

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	5
4	Conclusions	6
5	Appendix	7
5.1	R-scripts	7
5.1.1	Exploratory.r	7
5.1.2	Inference.r	9

1 Introduction

The main objective of this project involved 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 researchers 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 better understanding 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

A confidence interval estimate is a range of values, derived from sample data, that is likely to contain a true population parameter with a specified level of confidence. In this case, the sample data consist of the various parameters of the Iris dataset, with the confidence interval estimates specifically focused on their sample means. Since the sample size is large (greater than 30), the sample means will follow a normal distribution according to the Central Limit Theorem. This justifies the use of the Z-distribution for confidence interval estimation as seen in Eq. 1.

$$\bar{X} \sim N\left(\mu, \frac{S^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1) \quad (1)$$

After establishing that the sample means follow a normal distribution, the confidence intervals for each parameter in the Iris dataset were calculated using R. For each parameter, the corresponding confidence interval provides a range of values that are likely to contain the true population mean with a specified confidence level. These intervals are then used to assess the variability and reliability of the sample estimates.

2.3 Hypothesis Test

The next step was to use a hypothesis test to compare the means of two population samples. Since the variance is unknown for each population, the test for equality of variances was conducted first using a p-value test on an F-distribution. If the variances are found to be equal ($\sigma_1 = \sigma_2$), a pooled sample variance is used as seen in Eq. 2.

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

This result also leads to the test statistic following a t-distribution with degrees of freedom equivalent to

$n_1 + n_2 - 2$, where n_1 and n_2 are the sample sizes for each population. The test statistic, calculated using the pooled variance, is given by Eq. 3

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim T_{n_1+n_2-2} \quad (3)$$

If the variances are found to be not equal ($\sigma_1 \neq \sigma_2$), a different test statistic must be used. The test statistic, as shown in Eq. 4, will follow a t-distribution with degrees of freedom γ .

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_\gamma \quad (4)$$

When the variances of the two populations are unequal, the degrees of freedom (γ) are not simple the sum of the sample sizes minus two. Instead, the degrees of freedom are calculated using Eq. 5.

$$\gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2-1}} \quad (5)$$

After calculating the appropriate test statistic for the hypothesis test, the next step is to determine to reject or fail to reject the null hypothesis. The p-value test was used for this decision. If the p-value is less than the significance level α , the null hypothesis is rejected. If the p-value is greater than or equal to the significance level α , the null hypothesis is not rejected.

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.

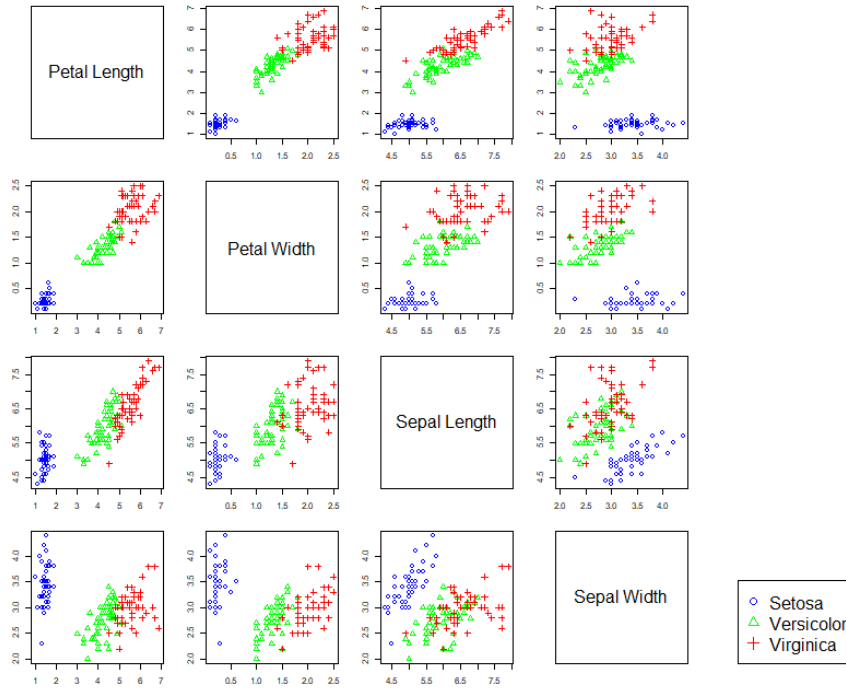


Figure 1: A visualization of the Iris dataset showing scatterplots of each pair of the features.

