Applied Statistics - Assignment 2

Thom Hurks *0828691*

June 27, 2016

Contents

1 Introduction

This is the report for assignment 2 for the course Applied Statistics. The results in this report regard set 3 and as such contain Exercise 3 about outlier tests, Exercise 4 about normality tests and Exercise 9 regarding the McNemar and Agreement tests.

1.1 Significance level

In our tests we will use a significance level of $\alpha = 0.05$. The choice of significance level is always somewhat arbitrary, but we pick 0.05 because it is a value that is widely used and accepted amongst statisticians and because our client recommends this.

1.2 Exploratory data analysis

Before explaining the methods that we use for the outlier and normality tests and analysing the results, we will explore the data for exercises 3 and 4 to get an idea of what we are dealing with. In table 1 we show some descriptive statistics.

n	mean	median	min	max	variance	st.dev
20	107.75	107.35	97.00	132.40	77.26	8.79

Table 1: Descriptive Statistics

In figure 1 we also provide a histogram with overlaid density plot in order to get a visual representation of the dataset.

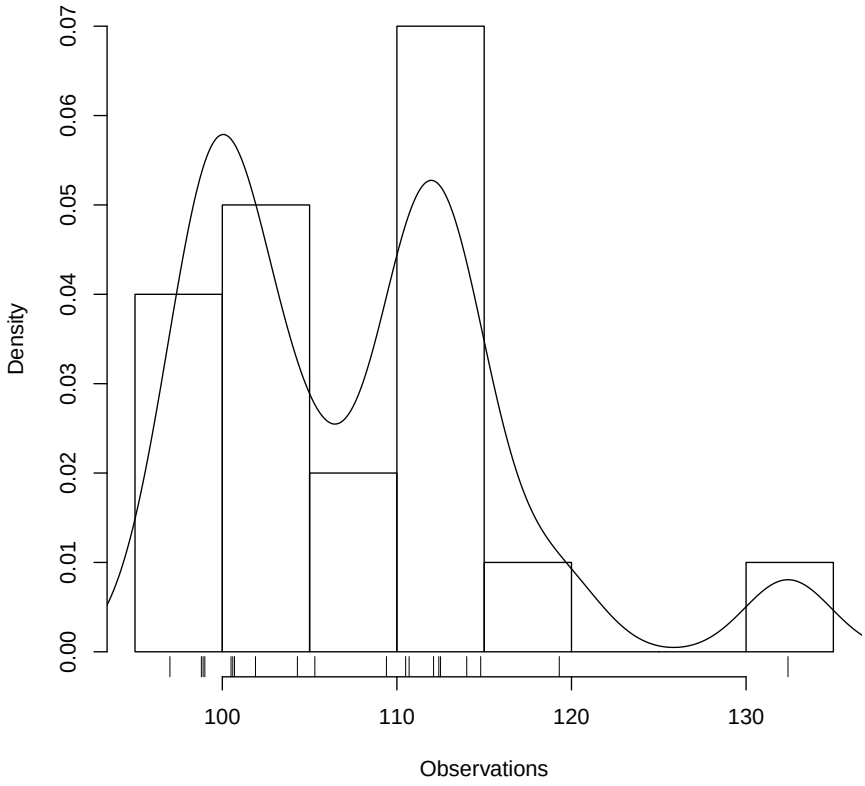

Figure 1: Histogram and density plot of the observations

We also explore the dataset for exercise 9, which are binary paired blood samples for measurements denoted C and K for a group of patients. Descriptive statistics are shown in tables 2 and 3 for the measurements C and K respectively. A joint scatterplot for C and K is shown in figure 2

