Case Studies

"Data Analytics"

Topic

Summer Term 2013

FirstName LastName

July 7, 2013

Contents

1 Introduction **1**
- 1.1 Normality as a requirement for statistical methods . . . 1
- 1.2 The glass data sample . . . 1
- 1.3 Aim and structure . . . 1

2 Preliminaries **2**
- 2.1 Test methods for normality . . . 2
  - 2.1.1 Q-Q-plot . . . 2
  - 2.1.2 Shapiro-Wilk test . . . 2
  - 2.1.3 Pearson's chi-squared test . . . 2
  - 2.1.4 Kolmogorov-Smirnov test . . . 6
- 2.2 Box-Cox-transformation . . . 12
- 2.3 Plot of a multivariate normal distribution . . . 15

3 Testing the data sample for normality **16**
- 3.1 Testing original data . . . 16
  - 3.1.1 Q-Q-plot . . . 16
  - 3.1.2 Shapiro-Wilk test . . . 16
  - 3.1.3 Pearson's chi-squared test . . . 16
  - 3.1.4 Kolmogorov-Smirnov test . . . 19
- 3.2 Testing transformed data . . . 20
  - 3.2.1 Q-Q-plot . . . 20
  - 3.2.2 Shapiro-Wilk test . . . 20
  - 3.2.3 Pearson's chi-squared test . . . 20
  - 3.2.4 Kolmogorov-Smirnov test . . . 20

4 Conclusion **21**

5 Section 1 **22**
- 5.1 First Subsection . . . 23
- 5.2 Second Subsection . . . 23

6 Section 2 **24**

A Appendix **i**

List of Figures **ii**

List of Tables **iii**

1 Introduction

1.1 Normality as a requirement for statistical methods

test test test

1.2 The glass data sample

1.3 Aim and structure

2 Preliminaries

2.1 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 is usually the hypothesis 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 and 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.

2.1.1 Q-Q-plot

2.1.2 Shapiro-Wilk test

2.1.3 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 1), 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 more area is not overlapping, 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 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 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

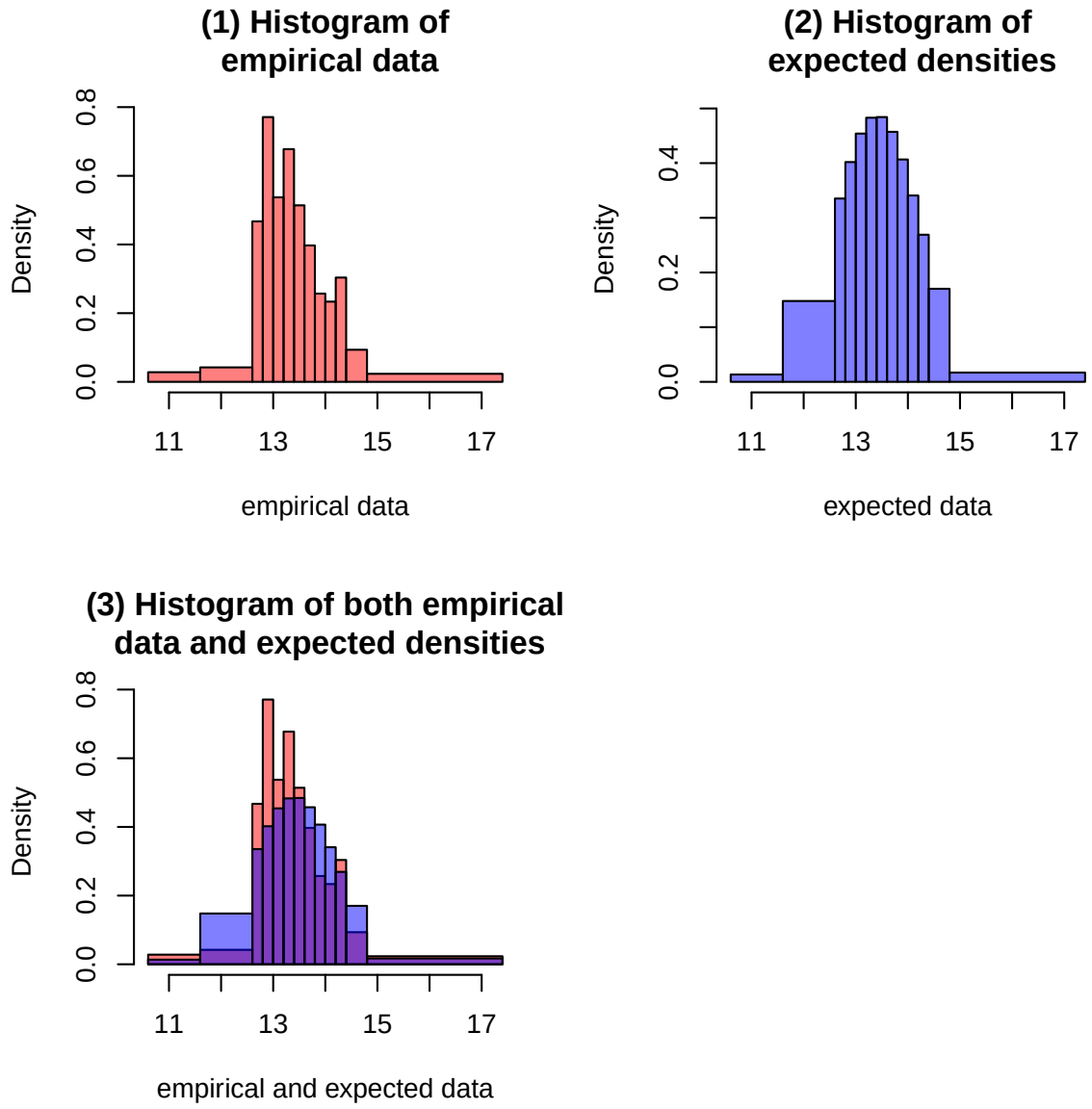


Figure 1: Exemplary histrograms of a data sample, expected densities for a normal distribution with parameters estimated from the sample and a combined histogram of these both histograms.

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 $Y_k = |\{i : X_i \in C_k\}|$.

