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

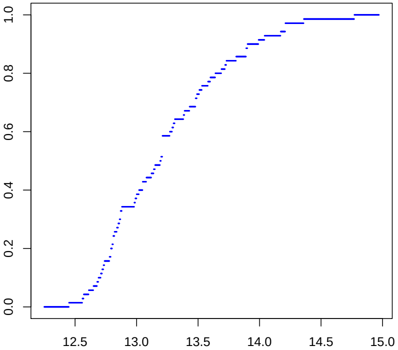
Let $x = (x_1, x_2, \dots, x_n)$ be a sample of unknown distribution $\mathbb{P}$.

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

Definition

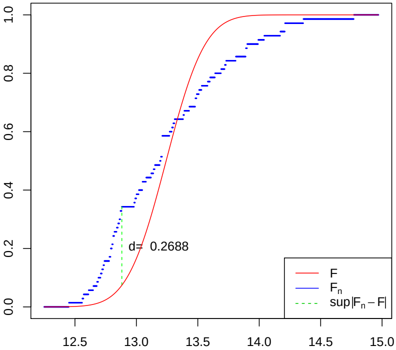
$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}(x)$

- 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

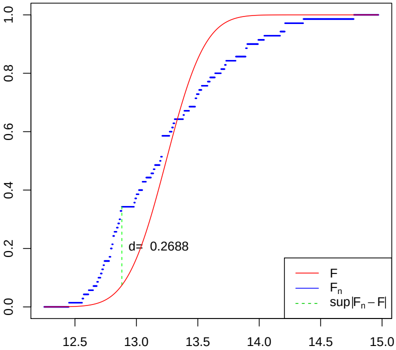
$$\mathbb{1}_{\{x_i \le x\}}(x) = \begin{cases} 1 & \text{if } x_i \le x \\ 0 & \text{otherwise.} \end{cases}$$

$F(x)$ - 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 $\implies$ 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 $\mathbb{P}_0$.

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

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 $\hat{D}_n := (D_1, D_2, \dots, D_n)$ does not depend on F $\implies$ 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

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

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:

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

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

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

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

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

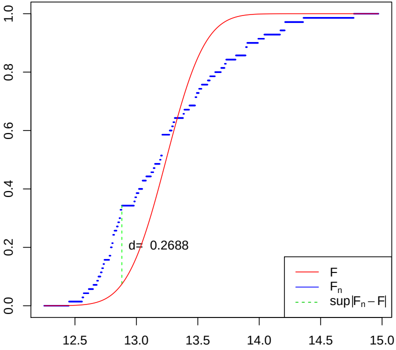
Quantitative Methods for Normality Testing

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)

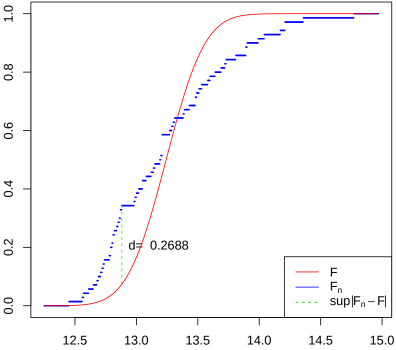
Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

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

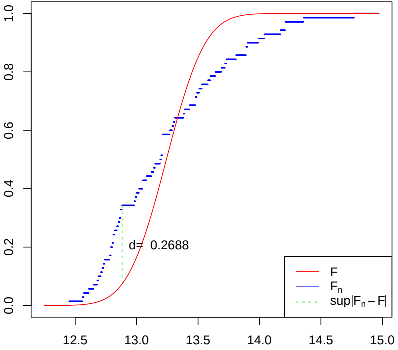
Kolmogorov-Smirnov Test

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

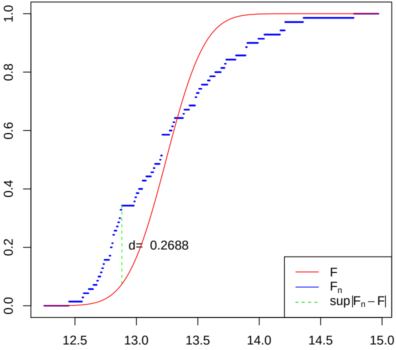
The KS test uses the decision rule for a given significance level α

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

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

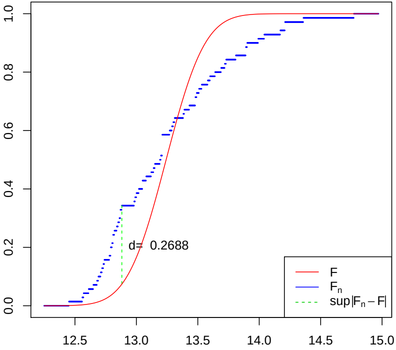
The KS test uses the decision rule for a given significance level α

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

Example:

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

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



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Kolmogorov-Smirnov Test

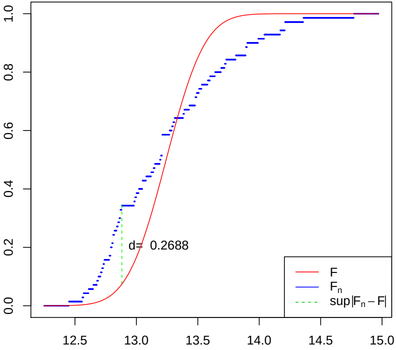
The KS test uses the decision rule for a given significance level α

