Normal Distribution

Andrey Chinnov, Sebastian Honermann, Carlos Zydorek

Case Studies
"Data Analytics"

Outline

Normality as a requirement for statistical methods

Data Set Overview

- `Glass` data set from package `mlbench`
- sample of 214 observations
- 7 types of glass (but only 6 present in this sample)
- Variables: refractive index (RI) and 8 elements (Na, Mg, Al, Si, K, Ca, Ba, Fe)

Box-Cox Transformation
Definition

Box-Cox Transformation

If data are not normally distributed, they can possibly be transformed by the parameterised power transformation

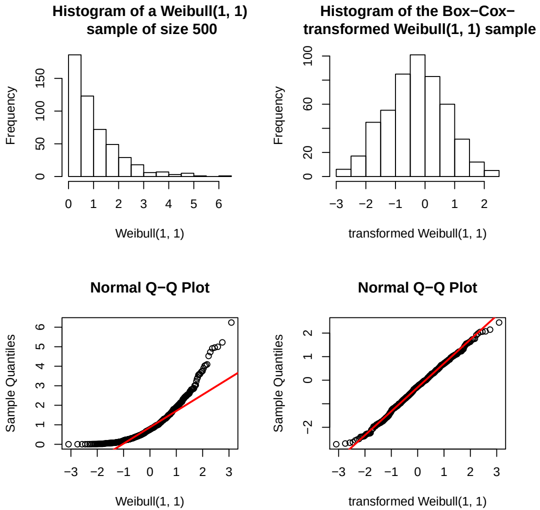
$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases} \qquad \text{for } x > 0$$

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

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

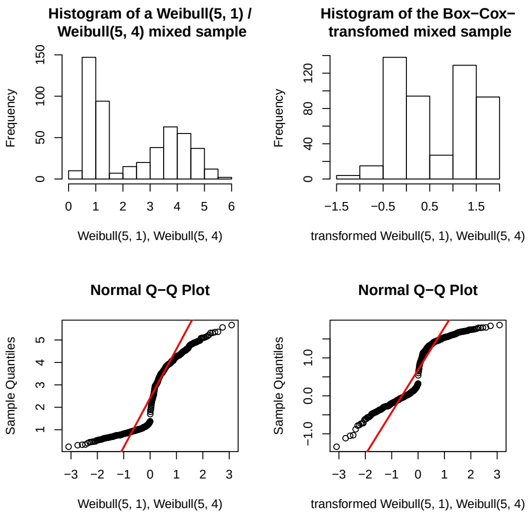
with $\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)}$

Box-Cox Transformation
Transformation issues



Box-Cox Transformation

Transformation issues



Q-Q-Plots

Sample:

$$x = (x_1, x_2, \ldots, x_n)$$

Empirical quantiles:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Theoretical quantiles:

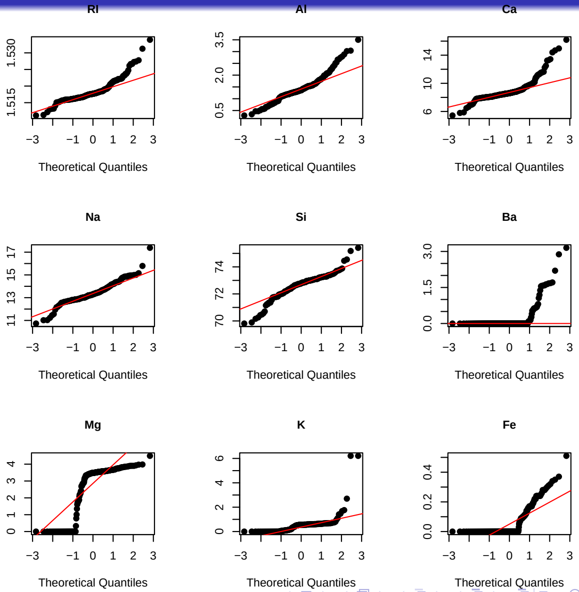
$$q_{(j)} = \Phi^{-1}(p_{(j)}),$$

where

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

Φ - $N(0,1)$ c. d. f. .

