**Quantitative Methods**

**List of Exercises N. 4**

**Selected Exercises from McClave (2014) – Chapters 5 and 6**

* 1. **The Concept of a Sampling Distribution**

**Exercise 1. (3). Consider the population described by the probability distribution shown below:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **X** | **1** | **2** | **3** | **4** | **5** |
| **p(x)** | **,2** | **0,3** | **0,2** | **0,2** | **0,1** |

**The random variable x is observed twice. If these observations are independent, verify that the different samples of size 2 and their probabilities are as shown in the next column.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample** | **Probability** | **Sample** | **Probability** |
| **1 , 1** | **0,04** | **3 ,4** | **0,04** |
| **1 , 2** | **0,06** | **3, 5** | **0,02** |
| **1 , 3** | **0,04** | **4, 1** | **0,04** |
| **1 , 4** | **0,04** | **4, 2** | **0,06** |
| **1 , 5** | **0,02** | **4, 3** | **0,04** |
| **2 , 1** | **0,06** | **4, 4** | **0,04** |
| **2 , 2** | **0,09** | **4, 5** | **0,02** |
| **2 , 3** | **0,06** | **5, 1** | **0,02** |
| **2 , 4** | **0,06** | **5, 2** | **0,03** |
| **2 , 5** | **0,03** | **5, 3** | **0,02** |
| **3 , 1** | **0,04** | **5, 4** | **0,02** |
| **3 , 1** | **0,06** | **5, 5** | **0,01** |
| **3 ,3** | **0,04** |  |  |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L4E1\_DATA1 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E1\_DATA1.xlsx")

View(L4E1\_DATA1)

attach(L4E1\_DATA1)

L4E1\_DATA2 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E1\_DATA2.xlsx")

View(L4E1\_DATA2)

attach(L4E1\_DATA2)

Create variables: x <- x

px <- px

Sample <- Sample

Probability <- Probability

1. **Find the sampling distribution of the sample mean .**

To find the means of each sample and the probability of them, we have to calculate. The probability are found by multiplying the p(x) of each value in the probability distribution table. The means are calculated by adding both values and dividing by n (n is equal to 2).

In this exercise, you can calculate it by hand in word, excel or in R. We will show one of the calculations, and then the table:

|  |  |
| --- | --- |
| x | px |
| 1 | 0.4 |

Sample 1,1

S1\_px <- (1+1)/2 = 1

S1\_x <- 0.2+0.2 = 0.4

Then you just need to do this for each of the samples:

|  |  |  |
| --- | --- | --- |
| Sample |  |  |
| 1, 1 | 1 | 0,04 |
| 1, 2 | 1,5 | 0,06 |
| 1, 3 | 2 | 0,04 |
| 1, 4 | 2,5 | 0,04 |
| 1, 5 | 3 | 0,02 |
| 2, 1 | 1,5 | 0,06 |
| 2, 2 | 2 | 0,09 |
| 2, 3 | 2,5 | 0,06 |
| 2, 4 | 3 | 0,06 |
| 2, 5 | 3,5 | 0,03 |
| 3, 1 | 2 | 0,04 |
| 3, 2 | 2,5 | 0,06 |
| 3, 3 | 3 | 0,04 |
| 3, 4 | 3,5 | 0,04 |
| 3, 5 | 4 | 0,02 |
| 4, 1 | 2,5 | 0,04 |
| 4, 2 | 3 | 0,06 |
| 4, 3 | 3,5 | 0,04 |
| 4, 4 | 4 | 0,04 |
| 4, 5 | 4,5 | 0,02 |
| 5, 1 | 3 | 0,02 |
| 5, 2 | 3,5 | 0,03 |
| 5, 3 | 4 | 0,02 |
| 5, 4 | 4,5 | 0,02 |
| 5, 5 | 5 | 0,01 |

The last step is now to sum this table. Sum up the table so you only have the values you need in a probability distribution table to answer the question.

Remember to add the probabilities so you have the total probability of a given mean. For example the mean of 1.5 appears twice as one can both have x=1 with x=2, and x=2 with x=1, meaning that a sample of 1,2 and 2,1 is the same. The mean of 1.5 thus appears twice, and you must add its probability of 0.06 + 0.06 to get =0.12, as is also seen in the table below.

|  |  |
| --- | --- |
|  |  |
| 1 | 0,04 |
| 1,5 | 0,12 |
| 2 | 0,17 |
| 2,5 | 0,2 |
| 3 | 0,2 |
| 3,5 | 0,14 |
| 4 | 0,08 |
| 4,5 | 0,04 |
| 5 | 0,01 |

We make sure, that the table above is constructed as a new dataset in R:

x\_bar <- c(1,1.5,2,2.5,3,3.5,4,4.5,5)

px\_bar <- c(.04,.12,.17,.20,.20,.14,.08,.04,.01)

DATA1A <- cbind(x\_bar,px\_bar)

Or we construct the table in excel, and import the data:

library(readxl)

L4E1A <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E1A.xlsx")

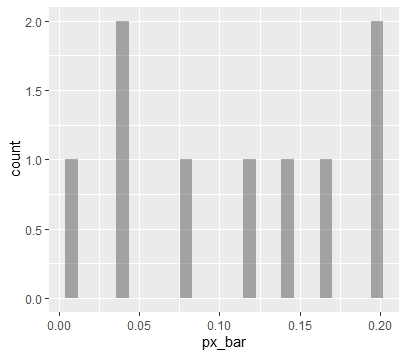
View(L4E1A)

attach(L4E1A)

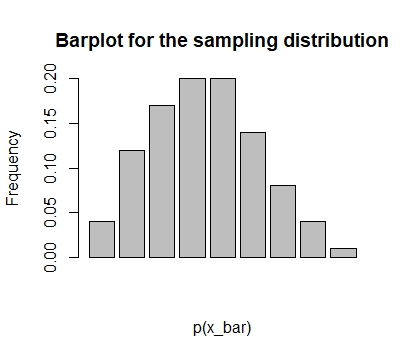
Now we can work with our new dataset.

1. **Construct a probability histogram for the sampling distribution .**

If we try to make a histogram of this data it will show, that the data is on the left side. This is because it will understand, that the means and everything below 1, such as 0.04 or 0.12 and so on will be binned into that category. That is not what we want.

gf\_histogram( ~ px\_bar, data = L4E1A)

As we can see, our Histogram is not looking correctly. We should therefore consider, if we could use another method.

Try to consider what the data represents. You want to see the distribution of your probabilities of your means. You have your means and your probabilities already in a table. So an x-y graph is fine for the purpose of seeing the distribution.

barplot(px\_bar, xlab="p(x\_bar)",ylab="Frequency", col="grey", main="Barplot for the sampling distribution", border="black")

1. **What is the probability that is 4,5 or larger?**

If we look at the table with and , we need to find the values of 4.5 and higher.

There is two values:

and

When then and then .

To find the probability of 4.5 or larger, we just need to make a simple calculation:

1. **Would you expect to observe a value of . Equal to 4.5 or larger? Explain.**

No. The probability of observing or larger is small (0.05).

**Exercise 2. (5). Refer to exercise 1. Assume that a random sample of n = 2 measurements is randomly selected from the population.**

1. **List the different values that the sample median m may assume and find the probability of each. Then give the sampling distribution of the sample median.**

The median and the mean are the same for each possible sample. Thus, the probability of the median would be the same as the probability of the mean.

1. **Construct a probability histogram for the sampling distribution of the sample median and compare it with the probability histogram for the sample mean (Exercise 1, question b).**

The probability histogram for the sample median is identical to that for the sample mean. See Exercise 1.

* 1. **Properties of Sampling Distributions: Unbiasedness and Minimum Variance**

**Exercise 3. (9). Consider the following probability distribution:**

|  |  |  |  |
| --- | --- | --- | --- |
| x | 2 | 4 | 9 |
| p(x) |  |  |  |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L4E3 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E3.xlsx")

View(L4E3)

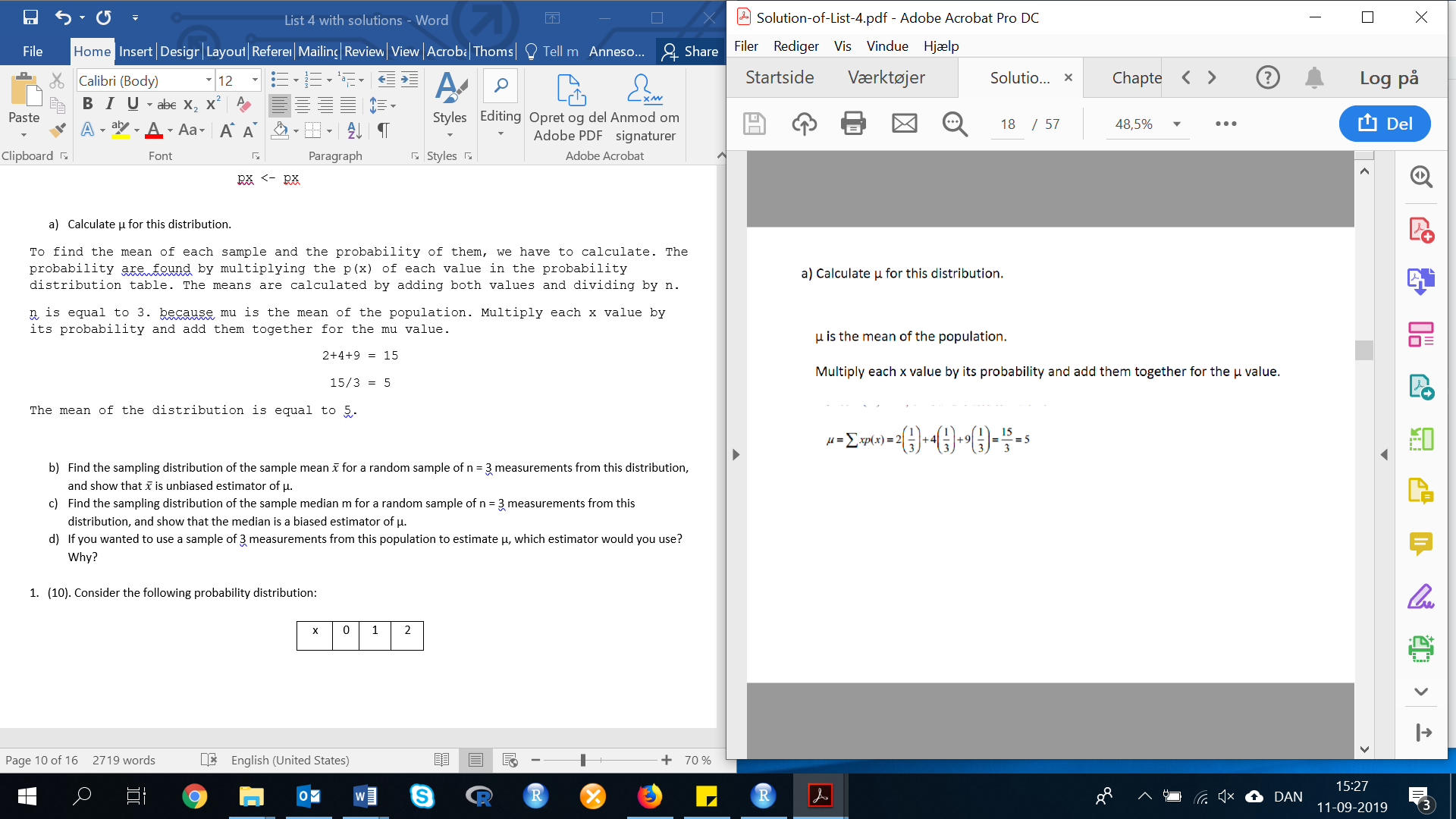
attach(L4E3)

Create variables: x <- x

px <- px

1. **Calculate μ for this distribution.**

To find the mean of each sample and the probability of them, we have to calculate. The probability are found by multiplying the p(x) of each value in the probability distribution table. The means are calculated by adding both values and dividing by n.

n is equal to 3. because mu is the mean of the population. Multiply each x value by its probability and add them together for the mu value.

