Victoria Haley
Homework 3

Exercise 2
      Names of variables: weight, Time, Chick, Diet
      The first number of dim(ChickWeight) is 578. This number is the total number of
      observations in the entire ChickWeight dataset.

Exercise 3
      summary(ChickWeight$weight)

        a) This command returns basic information about the weight column of the
           ChickWeight dataset.
           > Min. 1st Qu. Median Mean 3rd Qu.   Max.
             35.0   63.0   103.0   121.8 163.8   373.0
           The information returned from this command tells us that the weights of
           the chicks in the dataset range from 35 to 373, with an average weight of
           121.8, and with most chicks weighing between 63 and 163.8.

      head(ChickWeight$weight)

        a) This command returns the first few observations in the weight column of
           the ChickWeight dataset.
           > 42 51 59 64 76 93
           Without context, the information returned from this command does not
           tell us much about the overall data. However, knowing that the dataset
           records the weights of chicks across time, we can see that these are the
           first 6 weights recorded for the first chick. This chick weighed 42 at the
           first weigh-in and 93 at the 6th weigh.-in

      mean(ChickWeight$weight)

        a) This command returns the mean of the weights in the ChickWeight
           dataset.
           > 121.8183
           The output tells us that the average weight of all the weights in the
           weight column is 121.8. This matches with the mean that was returned
           from the summary() command above.

      myChkWts <- ChickWeight$weight

        a) This command stores the data in the weight column of the ChickWeight
           dataset into a new storage area called myChkWts.
           This command does not return any information into the console,
           however it does store myChkWts as a value in the global environment
           section of R Studio.
           With this new stored value, we can easily call the weights of the chicks in
           ChickWeight without having to type out the entire dataset name
           anymore. We can also make changes to the data if we want without
           altering the original dataset.

      quantile(myChkWts,0.50)

a) This command returns the 50% quantile of the observations in myChkWts.

> 50%
103

The information that this command outputs tells us that the median observation of the weights in myChkWts is 103. This matches the median value that was returned from the summary() command from earlier.

Exercise 4

R code:

```
> hist(myChkWts)
> abline(v=quantile(myChkWts,0.025))
> abline(v=quantile(myChkWts,0.975))
```

Output:

Histogram of myChkWts
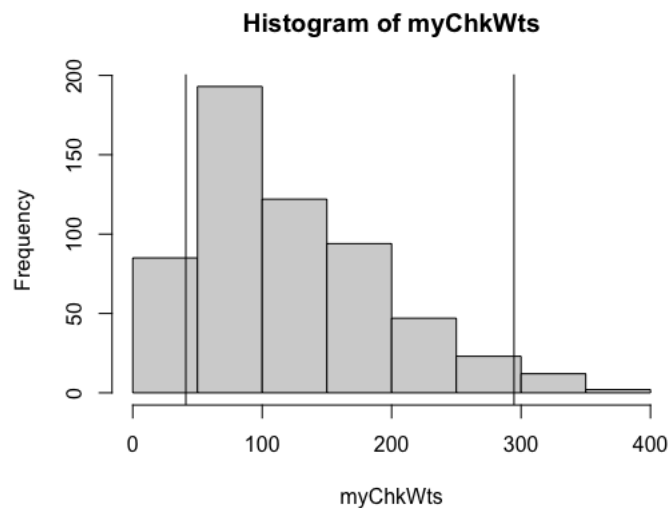
The histogram of myChkWts shows the distributions of the weights of the chicks in the dataset, with a skew to the left side. The median and the mean, 103 and 121.8 respectively, of the variable are not the tallest bin in this histogram. The shape of the histogram tells us that most of the observations fall between the 50-100 range. There are a good number of observations in the 2.5% quantile, however those above the 97.5% quantile are very rare. The 2.5% quantile signifies that there are few weights that fall in the uncommon but not exactly rare side of the data, and the 97.5% quantile signifies that the weights around 300 and above are really rare and do not occur very often.
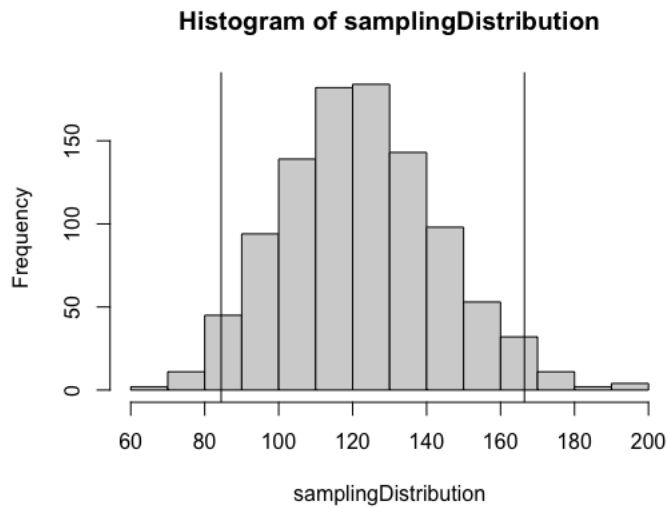
Exercise 5

