upGrad

Summary

Basics of Statistics

In this session, you were introduced to the concepts of statistics and basics of data visualisation.

Introduction to Statistics

Data, today, is one of the most important tools that, if read properly, can reveal many doors and solutions at the same time. In this session, you learnt the basics of statistics and how to use data visualisation effectively.

First, you learnt what statistics is and the basic concepts in statistics.

Statistics is basically collection, summarisation, analysis and reporting of data in a well-organised manner. There are two types of statistics:

1. **Descriptive statistics:** In this type of statistics, you analyse the data that has been collected and find patterns in them. It is often used to **analyse past performances** of the employees or marketing tactics.
2. **Inferential statistics:** In this type of statistics, you analyse the data and **predict future events**. For instance, predicting company sales and impact of marketing strategy.

Next, you explored the basic concepts in statistics, which included:

1. **Data**: Any information collected or generated from online activity or otherwise is known as data. Any attribute of an object can be recorded as data. For example, the most bought dress, the cost of a dress, etc.
2. **Dataset**: A collection of data for a particular study is known as a dataset. For example, various types of purchases made on christmas.
3. **Variable**: Any characteristics such as age, weight etc that can be measured or counted. Suppose you need to use some data repeatedly in a report. Here, you can define a variable, that is, a name that represents the data. You can then use the variable instead of describing the whole data repeatedly in a report.

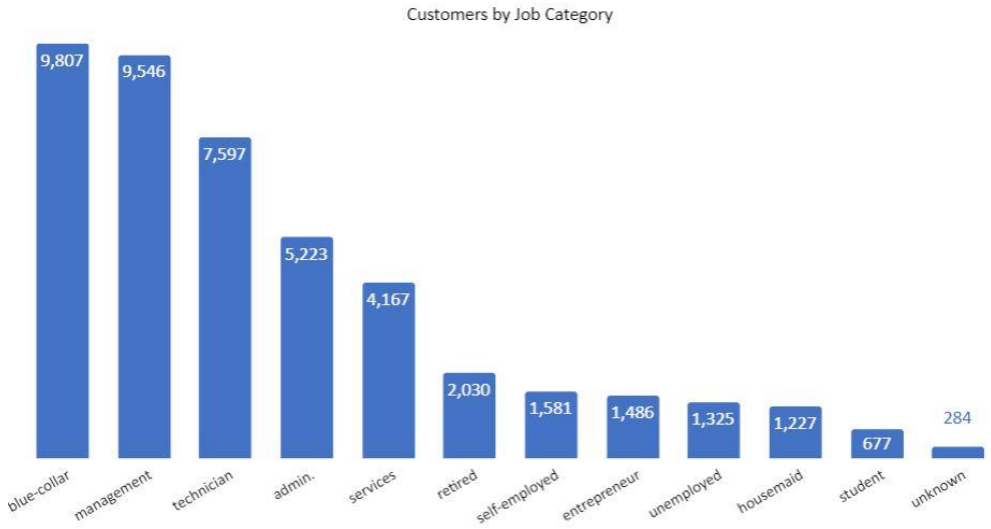
Types of variables:

- **Qualitative or categorical variables**
- **Quantitative variables**
    - **Discrete variables**
    - **Continuous variables**

Data Visualisation

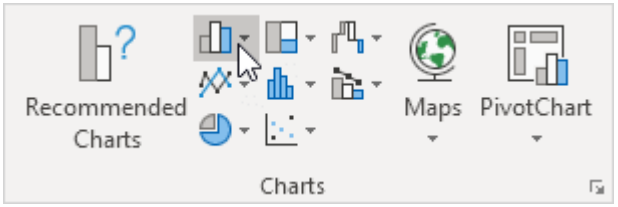
Data visualisation is a part of statistics that helps you build a narrative and present your data in the best possible way.

First, you learnt bar graphs as a data visualisation tool.
You learnt that bar graphs show relative data of various components of data.
Given below is an example of a data represented in the form of bar graph:

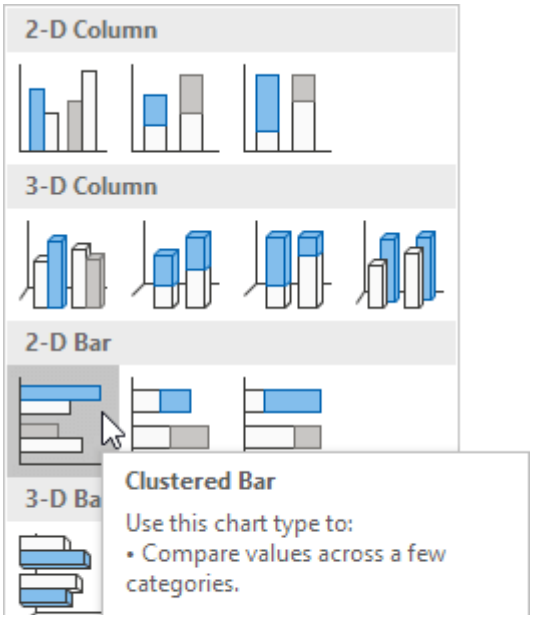


Next, you learnt how to create a bar chart in Excel:

1. Select the range. For example, F1:G16.
2. In the Insert tab, of the Charts group, click on the Column symbol.



3. Select the desired type of bar graph.



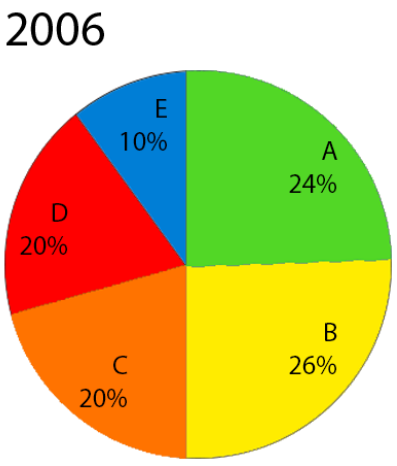
4. The resultant bar graph will appear on your sheet.

Pie Charts

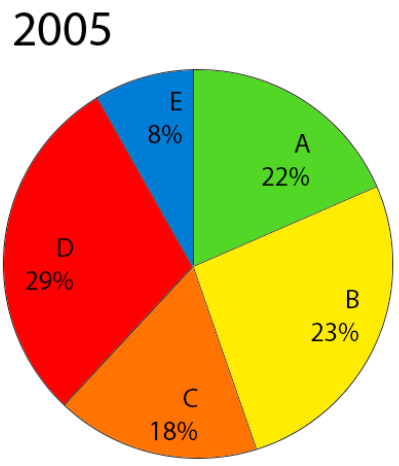
Next, you learnt another data visualisation tool, that is, **pie charts**.

Pie chart is the circular representation of data. A pie chart is used when 100% of data needs to be represented and the division of categories is within 100%.

Given below is an example of a data represented in the form of pie charts:



Number of Employees- 20,000



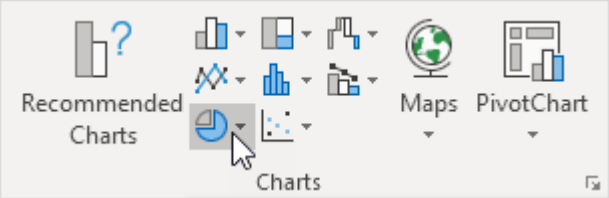
Number of Employees- 18,000

Next, you learnt how to make pie charts in Excel.

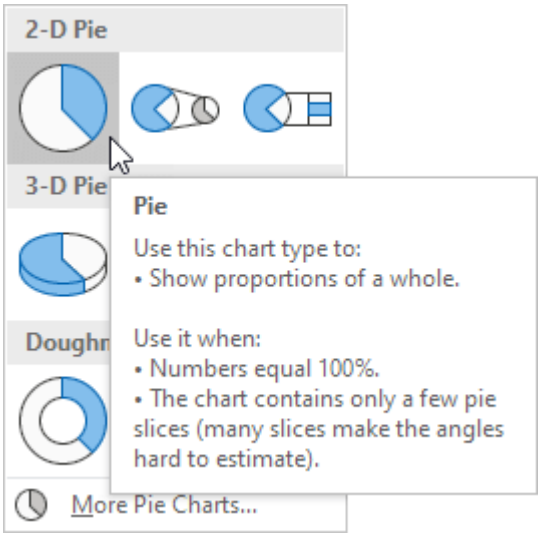
1. Select a range, say, A1:D2.

	A	B	C	D	E
1		Bears	Dolphins	Whales	
2	2017	8	150	80	
3	2018	54	77	54	
4	2019	93	32	100	
5	2020	116	11	76	
6	2021	137	6	93	
7	2022	184	1	72	
8					

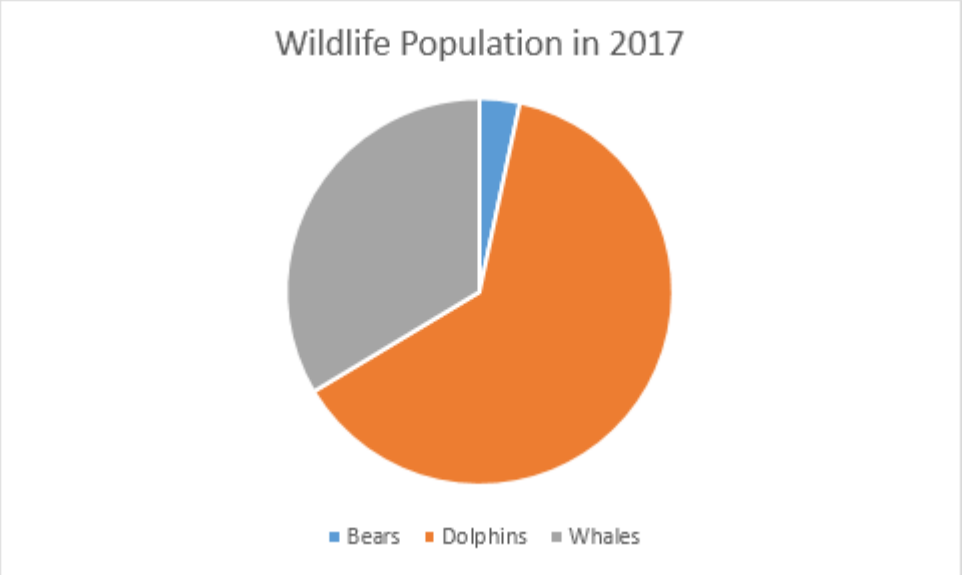
2. In the Insert option, select the Pie symbol.