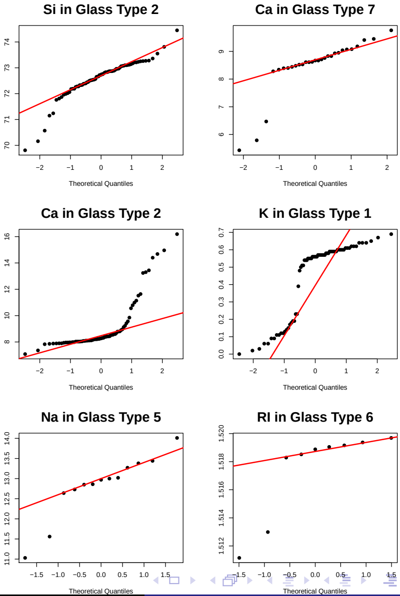
$\Longrightarrow$ Plot $x_{(i)}$ against $q_{(i)}$



Q-Q-Plots

QQ-Plots of the subdatasets:

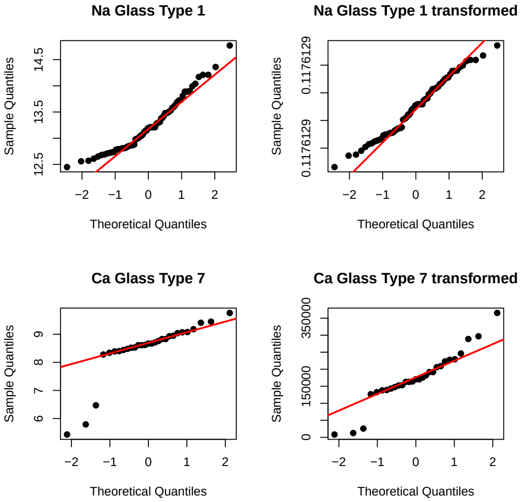
- Variables could be normally distributed within the subclasses
- For some cases there appear to be a linear relationships
- For other cases a linear relationship is questionable
- In some subdatasets a linear relationship seems plausible, however n is very small



Q-Q-Plots

Results of the Transformation of the Full Dataset :

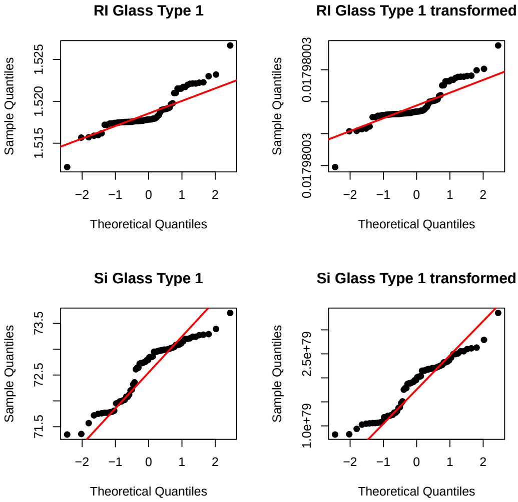
- For some of the cases there seems to be a slight improvement
- For non-unimodal cases the transformation does not show significant improvements towards normality



Q-Q-Plots

Results of the Transformation of the Subdatasets :

- For unimodal cases the transformation shapes the distribution closer to normality
- For non-unimodal cases the transformation does not show significant improvements towards normality



Shapiro-Wilk Test

The test statistic W indicates the deviation of the observed quantile values from the assumed cumulative distribution function quantiles

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

where

- a_i denotes the normalised "best linear unbiased" coefficients,
- y_i denotes the observations.

The critical value for W is obtained by the Monte Carlo Method
$\implies$ p-value is calculated

Important: If a variable contains only zeros the Shapiro-Wilk test is not applicable, since the term in the denominator sums up to zero.

Shapiro-Wilk Test

Testing the Full Dataset :

Null hypothesis is rejected for all variables at a 1 % significance level

After the Transformation :

The null hypothesis can be rejected for the four transformed variables
$\Longrightarrow$ Possible Explanation: Combination of different distributions in the different glass types

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

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

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

Test results of the Shapiro-Wilk test on the whole transformed data sample

Shapiro-Wilk Test

Testing the Subdatasets Example – Glass Type 1 :

Null hypothesis is rejected for all variables at a 1 % significance level

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

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

After the Transformation :

