# (MR20-1CS0306) DATA ANALYTICS WITH PYTHON

## III Year B. Tech – II Semester

**Dr.M.Narayanan**
**Professor / CSE**

**COURSE OBJECTIVES:**

- ❖ Learning about the Importance of Data and its importance
- ❖ Knowing Python fundamentals and Pandas essentials
- ❖ Learning the Principles of Probability and sampling Methods
- ❖ Getting knowledge about formulating and testing hypothesis
- ❖ Learning and analytical comparison's with ANOVA methods
- ❖ Learning about Performance indicators using ROC methods

# UNIT-III

## UNIT-III

**Hypothesis testing: Importance of Hypothesis testing, null and alternative hypotheses, Type-I and Type –II errors, approaches to Hypothesis testing, two sample testing**

### Reference Books:


1. McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".
2. Swaroop, C. H. (2003). A Byte of Python. Python Tutorial.
3. Ken Black, sixth Editing. Business Statistics for Contemporary Decision Making. “John Wiley & Sons, Inc”.
4. Anderson Sweeney Williams (2011). Statistics for Business and Economics. “Cengage Learning”.

## ❖ Hypothesis Testing

- ❖ Hypothesis testing is a statistical method used to make inferences about a population based on sample data.
- ❖ ***Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population***
- ❖ Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- ❖ **Example: You say an average height in the class is 30 or a boy is taller than a girl. All of these is an assumption that we are assuming, and we need some statistical way to prove these. We need some mathematical conclusion whatever we are assuming is true.**
- ❖ It provides a systematic framework for evaluating research hypotheses and determining the likelihood of observed results occurring due to chance.
- ❖ A hypothesis enables researchers not only to discover a relationship between variables, but also to predict a relationship based on theoretical guidelines and/or empirical evidence.

## General Information

- **Null Hypothesis ($H_0$)** and **Alternative Hypothesis (Ha / $H_1$)**.
- The **null hypothesis** represents the **default or conservative (traditional ) position**, assuming no significant difference or effect in the population.
- The **alternative hypothesis** contradicts the null hypothesis and suggests that there is a significant difference or effect.
- **Examples:**
- **Null Hypothesis: H0: There is no difference in the salary of factory workers based on gender.**
- **Alternative Hypothesis: Ha: Male factory workers have a higher salary than female factory workers.**
- NOTE: You want to reject the null hypothesis, but how and when can you do that?
- To start, you'll need to perform a statistical test on your data. Then has to decide the Null Hypothesis is rejected.

## Process of Hypothesis Testing

The process of hypothesis testing involves the following steps:

- Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis (Ho) and alternate hypothesis (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

**Step 1: State your null and alternate hypothesis**

- After developing your initial research hypothesis (the prediction that you want to investigate), it is important to restate it as a null (Ho) and alternate (Ha) hypothesis so that you can test it mathematically.
- The alternate hypothesis is usually your initial hypothesis that predicts a relationship between variables. The null hypothesis is a prediction of no relationship between the variables you are interested in.

**Hypothesis testing example**

- You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:
- H0: Men are, on average, not taller than women.
- Ha: Men are, on average, taller than women.

**Step 2: Collect data**

- For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis.
- If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

**Hypothesis testing example**

- To test differences in average height between men and women, your sample should have an equal proportion of men and women, and cover a variety of socio-economic classes and any other control variables that might influence average height.
- You should also consider your scope (Worldwide? For one country?)
- A potential data source in this case might be census data, since it includes data from a variety of regions and social classes and is available for many countries around the world.

**Step 3: Perform a statistical test**

- There are a variety of statistical tests available, but they are all based on the comparison of **within-group variance** (how spread out the data is within a category) versus **between-group variance** (how different the categories are from one another).
- If the **between-group variance** is substantial enough to result in minimal or no overlap between groups, your statistical test will indicate this with a **low p-value**. This implies that the observed differences between these groups are unlikely to have occurred by chance.
- Alternatively, If there is **high within-group variance** and **low between-group variance**, your statistical test will be characterized by a **high p-value**. This suggests that any observed difference between groups is likely attributable to chance. The selection of your statistical test should be guided by the nature of variables and the level of measurement in your collected data.

**The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.**

**Hypothesis testing example**

- Based on the type of data you collected, you perform a one-tailed t-test to test whether men are in fact taller than women. This test gives you:
- an estimate of the difference in average height between the two groups.
- a p-value showing how likely you are to see this difference if the null hypothesis of no difference is true.
- Your t-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2 cm to infinity. The p-value is 0.002.

**Step 4: Decide whether to reject or fail to reject your null hypothesis**

- Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.
- In most cases you will use the p-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.
- In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

**Hypothesis testing example**

- In your analysis of the difference in average height between men and women, you find that the p-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.

**Step 5: Present your findings**

- The results of hypothesis testing will be presented in the results and discussion sections of your research paper, dissertation or thesis.
- In the results section you should give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated p-value). In the discussion, you can discuss whether your initial hypothesis was supported by your results or not.
- In the formal language of hypothesis testing, we talk about rejecting or failing to reject the null hypothesis. You will probably be asked to do this in your statistics assignments.

**Stating results in a statistics assignment**

- In our comparison of mean height between men and women we found an average difference of 13.7 cm and a p-value of 0.002; therefore, we can reject the null hypothesis that men are not taller than women and conclude that there is likely a difference in height between men and women.

## Importance of Hypothesis Testing

- According to the San Jose State University Statistics Department, hypothesis testing is one of the most important concepts in statistics because it is how you decide if something really happened,
- or if certain treatments have positive effects,
- or if groups differ from each other
- or if one variable predicts another.
- **In short, you want to proof if your data is statistically significant and unlikely to have occurred by chance alone. In essence then, a hypothesis test is a test of significance.**
- Hypothesis testing allows the researcher to determine whether the data from the sample is statistically significant.
- Hypothesis testing is one of the most important processes for measuring the validity and reliability of outcomes in any systematic investigation

❖ **Here are some key reasons why hypothesis testing is important:**

❖ **Testing Research Questions:**

❖ Hypothesis testing helps researchers test specific research questions or hypotheses.