3. Select the desired type of pie chart.



4. The final result will appear on the Excel sheet. Add labels as necessary.



Histograms

Next, you learnt another visualisation tool, that is, histograms.

A histogram is a visual representation of the distribution of data. The data can be of two types, numerical and categorical. Two important points about histograms are as follows:

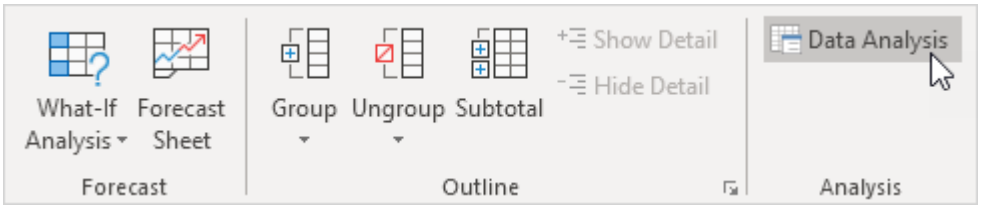
- A histogram is an extended form of a bar graph in which there are no gaps between adjacent bars.
- To construct a histogram, you need to first construct a frequency distribution table.

Finally, you learnt how to make a histogram in Excel.

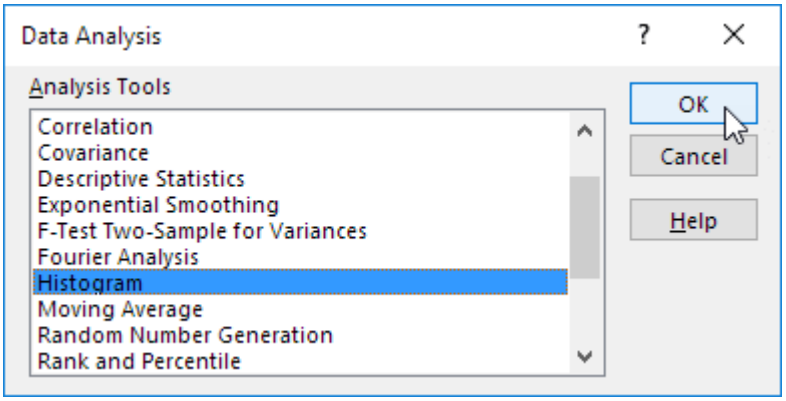
1. Enter the bin numbers (upper levels) in the desired range (say, C8:C15).

	A	B	C	D
1	Number of students			
2	22			
3	29			
4	40		20	
5	30		25	
6	48		30	
7	24		35	
8	21		40	
9	19			
10	24			
11	22			
12	25			
13	52			
14	35			
15	40			
16	31			
17	37			
18	21			
19	23			
20				

2. Click on Data Analysis in the Data tab.

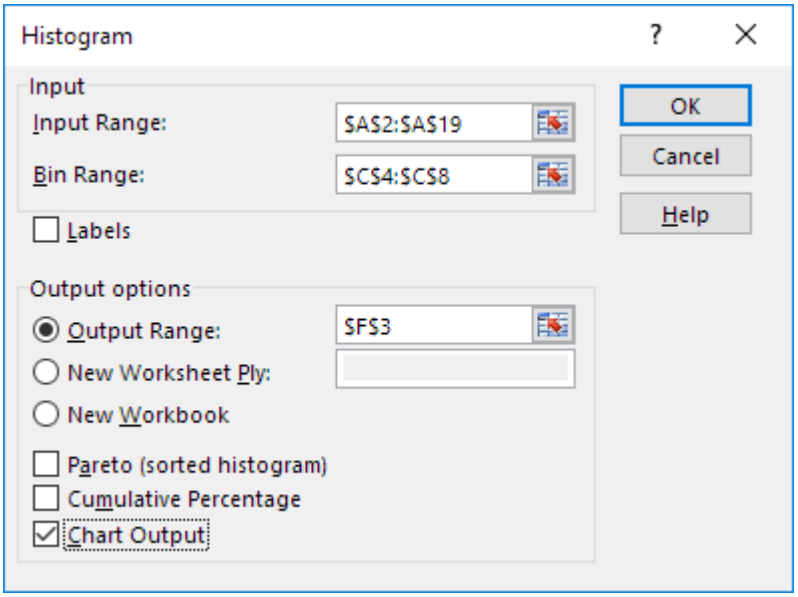


3. Select Histogram and click Okay.

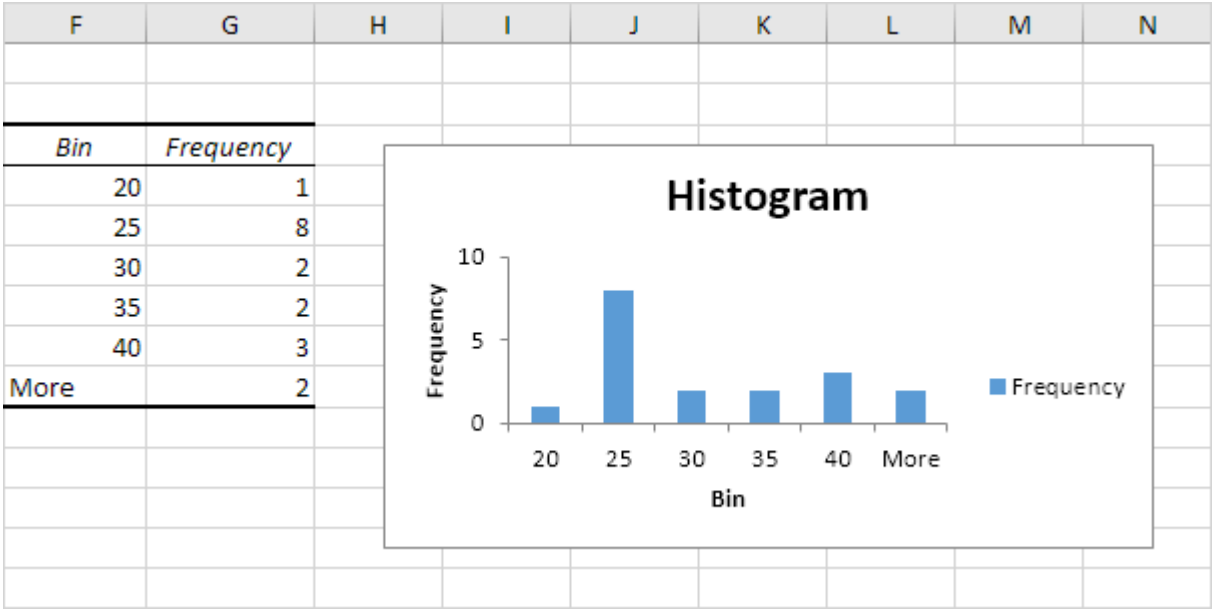


4. Select the Input range.
5. Click the Bin Range box and select the range you want.

6. Then, click on the Output Range option, click in the Output Range box and select any cell.



7. Check final Chart Output and label them.



upGrad

upGrad

Summary

Measures of Central Tendency

In this session, you learnt various measures of central tendency.

Measures of Central Tendency

Measures of central tendency help to understand data in a systematic manner. Each measure is an indication of a different tendency.

You first learnt the three basic measures of central tendency, which were:

- **Mean**
  - Mean is the sum of all the data values divided by the total number of sample values.
  - Mean is commonly represented by the symbol μ.
- **Median**
  - If you arrange the sample data in the ascending order of frequency, from left to right, the value in the middle is called the median.
  - For an odd number of values, you get one median.
  - For an even number of values, the median is the average of the two middle values.
- **Mode**
  - In a data set, the value with the highest frequency is referred to as the mode.
  - For qualitative data, it is not possible to measure the mean or median values, as there are no numerical values.
  - Thus, the variable with the highest frequency is considered as the measure of central tendency in such cases.

Next, you learnt how to calculate these measures using Excel. The formulas for mean and median would be as follows:

- **Mean** can be calculated using the Excel function =**AVERAGE(A1: A20)** if the data is distributed over A1: A20 in the Excel workbook.
- **Median** can be calculated using the Excel function **=MEDIAN(A1: A20)** if the data is distributed over A1: A20 in the Excel workbook.

upGrad

Measures of Dispersion

The second part you need to consider when calculating the measures of central tendency would be the effect of the outliers, which are called measures of dispersion.
First, you looked at the following parameters that are used as measures of dispersion:

a. **Variance:** Variance is the variability of individual data points.
b. **Standard deviation:** Standard deviation is the square root of variance.

