Homework - Week 4 (solution)

Michael Shyne

1. Consider the builtin dataset `iris`.

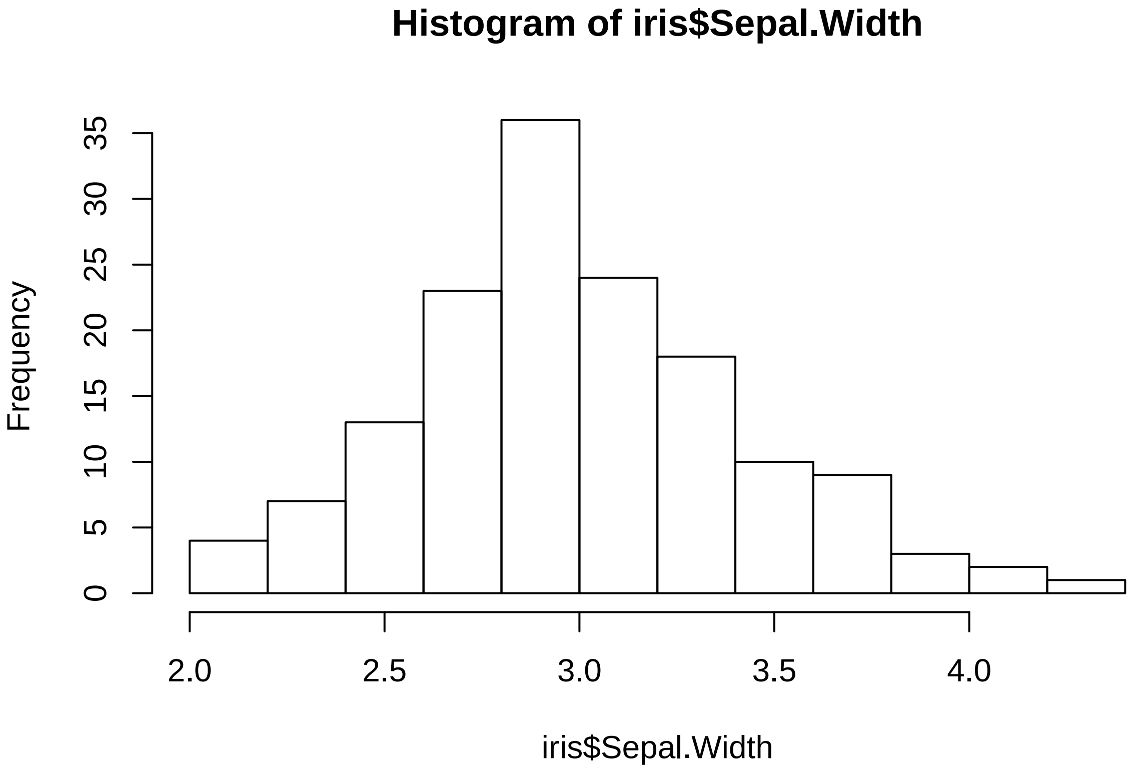
   a. What is the structure of the `iris` data frame?

