Laborator 5

Funcția de repartiție și cuantilele empirice

Obiectivul acestui laborator este de a ilustra noțiunea de funcție de repartiție empirică și de cuantile empirice și de a verifica câteva proprietăți asimptotice ale acestora.

1 Funcția de repartiție empirică

Fie $X_1, X_2, \dots, X_n$ un eșantion de talie n dintr-o populație a cărei funcție de repartiție este F. Funcția de repartiție empirică este definită, pentru toate valorile $x \in \mathbb{R}$, prin

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_{(i)})$$

unde $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ reprezintă statisticile de ordine. Observăm că, notând $X_{(n+1)} = +\infty$, avem

$$\hat{F}_n(x) = \sum_{i=1}^{n} \frac{i}{n}\mathbf{1}_{\left[X_{(i)}, X_{(i+1)}\right)}(x).$$

> Dacă $\hat{F}_n(x)$ este funcția de repartiție empirică asociată unui eșantion de talie n, dintr-o populație a cărei funcție de repartiție este F, atunci, pentru $x \in \mathbb{R}$:
>
> - variabila aleatoare $n\hat{F}_n(x)$ este repartizată binomial $\mathcal{B}(n, F(x))$
> - are loc convergența (LNM): $\hat{F}_n(x) \xrightarrow{a.s.} F(x)$
> - are loc proprietatea de normalitate asimptotică (TLC): $\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x)))$.
>
> Ilustrați grafic rezultatele de mai sus pentru o populație repartizată $\mathcal{N}(0,1)$ și respectiv $\mathcal{E}(3)$. Pentru proprietatea de normalitate considerați $x_0 = 2$ și respectiv $x_0 = 1.5$.

Fie $x \in \mathbb{R}$ fixat și definim variabilele aleatoare $Y_i = \mathbf{1}_{(-\infty,x]}(X_i)$, $1 \leq i \leq n$. Cum $X_1, X_2, \dots, X_n$ sunt i.i.d. deducem că $Y_1, Y_2, \dots, Y_n$ sunt i.i.d. și în plus $Y_i \sim \mathcal{B}(p)$ cu $p = \mathbb{P}(Y_1 = 1) = F(x)$.

Din definiția funcției de repartiție empirică avem

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

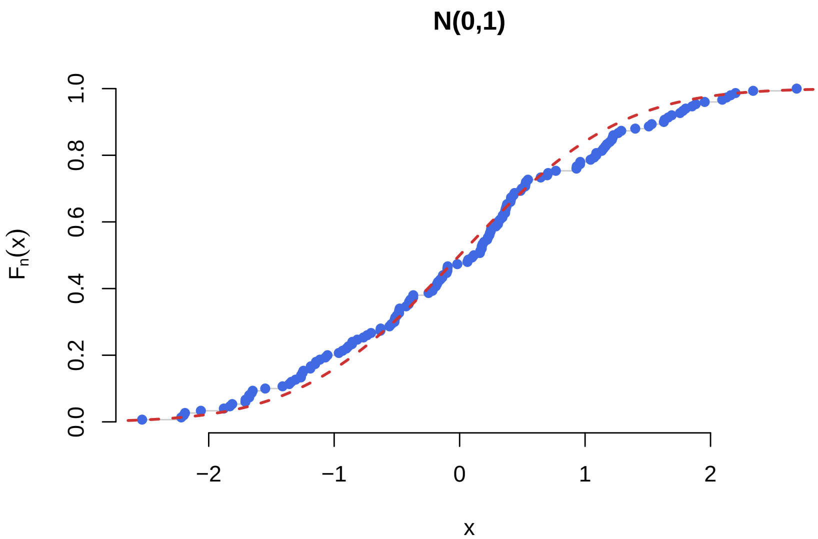
și aplicând *Legea Tare a Numerelor Mari* obținem

