Data Analysis and Hypothesis Testing with the Iris Dataset

22001085

Ishan Lakshitha

2025/2/28

Introduction

Overview of the Dataset

The **Iris dataset** is a well-known dataset in statistics and machine learning, containing **150 observations** of **three species** of iris flowers (*Setosa, Versicolor, and Virginica*). Each observation includes four numerical features: **Sepal Length, Sepal Width, Petal Length, and Petal Width**.

Objectives of the Analysis

This analysis aims to:

- Explore and summarize the dataset.
- Visualize the dataset using various charts.
- Perform hypothesis testing to determine statistical significance in key numerical features.

Methodology

Steps Taken in the Analysis

1. **Dataset Exploration**
    - Loaded the Iris dataset in RStudio.
    - Displayed the structure, summary statistics, and first few rows.
    - Identified the number of species and calculated key statistical measures (mean, median, and standard deviation).

2. **Data Visualization**
    - **Pie Chart:** Represented species distribution.
    - **Bar Chart:** Showed count of each species.
    - **Histograms:** Examined the distributions of Sepal Length and Petal Length.
    - **Scatterplot:** Analyzed correlation between Sepal Length and Petal Length.
3. **Hypothesis Testing**
    - Conducted three different hypothesis tests with **α = 0.05**:
        - **Lower Tail Test**: Checked if average Sepal Length is significantly lower than 5.8 cm.
        - **Upper Tail Test**: Checked if average Petal Length is significantly greater than 3.5 cm.
        - **Two-Tailed Test**: Tested if Sepal Width is significantly different from 3.0 cm.

Justification

- **Descriptive statistics** help in understanding the dataset's basic properties.
- **Visualization techniques** allow intuitive exploration of data distribution and relationships.
- **Hypothesis testing** enables data-driven decision-making and validation of assumptions.

