

Normal Distribution

Andrey Chinnov, Sebastian Honermann, Carlos Zydorek

Case Studies
"Data Analytics"

Outline

① Introduction

- ▶ Normality as a requirement for statistical methods
- ▶ Data Set Overview

② Normality Testing

- ▶ Graphical Methods for Normality Testing
 - ★ Q-Q-Plots
 - ★ Chi-Square Plot
- ▶ Quantitative Methods for Normality Testing
 - ★ Shapiro-Wilk Test
 - ★ Pearson's Chi-Squared Test
 - ★ Kolmogorov-Smirnov Test

③ Transformation to Normality

- ▶ Box-Cox Transformation
- ▶ Transformation Results Testing

④ Summary

Normality as a requirement for statistical methods

Data Set Overview

Outline

① Introduction

- ▶ Normality as a requirement for statistical methods
- ▶ Data Set Overview

② Normality Testing

- ▶ Graphical Methods for Normality Testing
 - ★ Q-Q-Plots
 - ★ Chi-Square Plot
- ▶ Quantitative Methods for Normality Testing
 - ★ Shapiro-Wilk Test
 - ★ Pearson's Chi-Squared Test
 - ★ Kolmogorov-Smirnov Test

③ Transformation to Normality

- ▶ Box-Cox Transformation
- ▶ Transformation Results Testing

④ Summary

Graphical Methods for Normality Testing

Q-Q-Plots

Graphical Methods for Normality Testing

Chi-Square Plot

Quantitative Methods for Normality Testing

Shapiro-Wilk Test

Quantitative Methods for Normality Testing

Pearson's Chi-Squared Test

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

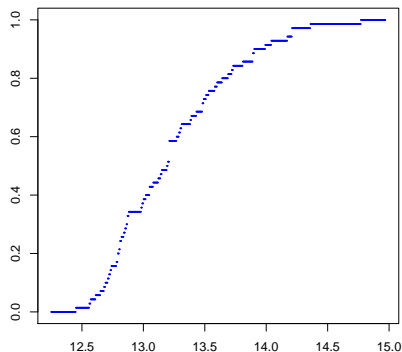
Let $x = (x_1, x_2, \dots, x_n)$ be a sample of unknown distribution \mathbb{P} .

Definition

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

- **empirical** c. d. f. , where

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



Glass Type 1, Natrium (Na)

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} .

Definition

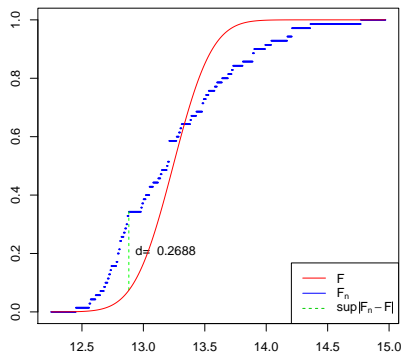
$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}}(x)$$

- **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 $x = (x_1, x_2, \dots, x_n)$ 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

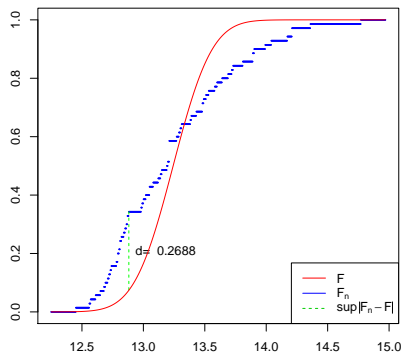
$$\mathbb{1}_{\{x_i \leq x\}}(x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{otherwise.} \end{cases}$$

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

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

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

- distance between them.



