Statistical Data Analysis
Assignment 2

Matt Kuodis & James Zoryk.
2640428 : 2663347
Group: 36
Due Date: 01.03.2022.

Question 2.2.a.

A QQ-plot yields a method to judge whether the data comes from a certain distribution by only looking at the plot itself. When the data and distribution are roughly equivalent the QQ-plot should yield a straight line of the form $y = x$. Moreover, any deviations from a line indicates that there are differences between the two sets of data.

In this exercise we plot the QQ-plot for; standard normal against exponential with rate 3, normal with mean -1 and variance 9 against t_4, and t_6 against chi-squared with 3 degrees of freedom. The plots of which can be found in Figure 1 respectively.

For the standard normal against exponential with rate 3, we observe that the qq-plot (Figure 1:Left) is approximately the form of an exponential curve, namely that the gradient increases as we move along the x -axis. This indicates that the data is skewed compared to the other data set. Therefore we can clearly see from the plot see that thees two distribution are not similar Add text talking about the tails.

For the normal with mean -1 and variance 9 against t_4, we can see from the qq-plot (Figure 1:Mid) that it takes the approximate form of a rotated logistic curve. Namely that there is a portion in the centre of the curve that is roughly a straight line, however at the tails we see that the gradients of the curve grows exponential as we move away from the centre in either direction. This form of curve, typical occurs when one data set has more extreme values than the other data set. Hence we can confidently say that these two distribution are not similar.

Lastly, for t_6 against chi-squared with 3 degrees of freedom. As we can see from the plot (Figure 1:Right), that it is similar that of the first plot since we have exponential grow for the left hand-side of the curve. However, it can be noticed that the gradient of the curve becomes approximately constant as we move along the x-axis, namely that the curve becomes a straight line. This suggests that the data skewed in comparison to each other.

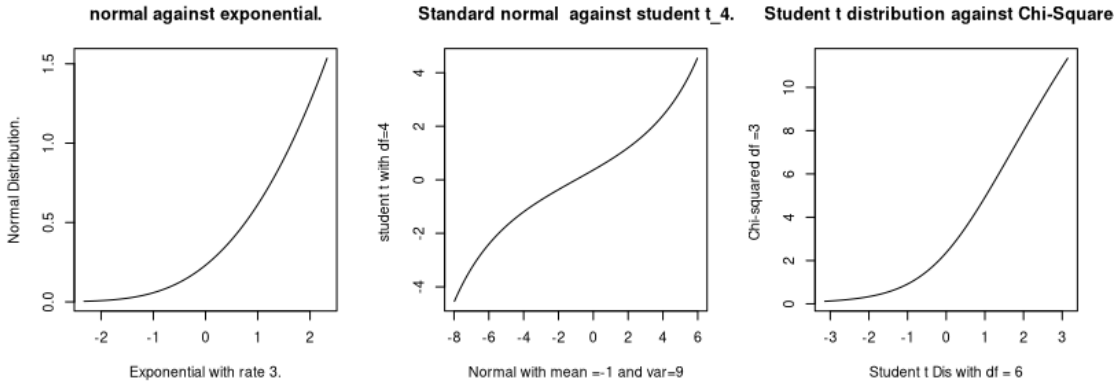

Figure 1: QQ-plots

Part b.

Here we shall investigate the data 'sample2022a' which is from 'sample2022.txt'. Our first step is to generate a numerical analysis of the data, which can be found in Table 1 . The purpose of this is to understand how the data is distributed, such as if it is skewed or not and the generally shape.

Mean:	7.04
Min:	5.03
1st qu.:	5.49
Median:	6.22
3rd qu.:	7.39
Max:	16.24

Table 1: Numerical analysis for sample2022a

Here we observe that the median of the data is less than the mean, which indicate that sample2022a has a positive skew. This can also be seen in the histogram in figure 2.

