Le processus empirique et ses applications.

Stage dirigé par Anne Philippe, Université de Nantes

Baptiste Huguet

Mai-Juin 2014

Table des matières

Introduction

Lorsque l'on étudie un phénomène aléatoire, on est amené à rechercher la loi qui régit ce phénomène. En se donnant un échantillon, sommes-nous capables de retrouver la probabilité qui se cache derrière ?

On travaille dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On considère une famille de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}^*}$ indépendantes identiquement distribuées (*iid*), de fonction de répartition F. Pour tout $n \in \mathbb{N}$ et pour tout $t \in \mathbb{R}$, on définit la variable aléatoire $F_n(t)$ par :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i)$$

On appelle *processus empirique* associé à $(X_1, \ldots, X_n)$ la fonction $t \mapsto F_n(t)$.

Nous allons commencer par étudier cet objet à t et n fixés, puis nous allons nous intéresser aux propriétés de convergence à t fixé, avant d'étudier la famille de fonctions $(t \mapsto F_n(t))_{n\in\mathbb{N}^*}$. Après cette approche probabiliste, nous essayerons de le mettre en application pour faire de la statistique.

1 Étude ponctuelle de la fonction de répartition empirique

1.1 Étude à t et n fixés

Dans cette partie, on fixe $n \in \mathbb{N}^*$ et $t \in \mathbb{R}$. Pour tout $i \in [\![1, n]\!]$, on pose $Y_i = \mathbb{1}_{]-\infty,t]}(X_i)$.

Proposition 1 (Loi de Y_i)
Pour tout i dans $[\![1, n]\!]$, Y_i suit une loi $\mathcal{b}(F(t))$.

Démonstration 1
Soit i dans $[\![1, n]\!]$. Y_i est une variable aléatoire à support inclus dans $\{0, 1\}$. Elle suit donc une loi de *Bernoulli* $\mathcal{b}(\theta)$.
On a : $\theta = \mathbb{P}(Y_i = 1)$
$= \mathbb{P}(X_i \leq t)$
$= F(t)$

□

Proposition 2 (Indépendance des Y_i)
La famille $(Y_i)_{i \in [\![1,n]\!]}$ est iid.

Démonstration 2
On a montré dans la proposition précédente que les Y_i étaient identiquement distribuées.
Soit $(\alpha_1, \dots, \alpha_n)$ dans $\{0, 1\}^n$, on s'intéresse à $\mathbb{P}(Y_1 = \alpha_1, \dots, Y_n = \alpha_n)$. On peut réécrire cette quantité en fonction des X_i :

$$\mathbb{P}(Y_1 = \alpha_1, \dots, Y_n = \alpha_n) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n)$$

$$\text{avec} \quad A_i = \begin{cases}]-\infty, t] & \text{si} \quad \alpha_i = 1 \\ ]t, +\infty[& \text{si} \quad \alpha_i = 0 \end{cases}$$

Puis, par indépendance mutuelle de la famille $(X_i)_{i \in [\![1,n]\!]}$ on sépare en produit de probabilité ne dépendant que d'un X_i : $\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \dots \mathbb{P}(X_n \in A_n)$.
On termine en ré-exprimant à l'aide des Y_i. On a alors :

$$\mathbb{P}(Y_1 = \alpha_1, \dots, Y_n = \alpha_n) = \mathbb{P}(Y_1 = \alpha_1) \dots \mathbb{P}(Y_n = \alpha_n)$$

Ce qui montre l'indépendance mutuelle de la famille des $(Y_i)_{i \in [\![1,n]\!]}$

□

Théorème 3 (Loi de $F_n(t)$)
$nF_n(t)$ suit une loi binomiale $\mathcal{B}(n, F(t))$.

$$\forall\, k \in [0,n],\ \mathbb{P}(F_n(t) = \frac{k}{n}) = C_n^k F(t)^k (1-F(t))^{n-k}$$

$$\mathbb{E}[F_n(t)] = F(t) \quad \text{et} \quad \mathbb{V}\text{ar}(F_n(t)) = \frac{1}{n} F(t)(1-F(t))$$

Démonstration 3
$nF_n(t)$ est la somme de n variables aléatoires de loi $\mathcal{b}(F(t))$. Elle suit donc la loi binomiale décrite.

$\mathbb{E}[nF_n(t)] = n\mathbb{E}[F_n(t)] = nF(t)$ donc $\mathbb{E}[F_n(t)] = F(t)$.
$\mathbb{V}\text{ar}(nF_n(t)) = n^2\mathbb{V}\text{ar}(F_n(t)) = nF(t)(1-F(t))$ donc $\mathbb{V}\text{ar}(F_n(t)) = \frac{1}{n}F(t)(1-F(t))$

□

1.2 Convergence de $F_n(t)$ à t fixé

On se donne toujours un t fixe dans $\mathbb{R}$ et on étudie la suite de variable aléatoire $(F_n(t))_{n\in\mathbb{N}^*}$. Converge-t-elle en probabilité ? En norme L^p ? Presque sûrement ?

Théorème 4 (Convergence presque sûre de $(F_n(t))_{n\in\mathbb{N}^*}$)
La suite $(F_n(t))_{n\in\mathbb{N}^*}$ converge presque sûrement vers $F(t)$.

$$F_n(t) \overset{ps}{\to} F(t)$$

Démonstration 4
On utilise de nouveau la notation $Y_i = \mathbb{1}_{]-\infty,t]}(X_i)$. Les Y_i sont *iid* et suivent des lois de *Bernoulli*. Elles sont donc L^1 et on peut appliquer la *loi des grands nombres forte*. Ainsi $(F_n(t))_{n\in\mathbb{N}^*}$ converge presque sûrement vers $\mathbb{E}[Y_1] = F(t)$.

□

On en déduit donc que la suite $(F_n(t))_{n\in\mathbb{N}^*}$ converge donc aussi en probabilité et en loi vers $F(t)$ grâce aux relations entre les modes de convergence.

Proposition 5 (Convergence en norme p de $(F_n(t))_{n\in\mathbb{N}^*}$)
Pour tout p dans $[1, +\infty[$, on :

$$F_n(t) \overset{\|.\|_p}{\to} F(t)$$

Démonstration 5
Soit p dans $[1, +\infty[$. Comme on a déjà montré la convergence en probabilité, il suffit de montrer que la suite $(F_n(t)^p)_{n\in\mathbb{N}^*}$ est uniformément intégrable. Pour cela, on montre qu'il existe un réel $\delta > 0$ tel que $\sup_{n\geq 1} \mathbb{E}[|F_n(t)^p|^{1+\delta}] < \infty$.

$$\begin{aligned}\sup_{n\geq 1} \mathbb{E}[|F_n(t)^p|^{1+\delta}] &= \sup_{n\geq 1} \sum_{k=0}^{n} (\frac{k}{n})^{p(1+\delta)} \mathbb{P}(F_n(t) = \frac{k}{n}) \\ &\leq \sup_{n\geq 1} \sum_{k=0}^{n} (\frac{k}{n})^{p(1+\delta)} \\ &\leq 1 + \int_1^{+\infty} \frac{\mathrm{d}x}{x^{p(1+\delta)}}\end{aligned}$$

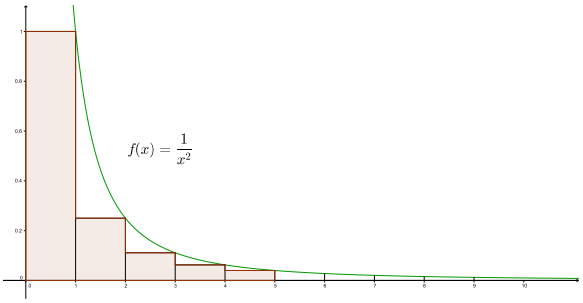

FIGURE 1 – Majoration de la somme par l'intégrale.

Cet intégrale converge si $p(1+\delta) > 1$, ce qui est toujours le cas. Donc il existe $\delta > 0$ tel que $\sup_{n\geq 1} \mathbb{E}[|F_n(t)^p|^{1+\delta}] < \infty$. La suite $(F_n(t)^p)_{n\in\mathbb{N}^*}$ est donc uniformément intégrable.
La fonction $x \mapsto x^p$ est continue et $F_n(t) \xrightarrow{ps} F(t)$. Donc $F_n(t)^p \xrightarrow{ps} F(t)^p$. Ainsi, elle converge aussi en probabilité. D'après le théorème de *Vitali*, la convergence a donc lieu en norme L^p.

□

Le cas de la convergence en norme 2 pouvait se faire plus simplement dès lors que l'on remarquait que la variance de $F_n(t)$ converge vers 0.

$$\begin{aligned}\text{Or} \quad \mathbb{V}\mathrm{ar}(F_n(t)) &= \mathbb{E}[(F_n(t) - \mathbb{E}[F_n(t)])^2] \\ &= \mathbb{E}[(F_n(t) - F(t))^2] \\ &= \|F_n(t) - F(t)\|_2^2 .\end{aligned}$$

Un autre critère intéressant est la vitesse de convergence.

Théorème 6 (Vitesse de convergence de $(F_n(t))_{n\in\mathbb{N}^*}$)
La convergence de la suite $(F_n(t))_{n\in\mathbb{N}^*}$ se fait en $\mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$ et on a :

$$\sqrt{n}(F_n(t) - F(t)) \Rightarrow \mathcal{N}(0, F(t)(1 - F(t))).$$

Démonstration 6
C'est une application du *théorème central limite* (TCL) à la suite $(Y_i)_{n\in\mathbb{N}^*}$. C'est une suite de variables aléatoire L^2, d'espérance $F(t)$ et de variance $F(t)(1 - F(t))$. D'où le résultat.

□

On utilise ici la notion de $\mathcal{O}_{\mathbb{P}}$, qu'il nous faut définir.

Définition 1 ($\mathcal{O}_{\mathbb{P}}(.)$)
Soient $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ deux suite de variables aléatoire.
On dit que $(X_n) = \mathcal{O}_{\mathbb{P}}(Y_n)$ si on a :

$$\forall\, \varepsilon > 0, \exists\, C > 0 \quad \text{telle que} \quad \sup_n \mathbb{P}(|X_n| > C\,|Y_n|) < \varepsilon.$$

Cette notion est très proche des $\mathcal{O}$ utilisés usuellement en analyse. Il vérifie des propriétés similaires. On peut notamment montrer que si $X_n \Rightarrow X$ alors $(X_n) = \mathcal{O}_{\mathbb{P}}(1)$.

2 Étude globale de la fonction de répartition empirique

On va maintenant traiter la dépendance en t. On se donne une réalisation $(X_n(\omega))_{n\in\mathbb{N}^*}$ avec $\omega \in \Omega$.

Théorème 7
Soit $n \in \mathbb{N}^*$, la fonction $t \mapsto F_n(t)$ est une fonction de répartition.

Démonstration 7
Soit $n \in \mathbb{N}^*$, on note $m = \min_{i\in[\![1,n]\!]}\{X_i(\omega)\}$ et $M = \max_{i\in[\![1,n]\!]}\{X_i(\omega)\}$. Ces quantités sont bien définies car on travaille sur un ensemble fini de n éléments.
Pour tout $t < m$, on a : $F_n(t) = 0$. Donc $\lim_{t\to-\infty} F_n(t) = 0$.
Pour tout $t \geq M$, on a : $F_n(t) = 1$. Donc $\lim_{t\to+\infty} F_n(t) = 1$.
Soit $i \in [\![1,n]\!]$, la fonction $t \mapsto \mathbb{1}_{]-\infty,t]}(X_i(\omega))$ est croissante et continue à droite et admet une limite à gauche en tout point (fonction cadlag). Une somme de fonctions cadlag étant encore une fonction cadlag, F_n est une fonction cadlag. Une somme de fonctions croissantes est croissante, donc F_n est une fonction croissante.
$F_n(t)$ est donc la fonction de répartition d'une variable aléatoire à support discret $\{X_1(\omega), \ldots, X_n(\omega)\}$.

□

2.1 Convergence uniforme, première approche

On a montré dans la partie précédente la convergence ponctuelle presque sûre des fonctions de répartition empiriques vers F. On peut montrer quelque chose de plus fort.

Commençons par rappeler la définition de la norme de la convergence uniforme.

Définition 2 ($\|.\|_\infty$)
On définit la norme infinie de f, une fonction définie sur $\mathbb{R}$, par :

$$\|f\|_\infty = \sup_{t\in\mathbb{R}} |f(t)|$$

Proposition 8 (Inégalité de Dvoretzky, Kiefer et Wolfowitz)
Il existe $C > 0$ telle que pour tout $\varepsilon > 0$ on ait : $\mathbb{P}(\|F_n - F\|_\infty > \varepsilon) \leq Ce^{-2n\varepsilon^2}$

La démonstration n'est pas accessible. On se doute qu'elle utilise des outils complexes, plus puissants que ceux à notre disposition.

Théorème 9 (Convergence uniforme presque sûrement de F_n)
$\|F_n - F\|_\infty$ converge presque sûrement vers 0.

Démonstration 8
D'après l'inégalité de Dvoretzky, Kiefer et Wolfowitz, pour tout $\varepsilon > 0$, on a :

$$\sum_{n=1}^{+\infty} \mathbb{P}(\|F_n - F\|_\infty > \varepsilon) < +\infty$$

D'après le lemme de *Borel-Cantelli* pour la convergence presque sûre, on a le résultat.

□

L'inégalité de DKW, qui rend la démonstration si simple, laisse une impression de miracle. On va s'en passer en utilisant une méthode plus technique.

2.2 Convergence uniforme, seconde approche

On va essayer de montrer les mêmes résultats avec des outils plus simples. Pour commencer, il faut comprendre ce qui nous empêche de passer de la convergence ponctuelle presque sûre à la convergence uniforme presque sûre.

Pour tout t, on sait que l'on dispose d'un événement certain $\Omega_t \subset \Omega$ tel que $\forall\ \omega \in \Omega_t, F_n^\omega(t) \to F(t)$. On recherche l'ensemble des ω tel que $F_n^\omega(t) \to F(t)$ pour tout t. C'est $\bigcap_{t\in\mathbb{R}} \Omega_t$. Le problème est donc de montrer que cet ensemble est un événement certain.

Proposition 10
Soit $(\Omega_n)_{n\in\mathbb{N}} \in \Omega^{\mathbb{N}}$ une suite d'événements certains ($\mathbb{P}(\Omega_n) = 1$ pour tout n), alors $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) = 1$.

Démonstration 9

$$\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) = 1 - \mathbb{P}(\bigcup_{n\in\mathbb{N}} \Omega_n^c)$$

Comme on travaille avec un nombre dénombrable d'événements, on peut utiliser la $\sigma-$sous-additivité de $\mathbb{P}$.

$$\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) \geq 1 - \sum_{n\in\mathbb{N}} \mathbb{P}(\Omega_n^c)$$

Puisque tous les Ω_n^c sont de probabilité nulle, on a : $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) \geq 1$. Donc $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) = 1$.

