Théorème central limite et ses applications

Caroline Robet

4 juin 2014

Table des matières

Introduction

Le problème statistique qui nous intéresse consiste à déterminer avec une marge de confiance un paramètre inconnu.

On considère X_1 ... X_n des variables aléatoires indépendantes et identiquement distribuées de densité commune f. On se donne de plus une famille $\mathcal{P}$ de densités telle que $f \in \mathcal{P}$ avec $\mathcal{P} = \{f_\theta \ , \ \theta \in \Theta\}$. On suppose qu'il existe un unique θ_0 tel que $f = f_{\theta_0}$. Autrement dit, que l'application $\theta \longrightarrow f_\theta$ soit injective.

Par exemple, on peut avoir $\Theta = (m, \sigma^2)$ et $\mathcal{P} = \{N(m, \sigma^2)\}$. On cherche alors f parmi l'ensemble des lois normales.

Pour cela, on utilise un estimateur de θ. Si on continue l'exemple avec $\sigma^2 = 1$, on a $f_\theta \sim N(\theta, 1)$ et $\theta = E[X_1]$ alors on sait d'après la loi des grands nombres que $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \theta$. On peut donc ici choisir comme estimateur de θ, $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Pour valider le choix de $\mathcal{P}$, on essaie également à partir de variables aléatoires dont on suppose qu'elles sont indépendantes et identiquement distribuées de comparer leur loi à des lois usuelles grâce à un test, appelé test de Kolmogorov-Smirnov.

1 Théorème central limite

1.1 énoncé du théorème central limite

Définition 1 (convergence en loi)**.**
Soit $(X_n)_n$ une suite de variables aléatoires réelles de fonction de répartition F_n et X une variable aléatoire réelle de fonction de répartition F. On dit que X_n converge en loi vers X si $\lim_{n\to+\infty} F_n(t) = F(t)$ pour tout t tel que F est continue en t. On note $X_n \Longrightarrow X$ ou $X_n \xrightarrow{\mathcal{L}} X$

Théorème 2 (théorème central limite, TCL)**.**
Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées (iid) de moyenne $m = E[X_1]$ et de variance $\sigma^2 = var(X_1) < +\infty$ alors on a

$$\frac{\sum_{i=1}^{n} X_i - nm}{\sqrt{n\sigma^2}} \Longrightarrow N(0,1)$$

ou de manière équivalente

$$\frac{\sqrt{n}(\overline{X} - m)}{\sigma} \Longrightarrow N(0,1)$$

avec $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

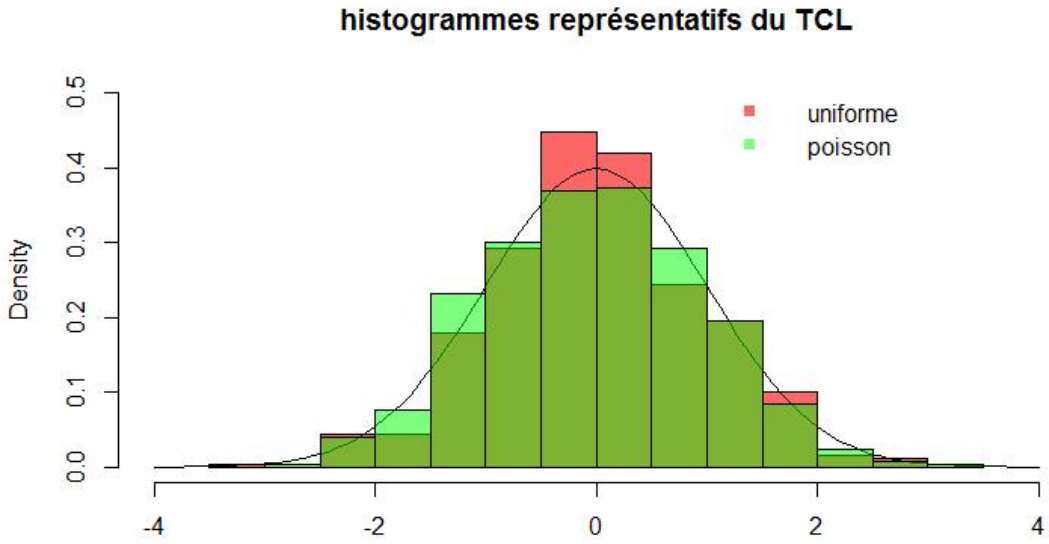
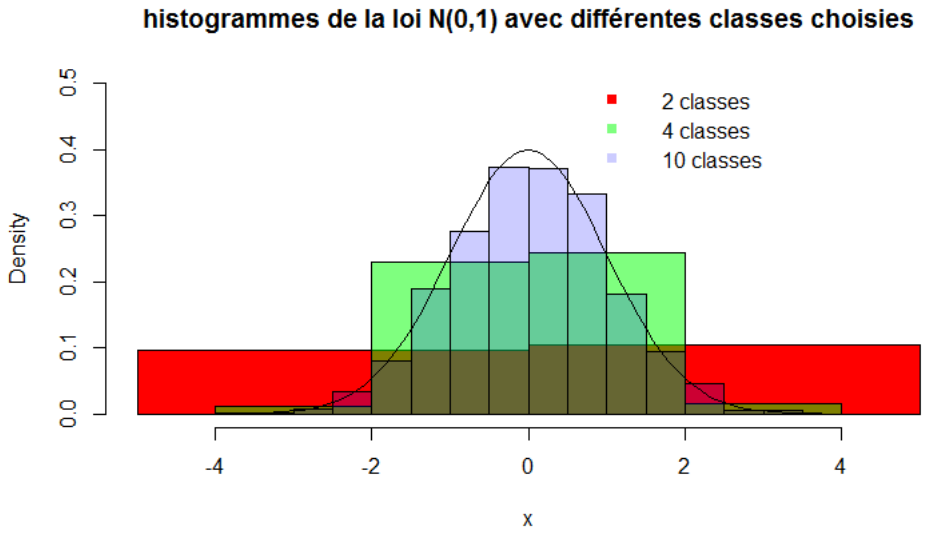


FIGURE 1 – histogrammes représentatifs du TCL pour la loi uniforme et la loi de poisson

Sur la figure 1 ci-dessus, on a illustré le TCL pour n=10000 en représentant pour les deux lois l'histogramme d'un échantillon simulé suivant la loi de $\frac{\sum_{i=1}^{n} X_i - nm}{\sqrt{n\sigma^2}}$. On observe ainsi la convergence vers la loi normale.

Le logiciel R choisit le nombre de classe de manière à optimiser la correspondance avec le TCL mais on constate que pour un nombre de classe fixée, on peut avoir parfois un histogramme peu ressemblant à la loi N(0,1).



Par la suite, on ne se souciera pas du nombre de classe choisi.