The histogram and numerical analysis helps us to narrowing down a suitable choice for selecting a distribution to compare it with in a QQ-plot. From personal experience we choose to investigate the exponential, Cauchy and binomial distribution since they

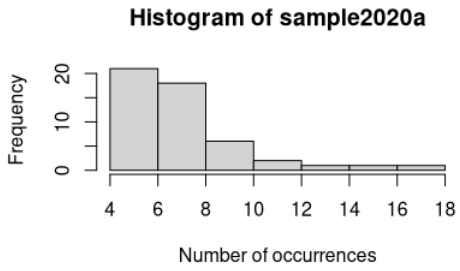

Figure 2: QQ-plots

have similar characteristics exhibited from our data. The results can be seen in figure 3.

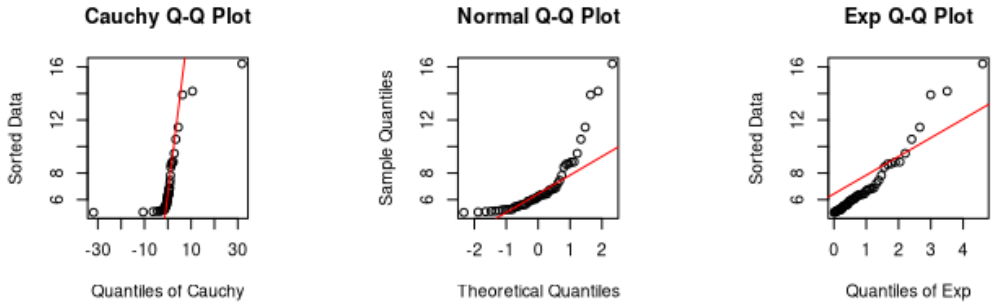

Figure 3: QQ-plots for the data 'sample2022a'

From Figure 3 we observe that both the QQ-plots against the Exponential and Normal distribution have large deviations from the line, with the Normal qq-plot having a large right hand tail. Moreover, we also see that these two QQ-plots are curved, which indicates that our date is not of the distribution of either the Normal or the Exponential.

The Cauchy QQ-plot can be observed having a majority of the data point on the line, indicating a good fit. However, a few points on both tails indicates slight differences to the extreme values of our data the the theoretical Cauchy distribution (possible max/min, aka range of data sets).

Therefor we conclude that there is evidence to suggest that the data from '2022a' is of the form of a Cauchy distribution with location 0 and scale 1.

Question 2.3.a.

Here, we shall explore the data 'sample2022b' from 'sample2022.txt', in order to find a suitable distribution family for the data. We begin by plotting a histogram of the data, which looks like a skewed Normal distribution. Therefore our next step is to see if we can transform the data of 'sample2022b' to be less skewed. We try by taking the log and the square root of the data, and comparing the histogram in Figure 4. Here we see that the transformation of the squared root is less skewed than the original data.

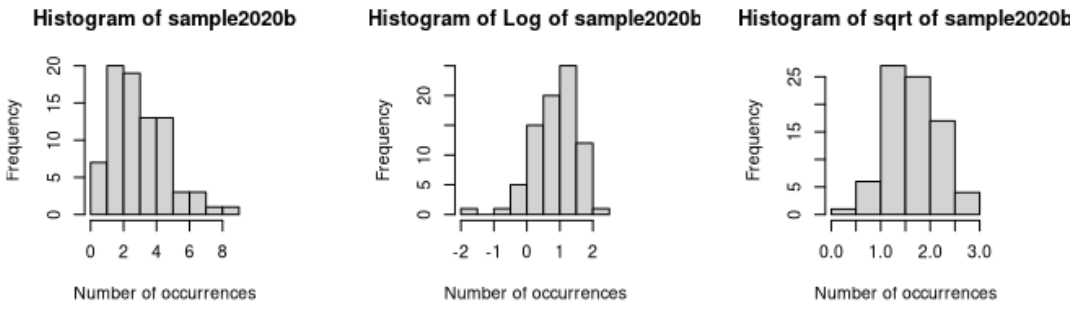

Figure 4: Comparison of histograms of the data 'sample2022b' and its transformation by *log* and sqrt.

