Homework 3

Kyle Walter

4/30/2021

Qustion 2

Prompt - Use the chick weight dataset. Use the summary command, you find that the set contains 4 varaiables. Also run the dim fuction and report what the first number signifies, as the 2nd number describes the number of variables?

```
data("ChickWeight")
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

```
dim(ChickWeight)
```

```
## [1] 578   4
```

The 578 signifies the number of times that the vaious chicks have had their weight observered.

Question 3

Prompt - R allows data sets with multiple variables to have the variable accessed by using the $ sign. Run the following commands and briefly explain each piece of output.

```
summary(ChickWeight$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    35.0    63.0   103.0   121.8   163.8   373.0
```

The summary of the weights shows us the minimum weight in the data set. Where the first, median, and third quartile are in the data and maximum observed weight. This is the result in a scenario where the value is numeric.

```
head(ChickWeight$weight)
```

```
## [1] 42 51 59 64 76 93
```

The head function shows the first 6 records of the dataset based on their order. In this case because we've called the weight it is the first 6 weights observed for the chickens.

```
mean(ChickWeight$weight)

## [1] 121.8183
```

The mean shows us the average weight in the chickweights data set, or more simply the sum of all the weights divided by the 578 (the number of observations.)

```
myChkWts <- ChickWeight$weight
```

This commands save the weights from the dataset in their own variable. From this point forward in the code I can just call the myChkWts to get the weights of the chickens rather than using the $ sign.

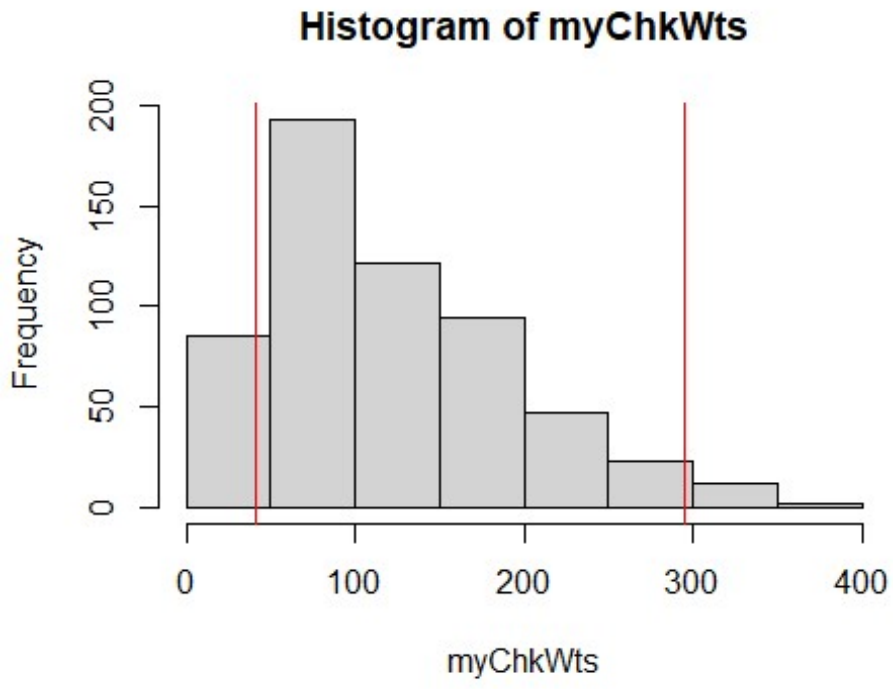
```
quantile(myChkWts, .5)

## 50%
## 103
```

Quantile when set to .5 shows us where the middle of the numbers are. It is equal to the median of the dataset and easily verified as we saw it earlier in the summary labeled as such.

Question 4

```
hist(myChkWts)
abline(v = c(quantile(myChkWts, .025),quantile(myChkWts, .975)), col="red")
```

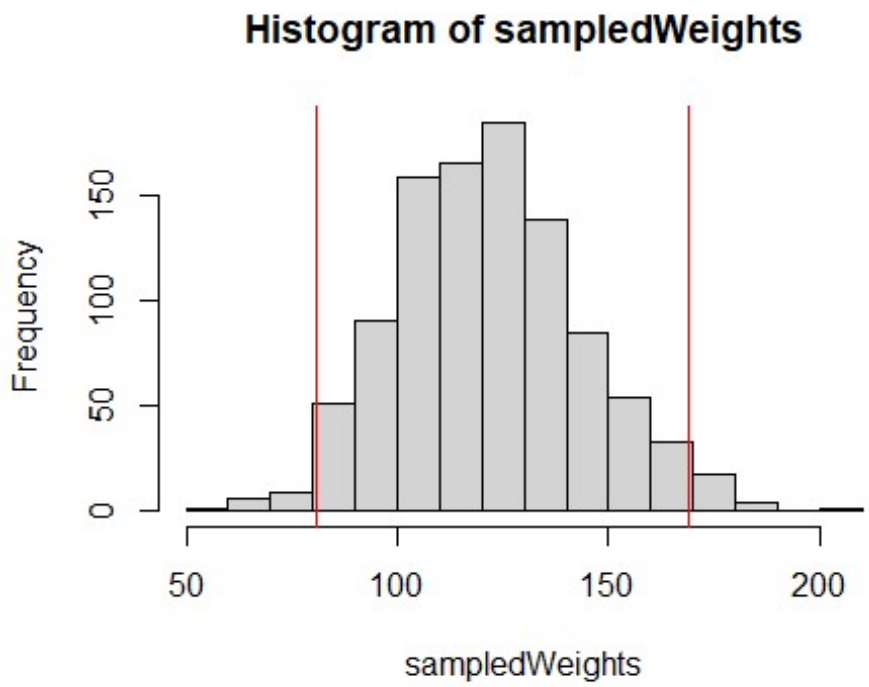


So looking at the data as histogram we can see the weight scewed to the right. 95% of our data is between the red lines on the graph, while the 5% of outlier are on the far sides of the right line, with larger outliers to right (hence the skew). The median, 103 is actually on the left side of the graph but the larger outliers are pulling a large mass of of the observations to the right.

Question 5

Prompt - Write R code that construction sampling distributions of at least 1,000 means and sample size of 11. Store the means in a new variable and create a histogram that will display the data with the 2.5% and 97.5% Quantitles

