Math 338 **Lab Assignment #20** Fall 2019

In this lab, we will test whether there is a difference in the mean sepal length among all three iris species.

To start, let's do some exploratory data analysis.

```
> library(ggplot2)
> iris_boxplot <- ggplot(iris, aes(x = Species, y = Sepal.Length)) +
geom_boxplot() + labs(title = "", x = "", y = "") # You finish the code!
Optional to add + coord_flip() to make horizontal boxplots instead
```

Question #1 Insert the final set of boxplots below.

Please see first attached PDF.

Question #2 Write the (null) hypothesis for a one-way ANOVA test of the claim that the three species have the same mean sepal length. Recall that this is a Fisher-type test and so only a null hypothesis is specified.

For our null hypothesis, there is no difference in petal length, therefore all of the species will have the same petal length.

Now let's check our variability.

```
> library(dplyr)
> iris %>% group_by(Species) %>% summarize(mean = mean(Sepal.Length), sd =
sd(Sepal.Length)) # No need to store in a variable – just output it
directly to console
```

Question #3 What are the three sample standard deviations? Is our rule of thumb about the population standard deviation satisfied?

The three standard deviations are 0.352, 0.516, 0.636 . These correspond with the setosa, versicolor and virginica flowers respectively. The lowest standard deviation is 0.352 and that doubled is going to be more than the largest standard deviation of 0.636. Therefore our rule of thumb applies.

We will now do the one-way ANOVA test in R:

```
> iris_anova <- aov(Sepal.Length ~ Species, data = iris)
> summary(iris_anova)
```

Question #4 Copy and paste the output from the summary function below. Note that this table does not include the "Total" row shown in lecture (we can derive all numbers in that row from the values in the table).

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species      2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question #5 From the output in **Question #4**, identify the test statistic, the sampling distribution it comes from (don't forget to include all relevant degrees of freedom parameters), and its observed value in this sample of 150 flowers.

The value of the test statistic is 119.3 and the f distribution of(2, 147)

Question #6 What is the p-value for this one-way ANOVA F test? At the 5% significance level, is the model assumption "all three populations have the same mean sepal length" reasonable?

If the p-value is really small then we should not assume that they are not the same.

Your answer to **Question #6** should be that the model assumption is not reasonable. In this case, we want to do *post hoc* tests to determine *which* means are different. Let's perform the Tukey Honestly Significant Difference post hoc test:

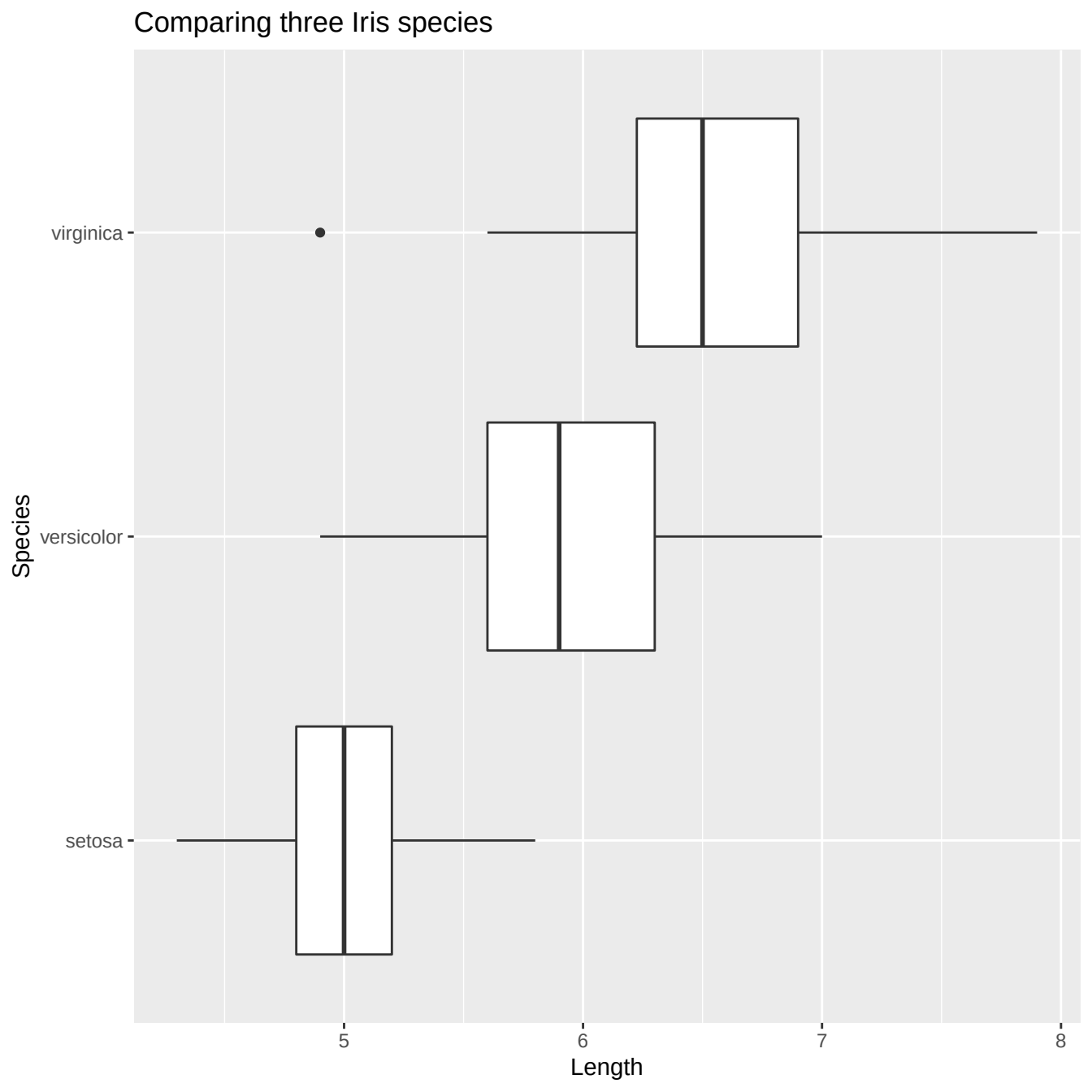
```
> TukeyHSD(iris_anova)
```

Question #7 Copy and paste below the part of the output starting at **`Fit: aov(formula = Sepal.Length ~ Species, data = iris)`**. Which pairs of means appear to be different?

```
Fit: aov(formula = Sepal.Length ~ Species, data = iris)

$Species
                     diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

We reject the null hypothesis because our p-value is less than our significance level.


Comparing three Iris species
virginica
versicolor
setosa
Species
5
6
7
8
Length