**WEEK 6**

**Task 1 : Advanced Statistics**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 1 : Advanced Statistics](https://dashboard.stige.in/index.php/lessons/task-1-advance-statistics/)

In Progress

Hey there, so we covered statistics & probability in the previous week, in this week we are going to cover advanced statistics. Before that, let’s watch some videos from an Industry standard and widely used platform – Udacity.

First Half Topics:

[Correlation Vs Causation](https://classroom.udacity.com/courses/st101/lessons/f4d4cf4c-a989-4118-b699-fca77c368900/concepts/d24ba322-7946-40c4-ba5d-5c945177dad8)

[Estimation](https://classroom.udacity.com/courses/st101/lessons/40d901fe-f73e-4d68-9aea-83461ffb95c3/concepts/44f741b5-8355-40d0-9d07-a3671d3af0de)

[Outliers](https://classroom.udacity.com/courses/st101/lessons/52c8a6c4-6c3f-4d60-b353-c395fbe6eec1/concepts/ec1eb23a-3213-4330-91d3-1e36725b0ca1)

[Binomial Distribution](https://classroom.udacity.com/courses/st101/lessons/284fa195-e1e0-4aae-af59-b96d7b2f32e1/concepts/df38fe35-2e9c-4227-999b-e3d9fe4a2298)

Second Half Topics:

[Central Limit Theorem](https://classroom.udacity.com/courses/st101/lessons/77d7d7b8-535c-4af1-b989-9f9468e55ccf/concepts/c9c1864c-34c8-42ff-8d5c-264cf80c8d74)

[Normal Distribution](https://classroom.udacity.com/courses/st101/lessons/fa282ec0-c16f-499b-8efb-edda434c6a3f/concepts/dff7bd86-d612-4d88-82bc-b26829508827)

[Manipulating Normal](https://classroom.udacity.com/courses/st101/lessons/2241074a-9b1a-4acd-bec3-5e7c94c5243a/concepts/d23f6ad8-f114-4b5e-9e2a-3f80a07e07dd)

[Most Better than Average](https://classroom.udacity.com/courses/st101/lessons/a642fca4-c868-4771-9024-55024bdf5b29/concepts/1b9ef5bb-97f8-4b32-8344-7681b9d87bd6)

[Confidence Interval](https://classroom.udacity.com/courses/st101/lessons/491fdde1-6f84-4f6c-96fa-63feee15935a/concepts/a540970e-7f0d-4ddc-938f-1930fb424e6f)

[Normal Quantiles](https://classroom.udacity.com/courses/st101/lessons/43b531eb-7452-4712-b9bc-4403e3d8bfb3/concepts/6a956558-b796-4caa-93df-655540a5ad68)

**Task 2 : Inferential Statistics**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/)

In Progress

Statistics is one of the key fundamental skills required for data science. Any expert in data science would surely recommend learning / upskilling yourself in statistics.

However, if you go out and look for resources on statistics, you will see that a lot of them tend to focus on the mathematics. They will focus on derivation of formulas rather than simplifying the concept. I believe, statistics can be understood in very simple and practical manner. That is why I have created this guide.

In this guide, I will take you through Inferential Statistics, which is one of the most important concepts in statistics for data science. I will take you through all the related concepts of Inferential Statistics and their practical applications.

This guide would act as a comprehensive resource to learn Inferential Statistics. So, go through the guide, section by section. Work through the examples and develop your statistics skills for data science.

# Why we need Inferential Statistics?

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Why we need Inferential Statistics?](https://dashboard.stige.in/index.php/topic/why-we-need-inferential-statistics/)

In Progress

Suppose, you want to know the average salary of Data Science professionals in India. Which of the following methods can be used to calculate it?

1. Meet every Data Science professional in India. Note down their salaries and then calculate the total average?
2. Or hand pick a number of professionals in a city like Gurgaon. Note down their salaries and use it to calculate the Indian average.

Well, the first method is not impossible but it would require an enormous amount of resources and time. But today, companies want to make decisions swiftly and in a cost-effective way, so the first method doesn’t stand a chance.

On the other hand, second method seems feasible. But, there is a caveat. What if the population of Gurgaon is not reflective of the entire population of India? There are then good chances of you making a very wrong estimate of the salary of Indian Data Science professionals.

Now, what method can be used to estimate the average salary of all data scientists across India?

#### **Enter Inferential Statistics**

In simple language, Inferential Statistics is used to draw inferences beyond the immediate data available.

With the help of inferential statistics, we can answer the following questions:

* Making inferences about the population from the sample.
* Concluding whether a sample is significantly different from the population. For example, let’s say you collected the salary details of Data Science professionals in Bangalore. And you observed that the average salary of Bangalore’s data scientists is more than the average salary across India. Now, we can conclude if the difference is statistically significant.
* If adding or removing a feature from a model will really help to improve the model.
* If one model is significantly better than the other?
* Hypothesis testing in general.

I am sure by now you must have got a gist of why inferential statistics is important. I will take you through the various techniques & concepts involved in Inferential statistics. But first, let’s discuss what are the prerequisites for understanding Inferential Statistics.

**Pre-requisites**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Pre-requisites](https://dashboard.stige.in/index.php/topic/pre-requisites/)

In Progress

To begin with Inferential Statistics, one must have a good grasp over the following concepts:

1. Probability
2. Basic knowledge of Probability Distributions
3. Descriptive Statistics

If you are not comfortable with either of the three concepts mentioned above, you must go through them before proceeding further.

Throughout the entire article, I will be using a few terminologies quite often. So, here is a brief description of them:

* **Statistic**– A Single measure of some attribute of a sample. For eg: Mean/Median/Mode of a sample of Data Scientists in Bangalore.
* **Population Statistic**– The statistic of the entire population in context. For eg: Population mean for the salary of the entire population of Data Scientists across India.
* **Sample Statistic** – The statistic of a group taken from a population. For eg: Mean of salaries of all Data Scientists in Bangalore.
* **Standard Deviation**– It is the amount of variation in the population data. It is given by σ.
* **Standard Error**– It is the amount of variation in the sample data. It is related to Standard Deviation as σ/√n, where, n is the sample size.

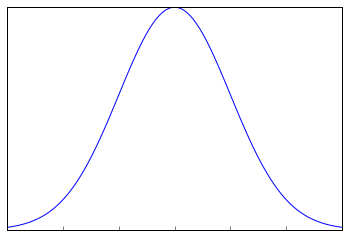
# Sampling Distribution and Central Limit Theorem

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Sampling Distribution and Central Limit Theorem](https://dashboard.stige.in/index.php/topic/sampling-distribution-and-central-limit-theorem/)

In Progress

Suppose, you note down the salary of any 100 random Data Science professionals in Gurgaon, calculate the mean and repeat the procedure for say like 200 times (arbitrarily).

When you plot a frequency graph of these 200 means, you are likely to get a curve similar to the one below.

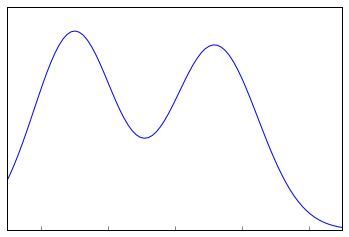


This looks very much similar to the normal curve that you studied in the Descriptive Statistics. This is called Sampling Distribution or the graph obtained by plotting sample means. Let us look at a more formal description of a Sampling Distribution.

A Sampling Distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population.

A Sampling Distribution behaves much like a normal curve and has some interesting properties like :

* The shape of the Sampling Distribution does not reveal anything about the shape of the population. For example, for the above Sampling Distribution, the population distribution may look like the below graph.



Population Distribution

* Sampling Distribution helps to estimate the population statistic.

But how ?

This will be explained using a very important theorem in statistics – **The Central Limit Theorem.**

#### 3.1 Central Limit Theorem

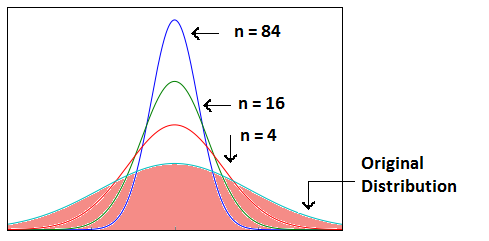
It states that when plotting a sampling distribution of means, the mean of sample means will be equal to the population mean. And the sampling distribution will approach a normal distribution with variance equal to σ/√n where σ is the standard deviation of population and n is the sample size.

Points to note:

1. Central Limit Theorem holds true irrespective of the type of distribution of the population.
2. Now, we have a way to estimate the population mean by just making repeated observations of samples of a fixed size.
3. Greater the sample size, lower the standard error and greater the accuracy in determining the population mean from the sample mean.

This seemed too technical isn’t it? Let’s break this down to understand this point by point.

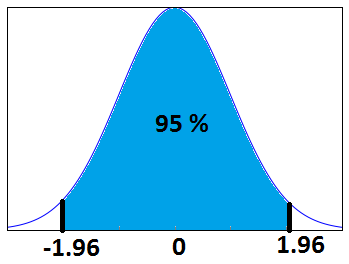
1. This means – No matter the shape of the population distribution, be it bi-modal, right skewed etc. The shape of the Sampling Distribution will remain the same (remember the normal curve- bell shaped). This gives us a mathematical advantage to estimate the population statistic – no matter the shape of the population.
2. The number of samples have to be sufficient (generally more than 50) to satisfactorily achieve a normal curve distribution. Also, care has to be taken to keep the sample size fixed since any change in sample size will change the shape of the sampling distribution and it will no longer be bell shaped.
3. As we increase the sample size, the sampling distribution squeezes from both sides giving us a better estimate of the population statistic since it lies somewhere in the middle of the sampling distribution (generally). The below image will help you visualize the effect of sample size on the shape of distribution.



Now, since we have collected the samples and plotted their means, it is important to know where the population mean lies with respect to a particular sample mean and how confident can we be about it. This brings us to our next topic – **Confidence Interval.**

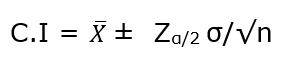
#### 3.2 Confidence Interval

The confidence interval is a type of interval estimate from the sampling distribution which gives a range of values in which the population statistic may lie. Let us understand this with the help of an example.



