**Quantitative Methods**

**List of Exercises N. 3**

**Selected Exercises from McClave (2014) – Chapter 4**

**4.3 Binomial distribution**

**Exercise 1. (55). *Bridge inspection ratings*. According to the National Bridge Inspection Standard (NBIS), public bridges over 20 feet in length must be inspected and rated every 2 years. The NBIS rating scale ranges from 0 (poorest rating) to 9 (highest rating). University of Colorado engineers used a probabilistic model to forecast the inspection ratings of all major bridges in Denver (*Journal of Performance of Constructed Facilities*, Feb. 2005). For the year 2020, the engineers forecast that 9% of all major Denver bridges will have ratings of 4 or below.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)

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

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

Library: library(mosaic)

**a)** **Use the forecast to find the probability that in a random sample of major 10 Denver bridges, at least 3 will have an inspection rating of 4 or below in 2020.**

Let x=number of major bridges in Denver that will have a rating of 4 or below in 2020 in 10 trials. Then x has an approximate binomial distribution with n=10 and p = 0.09.

There is two ways to measure this:

Method 1: Manually calculation

This means that the probability of, that at least 3 will have an inspection rating of 4 or below in 2020 is 0.055.

Method 2: with R

In R we can calculate it with dbinom(). We use this function as we will test for success or failure (we have two outcomes).

We know the following information from the text:

Size = n = 10

X = at least 3 (so from 3 to 10)

Probability = 9 % or 0.09

With this information, we can either calculate the dbinom() for each of the x values:

dbinom(3, size=10, prob=0.09) +

dbinom(4, size=10, prob=0.09) +

dbinom(5, size=10, prob=0.09) +

dbinom(6, size=10, prob=0.09) +

dbinom(7, size=10, prob=0.09) +

dbinom(8, size=10, prob=0.09) +

dbinom(9, size=10, prob=0.09) +

dbinom(10, size=10, prob=0.09)

Result: 0.05404002

or we can just sum it:

sum(dbinom(3:10,10,.09))

Result: 0.05404002

**b) Suppose that you actually observe 3 or more of the sample of 10 bridges with inspection ratings of 4 or below in 2020. What inference can you make? Why?**

Since the probability of seeing at least 3 bridges out of 10 with rating of 4 or less is so small, we can conclude that the forecast of 9% of all major Denver bridges will have ratings of 4 or less in 2020 is too small. There would probably be more than 9%.

**Exercise 2. (57). *FDA report on pesticides in food*. Every quarter, the Food and Drug Administration (FDA) produces a report called *The Total Diet Study*. The FDA’s report covers a variety of food items, each of which is analyzed for potentially harmful chemical compounds. A Total Diet Study reported that no pesticides at all were found in 65% of the domestically produced food samples (*FDA Pesticide Program: Residue Monitoring*, 2008). Consider a random sample of 800 food items analyzed for the presence of pesticides.**

1. **Compute μ and σ for the random variable *x*, the number of food items found that showed no trace of pesticide.**

n=800, p=0.65

The standard deviation is equal to 13.49.

**b) Based on a sample of 800 food items, is it likely you would observe less than half without any traces of pesticide? Explain.**

Half of the 800 food items would be 400. A value of x=400 have a z-core of:

So here we calculate the standard score for an observation, by taking the raw measurement, subtract the means, and then divide by the standard deviation.

X = the raw value of the measurement of interest.

Mu and Sigma represent the parameters for the population from which the observation was drawn.

We standardize our data, so we can place them within the standard normal distribution. In this manner, the standardization allow us to compare the different types of observations (in this case, with or without traces of pesticides) based on where each of the data points (values) falls within its own distribution.

If we compare the raw values, it is easy to see that no pesticides were found in 65% percent of the items. However, lets look at the standard score.

Since the z-score associated with 400 items is so small (-8.9), it would be virtually impossible to observe less than half with any pesticides if the 65% value was correct. This is far below the average.

**4.6 The Normal Distribution**

**Exercise 3. (105). *Optimal goal target in soccer.* When attempting to score a goal in soccer, where should you aim your shot? Should you aim for a goalpost (as some soccer coaches teach), the middle of the goal or some other target? To answer these questions, Chance (Fall 2009) utilized the normal probability distribution. Suppose the accuracy of x of a professional soccer player’s shots follows a normal distribution with a mean of 0 feet and a standard deviation of 3 feet. For example, if the player hits his target, x=0; if he misses his target 2 feet to the right, x=2; and if he misses 1 foot o the left, x=-1. Now, a regulation soccer goal is 24 feet wide. Assume that a goal keeper will stop (save) all shots within 9 feet of where he is standing; all other shots on goal will score. Consider a goalkeeper who stands in the middle of the goal.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)

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

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

Library: library(mosaic)

1. **If the player aims the right corner of the goalpost, what is the probability that he will score?**

The player will score if his ball is between the target (which is the right goal post) and 3 feet to the left of the target.

* The player's target is the goal post.
* If the player hits his target then x=0.
* If the player hits 3 feet to the left of the target, then x=3.
* 3 feet it also the standard deviation of the accuracy of the players scores.

This means that the probability that he will score is between -1 standard deviation and 0.

There is some different ways to calculate the probability:

Method 1: Manual calculation

We know that the probability of a value falling within 1 standard of the mean is 68%, therefore the probability of the player scoring within -1 standard deviation and 0 of the mean is 68/2 = 34%

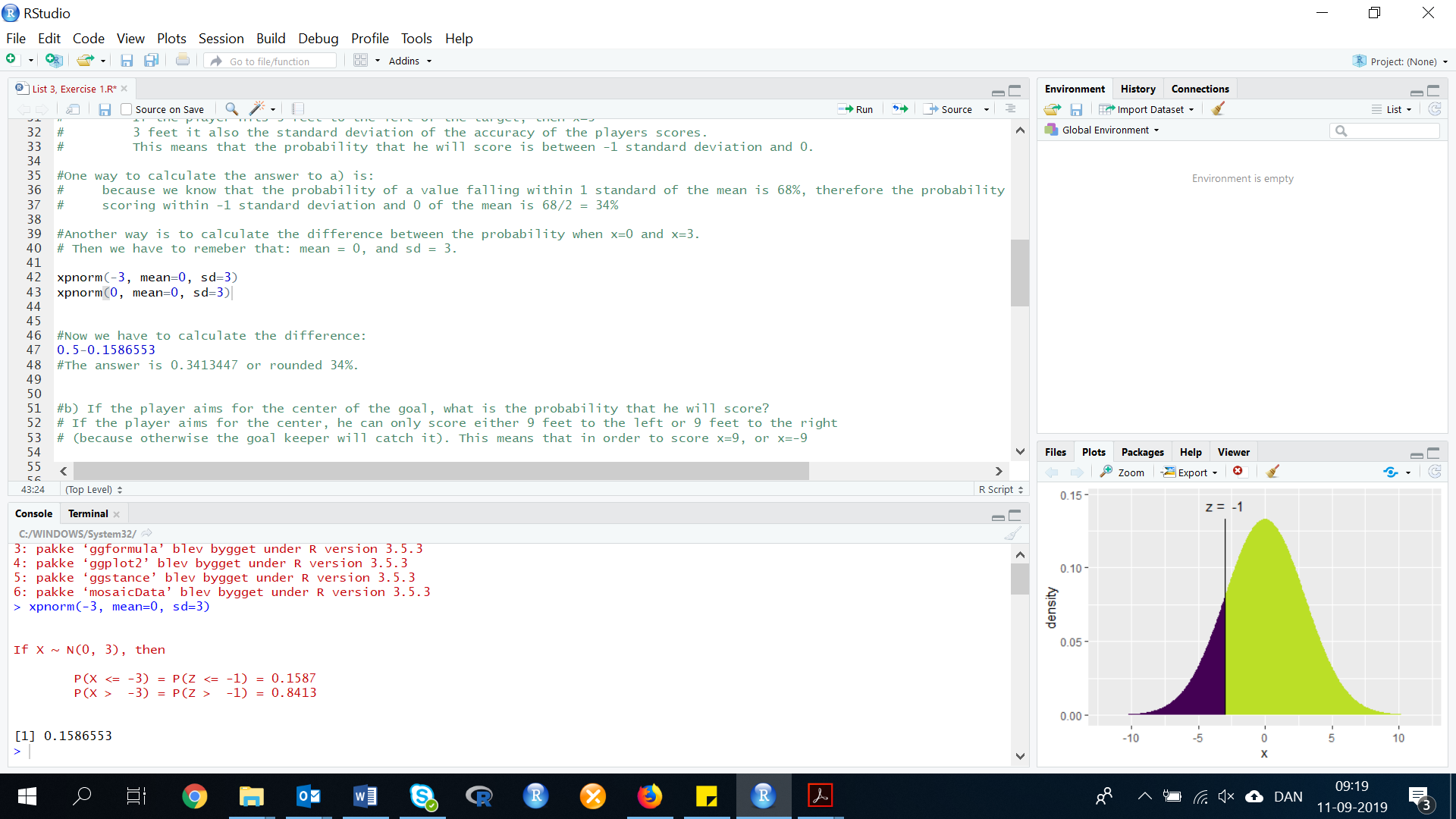
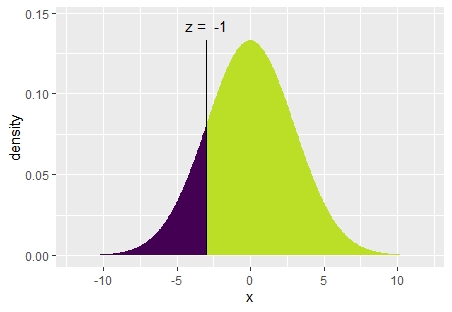
Method 2: With R

