SDA- Assignment 1

Jose Chacon (2699643)

1 Exercise 2.2

a) Comment on the heaviness of the QQ plots

The following section of the report will show the QQ plot of several pairs of distribution. In this particular analysis, we are focusing on the form of the QQ plot, especially on the heaviness of the QQ plots' tails to understand the relationship between the 2 distributions. In order to create the QQ plots, we first created a sequence from 0.01 to 0.99 at a step of 0.01, which is the sequence used to obtain the true quantiles of each distribution. Additionally, we used the given local functions (*qlnorm*, *qnorm*, *qexp*, and more) and the *plot* function to represent this QQ=plots graphically. First, we will show the resulting QQ plots of each pair of distributions. Then a comment on the heaviness of the tails will follow.

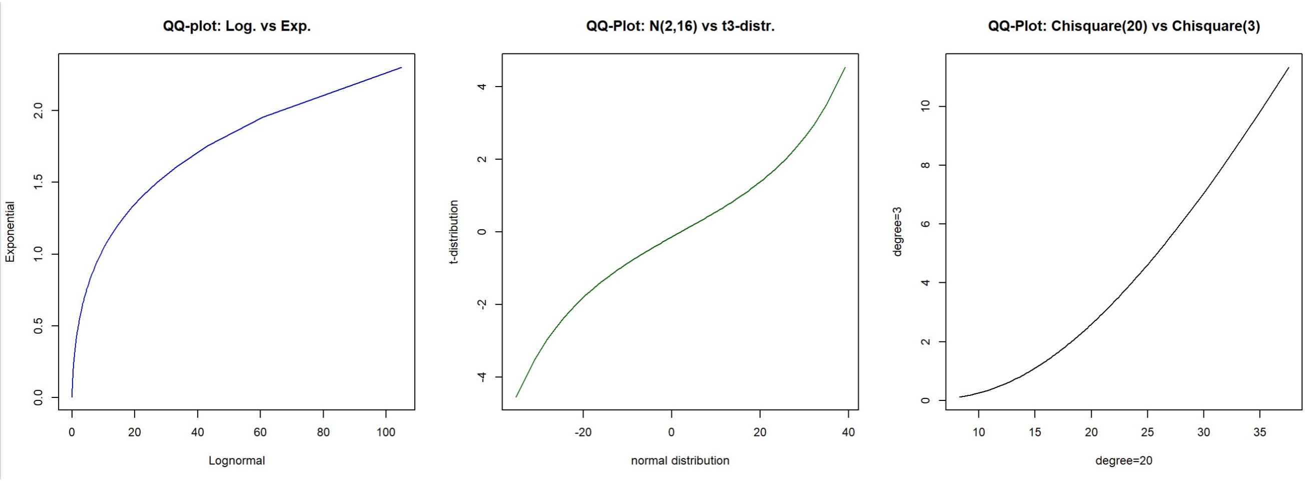

Figure 1: QQ-Plots of 3 different pairs of distributions

- **Lognormal with meanlog=0 and sdlog=2 vs. Exponential with rate 2**

  Comment: it is clear that there is no straight line in the QQ plot. Instead, we see a curve that starts bending downwards and towards the straight line. This is due to the heaviness of the lognormal distribution, which for both tails is heavier than on the exponential distribution. Especially, this is more evident in the left tail of the plot.

- **Normal with mean 2 and variance 16 vs t-distribution with 3 degrees of freedom**

  Comment: in this case, it is less obvious than before. However, we can also conclude that the normal distribution with a mean of 2 and standard deviation of 16 and the t distribution with 3 degrees of freedom does not belong to the same location-scale family. There are two tails in the plot. The right tail is a little heavier than the left tail. This is due to the fact that the t3 distribution has heavier tails (both the right and the left) than the normal distribution (especially when the degrees are low).

- **Chi-squared with 20 degrees of freedom vs Chi-squared with 3 degrees of freedom**

  Comment: Here we can see a small deviation from the straight line, especially on the lower left side of the plot. This is because there is a difference between the heaviness on the tails of both distributions. In this QQ plot, we see that line on the lower left side is above the QQ-line, which indicates that the chi-square distribution with degree=3 (on the vertical axis) has a lighter left tail than the other distribution. On the other hand, it is hard to tell if there is a heavier tail on the right side because there is an almost straight line.

By only analyzing these QQ plots, we can conclude that none of these pairs of distributions could be considered of the same location-scale family. This would only be the case of a straight line on the QQ plot.

b) Exploring Sample Data

In this part of the assignment, we are tasked in investigate the underlying distribution of a sample. In order, to successfully accomplish this task we used several graphical and numerical summaries, which helped us gain insights into the distribution.

