IST772 Problem Set 3

Abhijith Anil Vamadev

The homework for week 3 is based is based on exercises 2 through 7 on pages 50 and 51, but with changes as noted in the text in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor

Set the random number seed so that your results will match mine.

```
set.seed(772) #Setting a common seed
```

Chapter 3, Exercise 2

For the remaining exercises in this set, we will use one of R's built-in data sets, called the "trees" data set. According to the documentation for R, the trees data set contains information on the measurements of the girth, height and volume of timber in 31 felled black cherry trees. Use the summary(trees) command to reveal basic information about the trees data set. You will find that trees contains three different variables. Name the variables (1 pt). Use the dim(trees) command to show the dimensions of the trees data set. The second number in the output, 3, is the number of columns in the data set, in other words the number of variables. What is the first number (1 pt)? Report it and describe briefly what you think it signifies.

```
data("trees") #taking the data trees
trees<- trees #assigning the data set into a new variable
summary(trees) #summary of the tree variable
```

```
##      Girth           Height       Volume     
##  Min.   : 8.30   Min.   :63   Min.   :10.20  
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40  
##  Median :12.90   Median :76   Median :24.20  
##  Mean   :13.25   Mean   :76   Mean   :30.17  
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30  
##  Max.   :20.60   Max.   :87   Max.   :77.00
```

```
dim(trees) #dimensions of the tree variable
```

```
## [1] 31  3
```

```
trees[1,1] #displaying the first value from the first column, first row.
```

```
## [1] 8.3
```

- The three different variables are Girth, Height and Volume of trees. 31 rows, 3 columns.
- The first number in the first column is 8.3, which is the girth of 1 tree.

Chapter 3, Exercise 3

When a data set contains more than one variable, R offers a subsetting operator, $, to access each variable individually. (NB. the backslash is needed in the notebook file because a dollar sign by itself means to shift to math mode. In R code, you would just use the dollar sign, without the back slash.) For the exercises below, we are interested only in the contents of one of the variables in the data set, called Girth. We can access the Girth variable by itself, using the $, with this expression: trees$Girth. Run the following commands, add a comment to each line saying what each command does, report the output, and briefly explain each piece of output (1 pt for summary, head, and mean; 1 pt for new variable, and 0.50 quantile):

```
summary(trees$Girth) #summary of the Girth variable
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   11.05   12.90   13.25   15.25   20.60
```

```
head(trees$Girth) #first five rows of the Girth variable
```

```
## [1]  8.3  8.6  8.8 10.5 10.7 10.8
```

```
mean(trees$Girth) #mean of the girth column
```

```
## [1] 13.24839
```

```
myTreeGirth <- trees$Girth #assign the girth column to mTreeGirth
quantile(myTreeGirth,0.50) #taking the median of the myTreeGirth
```

```
##  50%
## 12.9
```

- Summary of the Girth variable gives thee min, max, 1st, 2nd and 3rd quartile.
- Head gives the first 5 numbers in the girth column, which are 8.3, 8.6, 8.8 etc
- Mean finds the average of the girth column which is 13.248
- myTreeGirth is the new variable that contains all the girth values from the trees
- quantile finds the median or the 50the percentile of the girth column from the new variable, which is 12.9, which is same as the median from the summary function.

Chapter 3, Exercise 4

In the second to last command of the previous exercise, you created a copy of the girth data from the trees data set and put it in a new vector called myTreeGirth You can continue to use this myTreeGirth 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 (1 pt for histogram and quantiles). Write an interpretation of the variable, including descriptions of the mean, median (1 pt for mean and median), shape of the distribution (1 pt), and the 2.5% and 97.5% quantiles. Make sure to clearly describe what the 2.5% and 97.5% quantiles signify (1 pt).

