,1,10)
```

Step 2: Calculate Mean for the data

```
#Calculate Mean
>mean = mean(data)
>print(mean)

[1] 5.996504
```

Step 3: Calculate the Median for the data

```
#Calculate Median
>median = median(data)
>print(median)

[1] 6.408853
```

Entropy

Entropy measures the impurity or uncertainty present in the data. *It can be measured by using the below formula:*

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Entropy – Statistics and Probability –

where:
S – set of all instances in the dataset
N – number of distinct class values
pi – event probability

Information Gain
Information Gain (IG) indicates how much "information" a particular feature/ variable gives us about the final outcome. It can be measured by using the below formula:

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A,S)$$

Information Gain – Statistics and Probability –

Here:

- *H(S) – entropy of the whole dataset S*
- *|Sj| – number of instances with j value of an attribute A*
- *|S| – total number of instances in dataset S*
- *v – set of distinct values of an attribute A*
- *H(Sj) – entropy of subset of instances for attribute A*
- *H(A, S) – entropy of an attribute A*

Information Gain and Entropy are important statistical measures that let us understand the significance of a predictive model. To get a more clear understanding of Entropy and IG, let's look at a use case.

Problem Statement: To predict whether a match can be played or not by studying the weather conditions.

Data Set Description: The following data set contains observations about the weather conditions over a period of time.

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Use Case Dataset – Statistics and Probability

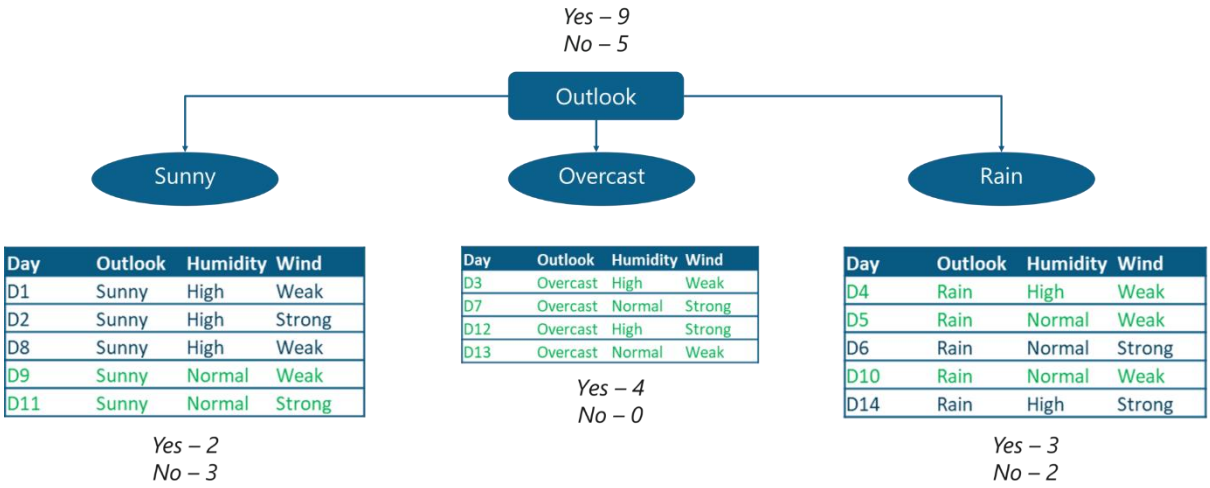
The predictor variables include:

1. Day
2. Outlook
3. Sunny
4. Wind

The target variable is the 'Play' variable which can be predicted by using the set of predictor variables. The value of this variable will decide whether or not a game can be played on a particular day.

To solve such a problem, we can make use of Decision Trees. Decision Trees are basically inverted trees that help us get to the outcome by making decisions at each branch node.

The below figure shows that out of 14 observations, 9 observations result in a 'yes', meaning that out of 14 days, the match can be played on 9 days. And if you notice, the decision was made by choosing the 'Outlook' variable as the root node (the topmost node in a Decision Tree).



Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

Use Case – Statistics and Probability

The outlook variable has 3 values,

1. Sunny
2. Overcast
3. Rain

These 3 values are assigned to the immediate branch nodes and for each of these values, the possibility of 'play= yes' is calculated. The 'sunny' and 'rain' branches give out an impure output, meaning that there is a mix of 'yes' and 'no'. But if you notice the 'overcast' variable, it results in a 100% pure subset. This shows that the 'overcast' variable will result in a definite and *certain* output.

This is exactly what entropy is used to measure. It calculates the impurity or the uncertainty and the lesser the uncertainty or the entropy of a variable, more significant is that variable.

In a Decision Tree, the root node is assigned the best attribute so that the Decision Tree can predict the most precise outcome. The 'best attribute' is basically a predictor variable that can best split the data set.

Now the next question in your head must be, "How do I decide which variable or attribute best splits the data?"

Well, this can be done by using Information Gain and Entropy.

We begin by calculating the entropy when the 'outlook' variable is assigned to the root node. From the total of 14 instances we have:

- 9 instances "yes"
- 5 instances "no"

The Entropy is:

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