Application : intervalle de confiance
Le théorème central limite nous donne pour l'estimateur empirique $T_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ que

$$\sqrt{n}\frac{(T_n - \theta)}{\sigma} \xrightarrow{Loi} N(0,1)$$

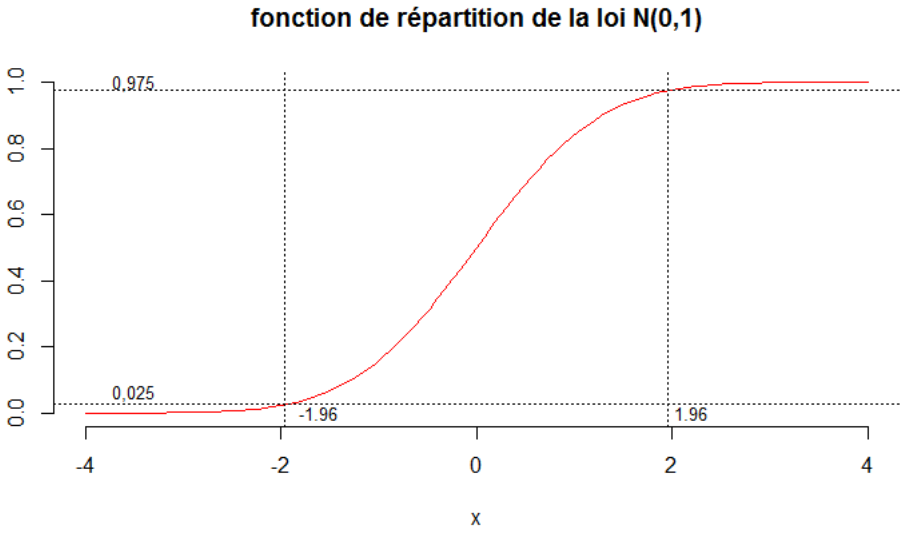
Si on appelle Φ la fonction de répartition de la loi N(0,1) alors on a

$$\mathbb{P}_\theta(\sqrt{n}\frac{(T_n - \theta)}{\sigma} \leq t) \to \Phi(t)$$

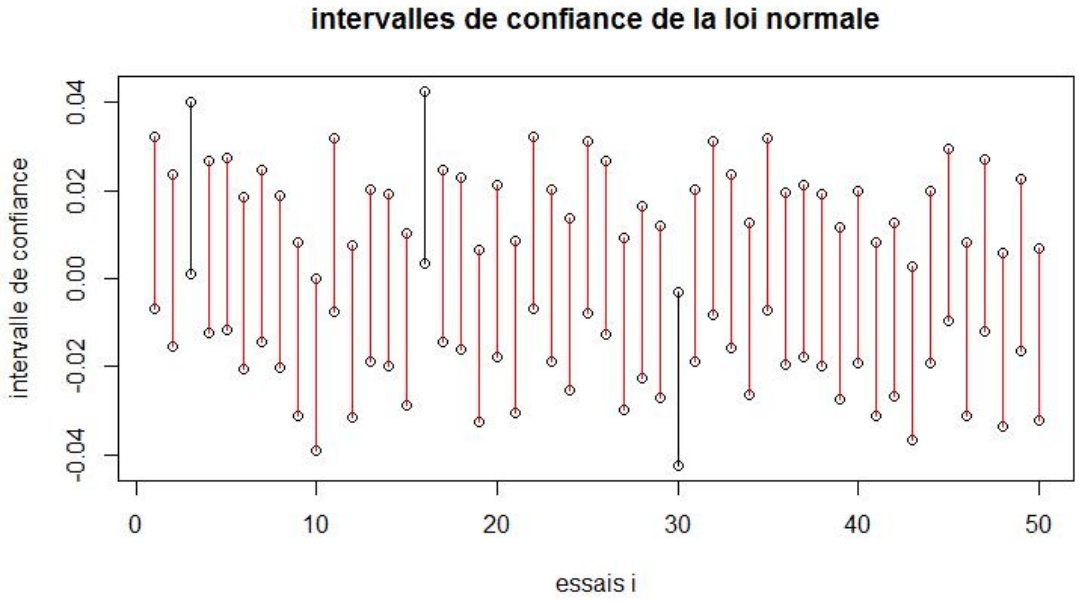
pour tout t $\in \mathbb{R}$ On appelle q_α le quantile d'ordre α c'est à dire la valeur pour laquelle on ait $\Phi(q_\alpha) = \alpha$. On a alors

$$\mathbb{P}_\theta(\sqrt{n}\frac{(T_n - \theta)}{\sigma} \in [q_\alpha, q_{1-\alpha}]) \to \Phi(q_{1-\alpha}) - \Phi(q_\alpha) = 1 - 2\alpha$$

donc $\forall\theta$, $\mathbb{P}_\theta(\theta \in [T_n - \frac{\sigma q_{1-\alpha}}{\sqrt{n}}; T_n - \frac{\sigma q_\alpha}{\sqrt{n}}]) \to 1 - 2\alpha$ avec α fixé on obtient ainsi un intervalle de confiance sur la valeur de θ qui est recherchée.
On cherche souvent un intervalle de confiance à 95%.
On choisit donc $\alpha = 0.025$ et on obtient ainsi les quantiles suivants :



On obtient alors $q_\alpha = -1.96$ et $q_{1-\alpha} = 1.96$. Si on applique désormais le TCL à une suite de variables aléatoires $(X_i)_{i\geq 1}$ avec des lois gaussiennes, on a $\sqrt{n}\frac{T_n - \theta}{\sigma}$ suit une loi N(0,1). La loi limite du TCL devient la loi exacte pour tout n. On a alors les intervalles de confiance de 50 simulations suivants :



Les intervalles de confiance qui contiennent notre valeur cherchée de θ c'est-à-dire 0 sont représentés en rouge alors que ceux qui ne contiennent pas θ sont en noirs. On obtient alors ici une fréquence de 0,96 qui est proche de la valeur théorique 1-2α

1.2 lien entre la loi normale et la loi du χ^2_n

On considère X une variable aléatoire suivant la loi du χ^2_n. La densité de X est alors $f(x,n) = \frac{(x/2)^{\frac{n}{2}-1}e^{-x/2}}{2\Gamma(\frac{n}{2})}\mathbb{1}_{\mathbb{R}^+}(x)$

Théorème 3. *Soit $(N_i)_{i\geq 1}$ des variables aléatoires i.i.d. suivant une loi N(0,1) alors $X = N_1^2 + ... + N_n^2$ suit la loi χ^2_n.*

Preuve

On va montrer que la fonction caractéristique pour une variable suivant la loi du χ^2_n est égale à celle pour une somme de n carrés de loi N(0,1) indépendantes.

- fonction caractéristique pour χ^2_n

  Soit $\phi_1(t) = E[e^{itx}] = \frac{1}{2\Gamma(n/2)}\int_0^{+\infty}\left(\frac{x}{2}\right)^{\frac{n}{2}-1}e^{-(\frac{1}{2}-it)x}dx$

On pose $y = (\frac{1}{2} - it)x$
$\phi_1(t) = \frac{1}{2\Gamma(n/2)} \int_0^{+\infty} (\frac{y}{1-2it})^{\frac{n}{2}-1} e^{-y} \frac{dy}{\frac{1}{2}-it}$
$\phi_1(t) = \frac{1}{2\Gamma(n/2)(1-2it)^{n/2}} \int_0^{+\infty} y^{\frac{n}{2}-1} e^{-y} dy$
$\phi_1(t) = \frac{1}{2\Gamma(n/2)(1-2it)^{n/2}} \Gamma(n/2)$
$\phi_1(t) = (1-2it)^{\frac{-n}{2}}$