Our hypothesis is that the data is from a Normal distribution, so we plot a QQ-plot against while also choosing another similar distribution, the Cauchy distribution to compare against. (5).

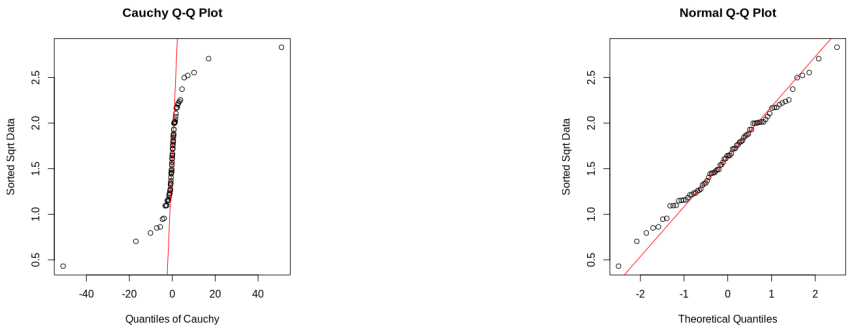

Figure 5: The QQ-plots for the square root of the data 'sample2022b'

In the Cauchy QQ-plot most of the data plots are on the line, however we see that the data points form a rotated logistic curve meaning that one of the two distributions has more extreme values than the other data set. While in the case for the Normal distribution we see that the data points form a roughly line with a constant gradient and also the deviation from the line is relatively small at the tails compared to the Cauchy distribution.

Hence, there is evidence to suggest that the data 'sample2022b' comes from the normal family with parameters $N(0, 1)$.

Part b.

Here we conduct both a Kologorov-Smirnov (KS) and Chi-Square (CS) test at a test level $\alpha = 5\%$ to see whether the data from 'sample2022b' originates from the Gompertz distribution. The p-value for the KS test is $p_{ks} = 0.052088$ hence we have $p_{ks} > \alpha$ .Thus in the KS test there is enough evidence to reject H_0, in other words 'sample2022b' is not from the Gompertz distribution. On the other hand, the p-value for the CS test is $p_{cs} = 0.036690 < \alpha$, hence there is not enough evidence to reject H_0, meaning that there is evidence to suggest that 'sample2022b' is from the Gompertz distribution.

Here we note that both test give a contradiction to each other. The differences can be attributed to the fact that the KS test looks at the data in continuous manner(Empirical distribution function), while the Chi-Square test puts the data into allocated bins. A result of which is that we lose some of the information about the data in the chi-square test.

Therefore it is possible to have contradiction with both of these test results.

Question 2.4.

For this question we had to analyze data regarding a person's weight, height, and ankle girth, and see whether they shared a similar distribution. The weight and height factors are often combined into the BMI (Body Mass Index) which is given by $\frac{weight}{height^2}$ with the weight given in kilograms, and the height in meters.

We extracted the relevant data from the given table, and plotted the BMI and ankle girth as histograms and boxplots. These give us the following figures.