❖ By formulating clear hypotheses, researchers can make predictions about the relationships between variables or the effects of interventions.

❖ Hypothesis testing provides a framework to evaluate these predictions objectively.

❖ **Statistical Inference:**

❖ Hypothesis testing enables researchers to draw conclusions about a population using sample data.

❖ By analyzing the sample data, researchers can determine whether the observed results are likely to occur due to chance or if they represent a true effect in the population.

❖ This allows for generalization from a smaller sample to a larger population.

❖ **Decision-Making:**

❖ Hypothesis testing provides a structured decision-making process. Researchers can define a null hypothesis ($H_0$), which represents no effect or no difference, and an alternative hypothesis ($H_1$), which represents an effect or difference.

❖ By comparing the sample data to the null hypothesis, researchers can make decisions such as rejecting the null hypothesis in favor of the alternative or failing to reject the null hypothesis due to insufficient evidence.

❖ **Objectivity and Reliability:**

❖ Hypothesis testing offers an objective approach to evaluating research hypotheses.

❖ It allows researchers to rely on statistical evidence rather than personal opinions or biases.

❖ The procedures followed in hypothesis testing are well-defined and can be replicated by other researchers, promoting transparency and the advancement of scientific knowledge.

## ❖ Null Hypotheses and Alternative Hypotheses

- ❖ In statistics and hypothesis testing, the **null hypothesis (H0)** and **alternative hypothesis (H1 or HA)** are two challenging statements about a population or a phenomenon under investigation.
- ❖ These hypotheses are used to make inferences and draw conclusions based on the available data.
- ❖ The **null hypothesis (H0)** is a statement of no effect or no relationship between variables.
- ❖ It represents the current situation or the commonly accepted belief before conducting a study or experiment.
- ❖ In other words, it assumes that any observed differences or effects in the data are due to chance or random variation.
- ❖ The null hypothesis is typically denoted as a statement of equality, such as "there is no difference," "there is no effect," or "the two variables are independent."

- The alternative hypothesis (H1 or HA) is a statement that contradicts the null hypothesis.
- It represents the researcher's or analyst's claim or the desired outcome.
- The alternative hypothesis suggests that there is a specific effect, relationship, or difference between variables, beyond what would be expected by chance alone.
- It is often denoted as a statement of inequality, such as "there is a difference," "there is an effect," or "the variables are dependent."
- When conducting a hypothesis test, the goal is to gather evidence from the data to support or reject the null hypothesis in favor of the alternative hypothesis.
- This is done by analyzing the data and calculating a test statistic (e.g., t-test, chi-square test, etc.) that quantifies the difference or effect observed.
- The test statistic is then compared to a critical value or p-value to determine the level of evidence against the null hypothesis

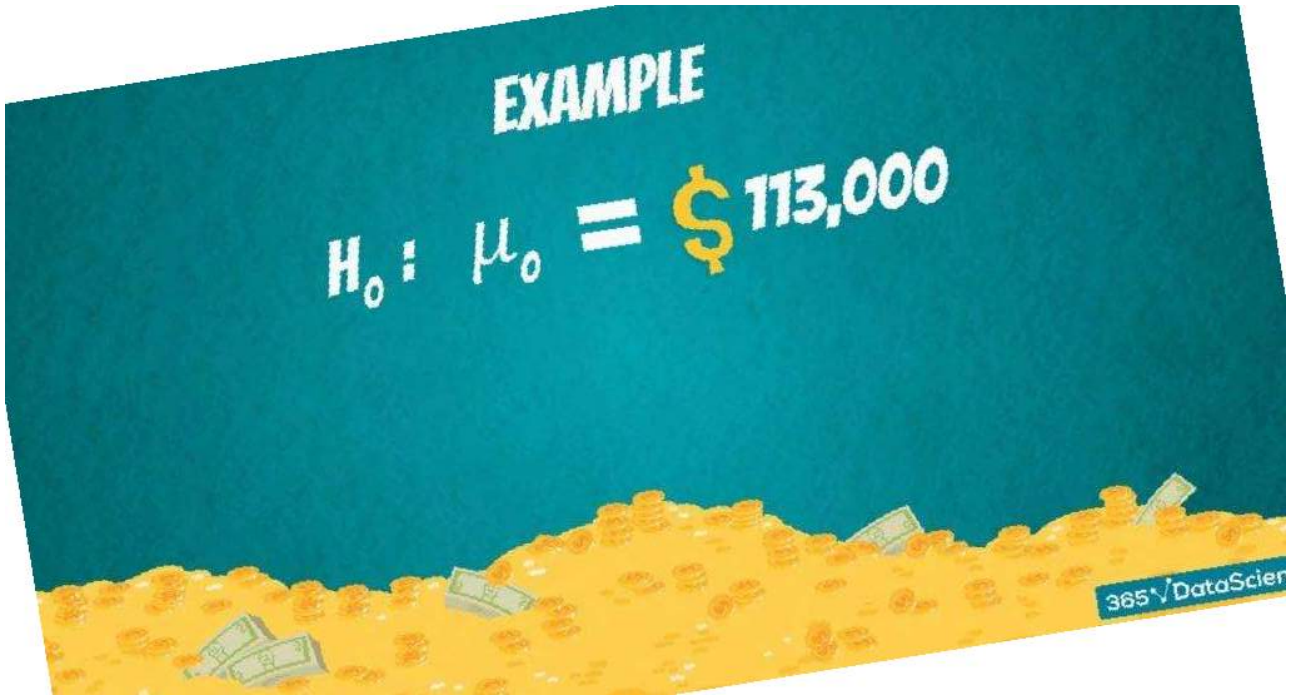
- If the test statistic falls in the critical region or the p-value is smaller than a predetermined significance level (e.g., 0.05), the null hypothesis is rejected in favor of the alternative hypothesis.
- This suggests that the observed data provide strong evidence against the null hypothesis and support the claim made in the alternative hypothesis.
- On the other hand, if the test statistic does not fall in the critical region or the p-value is larger than the significance level, there is insufficient evidence to reject the null hypothesis, and the alternative hypothesis is not supported.
- **It's important to note that failing to reject the null hypothesis does not necessarily prove it to be true; it simply suggests that there is not enough evidence to support the alternative hypothesis.**
- **Additionally, the choice of null and alternative hypotheses depends on the research question, the study design, and the desired outcome of the analysis.**

