IST 772 Homework 3

Maya Mileva

due date: Oct 24th, 2019

I did this homework by myself, with help from the book and the professor and the book "An introduction to data science" by J. Saltz and J. Stanton.

```
## Run these three functions to get a clean test of homework code
## dev.off() # Clear the graph window
cat('\014')  # Clear the console
```

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

Exercises

2. 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 hatch- ing. 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 col- umns 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) # 4 variables
```

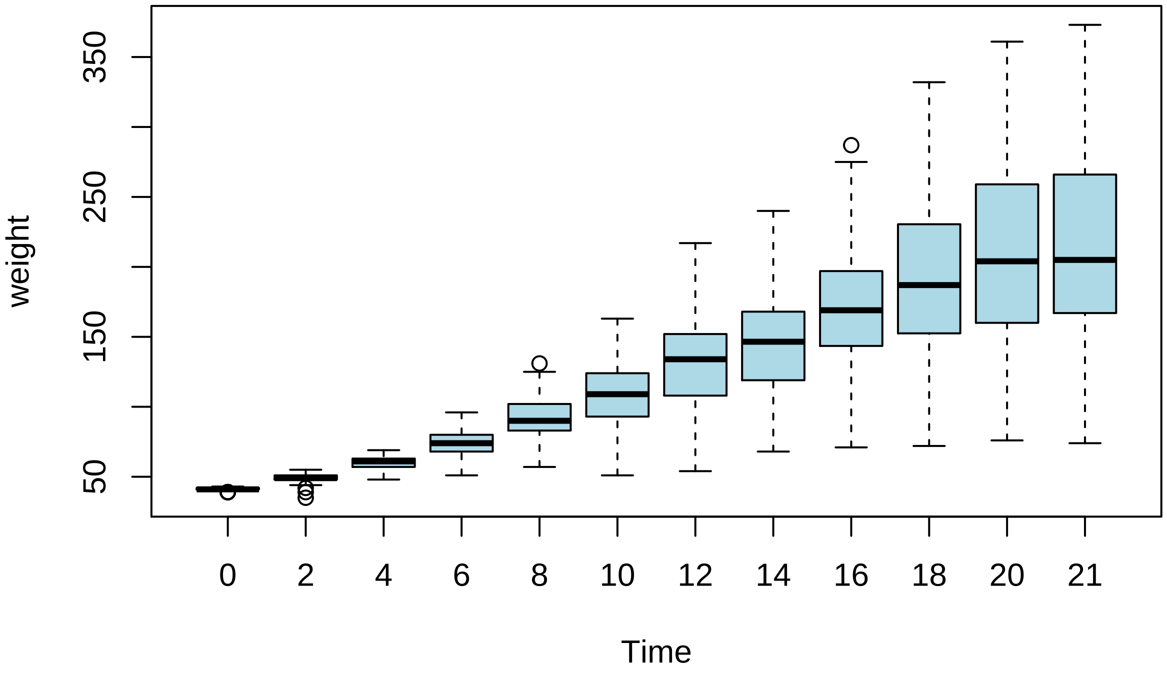
```
##      weight           Time           Chick     Diet   
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220  
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120  
##  Median :103.0   Median :10.00   20     : 12   3:120  
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118  
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12          
##  Max.   :373.0   Max.   :21.00   19     : 12          
##                                  (Other):506          
```

```
dim(ChickWeight) # shows the dimension of the data set(578 rows with 4 col)
```

```
## [1] 578   4
```

"ChickWeight" data set contains information on the weight of chicks in grams up to 21 days after hatch- ing. It contains 4 different variables: "weight", "Time", "Chick", "Diet" and total of 578 rows. The *dim* function of the R programming language returns the dimension (the number of columns and rows) of a matrix, array or data frame. "ChickWeight" consists of 578 rows and 4 columns.

