Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40
Assignment 2

Exercise 2.1
a) 120/300 = 0.4
b)
P(he or she is over age of 60|respondent gets new primarily from newspapers)

A = respondent gets new primarily from newspapers
B = he or she is over age of 60
P(A) = 202/600 = 0.3366
P(B) = 300/600 = 0.5

P(B|A) = P(A && B)/P(A)
P(B|A) = (120/600)/0.3366 = 0.59

c)
Because the probabilities are conditional they are calculated relatively to the other probability. For example if we have 2 types of people: people who own a Ferrari, and people who are rich. Almost all Ferrari owners are rich, but not that many rich people do own a Ferrari. Since they are relative to each other the outcome is different. The same holds for a and b, the order of the conditional probability is different, the outcome aswell.

d)
If the two events are independent the following statement should hold:
P(A&&B) = P(A) * P(B)
0.2 = 0.3366*0.5
0.2 = 0,1683
The above statement is NOT true, therefore the 2 events are dependent.

Exercise 2.2

a) If we have only one coin toss, we have a sample space of Ω{H, T}, and X = number of heads in one coin toss.

X(T) = 0, X(H) = 1

P(X = 0) = P({T}) = 0.3
P(X = 1) = P({H}) = 0.7

x	P(X = x)	Numerical value of P(X = x)
0	7/10	0.7
1	3/10	0.3

b) If we have two coin tosses, we have a sample space of Ω = {HH, HT, TH, TT}, and X = number of heads in two coin tosses.
X(TT) = 0
X(HT) = 1, X(TH) = 1
X(HH) = 2

P(X = 0) = P({TT}) = 0.3 * 0.3 = 0.09
P(X = 1) = P({HT, TH}) = (0.3 * 0.7) + (0.3 * 0.7) = 0.21 + 0.21 = 0.42
P(X = 2) = P({HH}) = 0.7 * 0.7 = 0.49

x	P(X = x)	Numerical value of P(X = x)
0	9/100	0.09
1	21/50 (42/100)	0.42
2	49/100	0.49

c) This table shows the probability distribution of X = number of heads in two coin tosses. (See 2.2b) To calculate the expected value, we need the following ‘formula’:

μ = E(X) = $\sum_{i=0}^{2} i * P(X=i)$. The table also shows the values of i * P(X = i) (last column).

x	P(X=x)	Numerical value of P(X = x)	x * P(X = x)
0	9/100	0.09	0
1	21/50	0.42	0.42
2	49/100	0.49	0.98

μ = E(X) = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2)
μ = E(X) = 0 + 0.42 + 0.98
μ = E(X) = 1.4

The expected number of heads in two coin tosses is 1.4

d) First, we have to calculate the expected value of numbers of heads in one coin toss. This is:

μ = E(X) = $\sum_{i=0}^{1} i * P(X = i)$

μ = E(X) = 0 * P(X = 0) + 1 * P(X = 1)

μ = E(X) = 0 + 0.7 (see 2.2a for table + graph)

μ = E(X) = 0.7

Then, we have to calculate the variance of X:

Var(X) = $\sum_{i=0}^{1} i^2 * P(X = i) - \mu^2$

Var(X) = (0^2 * P(X = 0) + 1^2 * P(X = 1)) - μ^2

Var(X) = (0 * 0 + 1 * 0.7) – 0.7^2

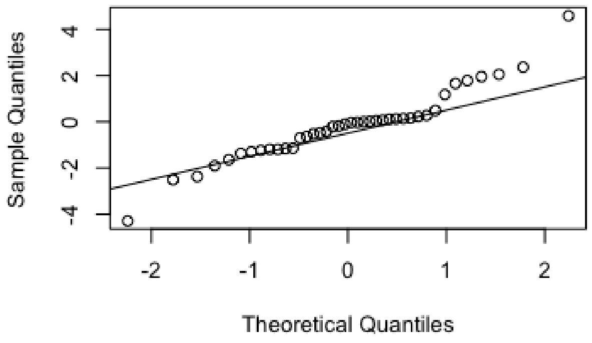
Var(X) = 0.7 – 0.49 = 0,21

The standard deviation of X is $\sqrt{Var(X)}$ -> $\sqrt{0.21}$ = 0.46.

