Chapitre 2: Statistique: généralités

Rabah Messaci

Département de Probabilités-Statistique
USTHB

Octobre 2011

Espaces de probabilités

En probabilité, un phénomène aléatoire est modélisé par un triplet $(\Omega, \mathfrak{A}, P)$ dit espace de probabilité.
Ω : ensemble des résultats possibles de l'expérience
$\mathfrak{A}$: ensemble des évènements
(tribu de parties de Ω)
P : mesure de probabilité unique et fixée
Exemples :

- Lancer d'une pièce de monnaie équilibrée : $(\{\pi, F\}, \mathcal{P}(\{\pi, F\}), B(\frac{1}{2}))$
- Note obtenue par un étudiant de ST choisi au hasard à l'épreuve de mathématiques $(R, \mathcal{B}_R, \{N(10, 16))$

Modèles statistiques

En statistique la loi de probabilité est inconnue. Cette modélisation se fait , à l'aide d'un modèle statistique :

Définition : Modéle statistique

On appelle modèle statistique un triplet $(\mathfrak{X}, \mathfrak{B}, \mathcal{P})$
$\mathfrak{X}$: ensemble des observations possibles de l'expérience
$\mathfrak{B}$: ensemble des évènements
(tribu de parties de $\mathfrak{X}$
$\mathcal{P}$: ensemble de mesures de probabilité

Exemples :

- $(\{\pi, F\}, \mathcal{P}(\{\pi, F\}), \{B(p)/p \in [0,1]\})$: Modéle associée à l'observation d'un lancer d'une pièce de monnaie
- $(R, \mathcal{B}_R, \{N(m, \sigma^2)/m \in R, \sigma^2 \in R_+^*\})^{(5)}$
  Modéle associée à l'observation de la note obtenue en mathématiques par un étudiant de ST choisi au hasard.

Modèles statistiques

Modèle statistique d'échantillon : $(\mathfrak{X}, \mathfrak{B}, \mathcal{P})^{(n)}$ associé à n observations d' une v.a X de loi dans des conditions d'indépendance. Exemples :

- $(\{\pi, F\}, \mathcal{P}(\{\pi, F\}), \{B(p)/p \in [0,1]\})^{(10)}$: Modéle associée à l'observation de 10 lancers d'une pièce demonnaie
- $(R, \mathcal{B}_R, \{N(m, \sigma^2)/m \in R, \sigma^2 \in R_+^*\})^{(5)}$
  Modéle associée à l'observation des notes obtenues par 5 étudiants de ST choisis au hasard.

Modèles statistiques

On distingue deux types de modèles statistiques :
modèles statistiques paramétriques :
$(\mathfrak{X}, \mathfrak{B}, (\mathcal{P}_\theta)_{\theta\in\Theta})$:
$(\mathfrak{X}, \mathfrak{B}, (f_\theta)_{\theta\in\Theta})$
modèles statistiques non paramétriques :
$(\mathfrak{X}, \mathfrak{B}, \{P/P \in \mathcal{P}\})$ $\mathcal{P}$: ensemble de toutes les lois de probabilités
$(\mathfrak{X}, \mathfrak{B}, \{F/F \in \mathcal{F}\})$ $\mathcal{F}$: ensemble de toutes les fonctions de répartition.
Conclusion :
Un modèle paramétrique est un modèle ou on fait l'hypothése que la loi de X appartient à une famille bien déterminée de lois, famille indéxée par un nombre fini de paramètres.
Un modèle non paramétrique est un modèle ou on ne fait aucune hypothése sur la loi de X ou une hypothèse trés large, par exemple que la loi est absolument continu.

Estimation

Définition : Statistique

On appelle statistique toute application mesurable définie sur $(\mathfrak{X}, \mathfrak{B}, (\mathcal{P}_\theta)_{\theta\in\Theta})$ à valeurs dans un espace mesurable $(\mathcal{Y}, \mathcal{C})$

$T : \mathfrak{X} \longrightarrow \mathcal{Y}$

$x \longrightarrow T(x)$

Notation : T,S,U...

