Assignment 2

by

Ronnie Nhalevilo (2730319), Yonatan Getachew (2660122) and Dimitar Bachvarov (2728704)

- group 87

2.1 a)

Bold values are the values given in the question

	Positive test	Negative test	Total
Cancer patient	**0.95 * 0.4 = 0.38 (95%)**	0.05*0.4 0.02 (5%)	**0.4**
Non cancer patient	0.05*99.6 4.98 (5%)	**0.95 * 96.6 =94.62 (95%)**	99.6
Total	5.36	94.64	100

$P(positive) = \frac{5.36}{100} 0.05,$

The probability asked in exercise 1.3 is conditioned on whether the individual tested has a positive result, whereas the question asked here is testing the probability of getting a positive result from the general population.

b) $P(Cancer) = 0.004$, 0.4% of the population has cancer

$P(Positive) = 0.0536$

$P(Positive|Cancer) = 0.95$, derived from the 95% accuracy

$P(Cancer|Positive) = \frac{0.004 \times 0.95}{0.0536} \approx 0.07$, using Bayes' Theorem

c) The two events are dependent, as the probability of receiving a positive test is higher if the population is restricted to cancer patients (95%), whereas the probability of finding a cancer patient isn't as likely if the population is restricted to people with a positive test result (7%).

The latter does however have a higher likelihood of finding patient with cancer when compared to picking a random person from the whole population

2.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

xi=WT	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
P(X)	0,07	0,13	0,20	0,27	0,33	0,40	0,47	0,53	0,60	0,67	0,73	0,80	0,87	0,93	1,00
P(xi)	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07
P(X=xi)* xi	1,00	0,07	0,01	0,80	0,09	0,01	0,60	0,13	0,01	0,40	0,20	0,02	0,20	0,50	0,07
W^2P(X=xi) -E(X)^ 2	-1,74	-3,68	-5,48	9,60	8,07	6,67	5,40	4,27	3,27	2,40	1,67	1,07	0,60	0,27	0,07

a) The maximum waiting time is 15, assuming the train arrives every 15 min and leaves within the minute.

   We are counting full rounded minutes. We have a sample space of x= {0,1,2,3,4,5,6,,8,9,10,11,12,13,14} and we have 15 possibilities. But there is only 1/15 chances that we are arrive simultaneously with the bus , thus P=1/15=0.67

b) We have a sample space at first of x= {0,1,2,3,4,5,6,,8,9,10,11,12,13,14}, but since we miss it by 5 we exclude 0,1,2,3,4 and are left with a sample space of x={5,6,7,8,9,10,11,12,13,14,}, but the condition also states that the waiting times is of 10 minutes or less therefore we exclude 11,12,13,14 thus leaving us with {5,6,7,8,9,10}, and the P=6/15=0.4=40%

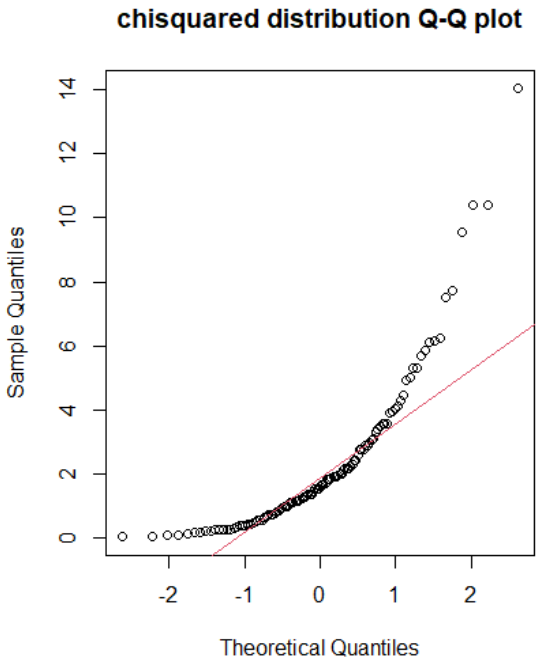
c) $E(X) = \sum_{i=1}^{15} P(X = x_i) \times x_i \approx 4.09$