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

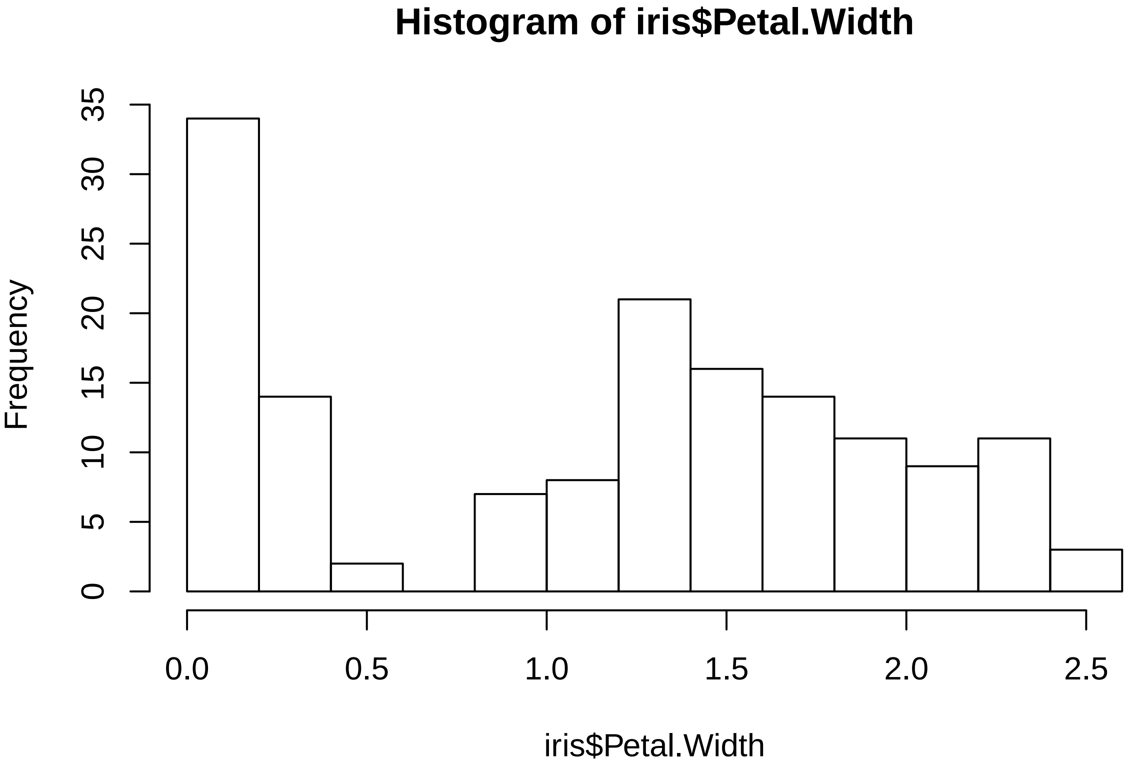
   b. Create a histogram of the `Sepal.Width` variable.

```
hist(iris$Sepal.Width)
```



   c. Create a histogram of the `Petal.Width` variable.

```
hist(iris$Petal.Width)
```



d. For both histograms, does the data appear normally distributed? Are they skewed?

   The sepal width histogram is approximately normal, with a slight right-skew.

   The petal width histogram is not normal. The data appear right-skewed, but since the distribution is bimodal defining a skew is not really appropriate.

e. For both histograms, does it appear that the data come from more than one populations?

   The sepal width histogram does not show any evidence of coming from two populations.

   The petal width histogram is bimodal, suggesting the data come from two populations.

f. What is the mean and median of `Sepal.Width`? What is the variance and standard deviation?

   Mean:

```
mean(iris$Sepal.Width)
```

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

   Variance and standard deviation:

```
var(iris$Sepal.Width)
```

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

```
sd(iris$Sepal.Width)
```

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

g. What is the mean and median of `Petal.Width`? What is the variance and standard deviation?

   Mean:

```r
mean(iris$Petal.Width)
```

```
## [1] 1.199333
```

Variance and standard deviation:

```r
var(iris$Petal.Width)
```

```
## [1] 0.5810063
```

```r
sd(iris$Petal.Width)
```

```
## [1] 0.7622377
```

