Enrico Dal Pos (2708533) & Osama Obaidat(2704856) & Justus Beck(2717047) - Group 64

Assignment 2

Exercise 2.1

a) The difference between what is asked here and what is asked in the additional exercise, is that here we are asked the probability of a random person getting a positive test result from the whole population, considering false and true positives. While what's asked from the exercise is the probability that a random person has cancer given that they have a positive result from the test. Which is the probability that the person didn't get a false positive and actually has cancer.

P(A) + P(¬A) and P(B) + P(¬B) = 1. From there we can use the probabilities we already have and plot a table using the multiplication rule, $P(A \cap B) = P(A|B) \cdot P(B)$.
Let A = {has cancer} and B = {positive}:

P(A) = 0.004	P(B) = ?	P(B|A) = 0.95	P(¬B|¬A) = 0.95
P(¬A) = 0.996	P(¬B) = ?	P(¬B|A) = 0.05	P(B|¬A) = 0.05

$P(A \cap B) = P(B|A) \cdot P(A) = 0.95x0.004 = 0.0038$
$P(A \cap \neg B) = P(\neg B|A) \cdot P(A) = 0.05x0.004 = 0.0002$
$P(\neg A \cap B) = P(B|\neg A) \cdot P(\neg A) = 0.05x0.996 = 0.0498$
$P(\neg A \cap \neg B) = P(\neg B|\neg A) \cdot P(\neg A) = 0.95x0.996 = 0.9462$

	Positive (B)	Negative (¬B)	Total
Has cancer (A)	0.0038	0.0002	0.0040
No cancer (¬A)	0.0498	0.9462	0.0060
Total	0.0536	0.9464	1

So P(B) = 0.0536

b) Using Bayes' theorem we can compute the probability of a person having cancer given that they have a positive result, P(A|B).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

$$= \frac{0.95 \times 0.004}{0.95 \times 0.004 + 0.05 \times 0.996} = 0.0708$$

c) Possibilities being dependant means that means that $P(A \cap B) \neq P(A) \cdot P(B)$. $P(A \cap B)$ is also $P(A|B) \cdot P(B)$. So if P(A) = P(A|B) they are dependent. P(A) = 0.004 and P(A|B) =

0.0708, P(A|B) is greater than P(A) meaning that P(A) and P(B) are dependent. The influence of B (having a positive result) made A (having cancer) greater than a random person from the population having cancer. So the risk of having cancer given that your test result is positive is greater than a random individual.

Exercise 2.2

a) Between each bus there is an interval of 15 minutes and we're rounding the arrival times to the nearest minutes. x being the time to the next bus, there are 15 possibilities. As each of them happening is random, they have an equal chance of happening.

x	P(X = x)
0	1/15
1	1/15
2	1/15
3	1/15
4	1/15
5	1/15
6	1/15
7	1/15
8	1/15
9	1/15
10	1/15
11	1/15
12	1/15
13	1/15
14	1/15

b) There are 5 instances where the waiting time is less than 4 minutes. By making the waiting time for those 5 instances 0 it makes the possibility of getting to the bus with 0 waiting time higher with the possibility of the times where the person waited 4,3,2, and 1 minutes being added to the possibility of the waiting time being 0. Making the waiting time of 0 have the possibility of 1/3 (5/15).
The rest of the times to wait is when the person is waiting for at least 5 minutes, which is the rest of the probabilities, so 2/3 (10/15).