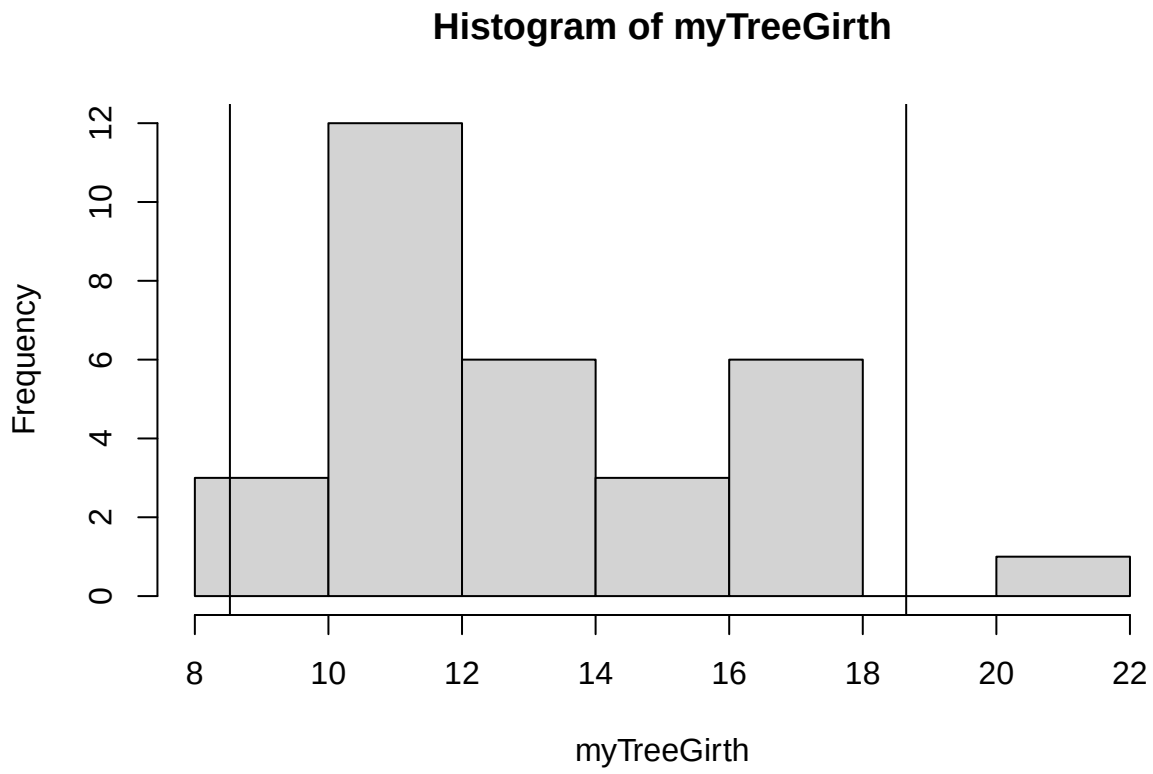
```
hist(myTreeGirth) #displaying the histogram
mean(myTreeGirth) #mean of the myTreeGirth variable
```

```
## [1] 13.24839
```

```r
median(myTreeGirth) #median of myTreeGirth
```

```
## [1] 12.9
```

```r
abline(v = quantile(myTreeGirth, 0.025)) #2.5% mark in the histogram
abline(v = quantile(myTreeGirth, 0.975)) #97.5% mark in the histgoram
```



* The 2.5 and 97.5 are the quantile of data. This means that 95% of the data lies within the two lines displayed above, and only 5% of the data lies outside the two lines. A value lying outside the two lines are considered to be rare and extreme values. * The shape of the distribution cannot be exactly determined, but it somewhat looks like a normal distribtuion with 1 peak and looking like a bell curve. * The mean of the histogram is 13.24 and the 50th percentile/median is 12.9. 50% of the data lies towards left of the 12.9 and 50% of the data lies towards the right of the 12.9 which is the median.

Chapter 3, Exercise 5

Write R code that will construct a sampling distribution of means from the girth data (as noted above, if you did exercise 3 you can use myTreeGirth instead of trees$Girth). 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 = 7 (sampling with replacement) (2 pts). 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 (1 pt).