The null hypothesis cannot be rejected for 3 of the transformed variables

$\Longrightarrow$ Appearently the transformation was successful

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

Test results of the Shapiro-Wilk test on the transformed type 1 glass

Graphical Methods for Normality Testing

Shapiro-Wilk Test

Pearson's Chi-Squared Test

Theoretical foundations

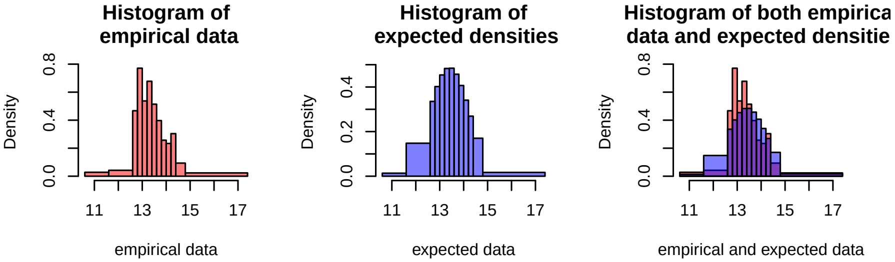
- Divide observations $X_1, \ldots, X_N$ into pairwise disjoint classes $C_1, \ldots, C_K$
- Common requirement: minimum class size of 5
- Compare observed class frequencies to expected theoretical class frequencies for a certain distribution

$$\text{test statistic:} \quad \chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}$$

- The test statistic is approximately χ^2-distributed with $K - 1$ degrees of freedom (minus one degree of freedom per estimated parameter)

Pearson's Chi-Squared Test

Theoretical foundations



Pearson's Chi-Squared Test

Theoretical foundations

- Test under the null hypothesis that the sample is drawn from a population with unknown distribution $\mathbb{P}$ which is equal to the assumed distribution $\mathbb{P}_0$:

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

Decision rule

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

(significance level α, number of estimated parameters p)

Pearson's Chi-Squared Test

Test results for the whole sample

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

Pearson's Chi-Squared Test

Test results for type 1 glass

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

Pearson's Chi-Squared Test

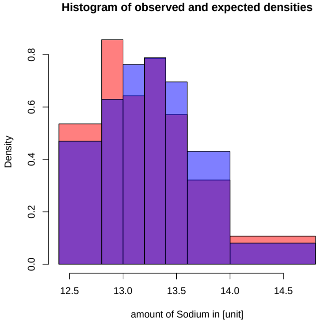
Test results for type 1 glass

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

Pearson's Chi-Squared Test

Test results for sodium of type 1 glass

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



Pearson's Chi-Squared Test

Test results for transformed type 1 glass

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

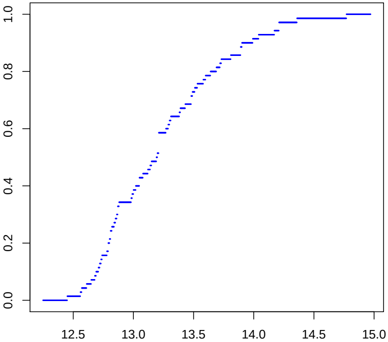
Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.

Definition

$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \le x\}}(x)$
- empirical c. d. f. , where

$$\mathbb{1}_{\{x_i \le x\}}(x) = \begin{cases} 1 & \text{if } x_i \le x \\ 0 & \text{otherwise.} \end{cases}$$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.

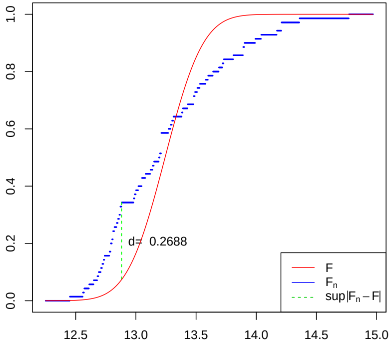
Definition

$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}(x)$
- empirical c. d. f. , where

$$\mathbb{1}_{\{x_i \leq x\}}(x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{otherwise.} \end{cases}$$

$F(x)$ - theoretical normal c. d. f.
with

$$\bar{x} = \frac{1}{n} \sum_i x_i, \quad \sigma_x^2 = \frac{1}{n}(x_i - \bar{x})^2$$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.