As a common requirement for a sufficient approximation, the minimum number of observations in each class should not fall below five. Hence, marginal or even inner classes have to be unified in some cases in order to achieve a sufficient class size. The following R-function is used here for this purpose.

```
# Calculates bounds of bins (classes) of a data sample.
# The initial bounds are given by initial_breaks,
# k denotes the minimum class size.
makebins = function(data, initial_breaks, k) {
  h = hist(data, breaks=initial_breaks, plot=FALSE)

  br = h$breaks
  changed = TRUE

  while(changed) {
    h = hist(data, breaks=br, plot=FALSE)
    br = h$breaks
    changed=FALSE

    for(i in 1:length(h$counts)) {
      if(h$counts[i] < k) {
        if(i > 1 && i < length(h$counts)) {
          if(h$counts[i-1] < h$counts[i+1]) {
            br = br[-i]
            changed = TRUE
            break
          }
          else {
            br = br[-(i+1)]
            changed = TRUE
            break
          }
        }
        # index on first class
        else if(i == 1) {
          br = br[-2]
          changed = TRUE
          break
        }
        # index on last class
        else {
          br = br[-(length(h$counts))]
          changed = TRUE
          break
        }
      }
    }
  }
  return(br)
}
```

Further functions are needed for calculating the expected frequencies, the test statistic

and the result of the test (since the mean and the standard deviation are estimated from the sample, two degrees of freedom are additionally lost):

```
# Calculates the expected probabilities of a normal distribution
# with the given parameters mean and sd
# for the given bin (class) bounds
probabilities.exp = function(bins, mean, sd) {
  result = rep(0, length(bins)-1)

  result[1] = pnorm(q=bins[2], mean=mean, sd=sd)

  for(i in 2:(length(bins)-1)) {
    result[i] <- pnorm(q=bins[i+1], mean=mean, sd=sd)
      - pnorm(q=bins[i], mean=mean, sd=sd)
  }

  result[length(bins)-1] = pnorm(q=bins[length(bins)-1],
    mean=mean, sd=sd, lower.tail=FALSE)

  return(result)
}
# Returns the chi squared test statistics
# for the given actual and expected values.
teststat.chi = function(actual, expected) {
  sum((actual - expected)^2 / expected)
}
# Performs a chi squared goodness of fit test on the given data
# for the assumption of a normal distribution.
# Returns true if the null hypothesis (sample drawn from a
# normal distributed population) is rejected, false otherwise.
# The parameters are estimated from the sample.
# The initial bounds for the classes are given by initial_breaks,
# min denotes the minimum class size.
# The significance level is determined by sig.
chisq.test.norm = function(data, initial_breaks, min, sig) {
  bins = makebins(data, initial_breaks, min)
  hist = hist(data, breaks=bins, plot=FALSE)
  expected_probabilities = probabilities.exp(bins, mean(data), sd(data))
  expected_frequencies = expected_probabilities * length(data)
  teststat = teststat.chi(hist$counts, expected_frequencies)

  # length(bin) - 1 classes, 2 estimated parameters (mean, sd)
  df=length(bins)-4
  critical_value = ifelse(df < 1, NA, qchisq(p=1-sig, df=df))

  print(teststat > critical_value)
```

```
  return(list(hist = hist,
              expected_probabilities = expected_probabilities,
              expected_frequencies = expected_frequencies,
              teststat = teststat,
              critical_value = critical_value,
              p_value =
                ifelse(df < 1, NA, 1 - pchisq(q=teststat, df=df)),
              rejected = teststat > critical_value))
}
```

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.

2.1.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS later on) test as all other tests is used for testing whether a given univariate sample $X_1, \dots X_2$ with unknown distribution $\mathbb{P}$ is distributed according to a completely determined distribution $\mathbb{P}_0$. It implies a decision making between the following two hypotheses:

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

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

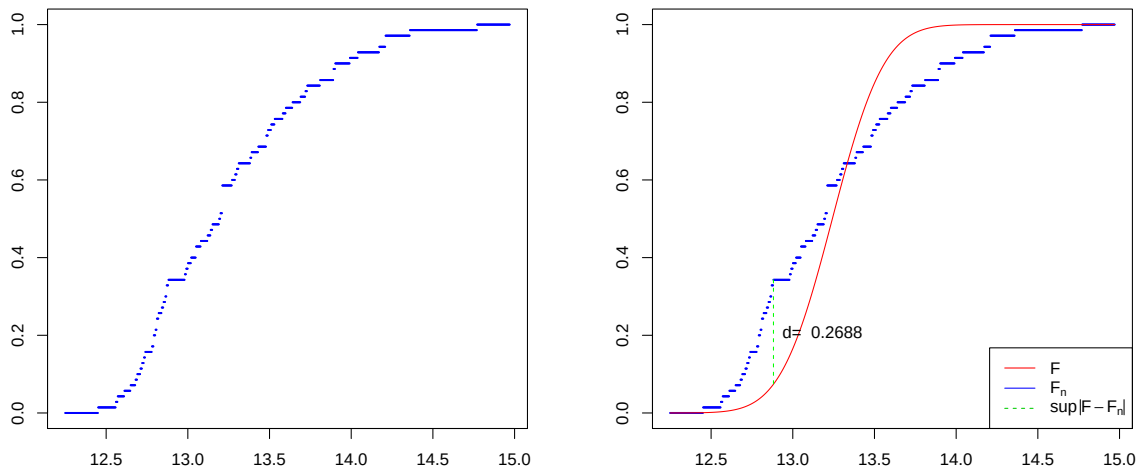
Definition 1 *For a given univariate sample $X_1, X_2, \dots, X_n$ the function*

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

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 2a.

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 2b illustrates the calculated distance between the empirical c. d. f. and the theoretical normal c. d. f. with parameters of sample mean and sample variance.



(a) **(b)**

Figure 2: Empirical c. d. f. for Natrium vector (a) and theoretical normal c. d. f. with sample mean and sample variance (b)

The KS 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 then 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 [**?**].

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 was mentioned above the last equality can be considered only 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 is has a huge drawback. Namely it considers only completely defined theoretical c. d. f. . In case of normality testing both parameters μ and σ 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. Further example of Natrium distribution illustrates this proposition.

