Homework 3

Daniel Caley

7/10/2021

Contents

Questions for this Week

The homework for week three is exercises 2 through 7 on pages 50 and 51.

Question 2

For the remaining exercises in this set, we will use one of R's built-in data sets, called the "ChickWeight" data set. According to the documentation for R, the `ChickWeight` data set contains information on the weight of chicks in grams up to 21 days after hatching.

Summary Chicken Weight

Use the `summary(ChickWeight)` command to reveal basic information about the `ChickWeight` data set.

```
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
```

Name the Four Variables

You will find that `ChickWeight` contains four different variables. Name the four variables. - Weight - Time - Chicks - Diet

Dimensions Chiken Weight

Use the `dim(ChickWeight)` command to show the dimensions of the `ChickWeight` data set.

```
dim(ChickWeight)
```

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

Observing the Data

The second number in the output, 4, is the number of columns in the data set, in other words the number of variables. What is the first number? Report it and describe briefly what you think it signifies.

- The first number, which is 578, is the total observation or rows for each of the 4 columns in the `ChickWeight` data set.
- Meaning that the the `ChickWeight` data set is a total of 578*4 = 2,312 data points

Question 3

Report the Output

When a data set contains more than one variable, R offers another subsetting operator, `$`, to access each variable individually. For the exercises below, we are interested only in the contents of one of the variables in the data set, called weight. We can access the weight variable by itself, using the `$`, with this expression: `ChickWeight$weight`. Run the following commands and report the Output

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

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

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

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

```
mean(ChickWeight$weight)
```

```
## [1] 121.8183
```

```
myChkWts <- ChickWeight$weight
quantile(myChkWts,0.50)
```

```
## 50%
## 103
```

What Does the Command Do & Briefly Explain the output?

- The summary command describes a vector or dataframe by providing the min, median, max along with the 1st quartile and 3rd quartile. In this instance the the summary command is being used on a single vector called `weight`. The data is right skewed with a median of 103 and a mean of 121, which the median is less than the mean.
- The head command returns the first 5 records of a vector or dataframe. In this case the head command is returning the 5 numbers of the vector `weight.` The output here is the weight of chicks, baby chickens.
- The mean command finds the average of a vector. This command cannot be used on an entire dataframe. The mean weight of a chick is 121.8.
- The assignment operator or `<-` is used to store a dataframe, vector, variable or a series of commands in an R console. In this case the `weigth` vector is being stored in `myChkWts`.
- The quantile command with the specified probability gives the corresponding number associated to that probability. In this case the `myChkWts` vector is being passed along with a probability of 0.50 to the `quantile` command to return 103, which is also our Median.

