Aim: To understand how to perform Hypothesis Testing using python

1.What is Null hypothesis?

A null hypothesis is a statement or assumption that there is no statistical significance or difference between two or more variables or groups being analyzed. It is a fundamental concept in statistics and scientific research, where it is used to test hypotheses and determine the likelihood of a relationship or effect. The null hypothesis is typically denoted as H0 and is often formulated as a null or default hypothesis that the researchers would like to reject when they perform a hypothesis test. In simpler terms, it is a hypothesis which assumes that there is no relationship or difference between variables that is being investigated.

2.What is alternate hypothesis?

An alternative hypothesis Is a statistical hypothesis that represents an alternative explanation for the relationship or effect being analyzed, in contrast to the null hypothesis. It proposes that there is a significant difference between two or more variables or groups being studied, rather than assuming that there is no difference, which is what the null hypothesis states. In hypothesis testing, the alternative hypothesis is used to evaluate the probability of the observed data occurring by chance alone, and it is typically what is being tested against the null hypothesis. The alternative hypothesis is usually denoted as H1 or Ha.

3.Define Type 1 and Type 2 errors.

In statistics and hypothesis testing, Type I and Type II errors are two different kinds of mistakes that can occur when evaluating a null hypothesis.

-Type I error, also known as a false positive, occurs when we reject a null hypothesis that is actually true . Its probability is denoted by alpha and is controlled using the significance level of a hypothesis test.

-Type II error, also known as a false negative , occurs when we fail to reject a null hypothesis that is actually false. Its probability is denoted by beta and is related to statistical power that is the ability to detect a true difference.

The probability of Type I and Type II errors are inversely proportional; that is, as we decrease the probability of Type I error, we increase the probability of Type II error and vice versa.

4.When to go for Anova instead of t-test or chi square test?

ANOVA (Analysis of Variance) is a statistical test that is used to compare the means of three or more groups or samples, whereas t-tests and chi-square tests are statistical tests that compare the means of two groups or test for association between two categorical variables, respectively.

5.What role does Anova play in a ML Project ppipeline?

In a machine learning project pipeline, ANOVA (Analysis of Variance) can be used as a feature selection method. Feature selection is an important step in the machine learning process, where we choose the most useful subset of features from the original set of features for our model. ANOVA can be used to compare the means of three or more groups or samples, and can help us determine which features are statistically significant in predicting the target variable. We can use the F-statistic and p-value from the ANOVA test to rank the features according to their importance and select the top features for our model. By doing so, we can improve the performance of our model by reducing the dimensionality of the feature space and removing redundant or irrelevant features from our data.