We know that 95% of the values lie within 2 (1.96 to be more accurate) standard deviation of a normal distribution curve. So, for the above curve, the blue shaded portion represents the confidence interval for a sample mean of 0.

Formally, Confidence Interval is defined as,



whereas,  = the sample mean

= Z value for desired confidence level α

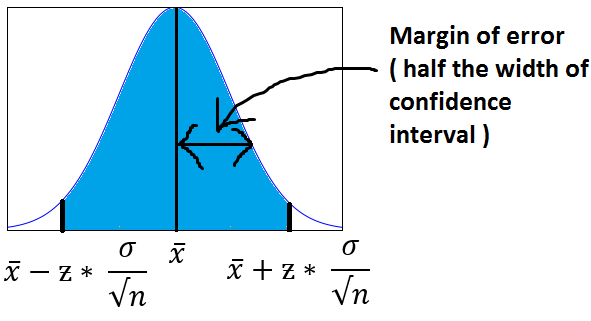
**σ** = the population standard deviation

For an alpha value of 0.95 i.e 95% confidence interval, z=1.96.

Now there is one more term which you should be familiar with, **Margin of Error.** It is given as {(z.σ)/√n} and defined as the sampling error by the surveyor or the person who collected the samples. That means, if a sample mean lies in the margin of error range then, it might be possible that its actual value is equal to the population mean and the difference is occurring by chance. Anything outside margin of error is considered statistically significant.

And it is easy to infer that the error can be both positive and negative side. The whole margin of error on both sides of the sample statistic constitutes the Confidence Interval. Numerically, C.I is twice of Margin of Error.

The below image will help you better visualize Margin of Error and Confidence Interval.



The shaded portion on horizontal axis represents the Confidence Interval and half of it is Margin of Error which can be in either direction of x (bar).

Interesting points to note about Confidence Intervals:

1. Confidence Intervals can be built with difference degrees of confidence suitable to a user’s needs like 70 %, 90% etc.
2. Greater the sample size, smaller the Confidence Interval, i.e more accurate determination of population mean from the sample means.
3. There are different confidence intervals for different sample means. For example, a sample mean of 40 will have a difference confidence interval from a sample mean of 45.
4. By 95% Confidence Interval, we do not mean that – The probability of a population mean to lie in an interval is 95%. Instead, 95% C.I means that 95% of the Interval estimates will contain the population statistic.

Many people do not have right knowledge about confidence interval and often interpret it incorrectly. So, I would like you to take your time visualizing the 4th argument and let it sink in.

#### 3.3 Practical example

Calculate the 95% confidence interval for a sample mean of 40 and sample standard deviation of 40 with sample size equal to 100.

**Solution:**

We know, z-value for 95% C.I is 1.96. Hence, Confidence Interval (C.I) is calculated as:

C.I= [{x(bar) – (z\*s/√n)},{x(bar) – (z\*s/√n)}]

C.I = [{40-(1.96\*40/10},{ 40+(1.96\*40/10)}]

C.I = [32.16, 47.84]

**Hypothesis Testing**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Hypothesis Testing](https://dashboard.stige.in/index.php/topic/hypothesis-testing/)

In Progress

Before I get into the theoretical explanation, let us understand Hypothesis Testing by using a simple example.

**Example:** Class 8th has a mean score of 40 marks out of 100. The principal of the school decided that extra classes are necessary in order to improve the performance of the class. The class scored an average of 45 marks out of 100 after taking extra classes. Can we be sure whether the increase in marks is a result of extra classes or is it just random?

Hypothesis testing lets us identify that. It lets a sample statistic to be checked against a population statistic or statistic of another sample to study any intervention etc. Extra classes being the intervention in the above example.

Hypothesis testing is defined in two terms – **Null Hypothesis** and **Alternate Hypothesis**.

* **Null Hypothesis** being the sample statistic to be equal to the population statistic. For eg: The Null Hypothesis for the above example would be that the average marks after extra class are same as that before the classes.
* **Alternate Hypothesis** for this example would be that the marks after extra class are significantly different from that before the class.

Hypothesis Testing is done on different levels of confidence and makes use of z-score to calculate the probability. So for a 95% Confidence Interval, anything above the z-threshold for 95% would reject the null hypothesis.

Points to be noted:

1. We cannot accept the Null hypothesis, only reject it or fail to reject it.
2. As a practical tip, Null hypothesis is generally kept which we want to disprove. For eg: You want to prove that students performed better after taking extra classes on their exam. The Null Hypothesis, in this case, would be that the marks obtained after the classes are same as before the classes.

**Types of Error in Hypothesis Testing**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Types of Error in Hypothesis Testing](https://dashboard.stige.in/index.php/topic/types-of-error-in-hypothesis-testing/)

In Progress

Now we have defined a basic Hypothesis Testing framework. It is important to look into some of the mistakes that are committed while performing Hypothesis Testing and try to classify those mistakes if possible.

Now, look at the Null Hypothesis definition above. What we notice at the first look is that it is a statement subjective to the tester like you and me and not a fact. That means there is a possibility that the Null Hypothesis can be true or false and we may end up committing some mistakes on the same lines.

There are two types of errors that are generally encountered while conducting Hypothesis Testing.

* **Type I error**: Look at the following scenario – A male human tested positive for being pregnant. Is it even possible? This surely looks like a case of False Positive. More formally, it is defined as the incorrect rejection of a True Null Hypothesis. The Null Hypothesis, in this case, would be – Male Human is not pregnant.
* **Type II error**: Look at another scenario where our Null Hypothesis is – A male human is pregnant and the test supports the Null Hypothesis.  This looks like a case of False Negative. More formally it is defined as the acceptance of a false Null Hypothesis.

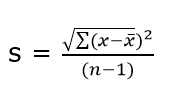
**T-tests**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [T-tests](https://dashboard.stige.in/index.php/topic/t-tests/)

In Progress

T-tests are very much similar to the z-scores, the only difference being that instead of the Population Standard Deviation, we now use the Sample Standard Deviation. The rest is same as before, calculating probabilities on basis of t-values.

The Sample Standard Deviation is given as:



where n-1 is the Bessel’s correction for estimating the population parameter.

Another difference between z-scores and t-values are that t-values are dependent on Degree of Freedom of a sample. Let us define what degree of freedom is for a sample.

**The Degree of Freedom –** It is the number of variables that have the choice of having more than one arbitrary value. For example, in a sample of size 10 with mean 10, 9 values can be arbitrary but the 1oth value is forced by the sample mean.

Points to note about the t-tests:

1. Greater the difference between the sample mean and the population mean, greater the chance of rejecting the Null Hypothesis. Why? (We discussed this above.)
2. Greater the sample size, greater the chance of rejection of Null Hypothesis.

# Different types of t-test

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Different types of t-test](https://dashboard.stige.in/index.php/topic/different-types-of-t-test/)

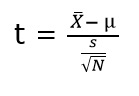
In Progress

#### 7.1 1-sample t-test

This is the same test as we described above. This test is used to:

* Determine whether the mean of a group differs from the specified value.
* Calculate a range of values that are likely to include the population mean.

For eg: A pizza delivery manager may perform a 1-sample t-test whether their delivery time is significantly different from that of the advertised time of 30 minutes by their competitors.



where, **X(bar)** = sample mean

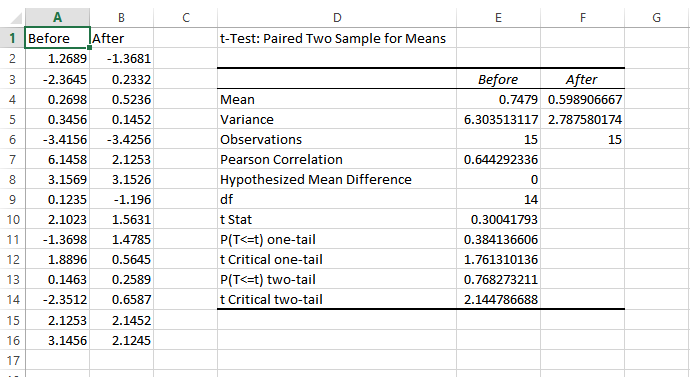
**μ** = population mean

**s** = sample standard deviation

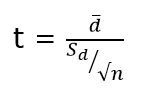
**N** = sample size

#### 7.2 Paired t-test

Paired t-test is performed to check whether there is a difference in mean after a treatment on a sample in comparison to before. It checks whether the Null hypothesis: The difference between the means is Zero, can be rejected or not.



The above example suggests that the Null Hypothesis should not be rejected and that there is no significant difference in means before and after the intervention since p-value is not less than the alpha value (o.o5) and t stat is not less than t-critical. The excel sheet for the above exercise is available [here](https://drive.google.com/open?id=0ByAvlBzuj2TgQ0M0U2lVZnZuams).



where, **d (bar)** = mean of the case wise difference between before and after,

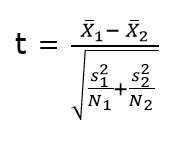
= standard deviation of the difference

**n** = sample size.

#### 7.3 2-sample t-test

This test is used to determine:

* Determine whether the means of two independent groups differ.
* Calculate a range of values that is likely to include the difference between the population means.



The above formula represents the 2 sample t-test and can be used in situations like to check whether two machines are producing the same output. The points to be noted for this test are:

1. The groups to be tested should be independent.
2. The groups’ distribution should not be highly skewed.

where, **X1 (bar)** = mean of the first group

**=**represents1st group sample standard deviation

= represents the 1st group sample size.

#### 7.4 Practical example

We will understand how to identify which t-test to be used and then proceed on to solve it. The other t-tests will follow the same argument.

**Example:** A population has mean weight of 68 kg. A random sample of size 25 has a mean weight of 70 with standard deviation =4. Identify whether this sample is representative of the population?

##### **Step 0: Identifying the type of t-test**

Number of samples in question = 1

Number of times the sample is in study = 1

Any intervention on sample = No

Recommended t-test = 1- sample t-test.

Had there been 2 samples, we would have opted for 2-sample t-test and if there would have been 2 observations on the same sample, we would have opted for paired t-test.`

##### **Step 1: State the Null and Alternate Hypothesis**

**Null Hypothesis:**The sample mean and population mean are same.

**Alternate Hypothesis:** The sample mean and population mean are different.

##### **Step 2: Calculate the appropriate test statistic**

df = 25-1 =24

t= (70-68)/(4/√25) = 2.5

Now, for a 95% confidence level, t-critical (two-tail) for rejecting Null Hypothesis for 24 d.f is 2.06 . Hence, we can reject the Null Hypothesis and conclude that the two means are different.