If the player aims at the right goal post, then the player will score if x is between -3 and 0. Find the difference between the probability when x=0 and x=-3. Then we have to remember that: mean = 0, and standard deviation = 3. We can calculate this in R, by using the following code:

xpnorm(x, mean=NUMBER, sd=NUMBER)

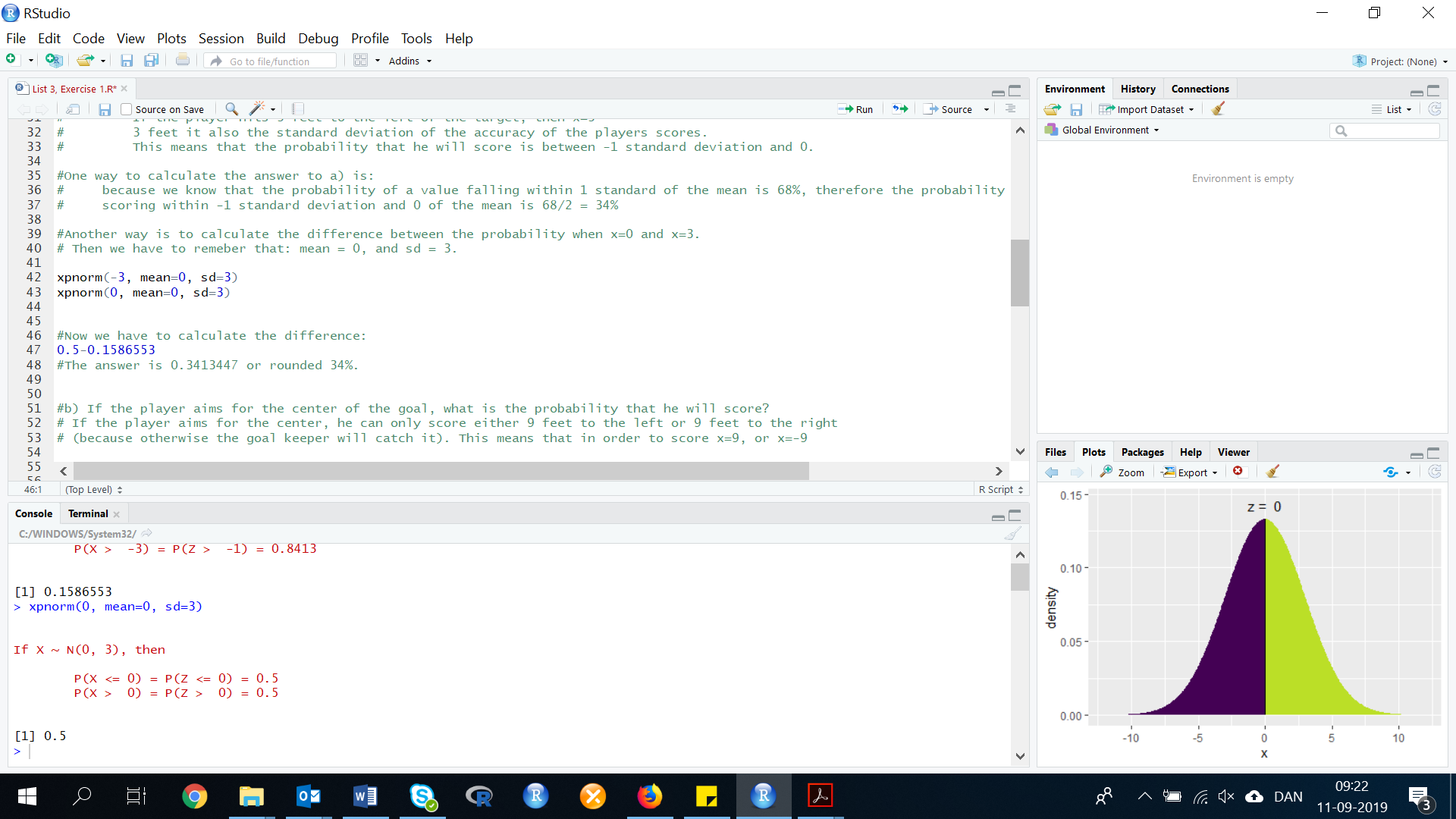
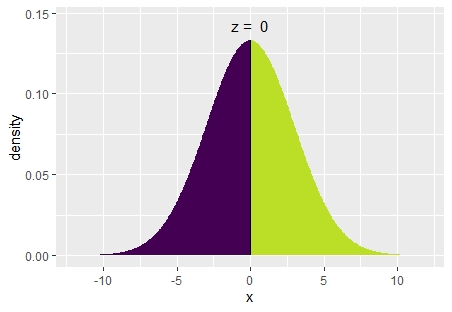
Insert data: When x = -3

xpnorm(-3, mean=0, sd=3)

The R Console gives the following result:

Insert data: When x = 0

xpnorm(0, mean=0, sd=3)



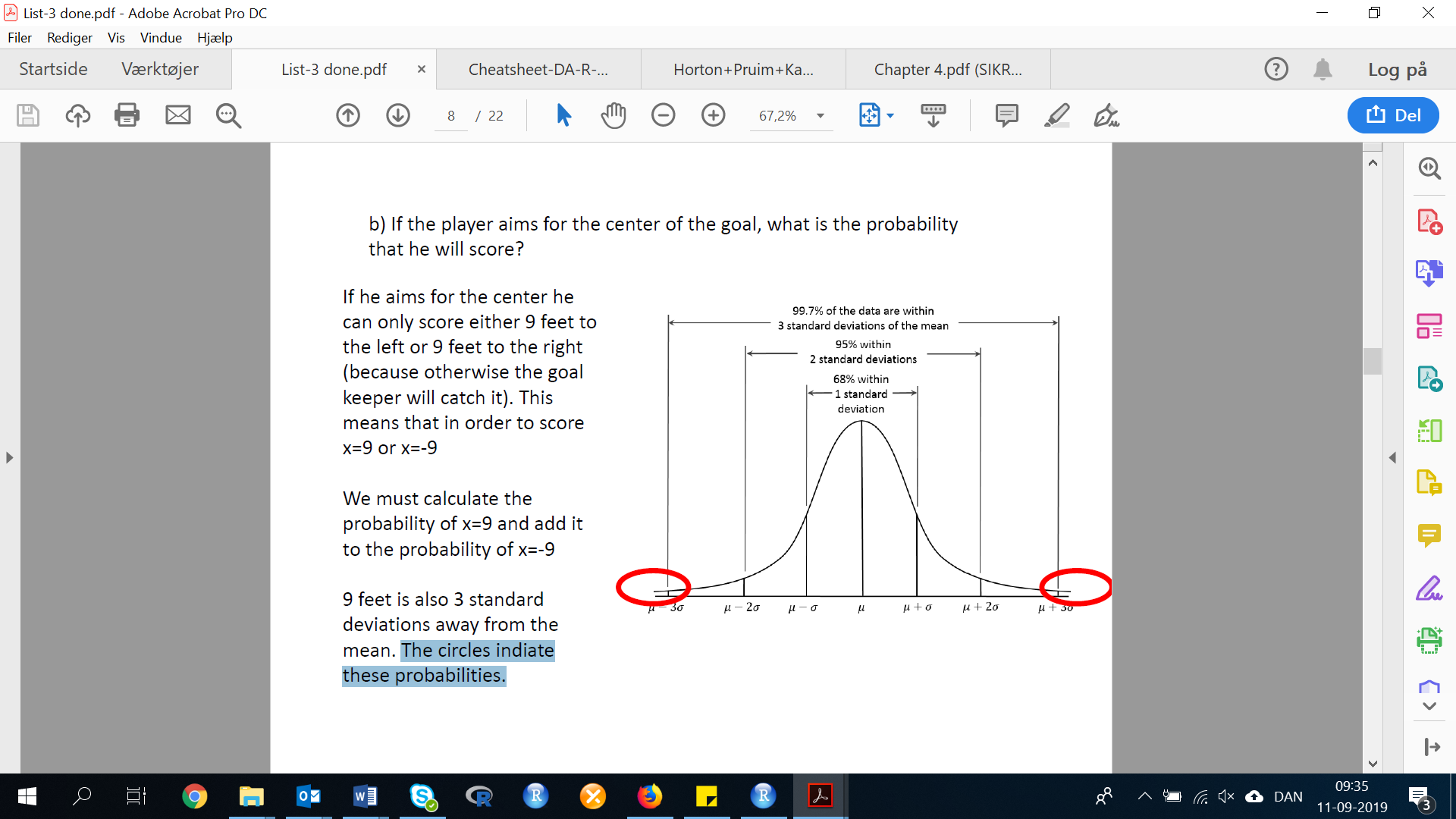
Now we can calculate the difference:

0.5-0.1586553

Result: 0.3413447

The answer is 0.3413447 or rounded 34%.