- **Let's consider an example related to a new drug being tested for its effectiveness in treating a particular medical condition.**
- **Null hypothesis (H0): The new drug has no effect on the medical condition.**
- **Alternative hypothesis (H1): The new drug has a positive effect on the medical condition.**
- In this example, the null hypothesis states that the new drug has no effect, meaning that the treatment does not lead to any improvements in the condition being studied.
- The alternative hypothesis, on the other hand, suggests that the new drug does have a positive effect, meaning that the treatment does result in improvements.
- To test these hypotheses, a clinical trial is conducted. A group of patients with the medical condition is randomly divided into two groups:
- **One group receives the new drug (treatment group) and the other receives a placebo or an existing standard treatment (control group).**
- The researchers then collect data on the participants' health outcomes, such as symptom severity or disease progression.

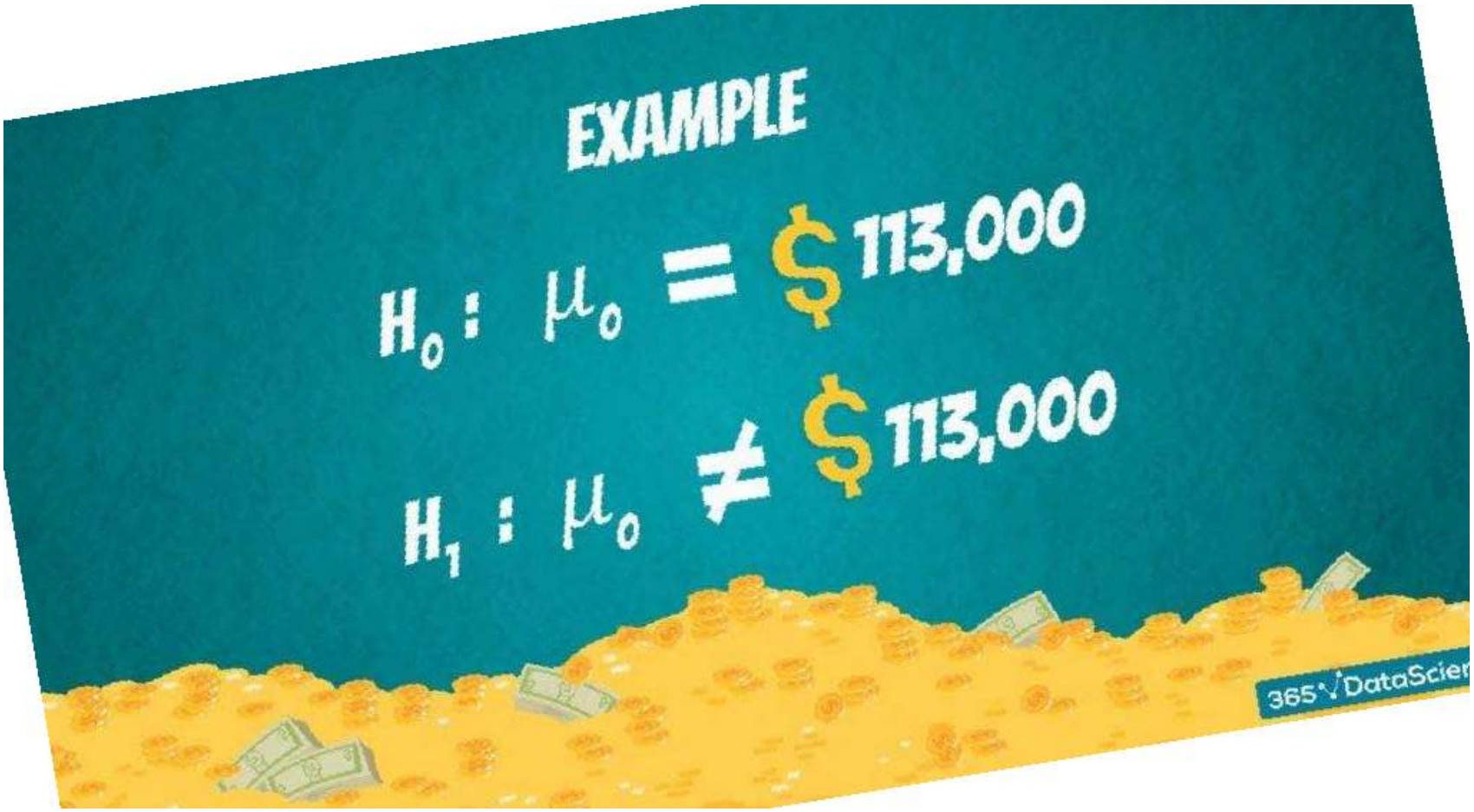
- After analyzing the data, the researchers compare the outcomes between the treatment and control groups.
- They calculate a test statistic, such as a t-statistic or a chi-square statistic, which quantifies the observed differences between the groups. The test statistic is then compared to a critical value or p-value to make a conclusion.
- If the test statistic falls in the critical region or the p-value is below a predetermined significance level (e.g., 0.05), the null hypothesis is rejected in favor of the alternative hypothesis.
- This would suggest that the new drug does have a positive effect on the medical condition, providing evidence to support its effectiveness.

- On the other hand, if the test statistic does not fall in the critical region or the p-value is above the significance level, there is insufficient evidence to reject the null hypothesis.
- This means that the data did not provide enough evidence to support the claim that the new drug has a positive effect on the medical condition.
- **Note:**
- **The null hypothesis in this example assumes that the new drug has no effect, while the alternative hypothesis suggests that the drug does have a positive effect.**
- **The goal is to gather evidence from the data to support or reject the null hypothesis in favor of the alternative hypothesis, using statistical tests and analyses.**

- The null hypothesis is the one to be tested and the alternative is everything else. In our example:
- The null hypothesis would be: The mean data scientist salary is 113,000 dollars.

