Islamic University – Gaza
Faculty of Information Technology
Department of Information Technology

الجامعة الإسلامية – غزة
كلية تكنولوجيا المعلومات
ماجستير تكنولوجيا المعلومات

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

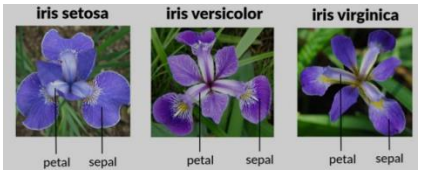
Dec, 2024

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.

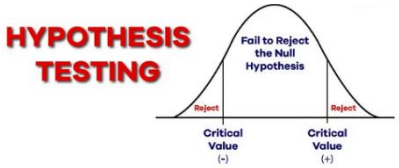
So we will explore the application of hypothesis testing as an EDA technique using the Iris dataset- data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length.



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*

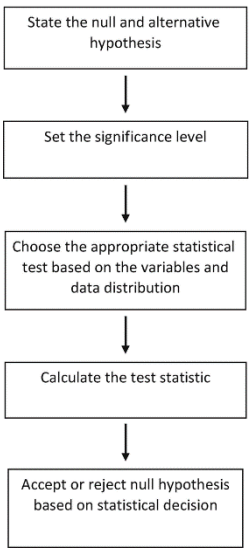
So The null hypothesis is the claim we test for possible rejection, while the alternate hypothesis is accepted when there's evidence against the null.



2. **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.

3. **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**.



4. **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.

5. **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

1. **Z-Tests not preferred because the sample size is relatively small**
2. **T-Tests can used**

```
[24]: from scipy import stats
      # used in Hypothesis Testing to excute for example t-tests, chi-square tests, ANOVA, and more.
```

One Sample T-Test

Null Hypothesis (H0)

$H0: \mu = 5.1$

The sample mean sepal_length is equal to the 5.1 => sepal_length=5.1

Alternative Hypothesis (Ha or H1)

$H1: \mu != 5.1$

The sample mean sepal_length is not equal to the 5.1 "complement of H0" => sepal_length != 5.1

mathematical equation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

t = Test statistic

$\bar{x}$ = sample mean

μ = population mean

s = standard deviation of the sample

n = sample size

Using Equation:

[26]:
```
# sample of sepal_lengths from the Iris dataset
sample = dataset[dataset["species"] == "Iris-setosa"]["sepal_length"] # Hypothes here that all Iris-setosa specie lenght = 5.2

hypothetical_value = 5.1  # Hypothetical mean
# to test whether the mean sepal_length of the Iris-setosa species is different from this hypothetical value

# Calculate sample statistics
sample_mean = np.mean(sample)
sample_std = np.std(sample, ddof=1)
sample_size = len(sample)
# Calculate the t_statistic using eqation  t_statistic =   x̄ - μ 0 / (s /√n)
t_statistic_manual = (sample_mean - hypothetical_value) / (sample_std / np.sqrt(sample_size))
# Calculate the p-value using eqation  p-value = 2 × CDF(t,df)
p_value_manual = 2 * stats.t.cdf(t_statistic_manual, sample_size - 1)

# Output results
print("Sample Mean:", sample_mean)
print("Sample Standard Deviation:", sample_std)
print("Sample Size:", sample_size)
print("T statistic:", t_statistic_manual)
print("P-value:", p_value_manual)
```

```
Sample Mean: 5.006
Sample Standard Deviation: 0.3524896872134512
Sample Size: 50
T statistic: -1.885673250669746
P-value: 0.06527445885090737
```

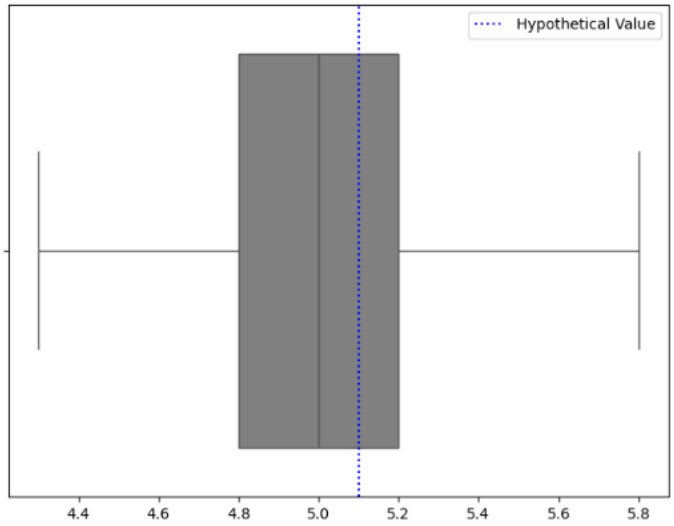
or using ttest_1samp directly from scipy library