For Natrium (Na) univariate sample of 70 observatons the sample mean $\bar{\mu}$ =13.2423 and the sample variance $\bar{\sigma} = 0.2493$ are calculated. For these values of theoretical c.d.f. is determined (Figure 2b). Then the value of KS statistics is calculated: $D_n = \sqrt{n} \cdot \sup_x |F_n(x) - F(x)|$ =2.2493. Critical value for the significance level $\alpha = 0.01$ is equal to $c = H^{-1}(1-\alpha)$=1.6276. To calculate this value the following R functions are used:

```
#Calculates KS distribution function
H= function(t) {
        i=1
        sum=0
        while(abs(f(t,i) - f(t,i+1))>0.000000000001) {
                sum=sum+f(t,i)
                i=i+1
        }
        1-2*sum
}
#Summand of H function
f=function(t,i) {
        (-1)^(i-1)*exp(-2*i^2*t^2)
}
#Returns value c such that H(c)=a
Inverse = function(H,a) {
        newFunction = function(t) {
                H(t)-a
        }
        #Returns the root of the equation H(x)-a=0
        uniroot(newFunction,c(0.2,4),tol=0.001)$root
}
```

Since $D_n > c$ the null hypothesis is rejected by KS test. But that does not mean that the sample is not normal distributed. That means only that it is not normal distributed with sample mean and sample variance as parameters of that distribution.

In order to manage that issue KS test is improved by solving the following optimization problem

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

When the new values $\hat{\mu}$ and $\hat{\sigma}$ that minimize the value $KS(\mu, \sigma)$ are found the general KS test is performed with respect to these values of parameters. For solving this optimization 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
```

```
        diff=theordat-empdat
        absdiff=abs(diff)
        max(absdiff)
}
#optim is a predifined R function in stats package
#defalut method of optimization 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 $\hat{\sigma}$ =0.4682 are taken as parameters of a new theoretical normal c. d. f. and a new distance is calculated (Figure 3).

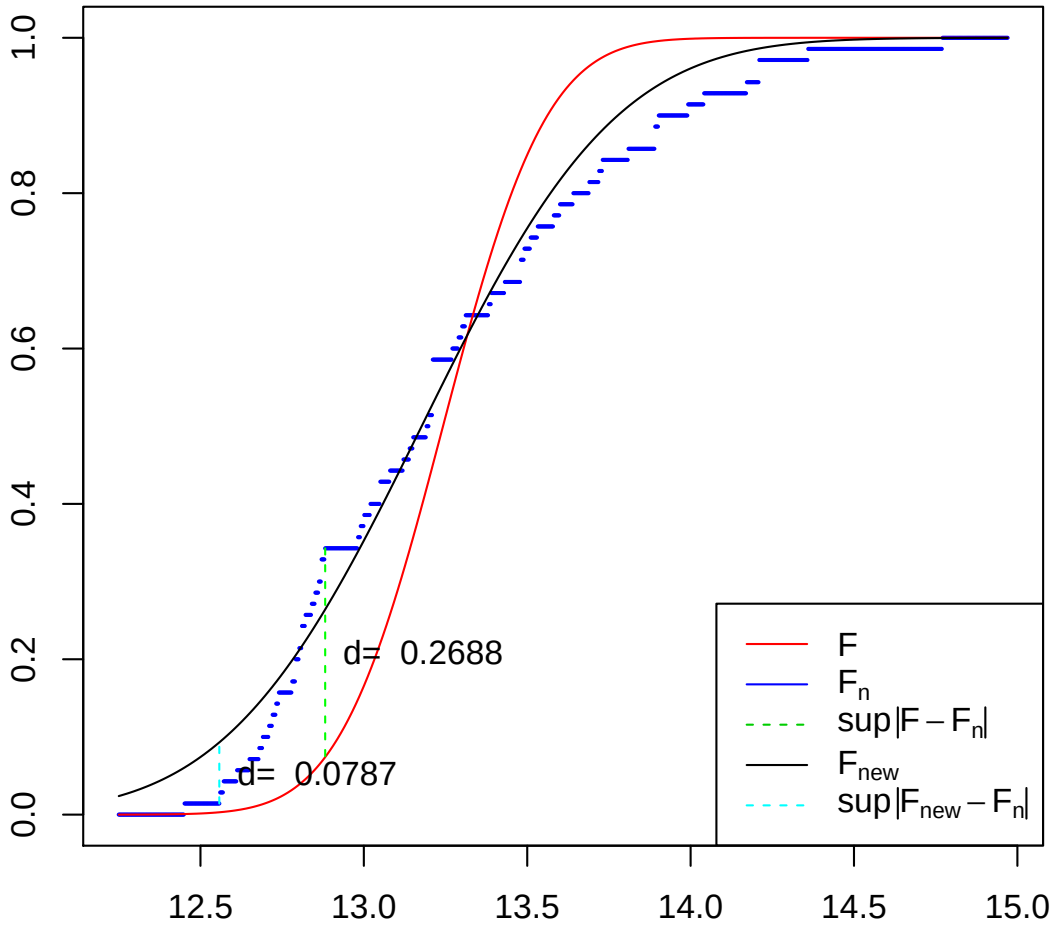


Figure 3: Normal c. d. f. with optimized parameters in comparison to the old c. d. f. with sample mean and sample variance.

The new value of KS statistics is $D_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x, \hat{\mu}, \hat{\sigma})|$ =0.6585 that is smaller than the critical value c =1.6276. Therefore the null hypothesis H_0 is accepted by KS-test. For all further samples the improved KS test is used. It was implemented in *R* and has the following code.

```
# Performs a Kolmogorov-Smirnov goodness of fit test on the given data
# for the assumption of a normal distribution.
# Returns true if the null hypothesis (sample drawn from a
# normal distributed population) is rejected, false otherwise.
# The parameters are optimized.
# The significance level is determined by sig.
KSimpr.test.norm = function(data, sig) {
```

```
        critical_value = Inverse(H,1-sig)
        #Sample mean and sample variance calculation
        mu = mean(data)
        sigma = var(data)
        #KS function that has to be minimized.
        KS= function(param) {
                #discretization of a line segment
                seq = seq(from = min(data)-0.2,
                                to = max(data)+0.2, length.out=1000)
                #values of empirical c.d.f.
                empdat = sapply(seq, function(x) {empiric(x,data)})
                #values of theoretical c.d.f.
                theordat = pnorm(seq,param[1],abs(param[2]))
                #difference between the values
                diff=theordat-empdat
                absdiff=abs(diff)
                max(absdiff)
                }
        #optim is a predifined R function in stats package
        #defalut method of optimization is Nelder and Mead (1965)
        KSoptim = optim(c(mu,sigma),KS)
        teststat = sqrt(length(data))*KSoptim$value
        p_value = 1-H(teststat)

        print(teststat > critical_value)

        return(list( mean = mu,
                                var = sigma,
                                mu_opt = KSoptim$par[1],
                                sigma_opt = KSoptim$par[2],
                                teststat = teststat,
                                critical_value = critical_value,
                                p_value = p_value,
                                rejected = teststat > critical_value))
 }
 KSimpr.test.norm(dat,0.01)
```

```
[1] FALSE
$mean
[1] 13.24229

$var
[1] 0.2493019

$mu_opt
[1] 13.17695
```

```
$sigma_opt
[1] 0.4682486

$teststat
[1] 0.6585078

$critical_value
[1] 1.627616

$p_value
[1] 0.7787185

$rejected
[1] FALSE
```

The problem of approximation with 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.

