# Chapter 6 Homework

Isaac Attuah

11/15/2020

---

## Import Iris Data

Attuah starts with 'A' hence this document will be an analysis of `Sepal.Width`.

```
mydata <- iris
#Extract sepal widths into variable
sepal_width = mydata$Sepal.Width
```

---

## Check Iris Data

The `head()` function displays 6 data columns by default.

```
head(mydata)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

### Check Iris Data (with parameter)

We can define a value `n` in `head(mydata, n)` where n is the number of columns to be displayed.

```
head(mydata,10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
```

```
## 6          5.4         3.9          1.7         0.4  setosa
## 7          4.6         3.4          1.4         0.3  setosa
## 8          5.0         3.4          1.5         0.2  setosa
## 9          4.4         2.9          1.4         0.2  setosa
## 10         4.9         3.1          1.5         0.1  setosa
```

---

# Descriptive Statistics

## Measures of Central Tendency

## Mean

The mean is the average of all the iris data sepal widths.

```
#Compute the Mean
mean(sepal_width)
```

```
## [1] 3.057333
```

## Median

The median is the middle of all the data when sorted by increasing sepal widths.

```
#Compute the Median
median(sepal_width)
```

```
## [1] 3
```

Since the median, 3, is less than the mean,3.0573333, the data will be skewed to the right.

## Mode

The mode is the most frequently occurring value in the data set.

```
#modeest library must be imported to use mfv() for mode
require(modeest)
```

```
## Loading required package: modeest
```

```
## Warning: package 'modeest' was built under R version 4.0.3
```

```
#Compute the Mode
mfv(sepal_width)
```

```
## [1] 3
```

---

## Measure of Variablity

Range

The range provides the maximum and minimum values of the sepal widths.

```
#Compute the Range
range(sepal_width)
```

```
## [1] 2.0 4.4
```

Interquartile Range

We will compute the various percentiles (relates to quantiles) in 5% intervals. The IQR is a measure of statistical dispersion, between the upper and lower quantiles (75% and 25% percentiles).

```
# Compute the quantiles/percentiles
quantile(sepal_width, seq(0, 1, 0.05))
```

```
##    0%    5%   10%   15%   20%   25%   30%   35%   40%   45%   50%   55%   60%
## 2.000 2.345 2.500 2.600 2.700 2.800 2.800 2.900 3.000 3.000 3.000 3.000 3.100
##   65%   70%   75%   80%   85%   90%   95%  100%
## 3.200 3.200 3.300 3.400 3.500 3.610 3.800 4.400
```

```
#Compute the Interquartile Range
IQR(sepal_width)
```

```
## [1] 0.5
```

### Variance

The variance helps to measure how far sepal widths are spread out from the mean, 3.0573333.

```
#Compute the Variance
var(sepal_width)
```

```
## [1] 0.1899794
```

Sepal widths are spread out by 0.1899794 from the mean, 3.0573333.

### Standard Deviation

The Standard Deviation measures the amount of variation or dispersion of sepal widths. It indicates the extent of deviation for the sepal width group as a whole.

```
#Compute the Standard Deviation
sd(sepal_width)
```

```
## [1] 0.4358663
```

The average deviation from the mean value is 0.4358663. Since the standard deviation is low, the data points tend to be very close to the mean.

---

### Summary Data

The `summary()` function provides a general summary of the entire data set.

```
#Compute the Sepal Width Summary
summary(mydata)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

---

# Plots

### BoxPlot

The median, upper and lower quantile are used to construct a box plot. The length of the box is equal to the IQR. The left and right whiskers represent the first and fourth quarters of the data, while the two middle quarters of the data are represented, respectively, by the two sections of the box, one to the left and one to the right of the median line.

```
message("Upper Quantile: ", quantile(sepal_width, 0.75)) #Upper Quantile (75) (Upper line)
```

```
## Upper Quantile: 3.3
```

```
message("Median: ", median(sepal_width)) #Median (Middle Line)
```

```
## Median: 3
```

```
message("Lower Quantile: ", quantile(sepal_width, 0.25))#Lower Quantile (25) (Lower line)
```

```
## Lower Quantile: 2.8
```