1. **If the player aims for the center of the goal, what is the probability that he will score?**

If a player aims at the center of the goal, then the player will score if x is greater than 9 or less than -9. So either 9 feet to the left or 9 feet to the right, because otherwise the goal keeper will catch it. This means that in order to score, the player needs to aim at x=9 or x=-9.

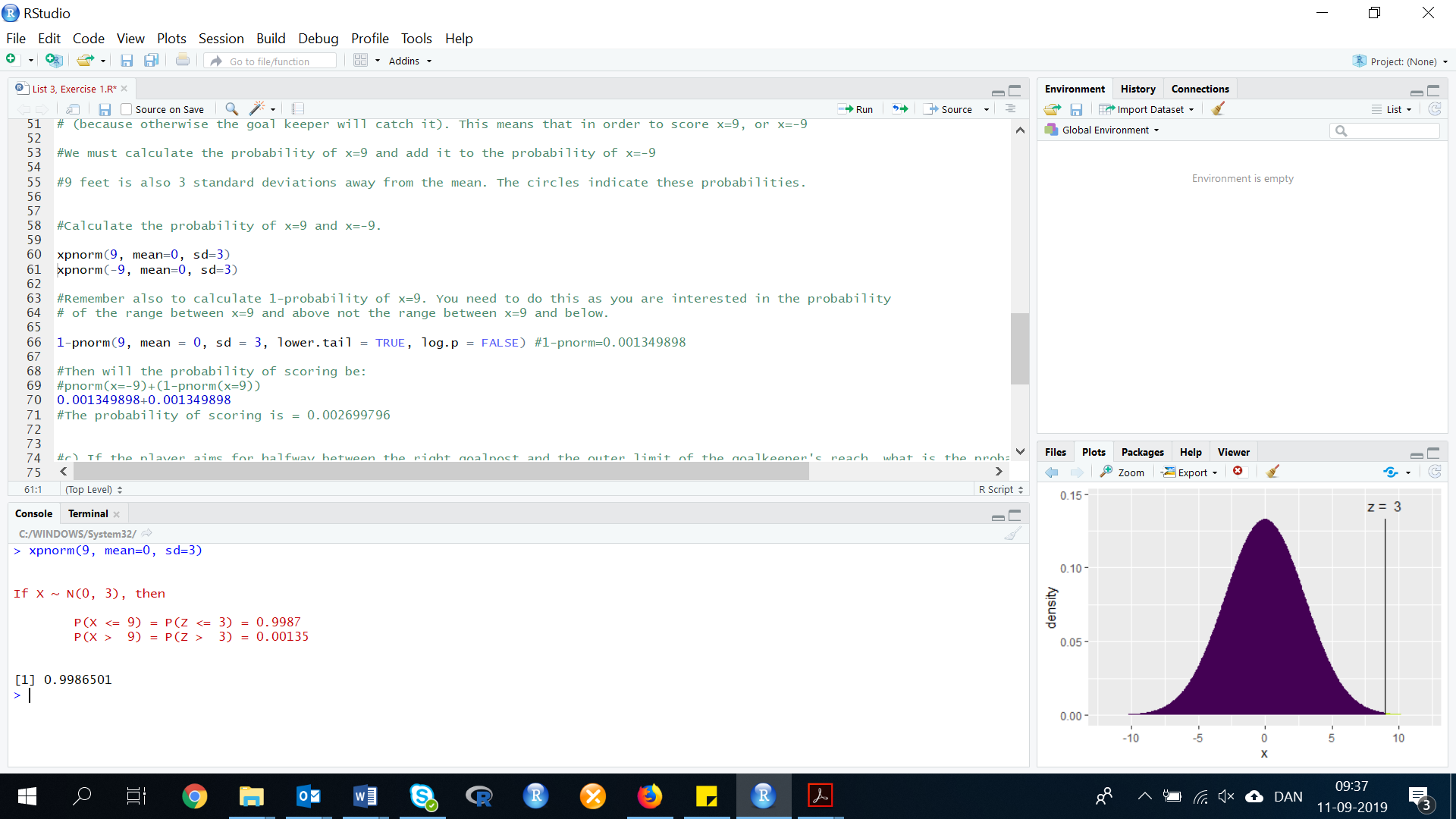
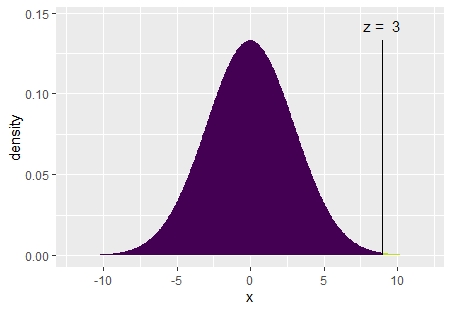
9 feet is also 3 standard deviations away from the mean. The circles indicate these probabilities.

To calculate the probability of this, we use method 2 as in exercise a.

xpnorm(x,mean=NUMBER, sd=NUMBER)

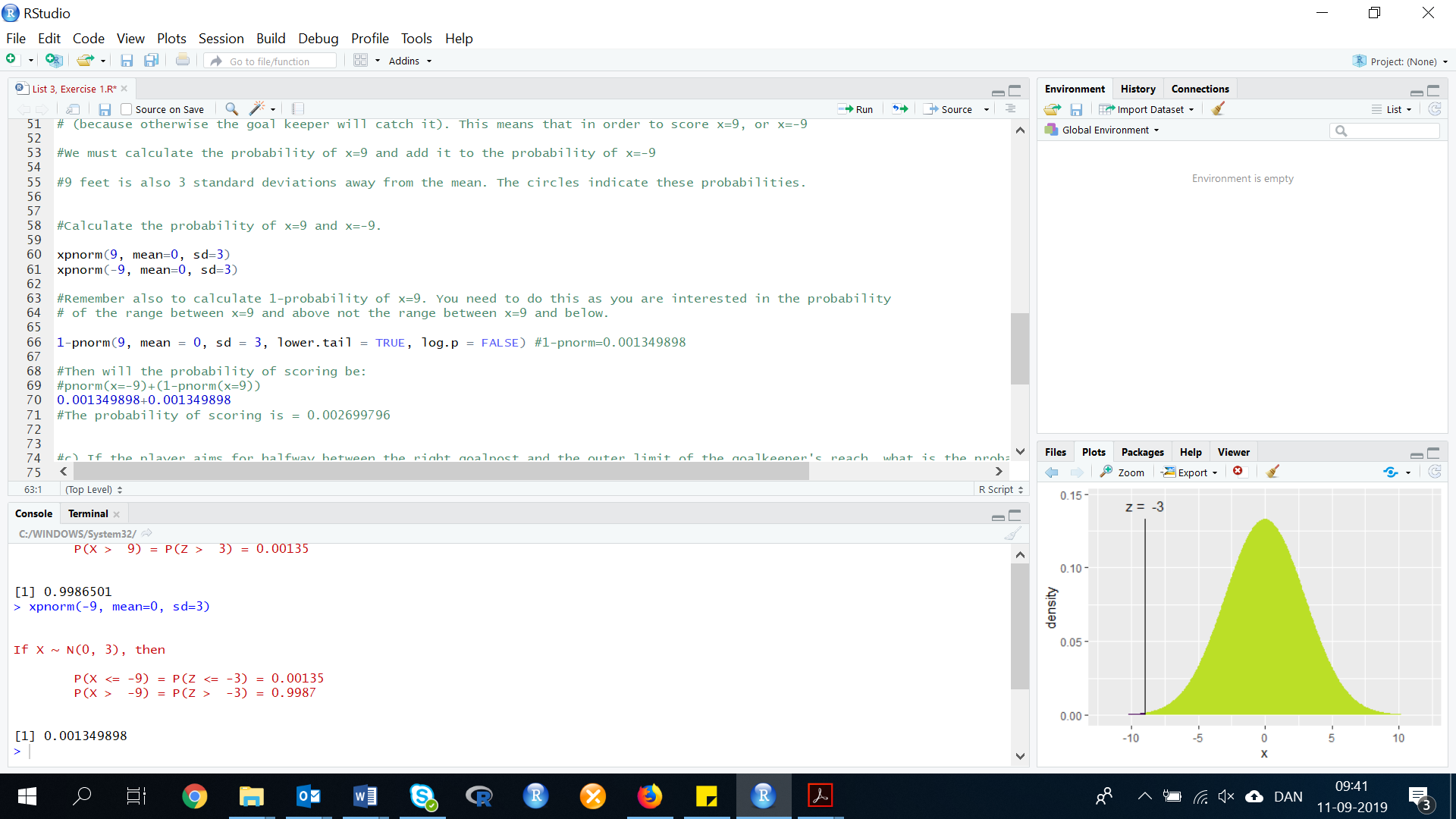
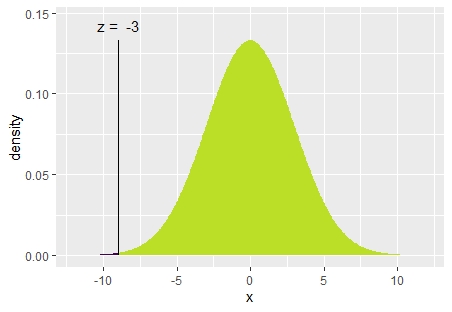
Insert data, when x=9:

xpnorm(9, mean=0, sd=3)