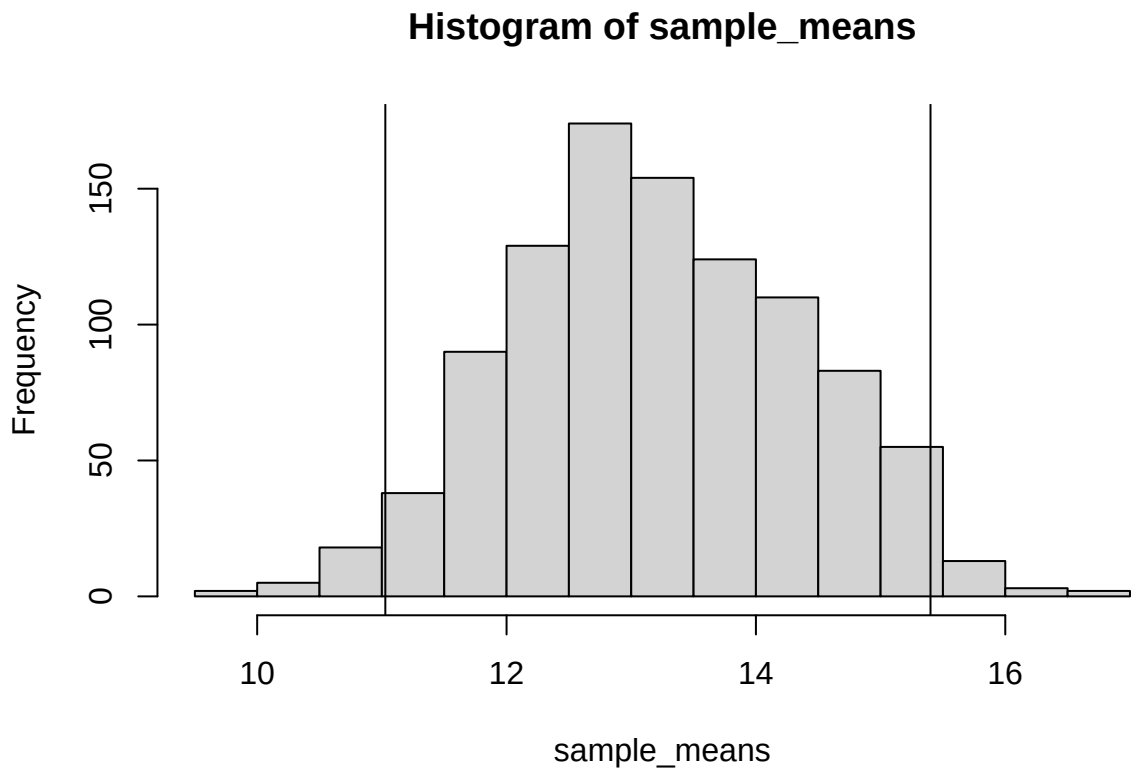
```r
set.seed(772)
sample_means <- replicate(1000, mean(sample(myTreeGirth, size = 7, replace = TRUE)))
#sampling the data with size 7 with replacement and replicated 1000 times
head(sample_means) #head of the sample_means
```

```
## [1] 14.48571 13.60000 12.11429 12.60000 13.97143 11.51429
```

```r
range(sample_means) #range of the sample_means
```

```
## [1]  9.957143 16.800000
```

```r
hist(sample_means) #histogram of the sampled data
abline(v = quantile(sample_means, 0.025)) #2.5% perctile mark
abline(v = quantile(sample_means, 0.975)) #97.5 percentile mark
```



Chapter 3, Exercise 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 (2 pts). 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 (2 pts). * The difference between a distribution of raw