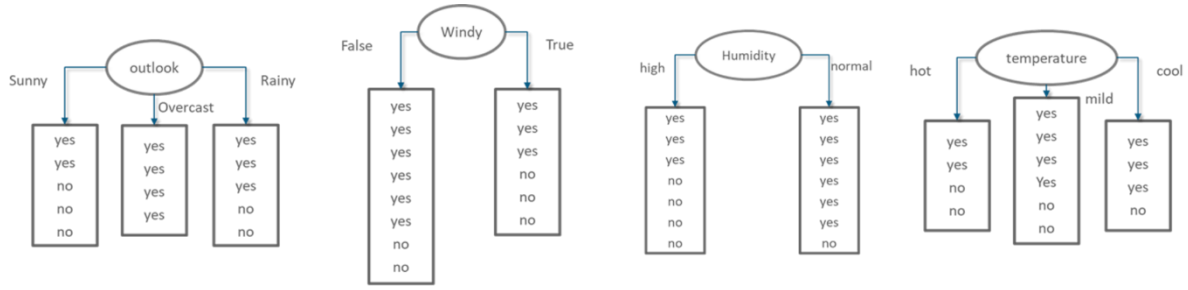
$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Calculating Entropy – Statistics and Probability

Thus, we get an entropy of 0.940, which denotes impurity or uncertainty.

Now in order to ensure that we choose the best variable for the root node, let us look at all the possible combinations.

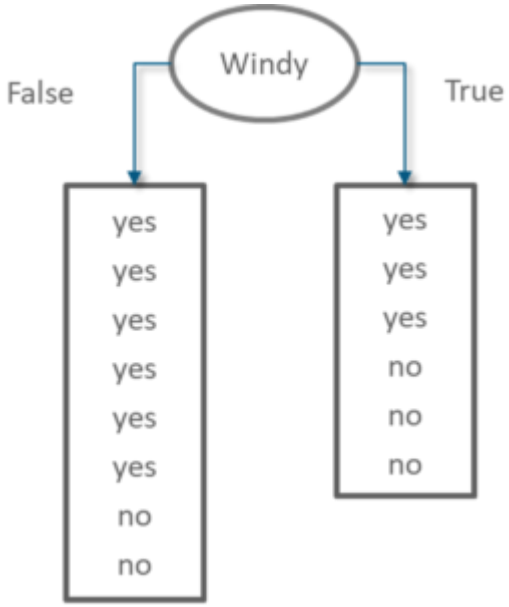
The below image shows each decision variable and the output that you can get by using that variable at the root node.



Possible Decision Trees- Statistics and Probability

Our next step is to calculate the Information Gain for each of these decision variables (outlook, windy, humidity, temperature). *A point to remember is, the variable that results in the highest IG must be chosen since it will give us the most precise output and information.*

Information Gain of attribute "windy"



Decision Tree Windy – Statistics and Probability

From the total of 14 instances we have:

- 6 instances “true”
- 8 instances “false”

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$Gain(A_{Windy}, S) = 0.940 - \frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048$$

Information Gain windy – Statistics and Probability

Information Gain of attribute “outlook”

Decision Tree Outlook – Statistics and Probability

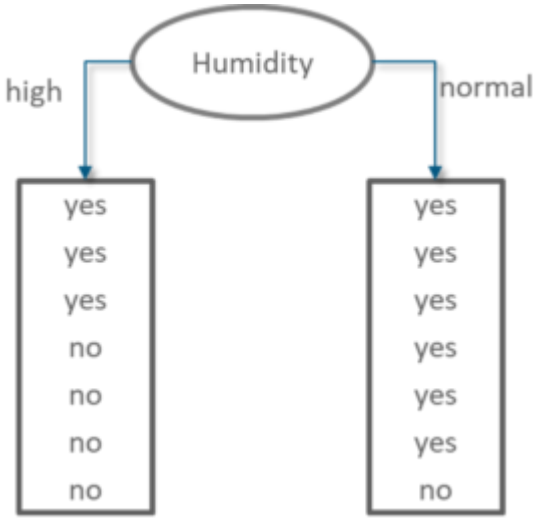
From the total of 14 instances we have:

- 5 instances “sunny”
- 4 instances “overcast”
- 5 instances “rainy”

$$Gain(A_{Outlook}, S) = 0.940 - \frac{5}{14} \cdot \left(- \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \right) + \frac{4}{14} \cdot \left(- \left(\frac{4}{4} \log_2 \frac{4}{4} \right) \right) + \frac{5}{14} \cdot \left(- \left(\frac{3}{5} \cdot \log_2 \frac{3}{5} + \frac{2}{5} \cdot \log_2 \frac{2}{5} \right) \right) = 0.247$$

Information Gain outlook – Statistics and Probability

Information Gain of attribute “humidity”



Decision Tree Humidity – Statistics and Probability

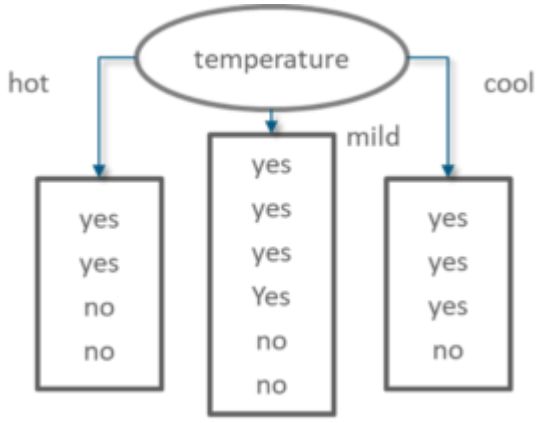
From the total of 14 instances we have:

- 7 instances “high”
- 7 instances “normal”

$$Gain(A_{Humidity}, S) = 0.940 - \frac{7}{14} \cdot \left(- \left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7} \right) \right) + \frac{7}{14} \cdot \left(- \left(\frac{6}{7} \cdot \log_2 \frac{6}{7} + \frac{1}{7} \cdot \log_2 \frac{1}{7} \right) \right) = 0.151$$

Information Gain humidity – Statistics and Probability

Information Gain of attribute “temperature”



Decision Tree Temperature – Statistics and Probability

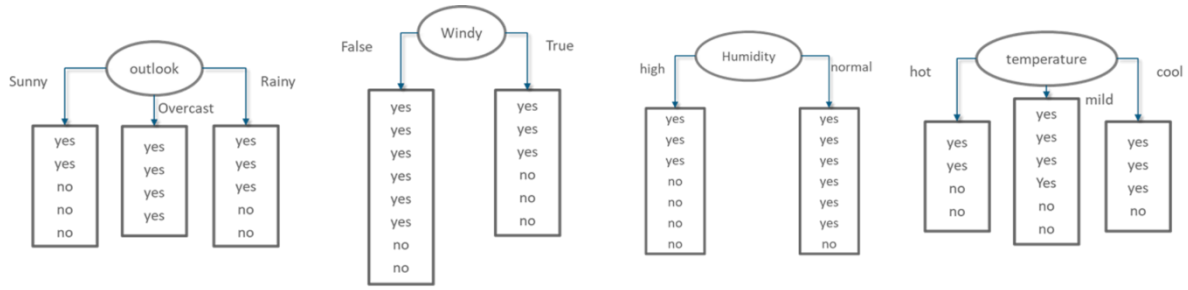
From the total of 14 instances we have:

- 4 instances “hot”
- 6 instances “mild”
- 4 instances “cool”

$$Gain(A_{Temperature}, S) = 0.940 - \frac{4}{14} \cdot \left(-\left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6} \right) \right) + \frac{4}{14} \cdot \left(-\left(\frac{3}{4} \cdot \log_2 \frac{3}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4} \right) \right) = 0.029$$

Information Gain temperature – Statistics and Probability

The below figure shows the IG for each attribute. The variable with the highest IG is used to split the data at the root node. The ‘Outlook’ variable has the highest IG, therefore it is assigned to the root node.



Gain = 0.247 Gain = 0.048 Gain = 0.151 Gain = 0.029

Information Gain Summary – Statistics and Probability –

So that was all about Entropy and Information Gain. Now let’s take a look at another important statistical method called Confusion Matrix.

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Basically, a Confusion Matrix will help you evaluate the performance of a predictive model. It is mainly used in classification problems.

Confusion Matrix represents a tabular representation of Actual vs Predicted values. You can calculate the accuracy of a model by using the following formula:

$$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}}$$

Confusion Matrix Formula – Statistics and Probability