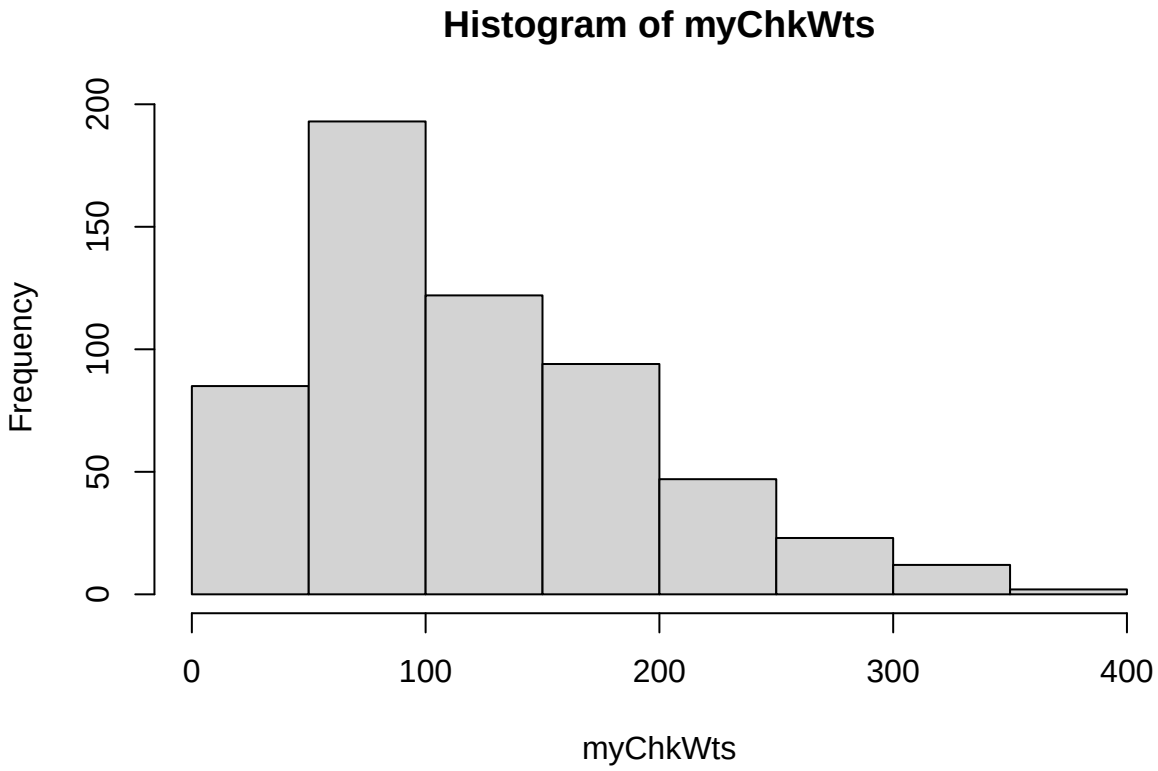
Question 4

In the second to last command of the previous exercise, you created a copy of the weight data from the `ChickWeight` data set and put it in a new vector called `myChkWts`. You can continue to use this `myChkWts` variable for the rest of the exercises below.

Chick Weight Histogram

Create a histogram for that variable.

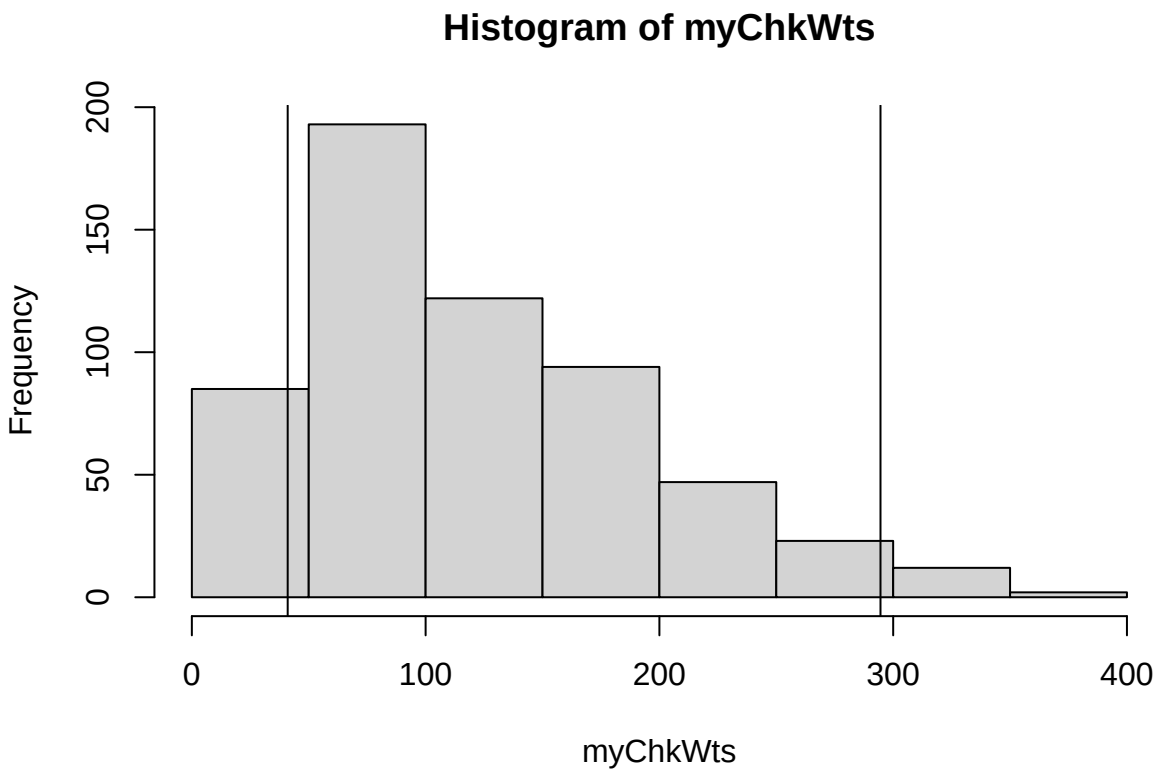
```
hist(myChkWts)
```



Chick Weight Quantiles

Then write code that will display the 2.5% and 97.5% quantiles of the distribution for that variable.