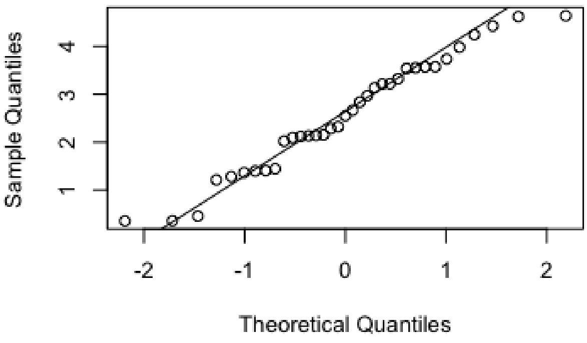
e) The law of large numbers states, that if you do the experiment for a large number of times (so n should be a large number), the mean will approach the expected value. Since we have an unfair coin with a chance of 0.7 for heads, the expected value of the random variable “the mean number of heads per coin toss after n tosses” is 0.7. The more tosses we will do, the more we will approach this expected value. This means that we are getting a smaller and smaller standard deviation, that will approach 0.

Exercise 2.3
a)

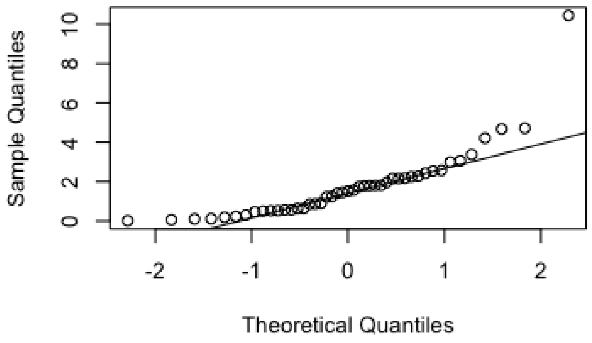
(i) t-distribution



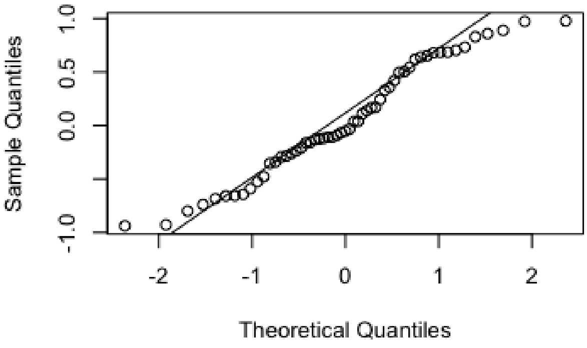
(ii) normal distribution



(iii) chi-squared distribution



(iv) uniform distribution



(i) The middle part of the QQ-plot follows approximately a straight line, x -1. However, the left and (most of all) the right tail, are a bit heavier than the normal distribution. So this distribution has a kind of bell-shaped form, but with heavier tails.§

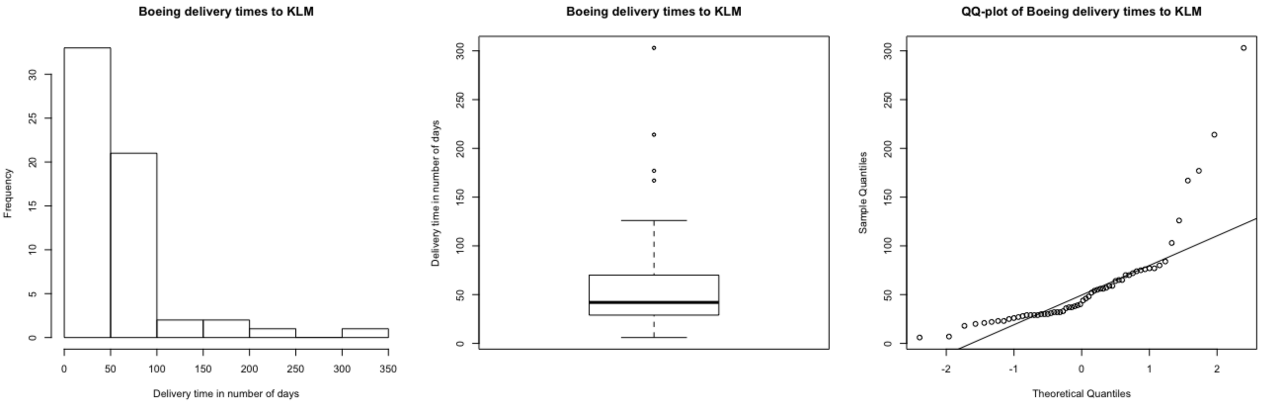
(ii) This plot approximately follows a straight line, x + 2,5. This means that we do actually have a normal distribution here.

