Statistical Data Analysis

Dr. Jana de Wiljes

9. November 2021

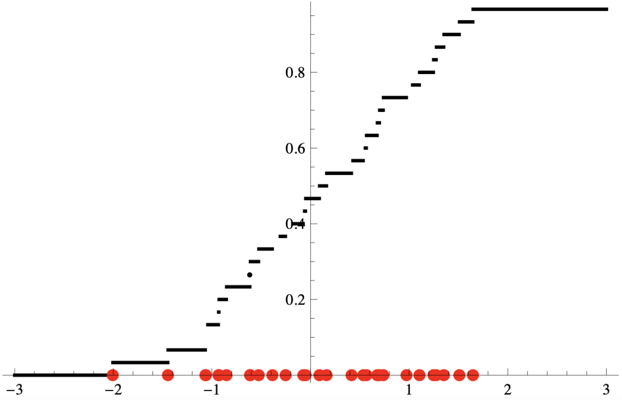
Universität Potsdam

Empirical cdf of a sample set

The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq 1} = \frac{1}{n} \#\{i \in \{1, \ldots, n\} : x_i \leq t\}, \quad t \in \mathbb{R} \quad (1)$$

Empirical cdf



Empirical cdf

The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq 1} \tag{2}$$

Proposition

Proposition: Let $(X_1, X_2 \ldots.)$ independent and identical distributed random variabel with cdf F. Then

- $$n\widehat{F}_n(t) \sim Bin(n, F(t)). \tag{3}$$

  This means
  $$\mathbb{P}\left[\widehat{F}_n(t) = \frac{k}{n}\right] = \binom{n}{k} F(t)^k (1 - F(t))^{n-k}, \quad k = 0, 1, \ldots, n.$$

- The expect value and variance of $\widehat{F}_n(t)$ are given by

  $$\mathbb{E}[\widehat{F}_n(t)] = F(t), \quad \mathrm{Var}[\widehat{F}_n(t)] = \frac{F(t)(1 - F(t))}{n} \tag{5}$$

  i.e., $\widehat{F}_n(t)$ is an unbiased estimator of F(t).

- For all $t \in \mathbb{R}$ it holds that
  $$\widehat{F}_n(t) \to F(t) \quad n \to \infty \text{ almost everywhere} \tag{6}$$

- For all $t \in \mathbb{R}$ with $F(t) \neq 0$ or 1 the following holds:

  $$\sqrt{n} \frac{\widehat{F}_n(t) - F(t)}{\sqrt{F(t)(1 - F(t))}} \to \mathcal{N}(0, 1) \text{ for } n \to \infty \text{ (in distribution)} \tag{7}$$

Proof

Proof

Proof

Proof

Theoretical distribution

Def: Let X be a random. The theoretical distribution of X is a probability measure μ on $(\mathbb{R}, \mathcal{B})$ with

$$\mu(A) = \mathbb{P}[X \in A] \text{ for every Borel set A} \subset \mathbb{R} \tag{8}$$

Note: the relationship between the theoretical distribution μ and the theoretical cdf F is;

$$F(t) = \mu((-\infty, t]), \quad t \in \mathbb{R} \tag{9}$$

Empirical distribution

Def: The empirical distribution of a sample set $(x_1, \dots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{\mu}_n(A) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \in A} = \frac{1}{n}\#\{i \in \{1, \dots, n\} : x_i \in A\}, \qquad (10)$$

for every Borel set $A \subset \mathbb{R}$

Dirac δ measure

Def: Let $x \in \mathbb{R}$ be a real number. The dirac-δ measure δ_x is a probability measure on $(\mathbb{R}, \mathcal{B})$ with with

$$\delta_x(A) \begin{cases} 1, & \text{for } x \in A \\ 0, & \text{for } x \notin A \end{cases} \tag{11}$$

for all Borel set $A \subset \mathbb{R}$

Remark: Then the empirical measure $\widehat{\mu}_n$ can be written as

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \tag{12}$$

and further note that

$$\widehat{F}_n(t) = \widehat{\mu}_n((-\infty, t]) \tag{13}$$

Proposition

Proposition: Let $(X_1, X_2 \ldots.)$ independent and identical distributed random variabel with distribution μ and let $A \subset \mathbb{R}$ a Borel set. Then

- $$n\widehat{\mu}_n(A) \sim Bin(n, \mu(A)). \tag{14}$$
- The expect value and variance of $\widehat{\mu}_n(A)$ are given by

  $$\mathbb{E}[\widehat{\mu}_n(A)] = \mu(A), \quad \mathrm{Var}[\widehat{\mu}_n(A)] = \frac{\mu(A)(1-\mu(A))}{n} \tag{15}$$

  i.e., $\widehat{\mu}_n(A)$ is an unbiased estimator of $\mu(A)$.
- Further it follows that $\widehat{\mu}_n$ is a consistent estimator, i.e.,

  $$\widehat{\mu}_n(A) \to \mu(A) \quad n \to \infty \text{ almost everywhere} \tag{16}$$

- For $\mu(A) \neq 0$ or 1 the following holds:

  $$\sqrt{n}\frac{\widehat{\mu}_n(A) - \mu(A)}{\sqrt{\mu(A)(1-\mu(A))}} \to \mathcal{N}(0,1) \text{ for } n \to \infty \text{ (in distribution)} \tag{17}$$

Plugin Estimator

Setting: Let $(X_1, \ldots, X_n)$ be independent and identical distributed random variables with the distribution μ. Further we assume that a reliasation $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of the respective random variables

Goal: approximate $\Psi(\mu)$ where $\Psi : \mathcal{M} \to \mathbb{R}$

Def: $\Psi(\widehat{\mu}_n)$ is called the plugin estimator of $\Psi(\mu)$.

Example

Example

Kolmogorov-distance

Def: The Kolmogorov-distance between the empirical cdf $\widehat{F}_n(t)$ and the theoretical cdf F is defined as follows

$$D_n := \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \tag{18}$$

Theorem of Gliwenko-Cantelli

Theorem: For the Kolmogorov-distance D_n the following holds

$$D_n \to 0 \text{ for } n \to \infty \text{ almost everywhere} \tag{19}$$

i.e.,

$$\mathbb{P}\Big[\lim_{n\to} D_n = 0\Big] = 1 \tag{20}$$

Proof

Proof

Proof

Proof