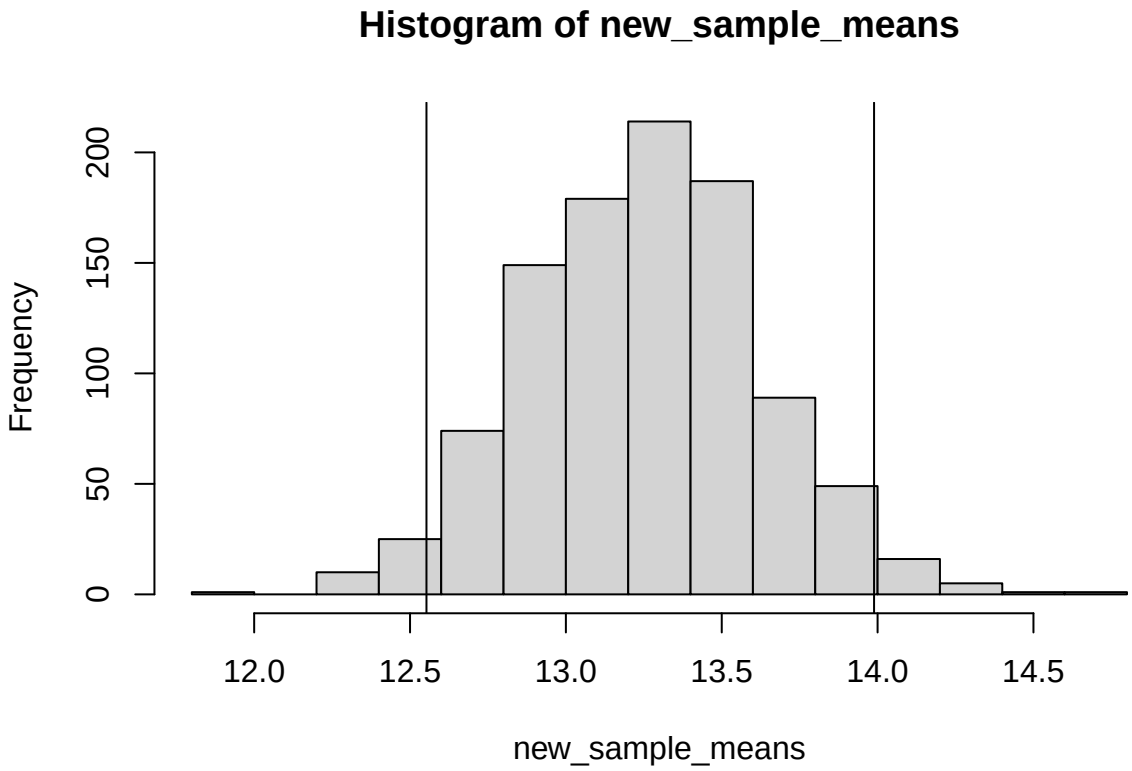
data and a distribution of sampling means, is that the more we sample and replicate the sample means, the distribution will progressively become a bell shaped normal distribution. The more times we sample the data the smoother the curve becomes and looks more like a bell-shaped.

- The quantile are different between the raw data and the sampling distribution of means as within the mean sample data, all the values are more likely to be closer to each other and will be within a range of numbers.The raw data quantile are just the raw quantile for 2.5% 97.5%, but for the sample means the quantile are for sample means, and all means are within a range of means which is between 9.95(min) and 16.8(max).

Chapter 3, Exercise 7

Redo Exercise 5, but this time use a sample size of n = 70 instead of the original sample size of n = 7 used in Exercise 5. (1 pt) Explain why the 2.5% and 97.5% quantiles are different from the results you got for Exercise 5 (1 pt). As a hint, be sure to specify what feature or characteristic of a sample makes it a "better" sample. (1 pt)

```
set.seed(772)
new_sample_means <- replicate(1000, mean(sample(myTreeGirth, size = 70, replace = TRUE)))
#Sampling the data with replacement sampling with 70 and replicating it 1000
hist(new_sample_means) #hist of the new data
abline(v = quantile(new_sample_means, 0.025)) #2.5% percentile mark
abline(v = quantile(new_sample_means, 0.975)) #97.5% percentile mark
```



* The quantile are different as the sample size is higher from 7 to 70. There is limited variability and so the

quantile are more or less closer to the quantile of the population. As the sample size grows, the more likely a sample is within the range of the quantile, as most of the data should be in the 95% area.

- The number of sample size is the characteristic in determining a better sample, the higher the sample size the better the sample will be and the better it represents the population.