Definition

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

- empirical c. d. f. , where

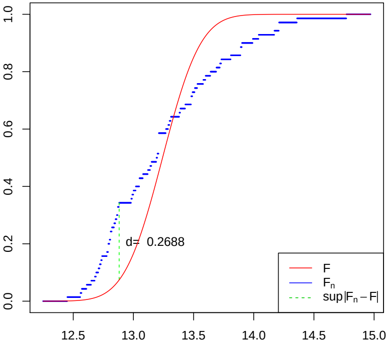
$$\mathbb{1}_{\{x_i \le x\}}(x) = \begin{cases} 1 & \text{if } x_i \le x \\ 0 & \text{otherwise.} \end{cases}$$

$F(x)$ - theoretical normal c. d. f.
with

$$\bar{x} = \frac{1}{n}\sum_i x_i, \quad \sigma_x^2 = \frac{1}{n}(x_i - \bar{x})^2$$

$$d = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

- distance between them.



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

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

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

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

KS test statistics:

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

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

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

KS test statistics:

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

Properties of D_n in case H_0 is TRUE:

- Distribution of $\hat{D}_n := (D_1, D_2, \ldots, D_n)$ does not depend on F

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \ldots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

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

KS test statistics:

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

Properties of D_n in case H_0 is TRUE:

- Distribution of $\hat{D}_n := (D_1, D_2, \ldots, D_n)$ does not depend on F

$\Longrightarrow$ tabulated

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

Let $x = (x_1, x_2, \dots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.
Theoretical c. d. f. F defines a distribution $\mathbb{P}_0$.

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

KS test statistics:

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

Properties of D_n in case H_0 is TRUE:

- Distribution of $\hat{D}_n := (D_1, D_2, \dots, D_n)$ does not depend on F $\Longrightarrow$ tabulated
- $\forall t > 0$:

$$P(D_n \leq t) \xrightarrow[n \to \infty]{} H(t) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp^{-2i^2 t^2}$$

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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 c - critical value

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

The KS test uses the decision rule

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

where c - critical value that
depends on a significance level α:

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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 c - critical value that
depends on a significance level α:

$$\alpha = P(\delta \neq H_0 | H_0)$$

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

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 c - critical value that
depends on a significance level α:

$$\alpha = P(\delta \neq H_0 | H_0) = P(D_n > c | H_0)$$

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

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 c - critical value that
depends on a significance level α:

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

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

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 c - critical value that
depends on a significance level α:

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

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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 c - critical value that
depends on a significance level α:

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

$$\Longrightarrow \; c \approx H_{1-\alpha}$$

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

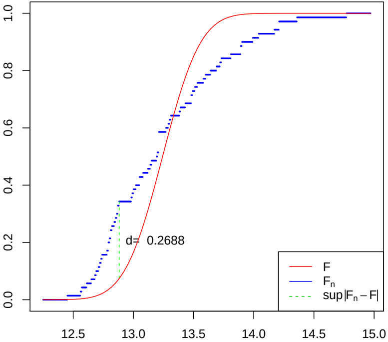
$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases} , \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

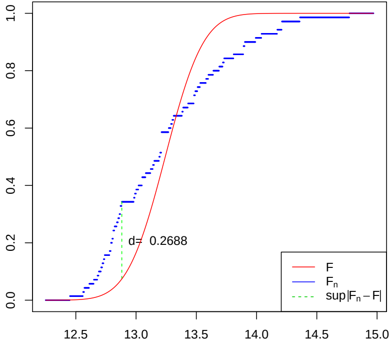
Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases} , \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:

- $n = 70$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

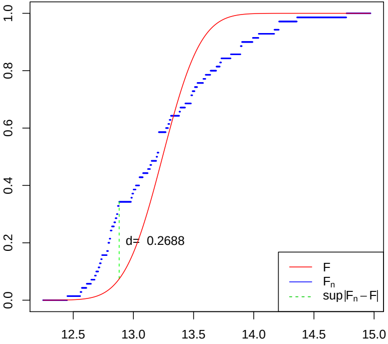
Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:

- $n = 70$
- $D_n = \sqrt{n}\sup|F_n - F| = 2.2493$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

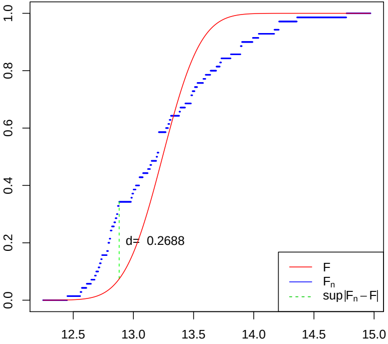
The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:

- $n = 70$
- $D_n = \sqrt{n}\sup|F_n - F| = 2.2493$
- $\alpha = 0.01$

$$\Longrightarrow \quad c = H_{1-\alpha} = 1.6276$$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

