**Group No.: Date:**

**EXPERIMENT NO. 4**

**AIM**: To understand various concepts about Null Hypothesis in python .

**SOFTWARE USED**: Jupyter Notebook

**THEORY**:

The theory of the null hypothesis is a fundamental concept in statistical hypothesis testing. It provides a framework for making decisions based on sample data regarding population parameters: The null hypothesis (H0) is a statement that there is no effect, no difference, or no relationship in the population. Hypothesis testing involves comparing observed data to what would be expected under the null hypothesis

.Let's delve into the theory in detail:

# Introduction to Hypothesis Testing:

Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It involves comparing observed data to what would be expected under a null hypothesis, which represents a default assumption about the population.

# Null Hypothesis (H0):

The null hypothesis, denoted as H0, is a statement that there is no effect, no difference, or no relationship in the population. It typically represents the status quo or the absence of an effect. For example:

H0: There is no difference in means between two groups. H0: The population proportion is equal to a specific value. H0: There is no correlation between two variables.

# Alternative Hypothesis (Ha or H1):

The alternative hypothesis, denoted as Ha or sometimes H1, is the opposite of the null hypothesis. It suggests that there is an effect, difference, or relationship in the population that differs from what is stated in the null hypothesis. For example:

Ha: There is a difference in means between two groups.

Ha: The population proportion is different from a specific value. Ha: There is a correlation between two variables.

# Significance Level (α):

The significance level, denoted by α (alpha), is the probability of rejecting the null hypothesis when it's actually true. Commonly used values for α are 0.05 (5%) or 0.01 (1%). It represents the threshold for determining whether the evidence against the null hypothesis is strong enough to reject it.

# Test Statistic and p-value:

In hypothesis testing, a test statistic is calculated from the sample data to assess the likelihood of observing the data under the null hypothesis. The p-value is the probability of obtaining a test statistic (or one more extreme) if the null hypothesis is true. A small p-value indicates strong evidence against the null hypothesis, leading to its rejection.

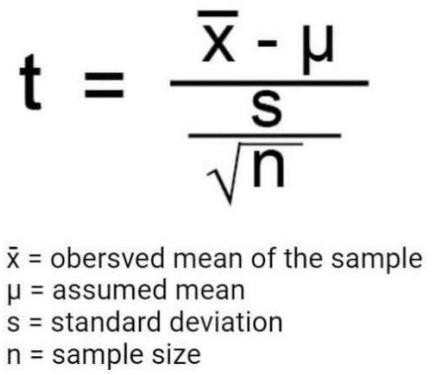
# Decision Rule:

Based on the calculated p-value and the chosen significance level, a decision is made whether to reject the null hypothesis or not:

* If the p-value is less than α, reject the null hypothesis.
* If the p-value is greater than or equal to α, fail to reject the null hypothesis.

# Calculation of the t-value:

The t-value is calculated as the difference between the sample mean and the hypothesized population mean (under the null hypothesis), divided by the standard error of the mean. Mathematically, it can be expressed as:



# Interpretation of the t-value:

The t-value represents the number of standard errors the sample mean is away from the hypothesized population mean under the null hypothesis. A large absolute t-value indicates that the sample mean is significantly different from the hypothesized population mean.

# Methodology:

Here's a methodology for conducting a hypothesis test regarding the mean of a normal distribution using Python:

# Formulate the Hypotheses:

Null Hypothesis (H0): There is no difference or no effect. For example, H0: μ = μ0, where μ is the population mean and μ0 is a specific value.

Alternative Hypothesis (H1 or Ha): There is a difference or effect. It's the opposite of the null hypothesis. For example, Ha: μ ≠ μ0 (two-tailed test), Ha: μ > μ0 (one-tailed test), or Ha: μ < μ0 (one-tailed test).

# Select the Significance Level (α):

The significance level, denoted by α, is the probability of rejecting the null hypothesis when it's actually true. Commonly used values for α are 0.05 or 0.01.

# Collect Data:

Collect a sample of data from the population of interest. Ensure that the sample is random and representative of the population.

# Conduct the Hypothesis Test:

Compute the Test Statistic:

Depending on the hypothesis test (e.g., z-test or t-test), calculate the appropriate test statistic.

For example, for testing the population mean with a known population standard deviation, you'd use the z- test. If the population standard deviation is unknown, you'd typically use the t-test.

Calculate the p-value:

Determine the probability of observing the test statistic (or one more extreme) under the null hypothesis. Make a Decision: Compare the p-value to the significance level (α). If the p-value is less than α, reject the null hypothesis in favor of the alternative hypothesis; otherwise, fail to reject the null hypothesis.

# Interpret the Results:

If the null hypothesis is rejected, conclude that there is sufficient evidence to support the alternative hypothesis.

If the null hypothesis is not rejected, conclude that there is not enough evidence to support the alternative hypothesis.