2. Consider the builtin dataset `trees`.

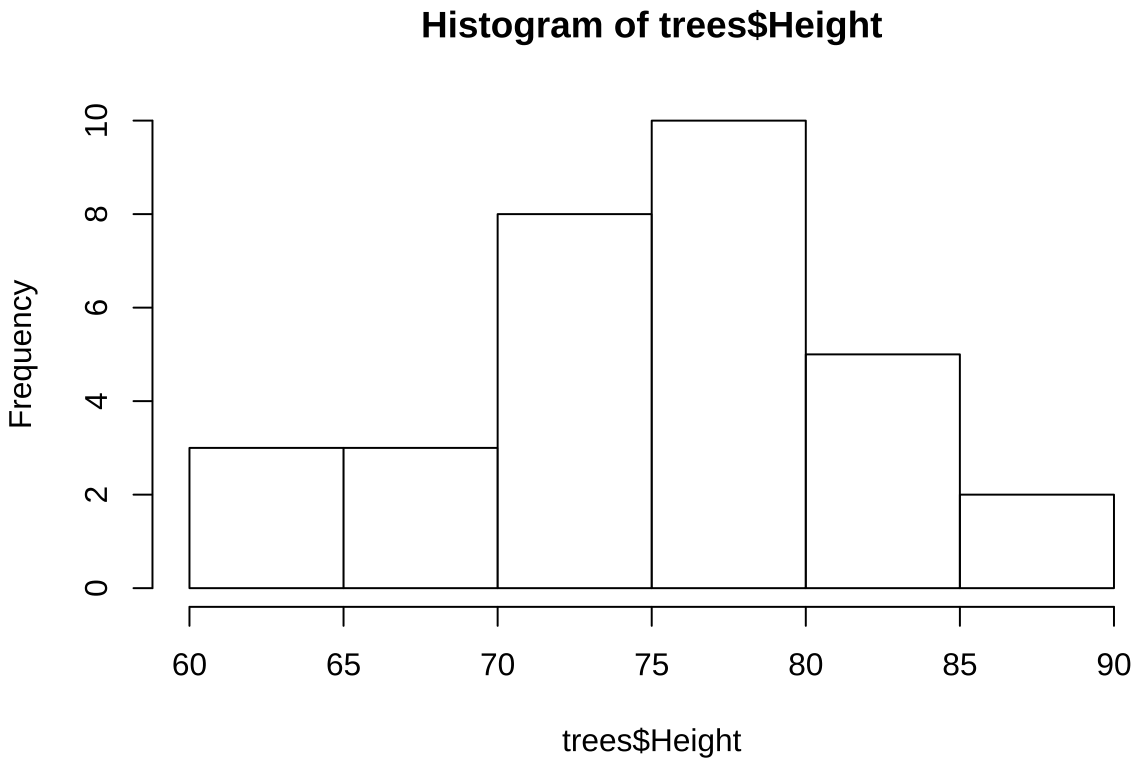
   a. What is the structure of the `trees` data frame?

```r
str(trees)
```

```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

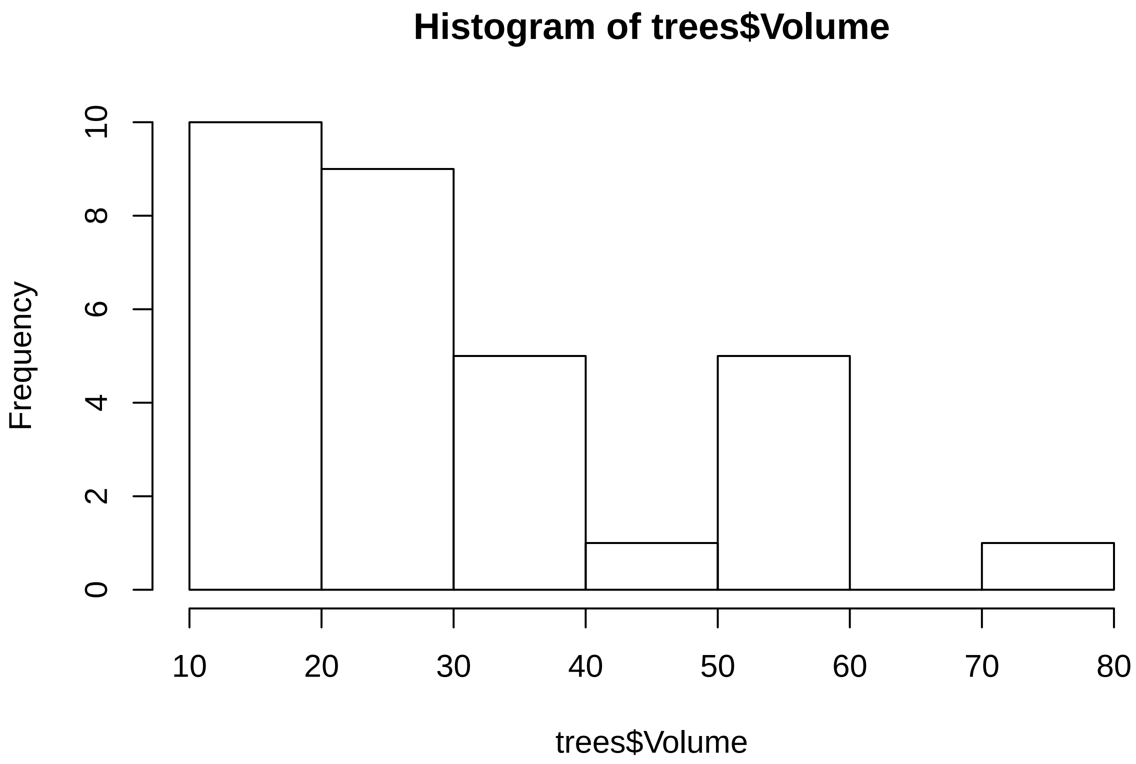
   b. Create a histogram of the `Height` variable.