– fonction caractéristique pour $Z = X_1^2 + ... + X_n^2$ avec $(X_i)_{i\geq 1}$ i.i.d. de loi N(0,1)
$\phi_2(t) = E[e^{itz}] = (E[e^{itx^2}])^n$ car on a n variables aléatoires indépendantes de même loi .
$\phi_2(t) = (\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{itx^2} dx)^n$
$\phi_2(t) = (\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}(1-2it)} dx)^n$
$\phi_2(t) = (\frac{1}{\sqrt{2\pi}} \sqrt{\frac{2\pi}{1-2it}})^n$
$\phi_2(t) = (1-2it)^{\frac{n}{2}}$

– on a bien $\phi_1 = \phi_2$ or les fonctions caractéristiques déterminent les lois donc $\chi_n^2 \sim X_1^2 + ... + X_n^2$ avec (X_i) i.i.d. de même loi N(0,1)

1.3 réciproque du théorème central limite

Lemme 4. *Soit X une variable aléatoire réelle de fonction caractéristique φ. Alors* $\lim_{u\to 0} \frac{2(1-Re\varphi(u))}{u^2} = \int_{\mathbb{R}} x^2 dP_X(x) = E[X_1^2] \in \overline{\mathbb{R}}_+$

Preuve
Pour $u \neq 0$ on a

$$I(u) := \frac{2(1-Re\varphi(u))}{u^2} = \frac{2}{u^2}(1 - \int_{\mathbb{R}} cos(ux) dP_X(x))$$
$$= 2\int_{\mathbb{R}} \frac{1-cos(ux)}{u^2} dP_X(x)$$

On remarque alors que :

$$\frac{1-cos(ux)}{u^2} = \frac{2sin^2(ux/2)}{u^2} \leq \frac{2}{u^2}\left(\frac{ux}{2}\right)^2 = \frac{x^2}{2}$$

On a toujours

$$\int_{\mathbb{R}} x^2 dP_X(x) \leq \liminf_{u\to 0} I(u) \leq \limsup_{u\to 0} I(u) \leq \int_{\mathbb{R}} x^2 dP_X(x)$$

où la première inégalité résulte du lemme de Fatou et la dernière de l'inégalité précédente. Le lemme est prouvé

Lemme 5. *Si X et Y sont deux variables aléatoires indépendantes telles que $X+Y \in L^2(\Omega)$, alors $X \in L^2(\Omega)$ et $Y \in L^2(\Omega)$*

Preuve Compte tenu de l'indépendance de X et Y, une simple application du théorème de Fubini nous donne

$$\int_{\mathbb{R}^{\not\vdash}} (x+y)^2 P_{X,Y}(dx,dy) = \int_{\mathbb{R}^{\not\vdash}} (x+y)^2 P_X(dx)P_Y(dy) = \int_{\mathbb{R}} P_X(dx) \int_{\mathbb{R}} (x+y)^2 P_Y(dy)$$

Par conséquent,

$$E[(X+Y)^2] = \int_{\mathbb{R}} P_X(dx) E[(x+Y)^2] < +\infty$$

d'où $E[(x+Y)^2] < +\infty$ pour P_X-presque tout x$\in \mathbb{R}$, donc au moins pour un x_0. Ainsi $x_0 + Y$ est dans $L^2(\Omega)$ et comme les constantes sont dans l'espace vectoriel $L^2(\Omega)$, il en va de même pour Y=$(x_0 + Y) - x_0$. Enfin l'égalité X=(X+Y)-Y donne l'appartenance de X à $L^2(\Omega)$

Théorème 6 (réciproque du théorème central limite)**.** *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. définies sur le même espace probabilisé et telle que $\frac{S_n}{\sqrt{n}}$ converge en loi vers $N(0,1)$. Alors X_1 est de carré intégrable, $E[X_1^2] = 1$ et $E[X_1] = 0$.*

Preuve

La démonstration comporte 5 étapes dont l'utilisation des deux lemmes précédents, le lemme 4 et le lemme 5. On commence par réduire le problème à la preuve de l'appartenance de X_1 à $L^2(\Omega)$. Le lemme 4 établit ensuite un lien entre le comportement de la fonction caractéristique de X_1 au voisinage de 0 et $E[X_1^2]$. On peut alors achever la preuve dans le cas particulier où X_1 est de loi symétrique. Le lemme 5 est utile pour le passage du cas symétrique au cas général.

- *Réduction de la preuve à $E[X_1^2] < +\infty$*

Supposons établie l'appartenance de X_1 à $L^2(\Omega)$. On a alors $E[|X_1|] < +\infty$

donc par la loi forte des grands nombres, $\frac{S_n}{n}$ converge p.s. vers $E[X_1]$, donc aussi en loi vers la même limite. D'autre part en écrivant $\frac{S_n}{n} = \frac{1}{\sqrt{n}}(\frac{S_n}{\sqrt{n}})$, l'hypothèse de convergence de $\frac{S_n}{\sqrt{n}}$ vers Z de loi N(0,1) et le lemme de Slutsky (sous une forme dégénérée) nous donnent la convergence en loi de $\frac{S_n}{n}$ vers $0 * Z = 0$. Par unicité de la loi limite, les deux variables aléatoires constantes $E[X_1]$ et 0 doivent avoir même loi. Ceci implique $E[X_1] = 0$ Ensuite par le T.L.C., $\frac{S_n}{\sqrt{n}}$ converge en loi vers N(0,$E[X_1^2]$) Par unicité de la loi limite, on doit donc avoir N(0, 1) = N(0,$E[X_1^2]$), d'où $E[X_1^2] = 1$

-*Cas où X_1 a une loi symétrique*

Notons $\varphi(u) := E[exp(iuX_1)]$ la fonction caractéristique de X_1 et φ_n celle de $S_n/\sqrt{n}$ Comme les X_k sont i.i.d., $\varphi_n(u) = \varphi(u/\sqrt{(n)})^n$. De plus, X_1 étant symétrique, φ et φ_n sont réelles. L'hypothèse de convergence en loi de $S_n/\sqrt{n}$ vers N(0,1) équivaut à la convergence ponctuelle de $\varphi_n(u)$ vers $exp(-u^2/2)$. En particulier pour u=1 on a $\varphi(1/\sqrt{n})^n \to exp(-1/2)$ et $\varphi(1/\sqrt{n})^n > 0$ pour $n \geq n_0$. On peut passer au logarithme et en déduire $ln\varphi(1/\sqrt{n}) \sim -1/(2n)$, puis $\varphi(1/\sqrt{n}) - 1 \sim -1/(2n)$. En réécrivant ceci sous la forme

$$\lim_{n\to+\infty} \frac{2(1-\varphi(1/\sqrt{n}))}{\frac{1}{n}} = 1$$

le lemme 4 nous donne alors $E[X_1^2] = 1$

-*Cas général*
Considérons les variables aléatoires

$$Z_n := \frac{(X_1 - X_2) + (X_3 - X_4) + ... + (X_{2n-1} - X_{2n})}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\left(\frac{S'_n}{\sqrt{n}} - \frac{S''_n}{\sqrt{n}}\right)$$

où

$$S'_n := \sum_{k=1}^{n} X_{2k-1}, \qquad S''_n := \sum_{k=1}^{n} X_{2k}$$

Par hypothèse, $\frac{S'_n}{\sqrt{n}}$ et $\frac{S''_n}{\sqrt{n}}$ convergent en loi vers N(0,1). Comme elles sont indépendantes, on en déduit la convergence en loi du *vecteur* aléatoire (S'_n, S''_n) vers (Z', Z'') où Z' et Z'' sont deux variables aléatoires *indépendantes* de même loi N(0,1). On en déduit par image continue que

$$Z_n = \frac{1}{\sqrt{2}}\left(\frac{S'_n}{\sqrt{n}} - \frac{S''_n}{\sqrt{n}}\right) \Longrightarrow \frac{1}{\sqrt{2}}(Z' - Z'')$$

On peut vérifier (grâce à la fonction caractéristique) que $(Z' - Z'')/\sqrt{2}$ a pour loi N(0,1). Notons maintenant

$$Y_k := \frac{X_{2k-1} - X_{2k}}{\sqrt{2}}, \quad k \geq 1$$

et remarquons que puisque X_{2k-1} et X_{2k} sont indépendantes et de même loi, la loi de Y_k est symétrique. On vient donc de montrer que

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k \Longrightarrow N(0,1)$$

Par le cas symétrique, on en déduit que $E[Y_1^2] < +\infty$. Le lemme 5 appliqué à $X = X_1$ et $Y = -X_2$ nous donne $E[X_1^2] < +\infty$, ce qui achève la preuve.