[27]:
```
# t-test on sample directly using ttest_1samp on scipy library
t_statistic, p_value = stats.ttest_1samp(sample, hypothetical_value)

# ttest_1samp(sample,  popmean) Calculate the T-test for the mean of sample from all data
# test for the null hypothesis that the expected value /mean of a sample of independent observations a is equal to the given population mean
# t_statistic: difference between the sample mean and the hypothetical mean relative to the variation in the sample
# p_value: probability of obtaining the observed data assuming the null hypothesis is true
print("T-statistic:", t_statistic)
print("P-value:", p_value)
```

```
T-statistic: -1.8856732506697453
P-value: 0.06527445885090742
```

[28]:
```
# Create a box plot
plt.figure(figsize=(8, 6)) # figure with a specified size
sns.boxplot(x=sample, color="gray") # box plot of the sample_data
plt.axvline(hypothetical_value, color="blue", linestyle="dotted" , label="Hypothetical Value" ) # add vertical line at the specified_value
plt.legend()
plt.xlabel("Sepal Length") # label for the x-axis
plt.show()
```



[29]:
```
# Define the significance level -- the probability of rejecting the null hypothesis when it the null hypothesis is true
significance_level = 0.05

# Compare the p-value to the significance level
if p_value < significance_level:
    print("Reject the null hypothesis μ = 5.1 , Significant difference in sepal length")
else:
    print("Accept the null hypothesis μ = 5.1 , No significant difference  in sepal length")
```

Strong evidence against null hypothesis

```
Accept the null hypothesis μ = 5.1 , No significant difference  in sepal length
```

- **Two sample T-Test**

$$t = \frac{\bar{x_1} - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $\bar{x_1} - \bar{x}_2$ are the sample means of the two groups
- s_p is the pooled standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

s_1 , s_2 standard deviations of the two samples

- n_1 and n_2 are the sample sizes of the two groups.

Null Hypothesis (H0)

H0: $\mu 1 = \mu 2$

The means of the two groups['Iris-setosa' , 'Iris-versicolor'] are equal

Alternative Hypothesis (Ha or H1)

H1: $\mu 1 != \mu 2$

The means of the two groups['Iris-setosa' , 'Iris-versicolor'] are not equal

[37]:
```
#  samples for setosa and versicolor species
setosa_sample = dataset[dataset['species'] == 'Iris-setosa']['petal_length']
versicolor_sample = dataset[dataset['species'] == 'Iris-versicolor']['petal_length']

# Perform two-sample t-test
t_statistic, p_value = stats.ttest_ind(setosa_sample, versicolor_sample)
print("T-statistic:", t_statistic)
print("P-value:", p_value)

# Define the significance level -- the probability of rejecting the null hypothesis when it the null hypothesis is true
significance_level = 0.05

# Compare the p-value to the significance level
if p_value < significance_level:
    print("Reject the null hypothesis μ1 = μ2 , Significant difference in petal lengths.")
else:
    print("Accept the null hypothesis  μ1 = μ2 , No difference in petal lengths.")
```

```
T-statistic: -39.46866259397272
P-value: 5.717463758170621e-62
Reject the null hypothesis μ1 = μ2 , Significant difference in petal lengths.
```

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

ANOVA TEST

Null Hypothesis (H0)

H0: μ Setosa = μ Versicolor = μ Virginica

There is no significant difference in mean petal lengths among the three species.

Alternative Hypothesis (Ha or H1)

H1: μ Setosa != μ Versicolor != μ Virginica

There is a significant difference in mean petal lengths among the three species.

[39]:
```
anova_result = stats.f_oneway(
    dataset[dataset['species'] == 'Iris-setosa']['petal_length'],
    dataset[dataset['species'] == 'Iris-versicolor']['petal_length'],
    dataset[dataset['species'] == 'Iris-virginica']['petal_length']
)

print("T-statistic:", anova_result.statistic)
print("P-value:", anova_result.pvalue)

# Define the significance level -- the probability of rejecting the null hypothesis when it the null hypothesis is true
significance_level = 0.05

# Compare the p-value to the significance level
if p_value < significance_level:
    print("Reject null hypothesis μ Setosa = μ Versicolor = μ Virginica : at least one group mean is different.")
else:
    print("Accept the null hypothesis μ Setosa = μ Versicolor = μ Virginica: all group means are the same")
```

```
T-statistic: 1179.0343277002194
P-value: 3.0519758018278374e-91
Reject null hypothesis μ Setosa = μ Versicolor = μ Virginica : at least one group mean is different.
```

References

- GeeksforGeeks. (2024a, May 16). *What is Exploratory Data Analysis?* GeeksforGeeks. https://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/