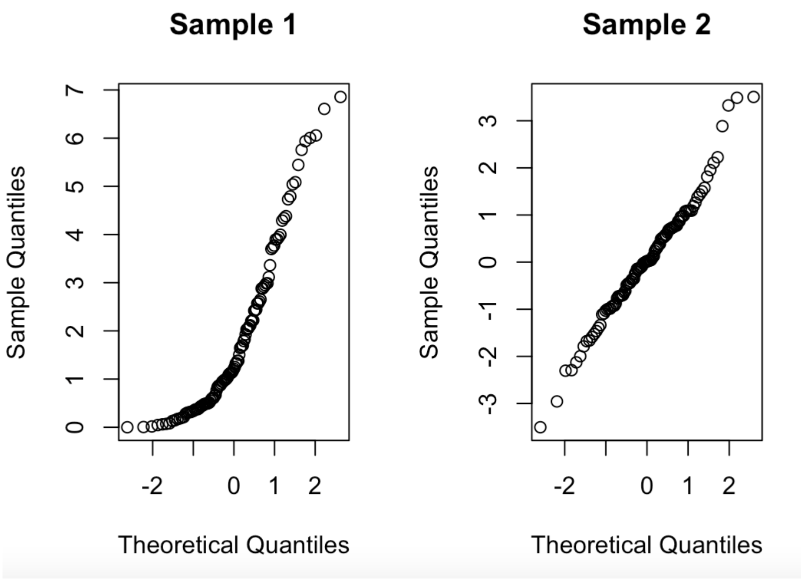
c) $E(X) = \sum x \cdot P(X=x) =$ $(0\times5/15)+(1\times0/15)+(2\times0/15)+(3\times0/15)+(4\times0/15)+(5\times1/15)+(6\times1/15)+(7\times1/15)+(8\times1/15)+(9\times1/15)+(10\times1/15)+(11\times1/15)+(12\times1/15)+(13\times1/15)+(14\times1/15)+(15\times1/15) =$ 7.8

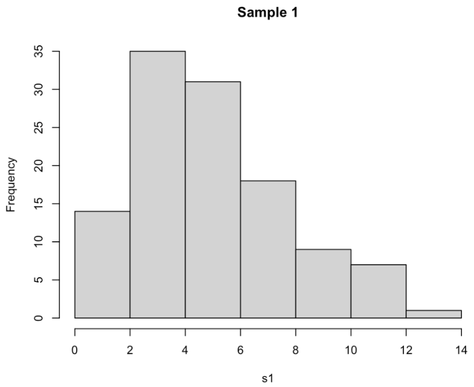
d) $Var(X) = \sum(x-E(x))^2 \cdot P(X=x) =$

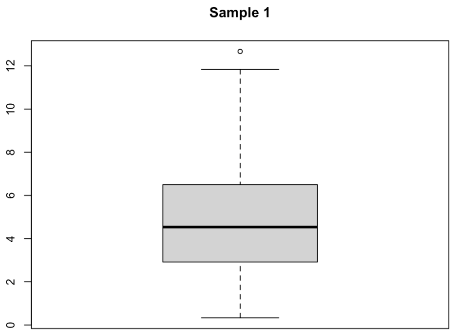
e) The value variance gets closer to the expected value as the data points increase.

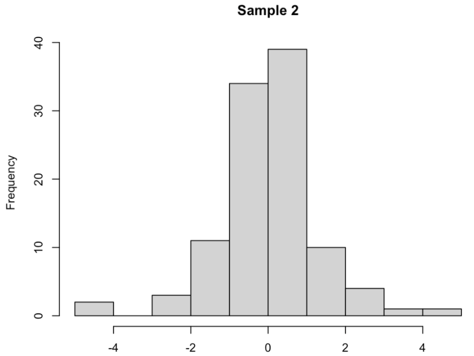
Exercise 2.3

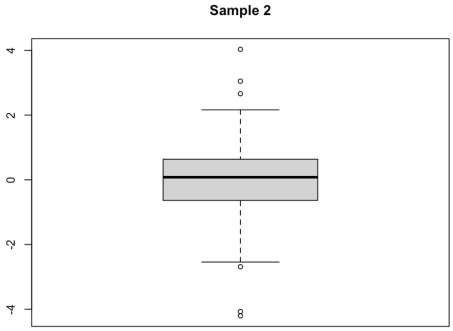
a) The qqplot is very useful to determine if a data is normally distributed or not by checking if it follows a straight line .The plot from the sample1 doesn't represent normal distribution and the second does represent a normal distribution. since it follows a straight line.