**CODE AND OUTPUT:**

In [57]:

**import** pandas **as** pd **import** scipy **import** numpy **as** np **from** scipy **import** stats

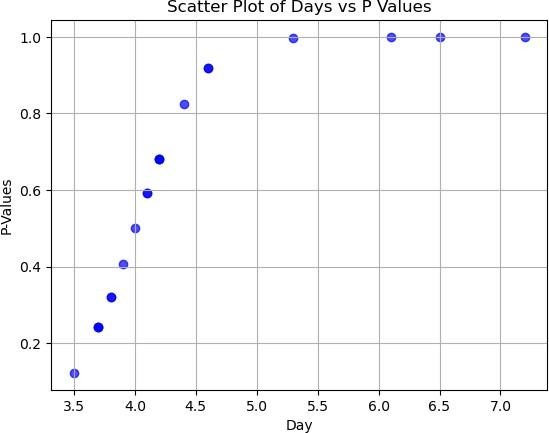
*# Read the Excel file with correct parameters* df **=** pd**.**read\_excel("Lab 4\_DSA1.xlsx", header**=**3) *# Skip the first three rows as the* df**.**drop(['Unnamed: 0','Unnamed: 1','Unnamed: 5'],axis**=**1,inplace**=True**) df **=** df**.**dropna() t\_value **=** df['T-VALUE'] p\_value**=** stats**.**t**.**cdf(t\_value,49) *#by default for two tail test, for one tail di* df['P- VALUE'] **=** p\_value df

Out[57]: **Day Mean Call Duration T-VALUE P-VALUE**

|  |  |  |  |
| --- | --- | --- | --- |
| **1** 1.0 | 3.7 | -0.707107 | 0.241425 |
| **2** 2.0 | 4.1 | 0.235702 | 0.592677 |
| **3** 3.0 | 3.5 | -1.178511 | 0.122142 |
| **4** 4.0 | 4.2 | 0.471405 | 0.680278 |
| **5** 5.0 | 3.9 | -0.235702 | 0.407323 |
| **6** 6.0 | 4.1 | 0.235702 | 0.592677 |
| **7** 7.0 | 4.2 | 0.471405 | 0.680278 |
| **8** 8.0 | 3.8 | -0.471405 | 0.319722 |
| **9** 9.0 | 3.7 | -0.707107 | 0.241425 |
| **10** 10.0 | 4.6 | 1.414214 | 0.918189 |
| **11** 11.0 | 3.7 | -0.707107 | 0.241425 |
| **12** 12.0 | 4.6 | 1.414214 | 0.918189 |
| **13** 13.0 | 4.0 | 0.000000 | 0.500000 |
| **14** 14.0 | 4.2 | 0.471405 | 0.680278 |

|  |  |  |  |
| --- | --- | --- | --- |
| **15** 15.0 | 3.8 | -0.471405 | 0.319722 |
| **16** 16.0 | 4.4 | 0.942809 | 0.824798 |
| **17** 17.0 | 5.3 | 3.064129 | 0.998228 |
| **18** 18.0 | 6.1 | 4.949747 | 0.999995 |
| **19** 19.0 | 7.2 | 7.542472 | 1.000000 |
| **20** 20.0 | 6.5 | 5.892557 | 1.000000 |

In [62]:



**import** matplotlib.pyplot **as** plt

*# Scatter plot*

plt**.**scatter(df['Mean Call Duration'], df['P-VALUE'], alpha**=**0.7, color**=**'blue') plt**.**xlabel('Day')

plt**.**ylabel('P-Values')

plt**.**title('Scatter Plot of Days vs P Values')

plt**.**grid(**True**)

*# Display the plot*

plt**.**show()

In [63]:

alpha**=**0.5 *#determined by company* df['Decision'] **=** np**.**where(df['P-VALUE'] **<** alpha, 'reject Null Hypothesis', 'accept df

Out[63]:

**Day Mean Call Duration T-VALUE P-VALUE Decision**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** 1.0 | 3.7 | -0.707107 | 0.241425 | reject Null Hypothesis |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2** 2.0 | 4.1 | 0.235702 | 0.592677 | accept Null Hypothesis |
| **3** 3.0 | 3.5 | -1.178511 | 0.122142 | reject Null Hypothesis |
| **4** 4.0 | 4.2 | 0.471405 | 0.680278 | accept Null Hypothesis |
| **5** 5.0 | 3.9 | -0.235702 | 0.407323 | reject Null Hypothesis |
| **6** 6.0 | 4.1 | 0.235702 | 0.592677 | accept Null Hypothesis |
| **7** 7.0 | 4.2 | 0.471405 | 0.680278 | accept Null Hypothesis |
| **8** 8.0 | 3.8 | -0.471405 | 0.319722 | reject Null Hypothesis |
| **9** 9.0 | 3.7 | -0.707107 | 0.241425 | reject Null Hypothesis |
| **10** 10.0 | 4.6 | 1.414214 | 0.918189 | accept Null Hypothesis |
| **11** 11.0 | 3.7 | -0.707107 | 0.241425 | reject Null Hypothesis |
| **12** 12.0 | 4.6 | 1.414214 | 0.918189 | accept Null Hypothesis |
| **13** 13.0 | 4.0 | 0.000000 | 0.500000 | accept Null Hypothesis |
| **14** 14.0 | 4.2 | 0.471405 | 0.680278 | accept Null Hypothesis |
| **15** 15.0 | 3.8 | -0.471405 | 0.319722 | reject Null Hypothesis |
| **16** 16.0 | 4.4 | 0.942809 | 0.824798 | accept Null Hypothesis |
| **17** 17.0 | 5.3 | 3.064129 | 0.998228 | accept Null Hypothesis |
| **18** 18.0 | 6.1 | 4.949747 | 0.999995 | accept Null Hypothesis |
| **19** 19.0 | 7.2 | 7.542472 | 1.000000 | accept Null Hypothesis |
| **20** 20.0 | 6.5 | 5.892557 | 1.000000 | accept Null Hypothesis |

In [60]:

*# by analysing the final column we find the range of call duration which is success*

*# call*

**import import**

*centre*

pandas scipy

*example*

**as**

pd

**import**

numpy **as** np **from** scipy

**import** stats (4**-** 4.6)**/**(3**/**np**.**sqrt(50))

-1.4142135623730943

Out[60]:

In [36]:

*#Q.An outbreak of Salmonella related illness was attributed to icecream produced at #Scientist measured the level of Salmonella in 9 randomly sampled batches of icecre #Levels of toxicity : 0.593, 0.142, 0.329, 0.691, 0.231,0.793, 0.519, 0.392, 0.418 #data=pd.Series([0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418]) (pvalues)*

*#Is there evidence that the mean level of Salmonella in the icecream is greater tha #H0<=0.3 ----> H0 = that the company isn't at fault for poisoning*

*#Ha>0.3*

data**=**pd**.**Series([0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418])

p**=**scipy**.**stats**.**ttest\_1samp(data,0.3)[1] *#1 sample 1 tail test* p\_value**=**p**/**2

*#for one tail test and this is for the first datapoint. We compare this value with #since it is less than 0.05, We accept the null hypothesis, So for this value, it i #entire data set, after comparing alpha values we find the range for which the faul #(we always assume null hypothesis to be opposite to our objective in question) #here we comapre the alpha-value which has 0.05 as constant value*

*#as it is a 1 sample 1 tail test*

*# in the above code cell the obtained value is for 0.593 which indicates we accept*

p\_value

Out[36]:

In [37]:

ialpha **=** 0.05

**if** p\_value **<**

alpha:

print("Reject null hypothesis at the 0.05 significance level")

**else**:

print("Accept null hypothesis at the 0.05 significance level")

0.029265164842448826

*#Q.6 subjects were given a drug (treatment group) and an additional 6 subjects a pl #Their reaction time to a stimulus was measured (milliseconds)*

*# control ={91,87,99,77,88,99}*

*# treat ={101,110,103,93,99,104}*

*# Problem statement: control=treat 2-sample 2 tail test # degree of freedom:(6-1)+(6-1)=10*

*#we want to perform a 2 sample t-test for comparing the means of treatment and cont #Ho= control=test*

*#Ha= control not equal to test* **import** numpy **as** np **from** scipy **import** stats

*# Define the data for the control and treatment groups*

control **=** np**.**array([91, 87, 99, 77, 88, 99]) treatment

**=** np**.**array([101, 110, 103, 93, 99, 104])

[44]:

In [52]:

*# Perform two-sample t-test*

t\_statistic, p\_value **=** stats**.**ttest\_ind(control, treatment)

Reject null hypothesis at the 0.05 significance level

In [53]:

*# Calculate degrees of freedom* df **=**

len(control) **+** len(treatment) **-** 2 df

10

Out[53]:

In [55]:

*# Print the*

statistic print("p-value

*results* print("t-

=",t\_statistic)

=",

p\_value)

print("Degrees of Freedom =", df)

t-statistic = -2.8052489172270088 p-value = 0.018626074653407012 Degrees of

Freedom = 10

*# Interpret the results*

alpha **=** 0.05 **if** p\_value

**<** alpha:

print("Reject the null hypothesis")

**else**:

print("Accept the null hypothesis")

In [56]:

Reject the null hypothesis

**CONCLUSION:**

In summary, the t-value serves as a pivotal metric in hypothesis testing, particularly when dealing with small sample sizes or unknown population standard deviations. Its calculation, based on the sample mean, population mean assumed under the null hypothesis, sample standard deviation, and sample size, provides a quantitative measure of how significantly the sample mean deviates from the null hypothesis. By comparing the t-value to critical values from the t-distribution at the chosen significance level, researchers can determine the statistical significance of their findings.