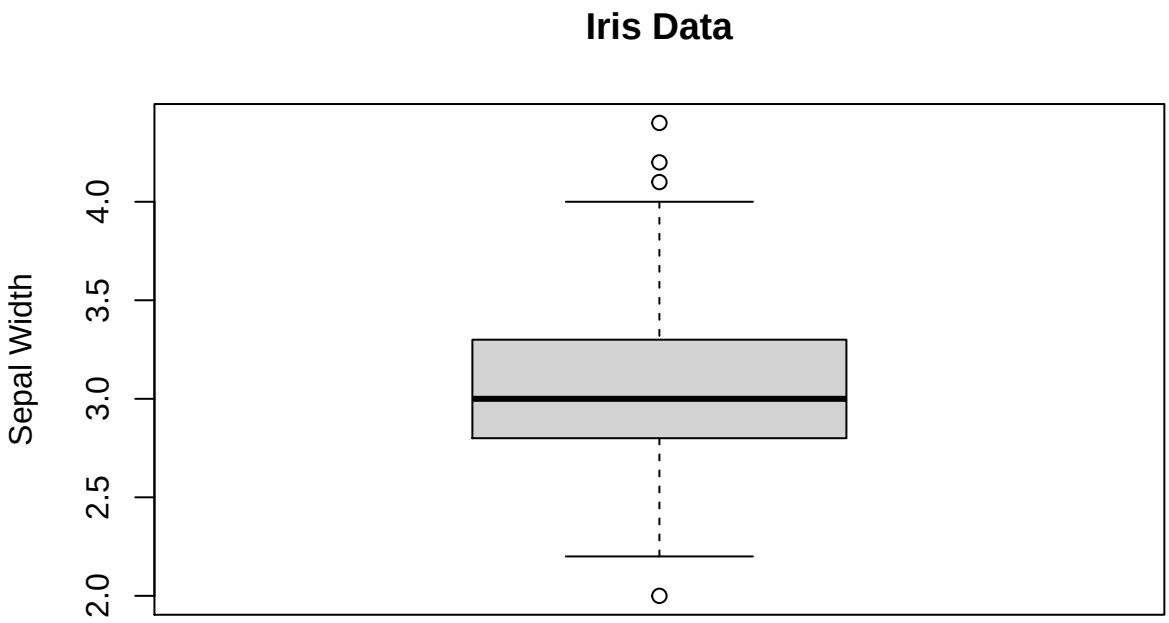
```
message("Box Length: ",IQR(sepal_width)) #Box Length
```

```
## Box Length: 0.5
```

```r
message("Whiskers: ", IQR(sepal_width) * 1.5) #Whiskers
```

```
## Whiskers: 0.75
```

```r
boxplot(sepal_width,
        data=mydata,
        main="Iris Data",
        ylab="Sepal Width")
```