You can use the t-test calculator [here](http://www.danielsoper.com/statcalc/calculator.aspx?id=98).

# ANOVA

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [ANOVA](https://dashboard.stige.in/index.php/topic/anova/)

In Progress

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises – Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

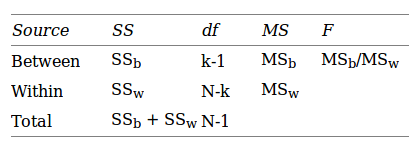
The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

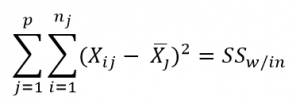
To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels. However, ANOVA procedures work quite well even if the normality assumption has been violated unless one or more of the distributions are highly skewed or if the variances are quite different.

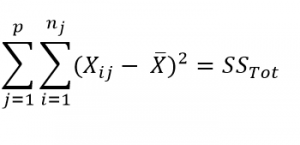
ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).

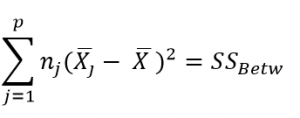
Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups)

Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)









Here, **p** = represents the number of groups

**n =**represents the number of observations in a group

=  represents the mean of a particular group

**X (bar)** = represents the mean of all the observations

Now, let us understand the degree of freedom for within group and between groups respectively.

Between groups : If there are k groups in ANOVA model, then k-1 will be independent. Hence, k-1 degree of freedom.

Within groups : If N represents the total observations in ANOVA (∑n over all groups) and k are the number of groups then, there will be k fixed points. Hence, N-k degree of freedom.

#### **8.1 Steps to perform ANOVA**

1. Hypothesis Generation
   1. Null Hypothesis : Means of all the groups are same
   2. Alternate Hypothesis : Mean of at least one group is different
2. Calculate within group and between groups variability
3. Calculate F-Ratio
4. Calculate probability using F-table
5. Reject/fail to Reject Null Hypothesis

There are various other forms of ANOVA too like Two-way ANOVA, MANOVA, ANCOVA etc. but One-Way ANOVA suffices the requirements of this course.

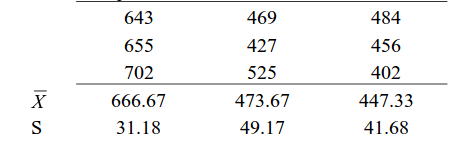
Practical applications of ANOVA in modeling are:

1. Identifying whether a categorical variable is relevant to a continuous variable.
2. Identifying whether a treatment was effective to the model or not.

#### **8.2 Practical Example**

Suppose there are 3 chocolates in town and their sweetness is quantified by some metric (S). Data is collected on the three chocolates. You are given the task to identify whether the mean sweetness of the 3 chocolates are different. The data is given as below:

                                                                 Type A                    Type B                   Type C



Here, first we have calculated the sample mean and sample standard deviation for you.

Now we will proceed step-wise to calculate the F-Ratio (ANOVA statistic).

#### Step 1: Stating the Null and Alternate Hypothesis

**Null Hypothesis:**Mean sweetness of the three chocolates are same.

**Alternate Hypothesis:**Mean sweetness of at least one of the chocolates is different.

#### Step 2: Calculating the appropriate ANOVA statistic

In this part, we will be calculating SS(B), SS(W), SS(T) and then move on to calculate MS(B) and MS(W). The thing to note is that,

Total Sum of Squares [SS(t)] = Between Sum of Squares [SS(B)] + Within Sum of Squares [SS(W)].

So, we need to calculate any two of the three parameters using the data table and formulas given above.

As, per the formula above, we need one more statistic i.e Grand Mean denoted by X(bar) in the formula above.

**X bar** = (643+655+702+469+427+525+484+456+402)/9 = 529.22

**SS(B)**=[3\*(666.67-529.22)^2]+ [3\*(473.67-529.22)^2]+[3\*(447.33-529.22)^2] = 86049.55

**SS (W)** = [(643-666.67)^2+(655-666.67)^2+(702-666.67)^2] + [(469-473.67)^2+(427-473.67)^2+(525-473.67)^2] + [(484-447.33)^2+(456-447.33)^2+(402-447.33)^2]= 10254

**MS(B)** = SS(B) / df(B) = 86049.55 / (3-1) = 43024.78

**MS(W)** = SS(W) / df(W) = 10254/(9-3) = 1709

**F-Ratio** = MS(B) / MS(W) = 25.17 .

Now, for a 95 % confidence level, F-critical to reject Null Hypothesis for degrees of freedom(2,6) is 5.14 but we have 25.17 as our F-Ratio.

So, we can confidently reject the Null Hypothesis and come to a conclusion that at least one of the chocolate has a mean sweetness different from the others.

You can use the F-calculator [here](http://stattrek.com/online-calculator/f-distribution.aspx).

**Note:** ANOVA only lets us know the means for different groups are same or not. It doesn’t help us identify which mean is different.To know which group mean is different, we can use another test know as Least Significant Difference Test.

# Chi-Square Goodness of Fit

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Chi-Square Goodness of Fit](https://dashboard.stige.in/index.php/topic/chi-square-goodness-of-fit/)

In Progress

Sometimes, the variable under study is not a continuous variable but a categorical variable. Chi-square test is used when we have one single categorical variable from the population.

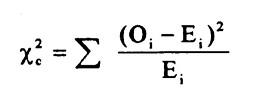
Let us understand this with help of an example. Suppose a company that manufactures chocolates, states that they manufacture 30% dairy milk, 60% temptation and 10% kit-kat. Now suppose a random sample of 100 chocolates has 50 dairy milk, 45 temptation and 5 kitkats. Does this support the claim made by the company?

Let us state our Hypothesis first.

Null Hypothesis: The claims are True

Alternate Hypothesis: The claims are False.

Chi-Square Test is given by:



where, = sample or observed values

= population values

The summation is taken over all the levels of a categorical variable.

= **[n \* ]**  Expected value of a level (i) is equal to the product of sample size and percentage of it in the population.

Let us now calculate the Expected values of all the levels.

E (dairy milk)= 100 \* 30% = 30

E (temptation) = 100 \* 60% =60

E (kitkat) = 100 \* 10% = 10

Calculating chi-square = [(50-30)^2/30+(45-60)^2/60+(5-10)^2/10] =19.58

Now, checking for p (chi-square >19.58) using [chi-square calculator](http://stattrek.com/online-calculator/chi-square.aspx), we get p=0.0001. This is significantly lower than the alpha(0.05).

So we reject the Null Hypothesis.

# Regression and ANOVA

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Regression and ANOVA](https://dashboard.stige.in/index.php/topic/regression-and-anova/)

In Progress

If you have studied some basic Machine Learning Algorithms, the first algorithm that you must have studied is Regression. If we  recall those lessons of Regression, what we generally do is calculate the weights for features present in the model to better predict the output variable. But finding the right set of feature weights or features for that matter is not always possible.

It is highly likely that that the existing features in the model are not fit for explaining the trend in dependent variable or the feature weights calculated fail at explaining the trend in dependent variable. What is important is knowing the degree to which our model is successful in explaining the trend (variance) in dependent variable.

Enter ANOVA.

With the help of ANOVA techniques, we can analyse a model performance very much like we analyse samples for being statistically different or not.

But with regression things are not easy. We do not have mean of any kind to compare  or sample as such but we can find good alternatives in our regression model which can substitute for mean and sample.

Sample in case of regression is a regression model itself with pre-defined features and feature weights whereas mean is replaced by variance(of both dependent and independent variables).

Through our ANOVA test we would like to know the amount of variance explained by the Independent variables in Dependent Variable VS the amount of variance that was left unexplained.

It is intuitive to see that larger the unexplained variance(trend) of the dependent variable smaller will be the ratio and less effective is our regression model. On the other hand, if we have a large explained variance then it is easy to see that our regression model was successful in explaining the variance in the dependent variable and more effective is our model. The ratio of Explained Variance uand Unexplained Variance is called F-Ratio.

Let us now define these explained and unexplained variances to find the effectiveness of our model.

**1. Regression (Explained) Sum of Squares** – It is defined as the amount of variation explained by the Regression model in the dependent variable.

Mathematically, it is calculated as:

where, [hat] = predicted value and

**y(bar)** = mean of the actual y values.

**Interpreting Regression sum of squares –**

If our model is a good model for the problem at hand then it would produce an output which has distribution as same to the actual dependent variable. i.e it would be able to capture the inherent variation in the dependent variable.

**2. Residual Sum of Squares** – It is defined as the amount of variation independent variable which is not explained by the Regression model.

Mathematically, it is calculated as:

where,  = actual ‘y ‘ value

**f(x)** = predicted value

**Interpretation of Residual Sum of Squares –**

It can be interpreted as the amount by which the predicted values deviated from the actual values. Large deviation would indicate that the model failed at predicting the correct values for the dependent variable.

Let us now  work out F-ratio step by step. We will be making using of the Hypothesis Testing framework described above to test the significance of the model.

While calculating the F-Ratio care has to be taken to incorporate the effect of degree of freedom. Mathematically, F-Ratio is the ratio of **[Regression Sum of Squares/df(regression)] and [Residual Sum of Squares/df(residual)].**

We will be understanding the entire concept using an example and [this excel sheet](https://drive.google.com/file/d/0ByAvlBzuj2TgV0RDM0FORkQ3YW8/view).

#### Step 0: State the Null and Alternate Hypothesis

Null Hypothesis: The model is unable to explain the variance in the dependent variable (Y).

Alternate Hypothesis: The model is able to explain the variance in dependent variable (Y)

#### Step 1:

Calculate the regression equation for X and Y using Excel’s in-built tool.

#### Step 2:

Predict the values of y for each row of data.

#### Step 3:

Calculate y(mean) – mean of the actual y values which in this case turns out to be 0.4293548387.

#### Step 4:

Calculate the Regression Sum of Squares using the above-mentioned formula. It turned out to be 2.1103632473

The Degree of freedom for regression equation is 1, since we have only 1 independent variable.