2.2 Box-Cox-transformation

If data is 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

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

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

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 4. 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 5. 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").

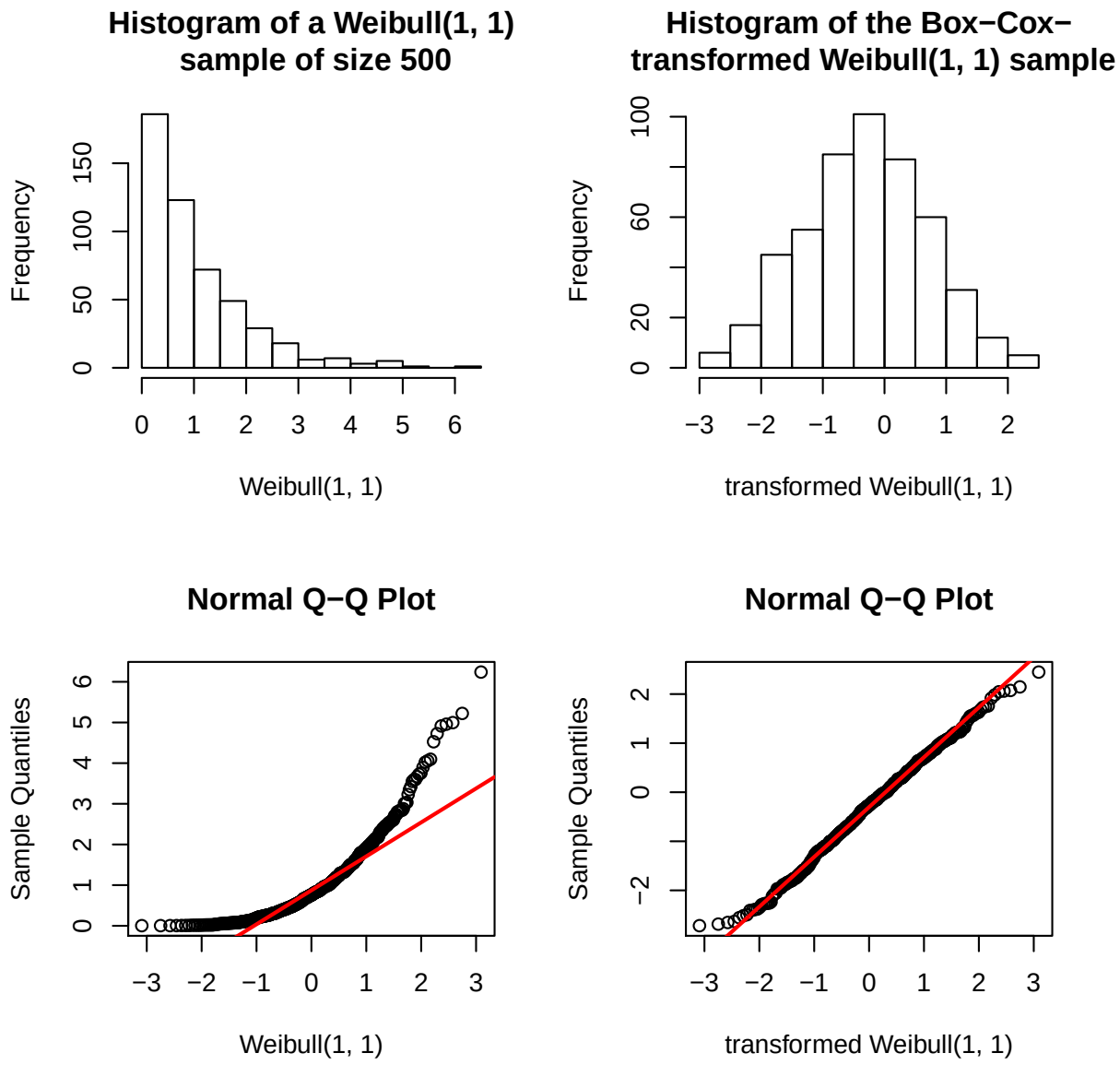


Figure 4: Histograms and QQ-plots of a Weibull(1, 1) simulated sample of size 500 and of the Box-Cox-transformed data