```
# Graphical overview
boxplot(weight~Time,data = ChickWeight, xlab="Time", ylab="weight", col="lightblue")
```



3. When a data set contains more than one variable, R offers another subsetting operator, dol sign, 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 dol sign 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)

```
summary(ChickWeight$weight) # report data models and data overview
```

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

summary() reveals some overlapping information to the str command about the weight variable(numeric) in the data set.

The minimun weight(lowest value) is 35g and the max is 373g - highest value of the weight(we can say that the range varies from 35 to 373).

1st Qu. refers to the dividing line at the top of the first quartile. If we look at all the weights and line them up in order, we can divide up the whole into 4 groups, where each group had the same number of observations(smallest on the left and largest on the right). So the 1st Qu is the value of the weight that divides the first quarter of the cases from the other three quarters.

Median refers to the value of the weight that splits the whole weight group in half(half having highter values and half - lower). We can alse state that the median is the line that separates the second and third Qu.

Mean shows the average weight - 121.8.

3rd Qu is the third quartile. It represents the third final dividing line that splits all the cases into 4 equal parts. Quartiles give a sense of the shape of the distrubution, can be used for comparisons too(if we want to know if a sample is drawn by specific data set).

```
head(ChickWeight$weight)
```

```
## [1] 42 51 59 64 76 93
```

head() returns the first 6 values of the column weight. Give us an idea what it is in the column.

```
mean(ChickWeight$weight)
```

```
## [1] 121.8183
```

Mean is the numeric average of all the values in that column. 121.8g is the averagne weight of the chicken.

```
myChkWts <- ChickWeight$weight # saving into new variable containing chicken weightd only
```

Creates a copy of the weight data from the ChickWeight data set and put it in a new vector called myChkWts. Its useful for plotting and fruther analysis.

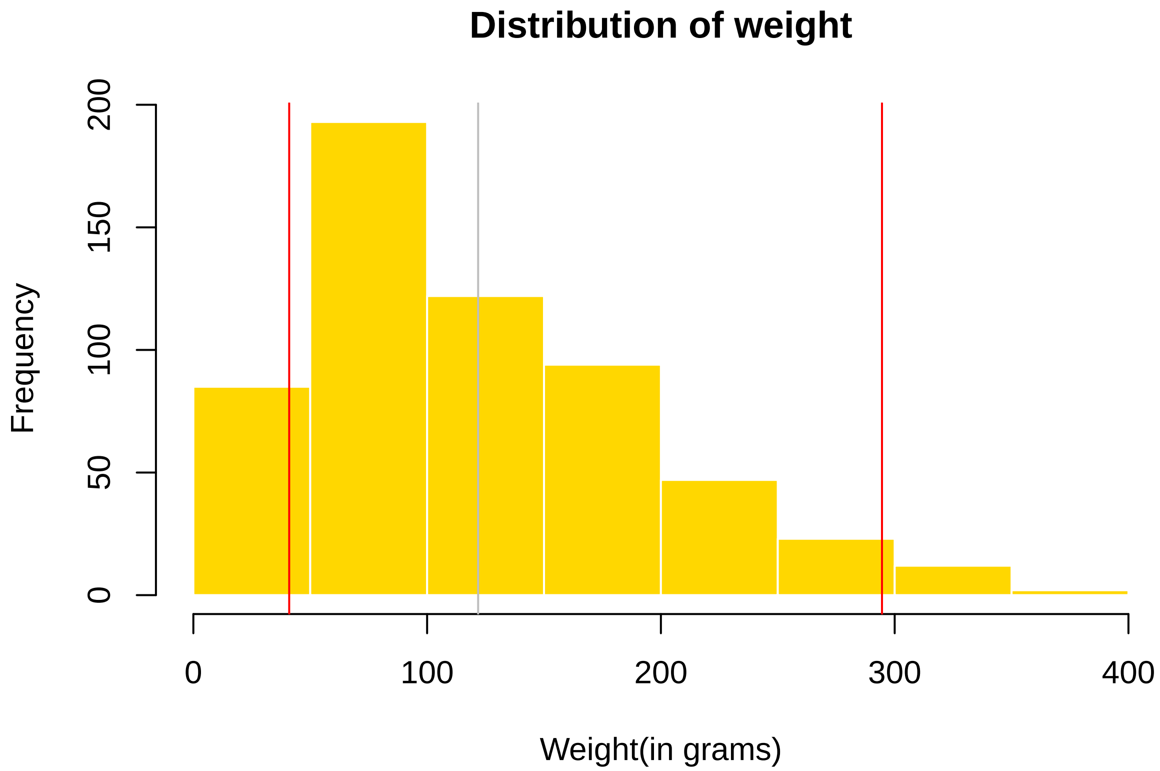
```
quantile(myChkWts,0.50)
```

