Title

Assignment Title: Data Analysis and Hypothesis Testing with the Iris Dataset
Name: Rashmi Prasadi
Index Number: 22001654
Date: 28/2/2025

---

1. Introduction

The Iris dataset is a widely used dataset in statistics and machine learning, containing measurements of **Sepal Length, Sepal Width, Petal Length, and Petal Width** for three species of iris flowers: *Setosa, Versicolor, and Virginica.* This report explores the dataset using various statistical and visualization techniques in **RStudio** and conducts hypothesis tests to derive meaningful insights.

---

2. Methodology

The following steps were undertaken for analysis:

1. **Dataset Exploration** - Load the dataset, inspect its structure, compute descriptive statistics.
2. **Data Visualization** - Generate pie charts, bar charts, histograms, and scatterplots.
3. **Hypothesis Testing** - Perform three hypothesis tests (Lower Tail, Upper Tail, Two-Tailed) using a significance level of $\alpha = 0.05$.

All computations were performed using **RStudio** with appropriate built-in functions.

---

3. Results

3.1 Dataset Exploration

- **Structure & Summary Statistics:**
    - The dataset contains **150 observations** and **5 variables (4 numerical, 1 categorical).**

- Species count:
  - Setosa: **50**
  - Versicolor: **50**
  - Virginica: **50**
- Mean, Median, and Standard Deviation:
  - Sepal Length: **Mean = 5.84, Median = 5.80, SD = 0.83**
  - Sepal Width: **Mean = 3.05, Median = 3.00, SD = 0.43**
  - Petal Length: **Mean = 3.76, Median = 4.35, SD = 1.76**
  - Petal Width: **Mean = 1.20, Median = 1.30, SD = 0.76**

3.2 Data Visualization

- **Pie Chart**: Displays the proportion of each species.
- **Bar Chart**: Shows the count of each species.
- **Histograms**: Visualize the distributions of Sepal Length and Petal Length.
- **Scatterplot**: Shows the correlation between Sepal Length and Petal Length.

3.3 Hypothesis Testing

Lower Tail Test: (H_0: $\mu \geq 5.8$, H_1: $\mu < 5.8$)

- **Test Statistic:** -0.3073
- **p-value:** 0.3791
- **Conclusion:** Since *p-value > 0.05*, we **fail to reject H_0**, meaning there is no significant evidence that Sepal Length is less than 5.8 cm.

Upper Tail Test: (H_0: $\mu \leq 3.5$, H_1: $\mu > 3.5$)

- **Test Statistic:** 2.924
- **p-value:** 0.0019
- **Conclusion:** Since *p-value < 0.05*, we **reject H_0** and conclude that Petal Length is significantly greater than 3.5 cm.

Two-Tailed Test: (H_0: $\mu = 3.0$, H_1: $\mu \neq 3.0$)

- **Test Statistic:** 1.6625
- **p-value:** 0.0985
- **Conclusion:** Since *p-value > 0.05*, we **fail to reject H_0**, meaning there is no significant evidence that Sepal Width is different from 3.0 cm.

---

4. Discussion

The exploratory analysis provided insights into the dataset, confirming the presence of three species and the variations in their attributes. The visualizations illustrated how different

attributes are distributed, while hypothesis tests validated or rejected assumptions about the mean values of Sepal Length, Sepal Width, and Petal Length.

---

5. Conclusion

This study demonstrated how **statistical analysis and hypothesis testing** can be applied to a real dataset using RStudio. The insights derived could be beneficial for understanding species differences and feature relationships. Future work could involve applying **machine learning techniques** for species classification.