□

Pour utiliser la convergence ponctuelle presque sûre, la démonstration doit donc se ramener à un nombre dénombrable de t.

Théorème 11 (*Glivenko-Cantelli*)
On suppose que F est inversible. Dans ce cas F_n converge uniformément vers F presque sûrement :

$$\mathbb{P}(\lim_{n\to+\infty} \|F_n - F\|_\infty = 0) = 1$$

Démonstration 10
On note G la fonction quantile associée à F. Soit $N \in \mathbb{N}^*$, on définit :

$$\begin{cases} x_0 = -\infty \\ x_i = G(\frac{i}{N}) \quad \forall i \in [\![1, N-1]\!] \\ x_N = +\infty \end{cases}$$

Les x_i ainsi définis vérifient donc $F(x_i) = \frac{i}{N}$ pour $1 \leq i \leq n-1$.

Soit $x \in \mathbb{R}$. On dispose de i tel que $x \in]x_i, x_{i+1}]$. F_n et F étant des fonctions croissantes, on a :

$$F_n(x) - F(x) \leq F_n(x_{i+1}) - F(x_i) = F_n(x_{i+1}) - F(x_{i+1}) + \frac{1}{N}$$

$$F_n(x) - F(x) \geq F_n(x_i) - F(x_{i+1}) = F_n(x_i) - F(x_i) - \frac{1}{N}$$

Donc $|F_n(x) - F(x)| \leq \max(|F_n(x_{i+1}) - F(x_{i+1})|, |F_n(x_i) - F(x_i)|) + \frac{1}{N}$.
Cette dernière quantité converge presque sûrement vers $\frac{1}{N}$ d'après le résultat de convergence ponctuelle.

$$\|F_n - F\|_\infty \leq \max_{i\in[\![1,N]\!]}(|F_n(x_i) - F(x_i)|) + \frac{1}{N}$$

On pose $\Delta_n = \{\omega \in \Omega / \forall i \in [\![1, N-1]\!] \; F_n^\omega(x_i) \to F(x_i)\}$, c'est un événement certain car c'est une intersection finie d'événement certain.
Sur cet ensemble on a $\lim_{n\to+\infty} \|F_n^\omega - F\|_\infty \leq \frac{1}{N}$.

On définit enfin $\Delta = \bigcap_{n\in\mathbb{N}^*} \Delta_n$. C'est un événement certain et pour tout $\omega \in \Delta$,
$\|F_n^\omega - F\|_\infty$ converge vers 0. On a donc montré qu'il y avait convergence uniforme presque sûrement.

□

Dans le cas de la convergence uniforme, on remarque que les résultats sont plus difficiles à obtenir. On doit ajouter des hypothèse restrictives aux théorèmes. On pourrait aussi montrer que la vitesse se fait à la même vitesse que pour la convergence ponctuelle, mais il y a des différences notoires entre le cas où F est continue, et le cas où l'on travaille avec une loi discrète.

2.3 Lois continues et loi de *Kolmogorov*

On suppose que la fonction de répartition F est continue et strictement croissante. Cela signifie que F admet une inverse et que la loi des X_i admet une densité que l'on notera f.

Proposition 12
Pour tout $n \in \mathbb{N}^*$, pour tout $i, j \in [\![1, n]\!]$, on a : $X_i(\omega) = X_j(\omega) \Leftrightarrow i = j$ presque sûrement.

Démonstration 11 Soit $n \in \mathbb{N}^*$ et $i \neq j \in [\![1, n]\!]$.

$$\mathbb{P}(X_i(\omega) = X_j(\omega)) = \int_{\{(x,y)/x=y\}} f(x)f(y)\mathrm{d}x\mathrm{d}y$$

Cette probabilité est nulle car on intègre une fonction continue sur une droite, qui est de mesure nulle pour la mesure de *Lebesgue* λ^2.

□

Théorème 13 (Loi associée à F_n)
F_n est la fonction de répartition associée à la loi équi-répartie sur l'ensemble $\{X_1(\omega), \ldots, X_n(\omega)\}$.

Démonstration 12
Un petit graphique peut nous aider à voir les choses.

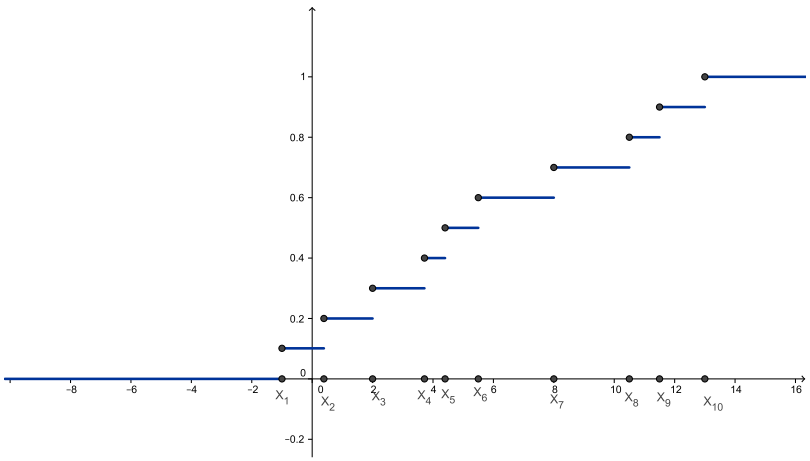

FIGURE 2 – Exemple de fonction de répartition empirique pour une loi continue.

On note U_n une variable aléatoire de fonction de répartition F_n. F_n admet n discontinuités en les X_i et chaque discontinuité est un saut de $\frac{1}{n}$ car les X_i sont tous distincts. Donc pour tout $i \in [\![1, n]\!]$, $\mathbb{P}(U = X_i) = \frac{1}{n}$.

□

La variable aléatoire définie par : $D_n = \|F_n - F\|_\infty$, converge presque sûrement vers 0. Afin de l'étudier, on introduit la loi de *Kolmogorov*.

Définition 3 (Loi de *Kolmogorov*)
On dit qu'une variable aléatoire X suit la loi de *Kolmogorov* si sa fonction de répartition est :

$$F(t) = \begin{cases} 0 & \text{si} \quad t \leq 0 \\ 1 + 2\sum_{n=1}^{+\infty} (-1)^n e^{-2n^2t^2} & \text{sinon} \end{cases}$$

La loi de *Kolmogorov* vérifie, $\mathbb{E}[X] \approx 0.87$ et $\mathbb{V}\text{ar}(X) \approx 0.26$. On donnera en annexe les procédures qui permettent d'avoir accès aux fonctions de répartition, de densité, ou aux quantiles.

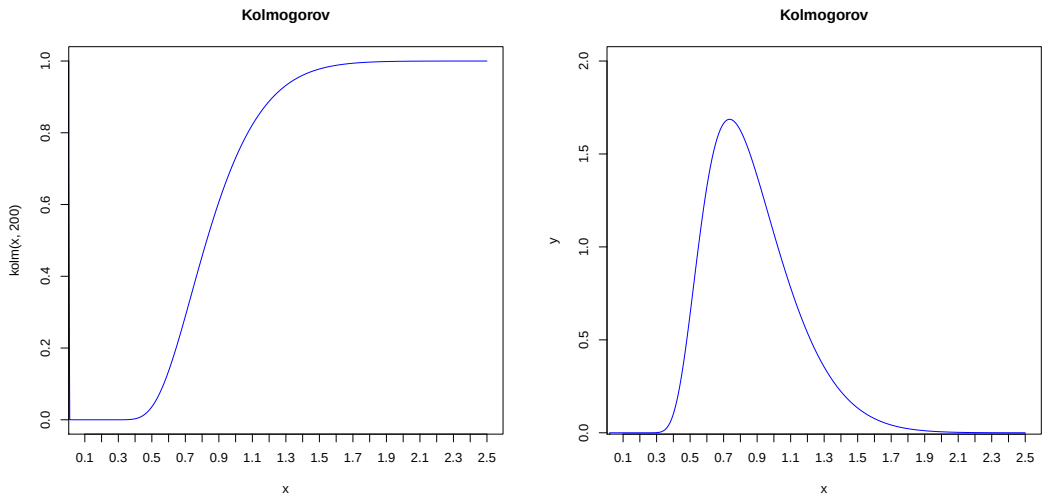

FIGURE 3 – fonction de répartition et densité de la loi de *Kolmogorov*.

Cette loi va nous permettre de préciser la vitesse de convergence de D_n vers 0.

Proposition 14 (Vitesse de convergence)
$\sqrt{n}D_n$ converge en loi vers une loi de Kolmogorov.

Les démonstrations de ce résultat sont très complexes et longues. L'une d'entre elles repose, en partie, sur le théorème 16 suivant qui permet de passer de la généralité des lois continues, au cas particulier de la loi uniforme.

Proposition 15
Soit X une variable aléatoire de loi F continue et strictement croissante. Alors $F(X)$ suit une loi $\mathcal{U}([0,1])$.

Démonstration 13
Comme la loi de F est continue et strictement croissante, son inverse existe.

$$\begin{aligned} \text{Soit } t \in \mathbb{R}, \quad \mathbb{P}(F(X) \leq t) &= \mathbb{P}(X \leq F^{-1}(t)) \\ &= F(F^{-1}(t)) \\ &= t \end{aligned}$$

Donc $F(X)$ suit bien la loi uniforme sur $[0,1]$.

□

Théorème 16 (Loi de D_n)
Si F est continue et strictement croissante, la loi de D_n ne dépend pas de F.

Démonstration 14
Pour tout $t \in \mathbb{R}$, on pose $u = F(t)$. $u \in [0,1]$.

$$D_n = \sup_{u \in [0,1]} \left| F_n(F^{-1}(u)) - u \right|$$

$$\begin{aligned} \text{Or} \quad F_n(F^{-1}(u)) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty, F^{-1}(u)]}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]0,1]}(F(X_i)). \end{aligned}$$

On reconnaît le processus empirique associé à la famille de variables aléatoires $(F(X_n))_{n \in \mathbb{N}^*}$ *iid* de loi $\mathcal{U}([0,1])$. Donc $F_n(F^{-1})$ ne dépend pas de F. Ainsi, K_n ne dépend pas de F.

□

Ce résultat est intéressant car si l'on travaille avec une loi continue, on peut supposer sans restriction que les $(X_n)_{n \in \mathbb{N}^*}$ suivent une loi uniforme (par exemple).

2.4 Lois à support discret

Dans le cas des lois à support discret, on ne peut plus montrer que les X_i sont deux à deux distincts. C'est même généralement faux. Les discontinuités de F_n ne sont plus toutes de la même taille. Ainsi F_n n'est plus la fonction de répartition d'une loi équi-répartie.

Par exemple, on suppose que les $(X_n)_{n\in\mathbb{N}^*}$ suivent une loi de *Bernoulli* $\mathcal{b}(p)$. Dans ce cas, $F = (1-p)\mathbb{1}_{[0,1[} + p\mathbb{1}_{[1,+\infty[}$ et $X_i \in \{0,1\}$.
On a alors

$$F_n(t) = \begin{cases} 0 & \text{si} \quad t < 0 \\ \frac{|\{i, X_i = 0\}|}{n} & \text{si} \quad 0 \leq t < 1 \\ 1 & \text{si} \quad t \geq 1 \end{cases}$$

A moins que $|\{i, X_i = 0\}|$ soit égal à $\frac{n}{2}$, on ne retrouve pas une fonction de répartition de loi équi-répartie.

Cependant, cela ne vaut pas dire que l'on ne peut pas caractériser la loi dont F_n est la fonction de répartition. On connaît son support et on peut donner les probabilités ponctuelles : $\forall\, i \in [\![1,n]\!], \mathbb{P}(U = X_i) = \frac{|\{j, X_j = X_i\}|}{n}$ (où U est une variable aléatoire dont la loi a pour fonction de répartition F_n).

En reprenant l'exemple de la loi de *Bernoulli*, on peut aussi montrer facilement que F_n converge uniformément vers F au sens usuel des fonctions :

$$\begin{aligned} \|F_n - F\|_\infty &= \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \\ &= |F_n(0) - F(0)| \\ &\longrightarrow 0 \end{aligned}$$

car la suite de fonctions converge ponctuellement presque sûrement. De plus, on obtient la même vitesse de convergence que pour l'erreur ponctuelle.

De manière générale, la convergence uniforme se fait à la même vitesse. Cependant la loi de la variable D_n dépende de la loi F choisie, et la loi limite n'est pas la loi de *Kolmogorov*.

3 Simulations

Les résultats démontrés dans les parties précédentes peuvent être illustrés par des simulations numériques. Nous utiliserons le logiciel $\mathcal{R}$, qui est le plus adapté pour manipuler des variables aléatoires ou réaliser des tests statistiques. Mais avant d'utiliser ce logiciel, nous allons nous intéresser aux générateurs pseudo-aléatoires comme ceux utilisés par $\mathcal{R}$.

3.1 Simuler une loi

Les premiers générateurs aléatoires nous viennent des jeux : les dés, la roulette ou le pile-ou-face. En théorie, un dé cubique réalise une loi équi-répartie sur $[\![1, 6]\!]$. Cependant, il est physiquement impossible de concevoir un dé parfaitement équilibré. Il est, en revanche plus aisé de favoriser une ou plusieurs faces... De plus, réaliser 1000 tirages successifs peut devenir très rébarbatif.

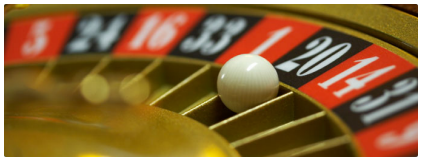

FIGURE 4 – La roulette de Monte Carlo, un des premiers générateurs aléatoires.

Un ordinateur est incapable de faire un choix au hasard. Il ne peut que réaliser des opérations déterministes. Il faut donc concevoir des algorithmes qui puissent approcher au mieux le hasard. Un théorème important permet, dans bien des cas, de se ramener à la simulation d'une loi $\mathcal{U}([0, 1])$.

Théorème 17
Soit U une variable aléatoire de loi $\mathcal{U}([0, 1])$, et F une fonction de répartition. On note G sa fonction quantile associée.
Alors la loi de $G(U)$ a pour fonction de répartition F.

Démonstration 15
Soit $t \in \mathbb{R}$, on a :

$$\begin{aligned} F(t) &= \mathbb{P}(U \leq F(t)) \\ &= \mathbb{P}(U < F(t)) \end{aligned}$$

Si $U < F(t)$ alors $G(U) < G(F(t) \leq t$. Donc $\{U < F(t)\} \subset \{G(U) \leq t\}$
Ainsi $F(t) \leq \mathbb{P}(G(U) \leq t)$

En même temps, par définition, $G(U) = \inf\{s \in \mathbb{R} / F(s) \geq U\}$.

Donc pour tout $s > t, \{G(U) \leq t\} \subset \{U < F(s)\}$.
On en déduit que :

$$\begin{aligned} \forall\, s > t, \mathbb{P}(G(U) \leq t) &\leq \mathbb{P}(U < F(s)) \\ &= \mathbb{P}(U \leq F(s)) \\ &= F(s) \end{aligned}$$

En faisant tendre s vers t, on montre que $\mathbb{P}(G(U) \leq t) = F(t)$.

