**BIFX 503: Statistics for Bioinformatics**

**Problem Set #6**

**Due: October 31, 2018**

*Instructions:*

*Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

The statistician and geneticist Ronald Fisher, inventor the Analysis of Variance method, did agricultural research at the Rothamsted Experimental Station in England. One of his research interests was classification based on quantitative traits. The dataset you are using for this problem set is extracted from his historical data on morphologic variation of three species of iris: setosa, virginica, and versicolor. It is the **iris** dataset that is part of the R datasets package.

Test the hypothesis that the four characteristics differ by species using analysis of variance (ANOVA).

1. What are the assumptions of ANOVA? Are they met?

There are three assumptions for ANOVA test.

1.The first assumption is data independence which means that the data we get and the way we set up the experiment should be independent from one another, and there should be no correlations.

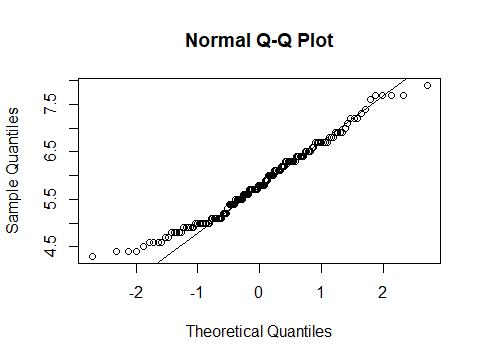
2. The second assumption is normal distribution which means that the data is spread normally around the mean.

3.Equality (or "homogeneity") of variances, called homoscedasticity which means that the variances are equal meaning that there is an equal spread of data.

I would say yes for iris dataset because it passes all three assumptions.

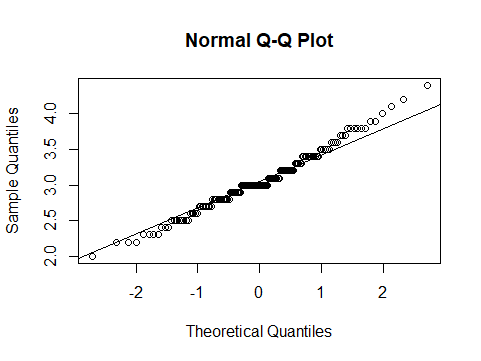
qqnorm(iris$Sepal.Length)

qqline(iris$Sepal.Length)



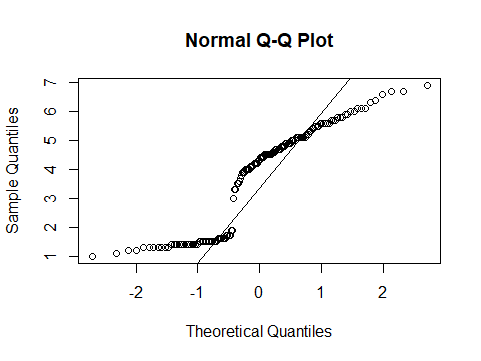
qqnorm(iris$Sepal.Width)

qqline(iris$Sepal.Width)



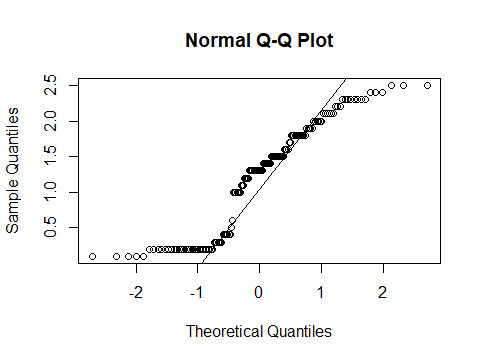
qqnorm(iris$Petal.Length)

qqline(iris$Petal.Length)



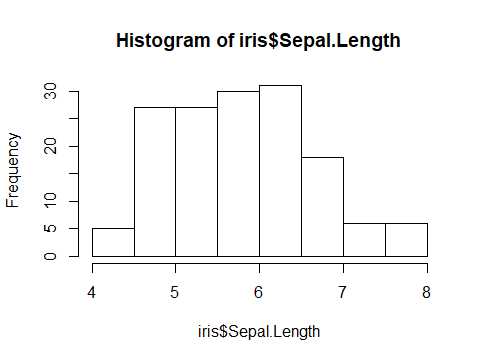
qqnorm(iris$Petal.Width)

qqline(iris$Petal.Width)



