UNIT II	**Linear Algebra & Statistical Modeling for Data Science**	**06 - Hrs**
	Hypotheses Testing, Type 1 and Type 2 errors. Introduction to the chisquared test. The concept of p-value.	

Recap

- Intro. Statistical Modeling
- Types of Statistical modeling
- What is sampling, population, event, trail etc.
- Reason for choose Statstical Modeling in DS.
- Techniques in Stastical Modeling.
- What is Probability Function.
- What is Probability Distribution etc.
- Random Variable etc.

Why hypothesis testing ?

- **Hypothesis testing** is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population.
- The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favor of a certain belief, or hypothesis, about a parameter.

  Examples:

  1. Is there statistical evidence, from a random sample of potential customers, to support the hypothesis that more than 10% of the potential customers will purchase a new product?

  2. Is a new drug effective in curing a certain disease?
- A sample of patients is randomly selected. Half of them are given the drug while the other half are given.The conditions of the patients are then measured and compared.

Steps Involved in Hypothesis Testing

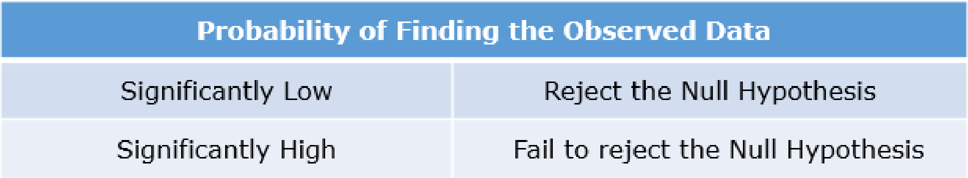
- In statistics, hypothesis is a statement about the population which is represented by some numerical values.
- Hypothesis testing deals with collecting enough evidence about the hypothesis. Then, based on the evidence collected, the test either accepts or rejects the hypothesis about the population.
- **For example**, post conducting surprise test for **100 students** and calculating the mean marks, the government proposes a statement/ hypothesis that the estimation of surveying agency may not be accurate.
- Hypothesis testing needs to be performed to find evidence in support of this hypothesis. Based on the evidence found, this hypothesis can be accepted or rejected.
- Let us understand a few terminologies and concepts used in hypothesis testing.

Null Hypothesis and Alternate Hypothesis

- Each hypothesis test usually includes two competing hypotheses about the population. They are Null Hypothesis and Alternate Hypothesis.
- **Null Hypothesis (H_0):** It is a statement/hypothesis that disapproves/rejects the observation based on which the hypothesis is made. You can start the hypothesis testing considering the null hypothesis to be true. It cannot be rejected until there is an evidence which suggests otherwise.
- **Alternate Hypothesis (H_a):** It is a statement/hypothesis which is contradictory to the null hypothesis. If you find enough evidence to reject the null hypothesis, then the alternate hypothesis is accepted.

Rejecting/Accepting the Null Hypothesis

- In hypothesis testing, assume the null hypothesis (which disapproves the observed data) to be true and find the probability of finding the observed data. Based on the probability of finding the observed data, accept/reject the null hypothesis. The condition is given in the table below.



Probability of Finding the Observed Data	
Significantly Low	Reject the Null Hypothesis
Significantly High	Fail to reject the Null Hypothesis

The distinction between the probability being significantly high or low is determined by Level Of Significance represented by the symbol **'α'.**

Usually, the level of significance is set as **0.05.** Assume that the null hypothesis is true. A level of significance **(of 0.05) i**mplies that:

Steps Involved in Hypothesis Testing

- you can reject the null hypothesis if the probability of occurrence of the given data is less than the level of significance (0.05 in this case)
- you fail to reject the null hypothesis if the probability of occurrence of the given data is greater than or equal to the level of significance (0.05 in this case)
- With this, let us see how the surveying agency performs hypothesis test to validate its estimations.

  Step 1: Define the null hypothesis, alternate hypothesis and the level of significance.

  Step 2: Calculate the probability of getting the observed data (P value) assuming the null hypothesis to be true. This involves two intermediate steps as given below.

  Step 2.1: Calculate the Z statistic.

  Step 2.2: Calculate the P value.

  Step 3: Conclude whether to reject the null hypothesis or not based on the P value i.e.

  -> If P value < significance level, then reject the null hypothesis

  ->If P value >= significance level, the null hypothesis cannot be rejected

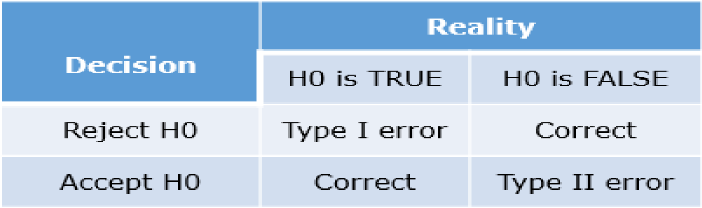
- **Step 4:** State the conclusion.

Errors in Hypothesis Testing

- While conducting hypothesis tests, two types of errors can be encountered. These errors are termed as:

  1. Type I

  2. Type II



Decision	Reality	
	H0 is TRUE	H0 is FALSE
Reject H0	Type I error	Correct
Accept H0	Correct	Type II error

Accepting or rejecting hypothesis is probabilistic. Therefore, you may fail to reject a hypothesis when it is false and reject it when it is true. These error conditions have special names.

- Type **I error** is said to occur when a null hypothesis is rejected when it is true.
- Type **II error** is said to occur when you fail to reject a null hypothesis when it is false.

Intro Chi-Square Test

- A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.
- Therefore, a chi-square test is an excellent choice to help us better understand and interpret the relationship between our two categorical variables.

1. Is a Chi-square test the same as a χ^2 test?

Ans: Yes, χ is the Greek symbol Chi.

2. What are my choices?

Ans: If you have a single measurement variable, you use a Chi-square goodness of fit test. If you have two measurement variables, you use a Chi-square test of independence. There are other Chi-square tests, but these two are the most common.

Formula For Chi-Square Test

$$x_c^2 = \frac{\Sigma \, (O_i - E_i)^2}{E_i}$$

- **Where,**

  c = Degrees of freedom

  O = Observed Value

  E = Expected Value
- The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation. The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid. These tests are frequently used to compare observed data with data that would be expected to be obtained if a particular hypothesis were true.

Types of Chi-square tests

- You use a Chi-square test for hypothesis tests about whether your data is as expected. The basic idea behind the test is to compare the observed values in your data to the expected values that you would see if the null hypothesis is true.
- There are two commonly used Chi-square tests:

  1. The Chi-square goodness of fit test.

  2. The Chi-square test of independence.
- Both tests involve variables that divide your data into categories. As a result, people can be confused about which test to use.

How to perform a Chi-square test:

- For both the Chi-square goodness of fit test and the Chi-square test of independence, **you perform the same analysis steps, listed below.**

1. Define your null and alternative hypotheses before collecting your data.
2. Decide on the alpha value. This involves deciding the risk you are willing to take of drawing the wrong conclusion. For example, suppose you set α=0.05 when testing for independence. Here, you have decided on a 5% risk of concluding the two variables are independent when in reality they are not.
3. Check the data for errors.
4. Check the assumptions for the test. (Visit the pages for each test type for more detail on assumptions.)
5. Perform the test and draw your conclusion.

What Is P-Value?

- In statistics, the **p-value** is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to or in addition to preselected confidence levels for hypothesis testing.