#### Step 5:

Calculate the Residual Sum of Squares using the above-mentioned formula. It turned out to be 0.672210946.

Degree of Freedom for residual = Total degree of freedom – Degree of freedom(regression)

=(62-1) – 1 = 60

#### Step 6:

F-Ratio = (2.1103632473/1)/(0.672210946/60) = 188.366

Now, for 95% confidence, F-critical to reject Null Hypothesis for 1,60 degrees of freedom in 4. But we have F-ratio as 188, so we can safely reject the Null Hypothesis and conclude that model explains variation to a large extent.

# Coefficient of Determination (R-Squared)

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Coefficient of Determination (R-Squared)](https://dashboard.stige.in/index.php/topic/coefficient-of-determination-r-squared/)

In Progress

It is defined as the ratio of the amount of variance explained by the regression model to the total variation in the data. It represents the strength of correlation between two variables.

We already calculated the Regression SS and Residual SS. Total SS is the sum of Regression SS and Residual SS.

Total SS = 2.1103632473+ 0.672210946 = 2.78257419

Co-efficient of Determination = 2.1103632473/2.78257419 = 0.7588

# Correlation Coefficient

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 2 : Inferential Statistics](https://dashboard.stige.in/index.php/lessons/task-2-inferential-statistics/) [Correlation Coefficient](https://dashboard.stige.in/index.php/topic/correlation-coefficient/)

In Progress

This is another useful statistic which is used to determine the correlation between two variables. It is simply the square root of coefficient of Determination and ranges from -1 to 1 where 0 represents no correlation and 1 represents positive strong correlation while -1 represents negative strong correlation.

# Task 3 : Hypothesis Testing

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/)

In Progress

Gathering data in itself is meaningless unless we can analyze it and draw powerful insights. What makes data interesting is the ability to evaluate and interpret it.

Hypothesis testing refers to a term in statistics where we, as the analysts, evaluate an assumption related to a data set parameter.

Based on the purpose of the analysis and the specific characteristics of the data, we can use different methodologies. In general, the technique gives us a standardized way to assess the plausibility of an assumption based on sample data.

This sample data can originate either from a larger population or a data-generating process.

**What is a Hypothesis**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [What is a Hypothesis](https://dashboard.stige.in/index.php/topic/what-is-a-hypothesis/)

In Progress

Essentially, it is an educated guess, which we can test with observations or by experimenting. It can be anything, so long as it is testable.

When we propose a hypothesis, we write a hypothesis statement.

Generally, we strive to keep this in the form of ‘If… then…’. More specifically,

*‘If A happens to an independent variable, then B will happen to the dependent variable.’*

There are some characteristics to a well-written statement:

* As mentioned, it’s an if-then statement;
* We can test the statement scientifically;
* It states both the independent and dependent variables.

First, we define the problem we are analyzing, and then we base our hypothesis statement on this problem.

It is crucial to remember that the underlying assumption can be about any parameter of the population, and it can either be true or not.

**What is Hypothesis Testing**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [What is Hypothesis Testing](https://dashboard.stige.in/index.php/topic/what-is-hypothesis-testing/)

In Progress

The best way to evaluate a hypothesis would be to review the entire population of the data we are analyzing. However, this usually proves to be highly impractical, if not wholly impossible. Therefore, we typically assess only a randomly selected sample instead of the entire population.

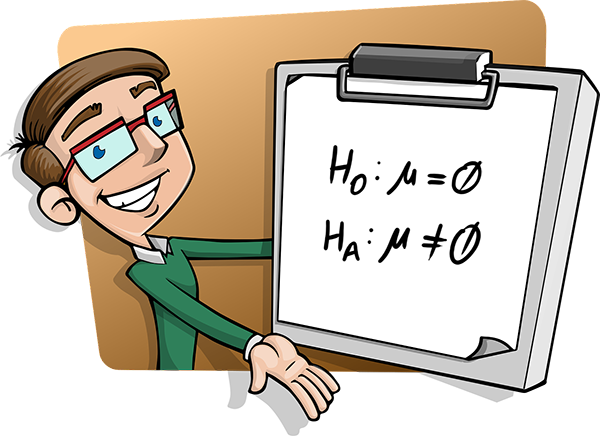
And if the data within the sample is not consistent with our hypothesis, we can reject it.

When we perform statistical analysis, we test a hypothesis by evaluating a random sample of the entire population. Practically, we test two hypotheses:

* The null hypothesis (H0)
* The alternative hypothesis (HA)

The null hypothesis usually assumes equality in parameters of the population, like the mean of the population is equal to zero. The alternative hypothesis then will be the exact opposite — the mean does not equal zero.

The null and alternative hypotheses have to be mutually exclusive. Only one can be correct, but one of the two is always right.



The null hypothesis is usually the accepted fact — the mean equals zero, smoking causes cancer, loud music hurts your ears, and others. When we look at the randomly selected sample, we usually consider that the null hypothesis is that the observations are simply the result of chance. The alternative view is then that they are affected by a non-random cause.

**The Four Steps of Hypothesis Testing**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [The Four Steps of Hypothesis Testing](https://dashboard.stige.in/index.php/topic/the-four-steps-of-hypothesis-testing/)

In Progress

We can present the process of data-driven decision making in four steps:

1. State the two hypotheses (null and alternative) in a way that only one can be true;
2. Plan how to evaluate the data and prepare the analysis plan, outlining how we will use the sample to assess the population. It is common to focus on a single parameter (e.g., mean, standard deviation, p-value, z-score, and others);
3. Evaluate the sample data and calculate the value of the test statistics, as described in the analysis plan;
4. Assess the results by applying the decision rules from the plan. Here we either accept the null hypothesis as plausible or reject it in favor of the alternative hypothesis

# Decision Rules

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Decision Rules](https://dashboard.stige.in/index.php/topic/decision-rules/)

In Progress

One of the most important things we need to define in our analysis plan is the set of decision rules for rejecting the null hypothesis, to be used in our assessment. In practice, these can be specified in two ways –referring to either a p-value or a region of acceptance.

A p-value measures the significance of the evidence in support of the null hypothesis. It is the probability of observing the test statistic with the assumption that the null is true. If the p-value is less than the significance level (our threshold), we reject the null hypothesis.

An acceptance region is a set of values for the test statistic. If it falls within those values, we fail to reject the null hypothesis. And values outside the region of acceptance fall within the region of rejection. If our test statistic ends up here, we reject the null. We can then say that we reject the null hypothesis at the **α**level of significance.

# Accepting or Failing to Reject

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Accepting or Failing to Reject](https://dashboard.stige.in/index.php/topic/accepting-or-failing-to-reject/)

In Progress

The testing has one of two outcomes — we accept the null hypothesis, or we reject it. However, most statisticians prefer to say they reject the null or fail to reject it instead of accepting it.

The idea behind is that saying we accept the null hypothesis means we deem it to be true while saying we fail to reject means we did not find the data to be persuasive enough to select the alternative over the null. Because we are performing a probabilistic test, there’s always a small chance of being wrong, and this different wording covers that.

# Errors

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Errors](https://dashboard.stige.in/index.php/topic/errors/)

In Progress

When we evaluate a hypothesis, we can end up with one of two types of errors:

### **Type I**

This is when we reject the null hypothesis, but it is true. The probability of making a Type I error is the significance level, also called alpha (denoted **α**). In financial modeling and analysis, we would usually set alpha at 5% or 0.05. A smaller alpha (like 1%, or 0.1%) suggests a more robust evaluation of the null hypothesis.

### **Type II**

We make this error when we fail to reject the null hypothesis, but it is false. The probability of making a Type II error is the beta, denoted **β**. On the other hand, the chance of not making such an error is called the **power** of the test.

# Interpreting the Results

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Interpreting the Results](https://dashboard.stige.in/index.php/topic/interpreting-the-results/)

In Progress

We evaluate the p-value to portrait a finding as statistically significant by comparing the value of the statistical test to the predefined alpha level. If the p-value is less than the predefined threshold, then it has statistical significance.

From the perspective of hypothesis testing, if the p-value is less than (or equal to) the alpha, we reject the null hypothesis (significant result). If the p-value is higher than the alpha, we fail to reject the null (insignificant result).

The confidence level of the hypothesis for the observed data can be calculated as one minus alpha (1 — α). Knowing this, we have two ways to write up our conclusions.

* Fail to reject the null hypothesis at a 5% significance level; or
* Fail to reject the null hypothesis at a 95% confidence level.

When we interpret the p-value, it does not mean the null is true or false. It only means we have chosen to reject (or fail to reject) the null hypothesis at a specific confidence level based on the sample observations of the data. We cannot make binary decisions as we only rely on a probabilistic approach.

### Critical Values

Instead of p-values, some tests may return a list of critical values, with their respective significance levels, and also a test statistic. We usually get such results in distribution-free hypothesis testing. However, the choice between p-value and critical values happens as part of the initial test design.

We similarly assess them by comparing the test statistic to the critical value at a chosen significance level. If the test statistic is higher than (or equal to) the critical value, we reject the null hypothesis. And conversely, if the test statistic is less than the critical value, we fail to reject the null.

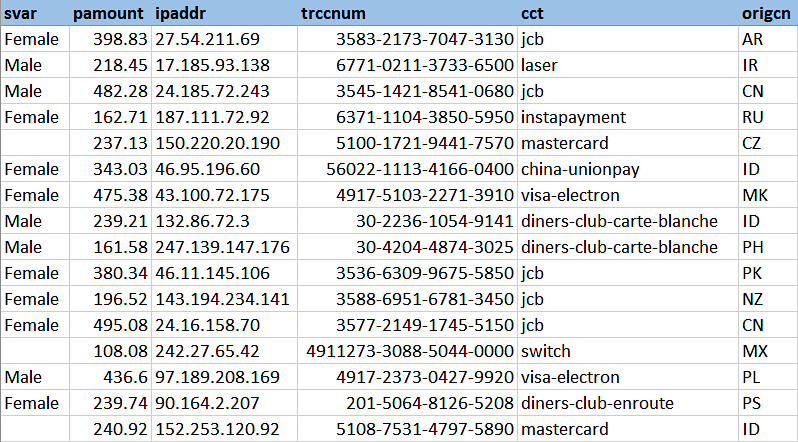
We present the results in the same way as with p-values.