```
## 50%
## 103
```

Quantiles mark a position within a collection of values; devide things up into arbitrary number of devisions. The first argumet(myChkWts) is the vector, the second argument(0.5) provides a cut-point for R to use over the provided data. In this case .50 is the median(103), which is really close/same to the median, produced by summary().

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. Cre- ate 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.

```
hist(myChkWts, col = "gold", border = "white", main = "Distribution of weight"
     , xlab = "Weight(in grams)") # create a hist for weight variable
abline(v = quantile(myChkWts, 0.025), col = 'red') # lower bound of central region
abline(v = quantile(myChkWts, 0.975), col = 'red')# upper bound of central region
m <- mean(myChkWts)
abline(v = m, col = 'gray')
```



```
summary(myChkWts)
```

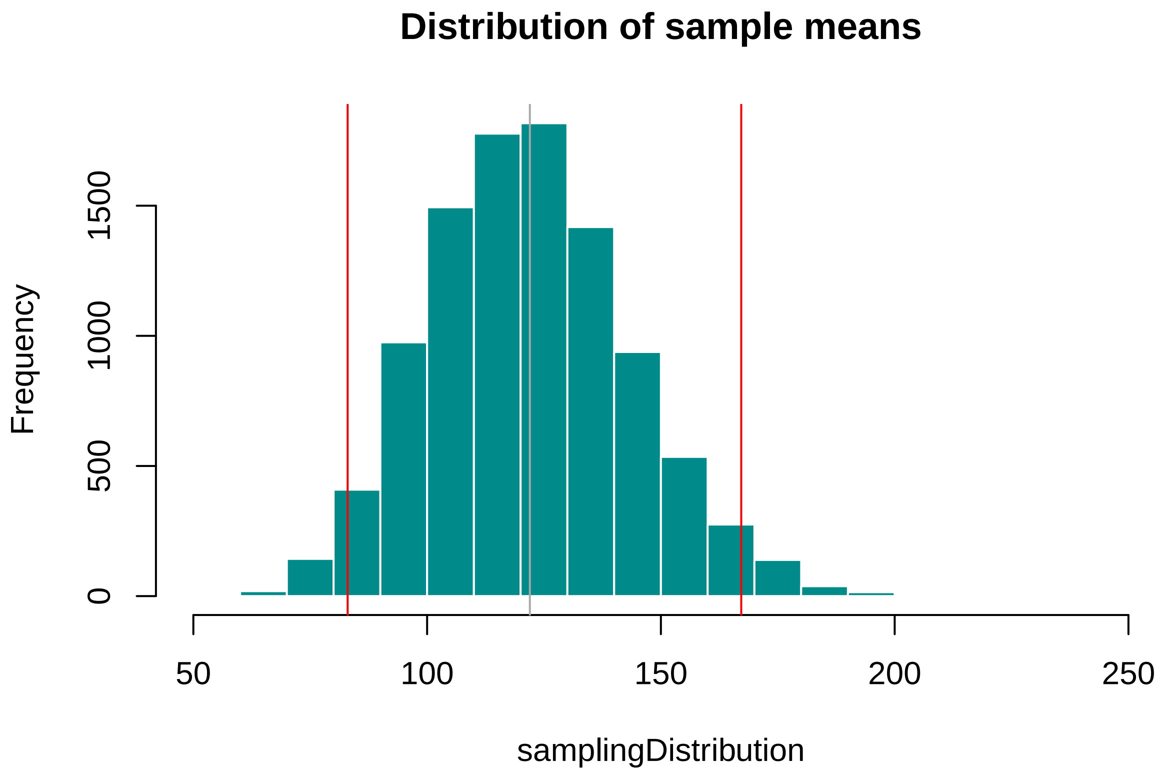
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     35.0    63.0   103.0   121.8   163.8   373.0
```