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[Y_1] = F(x).$$

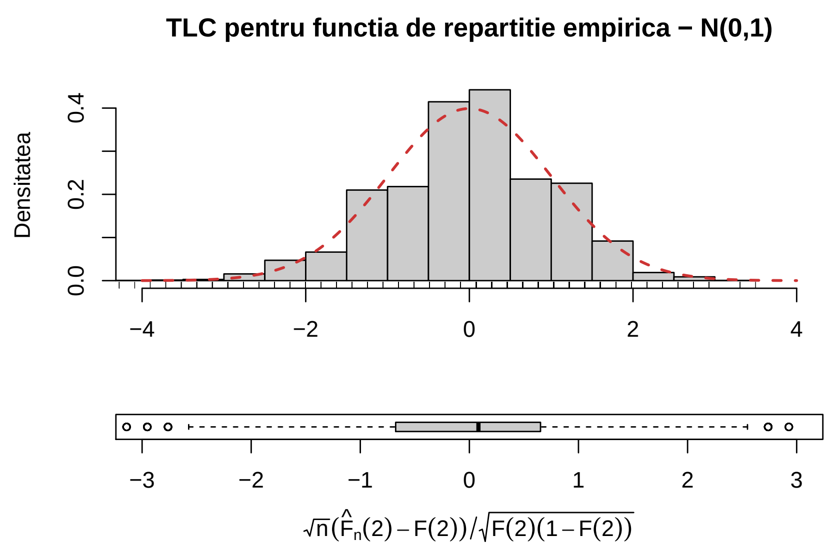
În mod similar aplicând *Teorema Limită Centrală* deducem

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, Var(Y_1)) = \mathcal{N}(0, F(x)(1 - F(x))).$$

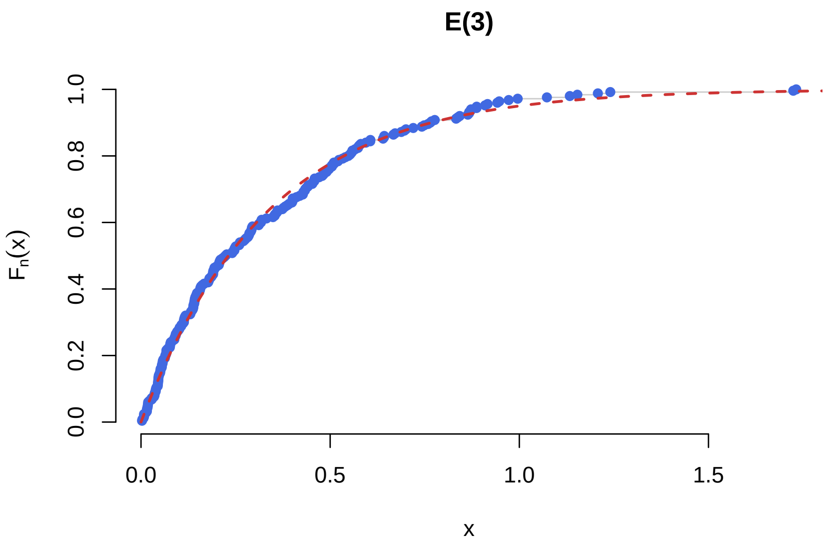
Pentru ilustrare, în cazul $\mathcal{N}(0,1)$ avem convergența



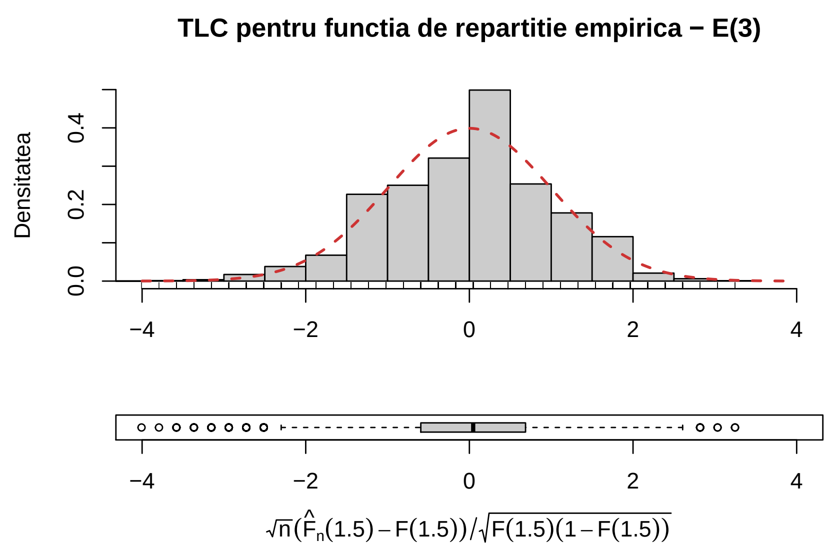
și proprietatea de normalitate (TLC)



