PROJECT REPORT

Statistical Hypothesis Testing for Exploring Patterns in Multivariate Data: A Case Study on the Iris Dataset

Jagnoor Singh Marok

29.11.2024

AIM

The aim of this project is to leverage statistical hypothesis testing techniques (Z-Test, T-Test, and Chi-Square Test) to analyze the Iris dataset, a classic multivariate dataset in data science. Through this analysis, we seek to identify significant differences in numerical features across species and explore relationships between categorical and numerical variables. By applying hypothesis testing, we aim to draw meaningful conclusions about the underlying patterns in the data and evaluate the strengths and limitations of different statistical methods.

ABSTRACT

Statistical hypothesis testing is a foundational tool in data analysis, enabling researchers to draw inferences about data characteristics and relationships. This project utilizes the Iris dataset, a classic multivariate dataset, to explore patterns using three statistical methods: Z-Test, T-Test, and Chi-Square Test. The primary objective is to compare numerical features (e.g., petal length, sepal width) across species using Z- and T-Tests and to investigate relationships between categorical variables using the Chi-Square Test.

The analysis includes visualizations and interprets the results at a 95% confidence level, highlighting significant differences between species. Key findings demonstrate the sensitivity of Z-Test and T-Test to assumptions about population standard deviation and the applicability of the Chi-Square Test for examining associations between categorical variables.

This project not only provides insights into the Iris dataset but also evaluates the strengths and limitations of these statistical methods, offering a comprehensive perspective on their use in real-world data analysis

TABLE OF CONTENTS:

Full Code, Additional Graphs and Tables are attached with the document.

INTRODUCTION

Background

Statistical hypothesis testing is a critical component of data analysis, providing a framework to test assumptions and draw meaningful inferences from datasets. It plays a pivotal role in research, helping analysts decide whether observed patterns in data are statistically significant or merely due to chance. With the ever-growing availability of data, applying statistical methods effectively is essential for uncovering trends, relationships, and insights.

This project applies statistical hypothesis testing techniques, including Z-Test, T-Test, and Chi-Square Test, to analyze the Iris dataset. The Iris dataset, often used in machine learning and statistical analysis, is an ideal choice for understanding and demonstrating these concepts due to its structured format and well-defined features.

Overview of the Dataset

The Iris dataset is a multivariate dataset introduced by the statistician Ronald Fisher in 1936. It contains 150 entries divided into three species of Iris flowers: *Setosa*, *Versicolor*, and *Virginica*. Each entry consists of four numerical features:

1. Sepal Length (in cm)
2. Sepal Width (in cm)
3. Petal Length (in cm)
4. Petal Width (in cm)

Additionally, a categorical variable, `Species`, identifies the flower's species. The dataset is widely used for statistical and machine learning tasks due to its simplicity and interpretability.

Objectives and Hypotheses

The primary objective of this project is to explore and analyze the Iris dataset using statistical hypothesis testing techniques. Specific objectives include:

1. Applying Z-Test and T-Test to compare numerical features across species.
2. Using Chi-Square Test to evaluate relationships between categorical and numerical variables.

3. Interpreting results to draw conclusions about species-specific differences and relationships in the dataset.

Hypotheses to be tested:

- **Z-Test:** Null Hypothesis (H0H_0H0): The mean petal length of *Setosa* and *Versicolor* is equal.
- **T-Test:** Null Hypothesis (H0H_0H0): The mean sepal width of *Versicolor* and *Virginica* is equal.
- **Chi-Square Test:** Null Hypothesis (H0H_0H0): There is no significant association between the `Species` and a binned version of `Sepal Length`.

These hypotheses provide a foundation for investigating the data and understanding patterns within the Iris dataset.

METHODOLOGY

Dataset Description

The Iris dataset, used for this project, is a well-known multivariate dataset comprising 150 entries of Iris flowers, equally distributed among three species: *Setosa*, *Versicolor*, and *Virginica*. Each entry includes the following attributes:

1. **Numerical Variables:**
    - **Sepal Length (cm)**
    - **Sepal Width (cm)**
    - **Petal Length (cm)**
    - **Petal Width (cm)**
2. **Categorical Variable:**
    - **Species:** Identifies the flower species (*Setosa*, *Versicolor*, or *Virginica*).

For this analysis, the dataset is preprocessed to ensure its suitability for statistical testing. This includes verifying data integrity, standardizing units, and binning numerical variables for Chi-Square analysis.

Statistical Tests Overview

The following statistical hypothesis testing techniques are applied:

1. **Z-Test:**
    - Compares the means of a numerical variable between two groups of a categorical variable.
    - Assumes the population standard deviation is known or the sample size is sufficiently large.
2. **T-Test:**
    - Performs a two-sample t-test to compare means when the population standard deviation is unknown.
    - More suitable for smaller samples or when the population variance must be estimated from sample data.
3. **Chi-Square Test:**
    - Evaluates the independence of two categorical variables.
    - Uses observed and expected frequency tables to calculate the Chi-Square statistic.

These tests are implemented with a significance level α\alpha of 0.05 to determine statistical significance.

