Introduction to statistics

Prof. Dr. Christoph Richard, Leonie Wicht, Anna Vandebosch and Theresa Schmid

January 4, 2024

T1. Empirical Cumulative Distribution Function

Let $x_1, x_2, \ldots, x_n$ be a data sequance with empirical cumulative distribution function $F_n(t)$ and relative interval frecuencies $h_n(I)$, i.e

$$F_n : \mathbb{R} \to [0,1], t \mapsto h_n((-\infty, t]) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(x_i)$$

Show that for any real numbers $a < b$ we have:

$$\begin{aligned} h_n((a,b]) &= F_n(b) - F_n(a) \\ h_n(\{a\}) &= F_n(a) - F_n(a-) \\ h_n([a,b]) &= F_n(b) - F_n(a) + h_n(\{a\}) \\ h_n([a,b)) &= F_n(b-) - F_n(a-) \\ h_n((a,\infty)) &= 1 - F_n(a) \end{aligned}$$

Here $F(t-) := \lim_{x \uparrow t} F(x)$ denotes the limit from the left. If you manipulate the indicator functions, give proof of any of these rules.

Solution T1

To prove these rules, we first have to notice the following manipulations of the indicator function:

a. $\mathbb{1}_{(a,b]}(x) = \mathbb{1}_{(-\infty,b]}(x) - \mathbb{1}_{(-\infty,a]}(x)$

Proof:

$$\mathbb{1}_{(a,b]}(x) = \begin{cases} 0 & \text{if } x \notin (a,b] \\ 1 & \text{if } x \in (a,b] \end{cases}$$

$$\mathbb{1}_{(-\infty,b]}(x) - \mathbb{1}_{(-\infty,a]}(x) = \begin{cases} 0 & \text{if } x \notin (-\infty,b] \wedge x \notin (-\infty,a] \quad \vee \quad x \in (-\infty,b] \cap (-\infty,a] \\ 1 & \text{if } x \in (-\infty,b] \wedge x \notin (-\infty,a] \\ -1 & \text{if } x \notin (-\infty,b] \wedge x \in (-\infty,a] \end{cases}$$

Since $a < b$, we have

$$\begin{aligned} \mathbb{1}_{(-\infty,b]}(x) - \mathbb{1}_{(-\infty,a]}(x) &= \begin{cases} 0 & \text{if } x \notin (-\infty,b] \vee x \in (-\infty,a] \\ 1 & \text{if } x \in (-\infty,b] \wedge x \notin (-\infty,a] \end{cases} \\ &= \begin{cases} 0 & \text{if } x \notin (a,b] \\ 1 & \text{if } x \in (a,b] \end{cases} \end{aligned}$$

b. $\mathbb{1}_{\{a\}}(x) = \mathbb{1}_{(-\infty,a]}(x) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty,t]}(x)$

Proof:

$$\mathbb{1}_{\{a\}}(x) = \begin{cases} 0 & \text{if } x \notin \{a\} \\ 1 & \text{if } x \in \{a\} \end{cases}$$

$$\mathbb{1}_{(-\infty,a]}(x) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty,t]}(x) = \lim_{t \uparrow a} (\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x))$$

Because of what we already showed in **a**, we have that $\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x) = \mathbb{1}_{(t,a]}(x)$, so we can continue as follows:

$$\begin{aligned} \lim_{t \uparrow a} (\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x)) &= \lim_{t \uparrow a} \mathbb{1}_{(t,a]}(x) \\ &= \begin{cases} 0 & \text{if } x \notin \{a\} \\ 1 & \text{if } x \in \{a\} \end{cases} \\ &= \mathbb{1}_{\{a\}}(x) \end{aligned}$$

Now that we have shown these two manipulations (**a** and **b**), we can start proving the rules.

1. $h_n((a,b]) = F_n(b) - F_n(a)$

$$\begin{aligned} h_n((a,b]) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(a,b]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,b]}(x_i) - \mathbb{1}_{(-\infty,a]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,b]}(x_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,a]}(x_i) \\ &= F_n(b) - F_n(a) \end{aligned}$$

2. $h_n(\{a\}) = F_n(a) - F_n(a-)$

$$\begin{aligned} h_n(\{a\}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{a\}}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,a]}(x_i) - \lim_{t\uparrow a} \mathbb{1}_{(-\infty,t]}(x_i) \\ &= F_n(a) - F_n(a-) \end{aligned}$$

3. $h_n([a,b]) = F_n(b) - F_n(a) + h_n(\{a\})$

$$\begin{aligned} h_n([a,b]) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[a,b]}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,b]}(x_i) - \mathbb{1}_{(-\infty,a)}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,b]}(x_i) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,a]}(x_i) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{a\}}(x_i) \\ &= F_n(b) - F_n(a) + h_n(\{a\}) \end{aligned}$$

4. $h_n([a,b)) = F_n(b-) - F_n(a-)$

$$\begin{aligned} h_n([a,b)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[a,b)}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,b)}(x_i) - \mathbb{1}_{(-\infty,a)}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \lim_{t\uparrow b} \mathbb{1}_{(-\infty,t]}(x_i) - \lim_{t\uparrow a} \mathbb{1}_{(-\infty,t]}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,b)}(x_i) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,a)}(x_i) \\ &= F_n(b-) - F_n(a-) \end{aligned}$$

5. $h_n((a,\infty)) = 1 - F_n(a)$

$$\begin{aligned} h_n((a,\infty)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(a,\infty)}(x_i) \\ &= \lim_{t\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(x_i) - \mathbb{1}_{(-\infty,a]}(x_i) \\ &= \lim_{t\to\infty} F_n(t) - F_n(a) \\ &= 1 - F_n(a) \end{aligned}$$

T2. Convergence of cumulative distribution functions

A function $F : \mathbb{R} \to [0, 1]$ is called a cumulative distribution function, if F is monotonically increasing and right continuous, and if we have $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$. Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of cumulative distribution functions, which converges uniformly to a function $F : \mathbb{R} \to \mathbb{R}$

1. Give the definition of uniform convergence. Recall from your calculus lecture notes the following result: Given a sequence of continuous functions that converges uniformly, then the limit function is continuous. Recall the proof of that statement.