Next, you learnt how to calculate variance using the following formula:

- **Variance = ∑(x−μ)^2/(n)** (for a population) **| ∑(x−μ)^2/(n-1)** (for a sample)

Next, you focussed on standard deviation.

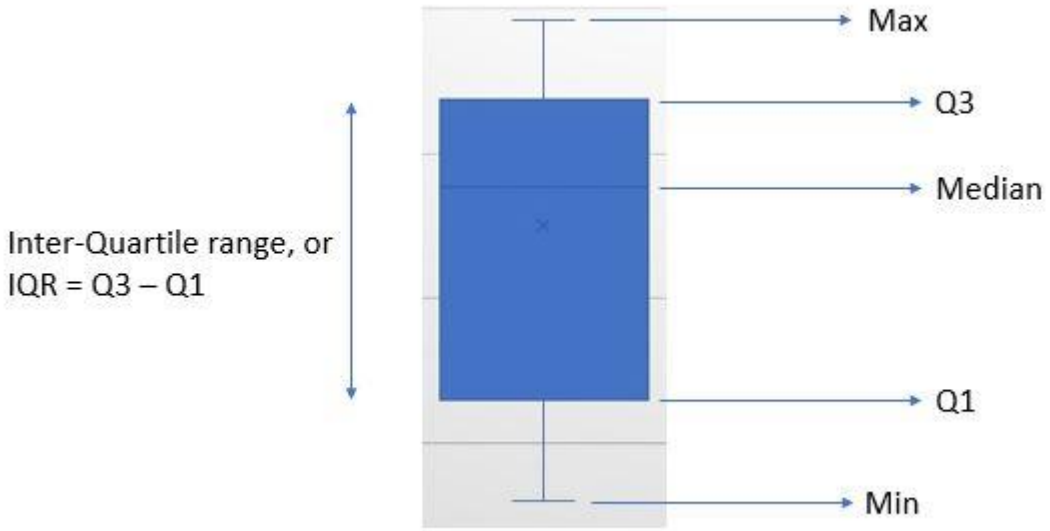
- **Standard deviation** is the square root of the variance. This metric serves the purpose of measuring variation without exaggerating its magnitude. It is popularly represented as σ. So, the variance is represented as σ^2.
- **Standard deviation** can be calculated using the Excel function **=STDEV.S(A1: A20)** if the data is distributed over A1: A20 in the Excel workbook.

Interquartile Range

Next, you learnt another measure of dispersion, that is, interquartile range. Unlike standard deviation, interquartile range is also considered as a measure of resistance.

The interquartile range is said to be a much better parameter to indicate the variation or spread in the data. Quartile values refer to the values in a sample at the 25th, 50th, 75th and 100th percentile.

Next, you learnt the visual representation of an interquartile range, as shown below:



Finally, you learnt how to calculate the interquartile scores using Excel.

- **QUARTILE** functions mentioned in the table below can be used in Excel.

Quartile	**Excel Function**
25th percentile or First Quartile	**QUARTILE (A1: A20, 1)**
50th percentile or Second Quartile	**QUARTILE (A1: A20, 2)**
75th percentile or Third Quartile	**QUARTILE (A1: A20, 3)**
100th percentile or Fourth Quartile	**QUARTILE (A1: A20, 4)**

upGrad

upGrad

Summary

Probability and Probability Distribution

In this session, you learnt all about probability and different concepts involved in probability.

Probability

When it comes to dealing with data, it is important to understand how probable it is for an event to occur or how likely a proposition is true, and probability does exactly that.

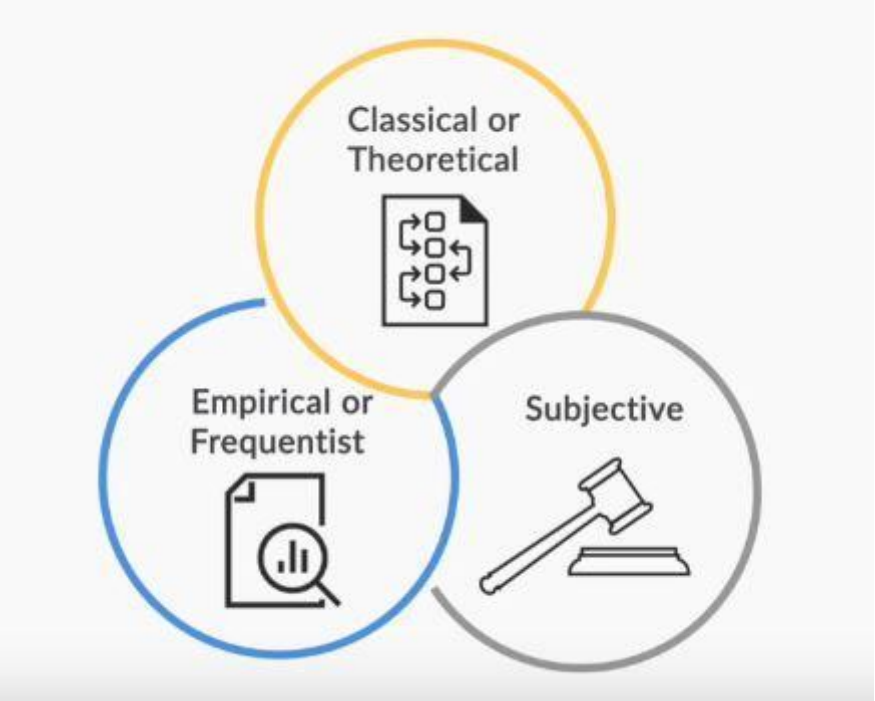
Probability helps to take well-informed and calculated decisions, which is essential when it comes to handling a business.
In this segment, you learnt about what exactly probability is.
Probability is a **measure of uncertainty** that helps us understand the chance that a certain event can occur out of all possible outcomes.

Calculation of Probability

Probability is a mathematical formula of **measuring your chances**.

After learning about what probability is, you moved on to learning how to calculate probability.

The formula for probability is as follows:
Probability of an event = Number of favourable outcomes/
Total number of equally likely outcomes
Next, you learnt the three ways of calculating probability:



1. **Classical/Theoretical approach** - Certain assumptions are made (for instance, all possible outcomes are equally likely) and the probability values are then calculated using a formula.

2. **Empirical/Frequentist approach** - According to this approach, probabilities are derived from observations.

3. **Subjective approach -** It reflects an individual's understanding and judgement of how likely the outcomes are.

Finally, you looked at some basic rules of probability:

1. The value of the probability of any event **always lies between 0 and 1**. This is fairly intuitive, as on the extreme ends, either the event is impossible (for instance, the probability of rolling a regular die and getting a '7' on the face) or it always occurs (for instance, the probability of tossing a regular coin and getting either heads or tails). Most events, however, lie somewhere in between. The range that the probability of an event A can take is represented mathematically as follows:
    - **$0 \leq P(A) \leq 1$**
2. According to the **addition rule of probability**, if two events, A and B, are **mutually exclusive** (both the events cannot occur at the same time), then the probability that either of them occurs is equal to the sum of their individual probabilities. This rule can be represented mathematically as follows:
    - **P(A 'or' B) = P(A) + P(B)**; if A and B are mutually exclusive events

The outcomes of tossing a coin or rolling a die are mutually exclusive. In other words, you cannot get both '3' and '5' on rolling a regular die once. Hence, the probability of getting either '3' or '5' is equal to the sum of the probability of getting '3' and the probability of getting '5'.

3. According to the **multiplication rule of probability**, if two events, A and B, are **independent** (the occurrence of one event does not affect the probability of the occurrence of the other), then the probability that both of them will occur is equal to the product of their individual probabilities. This rule can be represented mathematically as follows:
    - **P(A 'and' B) = P(A) * P(B)**; if A and B are independent events

Random Variables

A random variable maps the outcome of a random process to a numerical value.

Suppose you are tossing a coin, which is a random process. You can define a random variable X that takes the value 10.5 if the outcome is 'heads' and 15.7 if the outcome is 'tails'. In this case, the probability of getting heads will be denoted by **P(X = 10.5)** and that of getting tails will be denoted by **P(X = 15.7)**. If you use a fair coin, then both these values will be equal to ½.

It is important to note that even if a process inherently produces numerical outcomes, such as the roll of a die, a random variable can be defined such that it alters the numerical outcomes. For example, define a random variable Y as twice the number that appears on the face of a die when you roll it. In this case, your random variable Y takes the values 2, 4, 6, 8, 10 and 12 when the face of the die shows 1, 2, 3, 4, 5 and 6, respectively.

Discrete and Continuous Random Variable

After learning about random variables, you learnt about discrete and continuous variables.

A random variable is discrete if there are finite or countable values. And, random variables are continuous if the values have intervals and are normally infinite.

Next, you learnt how to calculate the mean, variance and standard deviation of a discrete random variable. A probability distribution could be any of the following:

- An equation
- P(x) = x/21
- (for x = 1, 2, 3, 4, 5 and 6)

The sum of the probabilities of all of the outcomes must be equal to 1.

- If the probability of getting heads on tossing a coin is 0.7, then that of not getting heads (i.e., getting tails) must be 0.3 so that the sum of the probabilities of all the possible outcomes equals 1. The mean of the random variable is also referred to as the expected value of the random variable.

- The expected value of a random variable X is the average value of X that you would 'expect' to obtain after performing the experiment for an infinite number of times.