2 Delta méthode

2.1 Énoncé de la méthode

Théorème 7 (delta méthode)**.**
Soit f une fontion C^1 et $(T_n)_n$ une suite de variables aléatoires.
Si $\sqrt{n}(T_n - \theta) \Longrightarrow N(0, \tau^2)$, $f'(\theta)$ existe et est non nul alors
$\sqrt{n}(f(T_n) - f(\theta)) \Longrightarrow N(0, \tau^2 f'(\theta)^2)$.

Preuve

- On pose $Z_n = \frac{\sqrt{n}}{\tau}(T_n - \theta) \Longrightarrow Z \sim N(0,1)$ T_n et Z sont définies sur le même espace probabilisé $(\Omega, A, \mathbb{P})$. Par le théorème de Skorokhod, il existe $(\Omega', A', \mathbb{P}')$ et Z'_n et Z' définies sur cet espace telles que Z_n soit de même loi que Z'_n , Z soit de même loi que Z' et $Z'_n \overset{\mathbb{P}'-p.s.}{\longrightarrow} Z'$
- On pose $Y'_n = \tau \frac{Z'_n}{\sqrt{n}} + \theta$, comme $T_n = \tau \frac{Z_n}{\sqrt{n}} + \theta$ et que Z_n et Z'_n ont même loi alors on a T_n et Y'_n ont même loi.
- De plus, $Z'_n \overset{\mathbb{P}'-p.s.}{\longrightarrow} Z'$ donc $\frac{\sqrt{n}}{\tau}(Y'_n - \theta) \overset{\mathbb{P}'-p.s.}{\longrightarrow} Z'$ comme $\frac{\sqrt{n}}{\tau} \underset{n\to+\infty}{\to} +\infty$ on a $\mathbb{P}' - p.s.$ $Y'_n - \theta \to 0$ donc $Y'_n \overset{\mathbb{P}'-p.s.}{\longrightarrow} \theta$
- f est définie et dérivable au voisinage de θ. $\exists \alpha > 0, \forall x \in]\theta - \alpha,\ \theta + \alpha[$ $f(x) - f(\theta) = f'(\theta)(x - \theta)(1 + \varepsilon(x))$ avec $\lim\limits_{x\to\theta} \varepsilon(x) = 0$

– Comme Y'_n converge $\mathbb{P}' - p.s.$ vers θ, $\exists n_0(\omega, \alpha)$ tel que $\forall n \geq n_0(\omega, \alpha)$ $Y'_n(\omega) \in]\theta - \alpha,\ \theta + \alpha[$ donc $\frac{f(Y'_n(\omega)) - f(\theta)}{f'(\theta)} = (Y'_n(\omega) - \theta)(1 + \varepsilon(Y'_n(w)))$
On en déduit que pour $\mathbb{P}'$-presque tout $\omega \in \Omega'$, $\forall n \geq n_0(\omega, \alpha)$, $\sqrt{n}\frac{f(Y'_n(\omega)) - f(\theta)}{\tau f'(\theta)} = Z'_n(w)(1 + \varepsilon(Y'_n(w)))$ par conséquent $\sqrt{n}\frac{f(Y'_n(\omega)) - f(\theta)}{\tau f'(\theta)} \overset{\mathbb{P}'-p.s.}{\longrightarrow} Z'$ or la convergence $\mathbb{P}'$-p.s. entraîne la convergence en loi donc $\sqrt{n}\frac{f(Y'_n(\omega)) - f(\theta))}{\tau f'(\theta)} \implies Z'$ or $Z' \sim Z$ et $Y'_n \sim T_n$ donc $\sqrt{n}\frac{f(Y'_n(\omega)) - f(\theta)}{\tau f'(\theta)} \implies N(0, 1)$ d'où $\sqrt{n}(f(T_n) - f(\theta)) \implies N(0, \tau^2 f'(\theta)^2)$

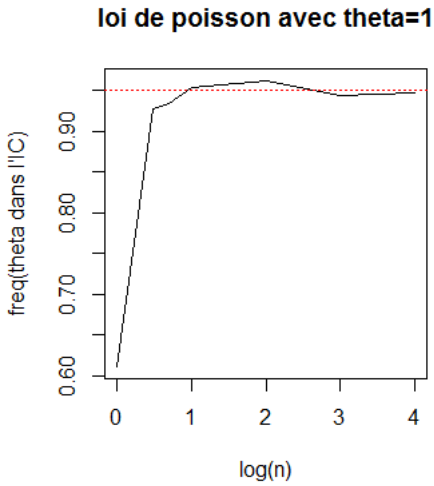
Théorème 8 (delta méthode d'ordre 2)**.**
Soit f une fontion C^2 et $(T_n)_n$ une suite de variables aléatoires. Si $\sqrt{n}(T_n - \theta) \implies N(0, \tau^2)$, $f'(\theta) = 0$ et $f''(\theta) \neq 0$ alors

$$n(f(T_n) - f(\theta)) \implies \frac{1}{2}\tau^2 f''(\theta)\chi_1^2$$

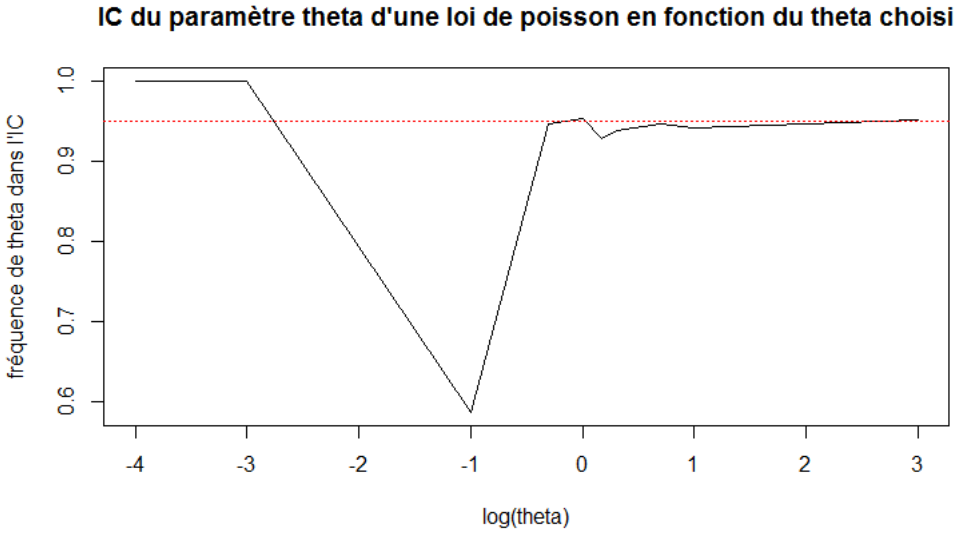
La preuve est similaire à celle du théorème 7