The weight variable describes the weight of the chickens in grams. It has poisson dristribution. The mean, which represents the average wight is 121.8g. Like we said before the range varies deom 35g to 373g, which represent the minumum and the maximum value. 1st Qu is signifies the 25% of all the values in the variable smaller than 63. 3rd Qu represent 25% bigger that 163.8g.

When we displayed 2.5% and 97.5% quantiles of the distribution of that variables we drew line in the histogram to display them. "Qant"(meaning number), is cutting somthing into number of pieces. The ablines, defined my those 2 quantiles(2.5% and 97.5%)designate the 95% cetral region, with lower(0.025) and upper(0.975) tails. The 50% quantile is the median. 95% of the Area under the curve is the "central region". The tails contains extreme events(5%). It is important to know about the extreme examples because they are quite unlikelly and not typical to be observed.

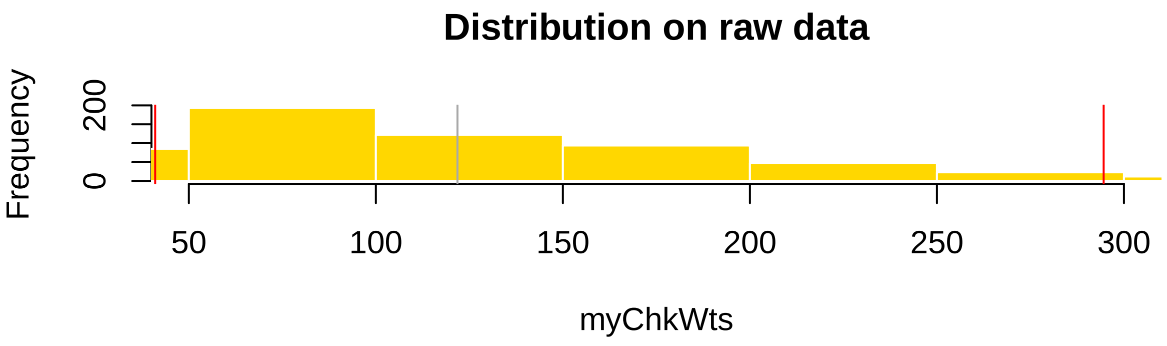
5. Write R code that will construct a sampling distribution of means from the weight data (as noted above, if you did exercise 3 you can use myChkWts instead of ChickWeight$weight to save yourself some typing). Make sure that the sampling distribution contains at least 1,000 means. Store the sampling 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 50 REASONING WITH DATA 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.