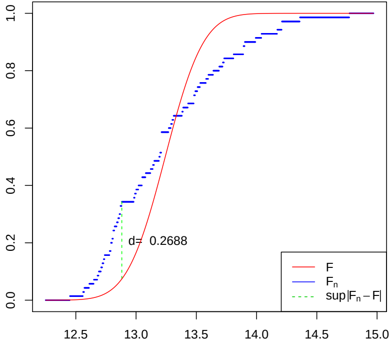
The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:

- $n = 70$
- $D_n = \sqrt{n}\sup|F_n - F| = 2.2493$
- $\alpha = 0.01$

  $\implies c = H_{1-\alpha} = 1.6276$
- $D_n > c \implies H_0$ rejected



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

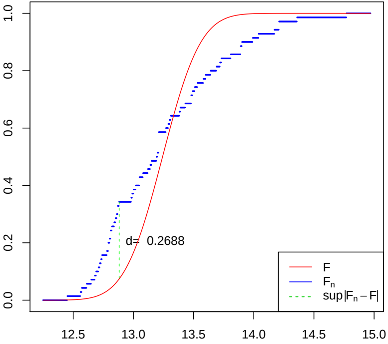
Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \le H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp^{-2i^2t^2}$$

Example:

- $n = 70$
- $D_n = \sqrt{n} \sup |F_n - F| = 2.2493$
- $\alpha = 0.01$
  $\implies \quad c = H_{1-\alpha} = 1.6276$
- $D_n > c \implies H_0$ rejected
- $\implies \mathbb{P} \neq \mathbb{P}_0$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

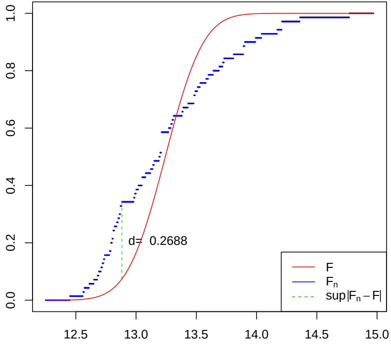
Kolmogorov-Smirnov Test

The KS test uses the decision rule for a given significance level α

$$\delta = \begin{cases} H_0 & : \quad D_n \leq H_{1-\alpha} \\ H_1 & : \quad D_n > H_{1-\alpha} \end{cases}, \qquad H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp^{-2i^2t^2}$$

Example:

- $n = 70$
- $D_n = \sqrt{n}\sup|F_n - F| = 2.2493$
- $\alpha = 0.01$
  $\implies \quad c = H_{1-\alpha} = 1.6276$
- $D_n > c \implies H_0$ rejected
- $\implies \mathbb{P} \neq \mathbb{P}_0$
- $\not\Longrightarrow$ data not normally distributed!!!



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

R code used:

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

R code used:

```
c(mean(dat),var(dat))
[1] 13.2422857  0.2493019
#optim is a predifined R function in stats package
#defalut method of optimization is Nelder and Mead
result = optim(c(mean(dat),var(dat)),KS)
result$par
[1] 13.1769501  0.4682486
result$value
[1] 0.07870673
```

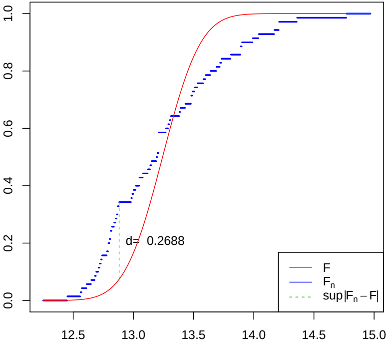
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$



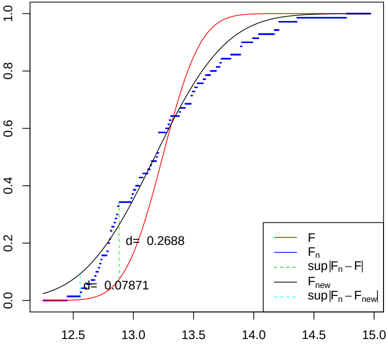
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$



Glass Type 1, Natrium (Na)