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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

$$H_0 : \mathbb{P} = \mathbb{P}_0,$$

$$H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

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 .

$$H_0 : \mathbb{P} = \mathbb{P}_0,$$

$$H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

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

$$H_0 : \mathbb{P} = \mathbb{P}_0,$$

$$H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

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

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 .

$$H_0 : \mathbb{P} = \mathbb{P}_0,$$

$$H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

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
 \implies **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 .

$$H_0 : \mathbb{P} = \mathbb{P}_0,$$

$$H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

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
 \implies **tabulated**
- $\forall t > 0 :$

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

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

The KS test uses the decision rule

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

where c - critical value

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

The KS test uses the decision rule

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\implies c \approx H_{1-\alpha}$$

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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

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

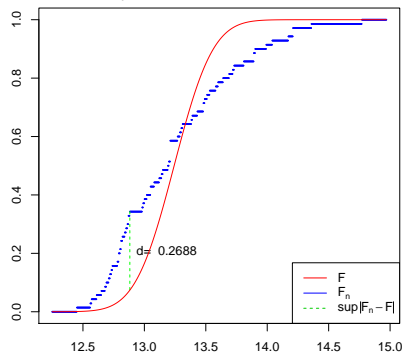
Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

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

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

Example:



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

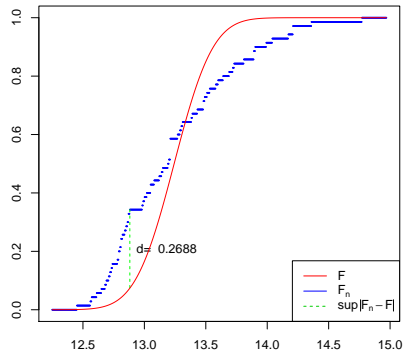
Kolmogorov-Smirnov Test

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

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

Example:

● $n = 70$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

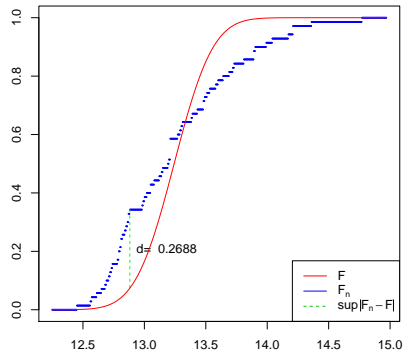
Kolmogorov-Smirnov Test

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

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

Example:

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



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

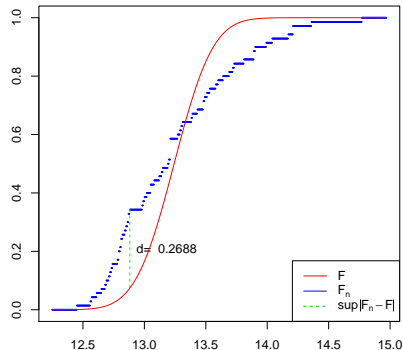
Kolmogorov-Smirnov Test

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

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

Example:

- $n = 70$
 - $D_n = \sqrt{n} \sup |F_n - F| = 2.2493$
 - $\alpha = 0.01$
- $$\Rightarrow c = H_{1-\alpha} = 1.6276$$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

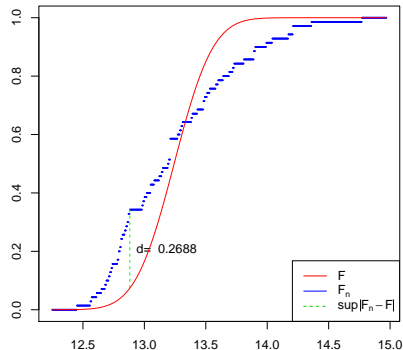
Kolmogorov-Smirnov Test

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

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

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

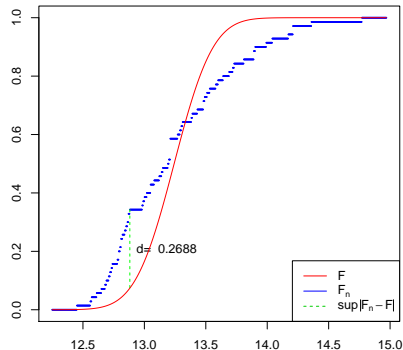
Kolmogorov-Smirnov Test

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

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

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**
- $\implies \mathbb{P} \neq \mathbb{P}_0$



