ESE 326: Final Project

Eric Stewart & Gabriel Minton

November 2024

Contents

1 Introduction **2**

2 Methods **2**
2.1 Exploratory Analysis 2
2.2 Confidence Interval Estimate 2
2.3 Hypothesis Test 2

3 Results and Observations **3**
3.1 Exploratory Analysis 3
3.2 Inference Analysis 6

4 Conclusions **6**

5 Appendix **7**
5.1 Extra Figures and Tables 7
5.2 R-scripts 7
5.2.1 Exploratory.r 7
5.2.2 Inference.r 9

!!! All content must end on page 6 !!!

1 Introduction

The main objective of this project involoved R's built-in dataset, Iris. This dataset contains measurements of 150 iris flowers, from three species: Setosa, Versicolor, and Virginica. Each specimen has measurements of petal length, petal width, sepal length, and sepal width. Through graphical exploration and mathematical analysis, the reseachers determine whether there are clear rules as to which of these features can determine the species of a given specimen. Through these two forms of analysis, the researchers will show if and how the three species of iris flowers can be differentiated by the four features.

2 Methods

2.1 Exploratory Analysis

Exploratory data analysis is meant for betterunderstanding and visualizing the data. This will also allow for researchers to form hypotheses and find patterns. Later in this report, these patterns will be proven or disproven given the evidence found from the mathematical Inference analysis.

For this project, the researchers have created multiple graphs showing how the features of the Iris dataset are separated between species. The first set of graphs, Figure 1, plots pairs of features separated by species. These plots begin to show separation between the features of the different species. This is not enough data to concretely show that there exist statistical differences between the features of each species. Next, the researchers created boxplots for each feature. These, shown in Figure 2, more discretely show separation between the three Iris species. These findings will be discussed further in the following sections.

2.2 Confidence Interval Estimate

2.3 Hypothesis Test

γ was found through Equation 4.

$\sigma_1 == \sigma_2$ T-test using Sp:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{{S_p}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim T_{n_1+n_2-2} \tag{1}$$

$${S_p}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{2}$$

$\sigma_1 \neq \sigma_2$ T-test using γ:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_\gamma \tag{3}$$

$$\gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4}$$

3 Results and Observations

3.1 Exploratory Analysis

Figure 2 depicts the 6 pairings of the four features. These features are petal length, petal width, sepal length, and sepal width. Several images are jumbled together, showing that the species have no statistical difference for that feature. For example, take the Sepal Length column. There is not a distinct horizontal separation between the colored species. This means that the sepal length is roughly the same for each species of iris. Similarly, the Sepal Width column shows that there is not a significant statistical separation between species for that feature. The Petal Length subplots show that Versicolor and Virginica are close, but the Setosa points have much smaller values. This could mean that Setosa is statistically smaller than the other two species. This is supported by Figure 2a. Likewise, the Petal Width subplots show the Versicolor and Virginica points very close, though perhaps still statistically different. The Setosa plots definitely have smaller values than those of the Versicolor and Virginica datapoints. The best plot to see these differences is the Petal Length/ Petal Width plot(s), as the blue setosa group is away from the green and red Versicolor and Virginica groups.

The researchers also prepared a set of boxplots (Figure 2). As a property of boxplots, the vertical separation represents a statistical difference between samples. Figures 2a and 2b show a significant amount of difference between the species for Petal Lengths and Petal Widths, respectively. For each of these features, there is a clear separation of the species. Setosa has the smallest petals, with Versicolor and Virginica being larger. As expected, this is the same result as found from Figure 1. The boxes for Sepal Length show the same pattern, though the Versicolor and Virginica boxes have some overlap (Figure 2c). Figure 2d changes the previous pattern of Setosa having the smallest features. This subplot also has overlap for all three boxes, meaning there is no statistical difference between species for sepal width. This is reflected in the Sepal Width column of Figure 1, where all three species populate the same horizontal region.