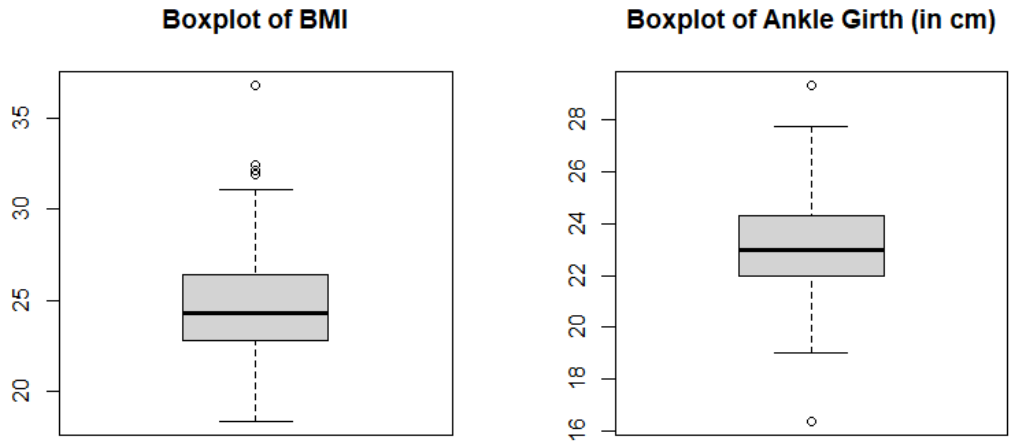

Figure 6: The boxplots for the BMI and Ankle Girth.

We note that the plots have very similar shapes, and both the interquartile range and the whiskers indicate large similarities in the distribution of the two samples. The outliers in the BMI boxplot are of noteworthy difference.

The histograms give us a general look at the distribution of the samples, but share less similarities between them than the boxplots do.

The histograms in figure 7 however show us that both plots look somewhat normal, and are in fact a bit right-skewed. The outliers in the BMI boxplot are translated to the histogram as an extension of the tail. In comparison, the Ankle Girth histogram is tightly nested around the center.

To see whether the plots come from the same location-scale family, regardless of distribution, we compare them using a QQ-plot. Overall, the sample data forms a clear line, meaning that they share a location-scale family.

We can use the data summary to generate some values for our data that might give us more constructive data to determine the distribution. From the plots it already strongly resembles the normal distribution, which is also supported by the clustering of the data points near the middle in the QQ-plots.

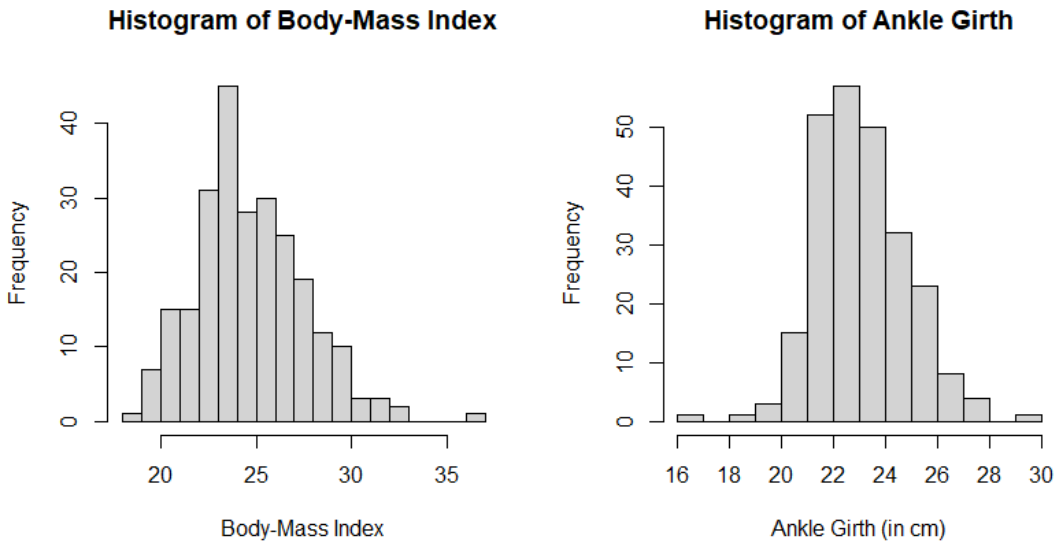

Figure 7: The histograms for the BMI and Ankle Girth.

Mean:	24.71
Min:	18.35
1st qu.:	22.81
Median:	24.30
3rd qu.:	26.43
Max:	36.82

Mean:	23.16
Min:	16.40
1st qu.:	22.00
Median:	23.00
3rd qu.:	24.30
Max:	29.30

Table 2: Numerical analysis for BMI sample (left), and Ankle Girth sample (right).