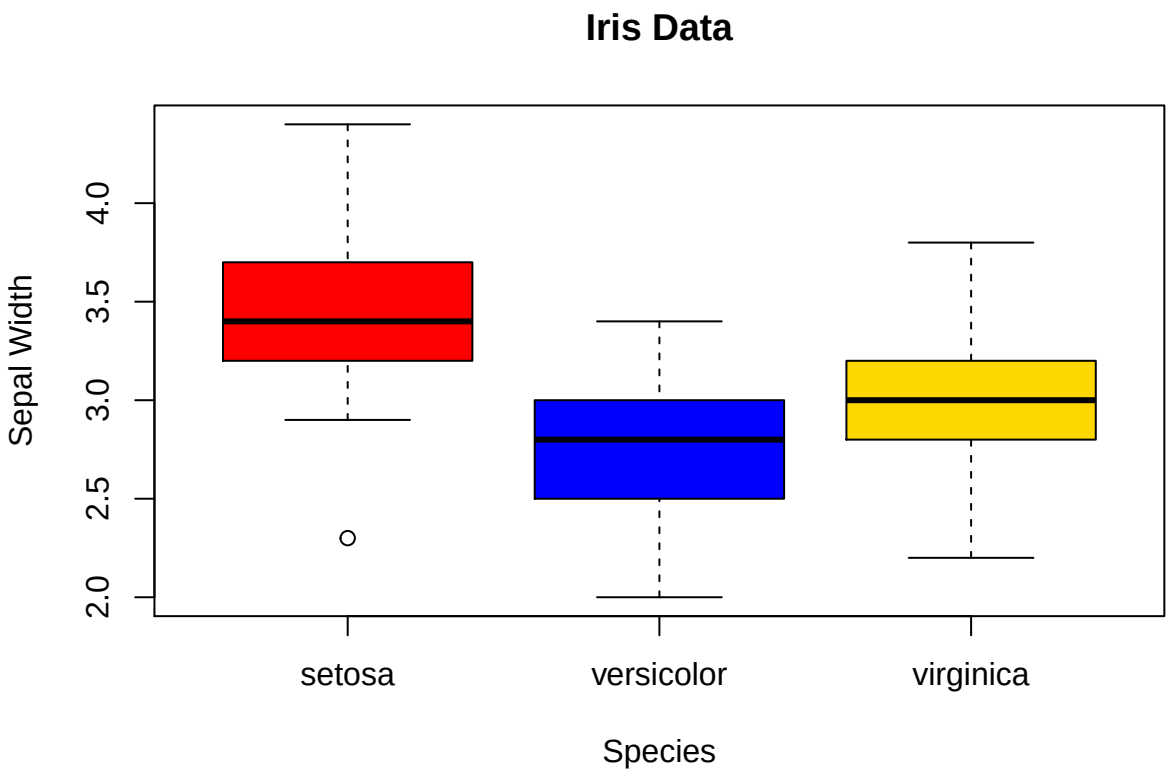



Since the upper whisker is slightly longer, the data is skewed to the right. This confirms our earlier hypothesis (mean > median). The BoxPlot also has an outlier below at 2.0 and three outliers from 4.0 upwards, this is indicated by the small circles.

We can hypothesize that longer sepal widths will be an expected outlier than shorter sepal widths.

### BoxPlot by Groups

```r
boxplot(sepal_width~Species,
        data=mydata,
        main ="Iris Data",
        col = c("red","blue","gold"),
        ylab="Sepal Width")
```




```
# '~' sign does the divisions by species
```

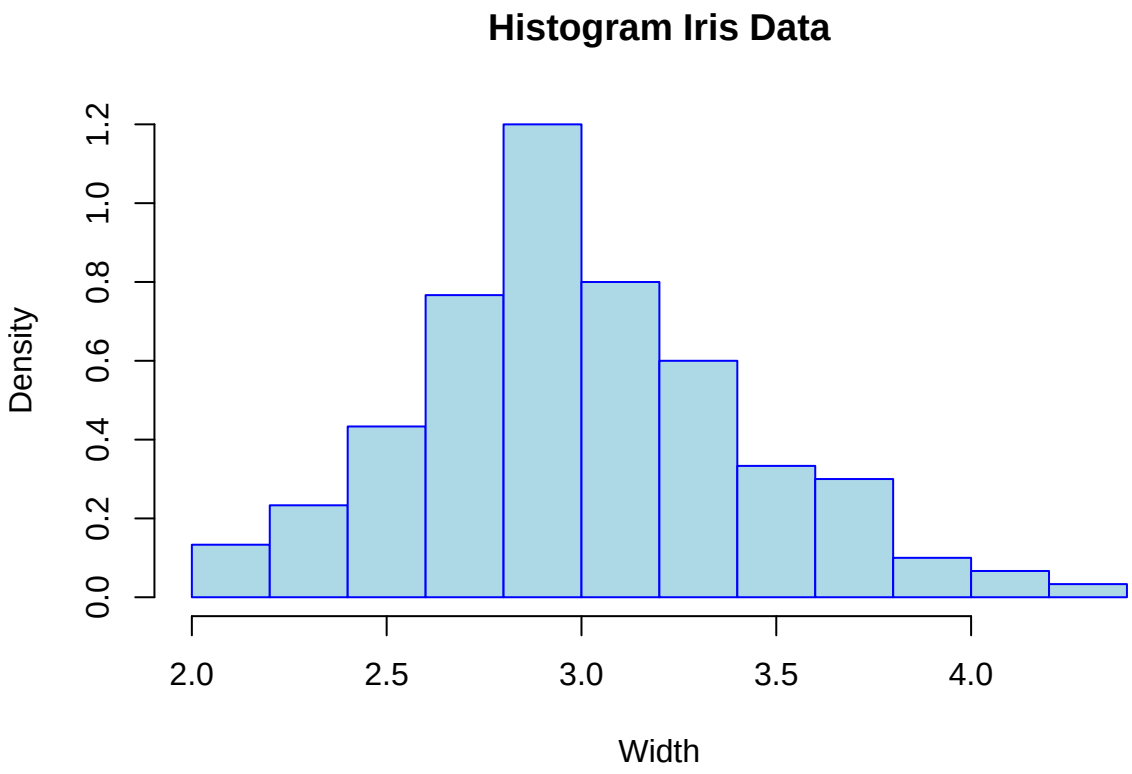
The Setosa sepal widths are skewed to the right (longer upper whisker) with an outlier between 2.0 and 2.5. The Versicolor seoal widths are skewed to the left (longer lower whisker) The Virginca sepal widths are normally skewed.