(iii) This plot also follows a straight line in the middle part, x + 2. However, we a flattened, slightly heavier left tail, and a right tail that is way heavier than the normal distribution. So we end up with a kind of line that can be found in an exponential formula. The distribution has a bell-shaped middle, with a heavy right tail and a shorter, but also slightly heavy left tail.

(iv) This line follows the line (probably) 0,5x + 0,25. We have kind of the same situation as (i) here, where we have both a heavier left and right tail. However, the differences are more subtle here, with the left and the right being only slightly heavier than the normal distribution.

b)

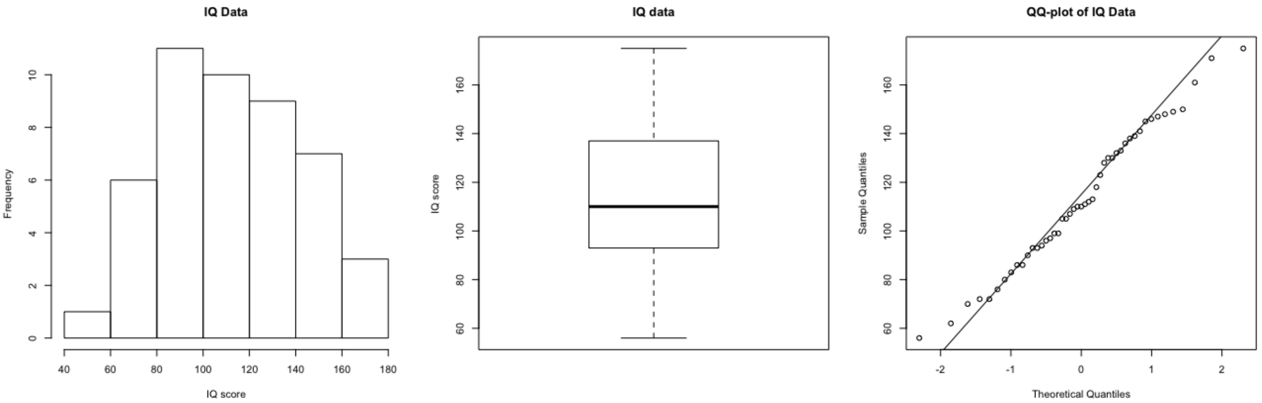
(i) KLM:



Obviously not from a normal distribution:

If we look at the histogram, we can already see that the shape is nothing like a bell-shaped form. We can see in the boxplot that we have some extreme values that are far away from the middle (1st to 3rd quantile) of the distribution. Also, if we take a look at the qq-plot, we see that it barely follows are straight line, and that it has a really heavy right tail when compared to the normal distribution.

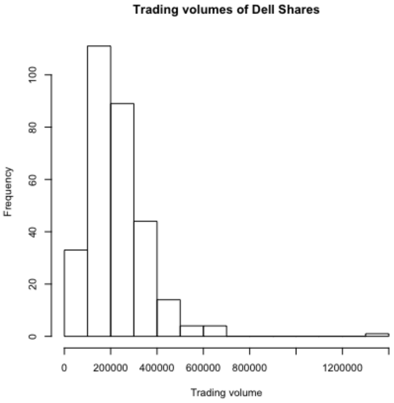
(ii) IQ data:

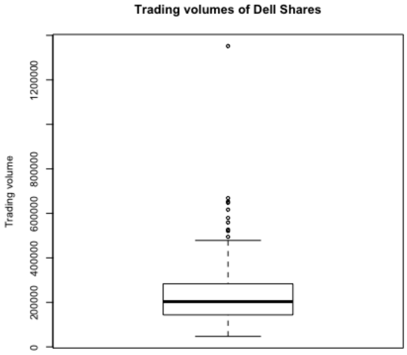


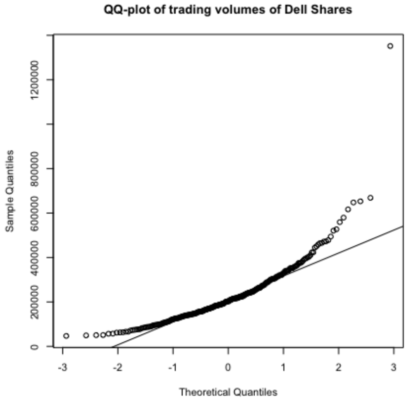
Normality cannot be excluded:

If we look at all three plots, we can see that they look quite nice and even. The histogram has a kind of bell shaped form, we have a boxplot that has a nice distribution, also the distance from minimum to the 1st quartile, and the 3rd quartile to the maximum look similar. Also, if we look at the QQ-plot, we see that it approximately follows a straight line 30x + 110. So we could say that this is probably a normal distribution, with a mean of 110, and a standard deviation of 30. (If we check the boxplot, we see that the mean actually is 110). However, we do have quite a small sample, so we cannot say that this will also hold for bigger, or other samples.

(iii) Trading volumes of dell shares:

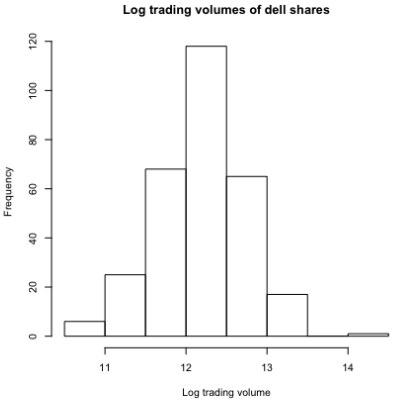


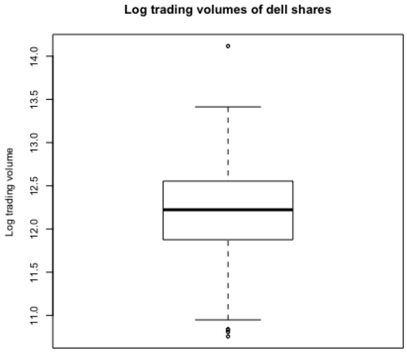


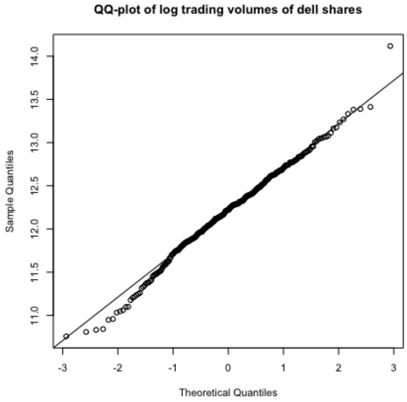


Obviously not from a normal distribution:
Although we do have a bell-shaped histogram, we have a big skew to the right. If we look at the boxplot and the qq-plot, you can see that we have a very high maximal value, and also a lot of values that lie above the 1,5 quantile range of the boxplot. The qq-plot follows a straight line for a while, but has a really heavy extreme right tail, and also a left tail that is 'lighter' than a normal distribution. This does not look like a normal distribution.

(iv) Log trading volumes of dell shares:





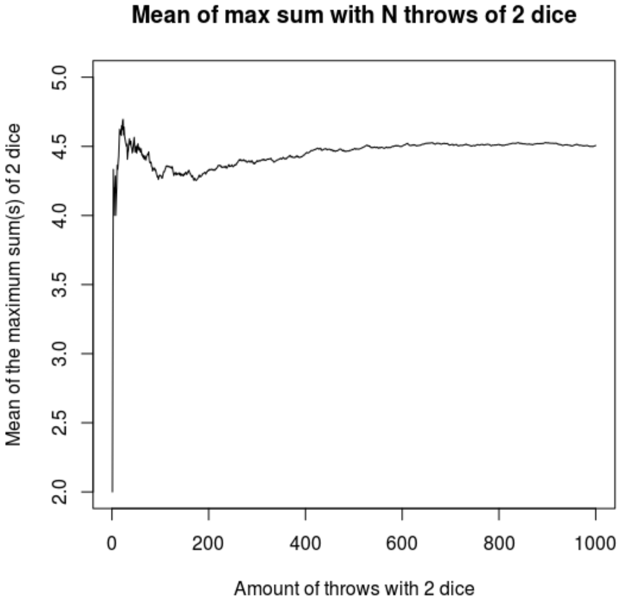


Normality cannot be excluded:
This seems like a normally distributed data set. We have a bell-shaped histogram, a nice and even boxplot (although we do have 3-4 more extreme values), and a qq-plot that follows approximately a straight line, which is probably 0,5x + 12,25. So we can say that this probably a normal distribution, with a mean of 12.25, and a standard deviation of 0.5. We do have quite a big sample of 300, so normality is highly probable.