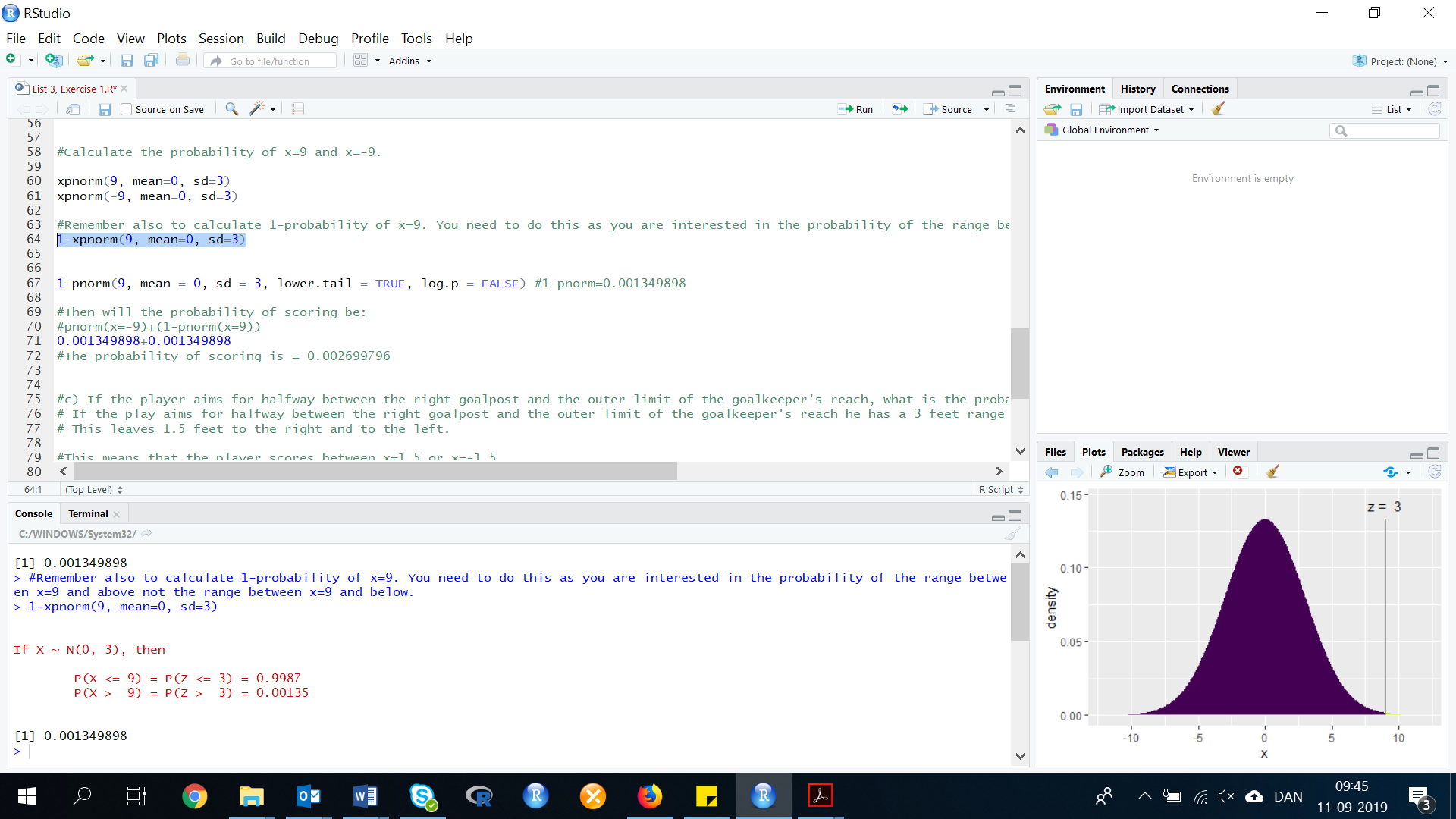
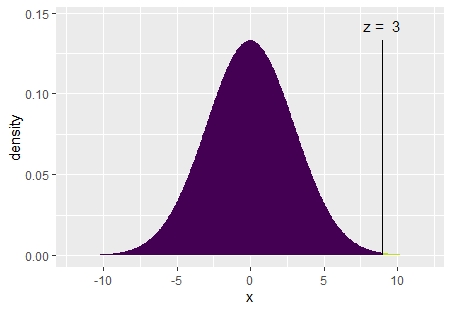
Insert data, when x=-9:

xpnorm(-9, mean=0, sd=3)



We also need to calculate the 1-probability of x=9. We need to do this, as we are interested in the probability of the range between x=9 and above not the range between x=9 and below.

1-xpnorm(9, mean=0, sd=3)



Now we can calculate the probability of scoring:

xpnorm(x=-9)+(1-xpnorm(x=9))

0.001349898+0.001349898

The probability of scoring is = 0.002699796

1. **If the player aims for halfway between the right goalpost and the outer limit of the goalkeeper’s reach, what is the probability that he will score?**

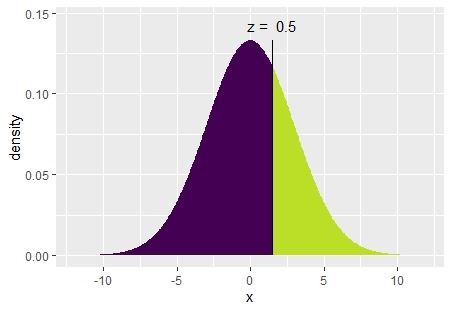
If a player aims halfway between the right goal post and the outer limit of the goal keeper’s reach, then the player will score if x is between -1.5 and 1.5.

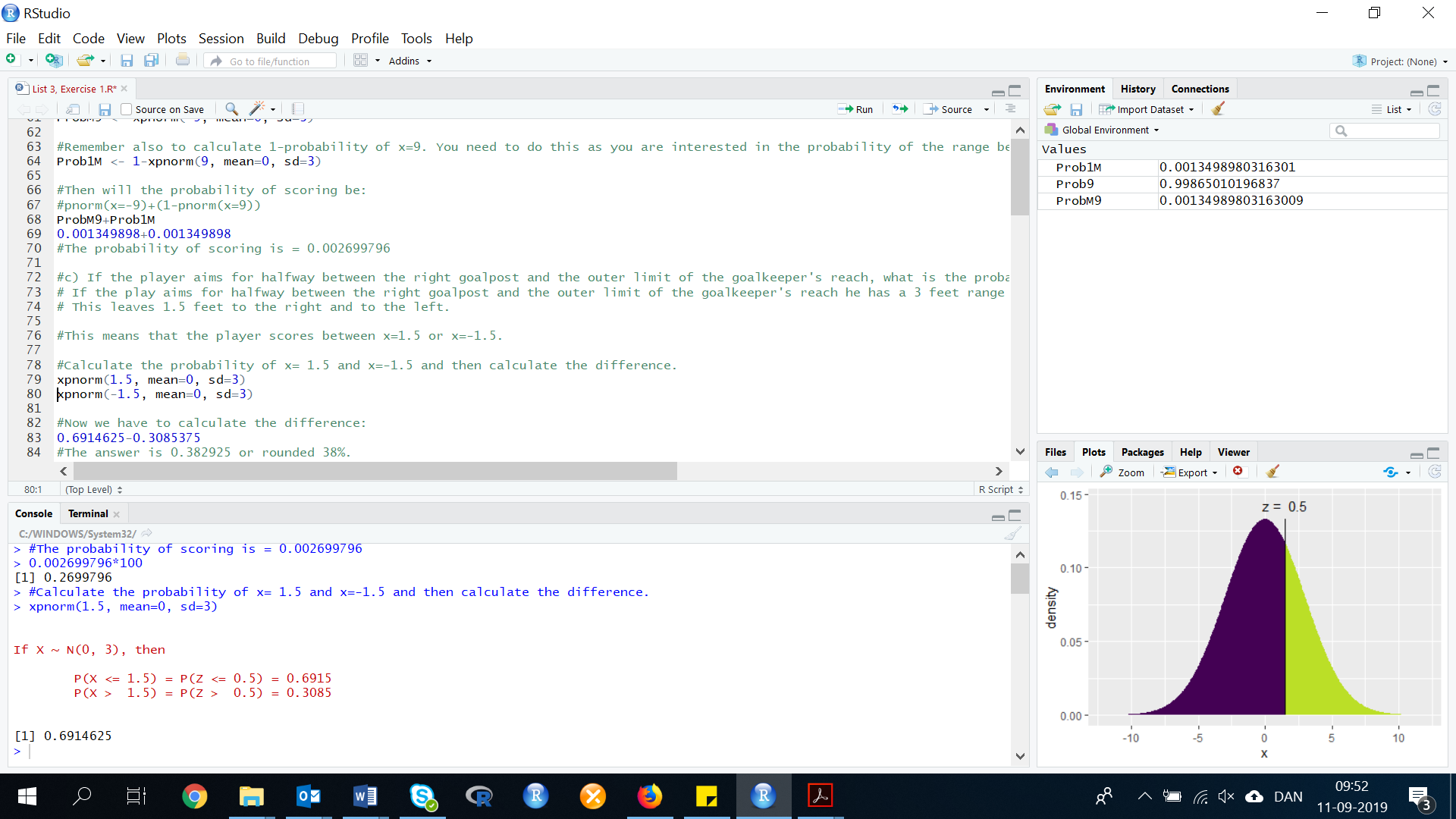
If the player aims for halfway between the right goalpost and the outer limit of the goalkeeper’s reach he has a 3 feet range to actually score in. This leaves 1.5 feet to the right and to the left. This means that the player scores between x= 1.5 or x= -1.5.