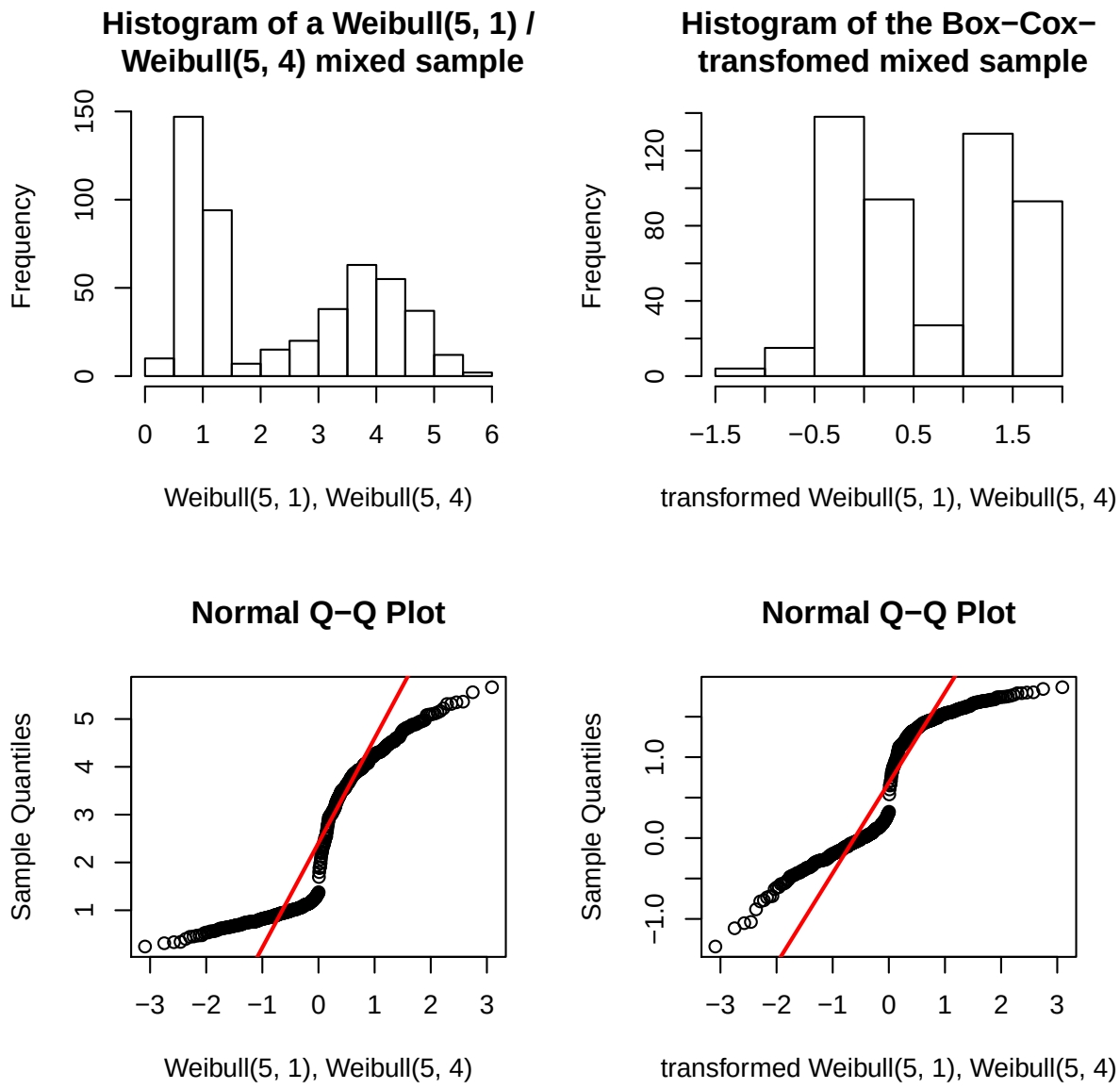


Figure 5: Histograms and QQ-plots of a mixed sample composed of a Weibull(5, 1) simulated sample and a Weibull(5, 4) simulated sample (each of size 250) and of the Box-Cox-transformed data

2.3 Plot of a multivariate normal distribution

Contour lines of the plot of a multivariate normal distribution

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

with mean vector μ and covariance matrix $\Sigma > 0$ 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.

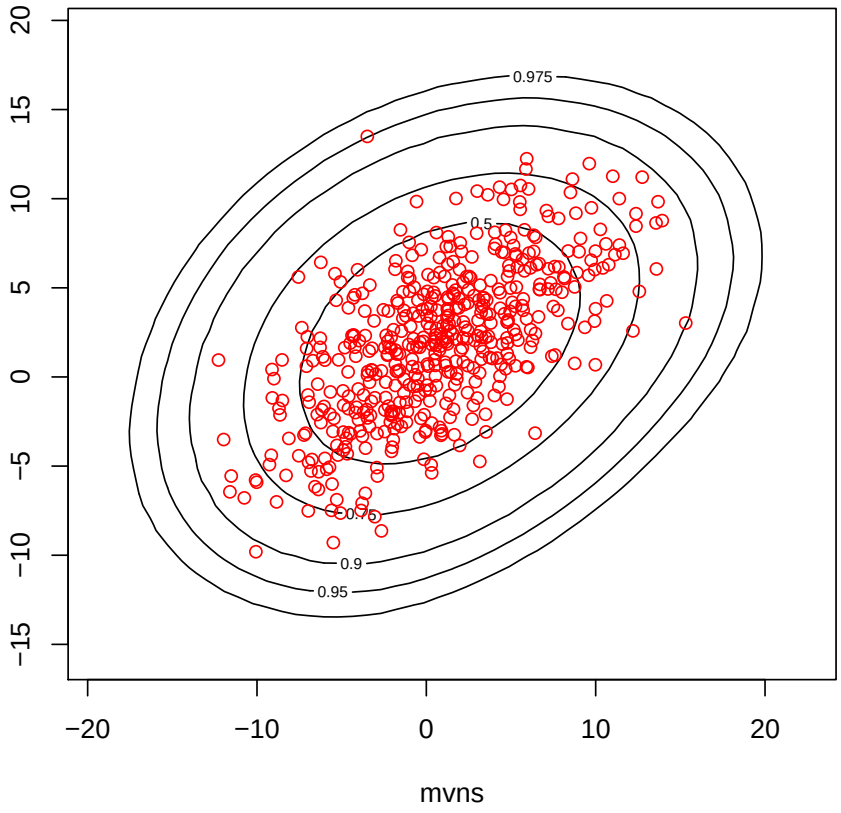


Figure 6: cap

3 Testing the data sample for normality

3.1 Testing original data

The test methods for normality that were introduced in section 2.1 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 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 1: Skipped variables for the particular glass types due to too many zero values

3.1.1 Q-Q-plot

3.1.2 Shapiro-Wilk test

3.1.3 Pearson's chi-squared test

As mentioned in section 2.1.3, 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 2. For two variables, it is not possible to determine a test result with the given parameters: The observations of the variables Potassium (K) and 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) are slightly different; in this case, the results 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 Natrium

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

(Na). The p-value for Natrium is comparably high amounting to approximately 0.52. It is well recognisable that the observed class frequencies for Natrium fluctuate around the expected class frequencies under the hypothesis of a normal distribution with the according parameters (table 4). The good compliance of empirical and hypothetical data for this variable is illustrated in figure 7. In general, the p-values for this part of the sample are 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: 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 4: Observed end expected frequencies of items in the classes for the variable Natrium of type 1 glass

The test results for observations of type 2 glass are summarised in table 5. 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).