□

Pour toutes les lois dont la fonction quantile est connue, on peut utiliser ce résultat. C'est le cas de la loi exponentielle, ou des lois discrètes usuelles.

3.2 Simuler une loi $\mathcal{U}([0,1])$

Définition 4 (Générateur pseudo aléatoire uniforme)
Un générateur pseudo aléatoire uniforme est un algorithme qui partant d'une donnée initiale u_0 et d'une transformation D, produit une suite $(u_i) = D(u_{i-1})$ de valeur dans [0,1].
Pour tout n, $(u_1, \ldots, u_n)$ doit reproduire le comportement d'une réalisation de la famille $(U_1, \ldots, U_n)$ *iid* selon la loi $\mathcal{U}([0,1])$ quand on le soumet aux tests usuels.

Pour simuler des nombres réels compris entre 0 et1, l'ordinateur doit simuler un suite (X_n) d'entiers entre 0 et $M (M \in \mathbb{N}^*)$ et diviser par M. Le générateur le plus simple est le générateur congruentiel, de la forme $X_i = aX_{i-1} + b[M]$. On peut montrer quels sont les choix de a, b et M les plus judicieux pour que le générateur soit le plus pertinent possible. Ce générateur est évidement périodique.

Un des tests utilisé pour contrôler la pertinence d'un générateur est graphique. On génère un échantillon de n données, un vecteur U, et on trace U_i en fonction de U_{i+1}. Si le générateur est mauvais, on voit apparaître des zones plus ou moins densément peuplées, ou des structures particulières. Par exemple, avec un générateur congruentiel on voit apparaître des droites. Sur cette illustration, les paramètres sont $a = 65, b = 1$ et $M = 2048$. Seul les 500 premiers couples ont été tracés. On remarque que les points sont ordonnés de manière très régulière. De plus la densité n'est pas du tout homogène : certain intervalle de [0,1] ne seront visités que tardivement.

```
 Z=NULL
Z[1]=1024
for (k in 1:2047){
  Z[k+1]=(65*Z[k]+1)%% 2048
}
Z=Z/2048
```

```
Y=c(Z[2:2048],Z[1])
plot(Z,Y)
```

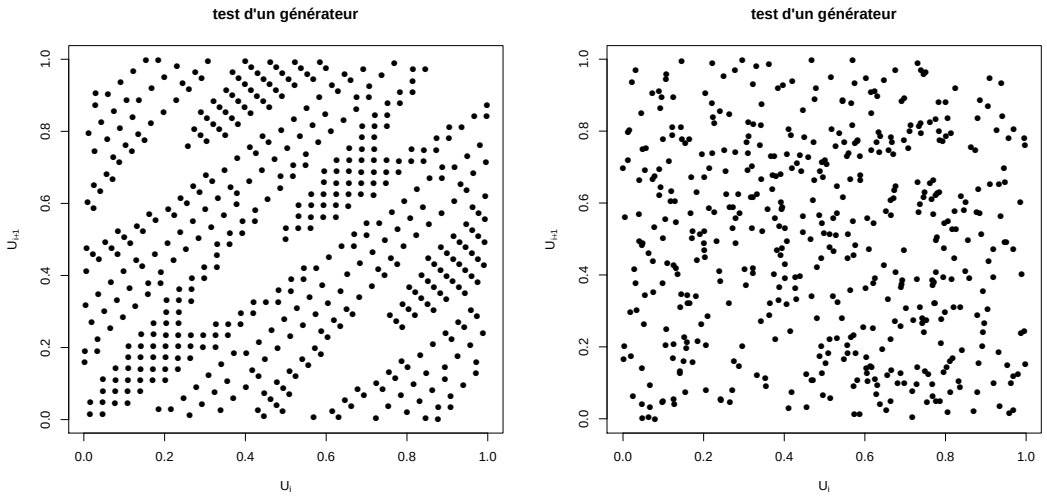

FIGURE 5 – Générateur congruentiel et générateur de $\mathcal{R}$.

On peut comparer avec un échantillon de même taille généré par le générateur de $\mathcal{R}$.

Le générateur congruentiel n'est donc pas très fiable, surtout si l'on veut générer des échantillons de petite taille. Aujourd'hui, un des générateurs les plus utilisés est le générateur *Kiss*[1], qui utilise en partie le générateur congruentiel.

3.3 Modéliser la loi normale

La fonction de répartition de la loi normale standard $\mathcal{N}(0,1)$ ne peut pas être exprimée à l'aide des fonctions usuelles. On ne peut donc pas utiliser le théorème 17 pour générer une loi normale à l'aide de loi uniforme. Cependant, il existe une transformation un peu plus compliquée qui permet de générer deux lois normales standards indépendantes à partir d'une loi exponentielle $\mathcal{E}(\frac{1}{2})$ et d'une loi uniforme $\mathcal{U}([0,2\pi])$ indépendantes. C'est l'algorithme de *Box-Muller*.

Proposition 18 (Loi exponentielle)
Soit X une variable aléatoire de loi $\mathcal{U}([0,1])$.
Alors la variable aléatoire $Y = -\frac{1}{\lambda}\ln(X)$ suit une loi exponentielle $\mathcal{E}(\lambda)$.

1. comme acronyme de ***K**eep **i**t **s**imple, **s**tupide !*

Démonstration 16

$$\begin{aligned}\text{Soit } t \in \mathbb{R}, \text{ on a : } \quad F_Y(t) &= \mathbb{P}(Y \leq t) \\ &= \mathbb{P}(-\frac{1}{\lambda}\ln(X) \leq t) \\ &= \mathbb{P}(\ln(X) \geq -\lambda t) \\ &= \mathbb{P}(X \geq e^{-\lambda t}) \\ &= (1 - e^{-\lambda t})\mathbb{1}_{\mathbb{R}_+}(t)\end{aligned}$$

On reconnaît la fonction de répartition de la loi exponentielle $\mathcal{E}(\lambda)$.

□

Théorème 19 (*Box-Muller*)
Soit U et V deux variables aléatoires *iid* de loi $\mathcal{U}([0,1])$.
On définit les variables $X = \sqrt{-2\ln(U)}\cos(2\pi V)$ et $Y = \sqrt{-2\ln(U)}\sin(2\pi V)$.
Alors X et Y sont *iid* et suivent la loi normale standard.

On peut comparer ici le résultat obtenu en appliquant cet algorithme avec le générateur congruentiel et le générateur de loi uniforme de $\mathcal{R}$ et le générateur de loi normale.

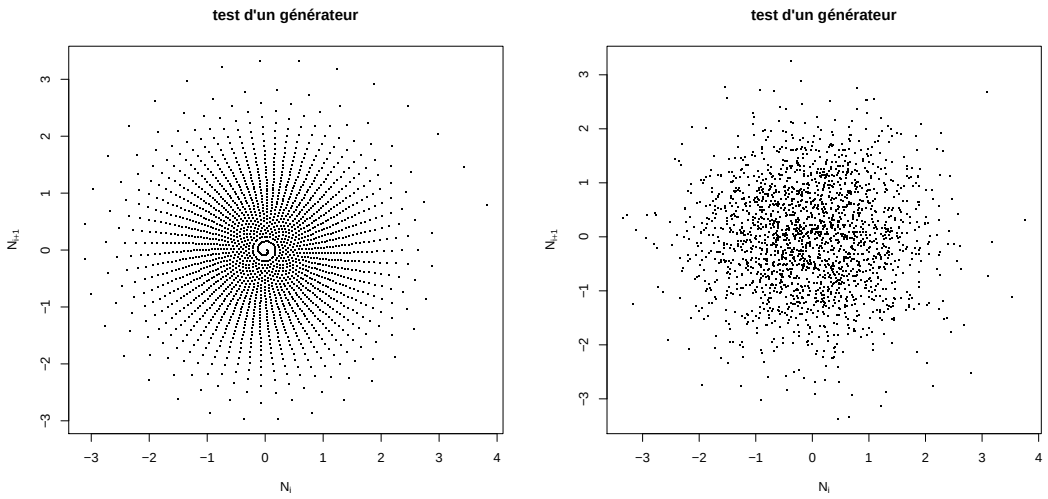

FIGURE 6 – *Box-Muller* appliqué au générateur congruentiel et au générateur de $\mathcal{R}$.

On remarque que les structures de droites du générateur congruentiel engendrent des structures de spirales après application de l'algorithme de *Box-Muller*. En revanche, on ne saurait différencier, à l'aide de ce test, le générateur de loi normale de $\mathcal{R}$ et l'algorithme de *Box-Muller* appliqué au générateur de loi uniforme de $\mathcal{R}$.

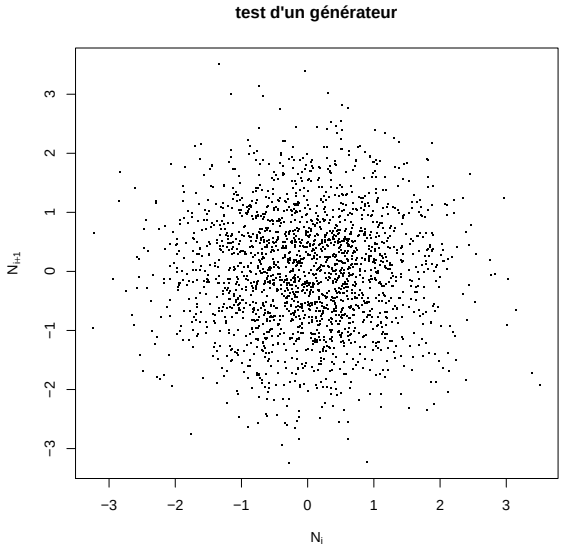

FIGURE 7 – Générateurs de loi normale de $\mathcal{R}$.

3.4 Convergence du processus

Pour illustrer la convergence de la suite de fonction de répartition empirique, on trace pour des échantillons de 20, 100 et 1000 données, la fonction de répartition et la fonction de répartition empirique associée, pour la loi normale standard.

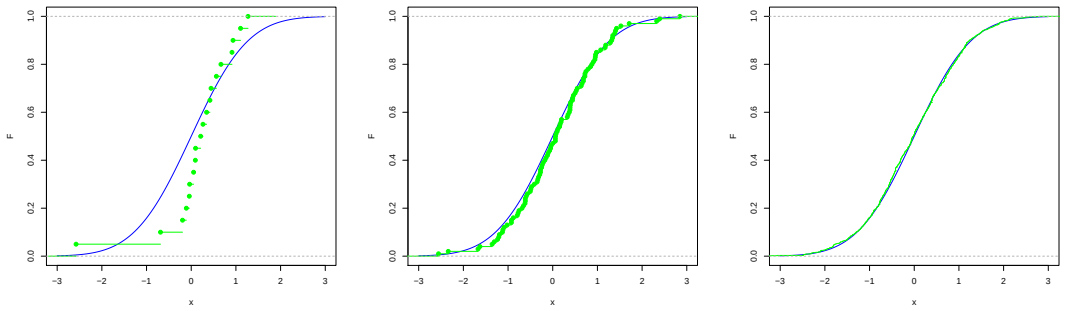

FIGURE 8 – processus empirique d'une loi normale standard

On remarque que pour 1000 données, la courbe de la fonction de répartition empirique colle celle de la fonction de répartition de très près. F_{1000} semble donc être une approximation raisonnable de de F.

On écrit une fonction `errponct(X,n,F,t)` qui, pour un échantillon X de réalisations de la loi F, permet de mesurer en t, l'erreur entre les n premières fonctions de répartition empiriques et la fonction de répartition. On trace ensuite cette erreur dans la même fenêtre graphique que la courbe $\frac{1}{\sqrt{n}}$ pour comparer la vitesse de conver-

gence. On utilise ici une loi uniforme $\mathcal{U}([0,1])$ et on regarde l'erreur en $t = 0,5$ pour les processus de 1 à 2500 données.

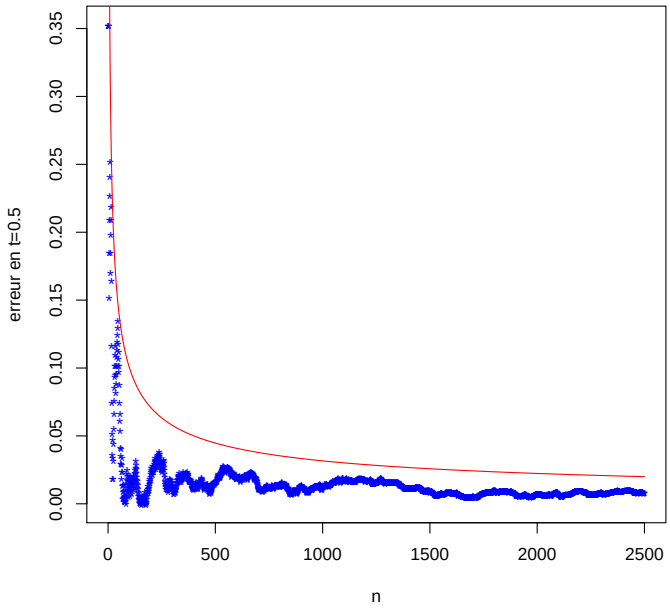

FIGURE 9 – Erreur en $t = 0,5$ en fonction du nombre de données (en bleu) et $n \mapsto \frac{1}{n}$ (en rouge).

On reconnaît bien l'allure en $\frac{1}{\sqrt{n}}$ de l'erreur, comme cela avait été prédit par le TCL.

3.5 Erreur uniforme

Nous avons vu que $D_n = \|F_n - F\|_\infty$ convergeait presque sûrement vers 0. A l'aide du logiciel $\mathcal{R}$, nous allons tester ce résultat et étudier la vitesse de convergence.

On définit la fonction `suppe(X,F,...)`, qui pour une réalisation X de la loi F donnée, renvoie le vecteur de l'erreur uniforme des itérations du processus empirique. On trace ensuite ce vecteur, en fonction de n. On peut alors comparer l'allure du graphe avec $\sqrt{n}$ par exemple.
Voici un exemple fait avec un échantillon de 2500 données pour une loi uniforme.