The medians of Species are in increasing size of Versicolor, Virginica and Setosa.

---

### Histogram

```
hist(sepal_width,
     main="Histogram Iris Data",
     xlab="Width",
     border="blue",
     col="lightblue",
     breaks=10,
     prob=TRUE)
```




The histogram of sepal widths is uniformly distributed. It has a few outliers after 4.0.

**Histogram by Groups**

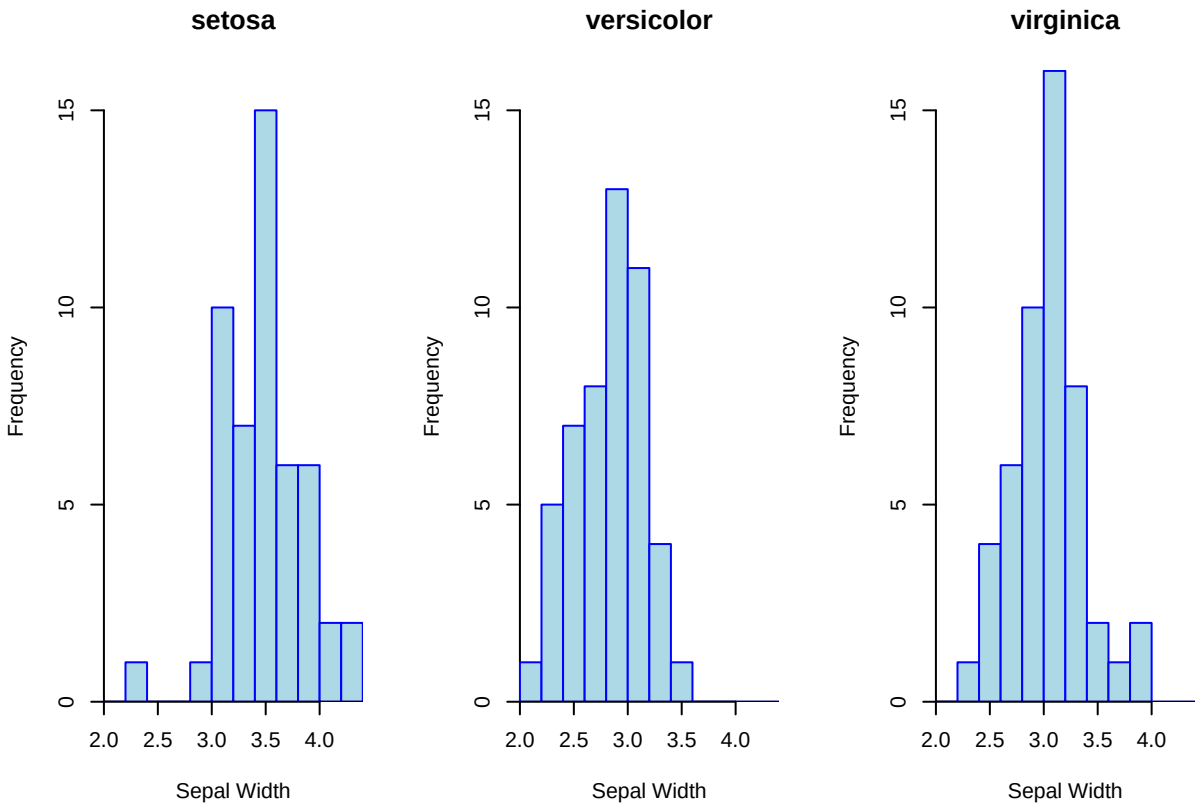
```
# install.packages("FSA")
require(FSA)
```

```
## Loading required package: FSA
```

```
## Warning: package 'FSA' was built under R version 4.0.3
```

```
## ## FSA v0.8.31. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
hist(sepal_width~Species,
     data=mydata,
     xlab="Sepal Width",
     col="lightblue",
     border="blue",
     nrow=1,
     ncol=3)
```