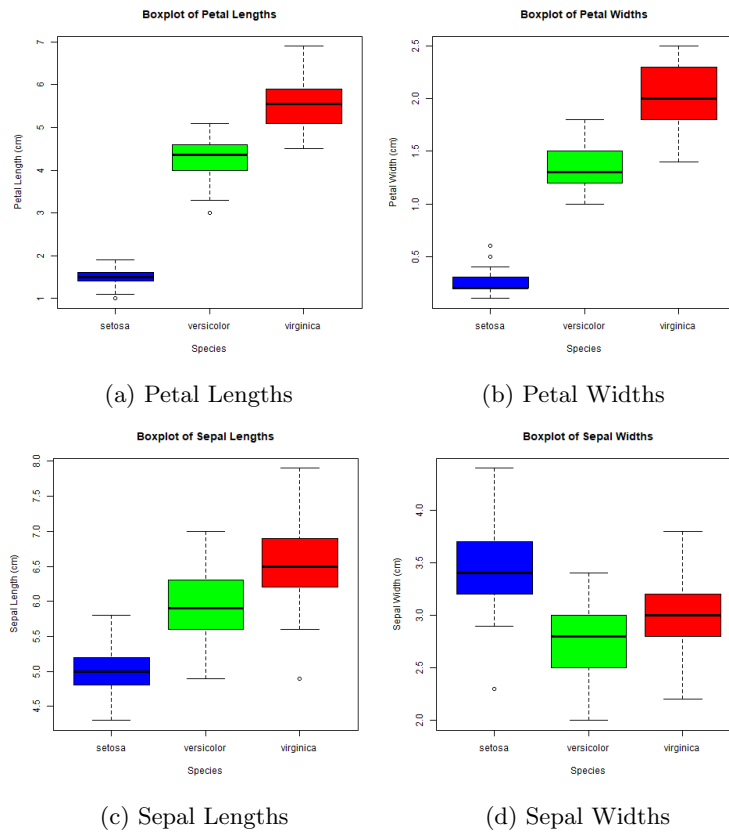


Figure 2: Boxplots for each of the four features.

3.2 Inference Analysis

The analysis begins with visualizing the distributions of the petal length and petal width using histograms. This allows for an initial assessment of the underlying patterns and potential differences between the groups before performing statistical tests. Similar to the exploratory analysis, it is noticeable in Figure 3 that the virginica species has a larger petal length and width with an approximate mean of 5.5 cm and 2 cm respectively. The results also show that the setosa species has the smallest petal length and width with sample means at 1.5 and 0.2 respectively.

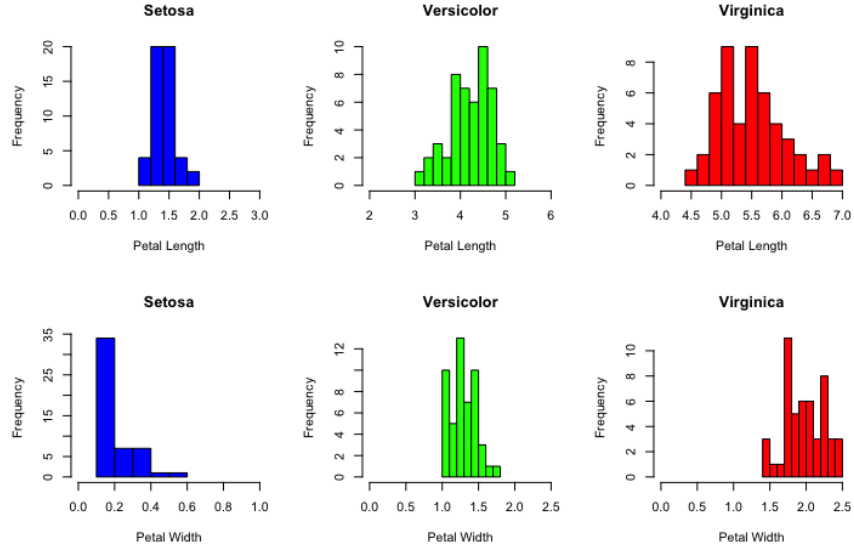


Figure 3: Histograms of Petal lengths and widths for Setosa, Versicolor, and Virginica