Calculate the probability of x= 1.5 and x= -1.5 and then calculate the difference:

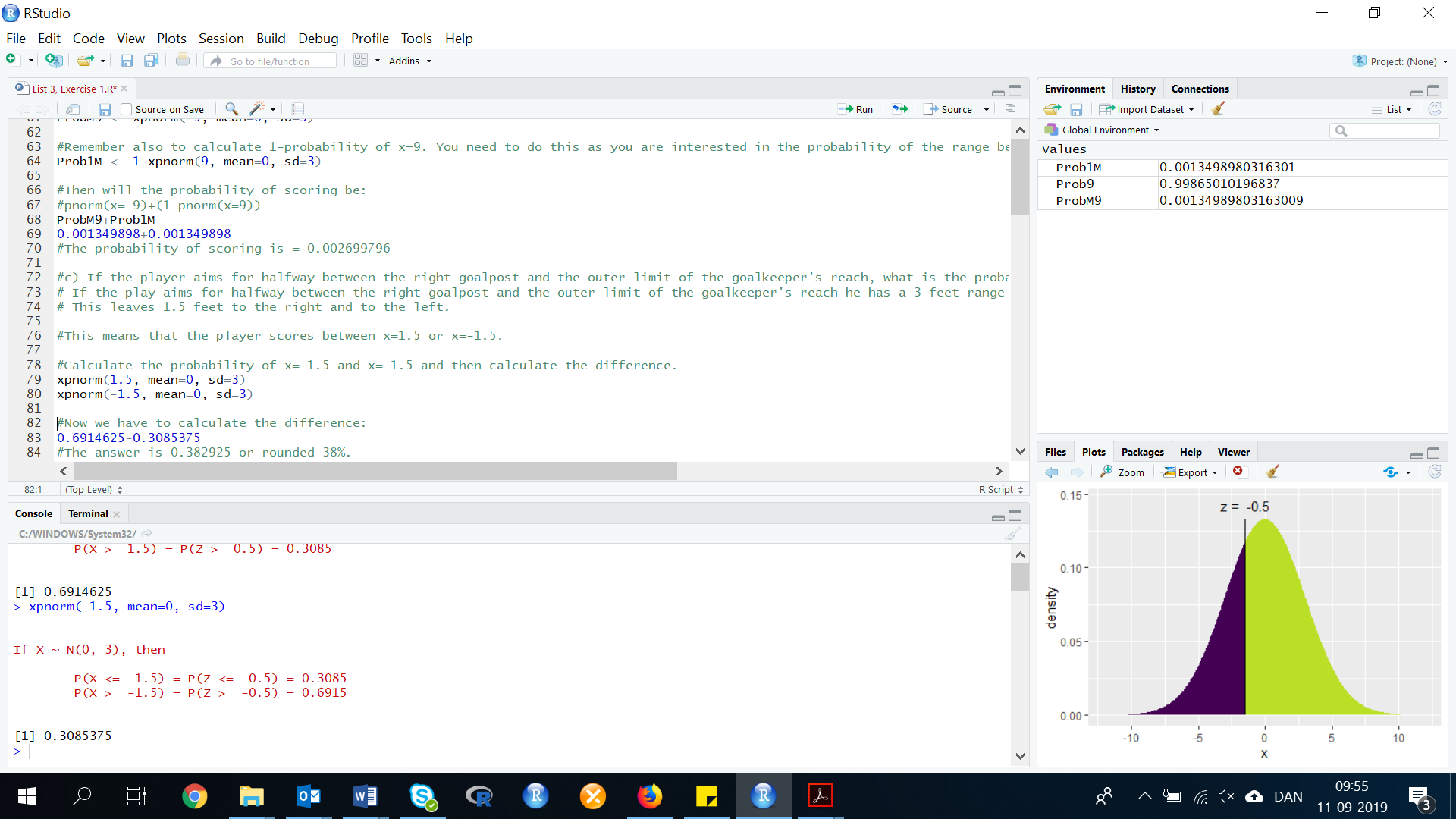
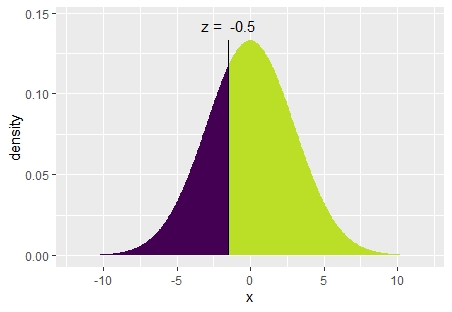
To calculate the probability, we use the same method as in exercise a and b:

xpnorm(1.5, mean=0, sd=3)





xpnorm(-1.5, mean=0, sd=3)

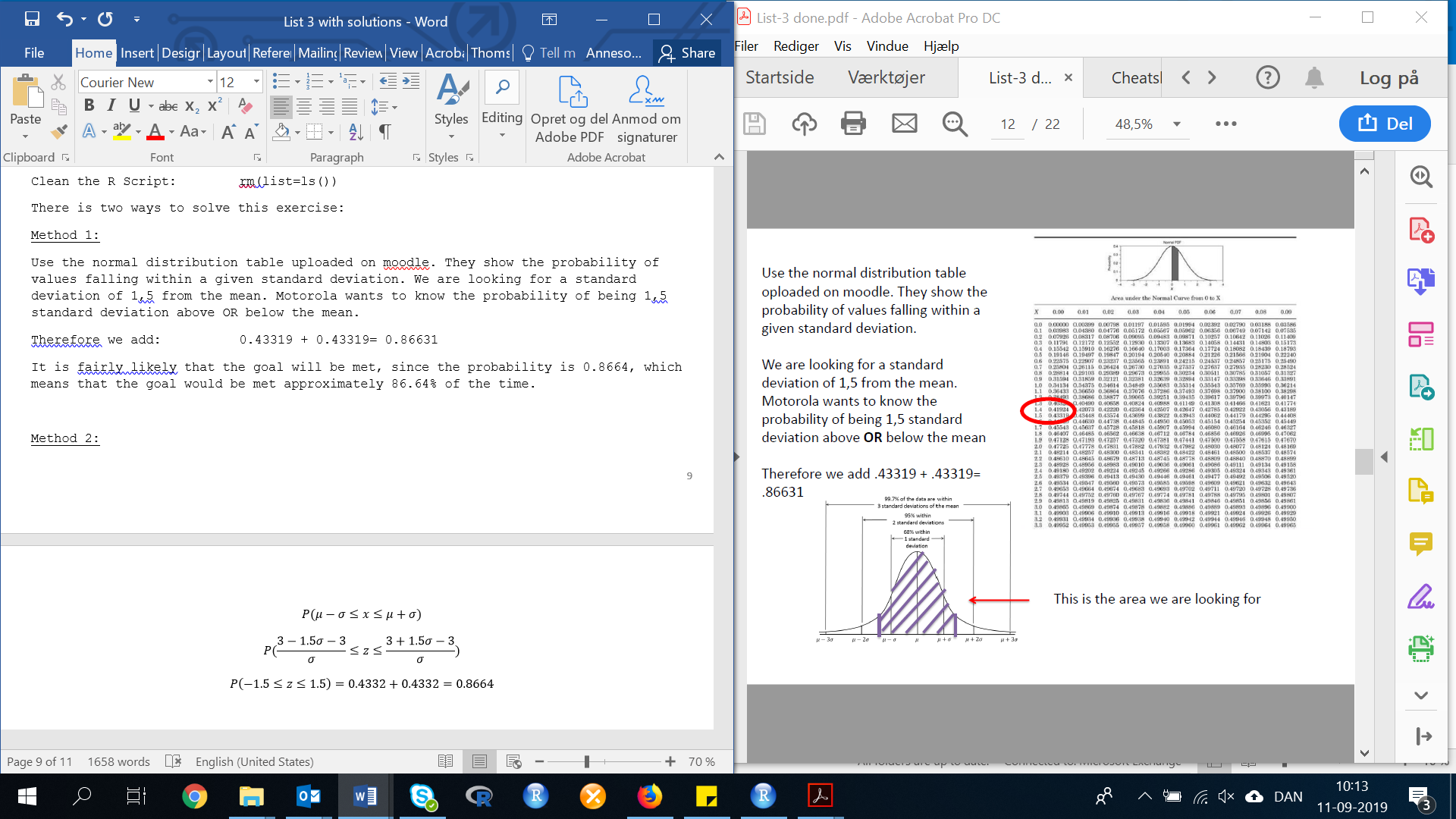


Now calculate the difference:

0.6914625-0.3085375

The answer is 0.382925 or rounded 38%.