# Example

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Example](https://dashboard.stige.in/index.php/topic/example/)

In Progress

You can take a look at the following example to illustrate the concept better. The breakdown contains a sample of 1000 rows from our client’s online website sales. The extract has various information, but we will focus mainly on **svar**, showing us if the client was Male, Female, or not specified (empty cells) and the purchase value, **pamount**.



The new sales director that our client hired is sure that statistically speaking, there’s no difference between male and female spending and believes the marketing targets should not reflect gender. However, our client has a long experience in the business and believes that men generally spend more on tech. This should be reflected by allocating a higher portion of the marketing budget towards advertising to men.

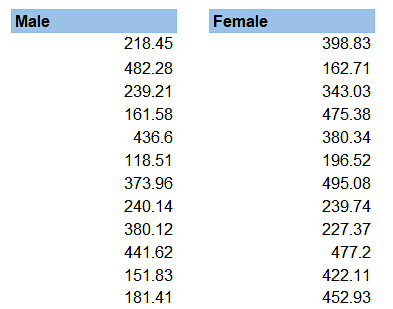
They have asked you to resolve their argument and suggest how to proceed.

Let’s start by writing our hypothesis statement:

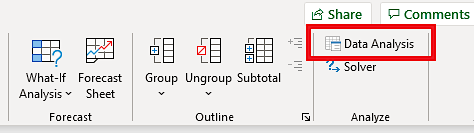
If an online tech shopper is a man, then their purchase value will be different (higher).

We can then denote our null hypothesis as the difference in means for both observation sets (Male and Female) is equal to zero. And the alternative hypothesis will be the exact opposite — the difference in the means is not zero, meaning the average purchase values for both sets are statistically different.

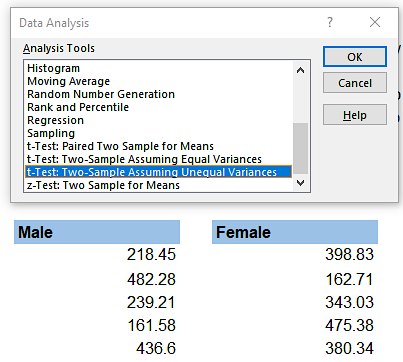
First, we can separate our observations into two groups, as we will be comparing the mean (average) purchase value by Males versus Females.



Now, let’s run the Data Analysis menu from the Data tab on the Ribbon:

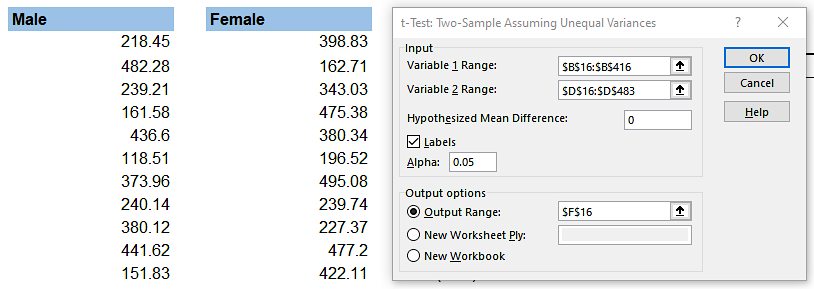


From the menu, select t-Test: Two-Sample Assuming Unequal Variances (as we don’t know the variances of the total populations):



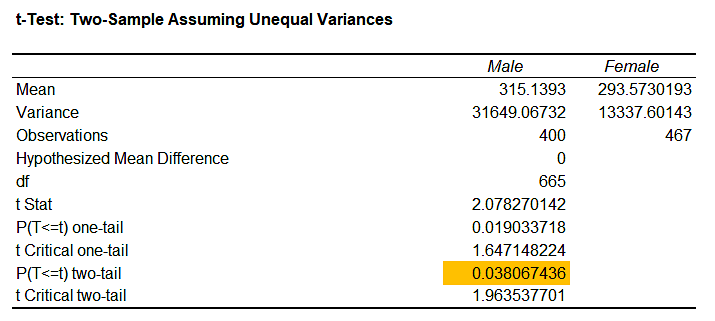
A dialog box opens up, where you will have to select the ranges for both variables (we now treat Males and Females as separate variables). Do not forget the check the Labels box if you select the headers of your data. The hypothesized mean difference will be zero, as per our null hypothesis.

In 99% of the cases, you can leave alpha at 0.05, as is, and you’ll be fine.

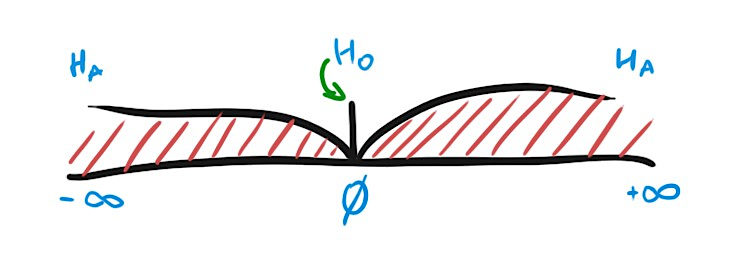


Now run the analysis, and let’s look at the results.

The t-Test gives us some statistical information, but we are most interested in the two-tail p-value.



But why do we pick two-tail and not one-tail? Well, our null hypothesis is that the difference in means is = to zero. Then if we reject the null, this means the difference can be either above zero or below zero. Because of this, we pick the two-tail p-value. The best way to fully understand this concept is to draw a simple chart with two ‘tails’ for the alternative hypothesis. The red areas form the rejection region, and the green arrow points to the single value that is our acceptance region.



Therefore, we use the two-tail p-value of 0.038 and compare it to the alpha level of 0.05. The p-value is less than the level of significance (alpha), which means it is of statistical importance. We can reject the null hypothesis at the 95% confidence level. In conclusion, the data suggest that the difference between the average purchase value by males and females is significant. Therefore, it might be a good idea to do marketing separately.

Keep in mind the test result does not tell us if men spend more or less, just that there’s a statistically meaningful difference. Looking at our t-Test above, we see the mean for the Male variable is higher, which suggests men spend more on tech than women.

You can [download the example model in Excel](https://magnimetrics.com/hypothesis-testing-for-complete-beginners/) in the original article.

# Conclusion

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 3 : Hypothesis Testing](https://dashboard.stige.in/index.php/lessons/task-3-hypothesis-testing/) [Conclusion](https://dashboard.stige.in/index.php/topic/conclusion-4/)

In Progress

Hypothesis testing is the process of evaluating a hypothesis either to reject it or to fail to reject it. It is important to remember that in this analysis, we rely on a probability-based approach, so we can never be 100% certain that our hypothesis is true or false. Coming up with a well-written hypothesis statement is one of the essential tasks of the process. It ensures that what we are testing relates to the problem we are analyzing.

# Task 4 : Hypothesis Testing Advanced

Video - <https://youtu.be/8JIe_cz6qGA>

# Task 5 : Sampling Techniques

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/)

In Progress

**Sampling** helps a lot in research. It is one of the most important factors which determines the accuracy of your research/survey result. If anything goes wrong with your sample then it will be directly reflected in the final result. There are lot of techniques which help us to gather sample depending upon the need and situation. This blog post tries to explain some of those techniques.

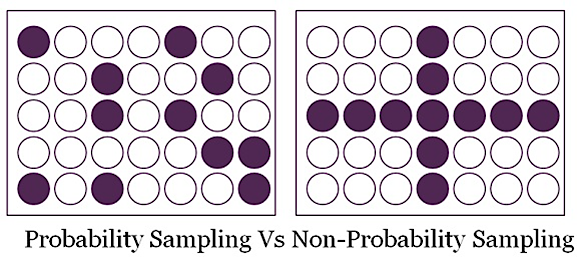
**Sampling Techniques Categories**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Sampling Techniques Categories](https://dashboard.stige.in/index.php/topic/sampling-techniques-categories/)

In Progress

There are lot of sampling techniques which are grouped into two categories as

* Probability Sampling
* Non- Probability Sampling



The difference lies between the above two is whether the sample selection is based on randomization or not. With randomization, every element gets equal chance to be picked up and to be part of sample for study.

# Probability Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Probability Sampling](https://dashboard.stige.in/index.php/topic/probability-sampling/)

In Progress

This Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It’s alternatively known as random sampling.

Simple Random Sampling

Stratified sampling

Systematic sampling

Cluster Sampling

Multi stage Sampling

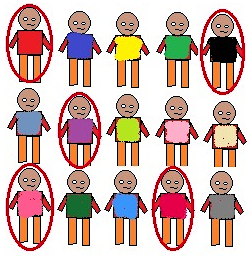
# Simple Random Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Simple Random Sampling](https://dashboard.stige.in/index.php/topic/simple-random-sampling/)

In Progress

Every element has an equal chance of getting selected to be the part sample. It is used when we don’t have any kind of prior information about the target population.

**For example:** Random selection of 20 students from class of 50 student. Each student has equal chance of getting selected. Here probability of selection is 1/50

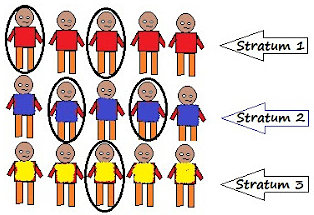
Single Random Sampling

# Stratified sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Stratified sampling](https://dashboard.stige.in/index.php/topic/stratified-sampling/)

In Progress

This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed. And then the elements are randomly selected from each of these strata. We need to have prior information about the population to create subgroups.

Stratified Sampling

# Systematic sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Systematic sampling](https://dashboard.stige.in/index.php/topic/systematic-sampling/)

In Progress

Here the selection of elements is systematic and not random except the first element. Elements of a sample are chosen at regular intervals of population. All the elements are put together in a sequence first where each element has the equal chance of being selected.

For a sample of size n, we divide our population of size N into subgroups of k elements.

We select our first element randomly from the first subgroup of k elements.

To select other elements of sample, perform following:

We know number of elements in each group is k i.e N/n

So if our first element is n1 then

Second element is n1+k i.e n2

Third element n2+k i.e n3 and so on..

Taking an example of N=20, n=5

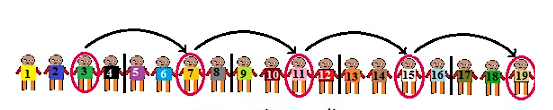
No of elements in each of the subgroups is N/n i.e 20/5 =4= k

Now, randomly select first element from the first subgroup.

