Homework 3

Kwasi Mensah

2022-10-28

Author: Kwasi Mensah Homwork_Number: 3 Due Date: “2022-29-21” Output: pdf_document Attribution statement:

I did the homework by myself, with help from the book and the professor #R Markdown #Run these three functions to get a clean test of homework code

```
dev.off() #Clear the graph window

## null device
##           1

cat('\014') #Clear the console

rm(list = ls()) #Clear user objects from the environment
```

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
###

2. For the remaining exercises in this set, we will use one of RVs 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. use the summary(ChickWeight) command to reveal basic information about the ChickWeight data set. You will find that ChickWeight contains four different variables.

   Name the four variables. Use the dim(ChickWeight) command to show the dimensions of the ChickWeight data set. 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.

```
# Load the data set
data("ChickWeight")

# Reveal basic information about the 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

#The 4 Variables are weight, Time, Chick, & Diet.

# Show the dimensions of the data set
dim(ChickWeight)

## [1] 578   4

# The number of rows in the data set is the other number shown. There are a
total of 578 observations.
```

3. When a data set contains more than one variable, R offers another subsetting operator,$, to access each variable individually. For the exercises below, we are interested only in the contents of one of the variables in the data set, called weight. We can access the weight variable by itself, using the ,$withthisexpression: ChickWeight$weight. Run the following commands, say what the command does, report the output, and briefly explain each piece of output: Report on a summary of the weight variable.

```
#Summary of the weight column
summary(ChickWeight$weight)

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

# Return the first six values of the weight column in the data set.
head(ChickWeight$weight)

## [1] 42 51 59 64 76 93

# Return the mean of the weight column
mean(ChickWeight$weight)

## [1] 121.8183

# Set the weights column to a variable
myChkWts <- ChickWeight$weight
```

```
#Shows the .50 Quantile of the variable or the median
quantile(myChkWts,0.50)

## 50%
## 103
```

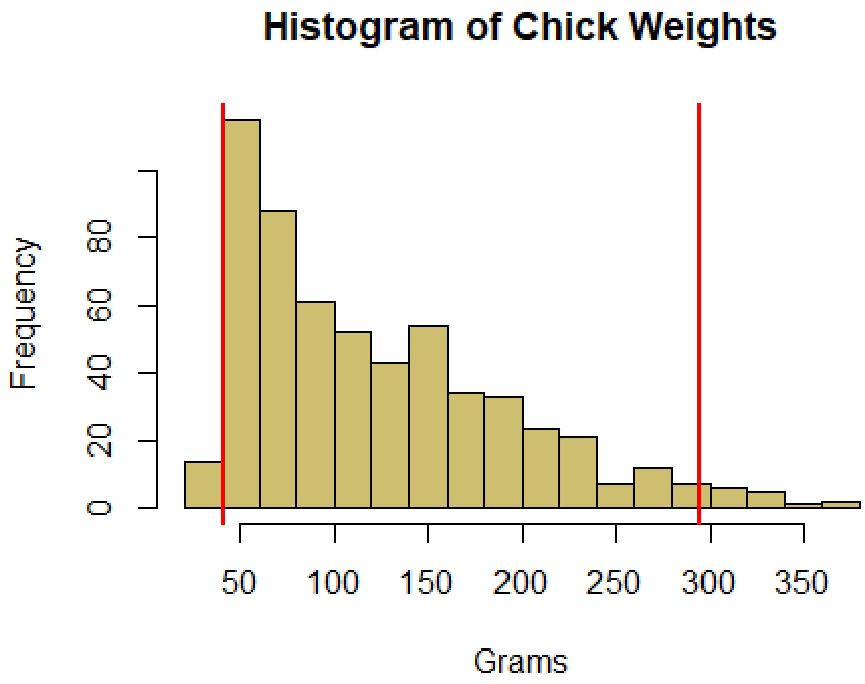
4. In the second to last command of the previous exercise, you created a copy of the weight data from the ChickWeight data set and put it in a new vector called myChkWts.

   You can continue to use this myChkWts variable for the rest of the exercises below. Create a histogram for that variable. Then write code that will display the 2.5% and 97.5% quantiles of the distribution for that 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 975% quantiles signify.

```
myChkWts <- ChickWeight$weight


# Create a histogram from the weights vector
hist(myChkWts, main="Histogram of Chick Weights"
     , ylab = "Frequency"
     , xlab="Grams"
     , col="lightgoldenrod3"
     , breaks=15)
abline(v=quantile(myChkWts,c(.025,.975)),col="red",lwd=2)
```



```
# Return quantile values
quantile(myChkWts,c(.025,.50,.975))

##    2.5%     50%   97.5%
##  41.000 103.000 294.575

mean(myChkWts)

## [1] 121.8183

# The histogram is takes on a shape of a bell curve that is skewed to the
right. The mean of the myChkWts variable is 121.8183, and the median is 103,
they are both to the right of the distribution. We are shown the quantiles of
the data that falls within 2.5% of the data which is 41 and the 97.5% of the
data which is 294.575. It is shown in the graph that most of the data falls
into the 2.5% quantile when compared to the 97.5% quantile.
```