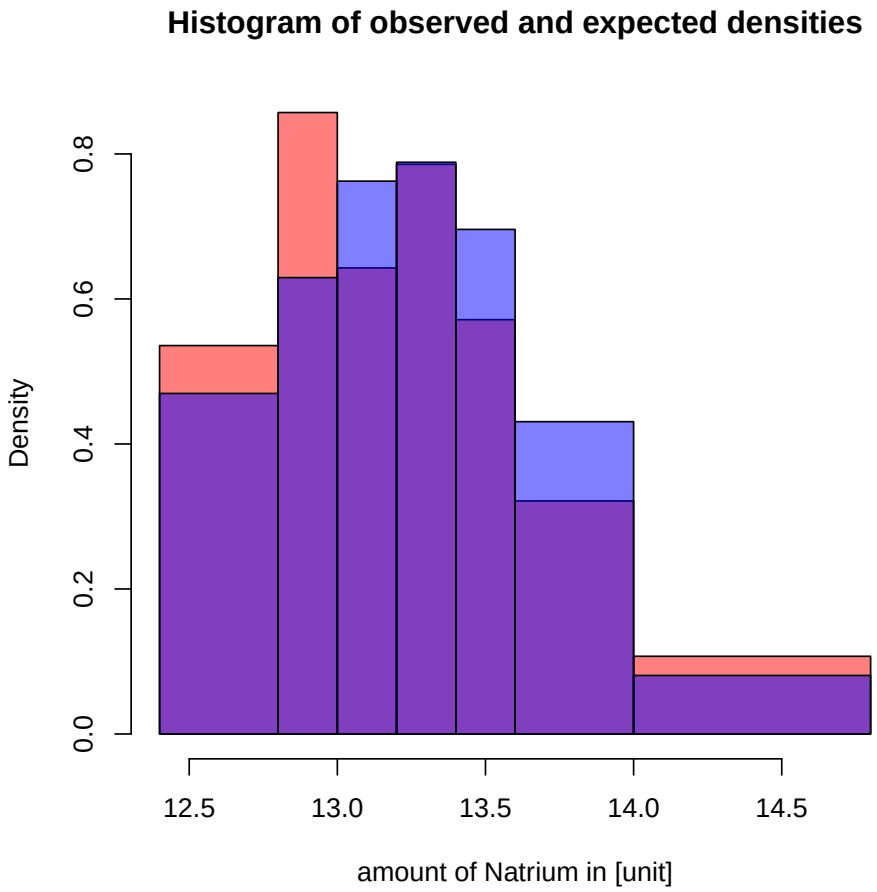


Figure 7: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass

For the observations of type 7 glass, test results (table 6) 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.

As mentioned in section 2.1.3, 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 13 to 16 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 Natrium (Na). In general, the p-values can alternate much with different classes; so the rather high p-value for Natrium 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 statis-

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 5: 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 6: Test results of the chi-squared test on type 7 glass with ten initial classes

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

3.1.4 Kolmogorov-Smirnov test

The first tests are conducted on the whole data set for each variable. The results are shown in table 7.

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 7: Test results of the improved KS test on the whole data sample with ten initial classes

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.

3.2.1 Q-Q-plot

3.2.2 Shapiro-Wilk test

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 8). 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.

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

Concerning the transformation of type 1 glass, two more variables (Al and Ca) are now tested positively on the hypothesis of normality (table 9). In both cases, the p-value is be increased substantially by the Box-Cox-transformation. For the variable Calcium, the frequencies of the original data are slightly shifted to lower values (figure 8) whereas the transformation fits the data approximately to an according normal distribution (figure 9).

Similar effects can be observed for the results of type 2 glass (table 10). Here, the hypothesis of normality is additionally not rejected for the variable Natrium.

For the observations of type 7 glass, the test can now even be conducted on data of the variable Natrium, which were not properly distributed to perform the test before, i. e. the data were not divided into a sufficient number of classes (table 11).

3.2.4 Kolmogorov-Smirnov test

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 9: Test results of the chi-squared test on the transformed data of type 1 glass with ten initial classes

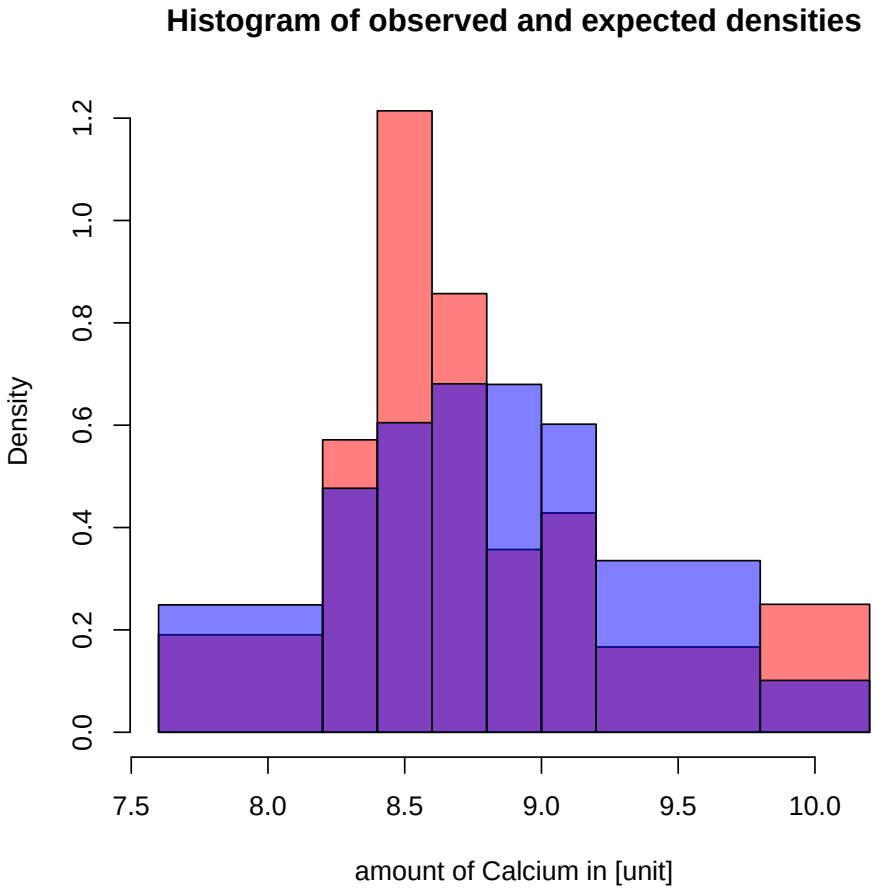


Figure 8: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Calcium of type 1 glass