3.6 Loi de *Kolmogorov*

Dans le cas où F est continu, $\sqrt{n}D_n$ doit converger en loi vers la loi de *Kolmogorov*, et ce de manière indépendante de F. On peut illustrer ce résultat sur plusieurs

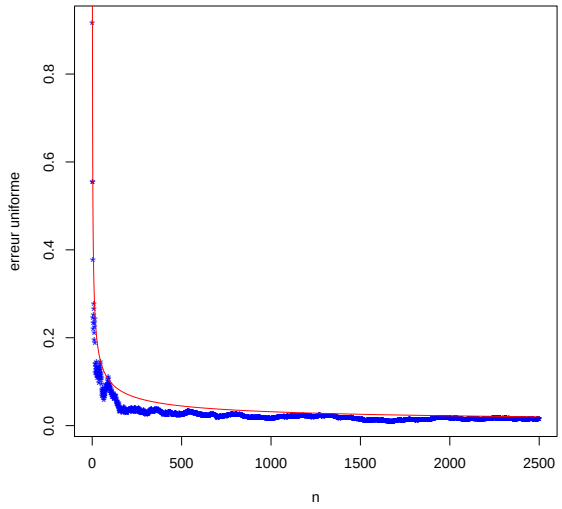

FIGURE 10 – $D_n = \|F_n - F\|_\infty$ pour une loi de Cauchy $\mathcal{C}(1)$

lois : la loi uniforme $\mathcal{U}([0,1])$, la loi normale $\mathcal{N}(0,1)$, et la loi de Cauchy $\mathcal{U}(1)$. Pour chacune d'elle, on effectue la démarche suivante :

1. On exécute la fonction suppe(n,...)

2. On conserve les résultats dans un vecteur colonne V

3. On recommence m fois ces 2 opérations (V est donc une matrice)

4. Pour chaque ligne de V, on trace la densité empirique associée à $\sqrt{i}V[i:]$

On utilise la fonction `errunif(n,m,rP=rnorm,pP=pnorm,...)` pour construire la matrice V. En fait chaque ligne i de V est une suite de réalisation de D_i. Pour plus de clarté, toutes les densités ne sont pas tracées.

La difficulté est d'avoir des paramètres m et n assez grands pour que la D_n soit suffisamment proche de 0 et qu'il y est assez de réalisation de D_i pour chaque i, pour que les densités empiriques soient assez proches de la densité de D_i. Mais il ne faut pas pour autant que le temps de calcul ne soit trop long. Ici on prendra les paramètres $m = 1000$ et $n = 1000$ qui permettent d'avoir une erreur relativement faible, tout en ayant un temps de calcul raisonnable (de l'ordre de 10 minutes).

Même si l'on "voit bien" que les densités convergent vers la densité de la loi de *Kolmogorov*, il peut-être intéressant de s'appuyer sur d'autres critères. Par exemple, la moyenne, la variance, ou la convergence des fonctions de répartition empiriques.

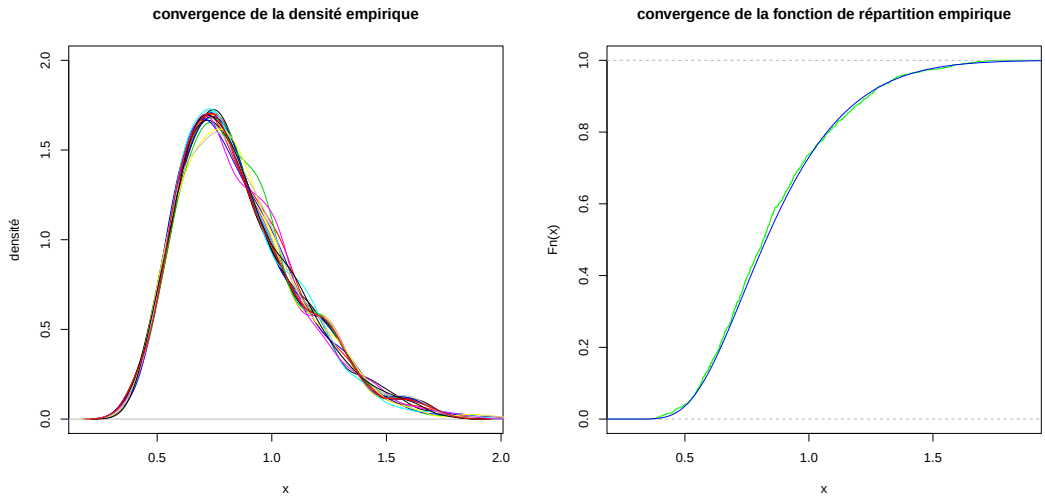

FIGURE 11 – convergence de la densité et la fonction de répartition de $\sqrt{n}D_n$ pour la loi $\mathcal{U}([0,1])$.

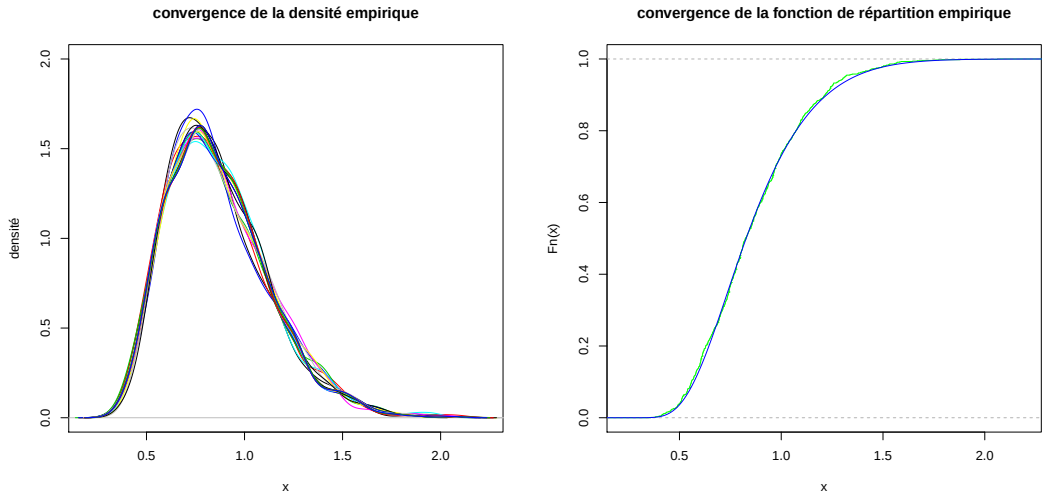

FIGURE 12 – convergence de la densité et la fonction de répartition de $\sqrt{n}D_n$ pour la loi $\mathcal{N}(0,1)$.

Les espérances et variances obtenues coïncident avec les valeurs que l'on a obtenues pour la loi de *Kolmogorov*. Le plus pertinent serait de faire un test statistique pour vérifier si la loi suivie est bien la loi de *Kolmogorov*.

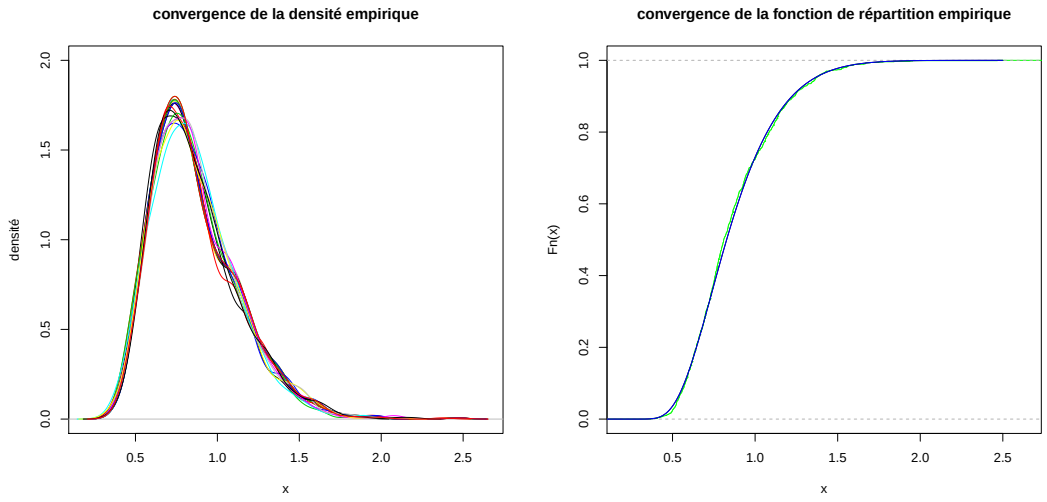

FIGURE 13 – convergence de la densité et la fonction de répartition de $\sqrt{n}D_n$ pour la loi $\mathcal{C}(1)$.

	$\mathbb{E}$	$\mathbb{V}$ar
$\mathcal{U}([0,1])$	0.862	0.263
$\mathcal{N}(0,1)$	0.864	0.256
$\mathcal{C}(1)$	0.869	0.266

TABLE 1 – Espérance et variance empirique de $\sqrt{1000}D_{1000}$ pour trois lois continues.

3.7 Cas des loi discrètes

On n'a vu que très peu de résultats dans le cas où la loi n'est pas continue. Qu'en est-il de la convergence de D_n ? Grâce à une simulation avec $\mathcal{R}$, on peut montrer que $\sqrt{n}D_n$ ne converge pas vers une loi de *Kolmogorov*. En effet, si l'on applique la démarche précédente avec des lois discrètes,comme une loi de *Poisson* $\mathcal{P}(\lambda)$, ou une loi géométrique $\mathcal{G}(p)$, les courbes des densités ou des fonctions de répartition empiriques obtenues ne se superposent pas du tout avec la densité ou a fonction de répartition de la loi de *Kolmogorov*.

On peut de la même manière montrer que les espérances et variances ne sont pas les mêmes que pour la loi de *Kolmogorov*. De plus les moments semblent indiquer que la convergence ne se fait pas vers une loi commune, mais vers une loi qui dépend de la loi de départ.

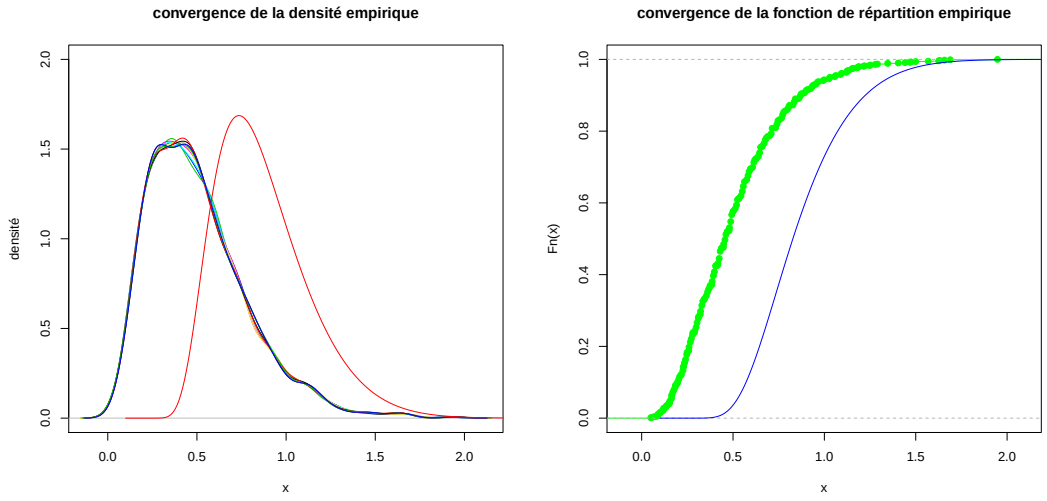

FIGURE 14 – Convergence de $\sqrt{n}D_n$ pour la loi $\mathcal{P}(0.7)$.

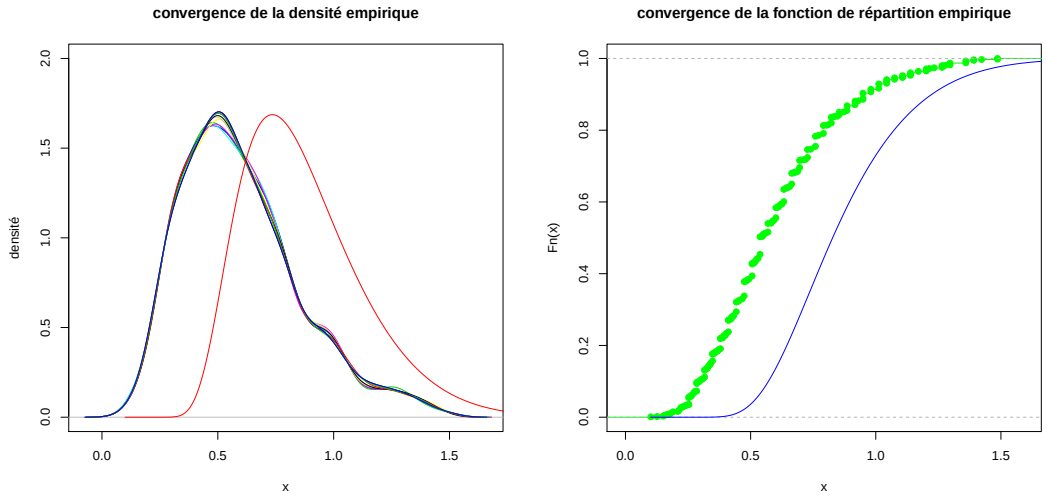

FIGURE 15 – Convergence de $\sqrt{n}D_n$ pour la loi $\mathcal{G}(0.4)$.

	$\mathbb{E}$	$\mathbb{V}$ar
$\mathcal{G}(0.4)$	0.599	0.259
$\mathcal{P}(0.7)$	0.494	0.276
$\mathcal{P}(1.3)$	0.555	0.257

TABLE 2 – Espérance et variance empirique de $\sqrt{1000}D_{1000}$ pour trois lois discrètes.

4 Processus empirique et tests statistiques

On va maintenant utiliser les résultats que nous venons de voir pour faire des tests statistiques. On se donne toujours $(X_1, \dots, X_n)$, des variables aléatoires *iid*, dont la fonction de répartition F est inconnue. On se demande si cette fonction F est égale ou non à la fonction de répartition F_0 (connue).

4.1 Vocabulaire.

On se donne une *structure statistique* $(\Omega, \mathcal{A}, \mathcal{P})$, où $\mathcal{A}$ est une tribu sur Ω et $\mathcal{P}$ une famille de probabilité sur l'espace mesurable $(\Omega, \mathcal{A})$. Soit $(\mathcal{P}_0, \mathcal{P}_1)$ une partition de $\mathcal{P}$, on cherche à déterminer, à partir d'une observation $(X_1(\omega), \dots, X_n(\omega)), \omega \in \Omega$ si la loi régissant le phénomène observé appartient à $\mathcal{P}_0$ ou à $\mathcal{P}_1$ et on cherche la stratégie qui occasionne le moins d'erreur possible.

Définition 5 (Hypothèse)
On appelle *hypothèse* une affirmation "la loi des X_i est $\mathcal{P}_0$" (ou $\mathcal{P}_1$). Pour $\mathcal{P}_0$ on parle de l'*hypothèse nulle* et de l'*hypothèse alternative* pour $\mathcal{P}_1$.

On appelle le *problème de test* de l'hypothèse nulle contre l'hypothèse alternative, le problème P :

$$\mathcal{H}_0 : "\mathbb{P} \in \mathcal{P}_0" \quad \text{contre} \quad \mathcal{H}_1 : "\mathbb{P} \in \mathcal{P}_1".$$

Définition 6 (Test)
Un test est une application mesurable de E^n à valeur dans [0,1] (avec les X_i à valeur dans E).