Pentru repartiția $\mathcal{E}(3)$ avem



și rezultatul de normalitate asimptotică



Conform rezultatului anterior putem spune că $\hat{F}_n(x)$ este un estimator *rezonabil* pentru funcția de repartiție $F(x)$ dat fiind o valoare $x \in \mathbb{R}$ fixată. Întrebarea care se pune este dacă $\hat{F}_n(x)$ este un estimator *rezonabil* pentru întreaga funcție de repartiție $F(x)$? Răspunsul la această întrebare este dat de *Teorema Glivenko-Cantelli*[1] de mai jos:

Teorema Glivenko-Cantelli. Fie $(X_n)_n$ un șir de variabile aleatoare independent și identic repartizate, cu funcția de repartiție comună F. Atunci are loc

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

[1]Pentru o demonstrație a acestei teoreme se poate consulta, spre exemplu, cartea lui Sidney Resnick *A probability path*, Springer, 1998 (pag 224)

2 Cuantile empirice

Reamintim că dată fiind o funcție de repartiție F, funcția *cuantilă* (inversa generalizată) asociată lui F, $F^{-1} : (0,1) \to \mathbb{R}$ este definită prin

$$F^{-1}(u) = \inf\{x \in \mathbb{R} \,|\, F(x) \geq u\}, \quad \forall u \in (0,1)$$

unde folosim convențiile $\inf \mathbb{R} = -\infty$ și $\inf \emptyset = +\infty$.

Funcția cuantilă F^{-1} verifică următoarele proprietăți:

1) Valoarea în 0: $F^{-1}(0) = -\infty$
2) Monotonie: F^{-1} este crescătoare
3) Continuitate: F^{-1} este continuă la stânga
4) Echivalență: pentru $\forall u \in [0,1]$ avem $F(x) \geq u \iff x \geq F^{-1}(u)$
5) Inversabilitate: $\forall u \in [0,1]$ avem $(F \circ F^{-1})(u) \geq u$. În plus
   a) dacă F este continuă atunci $F \circ F^{-1} = Id$ dar dacă nu este injectivă atunci există x_0 așa încât $(F^{-1} \circ F)(x_0) < x_0$
   b) dacă F este injectivă atunci $F^{-1} \circ F = Id$ dar dacă nu este continuă atunci există u_0 astfel că $(F \circ F^{-1})(u_0) > u_0$

Pentru a exemplifica punctul 5a, putem considera variabila aleatoare $X \sim \mathcal{U}[0,1]$ a cărei funcție de repartiție F este continuă dar nu injectivă și în plus $(F^{-1} \circ F)(2) = F^{-1}(1) = 1 < 2$. Pentru punctul 5b să considerăm variabilele aleatoare $Y \sim \mathcal{N}(0,1)$ și $B \sim \mathcal{B}(0.5)$ independente și să definim $X = BY$. Atunci funcția de repartiție a lui X verifică $F(0-) = \frac{1}{4}$ și $F(0) = \frac{3}{4}$, este injectivă dar nu și continuă în 0 și în plus avem $(F \circ F^{-1})(1/2) = F(0) = \frac{3}{4} > \frac{1}{2}$.

Se numește *cuantilă* de ordin $p \in (0,1)$ (sau p-cuantilă) asociată lui F valoarea

$$x_p = F^{-1}(p) = \inf\{x \in \mathbb{R} \,|\, F(x) \geq p\}.$$

Cuantila de ordin 0.5, $x_{\frac{1}{2}}$ se numește mediana lui F și se notează cu M sau Q_2, iar cuantilele de ordin $\frac{1}{4}$ și respectiv $\frac{3}{4}$ se numesc prima și respectiv a treia cuartilă și se notează cu Q_1 și respectiv Q_3.

Fie acum $X_1, X_2, \ldots, X_n$ un eșantion de talie n dintr-o populație a cărei funcție de repartiție este F și fie $\hat{F}_n$ funcția de repartiție empirică asociată. Pentru $p \in (0,1)$ definim cuantila empirică de ordin p și o notăm $\hat{x}_p = \hat{x}_p(n)$ valoarea

$$\hat{x}_p = \hat{F}_n^{-1}(p) = \inf\{x \in \mathbb{R} \,|\, \hat{F}_n(x) \geq p\}.$$

Folosind convenția $X_{(0)} = -\infty$, cunatila empirică de ordin p coincide cu una dintre statisticile de ordine:

$$\hat{x}_p = X_{(i)} \iff np \leq i < np + 1 \iff \hat{x}_p = X_{(\lceil np \rceil)},$$

unde $\lceil x \rceil$ reprezintă cea mai mică valoare întreagă mai mare sau egală cu x.