2.2 Application à la loi de poisson

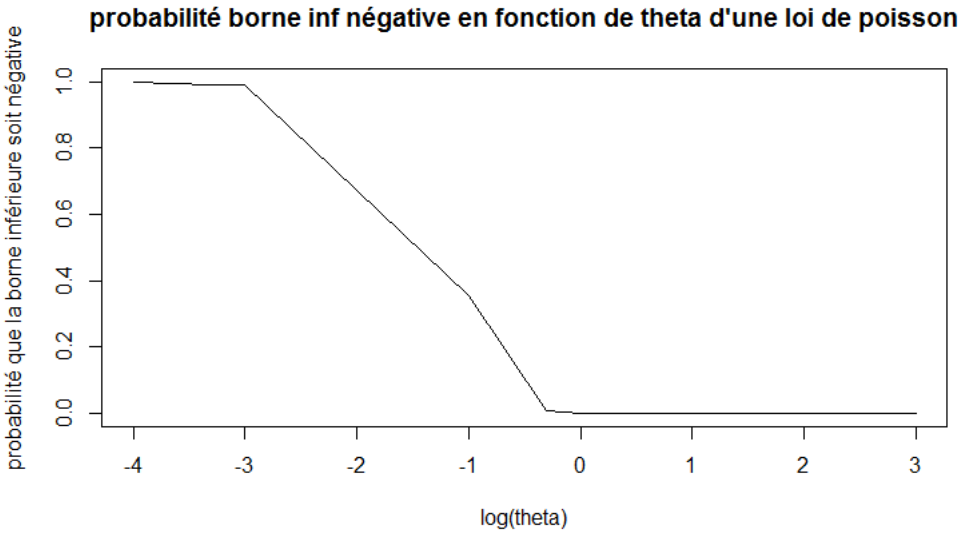
On utilise la delta-méthode pour la loi de poisson car d'après le TCL on a : $\sqrt{n}(\overline{X} - \theta) \implies N(0, \theta)$ et d'après la delta-méthode on sait que $\sqrt{n}(f(\overline{X}) - f(\theta)) \implies N(0, \theta f'(\theta)^2)$.
On cherche ainsi f tel que $\theta f'(\theta)^2 = 1$. On peut donc choisir f telle que $f(x) = 2\sqrt{x}$.
On a alors $2\sqrt{n}(\sqrt{\overline{X}} - \sqrt{\theta}) \implies N(0, 1)$
On obtient ainsi $\mathbb{P}(\sqrt{\theta} \in [\sqrt{\overline{X}} - \frac{1.96}{2\sqrt{n}}, \sqrt{\overline{X}} + \frac{1.96}{2\sqrt{n}}] \underset{n \to +\infty}{\longrightarrow} 0.95$
Pour $\theta = 1$ on obtient des fréquences suivant n avec une asymptote à 0.95 comme ce qui était prévu avec la delta méthode :



On fait désormais varier θ et on garde n fixé n=10



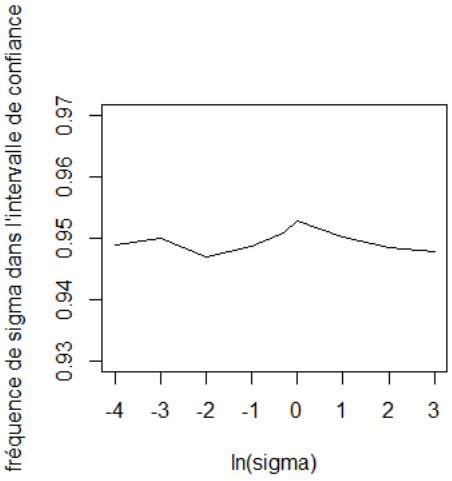
On observe alors un décrochage autour de $\theta = \frac{1.96}{2\sqrt{n}}$ On peut expliquer ce décrochage par le fait que pour θ petit la borne inférieure de l'intervalle est en moyenne négative comme on estime la racine de θ cette borne est inutile lorsqu'elle est négative. Le décrochage survient donc lorsque la borne inférieure devient positive car alors on impose une condition plus forte.



Pour un θ fixé, on peut toujours trouver un n assez grand pour éviter ce problème.

2.3 Application à la loi du χ^2

On considère $Y_i = X_i^2$ où les X_i sont i.i.d de loi N(0,σ^2). On a alors $E[Y_i] = \sigma^2$ et $var(Y_i) = 2\sigma^4$. D'après le TCL, on a $\sqrt{n}(\overline{Y} - \sigma^2) \Longrightarrow N(0, 2\sigma^4)$. On applique la delta méthode avec $f(x) = \frac{ln(x)}{\sqrt{2}}$.
On a alors $\sqrt{n/2}\ ln(\frac{\overline{Y}}{\sigma^2}) \Longrightarrow N(0,1)$

