Introduction au statistique non-paramétrique Chap02 : Estimation non-paramétrique d'une fonction de répartition

Hamel Elhadj

Département de mathématiques
Université Hassiba Benbouali-Chlef
Ce cours est distiné aux étudiants Master2 mathématiques
Option :Mathématique Apliquées et statistique
2020/2021

Plan de la présentation

Chap02 : Estimation non-paramétrique d'une fonction de répartition

Introduction générale

L'estimation de la fonction de répartition d'une variable aléatoire est un volet important de l'estimation non paramétrique. De nombreuses méthodes ont été proposées et étudiées afin de modifier efficacement l'outil brut qu'est la fonction de répartition. Dans ce chapitre nous abordons l'estimation de la fonction de répartition et ses propriétés statistique .

Définition

La fonction de répartition F d'une v.a. X réelle est définie par :

$$\begin{aligned} F : \mathbb{R} &\to [0,1] \\ x &\to \mathbb{P}[X \leq x] \end{aligned} \tag{1}$$

Fonction de répartition (cumulative distribution function)

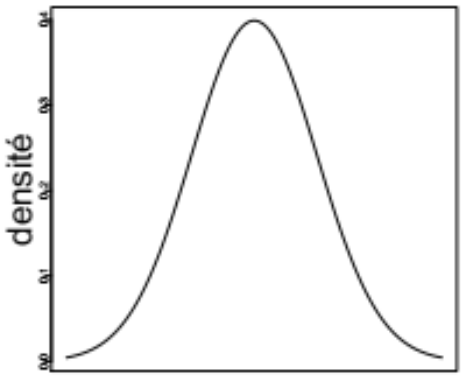
La fonction de répartition F d'une v.a. X réelle est définie par :

$$\begin{aligned} F : \mathbb{R} &\to [0,1] \\ x &\to \mathbb{P}[X \leq x] \end{aligned} \tag{2}$$

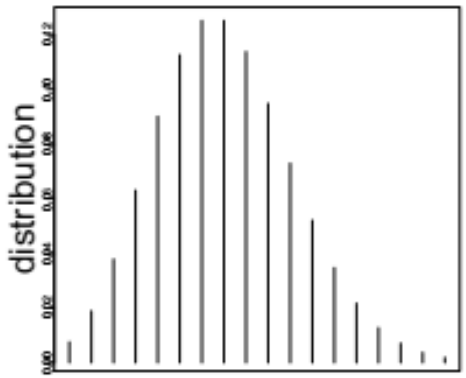
- F est croissante
- Si X est une v.a. continue et possède une densité f , alors F est dérivable et $F' = f$.
- F est définie pour toute variable aléatoire réelle, elle est cad-lag (continue à droite, dérivable à gauche)