To understand what is True Negative, True Positive and so on, let s consider an example.

Let's consider that you're given data about 165 patients, out of which 105 patients have a disease and the remaining 5o patients don't. So you build a classifier that predicts by using these 165 observations. Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.

Therefore, in order to evaluate the efficiency of the classifier, a Confusion Matrix is used:

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Confusion Matrix – Statistics and Probability

In the above figure,

- 'n' denoted the total number of observations
- Actual denotes the actual values in the data set
- Predicted denotes the values predicted by the classifier

The confusion matrix studies the performance of the classifier by comparing the actual values to the predicted ones. Below are some terms related to the confusion matrix:

1. **True Positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
2. **True Negatives (TN):** We predicted no, and they don't have the disease.
3. **False Positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
4. **False Negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

So those were the important concepts used in Descriptive Statistics. Now let's study all about Probability.

Probability

Before we understand what probability is, let me clear out a very common misconception. People often tend to ask this question:

What is the relation between Statistics and Probability?

Probability and Statistics and related fields. Probability is a mathematical method used for statistical analysis. Therefore we can say that probability and statistics are interconnected branches of mathematics that deal with analyzing the relative frequency of events.

Now let's understand what probability is.

What Is Probability?

Probability is the measure of how likely an event will occur. To be more precise probability is the ratio of desired outcomes to total outcomes: *(desired outcomes) / (total outcomes)*

The probabilities of all outcomes always sums up to 1. Consider the famous rolling dice example:

- On rolling a dice, you get 6 possible outcomes
- Each possibility only has one outcome, so each has a probability of 1/6
- For example, the probability of getting a number '2' on the dice is 1/6

Now let's try to understand the common terminologies used in probability.

Terminologies In Probability

Before you dive deep into the concepts of probability, it is important that you understand the basic terminologies used in probability:

- Random Experiment: An experiment or a process for which the outcome cannot be predicted with certainty.
- Sample space: The entire possible set of outcomes of a random experiment is the sample space of that experiment.

- Event: One or more outcomes of an experiment is called an event. It is a subset of sample space. There are two types of events in probability:

    - **Disjoint Event:** *Disjoint Events do not have any common outcomes.* For example, a single card drawn from a deck cannot be a king and a queen

    - **Non – Disjoint Event:** *Non-Disjoint Events can have common outcomes.* For example, a student can get 100 marks in statistics and 100 marks in probability

Probability Distribution

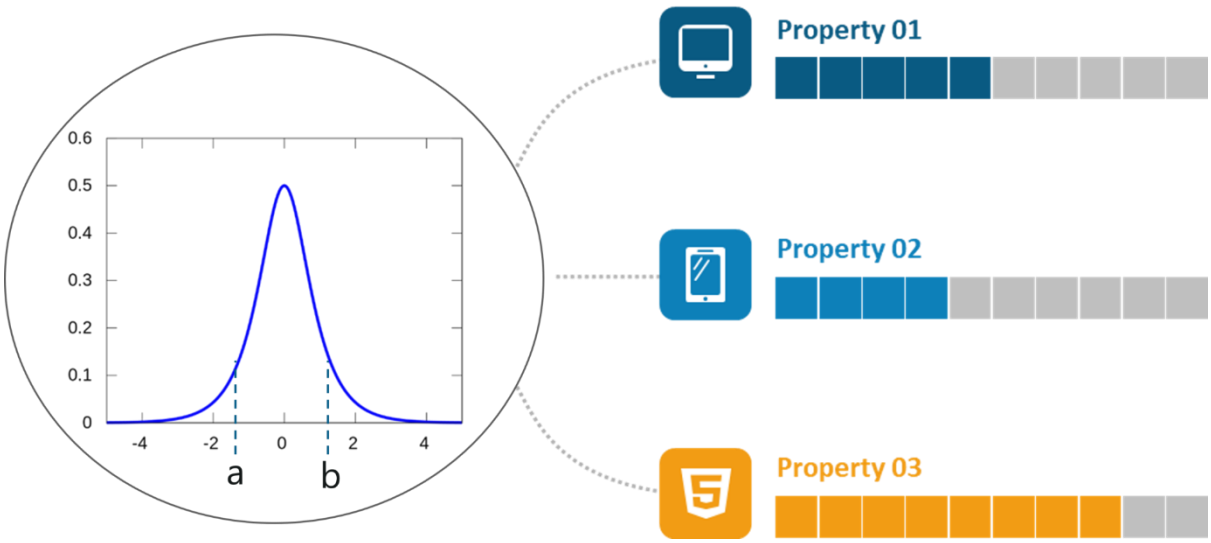
In this blog we shall focus on three main probability distribution functions:

1. Probability Density Function
2. Normal Distribution
3. Central Limit Theorem

Probability Density Function

The Probability Density Function (PDF) is concerned with the relative likelihood for a *continuous random variable* to take on a given value. The PDF gives the probability of a variable that lies between the range 'a' and 'b'.

The below graph denotes the PDF of a continuous variable over a range. This graph is famously known as the bell curve:



Probability Density Function – Statistics and Probability

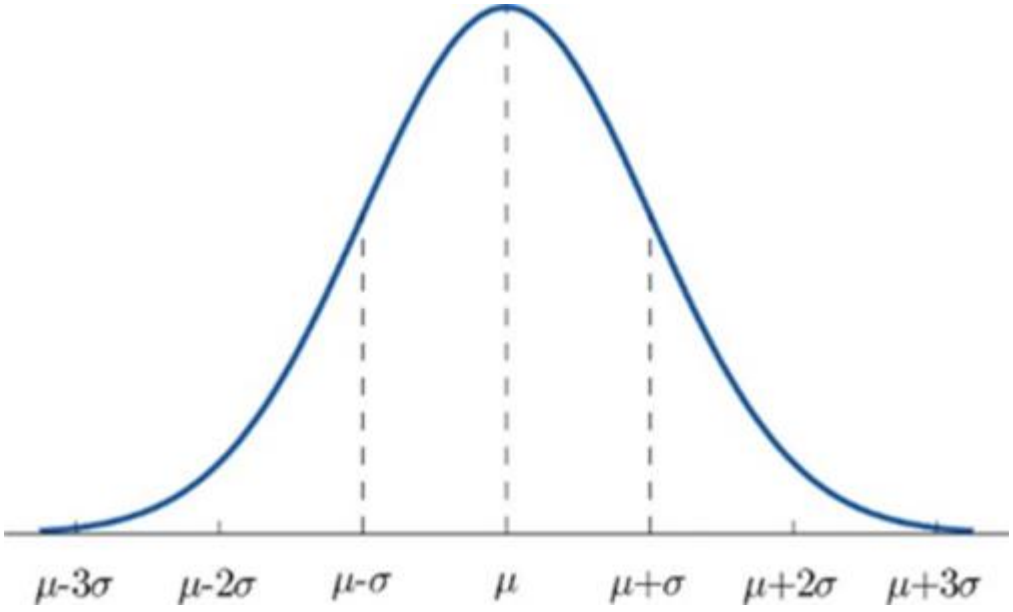
The following are the properties of a PDF:

- Graph of a PDF will be continuous over a range
- The area bounded by the curve of the density function and the x-axis is equal to 1
- The probability that a random variable assumes a value between a and b is equal to the area under the PDF bounded by a and b

Normal Distribution

Normal distribution, otherwise known as the Gaussian distribution, is a probability distribution that denotes the symmetric property of the mean. The idea behind this function is that the data near the mean occurs more frequently than the data away from the mean. It infers that the data around the mean represents the entire data set.

Similar to PDF, the normal distribution appears as a bell curve:

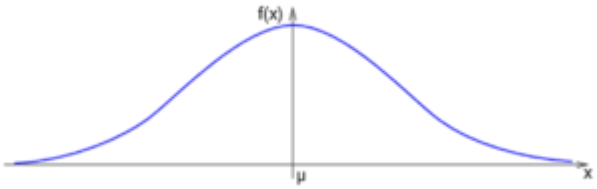


Normal Distribution – Statistics and Probability –

The graph of the Normal Distribution depends on two factors: the Mean and the Standard Deviation

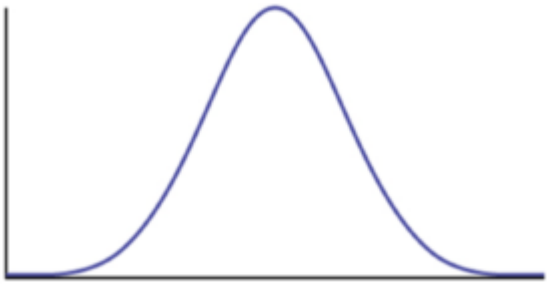
- Mean: Determines the location of the center of the graph
- Standard Deviation: Determines the height of the graph

If the standard deviation is large, the curve is short and wide:



Standard Deviation Curve – Statistics and Probability –

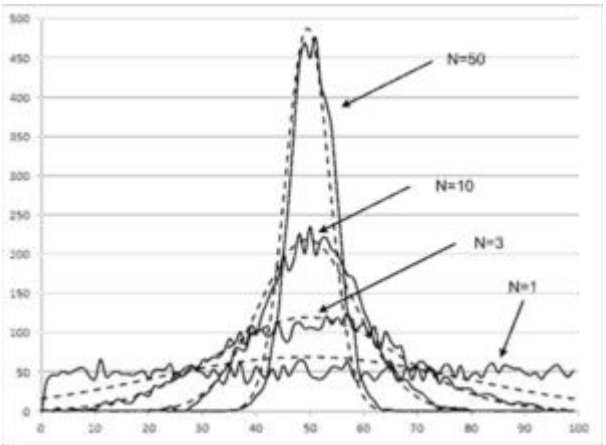
If the standard deviation is small, the curve is tall and narrow:



Standard Deviation Curve – Statistics and Probability –

Central Limit Theorem
The Central Limit Theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal if the sample size is large enough.

In simple terms, if we had a large population divided into samples, then the mean of all the samples from the population will be almost equal to the mean of the entire population. The below graph depicts a more clear understanding of the Central Limit Theorem:



Central Limit Theorem – Statistics and Probability –

The accuracy or resemblance to the normal distribution depends on two main factors:

1. Number of sample points taken
2. The shape of the underlying population

Now let's focus on the three main types of probability.

Types Of Probability

Marginal Probability
The probability of an event occurring (p(A)), unconditioned on any other events. For example, the probability that a card drawn is a 3 (p(three)=1/13).

It can be expressed as:

$$P(A) = \sum_{i=1}^{k} P(x_i)$$

Marginal Probability – Statistics and Probability –

Joint Probability
Joint Probability is a measure of two events happening at the same time, i.e., p(A and B), The probability of event A and event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written p(A ∩ B).

For example, the probability that a card is a four and red =p(four and red) = 2/52=1/26.

Conditional Probability
Probability of an event or outcome based on the occurrence of a previous event or outcome
Conditional Probability of an event B is the probability that the event will occur given that an event A has already occurred.

- p(B|A) is the probability of event B occurring, given that event A occurs.
- If A and B are dependent events then the expression for conditional probability is given by:
  P (B|A) = P (A and B) / P (A)
- If A and B are independent events then the expression for conditional probability is given by:
  P(B|A) = P (B)

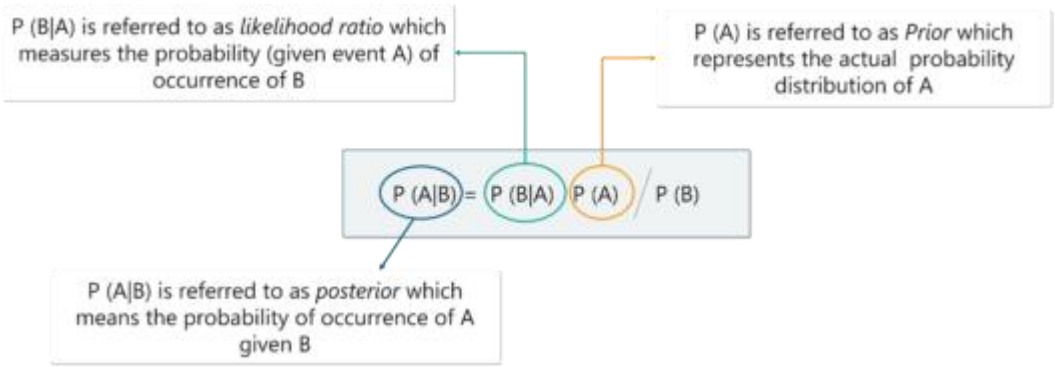
Example: Given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

Now let's look at the last topic under probability.

Bayes' Theorem

The Bayes theorem is used to calculate the conditional probability, which is nothing but the probability of an event occurring based on prior knowledge of conditions that might be related to the event.

Mathematically, the Bayes theorem is represented as:



Bayes Theorem – Statistics and Probability –

In the above equation:

P(A|B): Conditional probability of event A occurring, given the event B