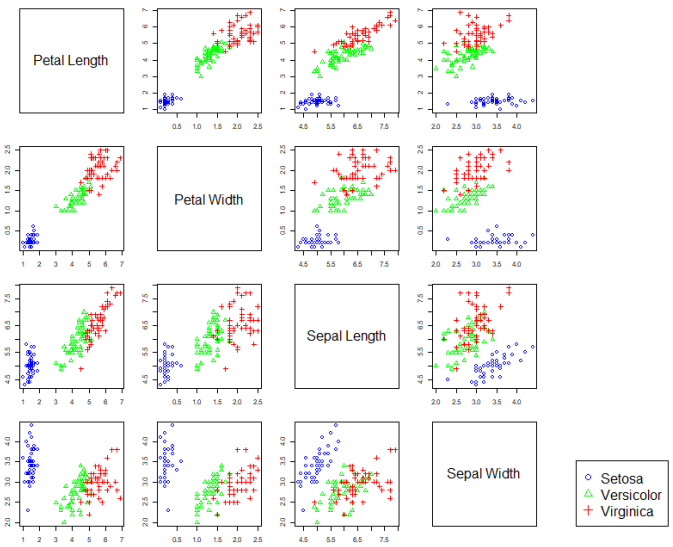
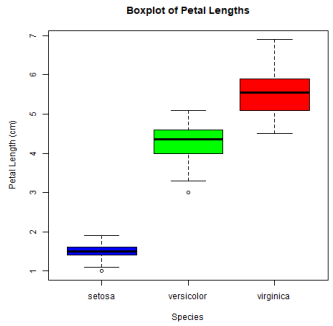
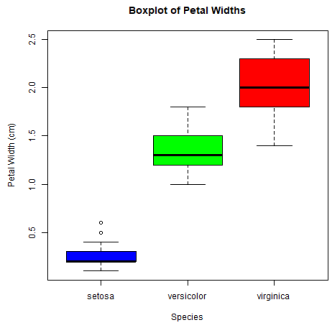


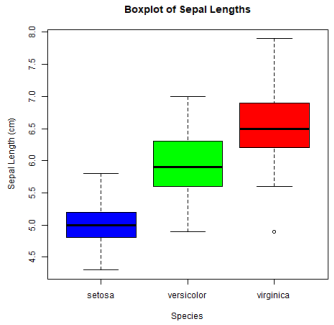
Figure 1: A visualization of the Iris dataset showing scatterplots of each pair of the features.



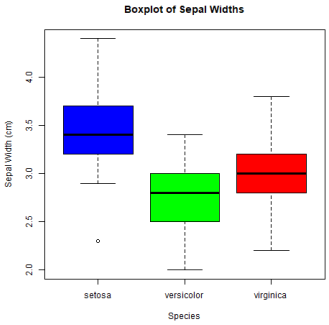
(a) Petal Lengths



(b) Petal Widths



(c) Sepal Lengths



(d) Sepal Widths

Figure 2: Boxplots for each of the four features.

3.2 Inference Analysis

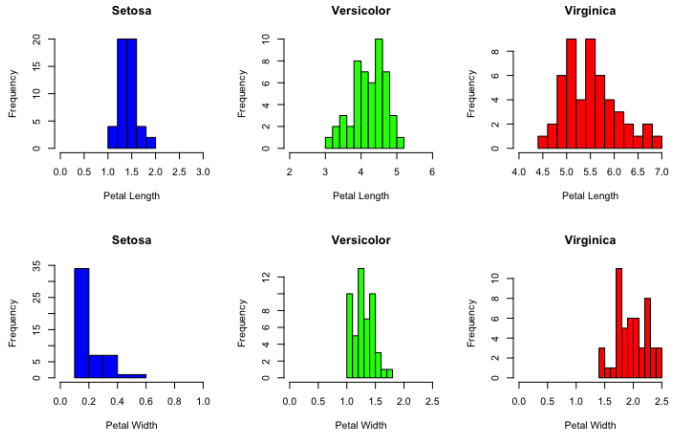


Figure 3: Histograms of Petal lengths and widths for Setosa, Versicolor, and Virginica