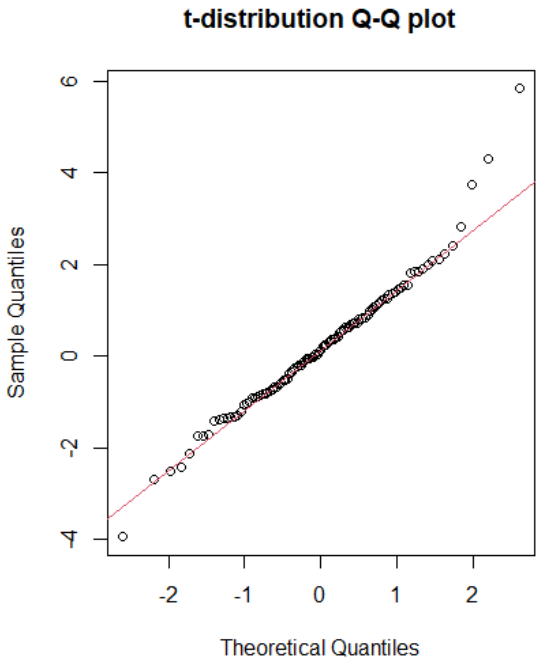
d) $Var(X) = \sum_{i=1}^{15} x_i^2 \times P(X = x_i) - \mu^2 \approx 32.43$

e)Based on the data set that we have , it can be deduced that we are presented with a uniform distribution in which the probability density function isn't dependant on the value of x, thus the probability of x is constant

3)

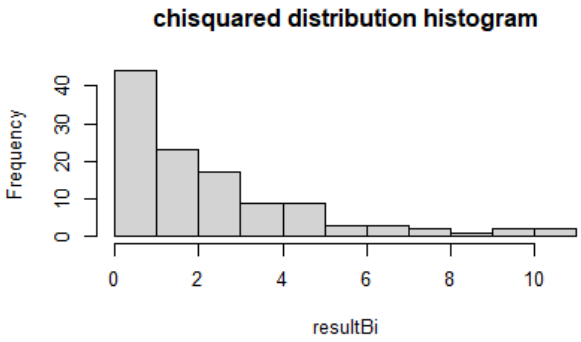
a)

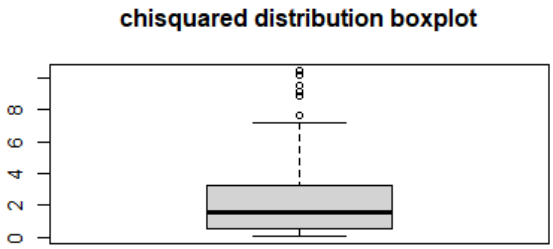


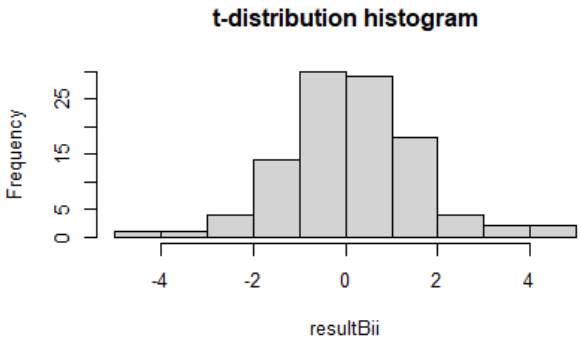


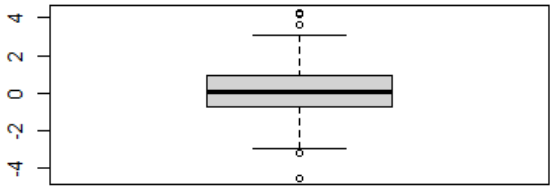
The usefulness of the normal distribution as a model distribution for the second qq-Plot is way more as we can see that if the data were normally distributed most of the circles are going to be on the straight red line (unlike the first plot where we need to find another more suitable distribution).