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



Interpretation of the variable

Write an interpretation of the variable, including descriptions of the mean, median, shape of the distribution, and the 2.5% and 97.5% quantiles. Make sure to clearly describe what the 2.5% and 97.5% quantiles signify.

- Mean: The average chick weight represented by the histogram or sample of chick weight is 121 and around the peak of the histogram.

- Median: The median chick weight is 103 from this histogram. By looking at the highest bar in the histogram, at 100, shows that some where around here is the median.
- Shape of the distribution: The histogram is right skewed and not bell shape or normal.
- Quantile: The 2.5% and 97.5% quantile shows how much of the chick weight data falls inside of that area in the histogram.

Question 5

Chicken Weight Sampling Distribution

Write R code that will construct a sampling distribution of means from the weight data (as noted above, if you did exercise 3 you can use `myChkWts` instead of `ChickWeight$weight` to save yourself some typing).

- Make sure that the sampling distribution contains at least 1,000 means.
- Store the sampling distribution in a new variable that you can keep using.
- Use a sample size of n = 11 (sampling with replacement).

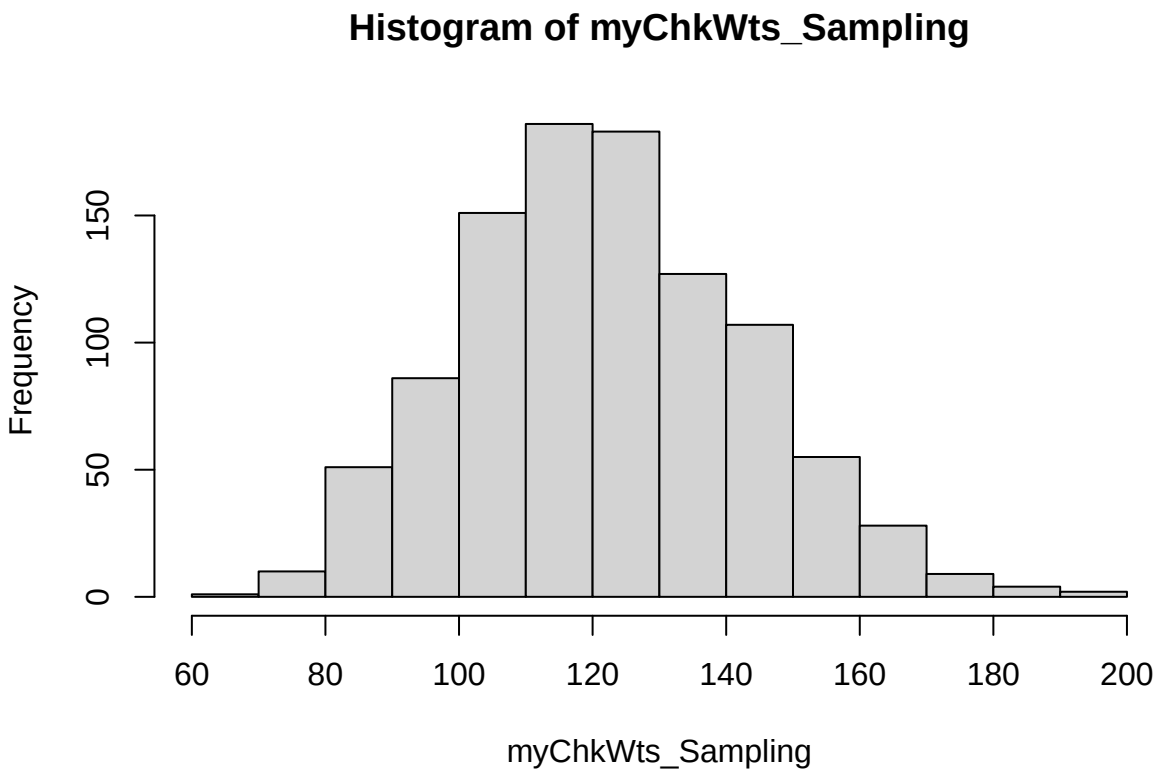
```
myChkWts_Sampling <- replicate(1000,mean(sample(myChkWts, 11, replace = TRUE)), simplify = TRUE)
summary(myChkWts_Sampling)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    69.0   106.7   120.6   122.0   136.4   191.7
```

Chicken Weight Sample Histogram

Show a histogram of this distribution of sample means.

