|  |  |  |
| --- | --- | --- |
| الجامعة الإسلامية – غزة  كلية تكنولوجيا المعلومات  ماجستير تكنولوجيا المعلومات |  | Islamic University – Gaza  Faculty of Information Technology  Department of Information Technology |
|  | شعار الجامعة.bmp |  |

**Using Hypothesis Testing**

**as Exploratory Data analysis Technique**

*A Case Study Using the Iris Dataset*

By:

*Abeer Yousef Abu Mosameh -220232641*

*Supervised By*

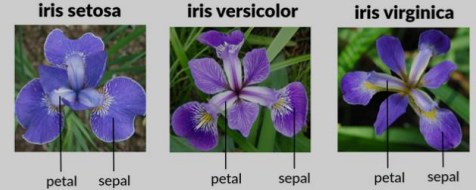
Dr. Iyad Husni Alshami

*Dec, 2024*

# Introduction

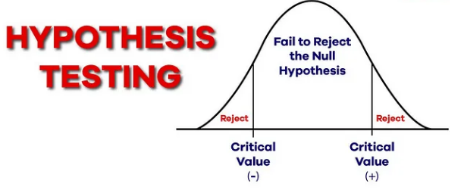
**Exploratory data analysis (EDA)** is used by data scientists to analyze and investigate data sets and declare main characteristics. For example, EDA use data visualization techniques to provide a clearer understanding of the data. Hypothesis testing plays an important role in EDA by validating patterns statistically.

**Hypothesis testing** is a structured approach or statistical method to validate or reject assumptions and claims about a dataset/population “It evaluates two mutually exclusive statements about a population to determine **which statement is best** supported by the sample data”. Therefore, it helps us make decisions based on data and evidence rather than intuition or assumption.

****So we will explore the application of hypothesis testing as an EDA technique using the Iris dataset- data sets consists of 3 different types of irises’ (Setosa, Versicolour, and Virginica) petal and sepal length.

# Hypothesis Testing Techniques Steps

1. **Defining Hypotheses:**

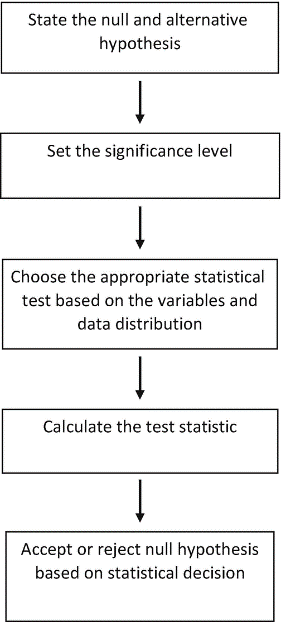
* ***Null Hypothesis (H0):***This represents the assumption to be tested. It typically states that there is *no effect, no difference between groups, or no relationship in the data*.
* *****Alternative Hypothesis (H1):***This represents the opposite/complement of the null hypothesis. It suggests there is *an effect, a difference, or a relationship in the data can expect to find*

So The null hypothesis is the claim we test for possible rejection, while the alternate hypothesis is accepted when there's evidence against the null.

1. **Choose Appropriate Test Tool**

There is many tool can used in Hypothesis Testing:

* **Z-Tests**: used If population means and standard deviations are known.
* **T-Tests:** used If population standard deviations are unknown. and sample size is small
* **Chi-Squared Tests:** Assess the association between categorical variables.
* **F-test "Analysis of Variance (ANOVA)":** Compare the means of three or more groups .. used if has multiple groups.

1. **Decide Significance Level (Alpha):** it refers to the degree of significance in which we accept or reject the null hypothesis

Common alpha values are **0.05** “which means output should be 95% confident to give a similar kind of result in other sample” and **0.01**.

1. **Calculate Test Statistic and P-value**

* Test Statistic: is a numerical value calculated from sample data during a hypothesis test, used to determine whether to reject the null hypothesis.
* P-value: is the probability of obtaining test results when the null hypothesis (H0) is true.

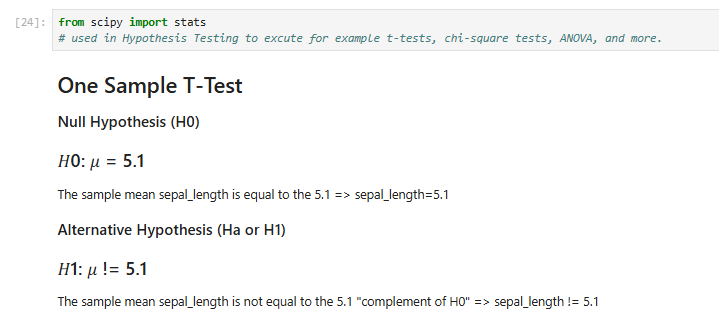
1. **Take Hypotheses decision:**

Based on the test results, decide to reject the null hypothesis or not.

If the p-value is less than alpha” significance level”, then reject the null hypothesis. If it’s greater, we usually accept the null hypothesis.