Si $\mathcal{Y} = R$: statistique réelle

Si $\mathcal{Y} = R^p$: statistique vectorielle

Exemples de statistiques

Moments empiriques :

- $T_1(X_1, X_2, ..., X_n) = \overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$: moyenne empirique
- $T_2(X_1, X_2, ..., X_n) = \dfrac{\sum_{i=1}^{n} X_i^2}{n}$: moment empirique d'ordre 2
- $T_k(X_1, X_2, ..., X_n) = \dfrac{\sum_{i=1}^{n} X_i^k}{n}$: moment empirique d'ordre k
- $S_n^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}{n}$: variance empirique
- $S_n'^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}{n-1}$: variance empirique corrigée

Exemples de statistiques

Si on a deux échantillons de deux v.a X et Y

- $S^2_{X,Y} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})(Y_i - \overline{Y_n})}{n}$ covariance empirique
- $S'^2_{X,Y} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})(Y_i - \overline{Y_n})}{n-1}$: covariance empirique corrigée
- $\widetilde{\rho}_{X,Y} = \dfrac{S^2_{X,Y}}{S_X S_Y}$: coefficient decorrélation empirique

Statistiques d'ordre :

- $U_1(X_1, X_2, ..., X_n) = \min(X_i)_{1 \leq i \leq n} = X_{(1)}$: minimum
- $U_2(X_1, X_2, ..., X_n) = \max(X_i)_{1 \leq i \leq n} = X_{(n)}$: maximum
- $R(X_1, X_2, ..., X_n) = \max(X_i)_{1 \leq i \leq n} - \min(X_i)_{1 \leq i \leq n} = X_{(n)} - X_{(1)}$: étendue

Inférence statistique

Soit $(\mathfrak{X}, \mathfrak{B}, (P_\theta)_{\theta \in \Theta})$ un modèle statistique associé à l'observation d'une v.a X de loi P_θ. θ est inconnu et on cherche à le connaitre.
On appelle inférence statistique le processus qui à partir d'observations de X permettent d'avoir une connaissance de θ.
L'hypothèse de base est que les observations faites contiennent de l'information sur θ.
On distingue trois grands problèmes d'inférence statistique :
- Estimation ponctuelle
- Estimation par régions de confiance
-Tests d'hypothèses

Estimation

Soit $(\mathfrak{X}, \mathfrak{B}, (\mathcal{P}_\theta)_{\theta\in\Theta})$ un modèle statistique, on veut estimer θ ou une fonction $g(\theta)$ où g est une fonction définie sur Θ.

Définition : estimateur

On appelle estimateur de $g(\theta)$ toute statistique à valeurs dans $g(\Theta)$:
$T : \mathfrak{X} \longrightarrow g(\Theta)$

$x \longrightarrow T(x)$
$T(x)$ est dite estimation de $g(\theta)$

Propriétés d'un estimateur

Définition : estimateur sans biais

Un estimateur T de $g(\theta)$ est dit sans biais si

$$E_\theta(T) = g(\theta)$$

Un estimateur est donc sans biais si en moyenne, il est égal à la quantité qu'il estime.

Définition : estimateur asymptotiquement sans biais

Une suite d' estimateurs $(T_n)_{n\in N}$ de $g(\theta)$ est dite asymptotiquement sans biais si