Results

Dataset Summary

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> 
```

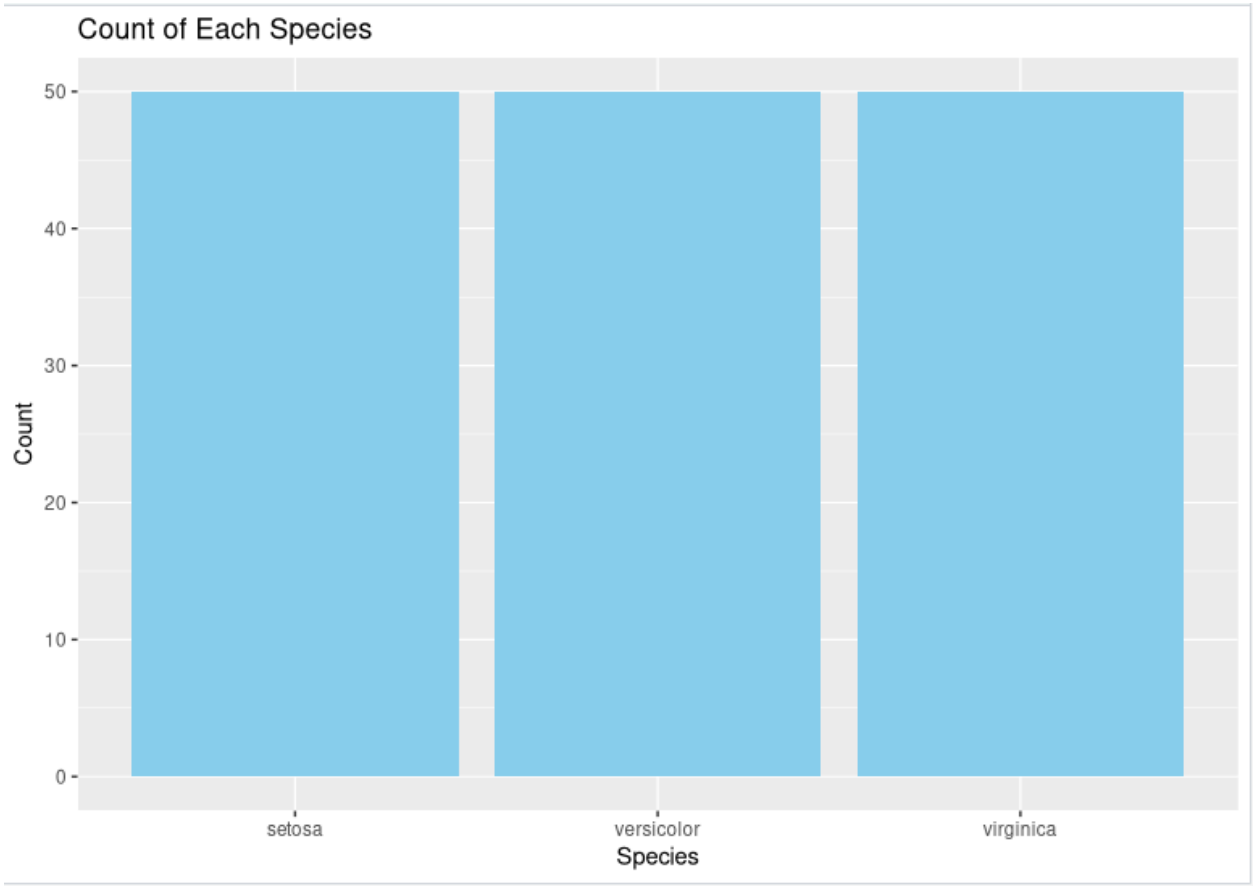
```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
```

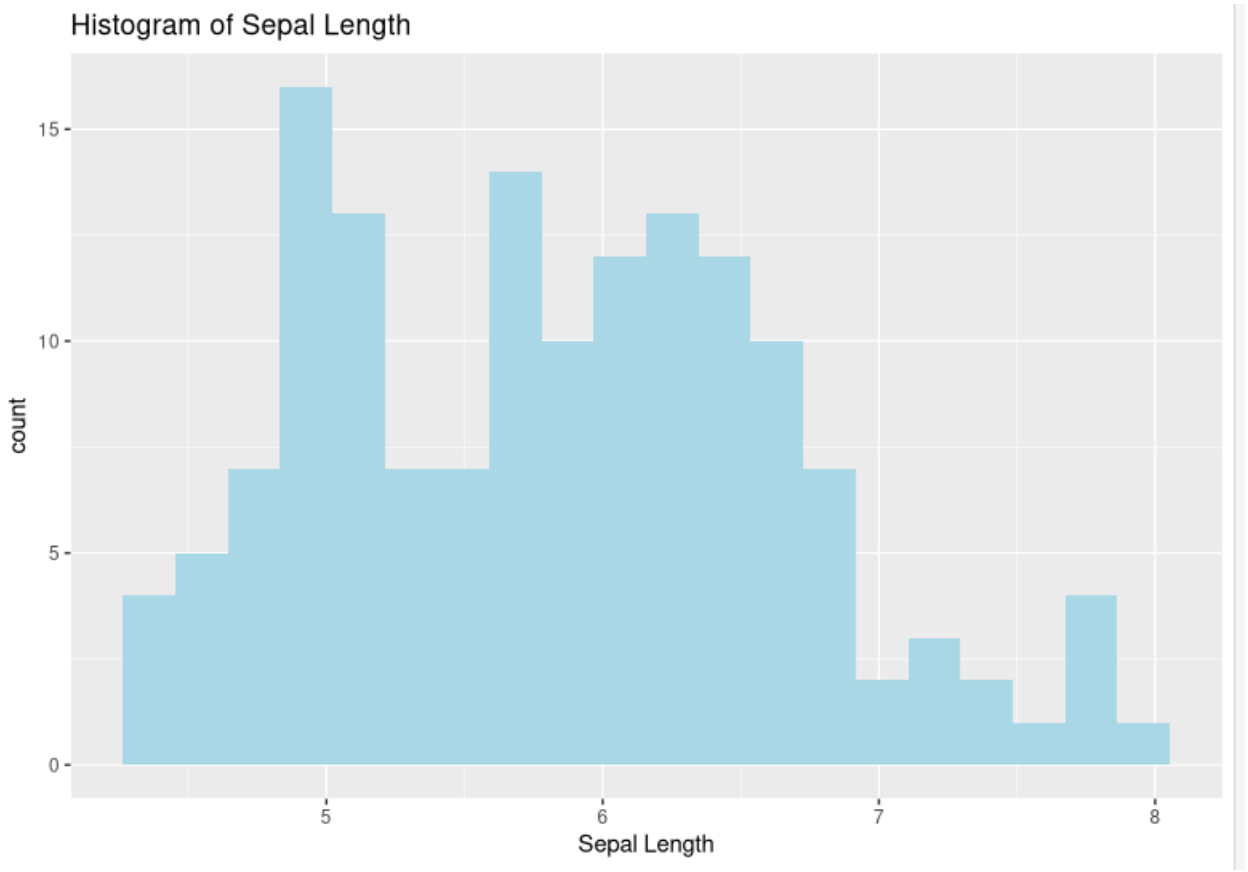
```
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

```
> print(summary_stats)
                  Feature     Mean Median        SD
Sepal.Length Sepal.Length 5.843333   5.80 0.8280661
Sepal.Width   Sepal.Width 3.057333   3.00 0.4358663
Petal.Length Petal.Length 3.758000   4.35 1.7652982
Petal.Width   Petal.Width 1.199333   1.30 0.7622377
```