If we select n1= 3

n2 = n1+k = 3+4 = 7

n3 = n2+k = 7+4 = 11

Systematic Clustering

# Cluster Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Cluster Sampling](https://dashboard.stige.in/index.php/topic/cluster-sampling/)

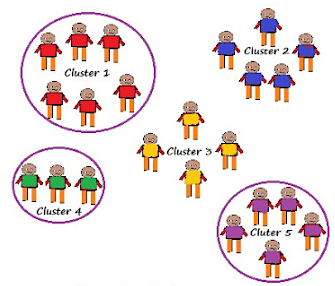
In Progress

Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc.

Cluster sampling can be done in following ways:

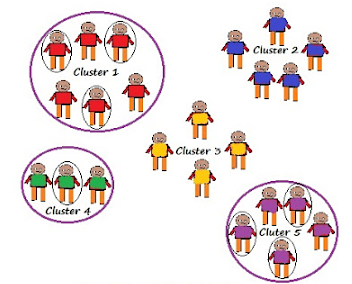
### **Single Stage Cluster Sampling**

Entire cluster is selected randomly for sampling.

 Single Stage Cluster Sampling

### **Two Stage Cluster Sampling**

Here first we randomly select clusters and then from those selected clusters we randomly select elements for sampling

Two Stage Cluster Sampling

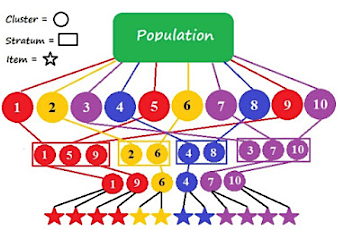
# Multi stage Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Multi stage Sampling](https://dashboard.stige.in/index.php/topic/multi-stage-sampling/)

In Progress

It is the combination of one or more methods described above.

Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity. One or more clusters can be randomly selected from each stratum. This process continues until the cluster can’t be divided anymore. For example country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.

Multi-Stage Sampling

# Non-Probability Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Non-Probability Sampling](https://dashboard.stige.in/index.php/topic/non-probability-sampling/)

In Progress

It does not rely on randomization. This technique is more reliant on the researcher’s ability to select elements for a sample. Outcome of sampling might be biased and makes difficult for all the elements of population to be part of the sample equally. This type of sampling is also known as non-random sampling.

Convenience Sampling

Purposive Sampling

Quota Sampling

Referral /Snowball Sampling

# Convenience Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Convenience Sampling](https://dashboard.stige.in/index.php/topic/convenience-sampling/)

In Progress

Here the samples are selected based on the availability. This method is used when the availability of sample is rare and also costly. So based on the convenience samples are selected.

**For example**: Researchers prefer this during the initial stages of survey research, as it’s quick and easy to deliver results.

# Purposive Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Purposive Sampling](https://dashboard.stige.in/index.php/topic/purposive-sampling/)

In Progress

This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study.

**For Example:** If we want to understand the thought process of the people who are interested in pursuing master’s degree then the selection criteria would be “Are you interested for Masters in..?”

All the people who respond with a “No” will be excluded from our sample.

# Quota Sampling

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Quota Sampling](https://dashboard.stige.in/index.php/topic/quota-sampling/)

In Progress

This type of sampling depends of some pre-set standard. It selects the representative sample from the population. Proportion of characteristics/ trait in sample should be same as population. Elements are selected until exact proportions of certain types of data is obtained or sufficient data in different categories is collected.

**For example:** If our population has 45% females and 55% males then our sample should reflect the same percentage of males and females.

# Referral /Snowball Sampling

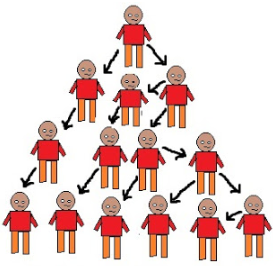
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 5 : Sampling Techniques](https://dashboard.stige.in/index.php/lessons/task-5-sampling-techniques/) [Referral /Snowball Sampling](https://dashboard.stige.in/index.php/topic/referral-snowball-sampling/)

In Progress

This technique is used in the situations where the population is completely unknown and rare.

Therefore we will take the help from the first element which we select for the population and ask him to recommend other elements who will fit the description of the sample needed.

So this referral technique goes on, increasing the size of population like a snowball.

Referral /Snowball Sampling

**For example**: It’s used in situations of highly sensitive topics like HIV Aids where people will not openly discuss and participate in surveys to share information about HIV Aids.

Not all the victims will respond to the questions asked so researchers can contact people they know or volunteers to get in touch with the victims and collect information

Helps in situations where we do not have the access to sufficient people with the characteristics we are seeking. It starts with finding people to study.

# Task 6 : A/B Testing

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/)

In Progress

### Overview

* A/B testing is a popular way to test your products and is gaining steam in the data science field
* Here, we’ll understand what A/B testing is and how you can leverage A/B testing in data science using Python

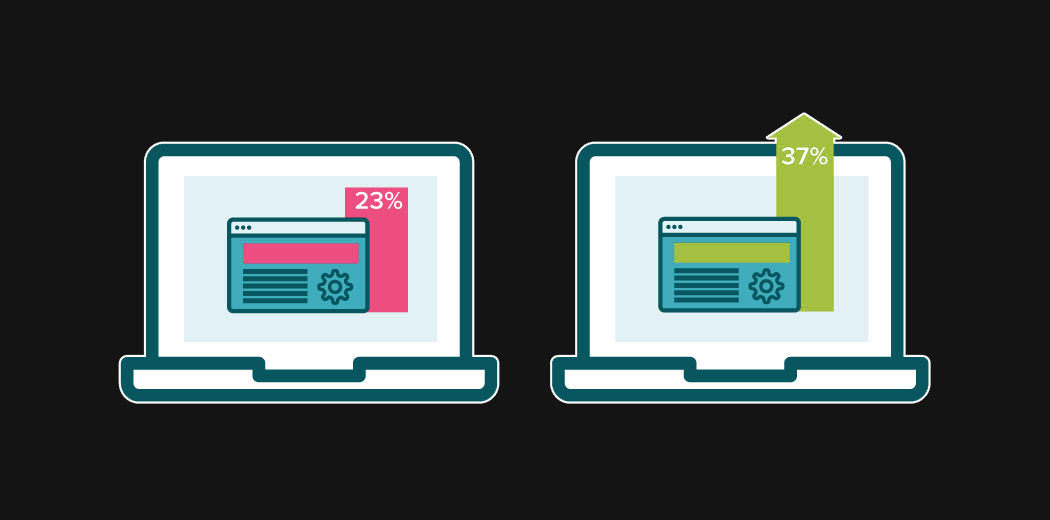
### Introduction

Statistical analysis is our best tool for predicting outcomes we don’t know, using the information we know.

Picture this scenario – You have made certain changes to your website recently. Unfortunately, you have no way of knowing with full accuracy how the next 100,000 people who visit your website will behave. That is the information we cannot know today, and if we were to wait until those 100,000 people visited our site, it would be too late to optimize their experience.

This seems to a classic Catch-22 situation!

This is where a data scientist can take control. A data scientist collects and studies the data available to help optimize the website for a better consumer experience. And for this, it is imperative to know how to use various statistical tools, especially the concept of A/B Testing.



A/B Testing is a widely used concept in most industries nowadays, and data scientists are at the forefront of implementing it. In this article, I will explain A/B testing in-depth and how a data scientist can leverage it to suggest changes in a product.

# What is A/B testing?

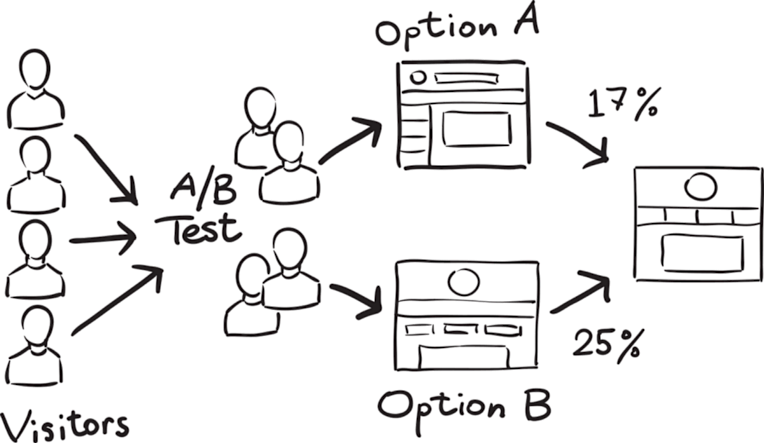
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [What is A/B testing?](https://dashboard.stige.in/index.php/topic/what-is-a-b-testing/)

In Progress

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

[*Source*](https://www.vippng.com/maxp/hxxximo/)

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The**population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

# How does A/B testing work?

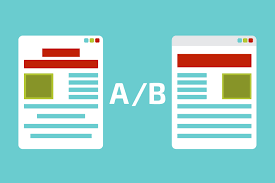
[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [How does A/B testing work?](https://dashboard.stige.in/index.php/topic/how-does-a-b-testing-work/)

In Progress

The big question!

In this section, let’s understand through an example the logic and methodology behind the concept of A/B testing.

Let’s say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.



### Objective

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

#### 1.  Make a Hypothesis

Before making a hypothesis, let’s first understand what is a hypothesis.

A hypothesis is a tentative insight into the natural world; a concept that is not yet verified but if true would explain certain facts or phenomena.

It is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation. In our example, the hypothesis can be “By making changes in the language of the newsletter, we can get more traffic on the website”.

In [hypothesis testing](https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/?utm_source=blog&utm_medium=ab-testing-data-science), we have to make two hypotheses i.e Null hypothesis and alternative hypothesis. Let’s have a look at both.

1. **Null hypothesis or H0:**The **null hypothesis** is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is **no** difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H0is ” there is no difference in the conversion rate in customers receiving newsletter A and B”.
2. Alternative Hypothesis or **H0**:

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

In our example, the Hais- “**the conversion rate of newsletter B is higher than those who receive newsletter A**“.

Now, we have to collect enough evidence through our tests to **reject the null hypothesis**.