b)



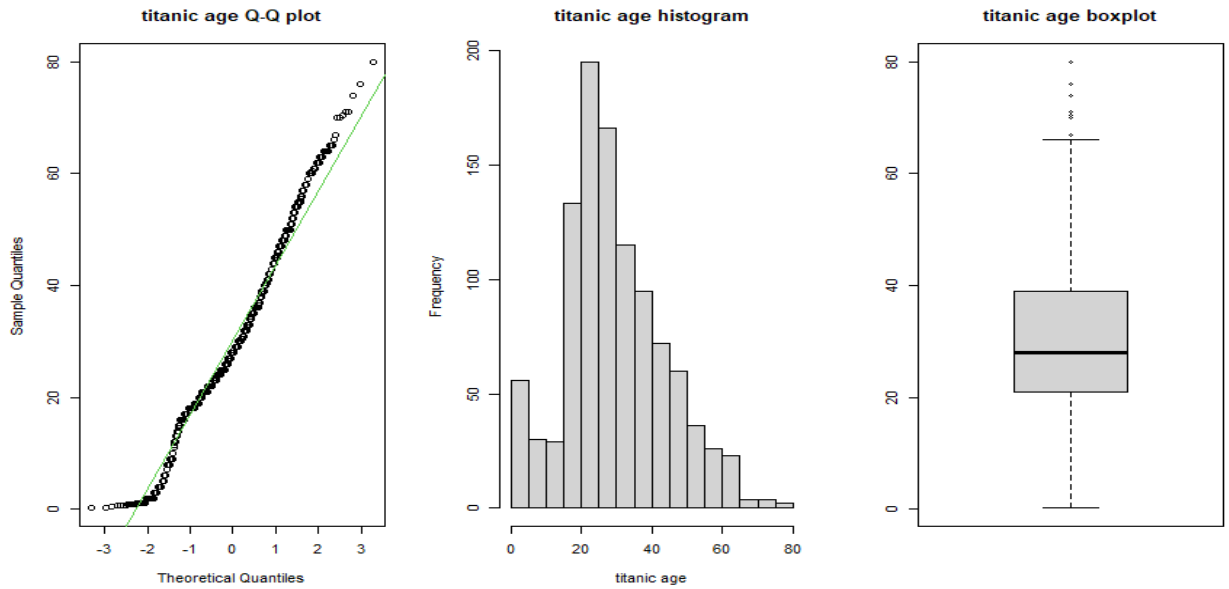




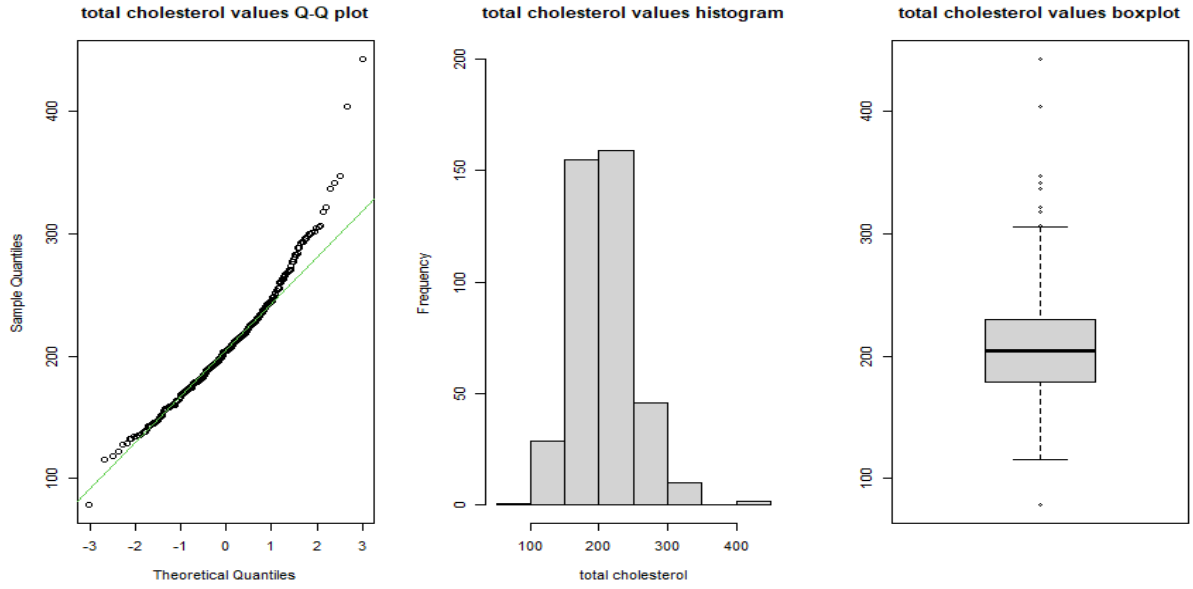


If we look at the pictures above we can see that we have very noticeable heavy tail in the first distribution which creates a lot of outliers and pushing the upper whisker/ max value away from the center while moving the median as well away from the “majority” (proving that the boxplot is not fully robust). On the other hand, we have a good symmetry in the second distribution which creates very correct mean in the boxplot while reducing the amount of outliers and creating more or less equal whiskers.

c)