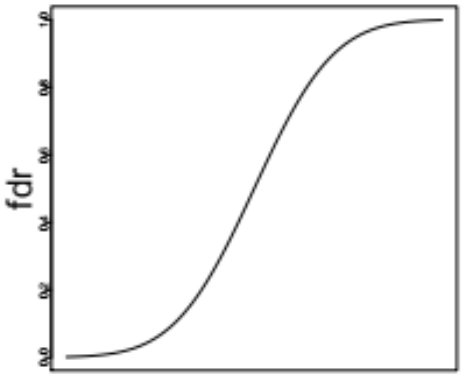
Exemples

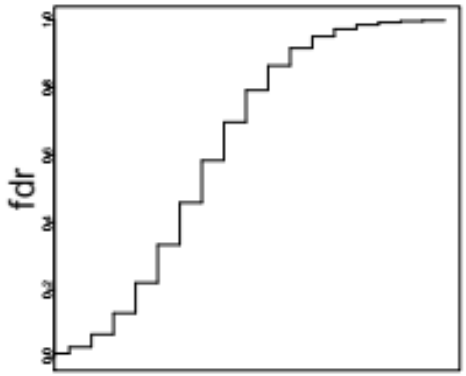
v.a. continue



v.a. discrète







La fonction de répartition

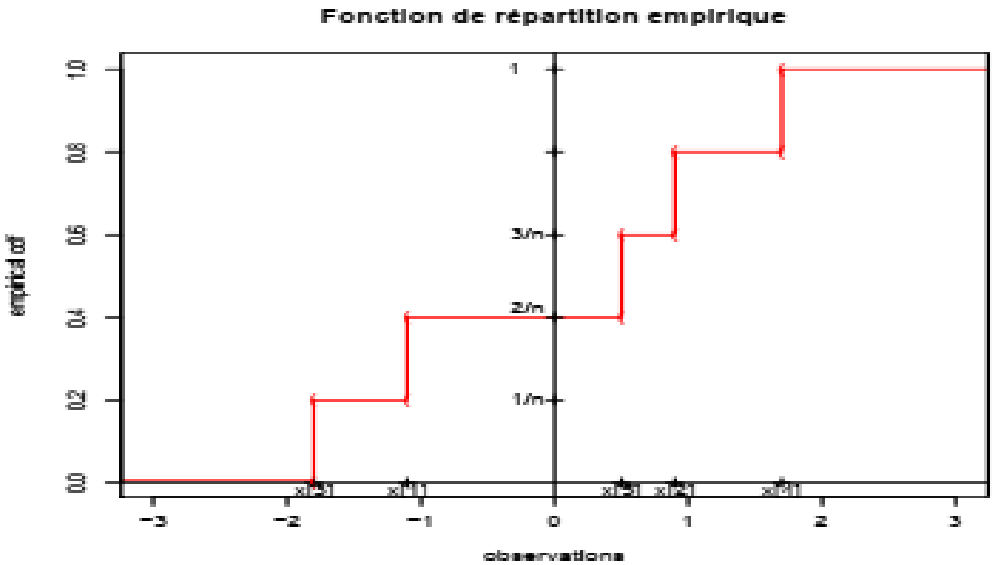
- Soit $X \sim F$, avec $F(x) = P\{X \leq x\}$ la fonction de répartition de X .
- Soit $X_1, X_2, ..., X_n$ un échantillon i.i.d. de F et $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ les observations ordonnées.
- Supposons que F soit complètement inconnue.
- Qt : Comment estimer F , en se basant sur les observations $X_1, ..., X_n$?

La fonction de répartition empirique

Un bon estimateur pour F est la fonction de répartition empirique, notée $\hat{F}_n$, et définie par :

$$\begin{aligned}
\hat{F}_n(x) &= \frac{\text{nombre d'observations } \leq x}{n} \\
&= \frac{\#\{i : X_i \leq x\}}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \\
&= \frac{1}{n}\sum_{i=1}^{n} I\{X_{(i)} \leq x\} \\
&= \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{k}{n} & \text{si } X_{(k)} \leq x < X_{(k+1)} \\ 1 & \text{si } x \geq X_{(n)} \end{cases}
\end{aligned} \tag{3}$$

La fonction de répartition empirique

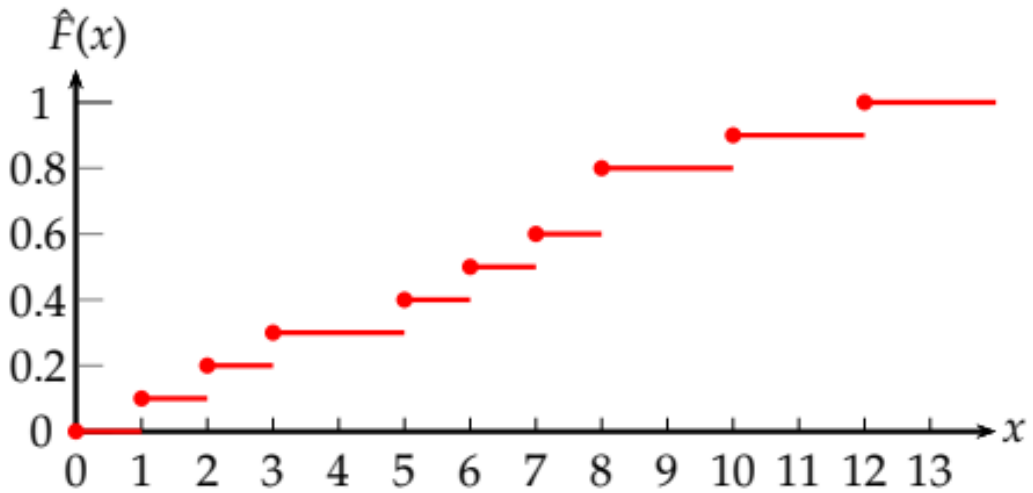


La fonction de répartition empirique

Exemple : soit les 10 réalisations suivantes : 8, 2, 6, 5, 3, 8, 10, 7, 1, 12.