First, we plotted the histogram of the sample to see how the data is distributed (we can also see a rough approximation of the mean, minimum and maximum). Next, we investigate the symmetry of the data by graphing the *symplot*. From this plot, we can conclude that the distribution of the sample is right-skewed because the area above the median is much larger than the area below it. Finally, we also plotted the empirical d=cumulative distribution function, so we have a complete understanding of the given data. A figure containing these three plots can be found in Appendix. The information gained from these plots helps us prioritize which distributions were going to be tested first and which ones to pay more attention to. Because of the skewness found in the histogram and the symmetry plot, we realized that the sample had to be exponential, log-normal, or chi-squared distributed. However, at this point, this was just a reference so we did not reject any other distribution yet.

After obtaining these insights, we proceeded to check if the sample distribution can be a location-scale family from a known distribution. Therefore, we plotted the QQ-plot of our sample distribution and the following distributions: normal, log-normal, uniform, exponential, Poisson, chi-squared (with different (varying) degrees of freedom), Laplace, logistic, t-distribution (also with different degrees of freedom), and Cauchy. From these QQ plots, we immediately discarded the following distribution: normal, Laplace, Cauchy, logistic, uniform, chi-square, and t-distribution. The main reason for discarding these distributions is that the qq-line is evidently not a straight line. A graph of these QQ plots will not be presented.

On the other hand, we have two possible candidates for the best distribution for this set: the exponential distribution and the log-normal distribution. As you can see from the QQ plots in the below-presented figure, both distributions are very strong candidates because there is an almost straight line in the plots. However, we conclude that the data set can be seen as a location-scale family of the log-normal distribution. The vector of the parameters (scale, location) of the log-normal distribution equals (mean(sample), sd(sample)), which in this case are 3.857 and 3.241, respectively. Note that the exponential distribution is a scale family (not a scale-location family). Regardless of the type of family of each distribution, we believe that both distributions represent the sample very well. However, we would consider the exponential distribution a better fit because it is capable of handling most outliers and because there is less deviation from the straight line than in the lognormal distribution.

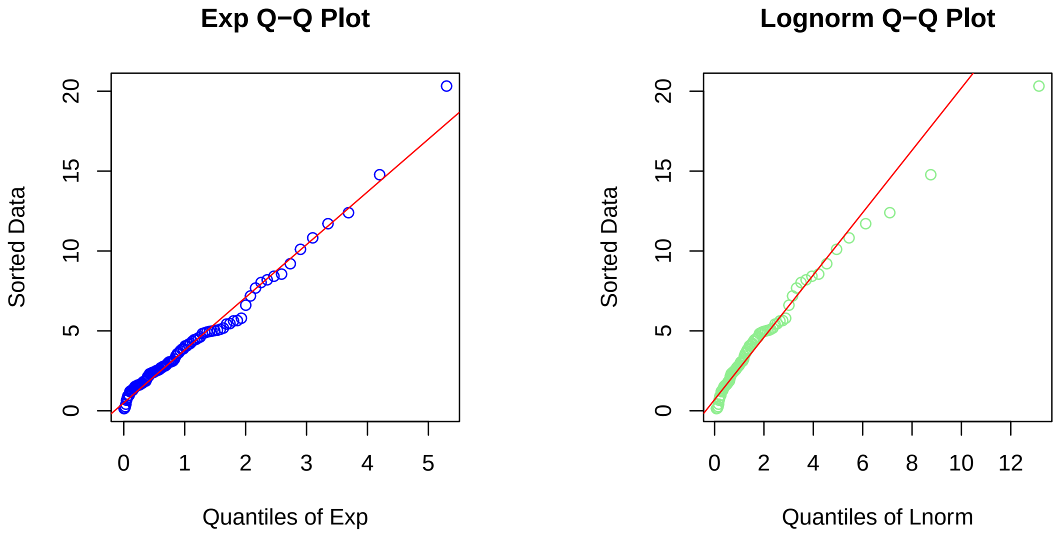

Figure 2: QQ-Plots for Exponential(2) and Lognormal Distributions(2,16)

2 Exercise 2.3

In this exercise, we obtained a data sample from a txt file named *sample2023b*. We have performed some analyses and tests to investigate the underlying distribution of the sample as well as other important features of the data.

a) Exploring the Sample Data

First, we have graphically explored the sample data. We have created a histogram, a boxplot, and an empirical cumulative distribution function. Some of these plots will be presented shortly. The histogram and the boxplot help us gain insights into how the sample data is distributed. From both plots, we see that the range, where most points are, is between 2.5 and 5, approximately. The mean is clear from the boxplot, approximately 3.5. Finally, from the symmetry plot, we can confirm that there is no perfect symmetry in the data but it is right skewed. This is because the area above the median is much higher than the area below it. We Just as in the previous exercise, these graphical summaries also help us focus on the exponential, chi-square, and lognormal distributions.