Tools and Technologies

This project leverages the following tools and technologies for data processing, statistical analysis, and visualization:

1. **Programming Language:** Python
    - Libraries:
        - **Pandas:** For data manipulation and analysis.
        - **NumPy:** For numerical computations.
        - **SciPy:** For statistical testing (Z-Test, T-Test, and Chi-Square Test).
        - **Matplotlib and Seaborn:** For visualizing data distributions and test results.
2. **Integrated Development Environment (IDE):** Google Colab
    - Provides an interactive environment for coding, visualization, and documentation.
3. **Dataset Source:**
    - The Iris dataset is accessed through Python's `sklearn.datasets` module.

IMPLEMENTATION

Z-Test

The Z-Test is used to compare the mean **Petal Length** between the species *Setosa* and *Versicolor*.

Null Hypothesis (H0H_0H0): The mean petal length of *Setosa* and *Versicolor* is equal.
Alternative Hypothesis (HaH_aHa): The mean petal length of *Setosa* and *Versicolor* is not equal.

Steps:

1. Extract samples for *Setosa* and *Versicolor*.
2. Compute the Z-statistic and p-value using SciPy.
3. Interpret results based on the significance level (α=0.05\alpha = 0.05α=0.05).

Visualization:

- The distribution of Petal Length is visualized using a histogram for *Setosa* and *Versicolor*.

```
sns.histplot(data=iris_df, x="petal_length", hue="species", kde=True)

plt.title("Petal Length Distribution by Species")

plt.show()
```

Results:

- **Z-Statistic:** 30.47
- **P-Value:** <0.0001< 0.0001<0.0001
- Conclusion: Reject H0H_0H0, significant difference in petal length.

T-Test

The T-Test is applied to compare the mean **Sepal Width** between the species *Versicolor* and *Virginica*.

Null Hypothesis (H0H_0H0): The mean sepal width of *Versicolor* and *Virginica* is equal.
Alternative Hypothesis (HaH_aHa): The mean sepal width of *Versicolor* and *Virginica* is not equal.

Steps:

1. Extract samples for *Versicolor* and *Virginica*.
2. Perform the T-Test using SciPy's `ttest_ind`.
3. Interpret results based on $\alpha = 0.05$\alpha = 0.05$\alpha=0.05$.

Visualization:

- Box plots are used to show the distribution of Sepal Width for *Versicolor* and *Virginica*.

```
sns.boxplot(data=iris_df[iris_df["species"].isin(["versicolor",
"virginica"])],
            x="species", y="sepal_width")
plt.title("Sepal Width Comparison: Versicolor vs Virginica")
plt.show()
```

Results:

- **T-Statistic:** -4.57
- **P-Value:** <0.0001< 0.0001<0.0001
- Conclusion: Reject H0H_0H0, significant difference in sepal width.

Chi-Square Test

The Chi-Square Test is used to check if the binned **Sepal Length** is independent of the species.

Null Hypothesis (H0H_0H0): Binned Sepal Length is independent of species.
Alternative Hypothesis (HaH_aHa): Binned Sepal Length is associated with species.

Steps:

1. Bin the Sepal Length into intervals (e.g., small, medium, large).
2. Create a contingency table of counts by species and bins.
3. Perform the Chi-Square Test using SciPy's `chi2_contingency`.
4. Interpret results based on $\alpha=0.05$\alpha = 0.05$\alpha=0.05.

Visualization:

- A stacked bar chart displays the distribution of Sepal Length bins across species.

```
contingency_table.plot(kind="bar", stacked=True)

plt.title("Binned Sepal Length by Species")

plt.xlabel("Species")

plt.ylabel("Count")

plt.legend(title="Sepal Length Bin")

plt.show()
```

Results:

- **Chi-Square Statistic:** 140.91
- **P-Value:** <0.0001< 0.0001<0.0001
- Conclusion: Reject H0H_0H0, significant association between binned Sepal Length and species.

RESULTS

Statistical Test Results

The results of the statistical hypothesis tests are summarized below:

Test	Variable(s)	Statistic	P-Value	Conclusion
Z-Test	Petal Length (*Setosa* vs. *Versicolor*)	30.47	< 0.0001	Reject H0H_0H0: Significant difference in means.
T-Test	Sepal Width (*Versicolor* vs. *Virginica*)	-4.57	< 0.0001	Reject H0H_0H0: Significant difference in means.
Chi-Square Test	Binned Sepal Length vs. Species	140.91	< 0.0001	Reject H0H_0H0: Significant association.

Visualizations of Findings

1. **Z-Test Visualization:**
    - **Histogram:** Overlaid distributions of Petal Length for *Setosa* and *Versicolor*.
2. **T-Test Visualization:**
    - **Box Plot:** Sepal Width distributions for *Versicolor* and *Virginica*.
3. **Chi-Square Test Visualization:**
    - **Stacked Bar Chart:** Distribution of binned Sepal Length across species.

Interpretation of Results

1. **Z-Test Interpretation:**
    - The large Z-statistic and extremely low p-value (<0.0001< 0.0001<0.0001) indicate a statistically significant difference in the mean Petal Length of *Setosa* and *Versicolor*.
2. **T-Test Interpretation:**
    - The negative T-statistic and very low p-value (<0.0001< 0.0001<0.0001)

suggest a significant difference in Sepal Width between *Versicolor* and *Virginica*.