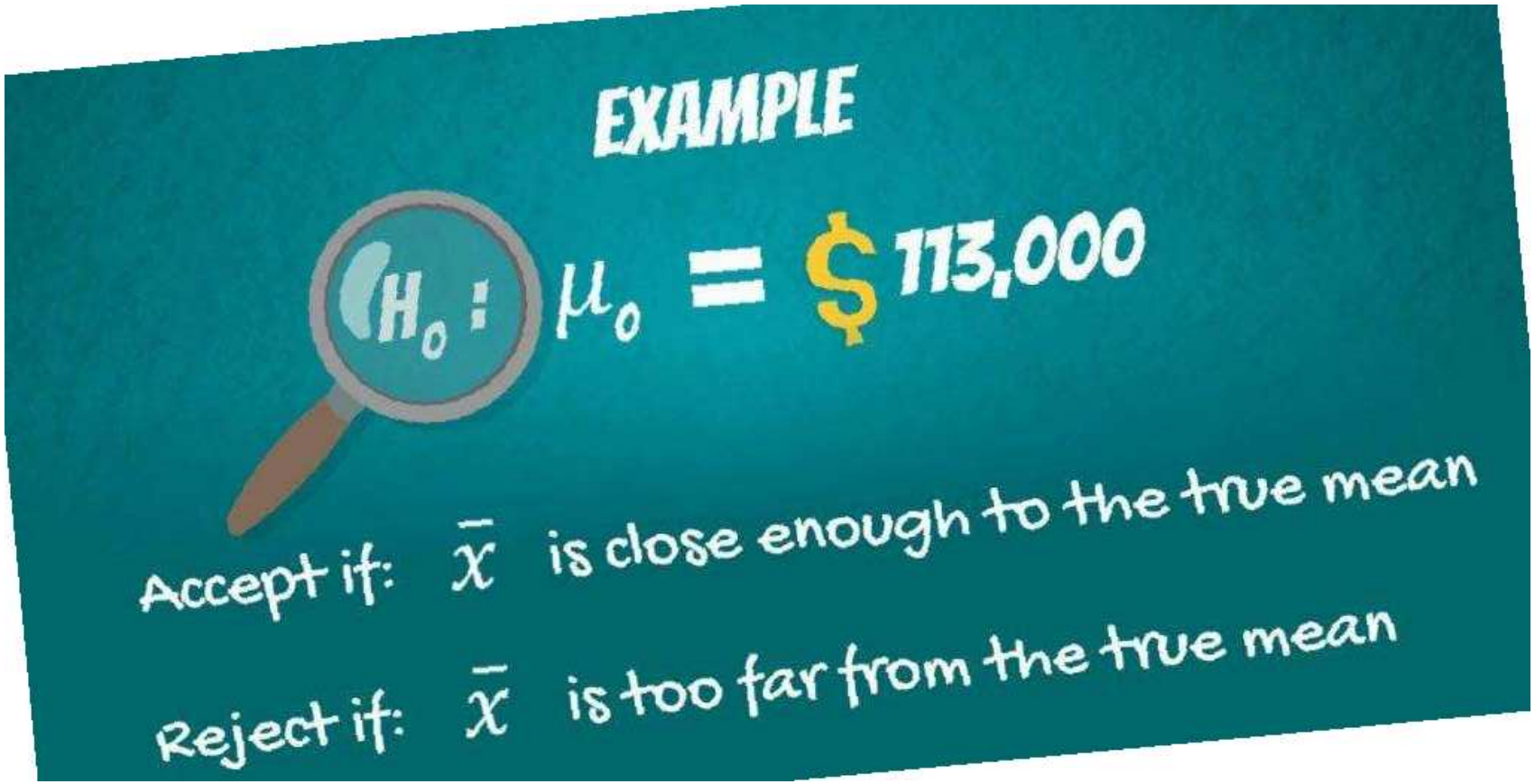


❖ While the alternative: The mean data scientist salary is not 113,000 dollars



**The Concept of the Null Hypothesis**

❖ Now, you would want to check if 113,000 is close enough to the true mean, predicted by our sample. In case it is, you would accept the null hypothesis. Otherwise, you would reject the null hypothesis.



❖ The concept of the null hypothesis is similar to: innocent until proven guilty. We assume that the mean salary is 113,000 dollars and we try to prove otherwise.

## TYPE-I ERRORS AND TYPE –II ERRORS

- In statistical hypothesis testing, we evaluate two types of hypotheses: the **null hypothesis (H0)** and the **alternative hypothesis (H1).** In this context, Type I and Type II errors are associated with the decisions made regarding these hypotheses.
- **Type I error (False Positive):**
- A Type I error occurs when the null hypothesis (H0) is incorrectly rejected, even though it is actually true.
- In other words, it is the error of claiming a relationship or effect exists when, in reality, there is no significant evidence to support that claim.
- It represents a false positive result. The probability of committing a Type I error is denoted as alpha (α), which is the predetermined significance level in hypothesis testing.
- A lower significance level reduces the likelihood of Type I errors but increases the chance of Type II errors.
- ***Note:***
- **A test result that indicates that a person has a specific disease or condition when the person actually does not have the disease or condition**

- The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true.
- It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type I error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.
- Example:
- In a clinical trial, the null hypothesis might be that a new drug has no effect, while the alternative hypothesis suggests that the drug is effective.
- Committing a Type I error would mean concluding that the drug is effective when, in reality, it has no significant impact.

