A **statistical hypothesis** is a [hypothesis](https://en.wikipedia.org/wiki/Hypothesis) that is testable on the basis of [observed](https://en.wikipedia.org/wiki/Observable_variable) data [modelled](https://en.wikipedia.org/wiki/Statistical_model" \o "Statistical model) as the realised values taken by a collection of [random variables](https://en.wikipedia.org/wiki/Random_variable). A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions.

A **statistical hypothesis test** is a method of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference). An [alternative hypothesis](https://en.wikipedia.org/wiki/Alternative_hypothesis) is proposed for the probability distribution of the data, either explicitly or only informally. The comparison of the two models is deemed [*statistically significant*](https://en.wikipedia.org/wiki/Statistically_significant) if, according to a threshold probability—the significance level—the data would be unlikely to occur if the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) were true.

A hypothesis test specifies which outcomes of a study may lead to a rejection of the null hypothesis at a pre-specified level of significance, while using a pre-chosen measure of deviation from that hypothesis (the test statistic, or goodness-of-fit measure). The pre-chosen level of significance is the maximal allowed "false positive rate". One wants to control the risk of incorrectly rejecting a true null hypothesis.

The process of distinguishing between the null hypothesis and the [alternative hypothesis](https://en.wikipedia.org/wiki/Alternative_hypothesis) is aided by considering two types of errors.

A [Type I error](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) occurs when a true null hypothesis is rejected.

A [Type II error](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) occurs when a false null hypothesis is not rejected.

Hypothesis tests based on statistical significance are another way of expressing [confidence intervals](https://en.wikipedia.org/wiki/Confidence_interval) (more precisely, confidence sets). In other words, every hypothesis test based on significance can be obtained via a confidence interval, and every confidence interval can be obtained via a hypothesis test based on significance.

A statistical test procedure is comparable to a criminal [trial](https://en.wikipedia.org/wiki/Trial_(law)); a defendant is considered not guilty as long as his or her guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough evidence for the prosecution is the defendant convicted.

In the start of the procedure, there are two hypotheses

H0{\displaystyle H\_{0}}: "the defendant is not guilty", and {\displaystyle H\_{1}}

**H1**: "the defendant is guilty".

The first one, **{\displaystyle H\_{0}}H0**, is called the [*null hypothesis*](https://en.wikipedia.org/wiki/Null_hypothesis).

The second one, **{\displaystyle H\_{1}}H1**, is called the *alternative hypothesis*. It is the alternative hypothesis that one hopes to support.

The hypothesis of innocence is rejected only when an error is very unlikely, because one doesn't want to convict an innocent defendant. Such an error is called [*error of the first kind*](https://en.wikipedia.org/wiki/Error_of_the_first_kind) (i.e., the conviction of an innocent person), and the occurrence of this error is controlled to be rare. As a consequence of this asymmetric behaviour, an [*error of the second kind*](https://en.wikipedia.org/wiki/Error_of_the_second_kind) (acquitting a person who committed the crime), is more common.

|  |  |  |
| --- | --- | --- |
|  | **H0 is true Truly not guilty** | **H1 is true Truly guilty** |
| **Do not reject the null hypothesis Acquittal** | Right decision | Wrong decision Type II Error |
| **Reject null hypothesis Conviction** | Wrong decision Type I Error | Right decision |

**Statistical hypothesis**

A statement about the parameters describing a [population](https://en.wikipedia.org/wiki/Statistical_population) (not a [sample](https://en.wikipedia.org/wiki/Statistical_sample)).

[**Statistic**](https://en.wikipedia.org/wiki/Statistic)

A value calculated from a sample without any unknown parameters, often to summarize the sample for comparison purposes.

**Simple hypothesis**

Any hypothesis which specifies the population distribution completely.

**Composite hypothesis**

Any hypothesis which does *not* specify the population distribution completely.

[**Null hypothesis**](https://en.wikipedia.org/wiki/Null_hypothesis)**(H0)**

A hypothesis associated with a contradiction to a theory one would like to prove.

**Positive data**

Data that enable the investigator to reject a null hypothesis.

[**Alternative hypothesis**](https://en.wikipedia.org/wiki/Alternative_hypothesis)**(H1)**

A hypothesis (often composite) associated with a theory one would like to prove.