$$\lim_{n\to+\infty} E_\theta(T_n) = g(\theta)$$

Estimation :illustration

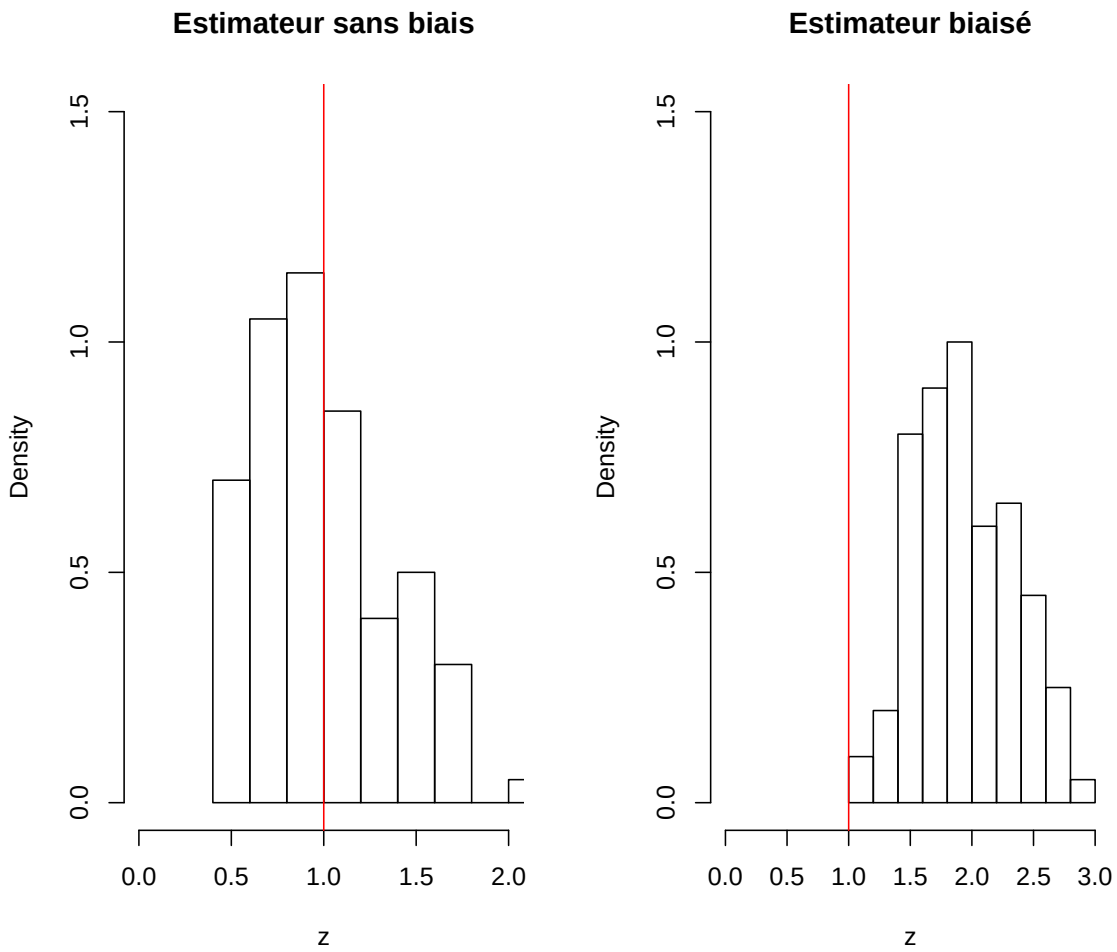


FIGURE: Estimateurs avec et sans biais

Propriétés d'un estimateur

Définition : estimateur convergent

Une suite d' estimateurs $(T_n)_{n\in N}$ de $g(\theta)$ est dite convergent en probabilité (respectivement p.s) si

$$T_n \underset{n\longrightarrow+\infty}{\overset{\text{Proba}}{\longrightarrow}} g(\theta)$$

(respectivement

$$T_n \underset{n\longrightarrow+\infty}{\overset{P.S}{\longrightarrow}} g(\theta)$$

)