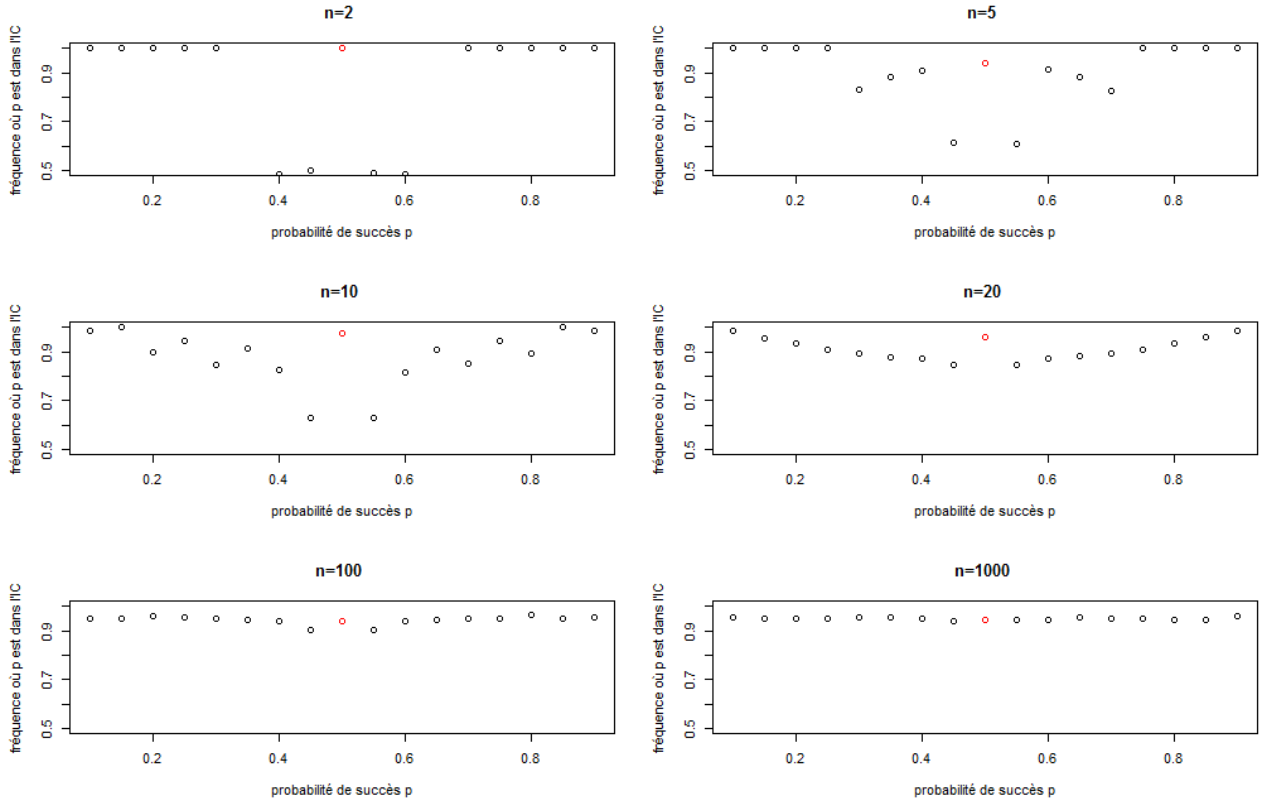


Pour toute les valeurs de σ, on observe une fréquence autour de 0,95 donc cette méthode converge bien pour la loi du χ^2

2.4 Application à la loi binomiale

On s'intéresse à la loi binomiale pour laquelle on veut estimer p avec un intervalle de confiance à 95%. Soit X suivant une loi binomiale B(n,p). On a alors d'après le TCL $\sqrt{n}\left(\frac{X}{n} - p\right) \Longrightarrow N(0, p(1-p))$
Pour $p \neq 0.5$, on prend f(x)=x(1-x) d'où $\sqrt{n}\left(\frac{X}{n}\left(1-\frac{X}{n}\right) - p(1-p)\right) \Longrightarrow N(0, p(1-p)(1-2p)^2)$. On peut alors appliquer la delta méthode avec f pour trouver la variance p(1-p)
Si $p = \frac{1}{2}$ on a $f'(p) = 0$, on doit donc appliquer la delta méthode d'ordre 2. On a $n\left(\frac{X}{n}\left(1-\frac{X}{n}\right) - p(1-p)\right) \Longrightarrow \frac{1}{4}\chi^2_1$.

On obtient alors une meilleure convergence car on converge en n alors que pour $p \neq \frac{1}{2}$ on a une convergence en $\sqrt{n}$.



On représente sur chacun des graphiques ci-dessus, la fréquence où la probabilité p de succès soit dans l'intervalle de confiance en fonction de la valeur p. On observe ainsi toujours une fréquence autour de 0.95 pour p=0.5 même pour des valeurs faibles de n alors que pour les autres valeurs, la fréquence descend parfois à seulement 0.5. En revanche pour les grandes valeurs de n, la différence entre les 2 ordres de la méthode n'est plus visible.

3 Estimation suivant trois méthodes

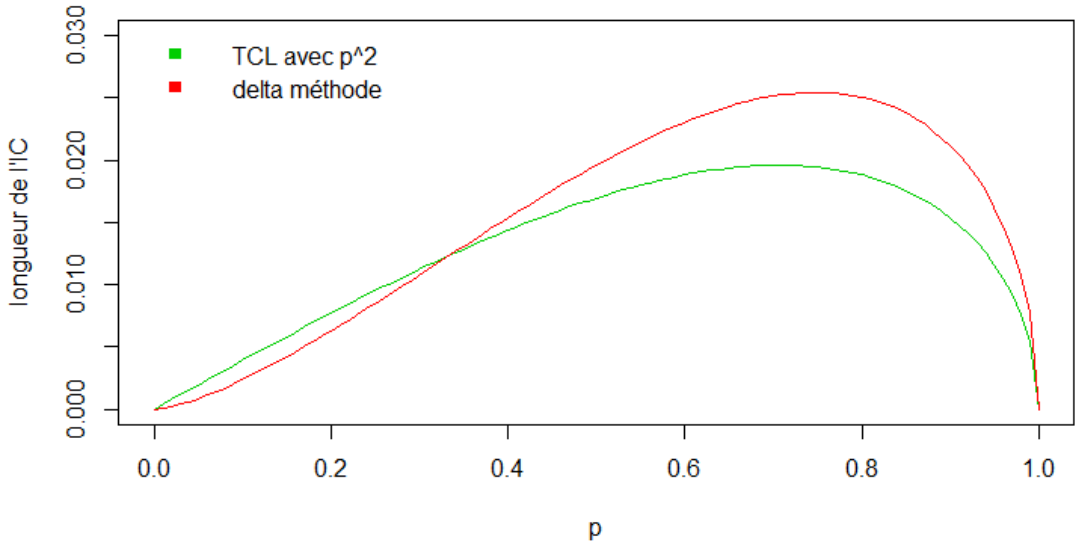
3.1 théorème de Slutsky

Théorème 9 (Slutsky). *Si X_n converge en loi vers X, et si Y_n converge en probabilité vers une constante c, alors le couple (X_n, Y_n) converge en loi vers le couple (X,c).*
Ce qui entraine $X_n + Y_n \Longrightarrow X + c$, $X_n Y_n \Longrightarrow cX$, $X_n/Y_n \Longrightarrow X/c$ si $c \neq 0$

3.2 comparaison des méthodes

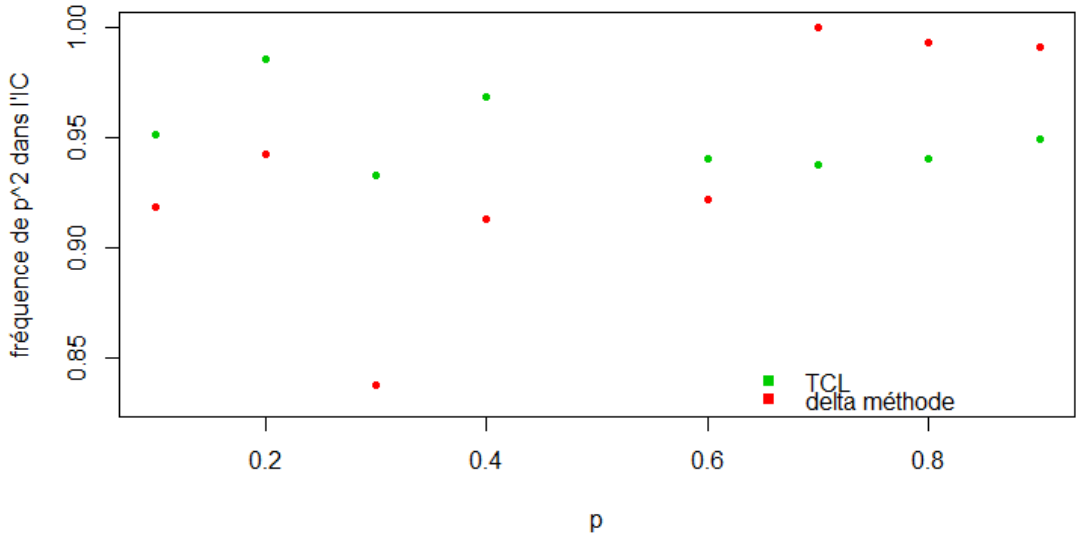
3.2.1 Deux méthodes pour estimer p^2 pour une binomiale