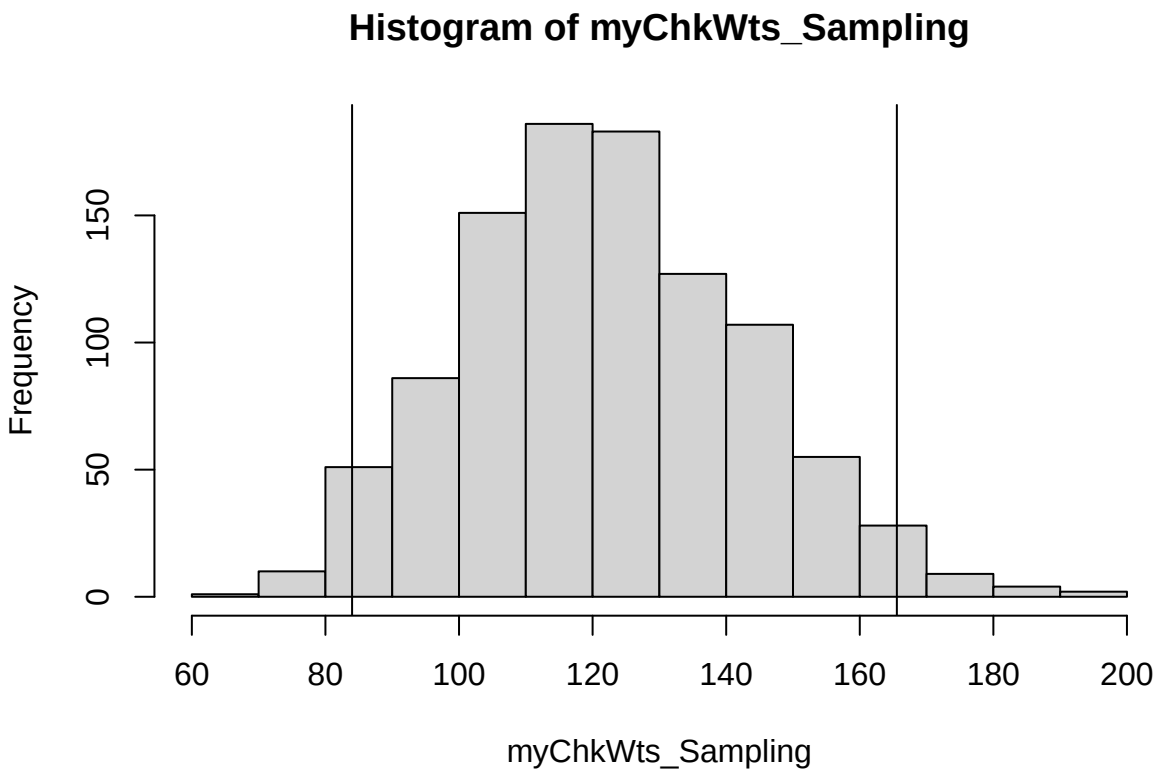
```
hist(myChkWts_Sampling)
```



Chiken Weight Sample Histogram Quartile

Then, write and run R commands that will display the 2.5% and 97.5% quantiles of the sampling distribution on the histogram with a vertical line.

```
hist(myChkWts_Sampling)
abline(v = quantile(myChkWts_Sampling, 0.025))
abline(v = quantile(myChkWts_Sampling, 0.975))
```



Question 6

Raw vs. Sampling Means

If you did Exercise 4, you calculated some quantiles for a distribution of raw data. If you did Exercise 5, you calculated some quantiles for a sampling distribution of means. Briefly describe, from a conceptual perspective and in your own words, what the difference is between a distribution of raw data and a distribution of sampling means.

