Matthew L. Pergolski
IST 772
Dr. Block
10/24/2021

Homework 3

Beginning Statement

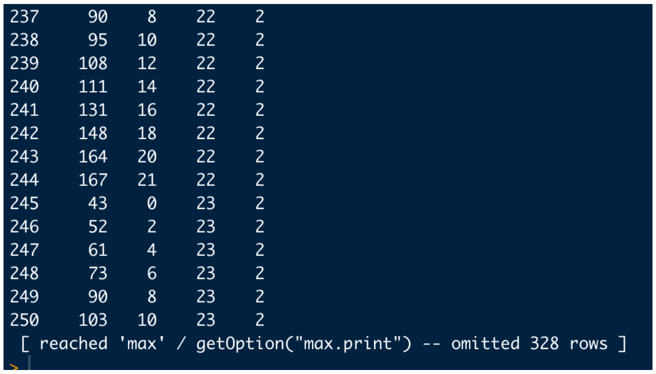
"I produced the material below with no assistance [direct quote from IST 772 class syllabus]." Note: Homework questions from the book may have been copied/pasted into the document for both the student and viewer's convenience.

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

Homework Question 2

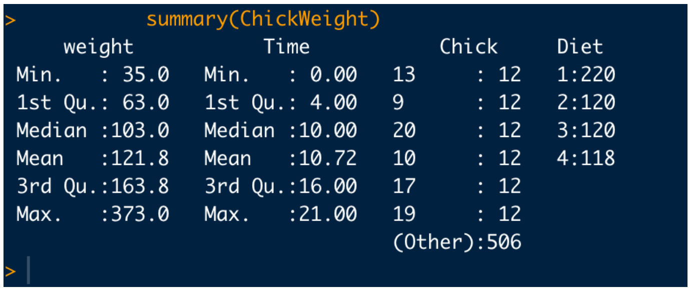
Question:
For the remaining exercises in this set, we will use one of R's built-in data sets, called the "ChickWeight" data set. According to the documentation for R, the ChickWeight data set contains information on the weight of chicks in grams up to 21 days after hatching. Use the summary(ChickWeight)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

Answer/Student Response:
Upon calling the 'ChickWeight' dataset, I get the following R output (please see R script file for more detail, only a snippet will be provided in this word/PDF file):



```
237     90    8    22    2
238     95   10    22    2
239    108   12    22    2
240    111   14    22    2
241    131   16    22    2
242    148   18    22    2
243    164   20    22    2
244    167   21    22    2
245     43    0    23    2
246     52    2    23    2
247     61    4    23    2
248     73    6    23    2
249     90    8    23    2
250    103   10    23    2
 [ reached 'max' / getOption("max.print") -- omitted 328 rows ]
>
```

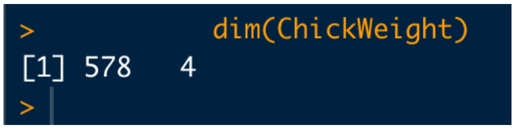
Upon calling the summary function for ChickWeight, I get the following output:



```
>          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
>
```

This summary command reveals four variables or attributes: (1) weight; (2) Time; (3) Chick; (4) Diet.

Upon running the dim function on this same dataset, I got the following output:



```
>          dim(ChickWeight)
[1] 578   4
>
```

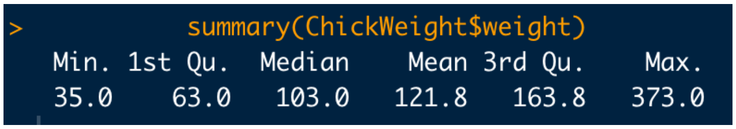
The dim function reveals the number of observations (578) as well as the variables or attributes (4); more specifically, the observations (i.e., rows) equate to the number of chick weights that were recorded.