Estimation :illustration

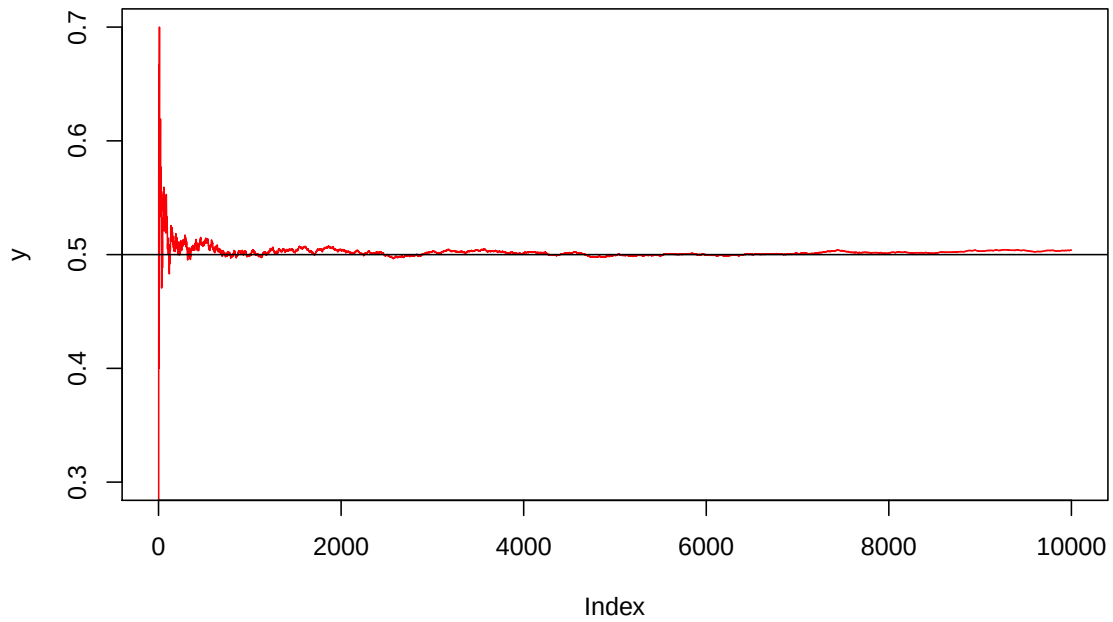


FIGURE: Estimateur convergent

Estimation

Théorème

Soit $X_1, X_2, .., X_n)$ un n-échantillon d'une v.a X, admettant un moment d'ordre k,

alors le moment emirique d'ordre k $m_k = \dfrac{\sum_{i=1}^{n} X_i^k}{n}$

est un estimateur sans biais et convergent presque sûrement de $E(X^k)$.

Démonstration : Rappels : Lois des grands nombres

Théorème : loi faible des grands nombres

Soit $(X_n)_{n \in N}$ une suite de v.a i.i.d admettant une espérance m finie. Alors

$$\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n} \underset{n \longrightarrow +\infty}{\overset{\text{Pr } oba}{\longrightarrow}} m$$

Estimation

Théorème : loi forte des grands nombres

Soit $(X_n)_{n\in N}$ une suite de v.a i.i.d admettant une espérance m finie. Alors

$$\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n} \underset{n\longrightarrow+\infty}{\overset{P.S}{\longrightarrow}} m$$

Le théorème découle par application immédiate de ces lois.

Estimation

Corollaire 1

$\overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ est un estimateur sans biais et convergent presque sûrement de $E(X)$.

Estimation

Corollaire 2

$S_n'^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}{n-1}$ est un estimateur sans biais et convergent presque sûrement de est un estimateur sans biais et convergent presque sûrement de $Var(X)$.

Démonstration :

On a

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}{n} = \frac{\sum_{i=1}^{n} {X_i}^2}{n} - \overline{X_n}^2$$

$$E(S_n^2) = E(\frac{\sum_{i=1}^{n} {X_i}^2}{n}) - E(\overline{X_n}^2) = E({X_i}^2) - E(\overline{X_n}^2)$$

Estimation

Démonstration (suite) : $E(\overline{X_n}^2) = E((\frac{\sum_{i=1}^{n} X_i}{n})^2) = E(\frac{\sum_{i=1}^{n} {X_i}^2}{n^2}) - E(\frac{\sum\sum_{i\neq j} X_i X_j}{n^2}) =$

$\frac{E({X_i}^2)}{n} - \frac{n(n-1)}{n^2} E(X_i X_j)$

$= \frac{E({X_i}^2)}{n} - \frac{(n-1)}{n} E(X_i)^2 \implies$

$E(S_n^2) = \frac{(n-1)}{n}(E({X_i}^2) - E(X_i)^2) = \frac{(n-1)}{n} Var(X) \implies$

$E(S_n'^2) = E(\frac{n}{n-1} S_n^2) = Var(X)$

D'autre part :

$S_n^2 = \frac{\sum_{i=1}^{n} {X_i}^2}{n} - \overline{X_n}^2 \underset{n\longrightarrow+\infty}{\overset{P.S}{\longrightarrow}} E(X^2) - E(X)^2 = Var(X)$

(Loi forte des grands nombres)

$\implies S_n'^2 = \frac{n}{n-1} S_n^2 \underset{n\longrightarrow+\infty}{\overset{P.S}{\longrightarrow}} Var(X)$

Estimation

Corollaire 2

$S'^2_{X,Y} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X_n})(Y_i - \overline{Y_n})}{n-1}$ est un estimateur sans biais et convergent presque sûrement de est un estimateur sans biais et convergent presque sûrement de $Cov(X, Y)$.

Démonstration identique que pour le corollaire 2.