**Statistical test**

A procedure whose inputs are samples and whose result is a hypothesis.

**Region of rejection / Critical region**

The set of values of the test statistic for which the null hypothesis is rejected.

[**Critical value**](https://en.wikipedia.org/wiki/Critical_value#Statistics)

The threshold value of the test statistic for rejecting the null hypothesis.

[**Power of a test**](https://en.wikipedia.org/wiki/Statistical_power)**(1 − *β*)**

The test's probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. The complement of the [false negative](https://en.wikipedia.org/wiki/False_negative) rate, *β*. Power is termed **sensitivity** in [biostatistics](https://en.wikipedia.org/wiki/Biostatistics). ("This is a sensitive test. Because the result is negative, we can confidently say that the patient does not have the condition.") See [sensitivity and specificity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity) and [Type I and type II errors](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) for exhaustive definitions.

[**Size**](https://en.wikipedia.org/wiki/Size_(statistics))

For simple hypotheses, this is the test's probability of *incorrectly* rejecting the null hypothesis. The [false positive](https://en.wikipedia.org/wiki/False_positive) rate. For composite hypotheses this is the supremum of the probability of rejecting the null hypothesis over all cases covered by the null hypothesis. The complement of the false positive rate is termed **specificity** in [biostatistics](https://en.wikipedia.org/wiki/Biostatistics). ("This is a specific test. Because the result is positive, we can confidently say that the patient has the condition.") See [sensitivity and specificity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity) and [Type I and type II errors](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) for exhaustive definitions.

**Significance level of a test (*α*)**

It is the upper bound imposed on the size of a test. Its value is chosen by the statistician prior to looking at the data or choosing any particular test to be used. It is the maximum exposure to erroneously rejecting H0 that they are ready to accept. Testing H0 at significance level *α* means testing H0 with a test whose size does not exceed *α*. In most cases, one uses tests whose size is equal to the significance level.

[***p*-value**](https://en.wikipedia.org/wiki/P-value)

What the probability of observing a test statistic at least as extreme as the one actually observed would be if the null hypothesis were true.

[**Statistical significance**](https://en.wikipedia.org/wiki/Statistical_significance)**test**

A predecessor to the statistical hypothesis test (see the Origins section). An experimental result was said to be statistically significant if a sample was sufficiently inconsistent with the (null) hypothesis. This was variously considered common sense, a pragmatic heuristic for identifying meaningful experimental results, a convention establishing a threshold of statistical evidence or a method for drawing conclusions from data. The statistical hypothesis test added mathematical rigor and philosophical consistency to the concept by making the alternative hypothesis explicit. The term is loosely used for the modern version which is now part of statistical hypothesis testing.

**Conservative test**

A test is conservative if, when constructed for a given nominal significance level, the true probability of *incorrectly* rejecting the null hypothesis is never greater than the nominal level.

[**Exact test**](https://en.wikipedia.org/wiki/Exact_test)

A test in which the significance level or critical value can be computed exactly, i.e., without any approximation. In some contexts this term is restricted to tests applied to [categorical data](https://en.wikipedia.org/wiki/Categorical_data) and to [permutation tests](https://en.wikipedia.org/wiki/Permutation_tests), in which computations are carried out by complete enumeration of all possible outcomes and their probabilities.

A statistical hypothesis test compares a test statistic (*z* or *t* for examples) to a threshold. The test statistic (the formula found in the table below) is based on optimality. For a fixed level of Type I error rate, use of these statistics minimizes Type II error rates (equivalent to maximizing power). The following terms describe tests in terms of such optimality:

**Most powerful test**

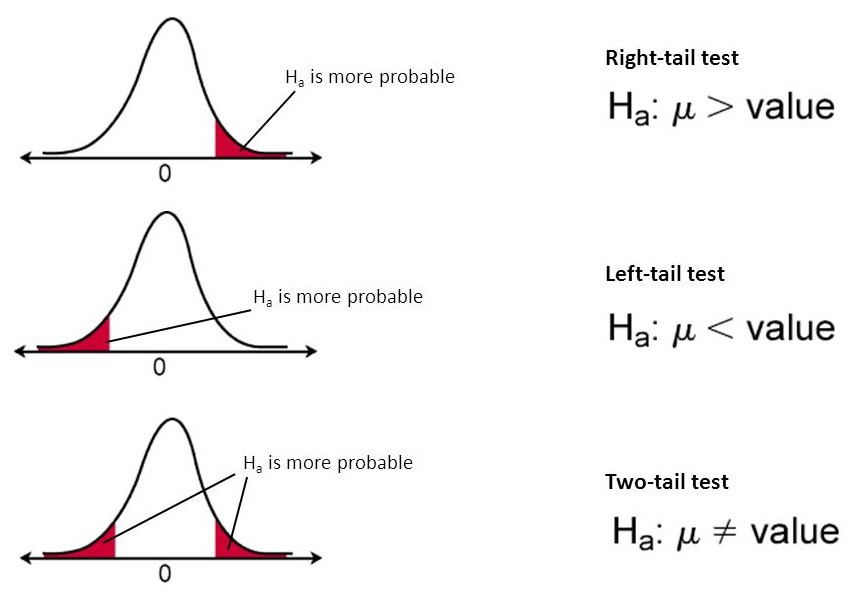
For a given *size* or *significance level*, the test with the greatest power (probability of rejection) for a given value of the parameter(s) being tested, contained in the alternative hypothesis.

[**Uniformly most powerful test**](https://en.wikipedia.org/wiki/Uniformly_most_powerful_test)**(UMP)**

A test with the greatest *power* for all values of the parameter(s) being tested, contained in the alternative hypothesis.

Statistics is all about data but data alone is not interesting. It is the interpretation of the data that we are interested in…

Statistics is all about data. Data alone is not interesting. It is the interpretation of the data that we are interested in. Using **Hypothesis Testing**, we try to interpret or draw conclusions about the population using sample data. A **Hypothesis Test**evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. Whenever we want to make claims about the distribution of data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests



the terminology that we should be aware of in **Hypothesis Testing**

1. Parameter and Statistic:

A **Parameter**is a summary description of a fixed characteristic or measure of the target population. A Parameter denotes the true value that would be obtained if a census rather than a sample were undertaken

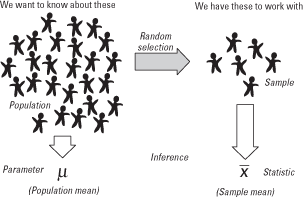
***Ex:*** Mean (μ), Variance (σ²), Standard Deviation (σ), Proportion (π)

*Population: Population is a collection of objects that we want to study/test. The collection of objects could be Cities, Students, Factories, etc. It depends on the study at hand.*

In the real world, it’s tough ask to get complete information about the population. Hence, we draw a sample out of that population and derive the same statistical measures mentioned above. And these measures are called Sample Statistics. In other words,

A **Statistic**is a summary description of a characteristic or measure of the sample. The Sample Statistic is used as an estimate of the population parameter.

***Ex:***Sample Mean (x̄), Sample Variance (S²), Sample Standard Deviation (S), Sample Proportion (*p)*



## 2. Sampling Distribution:

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

Ex: Suppose a simple random sample of five hospitals is to be drawn from a population of 20 hospitals. The possibilities could be, (20, 19, 18, 17, 16) or (1,2,4,7,8) or any of the 15,504 (using 20C₅ combinations) different samples of size 5 can be drawn.

In general, the mean of the sampling distribution will be approximately equivalent to the population mean i.e. E(*x̄) = μ*

## 3. Standard Error (SE):

The standard error (SE) is very similar to the standard deviation. Both are measures of spread. The higher the number, the more spread out your data is. To put it simply, the two terms are essentially equal — but there is one important difference. While the standard error uses **statistics**(sample data) standard deviation use **parameters**(population data)

*The standard error tells you how far your sample statistic (like the sample mean) deviates from the actual population mean. The larger your sample size, the smaller the SE. In other words, the larger your sample size, the closer your sample mean is to the actual population mean*

## 4. (a). Null Hypothesis (H₀):

A statement in which no difference or effect is expected. If the null hypothesis is not rejected, no changes will be made.

*The word “null” in this context means that it’s a commonly accepted fact that researchers to nullify. It doesn’t mean that the statement is null itself! (Perhaps the term should be called the “nullifiable hypotheiss” as that might cause less confusion)*

## 4. (b). Alternate Hypothesis (H₁):

A statement that some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions. It is the opposite of the null hypothesis.

