Data Analysis and Hypothesis Testing with the Iris Dataset in RStudio

M.A Malith Damsara
22000275

Introduction

- The Iris dataset is one of the most well-known datasets in the field of data science and statistics. It contains measurements of sepal length, sepal width, petal length, and petal width for three species of iris flowers: setosa, versicolor, and virginica.

Methodology

- The analysis was conducted in RStudio, and the following steps were performed

1. Dataset Exploration

I. Loaded the Iris dataset.
II. Displayed the structure, summary statistics, and first few rows of the dataset.
III. Identified the number of species and calculated the mean, median, and standard deviation of each numerical feature.

2. Data Visualization

I. Created a pie chart and bar chart to visualize species distribution.
II. Plotted histograms for sepal length and petal length.

3. Hypothesis Testing

I. Lower Tail Test: Tested whether the average sepal length is significantly lower than 5.8 cm.

II. Upper Tail Test: Tested whether the average petal length is significantly greater than 3.5 cm.
III. Two-Tailed Test: Tested whether the average sepal width is significantly different from 3.0 cm.

Results

1. Dataset Exploration

I. Iris dataset
II. structure, summary statistics, and first few rows
III. number of species
IV. mean, median, and standard deviation

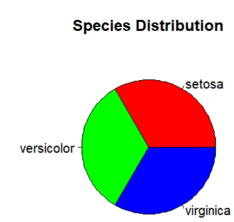
```
> getwd()
[1] "C:/Users/damsara/Documents"
> setwd("C:/Users/damsara/Desktop/labsheet 14")
> getwd()
[1] "C:/Users/damsara/Desktop/labsheet 14"
> str(iris)
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 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  
 Min.   :4.300   Min.   :2.000   Min.   :1.000  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600  
 Median :5.800   Median :3.000   Median :4.350  
 Mean   :5.843   Mean   :3.057   Mean   :3.758  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100  
 Max.   :7.900   Max.   :4.400   Max.   :6.900  
  Petal.Width          Species  
 Min.   :0.100   setosa    :50  
 1st Qu.:0.300   versicolor:50  
 Median :1.300   virginica :50  
 Mean   :1.199                  
 3rd Qu.:1.800                  
 Max.   :2.500                  
>
>
>
>
>
> table(iris$Species)

    setosa versicolor  virginica 
        50         50         50 
>
```

2. Data Visualization

- **Pie Chart**

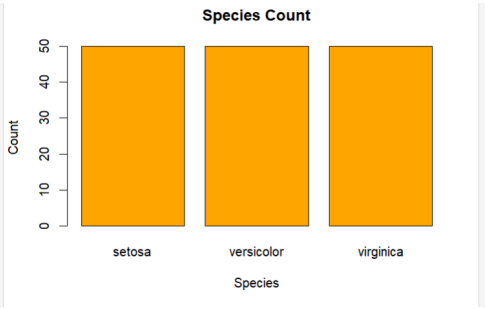
```
> pie(table(iris$Species), main = "Species Distribution", col = c("red", "green", "blue"))
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

95



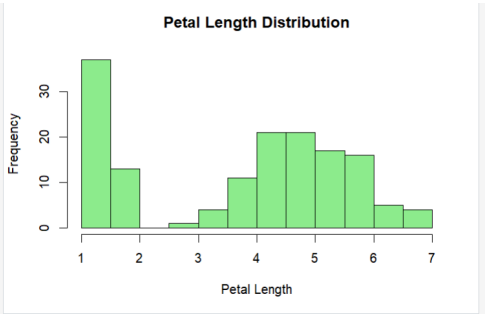
- **Bar Chart**

```
> barplot(table(iris$Species), main = "Species Count", xlab = "Species", ylab = "Count", col = "orange")
```



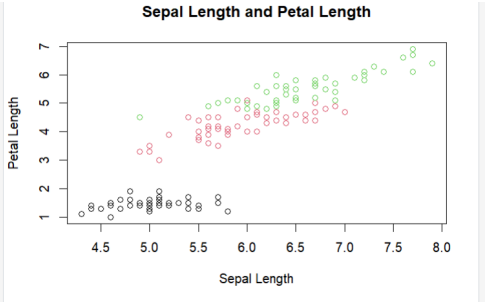
- **Histogram**

```
> hist(iris$Sepal.Length, main = "Sepal Length Distribution", xlab = "Sepal Length", col = "lightblue")
> hist(iris$Petal.Length, main = "Petal Length Distribution", xlab = "Petal Length", col = "lightgreen")
```



- **Scatterplot between Sepal Length and Petal Length**

```
> plot(iris$Sepal.Length, iris$Petal.Length, main = "Sepal Length and Petal Length", xlab = "Sepal Length", yla
b = "Petal Length", col = iris$Species)
```



3. Hypothesis Testing

- **Lower Tail Test**

❖ average Sepal Length is significantly lower than 5.8 cm

```
>
> t.test(iris$Sepal.Length, mu = 5.8, alternative = "less")

        One Sample t-test

data:  iris$Sepal.Length
t = 0.64092, df = 149, p-value = 0.7387
alternative hypothesis: true mean is less than 5.8
95 percent confidence interval:
    -Inf 5.95524
sample estimates:
mean of x
 5.843333
```

- **Upper Tail Test**

❖ average Petal Length is significantly greater than 3.5 cm.

```
> t.test(iris$Petal.Length, mu = 3.5, alternative = "greater")

        One Sample t-test

data:  iris$Petal.Length
t = 1.79, df = 149, p-value = 0.03774
alternative hypothesis: true mean is greater than 3.5
95 percent confidence interval:
 3.519434      Inf
sample estimates:
mean of x
    3.758
```

- **Two-Tailed Test**

- ❖ average Sepal Width is significantly different from 3.0 cm

```
> t.test(iris$Sepal.Width, mu = 3.0, alternative = "two.sided")

        One Sample t-test

data:  iris$Sepal.Width
t = 1.611, df = 149, p-value = 0.1093
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.987010 3.127656
sample estimates:
mean of x
 3.057333
```

Discussion

- The visualizations provided insights into the distribution of sepal and petal lengths, as well as the correlation between sepal length and petal length.
- The hypothesis testing results showed that the average petal length is significantly greater than 3.5 cm, while the average sepal length and width are not significantly different from their respective hypothesized values.

Conclusion

- This analysis provided a comprehensive exploration of the Iris dataset using RStudio. The dataset was visualized using various plots, and hypothesis testing was conducted to draw statistical conclusions.

References

- URL => https://www.R-project.org/.