Un test statistique T doit permettre de trancher entre les deux hypothèses :
- si $T(X_1(\omega), \dots, X_n(\omega)) = 0$, $\mathcal{H}_0$ est acceptée.
- si $T(X_1(\omega), \dots, X_n(\omega)) = 1$, $\mathcal{H}_0$ est rejetée.
- si $T(X_1(\omega), \dots, X_n(\omega)) = p \quad (0 < p < 1)$, $\mathcal{H}_0$ est rejetée avec la probabilité p (on lance une pièce déséquilibrée qui tombe sur 1 avec une probabilité p, si elle fait 1, on rejette $\mathcal{H}_0$).

On dira qu'un test est *déterministe*, s'il est à valeur dans {0,1}. On peut alors définir sa *zone critique* comme étant l'ensemble des observations pour lesquelles l'hypothèse nulle est rejetée, par : $R = T^{-1}(1)$.

On dira que $\mathcal{H}_0$ est *vrai* si $\mathbb{P}$ appartient effectivement à $\mathcal{P}_0$ (et de même avec $\mathcal{H}_1$). En résolvant le problème de test P à l'aide du test T, on peut commettre deux erreurs :

– rejeter $\mathcal{H}_0$ alors qu'elle est vraie

– accepter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie

Les probabilités de commettre ces erreurs peuvent être quantifiées.

Définition 7 (Erreur)

– L'*erreur de première espèce*, ou *niveau*, de T est $\alpha_T = \sup_{\mathcal{P}_0} \mathbb{E}[T]$.

– L'*erreur de seconde espèce* de T est $1 - \sup_{\mathcal{P}_1} \mathbb{E}[T]$.

Pour mieux comprendre ces quantités, on se place dans le cas où $\mathcal{P}_0$ et $\mathcal{P}_1$ sont deux singletons et où le test T est de la forme : $T = \mathbb{1}_A$ $(A \in \mathcal{A})$. On a alors $\alpha_T = \mathbb{P}_0(A)$, c'est la probabilité que $\mathcal{H}_0$ soit rejetée alors qu'elle est vraie. De même, l'erreur de seconde espèce est $\mathbb{P}_1(A^c)$, c'est la probabilité que $\mathcal{H}_0$ soit acceptée alors qu'elle est fausse.

Définition 8 (Puissance)
La puissance d'un test T est la fonction

$$\begin{array}{rccc} \beta_T & : & \mathcal{P}_1 & \to & [0,1] \\ & & \mathbb{P} & \mapsto & \mathbb{E}_{\mathbb{P}}[T] \end{array}$$

La puissance sert à comparer les tests entre eux et permet de redéfinir l'erreur de seconde espèce.

4.2 Test de *Kolmogorov-Smirnov*

On définit les deux hypothèses complémentaires de cette manière :

$$\mathcal{H}_0 : "F = F_0" \quad \text{et} \quad \mathcal{H}_1 : "F \neq F_0".$$

Pour tout entier $n \in \mathbb{N}^*$, on définit le test de *Kolmogorov-Smirnov* T_n de la manière suivante :

$$T_n = \mathbb{1}_{]c_n(\alpha), +\infty[}(D_n)$$

avec $D_n = \|F_n - F_0\|_\infty$ et la suite $(c_n(\alpha))_{n \in \mathbb{N}^*}$ que nous définirons plus tard, qui devra vérifier certaines propriétés afin de contrôler les erreurs de première et seconde

espèce. C'est un test déterministe. Sa zone critique est $R_n = \{D_n > c_n(\alpha)\}$. Son erreur de première espèce est :

$$\alpha_{T_n} = \mathbb{P}_0(D_n > c_n(\alpha))$$

avec $\mathbb{P}_0$ la probabilité associée à la F_0.

Puisque l'on ne peut pas minorer simultanément les deux erreurs, on décide de fixer l'erreur de première espèce et on cherche à minorer l'erreur de seconde espèce. On se fixe un *seuil* α, qui doit majorer le niveau de nos tests. On voudrait que la suite $(c_n(\alpha))_{n\in\mathbb{N}^*}$vérifie :

1. Pour tout n, $\alpha_{T_n} = \alpha$.

2. L'erreur de seconde espèce converge vers 0 (id est $\beta_{T_n} \to 1$).

A priori, on ne sait pas si l'on peut définir une telle suite pour un F_0 donné. On ne sait encore moins si l'en existe une, valable pour toutes les lois F_0. Cependant, si F_0 est continue, on sait que la loi de D_n ne dépend pas de F et est continue. La loi de T_n ne dépend donc pas non plus de F_0 et $\mathbb{P}_0(D_n = c_n(\alpha)) = 0$. On peut donc définir la suite $(c_n(\alpha))_{n\in\mathbb{N}^*}$. Cette suite est universelle.

Proposition 20
On pose $q_{1-\alpha}$ le quantile $1-\alpha$ de la loi de *Kolmogorov*.
La suite définie par $\tilde{c}_n(\alpha) = \frac{q_{1-\alpha}}{\sqrt{n}}$ vérifie $\alpha_{T_n} \to \alpha$ et $\beta_{T_n} \to 1$.

Démonstration 17
Pour le niveau, on a :

$$\begin{aligned}\alpha_{T_n} &= \mathbb{P}((D_n > \tilde{c}_n(\alpha)) \\ &= \mathbb{P}(D_n > \frac{q_{1-\alpha}}{\sqrt{n}}) \\ &= \mathbb{P}(\sqrt{n}D_n > q_{1-\alpha})\end{aligned}$$

Comme on a la convergence en loi, vers la loi de *Kolmogorov*, On a le résultat.
On se qui concerne l'erreur de seconde espèce, si $F \neq F_0$, il existe $t \in \mathbb{R}$ tel que :

$$|F(t) - F_0(t)| = \varepsilon > 0.$$

$$\begin{aligned}\text{On a :} \quad D_n &= \|F_n - F_0\|_\infty \\ &\leq \|F_n - F\|_\infty + \|F - F_0\|_\infty \\ &\leq \|F_n - F\|_\infty + \varepsilon\end{aligned}$$

Le terme $\sqrt{n}\|F_n - F\|_\infty$ converge en loi vers la loi de *Kolmogorov* et le terme $\sqrt{n}\varepsilon$ diverge vers $+\infty$. D'après le théorème de *Slutsky*, $\sqrt{n}D_n$ converge en loi vers $+\infty$. Donc $\mathbb{P}_{\mathcal{H}_1}(D_n > c_n(\alpha)) = \mathbb{P}_{\mathcal{H}_1}(\sqrt{n}D_n > q_{1-\alpha}) \to 1$. L'erreur de seconde espèce converge donc vers 0.

□

On dispose même d'un lien assez fort entre les deux suites.

Proposition 21

$$\text{On a l'équivalence :} \quad c_n(\alpha) \sim \frac{q_{1-\alpha}}{\sqrt{n}} \quad \text{quand} \quad n \to \infty.$$

Démonstration 18
On a : $\mathbb{P}_0(\sqrt{n}D_n > \sqrt{n}c_n(\alpha)) = \alpha$ et que $\sqrt{n}D_n$ converge en loi vers la loi de *Kolmogorov* K. Comme les fonction de répartition sont croissantes et bornées, d'après le théorème de *Dini*, la convergence de F_{D_n} vers F_K est uniforme. Pour tout réel $\varepsilon > 0$ on a :

$$\begin{aligned}
&\|(1 - F_{D_n}) - (1 - F_K)\|_\infty < \varepsilon \\
&\left|\mathbb{P}_0(\sqrt{n}D_n > \sqrt{n}c_n(\alpha)) - \mathbb{P}_0(K > \sqrt{n}c_n(\alpha))\right| < \varepsilon \\
&\left|\alpha - \mathbb{P}_0(K > \sqrt{n}c_n(\alpha))\right| < \varepsilon
\end{aligned}$$

Or la loi de K est continue, donc il existe un unique réel x tel que $\mathbb{P}_0(K > x) = \alpha$, c'est $q_{1-\alpha}$. Ainsi la suite $\sqrt{n}c_n(\alpha)$ converge vers $q_{1-\alpha}$. D'où l'équivalence.

□

On peut essayer de comparer les valeurs des deux suites. Pour cela, on commence par construire une fonction `quant(x,eps)` qui donne le quantile x de la loi de *Kolmogorov* avec une précision *eps*. On peut donc construire la suite $(\tilde{c}_n(\alpha))_{n\in\mathbb{N}^*}$. Pour la suite $(c_n(\alpha))_{n\in\mathbb{N}^*}$, c'est plus compliqué comme en atteste le cas de $c_1(\alpha)$.

Proposition 22

$$c_1(\alpha) = \frac{\alpha + 1}{2}.$$

Démonstration 19
Soit X_1 une variable aléatoire. On peut supposer que ça loi est $\mathcal{U}([0,1])$ car cela la loi de D_1 est indépendante de F. La fonction de répartition empirique associée est :

$$F_1 = \mathbb{1}_{[X_1,+\infty]}.$$

La différence entre les deux fonctions est :

$$|F_1(x) - F(x)| = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ x & \text{si} \quad 0 \leq x < X_1 \\ 1 - x & \text{si} \quad {}_1 \leq x < 1 \\ 0 & \text{si} \quad x \geq 1 \end{cases}$$

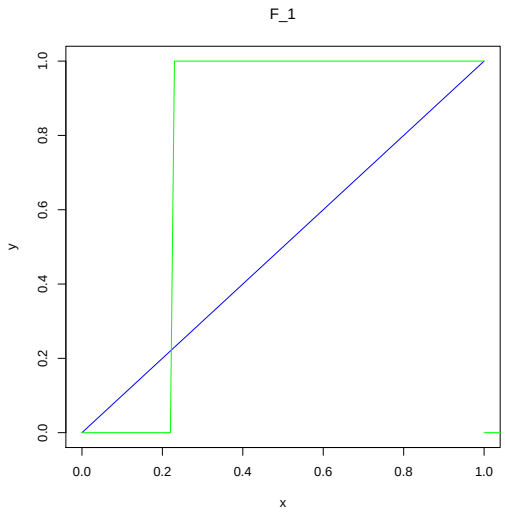

FIGURE 16 – courbe représentative de F et F_1.

On en déduit que D_1 est la variable aléatoire, à support dans $[\frac{1}{2}, 1]$ définie par $D_1 = \max(X_1, 1 - X_1)$. Pour tout $t \in \mathbb{R}$, on a :

$$\begin{aligned}
\mathbb{P}(D_1 \leq t) &= \mathbb{P}(X_1 \leq t \quad \text{et} \quad 1 - X_1 \leq t) \\
&= 1 - \mathbb{P}(X_1 \geq t \quad \text{ou} \quad 1 - X_1 \geq t) \quad \text{(car la loi est continue)} \\
&= 1 - \mathbb{P}(X_1 \geq t) - \mathbb{P}(1 - X_1 \geq t) \\
&= \mathbb{P}(X_1 \leq t) - \mathbb{P}(X_1 \leq 1 - t) \\
&= \begin{cases} 0 & \text{si} \quad t \leq t \\ 2t - 1 & \text{si} \quad 0 \leq x < \frac{1}{2} \\ 1 & \text{si} \quad t \geq 1 \end{cases}
\end{aligned}$$

On reconnaît une fonction de répartition. Donc $D_1 \sim \mathcal{U}([\frac{1}{2}, 1])$.

□

On comprend bien qu'il n'est pas raisonnable de chercher à calculer explicitement les $c_n(\alpha)$. On va donc en calculer des approximations : on effectue N processus empirique avec des échantillons de tailles n, on calcule les N $\sqrt{n}D_n$ associés et on considère que $c_n(\alpha)$ est le dernier échantillon inférieur à $1 - \alpha$. La figure 17 permet de comparer la suite $(c_n(\alpha))_{n \in \mathbb{N}^*}$ à celle des $(\tilde{c}_n(\alpha))_{n \in \mathbb{N}^*}$.

A partir de $n = 100$, on peut juger que $\tilde{c}_n(\alpha)$ est une approximation suffisamment juste de $c_n(\alpha)$. En effet $\tilde{c}_{100}(0.05) \approx 0.1358$ et $c_{100}(0.05) \approx 0.1340$.

4.3 La p-value

Pour l'instant, on a comparé notre statistique D_n aux quantiles de la fonction de répartition. Pour chaque seuil α et chaque n, il faut recalculer ces quantiles. C'est

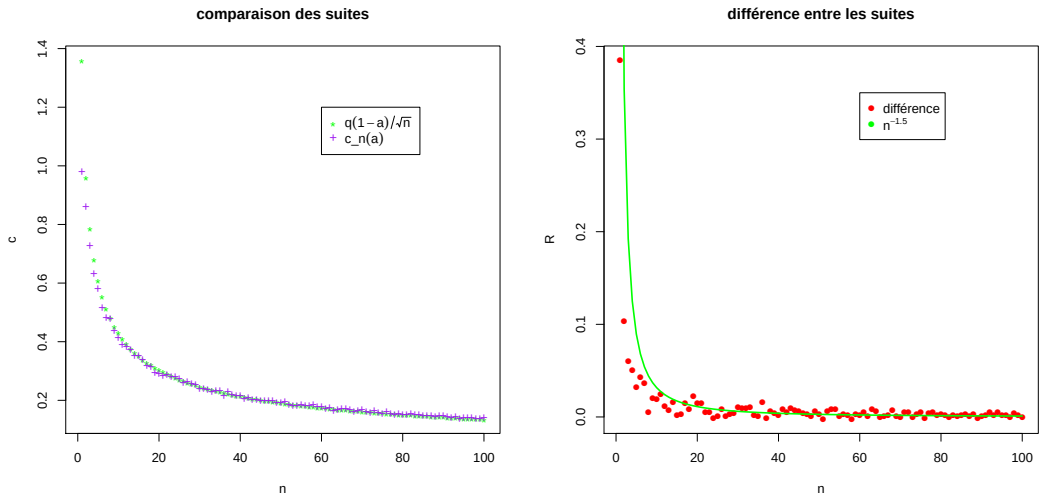

FIGURE 17 – Les 100 premiers termes des deux suites et leur différence pour $\alpha = 0.05$.

beaucoup de calcul. Dans le cas où n est grand, on peut considérer que $\sqrt{n}D_n$ suit la loi de *Kolmogorov*. On peut aussi se passer du calcul des quantiles pour chaque seuil en travaillant en terme d'aire : on introduit la *p-value*.

Définition 9 (p-value)
Soit T un test statistique et $(X_1, \ldots, X_n)$ une observation.
La p-value p_v est l'infimum des seuils tels que $\mathcal{H}_0$ soit rejeté.

En d'autres termes, pour un seuil α donné, la p-value est la quantité qui vérifie le critère suivant :

- si $p_v > \alpha$, alors $\mathcal{H}_0$ est vrai.
- si $p_v < \alpha$, alors $\mathcal{H}_1$ est vrai.

On remarque si l'on a la p-value, on peut déterminer quelle hypothèse est vérifiée à n'importe quel seuil donné.