The variance of X gives you an indication of the spread of the values from the mean that the random variable can take. The higher the variance, the higher the number of the values that are spread out from the mean of the distribution.

The standard deviation is also a measure of the dispersion of the values of the random variable from its mean.

Continuous Probability Distribution

In this segment, you learnt that the probability distribution of a continuous random variable is called a probability density function (PDF). It is important to understand that the value of a PDF at a given point is not the probability of obtaining that point. In fact, the probability of obtaining any single point is 0. Some properties of a PDF are listed below:

1. The probability of obtaining any single point is 0.
2. The probability of an interval is the area under the PDF curve in that interval.
3. The total area under the PDF curve is always equal to 1 since the total probability should equal 1.

Next, you learnt that the cumulative probability at a point is the probability of the occurrence of any value less than or equal to that point. It is denoted with F(x) and can be represented mathematically as follows:

- $F(x) = P(X \leq x)$

For a continuous random variable, the cumulative probability at a point is the area under the curve from $-\infty$ up to that point. Since the PDF never takes a negative value, the cumulative probability function is a non-decreasing function. This means that the value of the cumulative probability function either stays the same or increases as we move to the right in the graph.

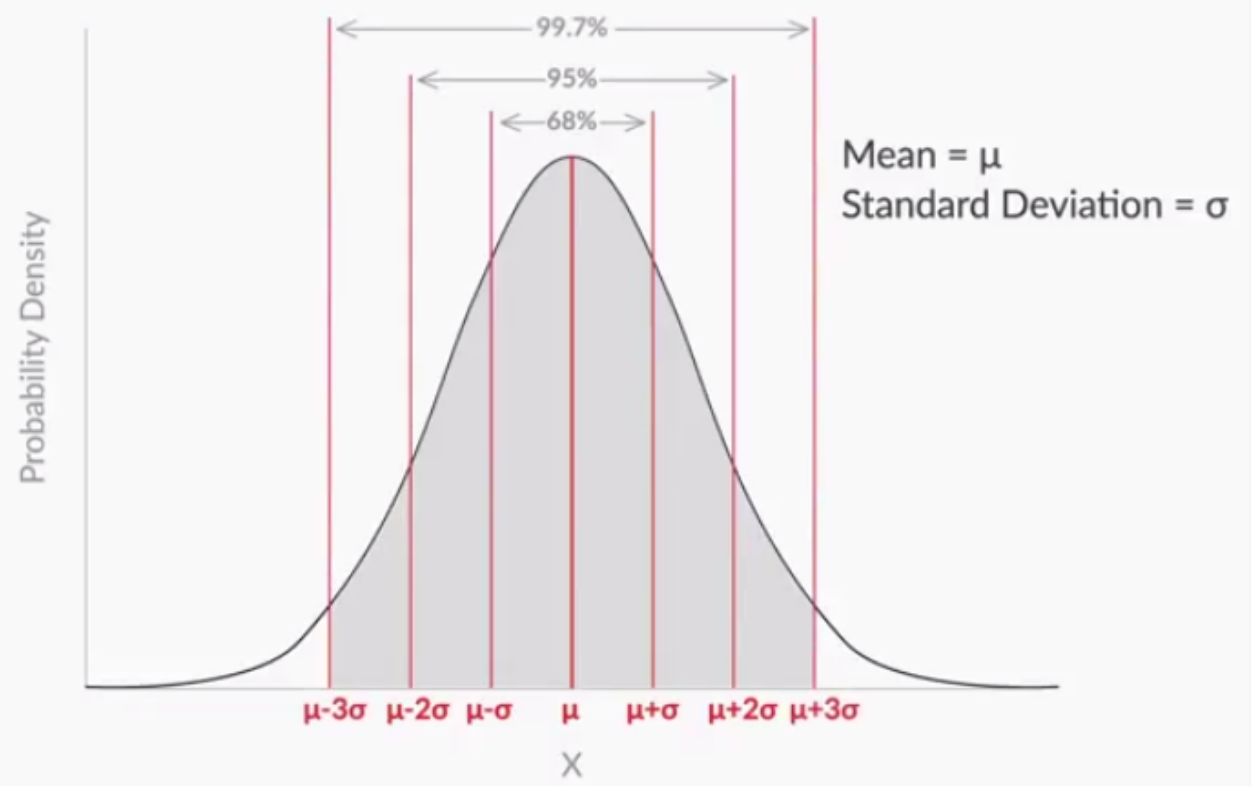
Normal Distribution

The most commonly occurring continuous probability distribution is a normal distribution. Several natural phenomena, such as the height of men/women of a certain age, blood pressure and IQ scores, follow a normal distribution. Normal distribution is symmetric with respect to its mean and extends infinitely on both sides.

In a normal distribution, the probability density is higher, close to the mean and decreases exponentially as you move further away from the mean. In simple terms, it means that there is a high probability that the value

of the random variable is close to the mean. As you move further away from the mean, the probability of the occurrence of such values decreases.

The empirical rule is illustrated in the image given below.



The empirical rule states that there is:

1. A 68% probability of the variable lying within one standard deviation of the mean,

2. A 95% probability of the variable lying within two standard deviations of the mean, and

3. A 99.7% probability of the variable lying within three standard deviations of the mean.

In other words, in case you look at the outcome of one trial at random, there is a 68% chance that the outcome lies within one standard deviation of the mean, a 95% chance that the outcome lies within two standard deviations of the mean and a 99.7% chance that the outcome lies within three standard deviations of the mean. It is important to note that the numbers specified in the empirical rule are only approximations.

upGrad

Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

The formula used to convert any point in a normal distribution into its equivalent point (referred to as the 'z-score') in the standard normal distribution is as follows:

- Z-score = x - μ / σ

An important point to note here is that the properties of the original point on a normal distribution are retained when it is translated on to the standard normal distribution. For instance, if the cumulative probability of x is 0.813 in a normal distribution, the cumulative probability of its equivalent z-score will also be 0.813 in the standard normal distribution.

The z-score tells you how many standard deviations away your observed value is from the mean. This is extremely helpful in probability calculations and for comparing points on two different normal distributions. The 'NORM.DIST' function in Excel can be used to calculate the cumulative probability of any point on a normal distribution. Similarly, the 'NORM.S.DIST' function can be used to calculate the cumulative probability of any point on the standard normal distribution.

upGrad

Summary

Sampling and Distribution

In this session, you learnt about the importance of sampling in decision-making.

Samples vs Populations

A sample is only a small part of a larger population.

Suppose you want to find the average number of times the people in India visited a doctor last year for a business application. Now, given India's population, that number must be in the millions, if not billions! You cannot possibly ask this question to every person, which would be a costly and time-consuming process.

A sample is a small part of a population. Often, the population is too large, making it infeasible or expensive to collect data from the entire population. This is where inferential statistics comes into play. By now, you know that inferential statistics involves making inferences about or deriving insights into a large population from a small sample.

The parameters of a sample (mean, variance, etc.) and a population are calculated using the same formulae. However, one major difference is that for a sample of size n, the formula for calculating the variance (S^2) has a denominator of 'n - 1', whereas in the case of a population of size N, the formula for calculating the variance (σ^2) has a denominator of 'N'. The table given below includes the formulae and notations related to populations and their samples.

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}$

Although it is not explicitly mentioned, it is assumed that when we say that we are collecting a sample, we are collecting a random one. Simply put, a random sample is chosen randomly without any bias.

Central Limit Theorem

The central limit theorem states that if you take a sufficiently large number of random samples (sample size 'n') from any distribution with a mean of μ and a standard deviation of σ, then the distribution of the sample means (or the 'sampling distribution of sample means') will be a normal distribution with a **mean** of **μ** and a **standard deviation** of **σ/√n.**

As the sample size increases, the sampling distribution of the sample means becomes narrower and better resembles a normal distribution.

The sample size should at least be 30 to apply the central limit theorem. One of the biggest implications of the theorem is that it can be applied irrespective of the probability distribution of the population.

Confidence Intervals

Here are the steps to arrive at a confidence interval within which the population mean (μ) will lie with a certain probability:

- First, the sampling distribution of sample means is constructed for a particular sample size (n):
  - The empirical rule states that 95% of all sample means lie within two standard errors of the mean of the sampling distribution (which is also the mean of the population according to the central limit theorem).
- Next, pick a random sample of a minimum sample size of 30, as this will allow you to apply the central limit theorem.
- If the standard deviation of the population is unknown, you need to estimate the sample standard deviation (S) as the population standard deviation (σ):
  - This can be done only if you assume that the population follows a normal distribution. Hence, if you cannot make the assumption that the population follows a normal distribution, then you need to be provided with an estimate of the population standard deviation before you proceed.
- According to the empirical rule, the sample mean of our randomly picked sample has a 95% chance of lying within two standard errors of the population mean. So, we can say that the population mean has a 95% chance of lying within two standard errors of the sample mean. Hence, the 95% confidence interval for the population mean is '$\bar{X}$ - 2 * σs/√n' to '$\bar{X}$ + 2 * σ/√n'.

The confidence interval for different confidence levels can be calculated using the following formula:

Confidence interval = μ ± (z-score * σ/√n)

Here, μ is the mean of the sample, s is the standard deviation of the sample and n is the sample size.

The z-score depends on the confidence level chosen, as shown in the table given below.

Confidence Level	Z-score
50%	0.674
80%	1.282
90%	1.645
95%	1.960
99%	2.576

upGrad

Summary

Hypothesis Testing: Z-Distribution

In business, you often have to make decisions with lakhs of rupees on stake. You cannot make such decisions purely based on intuition. You need to follow a more robust method to ensure that your decision is not biased. This is where hypothesis testing comes into the picture. It is important because it adds statistical rigour to decision-making.