4 Conclusion

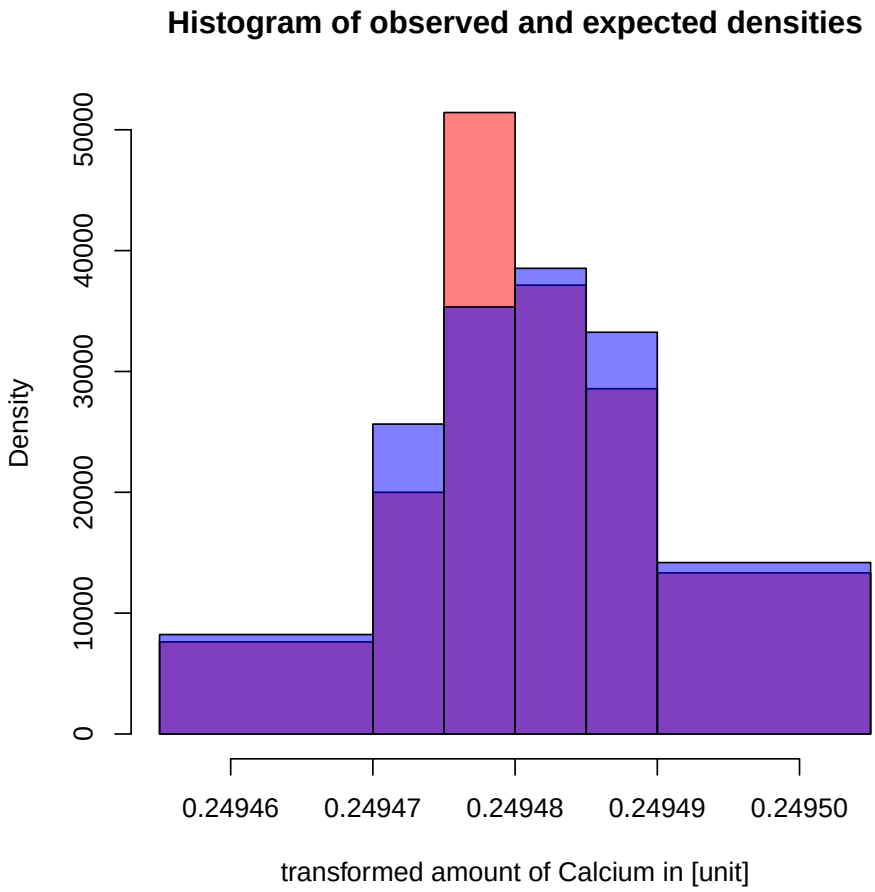


Figure 9: Histogram of observed densities (red) and expected densities (blue) within the classes for transformed values of the variable Calcium of type 1 glass

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 10: Test results of the chi-squared test on the transformed data of type 2 glass with ten initial classes

5 Section 1

Example for referring to a chapter: As written in section **??** ...

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 11: Test results of the chi-squared test on the transformed data of type 7 glass with ten initial classes

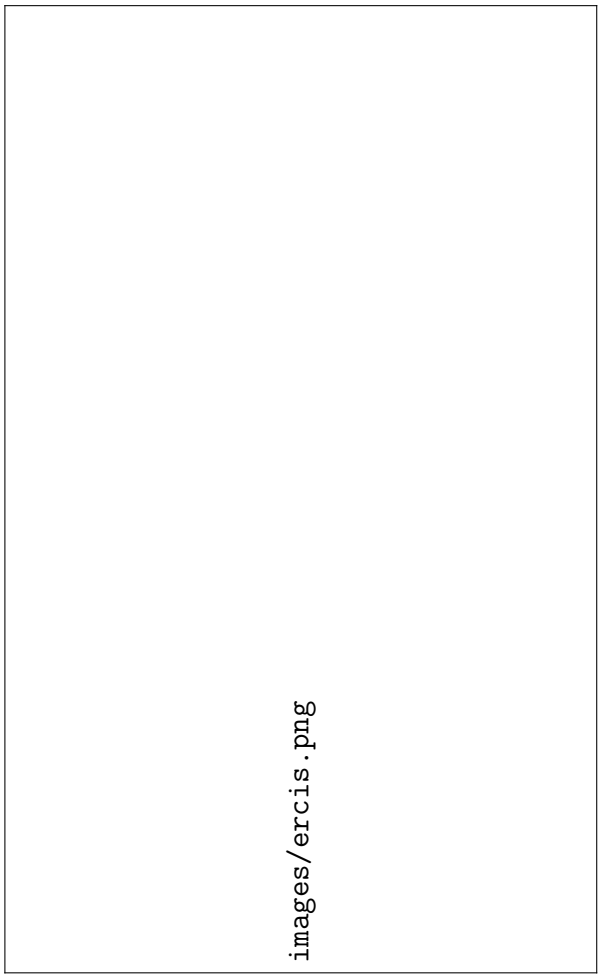


Figure 10: Logo of ERCIS as an example for figures

5.1 First Subsection

Example for a citation: [**?**],[**?**], [**?**]

5.2 Second Subsection

6 Section 2

Here could be a table, e.g. table 12 (which is on page 24):

Feature 1	Feature 2			
	case		studies	
	ca	te	go	ry
data	63,50%	9,56%	2,16%	1,17%
analytics	1,57%	0,41%	0,29%	0,41%

Table 12: This is the label of the table

If you want to relate to a figure or table from a different page, you could do it this way: Figure 10, see page 23, shows the ERCIS-Logo.

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 13: 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 14: 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 15: 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 16: Test results of the chi-squared test on type 7 glass with 30 initial classes

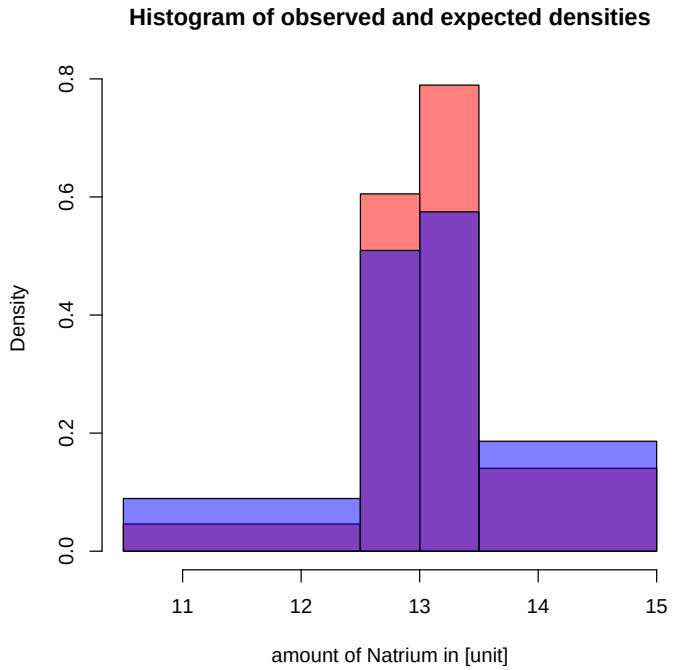


Figure 11: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass

List of Figures

1 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. 3
2 Empirical c. d. f. for Natrium vector (a) and theoretical normal c. d. f. with sample mean and sample variance (b) 7
3 Normal c. d. f. with optimized parameters in comparison to the old c. d. f. with sample mean and sample variance. 10
4 Histograms and QQ-plots of a Weibull(1, 1) simulated sample of size 500 and of the Box-Cox-transformed data 13
5 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 14