```r
hist(trees$Height)
```



   c. Create a histogram of the `Volume` variable.

```r
hist(trees$Volume)
```



d. For both histograms, does the data appear normally distributed? Are they skewed?

The height histogram is approximately normal, with a slight left-skew.

The volume histogram is not normal. It is heavily skewed to the right.

e. For both histograms, does it appear that there are outliers in the data?

The height histogram does not display evidence of outliers.

The volume histogram appears to have outliers.

f. What is the mean and median of `Height`? What is the variance and standard deviation?

```
height.mean <- mean(trees$Height)
height.var <- var(trees$Height)
```

The mean tree height is 76, the variance is 40.6 and the standard deviation is the square root of variance or 6.3718129.

g. What is the mean and median of `Volume`? What is the variance and standard deviation?

```
# The round() function is helpful if you don't want
#   to display many decimal places
volume.mean <- round(mean(trees$Volume), 3)
volume.var <- round(var(trees$Volume), 3)
volume.sd <- round(sd(trees$Volume), 3)
```

The mean tree volume is (to 3 decimal places) 30.171, the variance is 270.203 and the standard deviation is 16.438.

3. Load the dataset `bears.csv` from D2L.

a. What is the structure of the `bears` data frame?

```r
# Note: this assumes the `bears.csv` file is in the same folder as this
#   markdown file (*.Rmd). If the file is located elsewhere, the read.csv()
#   must be passed a path to the file
bears <- read.csv("bears.csv")