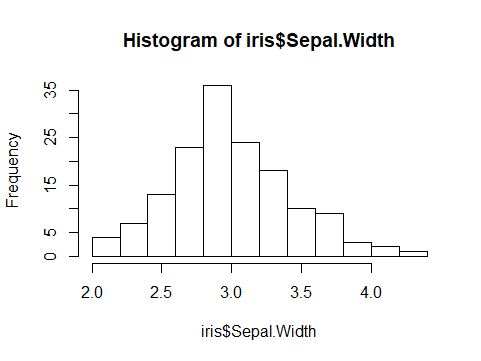
The QQ plots above are each forming lines that are roughly straight. The QQ-Plot checks the theoretical points of a normal distribution vs. the actual points for each of our flowers. While there are some deviations around the tails for setosa and virginica, the points are mostly along the curve. This is suggesting our data came from a normal distribution, but with some variance around the tails.

hist(iris$Sepal.Length)



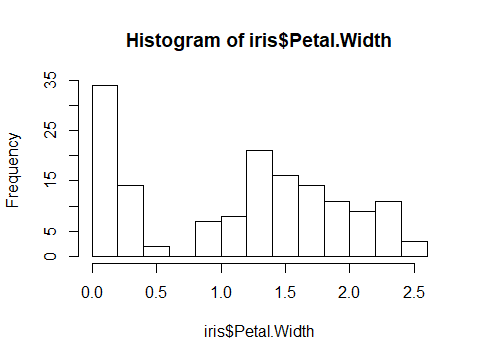
Histogram for Sepal.Length shows normal distribution.

hist(iris$Sepal.Width)



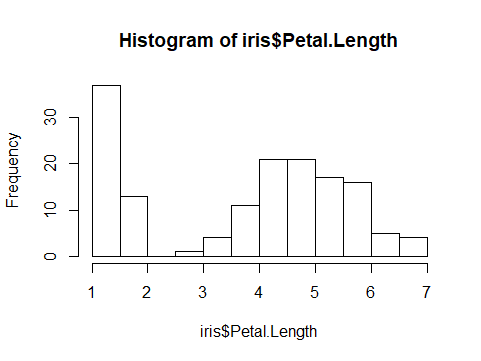
Histogram for Sepal.Width shows normal distribution.

hist(iris$Petal.Width)



Histogram for Petal.Width shows normal distribution.

hist(iris$Petal.Length)

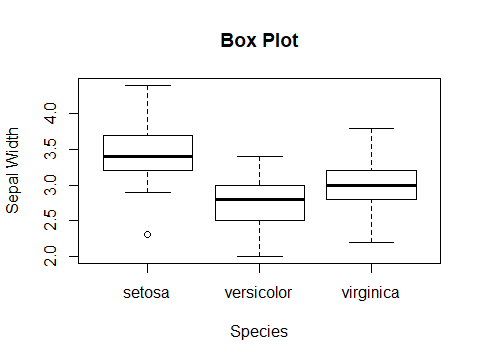


Histogram for Petal.Length shows normal distribution.

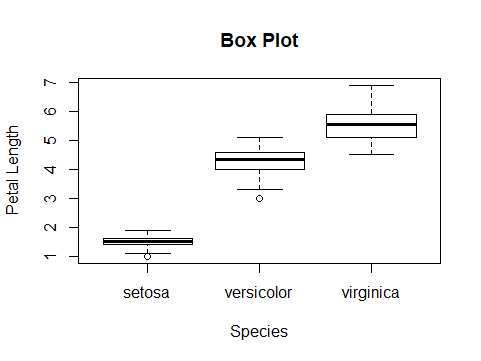
1. Generate box plots to illustrate how each of these four characteristics vary by species.



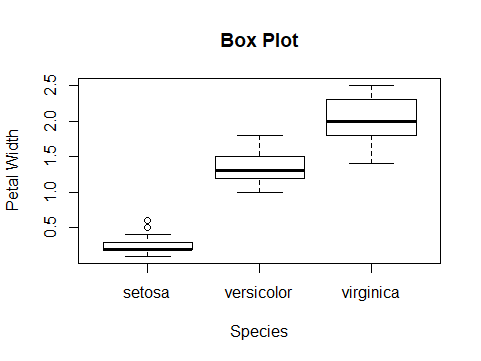
Boxplot for Sepal Length vary by species.



Boxplot for Sepal. Width vary by species.



Boxplot for Petal Length vary by species.



Boxplot for Pepal Width vary by species.

