IST772 Problem Set 3

Shivani Sanjay Mahaddalkar

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)
```

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.

```
summary(trees)

##      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  

#The three different variables are Girth, Height and Volume

dim(trees)

## [1] 31  3

#The data-set has 3 columns and 31 rows.
#The first number(31) indicates the entries of the 31 trees whose girth,
height and volume was measured.
```

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) #Gives the summary of the girth column of the data-set.
Gives the min, 1st quartile, median, 3rd quartile and max values.

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

head(trees$Girth) #Gives the first five entries in the girth column of the
data set

## [1]  8.3  8.6  8.8 10.5 10.7 10.8

mean(trees$Girth) #Gives the average of the girth column

## [1] 13.24839

myTreeGirth <- trees$Girth #Assigns the girth column as a list to a new
variable
quantile(myTreeGirth,0.50) # Gives the 50th percentile of the girth column.
Which means 50% of the values of girth lie below the output of the command.

##  50%
## 12.9
```

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).

```
median(myTreeGirth)

## [1] 12.9

mean(myTreeGirth)
```

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

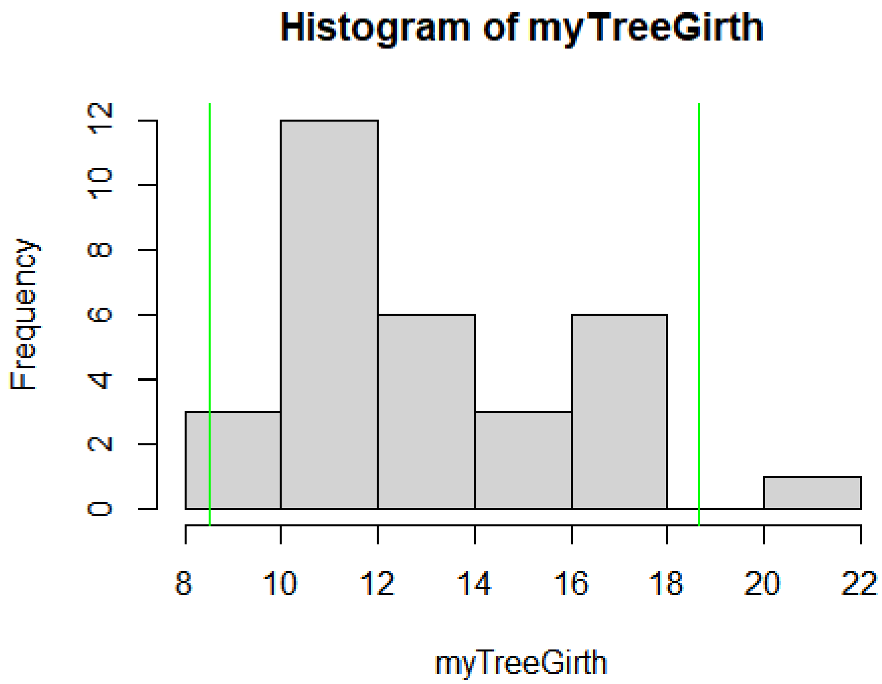
```r
quantile(myTreeGirth, probs=0.025)
```

```
##  2.5%
## 8.525
```

```r
quantile(myTreeGirth, probs=0.975)
```

```
## 97.5%
## 18.65
```

```r
hist(myTreeGirth)
abline(v=quantile(myTreeGirth, probs=0.025),col="green")
abline(v=quantile(myTreeGirth, probs=0.975),col="green")
```

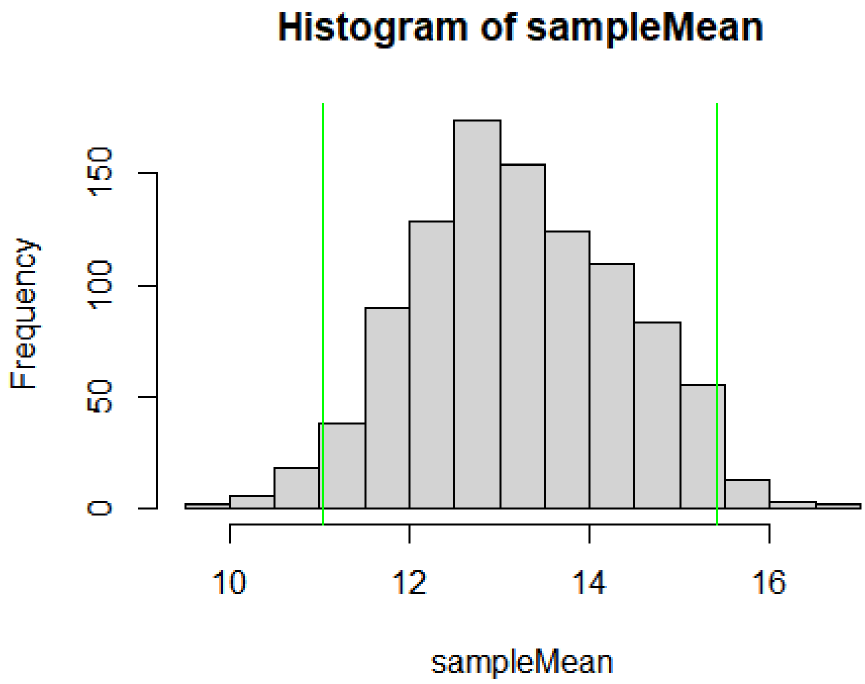


```r
#The mean and median suggest that they lie in the third bin. The histogram
looks right skewed and does not closely resemble a normal distribution.
#The values between 2.5% and 97.5% signify that 95% of the values in the data
lies between the range 8.25 and 18.65. Generally, when sampling it is taken
as the threshold quartiles to assess if the certain sample is taken from the
population.
```