Are loc următorul rezultat[2]:

[2] O demonstrație a acestui rezultat care nu necesită funcții caracteristice se regăsește în articolul lui Jan Wretman *A Simple Derivation of the Asymptotic Distribution of a Sample Quantile*, Scand. J. Statist., 5(2): 123-124, 1978.

Fie $X_1, X_2, \dots, X_n$ un eșantion de talie n dintr-o populație cu funcția de repartiție F, $p \in (0,1)$ fixat, x_p cuantila de ordin p asociată lui F și $\hat{x}_p(n)$ cuantila empirică de ordin p. Atunci

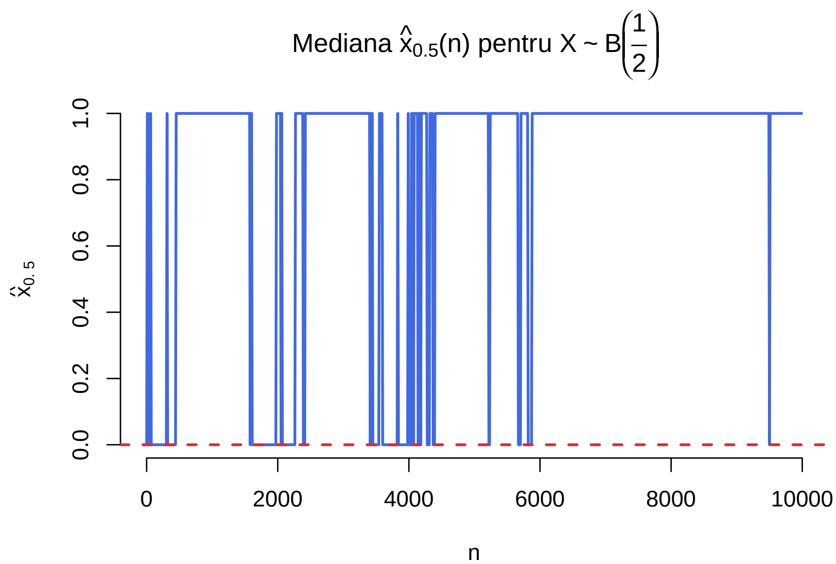
1) Convergența: dacă F este strict crescătoare în x_p are loc

$$\hat{x}_p(n) \xrightarrow[n\to\infty]{a.s.} x_p$$

2) Normalitatea asiptotică: dacă F este derivabilă în x_p cu derivata $f(x_p) > 0$, atunci

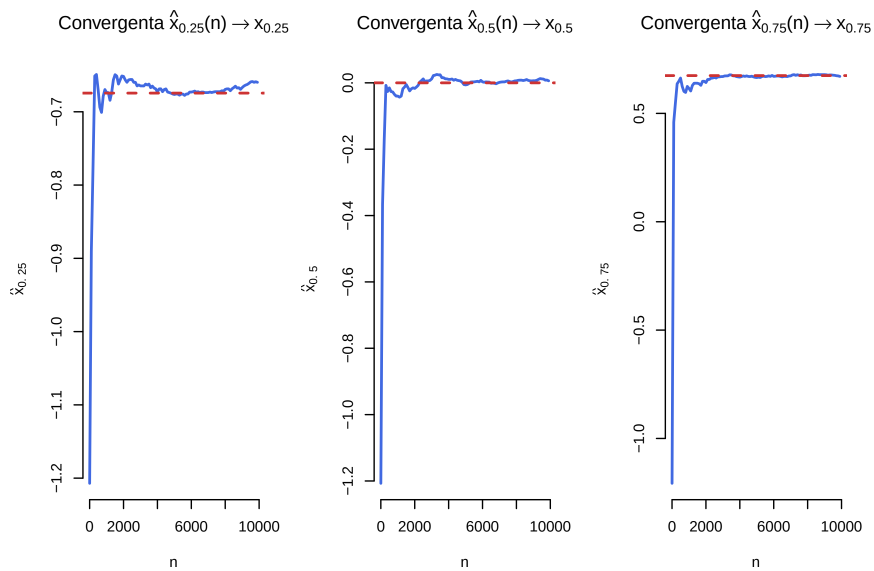
$$\sqrt{n}(\hat{x}_p(n) - x_p) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right).$$

Pentru a ilustra importanța condiției de la primul punct (F este strict crescătoare în x_p) să considerăm $X \sim \mathcal{B}(\frac{1}{2})$. Atunci mediana sa este $x_{\frac{1}{2}} = 0$ pe când mediana empirică $\hat{x}_{\frac{1}{2}}(n)$ va oscila mereu (dar neregulat) între valorile 0 și 1.