P(A): Probability of event A occurring
P(B): Probability of event B occurring
P(B|A): Conditional probability of event B occurring, given the event A
Formally, the terminologies of the Bayesian Theorem are as follows:

A is known as the proposition and B is the evidence
P(A) represents the prior probability of the proposition
P(B) represents the prior probability of evidence
P(A|B) is called the posterior
P(B|A) is the likelihood
Therefore, the Bayes theorem can be summed up as:

Posterior=(Likelihood).(Proposition prior probability)/Evidence prior probability

To better understand this, let's look at an example:

Problem Statement: *Consider 3 bags. Bag A contains 2 white balls and 4 red balls; Bag B contains 8 white balls and 4 red balls, Bag C contains 1 white ball and 3 red balls. We draw 1 ball from each bag. What is the probability to draw a white ball from Bag A if we know that we drew exactly a total of 2 white balls total?*

Soln:

- Let A be the event of picking a white ball from bag A, and let X be the event of picking exactly two white balls
- We want Probability(A|X), i.e. probability of occurrence of event A given X
- By the definition of Conditional Probability,

$$\Pr(A|X) = \frac{Pr(A \cap X)}{\Pr(X)}$$

- We need to find the two probabilities on the right side of the equal to symbol.

We can solve this problem in two steps:

Step 1: First find Pr(X). This can happen in three ways:

i. white from A, white from B, red from C
ii. white from A, red from B, white from C
iii. red from A, white from B, white from C

Step 2: Find Pr(A∩X).

- This is the sum of terms (i) and (ii) above

I just drew out a blueprint to solve this problem. Consider this as homework and let us know your answer in the comment section.

The following section will cover the concepts under Inferential statistics, also known as Statistical Inference. So far we discussed Descriptive Statistics and Probability, now let's look at a few more advanced topics.

Statistical Inference

As discussed earlier Statistical Inference is a branch of statistics that deals with forming inferences and predictions about a population based on a sample of data taken from the population in question.

The question you should ask now, is that how does one form inferences or predictions on a sample? The answer is through Point estimation.

What is Point Estimation?

Point Estimation is concerned with the use of the sample data to measure a single value which serves as an approximate value or the best estimate of an unknown population parameter.

Two important terminologies on Point Estimation are:

- **Estimator:** A function f(x) of the sample, that is used to find out the estimate.
- **Estimate:** The Realised value of an estimator.

For example, in order to calculate the mean of a huge population, we first draw out a sample of the population and find the sample mean. The sample mean is then used to estimate the population mean. This is basically point estimation.

Finding The Estimates

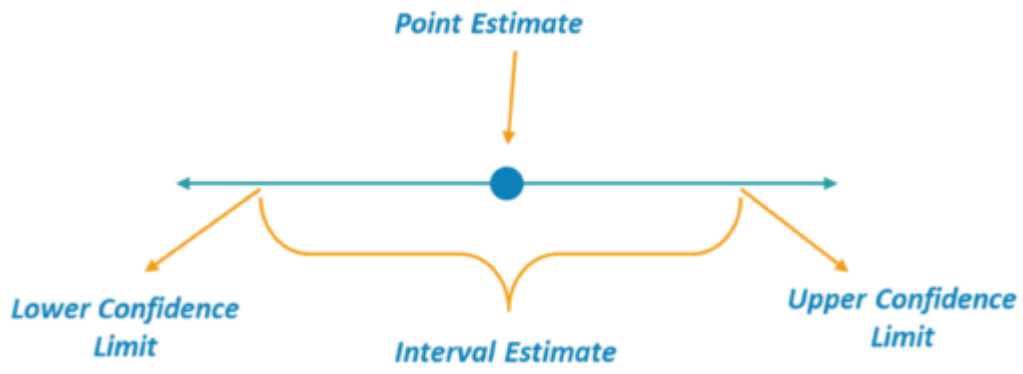
There are 4 common statistical techniques that are used to find the estimated value concerned with a population:

1. Method of Moments: It is a method used to estimate population parameters, like the population mean or the population variance. In simple terms this involves, taking down known facts about the population, and extending those ideas to a sample.
2. Maximum of Likelihood: This method uses a model and the values in the model to maximize a likelihood function. This results in the most likely parameter for the inputs selected.
3. Bayes' Estimators: This method works by minimizing the average risk (an expectation of random variables)
4. Best Unbiased Estimators: In this method, several unbiased estimators can be used to approximate a parameter (which one is "best" depends on what parameter you are trying to find)