- **Example: Medical Testing Scenario:**
- Imagine a new diagnostic test designed to detect a specific disease. The null hypothesis ($H_0$) in this case is that the patient is healthy and does not have the disease.
- Null Hypothesis ($H_0$): The patient is healthy.
- Alternative Hypothesis ($H_1$): The patient has the disease.
- **Type I Error Explanation:**
- A Type I error (False Positive) occurs when the test incorrectly rejects the null hypothesis ($H_0$), indicating that the patient has the disease when, in reality, they are healthy.
- **Outcome:**
- Test Result: Positive for the disease.
- Reality: The patient is actually healthy.
- **Example:**
- Suppose a patient takes the test, and the result comes back positive for the disease. If it's a Type I error:
- The test suggests the presence of the disease (rejects the null hypothesis).
- However, the patient is, in fact, disease-free.

- **Interpretation:**
- In this situation, the medical test has made an error by falsely identifying a healthy individual as having the disease.
- This can lead to unnecessary stress, further testing, and potentially invasive medical procedures for the patient who is, in reality, not afflicted by the disease.
- To minimize the risk of Type I errors, it's crucial to choose appropriate significance levels and conduct difficult validation studies when developing and evaluating medical tests.
- Additionally, healthcare professionals consider the overall context and may conduct further positive tests to reduce the impact of false positives in critical situations.

❖ **Type II error (False Negative):**

❖ A Type II error occurs when the null hypothesis (H0) is erroneously accepted, regardless of the alternative hypothesis (H1) being true.

❖ It is the error of failing to claim a relationship or effect that does exist.

❖ It represents a false negative result. The probability of committing a Type II error is denoted as beta (β).

❖ The complement of beta, known as the power of the test, is equal to (1 - β) and represents the ability to correctly reject the null hypothesis when it is false.

❖ **Note:**

**A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class**

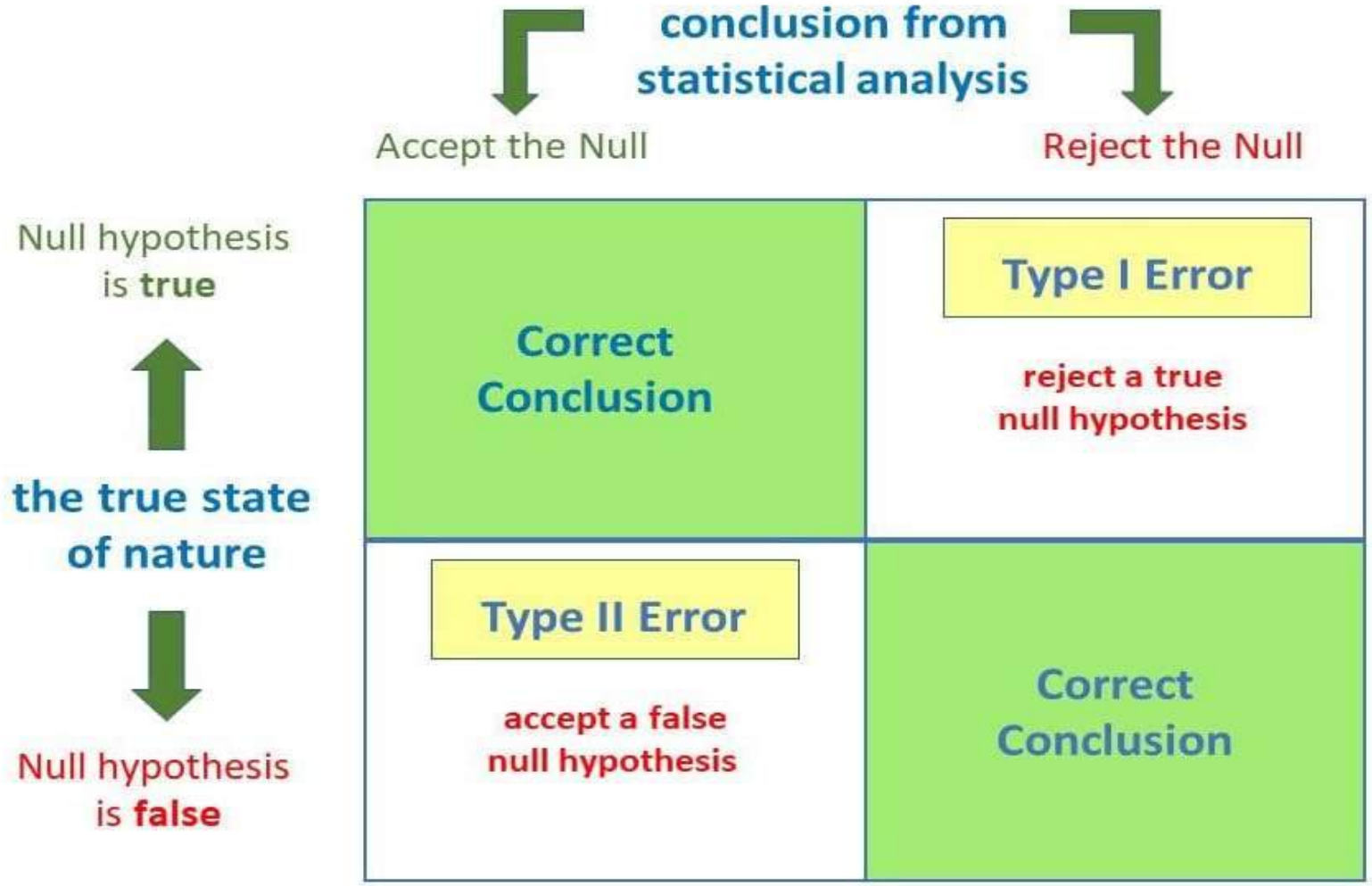
❖ **Example:**

❖ Scenario: A medical test is designed to detect a disease, and the null hypothesis is that the patient is healthy.

❖ Type II Error: The test fails to detect the disease in a patient who actually has it, leading to a false negative result.