Hypothesis testing involves several steps such as defining the hypothesis statements, choosing a significance level, collecting data points to construct the sample and calculating the test statistic, and finally making a decision. The decision can either be to reject or fail to reject the null hypothesis.

The Null and Alternative Hypotheses

The process of hypothesis testing starts with defining the two hypothesis statements.

The null hypothesis (H_0) is typically defined as the status quo or the commonly-held belief/assumption. It can be a popular belief/claim that you want to test. By convention, the null hypothesis contains equality in some form, i.e. either '=', '≤', or '≥'.

The alternative hypothesis (H_α or H_a) is defined as the perfect opposite of the null hypothesis; hence, it typically challenges the status quo. A common strategy is to define the alternative hypothesis as the one that you are trying to prove.

It is important to note that we always **begin with the assumption that the null hypothesis is true**. Then:

- If we have sufficient evidence to prove that the null hypothesis is false, we '**reject**' it. In this case, the alternative hypothesis is **proved to be true**.
- If we do NOT have sufficient evidence to prove that the null hypothesis is false, we '**fail to reject**' it. In this case, the assumption that the null hypothesis is true remains.

Remember that in hypothesis testing parlance, we never "prove" the null hypothesis. We can only say that we 'fail to reject' the null hypothesis based on the evidence that we have gathered.

Critical Value Method

You can follow these steps to conduct a hypothesis test using the critical value method:

- **Step 1**: Frame the appropriate null and alternative hypotheses.
- **Step 2**: Decide the confidence level (or the significance level).

The significance level (α) is defined as '1 - confidence level'. The significance level helps us identify the unlikely values of the sample statistic, assuming that the null hypothesis is true. It represents the area outside

the confidence interval (which is the rejection region). The higher the significance level, the higher are the chances of rejecting the null hypothesis. This means that the lower the significance level, the lower are the chances of rejecting the null hypothesis.
It is important to remember that 0.01, 0.05 and 0.10 are the most commonly used values of the significance level. If the significance level is not specified in the question, we choose a default 5% significance level.

- **Step 3**: Determine the critical z-score(s) (and the rejection region).

The critical z-score(s) depends on the significance level of the test. It defines the boundary of the region beyond which if the sample z-score lies, we reject the null hypothesis. This region is referred to as the 'rejection region'. The region of the graph that represents values closer to the hypothesised mean than the critical z-score(s) is called the 'acceptance region'.

- **Step 4**: Compute the sample z-score (or 'test statistic').

The sample z-score tells us how many standard deviations away from the mean the sample mean lies in the distribution of sample means.

$$\text{Sample z-score} = \frac{\text{Sample mean - Hypothesised mean}}{\text{Standard deviation of sample mean}}$$

Note that the standard deviation of the sample means distribution is also referred to as the 'standard error of the mean', or simply the 'standard error', and is denoted by 'SE'.

$$SE = \text{Standard deviation of population} / \sqrt{\text{Sample size}}$$

Therefore, the equation to calculate the sample z-score can be written as:

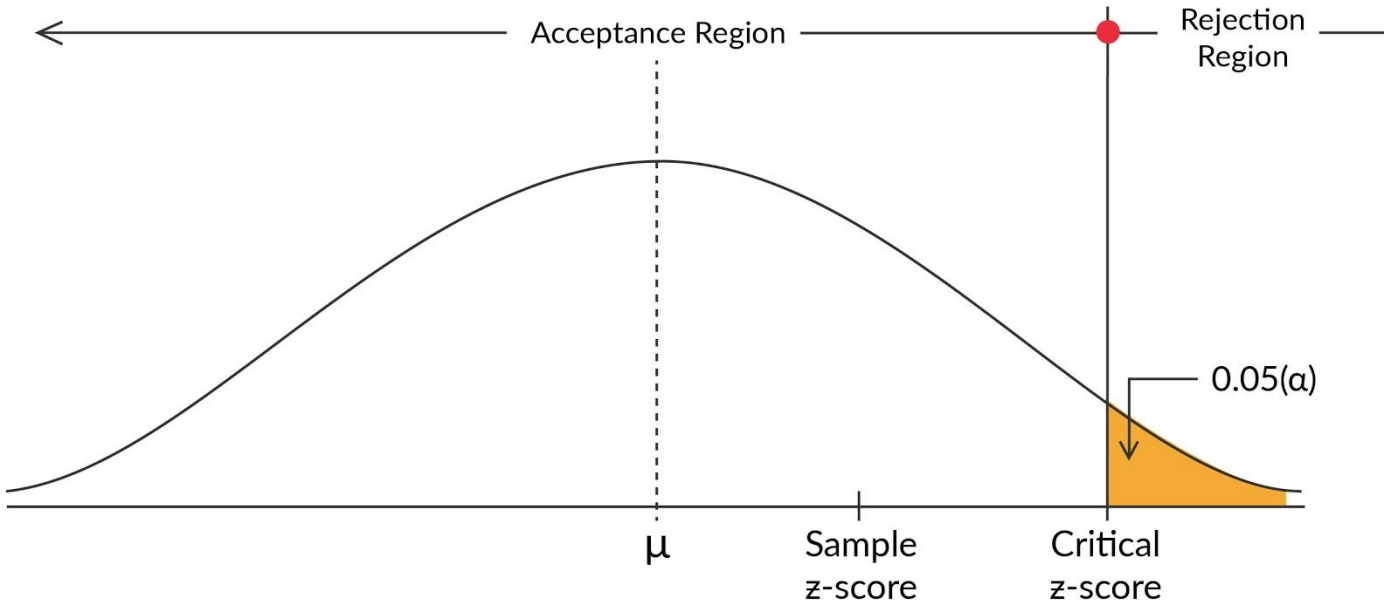
$$\text{Sample z-score} = \frac{\text{Sample mean - Hypothesised mean}}{\text{Standard deviation of population} / \sqrt{\text{Sample size}}}$$

If the sample size is large (≥30), the Central Limit Theorem allows us to approximate the population standard deviation (σ) as the sample standard deviation (S). Also, the mean of the sample means distribution is assumed to be equal to the population mean.

It is important to note that in the context of z-tests, the term 'sample z-score' is often used interchangeably with 'z-statistic'. The term 'test statistic' is also commonly used to refer to the standardized sample mean.

- **Step 5**: Reach a decision and interpret the result.

Finally, we compare the sample z-score with the critical z-score(s) of the test.



If the sample z-score lies in the acceptance region, we fail to reject the null hypothesis and conclude that the null hypothesis remains true. If the sample z-score lies in the rejection region, we reject the null hypothesis and conclude that the null hypothesis is false and the alternative hypothesis is true.

Types of Tests

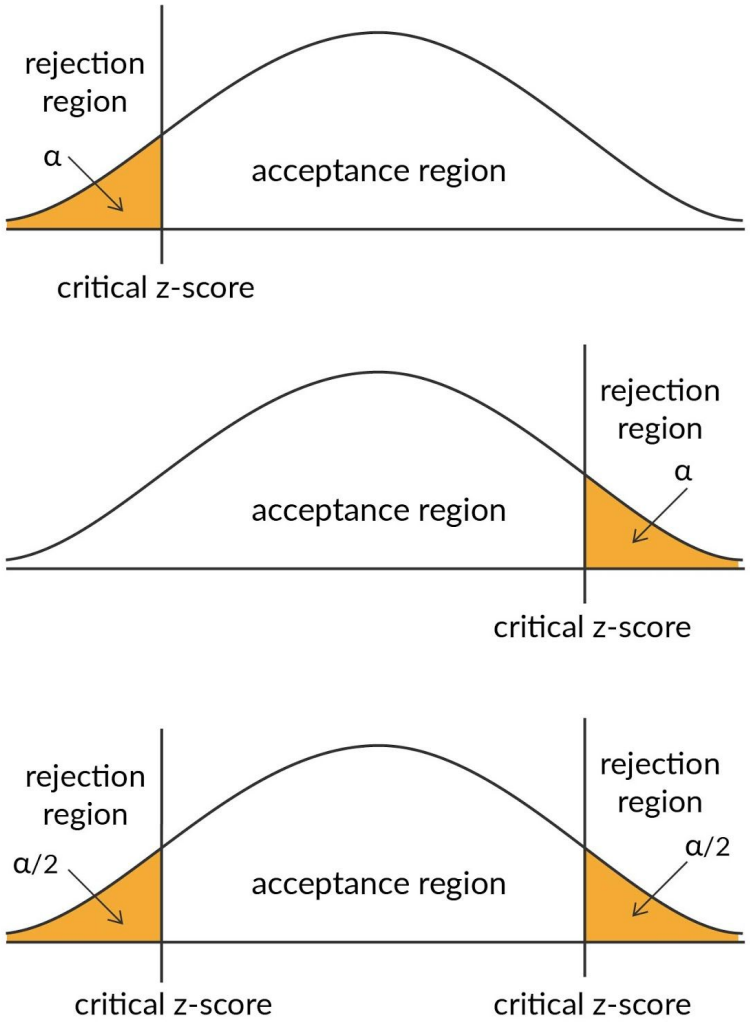
Depending on how the hypotheses are formulated, you test for deviation from the hypothesised mean on either one side of the mean (one-tailed tests) or on both sides of the mean (two-tailed tests).