The 95% confidence intervals for the petal lengths of the three iris species were calculated to estimate the population mean for each species. The confidence intervals for petal length were [1.41, 1.51] cm for the Setosa species, [4.13, 4.39] cm for the Versicolor species, and [5.4, 5.7] cm for the Virginica species. These intervals indicate that the petal lengths of each species are distinct from one another, as there is no overlap between the intervals. Although the intervals provide an estimate of the population mean for each species at a 95% confidence level, they do not conclusively confirm whether the populations are significantly different. A formal hypothesis test would be necessary to make this determination.

Since the population standard deviation (σ) was unknown for the iris species, a test was conducted to determine whether the variances of the two samples were equal, using a significance level of 0.05. The result of this variance equality test determined the appropriate t-distribution for the subsequent hypothesis test. For the first hypothesis test, the null hypothesis (H_0) stated that the petal length of Virginica iris is larger than or equal to that of Versicolor. The alternative hypothesis (H_1) stated that the petal length of Versicolor iris is larger than that of Virginica. The p-value for the variance equality test was 0.259, which is greater than the significance level ($\alpha = 0.05$), indicating that it is reasonable to assume the variances of the two populations are equal. Based on this result the t-distribution for the hypothesis test used the pooled variance, as given in Eq. 3. The resulting p-value for this hypothesis test was 1, indicating that there is no evidence to reject the null hypothesis with a significance value of 0.05. Therefore, it cannot be concluded that the petal length of Versicolor iris is larger than that of Virginica.

For the second hypothesis test, the null hypothesis (H_0) stated that the petal length of Versicolor iris is larger than or equal to that of Setosa iris, while the alternative hypothesis (H_1) stated that the petal length of Setosa iris is larger than that of Versicolor. The test for variance equality resulted in a p-value of 6.6×10^{-11} , indicating that the variances of the two populations are not equal. As a result, the t-distribution used the test statistic for unequal variances, as described in Eq. 4. The final p-value for this hypothesis

test was 1, showing no evidence to reject the null hypothesis with a significance value of 0.05. Therefore, it cannot be concluded that the petal length of Setosa iris is larger than that of Versicolor. These results align with the exploratory analysis and confidence interval findings, which also indicated differences in petal lengths among species but did not provide evidence to support the specific claims tested in these hypothesis tests.

4 Conclusions

5 Appendix

5.1 R-scripts

5.1.1 Exploratory.r

```
1 # Load packages
2 library(dplyr)
3
4 # Load data
5 data(iris)
6 summary(iris)
7
8 # Split dataset into different species
9 names(iris) <- tolower(names(iris))
10
11 # Colors:
12 # Setosa Versicolor Virginica
13 # Blue   Green   Red
14 # Shapes:
15 # Circle Triangle +
16 colors <- c('Setosa'='blue', 'Versicolor'='green', 'Virginica'='red')
17 shapes <- c('Setosa'=1, 'Versicolor'=2, 'Virginica'=3)
18
19 # *****
20 # PART 1: Plots
21 # *****
22
23 # IRIS Template
24 png('iris_gray.png')
25 plot(iris)
26 dev.off()
27
28 ## IRIS Colors
29 png('iris_colored.png', width = 800, height = 640)
30 #title()
31 par(mfrow=c(4, 5), mar=c(2, 2, 2, 2))
32 plot(x = 0:10, y = 0:10, ann=F, bty='o', type='n', xaxt='n', yaxt='n')
33 text(x = 5, y = 5, 'Petal Length', cex=2)
34 plot(iris$petal.width, iris$petal.length, col=colors[iris$species], pch=shapes[
  iris$species])
35 plot(iris$sepal.length, iris$petal.length, col=colors[iris$species], pch=shapes[
  iris$species])
36 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()