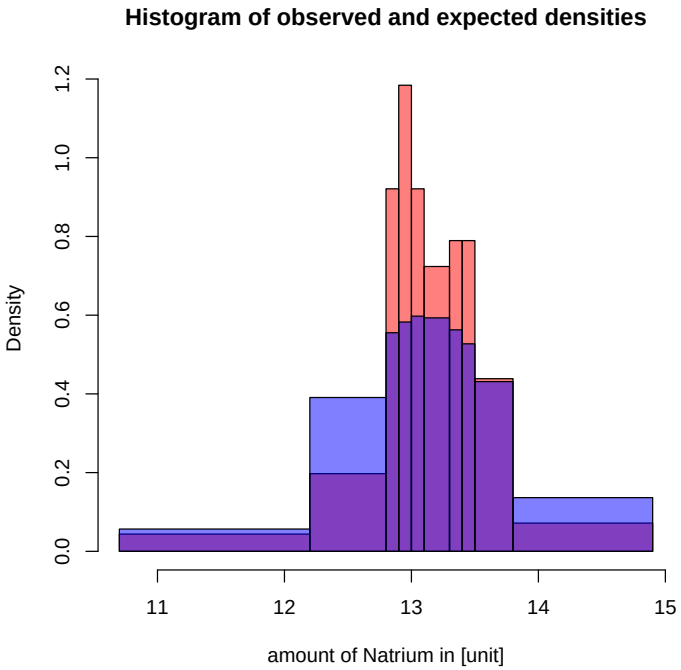


Figure 12: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass

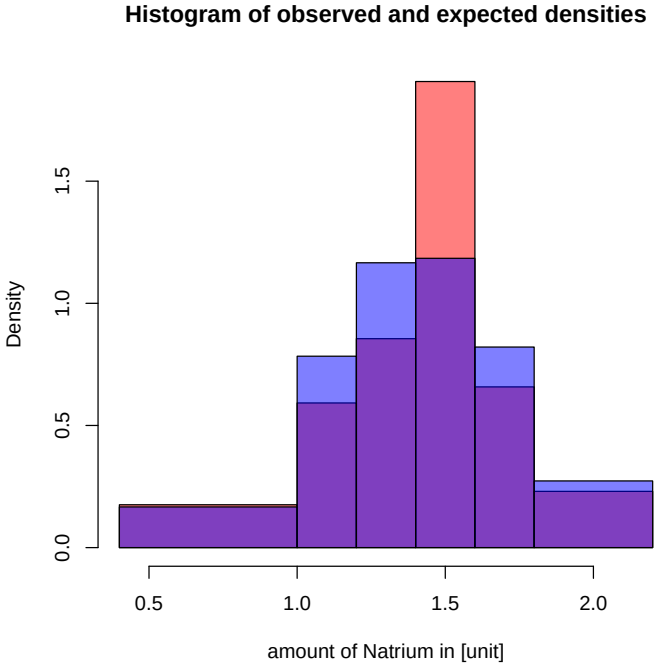


Figure 13: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass

6 cap 15

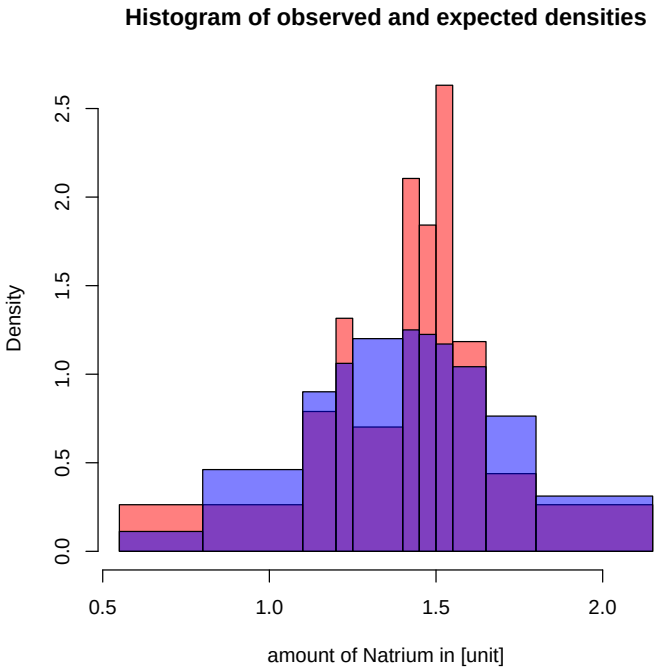


Figure 14: Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass

7 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass 18
8 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Calcium of type 1 glass 21
9 Histogram of observed densities (red) and expected densities (blue) within the classes for transformed values of the variable Calcium of type 1 glass . 22
10 Logo of ERCIS as an example for figures 23
11 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass ii
12 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass iii
13 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass iii
14 Histogram of observed densities (red) and expected densities (blue) within the classes for the variable Natrium of type 1 glass iv

List of Tables

1 Skipped variables for the particular glass types due to too many zero values 16
2 Test results of the chi-squared test on the whole data sample with ten initial classes 17
3 Test results of the chi-squared test on type 1 glass with ten initial classes . 17
4 Observed end expected frequencies of items in the classes for the variable Natrium of type 1 glass 17

5 Test results of the chi-squared test on type 2 glass with ten initial classes . 19
6 Test results of the chi-squared test on type 7 glass with ten initial classes . 19
7 Test results of the improved KS test on the whole data sample with ten initial classes 19
8 Test results of the chi-squared test on the whole transformed data sample with ten initial classes 20
9 Test results of the chi-squared test on the transformed data of type 1 glass with ten initial classes 21
10 Test results of the chi-squared test on the transformed data of type 2 glass with ten initial classes 22
11 Test results of the chi-squared test on the transformed data of type 7 glass with ten initial classes 23
12 This is the label of the table 24
13 Test results of the chi-squared test on the whole data sample with 30 initial classes i
14 Test results of the chi-squared test on type 1 glass with 30 initial classes . i
15 Test results of the chi-squared test on type 2 glass with 30 initial classes . i
16 Test results of the chi-squared test on type 7 glass with 30 initial classes . ii