1. For each of the four characteristics, calculate a 95% confidence interval for the mean of each group.

|  |
| --- |
| > confint(aov(Sepal.Length~Species == "setosa",iris))  2.5 % 97.5 %  (Intercept) 6.147619 6.376381  Species == "setosa"TRUE -1.454113 -1.057887  > confint(aov(Sepal.Length~Species == "versicolor",iris))  2.5 % 97.5 %  (Intercept) 5.6333304 5.9606696  Species == "versicolor"TRUE -0.1444841 0.4224841  > confint(aov(Sepal.Length~Species == "virginica",iris))  2.5 % 97.5 %  (Intercept) 5.3445721 5.597428  Species == "virginica"TRUE 0.8980204 1.335980  >  > confint(aov(Sepal.Width~Species == "setosa",iris))  2.5 % 97.5 %  (Intercept) 2.8030795 2.9409205  Species == "setosa"TRUE 0.4366263 0.6753737  > confint(aov(Sepal.Width~Species == "versicolor",iris))  2.5 % 97.5 %  (Intercept) 3.124612 3.277388  Species == "versicolor"TRUE -0.563308 -0.298692  > confint(aov(Sepal.Width~Species == "virginica",iris))  2.5 % 97.5 %  (Intercept) 3.0133757 3.18462425  Species == "virginica"TRUE -0.2733056 0.02330556  >  > confint(aov(Petal.Length~Species == "setosa",iris))  2.5 % 97.5 %  (Intercept) 4.771115 5.040885  Species == "setosa"TRUE -3.677628 -3.210372  > confint(aov(Petal.Length~Species == "versicolor",iris))  2.5 % 97.5 %  (Intercept) 3.1641767 3.849823  Species == "versicolor"TRUE 0.1592126 1.346787  > confint(aov(Petal.Length~Species == "virginica",iris))  2.5 % 97.5 %  (Intercept) 2.618462 3.103538  Species == "virginica"TRUE 2.270912 3.111088  >  > confint(aov(Petal.Width~Species == "setosa",iris))  2.5 % 97.5 %  (Intercept) 1.606310 1.745690  Species == "setosa"TRUE -1.550706 -1.309294  > confint(aov(Petal.Width~Species == "versicolor",iris))  2.5 % 97.5 %  (Intercept) 0.98591856 1.2860814  Species == "versicolor"TRUE -0.06994868 0.4499487  > confint(aov(Petal.Width~Species == "virginica",iris))  2.5 % 97.5 %  (Intercept) 0.6894677 0.8825323  Species == "virginica"TRUE 1.0728012 1.4071988 |
|  |
| |  | | --- | | > | |

1. Test the hypothesis that each characteristic is different for different species using ANOVA.

|  |
| --- |
| summary(aov(Sepal.Width~Species,iris))  Df Sum Sq Mean Sq F value Pr(>F)  Species 2 11.35 5.672 49.16 <2e-16 \*\*\*  Residuals 147 16.96 0.115  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > summary(aov(Petal.Length~Species,iris))  Df Sum Sq Mean Sq F value Pr(>F)  Species 2 437.1 218.55 1180 <2e-16 \*\*\*  Residuals 147 27.2 0.19  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > summary(aov(Petal.Width~Species,iris))  Df Sum Sq Mean Sq F value Pr(>F)  Species 2 80.41 40.21 960 <2e-16 \*\*\*  Residuals 147 6.16 0.04  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | | The output includes the columns F value and Pr(>F) corresponding to the p-value of the test. | |

1. Choose one of the four characteristics for which post-hoc testing is warranted, and perform post-hoc testing using one of the multiple comparisons procedures we went over in class.
2. pairwise.t.test(iris$Petal.Width,iris$Species,p.adjust.method = "BH")
3. Pairwise comparisons using t tests with pooled SD
4. data: iris$Petal.Width and iris$Species
5. setosa versicolor
6. versicolor <2e-16 -
7. virginica <2e-16 <2e-16
8. P value adjustment method: BH
9. The result is a table of p-values for the pairwise comparisons. Here, the p-values have been adjusted by the Benjamini-Hochberg method.
10. Explain, in your own words, how multiple testing can cause an increased type I error.

In statistical hypothesis testing, a **type I error** is the rejection of a true null hypothesis (also known as a "false positive" finding). In Statistics, **multiple testing** refers to the potential increase in **Type I error** that occurs when statistical **tests** are used repeatedly, for example while doing **multiple comparisons** to **test** null hypotheses stating that the averages of several disjoint populations are equal to each other (homogeneous).