2+4+9 = 15

15/3 = 5

The mean of the distribution is equal to 5.

1. **Find the sampling distribution of the sample mean for a random sample of n = 3 measurements from this distribution, and show that is unbiased estimator of μ.**

(not the population average, µ is the population average).

To show that is unbiased estimator of µ you must show that the average of your sample, where n=3, is equal to µ which is 5.

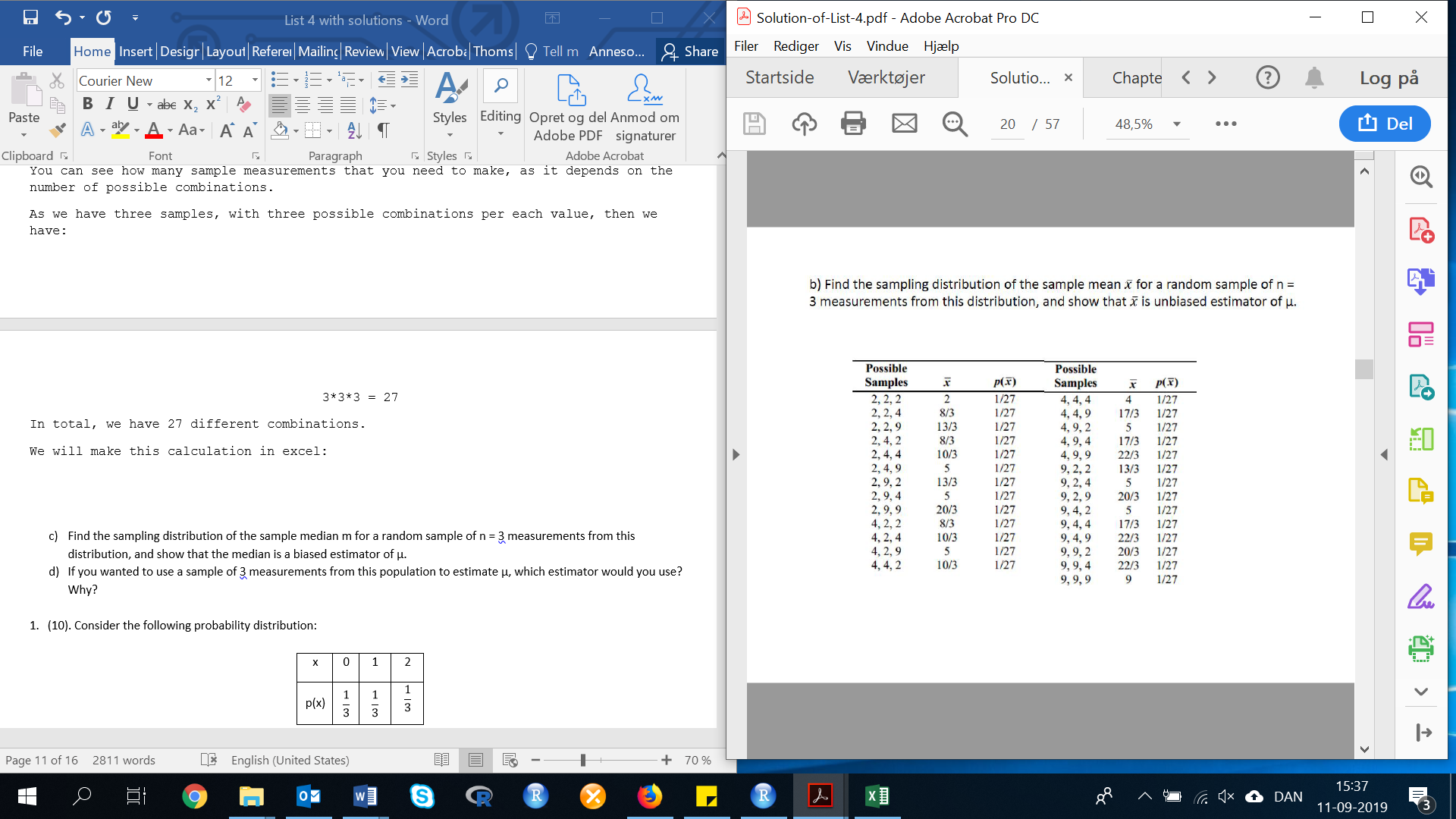
You can see how many sample measurements that you need to make, as it depends on the number of possible combinations.

As we have three samples, with three possible combinations per each value, then we have:

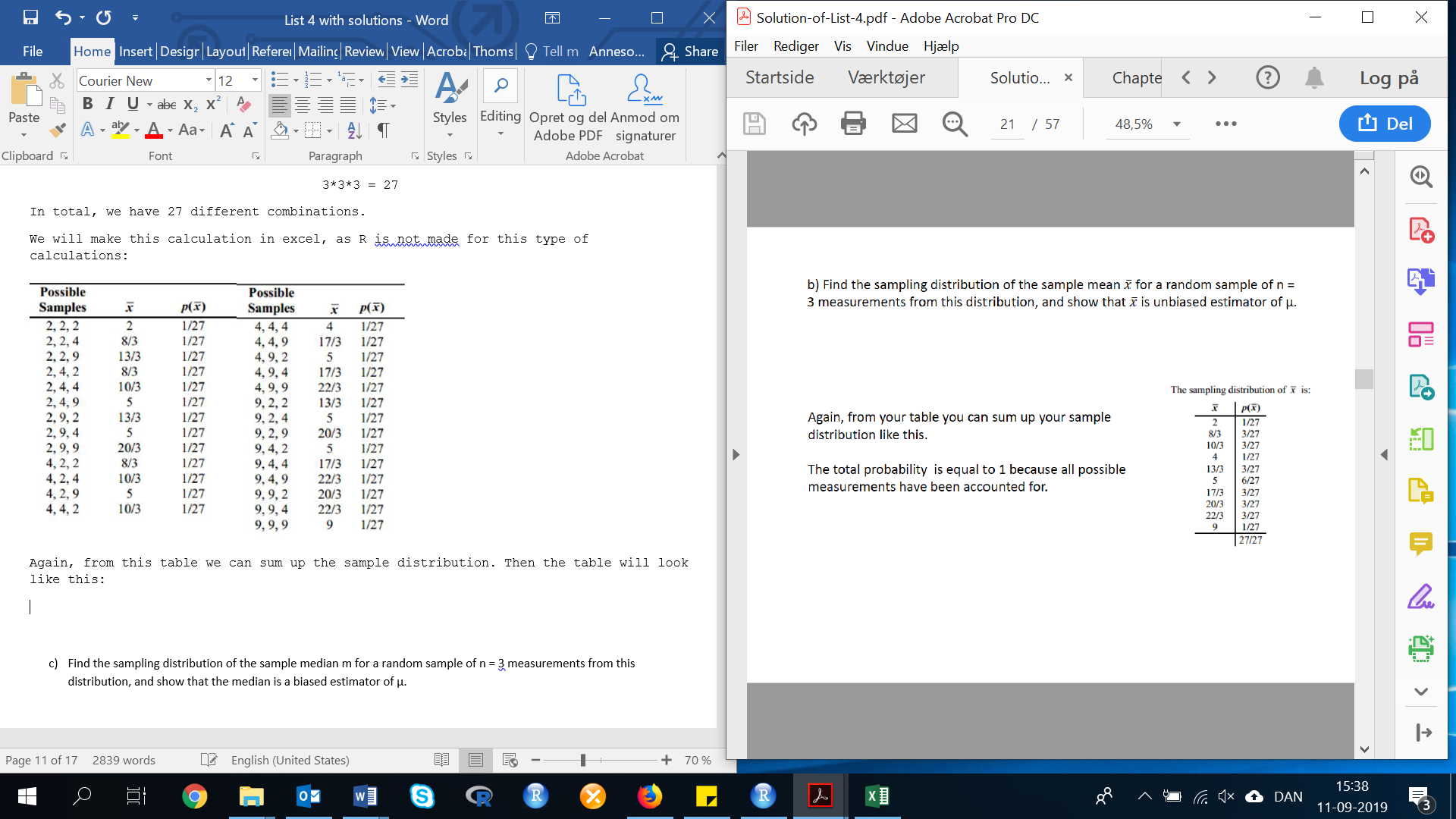
3\*3\*3 = 27

In total, we have 27 different combinations.

We will make this calculation in excel, as R is not made for this type of calculations:



Again, from this table we can sum up the sample distribution. Then the table will look like this:



The total probability is equal to 1, because all possible measurements have been accounted for.

In R: ((2)\*(1/27))+((8/3)\*(3/27))+((10/3)\*(3/27))+((4)\*(1/27))+((13/3)\*(3/27))+((5)\*(6/27))+((17/3)\*(3/27))+((20/3)\*(3/27))+((22/3)\*(3/27))+(9\*(1/27))