One-tailed tests can be either left-tailed or right-tailed, depending on which side of the curve the rejection region lies. The formulation of the null and alternative hypotheses determines the type of the test and the position of the rejection region(s) in the distribution of sample means.

You can tell the type of the test and the position of the rejection region(s) on the basis of the 'sign' in the alternate hypothesis.

- '<' in H_a → Left-tailed test → Rejection region on the left side of the distribution
- '>' in H_a → Right-tailed test → Rejection region on the right side of the distribution
- '≠' in H_a → Two-tailed test → Rejection region on both sides of the distribution

These can be better understood with the help of the image below:



In the image, you can see how the significance region defines the critical region for different types of tests (left-tailed, right-tailed and two-tailed, respectively). You can use the table provided below to arrive at the critical z-scores for different significance levels and types of test:

Significance Level (α)	One-Tailed Test (Left)	One-Tailed Test (Right)	Two-Tailed Test
0.01	-2.326	+2.326	±2.58
0.05	**-1.645**	**+1.645**	±1.96
0.10	-1.282	+1.282	**±1.645**

As the significance level is split equally between the two sides in a two-tailed test, we get half the value of the significance level on each side. This is evident from the fact that for a two-tailed test with a significance level of 0.10, the critical z-scores are ±1.645. Here, 0.10 is split into 0.05 on the left tail and 0.05 on the right tail. Hence, for one-tailed tests with a significance level of 0.05, we get critical z-scores of -1.645 and +1.645 for the left-tailed and right-tailed tests, respectively.

upGrad

P-Value Method

You can follow these steps to conduct a hypothesis test using the p-value method:

- **Step 1**: Frame the appropriate null and alternative hypotheses.
- **Step 2**: Decide the confidence level (or the significance level).
- **Step 3**: Compute the sample z-score and the p-value.

The p-value is the area that lies farther away from the hypothesised mean than the calculated sample z-score. Intuitively, it can be understood as the probability of getting a test statistic at least as extreme as the one observed under the assumption that the null hypothesis is true.

Simplifying this, one can also understand the p-value as the probability of the null hypothesis being true. Hence, a higher p-value indicates greater evidence in favour of the null hypothesis. Similarly, a lower p-value indicates greater evidence against the null hypothesis (and in favour of the alternative hypothesis).

If you enter the same sample z-score (say, +1.5) into the 'P Value Calculator', then you will notice that the p-value for a two-tailed test (0.1336) is twice that for a one-tailed test (0.06681). This is because, in a two-tailed test, the significance level is split into two equal halves - one half for each tail. Hence, we can either compare the probability value with $\alpha/2$, or we can multiple the probability value by 2 (which gives us the p-value for a two-tailed test) and compare it with α. The latter is what the calculator is doing for you.

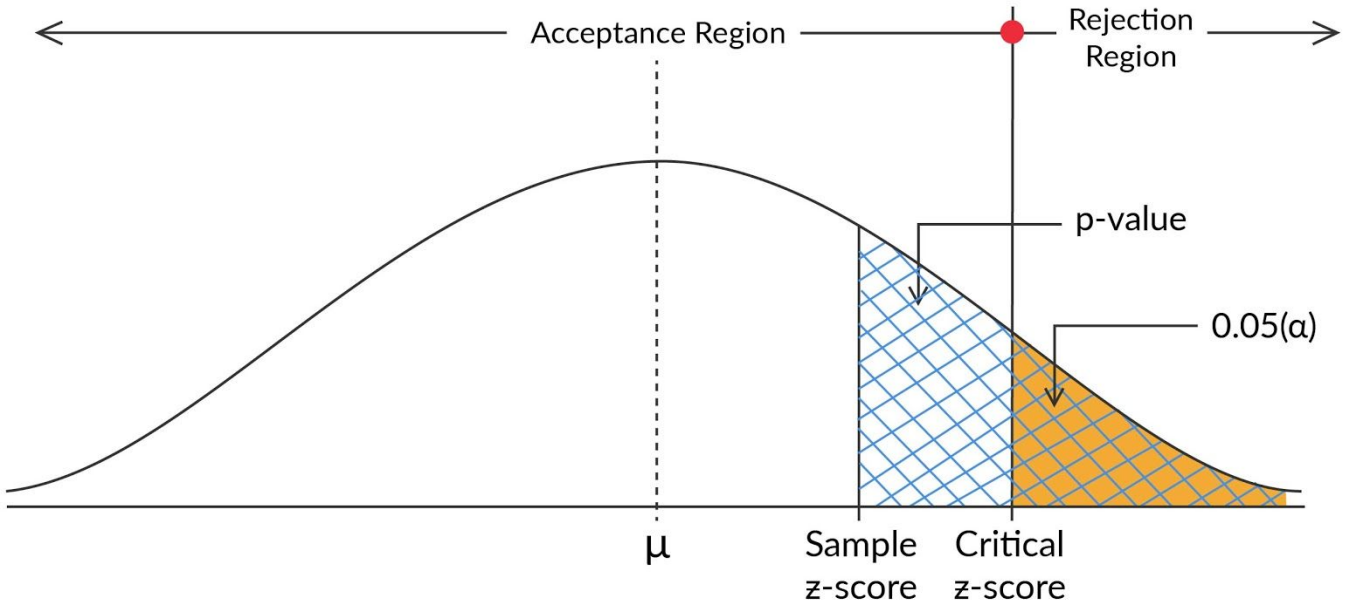
- **Step 4**: Reach a decision and interpret the result.

The p-value conveniently lets you know the confidence levels (or significance levels) at which you can reject a null hypothesis. The simple rule that you can follow while using the p-value method to test a hypothesis is:

- If p-value $< \alpha \rightarrow$ Reject the null hypothesis.
- If p-value $\geq \alpha \rightarrow$ Fail to reject the null hypothesis.

Hence, under the assumption that the null hypothesis is true, if the probability of obtaining the test statistic is less than the significance level of the test, we reject the null hypothesis in favour of the alternative hypothesis. The significance level, therefore, functions as a cutoff to determine what values of the sample statistic are too unlikely to have been observed if the null hypothesis were indeed true.

This rule can be understood intuitively as an extension of the critical value method. Take the example of a right-tailed test. We know that the p-value of the critical z-score will be nothing but the significance level of the test. If the sample z-score lies in the acceptance region (i.e., to the left of the critical z-score) its p-value will be greater than the significance level as a greater area of the curve lies to its right.



This reasoning can be extended to the other types of tests as well. Hence, using either the critical value method or the p-value method should give you the exact same result for identical sample and population parameters.

upGrad

Summary

Hypothesis Testing: t-Distribution

So far, you've had the luxury of collecting lots of data points, which meant your sample size was at least 30. Thus, the central limit theorem allowed you to assume that your population standard deviation is equal to the sample standard deviation. In case the sample size was less than 30, the population standard deviation was available to you.

However, you often have to deal with scenarios where the sample size is small and the population standard deviation is unknown. Here, a t-test is performed. Also, in several instances you need to compare the means of two samples. Here, a two-sample test is performed.

While hypothesis testing is statistically significant, there is still a possibility of committing an error. The types of errors and their implications need to be properly understood in order to make sensible choices.

t-test for a Single Sample

When the sample size is less than 30, a t-test is used to conduct hypothesis tests. A t-test uses the standard t-distribution, which is broader than the standard z-distribution. It is, however, similar to a z-distribution in most other respects; for example, it is symmetrical about its central tendency.

In a t-test, the term 't-value', 't-score', 't-statistic' or 'test statistic' is used instead of 'sample z-score'. The formula for calculating the t-statistic is as follows.

$$Test\ statistic = \frac{Sample\ mean - Hypothesised\ mean}{Standard\ deviation\ of\ the\ sample}$$

Note that the standard deviation of the sample means distribution is also referred to as the 'standard error of the mean', or simply the 'standard error', and is denoted by 'SE'.

$$Standard\ error\ of\ the\ mean\ (SE) = \frac{Sample\ standard\ deviation\ (S)}{\sqrt{Sample\ size\ (n)}}$$

Therefore, the formula for calculating the test statistic can be written as:

$$Test\ statistic = \frac{(Sample\ mean - Hypothesised\ mean)}{\frac{Sample\ standard\ deviation}{\sqrt{Sample\ size}}}$$

As you can see, the only difference between the formula for the test statistic for the t-test vis-a-vis the z-test is that the standard error is calculated using the sample standard deviation (S) in the t-test, as opposed to the population standard deviation (σ) in the z-test.

Also, the t-distribution depends on an additional parameter called the degrees of freedom (d.f.), which can be expressed as follows:

$$d.f. = Sample\ size\ (n) - 1$$

As the sample size (i.e., the degrees of freedom) increases, the t-distribution tends to become narrower and narrower. At sample sizes greater than or equal to 30, the t-distribution is essentially indistinguishable from a normal distribution. Hence, if the population standard deviation is unknown and the sample size is greater than or equal to 30, the t-test and z-test give the same results.

As a general practice, we shall use the t-test when the sample size is less than 30, irrespective of whether or not the population standard deviation is known. Similarly, we shall use the z-test when the sample size is greater than or equal to 30, irrespective of whether or not the population standard deviation is known.

We use the 'P Value Calculator' to compute the p-value and compare it to the significance level of the test in order to make a decision regarding the hypotheses.

t-test for Two-Sample Means

Two-sample means tests are of the following two types:

1. **Paired two-sample means test**, which is used when the two samples are related to each other. This is typically the case when repeated observations are made from the same population on different occasions.
2. **Unpaired two-sample means test**, which is used when your sample observations are independent of each other. This is typically the case when observations are made from two different populations.