```


5.1.2 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"],
19      main = "Versicolor", xlab = "Petal_Length", col = "green",
20      breaks = 10, xlim = c(2,6))
21
22 # virginica
23 hist(iris$Petal.Length[iris$Species == "virginica"],
24      main = "Virginica", xlab = "Petal_Length", col = "red",
25      breaks = 10, xlim = c(4,7))
26
27 # petal width histograms
28
29 # setosa
30 hist(iris$Petal.Width[iris$Species == "setosa"],
31      main = "Setosa", xlab = "Petal_Width", col = "blue",
32      breaks = 10, xlim = c(0,1))
33
34 # versicolor
35 hist(iris$Petal.Width[iris$Species == "versicolor"],
36      main = "Versicolor", xlab = "Petal_Width", col = "green",
37      breaks = 10, xlim = c(0,2.5))
38
39 # virginica
40 hist(iris$Petal.Width[iris$Species == "virginica"],
41      main = "Virginica", xlab = "Petal_Width", col = "red",
42      breaks = 10, xlim = c(0,2.5))
43
44 # Reset layout
45 par(mfrow = c(1, 1))
46
47
48 ### Comments on the distributions
49 # Since the sample statistics have unknown population variances and the sample
50 # sizes are larger than 30, they follow a normal distribution.
51 # Petal length parameters (approximations from histogram)
52 # * Setosa
53 # ** Sample Mean      = 1.5
54 # ** Sample Variance  = 0.5
55 # * Versicolor
56 # ** Sample Mean      = 4.25
57 # ** Sample Variance  = 0.45
```

```

57 # * Virginica
58 # ** Sample Mean      = 5.5
59 # ** Sample Variance  = 0.5
60 # Petal width parameters
61 # * Setosa
62 # ** Sample Mean      = 0.25
63 # ** Sample Variance  = 0.1
64 # * Versicolor
65 # ** Sample Mean      = 1.3
66 # ** Sample Variance  = 0.2
67 # * Virignica
68 # ** Sample Mean      = 2
69 # ** Sample Variance  = 0.25
70
71
72 # *****
73 # Part 2
74 # *****
75
76 # function to calculate the confidence interval on the mean of a normal
    distribution with an unknown variance
77 confidence_interval_unknown_variance <- function(sample, confidence = 0.95) {
78   n <- length(sample)
79   sample_mean <- mean(sample)
80   sample_sd <- sd(sample)
81   a <- 1 - confidence
82   z <- qnorm(1 - a/2,0,1)
83   margin_of_error <- z * sample_sd / sqrt(n)
84   lower_bound <- sample_mean - margin_of_error
85   upper_bound <- sample_mean + margin_of_error
86   return(c(lower_bound, upper_bound))
87 }
88
89 # *****
90 # Part 3
91 # *****
92
93 # calculate the confidence interval using the function
94 setosa_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$Species
    == "setosa"])
95 versicolor_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$
    Species == "versicolor"])
96 virginica_ci <- confidence_interval_unknown_variance(iris$Petal.Length[iris$
    Species == "virginica"])
97
98 # print results
99 print("Confidence Interval for Petal Length (Setosa):")
100 cat('Lower Bound:', setosa_ci[1], "\n")
101 cat('Upper Bound:', setosa_ci[2], "\n")
102
103 print("Confidence Interval for Petal Length (Versicolor):")
104 cat('Lower Bound:', versicolor_ci[1], "\n")
105 cat('Upper Bound:', versicolor_ci[2], "\n")
106
107 print("Confidence Interval for Petal Length (Verginica):")
108 cat('Lower Bound:', virginica_ci[1], "\n")
109 cat('Upper Bound:', virginica_ci[2], "\n")
110
111 # *****

```