Pour construire $\hat{F}_n$ à la main, le plus simple est évidemment d'ordonner les valeurs comme suit :

x	0	1	2	3	5	6	7	8	10	12	13
$\hat{F}(x)$	0/10	1/10	2/10	3/10	4/10	5/10	6/10	8/10	9/10	10/10	10/10



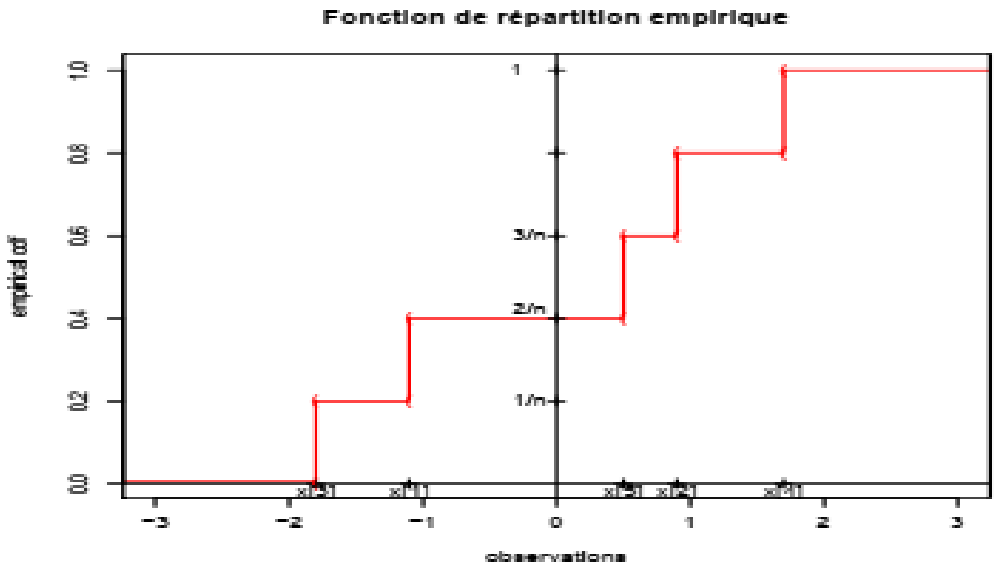
- Illustrer sous R ?

Estimer une fonction de répartition

On observe $X_1, ..., X_n$ variables aléatoires (v.a.) réelles, i.i.d. de fonction de répartition F.
L'estimateur naturel de la fdr F est la fdr empirique $\hat{F}_n$ définie par :

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} = \frac{1}{n}\#\{i, X_i \leq x\} \tag{4}$$

C'est un estimateur non paramétrique de la fdr F.



↪ Qualité de cet estimateur ?

Propriétés ponctuelles de $\hat{F}_n(x)$

- $\hat{F}_n$ est un estimateur sans biais de F :

$$\mathbb{E}\left[\hat{F}_n(x)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[1\mathrm{I}_{X_i\leq x}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left(X_i \leq x\right) = F(x) \tag{5}$$

- $\mathbb{V}\left[\hat{F}_n(x)\right] = \frac{F(x)(1-F(x))}{n}$.
- erreur en moyenne quadratique : $MSE = V[\hat{F}_n(x)] + biais[\hat{F}_n(x)]^2$

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{F}_n(x) - F(x)\right)^2\right] &= \mathbb{E}\left[\left(\hat{F}_n(x) - \mathbb{E}\left[\hat{F}_n(x)\right]\right)^2\right] \\
&= \mathrm{Var}\left(\hat{F}_n(x)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(1_{X_i\leq x}\right) \\
&= \frac{1}{n}\mathrm{Var}\left(1_{X_1\leq x}\right) = \frac{F(x)(1-F(x))}{n} \underset{n\to\infty}{\longrightarrow} 0
\end{aligned} \tag{6}$$