2. Show that F is a cumulative distribution function.

3. Why is the latter result important for our approach to statistics?

Solution T2

1. We say that a sequence of functions f_n, defined on a common domain A, converges uniformly to a function f on A, if for any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$ and for all $x \in A$ we have $|f_n(x) - f(x)| < \epsilon$.

   Given a sequence of continuous functions that converges uniformly, then the limit function is continuous.

   Proof:

   Let $\epsilon > 0$, $x \in A$.

   Uniform convergence implies that there exists a $N \in \mathbb{N}$ such that $\forall x' \in A$ we have

   $$|f_N(x') - f(x')| < \frac{\epsilon}{3}$$

   Since f_N is continuous, $\exists\delta > 0$ such that $\forall x' \in A$ with $|x' - a| < \delta$, $a \in A$ we have

   $$|f_N(x') - f_N(a)| < \frac{\epsilon}{3}$$

   Let $a \in A$ with $|x - a| < \delta|$. Then, by the triangle inequality, we have

   $$\begin{aligned} |f(x) - f(a)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(a)| + |f_N(a) - f(a)| \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &\leq \epsilon \end{aligned}$$

2. We need to show that F is

   (a) monotonically increasing,

   (b) right continuous, and

(c) $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$.

Proof:

(a) *Monotonicity:* Let $a < b \implies \forall n \in \mathbb{N} \quad F_n(a) \leq F_n(b)$.

$$F(a) = \lim_{n\to\infty} F_n(a) \leq \lim_{n\to\infty} F_n(b) = F(b)$$

(b) *Right Continuity:*
To establish right continuity of F, we must show that for every $t \in \mathbb{R}$,

$$\lim_{s\downarrow t} F(s) = F(t).$$

Since F_n is right continuous, we have

$$\lim_{s\downarrow t} F_n(s) = F_n(t).$$

By taking the uniform limit as $n \to \infty$, we get

$$\lim_{s\downarrow t} F(s) = \lim_{s\downarrow t} \lim_{n\to\infty} F_n(s) = \lim_{n\to\infty} \lim_{s\downarrow t} F_n(s) = \lim_{n\to\infty} F_n(t) = F(t).$$

(c) *Boundary Conditions:* The uniform convergence of F_n to F also guarantees that the boundary conditions at $-\infty$ and $+\infty$ will be preserved. Specifically, we consider the limits:

$$\begin{aligned}\lim_{t\to-\infty} F(t) &= \lim_{t\to-\infty} \lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \lim_{t\to-\infty} F_n(t) \\ &= \lim_{n\to\infty} 0 = 0,\end{aligned}$$

and,

$$\begin{aligned}\lim_{t\to\infty} F(t) &= \lim_{t\to\infty} \lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \lim_{t\to\infty} F_n(t) \\ &= \lim_{n\to\infty} 1 = 1.\end{aligned}$$

These steps are justified because uniform convergence allows us to switch the order of limits for functions, and F_n satisfy the CDF boundary conditions by definition.

By confirming the monotonicity, right continuity, and boundary conditions, we have shown that F is a cumulative distribution function.

3. *Importance of Result for Statistics:*

ECDF mostly converges uniformly and with the above results we know that the limit is also a CDF.

So by doing experiments, we can learn something about the random mechanisms behind the experiments. This is because the ECDF converges to the true CDF for a sufficiently large number of samples (n).

T3. Symmetric cumulative distribution function

We call a continuous cumulative distribution function F symmetric in c, if $F(c+t) = 1-F(c-t)$ for all $t \geq 0$. F is called *symmetric*, if there is a c such that F is symmetric in c.

1. Show that $F(c+t) = 1-F(c-t)$ for all $t \geq 0$ implies the continuity of the cumulative distribution function F. (Why is it enough to prove left continuity?)

2. Assume that there is a continuous function $f : \mathbb{R} \to [0, \infty)$ such that

$$F(t) = \int_{-\infty}^{t} f(x)dx \quad \forall t \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

i.e. F has density f.

Remark: Continuity of f is not necessary for the definition. A necessary condition is measurability, a concept that is treated in measure theory.

Show that F is symmetric in c if and only if $f(c+t) = f(c-t)$ for all $t \geq 0$.

Remark: This result stays true for piecewise continuous f.

3. For a continuous cumulative distribution function F, a median of F is any real number x such that $F(x) = 1/2$. Why is a median an important parameter of a distribution? For any symmetric F, give an example of a median. May there be more than one median for a given F ?

4. Give a median of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Is it unique?

Solution T3

1. Show that $F(c+t) = 1-F(c-t)$ for all $t \geq 0$ implies the continuity of the cumulative distribution function F. (Why is it enough to prove left continuity?)

   As F is a CDF, it is right continuous. We need to show that it is also left continuous. We have to show that for any $x \in \mathbb{R}$ and any sequence