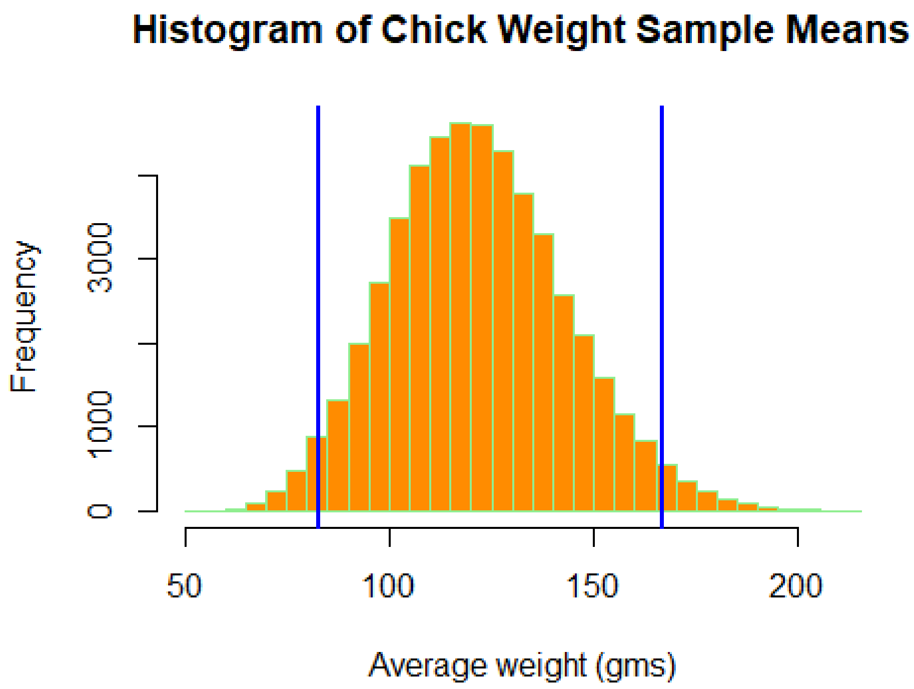
5. Write R code that will construct a sampling distribution of means from the weight data (as noted above, if you did exercise 3 you can use myChkWts instead of ChickWeight$weight to save yourself some typing). Make sure that the sampling distribution contains at least 1,000 means. Store the sampling distribution in a new variable that you can keep using.

   Use a sample size of n = 11 (sampling with replacement). Show a histogram of this distribution of sample means. 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.

```r
# Create a collection of sample means from the myChkWts variable.
SamplingDistribution <- replicate(50000,mean(sample(myChkWts,11,replace=T)))

# Display a histogram of the sampling distribution
hist(SamplingDistribution, main="Histogram of Chick Weight Sample Means"
     , ylab = "Frequency"
     , xlab="Average weight (gms)"
     , col="darkorange"
     , border = "lightgreen"
     , breaks=30)
abline(v=quantile(SamplingDistribution,c(.025,.975)),col="blue",lwd=2)
```



```r
# Quantile values
quantile(SamplingDistribution, c(.025, .975))

##      2.5%     97.5%
##  82.81818 166.81818
```

6. 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.

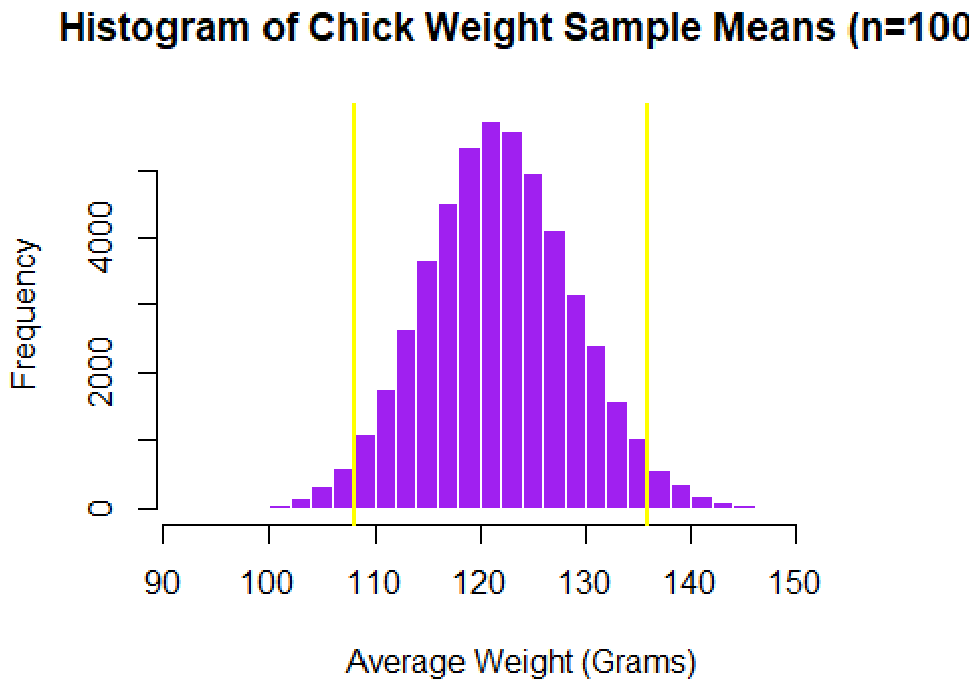
Briefly describe, from a conceptual perspective and in your own words, what the difference is between a distribution of raw data and a distribution of sampling means. Finally, comment on why the 2.5% and 97.5% quantiles are so different between the raw data distribution and the sampling distribution of means.

```
# The raw data distribution has a positive bell curve compared to the
distribution of sampling means has a normal bell curve. The raw data
distribution's mean comes from the original data set. When taking the Law of
Large Numbers and the Central Limit Theorem into account, the distribution of
sampling means start to create a normal distribution or bell shaped curve.
The quantiles are different because quantile of the raw data distribution are
using the original data, and the sampling distribution of mean's quantile is
using a subset of a data from the number of trials specified, which results
in more observations when compared to the raw data, to get the 2.5% and 97.5%
quantiles.
```