Exercise 2.4

a) The law of large numbers says, for large sample sizes the mean of the random variable values of the all the samples will tend to the expectation of the random variable: E(X).

Random variable X is “the maximum on 2 dice”.
If we plot the mean of the current outcome of X+ all previous outcomes of X (if any) against the amount of throws of the dice, we obtain the following graph:



As seen in the graph, the value grows fast when N is small. However, when N grows the law of large numbers can be seen clearly. The mean grows until it stabilizes to a mean of approximately 4.5. If we calculate the expected value of X for the same dataset we will get the following table:

```
     randomVar probability expectation
[1,]         1       0.027       0.027
[2,]         2       0.083       0.166
[3,]         3       0.127       0.381
[4,]         4       0.192       0.768
[5,]         5       0.262       1.310
[6,]         6       0.309       1.854
```

The sum of these expectations is:

```
[1] 4.506
```

As described before “ The mean grows until it stabilizes to a mean of approximately 4.5”. The expectation of 4.506 is accurately close to where the mean tends to grow to.

b) (I) To get an approximate value of expectation of “the maximum of 5 dice” we ran the commands listed in the appendix, section b1.

The expectation of the random variable is

```
[1] 5.432212
```

The expectation of “the maximum of 5 dice” is approximately 5.43 with a dataset of 1000000 throws.

(II)

The probability of the event “the maximum of 2 dice” is a 3 is obtained by running the code of appendix section b2.

Table with the each possible outcome of the random variable “the maximum of 2 dice” and its probability respectively:

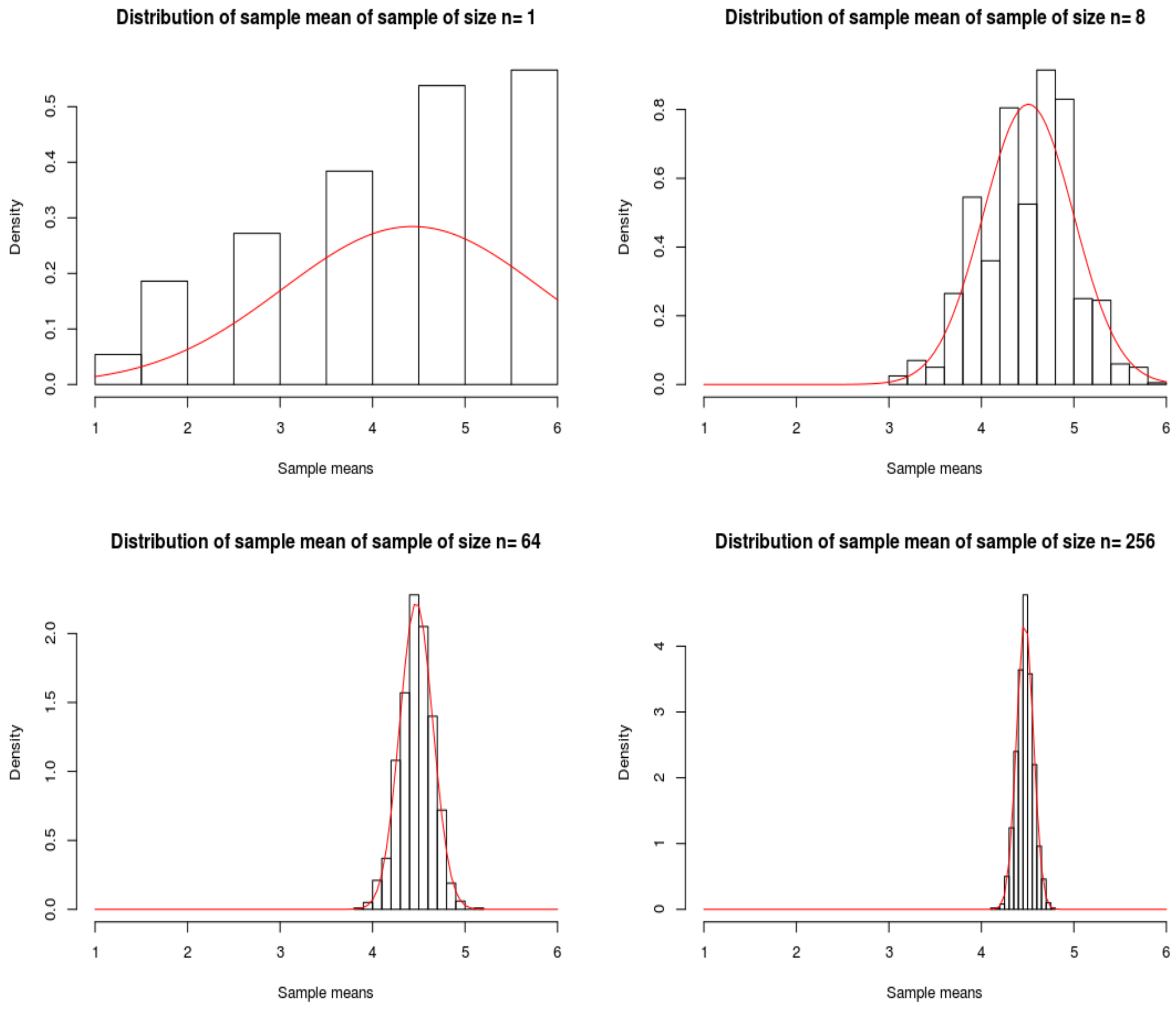
```
     randomVar probability
[1,]         1    0.027468
[2,]         2    0.083475
[3,]         3    0.138614
[4,]         4    0.194948
[5,]         5    0.250006
[6,]         6    0.305489
```