Result: 5

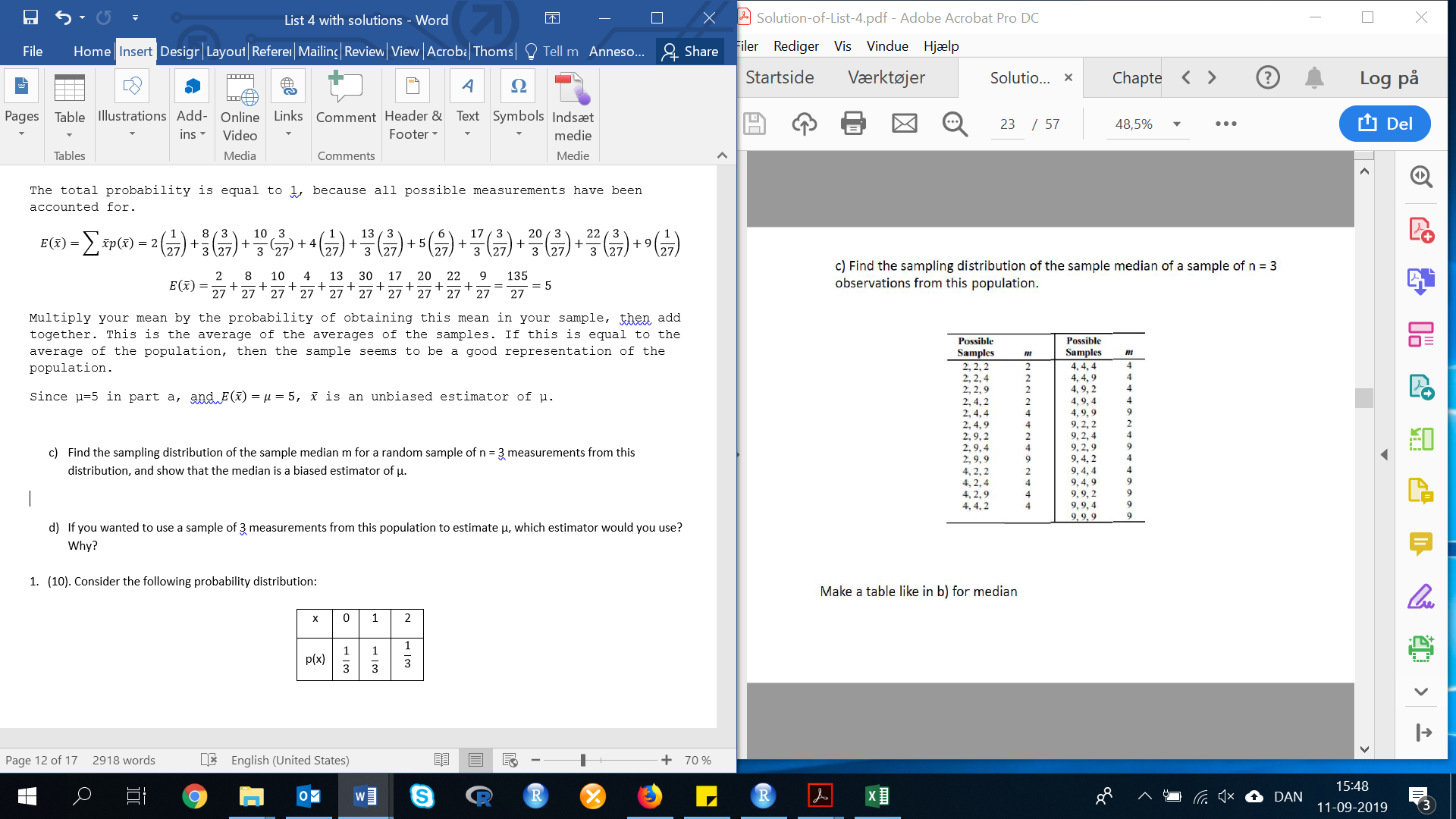
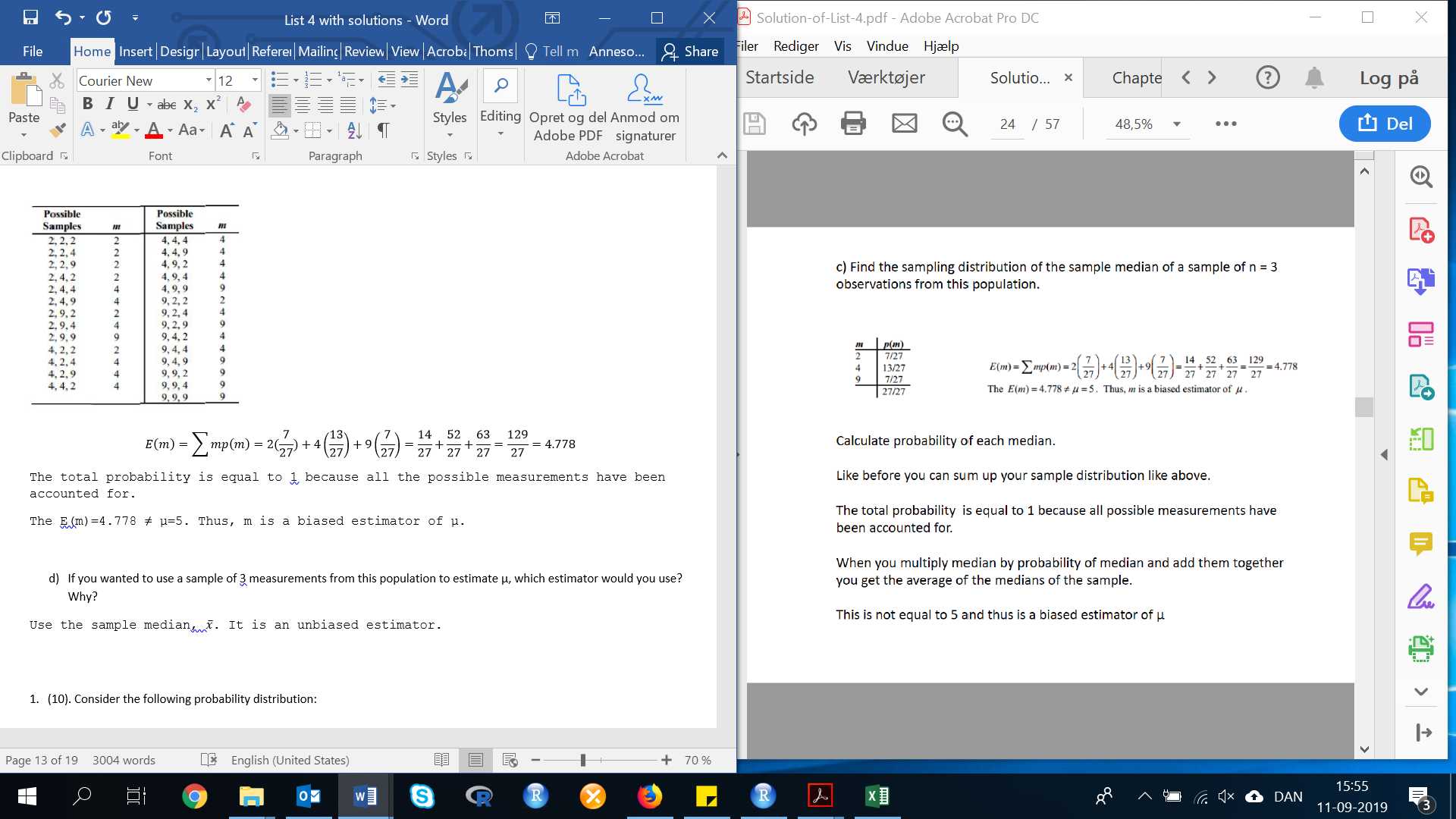
Multiply your mean by the probability of obtaining this mean in your sample, then add together. This is the average of the averages of the samples. If this is equal to the average of the population, then the sample seems to be a good representation of the population.

Since µ=5 in part a, and , is an unbiased estimator of µ.

1. **Find the sampling distribution of the sample median m for a random sample of n = 3 measurements from this distribution, and show that the median is a biased estimator of μ.**

In this exercise, we should make a table with the median for each sample. Below is a picture of how the table should look like.

The median is calculated for each sample and is shown in the table in part b. The sampling distribution of m is:



In R:

(2\*(7/27))+(4\*(13/27))+(9\*(7/27))

Result: 4.777778

The total probability is equal to 1 because all the possible measurements have been accounted for. When you multiply median by probability of median and add them together you get the average of the medians of the sample.

The E(m)=4.778 ≠ µ=5. Thus, m is a biased estimator of µ.

1. **If you wanted to use a sample of 3 measurements from this population to estimate μ, which estimator would you use? Why?**

Use the sample median, . It is an unbiased estimator.

**Exercise 4. (10). Consider the following probability distribution:**

|  |  |  |  |
| --- | --- | --- | --- |
| x | 0 | 1 | 2 |
| p(x) |  |  |  |

1. **Find μ.**

Remember as in Exercise 3, mu is the mean of the population. Therefore, we need to multiply each x value by its probability and add them together for the mu value.

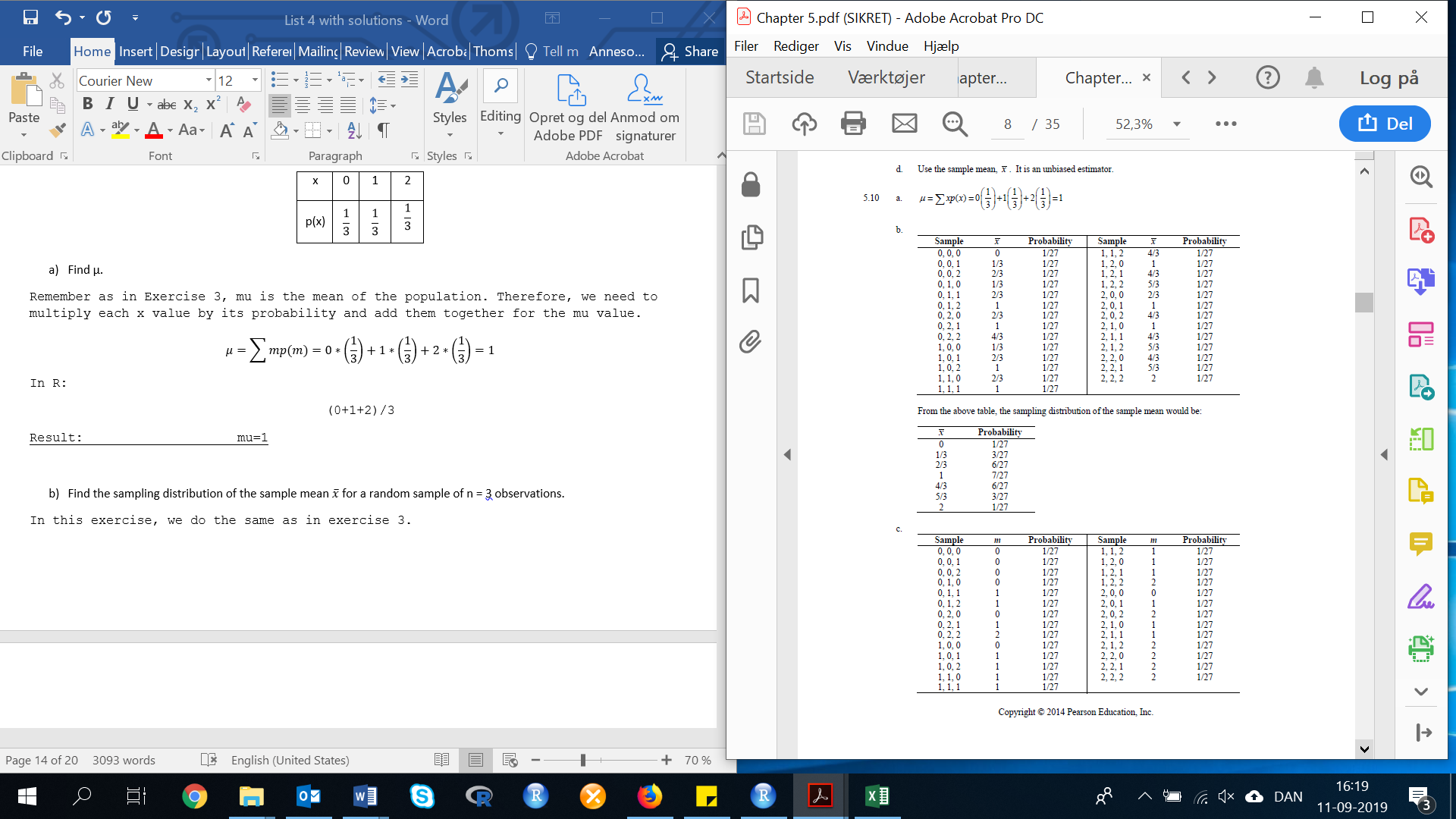
In R:

(0+1+2)/3

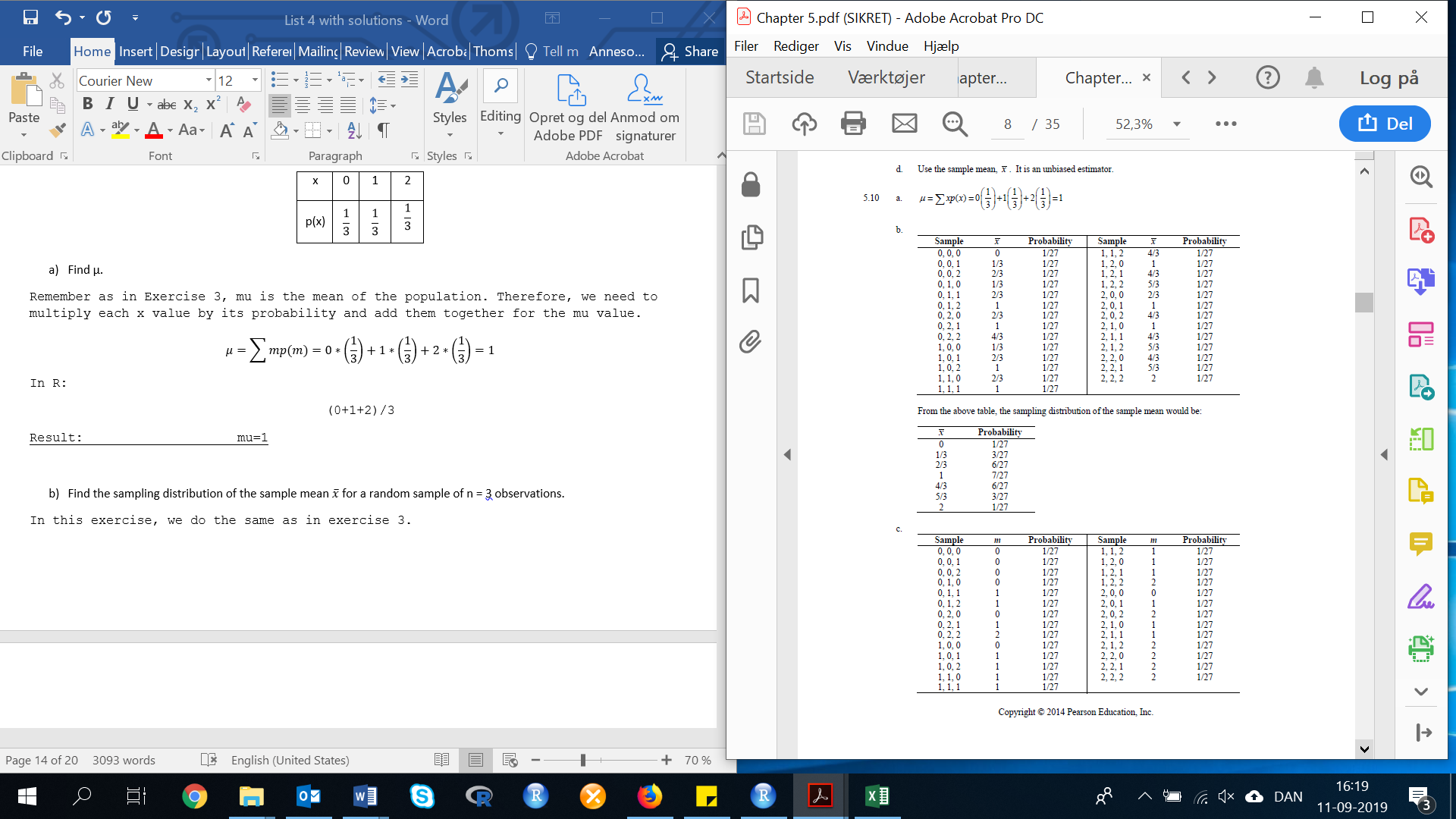
Result: µ=1

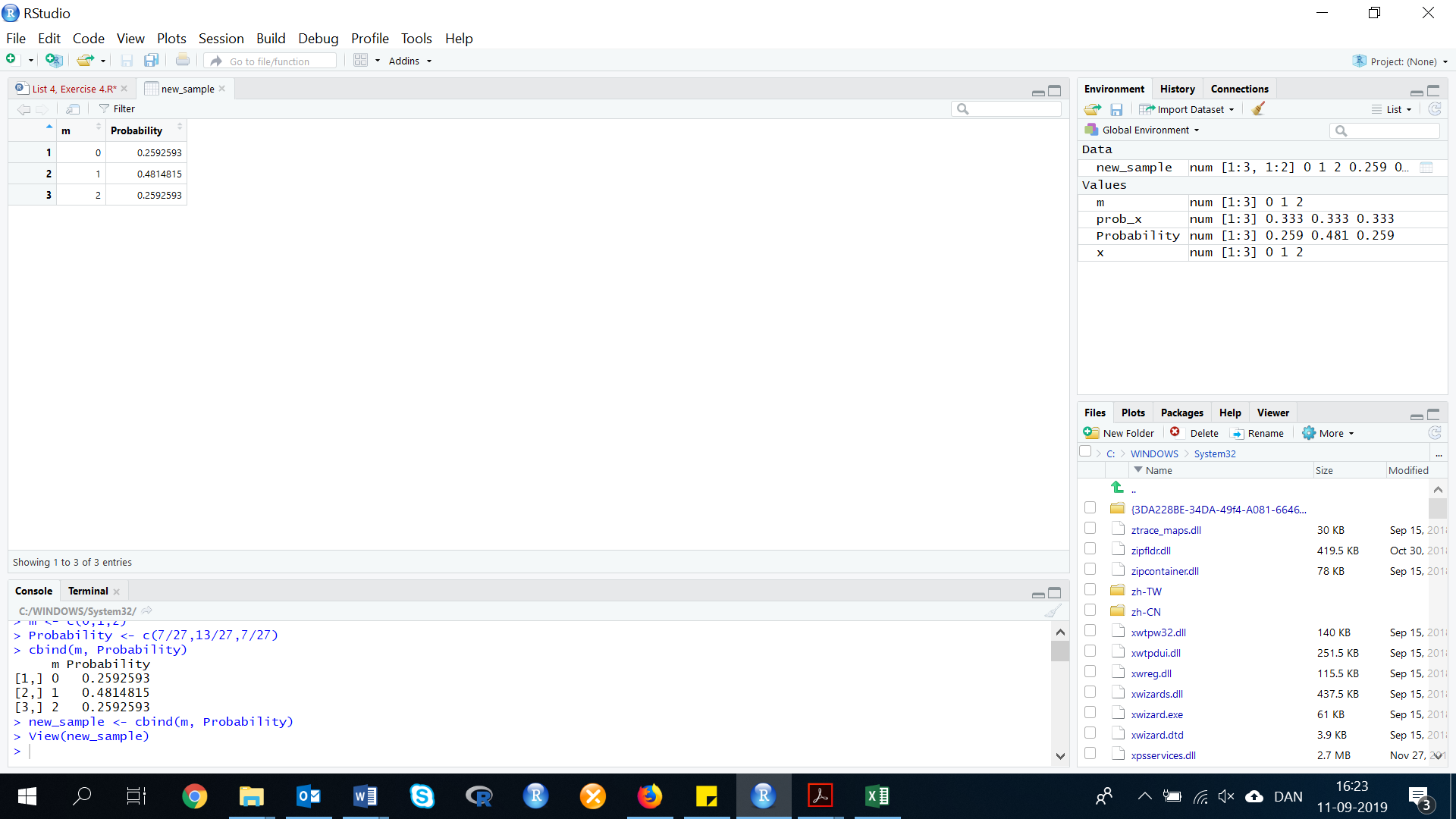
1. **Find the sampling distribution of the sample mean for a random sample of n = 3 observations.**

In this exercise, we do the same as in exercise 3.



1. **Find the sampling distribution of the sample median of a sample of n = 3 observations from this population.**



From the table, the sampling distribution of the sample median would be:

m <- c(0,1,2)

Probability <- c(7/27,13/27,7/27)

new\_sample <- cbind(m, Probability)

1. Refer to parts b and c, and show that both the mean and the median are unbiased estimators of μ for this population.

Since , is an unbiased estimator for µ.

Since , m is an unbiased estimator for µ.

In R:

Ex <- (0\*(1/27))+((1/3)\*(3/27))+((2/3)\*(6/27))+  
((1)\*(7/27))+((4/3)\*(6/27))+((5/3)\*(3/27))

Result: 0.9259259

EM <- (0\*(7/27))+(1\*(13/27))+(2\*(7/27))

Result: 1

1. **Find the variances of the sampling distributions of the sample mean and the sample median.**

In R:

Q2X <- ((0-1)^2)\*(1/27)+((((1/3)-1)^2)\*(3/27))+((((2/3)-1)^2)\*(6/27))+(((1-1)^2)\*(7/27))+((((4/3)-1)^2)\*(6/27))+((((5/3)-1)^2)\*(3/27))+(((2-1)^2)\*(1/27))

Result: 0.2222222

Q2M <- (((0-1)^2)\*(7/27))+(((1-1)^2)\*(13/27))+(((2-1)^2)\*(7/27))

Result: 0.5185185

1. **Which estimator would you use to estimate μ. Why?**

Since both the sample mean and median are unbiased estimators and the variance is smaller for the sample mean, the sample mean would be the preferred estimator of µ.

When the variance is smaller the number is more precise. Thus it is better to use the mean as an estimator. In real life, usually the mean is a better estimator of µ than the median is.

**5.3. The sampling distribution of the sample mean and the Central limit theorem**

**Exercise 5. (33, PHISH). *Phishing attacks to e-mail accounts*. *Phishing* describes an attempt to extract personal/financial information from unsuspecting people through fraudulent e-mail. Data from an actual phishing attack against an organization were presented in *Chance* (Summer 2007). The interarrival times, i.e., the time differences (in seconds), for 267 fraud box e-mail notifications, were recorded and are saved in the accompanying file. For this exercise, consider these interarrival times to represent the population of interest.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L4E5 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E5.xls")

View(L4E5)

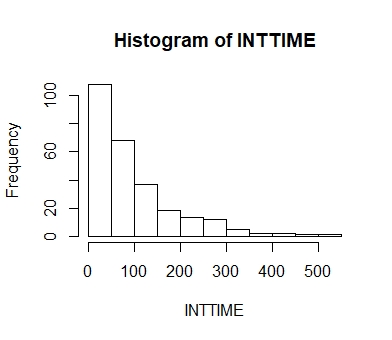
attach(L4E5)

Create variables: INTTIME <- INTTIME

**a) Construct a histogram for the interarrival times. Describe the shape of the population of interarrival times.**

We need to make a histogram:

hist(VARIABLE)