On se place dans le cas où n est suffisamment grand. Pour le test de *Kolmogorov-Smirinov*, la p-value peut être explicitée assez facilement. En effet un seuil α peut être interprété comme une aire sous la courbe représentative de la densité de la loi de *Kolmogorov*. Si $\sqrt{n}D_n$ est plus petit que $q(1-\alpha)$, alors l'aire $\mathcal{A}$ comprise sous la courbe entre les bornes $\sqrt{n}D_n$ et $+\infty$ est plus grande que α. Cette aire, c'est la p-value. On peut l'obtenir assez facilement à partir de la fonction de répartition de la loi de *Kolmogorov* :

$$\mathcal{A} = 1 - F_K(\sqrt{n}D_n).$$

La maîtrise du seuil ne se fait donc plus que sur un calcul de la fonction de répartition de la loi de *Kolmogorov*.

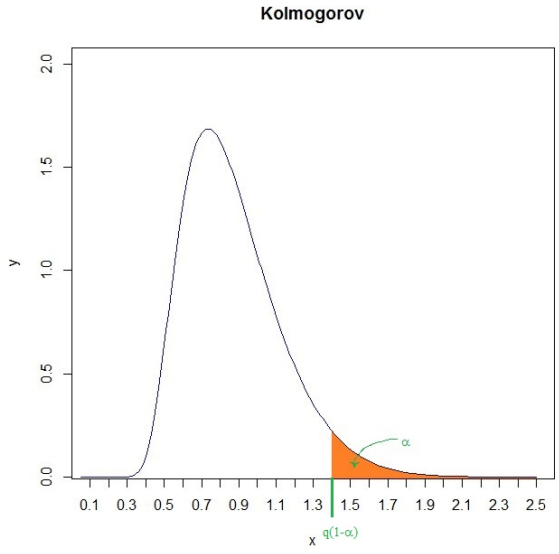

FIGURE 18 – lien entre seuil et aire

On note p_{v_n} la p-value associée à D_n. La p-value dépend de D_n, c'est donc elle même une variable aléatoire.

Théorème 23 (Loi de la p_{v_n})
Si $\mathcal{H}_0$ est vraie, p_{v_n} converge en loi vers $\mathcal{U}([0,1])$. Si $\mathcal{H}_1$ est vraie, p_{v_n} converge en probabilité vers la constante égale à 0.

Démonstration 20
Pour tout $n \in \mathbb{N}^*$, $p_{v_n} = 1 - F_K(\sqrt{n}D_n) \in [0,1]$. La fonction $t \mapsto 1 - F_K(t)$ est continue. Si $\mathcal{H}_1$ est vraie, $\sqrt{n}D_n$ diverge et donc p_{v_n} converge en loi vers 0, et donc la convergence ce fait aussi en probabilité. Si $\mathcal{H}_0$ est vraie, $\sqrt{n}D_n$ converge en loi vers la loi de *Kolmogorov*. D'après la proposition 15, $F_K(\sqrt{n}D_n)$ converge en loi vers $\mathcal{U}([0,1])$. Donc p_{v_n} converge en loi vers $\mathcal{U}([0,1])$.

□

Pour illustrer ce résultat, on simule N échantillons de taille n suivant la loi de *Student* à 1 degré de liberté, et on calcule les p-value. On crée un vecteur V avec l'hypothèse $\mathcal{H}_0$: "F_0 est une loi de *Student* à 1 degré de liberté", et un vecteur F avec l'hypothèse $\mathcal{H}_0$: "$F_0 = \mathcal{N}(0,1)$". On obtient ainsi deux échantillons de p-value. La figure 19 représente les fonctions de répartition empiriques associées, avec en rouge l'hypothèse "F_0 est une loi de *Student* à 1 degré de liberté" et en vert, l'hypothèse "$F_0 = \mathcal{N}(0,1)$".

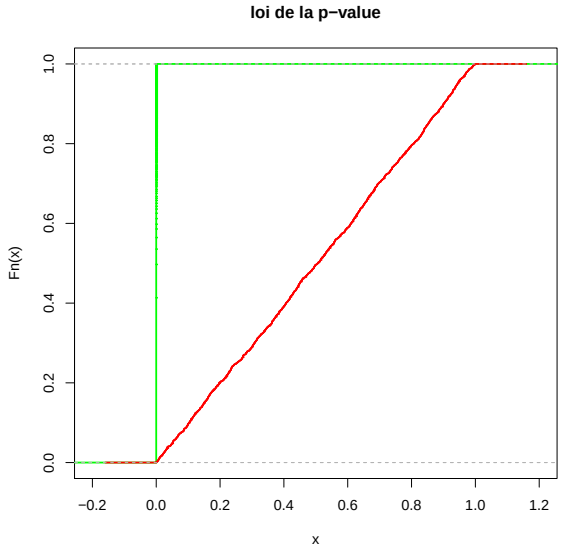

FIGURE 19 – Fonction de répartition de la p-value.

4.4 Loi de *Student*

Afin de mettre en évidence une des limites de ce test, nous allons le mettre en œuvre sur une famille de loi : les lois de *Student*.

Définition 10 (Loi de Student)
Soient $X_1, \ldots, X_\nu, Y$ des variables aléatoires *iid* de loi normale standard.
La variable aléatoire

$$t = \frac{Y}{\sqrt{\sum_{i=1}^{\nu} \frac{X_i^2}{\nu}}}$$

suit la loi de *Student* à ν degrés de liberté.

Proposition 24
Soit $(t_\nu)_{\nu \in \mathbb{N}^*}$ une suite de variables aléatoires de loi de *Student* à $\varphi(\nu)$ degrés de liberté, où φ est une extractrice.
Alors la suite $(t_\nu)_{\nu \in \mathbb{N}^*}$ converge en loi vers une loi normale standard.

Démonstration 21
On commence par le montrer dans le cas où l'on dispose de $(X_n)_{n \in \mathbb{N}^*}$ et Y *iid* de loi normale standard telles que Z_ν soit défini comme dans la définition.

D'après la *loi des grands nombres*, on a :

$$\sum_{i=1}^{\nu} \frac{X_i}{\nu} \xrightarrow{ps} \mathbb{E}[X_1^2] = \mathbb{V}\text{ar}(X_1) = 1.$$

La fonction $x \mapsto \frac{1}{\sqrt{x}}$ est continue. D'après le *Continuous Mapping Theorem*, on a :

$$\frac{1}{\sqrt{\sum_{i=1}^{\nu} \frac{X_i}{\nu}}} \Longrightarrow 1.$$

D'après le théorème de *Slutsky*, t_ν converge en loi vers Y, de loi normale standard.
On note $(F_n)_{n\in\mathbb{N}^*}$ la suite des fonctions de répartition des t_ν. Cette suite converge vers la fonction de répartition de la loi normale standard.

Maintenant pour une suite quelconque, il suffit de voir que la convergence en loi est une convergence seulement sur les fonctions de répartition, qui elles sont indépendantes des relations d'indépendances des variables aléatoires. La suite de fonctions de répartition associée est maintenant une sous-suite extraite de $(F_n)_{n\in\mathbb{N}^*}$. Elle converge donc vers la fonction de répartition d'une loi normale standard.
D'où la convergence en loi.

□

Que se passe-t-il lorsque l'on applique le test de *Kolmogorov-Smirnov* à un échantillon de loi de *Student* avec un degré de liberté élevé et avec les hypothèses suivantes :

$$\mathcal{H}_0 : "F_0 = F_{\mathcal{N}(0,1)}" \quad \text{contre} \quad \mathcal{H}_1 : "F_0 \neq F_{\mathcal{N}(0,1)}"?$$

On se fixe un seuil de $\alpha = 5\%$. On utilise la suite $(c_n)_{n\in\mathbb{N}^*}$ définie par $c_n = \frac{1{,}357}{\sqrt{n}}$. on ne pourra donc pas vraiment tirer de conclusion pour les petits échantillons. Pour différentes valeurs de ν, on calcule la puissance pour plusieurs valeurs de n, la taille de l'échantillon. La figure 20 représente ces puissances.

Toutes les courbes obtenues admettent pour asymptote la droite $y = 1$. Ainsi, le test fonctionne bien. Cependant, plus ν est élevé, plus la convergence de la courbe vers son asymptote est lente. Cela signifie que le test a besoin de plus d'observation pour différencier les lois de *Student* de fort degré de liberté d'une loi normale. On remarque même que pour une loi à 7 degrés de liberté, la puissance n'est que de 80% pour un échantillon de 3000 données. Le test n'est donc pas très performant dans ce cas.

On peut aussi illustrer ceci avec la p-value. La figure 21 représente la fonction de répartition de la p-value pour différentes valeurs du degré de liberté avec comme hypothèse $\mathcal{H}_0$: "$F_0 = \mathcal{N}(0,1)$".

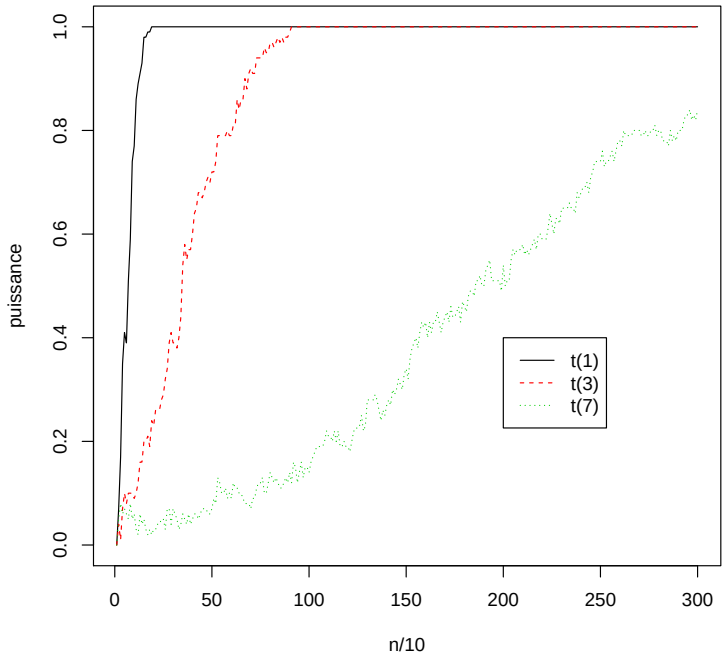

FIGURE 20 – puissance du test en fonction de la taille de l'échantillon

4.5 Lois Normales

Pour l'instant, on a toujours effectué le test en connaissant parfaitement la loi F_0. On veut maintenant poser les hypothèses suivante :

$$\mathcal{H}_0 : "F_0 \in \{\mathcal{N}(\mu, \sigma^2)/\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*\}" \quad \text{contre} \quad \mathcal{H}_1 : "F_0 \notin \{\mathcal{N}(\mu, \sigma^2)/\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*\}".$$

Il n'est pas possible de mettre en œuvre ces hypothèses telles quelles. Cependant, on sait que la moyenne empirique $\overline{X}$ et la variance empirique s^2 de notre échantillon convergent vers l'espérance et la variance. Si notre échantillon suit une loi normale, alors $\mathcal{N}(\overline{X}, s^2)$ converge vers cette loi. On pose donc $D_n = \left\| F_n - F_{\mathcal{N}(\overline{X}, s^2)} \right\|_\infty$ et on effectue le test comme précédemment.

On simule N échantillons de taille n suivant la loi normale standard. On calcule pour chacun la p-value avec les hypothèse citées ci-dessus. La figure 22 représente la fonction de répartition empirique de cet échantillon de p-value.

On remarque que le niveau de ce test est très faussé. Quand on pense faire un test à 30%, on réalise en fait un est à 5%. On rejette donc toujours l'hypothèse nulle.

Le test que l'on réalise n'est donc plus du tout pertinent. En réalité, la statistique $\sqrt{n}D_n$ ainsi définie ne converge plus en loi vers la loi de *Kolmogorov*, ce qui explique l'incohérence.

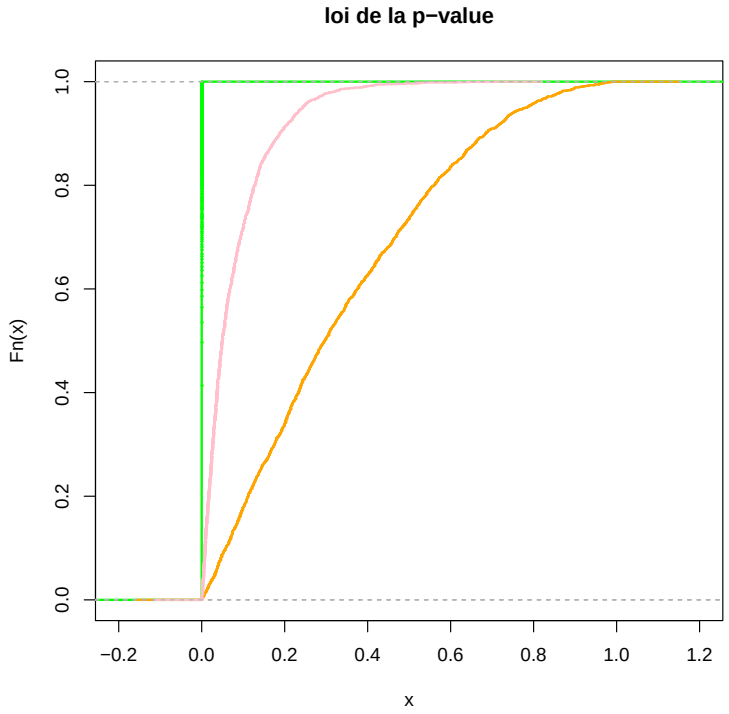

FIGURE 21 – p-value en pour 1, 5 et 8 degrés de liberté.

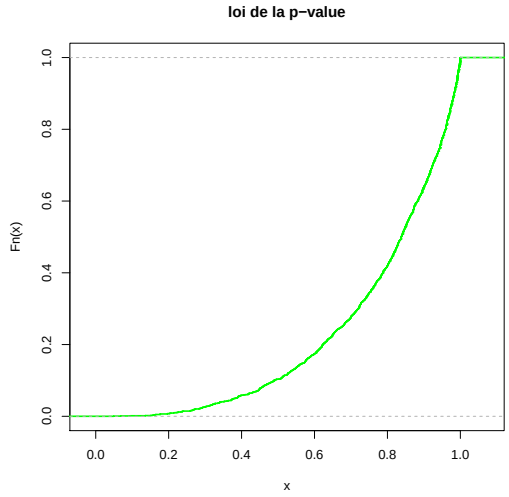

FIGURE 22 – Fonction de répartition de la p-value sous l'hypothèse $\mathcal{H}_0$ vraie.

5 Processus empirique et estimateurs statistiques

Dans cette partie, on va voir une nouvelle application du processus empirique, notamment à travers la méthode de *bootstrap*. Soient $(X_1, \dots, X_n)$ des variables aléatoires *iid* de fonction de répartition F. On s'intéresse à une application mesurable, θ de F, comme l'espérance. Comme F est inconnue, on ne peut pas avoir accès à $\theta(F)$. Cependant, on a vu que la fonction de répartition empirique caractérisait bien la loi. On étudie donc $\hat{\theta}_n = \theta(F_n)$ en espérant que cette nouvelle application converge bien vers $\theta(F)$.