We see in the numerical summary that the interquartile range is very narrow compared to the total range of the data. The distances between the Min and $1st\ qu$ are very close to those of Max and $3rd\ qu$ in our Ankle Girth sample. They veer off more strongly for the BMI data, but this is also apparent from the plots.

To further solidify our intuition, we performed a Shapiro-Wilk test on both data samples. For the BMI sample, the test values are $W = 0.97962$ and $p = 0.001285$, strongly implying that the sample distribution is normal. For the Ankle Girth, we get the values $W = 0.98169$ and $p = 0.002832$, also strongly implying normality.

For a smaller sample of the BMI, namely the first 50 samples, we get the values $W = 0.96073$ and $p = 0.09537$ which does not guarantee any certainty for normality. This means that the data approaches the normal distribution as the sample size increases. The histograms comparing the first 50 samples, and all of the BMI samples further show the smaller sample size lacking any convincing structure in its distribution, especially when compared to the normal distribution.

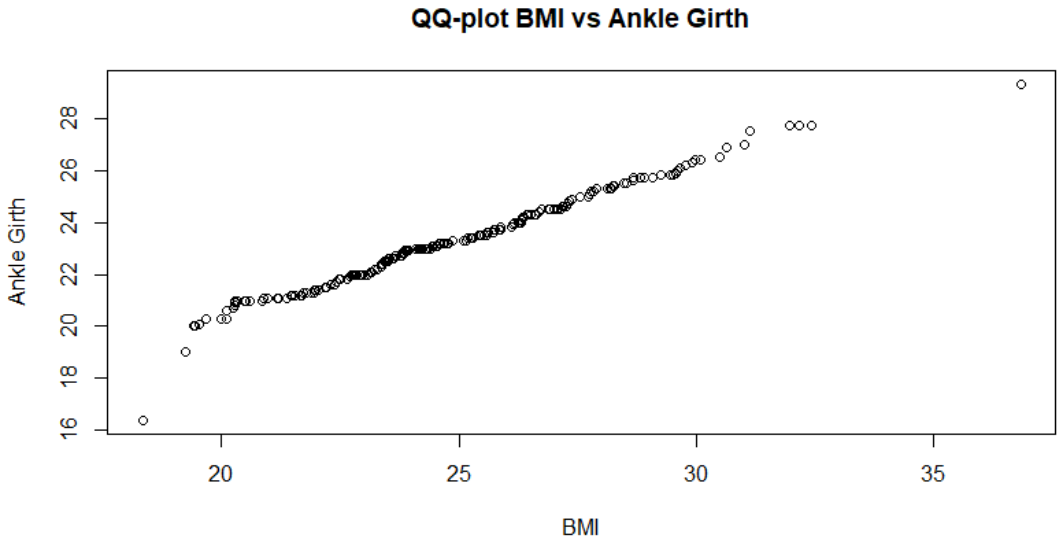

Figure 8: The QQ-Plot for the BMI and Ankle Girth.