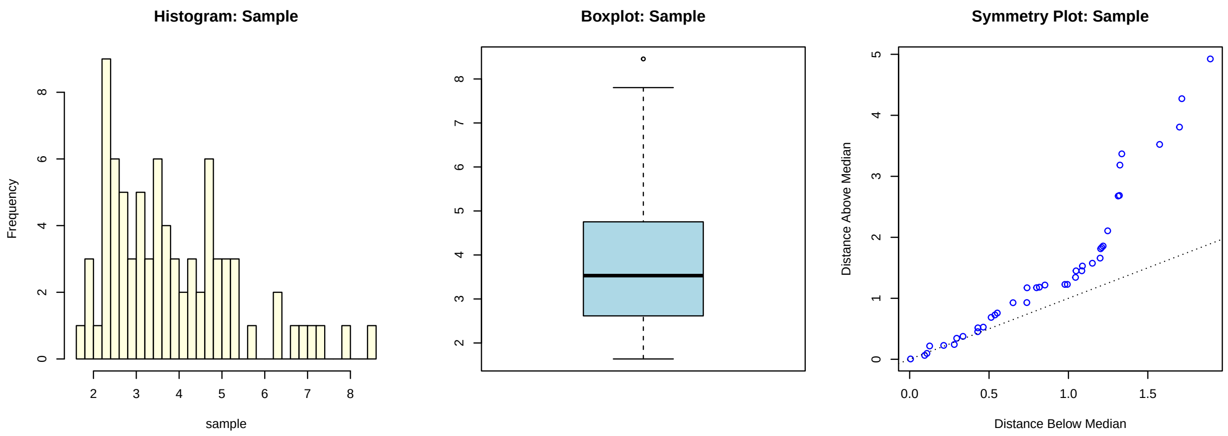

Figure 3: Graphical summaries of the sample

Additionally, we have created some QQ plots to verify the relationship between the sample and these other known distributions. From only the resulting QQ plots, the best candidates for the underlying distribution were the chi-square with degree 15, the normal distribution, and the exponential distribution, which are already sorted from the most probable distribution to the less probable one. We show the QQ plot of the best 2 candidates in the following figure.

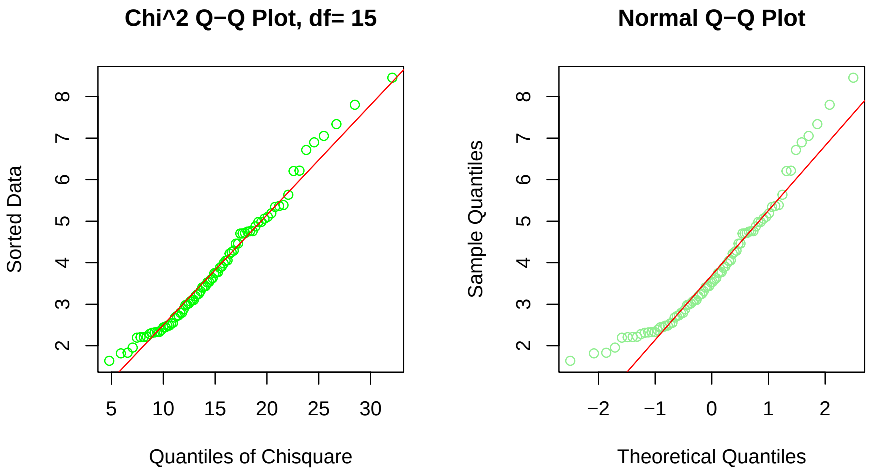

Figure 4: QQ-Plots for Chi-Square df=15 and for Normal Distribution

We performed a statistical test called Shapiro-Wilk Test (at level %5) to test for the normality of the set. However, the result of the test was to reject the hypothesis of normality for this data set. Therefore, we conclude from the graphical evidence gathered that the data set could have been drawn approximately from a chi-square distribution with a degree of freedom 15. Because the data can be modeled from this chi-square distribution, there are no scale and location parameters; the only parameter of the chi-square is the degree of freedom.

b) Kolmogorov-Smirnov Test for Gompertz distribuion

We have also performed the Kolmogorov-Smirnov test on the sample to check if its underlying distribution is the Gompertz distribution with parameters rate (b=0.7) and shape (n=0.07). The results of the Kolmogorov-Smirnov are the following: $p-value = 1.67 \times 10^{-12}$ and $score = 0,962$. We generally reject the hypothesis with a large value of the *score*, also denoted as Dn. However, because the p-value is 0; we can directly reject the hypothesis that the sample data come from the Gompertz distribution with shape (n=0.07) and rate (b=0.7).

c) Chi-Square Test for Gompertz distribuion

Just as in the previous exercise, we would like to test for the Gompertz distribution with the same parameters as before. However, now we would like to try it with the Chi-Square test. Following the rule of thumb, that there must be at least 5 observations per interval (also every np is above 5), we use the following vector to separate breaks of the test: c(1, 2.2, 2.6, 3, 3.3, 3.6, 4.3, 9). After performing the test, we conclude that we must reject the hypothesis again because we again obtained a p-value lower than 0.05 (in this case it was exactly 0), which is our alpha, and an extremely high chi-square score ($score = 1.1 \times 10^{15}$).

d) Are the results of b) and c) in agreement?