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).*

```
#Function for sampling for a size n
sampleGirthValues <- function(n) {sample(myTreeGirth, size=n, replace=TRUE)}
#Replicating the sampling function to find the means
sampleMean <- replicate(1000, mean(sampleGirthValues(7)))
quantile(sampleMean, probs=0.025)

##     2.5%
## 11.02786

quantile(sampleMean, probs=0.975)

##    97.5%
## 15.40036

#The 2.5% to 0.975% range of values ie 11.11429 to 15.53179 indicates that if
the mean of 7 samples lies between that range, it is possible that the
population of the sample was the myTreeGirth sample.

#Histogram of the mean of samples 1000 times.
hist(sampleMean)
abline(v=quantile(sampleMean, probs=0.025),col="green")
abline(v=quantile(sampleMean, probs=0.975),col="green")
```



Chapter 3, Exercise 6

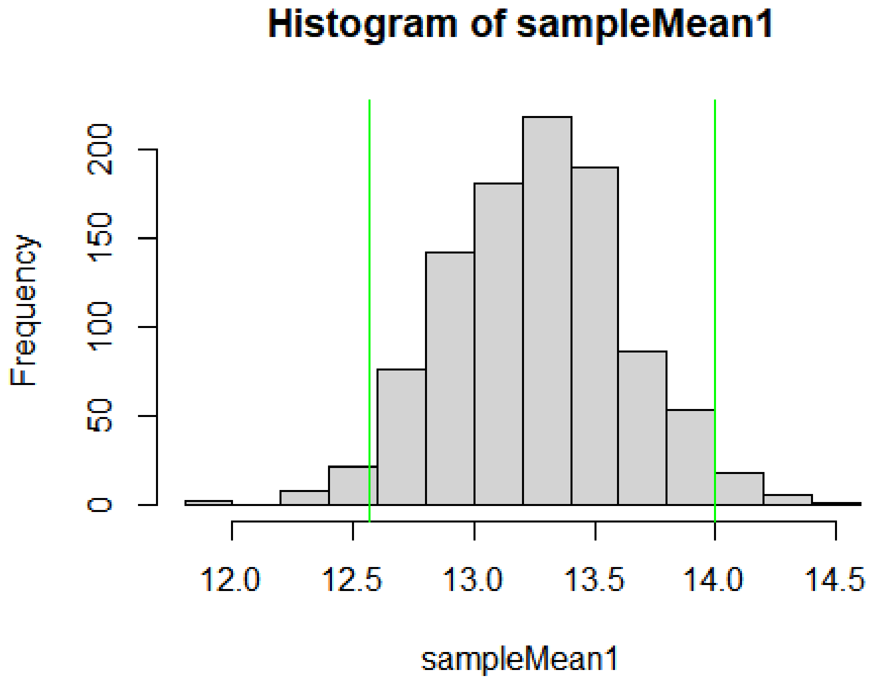
If you did Exercise 4, you calculated some quantiles for a distribution of raw data. If you did Exercise 5, you calculated some quantiles for a sampling distribution of means. Briefly describe, from a conceptual perspective and in your own words, what the difference is between a distribution of raw data 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 distribution for the raw data and the sampling gives very different
values. When a sample is taken, the chance of the sample having an even
dispersion around the median increases. Therefore the distribution of the
means of the samples generally tends to have lesser dispersion than the
original sample. The 2.5 to 97.5 percentile of data generally is the
threshold to determine whether a given sample is a part of the population. It
approximately covers two standard deviations from the median on either side.
With raw data, the dispersion could be very high, however, when a sample is
taken repeatedly, the distribution of their means gives a curve close to a
normal distribution and with dispersion lower than the raw data.
```

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 about a sample makes it "better." (1 pt)*

```
#Replicating the sampling function to find the means for sample size of 70
sampleMean1 <- replicate(1000, mean(sampleGirthValues(70)))
quantile(sampleMean1, probs=0.025)

##     2.5%
## 12.56821

quantile(sampleMean1, probs=0.975)

##    97.5%
## 13.99443

#Histogram of the mean of samples 1000 times.
hist(sampleMean1)
abline(v=quantile(sampleMean1, probs=0.025),col="green")
abline(v=quantile(sampleMean1, probs=0.975),col="green")
```



By looking at the scale it can be seen that the 2.5 and 97.5 percentile of the means of sample size 70 is 12.47 to 13.97. When a sample is taken multiple times, it tends to converge closer and closer to the median of the raw data. Hence, the dispersion was highest in the raw data. In the sample size of 7 it went down however, on increasing the sample size to 70 the standard deviation is expected to go down by a factor of 1/3 which happened.