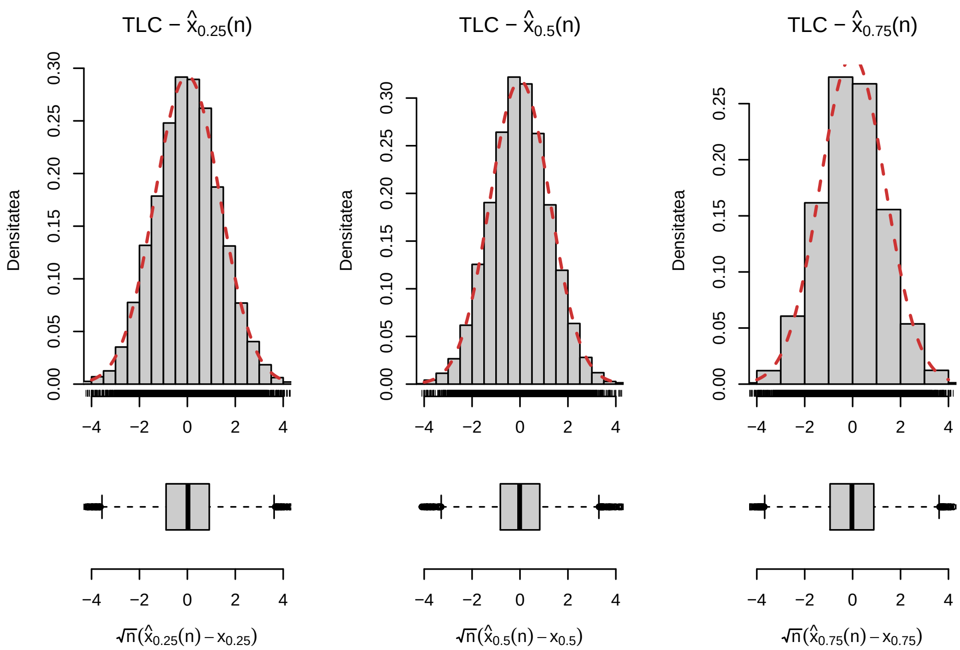


Ilustrați grafic în R proprietatea de convergență și de normalitate asiptotică (din rezultatul precedent) pentru o populație repartizată $\mathcal{N}(0,1)$ și respectiv $\mathcal{E}(3)$ și pentru $p \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$.

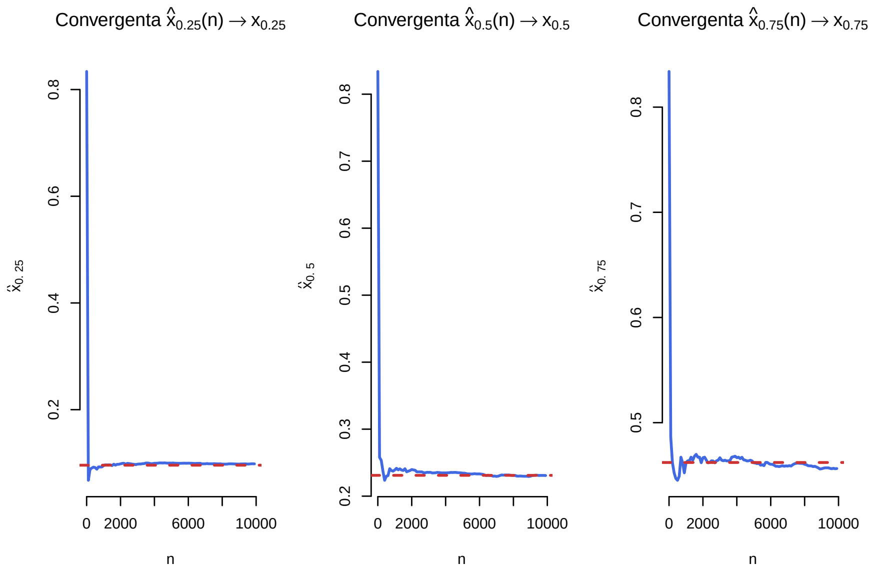
În cazul $\mathcal{N}(0,1)$ avem proprietatea de convergență a cuantilelor



și proprietatea de normalitate asimptotică



În cazul $\mathcal{E}(3)$ avem proprietatea de convergență a cuantilelor



și proprietatea de normalitate asimptotică