Math 338 **Lab Assignment #14** Fall 2019

In this lab we are going to investigate some properties of numerical data. We will use a dataset already available in R, called the *trees* dataset. To view information about the dataset, type:

```
> ?trees
```

Question #1 What are the cases in these data, and how many are there?

There are 31 felled black cherry trees.

Question #2 Name the three numerical variables in the dataset. Include the units of measurement.

Girth (inches), Height(feet), Volume(cubic feet).

Now, let's obtain a histogram of these data. We'll use the ggplot2 library again, with the more "standard" plotting syntax using the `ggplot` command.

```
> library(ggplot2)
> tree_height_plot <- ggplot(data = trees, mapping = aes(x = Height))
> tree_height_histogram <- tree_height_plot + geom_histogram(binwidth = 5,
center = 67.5) + labs(title = "Height of 31 Black Cherry Trees", x =
"Height (feet)", y = "Number of Trees")
> print(tree_height_histogra    m)
```

You can play around with the "binwidth" and "center" arguments. The idea is to specify that the bins in the histograms are centered at the values of *center + k(binwidth)*, where k is an integer, and have width *binwidth*. In our example, `binwidth = 5, center = 67.5` means to center the bins at 57.5, 62.5, 67.5, 72.5, 77.5, etc., and give each bin a width of 5. This corresponds to the bins 55-60, 60-65, 65-70, 70-75, 75-80, etc.

Question #3 Copy the histogram and paste it below.

Please see last page of document for attached graph.

Question #4 Is the distribution of height symmetric or skewed? If it is skewed, state the direction. Do you see any potential outliers on the histogram?

The data given is fairly symmetric and there does not appear to be any outliers in this data set.

Let's use a command in the dplyr library (`summarize`) to obtain the mean and standard deviation of the heights. Note that we could also use the Base R commands `mean(trees$Height)` and `sd(trees$Height)`.

```
> library(dplyr)
> summarize(trees, mean.height = mean(Height), sd = sd(Height))
```

Question #5 What is the mean height of these trees? What is the standard deviation of the height?

The mean height is going to be 76 and the standard deviation is 6.371813.

We could use the summarize command again to get the five-number summary, but instead, we'll use the ubiquitous R command `summary` to get the five-number summary of all three variables:

```
> summary(trees)
```

Question #6 Using the output, report the five-number summary of the height of the trees.

Min: 63
Q1: 72
Median: 76
Mean: 76
Q3: 80
Max: 87

Question #7 According to our 1.5 IQR rule of thumb, does this dataset contain any outlier heights? If so, which data points are outliers? Show your work.

According to the IQR rule of thumb, we do not have any outliers in our data set. This is because neither the min or the max exceed the lower/upper fence respectively.

# IQR = 80 - 72 = 8
# Lower fence = 72 - (1.5) * 8 = 60
# Upper fence = 80 + (1.5) * 8 = 92

Now, let's obtain a boxplot of these data using the ggplot2 library. Since we already set up our plot, it's sufficient to just add the new geom object (a boxplot) and new labels.

```
> tree_height_boxplot <- tree_height_plot + geom_boxplot(aes(x = "", y =
Height, group = NA))+labs(title = "31 Black Cherry Trees", x = "", y =
"Height (feet)")

> print(tree_height_boxplot)
```

Question #8 Copy the boxplot and paste it below. Did R identify any outliers based on the 1.5 IQR rule?

Please see last page for attached for boxplot.

Before the first exam, we discussed how linear transformations ($x_{\text{new}} = a + bx$) affect the mean and standard deviation of a random variable. In the last part of the lab, we will show that it works with actual data too. We will convert the measurements to meters (1 foot = 0.3048 meters) and discount the first 10m of the measurement. This results in the variable $Height.new = 0.3048 \times Height - 10$.

Question #9 Without doing anything in R, use the transformation rules and the values from Question **#5** to compute the values of the mean and standard deviation of *height_new*. Show all steps in your work.

Linear transformation:

h_new = 0.3048 x (Height – 10)

mean(Height) = 76

standard deviation(Height) = 6.371813

mean(h_new):

= mean(0.3048 x (Height)) -10

= 0.3048 x mean(Height) – 10

= 13.1648

standard deviation(h_new):

= standard deviation(0.3048 x (Height)-10)

= 0.3048 x standard deviation(Height)

= 1.94212

There are many different ways to do the actual transformation, but one of the most useful is using the `mutate` command in the dplyr library.

```
> trees_transform <- mutate(trees, height_new = 0.3048 * Height - 10)
```

This creates a new dataset with all the original variables in *trees* as well as a new variable, *height_new*. We can then create the summary that we want:

```
> summarize(trees_transform, mean_height_new = mean(height_new),
sd_height_new = sd(height_new))
```

Question #10 Copy your output and paste it below. Did you get the same mean and standard deviation for your transformed height as R computed?

```
  mean_height_new sd_height_new
1        13.1648      1.942129
```

Yes, this is the same as I computed above in question 9.