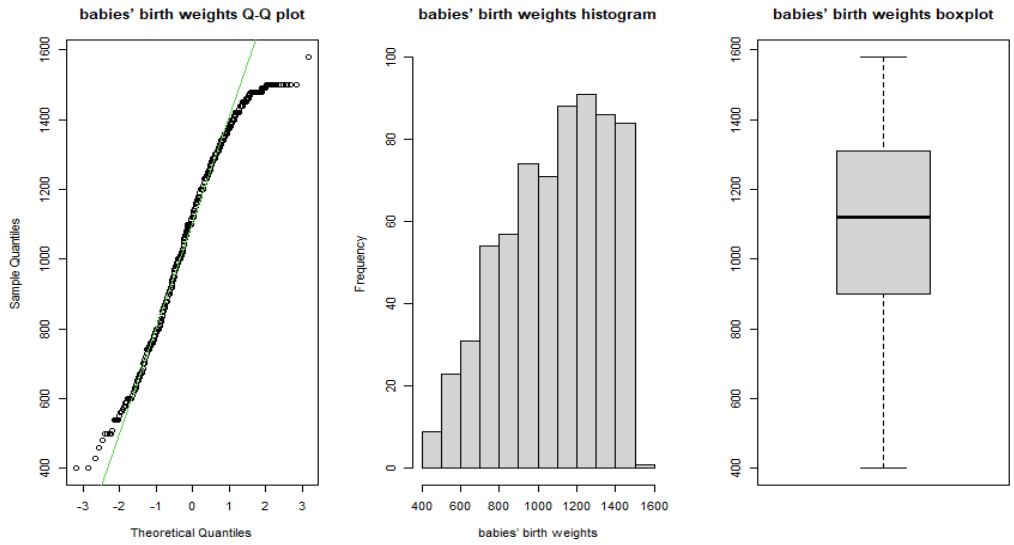


(“Normality cannot be excluded”) We think that is reasonable to believe that the data comes from a normal distribution as most of the circles are on top of the green line while the histograms show good symmetry and normal (expected for a normal distribution) representation. Not only that but the median seems to be close to the middle while the whiskers are more or less the same size.
Something peculiar is the fact that the median is slightly down from the centre as a result of the number of lower ages in the data is visible in the left tail of the histogram and at the beginning of the qq plot.