- **Example Scenario:**
- Consider a medical test designed to detect a specific disease, let's call it Disease X. The hypotheses in this scenario are:
- Null Hypothesis ($H_0$): The patient does not have Disease X.
- Alternative Hypothesis ($H_1$): The patient has Disease X.
- Type II Error in Detail:
- If the test result fails to detect Disease X in a patient who actually has it, this is a Type II error.
- **Outcome:**
- Test Result: Negative for Disease X.
- Reality: The patient actually has Disease X.
- **Implications and Consequences:**
- A Type II error in medical testing can have significant consequences, as it means the test has failed to identify a real condition. Some key points to understand: Missed Diagnoses, Impact on Patient Health,

- **Minimizing Type II Errors:**
- Increasing Sensitivity: Improving the sensitivity of the test is crucial to reduce the risk of Type II errors. This may involve refining the testing methodology or using more sensitive diagnostic tools.
- Validation and Calibration: Rigorous validation studies are essential to ensure that the test performs accurately across diverse populations and conditions.
- Clinical Judgment: Healthcare professionals often rely on a combination of test results and clinical judgment to make informed decisions, especially when there is suspicion of a condition despite a negative test result.
- **Type II errors in medical testing is critical for healthcare professionals and researchers to enhance the accuracy and reliability of diagnostic procedures, ultimately improving patient outcomes.**


conclusion from statistical analysis
Accept the Null
Reject the Null
Null hypothesis is true
the true state of nature
Null hypothesis is false
Correct Conclusion
Type I Error
reject a true null hypothesis
Type II Error
accept a false null hypothesis
Correct Conclusion

## Approaches to Hypothesis Testing:

### 1. Formulating Hypotheses:

- Hypothesis testing starts with formulating the null hypothesis (H0) and the alternative hypothesis (H1).
- The null hypothesis represents the assumption of no effect or no relationship, while the alternative hypothesis suggests the presence of an effect or relationship.

### 2. Setting the Significance Level:

- The significance level (alpha, α) is predetermined before conducting the test and represents the maximum probability of making a Type I error.
- Commonly used significance levels are 0.05 (5%) and 0.01 (1%).
- The choice of significance level depends on the field of study, consequences of errors, and desired level of confidence.

### 3. Selecting a Statistical Test:

- ❖ Based on the research question and characteristics of the data, an appropriate statistical test is chosen.
- ❖ Common tests include t-tests, chi-square tests, ANOVA, regression analysis, and more.

### 4. Analyzing the Data:

- ❖ Researchers collect a sample of data and perform the chosen statistical test.
- ❖ The test produces a test statistic and a p-value, which are used to make a decision regarding the null hypothesis.

### 5. Comparing p-value and Significance Level:

- ❖ The p-value represents the probability of obtaining the observed data or more extreme data assuming the null hypothesis is true.
- ❖ If the p-value is less than or equal to the significance level ($p \leq \alpha$), we reject the null hypothesis.
- ❖ If the p-value is greater than the significance level ($p > \alpha$), we fail to reject the null hypothesis.

## 6. Interpreting the Results:

- If the null hypothesis is rejected, we may conclude that there is a significant effect or relationship.
- If the null hypothesis is not rejected, we do not have sufficient evidence to support the alternative hypothesis.
- **It is essential to strike a balance between Type I and Type II errors by appropriately selecting the significance level, considering sample size calculations, using suitable statistical methods, and conducting independent replications of studies.**
- **The choice of significance level depends on the specific context, consequences of errors, and desired level of confidence.**

## Student's t-Test

- In the area of statistics, a student's t-test is mentioned as a method of testing the theory about the mean of a small sample drawn from a normally distributed population where the standard deviation of the given population is unknown.
- We can define the Student t-test as a method that tells you how significant the differences can be between different groups. **A Student t-test is defined as a statistic and this is used to compare the means of two different populations.**
- It is a method that is often used in hypothesis testing to find out whether a process or whether a given treatment actually has any effect on the population of interest, or whether or not two populations are different from each other.
- You wish to know whether the mean petal length of iris flowers differs according to their distinct species.
- You find two different species of iris flowers growing in a garden and they measure 25 petals of each species.
- You can test the difference between these two groups with the help of the Student t-test.

- The null hypothesis (H0) is one that tells the true difference between these groups.
- The alternate hypothesis (Ha) is one that tells the true difference is different from zero.

**Student t Test Introduction**

- In the year 1908, an Englishman named **William Sealy Gosset** developed the t-test as well as t distribution.
- **William** worked at the Guinness brewery in Dublin and found which existing statistical techniques using large samples were not useful for the small sample sizes which he encountered in his work).
- The **t distribution** belonging under a family of curves in which the number of degrees of freedom specifies a particular curve.
- As the sample size (and the degrees of freedom) increases, the t distribution approaches the bell shape of the standard normal distribution. In common, for tests involving the mean of a sample of size greater than 30, then the normal distribution is applied.

**Types of Student t-Test**

- When choosing a Student t-test, two things need to be kept in mind: whether the groups being compared are coming from a single population or two different populations, There are different types of t-tests, but the two most common ones are.
  1. **Independent Samples T-Test**
  2. **Paired Samples T-Test**
- The **independent samples t-test**, also known as the unpaired t-test, is a statistical test that determines if there is a significant difference between the means of two unrelated groups
- The Independent Samples t Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.
- A **paired samples t-test**, also known as a dependent samples t-test, compares the means of two measurements taken from the same individual, object, or related units
- The Paired-Samples T Test procedure compares the means of two variables for a single group. The procedure computes the differences between values of the two variables for each case and tests whether the average differs from 0. The procedure also automates the t-test effect size computation

**Student t-Test Formula**

- We have already discussed the t-test definition. The formula for the two-sample t-test (a.k.a. the Student's t-test) is shown below.

$$\text{Student t Test Formula, } t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}}$$

- In the formula given above, t is equal to the t-value, x1 and x2 are the means of the two groups being compared, $s^2$is the pooled standard error of the two groups, and n1 and n2 are the numbers of observations in each of the groups.
- A larger t-value denotes the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.
- You can compare your calculated t-value against the values in a critical value chart to determine whether your t-value is greater than what would be expected by chance. If so, you can reject the null hypothesis and you can conclude which two groups are in fact different.

**1. Independent Samples T-Test:**

- This test is used when comparing the means of two independent groups. For example, comparing the average scores of two different groups of participants in an experiment or comparing the means of two different treatment groups.
- The formula for the t-statistic in the independent samples t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Where X1 and X2 are the sample means s1 and s2 are the sample standard deviations, and n1 and n2 are the sample sizes.