After performing the Kolmogorov-Smirnov and the Chi-square tests, we have concluded that the sample does not come from the Gompertz distribution with shape (n=0.07) and rate (b=0.7). In both cases, the p-value was lower than 0.05 and the scores were in the rejection interval. It could be that for other parameters, there could be a disagreement but this is not the case here. To confirm through another method we have plotted the QQ line of the sample vs the Gompertz (shape=0.07 and rate=1). We can confirm also by the QQ plot that the sample probably does not belong to the Gompertz family. Therefore, we validate the results of both tests and reject the hypothesis for the Gompertz distribution.

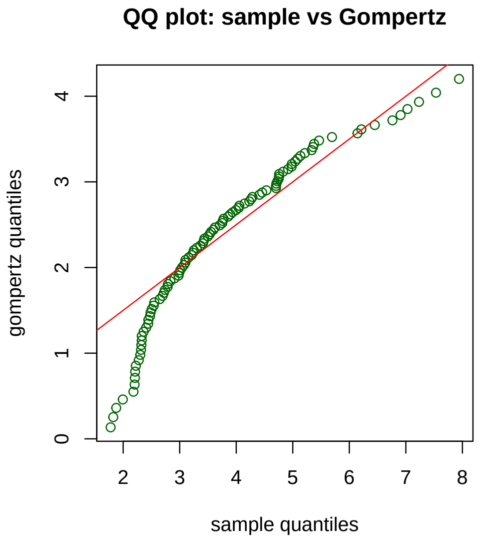

Figure 5: QQ-Plots for Gompertz

3 Exercise 2.4

In this exercise, we analyze the data set relating to several body measurements of female individuals. We have data points on 259 individuals. We are especially interested in the weight, height, and ankle circumference columns. A series of numerical and graphical analyses have been performed in order to understand relationships, trends, and more of the data set.

a) The shape of the Sample Distributions

First, we created a new column named BMI. This column consists of the body mass index calculated by dividing the weight by the squared height (in meters). After that, we compared the sample of ankle circumferences with the recently calculated column, the BMIs. The comparison starts with the creation of histograms and boxplots of both samples. The following figures will present these plots.

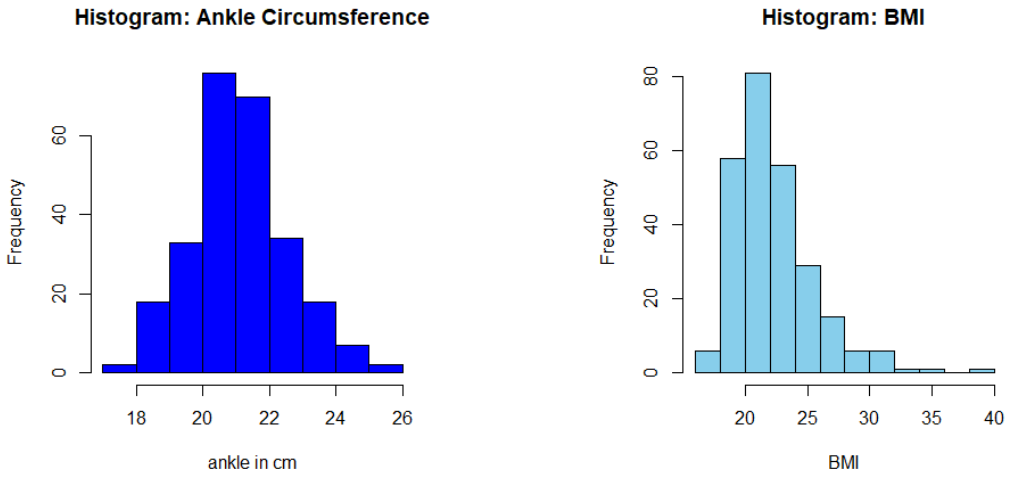

Figure 6: Comparison in the histograms of the ankle sample vs BMI

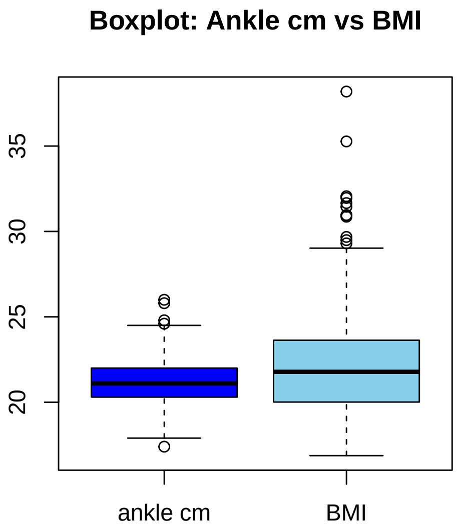

Figure 7: Comparison in the Boxplots of the ankle sample vs BMI

From these plots, we gained the following insights:

1. From the histograms, we can tell several things about the distribution. First, we can see that both distributions have a bell shape. This can indicate in some cases normality. However, it is clear that the BMI's histogram looks a little bit right-skewed.

2. Now, we compare the results of the boxplot. We can see that both samples have the same shape but in different proportions.

3. We can see from both sets of plots that the mean of both distributions is similar. However, we can easily distinguish a higher amount of outliers for the BMI sample.

Taking into consideration, that both histograms have a (similar) bell shape and that their corresponding boxplot has a similar shape, we can conclude that both data distributions have approximately the same shape.

b) QQ plot for the two samples: ankle circumferences vs. BMIs

With help with the function *qqplot*, we have plotted the QQ plot between the sample of the ankle circumferences and the BMIs. The plow will be presented below. Because most of all the data points are in a straight line, we can conclude that both distributions can belong to the same location-scale family.

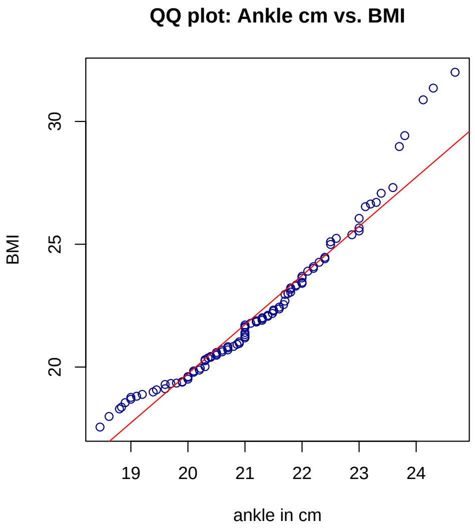

Figure 8: QQ plot: sample ankle circumferences vs. BMIs

c) Scale-Location Family (Normal/Chi-square and without statistical tests)

Without using any statistical tests, we can test the normality of each distribution by using the insights gains from the histograms and boxplots to plot its QQ plot against the theoretical values. We can repeat this process of creating a QQ plot for the Chi-square distribution. We had to test for different values for degrees of freedom. After trying different values, we concluded that 18 degrees of freedom were ideal for these two data sets. The QQ plots for both the normal and the chi-square distribution for both data sets are presented.

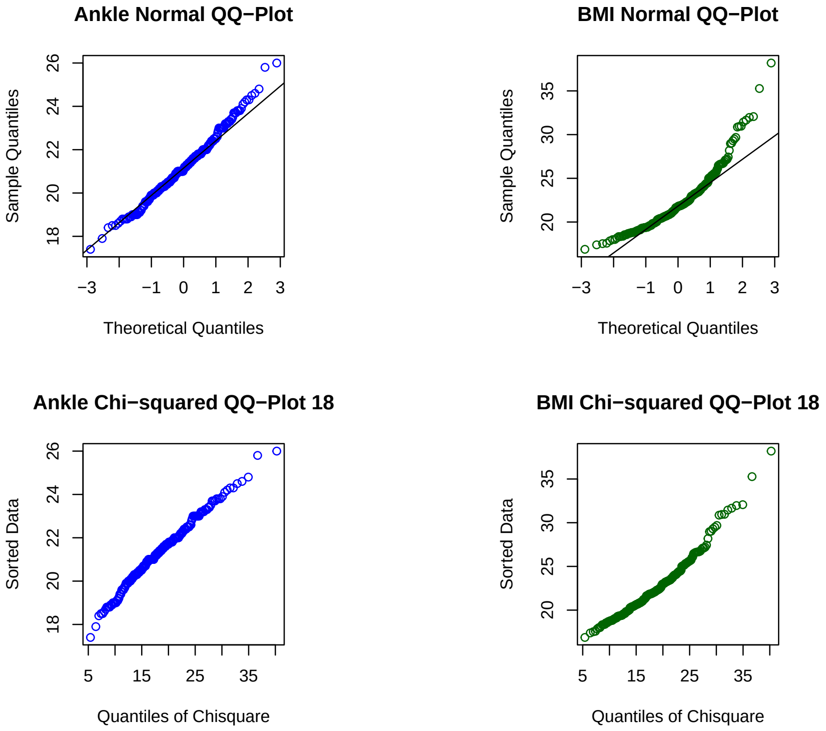

Figure 9: QQ Plots for the Normal and the Chi-Square Distributions

From these QQ plots, we can clearly see that the distribution for the ankle circumferences can be modeled as a location-scale family of both the normal and the chi-squared distribution with a degree of freedom 18. On the other hand, the sample for the BMI can only be modeled as a location-scale family of the chi-square with degree 18. This can be evaluated by the straightness of the line in the QQ plot.