The probability of the event “the maximum of 2 dice is a 3” :

```
[1] 0.138614
```

The probability of the event “the maximum of 2 dice is a 3” is around 0.139.

c)



In the plots we plotted a histogram. On the Y-axis of the histogram the densities and on the X-axis the means of the samples. The sample size varies per histogram. The amount of samples taken is 1000 per histogram.

d)

“The Central Limit Theorem (CLT)
Take a sample of size n > 30 from a population with mean μ and standard deviation σ.
Then the sample mean X has approximately a normal distribution with mean μ and
standard deviation σ/√n .”

The central limit theorem says that for sample sizes n>30 the sample mean X has an approximate normal distribution. So if we would plot means from sample sizes n>30 they would plot in an approximate normal distribution. In the first plot of part c we could see how the values are distributed. Since the sample size is 1, the mean is exactly the value of the sample.

When plotting the means of samples of size 8 we clearly see a bell curve forming. However, the standard deviation is too large for the normal distribution as stated in the CLT (because n<30).

The means of samples with size 64 and 256 ,respectively, should be (according to the CLT) approximate normal distributions. As we can see, the bottom plots do meet al requirements of a normal distribution and therefore illustrate the central limit theorem.

Appendix

2.2a)

```
> tdistribution = rt(40, df=3)
> normaldistribution = rnorm(35,2,1)
> chisquared = rchisq(45, df=2)
> uniformdistribution = runif(55,-1,1)

> par(mfrow=c(2,2))
> qqnorm(tdistribution, main="(i) t-distribution")
> qqline(tdistribution)
> qqnorm(normaldistribution, main = "(ii) normal distribution")
> qqline(normaldistribution)
> qqnorm(chisquared, main = "(iii) chi-squared distribution")
> qqline(chisquared)
> qqnorm(uniformdistribution, main="(iv) uniform distribution")
> qqline(uniformdistribution)
```

b)

```
> klm = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 2/Excersises/klm.txt")
> iqdata = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 2/Excersises/iqdata2.txt")
> dell = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 2/Excersises/dell.txt")
> logdell = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 2/Excersises/logdell.txt")

> par(mfrow= c(1,3))
> hist(klm, main = "Boeing delivery times to KLM", xlab = "Delivery time in
number of days", ylab = "Frequency")
> boxplot(klm, range = 1.5, main = "Boeing delivery times to KLM", ylab =
"Delivery time in number of days")
> qqnorm(klm, main= "QQ-plot of Boeing delivery times to KLM")
> qqline(klm)

> par(mfrow= c(1,3))
> hist(iqdata, main = "IQ Data", xlab = "IQ score", ylab = "Frequency")
> boxplot(iqdata, range = 1.5, main = "IQ data", ylab = "IQ score")
> qqnorm(iqdata, main = "QQ-plot of IQ Data")
> qqline(iqdata)
> par(mfrow= c(1,3))
> hist(dell, main = "Trading volumes of Dell Shares", xlab = "Trading volume",
ylab = "Frequency")
> boxplot(dell, range = 1.5, main = "Trading volumes of Dell Shares", ylab =
"Trading volume")
> qqnorm(dell, main = "QQ-plot of trading volumes of Dell Shares")
> qqline(dell)

> par(mfrow= c(1,3))
> hist(logdell, main = "Log trading volumes of dell shares", xlab = "Log trading
volume", ylab = "Frequency")
> boxplot(logdell, range = 1.5, main = "Log trading volumes of dell shares",
ylab = "Log trading volume")
> qqnorm(logdell, main = "QQ-plot of log trading volumes of dell shares")
> qqline(logdell)
```