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

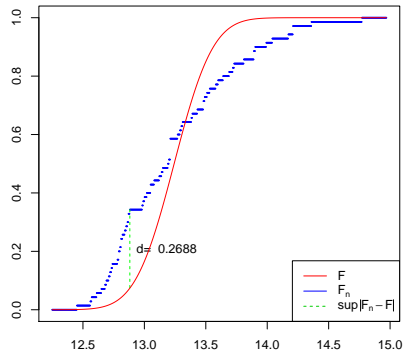
Kolmogorov-Smirnov Test

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

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

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**
- $\implies \mathbb{P} \neq \mathbb{P}_0$
- $\not\Rightarrow$ **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

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

R code used:

Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

R code used:

```
c(mean(dat), var(dat))
```

```
[1] 13.2422857 0.2493019
```

```
#optim is a predefined R function in stats package
```

```
#default 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

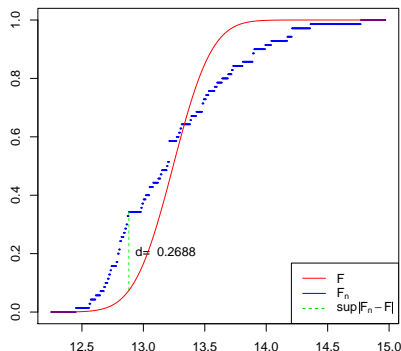
$$KS(\mu, \sigma) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x, \mu, \sigma)| \rightarrow \min.$$

- Initial vector of parameters

$$\mu = 13.2423, \quad \sigma^2 = 0.2493$$

- Optimized vector of parameters

$$\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$$



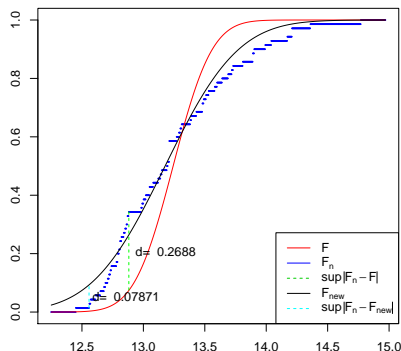
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

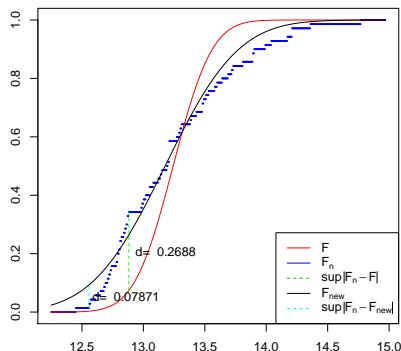
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

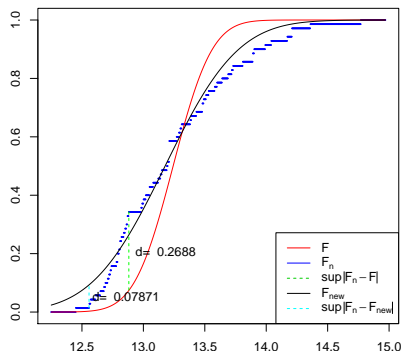
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

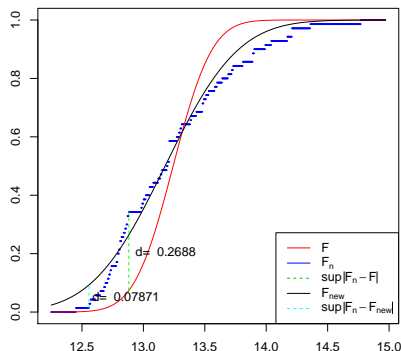
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

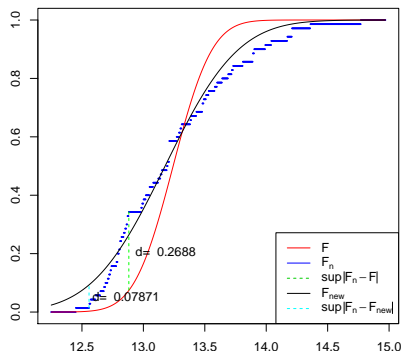
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

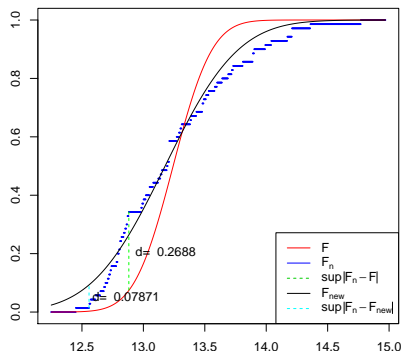
Quantitative Methods for Normality Testing

Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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-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-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:

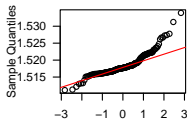
variable	rejected