#### 2. Create Control Group and Test Group

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Here we have two groups – **The Control group**, and **the Test (variant) group**.

The Control Group is the one that will receive newsletter A and the Test Group is the one that will receive newsletter B.

For this experiment, we randomly select 1000 customers – 500 each for our Control group and Test group.

Randomly selecting the sample from the population is called **random sampling**. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it’s important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.**

Another important aspect we must take care of is **the Sample size.** It is required that we determine the minimum sample size for our A/B test before conducting it so that we can eliminate **under coverage bias.** It is the bias from sampling too few observations.

#### 3. Conduct the A/B Test and Collect the Data

One way to perform the test is to calculate **daily conversion rates** for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

When we run our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%.

# Statistical significance of the Test

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [Statistical significance of the Test](https://dashboard.stige.in/index.php/topic/statistical-significance-of-the-test/)

In Progress

Now, the main question is – Can we conclude from here that the Test group is working better than the control group?

The answer to this is a simple No! For rejecting our null hypothesis we have to prove the **Statistical significance** of our test.

There are two types of errors that may occur in our hypothesis testing:

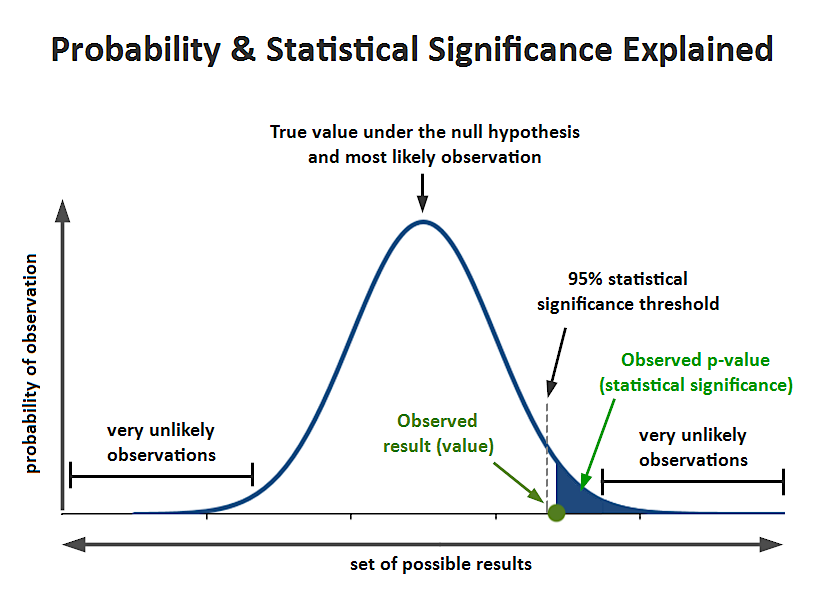
1. **Type I error**: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A
2. **Type II error**: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A

To avoid these errors we must calculate the statistical significance of our test.

An experiment is considered to be statistically significant when we have enough evidence to prove that the result we see in the sample also exists in the population.

That means the difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a [two-sample T-test](https://www.analyticsvidhya.com/blog/2019/05/statistics-t-test-introduction-r-implementation/?utm_source=blog&utm_medium=ab-testing-data-science).

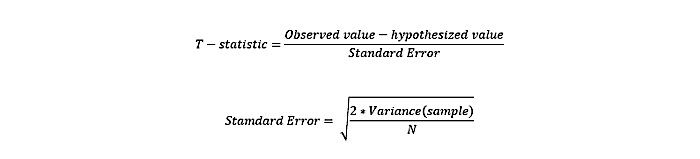
The **two**–**sample t**–**test** is one of the most commonly **used** hypothesis **tests**. It is applied to compare whether the average difference between **the two** groups.

[*Source*](http://blog.analytics-toolkit.com/2017/statistical-significance-ab-testing-complete-guide/2017-09-11-statistical-significance-p-value-2/)

To understand this, we must be familiar with a few terms:

1. **Significance level (alpha):** The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05
2. **P-Value:** It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value stronger the chances to reject the H0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis
3. **Confidence interval:** The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

Next, we can calculate our t statistics using the below formula:



### Let’s Implement the Significance Test in Python

Let’s see a python implementation of the significance test. Here, we have a dummy data having an experiment result of an A/B testing for 30 days. Now we will run a two-sample t-test on the data using Python to ensure the statistical significance of data.

import pandas as pd

import numpy as np

import seaborn as sns

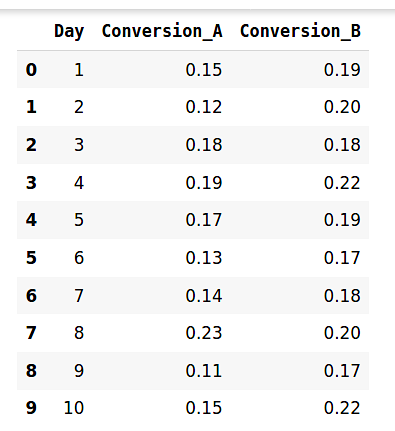
import scipy.stats as ss

data= pd.read\_csv("ab\_test.csv")

You can download the sample data [here](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/10/data.csv).

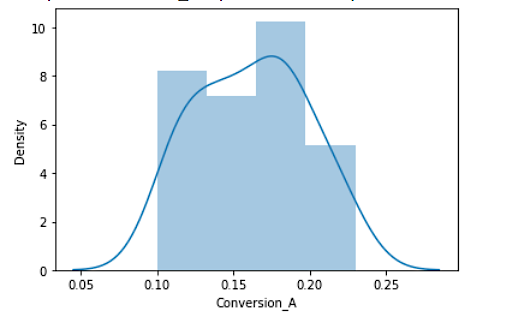
Let’s see the data:

data.head(10)

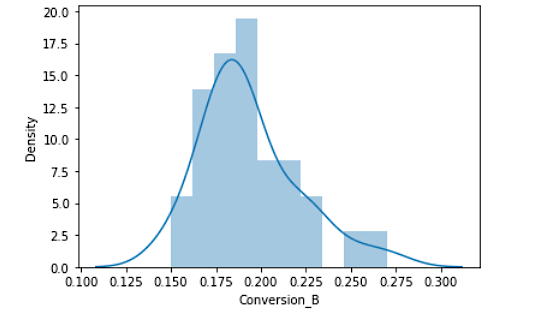


Let’s plot the distribution of target and control group:

sns.distplot(data.Conversion\_A)



sns.distplot(data.Conversion\_B)



At last, we will perform the t-test:

t\_stat, p\_val= ss.ttest\_ind(data.Conversion\_B,data.Conversion\_A)

t\_stat , p\_val

(3.78736793091929, 0.000363796012828762)

For our example, the observed value i.e the mean of the test group is 0.19. The hypothesized value (Mean of the control group) is 0.16. On the calculation of the t-score, we get the t-score as **.3787**. and the p-value is **0.00036**.

SO what does all this mean for our A/B Testing?

Here, our p-value is less than the significance level i.e 0.05. Hence, we can reject the null hypothesis. This means that in our A/B testing, newsletter B is performing better than newsletter A. So our recommendation would be to replace our current newsletter with B to bring more traffic on our website.

**Mistakes we must avoid while conducting the A/B test**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [Mistakes we must avoid while conducting the A/B test](https://dashboard.stige.in/index.php/topic/mistakes-we-must-avoid-while-conducting-the-a-b-test/)

In Progress

There are a few key mistakes I’ve seen data science professionals making. Let me clarify them for you here:

* **Invalid hypothesis**: The whole experiment depends on one thing i.e the hypothesis. What should be changed? Why should it be changed, what the expected outcome is, and so on? If you start with the wrong hypothesis, the probability of the test succeeding, decreases
* **Testing too Many Elements Together:** Industry experts caution against running too many tests at the same time. Testing too many elements together makes it difficult to pinpoint which element influenced the success or failure. Thus, prioritization of tests is indispensable for successful A/B testing
* **Ignoring Statistical Significance:**It doesn’t matter what you feel about the test. Irrespective of everything, whether the test succeeds or fails, allow it to run through its entire course so that it reaches its statistical significance
* **Not considering the external factor:**Tests should be run in comparable periods to produce meaningful results. For example, it is unfair to compare website traffic on the days when it gets the highest traffic to the days when it witnesses the lowest traffic because of external factors such as sale or holidays

# When to use A/B test

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [When to use A/B test](https://dashboard.stige.in/index.php/topic/when-to-use-a-b-test/)

In Progress

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn’t work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

**End Notes**

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 6 : A/B Testing](https://dashboard.stige.in/index.php/lessons/task-6-a-b-testing/) [End Notes](https://dashboard.stige.in/index.php/topic/end-notes-2/)

In Progress

To summarize, A/B testing is at least a 100-year-old statistical methodology but in its current form, it comes in the 1990s. Now it has become more eminent with the online environment and availability for big data. It is easier for companies to conduct the test and utilize the results for better user experience and performance.

There are many tools available for conducting A/B testing but being a data scientist you must understand the factors working behind it. Also, you must be aware of the statistics in order to validate the test and prove it’s statistical significance.

*To know more about hypothesis testing, I will suggest you read the following article:*

* [*Statistics for Analytics and Data Science: Hypothesis Testing and Z-Test vs. T-Test*](https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/?utm_source=blog&utm_medium=ab-testing-data-science)

# ask 7 : A/B Testing Summary

[Data Analytics](https://dashboard.stige.in/index.php/courses/lms-data-analytics/) [Task 7 : A/B Testing Summary](https://dashboard.stige.in/index.php/lessons/task-7-a-b-testing-summary/)

In Progress

## First, why do we do A/B tests?

The answer is testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from “we think” to “we know.” By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results. Nowadays it’s very common for companies to do A/B tests on web page versions, personalized recommendations and new features.

## Can we test everything?

NO. There are situations we cannot analyze through A/B test. For example, if you are adding a new experience and want to test it, old users may resist against the new version (change aversion), or old users may all go for the new experience, then the test set has everything (novelty effect). Two issues to consider when it comes to new experience: (1) what is the base of your comparison? (2) how much time you need in order for your users to adapt to the new experience, so that you can actually say what is the plateaued experience and make a robust decision? Except for new experience, long term effect is hard to test too. For example, a home rental website test its referral effect, but a customer may not return even in six months, it’s very hard to measure through A/B testing. If this is the case, what shall we do?