Remember that we usually select the option for conducting the two-sample unpaired test '**assuming unequal variances**' as it is safer to do so when we do not have any information about the variances.

We use the 'Data Analysis Toolpak' add-in in Microsoft Excel to perform the test. We follow the rejection rule comparing the p-value to the significance level of the test in order to make a decision regarding the hypotheses.

Types of Errors

The two types of errors that can occur during a hypothesis test are as follows:

1. A **type 1 error** is committed when you reject a null hypothesis that is actually true. The probability of committing a type 1 error is nothing but the significance level (α) of the test.
2. A **type 2 error** is committed when you fail to reject a null hypothesis that is actually false. The probability of committing a type 2 error is denoted by β .

Increasing the significance level of the test increases the size of the rejection region, thereby increasing the chances of rejecting the null hypothesis (irrespective of whether it is false or not). Hence, increasing the significance level (α) of the test results in an increase in the chances of committing a type 1 error.

The probabilities of committing type 1 and type 2 errors are inversely related to each other.

Also, keeping all other things constant, increasing the sample size has the effect of decreasing the probability of committing a type 2 error without affecting the probability of committing a type 1 error (as it is the chosen significance level).

Depending on the circumstances, the costs of committing either of these two errors vary. The choice of significance level affects the probabilities of the types of error that a business can encounter and the implications of this choice. Depending on the context, you can often decide which type of error is costlier and which is more tolerable.

upGrad

Summary

A/B Testing

A/B testing is a controlled experiment where statistical hypothesis testing is used to determine which variant of an experience performs better for a predetermined goal. It is especially popular in the business world.

For example, while developing an e-commerce website, there could be diverse opinions about the choices of various elements, such as the shapes of buttons, the text on the call-to-action buttons, the colour of various UI elements and numerous other such things.

Often, the choice of these elements is quite subjective and it is difficult to predict which option would perform better. To resolve such conflicts, you can use A/B testing. It provides a means for you to test two different versions of the same element/experience and check which one performs better.

A/B Testing: Process

In business, A/B Testing is predominantly used to deal with tests that generate categorical data. The categories could be True/False, 1/0, Yes/No, Male/Female, Success/Failure, etc.

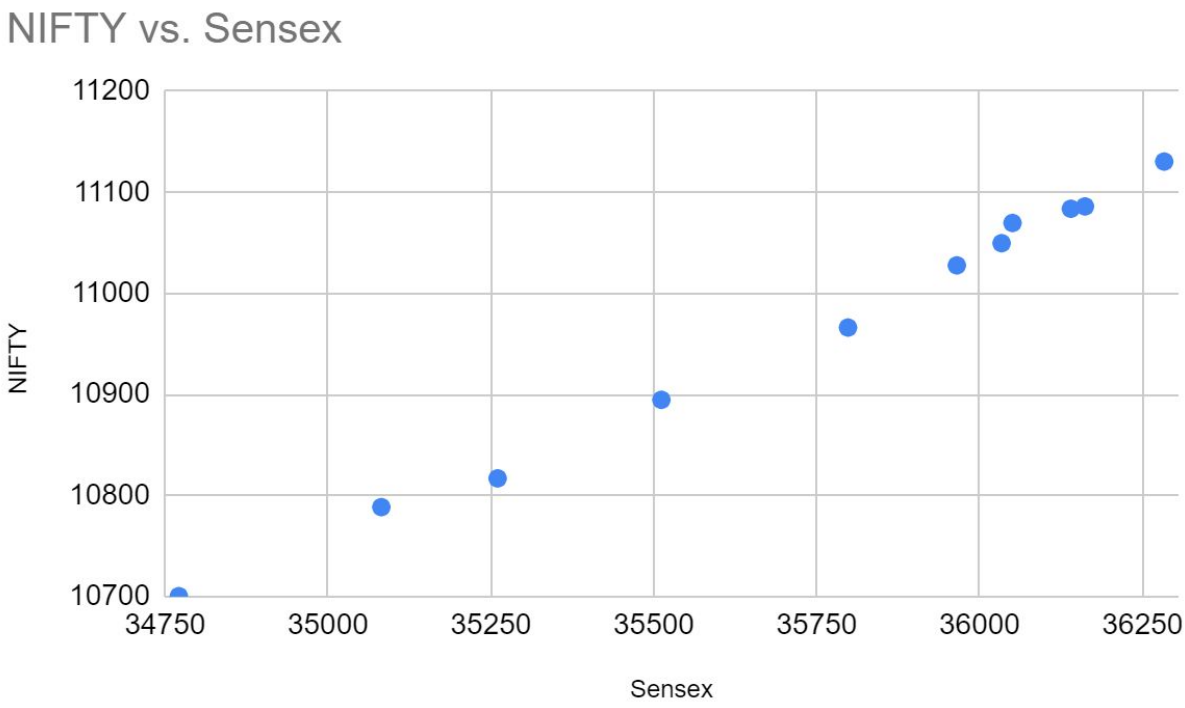
We use SurveyMonkey's AB testing calculator to perform the test. We follow the rejection rule comparing the p-value to the significance level of the test in order to make a decision regarding the hypotheses.

Session Summary

Covariance and Correlation

Decision Science

In this segment, you learnt how to make scatter- plots using Excel, and how we can interpret data in a scatter-plot.



Using this scatter- plot, you learnt how to build and decipher a relationship between NIFTY and Sensex. Each dot on the graph represents a value on the X-axis and Y-axis respectively.

- For every value of Sensex on the X-axis, the dot is the corresponding value of NIFTY on the Y-axis.
- Here, you can see a relationship between Sensex and NIFTY. Almost all the dots on the scatter-plot can be connected using a straight line. This hints at a linear relationship between the two data columns.

Next,

You learnt about Covariance and Correlation:

Covariance: It measures the directional association between two variables

- A positive value of covariance indicates that one of the variables increases with an increase in the other one.
- A negative value indicates that one of the variables decreases with an increase in the other one.
- The magnitude of covariance cannot give you any idea of the strength of the association between the two variables. Also, the covariance value doesn't have any bounds and can take any value.

You can calculate the covariance between two variables using the Data Analysis ToolPak or COVARIANCE.P function in Excel.

Correlation: It measures the extent of the association between two variables — the directional association and its strength.

- The value of the correlation coefficient always ranges from -1 to 1
- If the correlation coefficient is —
    - 1, then the two variables are perfectly positively correlated with each other
    - 0, then the two variables are not correlated with each other
    - -1, then the two variables are perfectly negatively correlated with each other

You can calculate the correlation between two variables using the Data Analysis ToolPak or CORREL function in Excel.

You solved a case study, and learnt the following observations about Covariance and Correlation:

- Correlation is not Causation
    - For example: In a college, it was observed that students who slept with their shoes on, woke up with severe headaches.
    - So, can we conclude that sleeping with your shoes on causes headaches? Is this correlation between sleeping with shoes on and severe headaches justified?
    - We should note that this conclusion ignores that the students are going to bed drunk, which causes the students to sleep with their shoes on in the first place!
    - In this case, the correlation between going to sleep with shoes on and headaches is known as a **spurious correlation.** The fact that the students are going to bed drunk is known as the **confounder effect.**
- If the correlation between two different variables is zero, it is not necessary that they are independent. This example was explained through the Credit score of people of different ages. In that case, there was no correlation between the two variables, but there was a relationship between them, and they were dependent on each other.

Session Summary

Simple Linear Regression

While covariance and correlation reveal the existence of a linear relationship between two variables, it cannot estimate future values in data.
Simple linear regression helps us to relate two variables with a mathematical equation and allows us to predict one value if the other is known.

Linear Equations

A linear relationship between two variables suggests that when one variable changes, the other will too. Now, the extent to which one variable will affect the other is defined through a mathematical equation, which is known as a linear equation.

Every mathematical equation is not a linear equation.

For instance, equations like **X * Y = 10 or $Y = X^2$**, when plotted on a graph, do not give a straight line.

Only equations that form a straight line on a graph are said to be linear equations and they can be represented in the general form:

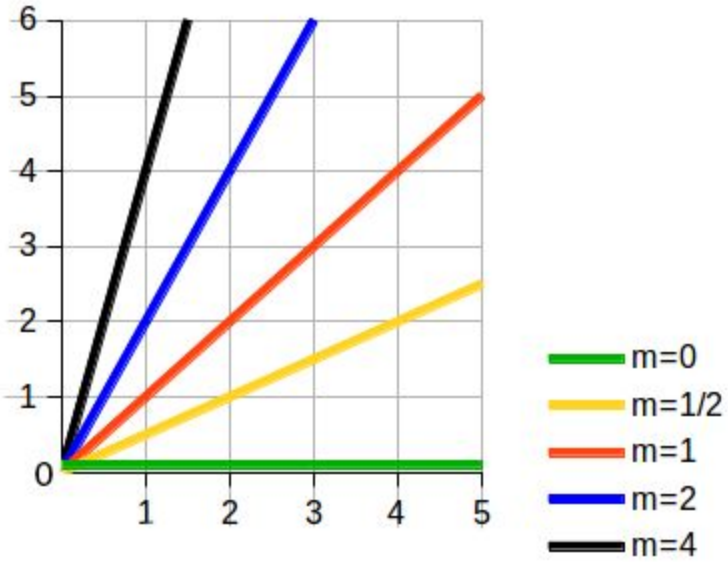
$$Y = mX + C$$

Here, **X** is known as the independent variable, which means it can take any value.