**Paired Samples T-Test:**

- This test is used when comparing the means of two related groups. For example, comparing the scores of the same group of participants before and after a treatment.
- The formula for the t-statistic in the paired samples t-test is:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

- where d is the mean of the differences between paired observations, s d is the standard deviation of the differences, and n is the number of pairs.

## Example Program: Paired t-Test Program for Dependent Samples

1. **Import the stats module from scipy:**

```
from scipy import stats
```

❖ This imports the necessary statistical functions from the scipy library.

2. **Define the paired_t_test function**

```
def paired_t_test(before, after):
    # Perform paired t-test
    t_statistic, p_value = stats.ttest_rel(before, after)

    return t_statistic, p_value
```

❖ This function takes two lists (before and after) as input, representing paired measurements before and after a treatment.

❖ It then uses stats.ttest_rel to perform a paired t-test and returns the t-statistic and p-value.

**3. Example usage:**

```
before_treatment = [28, 30, 32, 34, 36]
after_treatment = [25, 29, 31, 33, 35]
```

❖ These lists represent the measurements taken before and after a treatment.

**4. Perform the paired t-test:**

```
t_statistic, p_value = paired_t_test(before_treatment, after_treatment)
```

❖ This line calls the paired_t_test function with the provided lists and stores the results in t_statistic and p_value.

**5. Print the results:**

```
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

❖ These lines print the calculated t-statistic and p-value.

6. Interpret the result:

**alpha = 0.05**

**if p_value < alpha:**

**print("Reject the null hypothesis. There is a significant difference before and after treatment.")**

**else:**

**print("Fail to reject the null hypothesis. There is no significant difference before and after treatment.")**

❖ These lines interpret the result by comparing the p-value to a significance level (alpha). If the p-value is less than alpha, the null hypothesis is rejected, indicating a significant difference between before and after treatment. Otherwise, the null hypothesis is not rejected. Adjust the alpha level based on the desired significance threshold.

OUTPUT