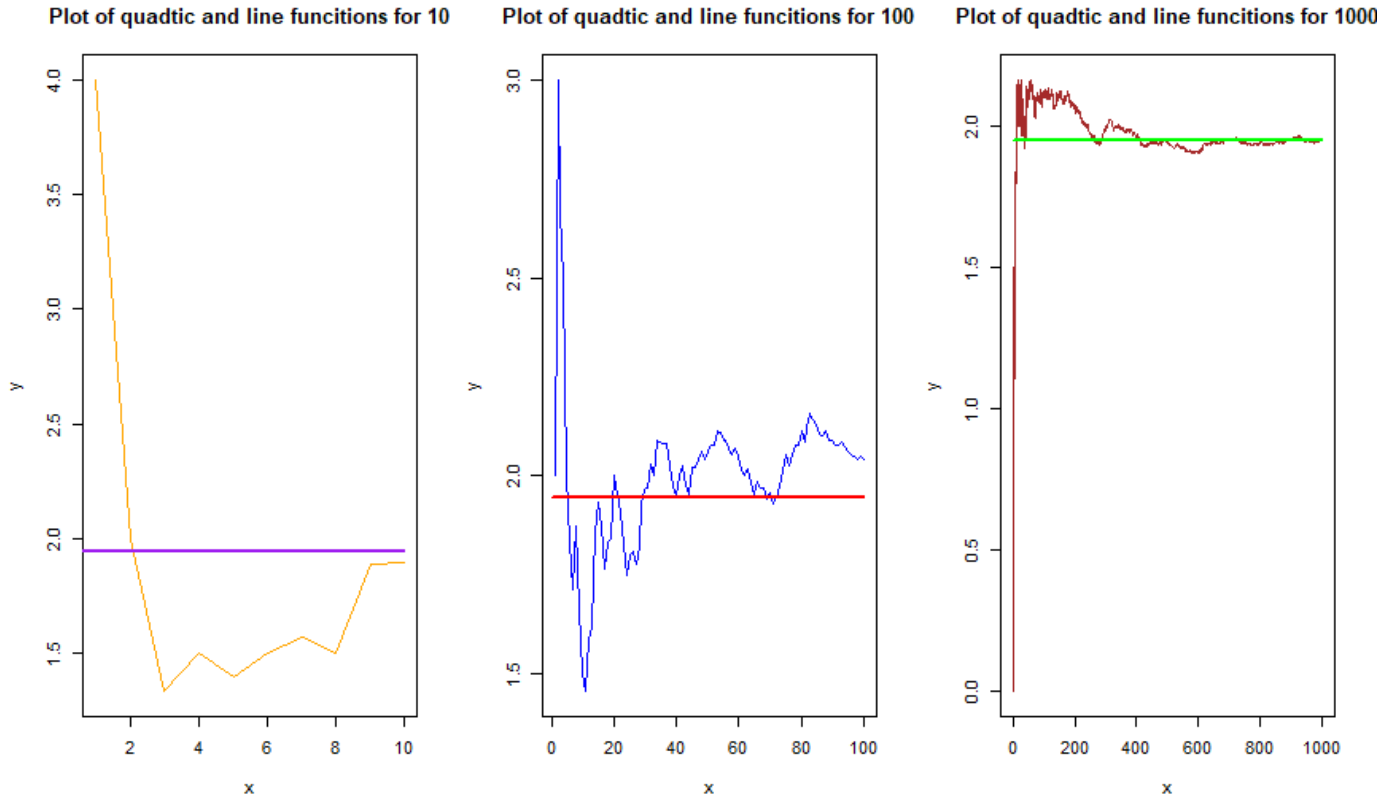
(“Normality cannot be excluded”) We think that is reasonable to believe that the data comes from a normal distribution as most of the circles are on top of the green line while the histograms shows

good symmetry and normal (expected for a normal distribution) representation. Not only that but the median seems to be close to the middle while the whiskers are more or less the same size. Something peculiar is the fact that we have a lot of outliers easily noticeable at the end of the qq plot which we can follow on the far right of the histogram and far out above in the boxplot.