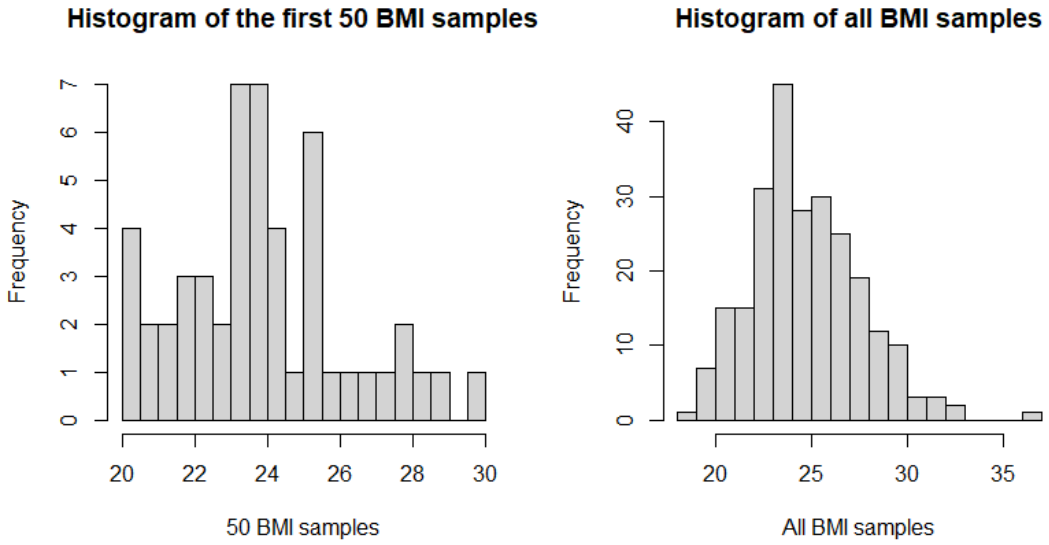

Figure 9: The histogram plots for the BMI sample: 50 (left), and total (right).

R code for exercise 2.2a

```
#########################################################
# SDA HW2
# Exercise: 2a
#########################################################
# set working DIR
setwd("~/Maths/SDA/Assignment_2")
# Pull data from disk
assignment_data = source("sample2022.txt")
source("functions_Ch3.txt")
# Set up a row of 3 figures
par(mfrow=c(1,3))
par(pty="s")
#Create an array of values
s = seq(1/100, 99/100, length.out = 10000)

## Start ploting
# qq-plot: Norm  v Exp rate=3
x = qnorm(s)
y = qexp(s, rate=3)
plot( x,y , type="l",
      main="normal against exponential.",
      ylab="Standard Normal.",
      xlab="Exponential with rate 3.")
# qq-plot: Norm mean=-1 sd=3 v t_4 var=9
x2 = qnorm(s, mean=-1, sd=3)
y2 = qt(s, df=3)
plot(x2, y2, type="l",
     main="Standard normal against student.",
     ylab="student t with df=4 ",
     xlab="Normal with mean =-1 and var=9")
# qq-plot: t_6  v Chi-square df=3
x3= qt(s, df=6)
y3 = qchisq(s, df=3)
plot(x3,y3, type='l',
     main="Student t distribution against Chi-Squared.",
     ylab="Chi-squared df =3 ",
     xlab="Student t Dis with df = 6")
```

code/SDA_hw2_ex2a.R

R code for exercise 2.2b

```
#########################################################
# SDA HW2
# Exercise: 2b
#########################################################
# set working DIR
setwd("~/Maths/SDA/Assignment_2")
# Pull data from disk
assignment_data = source("sample2022.txt")
source("functions_Ch3.txt")
# Grab data set sample2022
data_seta <- assignment_data[[1]][[1]]
# give numerical summary
summary(data_seta)
# Histogram of data set a
par(mfrow=c(1,1))
hist(data_seta,
     main="Histogram of sample2020a",
     xlab="Number of occurrences")
# set row of 3 figures
par(mfrow=c(1,3))
par(pty="s")
# QQ-plot against cauchy dis
qqcauchy(data_seta)
qqline(data_seta, col="red")
# QQ-plot against norm
qqnorm(data_seta)
qqline(data_seta, col="red")
# QQ-plot against exp
qqexp(data_seta)
qqline(data_seta, col="red")
```