Dans le cas de l'espérance, on utilise la moyenne empirique :

$$\hat{\theta_n} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

5.1 Estimation

On se donne une structure statistique $(\Omega, \mathcal{A}, \mathcal{P})$. On s'intéresse à θ, une application mesurable de $\mathbb{P}$. En général, on s'intéresse à l'espérance :

$$\theta(\mathbb{P}) = \int x \mathrm{d}\mathbb{P}(x),$$

la variance

$$\theta(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$$

ou à la médiane.

Définition 11 (Estimateur)
Un *estimateur* $\hat{\theta}$ de θ est une application mesurable de E^n, qui ne dépend pas de θ. Pour tout ω, $T(X_1(\omega), \dots, X_n(\omega))$ est une estimation de θ.

Cette définition est très générale. Il nous faut des critères pour distinguer les "bons" estimateurs des autres. Le plus intuitif est la notion de convergence. Quelle signification donne-t-on à "$\hat{\theta}$ est une bonne approximation de θ" ? On utilise les modes de convergence probabiliste usuels : convergence presque sûre, convergence en probabilité. On peut aussi s'intéresser à l'erreur moyenne de notre estimateur :

Définition 12 (Biais)
Soit $\hat{\theta}$ un estimateur de θ. Le *biais* de $\hat{\theta}$ est la fonction

$$\begin{array}{rccc} b_{\hat{\theta}} : & \mathcal{P} & \to & \mathbb{R} \\ & \mathbb{P} & \mapsto & \mathbb{E}_{\mathbb{P}}[\hat{\theta} - \theta(\mathbb{P})]. \end{array}$$

Un estimateur *sans biais* est un estimateur dont le biais est nul pour tout $\mathbb{P}$.

Même s'il est plus intéressant de travailler avec un estimateur sans biais, cela n'est pas toujours possible.

Une estimation n'a que peu de sens s'il l'on ne connaît pas la probabilité que la valeur exacte soit dans un intervalle autour de cette estimation.

Définition 13 (Intervalle de confiance)
On appelle I *intervalle de confiance de niveau* β un intervalle, dont les bornes sont aléatoires et tel que

$$\mathbb{P}(\theta \in I) = \beta.$$

En général β prend des valeurs comme 95% ou 90%. Un intervalle de confiance n'est pas unique. On peut, par exemple, choisir ou non de le centrer en $\hat{\theta}_n$.

5.2 Espérance d'une loi normale

On suppose que l'on dispose de $X_1, \ldots, X_n$ des variables aléatoires *iid* de loi normale $\mathcal{N}(\theta, 1)$. On cherche à déterminer le paramètre θ, qui est l'espérance de nos variables. On utilise donc la moyenne empirique.

Proposition 25
La moyenne empirique est un estimateur sans biais qui converge presque sûrement vers l'espérance.

Démonstration 22
La convergence découle de la *loi de grands nombres*.
Soit $n \in \mathbb{N}^*$, on a :

$$\mathbb{E}[\,\overline{X}_n - \theta] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] - \theta$$

Comme les variables sont indépendantes, on a :

$$\mathbb{E}[\,\overline{X}_n - \theta] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] - \theta.$$

Ainsi, on en déduit que le biais est nul.

□

On remarque dans cette démonstration que le biais est nul pour toutes les lois. En revanche la convergence presque sûre découle de l'intégrabilité de la loi normale.

Maintenant que l'on sait que notre estimateur converge, on doit déterminer quel niveau de confiance on peut lui accorder.

Proposition 26
Pour tout $\theta \in \mathbb{R}$, l'intervalle de confiance de niveau $1-\alpha$ est $\left[\overline{X}_n - \frac{q}{\sqrt{n}}, \overline{X}_n + \frac{q}{\sqrt{n}} \right]$, où q est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale standard.

Démonstration 23
Comme somme de lois normales *iid*, on a :

$$\sqrt{n}(\overline{X}_n - \theta) \sim \mathcal{N}(0,1).$$

En utilisant les quantiles de la loi normale standard, on a :

$$\mathbb{P}(\sqrt{n}(\overline{X}_n - \theta) \in [-q, q]) = 1 - \alpha.$$

D'où le résultat.

□

Pour illustrer ce résultat, on fait une simulation. On crée un échantillon de taille 10 suivant la loi normale standard.

0.4925046	0.8616297	0.5492627	0.9850950	0.3296270
0.7108167	3.1333777	1.2548722	-0.8043355	0.4583930

TABLE 3 – Échantillon suivant $\mathcal{N}(0,1)$.

On calcule la moyenne empirique associée : $\overline{X}_{10} = 0.7971243$. L'intervalle de confiance à 95% est :

$$I_{TCL} = [0.1773293, 1.416919].$$

Dans cet exemple, la valeur 0, n'est pas comprise dans l'intervalle de confiance. On se retrouve dans un des 5 cas sur 100 où la valeur estimée n'est pas dans l'intervalle. Il faut construire l'intervalle de confiance à 1% pour qu'il contiennent la valeur 0, mais quand on augmente le niveau de confiance, on étend l'intervalle.

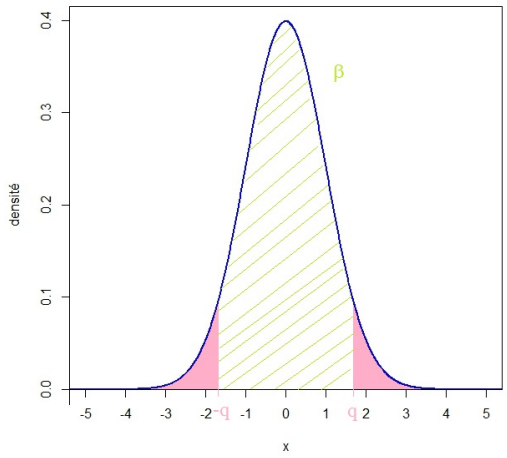

FIGURE 23 – Quantiles et intervalle de confiance.

5.3 Espérance d'une loi L^2

Dans le cas des lois gaussiennes, on obtient donc facilement les intervalles de confiance. Mais si l'on ne peut pas définir la loi de $\sqrt{n}(\overline{X}_n - \theta)$, ainsi que les quantiles associés à cette loi, on est incapable de déterminer les intervalles de confiance, et l'estimation perd de son sens. Dans le cadre L^2, le TCL permet d'aller un peu plus loin.

Soit $X_1, \ldots, X_n$ des variables aléatoires *iid* de loi L^2. On note θ son espérance et σ^2 sa variance, toutes les deux inconnues. On cherche encore à déterminer θ. Comme on l'a vu dans la partie précédente, la moyenne empirique est encore une fois adaptée.

On rappelle que $\sigma^2 = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$. Dans le même esprit que la moyenne empirique, on définit une variance empirique ainsi :

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2.$$

Proposition 27

$$\hat{\sigma}_n^2 \xrightarrow{ps} \sigma^2$$

Démonstration 24
Les X_i^2 sont des variables aléatoires L^1. D'après la *loi des grand nombre*, on a :

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{ps} \mathbb{E}[X_1^2].$$

De même, on a :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{ps} \mathbb{E}[X_1].$$

Comme la fonction $t \mapsto t^2$ est continue, on a :

$$(\frac{1}{n}\sum_{i=1}^{n} X_i)^2 \xrightarrow{ps} \mathbb{E}[X_1]^2.$$

On en déduit que $\hat{\sigma}_n^2 \xrightarrow{ps} \sigma^2$.

□

On peut maintenant définir un intervalle de confiance asymptotique.

Proposition 28
En gardant les mêmes notations, la probabilité

$$\mathbb{P}(\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\hat{\sigma}_n^2}} \in [-q, q])$$

converge vers β.

Démonstration 25
D'après le *théorème central limite*, $\sqrt{n}(\overline{X}_n - \theta)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$. D'après le théorème de *Slutsky*,

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\hat{\sigma}_n^2}} \Longrightarrow \mathcal{N}(0, 1).$$

D'où le résultat.

□

5.4 Le *bootstrap*

Ce qui nous empêche de construire nos intervalles de confiance, dans le cas général, c'est le fait que la loi $\mathcal{L}$ de $\hat{\theta}_n - \theta$, dépendant de θ et de n, est inconnue. Pour contourner ce problème, on va essayer de construire une loi que l'on connaisse parfaitement, et qui doit approcher $\mathcal{L}$.

On dispose d'un échantillon $X_1, \ldots, X_n$ suivant la loi F de classe L^2. On sait que dans ce cas que $\hat{\theta}_n = \theta(F_n)$ est un bon estimateur de θ. On considère $X_1^*, \ldots, X_n^*$ un échantillon suivant la loi F_n et on note $\hat{\theta}_n^*$ l'estimateur de $\hat{\theta}_n$ associé à cet échantillon $*$. On estime $\hat{\theta}_n$ par $\hat{\theta}_n^*$ de la même manière que l'on estime θ par $\hat{\theta}_n$.

Définition 14 (Principe du *bootstrap*)
Le principe du *bootstrap* consiste à dire que la loi $\mathcal{L}$ de $\hat{\theta}_n - \theta$ doit être la même que la loi $\mathcal{L}^*$ de $\hat{\theta}_n^* - \hat{\theta}_n$.

La loi $\mathcal{L}^*$ est bien sûr une loi conditionnellement à $(X_1, \ldots, X_n)$. Si l'idée semble assez simple, il faut quand même le démontrer au cas par cas.

On suppose que l'analogie a été démontrée de manière rigoureuse et que l'on a

$$\mathbb{P}(\hat{\theta}_n - \theta \leq x) = \mathbb{P}^*(\hat{\theta}_n^* - \hat{\theta}_n \leq x) + R_n,$$

où R_n est un reste qui converge vers 0 en probabilité. La fonction de répartition de $\hat{\theta}_n^* - \hat{\theta}_n$ est connue : F_n est la fonction de répartition de la loi équi-répartie sur $\{X_1, \ldots, X_n\}$. On a donc :

$$\begin{aligned}\mathbb{P}^*(\hat{\theta}_n^* - \hat{\theta}_n \leq x) &= \mathbb{P}^*(\hat{\theta}_n^* \leq x + \hat{\theta}_n) \\ &= \frac{1}{n^n} \sum_{(X_1^*, \ldots, X_n^*) \in \{X_1, \ldots, X_n\}^n} \mathbb{1}_{]-\infty, x+\hat{\theta}_n]}(\hat{\theta}_n^*).\end{aligned}$$

La loi $\mathcal{L}^*$ étant connue parfaitement, on peut calculer les quantiles q^* associés. On note q_α le quantile d'ordre α de la loi $\mathcal{L}$. On a l'intervalle de confiance suivant :

$$\mathbb{P}(\theta \in [\hat{\theta}_n - q_{1-\alpha}, \hat{\theta}_n - q_\alpha]) = 1 - 2\alpha.$$

Avec la méthode de *bootstrap*, on peut remplacer les q_α par le q_α^*.

5.5 Application de la méthode de *bootstrap*

On se donne un échantillon $X_1, \ldots, X_n$ de loi F continue. On souhaite déterminer l'espérance de F, ainsi que l'intervalle de confiance à 95% associé à cette estimation. On simule un échantillon $X_1^*, \ldots, X_n^*$ suivant la loi F_n N fois. Pour chaque simulation, on calcule l'estimation $\hat{\theta}_n$ et la différence $\hat{\theta}_n - \theta$. On obtient donc un échantillon de taille N de la loi $\mathcal{L}^*$. Pour N assez grand, les quantiles empiriques approchent les quantiles q^* ; On peut donc construire notre intervalle de confiance. L'erreur causée ici est une erreur numérique et non une erreur statistique.

En premier exemple, on peut revenir sur notre échantillon suivant la loi normale du tableau 3. A l'aide de la fonction `boot(X,N)`, qui pour un échantillon `X` donne l'intervalle de confiance à 95% de la méthode *bootstrap*. On obtient alors :

$$I_{Boot} = [-2.336253, 1.317346].$$

Cet intervalle contient bien la valeur 0, mais il est beaucoup plus grand. En fait, pour les lois normales, la méthode du *théorème central limite* est presque exacte, et la

convergence se fait très rapidement. On la préférera au *bootstrap* dans ce cas. Cependant, on ne peut pas condamner la méthode de *bootstrap* avec un seul intervalle. Ce qui compte c'est que $\mathbb{P}(0 \in I)$ soit proche de 95%, mais rien ne garantie que l'on ait pas $\mathbb{P}(0 \in I_{TCL}) \leq \mathbb{P}(0 \in I_{Boot}) < 95\%$.

5.6 Comparaison de la loi exacte et de la loi *bootstrap*

Dans certain cas particulier, on peut calculer la loi de $\hat{\theta}_n - \theta$ de manière exacte. C'est le cas des loi normale que nous avons déjà traités, mais aussi des lois exponentielles. Pour commencer, rappelons l'expression de la loi Gamma [2].

Définition 15 (La loi Gamma)
Une variable aléatoire suit la loi Gamma $\Gamma(k, \theta)$, si sa densité est :

$$\forall\, t \in \mathbb{R}, \quad f(t) = \frac{t^{k-1} e^{-\frac{t}{\theta}}}{\Gamma(k)\theta^k} \mathbb{1}_{\mathbb{R}_+}(t).$$

Proposition 29
Soit $X_1, \ldots, X_n$ des variables aléatoires *iid* suivant la loi $\mathcal{E}(\frac{1}{\theta})$ (avec $\theta \in \mathbb{R}_+^*$). Alors $\frac{1}{n}\sum_{i=1}^{n} X_i$ suit la loi $\Gamma(n, n\theta)$.

Démonstration 26
Soit X et Y deux variables aléatoires *iid* suivant la loi $\mathcal{E}(\frac{1}{\theta})$. Les densités vérifient :

$$\begin{aligned}
\forall\, t \in \mathbb{R}, \quad f_{X+Y}(t) &= f_X * f_y(t) \\
&= \int_{\mathbb{R}} f_Y(u) f_X(t-u) \mathrm{d}u \\
&= \frac{1}{\theta^2} \int_{\mathbb{R}} e^{-\frac{u}{\theta}} e^{-\frac{t-u}{\theta}} \mathbb{1}_{\mathbb{R}_+}(u) \mathbb{1}_{\mathbb{R}_+}(t-u) \mathrm{d}u \\
&= \frac{e^{-\frac{t}{\theta}}}{\theta^2} \mathbb{1}_{\mathbb{R}_+}(t) \int_0^t \mathrm{d}u \\
&= \frac{t e^{-\frac{t}{\theta}}}{\theta^2}\, \mathbb{1}_{\mathbb{R}_+}(t)
\end{aligned}$$

On reconnaît la densité d'une loi $\Gamma(2, \theta)$.

