Homework Chapter 3 - Mark Cappiello

1. 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. 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.

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 four variables for the ChickWeight dataset are weight, Time, Chick and Diet

dim(ChickWeight)

## [1] 578 4

the first number in the output of dim(ChickWeight) is the number of rows or observations in the data set

1. 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 weight. Run the following commands, say what the command does, report the output, 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

summary(ChickWeight$weight) gives us the minimum, first quartile, median, mean, third quartile and maximum values for the “weight” variable or column in the “ChickWeight” data set.

head(ChickWeight$weight)

## [1] 42 51 59 64 76 93

the head() function is used when you want to see the first values in a data set. By default it will show the first six values but this can be changed by adding a number as a second argument in the parenthesis. for example head(dataset, 10) will show the first 10 values.

head(ChickWeight$weight) returns the first six values in the ChickWeight weight column.

mean(ChickWeight$weight)

## [1] 121.8183

mean(ChickWeight$weight) calculates the average value in the ChickWeight weight column. It takes the total sum of the data and de=ivides it by the number of observations.

myChkWts <-ChickWeight$weight   
#myChkWts

myChkWts <-ChickWeight$weight takes the values from data set ChickWeight variable weight and assigns these values to a new variable called myChkWts. myChkWts data is in the form of a vector.

quantile(myChkWts,0.50)

## 50%   
## 103

quantile(myChkWts,0.50) The quantile function calculates a value where a percentage of data falls below it. In this case it calculates the value where 50 percent of data will fall below. Fifty percent represents the median.

1. 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 97.5% quantiles signify.

# Calculate mean, median, and quantiles  
mean\_weight <- mean(myChkWts)  
median\_weight <- median(myChkWts)  
quantiles <- quantile(myChkWts, c(0.025, 0.975))  
  
# Create the histogram  
hist(myChkWts, breaks = 10, col = "lightblue", main = "Histogram of Chick Weights",  
 xlab = "Weights", ylab = "Frequency")  
  
# Add vertical lines for mean, median, and quantiles  
abline(v = mean\_weight, col = "red", lwd = 2)  
abline(v = median\_weight, col = "blue", lwd = 2)  
abline(v = quantiles[1], col = "purple", lwd = 2, lty = 2)  
abline(v = quantiles[2], col = "purple", lwd = 2, lty = 2)  
  
# Add text labels for quantiles, mean, and median  
text(x = mean\_weight, y = max(table(cut(myChkWts, breaks = 10))), labels = "Mean", col = "red", pos = 2, srt = 90)  
text(x = median\_weight, y = max(table(cut(myChkWts, breaks = 10))) - 5, labels = "Median", col = "blue", pos = 2, srt = 90)  
text(x = quantiles[1], y = max(table(cut(myChkWts, breaks = 10))) - 10, labels = "Q 0.025", col = "purple", pos = 2, srt = 90)  
text(x = quantiles[2], y = max(table(cut(myChkWts, breaks = 10))) - 15, labels = "Q 0.975", col = "purple", pos = 2, srt = 90)



mean\_weight

## [1] 121.8183

median\_weight

## [1] 103

quantiles

## 2.5% 97.5%   
## 41.000 294.575

Interpretation of myChkWts:

the mean of myChkWts = 121.8183 the median of myChkWts = 103 myChkWts Q 0.025 = 41 myChkWts Q 0.975 = 295

Since the mean is greater than the median this indicates that the data is right-skewed. This means that there are values in the data that are skewing the results higher. 2.5% of the weights are below 41 and 2.5% are above 295. This means that 95% of the chick weights fall between 41 and 295.

1. 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.

# Set parameters for the sampling distribution  
sample\_size <- 11  
num\_samples <- 1000  
#set.seed(123) # Setting seed for reproducibility  
  
# Create a function to draw a sample and calculate the mean  
sample\_mean <- function(data, size) {  
 mean(sample(data, size, replace = TRUE))  
}  
  
# Generate the sampling distribution of means  
sampling\_distribution <- replicate(num\_samples, sample\_mean(myChkWts, sample\_size))  
  
# Display a histogram of the sampling distribution  
hist(sampling\_distribution, breaks = 30, col = "lightblue", main = "Sampling Distribution of Sample Means",  
 xlab = "Sample Mean", ylab = "Frequency")  
  
# Calculate the 2.5% and 97.5% quantiles  
quantiles <- quantile(sampling\_distribution, c(0.025, 0.975))  
  
# Add vertical lines for the quantiles  
abline(v = quantiles[1], col = "red", lwd = 2, lty = 2)  
abline(v = quantiles[2], col = "red", lwd = 2, lty = 2)  
  
# Add text labels for the quantiles  
text(x = quantiles[1], y = max(table(cut(sampling\_distribution, breaks = 30))), labels = "Q 0.025", col = "red", pos = 2, srt = 90)  
text(x = quantiles[2], y = max(table(cut(sampling\_distribution, breaks = 30))) - 5, labels = "Q 0.975", col = "red", pos = 2, srt = 90)