str(bears)
```

```
## 'data.frame':    54 obs. of  9 variables:
##  $ AGE    : int  19 55 81 115 104 100 56 51 57 53 ...
##  $ MONTH  : int  7 7 9 7 8 4 7 4 9 5 ...
##  $ SEX    : int  1 1 1 1 2 2 1 1 2 2 ...
##  $ HEADLEN: num  11 16.5 15.5 17 15.5 13 15 13.5 13.5 12.5 ...
##  $ HEADWTH: num  5.5 9 8 10 6.5 7 7.5 8 7 6 ...
##  $ NECK   : num  16 28 31 31.5 22 21 26.5 27 20 18 ...
##  $ LENGTH : num  53 67.5 72 72 62 70 73.5 68.5 64 58 ...
##  $ CHEST  : num  26 45 54 49 35 41 41 49 38 31 ...
##  $ WEIGHT : int  80 344 416 348 166 220 262 360 204 144 ...
```

b. Create a frequency table for the variable `MONTH`. What is the mode, if any?

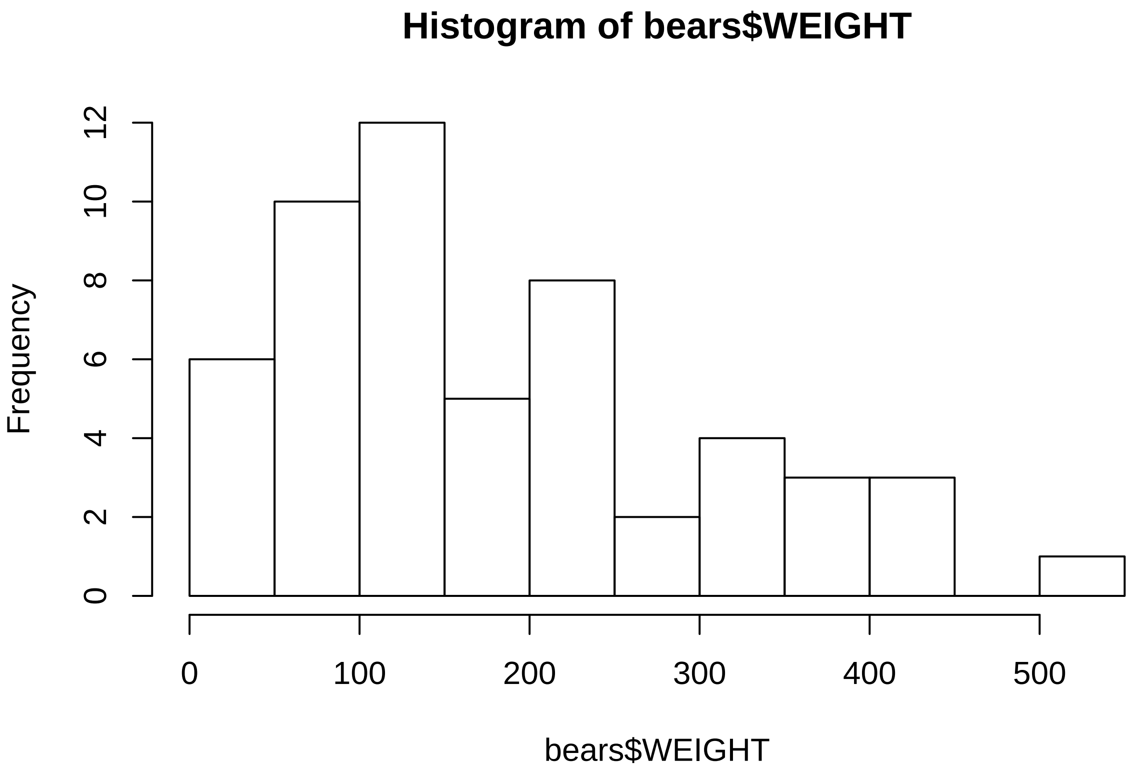
```r
table(bears$MONTH)
```

```
## 
##  4  5  6  7  8  9 10 11 
##  4  4  1  4  8 16 11  6
```

Month 9 (September) has the highest frequency at 16, and thus is the mode.

c. Create a histogram of the `WEIGHT` variable.

```r
hist(bears$WEIGHT)
```



d. Is there distribution of `WEIGHT` data normal? Is it skewed? Are there outliers?

The weight histogram shows that the data are not normal. They are strongly right-skewed and there do appear to be outliers (weight > 500).

e. Based on your answers to part (d), do you expect the mean and median to be the same (or very close)? If not, which do you expect to be greater?

   Because the data are skewed, the mean and median are likely not close to the same. Because the data are right-skewed with high value outliers, the mean will be pulled to a higher value. Thus, the mean should be greater than the median.

f. What is the mean and median of `WEIGHT`?

   Mean:

```r
mean(bears$WEIGHT)
```

```
## [1] 182.8889
```

   Median:

```r
median(bears$WEIGHT)
```

```
## [1] 150
```

g. Based on the histogram in part (c), what would you expect the mode to be, approximately?

   The tallest bar in the histogram is for the range from 100 to 150. Thus, we would expect the mode to be in this range.

h. What is the mode, if any?

```r
sort(table(bears$WEIGHT))
```

```
##
##  26  29  34  40  46  48  60  62  64  65  76  79  80  86  90  94 105 114
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 116 120 125 132 144 148 154 180 182 212 236 262 270 316 332 344 348 356
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 360 365 416 436 446 514 140 150 166 202 204 220
##   1   1   1   1   1   1   2   2   2   2   2   2
```

   Technically, there are 6 modes, {140, 150, 166, 202, 204, 220}. In a practical sense, the fact that six values occur twice, while the rest occur once, is not very useful information. The peak of the histogram is a more useful estimate than the literal modes. This is often the case with continuous variables.