Rl	no
Na	no
Mg	yes
Al	no
Si	no
K	yes
Ca	no
Ba	yes
Fe	yes

Quantitative Methods for Normality Testing

Test Results:

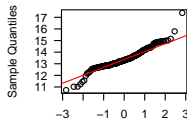
variable	rejected
RI	no
Na	no
Mg	yes
Al	no
Si	no
K	yes
Ca	no
Ba	yes
Fe	yes

QQ-Plot of iris (RI)



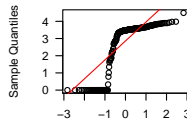
Theoretical Quantiles

QQ-Plot of iris (Na)



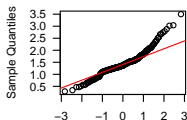
Theoretical Quantiles

QQ-Plot of iris (Mg)



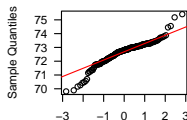
Theoretical Quantiles

QQ-Plot of iris (Al)



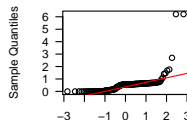
Theoretical Quantiles

QQ-Plot of iris (Si)



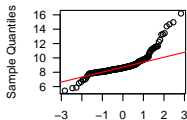
Theoretical Quantiles

QQ-Plot of iris (K)



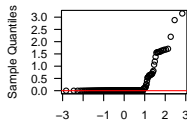
Theoretical Quantiles

QQ-Plot of iris (Ca)



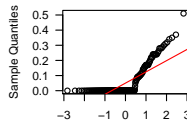
Theoretical Quantiles

QQ-Plot of iris (Ba)



Theoretical Quantiles

QQ-Plot of iris (Fe)



Theoretical Quantiles

Outline

① Introduction

- ▶ Normality as a requirement for statistical methods
- ▶ Data Set Overview

② Normality Testing

- ▶ Graphical Methods for Normality Testing
 - ★ Q-Q-Plots
 - ★ Chi-Square Plot
- ▶ Quantitative Methods for Normality Testing
 - ★ Shapiro-Wilk Test
 - ★ Pearson's Chi-Squared Test
 - ★ Kolmogorov-Smirnov Test

③ Transformation to Normality

- ▶ Box-Cox Transformation
- ▶ Transformation Results Testing

④ Summary

Box-Cox Transformation

Transformation Results Testing

Outline

① Introduction

- ▶ Normality as a requirement for statistical methods
- ▶ Data Set Overview

② Normality Testing

- ▶ Graphical Methods for Normality Testing
 - ★ Q-Q-Plots
 - ★ Chi-Square Plot
- ▶ Quantitative Methods for Normality Testing
 - ★ Shapiro-Wilk Test
 - ★ Pearson's Chi-Squared Test
 - ★ Kolmogorov-Smirnov Test

③ Transformation to Normality

- ▶ Box-Cox Transformation
- ▶ Transformation Results Testing

④ Summary

Summary