```
t-statistic: 0.5741692517632145
p-value: 0.5816333668955778
Fail to reject the null hypothesis. There is no significant difference between the groups.
```

## Week 6: Build Exploratory Data Analysis on Automobile data

**EDA tasks on the automobile dataset, including:**

1. Getting basic information about the dataset
2. Checking for missing values
3. Visualizing the missing values – Boys: Seaborn, Plotly, Girls: Matplotlib, Bokeh
4. Summary Statistics:
    1. Mean, median, mode
    2. Variance, standard deviation
    3. Minimum, maximum, range
5. Replacing the missing values with the mean of the column
6. Checking for outliers
7. Identifying correlated variables
8. Creating a scatter plot of the price and horsepower Boys: Seaborn, Plotly, Girls: Matplotlib, Bokeh
9. Creating a histogram of the price Boys: Seaborn, Plotly, Girls: Matplotlib, Bokeh
10. Creating a pie chart of the car types
11. Creating a bar chart of the average price by car type Boys: Seaborn, Plotly, Girls: Matplotlib, Bokeh
12. Creating a line chart of the average price by year Boys: Seaborn, Plotly, Girls: Matplotlib, Bokeh

## CHI-SQUARE TEST

- A chi-square test is a statistical test that is used to compare observed and expected results.
- A chi-square test is a statistical hypothesis test that examines whether two categorical variables are independent in influencing the test statistic.
- It's used to compare observed results with expected results to determine if a difference is due to chance or a relationship between the variables.
- **The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration**.
- As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.
- A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable.
- Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

- It is used to calculate the difference between two categorical variables, which are:
  - As a result of chance
  - Because of the relationship
- 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.
- The Observed values are those you gather yourselves.
- The expected values are the frequencies expected, based on the null hypothesis.

## Categorical Variables

- Categorical variables belong to a subset of variables that can be divided into discrete categories. Names or labels are the most common categories. These variables are also known as qualitative variables because they depict the variable's quality or characteristics.
- Categorical variables can be divided into two categories:
- **Nominal Variable:** A nominal variable's categories have no natural ordering. Example: Gender, Blood groups
- **Ordinal Variable:** A variable that allows the categories to be sorted is ordinal variables. Customer satisfaction (Excellent, Very Good, Good, Average, Bad, and so on) is an example.

**Use of Chi-Square Test:**

- Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting.
- Here are some of the uses of the Chi-Squared test:
- The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
- The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.
- **Karl Pearson introduced this test in 1900 for categorical data analysis and distribution. This test is also known as 'Pearson's Chi-Squared Test'.**
- Chi-Squared Tests are most commonly used in hypothesis testing. A hypothesis is an assumption that any given condition might be true, which can be tested afterwards.
- The Chi-Square test estimates the size of inconsistency between the expected results and the actual results when the size of the sample and the number of variables in the relationship is mentioned.

- These tests use degrees of freedom to determine if a particular null hypothesis can be rejected based on the total number of observations made in the experiments. Larger the sample size, more reliable is the result.
- There are two main types of Chi-Square tests namely –
  - Independence
  - Goodness-of-Fit

**Independence**

- The Chi-Square Test of Independence is a derivable ( also known as inferential ) statistical test which examines whether the two sets of variables are likely to be related with each other or not.
- This test is used when we have counts of values for two nominal or categorical variables and is considered as non-parametric test.
- A relatively large sample size and independence of observations are the required criteria for conducting this test.

**For Example-**

- In a movie theatre, suppose we made a list of movie genres.
- Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre.
- Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unreliable.
- If this is true, the movie genres don't impact snack sales.

**Goodness-Of-Fit**

- In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not.
- We must have a set of data values and the idea of the distribution of this data.
- We can use this test when we have value counts for categorical variables.
- This test demonstrates a way of deciding if the data values have a " good enough" fit for our idea or if it is a representative sample data of the entire population.

For Example-

- Suppose we have bags of balls with five different colours in each bag. The given condition is that the bag should contain an equal number of balls of each colour. The idea we would like to test here is that the proportions of the five colours of balls in each bag must be exact.

## Write a Python Program to implement Chi-Square Test

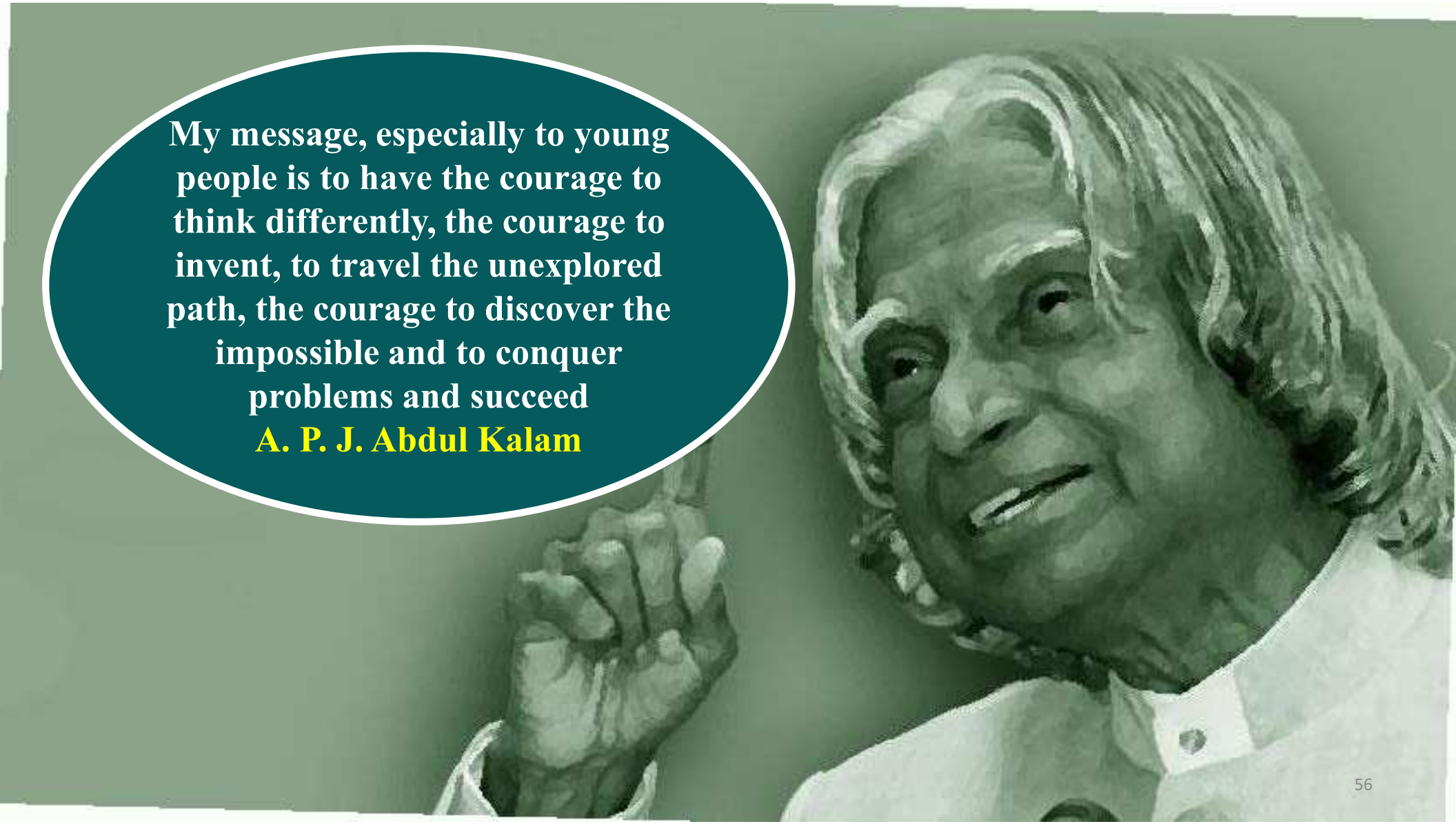
```
import numpy as np
from scipy.stats import chi2_contingency
# Define your contingency table (replace with your data)
observed_data = np.array([
    [10, 20, 30],
    [15, 25, 40] ])
# Calculate the Chi-square statistic, p-value, degrees of freedom, and expected table
chi2_statistic, p_value, degrees_of_freedom, expected_data = chi2_contingency(observed_data)
# Print the results
print("Chi-Square Statistic:", chi2_statistic)
print("P-value:", p_value)
print("Degrees of Freedom:", degrees_of_freedom)
print("Expected Table:\n", expected_data)
# Interpretation
if p_value < 0.05:
    print("Reject null hypothesis: There is a statistically significant relationship between the variables.")
else:
    print("Fail to reject null hypothesis: There is no evidence of a statistically significant relationship.")
```

## Output

```
Chi-Square Statistic: 0.1296296296296296
P-value: 0.9372410104578182
Degrees of Freedom: 2
Expected Table:
 [[10.71428571 19.28571429 30.        ]
 [14.28571429 25.71428571 40.        ]]
Fail to reject null hypothesis: There is no evidence of a statistically significant relationship.
```

Explanation:

1. **Import libraries:** We import numpy for numerical operations and chi2_contingency from scipy.stats for chi-square test calculations.
2. **Define observed data:** Replace observed_data with your actual contingency table containing observed counts for each category combination.
3. **Calculate Chi-square test:** chi2_contingency function takes the observed data as input and returns the chi-square statistic, p-value, degrees of freedom, and expected table.
4. **Print results:** The program prints the calculated values and interprets the results based on the p-value.
    1. If p-value < 0.05 (common significance level), we reject the null hypothesis and conclude that there is a statistically significant relationship between the variables.
    2. Otherwise, we fail to reject the null hypothesis and say there's no evidence of a significant relationship.


My message, especially to young people is to have the courage to think differently, the courage to invent, to travel the unexplored path, the courage to discover the impossible and to conquer problems and succeed
A. P. J. Abdul Kalam


Thank
You!