```

112 # Part 4
113 # *****
114
115 # confidence intervals
116 # Setosa has confidence intervals of [1.41, 1.51], which means that with 95
    percent confidence the populaiton mean lies within these bounds.
117 # Versicolor has confidence intervals of [4.14, 4.39], which means that with 95
    percent confidence the populaiton mean lies within these bounds.
118 # Virginica has confidence intervals of [5.4, 5.7], which means that with 95
    percent confidence the populaiton mean lies within these bounds.
119
120 # *****
121 # Part 5
122 # *****
123
124 sample1 = iris$Petal.Length[iris$Species == "virginica"]
125 sample2 = iris$Petal.Length[iris$Species == "versicolor"]
126 conf_level <- 0.95
127
128
129 mean_hypothesis_test <- function(sample1, sample2, conf_level = 0.95) {
130     n1 <- length(sample1)          # size of sample1
131     n2 <- length(sample2)          # size of sample2
132     mean1 <- mean(sample1)          # sample mean for sample1
133     mean2 <- mean(sample2)          # sample mean for sample2
134     sd1 <- sd(sample1)              # sample sd for sample1
135     sd2 <- sd(sample2)              # sample sd for sample2
136
137     ### check for equality in variances by testing H0: sigma1=sigma2 : H1: sigma1 !=
        sigma2
138     # test statistic
139     t_stat <- (sd1^2)/(sd2^2)
140     # significance level
141     alpha <- 1 - conf_level
142     # p value
143     p_value <- 2*(1-pf(t_stat,n1,n2))
144
145     # sigma values are unknown and equal
146     if (p_value > alpha) {
147         print("sigma1=simga2")
148         # degrees of freedom
149         df <- n1 + n2 - 2
150         # pooled valiance
151         Sp <- ((n1-1)*sd1^2 + (n2-1)*sd2^2)/(n1 + n2 - 2)
152         # observed value
153         T_obs <- (mean1 - mean2)/(sqrt(Sp*((1/n1) + (1/n2))))
154         # p value
155         pt_value <- pt(T_obs, df)
156
157     # sigma values are unknown and unequal
158     } else if (p_value < alpha) {
159         print("simga1!=simga2")
160         # degree of freedom
161         gamma <- (((sd1^2/n1) + (sd2^2/n2))^2) / (((sd1^2/n1)^2 / (n1 - 1)) + ((sd2^2/
            n2)^2 / (n2 - 1)))
162         # observed value
163         T_obs <- (mean1 - mean2)/(sqrt(((sd1^2/n1) + (sd2^2/n2))))
164         # p value
165         pt_value <- pt(T_obs, gamma)

```

```

166 }
167
168 print('variance_p_value')
169 print(p_value)
170 print('p_value_final')
171 print(pt_value)
172 if (pt_value > alpha) {
173     result <- "Accept_H0"
174 } else if (pt_value < alpha) {
175     result <- "Reject_H0"
176 }
177 return(result)
178 }
179
180 # *****
181 # Part 6
182 # *****
183
184 first_test <- mean_hypothesis_test(iris$Petal.Length[iris$Species == "virginica"],
185                                   iris$Petal.Length[iris$Species == "versicolor"],
186                                   0.95)
187 second_test <- mean_hypothesis_test(iris$Petal.Length[iris$Species == "versicolor"
188 ],
189                                   iris$Petal.Length[iris$Species == "setosa"],
190                                   0.95)
191 print("Petal_length_of_Virginica_iris_is_larger_than_that_of_Versicolor")
192 print(first_test)
193 print("Petal_length_of_Versicolor_iris_is_larger_than_that_of_Setosa")
194 print(second_test)
195
196 # *****
197 # Part 7
198 # *****
199
200 # Petal length of Virginica iris is larger than that of Versicolor
201 # H0:  $\mu_1 \geq \mu_2$  : H1:  $\mu_2 > \mu_1$ 
202 # population variance test results in pooled variance (p-value = 0.259)
203 # failed to reject null hypothesis because p-value = 1
204
205 # Petal length of Versicolor iris is larger than that of Setosa
206 # H0:  $\mu_1 \geq \mu_2$  : H1:  $\mu_2 > \mu_1$ 
207 # population variance test results in unequal population variance (p-value = 6.6e
    -11)
    # failed to reject null hypothesis because p-value = 1

```