d) Normality in new ratio measurement (without statistical tests)

For this part of the assignment, we created a new column named new ratio, which consisted of the value obtained by dividing the BMIs by the squared ankle circumferences. Now, we investigate whether this new dataset can be modeled as normally distributed or not.

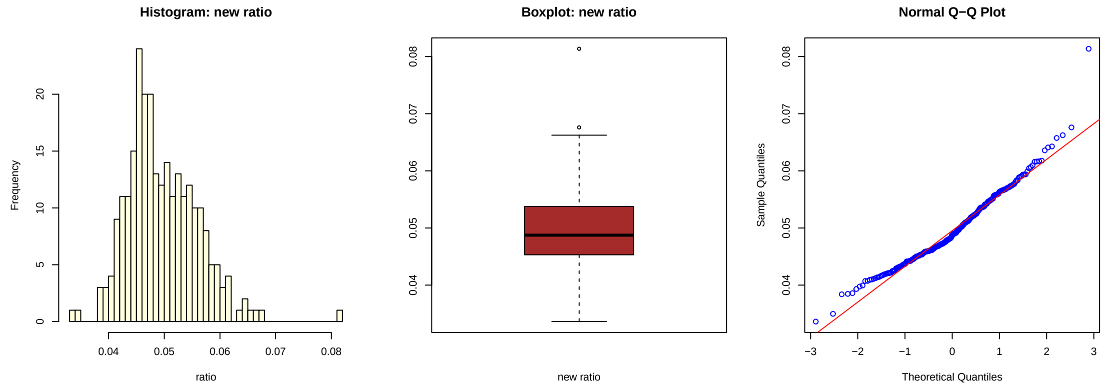

Figure 10: QQ Plots for the Normal and the Chi-Square Distributions

We have created a series of graphs to help us investigate the normality of the new ratio. The most relevant graphical summaries and their consequences are:

1. Histogram: has a bell shape, which is a good indication of normality.
2. Boxplot: has a very symmetric shape and the box is almost evenly separated. This is also a sign of normality.
3. QQ plot vs normal distribution: there is an almost straight line in this plot. This is a very strong indication of normality

Without having performed any statistical test, we can conclude by the arguments and graphs mentioned before that this new ratio can be modeled as a normal distribution.

e) Shapiro-Wilk Test at level 0.05

We complete the investigation of normality in the new ratio by applying the Shapiro-Wilk test at a significance level of 0.05. We applied the test with the function *shapiro.test*. The results of the test are the following: W= 0.966 and $p - value = 7.22 \times 10^{-6}$. Therefore, we reject the hypothesis that the sample can be modeled as a normal distribution, which is against our findings in the previous part.

The reason behind this disagreement is that the methods used before can only give us indications or hints of normality in the data. However, it cannot confirm that it indeed comes from a normally distributed set. On the other hand, the test does exactly that; it reassures that the sample distribution is not normally distributed.

f) Goodness-of-fit: full sample vs 50-pts sample

For the last part of the exercise, we have to compare the results of the goodness-of-fitness tests for normality and the histograms of the full sample and the first 50 data points of this sample. First, we have to create a new data frame with only the first 50 data points of the new ratio. After that, we plotted the histograms of both scenarios to see the differences.

We performed the Shapiro-Wilk test on the new data frame consisting of on;y the first 50 data points. The results of the test tell us that we cannot reject the hypothesis that the data come from the normal distribution, because the score is 0.98 and the p-value is 0.41. This is contrary to the results of this same test for the full sample, which rejects the hypothesis. These results are the same as in part e) of this exercise. If we see the below histograms, we can see that the 50-data points histogram has a more symmetric bell shape than the full sample; this supports the results for the normality of the short sample.

There are several possible explanations for this discrepancy. First and probably the best explanation, the 50-point sample does not represent the full sample. This means that the full sample and the subset have different distributions. Second and still applicable to this situation, the Shapiro-Wilk test can be sensitive to the sample size (even though it is considered to be one of the most robust tests). So the difference in sample size could be affecting the results. However, this second explanation is less likely due to the robustness of the test.

