Empirical Distribution Function

Exercise 1. We first consider univariate quantitative data where $x_1, \ldots, x_n$ are n real observed values. We consider in the following examples variables obtained from the dataset 'diamonds' from the package 'ggplot2' in R.

Example 1. *The first following dataset gives the width of the twelve first diamonds.*

Width (mm) :	*3.98*	*3.84*	*4.07*	*4.23*	*4.35*	*3.96*	*3.98*	*4.11*	*3.78*	*4.05*	*4.28*	*3.90*

Example 2. *The following dataset gives the quality of the cut of the twelve first diamonds with the following correspondence :* 1 *meaning ideal,* 2 *premium,* 3 *very good,* 4 *good and* 5 *fair.*

Quality of the cut :	*1*	*2*	*4*	*2*	*4*	*3*	*3*	*3*	*5*	*3*	*4*	*1*

1. Compute the empirical mean and the median of the data set of Example 1.
2. Draw the empirical cumulative distribution function of the datasets of Examples 1 and 2.

Solution

1. mean : $\bar{x}_n = 4.0442$; median : $n = 12$, $\lceil 1/2 \times 12 \rceil = \lceil 6 \rceil = 6$; the ordered values are

$$\begin{array}{llllll} x_{(1)} = 3.78, & x_{(2)} = 3.84, & x_{(3)} = 3.90, & x_{(4)} = 3.96, & x_{(5)} = 3.98, & x_{(6)} = 3.98, \\ x_{(7)} = 4.05, & x_{(8)} = 4.07, & x_{(9)} = 4.11, & x_{(10)} = 4.23, & x_{(11)} = 4.28, & x_{(12)} = 4.35 \end{array}$$

   hence $x_{(6)} = 3.98$.

2. The empirical distribution function is a step function. For a n -sample, it jumps by $1/n$ at each point of the sample. For the two datasets we get the following figures :

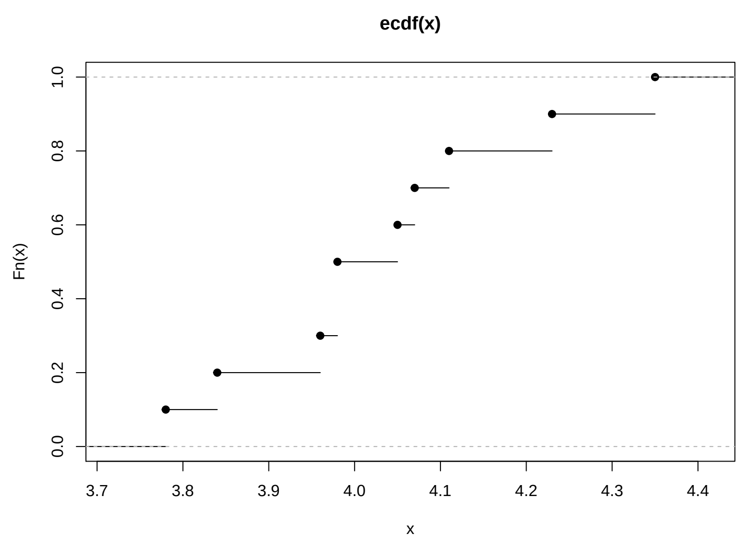


FIGURE 1 – Empirical Distribution Function, Example 1 (width of the diamonds)

Exercise 2. Empirical cumulative distribution function Let X_i be i.i.d. observations with c.d.f. F and $X_{1:n} = (X_1, \ldots, X_n)$.