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

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



Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing

Kolmogorov-Smirnov Test

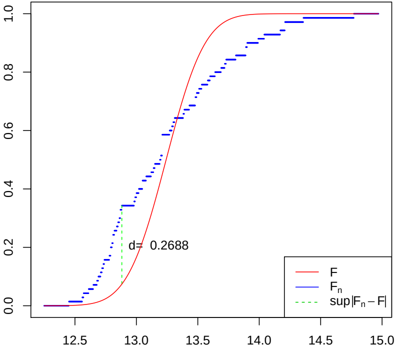
The KS test uses the decision rule for a given significance level α

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

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
- $\implies \mathbb{P} \neq \mathbb{P}_0$
- $\not\Longrightarrow$ 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

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

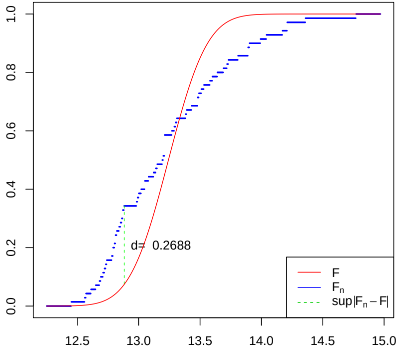
Quantitative Methods for Normality Testing

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



Glass Type 1, Natrium (Na)

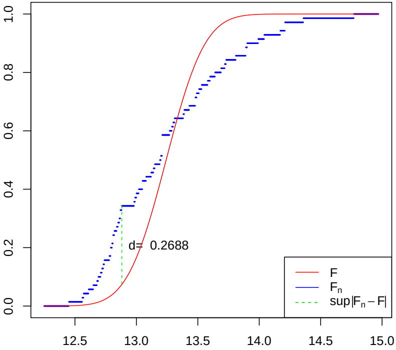
Quantitative Methods for Normality Testing

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



Glass Type 1, Natrium (Na)

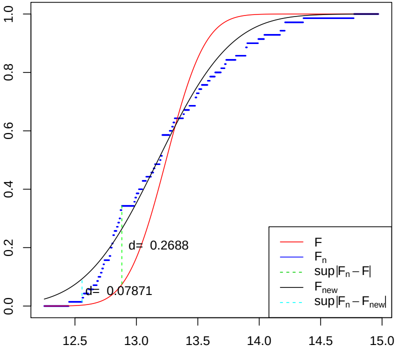
Quantitative Methods for Normality Testing

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
  $\hat{\mu} = 13.1770, \quad \hat{\sigma}^2 = 0.4682$
- $D_n = \sqrt{n} \sup |F_n - F_{new}| = 0.6585$



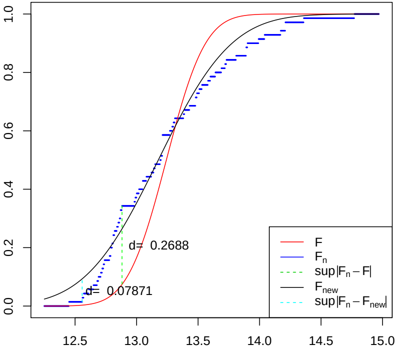
Glass Type 1, Natrium (Na)

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)

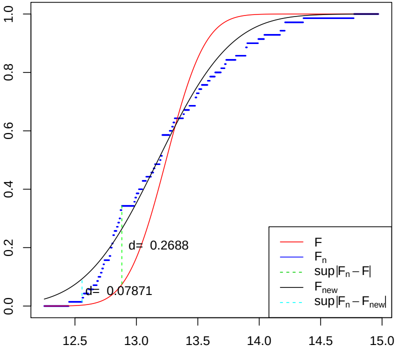
Quantitative Methods for Normality Testing

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



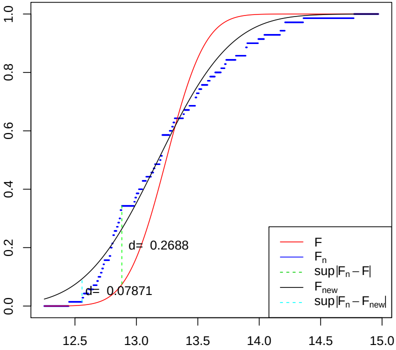
Glass Type 1, Natrium (Na)

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
- $\implies \mathbb{P} = \mathbb{P}_0$



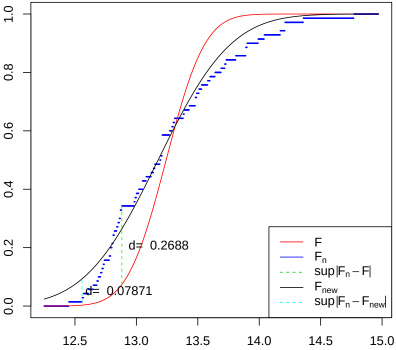
Glass Type 1, Natrium (Na)

Quantitative Methods for Normality Testing
Improved Kolmogorov-Smirnov Test

KS test is improved by solving the following optimization problem

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

- 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)

Outline

Box-Cox Transformation

Transformation Results Testing

Outline

Summary