**Exercise 4. (106). *Mean shifts on a production line.* Six Sigma is a comprehensive approach to quality goal setting that involves statistics. An article in Aircraft Engineering and Aerospace Technology (Vol. 76, No. 6, 2004) demonstrated the use of the normal distribution in Six Sigma goal setting at Motorola Corporation. Motorola discovered that the average defect rate for arts produced on an assembly line varies from run to run and is approximately normal distributed with a mean equal to 3 defects per million. Assume that the goal at Motorola is for the average defect rate to vary no more than 1.5, standard deviations above or below the mean of 3. How likely is it that the goal will be met?**

Use the normal distribution table uploaded on moodle. They show the probability of values falling within a given standard deviation. We are looking for a standard deviation of 1,5 from the mean. Motorola wants to know the probability of being 1,5 standard deviation above OR below the mean.

Therefore we add: 0.43319 + 0.43319= 0.86631

It is fairly likely that the goal will be met, since the probability is 0.8664, which means that the goal would be met approximately 86.64% of the time.

**Exercise 5. (112). *Hotel guest satisfaction*. Refer to the North American Hotel Guest Satisfaction Index Study, Exercise 48. You determined that the probability of a hotel guest participating in the hotel’s “green\* conservation program by reusing towels and bed linens is .45. Suppose a large hotel chain randomly samples 200 of its guests. the chain’s national director claims that more than 110 of these guests participated in the conservation program. Do you believe this claim? Explain.**

Let x = number of guests who participate in the conversation program in 200 trials. Then x is a binominal random variable in n = 200 and p = 0.45.

The mean of the distribution is:

µ=np=200\*(0.45)=90.

The standard deviation is:

Or

sqrt(npq)=sqrt(200\*0.45\*0.55)=sqrt(49.5)=7.0356

In order to approximate the binomial distribution with the normal distribution the interval (we do this when we have a normal distribution to approximate binomial probabilities, see page 242 for more information):

🡪

The interval should lie in the range 0 to n.

When n = 200 and p = 0.45.

Interval: (68.8932, 111.1068)

Since the interval calculated does lie in the range 0 to 200, we can use the normal approximation.

See page 242 in the book (12th edition), for explanation of the equation.

Result: 0.0018

Since the probability is so low, it is very unlikely that the claim is true.

**4.7 Descriptive Methods for Assessing Normality**

**Exercise 6. (126, PANEL). *Wear-out of used display panels.* Wear-out failure of electronic components is often assumed to have a normal distribution. Can the normal distribution be applied to the wear-out of used manufactured products, such as colored display panels? A lot of 50 used display panels was purchased by an outlet store. Each panel displays 12 to 18 color characters. Prior to acquisition, the panels had been used for about one-third of their expected lifetimes. The data in the accompanying table (saved in the file) give the failure times (in years) of the 50 used panels. Use the techniques learned in class to determine whether the used panel wear-out times are approximately normally distributed.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.01 | 1.21 | 1.71 | 2.30 | 2.96 | 0.19 | 1.22 | 1.75 | 2.30 | 2.98 | 0.51 |
| 1.24 | 1.77 | 2.41 | 3.19 | 0.57 | 1.48 | 1.79 | 2.44 | 3.25 | 0.70 | 1.54 |
| 1.88 | 2.57 | 3.31 | 0.73 | 1.59 | 1.90 | 2.61 | 1.19 | 0.75 | 1.61 | 1.93 |
| 2.62 | 3.50 | 0.75 | 1.61 | 2.01 | 2.72 | 3.50 | 1.11 | 1.62 | 2.16 | 2.76 |
| 3.50 | 1.16 | 1.62 | 2.18 | 2.84 | 3.50 |  |  |  |  |  |

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)

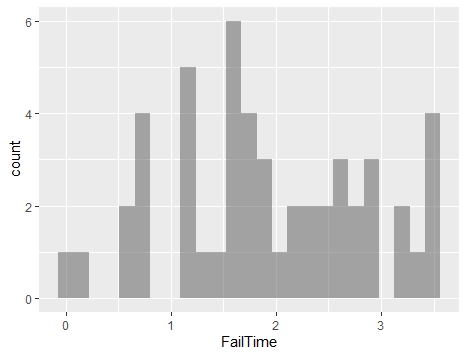
L3E6 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List3/L3E6.xlsx")

View(L3E6)

attach(L3E6)

Create variables: FailTime <- FailTime

Withdrawal <- Withdrawal

To solve this question you need to understand the data and the best way to do so is, by first visualizing the data through a histogram. Then we need to take a closer look at the data, therefore we make some descriptive statistics.

First, you need to make a histogram:

Method 1: HISTOGRAM

gf\_histogram( ~ VARIABLE, data = DATA)

gf\_histogram( ~ FailTime, data = L3E6)

From the histogram, we can see that the data appear to have a somewhat normal distribution, but we are not confident enough to declare that the data are normal distributed.

METHOD 2: EMPIRICAL RULE

sd(FailTime)

mean(FailTime)

Moving on further we shall work with the intervals to the right (it should be noted that the observations are approximately normal if they fall within 0.68, 0.95, 1.)

mean(FailTime)+sd(FailTime)

mean(FailTime)-sd(FailTime)

Interval: (1.006349, 2.863651)

by filtering the data, we get the data points (number of values) within the interval:

INT1 <- dplyr::filter(L3E6, FailTime<2.8636511 & FailTime>1.0063491)

If we look at the values 33 of the 50 values fall in this interval. The proposition is 33/50=.66. This is fairly close to the .68 we would expect if the data were normal.

mean(FailTime)+2\*sd(FailTime)

mean(FailTime)-2\*sd(FailTime)

Interval: (0.07769781, 3.792302)

INT2 <- dplyr::filter(L3E6, FailTime<3.7923021 & FailTime>0.077697811)

If we look at the values 49 of the 50 values fall in this interval. The proportion is 49/50=.98. This is a fair amount above the .95 we would expect if the data were normal.

mean(FailTime)+3\*sd(FailTime)

mean(FailTime)-3\*sd(FailTime)

Interval: (-0.8509533, 4.720953)

INT3 <- dplyr::filter(L3E6, FailTime<4.7209531 & FailTime>-0.85095331)

If we look at the values 50 of the 50 values fall in this interval. The proportion is 50/50=1.00. This is equal to the 1.00 we would expect if the data were normal. From this method, it appears that the data may be normal.

METHOD 3: IQR/SD

Now we need to look at the ratio of the IQR to the standard deviation. Please note that IQR means Inter quartile range. Hence IQR = Qu (upper) - QL (Lower).

IQR = Q3-Q1

IQR = 2.6175-1.225

IQR = 1.3925

Or we use the code in R to calculate IQR:

IQR(FailTime)

IQR=1.3925

The ratio of the IQR to sd:

IQR(FailTime)/sd(FailTime)

Result: 1.499487

This is somewhat larger than the 1.3 we would expect if the data were normal. This method indicates the data may be or could be normal.

METHOD 4: MEAN=MODE=MEDIAN

mean(FailTime)

median(FailTime)

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mode <- getmode(FailTime)

print(mode)

Mean=1.935, Median=1.835, Mode=3.5

From this we can see that the Mean, median and mode are not equal. This method do not indicate a normality.

METHOD 5: SKEWNESS AND KURTOSIS

If skewness = 0, and Kurtosis = 3, then the data is normal.

kurtosis(FailTime)

skewness(FailTime)

Kurtosis=-0.8861205, Skewness = -0.007239898

The skewness are close to 0, and therefore indicates a symmetric variable. However, the kurtosis is much less than expected.