code/SDA_hw2_ex2b.R

R code for exercise 3

```
#####################################################
# SDA HW2
# Exercise: 3
#####################################################
# Pull data from disk
setwd("~/Maths/SDA/Assignment_2")
assignment_data = source("sample2022.txt")
source("functions_Ch3.txt")
# Grab data set sample2022b
data_setb <- sample2022[["sample2022b"]]

####### Part A ########################
# Numerical analysis of the data set
summary(data_setb)
data_b_sqrt = sqrt(data_setb)
data_b_log = log(data_setb)
summary(data_b_sqrt)
summary(data_b_log)

# set up row of 3 figures
par(mfrow=c(1,3))
par(pty="s")
# Histograms
hist(data_setb,
     main="Histogram of sample2020b",
     xlab="Number of occurrences")
hist(data_b_log,
     main="Histogram of Log of sample2020b",
     xlab="Number of occurrences")
hist(data_b_sqrt,
     main="Histogram of sqrt of sample2020b",
     xlab="Number of occurrences")

# set up row of 2 figures
par(mfrow=c(1,2))
# par(pty="s")
# qq plot of data
qqcauchy(data_setb)
qqline(data_setb, col="red")
qqnorm(data_setb)
qqline(data_setb, col="red")
# qq plot of sqrt of data
qqcauchy(data_b_sqrt,
         ylab = "Sorted Sqrt Data")
qqline(data_b_sqrt, col="red")
qqnorm(data_b_sqrt,
       ylab = "Sorted Sqrt Data")
qqline(data_b_sqrt, col="red")

####### Part B, C ########################
# set test level
test_alpha = 0.05
# KS test
```

```
ks_test_data = ks.test(data_setb, pgompertz, shape=1, scale=0.2)
KS_score= ks_test_data[[1]]
KS_p_value= ks_test_data[[2]]
if(KS_p_value > test_alpha )  KS_reject=FALSE else KS_reject=TRUE


# chisq test
range(data_setb)
length(data_setb)
#compute the breaks with
b = quantile(
    qgompertz(seq(1/100, 99/100, length.out = 10000),shape=1, scale=0.2  ),
    seq(0,1,1/10))
b[1] <- 0
b[length(b)] <- Inf

chisq_test_data = chisquare(
  data_setb, pgompertz,length(b),0,9,b,shape=1, scale=0.2)
chisq_breaks = b
chisq_score = chisq_test_data["chisquare"]
chisq_p_value = chisq_test_data["pr"]
if(chisq_p_value > test_alpha )  chisq_reject=FALSE else chisq_reject=TRUE

# put allcomputed variables into a list
mylist=list( KS_score, KS_p_value, KS_reject, chisq_breaks, chisq_score,
    chisq_p_value,chisq_reject)

# Save to disk the return of the list
save(mylist, file="~/Maths/SDA/Assignment_2/myfile1_36.RData")
```

code/SDA_hw2_ex3.R

R code for exercise 4

```
# Extract the data from the given table
body_data = read.table("body.dat.txt")

# Extract significant information from the relevant columns
# Note that we extract the first 247 rows to study the male sample only
ankle_girth = body_data[1:247,20] # in cm
body_weight = body_data[1:247,23] #in kg
body_height = body_data[1:247,24] #in cm

# Compute the BMI
bmi = body_weight/((body_height/100)^2)

# Give a numerical summary
summary(bmi)
summary(ankle_girth)

# Set up the plot parameters
par(mfrow=c(1,2))

# Boxplots
boxplot(bmi,
        main="Boxplot of BMI")
boxplot(ankle_girth,
        main="Boxplot of Ankle Girth (in cm)")

# Histograms
hist(bmi, breaks=15,
     main="Histogram of Body-Mass Index",
     xlab="Body-Mass Index")

hist(ankle_girth, breaks=15,
     main="Histogram of Ankle Girth",
     xlab="Ankle Girth (in cm)")

# QQ-plot of the two samples
par(mfrow=c(1,1))
qqplot(bmi, ankle_girth,
       main="QQ-plot BMI vs Ankle Girth",
       xlab="BMI",
       ylab="Ankle Girth")

# Tests for normality
shapiro.test(bmi)
shapiro.test(ankle_girth)

# Plot comparisons between the first 50, and the total sample size
par(mfrow=c(1,2))

bmi_50 = bmi[1:50]

hist(bmi_50, breaks=15,
     main="Histogram of the first 50 BMI samples",
     xlab="50 BMI samples")
```

```
hist(bmi, breaks=15,
     main="Histogram of all BMI samples",
     xlab="All BMI samples")

# Normality test for the first 50 samples
shapiro.test(bmi_50)
```

code/SDA_hw2_ex4.R