5. (a). One-Tailed Test:

A one-tailed test is a statistical hypothesis test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

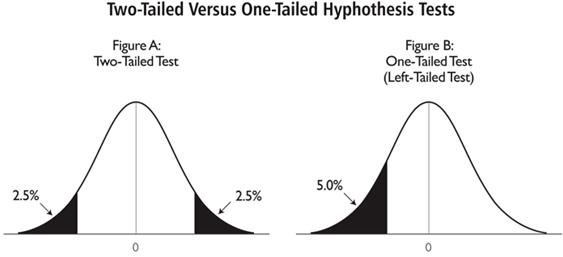
*A one-tailed test is also known as a directional hypothesis or directional test.*

***Critical Region:****The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.*

5. (b). Two-Tailed Test:

A two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

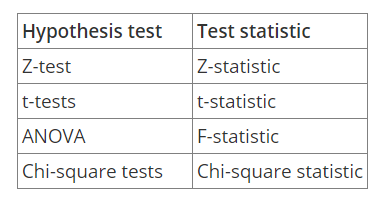
*By convention, two-tailed tests are used to determine significance at the 5% level, meaning each side of the distribution is cut at 2.5%*



6. Test Statistic:

The **test statistic**measures how close the sample has come to the null hypothesis. Its observed value changes randomly from one random sample to a different sample. A test statistic contains information about the data that is relevant for deciding whether to reject the null hypothesis or not.

Different hypothesis tests use different test statistics based on the probability model assumed in the null hypothesis. Common tests and their test statistics include:



*In general, the sample data must provide sufficient evidence to reject the null hypothesis and conclude that the effect exists in the population. Ideally, a hypothesis test fails to reject the null hypothesis when the effect is not present in the population, and it rejects the null hypothesis when the effect exists.*

By now we understand that the entire hypothesis testing works on based on the sample that is at hand. We may come to a different conclusion if the sample is changed. There are two types of errors that relate to incorrect conclusions about the null hypothesis.

7. (a). Type-I Error:

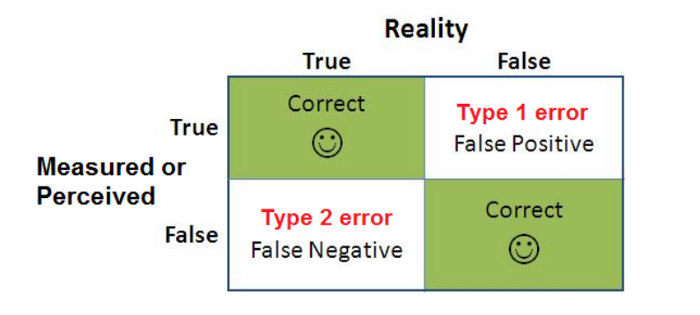
***Type-I***error occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true. **Type-I**errors are equivalent to false positives.

**Type-I**errors can be controlled. The value of alpha, which is related to the **level of Significance**that we selected has a direct bearing on **Type-I**errors.

7. (b). Type-II Error:

***Type-II***error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false. **Type-II**errors are equivalent to false negatives.

https://miro.medium.com/max/60/1*5GiMt-wyccIDq2XnSi2fug.png?q=20



Level of Significance (α):

The probability of making a **Type-I**error and it is denoted by ***alpha (α)***. Alpha is the maximum probability that we have a **Type-I**error. For a 95% confidence level, the value of alpha is 0.05. This means that there is a 5% probability that we will reject a true null hypothesis.

P-Value:

The ***p-value***is used all over statistics, from t-tests to simple regression analysis to tree-based models almost in all the machine learning models. We all use ***P-values***to determine statistical significance in a hypothesis test. Despite being so important, the ***P-value***is a slippery concept that people often interpret incorrectly.

***P-values***evaluate how well the sample data support the devil’s advocate argument that the null hypothesis is true. It measures how compatible your data are with the null hypothesis. How likely the effect observed in your sample data if the null hypothesis is true?