Extra tests for normality:

METHOD 6: JARQUEBERATEST

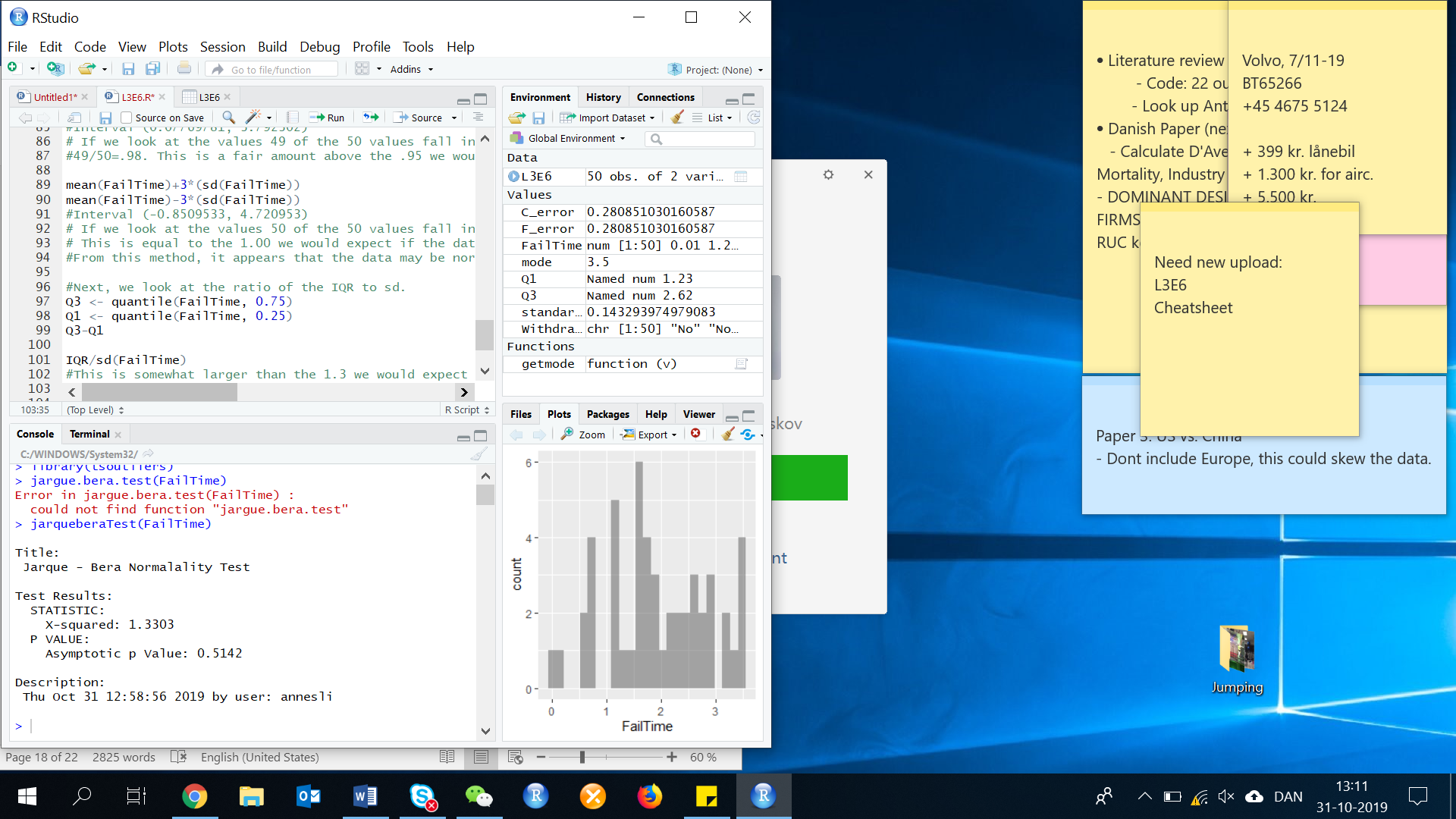
As an extra step, you can test for normality. There are some different methods to test for normality, such as the Kolmogorov-Smirnov, Shapiro-Wilk, and Jarque-Bera.

In this exercise, we will show is the Jargue-Bera test. Normality is one of the assumptions for many statistical test, such as the t-test or F-test. The Jargue-Bera test matches the skewness and kurtosis of data, to see if it matches a normal distribution.

Remember, a normal distribution has a skewness of zero (i.e. it is perfectly symmetrical around the mean) and a kurtosis of three. However, it is not necessary to know the mean or standard deviation for the data, to run this test.

jarqueberaTest(VARIABLE)

jarqueberaTest(FailTime)

Result:

Test interpretation:

H0: The sample follows a Normal distribution

Ha: The sample does not follow a Normal distribution

According to this test, we would expect the p-value > 0.05 or 0.5% if it follows a normal distribution. In this case, as the computed p-value>0.5142 is greater than the significance level of alpha=0.05, we cannot reject the null hypothesis (H0), and therefore it follows a normal distribution.

This result shows that our p-value = 0.5142 is a lot larger than 0.05, therefore we conclude that the skewness and kurtosis of the FailureTime is not significantly different from the values expected for a normal distribution.

METHOD 7: NORAMLITY PROBABILITY PLOT

First we apply the linear regression model to describe the variable Failtime by the number of withdrawals, and save the linear regression model in a new variable FailTime.lm. Then we compute the standardized residual with the rstandard() function.

FailTime.lm <- lm(FailTime~Withdrawal, data=L3E6)

FailTime.stdres <- rstandard(FailTime.lm)

Now we create the normal probability plot with the qqnorm function, and add the qqline for further comparison:

qqnorm(FailTime.stdres, ylab="Standardized Residuals", xlab="Normal scores", main="Fig 2: Normality plot of FailTime")

qqline(FailTime.stdres)



With this figure, we note that the points in the plot fall reasonable close to the line. Therefore, the variable seems to be somewhat normal distributed.

**Exercise 7. (128, SANIT). *Sanitation inspection of cruise ships.* Refer to the data on the Dec.2008 sanitation scores for 186 cruise ships. The data are saved in the accompanying file. Assess whether the sanitation scores are approximately normally distributed.**

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)

L3E7 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List3/L3E7.xlsx")

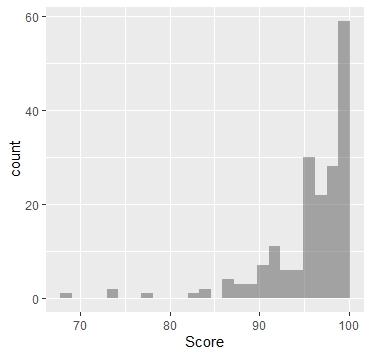
View(L3E7)

attach(L3E7)

Create variables: ShipName<- ShipName

CruiseLine <- CruiseLine

Score <- Score



To solve this question you need to understand the data and the best way to do so is by first visualizing the data through a histogram.

METHOD 1: HISTOGRAM

gf\_histogram( ~ VARIABLE, data = DATA)

gf\_histogram( ~ Score, data = L3E7)

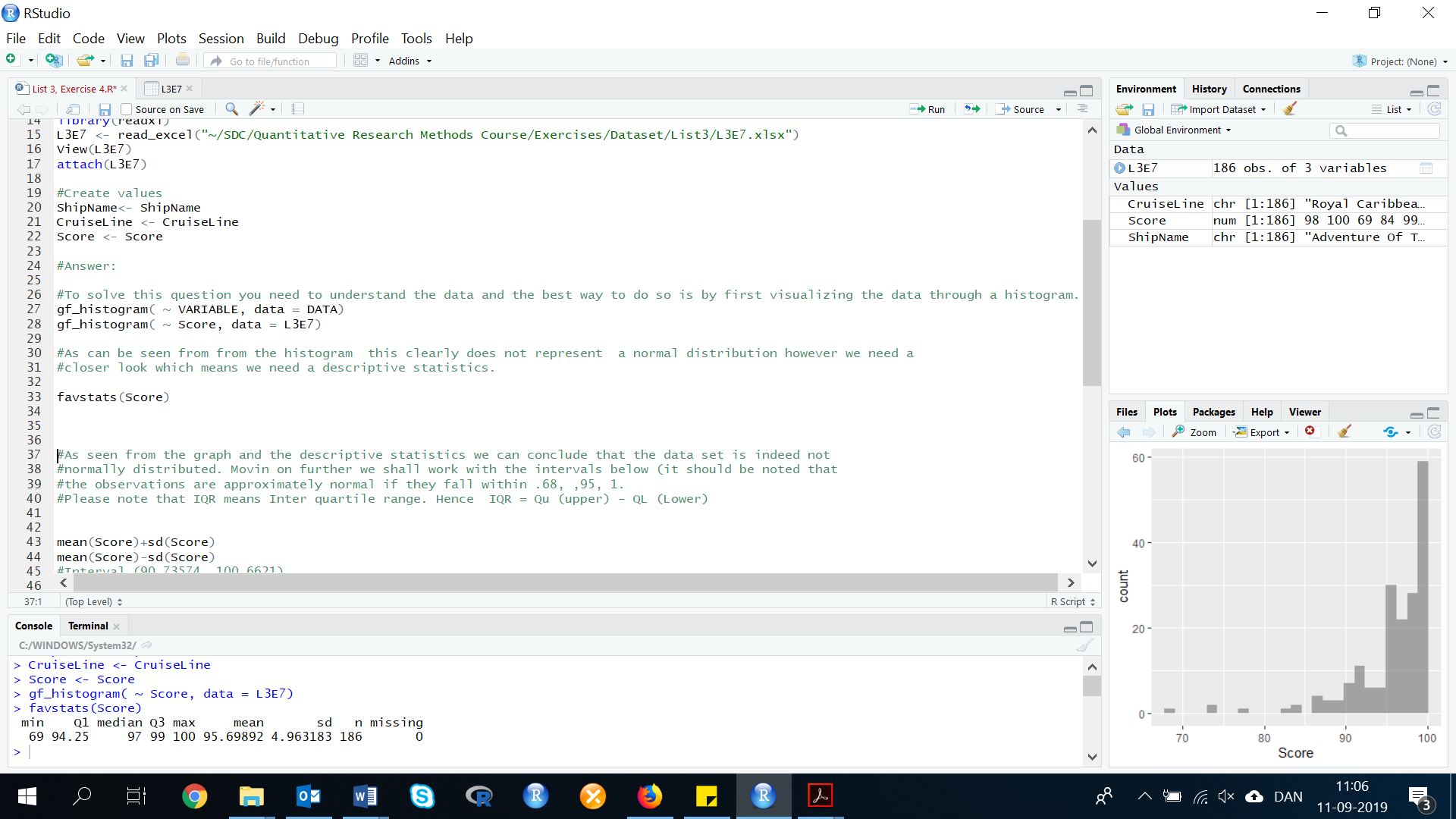
As can be seen from the histogram this clearly does not represent a normal distribution. However, we need a closer look which means we need a descriptive statistics.

