

Everything You Need To Know about Hypothesis Testing — Part I

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

Data Science field is evolving like never before. Many companies are now looking for professionals who can sift their goldmine data and help them drive swift business decisions efficiently. It also gives the edge to many working professionals to switch their careers to the Data Science field.

Having this AI, Data Science buzz around many college students also wants to pursue their careers in the Data Science field. And this hype around Data Science is correctly proclaimed by Thomas H. Davenport and D.J. Patil in one of the Harvard Business Review articles that,

“Data Scientist: The Sexiest Job of the 21st Century”

In today's analytics world building machine learning models has become relatively easy (thanks to more robust and flexible tools and algorithms), but still the fundamental concepts are very confusing. One of such concepts is Hypothesis Testing.

In this post, I'm attempting to clarify the basic concepts of Hypothesis Testing with illustrations.

What is Hypothesis Testing? What are we trying to achieve? Why do we need to perform Hypothesis Testing? We must know the answers to all these questions before we proceed.

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.

There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery — Enrico Fermi

Let's look at 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 (σ^2), 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 (\bar{x}), Sample Variance (S^2), 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 $20C_5$ 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. $E(\bar{x}) = \mu$

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 (H_0):

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 (H_1):

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.

To know more about Null and Alternate hypotheses please watch this below video

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:

Hypothesis test	Test statistic
Z-test	Z-statistic
t-tests	t-statistic
ANOVA	F-statistic
Chi-square tests	Chi-square statistic

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.

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 $\bar{x} = 25$

$$H_0: \mu = 20$$

$$H_1: \mu > 20$$

The interpretation of the p-value as follows:

We saw above that α is also known as committing **Type-I** error. When we say $\alpha=5\%$, 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 ($\bar{x} \geq 25$) given that H_0 is true is very less. In other words, we can't obtain our sample statistic as long as we assume H_0 is true. Hence, we reject H_0 and accept H_1 . 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 H_0 , 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.

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:

$$H_0: \pi \leq 0.40$$

$$H_1: \pi > 0.40$$

If the null hypothesis H_0 is rejected, then the alternative hypothesis H_1 will be accepted and the new Internet shopping service will be introduced. On the other hand, if we fail to reject H_0 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.

$$z = (p - \pi) / \sigma_p, \text{ where } \sigma_p = \sqrt{\pi(1 - \pi) / n}$$

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 $\sigma_p = \sqrt{((0.40)(0.60)/30)} = 0.089$.

The test statistic z can be calculated as

$$z = (p - \pi) / \sigma_p = (0.567 - 0.40) / 0.089 = 1.88$$

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. $P(z \leq 1.88) = 0.96995$. 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 $\alpha/2$. It is α for a one-tailed test and $\alpha/2$ for a two-tailed test. Our example is a one-tailed test.

If you wanted to understand how to look up for the critical value of α , Please watch below video:

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 $\pi = 0.40$. 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 H_0 .

But, if $|TSCAL| > |TSCR|$, then reject H_0

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.