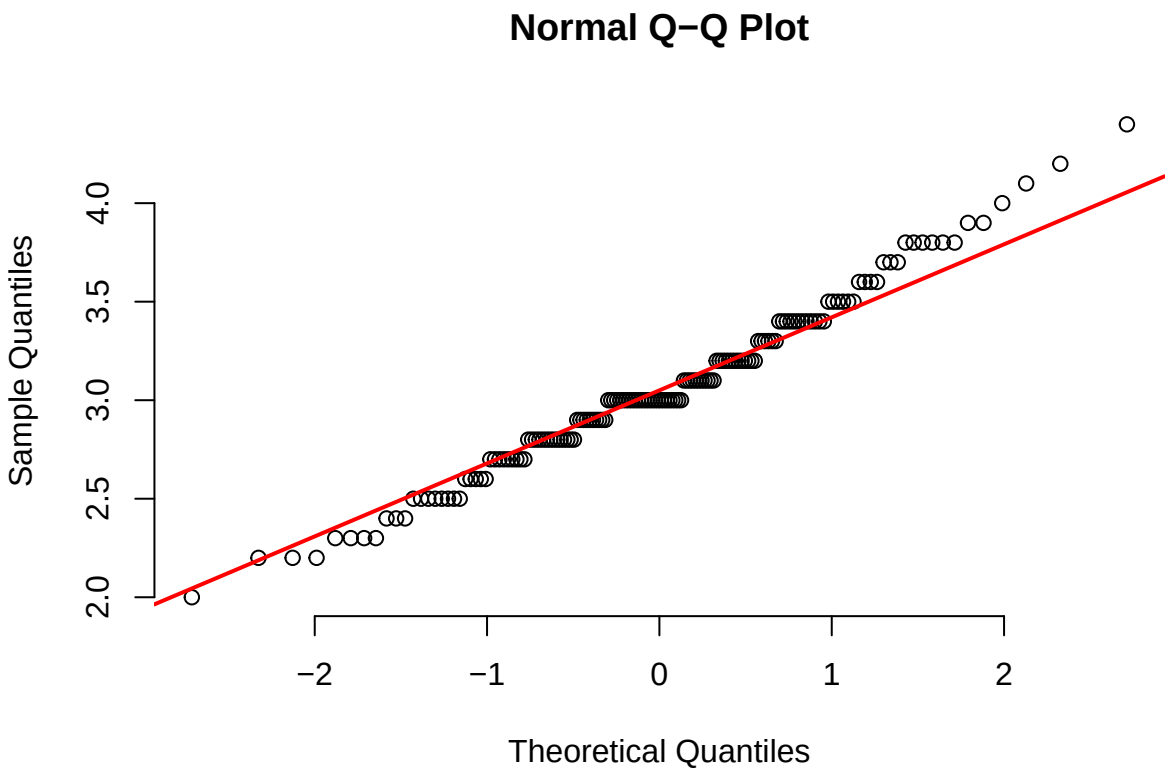



The Setosa sepal widths are skewed to the right with an outlier between 2.0 and 2.5. The Versicolor sepal widths are skewed to the left. The Virginca sepal widths are normally skewed.

---

## Q-Q Plot

```
qqnorm(sepal_width, pch = 1, frame = FALSE)
qqline(sepal_width, col = "red", lwd = 2)
```




Since a significant number of points fall along the red line, we can be confirm the data was collected from a normal distribution.