When A/B testing is not useful, we can:

* Analyze the user activity logs
* Conduct retrospective analysis
* Conduct user experience research
* Focus groups and surveys
* Human evaluation

## Then, how to do an A/B test?

In practice, an A/B test can be summarized into the 5 steps below:

1. Choose and characterize metrics to evaluate your experiments, i.e. what do you care about, how do you want to measure the effect
2. Choose significance level (alpha), statistical power (1-beta) and practical significance level you really want to launch the change if the test is statistically significant
3. Calculate required sample size
4. Take sample for control/treatment groups and run the test
5. Analyze the results and draw valid conclusions

In this Udacity course, the five steps are expanded into detailed explanation with numerous real-world examples:

## Step 1: Choose and characterize metrics for both sanity check and evaluation

The metrics we choose for sanity check are called as invariant metrics. They are not supposed to be affected by the experiment. They should not change across control and treatment groups. Otherwise, the experiment setup is incorrect.

The evaluation metrics we choose are used to measure which variation is better. For example, we could use daily active users (DAU) to measure user engagement, use click through rate (CTR) to measure a button design on a webpage, etc. In general, there are **four categories of metrics**that you should keep in mind:



* Sums and counts
* Distribution (mean, median, percentiles)
* Probability and rates (e.g. Click-through probability, Click-through rate)
* Ratios: any two numbers divide by each other

Other than choosing the category of metrics, you should also consider **sensitivity**and**robustness**. You want to choose a metric that is has high sensitivity, that means the metric can pick up the change you care about. You also want the metric to be robust against changes you don’t care about. It means the metric doesn’t change a lot when nothing you’re interested happened. If a metric is too sensitive then it is not robust enough, thus there’s a balance between these two and you need to look into the data to find out which metric to use.

**How to measure the sensitivity and robustness?**

* Run experiments
* Use A/A test to see if metrics pick up difference (if yes, then the metric is not robust)
* Retrospective analysis

## Step 2: Choose significance level, statistical power and practical significance level

Usually the significance level is 0.05 and power is set as 0.8. Practical significance level varies depends on each individual tests, it tells you how much change the test detects that makes you really want to launch the change. You may not want to launch a change even if the test is statistically significant because you need to consider the business impact of the change, whether it is worthwhile to launch considering the engineering cost, customer support or sales issue, and opportunity costs.

## Step 3: Calculate required sample size

Overview: Need to consider the choice of metric, the choice of unit of diversion, and the choice of population into account because they all affect the variability of your metrics. Then decide on the size of experiment.

* **Subject**: What is the subject (**unit of diversion**) of the test? I.e. what are the units you are going to run the test on and comparing. Unit of diversion can be event based (e.g. pageview) or anonymous ID(e.g. cookie id) or user ID. These are commonly used unit of diversion. For user visible changes, you want to use user\_id or cookie to measure the change. If measuring latency change, other metrics like event level diversion might be enough.
* **Population**: What subjects are eligible for the test? Everyone? Only people in the US? Only people in certain industry?
* **How to reduce the size of an experiment to get it done faster?**You can increase significance level alpha, or reduce power (1-beta) which means increase beta, or change the unit of diversion if originally it is not the same with unit of analysis (unit of analysis: denominator of your evaluation metric) .

## Step 4: Take sample for control/treatment groups and run the test

Several things to keep in mind:

* **Duration**: What’s the best time to run it? Students going back to college? Holidays? Weekend vs. weekdays?
* **Exposure**: What fraction of traffic you want to expose the experiment to? Suggestion is take a small fraction, run multiple tests at the same time (different days: weekend, weekday, holiday).
* **Learning effect**: When there’s a new change, in the beginning users may against the change or use the change a lot. But overtime, user behavior becomes stable, which is called plateau stage. The key thing to measure learning effect is time, but in reality you don’t have that much luxury of taking that much time to make a decision. Suggestion: run on a smaller group of users, for a longer period of time.

## Step 5: Analyze the results and draw conclusions

### **First step, sanity check.**

Before analyzing result the first step is to do sanity check — check if your invariant metrics have changed. If your sanity check failed, do not proceed. Instead, go analyze why your sanity check failed. You can do either: (1) retrospective analysis, or (2) look into if there’s learning effect.

### **Second step, analyze the results.**

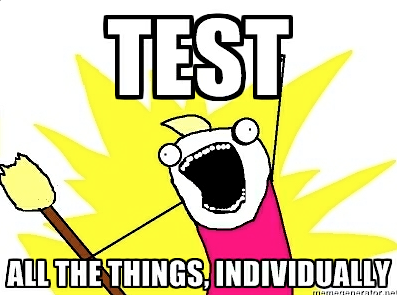
**a. If you have one single metric for evaluation, and if it is not significant:**

First round: Inspect result, to see if there is really not significant difference. e.g. Break down into different platforms, or day of the week. This may help you find out bug in the system, and may also help you find insight about how users react to your experiment.

Second round: Cross checking by using different methods. e.g. Compare with non-parametric sign test with parametric hypothesis test. What do you do if your hypothesis test and sign test does not agree? You should look into your data critically because you may be suffering from Simpson’s paradox (a trend appears in different groups of data but disappears or reverses when these groups are combined). The reasons for Simpson’s paradox happening could be: (1) The setup of your experiment is incorrect; (2) The change affects the new user and experienced users differently.

**b. If you are measuring multiple metrics at the same time**

(1) One potential problem is, you might see a significant result by chance. (Check out this [xkcd: significant](https://xkcd.com/882/))



For example, you are running a tests with 20 variants, and you test each hypothesis separately:

P(one significant result) = 1−P(no significant results)

P(one significant result) = 1−(1−0.05)^20 = 0.64

There’s very high chance you’ll see a significant result by chance!! Luckily, there are **several ways to solve this problem**:

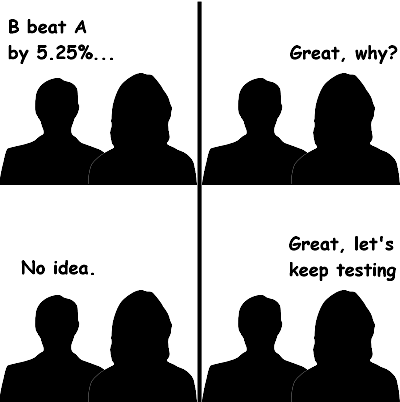
* **Bootstrap** and run experiments again and again, the significant metric should disappear if it occurred by chance.
* **Bonferroni correction**: divide the significance level 0.05 by the number of tests in the multiple testing. Say if you are measuring 20 tests, then your significance level for the test should be 0.05/20 = 0.0025. The problem of Bonferroni correction is it tends to be too conservative. If many metrics are tested at the same time, maybe none of them turned out to be significant.
* Control **Familywise Error Rate (FWER)**: probability that any metric shows false positive.
* Control **false discovery rate (FDR)**: FDR = # false positives / # total rejections.

(2) Another potential problem is, what if metrics are **not moving at the same direction** as you thought? For example, you expect DAU and average length of time users use your app both increase. However you observe DAU decrease while average length of time increase. WTH??

You should **dive deeper, and figure out why**. And this is also why people usually want to have one OEC (Overall Evaluation Criterion). A good OEC gives you a balance between short-term and long-term goal, or the balance between different metrics. However you also need to keep in mind, having an OEC helps you understand what your business cares about, and how do you balance metrics such as stay time and click, but it does not help you make a product change decision.

## **Last step, draw conclusions.**

If you have a significant result from the test…There comes two questions: Do you understand the change? Do you want to launch the change? What if your change has a positive impact on one slice of users, but no impact or negative impact for other slices of users? Do you understand WHY?? (e.g. Bolded words in English vs. in Chinese have different test results, because bolded Chinese is hard to read).



Then how do I decide whether to launch the change or not?

Ask yourself a few questions: Do I have statistically significant and practically significant result in order to justify the change? Do I understand what the change actually done to our user experience? Last but not the least, is it worth it to launch?

### Gotchas

Always do a ramp up when you want to launch a change after the A/B test. Because you want to know if there’s any incidental impact to unaffected users that you didn’t test in the original experiment.

When you are ramping up the change, you may see the effect flatten out. Thus making the tested effect not repeatable. There are many reasons for this phenomenon.

1. **Seasonality effect**: Social network platform user behavior changes a lot when students start summer vacation or going back to school. Holidays affect users’ shopping behavior a lot. Solution: use hold-back method , launch the change to everyone except for one small hold-back group of users, and continue comparing their behavior to the control group.
2. **Novelty effect** or**change aversion**: cohort analysis may be helpful.

### Lessons learned

1. Double check, triple check that your experiment is set up properly.
2. Not only think about statistically significant but also business impact. Think about the engineering cost, customer support or sales issue, what’s the opportunity cost, etc.
3. If you are running your first experiment that have a big impact, you might want to run a couple of experiments and check the results to see if you are comfortable launching it.

### Other thing to consider: **Politics and ethics for experiments**

Experiments involves real people, it is important to protect the users and follow the ethics. However, there were many problematic examples of experiments in the past. For example, [Tuskegee syphilis experiment](https://en.wikipedia.org/wiki/Tuskegee_syphilis_experiment), [Milgram experiment](https://en.wikipedia.org/wiki/Milgram_experiment) in history and a recent [Facebook emotion experiment](https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html). To conduct an A/B test in an ethical way, there are four principles to keep in mind:

**1.Risk: What risk are the participants exposed to?**

The main threshold is whether the risk exceeds that of “minimal risk”. Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. The harm considered encompasses physical, psychological and emotional, social, and economic concerns. If the risk exceeds minimal risk, then informed consent is required.

**2. Benefit: What’s the potential benefit of the outcome of the study?**

It is important to be able to state what the benefit would be from completing the study.

**3. Choice: What other choices do participants have?**

In online experiments, the issues to consider are what the other alternative services that a user might have, and what the switching costs might be, in terms of time, money, information, etc.

**4. Privacy: What privacy do participants have?**

For new data being collected and stored, how sensitive is the data and what are the internal safeguards for handling that data? Then, for that data, how will it be used and how will participants’ data be protected? How are participants guaranteed that their data, which was collected for use in the study, will not be used for some other purpose?