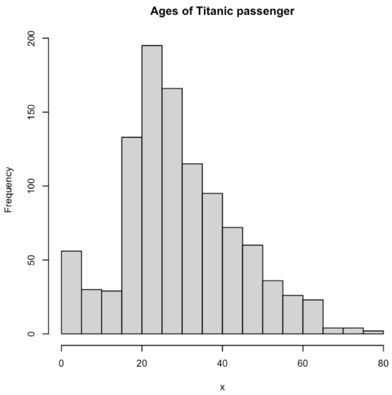


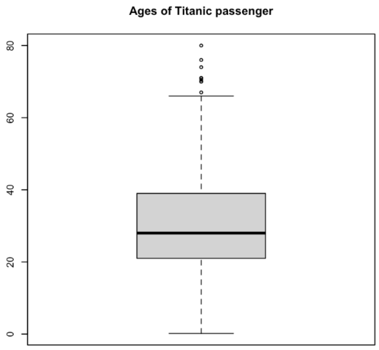
b)

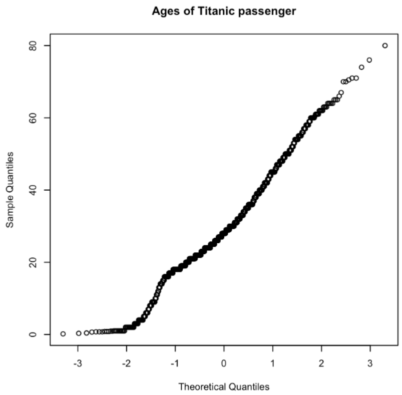
In the Histogram of sample1, we can see that is right-skewed(not symmetric) and doesn't have a bell shape, this is clearly visible on the boxplot since the median is in the lower part of the plot. From both the boxplot and the histogram the data doesn't look like is heavy-tailed.

In the Histogram of sample2, we can see it is symmetric with a bell shape. On the boxplot this is noticeable from the fact that the median is pretty centered and the boxplot quartile 1 and 3 are are both close to the median in the center. Around 2 and 4 there are some visible outliers that we can also clearly see on top of the boxplot.

c) i)