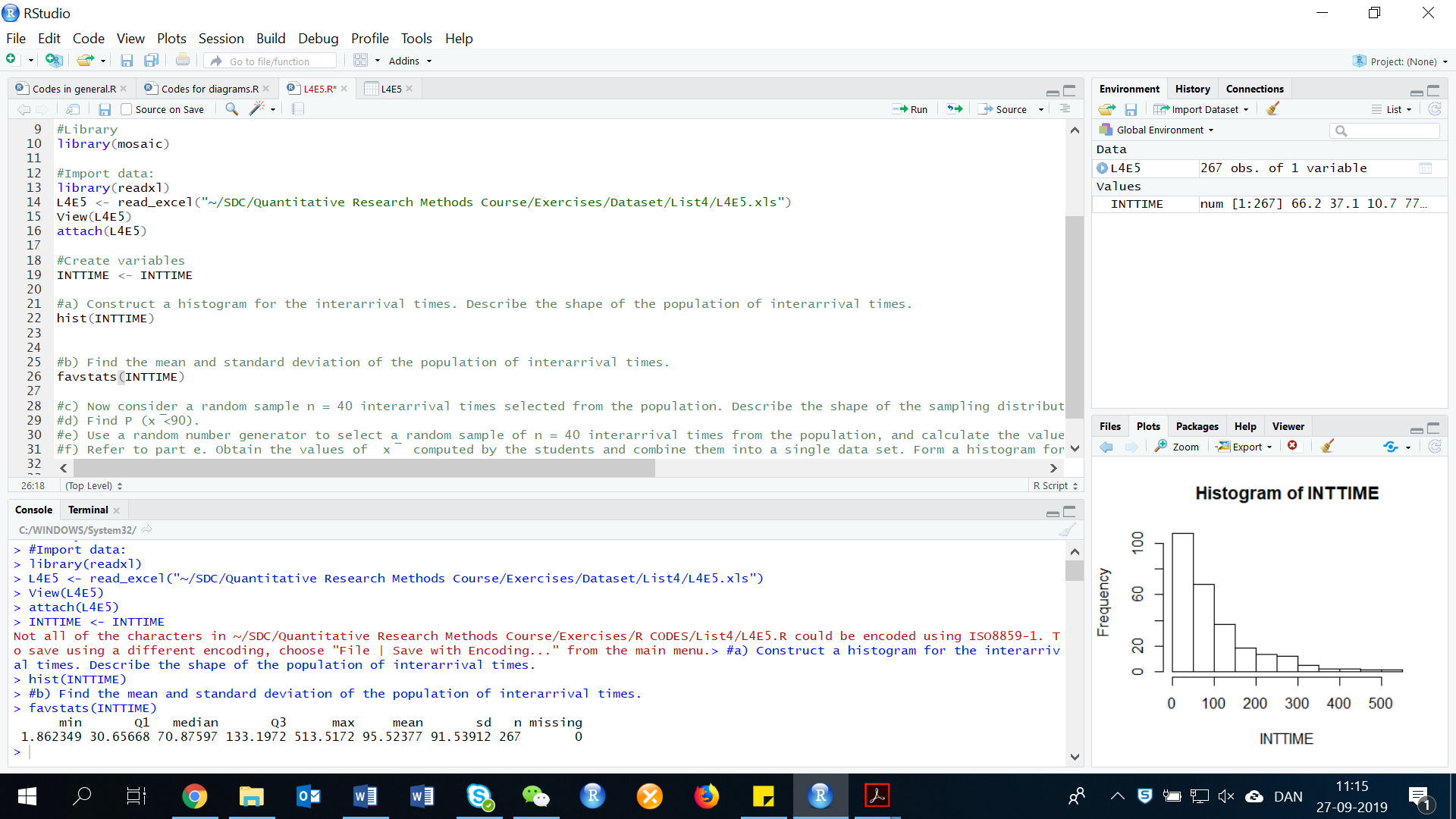
hist(INTTIME)

The population of interarrival times is skewed to the right.

**b) Find the mean and standard deviation of the population of interarrival times.**

In this exercise, we need to find the mean and standard deviation. We can use many different codes to do this, but we will use the favstats() function in this exercise:

favstats(INTTIME)



We can see that:

Mean=95.52377

Sd=91.53912

**c) Now consider a random sample *n =* 40 interarrival times selected from the population. Describe the shape of the sampling distribution of , the sample mean. Theoretically, what are and ?**

By the Central Limit Theorem, the sampling distribution of will be approximately normal. Theoretically:

In R:

(91.53912/(sqrt(40)))=14.47361

**d) Find *P* ().**

To find the probability, we do as followed:

If we look in the table, we can se that the tail area corresponding to the probability is:

**e) Use a random number generator to select a random sample of *n* = 40 interarrival times from the population, and calculate the value of .**

**f) Refer to part e. Obtain the values of computed by the students and combine them into a single data set. Form a histogram for these values of . Is the shape approximately normal?**

This will answer both e) and f).

NOTE: The answers here will vary.

To generate a random sampling distribution, we use the following code:

sample\_means <- rep(NA, SAMPLE)

for(i in SAMPLES){

samp <- sample(VARIABLE, n)

sample\_means[i] <- mean(samp)

}

Then calculate the histogram of the sample means:

hist(sample\_means50)

We now insert the data:

sample\_means <- rep(NA, 100)

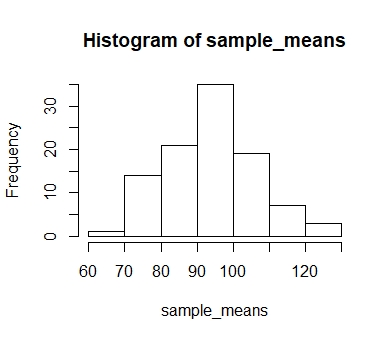
for(i in 1:100){

samp <- sample(L4E5$INTTIME, 40)

sample\_means[i] <- mean(samp)

}

hist(sample\_means)



Here we use R to take 100 samples of size 40 from the population, calculate the mean of each sample, and store each result in a vector called ‘sample\_means’. This shape is normal.

Explanation of the code:

Lets go through the code line by line to understand what it does. In the first line we initialized a vector. In this case, we created a vector of 100 zeros called ‘sample\_means’. This vector will store values generated within the ‘for’ loop.

The second line calls the ‘for’ loop itself. The line means:

*“For every element* ‘i' *from 1 to 100”*

You can think of ‘i’ as the counter that keeps track of which loop your are on. More precisely, the loop will run once when i=1, then once when i=2, and so forth up until i=100.

The body of the ‘for’ loop is the part inside the curly braces, and this set of code is run for each value of ‘i’. Here, on every loop, we take a random sample of size 40 from area, take its mean, and store it as the ith element of sample\_means50.

In order to display that this is really happening, we asked R to print ‘i’ at each iteration. This line of code is optional and is only used for displaying what’s going on while the for loop is running.

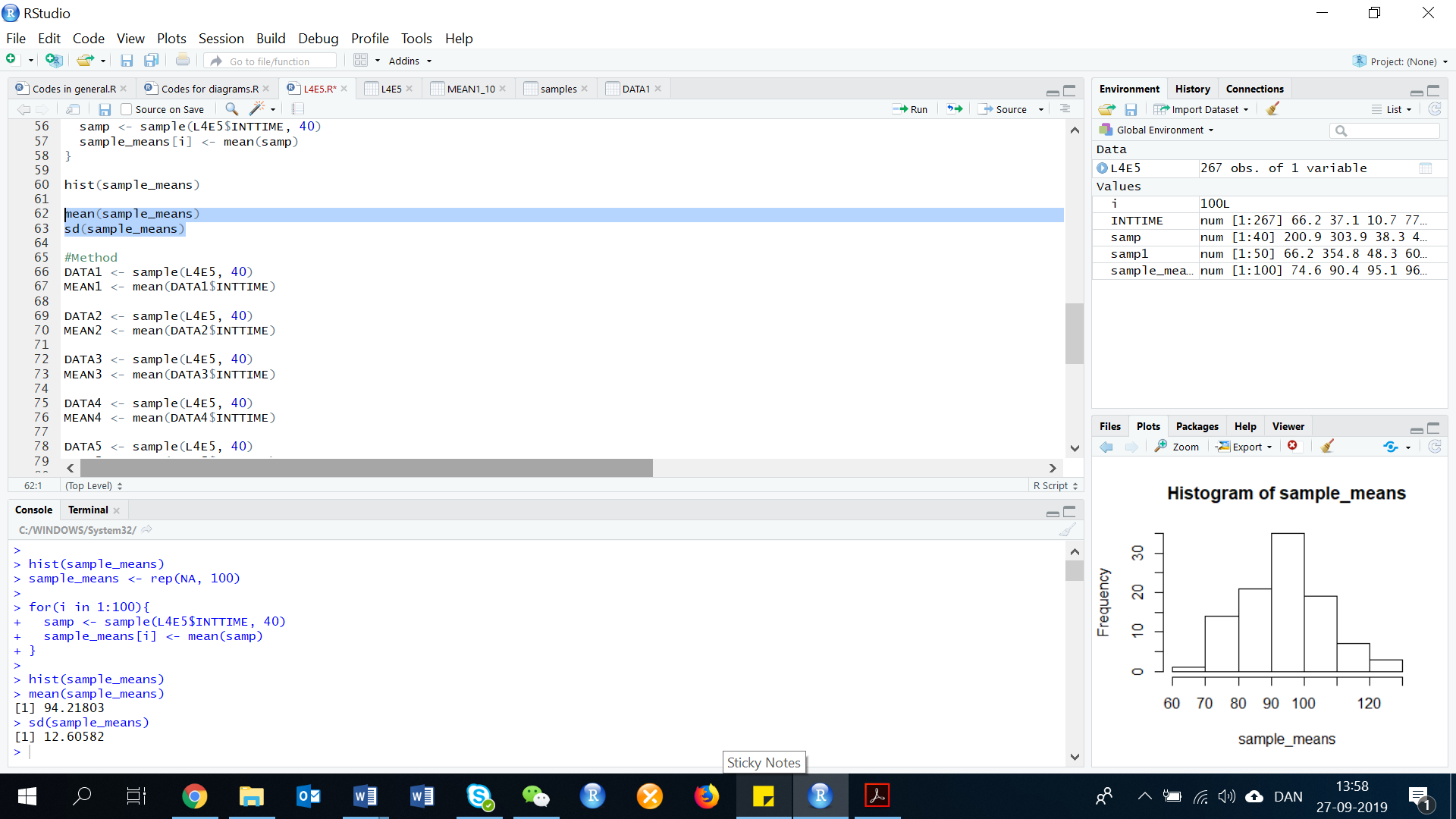
The ‘for’ loop allows us to not just run the code 5000 times, but to neatly package the results, element by element, into the empty vector that we initialized at the outset.

**g) Refer to part f. Find the mean and standard deviation of the -values. Do these values approximate and respectively?**

In this exercise, we need to calculate the mean and standard deviation of the 100 means:

mean(sample\_means)

sd(sample\_means)



The mean of these 100 means is 94.21803, this is very close to the . The standard deviation of these 100 means is 12.60582, this is somewhat close to the found in part c.

* 1. **Confidence Interval for a Population Mean: Normal (z) Statistic**

**Exercise 6. (16, BLKFRI). *Shopping on Black Fridays*. The day after Thanksgiving – called Black Friday – is one of the largest shopping days in the USA. Winthrop University researchers conducted interviews with a sample of 38 women shopping on Black Friday to gauge their shopping habits and reported the results in the International Journal of Retail and Distribution Management (Vol. 39, 2011). One question was “How many hours do you usually spend shopping on Black Friday?” Data for the 38 shoppers are listed in the accompanying table.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | 6 | 4 | 4 | 3 | 16 | 4 | 4 | 5 | 6 | 6 | 5 | 5 | 4 |
| 6 | 5 | 6 | 4 | 5 | 4 | 4 | 4 | 7 | 12 | 5 | 8 | 6 | 10 |
| 5 | 8 | 8 | 3 | 3 | 8 | 5 | 6 | 10 | 11 |  |  |  |  |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L4E6 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E6.xlsx")

View(L4E6)

attach(L4E6)

Create variable: RAW <- RAW

1. **Describe he population of interest to the researchers.**

The population of interest is all U.S. women who shop on Black Friday.