R code:

```
> samplingDistribution <-
replicate(1000,mean(sample(myChkWts,size=11,replace=TRUE)))
> hist(samplingDistribution)
> abline(v=quantile(samplingDistribution,0.025))
```

```
> abline(v=quantile(samplingDistribution,0.975))
```
Output:

Histogram of samplingDistribution

Exercise 6

The difference between the distribution of raw data and the distribution of sampling means is that the means of the samples will basically eliminate the extreme/rare observations and will bring the distributions closer to the population mean.
The 2.5% and 97.5% quantiles are so different from the raw data distribution and sampling distribution of means is because the raw data is really spread out, from 35 to 373, while the quantiles in the sampling distribution breaks the curve up so that the common events are separated from the rare ones.

Exercise 7

R code:
```
> samplingDistribution <-
replicate(1000,mean(sample(myChkWts,size=100,replace=TRUE)))
> hist(samplingDistribution)
> abline(v=quantile(samplingDistribution,0.025))
> abline(v=quantile(samplingDistribution,0.975))
```
Output:

Histogram of samplingDistribution
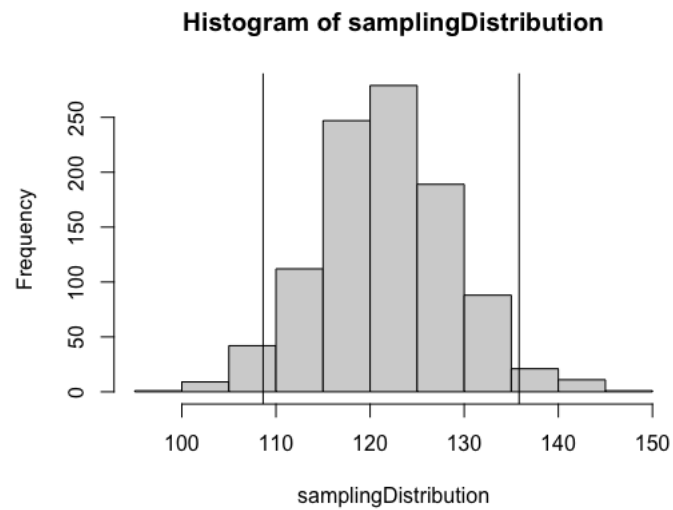
The quantiles are different in the distribution of sample means of a sample that has a sample size of 100 because the increased sample size allows for the mean of the sampling distribution to match the mean of the underlying population. As we can see here, the middle of this histogram is right about 120, which is really close to the mean from earlier of 121.8, while the middle of the histogram from exercise 5 is about 200. As such, the 2.5 and 97.5 quantiles in this histogram are able to be more precise when dividing the histogram into a central region of common events and two tails of rare events.