("Obviously not from a normal distribution"). Although a good part of the qq plot shows us the fact that most of the circles are again on the line in the end we can see a big disconnection, which is further proven by the lack of symmetry in the histogram and the uneven whiskers on the boxplot.
Something peculiar is the fact that we can see the exact moment in the qq graph where the distribution stopped being normal which corresponds to the lack of downward movement in the histogram.

4.
a)



b)

The random values x in the absolute difference of two die rolls are 0, 1, 2, 3, 4, 5

Samples: {11,12,13,14,15,16,21,22,23,24,25,26,31,32,33,34,35,36,41,42,43,44,45, 46,51,52,53,54,55,56,61,62,63,64,65,66} = 36

p(0) = {11,22,33,44,55,66} = 6/36 ~ 0.17

p(1) = {21,32,43,54,65,12,23,34,45,56} = 10/36 ~0.28

p(2) = {13,24,35,46,31,42,53,64} = 8/36 ~0.22

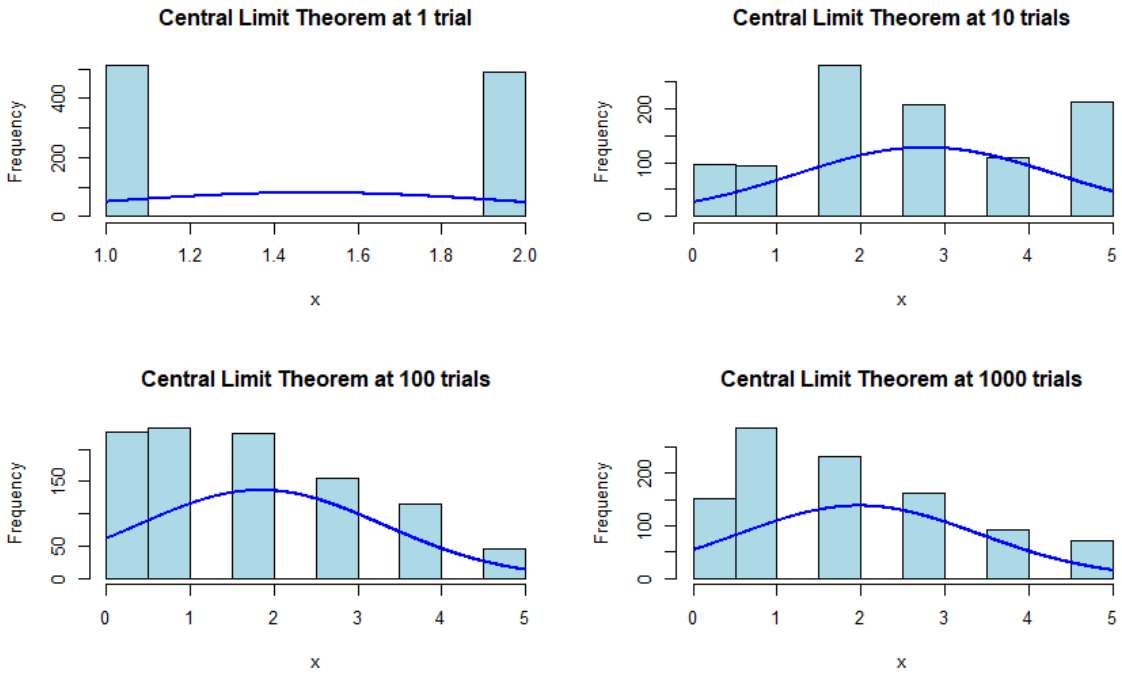
p(3) = {14,25,36,41,52,63} = 6/26 ~0.17

p(4) = {15,26,51,62} = 4/36 ~0.11

p(5) = {16,61} = 2/36 ~0.06

E(X) = 0*0.17 + 1*0.28 + 2*0.22 + 3*0.17 + 4*0.11+ 5*0.06 ~ 1.97

c)



d)

We know that the central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger. So as the number of trials grows we are getting closer and closer to the standard distribution. At first, we do not have the means of enough different samples so the distribution is stretched which changes with the number of samples that we continue to add.