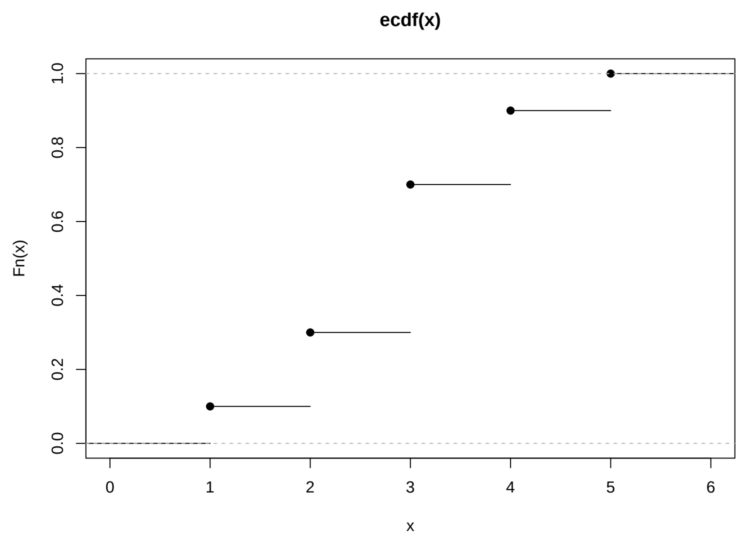


FIGURE 2 – Empirical Distribution Function, Example 2 (quality of the cut)

1. Show that for all $\alpha \in (0,1)$
$$X_\alpha(n) = \inf\{t \in \mathbb{R},\, \widehat{F}_{X_{1:n}}(t) \geqslant \alpha\} =: \widehat{F}^{-1}_{X_{1:n}}(\alpha),$$
where $\widehat{F}^{-1}_{X_{1:n}}$ is the generalized inverse of the empirical cumulative distribution function.
2. Fix $t \in \mathbb{R}$, what is the distribution of $n\widehat{F}_{X_{1:n}}(t)$? Can you complete the following limits :
$$\widehat{F}_{X_{1:n}}(t) \xrightarrow[n\to\infty]{F-\text{proba}} ?? \quad \text{and} \quad \sqrt{n}\left(\widehat{F}_{X_{1:n}}(t) - ??\right) \xrightarrow[n\to\infty]{F-\text{dist.}} \mathcal{N}(0, ??)\,?$$

Solution

1. By definition, the generalized inverse of the empirical cumulative distribution function $\widehat{F}_{X_{1:n}}$, denoted by $\widehat{F}^{-1}_{X_{1:n}}$ takes the following value at the point α :
$$\widehat{F}^{-1}_{X_{1:n}}(\alpha) = \inf\{t \in \mathbb{R}; \widehat{F}_{X_{1:n}}(t) \geq \alpha\} = X_\alpha(n).$$

2. For a fixed $t \in \mathbb{R}$ we have :
$$n\widehat{F}_{X_{1:n}}(t) = \sum_{i=1}^{n} I_{\{X_i \leq t\}}.$$
Then, the random variable $n\widehat{F}_{X_{1:n}}(t)$ takes values at $\{0, \ldots, n\}$. For any $k \in \{0, \ldots, n\}$, we have :
$$\mathbb{P}(n\widehat{F}_{X_{1:n}}(t) = k) = \mathbb{P}(\{\cap_{i=1}^{k}\{X_{(i)} \leq t\}\} \cap \{X_{(k+1)} > t\}),$$
where $X_{(i)}$ are the order statistics $X_{(1)} \leq \ldots \leq X_{(n)})$: the first k smaller data points are less than or equal to t, and the $n-k$ following data points are greater than t. using the independence of X_i, we get :
$$\mathbb{P}(n\widehat{F}_{X_{1:n}}(t) = k) = \binom{n}{k}\mathbb{P}(\{X_1 \leq t\})^k\,\mathbb{P}(\{X_1 > t\})^{n-k} = \binom{n}{k}F(t)^k(1-F(t))^{n-k}.$$
Then, $n\widehat{F}_{X_{1:n}}(t)$ follows a binomial law of parameters n and $F(t)$. Moreover, as seen in class, we have the following two convergences :
$$\widehat{F}_{X_{1:n}}(t) \xrightarrow[n\to\infty]{\text{proba}} F(t) \quad \text{et} \quad \sqrt{n}\left(\widehat{F}_{X_{1:n}}(t) - F(t)\right) \xrightarrow[n\to\infty]{\text{dist.}} \mathcal{N}(0, F(t)(1-F(t))).$$