4 Conclusions

5 Appendix

5.1 Extra Figures and Tables

5.2 R-scripts

5.2.1 Exploratory.r

```r
# Load packages
library(dplyr)

# Load data
data(iris)
summary(iris)

# Split dataset into different species
names(iris) <- tolower(names(iris))

# Colors:
# Setosa  Versicolor  Virginica
# Blue    Green       Red
# Shapes:
# Circle  Triangle    +
colors <- c('Setosa'='blue', 'Versicolor'='green', 'Virginica'='red')
shapes <- c('Setosa'=1,      'Versicolor'=2,       'Virginica'=3)

# **********************
# PART 1: Plots
# **********************

# IRIS Template
png('iris_gray.png')
plot(iris)
dev.off()

## IRIS Colors
png('iris_colored.png', width = 800, height = 640)
#title()
par(mfrow=c(4, 5), mar=c(2, 2, 2, 2))
plot(x = 0:10, y = 0:10, ann=F, bty='o', type='n', xaxt='n', yaxt='n')
text(x = 5, y = 5, 'Petal Length', cex=2)
plot(iris$petal.width, iris$petal.length, col=colors[iris$species], pch=shapes[iris$species])
plot(iris$sepal.length, iris$petal.length, col=colors[iris$species], pch=shapes[iris$species])
plot(iris$sepal.width, iris$petal.length, col=colors[iris$species], pch=shapes[iris$species])
```

```
37 plot(x = 0:10, y = 0:10, ann=F, bty='n', type='n', xaxt='n',
       yaxt='n')
38
39 plot(iris$petal.length, iris$petal.width, col=colors[iris$
       species], pch=shapes[iris$species])
40 plot(x = 0:10, y = 0:10, ann=F, bty='o', type='n', xaxt='n',
       yaxt='n')
41 text(x = 5, y = 5, 'Petal Width', cex=2)
42 plot(iris$sepal.length, iris$petal.width, col=colors[iris$
       species], pch=shapes[iris$species])
43 plot(iris$sepal.width, iris$petal.width, col=colors[iris$
       species], pch=shapes[iris$species])
44 plot(x = 0:10, y = 0:10, ann=F, bty='n', type='n', xaxt='n',
       yaxt='n')
45
46
47 plot(iris$petal.length, iris$sepal.length, col=colors[iris$
       species], pch=shapes[iris$species])
48 plot(iris$petal.width, iris$sepal.length, col=colors[iris$
       species], pch=shapes[iris$species])
49 plot(x = 0:10, y = 0:10, ann=F, bty='o', type='n', xaxt='n',
       yaxt='n')
50 text(x = 5, y = 5, 'Sepal Length', cex=2)
51 plot(iris$sepal.width, iris$sepal.length, col=colors[iris$
       species], pch=shapes[iris$species])
52 plot(x = 0:10, y = 0:10, ann=F, bty='n', type='n', xaxt='n',
       yaxt='n')
53
54 plot(iris$petal.length, iris$sepal.width, col=colors[iris$
       species], pch=shapes[iris$species])
55 plot(iris$petal.width, iris$sepal.width, col=colors[iris$
       species], pch=shapes[iris$species])
56 plot(iris$sepal.length, iris$sepal.width, col=colors[iris$
       species], pch=shapes[iris$species])
57 plot(x = 0:10, y = 0:10, ann=F, bty='o', type='n', xaxt='n',
       yaxt='n')
58 text(x = 5, y = 5, 'Sepal Width', cex=2)
59 plot(x = 0:10, y = 0:10, ann=F, bty='n', type='n', xaxt='n',
       yaxt='n')
60
61 legend('bottom', legend=c('Setosa', 'Versicolor', 'Virginica
       '), col=colors, pch=shapes, cex = 2)
62 dev.off()
63
64
65 # **********************
66 # PART 2: Boxplots
67 # **********************
68
69 # Sepal Length
```

```
70 png('box_sepal_length.png')
71 boxplot(sepal.length ~ species, data=iris, col=colors,
72         main='Boxplot of Sepal Lengths', xlab='Species',
             ylab='Sepal Length (cm)')
73 dev.off()
74
75
76 # Sepal Width
77 png('box_sepal_width.png')
78 boxplot(sepal.width ~ species, data=iris, col=colors,
79         main='Boxplot of Sepal Widths',
80         xlab='Species', ylab='Sepal Width (cm)')
81 dev.off()
82
83 # Petal Length
84 png('box_petal_length.png')
85 boxplot(petal.length ~ species, data=iris, col=colors,
86         main='Boxplot of Petal Lengths',
87         xlab='Species', ylab='Petal Length (cm)')
88 dev.off()
89
90 # Petal Width
91 png('box_petal_width.png')
92 boxplot(petal.width ~ species, data=iris, col=colors,
93         main='Boxplot of Petal Widths',
94         xlab='Species', ylab='Petal Width (cm)')
95 dev.off()
```