Estimation

Corollaire 4

Soit A un évènement, lié à une espérience aléatoire, de probabilité $P(A)$.
$F_n(A) = \frac{\sum_{i=1}^{n} 1_{\{X_i \in A\}}}{n}$, le nombre de réalisations de A, lors de n répétitions de cet expérience dans des conditions indépendantes est dit fréquence de A.
$F_n(A)$ est un estimateur sans biais et convergent presque sûrement de $P(A)$.

Démonstration :

$\forall i$, $1_{\{X_i \in A\}} \backsim B(P(A))$ et d'après le corollaire 1, $\frac{\sum_{i=1}^{n} 1_{\{X_i \in A\}}}{n}$ est un estimateur sans biais et convergent de $P(A)$.

Estimation

Définition

Soit $(x_1, x_2, ..., x_n)$ un n-échantillon d'observations de X, on appelle fonction de répartition empirique de X la fonction F_n de R dans [0,1] définie par :

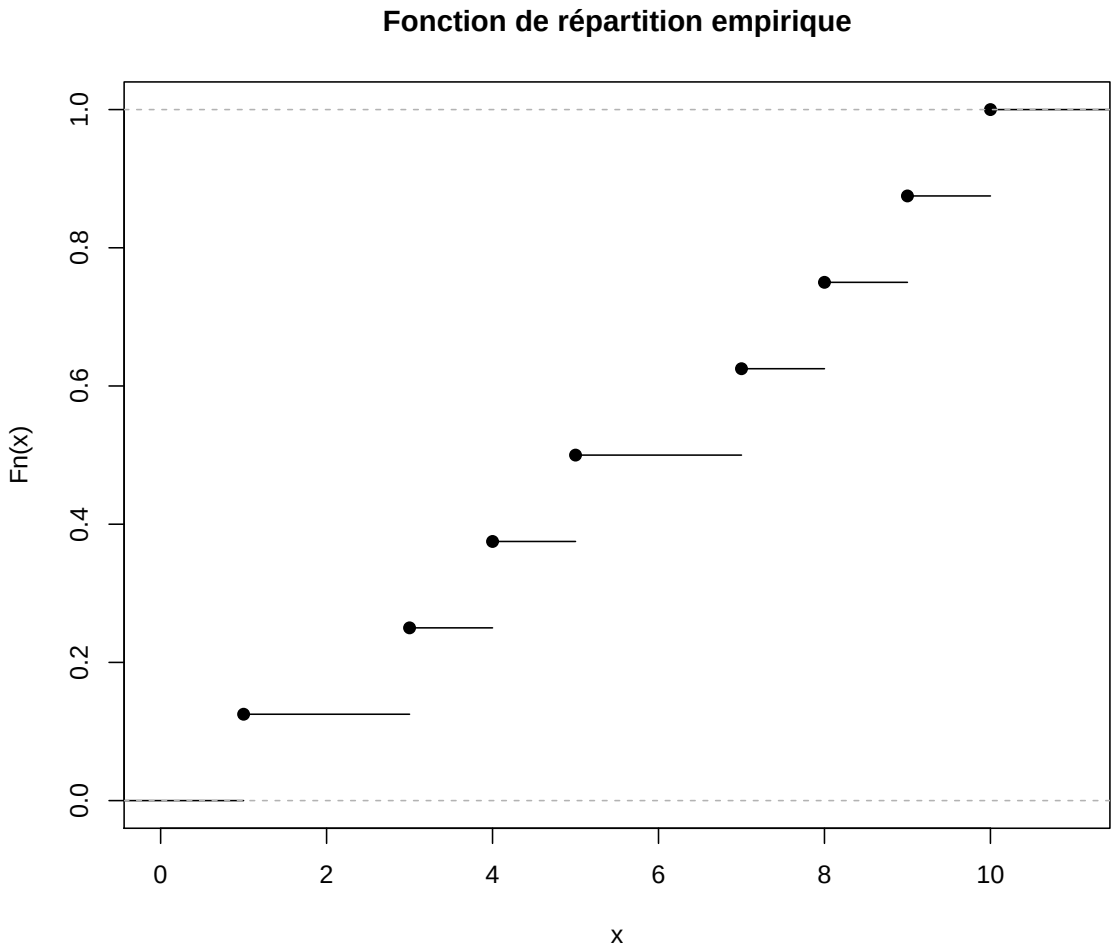
$$F_n(x) = \frac{\sum_{i=1}^{n} 1_{\{x_i \leq x\}}}{n}$$

Corollaire 5

$F_n(x) = \frac{\sum_{i=1}^{n} 1_{\{x_i \leq x\}}}{n}$ est un estimateur sans biais et convergent de $F(x)$, la fonction de répartition de X

Démonstration : $F(x) = P(X < x)$, le corollaire 4 permet de déduire le résultat.