Par récurrence, on montre que $\sum_{i=1}^{n} X_i$ suit la loi $\Gamma(n, \theta)$.

2. il existe un autre convention. C'est pourquoi, on précise celle adoptée.

Puis on identifie les lois grâce à la fonction de répartition :

$$\begin{aligned}\forall\, t \in \mathbb{R}, \quad \mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i \leq t) &= \mathbb{P}(\sum_{i=1}^{n} X_i \leq nt) \\ &= \int_0^{nt} \frac{u^{n-1}e^{-\frac{u}{\theta}}}{\Gamma(n)\theta^n}\mathrm{d}u \\ &= \int_0^{t} \frac{x^{n-1}e^{-\frac{x}{n\theta}}}{\Gamma(n)(n\theta)^n}\mathrm{d}x\end{aligned}$$

On reconnaît la densité de la loi $\Gamma(n, n\theta)$.

□

L'espérance d'une variable aléatoire suivant la loi $\mathcal{E}(\frac{1}{\theta})$ est θ. On connaît donc la loi de $\hat{\theta}_n - \theta$. Pour différentes valeurs de n, on peut comparer la loi exacte de $\hat{\theta}_n - \theta$ et la comparer avec la loi *bootstrap* de $\hat{\theta}_n^* - \hat{\theta}_n$.

Les figures 24, 25 et 26 représentent les densités des lois exactes et celles obtenues par méthode de *bootstrap* pour de échantillons de taille respectivement 5, 10 et 100.

On remarque que les densités collent de plus en plus. On peut aussi comparer directement les quantiles. Le tableau 4 donne les quantiles exacts de la loi $\Gamma(10, 10)$ et ceux de la loi *bootstrap* pour des échantillons de taille 10. Les quantiles choisis sont ceux qui nous intéressent lorsque l'on souhaite faire des intervalles de confiance à 5% ou bien à 10%.

Quantile	2.5%	5%	95%	97.5%
Exacts	-0.5204611	-0.4574594	0.5705216	0.7084803
bootstrap	-0.3839061	-0.3368135	0.3445148	0.4074367

TABLE 4 – comparaison des quantiles pour des échantillons de taille 10.

On se rend alors mieux compte que la méthode de *bootstrap* a besoin d'échantillons suffisamment grands pour être efficace. Les quantiles trouvés diffèrent trop de quantiles exacts pour les échantillons de taille 10.

5.7 Comparaison de la méthode *TCL*, et de la méthode *bootstrap*

On peut essayer de comparer les intervalles de confiance donnés par le *théorème central limite* et le *bootstrap*, en fonction de la taille des échantillons disponibles. Pour cela on calcule à quelle fréquence la valeur exacte se trouve dans l'intervalle de confiance, en fonction de la taille de l'échantillon. Théoriquement, avec des échantillons infinis, cette fréquence est le niveau de confiance de notre intervalle, ici 95%.

On prend ici l'exemple d'une loi exponentielle $\mathcal{E}(1)$, dont l'espérance est 1.

On remarque que de manière générale, les intervalles du *théorème central limite* sont plus rapidement fiable que ceux du *bootstrap*. C'est important de le remarquer

n	5	10	20	100	$+\infty$
TCL	0.888	0.922	0.934	0.946	0.95
bootstrap	0.760	0.836	0.887	0.928	0.95

TABLE 5 – Fréquence de validité de l'intervalle de confiance à 95%, en fonction de la taille de l'échantillon.

si l'on dispose d'échantillons très faibles comme cela peut être le cas en médecine, ou plus généralement lorsque l'échantillonnage est coûteux, ou naturellement limité.

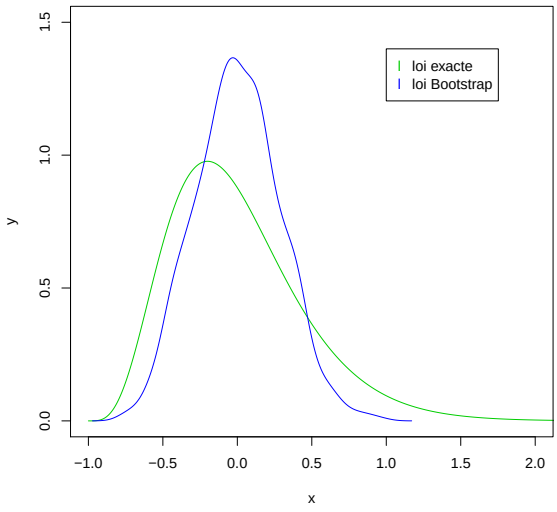

FIGURE 24 – comparaison des lois pour des échantillons de taille 5.

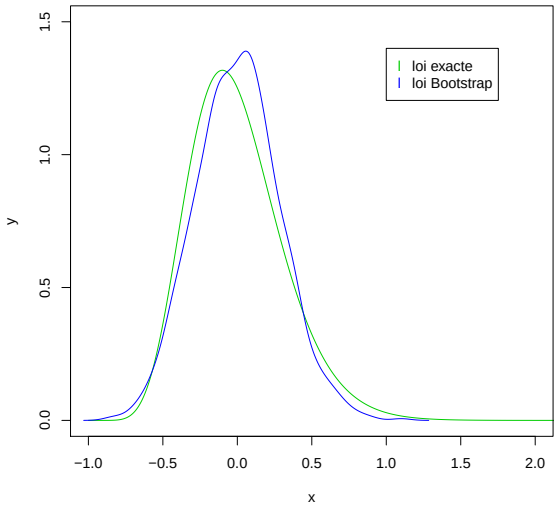

FIGURE 25 – comparaison des lois pour des échantillons de taille 10.

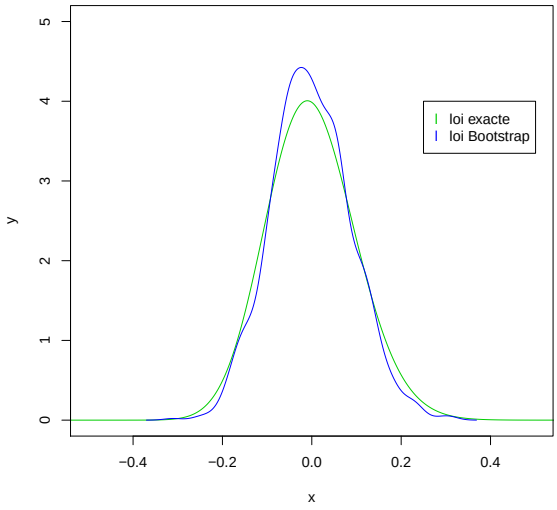

FIGURE 26 – comparaison des lois pour des échantillons de taille 100.

A Algorithmes sous $\mathcal{R}$

Voici le code des différentes fonctions utilisées.

A.1 Loi de *Kolmogorov*

Comme la fonction de répartition de la loi de *Kolmogorov* n'est connue que sous la forme de série, on ne peut qu'en avoir une estimation. Les fonctions `kolmF(t,n)` et `kolmf(t,n)`, qui donnent accès à une estimation de la fonction de répartition et à la densité, respectivement, en t, grâce à une troncature à partir du n-ième terme de la série.

```
kolmF=function(t,n){
  s=0
  for(i in 1:n){
    s=s+(-1)^(i)*exp(-2*i^2*t^2)
  }
  F=1+2*s
  F
}
```

```
kolmf=function(t,n){
  s=0   for(i in 1:n){
    s=s+(-1)^(i)*i^2*exp(-2*i^2*t^2)
  }
  F=-8*t*s
  F
}
```

On peut montrer qu'en choisissant n supérieur à 12, l'erreur est inférieure à $\frac{1}{\sqrt{1000}}$ pour tout $t \geq 0,1$.

A.2 Quantiles et test de *Kolmogorov-Smirinov*

La fonction `quant(a, eps)` permet de calculer le quantile `a` de la loi de *Kolmogorov* en se donnant une erreur `eps`.

```
quant=function(x,eps){
  a=0   b=3   while(abs(kolmF(c,1000)-x)>eps){
    c=(a+b)/2
    if(kolmF(c,1000)>x){b=c} else {a=c}
  }
```

```
  c
}
```

Le paramètre `n=1000` dans l'appel de la fonction `kolmF` a été choisi de manière arbitraire, il ne sert à rien de choisir un `eps` trop faible car l'erreur sera alors portée par `kolmF`.

La fonction `Cn(n,a,m)` calcule une approximation du quantile `a` exact du test de *Kolmogorov-Smirinov* pour des échantillons de taille `n`. Le paramètre `m` contrôle la précision.

```
Cn=function(n,a,m){
  rd=NULL
  for(i in 1:m){
    u=runif(n)
    u=sort(u)
    rd[i]=max(u[1], 1-u[n])
    for(k in 1:n-1){rd[i]=max(rd[i], abs(k/n-u[k]), abs(k/n-u[k+1]))}
  }
  rd=sort(rd)
  x=rd[floor(a*m)]
  x
}
```

A.3 Erreur du processus empirique

La première erreur à laquelle nous nous sommes intéressés est l'erreur ponctuelle. Pour cela, on utilise la fonction `errponct(X,n,F,t)`.

```
errponct=function(X,n,F,t){
 R=1:n
 for(i in 1:n){
   f=ecdf(X[1:i])
   R[i]=abs(f(t)-F(t))
 }
 R
}
```

La fonction `suppe(X,F=pnorm,...)` permet de calculer la distance uniforme entre la fonction de répartition empirique et la fonction de répartition, dans le cas où F est continue.

```
suppe=function(X,F=pnorm,...){
  n=length(X)
  R=rep(0,n)
  for(i in seq(1,n)){
```

```
    Y=X[1:i]
    Y=sort(Y)
    Z=(1:i)/i
    R[i]=max(abs(Z-F(Y,...)),abs(Z[1:i-1]-F(Y[2:i],...)),
                                        1-F(Y[i],...),F(Y[1],...))
  }
  R
}
```

On est ensuite amené à créer des échantillons de la loi D_n. Pour cela, on utilise la fonction `errunif(n,m,rP=rnorm,pP=pnorm,...)`. Elle aussi ne fonctionne que dans le cas des fonctions de répartition continues.

```
errunif=function(n,m,rP=rnorm,pP=pnorm,...){
 V=NULL
 for (k in 1:m){
   X=rP(n,...)
   R=suppe(X,pP,...)
   V=cbind(V,R)
 }
 matplot(V,type="l")
 V
}
```

Cette fonction possède un équivalent dans le cas des lois discrètes, la fonction `errunifdisc(n,m,rP=geom,pP=geom,...)` :

```
errunifdisc=function(n,m,rP=geom,pP=geom,...){
 V=NULL
 for (k in 1:m){
   X=rP(n,...)
   R=1:n
   for(i in seq(1,n)){
     Y=sort(X[1:i])
     f=ecdf(X[1:i])
     R[i]=max(abs(f(Y[1:i])-pP(Y[1:i],...)),1-pP(Y[i],...),
                                        pP(Y[1]-0.001,...))
   }
   V=cbind(V,R)   }
 matplot(V,type="l")
 V }
```

A.4 *bootstrap*

Le calcul des intervalles de confiance par la méthode *bootstrap* se fait au moyen de deux fonctions. La fonction `boot(X,N)` calcule une approximation de l'intervalle

de confiance par la méthode *bootstrap* de l'échantillon `X`, en réalisant un échantillon de taille`N` de $\hat{\theta}_n^* - \theta_n$. Cette fonction est spécifique à l'étude de l'espérance.

```
boot=function(X,N){
 n=length(X)
 m=mean(X)
 B=replicate(N, fun(X,m,n))
 I=c(m-quantile(B,0.975, names = FALSE),m-quantile(B,0.025, names
  = FALSE))
 I
}
```

La méthode de *bootstrap* à proprement parlé est réalisé par la fonction auxiliaire `fun(X,m,n)`, qui crée un échantillon *bootstrap* de l'échantillon `X`, de taille `n` et de moyenne `m`.

```
fun=function(X,m,n){
 Y=sample(X, n, replace = TRUE, prob = NULL)
 mean(Y)-m
}
```

La fonction `freqboot(n)` génère 10000 échantillons de taille `n` suivant la loi $\mathcal{E}(1)$. Elle calcule l'intervalle de confiance à 95% et permet donc d'approcher la fréquence de pertinence des intervalles.

```
freqboot=function(n){
 f=0
 for(i in 1:10000){
   x=rexp(n,1)
   I=boot2(x,1000)
   if((I[1]<=1) && (I[2]>=1)){f=f+1}
 }
 f=f/10000
 f
}
```

Pour faire la comparaison entre les deux méthodes, voici la même fonction adaptée au TCL.

```
freqtcl=function(n){
 q=qnorm(0.975)
 f=0
 for(i in 1:10000){
   x=rexp(n,1)
   if((mean(x)-q*sqrt(sd(x))/sqrt(n)<=1)&& (mean(x)+q*
                                    sqrt(sd(x))/sqrt(n)>=1)){f=f+1}
 }
```

```
  f=f/10000
  f
}
```

Conclusion

En définitive, le processus empirique possède des bonnes propriétés de convergence. Qu'il s'agisse de loi discrète ou de loi continue, des échantillons de taille de l'ordre de 10^3 apportent des approximations très pertinentes. Dans le cas des lois continues, on est même capable de maîtriser la répartition des erreurs à la manière du TCL.

Le processus empirique peut être utilisé à des fins statistiques. Le test de *Kolmogorov-Smirinov* en est une application directe dans le cas des lois continues. Ce test permet alors de vérifier si un échantillon suit une loi continue donnée. La méthode de *bootstrap* quant à elle se base sur la convergence établies de la fonction de répartition empirique pour obtenir des intervalles de confiance aux estimateurs statiques.

Le cas des lois discrètes reste encore très flou. Notamment en ce qui concerne les lois limites de D_n ou de $\sqrt{n}D_n$. Peut-on décrire ces lois dans le cas général, ou même sur des cas particuliers simples ? Comment dépendent-elles de la loi F ? Peut-on en tirer des outils en statistique ?

Remerciements

Je tiens à remercier Anne Philippe de m'avoir pris en stage, de m'avoir accordé du temps et d'avoir guidé mon travail durant six semaines. Je voudrais aussi remercier toute l'équipe des doctorants de l'université de Nantes qui m'a fait découvrir un peu leur monde et avec qui j'ai passé des bons moments.

Références

[1] Jun SHAO. *Mathematical Statistics*. Springer, 1999.

[2] Michael R. KOSOROK. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

[3] B. RIPLEY. *Stochastic Simulation*. J. Wiley, New-York, 1987.

[4] Christian P. ROBERT et George CASELLA. *Monte Carlo Statistical Methods*. Springer, 1999.

[5] X. MILHAUD. *Statistque*. Belin, 2001.

[6] Marek FISZ. *Probability Theory and Mathematical Statistics*. John Wiley and sons, Inc. troisième édition, 1963.

[7] A.C. DAVISON et D.V. HINKLEY. *bootstrap Methods and their Application*. Cambridge university press, 1997