n	mean	median	min	max	variance	st.dev
21	128.05	128	105	155	160.55	12.67

Table 2: Descriptive Statistics for measurement C

n	mean	median	min	max	variance	st.dev
21	121.00	123	56	140	311.60	17.65

Table 3: Descriptive Statistics for measurement K

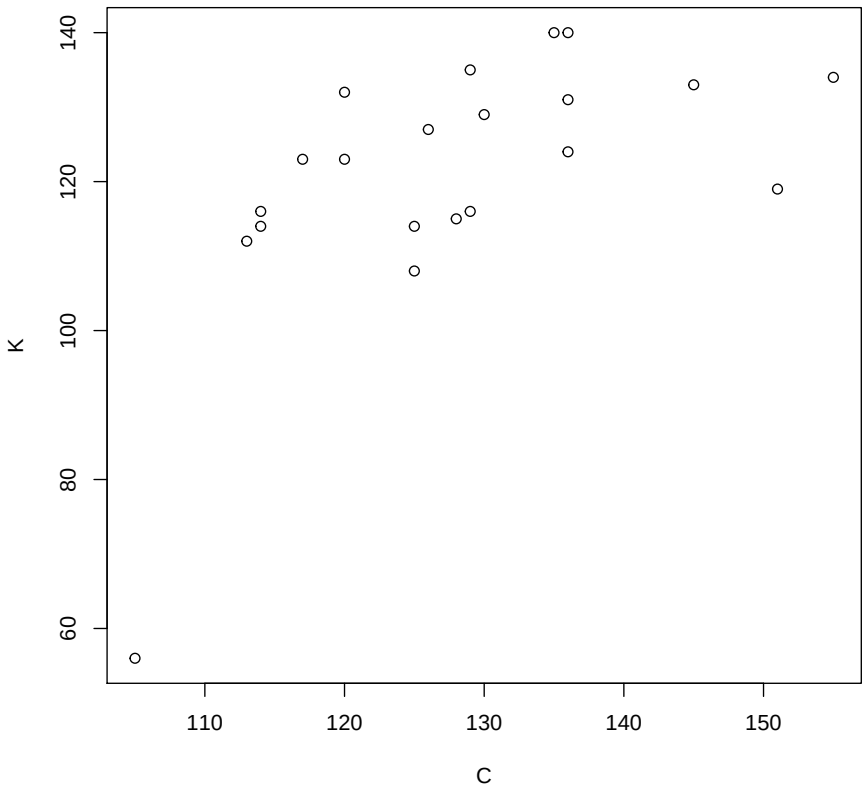

Figure 2: Scatterplot of measurements C and K

1.3 Reproducible Research

We value the idea of reproducible research, which is publishing data analyses together with the software code, so others can reproduce and verify our findings and possibly extend our research. For this reason this report has been created using Knitr, which is a dynamic report generation tool for the statistical programming language R. The Knitr sourcecode for this report is available from the authors if others are interested in verifying our results.

2 Exercise 3: Outlier Tests

2.1 Methods

We will perform various outlier tests on the data, which we will first describe.

2.1.1 Grubbs test

The grubbs test is also known as the extreme studentized deviate test or the maximum normed residual test. The test assumes the data comes from a normal distribution. The Grubbs test has three variants: we can test for a single outlier, for two outliers on opposite tails or for two outliers in one tail. We can also treat the test as a one-sided or two-sided test. We will perform all variants, which means we will perform six tests.

2.1.2 Dixon test

We also perform the Dixon test, which can find up to two outliers in the data, one in the lower tail and one in the upper tail. Dixon's test also assumes that the data comes from a normal distribution. The Dixon test has various variants based on the sample size, and we will use the one appropriate for a sample size of $n = 20$. We will also perform it as a one-sided and two-sided test.

2.1.3 Hampel's rule

Hampel's rule is considered a non-parametric outlier test.

2.1.4 Tukey's method

Tukey's method is a non-parametric outlier test that is often implemented in the box plot. As such, we will give a box plot based on work by Tukey and accordingly present the outliers.

2.1.5 Doornbos test

The Doornbos test uses the externally studentized values to investigate the existence of a single outlier. This test is similar to outlier detection in linear models.

2.2 Results

Some of the described outlier tests assume normality. We will see in the section for Exercise 4 about normality tests that this is a valid assumption for this data set, especially for the log transformed data.

2.2.1 Grubbs test

The Grubbs test has as a null hypothesis that the dataset does not have outliers. For each variant of the test we will note if that null hypothesis can be rejected at a significance level of $\alpha = 0.05$ and what the hypothesized outlier(s) of the alternative hypothesis are.
One-sided tests:
Test for one outlier: p-value is $0.0154391 < 0.05$, alternative hypothesis is that the highest value 132.4 is an outlier.
Test for two outliers on opposide tails: p-value is $0.3394144 > 0.05$, alternative hypothesis is that 97 and 132.4 are outliers.
Test for two outliers on the same tail: p-value is $0.0281379 < 0.05$, alternative hypothesis is that the highest values 119.3 , 132.4 are outliers.
Two-sided tests:
Test for one outlier: p-value is $0.0308782 < 0.05$, alternative hypothesis is that the highest value 132.4 is an outlier.
Test for two outliers on opposide tails: p-value is $0.6788288 > 0.05$, alternative hypothesis is that 97 and 132.4 are outliers.
Test for two outliers on the same tail: p-value is $0.0562757 > 0.05$, alternative hypothesis is that the highest values 119.3 , 132.4 are outliers.

The result of the Grubbs test seems to be that 132.4 is an outlier and perhaps 119.3 is an outlier too but that is less certain. The data does not seem to have outliers on opposite tails.

2.2.2 Dixon test

One-sided Dixon test: p-value is $0.0123036 < 0.05$, the alternative hypothesis is that the highest value 132.4 is an outlier.
Two-sided Dixon test: p-value is $0.0246071 < 0.05$, the alternative hypothesis is that the highest value 132.4 is an outlier.

The result of the Dixon test is that 132.4 may be an outlier.

2.2.3 Hampel's rule

If we compute the absolute normalized value according to Hampel's rule for each observation, then we see that none of the absolute normalized values are larger than 3.5 and as such none are considered outliers according to Hampel's rule. The largest computed absolute normalized value is $2.5407509 < 3.5$ for observation 132.4.

2.2.4 Tukey's method

In figure 3 you can see the box-and-whisker plot for the dataset. This boxplot is based on the work of Tukey. The outlier according to Tukey's method is as such 132.4

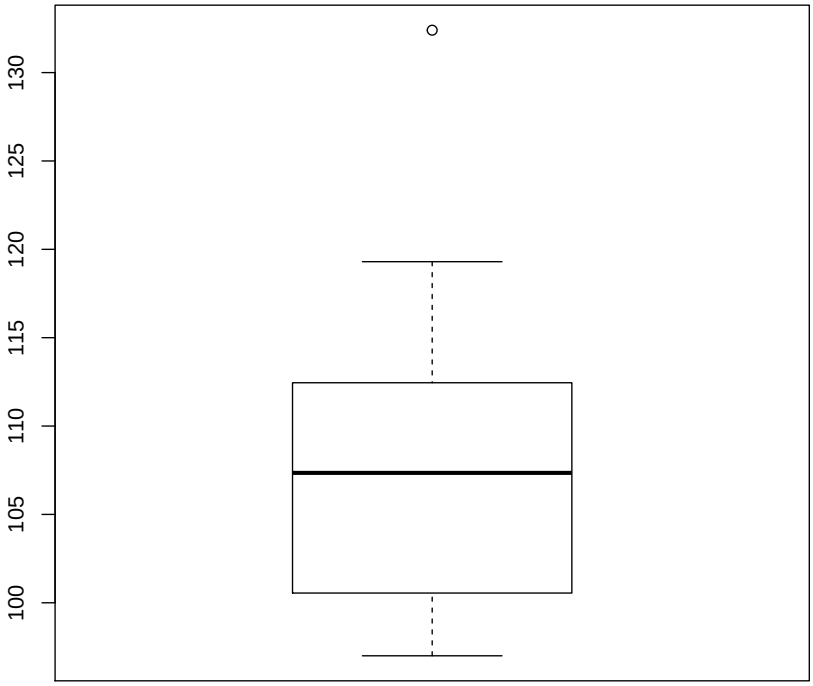

Figure 3: Box-and-whisker plot of the observations

2.2.5 Doornbos test

At a critical level of $\alpha = 0.05$ we compute the Doornbos test criterion to be 3.5101041. The result of the Doornbos test is that the hypothesis that the data has no outliers can be rejected, since the Doornbos test statistic for the value 132.4 is $3.7267405 > 3.5101041$. None of the other values in the dataset produce a Doornbos test statistic that is larger than the computed Doornbos criterion.

2.3 Results for the log-transformed data

We will now present the results again, but for the log-transformed data.

2.3.1 Grubbs test

The Grubbs test has as a null hypothesis that the dataset does not have outliers. For each variant of the test we will note if that null hypothesis can be rejected at a significance level of $\alpha = 0.05$ and what the hypothesized outlier(s) of the alternative hypothesis are.
One-sided tests:
Test for one outlier: p-value is $0.0336257 < 0.05$, alternative hypothesis is that the highest value 4.88582764350291 is an outlier.
Test for two outliers on opposide tails: p-value is $0.4612322 > 0.05$, alternative hypothesis is that 4.57471097850338 and 4.88582764350291 are outliers.
Test for two outliers on the same tail: p-value is $0.0609697 > 0.05$, alternative hypothesis is that the highest values 4.78164132910387 , 4.88582764350291 are outliers.
Two-sided tests:
Test for one outlier: p-value is $0.0672514 > 0.05$, alternative hypothesis is that the highest value 4.88582764350291 is an outlier.
Test for two outliers on opposide tails: p-value is $0.9224644 > 0.05$, alternative hypothesis is that 4.57471097850338 and 4.88582764350291 are outliers.
Test for two outliers on the same tail: p-value is $0.1219395 > 0.05$, alternative hypothesis is that the highest values 4.78164132910387 , 4.88582764350291 are outliers.

The result of the Grubbs test seems to be that 4.88582764350291 is an outlier and that there are not two outliers. This result only holds for the one-sided test, for the two-sided tests the results are non-significant although only slightly. The data does not seem to have outliers on opposite tails. Note that the exponent of this outlier is 132.4 from the original data.

2.3.2 Dixon test

One-sided Dixon test: p-value is $0.0260165 < 0.05$, the alternative hypothesis is that the highest value 4.88582764350291 is an outlier.
Two-sided Dixon test: p-value is $0.0520331 > 0.05$, the alternative hypothesis is that the highest value 4.88582764350291 is an outlier.

The result of the Dixon test is that 4.88582764350291 may be an outlier, but only the one-sided test is significant, the two-sided test is slightly beyond significance.

2.3.3 Hampel's rule

If we compute the absolute normalized value according to Hampel's rule for each observation, then we see that none of the absolute normalized values are larger than 3.5 and as such none are considered outliers according to Hampel's rule. The largest computed absolute normalized value is $2.2826769 < 3.5$ for observation 4.8858276.

2.3.4 Tukey's method

In figure 4 you can see the box-and-whisker plot for the dataset. This boxplot is based on the work of Tukey. The log-transformed dataset has no outliers according to Tukey's method.

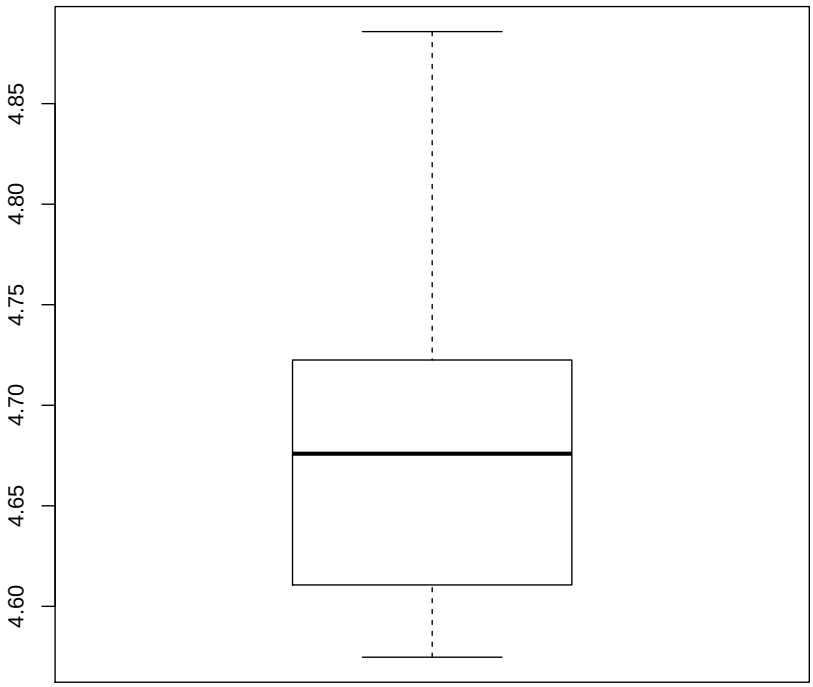

Figure 4: Box-and-whisker plot of the log-transformed observations

2.3.5 Doornbos Test

At a critical level of $\alpha = 0.05$ we compute the Doornbos test criterion to be 3.5101041. The result of the Doornbos test is that the hypothesis that the data has no outliers can not be rejected, since the Doornbos test statistic is < 3.5101041 for all values. The only value that comes close is 4.8858276 with a test statistic of 3.3764167.

2.4 Analysis

First, let us note that the Grubbs and Dixon's test assume normality of the data, which we will see in Exercise 4 is a valid assumption, especially for the log-transformed data. Based on our results we can be fairly confident that 132.4 is an outlier, since this is reported by the Grubbs test, the Dixon test, Tukey's method and the Doornbos test and both for the one-sided and two-sided tests. The grubbs test seems to suggest that 119.3 can possibly also be an outlier, but this is less certain since the p-value for that hypothesis is greater than $\alpha = 0.05$ for the two-sided test, but only slightly, and this possible outlier is not reported by the other tests.

The analysis on the log-transformed data yields the same outlier, the logarithm of 132.4, but only for the one-sided Grubbs test and the one-sided Dixon test and not according to Tukey's method. It must be noted that for that single outlier the two-sided tests are near significance.

Hampel's rule does not report outliers for both the original data and the log-transformed data.

3 Exercise 4: Normality Tests

We are given a dataset of which we are to determine if the data has a normal distribution. We will also determine the normality of the logarithmically transformed data.
In this section we will describe the various approaches at testing normality that we use. After presenting the approaches, we will give the results of the tests and provide an analysis of the results.

3.1 Methods

3.1.1 Graphical method

The first method that we will use to determine the normality of the data is a graphical method: the QQ-plot, a plot where a graphical representation of the hypothesized, in our case the normal, distribution is plotted against the given dataset. It must be noted that the interpretation of a QQ-plot can be highly subjective, so more formal tests must be performed.

3.1.2 Empirical distribution tests

We will apply three empirical distribution tests: the Kolmogorov-Smirnov test, the Anderson-Darling test and the Cramer-Von Mises test. These tests compare the empirical distribution function of the data with the hypothesized distribution. Our null hypothesis is that the data comes from a normal distribution, but we do not specify the mean and variance of the hypothesized distribution,

so in the case of the Kolmogorov-Smirnov test we actually use the Lilliefors variant of the test which does not make these assumptions. The Lilliefors test that we use computes the p-value using methods by Dallal-Wilkinson (1986) and Stephens (1974) in order to correct for the composite null hypothesis. The Anderson-Darling test and the Cramer-Von Mises tests are more appropriate when parameters of the hypothesized distribution need to be estimated, as in our case.

3.1.3 Regression techniques

As a regression technique we will use the well-known and powerful Shapiro-Wilk normality test. This test is probably the most powerful test that we perform. The Shapiro-Wilk test is sensitive to ties in the data, but our dataset does not contain ties.

3.1.4 Moment Tests

Lastly we will test the data's skewness and kurtosis. There are various "rules of thumb" that statisticians use as to what skewness and kurtosis values are acceptable to indicate that a distribution may be near normality, but this test has also been formalized in the D'Agostino test and the Jarque-Bera test, which we will both apply. These tests are not very strong, since they only test if the distribution of the data has similar skewness and kurtosis to the normal distribution, whereas the previous described tests are omnibus tests. The advantage of these two tests, though, are that they can show you where the departure from normality happens.

3.1.5 Chi-square tests

Chi-square tests for normality are not recommended when distributional parameters need to be estimated and when the data is numerical. Since this is the case for us, we choose not to perform chi-square tests.

3.2 Results

We now perform the described tests on the dataset and present the results. The tests were executed using the R statistical computing environment and we also used the following R packages: nortest (for the Anderson-Darling and Lilliefors tests), e1071 (for computing the skewness and kurtosis in the same way as SAS does), moments (for the Jarque-Bera test) and fBasics (for the D'Agostino test).

3.2.1 Graphical method

We can observe from the QQ-plot in figure 5 that the data points follow a rough "S" shape, which may indicate that the skewness or kurtosis of the data differs from that of the normal distribution. However, aside from one outlier in the

top-right of the plot there do not seem to be very major deviations from the normal distribution. It is clear that further testing is required.

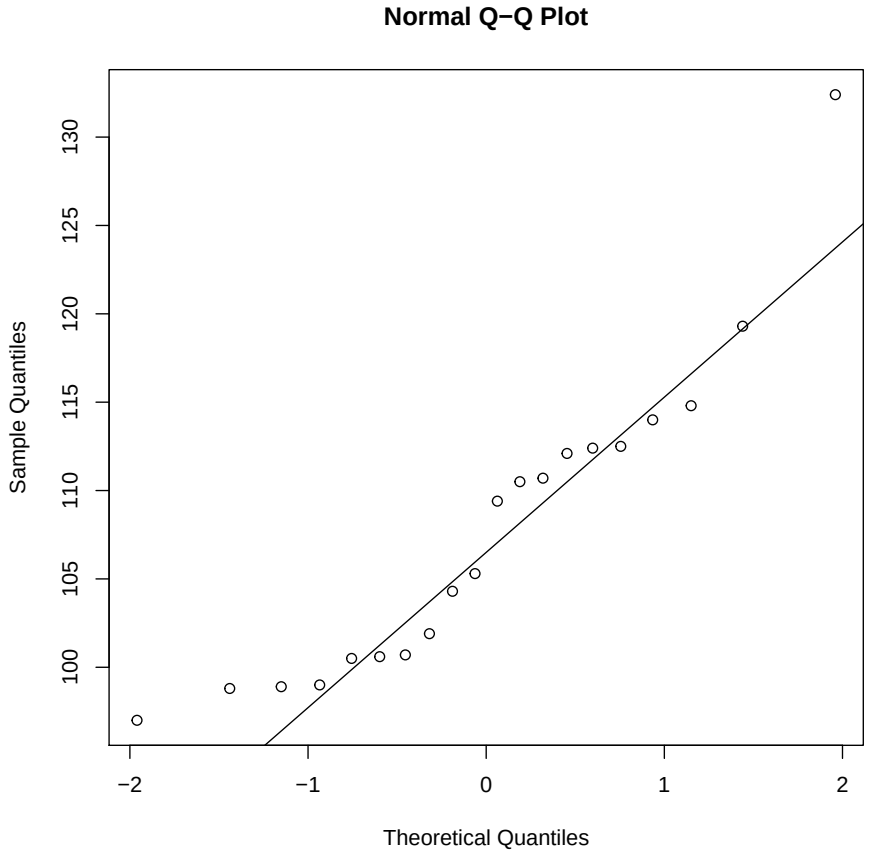

Figure 5: QQ Plot of the observations

3.2.2 Empirical distribution tests

The result of the empirical distribution tests is that the p-value of the Anderson-Darling test is $0.0814107 > 0.05$ and that the p-value of the Cramer-Von Mises test is $0.1533603 > 0.05$. As explained, we use the Lilliefors test with the corrected p-value for the composite hypothesis of normality as the variant of the Kolmogorov-Smirnov test. For this Kolmogorov-Smirnov test the p-value is $0.3062044 > 0.05$. This means none of the empirical distribution tests can reject the null hypothesis that the sample comes from a normal distribution at a significance level of $\alpha = 0.05$.

3.2.3 Regression techniques

The p-value of the Shapiro-Wilk test, likely the strongest test that we perform, is $0.0332668 < 0.05$. This means the Shapiro-Wilk test can reject the null hypothesis that the data comes from a normal distribution at a significance level of $\alpha = 0.05$, but it must be noted that the p-value is not very small and is still somewhat near our value of α.

3.2.4 Moment tests

We first examine the skewness and excess kurtosis of the sample dataset and compute that they are 1.0985356 and 1.7031546 respectively. The skewness and excess kurtosis are computed using the convention of SAS. An often used rule-of-thumb for the skewness and excess kurtosis is that values between -2 and 2 are acceptable for an indication that the samples may resemble a normal distribution. The computed values are within this range, but we can test this in a more formal way and use the Jarque-Bera test. The Jarque-Bera test tests the null hypothesis that the skewness is close to zero and the kurtosis close to three. The computed p-value of the Jarque-Bera test is $0.1166248 > 0.05$. Then we compute the p-value of the D'Agostino test, which is also based on testing the skewness and kurtosis. The p-value of this test on our dataset is $0.0330004 < 0.05$. This means that the Jarque-Bera test cannot reject the hypothesis that the skewness and kurtosis of the data is similar to that of the normal distribution, but the D'Agostino test can. Again, it must be noted that the p-value for the D'Agostino test is not very small and still somewhat near to our significance level of $\alpha = 0.05$, just as with the Shapiro-Wilk test.

3.3 Results for the log-transformed data

We now perform the described tests on the log-transformed dataset and present the results again.

3.3.1 Graphical method

We can see that the QQ-plot in figure 6 is almost exactly the same as the QQ-plot for the non-transformed data in figure 5, the datapoints seem only slightly shifted. Hence the same observation applies. Note that the log transformation is often used for skewed datasets, so a QQ-plot that hardly changes may indicate that our dataset is simply not that skewed. Note that the computed skewness for the non-transformed dataset is 1.0985356.

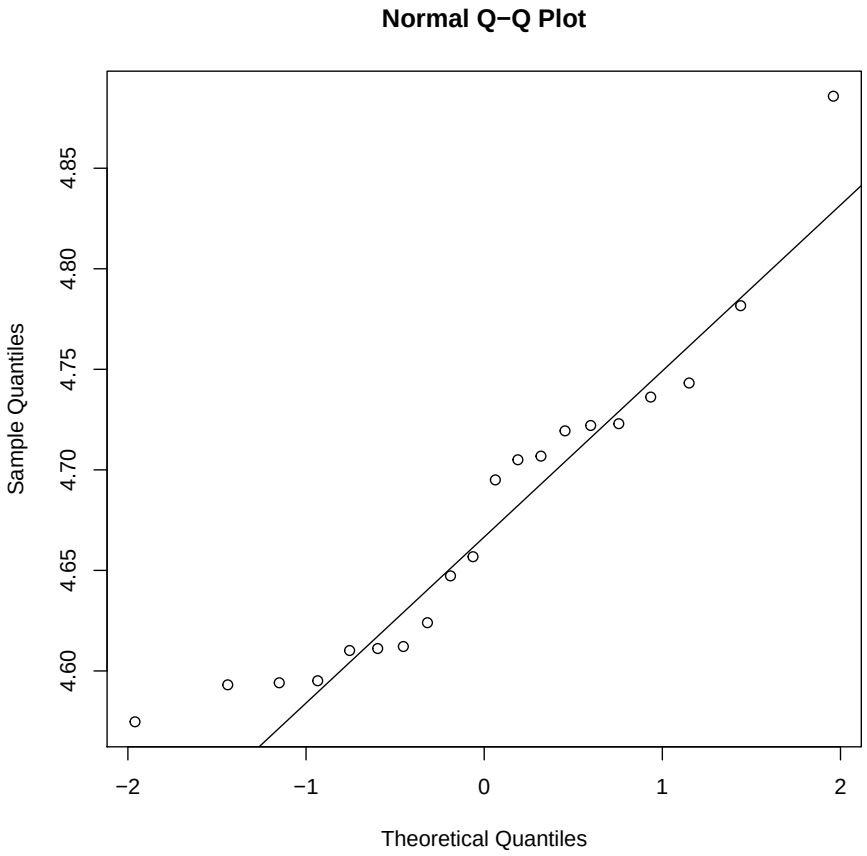

Figure 6: QQ Plot of the log-transformed observations

3.3.2 Empirical distribution tests

The result of the empirical distribution tests is that the p-value of the Anderson-Darling test is $0.1177284 > 0.05$ and that the p-value of the Cramer-Von Mises test is $0.1768264 > 0.05$. As explained, we use the Lilliefors test with the corrected p-value for the composite hypothesis of normality as the variant of the Kolmogorov-Smirnov test. For this Kolmogorov-Smirnov test the p-value is $0.2974229 > 0.05$. This means none of the empirical distribution tests can reject the null hypothesis that the log-transformed sample comes from a normal distribution at a significance level of $\alpha = 0.05$.

3.3.3 Regression techniques

The p-value of the Shapiro-Wilk test, likely the strongest test that we perform, is $0.0730852 > 0.05$. This means the Shapiro-Wilk test cannot reject the null hypothesis that the data comes from a normal distribution at a significance level of $\alpha = 0.05$. As with the non-transformed dataset, the p-value is close to our significance level of $\alpha = 0.05$, but now it is *greater* than α.

3.3.4 Moment tests

We first examine the skewness and excess kurtosis of the sample dataset and compute that they are 0.8669222 and 0.8898526 respectively. The skewness and excess kurtosis are computed using the convention of SAS. We can immediately observe that these values are smaller than the original skewness and kurtosis values of 1.0985356 and 1.7031546, which means that the log-transformation lessened the skewness and kurtosis of the data. This will have an effect on the outcome of the Jarque-Bera and D'Agostino tests, since they test the skewness and kurtosis. The computed p-value of the Jarque-Bera test is $0.3218793 > 0.05$. The p-value of the D'Agostino test on our log-transformed dataset is $0.1379792 > 0.05$. This means that both the Jarque-Bera test and the D'Agostino test cannot reject the hypothesis that the skewness and kurtosis of the log-transformed data is similar to that of a normal distribution.

3.4 Analysis

In our analysis, we must first note that our dataset consists of $n = 20$ values. This is not very large, and a general recommendation with the applied normality tests is to be careful when using them on datasets of < 30 observations. It must also be noted that we are testing multiple null hypotheses, which increases the likelihood of incorrectly rejecting a null hypothesis. However, a full Bonferroni correction is likely to be very conservative, since our tests are all normality tests and as such are positively correlated and not independent. This simply means that we must keep in mind that when a test rejects the null hypothesis but is still close to our significance level of $\alpha = 0.05$ that this may be a Type I error.

First looking at our results for the original non-transformed dataset, we can see that the QQ-plot did not provide significant visual clues that the dataset may not be normal. None of the empirical distribution tests can reject the normality null hypothesis. The Shapiro-Wilk test did reject the null hypothesis, but the p-value for that test is 0.0332668 which is still close enough to $\alpha = 0.05$ that we must be wary of Type I errors. The kurtosis and skewness values seem to be within acceptable ranges, and the Jarque-Bera test could also not reject the normality null hypothesis. The D'Agostino test did reject the null hypothesis, but the p-value for that test is 0.0330004 which again is still somewhat close to $\alpha = 0.05$. Based on these results and on our concerns noted above, we must conclude that there is not strong evidence to reject the hypothesis that the

data comes from a normal distribution. However, if more observations become available for this dataset, this result should definitely be re-evaluated.

Lastly, looking at our results for the log-transformed dataset, we can see that the QQ-plot is hardly different from the original QQ-plot. None of the empirical distribution tests can reject the normality null hypothesis. The Shapiro-Wilk test can also not reject the null hypothesis anymore. The log transformation decreases the dataset's skewness and kurtosis values, as would be expected, and the Jarque-Bera and D'Agostino tests can both not reject the null hypothesis. Since none of the normality tests can reject the null hypothesis at a significance level of $\alpha = 0.05$, we must conclude that there is not sufficient evidence to reject the null hypothesis that the log transformed data comes from a normal distribution. Since a log-transformation is generally used to remove skewness in datasets and since we can see that the p-values for the Jarque-Bera and D'Agostino tests become significantly higher after the log transformation, we may hypothesize that normality was previously rejected in the Shapiro-Wilk and D'Agostino tests due to the skewness and kurtosis of the data, which has been lessened due to the log transformation. Since the size of our dataset is only $n = 20$ this skewness could be due to chance. Interestingly, the QQ-plot of the data hardly changed after the transformation. We may conclude that when a test is very sensitive to departures from normality, it may be useful to use the log-transformed data instead of the original data, and that normality was rejected in the original dataset on the Shapiro-Wilk test and the D'Agostino test due to a skewness and kurtosis that was slightly too high.

4 Exercise 9: McNemar and Agreement

We have a dataset of 21 patients of which measurements C and K are recorded. The measurements are numerical and a value below 120 means the blood of the patient is too thin. The question is if the two measurements C and K are the same on decision. To test this we will perform the McNemar and Kappa agreement tests. The dataset is described in the introduction of this report.

4.1 Methods

4.1.1 McNemar chi-squared test

To perform the McNemar test we first create two vectors of boolean values for measurements C and K. In our case a boolean measurement is true when the measurement is < 120 and false otherwise. When we have these two vectors of boolean values we compute a contingency table. From this contingency table we can compute the McNemar test statistic with continuity correction using R. The null hypothesis for the McNemar test is that the probabilities of being classified into cells [i,j] and [j,i] are the same. In simpler terms, the null hypothesis represents marginal homogeneity and means that the measurements agree.

4.1.2 Agreement test

For the agreement test we compute the Cohen's Kappa test statistic using R's "irr" package for interrater reliability and agreement. The Kappa test also uses a contingency table to compute a statistic. We will classify the Kappa statistic using the magnitude guidelines by Landis and Koch.

4.2 Results

4.2.1 McNemar chi-squared test

In table 4 you can see the computed contingency table for measurements C and K. In the table x is measurement C and y is measurement K.

y / x	FALSE	TRUE
FALSE	11	5
TRUE	1	4

Table 4: Contingency table for exercise 9. x=C, y=K

The computed p-value of the McNemar chi-squared test is $0.2206714 > 0.05$ which means that we cannot reject the null hypothesis of marginal homogeneity, meaning that we cannot reject that the measurements agree.

4.2.2 Agreement test

We compute the Kappa statistic to be 0.3823529 which means a "fair" agreement, but it is near Landis and Koch's 0.4 threshold for a "moderate" agreement.

4.3 Analysis

Looking at our results, we can see that the McNemar test cannot reject the null hypothesis of marginal homogeneity. The guidelines by Landis and Koch classify the computed Kappa value as representing a "fair" to "moderate" agreement. Since both the Kappa test and the McNemar test suggest that there is agreement we can conclude that there is agreement between measurements C and K regarding the patient's blood being too thin.