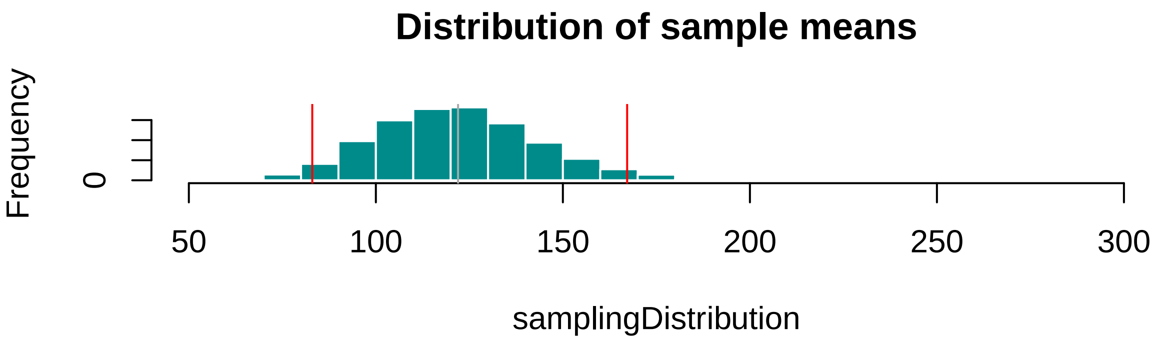
```r
## Run 10,000 replications(create a vector of 10,000 means)
## and place in in storage location  called "samplingDistribution"
samplingDistribution <- replicate(10000,mean(sample(myChkWts,
                                                    size = 11,
                                                    replace = TRUE)),
                                  simplify = TRUE)

hist(samplingDistribution   # hist of the distribution of the sample means
     , main = "Distribution of sample means", border = "white", col = "cyan4")

# lower bound of central region
abline(v = quantile(samplingDistribution, 0.025), col = 'red')

# upper bound of central region
abline(v = quantile(samplingDistribution, 0.975), col = 'red')
m1 <- mean(samplingDistribution)
abline(v = m1, col = 'darkgray')
```



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 differ- ence 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.

```r
par(mfrow=c(2,1))
hist(myChkWts, xlim = c(50,300)
     , col = "gold"
     , main = "Distribution on raw data"
     ,border = "white")

# lower bound of central region
abline(v = quantile(myChkWts, 0.025), col = 'red')

# upper bound of central region
abline(v = quantile(myChkWts, 0.975), col = 'red')

m <- mean(myChkWts)
abline(v = m, col = 'darkgray')

hist(samplingDistribution, xlim = c(50,300)
     , col = "cyan4", main = "Distribution of sample means"
     , border = "white")
abline(v = quantile(samplingDistribution, 0.025)
       , col = 'red') # lower bound of central region
abline(v = quantile(samplingDistribution, 0.975)
       , col = 'red')# upper bound of central region
m1 <- mean(samplingDistribution)
abline(v = m1, col = 'darkgray')
```