- La vitesse de convergence de $\hat{F}_n$ pour le risque quadratique en un point fixé est $1/n$

Rappel : Vitesse de convergence

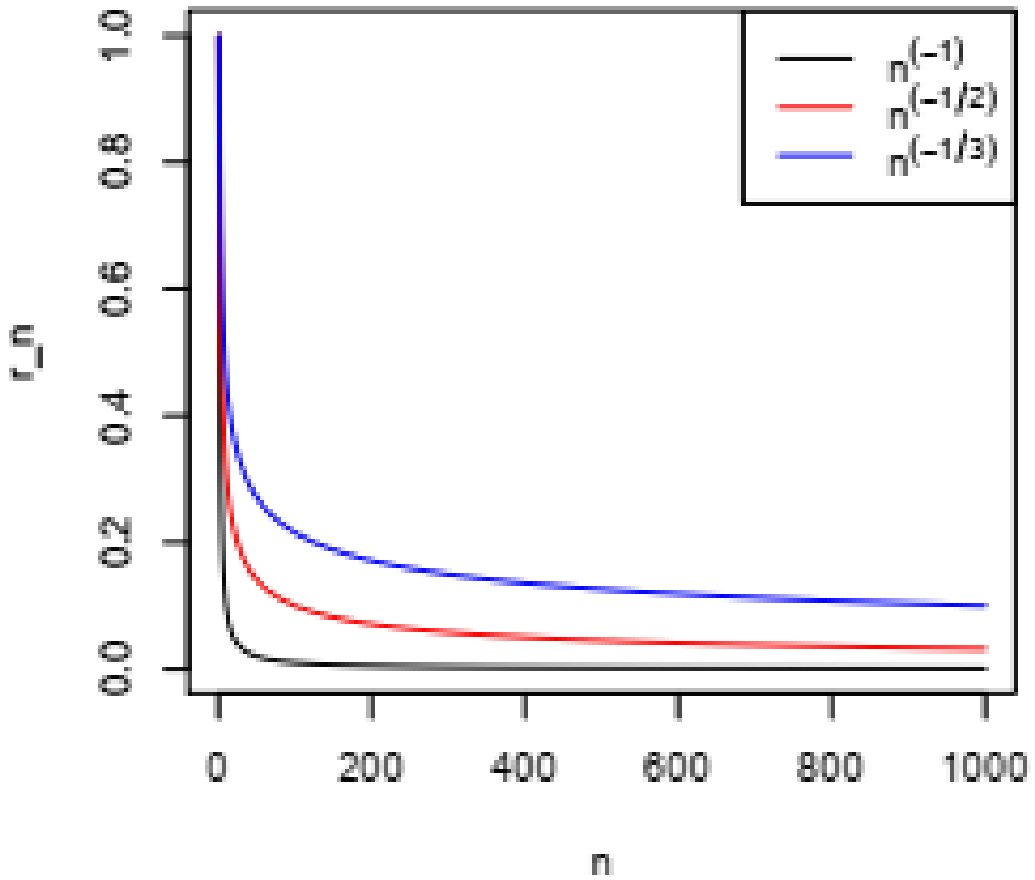
- $\hat{g}_n$ un estimateur d'une fonction g calculé à partir d'un échantillon de taille n
- d une distance fonctionnelle,
- $(r_n)_{n\in\mathbb{N}}$ une suite positive décroissante,

On dit que g^n converge vers g à la vitesse r_n pour la distance d si il existe une constante C telle que :

$$\mathbb{E}\left[d\left(\hat{g}_n(x) - g(x)\right)\right] \leq Cr_n$$

- $1/n$ vitesse de convergence des estimateurs paramétriques.
- En estimation non paramétrique, la vitesse est usuellement moins bonne, sauf pour la fdr.