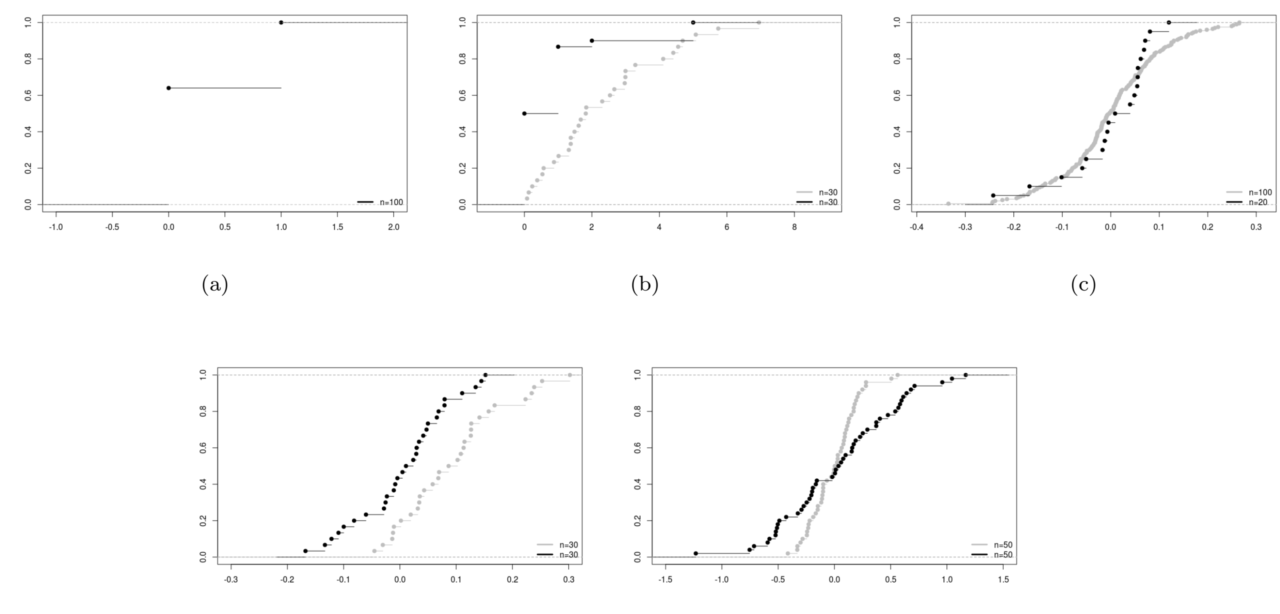
Exercise 3. Description of data and ecdf

1. Figure 3a represents the ecdf of some sample of size 100. Deduce the characteristics of the distribution of the sample and propose a distribution that is likely to have generated the data.
2. Each sub-figure 3(b-d) represents the ecdfs of two samples. For each sub-figure, compare the characteristics of the distributions of each sample.

Solution

(a) $X \sim$Bernoulli(p), $1-p = \mathbb{P}(X=0) \approx 0.65$, $p = \mathbb{P}(X=1) \approx 0.35$. Range $= 1$.

(b) — (b) : Both are discrete, black is concentrated on fewer points $X \in \{0, 1, 2, 5\}$; alternatively gray is continuous on $\mathbb{R}_{>0}$. Range $= 5$.
— (c) : Both are coming from normal distribution, with zero mean, gray finer 'resolution' due to the larger number of samples. Range $= 0.6$.
— (d) : Both are uniform with the same support size, black is concentrated on smaller numbers (=pdf shifted to the left).
— (e) : Both are coming from the same distribution, but the black has bigger variance.



(a) (b) (c)

(d) (e)

FIGURE 3 – ECDFs for some samples