Statistical Data Analysis

Dr. Jana de Wiljes

9. November 2021

Universität Potsdam

Empirical cdf of a sample set

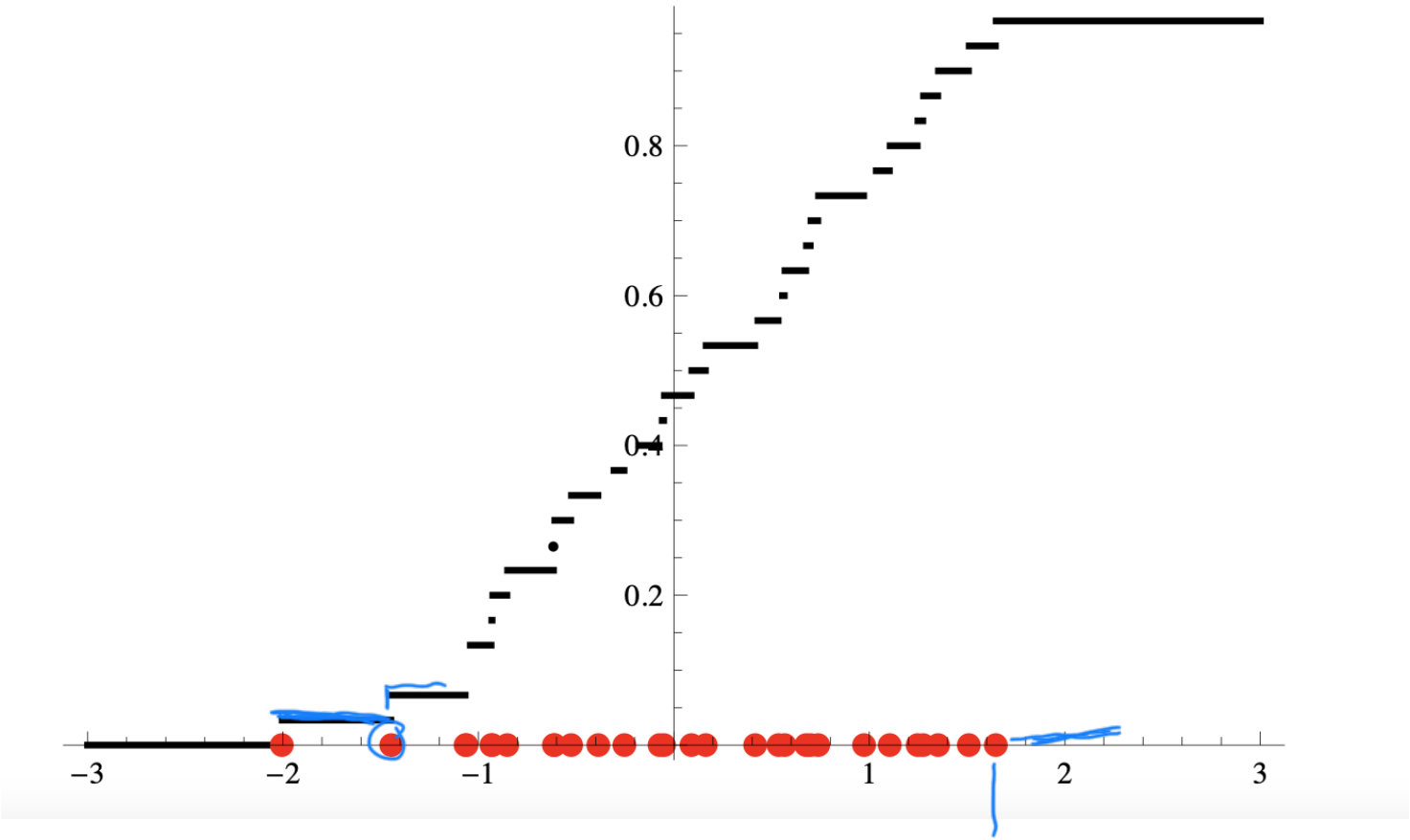
The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

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

Remark: the above defined empiral distribution can be approximated as follows via the order statistics $x_{(1)}, \ldots, x_{(n)}$

$$\hat{F}_n(t) = \begin{cases} 0 & \text{for } t < x_{(1)} \\ 1/n & x_{(1)} \le t < x_{(2)} \\ 2/n & x_{(2)} \le t < x_{(3)} \\ \vdots & \vdots \\ n-1/n & \text{for } x_{(n-1)} \le t < x_{(n)} \\ 1 & \text{for } x_{(n)} \le t \end{cases}$$

Empirical cdf



Empirical cdf

$X_1, \ldots, X_n$ are i.i.d.