For the distribution of raw data, a mean signifies the average from the original dataset. When sampling, the process pulls at random 11 values and finds the mean of that random sampled and performs this 1000 times. The reason for doing this, is so that a histogram will converge to be normally distributed and therefore we can leverage central limit theorem.

Raw vs. Sampling Quantiles

Finally, comment on why the 2.5% and 97.5% quantiles are so different between the raw data distribution and the sampling distribution of means.

The two quantiles have changed between the raw data and the distribution of sampling means as the process of sampling will normalize the tails of a distribution.

Question 7

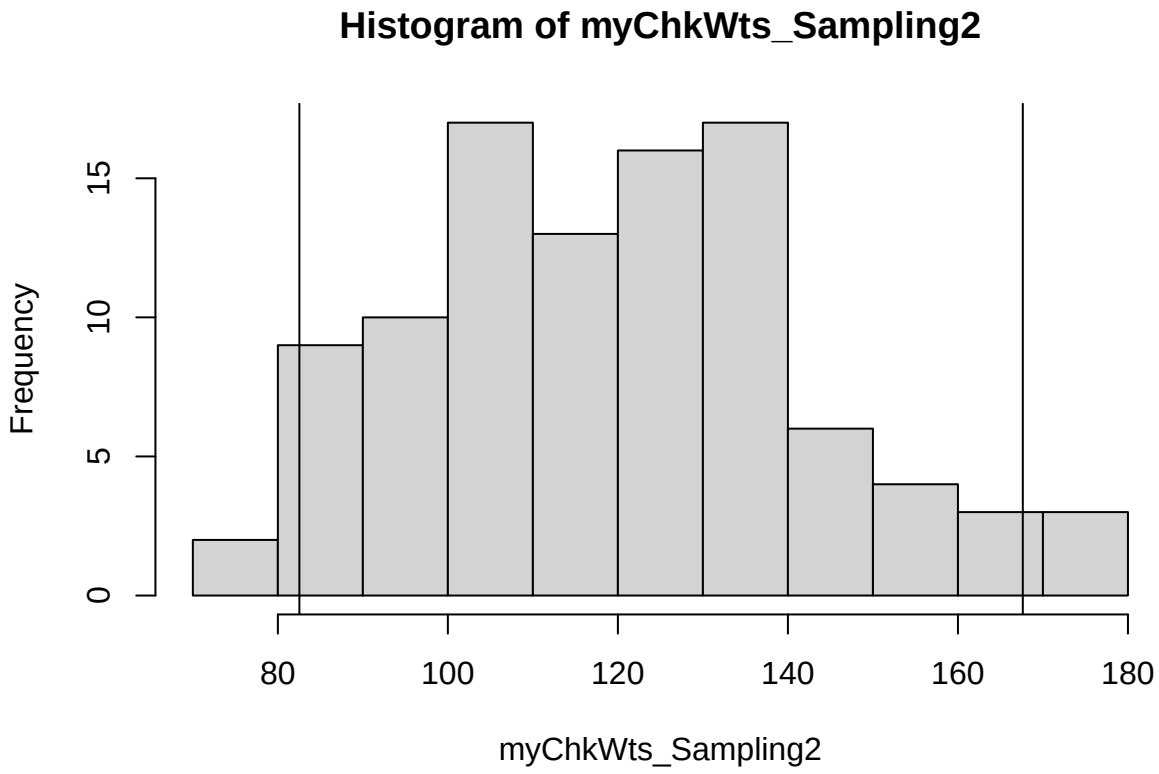
Redoing Exercise 5 Using n = 100

Redo Exercise 5, but this time use a sample size of n = 100 (instead of the original sample size of n = 11 used in Exercise 5).

```
myChkWts_Sampling2 <- replicate(100,mean(sample(myChkWts, 11, replace = TRUE)), simplify = TRUE)
summary(myChkWts_Sampling2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75.09  102.07  117.64  119.37  133.64  177.55
```

```
hist(myChkWts_Sampling2)
abline(v = quantile(myChkWts_Sampling2, 0.025))
abline(v = quantile(myChkWts_Sampling2, 0.975))
```



Explaining the Results

Explain why the 2.5% and 97.5% quantiles are different from the results you got for Exercise 5. As a hint, think about what makes a sample "better."

Law of large numbers really do help normalizing the data or making the distribution a normal distribution. By reducing the size of sampling by 90% or to 100 samples the data will be less normal.