*In other words, given the null hypothesis is true, a****P-Value****is a probability of getting a result as or more extreme than the sample result by random chance alone.*

***High P-Values:****Your data are likely with a true null*

***Low P-Values:****Your data are unlikely with a true null*

Ex: Suppose you are testing the following hypothesis at a significance level (α) of 5% and you got the p-value as 3%, and your sample statistic is *x̄* *=* *25*

H₀: μ = 20

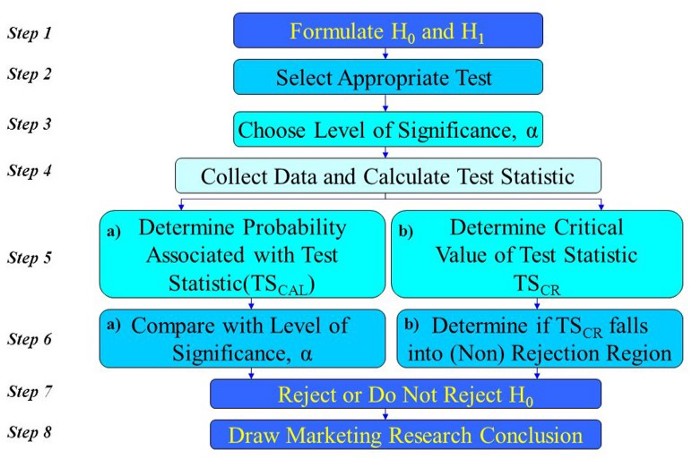
H₁: μ > 20

*The interpretation of the p-value as follows:*

*We saw above that****α****is also known as committing****Type-I****error. When we say α*=*5%, we are fine to reject our null hypothesis 5 out of 100 times even though it is true. Now that our****P-value****is 3% which is less than****α****(we are definitely below the threshold of committing****Type-I****error)****,****means obtaining a sample statistic as extreme as possible (x̄* >*=* *25) given that H₀ is true is very less. In other words, we can’t obtain our sample statistic as long as we assume H₀ is true. Hence, we reject H₀ and accept H₁. Suppose you get****P-Value****as 6% i.e. the probability of obtaining the sample statistic as extreme as possible is higher given that the null hypothesis is true. So we fail to reject H₀, comparing with****α****we can’t take risk of committing****Type-I****error more than the agreed level of significance. Hence, we fail to reject the null hypothesis and reject the alternative hypothesis.*

Now that we understood the basic terminology in the **Hypothesis Testing,**now let’s look at the steps involved in the Hypothesis Testing and an illustration with an example.

https://miro.medium.com/max/60/1*NgMLihgMLAcZdHr0ISecgg.jpeg?q=20



For example, a major department store is considering the introduction of an Internet shopping service. The new service will be introduced if more than 40 percent of the Internet users shop via the Internet.

***Step1: Formulate the Hypotheses:***

The appropriate way to formulate the hypotheses is:

H₀: π ≤ 0.40

H₁: π > 0.40

If the null hypothesis H₀ is rejected, then the alternative hypothesis H₁ will be accepted and the new Internet shopping service will be introduced. On the other hand, if we fail to reject H₀ then the new service should not be introduced unless additional evidence is obtained. This test of the null hypothesis is a **one-tailed**test, because the alternative hypothesis is expressed directionally: The proportion of Internet users who use the Internet for shopping is greater than 0.40.

***Step2: Select an appropriate Test:***

To test the null hypothesis, it is necessary to select an appropriate statistical technique. For this example, the ***z***statistic, which follows the standard normal distribution would be appropriate.

*z = (p-π)/σₚ, where σₚ=sqrt(π(1-π)/n)*

***Step3: Choose Level of Significance, α:***

We understood that ***Level of Significance***refers to ***Type-I***error. In our example, a Type-I error would occur if we concluded, based on the sample data, that the proportion of customers preferring the new service plan was greater than 0.40, when in fact it was less than or equal to 0.40.

The Type-II error would occur if we concluded, based on the sample data, that the proportion of customers preferring the new service plan was less than or equal to 0.40 when, in fact, it was greater than 0.40.

It is necessary to balance the two types of errors. As a compromise, α is often set at 0.05; sometimes it is 0.01; other values of α are rare. We will consider 0.05 for our example.

***Step4: Collect Data and Calculate Test Statistic:***

Sample size is determined after taking into account the desired α and other qualitative considerations, such as budget constraints to collect the sample data. For our example, let's say, 30 users were surveyed and 17 indicated that they used the Internet for shopping.