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

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

Remark (2): The empirical distribution $\widehat{F}_n$ has all the properties of cdf as following holds:

1) $\lim_{t \to -\infty} \widehat{F}_n(t) = 0$ $\quad \lim_{t \to \infty} \widehat{F}_n(t) = 1$

2) $\widehat{F}_n$ is mondonly increasing

3) $\widehat{F}_n$ is right continuous

Proposition

Proposition: Let $(X_1, X_2 \ldots\ldots)$ 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

$$\widehat{F}_n(t) \xrightarrow{a.s.} F(t) \quad n \to \infty \text{ almost everywhere} \tag{6}$$

- 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

Remark: Part 4 can be interpreted as the distribution of the error $\hat{F}_n(t) - F(t)$ is for large n approximately distributed according

$$N\left(0, \frac{F(t)\,(1-F(t))}{n}\right)$$

Proof (1) Let us consider n experiments. For every i-th experiment we check if $X_i \leq t$. In case $X_i \leq t$ we say that the experiments a success, all the experiments are independent as the RV $X_1, \ldots, X_n$ are independent

$\Rightarrow$ the probability for success is $P[X_i \leq t] = F(t)$

$\Rightarrow$ the number of successes in the n experiments is described by the random Variable

$$n\hat{F}_n(t) = \sum_{i=1}^{n} \mathbf{1}_{X_i \leq t}$$

which is binomially distributed with parameters n and $F(t)$

Proof

Proof of (2): We showed that $n\hat{F}_n(t)$ distributed according to $Bin(n, F(t))$

We know that the expected value and the variance of a random variable is that is binomial distributed (for n experiments and probably of success equal to $F(t)$)

$$\mathbb{E}[n\hat{F}_n(t)] = n \cdot F(t) \quad \Big| \text{ take } n \text{ out of } \mathbb{E}[\cdot] \text{ and divide by } n$$

$$\Rightarrow \mathbb{E}[\hat{F}_n(t)] = F(t)$$

The variance of a $Bin(n, p)$ distributed random is

$$Var(X) = n \cdot p(1-p) \quad \Rightarrow Var[n\hat{F}_n(t)] = n \cdot F(t) \cdot (1 - F(t))$$

$\rightsquigarrow$ we can pull out n out of the variance and get a factor of n^2 and then we divide by n^2. That yields

$$Var[\hat{F}_n(t)] = \frac{F(t) \cdot (1 - F(t))}{n}$$

□

Proof

Proof of (3) We introduce $Y_i = \mathbb{1}_{X_i \leq t}$, Y_i are independent and identical distributed as X_i are.

with $\mathbb{P}[Y_i = 1] = \mathbb{P}[X_i \leq t] = F(t)$

$$\mathbb{P}[Y_i = 0] = 1 - \mathbb{P}[X_i \leq t] = 1 - F(t)$$

$\rightsquigarrow$ Bernoulli distributed

$\mathbb{E}[Y_i] = F(t)$

We can apply the law of large numbers to $Y_1, Y_2, \ldots$

$$\Rightarrow \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow[n \to \infty]{\text{almost everywhere}} \mathbb{E}[Y_1] = F(t)$$

$$\frac{1}{n} \sum X_i \to \mu$$

$$\mathbb{E}[X_i] = \mu$$

□

Proof

Proof of (4): Using the definition for Y_i (from proof of (3))

$\mathbb{E}[Y_i] = F(t)$, $\mathrm{Var}(Y_i) = F(t)(1 - F(t))$

(using the fact that Y_i are Bernoulli distributed)

And we make use of the central limit theorem for the series $Y_1, Y_2, \ldots$

$$\sqrt{n}\,\frac{\hat{F}_n(t) - F(t)}{\sqrt{F(t)\cdot(1-F(t))}} = \sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1]}{\sqrt{\mathrm{Var}(Y_1)}} = \frac{\sum_{i=1}^{n} Y_i - n\,\mathbb{E}[Y_1]}{\sqrt{n\,\mathrm{Var}\,Y_1}}$$

$\xrightarrow[n \to \infty]{\text{convergence in probability}} N(0,1)$

□

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 $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

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

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

$F(t) = \mu((-\infty, t])$

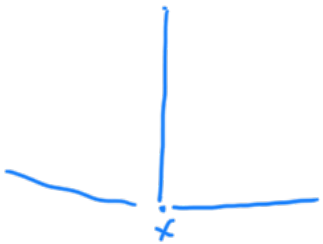
$[0, t]$

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

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

  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: (empirical moments). Let $m_k := \psi(\mu) = \int_{\mathbb{R}} x^k \cdot \mu(dx) = \mathbb{E}[X_i^k]$ the k-th moment of μ. The plugin estimator $\hat{m}_k$ is given by

$$\hat{m}_k = \int_{\mathbb{R}} x^k \, \hat{\mu}_n(dx) = \frac{x_1^k + \dots + x_n^k}{n}$$

$\hat{m}_k$ is called k-th empirical moment of a sample set $(x_1, \dots, x_n)$

Note that $\overline{X}_n = \hat{m}_1$

Example: Let $\psi(\mu) = \mathrm{Var}\, X_i = \int_{\mathbb{R}} x^2 \, \mu(dx) - \left(\int_{\mathbb{R}} x \mu(dx) \right)^2$ the variance associated with μ. The plugin estimator is given by

$$\hat{\sigma}^2_{\text{Plugin}} = \int_{\mathbb{R}} x^2 \, \hat{\mu}_n(dx) - \left(\int_{\mathbb{R}} x \, \hat{\mu}_n(dx) \right)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \overline{X}_n^2 = \frac{n-1}{n} S_n^2$$

$\uparrow$ $n-1$

where s_n^2 is the empirical variance

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

$$D_n \xrightarrow{a.s.} 0 \text{ for } n \to \infty \text{ almost everywhere} \tag{19}$$

i.e.,

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

Remark ① Note that before we showed that $\hat{F}_n(t) \to F(t)$ a.s for every fixed $t \in \mathbb{R}$

~) pointwise convergence

Now we show that is also converges uniformly

Remark ② From the almost sure convergence follows the convergence in Distribution which means:

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| > \varepsilon\right] = 0$$

Proof

Proof

Proof

Proof