Rappel : Vitesse de convergence



Propriétés de $\hat{F}_n$

- Asymptotiquement, l'erreur quadratique moyenne de $\hat{F}_n(x)$ est nulle, ce qui implique la convergence en probabilité de cet estimateur :

$$\forall \epsilon > 0, \Pr[|\hat{F}(x) - F(x)| > \epsilon] \underset{n\to\infty}{\to} 0 \tag{7}$$

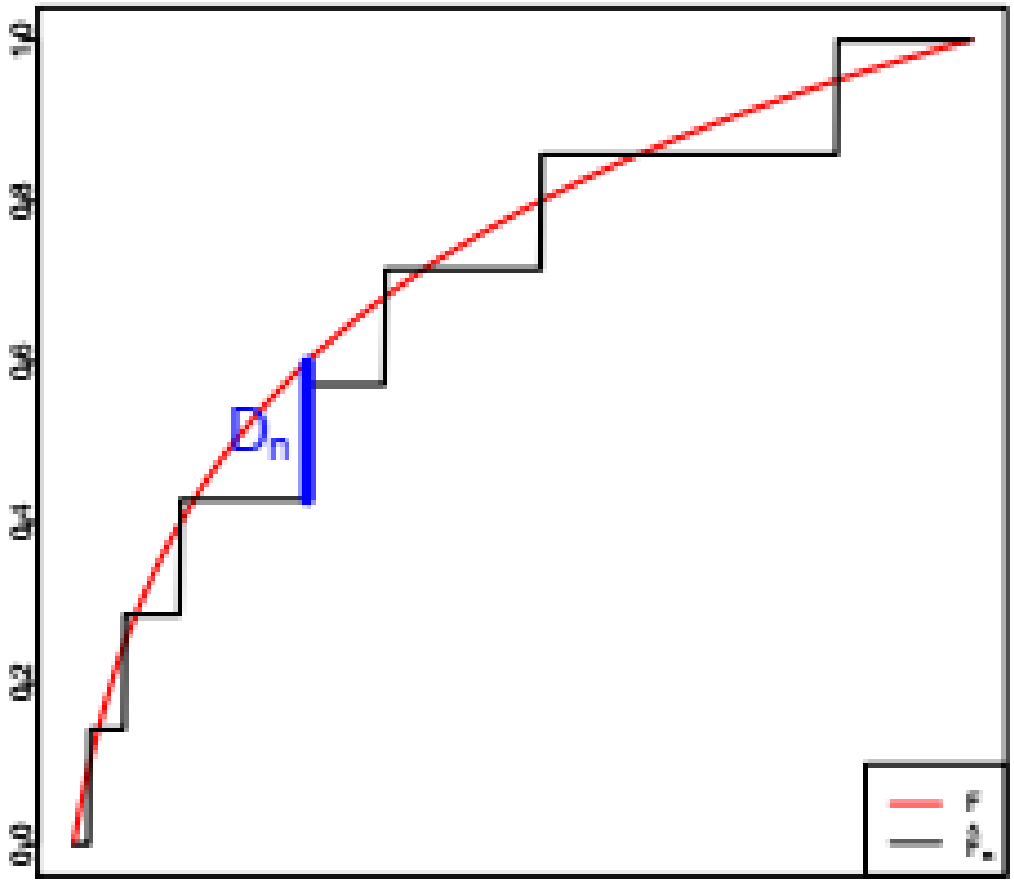
- La loi des grands nombres nous donne $\forall x \in \mathbb{R}, \qquad \hat{F}(x) \longrightarrow^{P}_{n\to\infty} F(x)$
- D'après le théorème central limite (TCL), on a :

$$\begin{aligned} &\frac{nF_n(x) - nF(x)}{\sqrt{nF(x)(1-F(x))}} \xrightarrow{L} \mathrm{N}(0;1) \\ \Longrightarrow \quad &\sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{L} \mathrm{N}(0; F(x)(1-F(x)) \end{aligned} \tag{8}$$

Propriétés uniformes de $\hat{F}_n$: (i.e. pour le sup sur x)