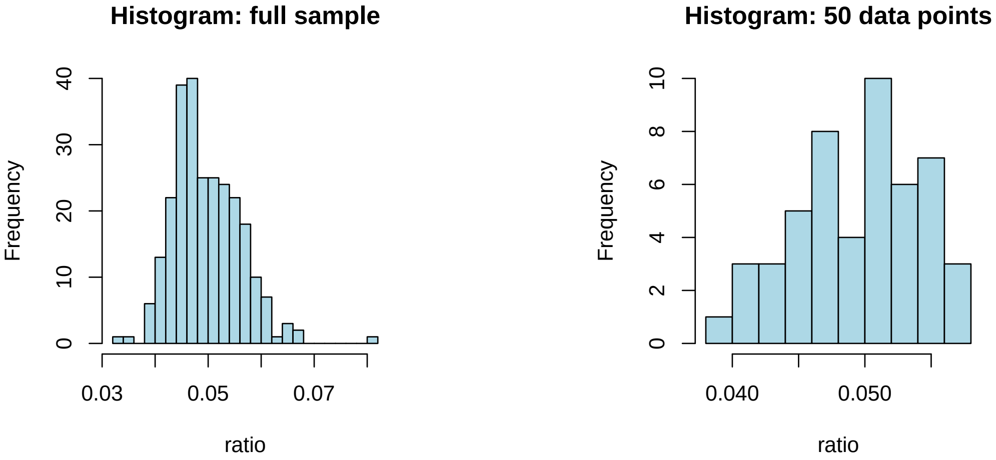

Figure 11: QQ Plots for the Normal and the Chi-Square Distributions

4 Appendix

The Appendix mostly consists of the code used to analyze each exercise and some extra visualization to gain a better understanding of each problem.

Code exercise 2.2 a

```
par(mfrow=c(1,3), pty='s')
x = seq(0.01, 0.99, 0.01)

lognorm_q = qlnorm(x, meanlog = 0, sdlog = 2)
exp_q = qexp(x, 2)
plot(lognorm_q, exp_q, pch = 20, main="QQ-plot: Log. vs Exp.",
     xlab = "Lognormal", ylab = "Exponential", type='l', col= 'blue')

norm_q = qnorm(x, mean=2, sd=16)
t3_q = qt(x, df=3)

plot(norm_q, t3_q, type='l', main='QQ-Plot: N(2,16) vs t3-distr.',
     xlab='normal distribution', ylab='t-distribution', col= 'dark green')

chis_q = qchisq(x, df=20)
chis_q2 = qchisq(x, df=3)
plot(chis_q, chis_q2, type='l', main = 'QQ-Plot: Chisquare(20) vs Chisquare(3)',
     xlab='degree=20', ylab='degree=3')

par(mfrow=c(1,3), pty='s') #for plotting the 3 graphs in one figure
```

Plots 2.2 b

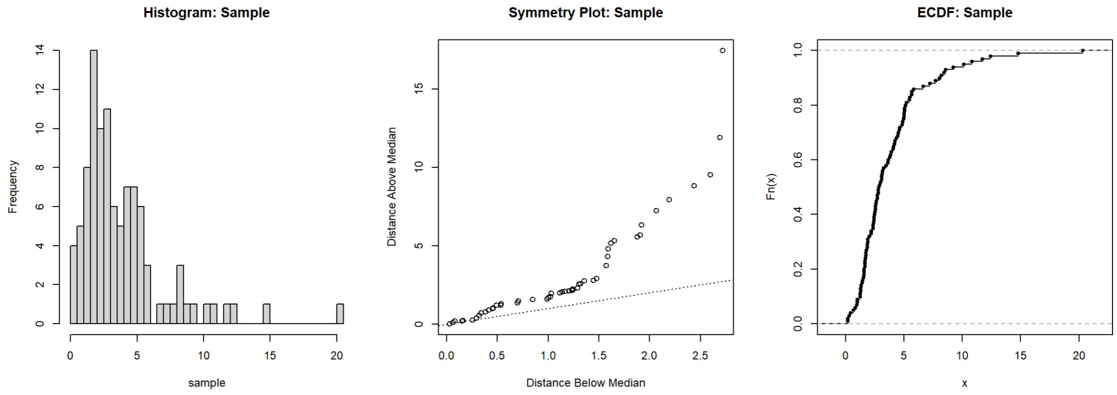

Figure 12: Supporting plots: histogram, symmetry plot, and ECDF of sample

Code 2.2 b

```
sample = scan("sample2023a.txt")
sample_quantiles = quantile(sample, probs = seq(0.01,0.99,0.01))

hist(sample, breaks=30, main='Histogram: Sample')
symplot(sample, main='Symmetry Plot: Sample')
plot(ecdf(sample), main='ECDF: Sample')

par(mfrow=c(1,2), pty='s')
qqexp(sample)   #testing for exponential
abline(a=0.5, b=3.3, col='red')
qqlnorm(sample)
```

```
abline(a=0.7, b=1.95, col='red')

qqnorm(sample) #testing for normal distribution
qqline(sample)
shapiro.test(sample)
qqlaplace(sample)
qqchisq(sample, df=6)
qqcauchy(sample)
qqlogis(sample)
qqunif(sample)
qqt(sample, df=2)
```

Code 2.3 a

```
sample = scan("sample2023b.txt")
summary(sample)

par(mfrow=c(1,3))
hist(sample, breaks= 25, main="Histogram: Sample")
boxplot(sample, main="Boxplot: Sample")
symplot(sample, main="Symmetry Plot: Sample")
plot(ecdf(sample))

par(mfrow=c(1,2), pty='s')
qqchisq(sample, df=15) #this is possible (best candidate)
#abline(0.15,0.25, col='red')
abline(-0.15,0.265, col='red')
qqnorm(sample) #possibly this one
qqline(sample, col='red')
shapiro.test(sample) #normality discarded with the Shapiro-Wilk test
qqexp(sample) #less possible
```

Code 2.3 b

```
rate=0.7
tscale = 1/rate

test1 = ks.test(sample, pgompertz, scale=0.07, shape=tscale)
ks_pvalue = test1$p.value
ks_score = test1$statistic
ks_reject = test1$exact
```

Code 2.3 c

```
br = c(1, 2.2, 2.6, 3, 3.3, 3.6, 4.3, 9)
test2 = chisquare(sample, pgompertz, shape=0.07, scale=tscale, breaks=br)
chisq_breaks = 7
chisq_score = test2$chisquare
chisq_pvalue= test2$pr
chisq_reject = 'true'
```

Code 2.2 d

```
#qqplot for the Gompertz Distribution
sample_quantiles = quantile(sample, probs=seq(0.01,0.99,0.01))
gompertz_quantiles = qgompertz(seq(0.01,0.99,0.01),shape=0.07, scale=1)
```

```
plot(sample_quantiles, gompertz_quantiles, main='QQ plot: sample vs Gompertz')
abline(0.5,0.5, col='red')
```

Code 2.4 a

```
table = read.table("body.dat.txt", header=FALSE)
df = table[248:507, c(20, 23, 24)]
names(df)[c(1,2,3)] = c("a_cm","weight_kg","height_cm")
df$BMI = df$weight_kg/((df$height_cm/100)^2)

par(mfrow=c(1,2))
hist(df$a_cm, xlab='BMI', main="Histogram: Ankle Circumsference")
hist(df$BMI, xlab= 'ankle in cm', main="Histogram: BMI")

par(mfrow=c(1,1))
boxplot(df$a_cm, df$BMI,main='boxplot: ankle cm vs BMI', names= c("ankle cm","BMI"))
```

Code 2.4 b

```
par(mfrow=c(1,1),pty='s')
qqplot(df$a_cm, df$BMI, xlab='ankle in cm', ylab='BMI', main="QQ plot: Ankle cm vs. BMI")
abline(-20.,2, col='red')
```

Code 2.4 c

```
par(mfrow=c(1,4), pty='s')

qqnorm(df$a_cm, main="Ankle Normal QQ-Plot", col='blue')
qqline(df$a_cm)
qqnorm(df$BMI,main="BMI Normal QQ-Plot", col='dark green')
qqline(df$BMI)
shapiro.test(df$a_cm)
shapiro.test(df$BMI)

degree=18
par(mfrow=c(1,2))
qqchisq(df$a_cm,df=degree, col='blue', main="Ankle Chi-squared QQ-Plot")
qqchisq(df$BMI,df=degree, col='dark green', main="BMI Chi-squared QQ-Plot")
```

Code 2.4 d

```
df$ratio = df$BMI/(df$a_cm^2)

par(mfrow=c(1,3))
hist(df$ratio, breaks=40,xlab='ratio', main='Histogram: new ratio', col='lightyellow')
boxplot(df$ratio, main='Boxplot: new ratio', xlab='new ratio', col='brown')
qqnorm(df$ratio, col='blue')
qqline(df$ratio, col='red')
```

Code 2.4 e

```
shapiro.test(df$ratio)
```

Code 2.4 f

```
df50 = df[1:50,]

shapiro.test(df$ratio)
shapiro.test(df50$ratio)

par(mfrow=c(1,2))
hist(df$ratio,breaks=20,xlab='ratio', main='Histogram: full sample', col='lightblue')
hist(df50$ratio,breaks=7,xlab='ratio', main='Histogram: 50 data points', col='lightblue')

par(mfrow=c(1,1))
boxplot(df$ratio, df50$ratio)
```