# Apply Hypothesis Testing Technique in Iris Dataset using Python code

### **Z-Tests not preferred because the sample size is relatively small**

1. **T-Tests can used**

t = Test statistic

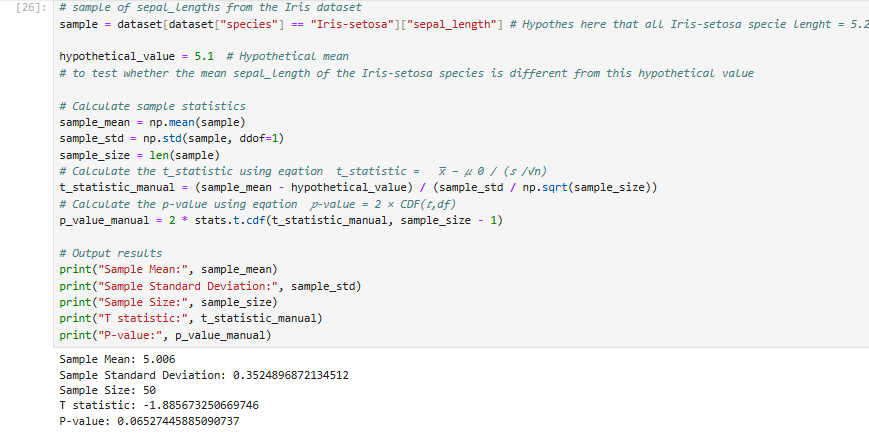
x̄ = sample mean

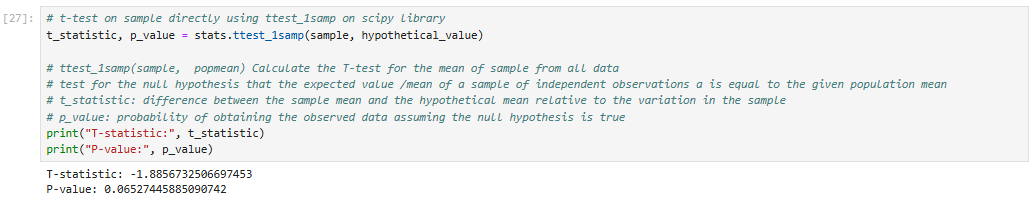
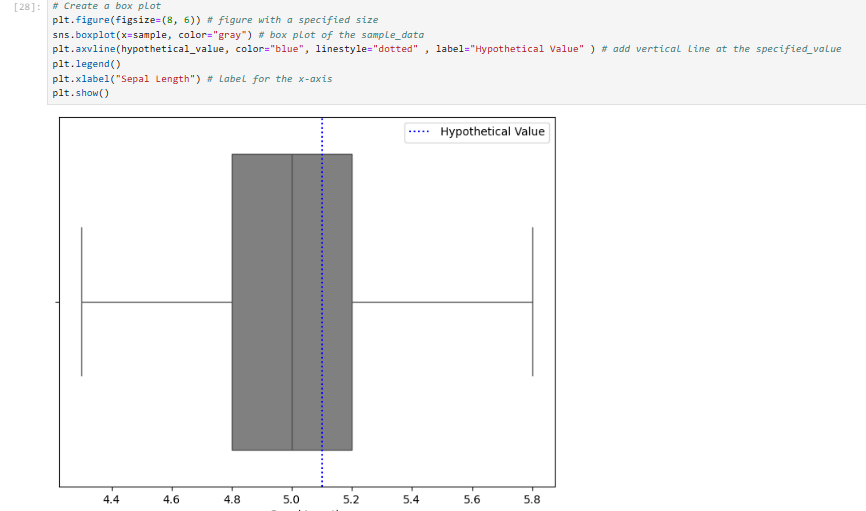
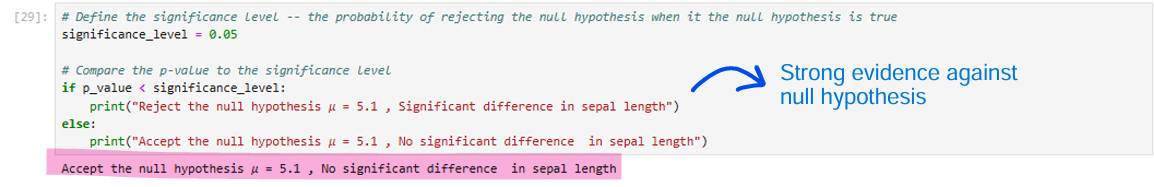
μ = population mean

s = standard deviation of the sample

n = sample size

Using Equation:

or using ttest\_1samp directly from scipy library



*  **Two sample T-Test**
* are the sample means of the two groups
* is the pooled standard deviation

standard deviations of the two samples

* and are the sample sizes of the two groups.

1. **Chi-Square Test can used but I will not mentioned**
2. **Analysis of Variance (ANOVA) :** suit for iris database because it’s 3 group ['Iris-setosa' , 'Iris-versicolor',’ Iris-virginica’]



Video Link

<https://youtu.be/eIhFP4xzQ3M>

References

* GeeksforGeeks. (2024a, May 16). *What is Exploratory Data Analysis?* GeeksforGeeks. [https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/](%20https:/www.geeksforgeeks.org/what-is-exploratory-data-analysis/)
* GeeksforGeeks. (2024c, October 8). *Understanding hypothesis testing*. GeeksforGeeks. <https://www.geeksforgeeks.org/understanding-hypothesis-testing/>
* Slidescope. (2023, February 22). *ANOVA Example using Python Pandas on Iris Dataset*. Slidescope. <https://slidescope.com/anova-example-using-python-pandas-on-iris-dataset/>
* GeeksforGeeks. (2024, September 23). *Exploratory data analysis on IRIS Dataset*. GeeksforGeeks. <https://www.geeksforgeeks.org/exploratory-data-analysis-on-iris-dataset/>