Homework Question 3

Question:
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, say what the command does, report the output, and briefly explain each piece of output:

summary(ChickWeight$weight)
head(ChickWeight$weight)
mean(ChickWeight$weight)
myChkWts <- ChickWeight$weight
quantile(myChkWts,0.50)

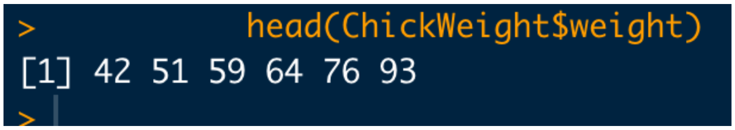
Answer/Student Response:
Upon running the summary command, I observed the following output:



```
>          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 function provides insight into the min, max, mean, median, and 1st/3rd quartiles of the dataset. We are able to gain 'some' insight into the spread of our data with this command.

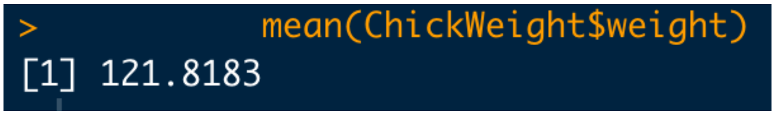
Upon running the head command, I observed the following output:



```
>           head(ChickWeight$weight)
[1] 42 51 59 64 76 93
>
```

The head function reveals the values of the dataset towards the top of the provided vector (i.e., ChickWeight$weight).

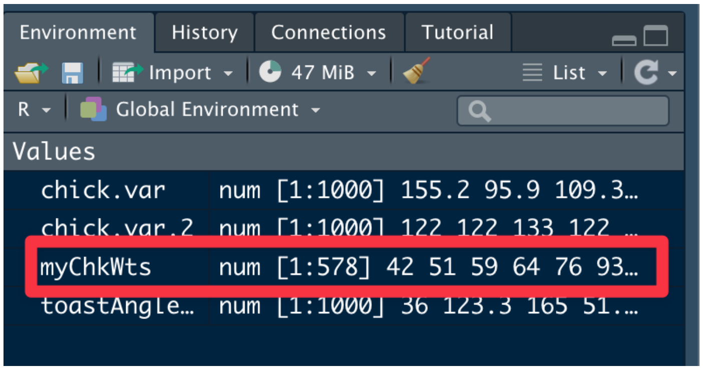
Upon running the mean command, I observed the following output:



```
>           mean(ChickWeight$weight)
[1] 121.8183
```

The mean function calculates the average of the values for the provided vector (i.e., ChickWeight$weight).

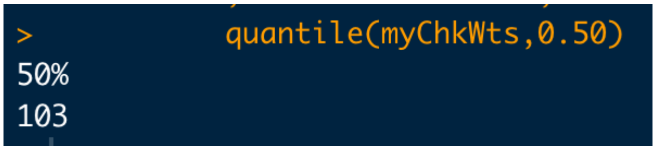
Upon creating the ‘myChkWts’ variable, I observed the following in the ‘Environment’ section of RStudio:



```
Values
  chick.var      num [1:1000] 155.2 95.9 109.3…
  chick.var.2    num [1:1000] 122 122 133 122 …
  myChkWts       num [1:578] 42 51 59 64 76 93…
  toastAngle…    num [1:1000] 36 123.3 165 51.…
```

'myChckWts' is a variable where information is stored; in this instance, the weight vector of the ChickWeight data is stored in this variable.

Upon running the quantile command, I observed the following output:



```
>           quantile(myChkWts,0.50)
50%
103
```

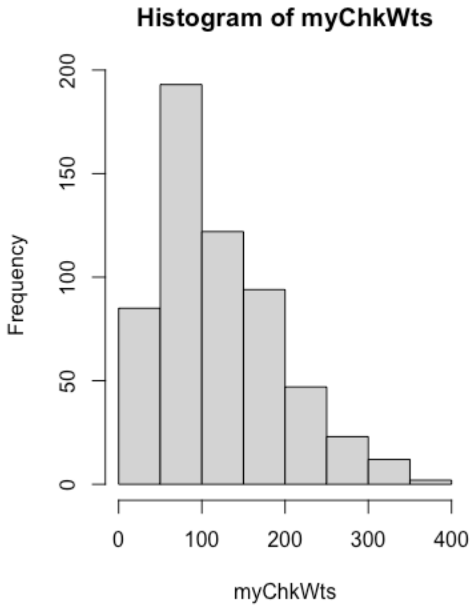
The quantile function provides a value corresponding to a specific percentage of the data specified from the user; in this instance, the value associated with the 50% mark of the data is returned (i.e., 103).

Homework Question 4

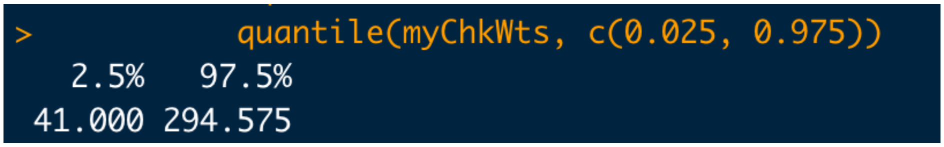
Question:
In the second to last command of the previous exercise, you created a copy of the

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.

Answer/Student Response:
Upon creating a histogram for 'myChkWts,' the following was observed:



Also, upon running the quantile function on this dataset for 2.5% and 97.5%, the following was observed:



```
>            quantile(myChkWts, c(0.025, 0.975))
   2.5%   97.5%
 41.000 294.575
```

'myChkWts' is a variable that contains the column data of chick weights from the 'ChickWeight' dataset. A histogram function was put forth on the myChkWts variable and produced an output of a positively skewed distribution. We can tell it's positively skewed by the 'tail' that forms on the right-hand side of the illustration

The shape of the distribution can also be described with the summary function; this informs us that the mean (121.8) is higher than the median (103), which makes sense since the mean is more 'vulnerable' to outliers vs the median

For the quantiles, we get a value of 41 for .025; this means that 2.5% of the data in this dataset is less than or equal to 41 units. Similarly, we get a value of 294.575 for the value corresponding to the quantile of

.975, meaning that 97.5% of the data is less than or equal to 294.575 units (and only 100%-97.5% of the data is larger than this value).

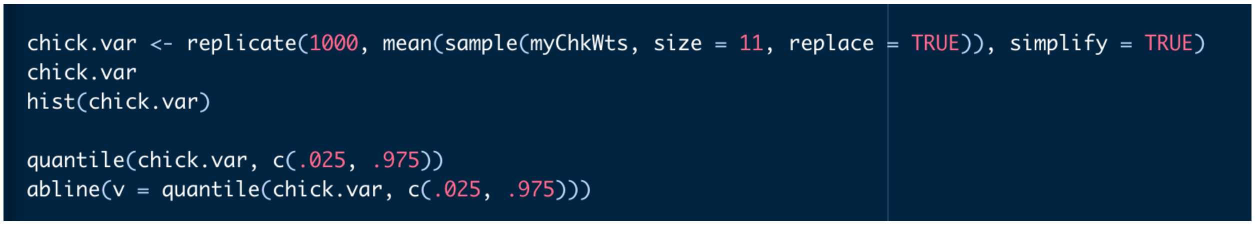
Homework Question 5

Question:

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.

Answer/Student Response:

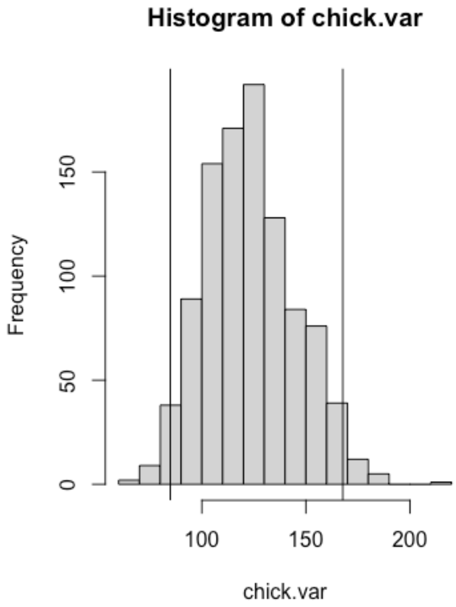
The following code was developed/constructed for this question:



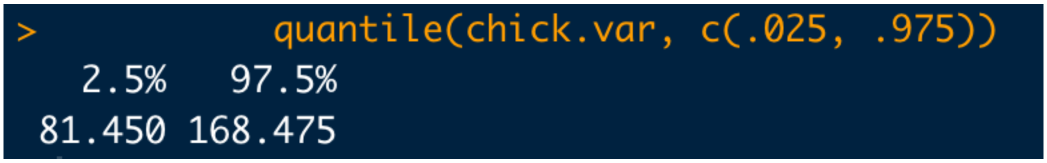
```
chick.var <- replicate(1000, mean(sample(myChkWts, size = 11, replace = TRUE)), simplify = TRUE)
chick.var
hist(chick.var)

quantile(chick.var, c(.025, .975))
abline(v = quantile(chick.var, c(.025, .975)))
```

A histogram of the data was observed (reference above screenshot and R script attached to the homework submission on 2SU for more detail: hist(chick.var)). Note this histogram has vertical lines in place for the quantiles of 2.5% as well as 97.5% via the abline(v = quantile(chick.var, c(.025, .975))) R code:



The following was observed from the quantile function:



```
>          quantile(chick.var, c(.025, .975))
   2.5%   97.5%
 81.450 168.475
```

Homework Question 6

Question:
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.

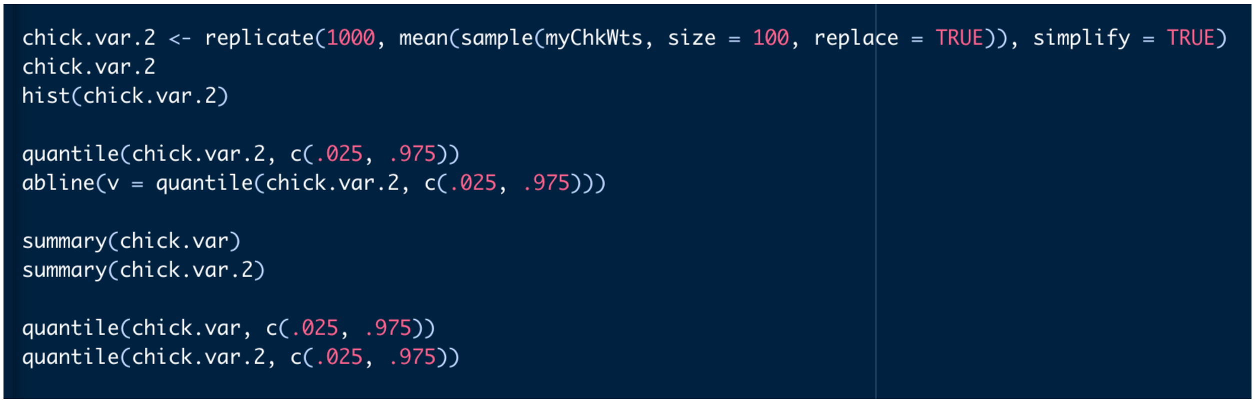
Answer/Student Response:
In this instance, we are comparing a histogram of raw data versus sampling data that was replicated many times over (1,000 to be exact). The raw data showed a skewed distribution while the replicated sampling distribution showed what appears to be more of a 'normal' curve. My initial thoughts are that the raw data is not subject to the central limit theorem (and law of large numbers) like the replicated sampling process is. As the sample size (assuming with replacement) increases – along with a larger amount of 'trials' – we can observe a dataset reflect something similar to a normal distribution. Since the raw data is not replicated, and is considered one 'trial,' we would not necessarily expect to see normal curve when plotting the dataset in a histogram.

The quantiles of 2.5 and 97.5% convey different values for the two distributions since one is a normal distribution (i.e., sample distribution) and the other is a positively skewed distribution (i.e., raw data distribution).

Homework Question 7

Question:
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."

Answer/Student Response:
The following code was developed for this problem:



```
chick.var.2 <- replicate(1000, mean(sample(myChkWts, size = 100, replace = TRUE)), simplify = TRUE)
chick.var.2
hist(chick.var.2)

quantile(chick.var.2, c(.025, .975))
abline(v = quantile(chick.var.2, c(.025, .975)))

summary(chick.var)
summary(chick.var.2)

quantile(chick.var, c(.025, .975))
quantile(chick.var.2, c(.025, .975))
```

In this instance, we are comparing two different distributions of replicated sample means. One has a sample size of 11 (with 1,000 trials) while the other has a sample size of 100 (with 1,000 trials). The output for quantiles 2.5 and 97.5% are different – and the reasoning for this would simply be the law of large numbers. As the number of samples increase, we would expect to see a more ‘perfect’ normal distribution, essentially a more ‘refined’ or ‘accurate’ illustration showing the mean, median, mode, etc.