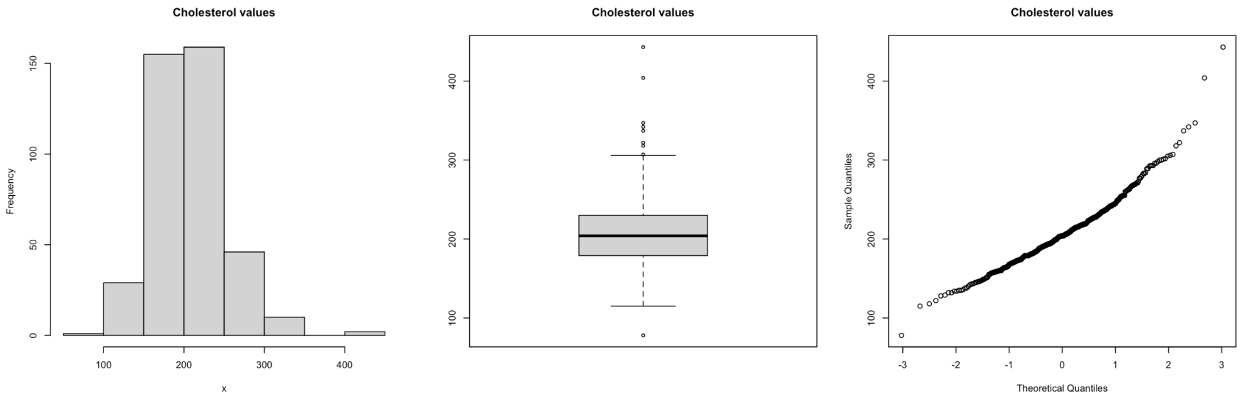






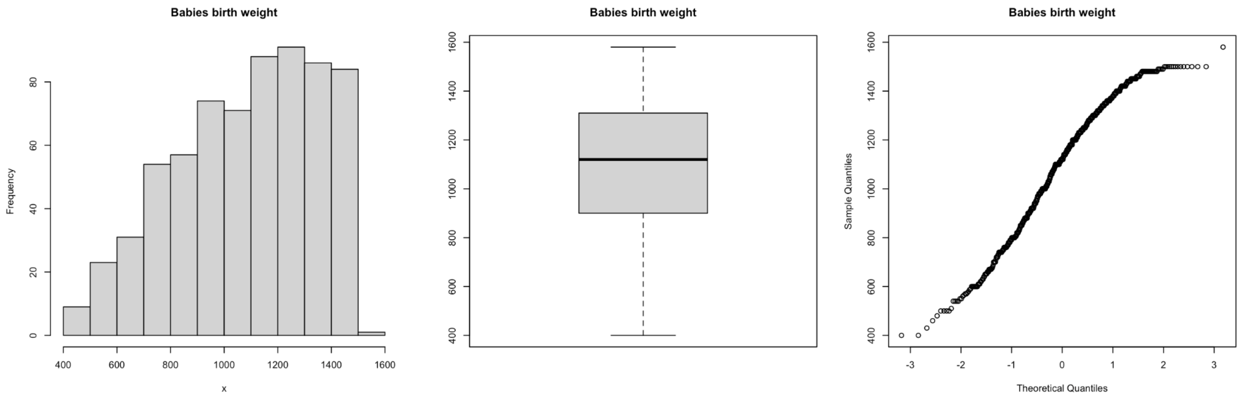
In these plots the data doesn't look like is from a normal distribution since the histogram is definitely not symmetrical(is right-skewed) and the mean is not really in the center. Also, from the QQ plot we can notice that that the points don't follow a straight line.

ii)



In this dataset normality cannot be excluded since the plot is symmetrical, the mean in the boxplot is very centered and the qqplot looks like it could follow a straight line.

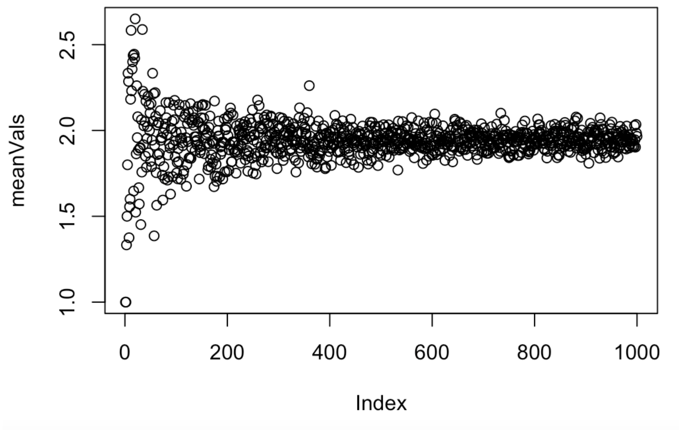
iii)



The distribution of this dataset is definitely not normal since the histogram is very left-skewed and the qqplot doesn’t really follow a straight line.

Exercise 2.4

a)



Appendix

Exercise 2.3

Relevant code:

a)

```
par(mfrow=c(1,2))
x1 = rchisq(115, df = 2)
x2 = rt(105, df = 4)

qqnorm(x1, main = "Sample 1")
qqnorm(x2, main = "Sample 2")
```

b)

```
s1 = rchisq(115,df=5)
s2 = rt(105,df=4)

par(mfrow= c(2,2))
hist(s1, main = "Sample 1")
boxplot(s1, main = "Sample 1")

hist(s2, main = "Sample 2")
```

```
boxplot(s2, main = "Sample 2")
```

c)

```
titanic <- read.csv("titanic3.csv")
diabetes <- read.csv("diabetes.csv")
vlbw <- read.csv("vlbw.csv")

plotData <- function(x, title) {
  par(mfrow= c(1,3))
  hist(x, main = title)
  boxplot(x, main = title)
  qqnorm(x, main = title)
}

plotData(titanic$age, "Ages of Titanic passenger")
plotData(diabetes$chol, "Cholesterol values")
plotData(vlbw$bwt, "Babies birth weight")
```

Exercise 2.4

a)

```
source("/Users/anonymous/Documents/Uni/Stats/function02.txt")
meanVals = numeric(1000)
for(i in 1:1000) {
  meanVals[i] = mean(diffdice(i))
}
plot(meanVals)
```

b)