Thus, the value of the sample proportion is *p=17/30=0.567.*

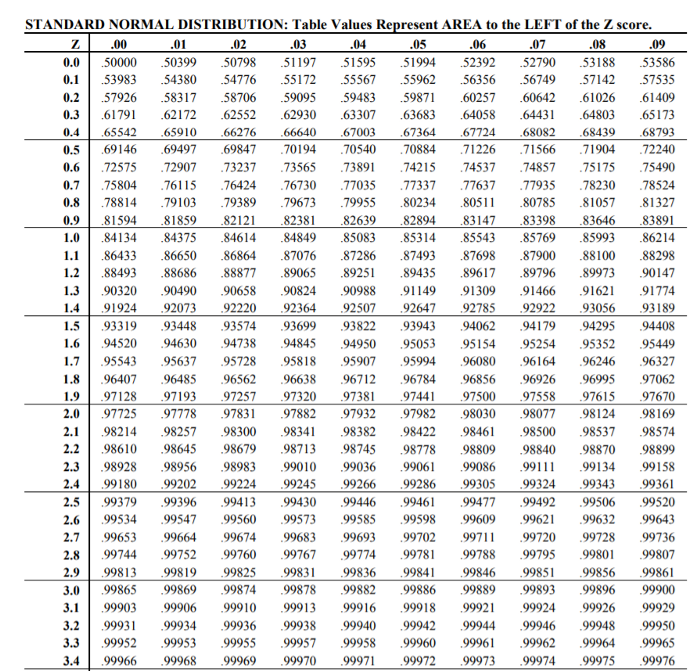
The value of *σₚ=sqrt((0.40)(0.60)/30)=0.089.*

The test statistic *z*can be calculated as

*z=(p-π)/σₚ=(0.567–0.40)/0.089=1.88*

**Step5: Determine the Probability (or Critical Value):**

***Step5: Determine the Probability (or Critical Value):***



Using standard normal tables from the above, the probability of obtaining a *z*value of 1.88 is 0.96995 i.e. **P(z≤1.88)=0.96995***.*But we wanted to calculate the probability to the right of *z (because we are interested in obtaining the probability value that falls in the rejection region or critical region),*i.e. **1–0.96995**=**0.03005**. This Probability is directly comparable to α *(since α is committing a Type-I error and the probability value that we calculated also falls in the critical region)*

If you wanted to understand how to look up for the probability values for the given z scores, please watch below video:

Alternatively, the critical value of *z,*which will give an area to the right side of the critical value of 0.05, is between 1.64 *(at 1.64 the probability is 0.94950)* and 1.65 *(at 1.65 the probability is 0.95053)* and equals 1.645 *(the probability is 0.95, i.e. from the left of the normal distribution, which means to the right it is 0.05)*.

Note that in determining the critical value of the test statistic, the area in the tail beyond the critical value is either *α or α/2.*It is *α*for a one-tailed test and *α/2*for a two-tailed test. Our example is a one-tailed test.

***Step 6 and 7: Compare the probability (or Critical value) and make the decision:***

The probability associated with the calculated or observed value of the test statistic is 0.03005. This is the probability of getting a ***P-Value***of 0.567 (sample proportion = *p)*when π=0.40. This is less than the level of significance of 0.05. Hence, the null hypothesis is rejected.

Alternatively, the calculated value of the test statistic ***z=1.88***lies in the rejection region, beyond the value of 1.645. Again, the same conclusion to reject the null hypothesis is reached.

Note that two ways of testing the null hypothesis are equivalent but mathematically opposite in the direction of comparison. If the probability associated with the calculated or observed value of the test statistic (TSCAL) is ***less than***the level of significance (α), the null hypothesis is rejected. However, if the absolute value of the calculated value of the test statistic is ***greater than***the absolute value of the critical value of the test statistic (TSCR), the null hypothesis is rejected. The reason for this sign shift is that the larger the absolute value of TSCAL, the smaller the probability of obtaining a more extreme value of the test statistic under the null hypothesis.

*if the probability of TSCAL < significance level (α), then reject H₀.*

*But, if |TSCAL| > |TSCR|, then reject H₀*

***Step8: Conclusion:***

In our example, we conclude that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the recommendation to the department store would be to introduce the new Internet shopping service.

This example refers to one sample test of proportions. However, there are several types of tests exist depends on the knowledge about the population and the problem at hand.

For Example, We have a t-test, Z-test. Chi-Square Test, Mann-Whitney Test, Wilcoxon Test, etc.