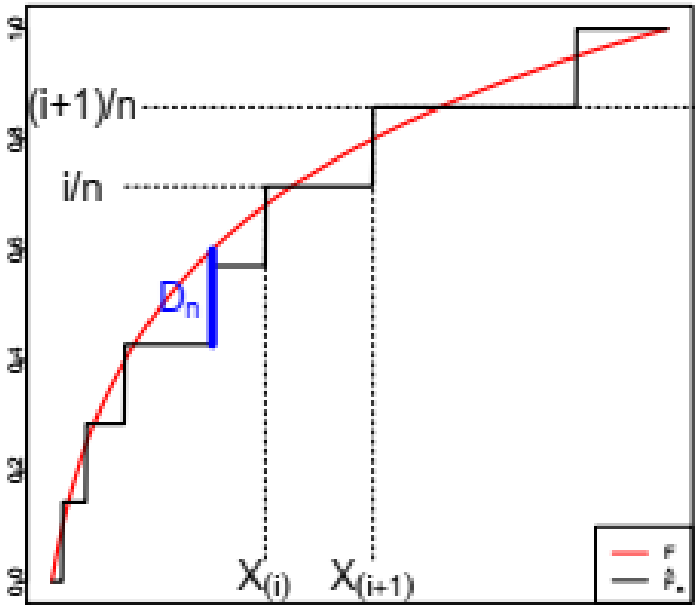
La distance de Kolmogorov-Smirnov

- Soit la distance $D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$.



Propriétés uniformes de $\hat{F}_n$: (i.e. pour le sup sur x)

- Soit la distance $D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$.



- D_n est calculable car le sup est néecessairement atteint en un point X_i.

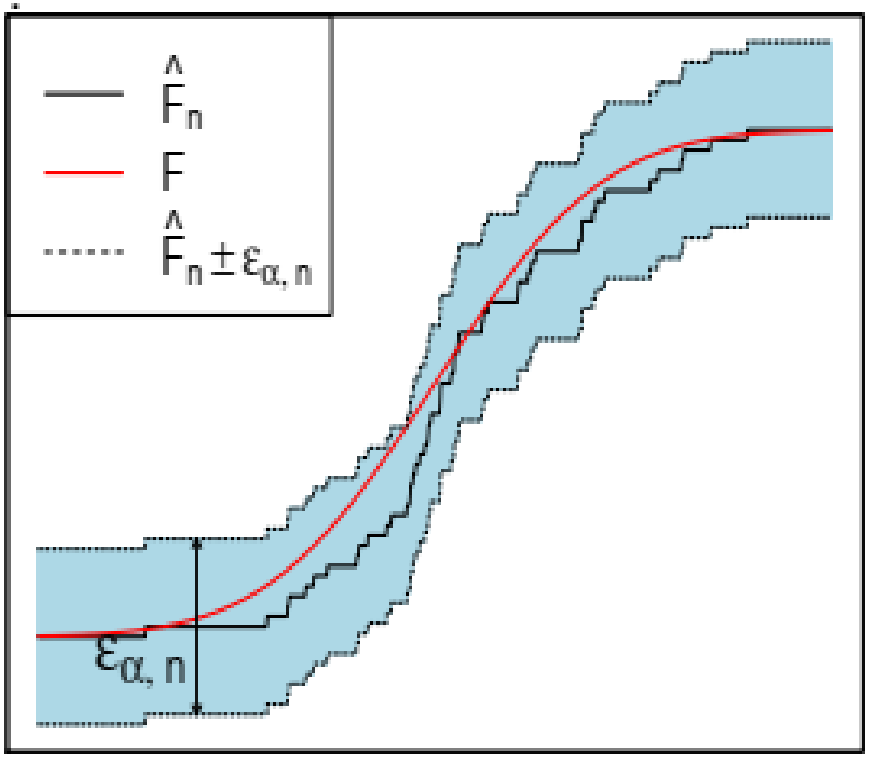
$$D_n = \max_{1\leq i\leq n} \left\{ \left| F\left(X_{(i)}\right) - \frac{i}{n} \right|, \left| F\left(X_{(i)}\right) - \frac{i-1}{n} \right| \right\} \tag{9}$$

Construction de bandes de confiance

Définition

$\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ *est une bande de confiance de niveau* α *pour* F *si :*

$$\mathbb{P}\left[D_n > \varepsilon_{\alpha,n}\right] = \mathbb{P}\left[F(x) \in \left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right], \forall x \in \mathbb{R}\right] \geq 1 - \alpha$$



