**Hypothetical Testing**

**Hypothesis Testing in Machine Learning**

The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample. The same inferences are drawn for different machine learning models through **T-test** .

For drawing some inferences, we have to make some assumptions that lead to two terms that are used in the hypothesis testing.

**Null hypothesis**: It is regarding the assumption that there is no anomaly pattern or believing according to the assumption made.

**Alternate hypothesis**: Contrary to the null hypothesis, it shows that observation is the result of real effect.

**P value:**It can also be said as evidence or level of significance for the null hypothesis or in machine learning algorithms. It’s the significance of the predictors towards the target.

Generally, we select the level of significance by 5 %, but it is also a topic of discussion for some cases. If you have a strong prior knowledge about your data functionality, you can decide the level of significance.

On the contrary of that if the p-value is less than 0.05 in a machine learning model against an independent variable, then the variable is considered which means there is heterogeneous behavior with the target which is useful and can be learned by the machine learning algorithms.

The steps involved in the hypothesis testing are as follow:

Assume a null hypothesis, usually in machine learning algorithms we consider that there is no anomaly between the target and independent variable.

Collect a sample

Calculate test statistics

Decide either to accept or reject the null hypothesis

Calculating test or T statistics

For Calculating T statistics, we create a scenario.

Suppose there is a shipping container making company which claims that each container is 1000 kg in weight not less, not more. Well, such claims look shady, so we proceed with gathering data and creating a sample.

After gathering a sample of 30 containers, we found that the average weight of the container is 990 kg and showing a standard deviation of 12.5 kg.

So calculating test statistics:

T = (Mean - Claim)/ (Standard deviation / Sample Size^(1/2))

Which is -4.3818 after putting all the numbers.

Now we calculate t value for 0.05 significance and degree of freedom.

Note: Degree of Freedom = Sample Size - 1

From T table the value will be -1.699.

After comparison, we can see that the generated statistics are less than the statistics of the desired level of significance. So we can reject the claim made.

You can calculate the t value using stats.t.ppf() function of stats class of scipy library.

**Errors**

As hypothesis testing is done on a sample of data rather than the entire population due to the unavailability of the resources in terms of data. Due to inferences are drawn on sample data the hypothesis testing can lead to errors, which can be classified into two parts:

Type I Error: In this error, we reject the null hypothesis when it is true.

Type II Error: In this error, we accept the null hypothesis when it is false.

Other Approaches

A lot of different approaches are present to hypothesis testing of two models like creating two models on the features available with us. One model comprises all the features and the other with one less. So we can test the significance of individual features. However feature inter-dependency affect such simple methods.

In regression problems, we generally follow the rule of P value, the feature which violates the significance level are removed, thus iteratively improving the model.

Different approaches are present for each algorithm to test the hypothesis on different features.

If you would like to learn more about Bayesian inferences fundamentals, take DataCamp's Fundamentals of Bayesian Data Analysis in R course.

**Other Example**:

We toss a coin. now we are sure that either it will be heads or tails. We do not assume that the coin is bias towards head or tails.

First Toss : Heads appeared ! Now still this does not change our assumption that our coin is not biased.

Second Toss : Heads appeared ! Now still you are not sure that coin is biased as it is very likely to land up in heads again.

Now Hundredth toss : Head appears. Now you start being skeptical about that coin might actually be biased and you were wrong in first place that con wasn’t biased.

Now,

The assumption that you made that coin isn’t biased is called as **Null Hypothesis** and the assumption where you assume coin to be biased is **Alternate Hypothesis**.

**P-value h**elps us to switch between these two hypotheses. There is some significant score of 5%. And if your P-value is lesser than this significant value then you choose to reject your null hypothesis and go along with alternate hypothesis.

But how it is related to machine learning ?

We tend to go with alternate hypothesis as the results will be something we don’t expect off. This will give a meaning inferential capability to your model to learn nicely.

Now consider a case where you are stuck with null hypothesis ( i.e. you are stuck with things that you already know) , this will not help our machine learning model to generalize better or perform better.

**For ex:------**

In Statistics, a hypothesis is accepted or rejected based on the observations. Generally, we have two hypothesis H0and H1

H0:The new drug performs the same as the previous one.

H1:The new drug performs better than the previous one.

It is to be noted that, one of the hypothesis is accepted or rejected in favor of the other.

p−value --lies between 0and 1.It indicates the credibility of the H0.

Ex: if the p-value is 0.05, then you have a 5% chance of making error when you reject H0.

**T-Test:**

It is a parametric test which tells you how significant the differences between groups are; In other words, it lets you know if those differences (measured in means/averages) could have happened by chance.

## Chi-Square Test:

## Normal Deviate Z Test:

T-tests are called so, because the test results are all based on t-values.

A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the probability of difference between two sets of data.

## T values : T test Statistics

T-values are an example of test statistics. A test statistic is a standardized value that is calculated from sample data during a hypothesis test. The procedure that calculates the test statistic compares your data to what is expected under the null hypothesis.

To perform a t-test calculation we require three key data values.

* The Difference between the mean values from each data set (called the mean difference),
* The standard deviation of each group
* The number of data values of each group.

## As per Investopedia:

The t-value is a ratio of the difference between the mean of the two sample sets and the difference that exists within the sample sets. While the numerator value (the difference between the mean of the two sample sets) is straightforward to calculate, the denominator (the difference that exists within the sample sets) can become a bit complicated depending upon the type of data values involved.

The Denominator in t value measures, how the data is the dispersion or variability.

Higher values of the t-value, also called t-score, indicate that a large difference exists between the two sample sets

## Degrees of Freedom In T Test:

Degrees of freedom refers to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

## T-Test Assumptions:

1. The first assumption is concerned with the scale of measurement. Here assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale.
2. The second assumption is regarding simple random sample. The Assumption is that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The fourth assumption is a that reasonably large sample size is used for the test. Larger sample size means the distribution of results should approach a normal bell-shaped curve.
5. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

# What Are The Types Of T Tests & How To Select One Of Them ?

# There are three types of t-test:



