Case Studies

"Data Analytics"

Topic

Summer Term 2013

Andrey Chinnov, Sebastian Honermann, Carlos Zydorek

July 11, 2013

Contents

1 Introduction

1.1 Normality as a requirement for statistical methods

Since the normal distribution has some convenient analytical properties, many statistical tests (e. g. Two-sample z-test, One-sample t-test, or Chi-squared test for variance) are based on the assumption that data were drawn from a normal distribution. Furthermore, the central limit theorem for example is applied in many cases in practice. It states that, when under certain conditions a large number of observations stem from the same distribution, their mean will be approximately normally distributed. This statement holds independent of the original distribution the observations are drawn from. Hence, methods for testing sample data on normality are of major importance.

This report aims at describing test methods for normality and conducting them on the Glass data set. Since the hypothesis of normality is rejected for most of the variables, this report is focused on methods for testing on univariate normal distribution. A transformation is applied to convert the data so that it complies with a normal distribution if possible. Nevertheless, a small part of this report deals with multivariate normal distribution.

Test methods for univariate normal distribution are introduced in section 2. Furthermore, a transformation method and a technique for identifying multivariate (more precisely: bivariate) normally distributed data are described. In section 3, the introduced test methods are conducted on the original data as well as on the transformed data. Subsequently, the technique for checking on bivariate normal distribution is demonstrated on an example from the Glass data set. The findings are shortly concluded in section 4.

1.2 The Glass data set

The data set that is investigated in this report is the `Glass` data set from the R package `mlbench`, originally taken from the UCI Repository of Machine Learning Databases (Bache, K., Lichman, M. 2013). It contains 214 observations on 10 variables of glass specimen. The first variable is the refractive index (RI) followed by weight percent in the corresponding oxide of Sodium (Na), Magnesium (Mg), Aluminium (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba) and Iron (Fe). One variable specifies the type of glass as a class attribute. There are seven different types of glass but only six of them are contained in the present data set.

2 Preliminaries

2.1 The Normal Distribution

The (univariate) normal distribution $N(\mu, \sigma^2)$ is defined to have the density

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$.

The univariate normal distribution is generalised by the multivariate normal distribution for higher dimensions. X has a multivariate normal distribution with mean vector μ and covariance $\Sigma > 0$, i.e. $X \sim N_p(\mu, \Sigma)$, if

$$Pr(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$

Contour lines of the plot of a multivariate normal distribution are shaped elliptically. Those ellipsoids are centered at $\mu : \{x : (x-\mu)'\,\Sigma^{-1}(x-\mu) = c^2\}$ with some constant c.

A first approach to check a sample of several variables on multivariate normal distribution is to examine the plot on the appearance of elliptical contour lines. An exemplary contour plot is depicted in figure 2.1. The data is simulated from the mulitvariate normal distribution with the parameters

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{bmatrix} 27 & 15 \\ 15 & 18 \end{bmatrix}$$

and a sample size of 500. The contour lines for this simulated sample are apparently elliptically shaped. Obviously, this graphical method only works for two variables respectively, i. e. for bivariate distributions.

2.2 Test methods for normality

A statistical hypothesis test which tests empirical data on conformance with a certain distribution (or a family of distributions) is called a goodness of fit test. The null hypothesis usually refers to the position that the tested sample has been drawn from a population which is distributed according to the given distribution. Consequently, the alternative hypothesis states that the sample was drawn from a population of any other distribution. In every test, a certain method is used to calculate a test statistic from the data. If the test statistic exceeds a critical value which is computed for the particular distribution at a certain significance level, the null hypothesis is rejected. The p-value is the lowest significance level for which the null hypothesis would still be rejected. It can be interpreted as the probability of getting a result like the present one or an even more extreme result if the null hypothesis is true.

2.2.1 Q-Q-plot

Quantile-Quantile-Plots provide a graphical comparison of the quantiles of two probability distributions. The observed values are plotted against the quantiles of a theoretical distribution. To check for normality, the observations $x_1, \ldots, x_n$ from a sample with size

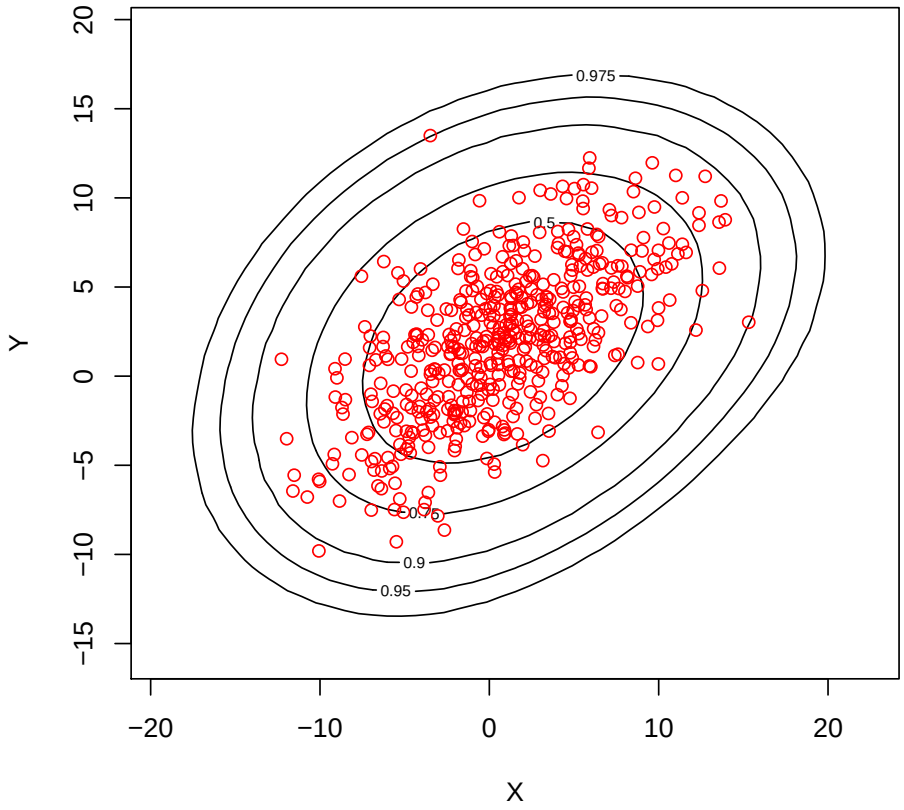


Figure 2.1: Data plot with contour lines of a simulated sample from a bivariate normal distribution

n are ordered ($x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$) and plotted against an assumed cumulative distribution function. Let $x_{(i)}$ denote the sample quantiles. For each sample quantile $x_{(j)}$ with j observations to the left, the proportion of these j observations is approximated by

$$p_{(j)} = \frac{j - \frac{1}{2}}{n}$$

and the theoretical quantiles $q_{(j)}$ are defined by

$$q_{(j)} = \Phi^{-1}(p_{(j)})$$

where Φ is the cumulative distribution function of the standard normal distribution.

In the QQ-Plot these theoretical quantiles $q_{(j)}$ are then plotted against the sample quantiles $x_{(j)}$. Given a sufficient sample size, in the case of normally distributed data the observed quantiles will be linearly related to the theoretical normal distribution quantiles in the QQ-plot.

2.2.2 Shapiro-Wilk test

The Shapiro-Wilk test is a statistical procedure for testing a complete sample for normality. Basically, the test statistic, denoted as W, indicates correlation of the observed

quantile values with the assumed cumulative distribution function quantiles and is therefore close to the idea of QQ-plots or probability-plots.

$$W = \frac{\sum_{i=1}^{n}(a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where

a_i denotes the normalized "best linear unbiased" coefficients and

x_i denotes the ordered observations.

In detail, W is obtained by dividing the best linear estimate of the slope of a linear regression of the ordered observations x_i by the actual variance of the ordered observed sample values.

The resulting test statistic W is then compared with a critial W-value for a given sample size n. The resulting significance level can be looked up in existing tables, however, most statistic software (including R) automatically determines the according p-value via the Monte Carlo method when displaying the test statistic W.

In general, the statistical power of tests decreases with smaller sample sizes. Compared to other statistical tests, like the Kolmogorov-Smirnov test or the Chi-Squared-Test, the Shapiro-Wilk test's statistical power decreases less with smaller samples. Since 4 of 6 glass type samples (glass type 3, 5, 6, and 7) have a size of $n < 50$, the test result is quite interesting for the tests of the subdatasets.

In order to conduct the Shapiro-Wilk test in R the function `shapiro.test()` from the package `stats` is used.

In some cases all observations of certain variables are identical. In these cases the Shapiro-Wilk-Statistic is not applicable for these specific variables since the expression

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

in the denominator, denoting the variance, sums up to 0. Tests can therefore only be run if such variables are excluded from the testing procedure.

2.2.3 Pearson's chi-squared test

Pearson's chi-squared goodness of fit test is used to test whether data from a sample are distributed according to an arbitrary theoretical distribution. The main idea of this test is to divide the observations $x_1, \ldots, x_n$ into several pairwise disjoint classes $C_1, \ldots, C_K$ and compare the empirical frequencies within these classes to the theoretical frequencies, which are expected if the data complies to the hypothetical distribution. If the histograms of the sample data and the expected densities are plotted together (see figure 2.2), the area of density that is not overlapped by both histograms can be understood as a kind of indicator for the likelihood that the sample is drawn from a population which is distributed according to the hypothetical distribution: The larger the non-overlapping area, the less likely it is that the sample is drawn from a population with the assumed distribution. However, the test statistic of the chi-squared test is calculated differently, namely by the sum of the squared differences between observed frequencies O_k and expected frequencies

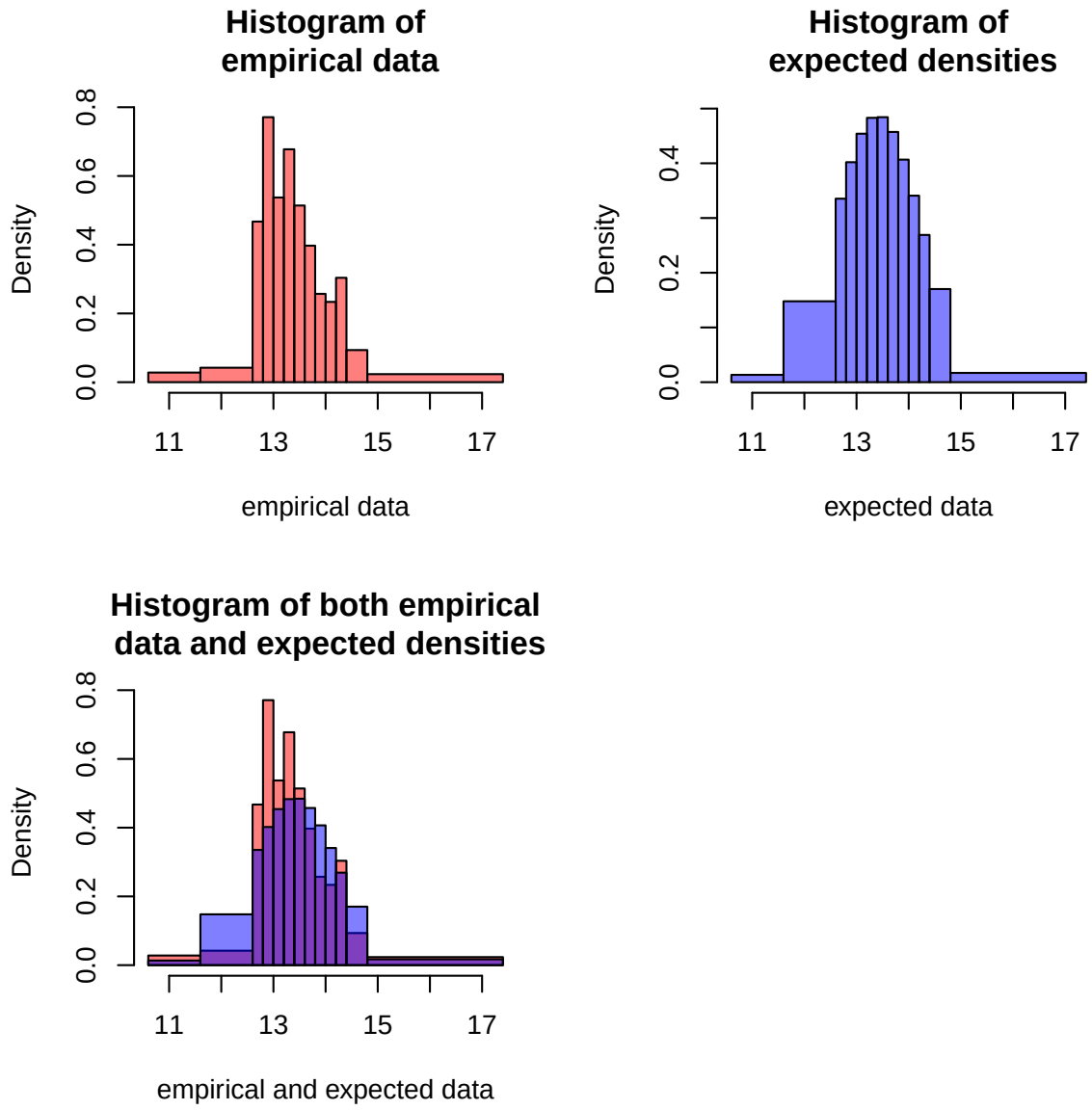


Figure 2.2: Exemplary histrograms of a data sample, expected densities for a normal distribution with parameters estimated from the sample and a combined histogram of these both histograms.

E_k divided by the expected frequencies for each class k of the overall K classes. Thus, the test statistic is calculated by

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}$$

Larger differences of observed and expected values indicate a lower compliance to the assumed distribution. However, the addends are not weighted (neither by the size of a class nor by the frequencies within a class nor by any other means). Therefore, the class bounds should be chosen equidistant or in such a way that the classes contain preferably the same number of observations, or should be chosen according to similar reasonable rationales. The test statistic is approximately χ^2-distributed with $K-1$ degrees of freedom – the larger the sample size, the better the approximation. A sample size that is too small

can be a reason for the approximation being insufficient. Moreover, for each parameter of the hypothetical distribution which is estimated from the data sample, one degree of freedom is lost; the number of estimated parameters is denoted by p. The test statistic is determined under the null hypothesis that the sample is distributed according to the assumed distribution and the chi-squared test is defined as

$$\delta(Y) = \begin{cases} 1 & \text{if } \chi^2 > F^{-1}(1-\alpha) \\ 0 & \text{otherwise} \end{cases} \qquad \text{with } F = \chi^2_{K-1-p}$$

for a given significance level α where Y is a multinomial distributed random variable denoting the counts of observations in each class with $Y_k = |\{i : X_i \in C_k\}|$.

A common requirement for a sufficient approximation demands the minimum number of observations in each class not to fall below five (Steel, R.G.D., Torrie, J.H. 1960, p. 350). Hence, marginal or even inner classes have to be unified in some cases in order to achieve a sufficient class size. A drawback of Pearson's chi-squared test is its inconsistency caused by information reduction, i. e. information about the data sample is lost in the process of categorising the observations in classes. As a consequence, different class bounds can lead to different test results. Furthermore, this test is rather suited for large sample sizes.

2.2.4 Kolmogorov-Smirnov test

Like the other tests introduced in the previous chapters, the Kolmogorov-Smirnov (subsequently abbreviated as KS) is used for testing whether a given univariate sample $x = (x_1, x_2 \dots x_n)$ with unknown distribution $\mathbb{P}$ is distributed according to a completely determined distribution $\mathbb{P}_0$. Based on the test a decision is made between the following two hypotheses:

$$\begin{aligned} H_0 &: \mathbb{P} = \mathbb{P}_0, \\ H_1 &: \mathbb{P} \neq \mathbb{P}_0. \end{aligned}$$

This decision is made according to the KS test statistics and a given significance level α.

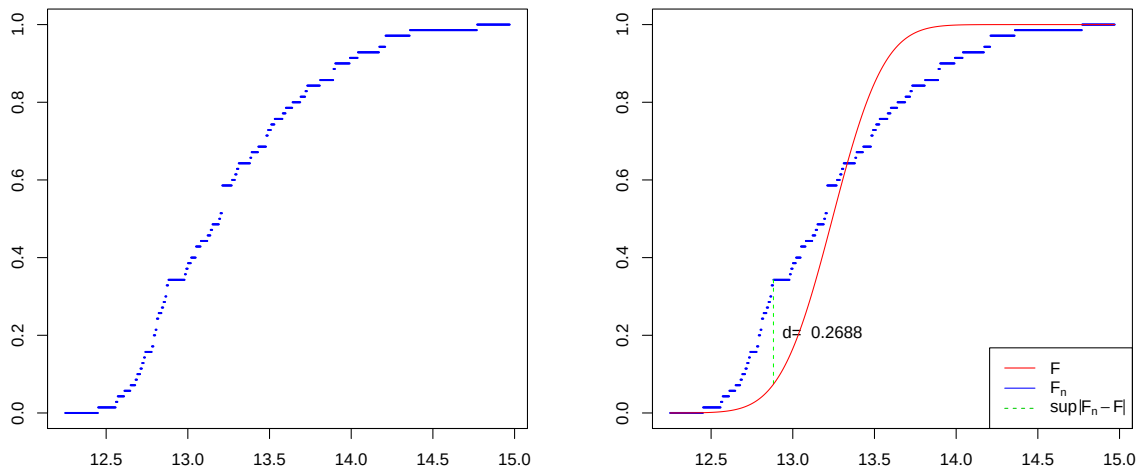
Definition 1 *For a given univariate sample $x = (x_1, x_2, \dots, x_n)$ the function*

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$$

is called empirical cumulative distribution function (c. d. f.), where $\mathbb{1}_{\{x_i \leq x\}}$ is an indicator function defined as follows: $\mathbb{1}_{\{x_i \leq x\}}(x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{otherwise.} \end{cases}$

The exemplary graph of such a function is depicted in the Figure 2.3a.

The main idea of the KS test is the analysis of the difference between the given cumulative distribution function (c. d. f.) F and the empirical c. d. f. F_n. Since both theoretical and empirical functions belong to normed space of bounded functions $\mathbb{B}(\mathbb{R})$ (all values are between 0 and 1), this difference can be measured as a distance $\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$. Figure 2.3b illustrates the calculated distance between the empirical c. d. f. and the theoretical normal c. d. f. with respect to the given parameters *sample mean* and *sample variance*.



(a) **(b)**

Figure 2.3: Empirical c. d. f. for Sodium (Na) vector (a) and theoretical normal c. d. f. with sample mean and sample variance (b)

The KS test statistics is defined as follows:

$$D_n = \sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

If value D_i is considered for different $1 \leq i \leq n$, then the sample $\hat{D}_n = (D_1, \ldots, D_n)$ is obtained that also complies with some distribution $\mathbb{D}_n$. It can be shown that if hypothesis H_0 is true, then this distribution does not depend on the c. d. f. F and therefore can be tabulated. Moreover, Kolmogorov proved that if n is large enough, the distribution function of $\mathbb{D}_n$ can be approximated by Kolmogorov-Smirnov distribution function $H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} \exp^{-2i^2t^2}$, i. e. for each positive value t the probability $P(D_n \leq t) \to H(t)$ when $n \to \infty$. More details can be found in the textbook (DeGroot, Morris, H., Mark, J. 2003).

The KS test uses the decision rule

$$\delta = \begin{cases} H_0 & : \quad D_n \leq c \\ H_1 & : \quad D_n > c \end{cases},$$

where the critical value c depends on the significance level α and can be calculated from the following equations:

$$\alpha = P(\delta \neq H_0 | H_0) = P(D_n > c | H_0) = 1 - P(D_n \leq c | H_0) \approx 1 - H(c).$$

As mentioned above the last expression can only be considered appriximately equal when n is relatively high. Otherwise the table values for $\mathbb{D}_n$ distribution should be used. Hence, $c \approx H^{-1}(1 - \alpha)$.

Although the KS test is commonly used, it has a huge drawback. Namely it considers only completely defined theoretical c. d. f. . When testing for normality both parameters μ and σ^2 have to be predefined. But usually they are a priori unknown when a sample is going to be tested. Of course it can be managed by assigning sample mean and sample

variance as unknown parameters (and initial KS test suggests to do that) but they are not always the best choice. The following example of the Sodium (Na) distribution illustrates this proposition.

For the univariate Sodium (Na) sample of 70 observations the sample mean $\bar{\mu}$ =13.2423 and the sample variance $\bar{\sigma^2}$ = 0.2493 are calculated. For these values the theoretical c.d.f. is determined (Figure 2.3b). Then the value of KS statistic is calculated: $D_n = \sqrt{n} \cdot \sup_x |F_n(x) - F(x)|$ =2.2493. The critical value for the significance level $\alpha = 0.01$ is equal to $c = H^{-1}(1-\alpha)$=1.6276.

Since $D_n > c$, the null hypothesis is rejected by the KS test. But that does not mean that the sample is not normally distributed. It only means that the hypothesis is rejected for the normal distribution with these specific parameters sample mean and sample variance.

In order to manage this issue, the KS test is improved by solving the following optimisation problem

$$KS(\mu, \sigma^2) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x, \mu, \sigma^2)| \to \min .$$

When the new values $\hat{\mu}$ and $\hat{\sigma^2}$ that minimise the value $KS(\mu, \sigma^2)$ are found, the general KS test is performed with respect to these values of parameters. For solving this optimisation problem the following R code is used:

```
KS= function(param) {
        #discretization of a line segment
        seq = seq(from = min(dat)-0.2, to = max(dat)+0.2, length.out=1000)
        #values of empirical c.d.f.
        empdat = sapply(seq, function(x) {empiric(x,dat)})
        #values of theoretical c.d.f.
        theordat = pnorm(seq,param[1],abs(param[2]))
        #difference between the values
        dif=theordat-empdat
        absdiff=abs(dif)
        max(absdiff)
}
#optim is a predifined R function in stats package
#defalut method of optimisation is Nelder and Mead (1965)
KSoptim = optim(c(mean,Cov),KS)
KSoptim$par

[1] 13.1769501  0.4682486

KSoptim$value

[1] 0.07870673
```

These new parameters $\hat{\mu}$ =13.1770 and $\hat{\sigma}^2$ =0.4682 are taken as parameters of a new theoretical normal c. d. f. and a new distance is calculated (Figure 2.4).

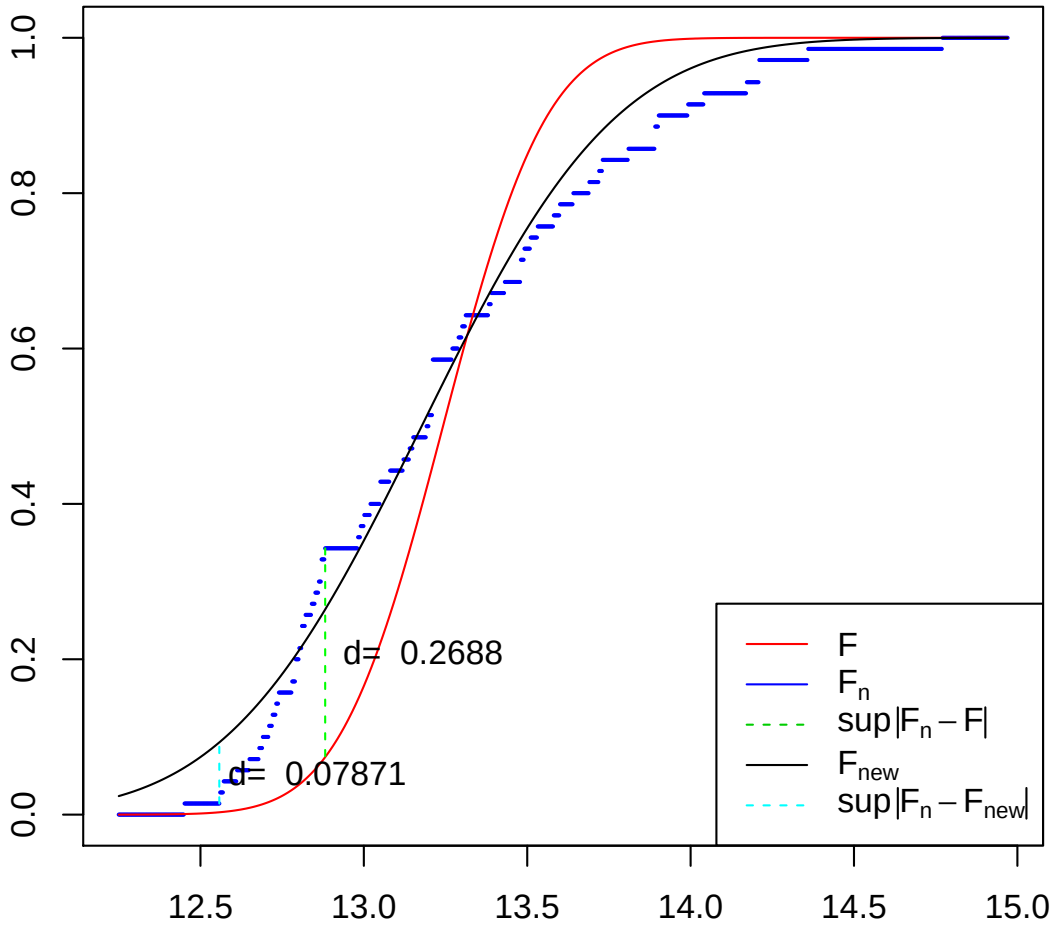


Figure 2.4: Normal c. d. f. with optimised parameters in comparison to the old c. d. f. with sample mean and sample variance.

The new value of the KS statistics is $D_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x, \hat{\mu}, \hat{\sigma}^2)|$ =0.6585 which is smaller than the critical value c =1.6276. Therefore, the null hypothesis H_0 cannot further be rejected by KS-test. For all further samples, the improved KS test is used.

The approximation problem with the H function remains in the improved KS test as well. Therefore this approximation is only applicable when the sample size is relatively high. Otherwise, the tabular values for $\mathbb{D}_n$ distribution should be used.

2.3 Box-Cox-transformation

If data are not normally distributed, it can still be transformed to fit to a normal distribtion in some cases. One possibility is the Box-Cox-transformation. It is a family of

parameterised power tranformations:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases} \quad \text{for } x > 0$$

The optimal parameter for specific observations $x_1, \ldots, x_n$ can be determined by a maximum-likelihood estimation, maximising the log likelihood

$$l(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln(x_j)$$

$$\text{with } \overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)}$$

However, a Box-Cox-transformation does not ensure that the data is normally distributed thereafter. One reason that a sample cannot be properly transformed could be that it is not unimodal. Histograms and QQ-plots of a sample from a unimodal distribution are depicted in figure 2.5. Data that is generated from a Weibull distribution can be transformed to approximately normally distributed values quite well as can be recognised by the histogram and the QQ-plot. In contrast, it is not possible to properly transform a sample that is combined from two different distributions (here with different scale parameters of the Weibull distribution) as shown in figure 2.6. By the combination of two samples with different mean values a bimodal sample emerges preventing the underlying data to be transformed to a unimodal sample (namely a normally distributed sample) by a simple function. Furthermore, noisy data is not suited for Box-Cox-transformation either because the Box-Cox-function is applied on the whole sample (and not only the "noisy parts").

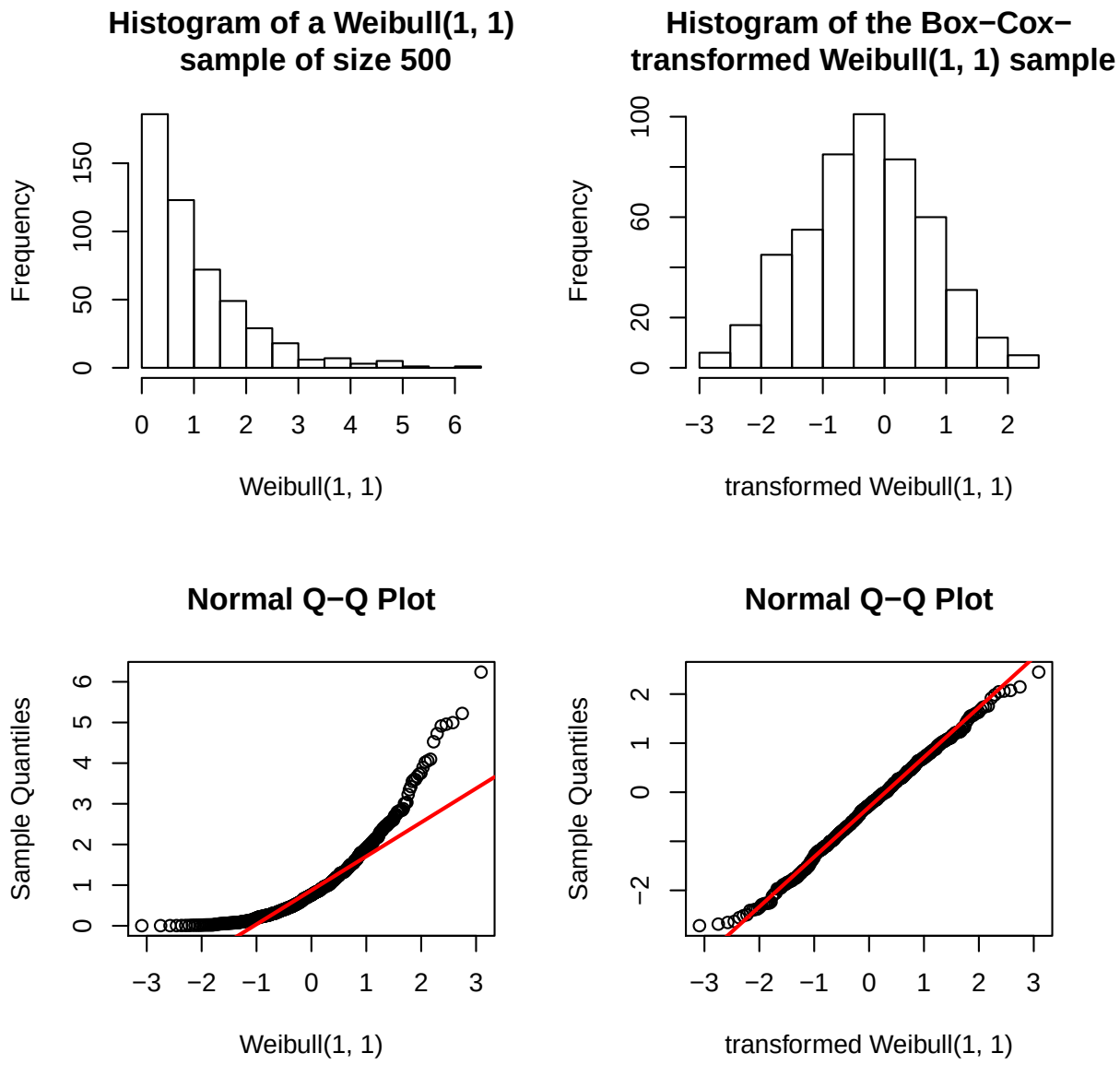


Figure 2.5: Histograms and QQ-plots of a Weibull(1, 1) simulated sample of size 500 and of the Box-Cox-transformed data

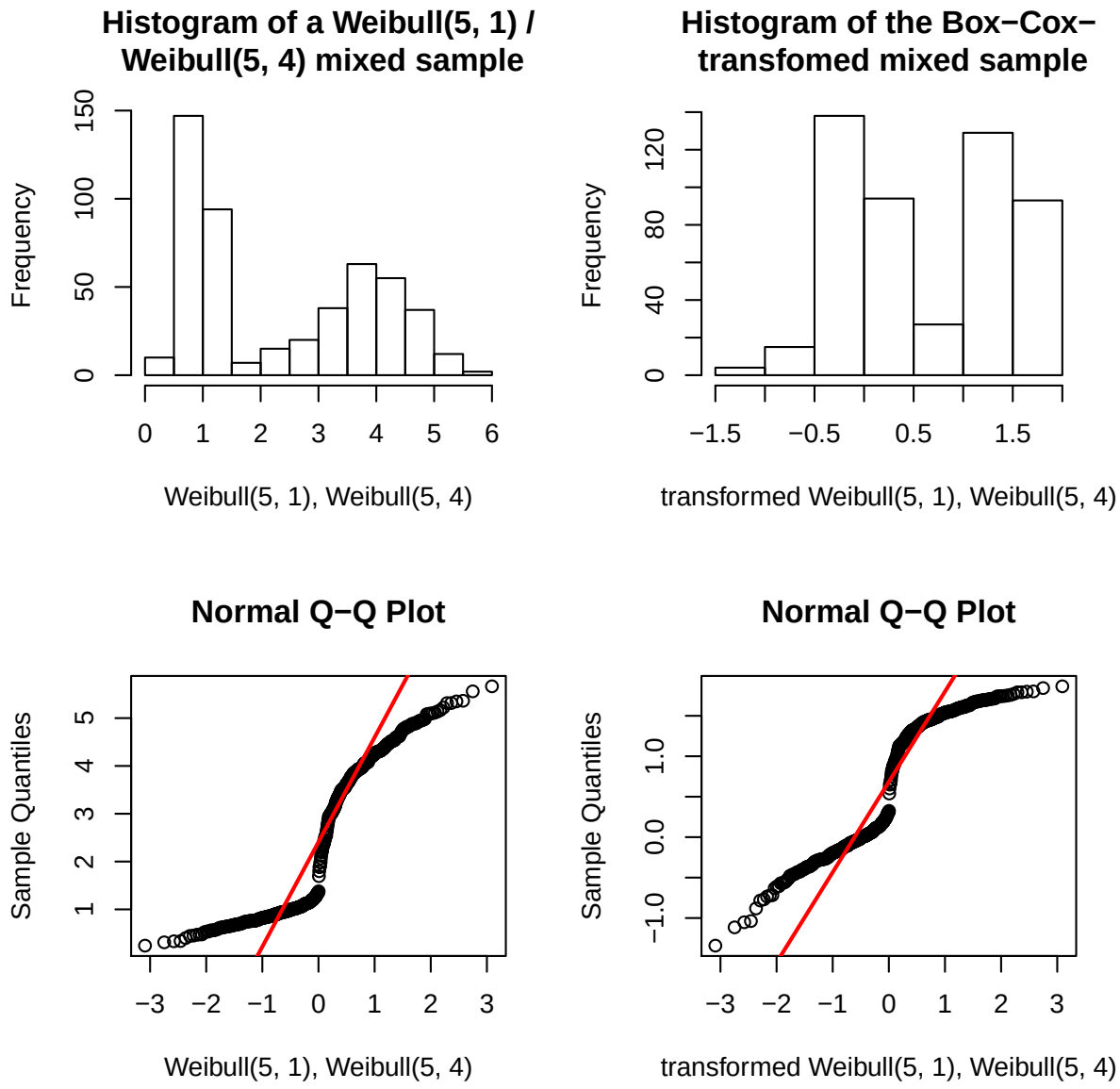


Figure 2.6: Histograms and QQ-plots of a mixed sample composed of a Weibull(5, 1) simulated sample and a Weibull(5, 4) simulated sample (each of size 250) and of the Box-Cox-transformed data

3 Testing the data sample for normality

3.1 Testing original data

The test methods for normality that were introduced in section 2.2 are now applied on the glass data set. For each method, the whole sample is tested first. Since it can be assumed that the different glass types are distinct in terms of the underlying distribution of variable values, the tests are conducted on the individual types as well (where applicable). It appears reasonable to skip particular variables whose values predominantly consist of zeros (more than 50 % of the observations, see table 3.1) because this indicates that those variables are not normally distributed anyway (moreover, this can lead to complications for some methods).

type	skipped variables
1	Ba
2	Ba
3	Ba, Fe
5	Ba, Fe
6	K, Ba, Fe
7	Mg

Table 3.1: Skipped variables for the particular glass types due to too many zero values

3.1.1 Q-Q-plot

For the graphical analysis the quantiles for all elements in the glass-dataset where plotted for the whole dataset and for each glass type separately. As depicted in Figure 3.1, the QQ-Plots for the five variables refractive index (RI), Sodium (Na), Aluminium (Al), Silicon (Si), and Calcium (Ca) do suggest that linear relationships could be assumed. The other four variables Magnesium (Mg), Barium (Ba), Potassium (K), and Iron (Fe) clearly are not linearly related to the hypothetical quantiles.

The graphical comparison of the subdatasets shows that for most of the cases there seems to be no linear relationship between the theoretical normally distributed quantiles and the observed values. For some cases in the subdatasets, however, based on a graphical inspection a linear relationship can be assumed given some outliers. As shown in Figure 3.3 Magnesium (Mg), Aluminium(Al), and Silicon (Si) in glass type 1 and Ca in glass type 7 show approximately linear relationships with just a few outliers. For some observations (e.g. Na in glass type 5 and RI in glass type 6 in Figure 3.4) a linear relationship seems to be plausible, however, the total number of observations in these cases is considered too small to make conclusions about a hypothetical linear relationship (glass type 5: $n = 13$; glass type 6: $n = 9$).

3.1.2 Shapiro-Wilk test

Performing the Shapiro-Wilk test on the complete dataset at hand, according to the p-values the null hypothesis (the data the sample is taken from is normally distributed) is rejected for all the elements. Testing the different glass types separately, the p-values of majority of the variables are still very small. Consequently, the null hypotheses can be

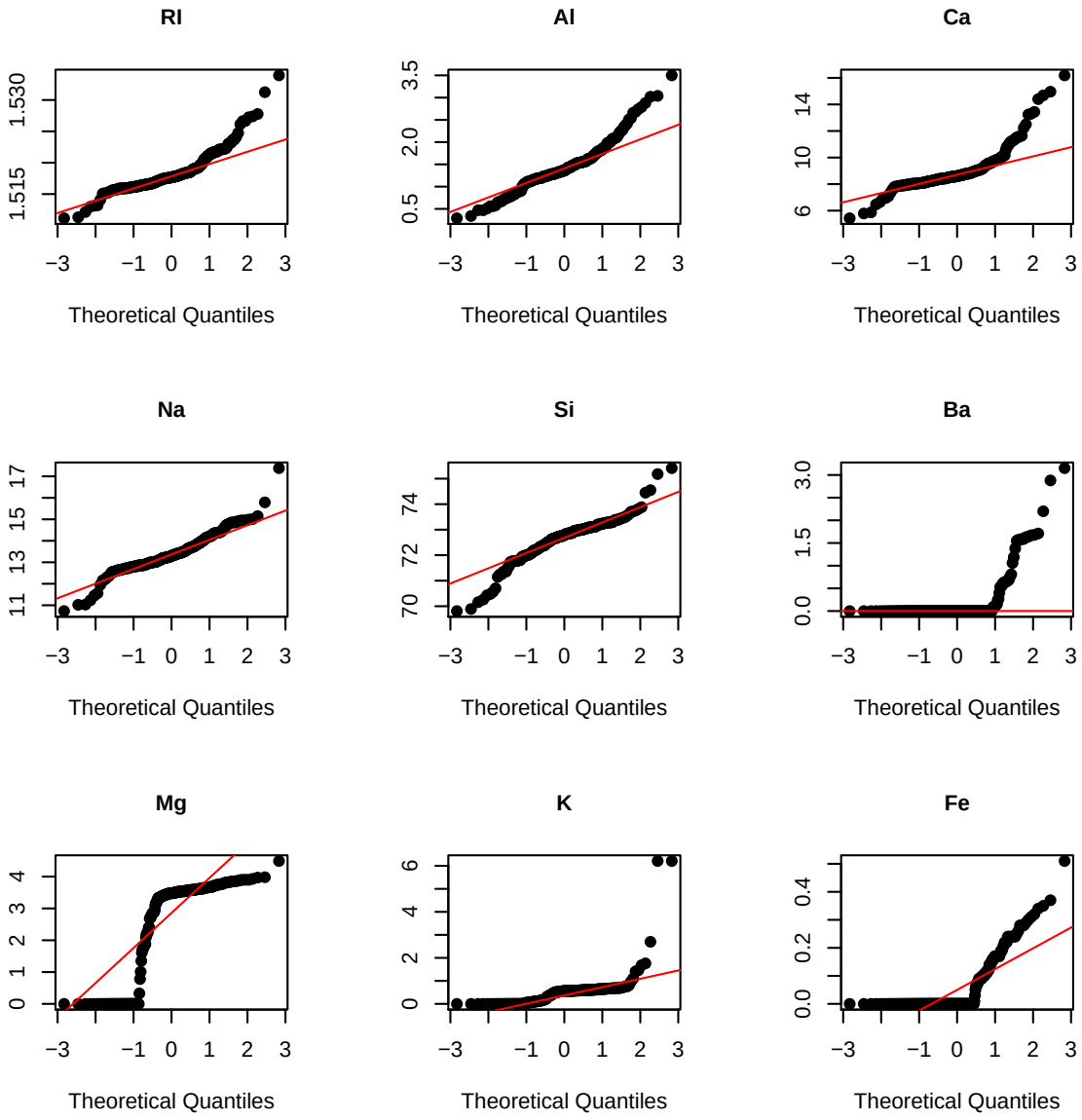


Figure 3.1: Exemplary QQ-Plots from the full data sample

rejected for almost all separate cases. For a total 16 cases (over all 6 glass types) the test returned a p-value above or equal to the set significance level of 1 % (Tables 3.3 to 3.8). However, the majority of the cases for which the null hypothesis holds stems from the three subdatasets containing few observations (Tables 3.5 to 3.7). For glass type 1, H_0 is rejected for all the variables (see Table 3.3). For glass type 2, it only holds for Aluminium (Al) and for glass type 7 only for Aluminium (Al) and Barium (Ba) (see Tables 3.4 and 3.8). More than 80 % of the total rejections in the 6 tests (13 out of 16) stem from the three smaller subdatasets glass type 3, glass type 5, and glass type 6. These results raise the question if the low number of rejections is due to the smaller sample size in these cases. Looking only at the glass types 1, 2, and 7, a normal distribution can be rejected for most of the cases.

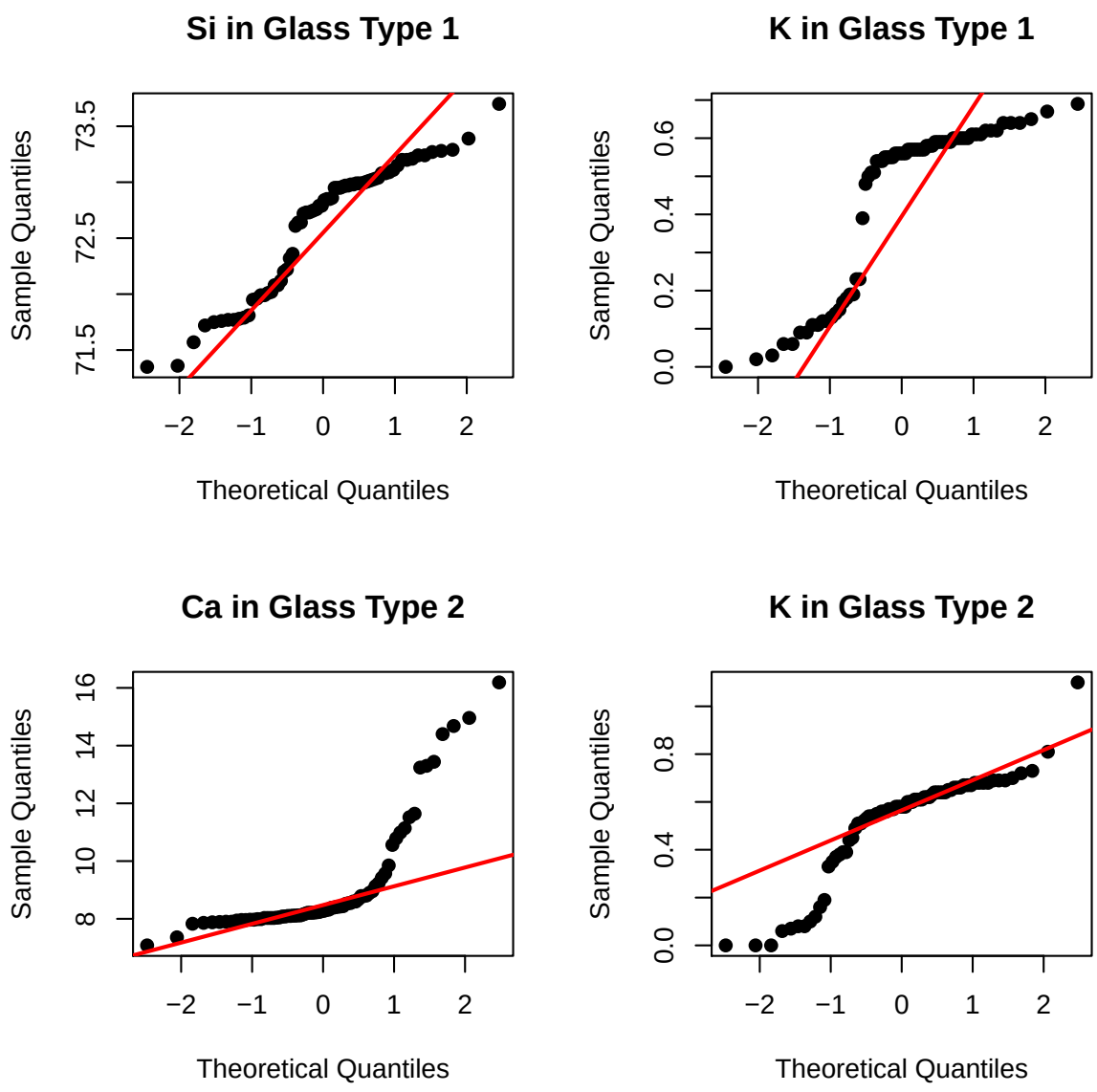


Figure 3.2: Exemplary QQ-Plots from Glass Type 1 where a graphical inspection does not suggest a linear relationship

3.1.3 Pearson's chi-squared test

As mentioned in section 2.2.3, Pearson's chi-squared test is not suited for rather small sample sizes because of the approximation via the chi-squared distribution. Concerning the given data, the samples of type 3 glass (17 observations), type 5 glass (13 observations) and type 6 glass (9 observations) are not large enough to ensure a viable test result. Hence, the data belonging to those types will not be considered for separate tests. However, it will remain in the overall data sample of all types. The minimum size of observations in each class is set to five and the number of initial classes (i. e. number of classes before unifying) will be ten. The first tests are conducted on the whole data set for each variable. The results are shown in table 3.9. For two variables, it is not possible to determine a test result with the given parameters: The observations of the variables Potassium (K) and

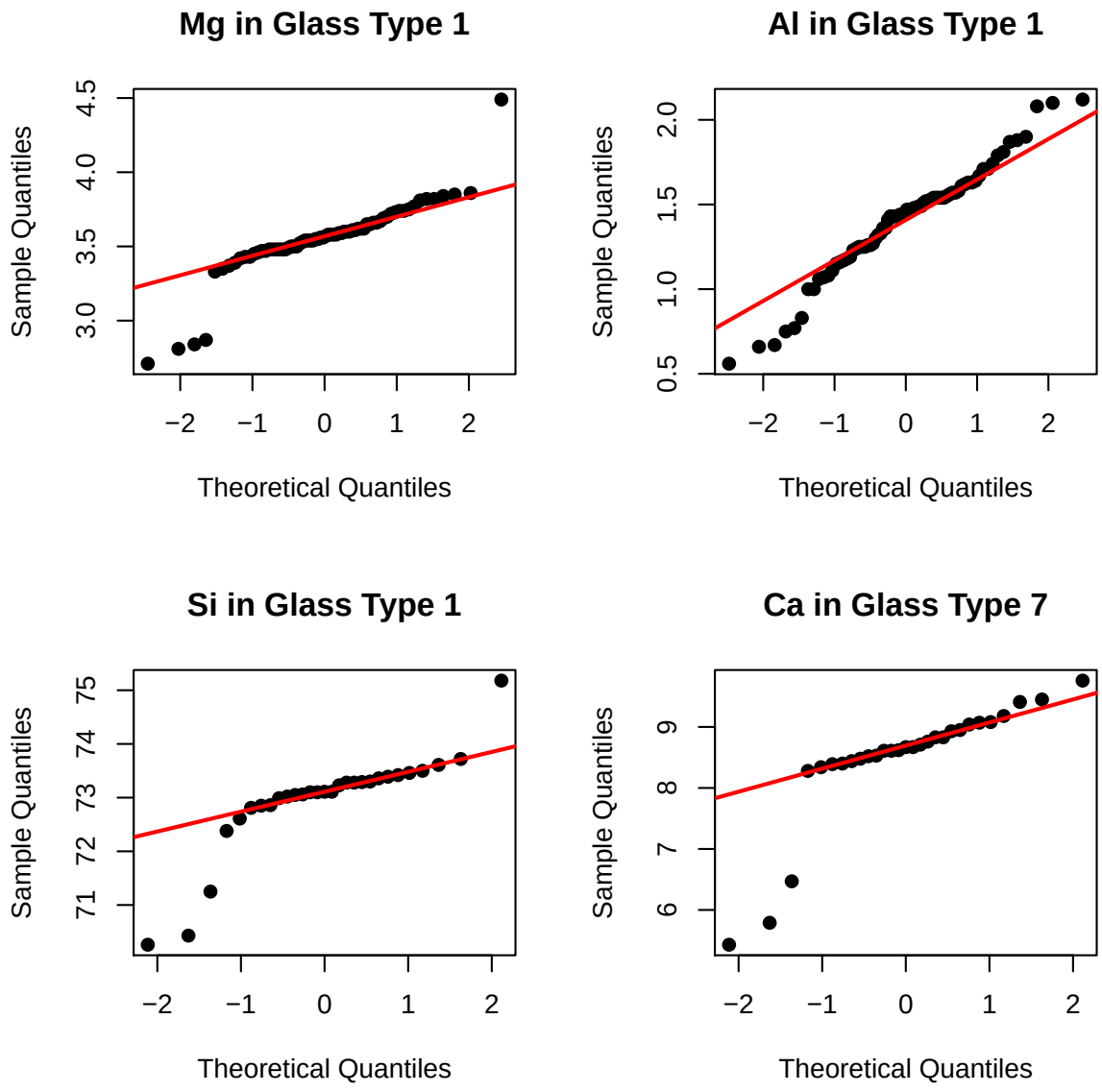


Figure 3.3: QQ-Plots of the cases where a linear relationship seems plausible

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.87	0.01	NA	1.0766713449726e-12	yes
Na	0.95	0.01	NA	3.4655430546966e-07	yes
Mg	0.7	0.01	NA	< 1.0e-15	yes
Al	0.94	0.01	NA	2.08315629600399e-07	yes
Si	0.92	0.01	NA	2.17503176825416e-09	yes
K	0.44	0.01	NA	< 1.0e-15	yes
Ca	0.79	0.01	NA	< 1.0e-15	yes
Ba	0.41	0.01	NA	< 1.0e-15	yes
Fe	0.65	0.01	NA	< 1.0e-15	yes

Table 3.2: Test results of the Shapiro-Wilk test on the whole data sample

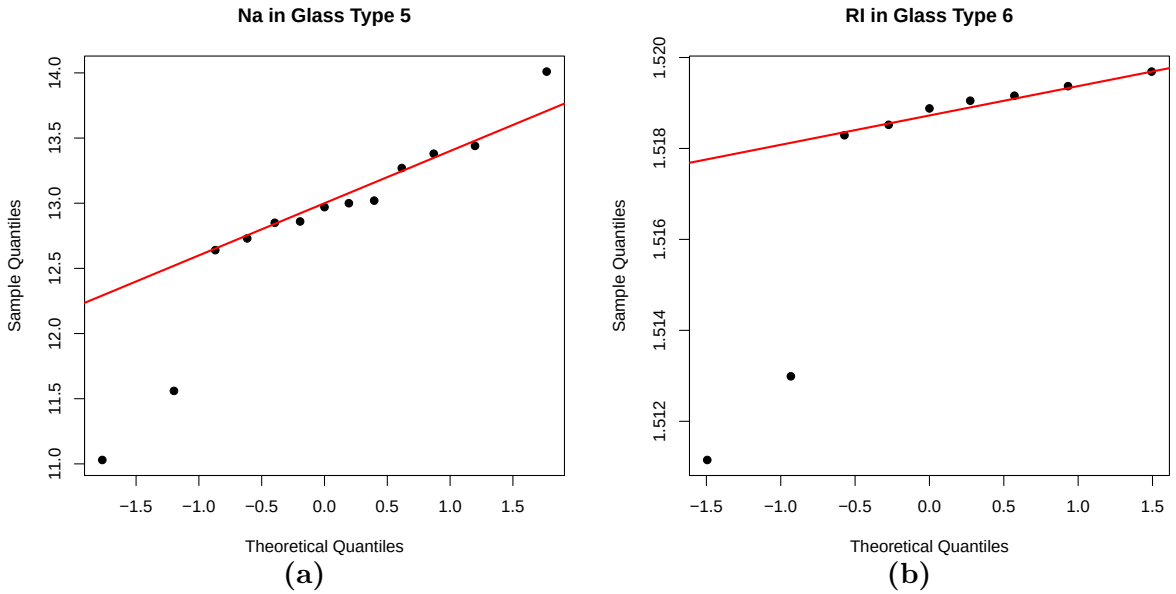


Figure 3.4: QQ-Plots of possible linear relationships with very small sample sizes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.88	0.01	NA	6.36192013015468e-06	yes
Na	0.95	0.01	NA	0.00459078607995831	yes
Mg	0.82	0.01	NA	8.02702432879544e-08	yes
Al	0.9	0.01	NA	5.42971629496434e-05	yes
Si	0.91	0.01	NA	0.000117060780025464	yes
K	0.77	0.01	NA	3.14049093233846e-09	yes
Ca	0.93	0.01	NA	0.00103561283726753	yes

Table 3.3: Test results of the Shapiro-Wilk test on type 1 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.72	0.01	NA	1.08673299441225e-10	yes
Na	0.9	0.01	NA	1.82678898797777e-05	yes
Mg	0.64	0.01	NA	2.38749395349814e-12	yes
Al	0.97	0.01	NA	0.0591136375348645	no
Si	0.88	0.01	NA	3.75379726827069e-06	yes
K	0.84	0.01	NA	1.54088105687157e-07	yes
Ca	0.67	0.01	NA	7.17521765879123e-12	yes

Table 3.4: Test results of the Shapiro-Wilk test on type 2 glass

Barium (Ba) are divided only into three classes respectively after the unification of classes in order to fulfill the requirement of minimum class size. Since one degree of freedom is subtracted always and two degrees of freedom are subtracted for the estimation of the mean value and the standard deviation, zero degrees of freedom remain and so the critical value cannot be calculated. For each of the other variables, the hypothesis of normality is clearly rejected for the given significance level.

The results for type 1 glass (table 3.10) are slightly different; in this case, the results

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.83	0.01	NA	0.00554477459285981	yes
Na	0.95	0.01	NA	0.42694777005566	no
Mg	0.93	0.01	NA	0.2109500928052	no
Al	0.95	0.01	NA	0.427341606866833	no
Si	0.89	0.01	NA	0.0442493392726916	no
K	0.76	0.01	NA	0.000539393334400821	yes
Ca	0.92	0.01	NA	0.147928579200966	no

Table 3.5: Test results of the Shapiro-Wilk test on type 3 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.9	0.01	NA	0.121477380963252	no
Na	0.88	0.01	NA	0.0623633158882554	no
Mg	0.75	0.01	NA	0.00177054101772265	yes
Al	0.79	0.01	NA	0.00461960768904678	yes
Si	0.9	0.01	NA	0.134368361027878	no
K	0.59	0.01	NA	5.18709751538094e-05	yes
Ca	0.85	0.01	NA	0.0259126903932093	no

Table 3.6: Test results of the Shapiro-Wilk test on type 5 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.69	0.01	NA	0.00107643062031905	yes
Na	0.7	0.01	NA	0.00125686447739013	yes
Mg	0.8	0.01	NA	0.0192642436690133	no
Al	0.88	0.01	NA	0.141250903376574	no
Si	0.78	0.01	NA	0.0119496897911561	no
Ca	0.92	0.01	NA	0.404068518356898	no

Table 3.7: Test results of the Shapiro-Wilk test on type 6 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.82	0.01	NA	0.000228738547251759	yes
Na	0.87	0.01	NA	0.00202603090330278	yes
Al	0.96	0.01	NA	0.289110586782862	no
Si	0.78	0.01	NA	3.1655457932838e-05	yes
Ca	0.72	0.01	NA	3.98543372675882e-06	yes
Ba	0.91	0.01	NA	0.0206467993594091	no

Table 3.8: Test results of the Shapiro-Wilk test on type 7 glass

can be determined for each variable (except for Barium (Ba), which has been dropped beforehand) and the hypothesis of normality is rejected for each variable but Sodium (Na). The p-value for Sodium is comparably high amounting to approximately 0.52. It is well recognisable that the observed class frequencies for Sodium fluctuate around the expected class frequencies under the hypothesis of a normal distribution with the according parameters (table 3.11). The good compliance of empirical and hypothetical data for this variable is illustrated in figure 3.5. In general, the p-values for this part of the sample are

variable	test statistic	sig. level	critical value	p-value	rejected
RI	64.95	0.01	13.28	2.64011035255862e-13	yes
Na	36.99	0.01	13.28	1.80797974702607e-07	yes
Mg	158.3	0.01	11.34	< 1.0e-15	yes
Al	27.2	0.01	9.21	1.24084046404516e-06	yes
Si	38.85	0.01	13.28	7.4876188027595e-08	yes
K	95.97	0.01	NA	NA	NA
Ca	131.13	0.01	13.28	< 1.0e-15	yes
Ba	31.37	0.01	NA	NA	NA
Fe	70.96	0.01	13.28	1.4210854715202e-14	yes

Table 3.9: Test results of the chi-squared test on the whole data sample with ten initial classes

higher than those for the whole sample.

variable	test statistic	sig. level	critical value	p-value	rejected
RI	28.01	0.01	9.21	8.26265138420545e-07	yes
Na	3.25	0.01	13.28	0.51688441877949	no
Mg	18.81	0.01	6.63	1.44068580684165e-05	yes
Al	23.55	0.01	11.34	3.10284613768141e-05	yes
Si	23.68	0.01	13.28	9.26014020323773e-05	yes
K	114.86	0.01	11.34	< 1.0e-15	yes
Ca	22.58	0.01	15.09	0.000405198755082603	yes
Fe	18.65	0.01	9.21	8.91413549507503e-05	yes

Table 3.10: Test results of the chi-squared test on type 1 glass with ten initial classes

class	frequencies	
(interval)	observed	expected
]12.4, 12.8]	15	13.15
]12.8, 13]	12	8.81
]13, 13.2]	9	10.68
]13.2, 13.4]	11	11.04
]13.4, 13.6]	8	9.74
]13.6, 14]	9	12.06
]14, 14.8]	6	4.52

Table 3.11: Observed end expected frequencies of items in the classes for the variable Sodium of type 1 glass

The test results for observations of type 2 glass are summarised in table 3.12. The null hypothesis is rejected for all variables except for Aluminium (Al) and Silicon (Si). The p-values for these variables are however rather small (approximately 0.02 and 0.04).

For the observations of type 7 glass, test results (table 3.13) are only available for the variable Aluminium (Al). Due to the small sample size of 29 observations, most of the initial classes are joined so that no degree of freedom remains for the chi-squared distribution function. The hypothesis of normality is not rejected for the data of Aluminium.

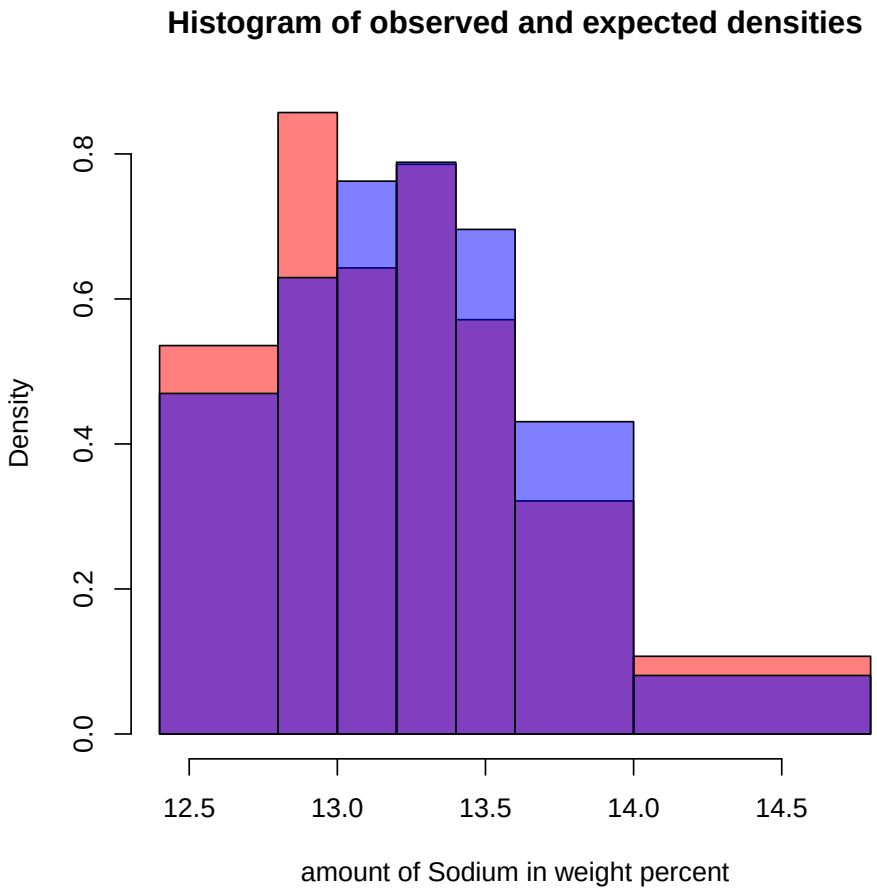


Figure 3.5: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Sodium of type 1 glass

As mentioned in section 2.2.3, Pearson's chi-squared test is inconsistent when the number or bounds of classes are changed. This inconsistency can also be observed with the present data set. The test have also been conducted with 30 initial classes each (see tables A.1 to A.4 in the appendix) with partly different results. Whereas with ten initial classes, there are not enough classes left for most of the variables of type 7 glass to be tested, the data is divided in a sufficient number of classes when using 30 initial classes. Above all, the null hypothesis is not rejected for Aluminium (Al) of type 2 glass with ten initial classes but it is rejected with 30 initial classes while the opposite is true for Sodium (Na). In general, the p-values can alternate much with different classes; so the rather high p-value for Sodium of type 1 glass (~ 0.52) with ten initial classes decreases to approximately 0.02 with 30 initial classes. On the contrary, the p-value for Aluminium of type 7 glass (~ 0.06) increases to approximately 0.19. These different impacts on the test results are due to two opposing effects: First, with more classes there are more degrees of freedom for the chi-squared distribution and thus the critical value increases. Second, the test statistic tends to increase as well because the observations have to fit to smaller classes more precisely; or in other words, observations may be distorted (relatively to the hypothetical expectations) within a large class so that differences between empirical and hypothetical data do not raise the test statistic as much as the same observations would

variable	test statistic	sig. level	critical value	p-value	rejected
RI	27.92	0.01	9.21	8.6430973300633e-07	yes
Na	8.2	0.01	6.63	0.00418393039163056	yes
Mg	66.57	0.01	6.63	< 1.0e-15	yes
Al	9.41	0.01	11.34	0.024332262426528	no
Si	6.24	0.01	9.21	0.0441247638253744	no
K	41.06	0.01	6.63	1.47495904379014e-10	yes
Ca	71.68	0.01	9.21	< 1.0e-15	yes
Fe	16.75	0.01	11.34	0.000794876178432768	yes

Table 3.12: Test results of the chi-squared test on type 2 glass with ten initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	19.93	0.01	NA	NA	NA
Na	1.4	0.01	NA	NA	NA
Al	3.42	0.01	6.63	0.0644860281274806	no
Si	4.84	0.01	NA	NA	NA
K	13.14	0.01	NA	NA	NA
Ca	11.93	0.01	NA	NA	NA
Ba	0.2	0.01	NA	NA	NA

Table 3.13: Test results of the chi-squared test on type 7 glass with ten initial classes

if they were divided into smaller classes (making the distortion "measurable").

3.1.4 Kolmogorov-Smirnov test

First and foremost, the Kolmogorov-Smirnov test is conducted on the whole dataset for each variable. As depicted in table 3.14, the results indicate that only for Magnesium (Mg), Potassium (K), Barium (Ba), and Iron (Fe) H_0 can be rejected. For all other variables, the hypothesis that the data the sample stems from is normally distributed cannot be rejected.

Performing the test on the subclasses type 1 glass, type 2 glass, and type 7 glass for each variable, the results further support the notion that data in the different subclasses is normally distributed. While for type 1 glass H_0 can only be rejected for Potassium (K) and Iron (Fe) (see table 3.15) and for type 2 glass H_0 can only be rejected for Iron (Fe) (see table 3.16), the tests on type 7 glass show that for none of the contained elements the null hypothesis can be rejected at the 1 % significance level.

3.2 Testing transformed data

The same tests are now conducted on the data that have been Box-Cox-transformed with a parameter that is estimated by the maximum likelihood method. For some variables, an estimation is not possible because the algorithm does not converge or, as in most cases, not all of the observations of one variable are strictly positive.

variable	test statistic	sig. level	critical value	p-value	rejected
RI	1.34	0.01	1.63	0.0561963016778131	no
Na	0.87	0.01	1.63	0.43825271603342	no
Mg	2.94	0.01	1.63	6.18457917100912e-08	yes
Al	0.84	0.01	1.63	0.474757887353829	no
Si	0.96	0.01	1.63	0.314710019077325	no
K	2.14	0.01	1.63	0.000212776619708754	yes
Ca	1.33	0.01	1.63	0.057710602872685	no
Ba	2.6	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	< 1.0e-15	yes

Table 3.14: Test results of the improved KS test on the whole data sample

variable	test statistic	sig. level	critical value	p-value	rejected
RI	1.31	0.01	1.63	0.0630043926883292	no
Na	0.66	0.01	1.63	0.77871853343362	no
Mg	0.49	0.01	1.63	0.967729719418776	no
Al	0.92	0.01	1.63	0.366854549713195	no
Si	1.06	0.01	1.63	0.208027646546284	no
K	1.73	0.01	1.63	0.00491847745617136	yes
Ca	0.84	0.01	1.63	0.48064266616439	no
Fe	2.65	0.01	1.63	1.63244296669252e-06	yes

Table 3.15: Test results of the improved KS test on type 1 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	1.08	0.01	1.63	0.190264315474967	no
Na	0.49	0.01	1.63	0.969807662718898	no
Mg	1.37	0.01	1.63	0.0477938096655182	no
Al	0.64	0.01	1.63	0.807950624660872	no
Si	0.55	0.01	1.63	0.918823986509849	no
K	1.18	0.01	1.63	0.125440771961193	no
Ca	1.43	0.01	1.63	0.0327944325987	no
Fe	2.5	0.01	1.63	7.20083142169425e-06	yes

Table 3.16: Test results of the improved KS test on type 2 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.67	0.01	1.63	0.767837508224946	no
Na	0.52	0.01	1.63	0.951658259816235	no
Al	0.49	0.01	1.63	0.968495966845634	no
Si	0.56	0.01	1.63	0.915628110530136	no
K	1.43	0.01	1.63	0.0340401962712393	no
Ca	0.56	0.01	1.63	0.91558957374917	no
Ba	0.76	0.01	1.63	0.61288183743927	no

Table 3.17: Test results of the improved KS test on type 7 glass

3.2.1 Q-Q-plot

Before testing statistically, the results of the transformation can be examined graphically by again plotting the quantiles of the transformed data and comparing these plots with the QQ-plots from the original data in section 3.1.1. For some of the elements in the dataset, the transformation leads to a slight approximation towards normality (figure 3.6). However, for rather complicated cases (e. g. where several distributions seem to be combined) the transformation appearently does not produce normally distributed data (figure 3.7). In these cases the initial distribution is not unimodal and therefore the Box-Cox-transformation is not very helpful (see section 2.3 for a detailed explanation).

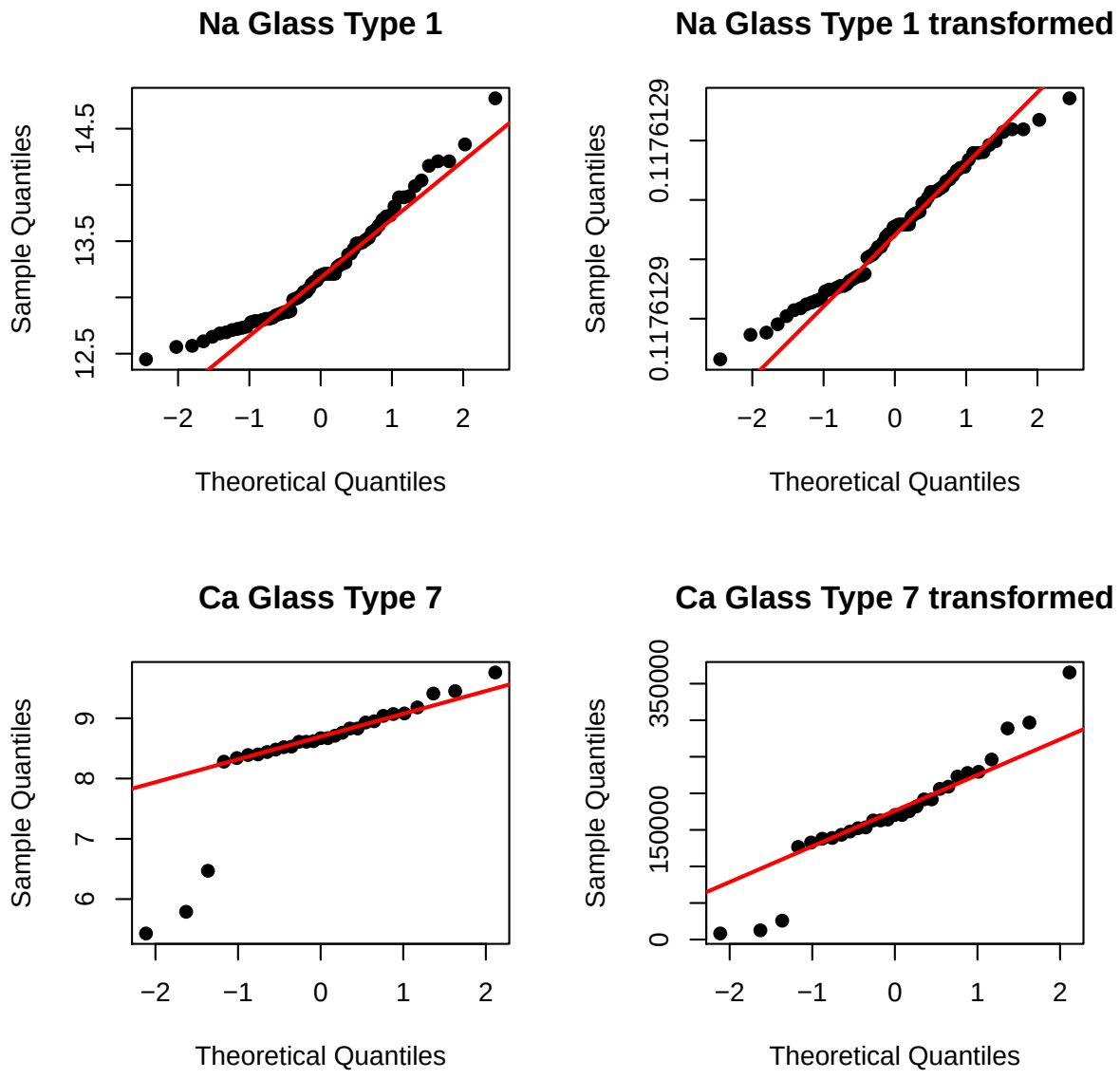


Figure 3.6: QQ-Plots of the cases where the transformation shows slight approximations to normality

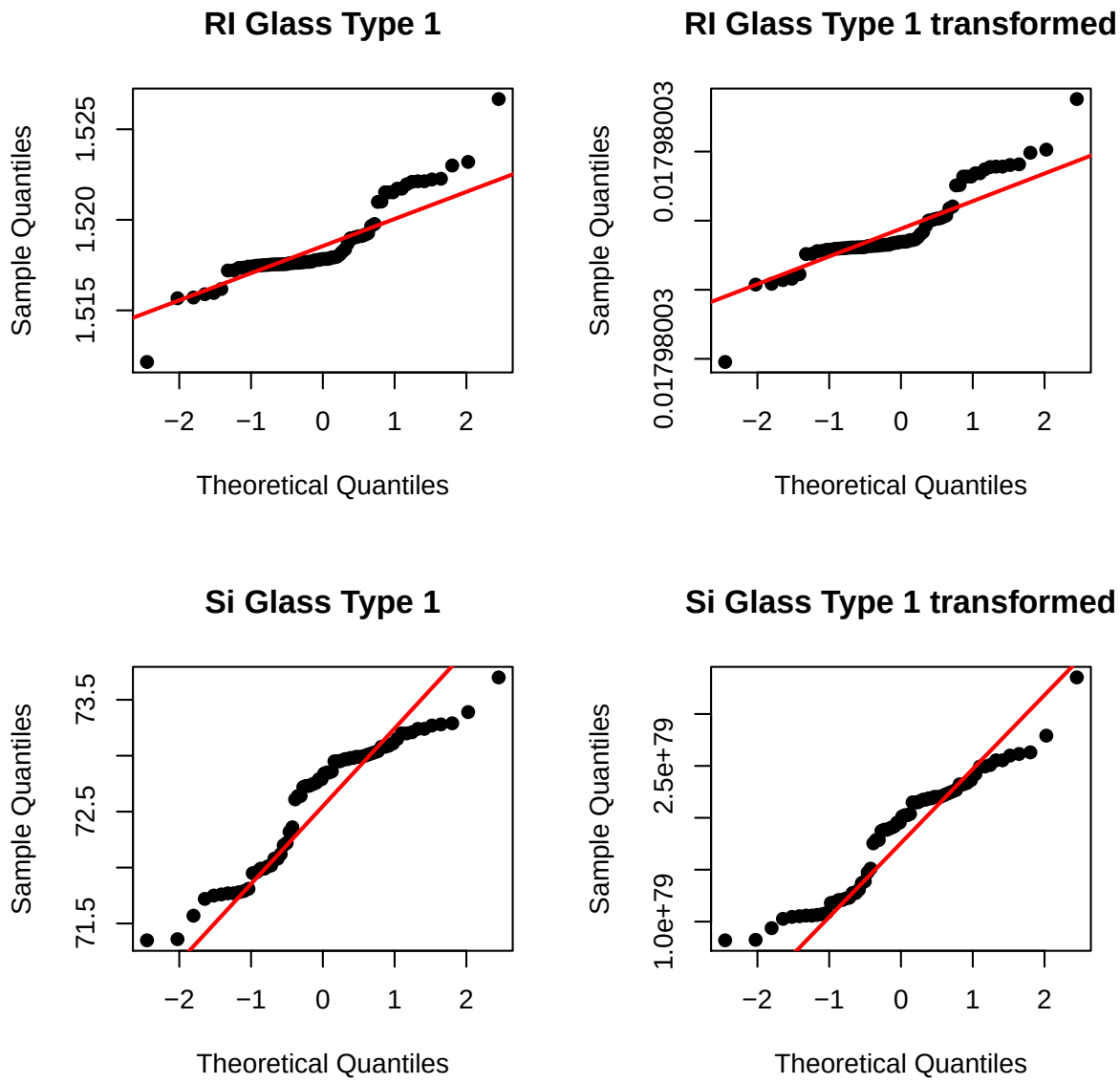


Figure 3.7: QQ-Plots of the non-unimodal cases where the transformation does not lead to normal distributions

3.2.2 Shapiro-Wilk test

When performing the Shapiro-Wilk test on the whole transformed data, at the first glance the transformation does not seem to have had the desired impact on the data. As displayed in table A.5 for all elements the null hypothesis is rejected. Although the p-values increased, they are still below the 1 % significance level, so there is no fundamental difference to the results from testing the original data.

However, proceeding analogously to section 3.1.2 and testing the different glass types separately, these separate results reveal some improvements.

Glass type 1 shows the largest number of improvements, since in the original dataset the null hypothesis was rejected for all variables. The test results of the transformed type 1 data from table A.6 show heavily increased p-values. For the three variables Sodium

(Na), Aluminium (Al), and Calcium (Ca) now a normal distribution cannot be rejected any more.

For the glass types 2 and 3 there are no changes in the number of rejections at the 1 % significance level (tables A.7 and A.8). In fact, the p-values of the variables increased considerably with the transformation, however, they still lie below the significance level.

The test on the transformed glass type 5 data shows that, in addition to refractive index (RI), Sodium (Na), Silicon (Si), and Calcium (Ca), now for Aluminium (Al) and Potassium (K) a normal distribution cannot be rejected any more (see table A.10).

Furthermore, for glass type 6 the p-values increase substantially as well, except for Silicon (Si) which only slightly increases (see table A.10). Consequently, H_0 is, analogously to the original data, not rejected for Aluminium (Al), Silicon (Si), and Calcium (Ca), plus the refractive index (RI) for which H_0 cannot be rejected after the transformation.

The test results of the transformed glass type 7 data show that for Aluminium (Al) the Null hypothesis still cannot be rejected. Due to contained zero-values Barium (Ba) is excluded from the transformation (see section 3.2) and therefore nothing is concluded about this specific variable. Furthermore, the result shows that the p-values of Sodium (Na) and Calcium (Ca) considerably increase and exceed the 1 % significance level after the transformation.

In a nutshell, the Shapiro-Wilk test shows shows considerable improvements towards a normal distribution of the variables when the Box-Cox-transformation is used and the different glass types are analysed separately. Analysing the full dataset also shows some minor improvements of the p-values, however, H_0 is still rejected at the chosen significance level. An explanation for this result could be that in the full dataset different distributions of the different distinctive glass types are combined and therefore the Box-Cox-transformation does not improve the distribution towards normality (see section 2.3).

3.2.3 Pearson's chi-squared test

Although for all variables for which a transformation is possible the p-value is higher than for the non-transformed data, the hypothesis of normality is still rejected for the whole sample (table A.12). The data of all types of glass is presumably too heterogenuous so that it comprises samples from several distributions within the overall sample of particular variables.

Concerning the transformation of type 1 glass, for two more variables (Al and Ca) the hypothesis of normality now cannot be rejected (table A.13). In both cases, the p-value increases substantially through the Box-Cox-transformation. For the variable Calcium, the frequencies of the original data are slightly shifted to lower values (figure 3.8) whereas the transformation fits the data approximately to an according normal distribution (figure 3.9).

Similar effects can be observed for the results of type 2 glass (table A.14). Here, the hypothesis of normality is additionally not rejected for the variable Sodium.

For the observations of type 7 glass, the test can now even be conducted on data of the variable Sodium, which were not properly distributed to perform the test before, i. e. the data were not divided into a sufficient number of classes (table A.15).

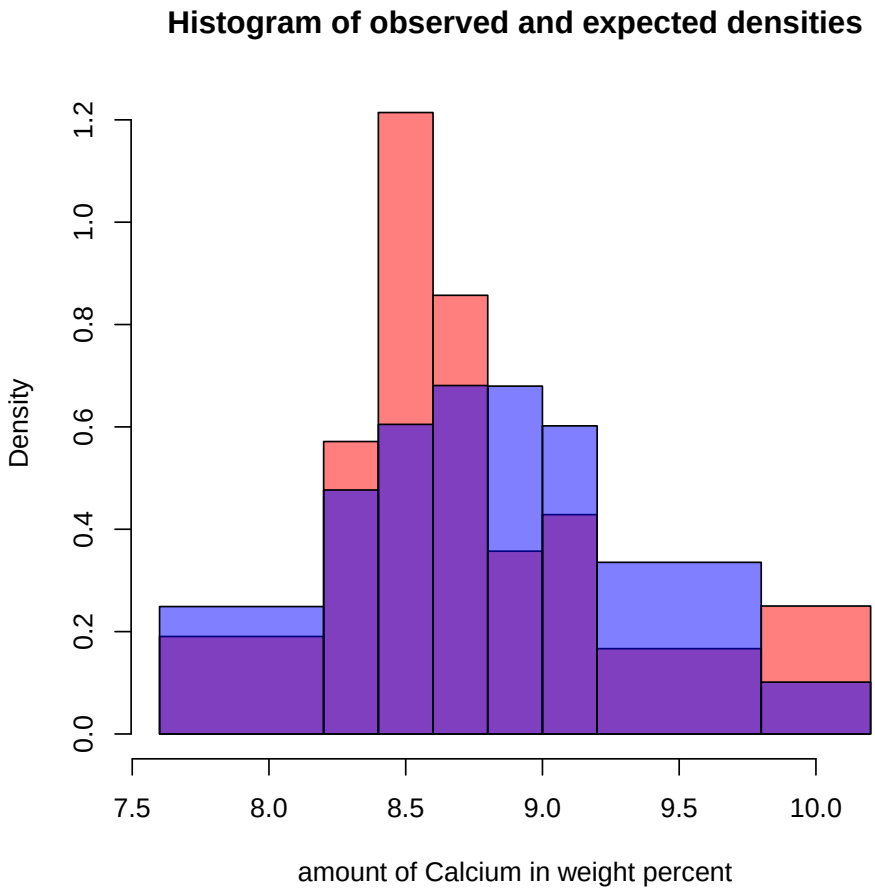


Figure 3.8: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Calcium of type 1 glass

3.2.4 Kolmogorov-Smirnov test

Since in the KS test of the original dataset the null hypotheses was only rejected for the variables Magnesium (Mg), Potassium (K), Barium (Ba), and Iron (Fe), transforming the data would aim at achieving a normal distribution for these four variables. However, exactly these four the transformation is not possible (see section 3.2). Therefore KS testing the transformed data will not provide further insights.

3.3 Contour plots of selected variables

The search for bivariate normally distributed variables shall be now demonstrated by an example from the Glass sample. Concerning the complete sample, the p-values for the variables Na and Al are highest among all used test methods. Therefore, these two variables are selected for the contour plot (3.10).

From the shape of the plot, no clear insight can be deducted about multivariate normality. In the center, the contour lines have an elliptical shape whereas an accumulation of data points can be observed at the upper right corner. This accumulation is presumably due to different distributions of variables among particular glass types. Indeed, it can be ascribed to the data of type 7 glass, which can be seen in figure 3.11.

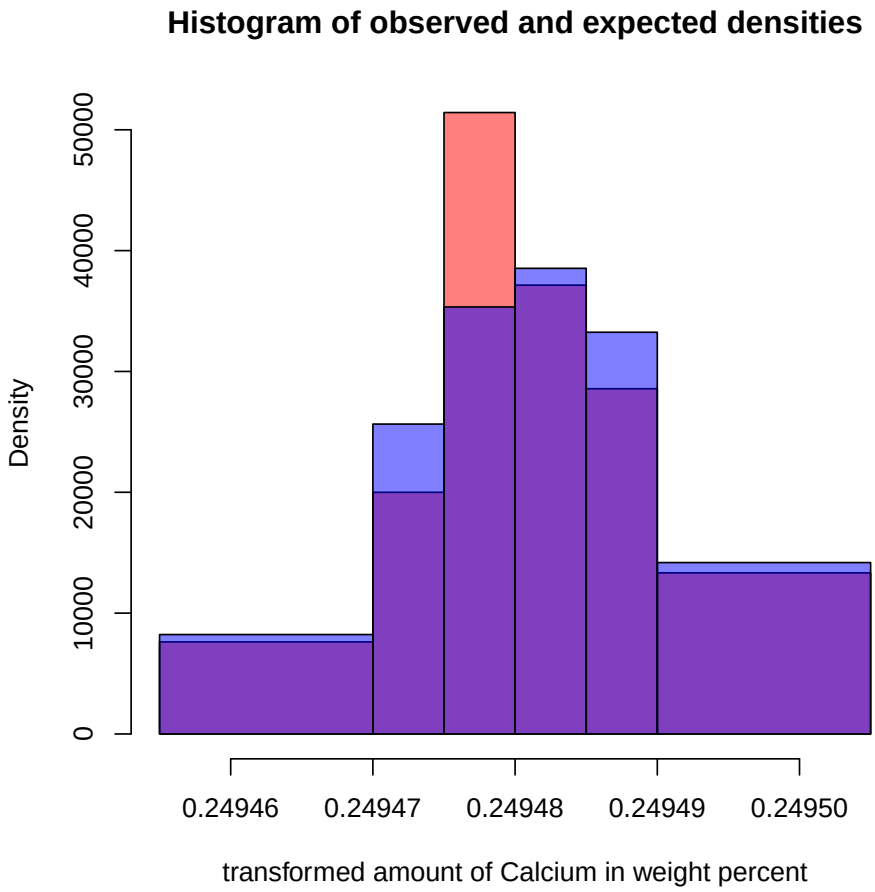


Figure 3.9: Histogram of observed densities (red) and expected densities (blue) within the classes for transformed values of the variable Calcium of type 1 glass

If the same plot is drawn for the whole sample without data of type 7 glass (figure 3.12), an elliptical shape is recognisable, except for some outliers. Quantitative tests should be conducted next in order to make a more accurate point about the compliance of the data with a multivariate normal distribution.

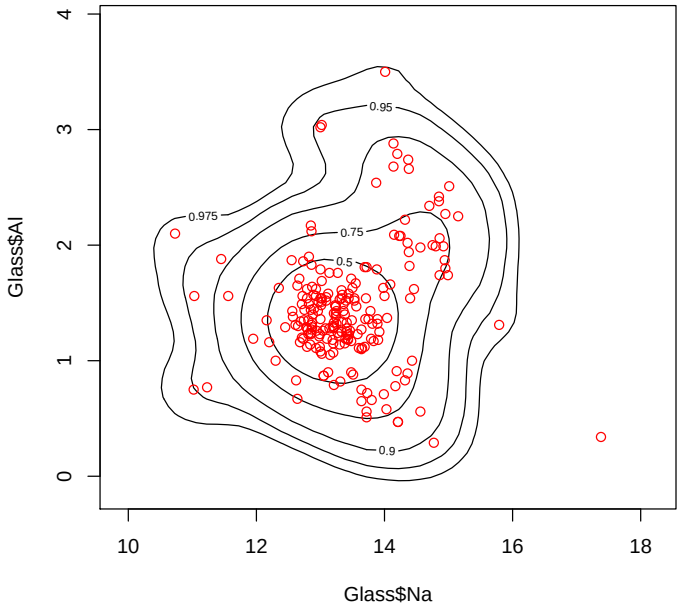


Figure 3.10: Data plot with contour lines of the variables Na and Al of the whole sample

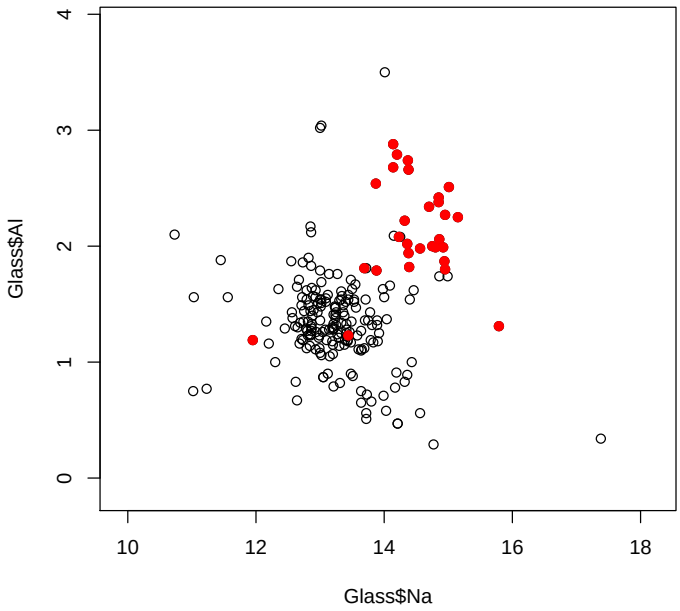


Figure 3.11: Data plot of the variables Na and Al of the whole sample. Data points of type 7 glass are highlighted in red.

4 Conclusion

For many statistical methods, the assumption of a normal distribution is a common requirement. Data samples can be tested on the normal distribution by a variety of tests. A

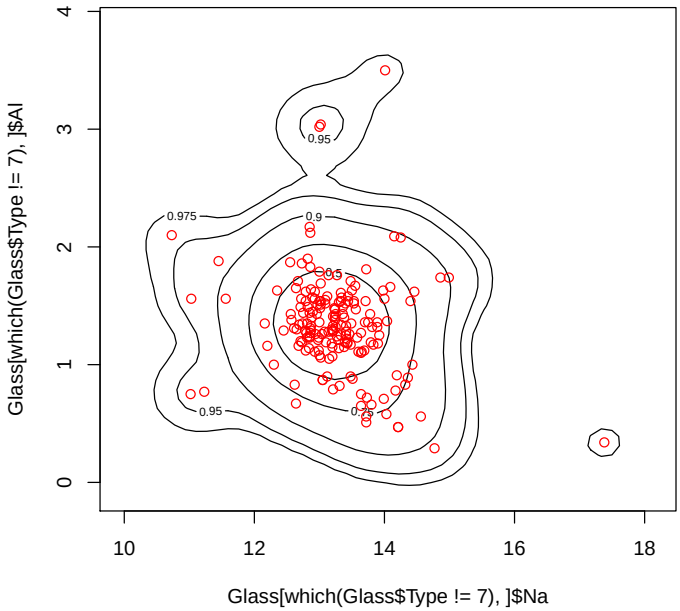


Figure 3.12: Data plot with contour lines of the variables Na and Al of the whole sample except data of type 7 glass

graphical method to check if a sample is likely to be drawn from a normally distributed population is the Q-Q-plot. In such a diagram, empirical quantiles are plotted against theoretical quantiles of the assumed distribution (the normal distribution in this case). A linear relation between the quantiles suggests an equivalent distribution. A quantitative method for testing on normality is the Shapiro-Wilk test. It refers to Q-Q-plots by calculating the squared correlation between empirical and theoretical quantiles as test statistic. It is specifically suited for the normal distribution and does not need an estimation of parameters. Furthermore, it is also suitable for rather small sample sizes. The Chi-squared test is based on the approximation of the test statistic to the Chi-squared distribution. Therefore, this test is not very suited for small sample sizes. The test statistic is calculated by dividing the observations in pairwise disjoint classes and comparing the observed class frequencies with the expected class frequencies for a specific distribution (e. g. the normal diestribution). The results can be inconsistent if different classes are used (which has been shown for testing the data with ten initial classes on the one hand and 30 initial classes on the other hand). Another method for testing on certain distributions is the Kolmogorov-Smirnov test. By applying this test, the empirical distribution function is compared with the theoretical distribution function and the supremum of the distances between these functions constitutes the test statistic. Like the Shapiro-Wilk test, the Kolmogorov-Smirnov test is suited for rather small sample sizes as well. The differences of the presented test methods are reflected in the different test results for the `Glass` data set. Nevertheless, for most variables the tests tend to return similar results when comparing them to test results of other variables.

Data that appear not to be normally distributed can possibly be transformed by a power transformation such as the Box-Cox transformation. A successful application of

this method depends on the structure of the original data. For non-unimodal data e. g. , there is not much of an improvement to be expected. Testing on the whole data set as well es on the separate glass types has forced the assumption that the different types of glass are drawn from specifically distributed populations such that different distributions are mixed in the overall sample.

A first step to check samples on multivariate normally distributed population origins is to investigate contour lines of data point plots of two variables respectively. Elliptical shapes should be recognised for a bivariate normal distribution. However, no clear clues for a multivariate normal distribution could be found in this report.

A Appendix

variable	test statistic	sig. level	critical value	p-value	rejected
RI	86.4	0.01	20.09	2.44249065417534e-15	yes
Na	49.04	0.01	23.21	3.99921126548186e-07	yes
Mg	668.81	0.01	16.81	< 1.0e-15	yes
Al	61.72	0.01	29.14	5.84218540211623e-08	yes
Si	92.86	0.01	23.21	1.4432899320127e-15	yes
K	178.04	0.01	11.34	< 1.0e-15	yes
Ca	149.63	0.01	18.48	< 1.0e-15	yes
Ba	301.07	0.01	11.34	< 1.0e-15	yes
Fe	137.37	0.01	18.48	< 1.0e-15	yes

Table A.1: Test results of the chi-squared test on the whole data sample with 30 initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	102	0.01	13.28	< 1.0e-15	yes
Na	16.61	0.01	18.48	0.0200763340092525	no
Mg	28.9	0.01	16.81	6.36733763034192e-05	yes
Al	27.56	0.01	16.81	0.000113486165164822	yes
Si	35.74	0.01	15.09	1.07201048937799e-06	yes
K	129.69	0.01	18.48	< 1.0e-15	yes
Ca	20.22	0.01	18.48	0.00510775321781454	yes
Fe	45.28	0.01	9.21	1.46922363164492e-10	yes

Table A.2: Test results of the chi-squared test on type 1 glass with 30 initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	77.84	0.01	15.09	2.33146835171283e-15	yes
Na	17.26	0.01	18.48	0.0157897896300554	no
Mg	306.84	0.01	15.09	< 1.0e-15	yes
Al	20.77	0.01	20.09	0.00778471801730796	yes
Si	13.38	0.01	16.81	0.0374408380909446	no
K	54.24	0.01	13.28	4.67884619936854e-11	yes
Ca	106.11	0.01	11.34	< 1.0e-15	yes
Fe	49.63	0.01	11.34	9.60126422810959e-11	yes

Table A.3: Test results of the chi-squared test on type 2 glass with 30 initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	22.42	0.01	6.63	2.18961874765e-06	yes
Na	12.16	0.01	6.63	0.000489125271576851	yes
Al	3.37	0.01	9.21	0.185460070738202	no
Si	21.68	0.01	6.63	3.2242744996136e-06	yes
K	15.8	0.01	NA	NA	NA
Ca	11.55	0.01	NA	NA	NA
Ba	19.75	0.01	9.21	5.14515780526414e-05	yes

Table A.4: Test results of the chi-squared test on type 7 glass with 30 initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	NA	NA	NA	NA
Na	0.95	0.01	NA	8.75605777309153e-07	yes
Mg	NA	NA	NA	NA	NA
Al	0.97	0.01	NA	0.000244326513056066	yes
Si	0.93	0.01	NA	1.58998125691823e-08	yes
K	NA	NA	NA	NA	NA
Ca	0.89	0.01	NA	1.13880689831982e-11	yes
Ba	NA	NA	NA	NA	NA
Fe	NA	NA	NA	NA	NA

Table A.5: Test results of the Shapiro-Wilk test on the whole transformed data sample

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.89	0.01	NA	1.62433657125306e-05	yes
Na	0.98	0.01	NA	0.353792914291578	no
Mg	0.83	0.01	NA	1.40023833110547e-07	yes
Al	0.96	0.01	NA	0.0459207068393172	no
Si	0.94	0.01	NA	0.00269629206710463	yes
K	NA	NA	NA	NA	NA
Ca	0.97	0.01	NA	0.148237775100495	no

Table A.6: Test results of the Shapiro-Wilk test on the transformed type 1 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	NA	NA	NA	NA
Na	0.94	0.01	NA	0.00209380292451352	yes
Mg	NA	NA	NA	NA	NA
Al	0.98	0.01	NA	0.180941341568135	no
Si	0.95	0.01	NA	0.00272685181069014	yes
K	NA	NA	NA	NA	NA
Ca	0.91	0.01	NA	4.61802518307857e-05	yes

Table A.7: Test results of the Shapiro-Wilk test on the transformed type 2 glass

List of Figures

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	NA	NA	NA	NA
Na	0.96	0.01	NA	0.64635351843862	no
Mg	0.95	0.01	NA	0.528768845683765	no
Al	0.96	0.01	NA	0.600570237058731	no
Si	0.93	0.01	NA	0.249637851485608	no
K	NA	NA	NA	NA	NA
Ca	0.97	0.01	NA	0.770271104449462	no

Table A.8: Test results of the Shapiro-Wilk test on the transformed type 3 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.94	0.01	NA	0.514552235425043	no
Na	0.94	0.01	NA	0.507118972628293	no
Mg	NA	NA	NA	NA	NA
Al	0.93	0.01	NA	0.340672646826333	no
Si	0.95	0.01	NA	0.593358086962012	no
K	0.95	0.01	NA	0.547078669634891	no
Ca	0.92	0.01	NA	0.2231352980081	no

Table A.9: Test results of the Shapiro-Wilk test on the transformed type 5 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	0.87	0.01	NA	0.126757540911585	no
Na	NA	NA	NA	NA	NA
Mg	NA	NA	NA	NA	NA
Al	0.92	0.01	NA	0.377444331641712	no
Si	0.8	0.01	NA	0.0180983710066516	no
Ca	0.95	0.01	NA	0.729143379606271	no

Table A.10: Test results of the Shapiro-Wilk test on the transformed type 6 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	NA	NA	NA	NA
Na	0.95	0.01	NA	0.213808937207852	no
Al	0.96	0.01	NA	0.430838611416679	no
Si	0.81	0.01	NA	0.000142333041740501	yes
Ca	0.93	0.01	NA	0.0686924536083304	no
Ba	NA	NA	NA	NA	NA

Table A.11: Test results of the Shapiro-Wilk test on the transformed type 7 glass

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	0.01	NA	NA	NA
Na	29.94	0.01	9.21	3.14878793927775e-07	yes
Mg	NA	0.01	NA	NA	NA
Al	24.44	0.01	18.48	0.000953232079632826	yes
Si	34	0.01	13.28	7.44688283815798e-07	yes
K	NA	0.01	NA	NA	NA
Ca	44.87	0.01	11.34	9.8475638754536e-10	yes
Ba	NA	0.01	NA	NA	NA
Fe	NA	0.01	NA	NA	NA

Table A.12: Test results of the chi-squared test on the whole transformed data sample with ten initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	27.81	0.01	6.63	1.33864150764218e-07	yes
Na	1.59	0.01	13.28	0.810360513797024	no
Mg	17.87	0.01	NA	NA	NA
Al	6.41	0.01	11.34	0.093110657016404	no
Si	16.87	0.01	13.28	0.00205136639513992	yes
K	NA	0.01	NA	NA	NA
Ca	3.35	0.01	11.34	0.341234021909645	no
Fe	NA	0.01	NA	NA	NA

Table A.13: Test results of the chi-squared test on the transformed data of type 1 glass with ten initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	0.01	NA	NA	NA
Na	4.56	0.01	9.21	0.102375873496938	no
Mg	NA	0.01	NA	NA	NA
Al	2.61	0.01	13.28	0.62553471134666	no
Si	2.37	0.01	6.63	0.123314180710608	no
K	NA	0.01	NA	NA	NA
Ca	15.61	0.01	6.63	7.78625904057639e-05	yes
Fe	NA	0.01	NA	NA	NA

Table A.14: Test results of the chi-squared test on the transformed data of type 2 glass with ten initial classes

variable	test statistic	sig. level	critical value	p-value	rejected
RI	NA	0.01	NA	NA	NA
Na	1.83	0.01	6.63	0.176007215391275	no
Al	1.53	0.01	9.21	0.465771718543797	no
Si	12.46	0.01	NA	NA	NA
K	NA	0.01	NA	NA	NA
Ca	2.39	0.01	NA	NA	NA
Ba	NA	0.01	NA	NA	NA

Table A.15: Test results of the chi-squared test on the transformed data of type 7 glass with ten initial classes

List of Tables

References

[1] Bache, K., Lichman, M. (2013), UCI Machine Learning Repository, URL: `http://archive.ics.uci.edu/ml/datasets/Glass+Identification`, Irvine, CA, University of California, School of Information and Computer Science.

[2] DeGroot, Morris, H., Mark, J. (2003) Probability and Statistics. 3rd ed. Boston, MA: Addison-Wesley.

[3] Steel, R.G.D., Torrie, J.H. (1960), Principles and Procedures of Statistics with Special Reference to the Biological Sciences, McGraw Hill, p. 350.