```
set.seed(1347)
sampledWeights <- replicate(1000, mean(sample(myChkWts, size = 11, replace =
T)))
hist(sampledWeights)
abline(v = c(quantile(sampledWeights, .025),quantile(sampledWeights, .975)),
col="red")
```



Question 6

Describe the different between the histogram in 4 vs 5

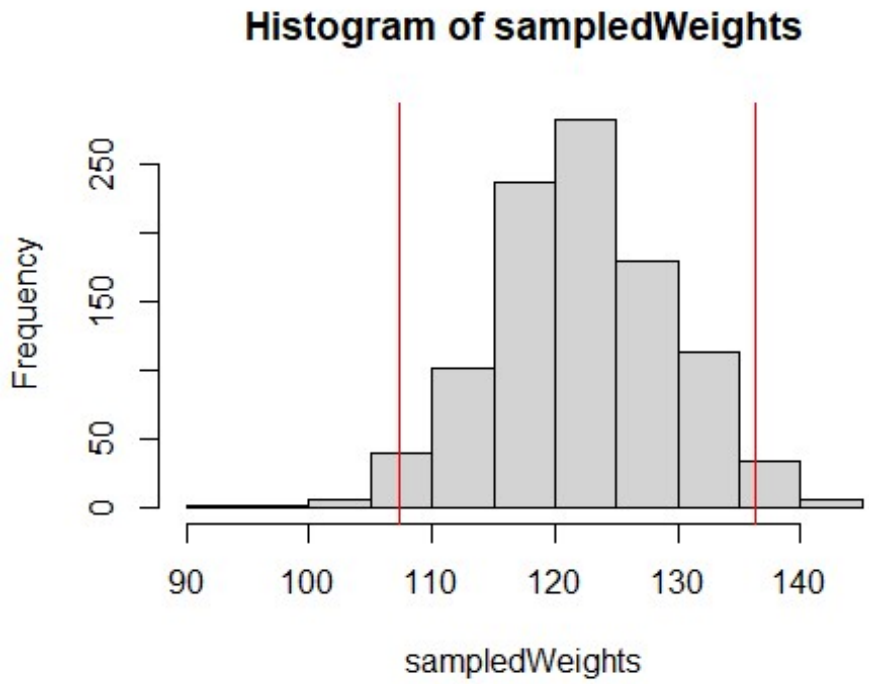
The histogram of the raw data appears a bit more random in it shape and pulled substantially by outlier. Where the shape of the histogram generated in 5 is more resembling a bell shaped cuver of the normal distribution. It still has a slightly longer tail to the right than the left, but the data because it is the means of the samples is more controlled by the central limit theory and we can it more closely align over the mean seen in the sumamry of 121.8 although it won't be there exactly.

In addtion the tails are more distinctly visible now that the data is more centered. The 2.5% of rare values to right and left actually look rare in this view where in the raw data expecially those on the left looked like the line was more or less arbitrarily drawn.

Question 7

Recreate the code from 5 expect increase the size of sample to 1000. Compare the results vs 5 and describe why the 100 for this particular data set actually works.

```
set.seed(1347)
sampledWeights <- replicate(1000, mean(sample(myChkWts, size = 100, replace =
T)))
hist(sampledWeights)
abline(v = c(quantile(sampledWeights, .025),quantile(sampledWeights, .975)),
col="red")
```



So repeating 5 with a larger sample size we can more easily see that the data has converged on the mean around 121 as it is where the highest point in the data is. This is happening because as we run the sampling with larger numbers, the sample becomes more representative of the population, and the outlier effect generally starts to cancel out and the normal curve and the mean becomes more distinct.For example based on the graph in 5 the center of the data could be anywhere between 110 and 140, although a small spike is seen around the 120 mark. With the larger samples we can more clearly see the mean around 120 mark as the breaks become obviously smaller.