- on peut choisir de réaliser n expériences de Bernoulli de probabilité p^2
  On a alors $\sqrt{n}\left(\frac{X}{n} - p^2\right) \Longrightarrow N(0, p^2(1-p^2))$ avec X la somme des n variables aléatoires de Bernoulli donc une binomiale B(n ;p^2)
- on peut choisir de réaliser n expériences de Bernoulli de probabilité p de succès. On note Y la somme.
  On a alors $\sqrt{n}\left(\left(\frac{Y}{n}\right)^2 - p^2\right) \Longrightarrow N(0, p(1-p)4p^2)$ d'après la delta-méthode



L'intervalle de confiance pour l'estimation de p^2 est plus ou moins grand suivant la valeur de p ainsi que suivant la méthode. On a un changement pour

$p = \frac{1}{3}$: en dessous de $\frac{1}{3}$, la delta-méthode semble plus précise car on estime p^2 dans un intervalle plus petit. Et inversement, au-dessus de $\frac{1}{3}$, le TCL est plus précis.
Cependant, bien que l'intervalle soit plus précis, on constate sur le graphique ci-dessous des différences de fréquence où p^2 est dans l'intervalle de confiance. Par exemple, pour n=5 on est censé avoir une meilleure estimation pour p=0.3 avec la delta-méthode mais on constate que la fréquence est en réalité plus faible que pour le TCL. En calculant des intervalles de confiance avec le TCL ou plus généralement à partir de convergence en loi, on n'a pas d'information sur la qualité de la méthode pour n petit.

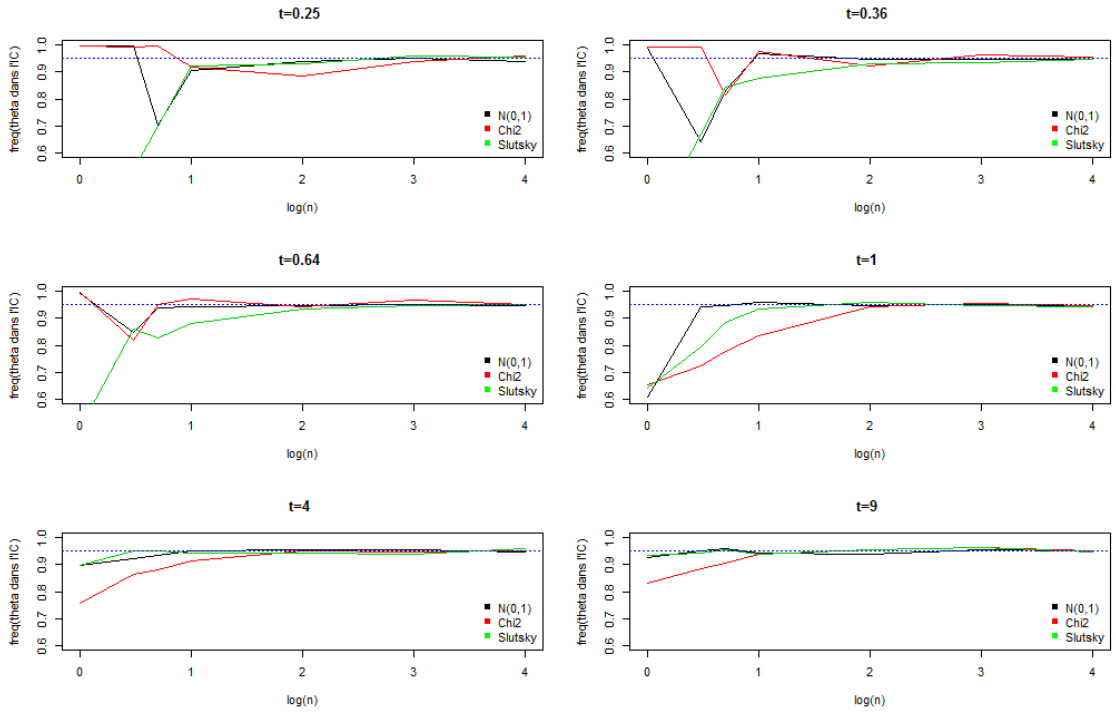


3.2.2 estimation du paramètre t d'une loi de poisson

Pour une loi de poisson, on connait d'après le TCL que $\sqrt{n}(\overline{X}_n - t) \Longrightarrow N(0,t)$. On possède trois méthodes différentes pour estimer la paramètre t d'une loi de Poisson :

- la delta-méthode pour se ramener à une convergence vers la loi N(0,1). En effet avec $f(x) = 2\sqrt{x}$, on a $2\sqrt{n}(\sqrt{\overline{X}_n} - \sqrt{t}) \Longrightarrow N(0,1)$
- la méthode du χ^2 pour se ramener à une convergence vers la loi χ^2_1. On passe au carré dans l'expression précédente, d'où $4n(\sqrt{\overline{X}_n} - \sqrt{t})^2 \Longrightarrow \chi^2_1$
- la méthode de Slutsky pour se ramener à une convergence vers la loi N(0,1). On a $\sqrt{\overline{X}_n} \xrightarrow{p.s.} \sqrt{t}$ et $\sqrt{n}\frac{(\overline{X}_n - t)}{\sqrt{t}} \Longrightarrow N(0,1)$ donc d'après

Slutsky, $\sqrt{n}\frac{(\overline{X}_n - t)}{\sqrt{\overline{X}_n}} \Longrightarrow N(0,1)$



On observe que la delta méthode et la méthode de Slutsky converge rapidement à 95% alors que la méthode du χ^2 converge mal. De plus, pour $t > 1$, la méthode de Slutsky est plus rapide que la delta-méthode et inversement pour $t < 1$.

4 Test de Kolmogorov-Smirnov

4.1 fonction de répartition empirique

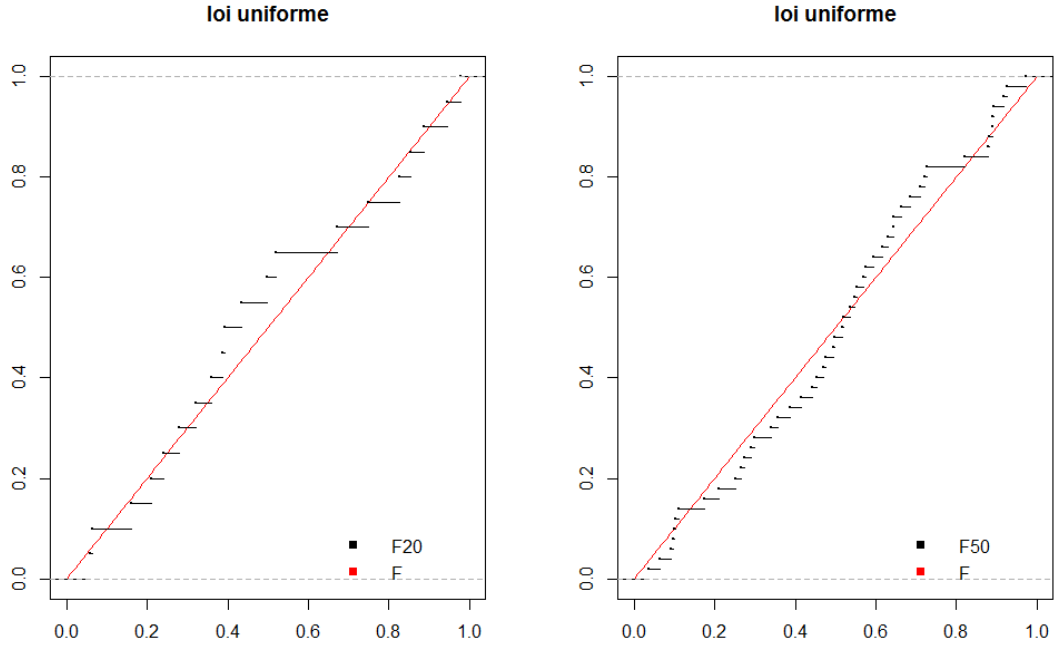
Définition 10. *Soit X_1,...X_n des variables aléatoires i.i.d. de fonction de répartition F. On appelle alors F_n fonction de répartition empirique avec*