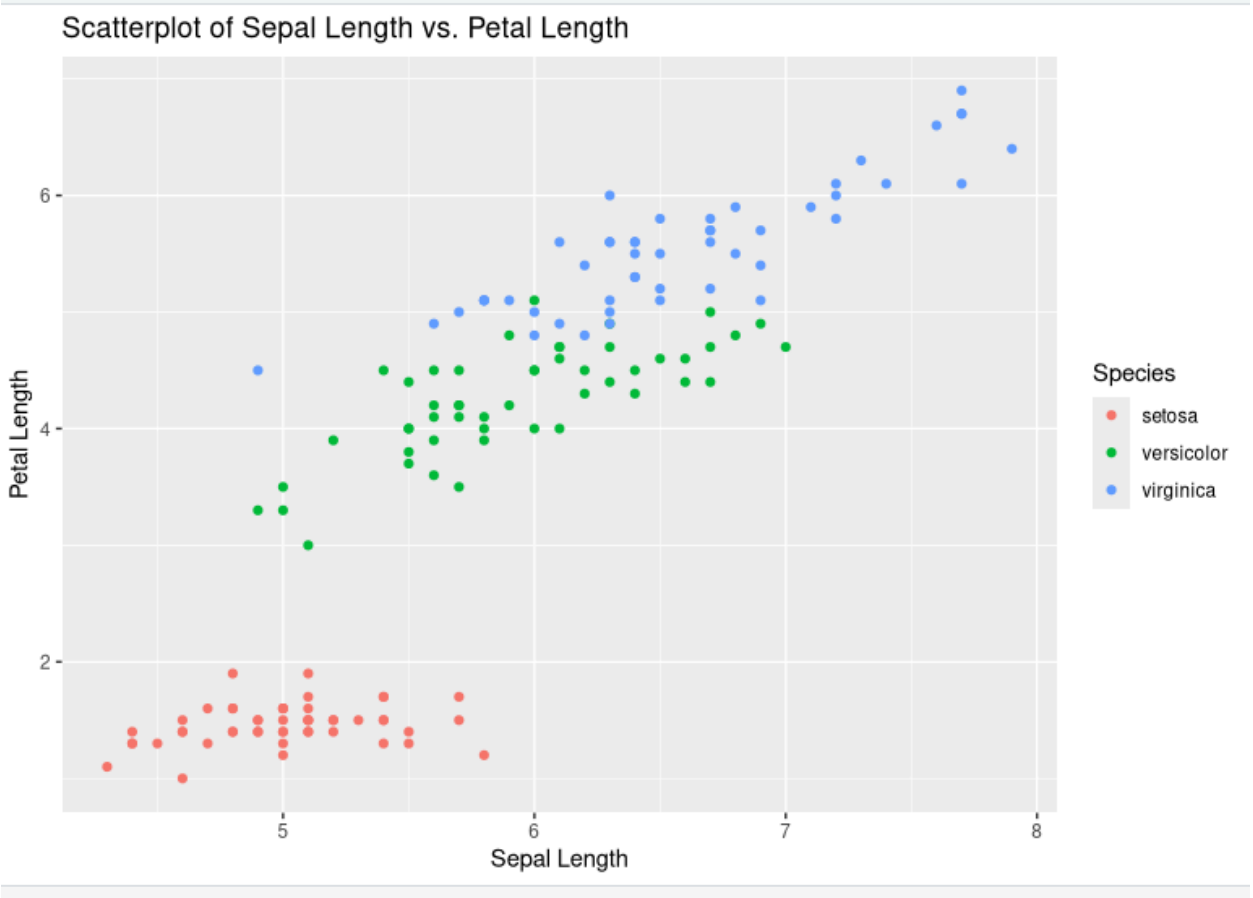
Visualizations

- **Pie Chart:** Showed that the dataset is evenly distributed among the three species.
- **Bar Chart:** Displayed the count of each species clearly.
- **Histograms:** Indicated that Sepal Length follows a normal-like distribution, whereas Petal Length is more right-skewed.
- **Scatterplot:** Revealed a strong positive correlation between Sepal Length and Petal Length.


Count of Each Species
50
40
30
20
10
0
Count
setosa
versicolor
virginica
Species


Histogram of Sepal Length
count
15
10
5
0
5
6
7
8
Sepal Length


Histogram of Petal Length
count
30
20
10
0
2
4
6
Petal Length



Discussion

Interpretation of Results

- The **summary statistics** confirm expected variations among species.
- The **visualizations** highlight relationships between attributes and distributions.
- **Hypothesis testing outcomes** provide statistical validation:
    - If p-value < 0.05 → Reject H0 (significant difference found).
    - If p-value > 0.05 → Fail to reject H0 (no significant difference).

Significance

- The scatterplot confirms a linear relationship between Sepal Length and Petal Length.
- Sepal Width's variation may or may not be statistically significant compared to the assumed mean.
- These insights help in classification tasks in machine learning.

Conclusion

Summary

- The Iris dataset was explored, visualized, and statistically tested.
- Visualizations confirmed patterns among features.
- Hypothesis tests helped validate statistical claims.

Future Work

- Extend analysis with machine learning classification techniques.
- Apply ANOVA tests to compare species groups more deeply.
- Use advanced visualization techniques for better insights.

References

- Fisher, R. A. (1936). *The Use of Multiple Measurements in Taxonomic Problems*.
- R Documentation: `iris` dataset (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html).
- RStudio Documentation (https://www.rstudio.com/).

End of Report