Appendix:

```
setwd("C:/Users/dbach/OneDrive/Desktop/C++VU/Statistics/R statistics")

par(mfrow=c(1,2))
set.seed(187)

#2.3.a.i
resultAi = rchisq(115, df = 2)
qqnorm(resultAi, main = "chisquared distribution Q-Q plot")

#2.3.a.ii
resultAii = rt(105, df = 4)
qqnorm(resultAii, main = "t-distribution Q-Q plot")

par(mfrow=c(2,2))
#2.3.b.i
resultBi = rchisq(115, df = 2)
```

```
hist(resultBi, main = "chisquared distribution histogram")
boxplot(resultBi, main = "chisquared distribution boxplot")

#2.3.b.ii
resultBii = rt(105, df = 4)
hist(resultBii, main = "t-distribution histogram")
boxplot(resultBii, main = "t-distribution boxplot")

#2.3.c.i
par(mfrow=c(1,3))
titanic = read.csv("titanic3.csv")
qqnorm(titanic$age, main = "titanic age Q-Q plot")
qqline(titanic$age, col = 3)
hist(titanic$age, xlab = "titanic age", main = "titanic age histogram")
boxplot(titanic$age, main = "titanic age boxplot")

#2.3.c.ii
par(mfrow=c(1,3))
diabetes = read.csv("diabetes.csv")
qqnorm(diabetes$chol, main = "total cholesterol values Q-Q plot")
qqline(diabetes$chol, col = 3)
hist(diabetes$chol, ylim = c(0,200), xlab = "total cholesterol", main = "total cholesterol values histogram")
boxplot(diabetes$chol, main = "total cholesterol values boxplot")

#2.3.c.iii
par(mfrow=c(1,3))
vlbw = read.csv("vlbw.csv")
qqnorm(vlbw$bwt, main = "babies' birth weights Q-Q plot")
qqline(vlbw$bwt, col = 3)
hist(vlbw$bwt, ylim = c(0,100), xlab = "babies' birth weights", main = "babies' birth weights histogram")
boxplot(vlbw$bwt, main = "babies' birth weights boxplot")

source("function02.txt")
library(rcompanion)
par(mfrow=c(1,3))
#2.4.a
set.seed(87)
expectedValue = 1.9444

resultFor1 = diffdice(1)
```

```
resultFor10 = diffdice(10)
resultFor100 = diffdice(100)
resultFor1000 = diffdice(1000)

SUM10 = cumsum(resultFor10)
SUM100 = cumsum(resultFor100)
SUM1000 = cumsum(resultFor1000)
avg10 = SUM10/(1:10)
avg100 = SUM100/(1:100)
avg1000 = SUM1000/(1:1000)
options(scipen = 10)

plot(avg10, xlab="x", ylab="y", main ="Plot of quadtic and line funcitions for 10", type="l", col = "orange")
lines(c(0,10), c(expectedValue, expectedValue), col = "purple", lwd = 2)

plot(avg100, xlab="x", ylab="y", main ="Plot of quadtic and line funcitions for 100", type="l", col = "blue")
lines(c(0,100), c(expectedValue, expectedValue), col = "red", lwd = 2)

plot(avg1000, xlab="x", ylab="y", main ="Plot of quadtic and line funcitions for 1000", type="l", col = "brown")
lines(c(0,1000), c(expectedValue, expectedValue), col = "green", lwd = 2)

#2.4.c
sample1 = sample(resultFor1, 1000, replace = TRUE)
sample10 = sample(resultFor10, 1000, replace = TRUE)
sample100 = sample(resultFor100, 1000, replace = TRUE)
sample1000 = sample(resultFor1000, 1000, replace = TRUE)

par(mfrow=c(2,2))

plotNormalHistogram( sample1, prob = FALSE, main = "1")
plotNormalHistogram( sample10, prob = FALSE, main = "10")
plotNormalHistogram( sample100, prob = FALSE, main = "100")
plotNormalHistogram( sample1000, prob = FALSE, main = "1000")
```