Estimation


Fonction de répartition empirique
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

CV des f.r.empiriques : illustration

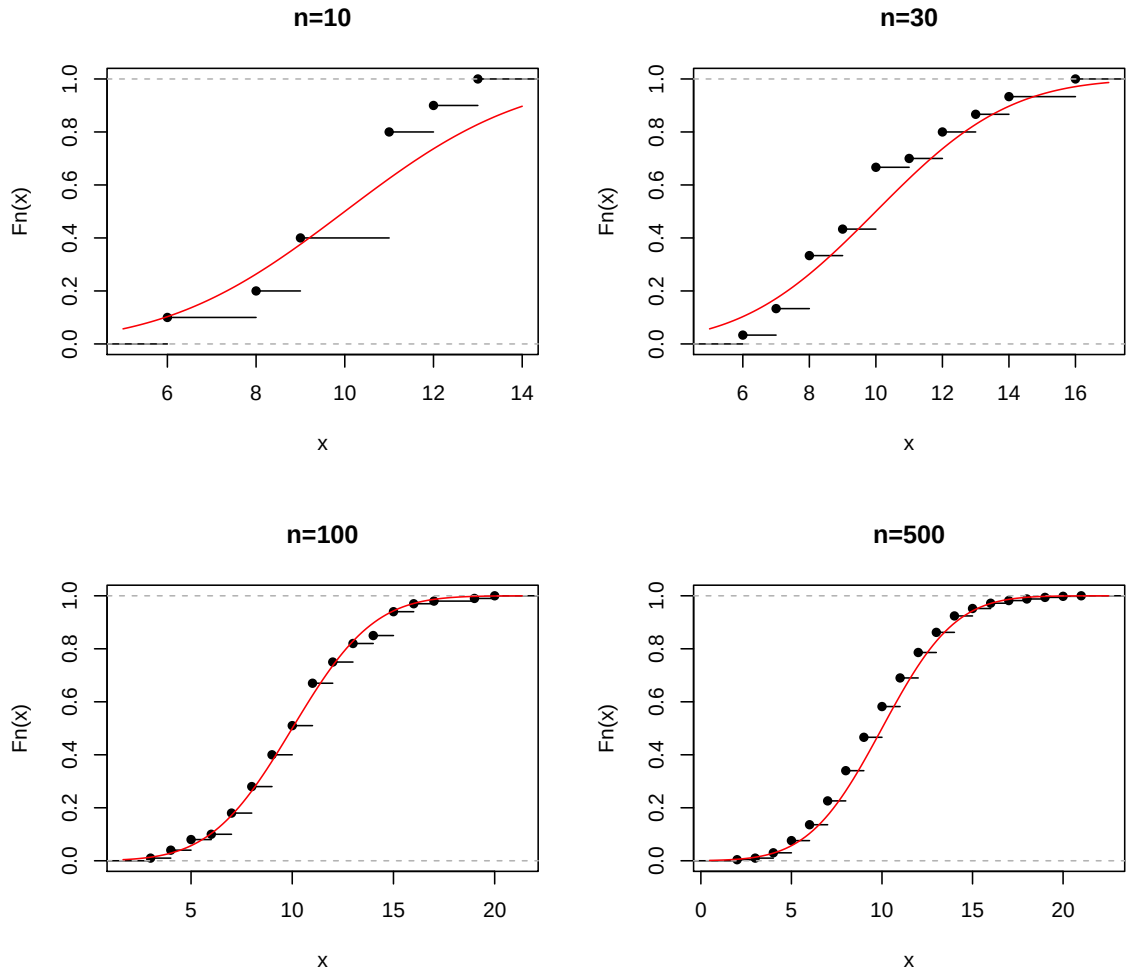


FIGURE: Evolution des f.r empiriques vers la f.r théorique