Apart from these four estimation methods, there is yet another estimation method known as Interval Estimation (Confidence Interval).

What Is Interval Estimation?

An Interval, or range of values, used to estimate a population parameter is known as Interval Estimation. The below image clearly shows what an Interval Estimation is as opposed to point estimation. The estimated value must occur between the lower confidence limit and the upper confidence limit.



Interval Estimate – Statistics and Probability –

For example, if I stated that I will take 30 minutes to reach the theater, this is Point estimation. However, if I stated that I will take between 45 minutes to an hour to reach the theater, this is an example of Interval Estimation.

Interval Estimation gives rise to two important Statistical terminologies: Confidence Interval and Margin of Error.

What Is Confidence Interval?

- Confidence Interval is the measure of your confidence, that the interval estimate contains the population mean, μ.
- Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter.
- Technically, a range of values so constructed that there is a specified probability of including the true value of a parameter within it.

For example, you survey a group of cat owners to see how many cans of cat food they purchase a year. You test your statistics at the 99 percent confidence level and get a confidence interval of (100,200). This means that you think they buy between 100 and 200 cans a year. And also since the Confidence Level is 99%, it shows that you're very confident that the results are correct.

What is Margin Of Error?

- The difference between the point estimate and the actual population parameter value is called the Sampling Error.
- When μ is estimated, the sampling error is the difference $\mu - \bar{x}$. Since μ is usually unknown, the maximum value of the error can be calculated by using the Level of Confidence.

- *The margin of Error E, for a given level of confidence, is the greatest possible distance between the point estimate and the value of the parameter it is estimating.*

The Margin Of Error E can be calculated by using the below formula:

$$E = Z_c \frac{\sigma}{\sqrt{n}}$$

Margin Of Error – Statistics and Probability –

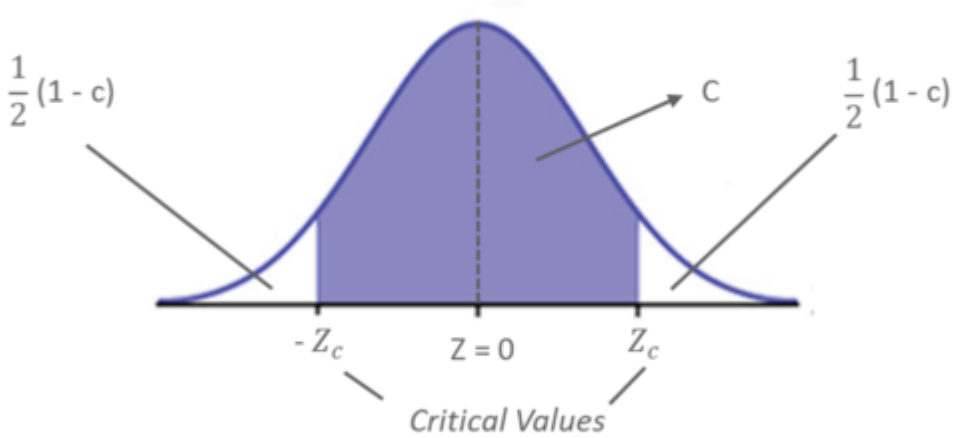
Here,

- Z_c denotes the critical value or the confidence interval
- σ denotes the Standard Deviation
- n denotes sample size

Now let's understand how to estimate the Confidence Intervals.

Estimating Level Of Confidence

The level of confidence 'c', is the probability that the interval estimate contains the population parameter. Consider the below figure:

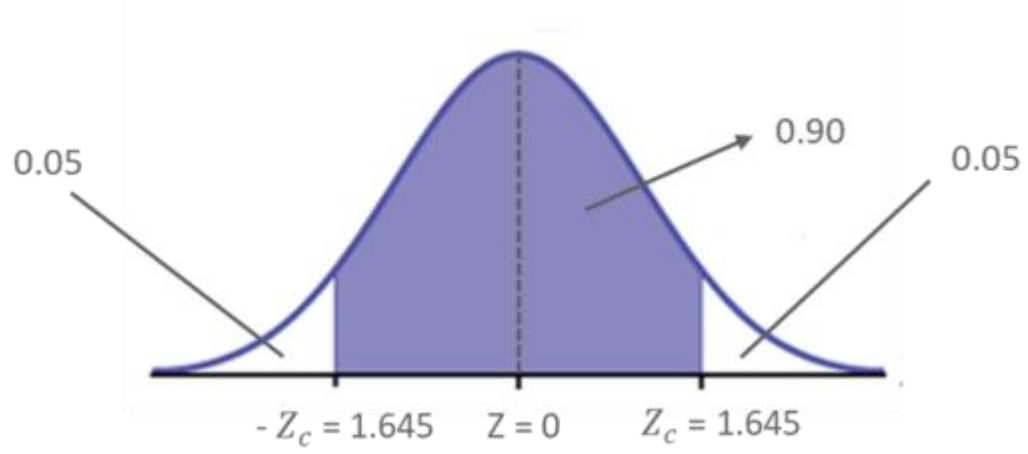


Estimating Level Of Confidence – Statistics and Probability –

- C is the area beneath the normal curve between the critical values
- Corresponding Z score can be calculated using the standard normal table

For example, if the level of confidence is 90%, this means that you are 90% confident that the interval contains the population mean, μ. The remaining 10% is

equally distributed (0.05) on either side of 'c' (the area that contains the estimated population parameter)



Estimating Level Of Confidence Example – Statistics and Probability –

The Corresponding Z – scores are ± 1.645 as per the Z table.

Construction Of Confidence Interval

The Confidence Interval can be constructed by following the below steps:

1. **Identify a Sample Statistic:** Choose the statistic that you will use to estimate a population parameter (ex: mean of the sample)
2. **Select a Confidence Level:** The confidence level describes the uncertainty of a sampling method.
3. **Find the Margin of Error:** Find the margin of error based on the previous equation explained
4. **Specify the Confidence Interval:** The Confidence Interval can be found out by: *Confidence Interval = Sample Statistic ± Margin of Error*