5.2.2 Inference.r

```
1  data(iris)
2
3  # **********************
4  # Part 1
5  # **********************
6
7  # create graph layout
8  par(mfrow = c(2, 3))
9
10 # petal length histograms
11
12 # setosa histogram
13 hist(iris$Petal.Length[iris$Species == "setosa"],
14      main = "Setosa", xlab = "Petal Length", col = "blue",
15      breaks = 5, xlim = c(0,3))
16
17 # versicolor
18 hist(iris$Petal.Length[iris$Species == "versicolor"],
```

```
     main = "Versicolor", xlab = "Petal Length", col = "green",
     breaks = 10, xlim = c(2,6))

# virginica
hist(iris$Petal.Length[iris$Species == "virginica"],
     main = "Virginica", xlab = "Petal Length", col = "red",
     breaks = 10, xlim = c(4,7))

# petal width histograms

# setosa
hist(iris$Petal.Width[iris$Species == "setosa"],
     main = "Setosa", xlab = "Petal Width", col = "blue",
     breaks = 10, xlim = c(0,1))

# versicolor
hist(iris$Petal.Width[iris$Species == "versicolor"],
     main = "Versicolor", xlab = "Petal Width", col = "green",
     breaks = 10, xlim = c(0,2.5))

# virginica
hist(iris$Petal.Width[iris$Species == "virginica"],
     main = "Virginica", xlab = "Petal Width", col = "red",
     breaks = 10, xlim = c(0,2.5))

# Reset layout
par(mfrow = c(1, 1))

# **********************
# Part 2
# **********************

# function to calculate the confidence interval on the mean of a normal distribution with an unknown variance
confidence_interval_unknown_variance <- function(sample, confidence = 0.95) {
  n <- length(sample)
  sample_mean <- mean(sample)
  sample_sd <- sd(sample)
  a <- 1-confidence
  t <- qt(1-a/2, n-1)
  z <- t*sample_sd/sqrt(n)
  lower_bound <- sample_mean - z
  upper_bound <- sample_mean + z
  return(c(lower_bound, upper_bound))
}

# calculate the confidence interval using the function
```

```r
setosa_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$Species == "setosa"])
versicolor_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$Species == "versicolor"])
virginica_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$Species == "virginica"])

# print results
print("Confidence Interval for Petal Length (Setosa):")
cat('Lower Bound:', setosa_ci[1], "\n")
cat('Upper Bound:', setosa_ci[2], "\n")

print("Confidence Interval for Petal Length (Versicolor):")
cat('Lower Bound:', versicolor_ci[1], "\n")
cat('Upper Bound:', versicolor_ci[2], "\n")

print("Confidence Interval for Petal Length (Verginica):")
cat('Lower Bound:', virginica_ci[1], "\n")
cat('Upper Bound:', virginica_ci[2], "\n")

# # **********************
# # Part 5
# # **********************
#
#
# # function to preform a hypothesis on two samples
# mean_hypothesis_test <- function(sample1, sample2, conf_level = 0.95) {
#
#   n1 <- length(sample1)        # size of sample1
#   n2 <- length(sample2)        # size of sample2
#   mean1 <- mean(sample1)       # sample mean for sample1
#   mean2 <- mean(sample2)       # sample mean for sample2
#   sd1 <- sd(sample1)           # sample sd for sample1
#   sd2 <- sd(sample2)           # sample sd for sample2
#
### check for equality in variances by testing H0: $\sigma_1 = \sigma_2$: H1: $\sigma_1 \neq \sigma_2$
# }

# **********************
# Part 6
# **********************

sample1 = iris$Petal.Length[iris$Species == "virginica"]
sample2 = iris$Petal.Length[iris$Species == "versicolor"]
conf_level <- 0.95


```

```r
mean_hypothesis_test <- function(sample1, sample2, conf_level = 0.95) {
  n1 <- length(sample1)        # size of sample1
  n2 <- length(sample2)        # size of sample2
  mean1 <- mean(sample1)       # sample mean for sample1
  mean2 <- mean(sample2)       # sample mean for sample2
  sd1 <- sd(sample1)           # sample sd for sample1
  sd2 <- sd(sample2)           # sample sd for sample2

  ### check for equality in variances by testing H0: σ1 = σ2: H1: σ1 ≠ σ2
  # test statistic
  t_stat <- (sd1^2)/(sd2^2)
  # significance level
  alpha <- 1 - conf_level
  # p value
  p_value <- 2*(1-pf(t_stat,n1,n2))

  # sigma values are unknown and equal
  if (p_value > alpha) {
    print("␣ σ1 = σ2 ")
    # degrees of freedom
    df <- n1 + n2 - 2
    # pooled valiance
    Sp <- ((n1-1)*sd1^2 + (n2-1)*sd2^2)/(n1 + n2 - 2)
    # observed value
    T_obs <- (mean1 - mean2)/(sqrt(Sp*((1/n1) + (1/n2))))
    # p value
    pt_value <- pt(T_obs, df)

    # sigma values are unknown and unequal
  } else if (p_value < alpha) {
    print("␣ σ1 = σ2 ")
    # degree of freedom
    gamma <- (((sd1^2/n1) + (sd2^2/n2))^2) / (((sd1^2/n1)^2 / (n1 - 1)) + ((sd2^2/n2)^2 / (n2 - 1)))
    # observed value
    T_obs <- (mean1 - mean2)/(sqrt(((sd1^2/n1) + (sd2^2/n2))))
    # p value
    pt_value <- pt(T_obs, gamma)

  }

  print('p␣value␣final')
  print(pt_value)
  if (pt_value > alpha) {
    result <- "Accept␣H0"
  } else if (pt_value < alpha) {
    result <- "Reject␣H0"
```

```
  }
  return(result)
}

first_test <- mean_hypothesis_test(iris$Petal.Length[iris$Species == "virginica"],
                                   iris$Petal.Length[iris$Species == "versicolor"],
                                   0.95)
second_test <- mean_hypothesis_test(iris$Petal.Length[iris$Species == "versicolor"],
                                    iris$Petal.Length[iris$Species == "setosa"],
                                    0.95)
print("Petal length of Virginica iris is larger than that of Versicolor")
print(first_test)
print("Petal length of Versicolor iris is larger than that of Setosa")
print(second_test)
```