On the other hand, **Y** is known as the dependent variable, which means its value is calculated based on the value of X.

Here, **m** is called the slope of the straight line, or, in other words, **m** is the rate at which **Y** increases with an increase in **X**.

As the value of the slope (m) increases, the steepness of the straight line increases as well.



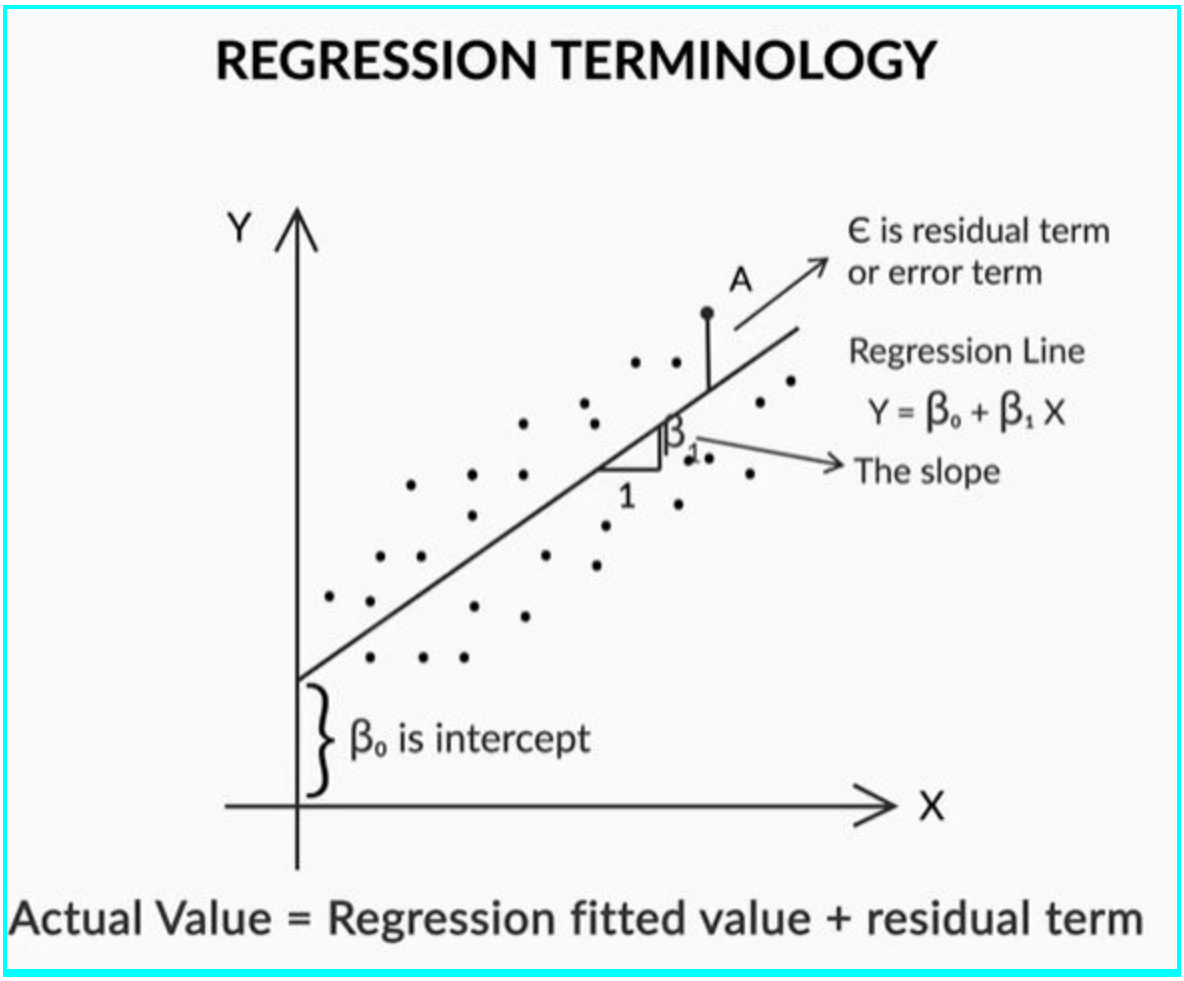
Finally, in the equation **Y = mX + C**, **C** is known as the intercept of the straight line. In other words, it is the value of **Y** when **X** is equal to 0.

Simple Linear Regression

Simple linear regression is a statistical method that is used to study the relationship between two continuous quantitative variables in the form of a linear equation.

Y= mX + C. In this equation:

- Y is known as the **dependent variable**. The value of Y depends on the value of X. If you input different values of X in the equation, then you will get different values of Y. Y is also known as **the response variable** or **the outcome variable** in this equation.
- In the linear equation above, X is known as the independent variable. The value of X is used to determine the value of Y. X is sometimes also called the **predictor variable** or the **explanatory variable**.



All of the data points do not lie on the regression line, and the difference between the data point and the regression line is known as the **residual term or the error term.**

Applying Simple Linear Regression

Hypothesis Testing in Simple Linear Regression

In simple linear regression, a regression model is in the form of a straight line, which can be expressed as:

$Y = \beta_0 + \beta_1 X$

Null Hypothesis

The null hypothesis in our example would be that the independent variable does not affect the dependent variable significantly, and this can be represented mathematically as shown below:

$\beta_1=0$, where β_1 is the coefficient of **X** in the equation $Y = \beta_0 + \beta_1 X$.

Alternative Hypothesis

The alternative hypothesis would be that the independent variable does affect the dependent variable significantly, and this can be represented mathematically as shown below:

$\beta_1 =/= 0$, where β_1 is the coefficient of **X** in the equation $Y = \beta_0 + \beta_1 X$.

R-squared is a metric that is used to evaluate the fit of a simple linear regression model or the straight line with your data set.

The R-squared value ranges from 0 to 1.

An R-squared value of 1 indicates that the regression model completely explains the variation in the data. A higher value of R-squared shows how closely the regression line is passing through the data point.

Usually, in practical situations, an R-squared value above 60% is also considered to represent a good model.

In this session, you also saw how to perform calculate the ANOVA table using Data Analysis ToolPak in Excel.

You also drew insights from data using different values in the ANOVA table.

upGrad

upGrad

Session Summary

Multiple Linear Regression

In simple linear regression, the dependent variable is predicted or explained using a single independent variable. But, in real-life scenarios the variable of our interest may have dependencies on several different variables. Therefore, it is necessary to determine the relationship between all the variables in order to know how a change in one of the variables affects the dependent variable.

Multiple Linear Regression

Multiple Linear Regression is one of the methods to determine the relationship between a dependent variable and more than one independent variables. In this regression, a straight-line form is extended to a linear equation to determine the relationship of between dependent and several independent variables.

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \ldots$$

This linear equation in multiple linear regression is estimated from the sample data using DataAnalysis ToolPak. This tool uses hypothesis testing in order to test the significance of the regression equation and determine the coefficients β_0, β_1, β_2, β_3 and so on …

There are two sets of hypothesis testing in Multiple linear regression:

1. Overall Significance Test

A. **Null Hypothesis(H0):** None of the independent variables has any significant influence on the dependent variable.

   Mathematically, $\beta_1 = \beta_2 = \beta_3 = 0$

B. **Alternative Hypothesis (H0):** At least one independent variable has a significant influence on the dependent variable.

   Mathematically, At least one of the β_1, β_2, β_3 is non-zero

If the null hypothesis of overall significance is failed to reject, then it implies that all the independent variable doesn't have a significant influence on dependent variable i.e. Mathematically, $\beta_1 = \beta_2 = \beta_3 = 0$.

If the null hypothesis of overall significance is rejected, then it implies that at least one of the independent variables has a significant influence on dependent variable i.e. Mathematically, At least one of the □1, □2, □3 is non-zero. Individual significance test helps in determining the coefficients which take a non-zero value and the value of that coefficient.

2. Individual Significance Test: The individual significance test is performed on each independent variable to determine their regression coefficients respectively. The hypothesis for the independent variable X1 is as follows.

A. **Null Hypothesis:** The independent variable X1 doesn't have a significant influence on the dependent variable.

   Mathematically, □1= 0

B. **Alternative Hypothesis:** The independent variable X1 has a significant influence on the dependent variable.

   Mathematically, □1 ≠ 0

The regression coefficient of the independent variable is determined by looking at the p-values of the independent variable.

- If p-value is less than the critical value (depends on the confidence level), Null hypothesis is rejected, which implies independent variable X1 does have a significant influence on dependent variable and □1 ≠ 0 and the value of □1 can be found from the regression coefficient in the regression results. If p-value is greater than the critical value (depends on the confidence level), Null hypothesis is failed to reject, which implies independent variable doesn't have a significant influence on dependent variable and □1 = 0.

Evaluation of Multiple Linear Regression

R-Square: R-square or coefficient of determination is a metric used to quantify the goodness of fit for a regression model.

- R-square explains how close the actual data points are to the fitted regression line.
- R-Square measures the percentage of variability in the dependent variable that is explained by the regression line.

$$\text{R-Square} = \frac{\text{Variation of dependent variable explained by regression line}}{\text{Total variation of dependent variable}}$$

Adjusted R-Square: R-Square does not consider the effect of multiple independent variables. Therefore, it has to be adjusted to the number of independent variables. So, a better measure of goodness of fit in the case of multiple regression is adjusted R-square.