The histogram shows, that the data appear to be skewed to the left. This indicates that the data are not normal.

METHOD 2: EMPIRICAL RULE

Next, we need to look at the intervals: , and . If the proportions of observations falling in each interval are approximately 0.68, 0.95, and 1.00, then the data are approximately normal. We use the same method as in List 3, Exercise 6.

favstats(Score)



mean(Score)+sd(Score)

mean(Score)-sd(Score)

Interval: (90.73574, 100.6621)

By filtering the data, we find the values within this interval above:

INT1 <- dplyr::filter(L3E7, Score<100.66211 & Score>90.735739)

If we look at the values, 166 of the 186 values falls in this interval. The proportion is 0.9120879 or 91.21% This is much larger than the 0.68 or 68% we would expect if the data were normal.

mean(Score)+2\*sd(Score)

mean(Score)-2\*sd(Score)

Interval: (85.77256, 105.6253)

INT2 <- dplyr::filter(L3E7, Score<105.62531 & Score>85.772559)

If we look at the values, 179 of the 186 values falls in this interval. The proportion is .96. This is slightly larger than the .95 we would expect if the data were normal.

mean(Score)+3\*sd(Score)

mean(Score)-3\*sd(Score)

Interval: (80.80937, 110.5885)

INT3 <- dplyr::filter(L3E7, Score<110.58851 & Score>80.809369)

If we look at the values, 186 of the 186 values fall in this interval. The proportion is 1, this is a bit higher than what we would expect if the data were normal.

From this method, it appears that the data are not normal. However, we are not finished yet. Next, we look at the ratio of the IQR to the standard deviation.

METHOD 3: IQR/SD

IQR <- Q3-Q1

IQR = 99-94.25

IQR = 4.75

Or by calculating in R:

IQR(Score)

IQR=4.75

Calculation of the ratio:

IQR/sd(Score) = 0.9570471

This is much smaller than the 1.3 we would expect if the data were normal. This method indicates the data are not normal.

METHOD 4: MEAN=MEDIAN=MODE

mean(Score)

median(Score)

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mode <- getmode(Score)

print(mode)

Mean= 95.69892, Median=97, Mode=99

These are not equal, but somewhat close.

METHOD 5: SKEWNESS AND KURTOSIS

# If skewness = 0, and Kurtosis = 3, then the data is normal.

kurtosis(Score)

skewness(Score)

Kurtosis= 7.873135, Skewness=-2.435988

Not equal to 0 or 3 as we would expect.

EXTRA METHODS:

METHOD 6: JARQUEBERA TEST (TEST FOR NORMALITY)

The Jargue-Bera test matches the skewness and kurtosis of data, to see if it matches a normal distribution.

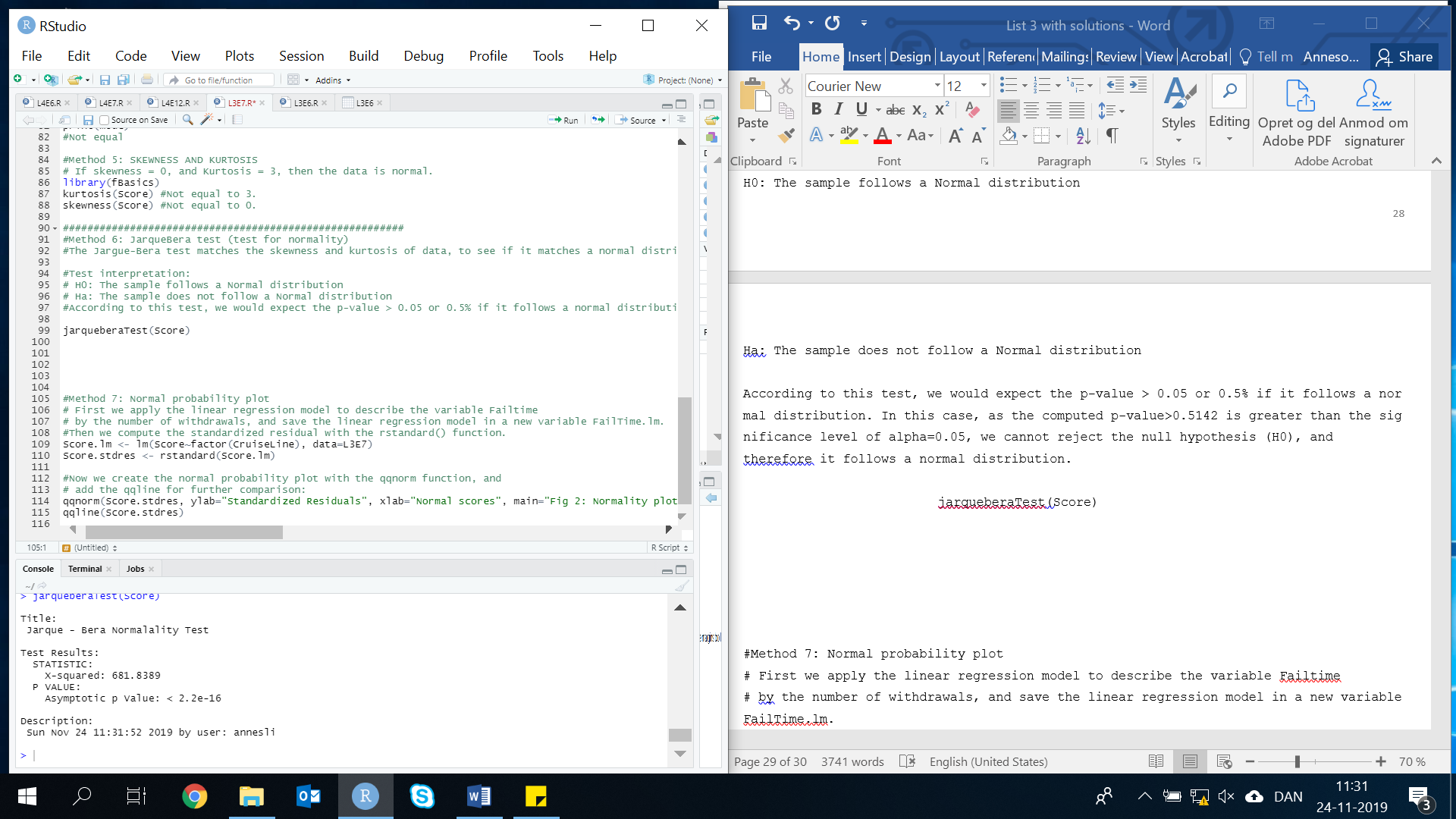
H0: The sample follows a Normal distribution

Ha: The sample does not follow a Normal distribution

According to this test, we would expect the p-value > 0.05 or 0.5% if it follows a normal distribution. In this case, as the computed p-value>0.001 is greater than the significance level of alpha=0.05, we reject the null hypothesis (H0), and

therefore it do not follow a normal distribution.

jarqueberaTest(Score)



METHOD 7: NORMAL PROBABILITY PLOT

First we apply the linear regression model to describe the variable Failtime by the number of withdrawals, and save the linear regression model in a new variable FailTime.lm.

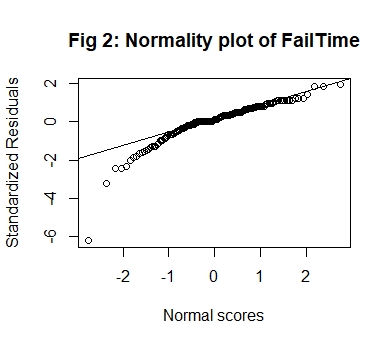
Then we compute the standardized residual with the rstandard() function.

Score.lm <- lm(Score~factor(CruiseLine), data=L3E7)

Score.stdres <- rstandard(Score.lm)

Now we create the normal probability plot with the qqnorm function, and add the qqline for further comparison:

qqnorm(Score.stdres, ylab="Standardized Residuals", xlab="Normal scores", main="Fig 2: Normality plot of FailTime")

qqline(Score.stdres)