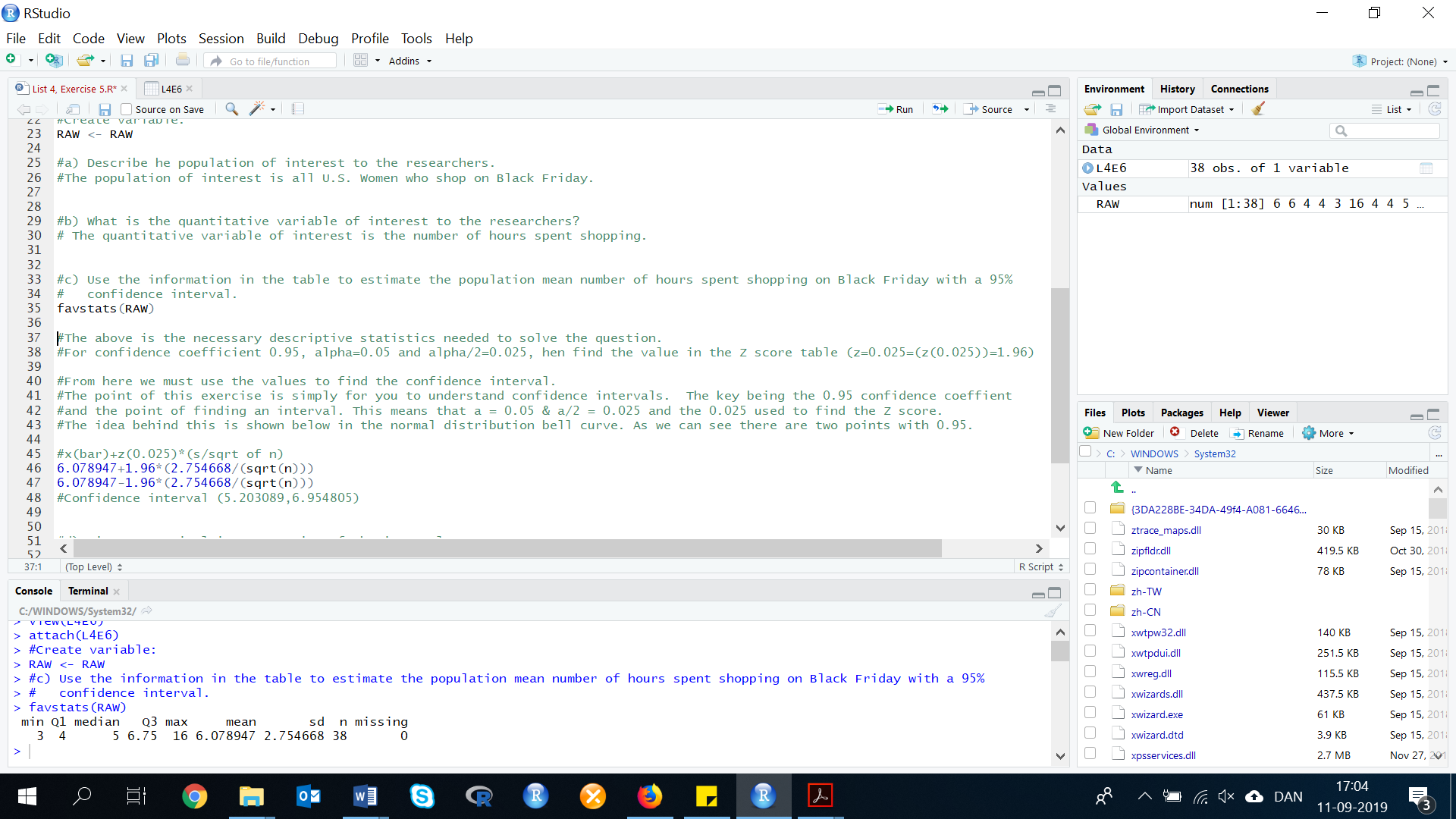
1. **What is the quantitative variable of interest to the researchers?**

The quantitative variable of interest is the number of hours spent shopping.

1. **Use the information in the table to estimate the population mean number of hours spent shopping on Black Friday with a 95% confidence interval.**

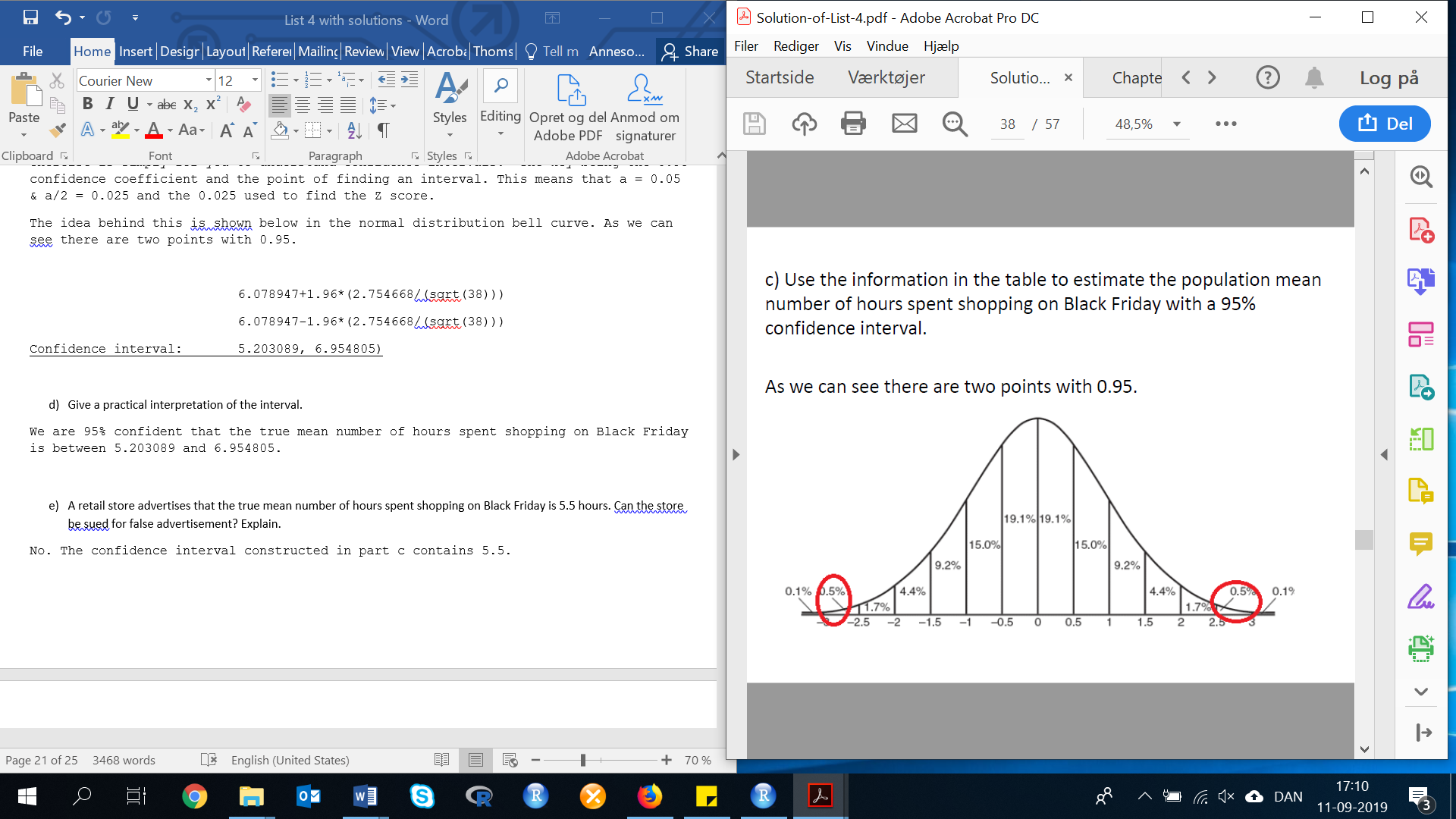
Now we need to calculate descriptive statistics:

favstats(RAW)



The above is the necessary descriptive statistics needed to solve the question. For confidence coefficient 0.95, alpha=0.05 and alpha/2=0.025, hen find the value in the Z score table (z=0.025=(z(0.025))=1.96).

From here we must use the values to find the confidence interval. The point of this exercise is simply for you to understand confidence intervals. The key being the 0.95 confidence coefficient and the point of finding an interval. This means that a = 0.05 & a/2 = 0.025 and the 0.025 used to find the Z score.



The idea behind this is shown below in the normal distribution bell curve. As we can see there are two points with 0.95.

6.078947+1.96\*(2.754668/(sqrt(38)))

6.078947-1.96\*(2.754668/(sqrt(38)))

Confidence interval: 5.203089, 6.954805)

1. **Give a practical interpretation of the interval.**

We are 95% confident that the true mean number of hours spent shopping on Black Friday is between 5.203089 and 6.954805.

1. **A retail store advertises that the true mean number of hours spent shopping on Black Friday is 5.5 hours. Can the store be sued for false advertisement? Explain.**

No. The confidence interval constructed in part c contains 5.5. Therefore the store can not be sued for false advertisement.

**Exercise 7. (20). *Facial structure of CEOs.* In Psychological Science (Vol. 22, 2011), researchers reported that a chief executive officer’s facial structure can be used to predict a firm’s financial performance. The study involved measuring the facial width-to-height ratio (WHR) for each in the sample of 55 CEOs at publicly traded Fortune 500 firms. These WHR values (determined by a computer analyzing a photo of the CEO’s face) had a mean of = 1,96 and a standard deviation of s = 0,15.**

1. **Find and interpret a 95% confidence interval for μ, the mean facial WHR for all CEOs at publicly traded Fortune 500 firms.**

For confidence coefficient .95, α=.05 and α/2=.05/2=.025. From the Z score Table, z.025=1.96. The confidence interval is:

Confidence interval: (1.92, 2.00)

1. **The researchers found that CEOs with wider faces (relative to height) tended to be associated with firms that had greater financial performance. They based their inference on an equation that uses facial WHR to predict financial performance. Suppose an analyst wants to predict the financial performance of a Fortune 500 firm based on the value of the true mean facial WHR of CEOs. The analyst wants to use the value of μ = 2,2. Do you recommend he use this value?**

No. The value of 2.2 does not fall in the 95% confidence interval. Therefore, it is not a likely value for the true mean facial WHR.