Now let's look at a problem statement to better understand these concepts.

Problem Statement: A random sample of 32 textbook prices is taken from a local college bookstore. The mean of the sample is $\bar{x}$ = 74.22, and the sample standard deviation is S = 23.44. Use a 95% confidence level and find the margin of error for the mean price of all textbooks in the bookstore

You know by formula, E = Z_c * ($\sigma/\sqrt{n}$)
*E = 1.96 * (23.44/√32) ≈ 8.12*

Therefore, we are 95% confident that the margin of error for the population mean (all the textbooks in the bookstore) is about 8.12.

Now that you know the idea behind Confidence Intervals, let's move ahead to the next topic, Hypothesis Testing.

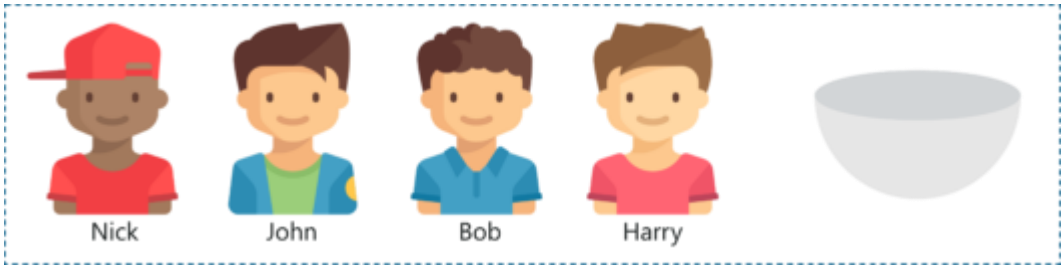
Hypothesis Testing

Statisticians use hypothesis testing to formally check whether the hypothesis is accepted or rejected. Hypothesis testing is an Inferential Statistical technique used to determine whether there is enough evidence in a data sample to infer that a certain condition holds true for an entire population.

To under the characteristics of a general population, we take a random sample and analyze the properties of the sample. We test whether or not the identified conclusion represents the population accurately and finally we interpret their results. Whether or not to accept the hypothesis depends upon the percentage value that we get from the hypothesis.

To better understand this, let's look at an example.

Consider four boys, Nick, John, Bob and Harry who were caught bunking a class. They were asked to stay back at school and clean their classroom as a punishment.



Hypothesis Testing Example – Statistics and Probability –

So, John decided that the four of them would take turns to clean their classroom. He came up with a plan of writing each of their names on chits and putting them in a bowl. Every day they had to pick up a name from the bowl and that person must clean the class.

Now it has been three days and everybody's name has come up, except John's! Assuming that this event is completely random and free of bias, what is the probability of John not cheating?

Let's begin by calculating the probability of John not being picked for a day:

P(John not picked for a day) = 3/4 = 75%

The probability here is 75%, which is fairly high. Now, if John is not picked for three days in a row, the probability drops down to 42%

P(John not picked for 3 days) = 3/4 ×3/4× 3/4 = 0.42 (approx)

Now, let's consider a situation where John is not picked for 12 days in a row! The probability drops down to 3.2%. Thus, the probability of John cheating becomes fairly high.

P(John not picked for 12 days) = (3/4) ^12 = 0.032 <?.??

In order for statisticians to come to a conclusion, they define what is known as a threshold value. Considering the above situation, if the threshold value is set to 5%, it would indicate that, if the probability lies below 5%, then John is cheating his way out of detention. But if the probability is above the threshold value, then John is just lucky, and his name isn't getting picked.

The probability and hypothesis testing give rise to two important concepts, namely:

- **Null Hypothesis:** Result is no different from assumption.
- **Alternate Hypothesis:** Result disproves the assumption.

Therefore, in our example, if the probability of an event occurring is less than 5%, then it is a biased event, hence it approves the alternate hypothesis.

To better understand Hypothesis Testing, we'll be running a quick demo in the below section.

Hypothesis Testing In R

Here we'll be using the gapminder data set to perform hypothesis testing. The gapminder data set contains a list of 142 countries, with their respective values for life expectancy, GDP per capita, and population, every five years, from 1952 to 2007.

The first step is to install and load the gapminder package into the R environment:

```
#Install and Load gapminder package
install.packages("gapminder")
library(gapminder)
data("gapminder")
```

Next, we'll display the data set by using the View() function in R:

```
#Display gapminder dataset
View(gapminder)
```

Here's a quick look at our data set:

	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	1952	28.801	8425333	779.4453
2	Afghanistan	Asia	1957	30.332	9240934	820.8530
3	Afghanistan	Asia	1962	31.997	10267083	853.1007
4	Afghanistan	Asia	1967	34.020	11537966	836.1971
5	Afghanistan	Asia	1972	36.088	13079460	739.9811
6	Afghanistan	Asia	1977	38.438	14880372	786.1134
7	Afghanistan	Asia	1982	39.854	12881816	978.0114
8	Afghanistan	Asia	1987	40.822	13867957	852.3959
9	Afghanistan	Asia	1992	41.674	16317921	649.3414
10	Afghanistan	Asia	1997	41.763	22227415	635.3414
11	Afghanistan	Asia	2002	42.129	25268405	726.7341

Data Set – Statistics And Probability –

The next step is to load the famous dplyr package provided by R.

```

#Install and Load dplyr package
install.packages("dplyr")
```

```
library(dplyr)
```

Our next step is to compare the life expectancy of two places (Ireland and South Africa) and perform the t-test to check if the comparison follows a Null Hypothesis or an Alternate Hypothesis.

```
#Comparing the variance in life expectancy in South Africa & Ireland
df1 <-gapminder %>%
select(country, lifeExp) %>%
filter(country == "South Africa" | country =="Ireland")
```

So, after you apply the t-test to the data frame (df1), and compare the life expectancy, you can see the below results:

```
#Perform t-test

t.test(data = df1, lifeExp ~ country)

Welch Two Sample t-test

data: lifeExp by country

t = 10.067, df = 19.109, p-value = 4.466e-09

alternative hypothesis: true difference in means is not equal to 0


95 percent confidence interval:

15.07022 22.97794

sample estimates:

mean in group Ireland mean in group South Africa

73.01725 53.99317
```

Notice the mean in group Ireland and in South Africa, you can see that life expectancy almost differs by a scale of 20. Now we need to check if this difference in the value of life expectancy in South Africa and Ireland is actually valid and not just by pure chance. For this reason, the t-test is carried out.

Pay special attention to the p-value also known as the probability value. The p-value is a very important measurement when it comes to ensuring the significance of a model. A model is said to be statistically significant only when the p-value is less than the pre-determined statistical significance level, which is ideally 0.05. As you can see from the output, the p-value is 4.466e-09 which is an extremely small value.

In the summary of the model, notice another important parameter called the t-value. A larger t-value suggests that the alternate hypothesis is true and that the difference in life expectancy is not equal to zero by pure luck. Hence in our case, the null hypothesis is disapproved.

So that was the practical implementation of Hypothesis Testing using the R language.

With this, we come to the end of this blog. If you have any queries regarding this topic, please leave a comment below and we'll get back to you.

Stay tuned for more blogs on the trending technologies.