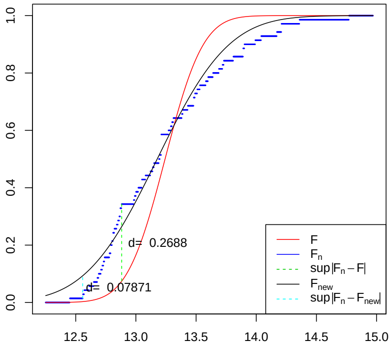
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$



Glass Type 1, Natrium (Na)

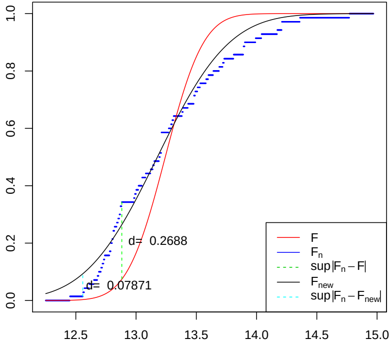
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$
- $c = 1.6276$



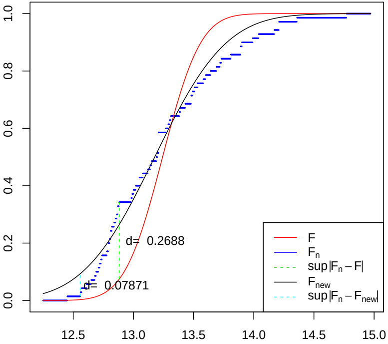
Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$
- $c = 1.6276$
- $D_n < c \implies H_0$ accepted



Glass Type 1, Natrium (Na)

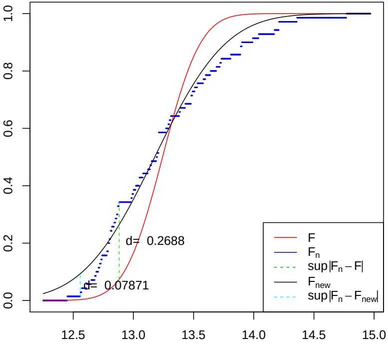
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$
- $c = 1.6276$
- $D_n < c \implies H_0$ accepted
- $\implies \mathbb{P} = \mathbb{P}_0$



Glass Type 1, Natrium (Na)

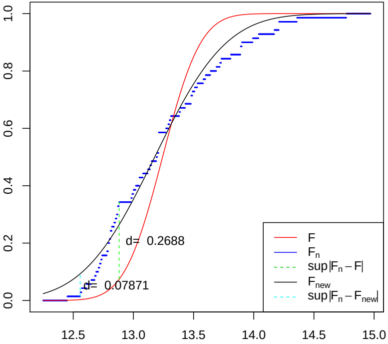
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- Initial vector of parameters
  $\mu = 13.2423, \quad \sigma^2 = 0.2493$
- Optimized vector of parameters
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$
- $c = 1.6276$
- $D_n < c \implies H_0$ accepted
- $\implies \mathbb{P} = \mathbb{P}_0$
- $\implies$ data normally distributed!



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Improved Kolmogorov-Smirnov Test Results

Results of Improved KS test on the whole data set:

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	2.60	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	$< 1.0e\text{-}15$	yes

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test Results

Results of Improved KS test on the whole data set:

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	2.60	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	$<$ 1.0e-15	yes

- 5 variables are normaly distributed (RI,Na,Al,Si,Ca)

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test Results

Results of Improved KS test on the whole data set:

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	2.60	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	$<$ 1.0e-15	yes

- 5 variables are normaly distributed (RI,Na,Al,Si,Ca)
- 4 variables are not (Mg,K,Ba,Fe)

Quantitative Methods for Normality Testing
Improved Kolmogorov-Smirnov Test Results

Results of Improved KS test on the whole data set:

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	2.60	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	$<$ 1.0e-15	yes

- 5 variables are normaly distributed (RI,Na,Al,Si,Ca)
- 4 variables are not (Mg,K,Ba,Fe)
- The best statistics test value for Al

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test Results

Results of Improved KS test on the whole data set:

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	2.60	0.01	1.63	2.75476085742632e-06	yes
Fe	4.68	0.01	1.63	$< 1.0e\text{-}15$	yes

- 5 variables are normaly distributed (RI,Na,Al,Si,Ca)
- 4 variables are not (Mg,K,Ba,Fe)
- The best statistics test value for Al
- The worst statistic test value for Fe