* 1. **Confidence interval for a Population Mean: Student’s t-Statistic**

**Exercise 8. (31). *Assessing the bending strength of a wooden roof.* The white wood material used for the roof of an ancient Japanese temple is imported from Northern Europe. The wooden roof must withstand as much as 100 centimeters of snow in the winter. Architects at Tohoku University (Japan) conducted a study to estimate the mean bending strength of the white wood roof (Journal of the International Association for Shell and Spatial Structures, Aug. 2004). A sample of 25 pieces of the imported wood was tested and yielded the following statistics on breaking strength (MPa): = 74,5, s = 10,9. Estimate the true mean breaking strength of the white wood with a 90% confidence interval. Interpret the result.**

So the question, that we should estimate the true means breaking strength of the white wood with a 90% confidence interval. Interpret the result. For confidence coefficient .90, α=.10 and α/2=.10/2=.05.

We known that:

Mean = 74.5

Standard Deviation = 10.9

n = 25

Therefore, the degrees of freedom (df) is:

df = n-1 = 25-1 = 24

Then we look at the table for t and find t 0.05 = 1.711.

We use almost the same procedure as in the last exercise:

Equation:

Calculated in R:

74.5+1.711\*(10.9/(sqrt(25)))

74.5-1.711\*(10.9/(sqrt(25)))

Interval: (70.77, 78.23)

The solution to the question is that we are 90% confident that the true mean breaking strength of the white wood is between 70.77 and 78.23.

**Exercise 9. (38, OVERBK). *Overbooking policies for major airlines*. Airlines overbook flights in order to reduce the odds of flying with unused seats. An article in Transportation Research (Vol. 38, 2002) investigated the optimal overbooking policies for major airlines. One of the variables measured for each airline was the compensation (in dollars) per bumped passenger required to maximize future revenue. Consider the threshold levels of compensation for a random sample of 10 major airlines shown in the table.**

**Estimate the true mean threshold compensation level for all major worldwide airlines using a 90% confidence interval. Interpret the result practically.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 825 | 850 | 1210 | 1370 | 1415 |
| 1500 | 1560 | 1625 | 2155 | 2220 |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

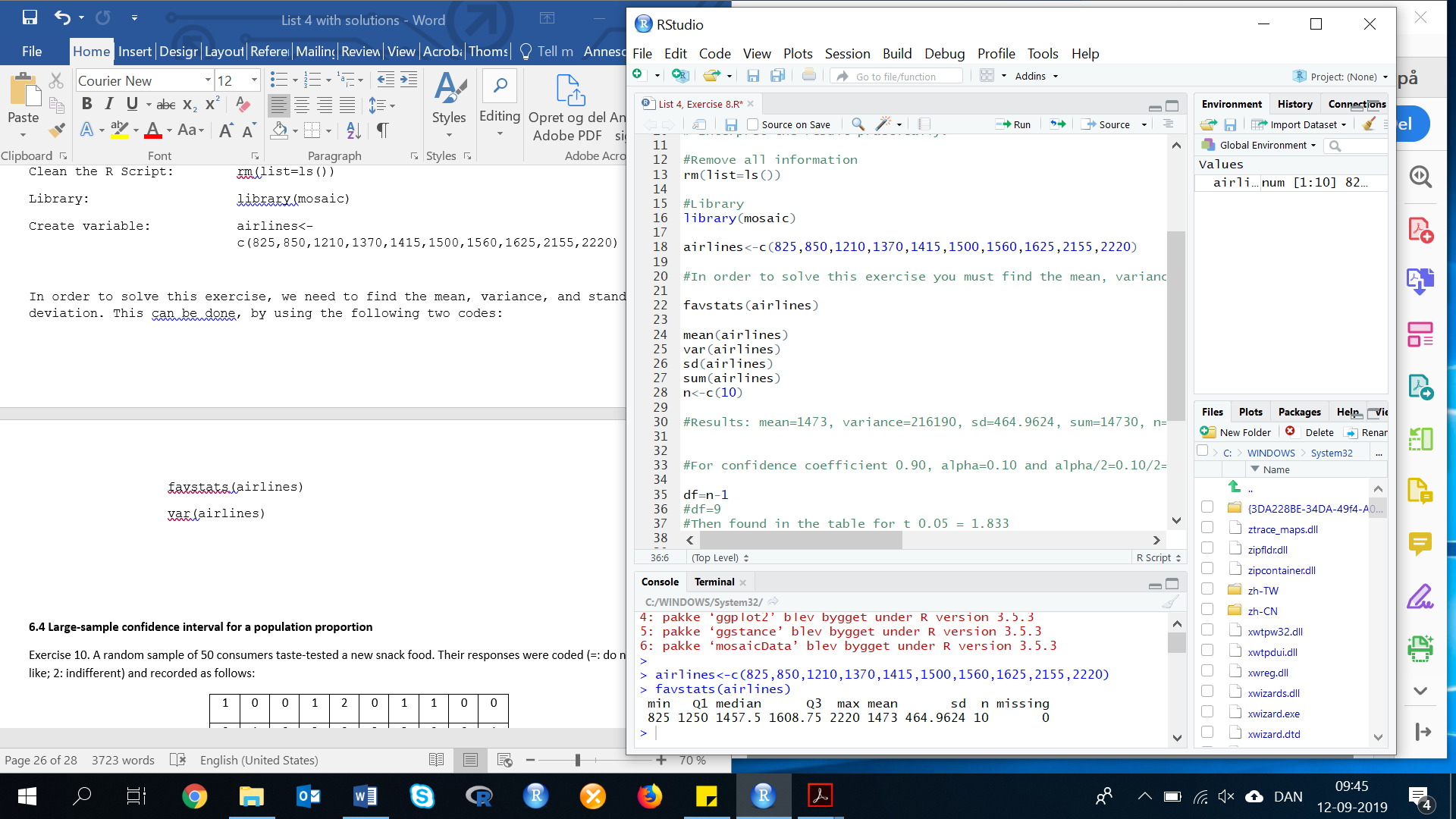
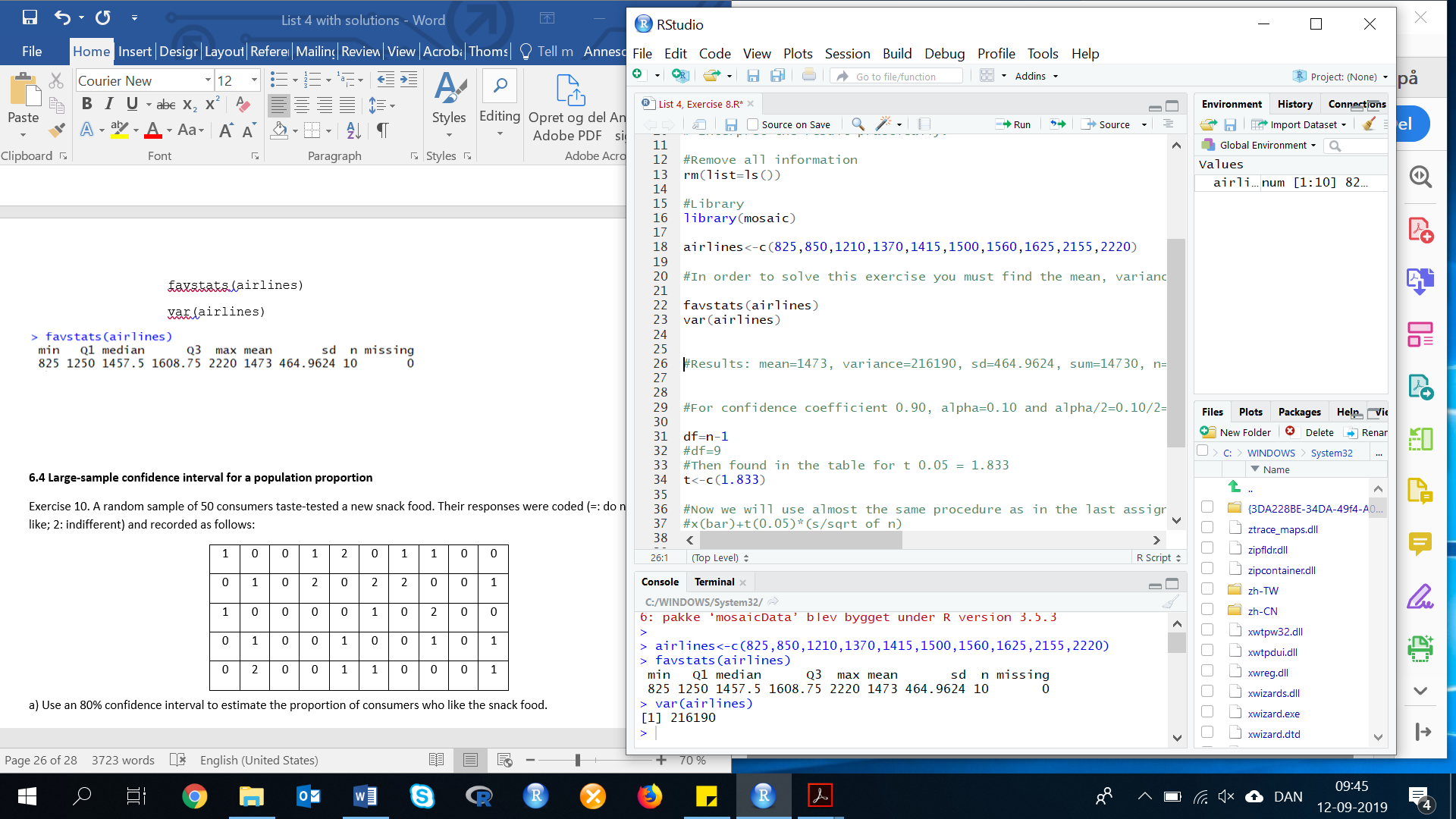
Either import the data, or create the variable manually. We created the variable:

Create variable: airlines<-c(825,850,1210,1370,1415,1500,1560,1625,2155,2220)

In order to solve this exercise, we need to find the mean, variance, and standard deviation. This can be done, by using the following two codes:

favstats(airlines)

var(airlines)



We now have these values.

For confidence coefficient 0.90, alpha=0.10 and alpha/2=0.10/2=0.05

df=n-1 = 10-1 = 9

Then found in the table for t 0.05 = 1.833

t <- c(1.833)

Now we will use almost the same procedure as in the last assignment, and the thought process is reproduced.

x(bar)+t(0.05)\*(s/sqrt of n)

mean(airlines)+t\*(sd(airlines)/(sqrt(10)))

mean(airlines)-t\*(sd(airlines)/(sqrt(10)))

Confidence interval: (1203.487, 1742.513)

The solution to the question is that we are 90% confident that the true mean threshold compensation level for all major airlines is between (1203.487 and 1742.513).

* 1. **Large-sample confidence interval for a population proportion**

**Exercise 10. (44). A random sample of 50 consumers taste-tested a new snack food. Their responses were coded (0: do not like; 1: like; 2: indifferent) and recorded as follows:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 1 | 2 | 0 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 0 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L4E10 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List4/L4E10.xlsx")

View(L4E10)

attach(L4E10)

Create values: Sample <- Sample

**a) Use an 80% confidence interval to estimate the proportion of consumers who like the snack food.**

In this exercise, the object is to estimate a population proportion, p, using a sample proportion, plus or minus a margin of error. The result will then be: the confidence interval for the population proportion, p.

The formula for CI for a population proportion is:

, where is the sample proportion, n is the sample size, and z\* is the appropriate value from the standard normal distribution for the desired confidence level.