7. 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). 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.” Create a collection of sample means from the weight attribute.

```
SamplingDistribution1 <-
replicate(50000,mean(sample(myChkWts,100,replace=T)))

# Display a histogram of the sampling distribution
hist(SamplingDistribution1, main="Histogram of Chick Weight Sample Means
(n=100)"
     , ylab = "Frequency"
     , xlab="Average Weight (Grams)"
     , col="purple"
     , border = "white"
     , breaks=30)
abline(v=quantile(SamplingDistribution1,c(.025,.975)),col="yellow",lwd=2)
```

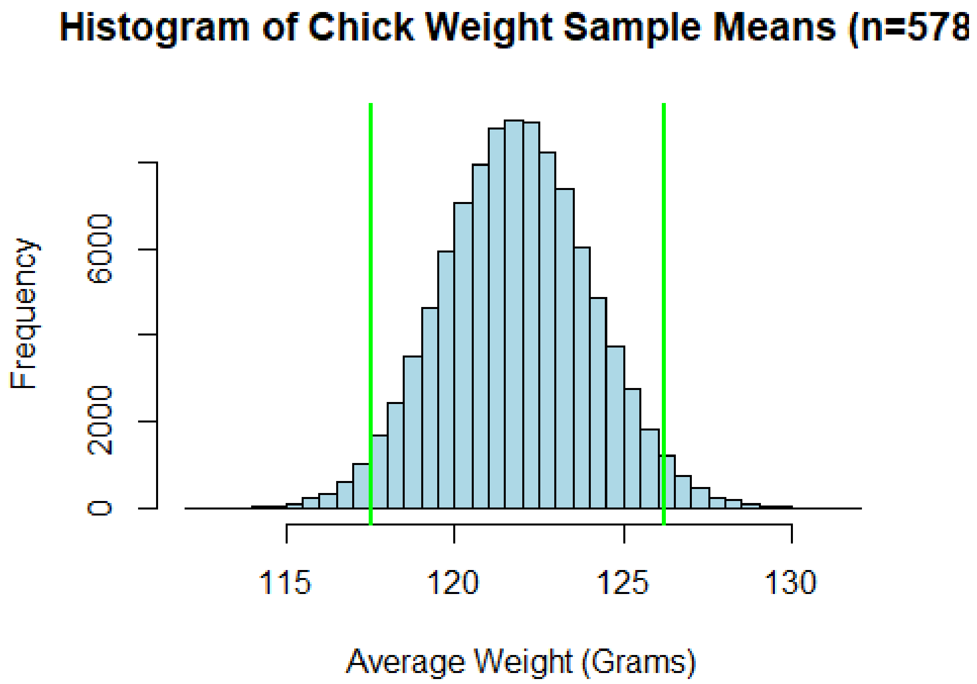


```
# Return quantile values
quantile(SamplingDistribution1, c(.025, .975))

##   2.5%  97.5%
## 108.25 135.99

# Create a collection of sample means from the weight attribute.
SamplingDistribution2 <-
replicate(100000,mean(sample(myChkWts,1028,replace=T)))

# Display a histogram of the sampling distribution
hist(SamplingDistribution2, main="Histogram of Chick Weight Sample Means
(n=578)"
     , ylab = "Frequency"
     , xlab="Average Weight (Grams)"
     , col="lightblue"
     , border = "black"
     , breaks=30)
abline(v=quantile(SamplingDistribution2,c(.025,.975)),col="green",lwd=2)
```



```
# Return quantile values
quantile(SamplingDistribution2, c(.025, .975))

##     2.5%    97.5%
## 117.5477 126.1800

# The 2.5% and 97.5% quantiles are different because the sample size
increased, which as a result creates more observations and the data that is
subset for the specified quantile have different values. As the sample size
increases the mean gets closer to the actual mean of the population in
accordance with the Law of Large Numbers. This is because as the sample size
increase it becomes more representative of the population.
```