Quantitative Methods for Normality Testing

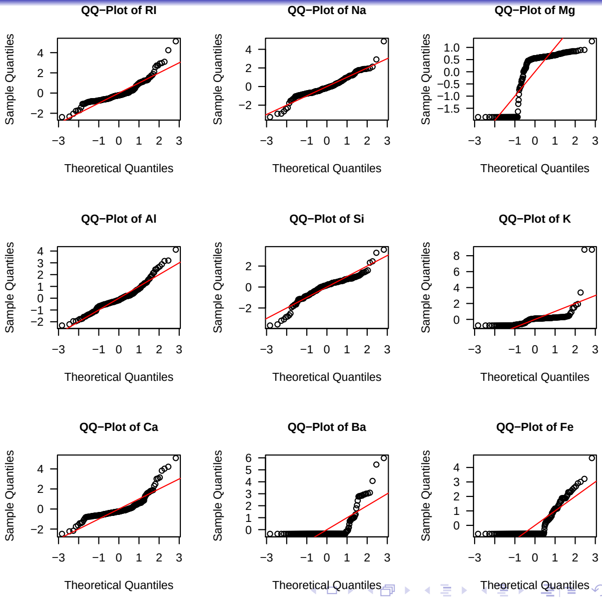
Test Results:

variable	rejected
RI	no
Na	no
Mg	yes
Al	no
Si	no
K	yes
Ca	no
Ba	yes
Fe	yes

Quantitative Methods for Normality Testing

Test Results:

variable	rejected
RI	no
Na	no
Mg	yes
Al	no
Si	no
K	yes
Ca	no
Ba	yes
Fe	yes



Plot of multivariate normal distribution

Theoretical foundations

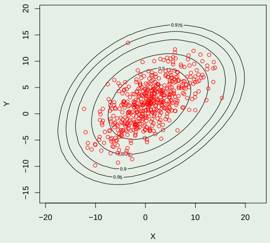
- Contour lines of the plot of a multivariate normal distribution are shaped elliptically
- Ellipsoids are centered at $\mu : \left\{x : (x-\mu)' \Sigma^{-1}(x-\mu) = c^2\right\}$ with some constant c.

Example

Multivariate normal distribution with sample size 500 and parameters

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

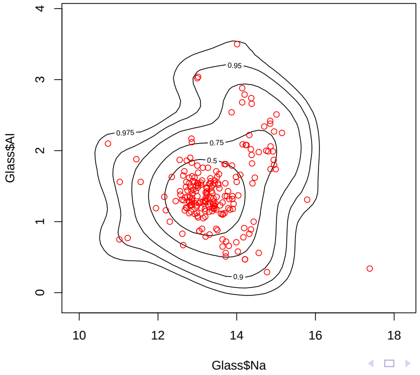
$$\Sigma = \begin{bmatrix} 27 & 15 \\ 15 & 18 \end{bmatrix}$$



Plot of multivariate normal distribution

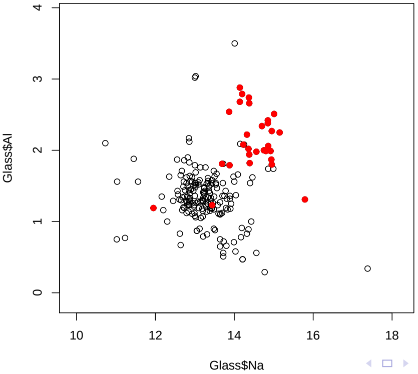
Application

- Concerning the complete sample, the p-values for the variables Na and Al are highest among all used test methods.
- Plot data points and determine contour lines.



Plot of multivariate normal distribution
Application

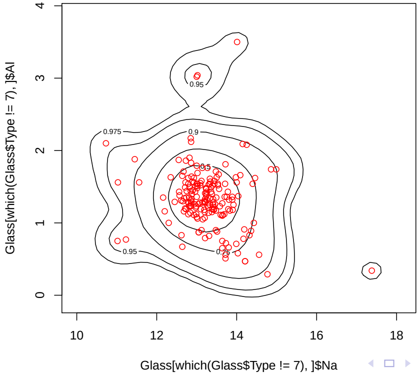
- Concerning the complete sample, the p-values for the variables Na and Al are highest among all used test methods.
- Plot data points and determine contour lines.



Plot of multivariate normal distribution

Application

- Concerning the complete sample, the p-values for the variables Na and Al are highest among all used test methods.
- Plot data points and determine contour lines.



Summary

Pearson's Chi-Squared Test
Theoretical foundations