To calculate a Confidence Interval for a population proportion, we need to do as followed:

Step 1: Determine the confidence level and find the appropriate z\*-value.

As we in this exercise, are aiming to calculate the 80% confidence interval. We know, that the z\*-value is equal to 1.28.

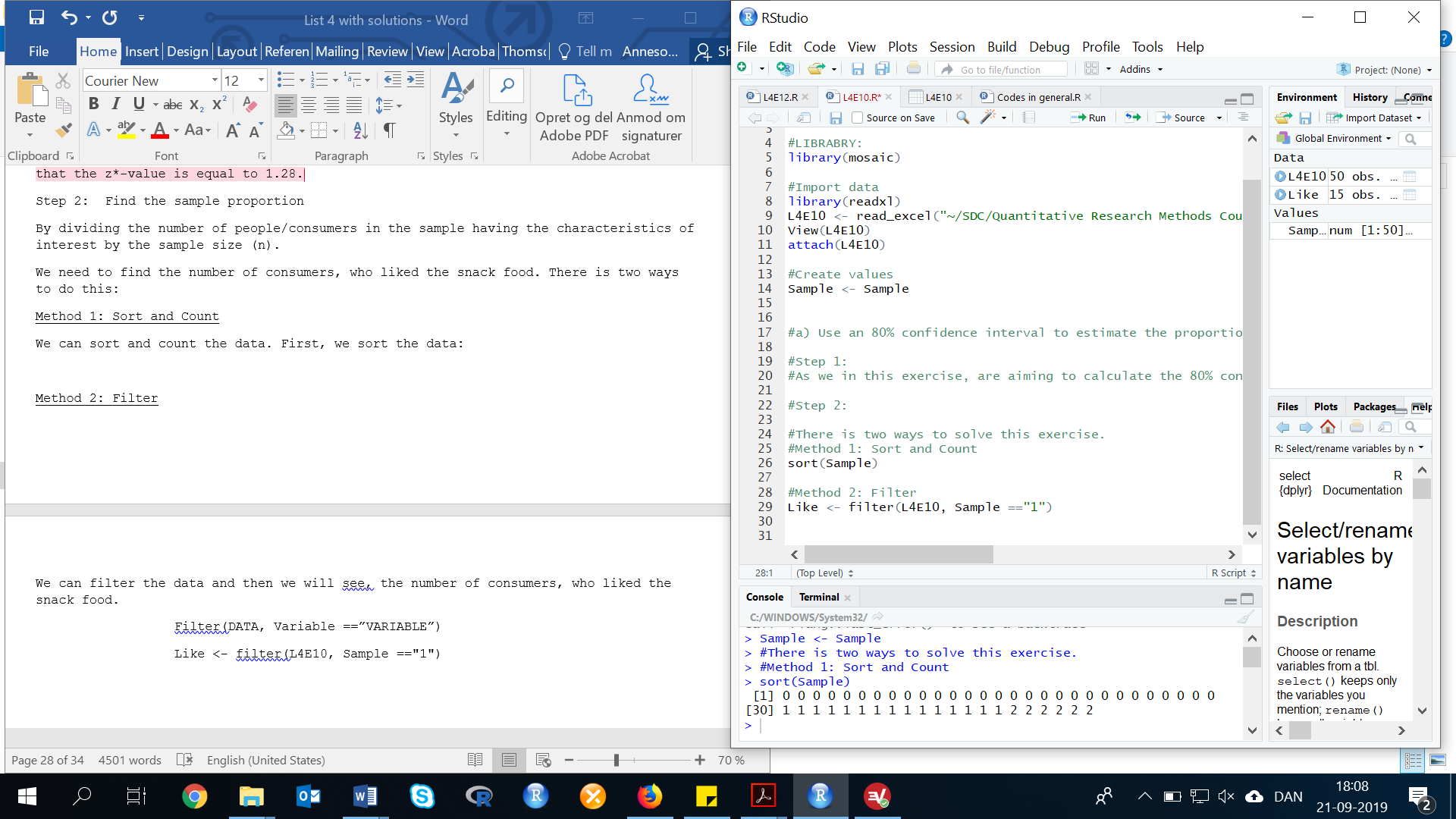
Step 2: Find the sample proportion

By dividing the number of people/consumers in the sample having the characteristics of interest by the sample size (n).

We need to find the number of consumers, who liked the snack food. There is two ways to do this:

Method 1: Sort and Count

We can sort and count the data. First, we sort the data:

sort(Sample)

Now, we can count the number of ‘liked’ values (numbers equal to 1).



By manual counting, we can see that 15 consumers out of 50 liked the food.

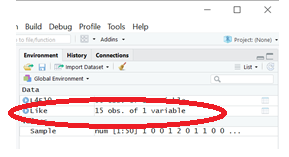
Now we can calculate :

Method 2: Filter

We can filter the data and then we will see, the number of consumers, who liked the snack food.

Filter(DATA, Variable ==”VARIABLE”)

Like <- filter(L4E10, Sample =="1")

Then we can see, that the number observations (also called the number of consumers, who liked the food) is 15 out of the 50 observation in the whole dataset.

Now we can calculate :

Step 3: Multiply

In R:

0.15\*((1-0.30)/50)=0.0021

Step 4: Take the square root of the result from step 3.

sqrt(0.0021)

Result: 0.04582576

The margin of error is, therefore, plus or minus z\*\*sqrt(0.0021):

1.28\*0.04582576

Result: 0.05865697 5.8565697 %

Step 5: Multiply your answer by z\*

Your 80% confidence interval for the percentage of times you will ever hit a red light at that particular intersection is 0.30 (or 30%), plus or minus 0.05865697 (rounded to 0.06 or 6%).

The lower end of the interval is:

0.30-0.05865697 = 0.241343 24.1343%

The higher end of the interval is:

0.30+0.05865697 = 0.358657 35.8657%

**b) Provide a statistical interpretation for the confidence interval you constructed in part a.**

To interpret these results within the context of the problem, you can say that with 80% confidence the percentage of the consumers who liked the food is somewhere between 24.13% and 35.87% based on the sample.

* 1. **Determining the Sample Size**

**Exercise 11. (72). *Shopping on Black Fridays*. Refer to the International Journal of Retail and Distribution Management (Vol. 39, 2011) survey of Black Friday shoppers. One question was “How many hours do you usually spend shopping on Black Friday?”**

1. **How many Black Friday shoppers should be included in a sample designed to estimate the average number of hours spent shopping on Black Friday if you want the estimate to deviate no more than 0.5 hour from the true mean (use a confidence level of 95% and suppose the sample standard deviation “s” is equal to 2.755)?**

ME = is to be understood as the maximum amount of deviation (0.5). Standard deviation = 2.755. For confidence coefficient 0.95, α=0.05 and α/2=0.025 then, find the value in the Z score table (Z= 0.025 = Z0.025 = 1.96.

This exercise is an extension of, what was done is Exercise 8 & 9. However, here instead we are looking to find the sample (n). In other words, we re-structure the formula to get:

In R:

((1.96\*2.755)/0.5)^2

Result: 116.6314

1. **Devise a sampling plan for collecting the data that will likely result in a representative sample.**

Answers will vary. A plan would need to be devised so that the selected shoppers were selected from a variety of different stores in a variety of locations so that the sample would be representative of the entire population.

* 1. **Confidence interval for a population variance**

**Exercise 12. (104, HCOUGH) *Is honey a cough remedy?* Archives of pediatrics and adolescent medicine (Dec. 2007) did a study of honey as a remedy for coughing. The 105 children ill in the sample were randomly divided into groups. One group received a dosage of an over-the-counter cough medicine (DM); another group received a dosage of honey (H). The coughing improvement scores (as determined by the children’s parents) for the patients in the two groups are reproduced in the accompanying table. The pediatric researchers desire information on the variation in coughing improvement scores for each of the two groups.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Honey dosage: | 12 | 11 | 15 | 11 | 10 | 13 | 10 | 4 | 15 | 16 | 9 | 14 | 10 | 6 | 10 | 8 | 11 | 12 | 12 | 8 |
| 9 | 11 | 15 | 10 | 15 | 9 | 13 | 8 | 12 | 10 | 8 | 9 | 5 | 12 | 12 |  |  |  |  |  |
| DM dosage: | 4 | 6 | 9 | 4 | 7 | 7 | 7 | 9 | 12 | 10 | 11 | 6 | 3 | 4 | 9 | 12 | 7 | 6 | 8 | 12 |
| 12 | 4 | 12 | 13 | 7 | 10 | 13 | 9 | 4 | 4 | 10 | 15 | 9 |  |  |  |  |  |  |  |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Either import data or create variables manually:

Create variable: HONEY <- c(12, 11, 15, 11, 10, 13, 10, 4, 15, 16, 9, 14, 10, 6, 10, 8, 11, 12, 12, 8, 9, 11, 15, 10, 15, 9, 13, 8, 12, 10, 8, 9, 5, 12, 12)

DM <- c(4, 6, 9, 4, 7, 7, 7, 9, 12, 10, 11, 6, 3, 4, 9, 12, 7, 6, 8, 12, 12, 4, 12, 13, 7, 10, 13, 9, 4, 4, 10, 15, 9)

1. **Find a 90% confidence interval for the standard deviation in improvement scores for the honey dosage group.**

First, we need to calculate the descriptive statistics:

length(HONEY)

length=35

var(HONEY)

var=8.151261

Now we have the needed values, to calculate the confidence level. For the confidence level .90, α=.10 and α/2=.10/2=.05, and df = n-1 = 35-1 = 34.

METHOD 1: CALCULATE IN R WITH CODES

n<- 35

We use the following code:

sqrt(((n-1)\*var(VARIABLE))/(qchisq(ALPHA, DF)))

Insert variable:

INT1 <- sqrt(((n-1)\*var(HONEY))/(qchisq(0.05, 34)))

INT2 <- sqrt(((n-1)\*var(HONEY))/(qchisq(0.95, 34)))

Interval: (3.576677, 2.387938)

METHOD 2: BY HAND

If we look at CHI-SQUARE table, we can see that and .

sqrt(((35-1)\*8.151261)/48.6024)

sqrt(((35-1)\*8.151261)/21.6643)

We now need to calculate the 90% confidence interval for the standard deviations:

Result:

**b) Repeat part a for the DM dosage group.**

First, we need to calculate the descriptive statistics:

length(DM)

length(33)

var(DM)

var=10.60417

Now we have the needed values, to calculate the confidence level. For the confidence level .90, α=.10 and α/2=.10/2=.05. We need to calculate the degrees of freedom:

df = n-1 = 33-1 = 32.

METHOD 1: WITH CODES

n1 <- 33

INT3 <- sqrt(((n1-1)\*var(DM))/(qchisq(0.05, 33)))

INT4 <- sqrt(((n1-1)\*var(DM))/(qchisq(0.95, 33)))

Interval: (4.032627, 2.675621)

METHOD 2: BY HAND

If we look at the table, we can see that and .

sqrt(((33-1)\*10.60417)/46.1943)

sqrt(((33-1)\*10.60417)/20.0719)

We now need to calculate the 90% confidence interval for the standard deviations:

Result:

**c) Based on the results, parts a and b, what conclusions can the pediatric researchers draw about which group has the smaller variation in improvement scores?**

Since the confidence intervals overlap, the researchers cannot conclude, that the variances of the two groups differ.