3. **Chi-Square Test Interpretation:**
    - The high Chi-Square statistic and near-zero p-value (<0.0001< 0.0001<0.0001) reveal a strong association between binned Sepal Length and species, indicating that Sepal Length distributions differ significantly across species.

DISCUSSION AND ANALYSIS

Comparison of Z-Test and T-Test

1. **Purpose:**
    - The Z-Test compares means assuming the population standard deviation (σ\sigmaσ) is known or that the sample size is large enough to approximate it.
    - The T-Test is used when σ\sigmaσ is unknown, and the sample standard deviation (sss) is used as an estimate.
2. **Statistical Behavior:**
    - The Z-Test relies on the standard normal distribution (N(0,1)N(0, 1)N(0,1)).
    - The T-Test uses the t-distribution, which is wider and accounts for variability in sss, especially for smaller samples.
3. **Findings:**
    - Both tests yielded significant results (p-value <0.0001< 0.0001<0.0001), but the T-Test accounted for greater variability due to unknown population standard deviation.
    - For large datasets like the Iris dataset, differences in results between the Z-Test and T-Test are minimal.
4. **Conclusion:**
    - The choice of test impacts the interpretation, particularly for smaller datasets or when σ\sigmaσ is unknown.

Categorical vs. Numerical Variable Analysis

1. **Z-Test/T-Test (Numerical Variables):**
    - These tests focus on evaluating differences in means of numerical variables across categories.
    - Example: Comparing Petal Length or Sepal Width for two species highlights quantitative differences.
2. **Chi-Square Test (Categorical Variables):**
    - This test determines associations between categorical variables, such as species and binned Sepal Length.
    - It evaluates the relationship rather than mean differences.
3. **Key Differences:**
    - Z/T-Tests: Analyze quantitative measures (e.g., Petal Length).

    - Chi-Square Test: Identifies dependencies or associations between categorical attributes.
4. **Practical Insights:**
    - Z/T-Tests answer "How much do groups differ?"
    - Chi-Square answers "Are these categories related?"

Impact of Assumptions on Results

1. **Population Standard Deviation Assumption:**
    - The Z-Test assumes known or approximable σ\sigmaσ. Deviations from this assumption can lead to inaccuracies.
    - The T-Test relaxes this assumption, using sss to accommodate variability.
2. **Sample Size:**
    - Larger sample sizes make Z-Test and T-Test results more similar, as sss becomes a reliable proxy for σ\sigmaσ.
    - For small samples, the T-Test is preferred due to its robustness.
3. **Dataset Dependence:**
    - For the Iris dataset with over 50 samples per group, both tests provided consistent results. However, for smaller datasets, the T-Test would likely yield more conservative (wider) confidence intervals.
4. **Conclusion:**
    - Assumptions significantly influence test choice and results. The T-Test is more adaptable for real-world datasets with unknown variability.

CONCLUSION

Summary of Key Findings

- **Z-Test**: There was a significant difference in the means of Petal Length between *Setosa* and *Versicolor*, as indicated by the high Z-statistic and p-value (<0.0001< 0.0001<0.0001).
- **T-Test**: A significant difference in Sepal Width was found between *Versicolor* and *Virginica*, with a T-statistic of -4.57 and p-value (<0.0001< 0.0001<0.0001).
- **Chi-Square Test**: A strong association between binned Sepal Length and species was identified, with a high Chi-Square statistic and p-value (<0.0001< 0.0001<0.0001).
- The statistical tests provided consistent and meaningful results, confirming the relationships and differences within the Iris dataset.

Implications of Results

- The results support the use of hypothesis testing as a powerful tool for identifying significant differences and relationships in data.
- The significant differences in Petal Length and Sepal Width across species suggest that these variables can help differentiate between species, which could be valuable for machine learning classification models.
- The Chi-Square Test's significant result shows that certain categorical variables, like Sepal Length bins, are associated with specific species, which may also provide insights into data grouping for classification tasks.

These findings have implications for data-driven decision-making and model-building in fields like biology, agriculture, and data science, where understanding variable relationships is crucial for classification and prediction tasks.

REFERENCES

UCI Machine Learning Repository. (2019). *Iris Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/iris

- The Iris dataset, used in this analysis, is widely recognized in the field of data science for classification tasks and statistical analysis.

SciPy Documentation. (2024). *Stats Module - Hypothesis Tests*. Retrieved from https://docs.scipy.org/doc/scipy/reference/stats.html

- Provides detailed guidance and examples for implementing statistical tests, including Z-test, T-test, and Chi-square test in Python.

Matplotlib Documentation. (2024). *Matplotlib - Plotting in Python*. Retrieved from https://matplotlib.org/

- Matplotlib was used for generating plots and visualizations in the report, as it provides a versatile and flexible way to display statistical results.

Seaborn Documentation. (2024). *Seaborn - Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org/

- This source was used for the visualizations created during the analysis, offering functions for creating attractive and informative plots.