2.4a)

```
source("function2.txt")
par(mfrow=c(1,1))
maximaOfDice = maxdice(1000,2)
meanArray = c()
for(i in 1:length(maximaOfDice)){
  meanArray[[i]]= mean(maximaOfDice[1:i])
}
names(meanArray) = c(1:length(meanArray))
plot(meanArray,names.arg =names(meanArray),type = "l",ylab = "Mean of the
maximum of 2 dice",xlab = "Amount of throws with 2 dice", main = "Mean of max on
2 dice with N throws ",ylim = c(1,6))

probability = c()
randomVar = c()
uniques = unique(maximaOfDice,incomparables = FALSE)


for(i in sort(uniques)){
  probability = c(probability,length(which(maximaOfDice ==
i))/length(maximaOfDice))
  randomVar = c(randomVar,i)
}

expectation = c(randomVar*probability)
table = cbind(randomVar,probability,expectation)
sum(expectation)
```

b1)

```
source("function2.txt")
par(mfrow=c(1,1))
maximaOf5Dice = maxdice(1000000,5)
probability = c()
randomVar = c()
uniques = unique(maximaOf5Dice,incomparables = FALSE)

for(i in sort(uniques)){
  probability = c(probability,length(which(maximaOf5Dice ==
i))/length(maximaOf5Dice))
  randomVar = c(randomVar,i)
}

expectation = c(randomVar*probability)
cbind(randomVar,probability,expectation)
sum(expectation)
```

b2)

```
maximaOf2Dice = maxdice(1000000,2)
probability = c()
randomVar = c()
uniques = unique(maximaOf2Dice,incomparables = FALSE)

for(i in sort(uniques)){
  probability = c(probability,length(which(maximaOf2Dice ==
i))/length(maximaOf2Dice))
  randomVar = c(randomVar,i)
}
cbind(randomVar,probability)
probability[[3]]
```

c)

```
source("function2.txt")
par(mfrow=c(2,2))

meanArray1= c()
meanArray8= c()
meanArray64= c()
meanArray256= c()


for (i in 1:1000) {
  meanArray1[[i]] = maxdice(1,2)
  array8 = c(array8,maxdice(8,2))
  meanArray8[[i]] = mean(array8[[i]])
  array64 = c(array64,maxdice(64,2))
  meanArray64[[i]] = mean(array64)
  array256 = c(array256,maxdice(256,2))
  meanArray256[[i]] =  mean(array256)
}


hist(meanArray1,main = "Distribution of sample mean of sample of size n= 1",xlim
= c(1,6),xlab = "Sample means", ylab = "Density",freq = FALSE)
curve(dnorm(x, mean=mean(meanArray1), sd=sd(meanArray1)), add=TRUE,col = "red")

hist(meanArray8,main = "Distribution of sample mean of sample of size n= 8",xlim
= c(1,6),xlab = "Sample means", ylab = "Density",freq = FALSE)
curve(dnorm(x, mean=mean(meanArray8), sd=sd(meanArray8)), add=TRUE,col = "red")

hist(meanArray64,main = "Distribution of sample mean of sample of size n=
64",xlim = c(1,6),xlab = "Sample means", ylab = "Density",freq = FALSE)
curve(dnorm(x, mean=mean(meanArray64), sd=sd(meanArray64)), add=TRUE,col =
"red")

hist(meanArray256,main = "Distribution of sample mean of sample of size n=
256",xlim = c(1,6),xlab = "Sample means", ylab = "Density",freq = FALSE)
curve(dnorm(x, mean=mean(meanArray256), sd=sd(meanArray256)), add=TRUE,col =
"red")
```