Construction de bandes de confiance

- Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW) :

$$\forall n \in \mathbb{N}, \forall \varepsilon > 0, \quad \mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2} \tag{10}$$

- Résultat :

$$\begin{aligned}\mathbb{P}\left(F(x) \in \left[\hat{F}_n(x) - \varepsilon; \hat{F}_n(x) + \varepsilon\right]\right) &= 1 - \mathbb{P}\left(\left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \\ &\geq 1 - \mathbb{P}\left(\sup_x \left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \\ &\geq 1 - 2e^{-2n\varepsilon^2}\end{aligned} \tag{11}$$

- Pour un niveau de seuil $\alpha > 0$, soit $\varepsilon_{\alpha,n}$ tel que $2e^{-2n\varepsilon^2} = \alpha$, i.e. $\varepsilon = \sqrt{\log(2/\alpha)/(2n)}$. Alors $\left[\hat{F}_n - \varepsilon_{\alpha,n}, \hat{F}_n + \varepsilon_{\alpha,n}\right]$ est une bande de confiance de niveau α pour $F(x)$.

Construction de bandes de confiance avec TCL

$$\begin{aligned} &\frac{nF_n(x)-nF(x)}{\sqrt{nF(x)(1-F(x))}} \xrightarrow{L} \mathrm{N}(0;1) \\ \Longrightarrow \quad &\sqrt{n}\left(F_n(x)-F(x)\right) \xrightarrow{L} \mathrm{N}(0;F(x)(1-F(x)) \end{aligned} \tag{12}$$

Convergence en distribution

- Convergence vers la distribution de Kolmogorov : Quel que soit F, D_n converge en loi vers la distribution de Kolmogorov α_K indépendante de F

$$\mathbb{P}\left[D_n > \frac{c}{\sqrt{n}}\right] \underset{n \to \infty}{\longrightarrow} \alpha_K(c) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2c^2} \tag{13}$$

- Application : test d'adéquation à une distribution donnée
  - On dispose d'un échantillon $X_1, ..., X_n$ de fdr F et on veut savoir si $F = F_0$ avec F_0 la fdr d'une distribution de référence(connue).
  - L'hypothèse $H_0 : F = F_0$
  - Test : on calcule $\hat{F}_n$, et sous H_0, $\hat{F}_n - F_0$ suit approximativement la distribution 13.
- A partir de l'equation de kolmogorov, on ne peut pas construire une bande de confiance exacte mais seulement une bande de confiance asymptotique (qui peut être bien meilleure que la bande exacte !)

Convergence uniforme (théoreme de Glivenko Cantelli

(Ce théorème (que nous admettrons) est essentiel car il montre **la convergence uniforme**, pour tout $x \in \mathbb{R}$, de $\hat{F}_n$ vers F.

thèoréme

(Glivenko-Cantelli) Soit un échantillon aléatoire $X_1, X_2, ..., X_n$ issu de la loi de fonction de répartition F et $\hat{F}_n$ sa fonction de répartition empirique. Alors, quand $n \longrightarrow \infty$:

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{p.s} 0. \tag{14}$$

Excercice

Faire la preuve du théoreme ? ?

Remarque

Pour voir les choses concrètement, ce théorème nous dit que l'on peut etre assuré que l'écart maximal entre $\hat{F}_n$ et F va tendre vers 0 si l'on augmente la taille de l'échantillon à l'infini ou encore que partout, simultanément, la fonction de répartition empirique va se rapprocher de la vraie fonction de

La fonction quantile empirique $\hat{Q}_{n,p}$

- Le p^{eme} quantile (ou quantile d'ordre p) de la population

$$Q_p = F^{-1}(p) = \inf x : F(x) \geq p 0 < p < 1$$

- peut être estimé par

$$\hat{Q}_{n,p} = \hat{F}^{-1}(p) = \inf x : \hat{F}^{-1}(p) \geq p 0 < p < 1$$

- en utilise le théoreme de GC :

$$\hat{Q}_{n,p} \longrightarrow^{ps} Q_p$$

- Exemple :

Illustration sous R

Illustration sous R !