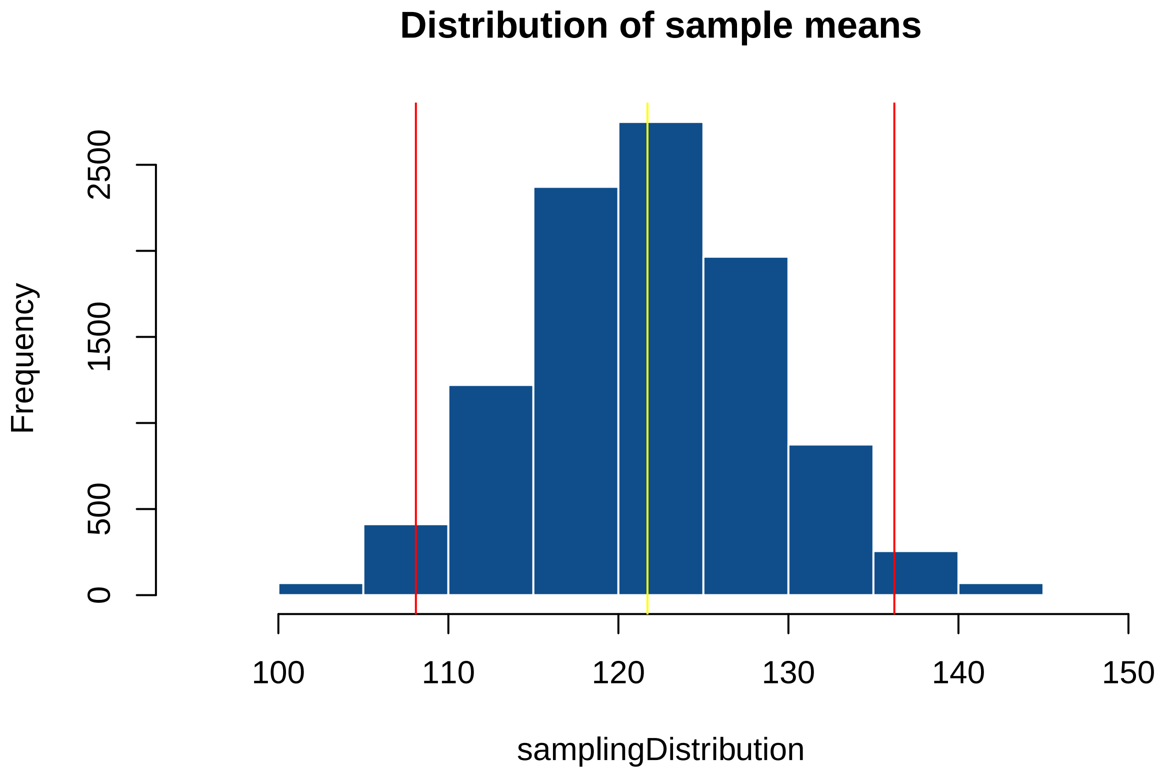
```r
par(mfrow = c(1,1))
```

From the hist we can see that the mean of the raw data and the mean of the sample means are so similar to one another. We can conclude that although any of one sample of 11 means will probably not have a mean that is exacly the same(121.8), and few samples will have means that are higher or lower, over the long run, the mean of the sampling distribution matches the mean of the original raw data. The low of the large number(if you rin a stat. process like sampling a large number of times, it will generally converge on particular result) and the central limit theorem are partially demonstrated. When we take both of this laws into account , we find that a distribution of sampling means starts to create a bell-shaped(normal)distribution over time, with mean values close to the population mean. That si because, when there is a big enought sample, repeated many times the high and low value tend to cancell themselve and the bell-shape curve starts forming. Also the smapling error get smaller with the bigger size.

Like we said the 2.5% and 97.5% represent the extreme values. They are different in both histrograms. With the distribution on the sample mean the extreme value cancell themselves, there is small variation. If we think about what sample is: drawing elements from population with all kinds of values. So the raw data has higher variation, when drawing sample means we cancell the high and the lows and that get even better overtime.

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."

```r
## Run 10,000 replications(create a vector of 10,000 means)
## and place in in storage location  called "samplingDistribution"
samplingDistribution <- replicate(10000,mean(sample(myChkWts, size = 100
                                                    ,replace = TRUE))
                                                    ,simplify = TRUE)

hist(samplingDistribution   # hist of the distribution of the sample means
     , main = "Distribution of sample means"
     , border = "white"
     , col = "dodgerblue4")

abline(v = quantile(samplingDistribution, 0.025)
       , col = 'red') # lower bound of central region
abline(v = quantile(samplingDistribution, 0.975)
       , col = 'red')# upper bound of central region
m1 <- mean(samplingDistribution)
abline(v = m1, col = 'yellow')
```



Bigger sample size give smaller error, less variation and we if run sampling a large number of time we get a stable, consistent pattern of results. Also when we were looking at the sample means, an take the law of the large numbers into account, we found that the distribution of sampling means starts to create a bell-shaped or normal distribution, and the center of that distribution(the mean of all those means) gets really close to the actual population mean. It gets closer faster for larger samples; for smaller samples we have to draw lots of them to get relly close. We can conclude that “better” sample is determined by confidence levels and margins of error, power and effect sizes. We already explained in exercise #6 why the extreme values change with the sample size.