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

Théorème 11. *Soit x fixé, on a alors $F_n(x) \xrightarrow{p.s.} F(x)$*

Preuve
En effet, on applique la loi des grands nombres à $Y_i = \mathbb{1}_{]-\infty,t]}(Xi)$. On obtient alors $\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p.s.} E[Y_1]$, or
$E[Y_1] = E[\mathbb{1}_{X_i \leq x}] = \mathbb{P}(X_i \leq x) = F(x)$ d'où le résultat.

Exemple avec la loi uniforme



Application : intervalle de confiance sur F(x)
On considère X_1,...X_n des variables aléatoires i.i.d. de fonction de répartition F. On peut utiliser le TCL et la fonction de répartition empirique pour trouver un intervalle de confiance sur la fonction de répartition F. En effet, on applique le TCL à $Y_i = \mathbb{1}_{]-\infty,t]}(Xi)$ de moyenne $E[Y_1] = E[\mathbb{1}_{X_i \leq x}] = \mathbb{P}(X_i \leq x) = F(x)$. Y_i suit une loi de Bernoulli de probabilité de succès p=F(x). On obtient alors

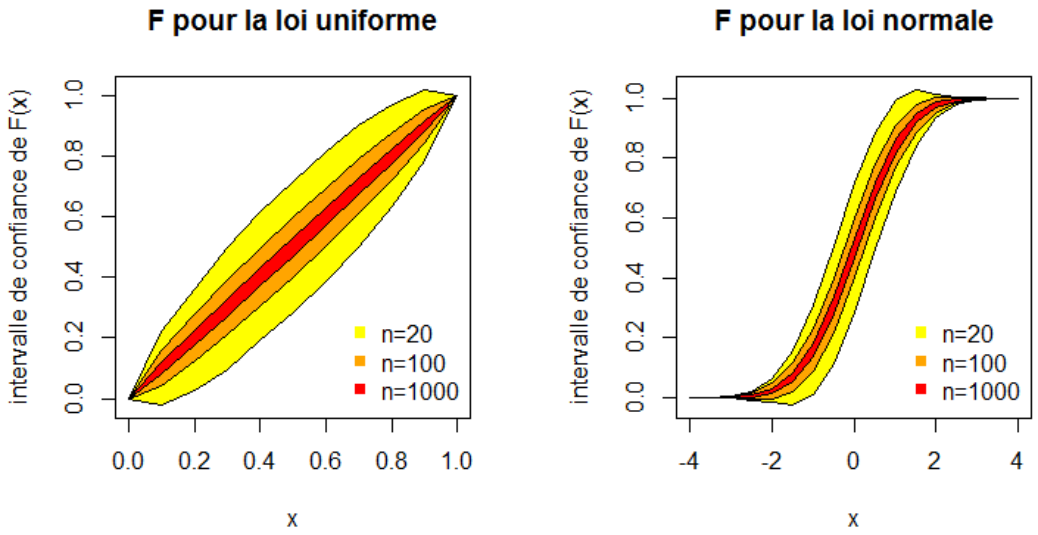
$$\sqrt{n}(F_n(x) - F(x)) \Longrightarrow N(0, (1 - F(x))F(x))$$

On peut alors appliquer Slutsky pour obtenir

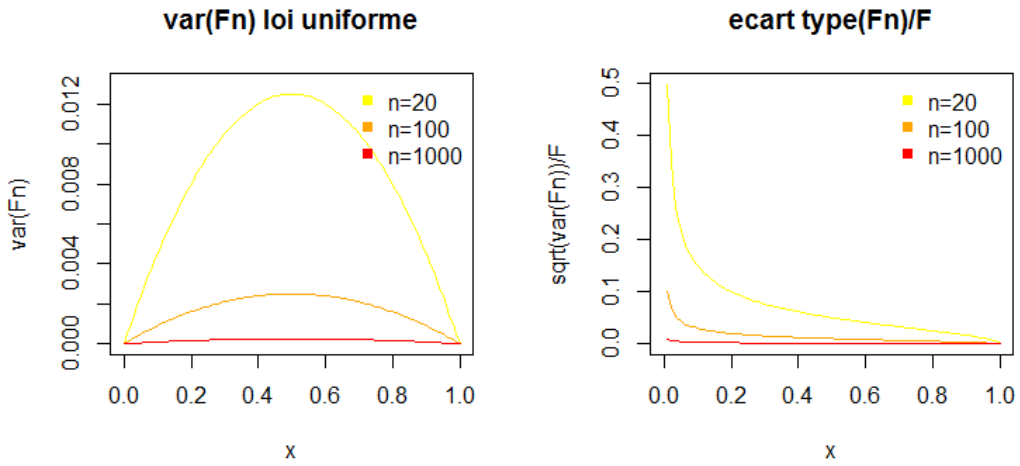
$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_n(x)(1 - F_n(x))}} \Longrightarrow N(0,1)$$

On en déduit que

$$\mathbb{P}(F(x) \in [F_n(x) - \frac{1.96}{\sqrt{n}}\sqrt{F_n(1-F_n)}, F_n(x) + \frac{1.96}{\sqrt{n}}\sqrt{F_n(1-F_n)}]) \longrightarrow 0.95$$



On s'intéresse à la largeur de l'intervalle afin de déterminer le x pour lequel l'intervalle est le plus grand. On observe une largeur plus importante pour $x = \frac{1}{2}$ mais si on considère l'écart relatif : $\frac{\sqrt{F_n}}{F}$ on s'aperçoit que le plus grand écart relatif est en x=0



4.2 Loi de Kolmogorov

Théorème 12.
On a, à condition que F^{-1} existe :

1. *Si $U \sim U_{[0,1]}$ alors $F^{-1}(U)$ a pour fonction de répartition F*
2. *Si X a pour fonction de répartition F alors $F(X) \sim U_{[0,1]}$*

Théorème 13.
Pour toute fonction de répartition continue F telle que F^{-1} existe, la loi de $\sup_x |F_n(x) - F(x)| = W_n$ ne dépend pas du F choisi.
En d'autres termes, on a $\sup_x |F_n(x) - F(x)| = \sup_{t\in[0,1]} |U_n(t) - U(t)|$ avec U la fonction de répartition de la loi uniforme et U_n la fonction de répartition empirique de la loi uniforme

Preuve
Supposons F bijective de $\mathbb{R}$ dans [0,1]
On a alors $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{u\in[0,1]} |F_n(F^{-1}(u)) - F(F^{-1}(u))|$

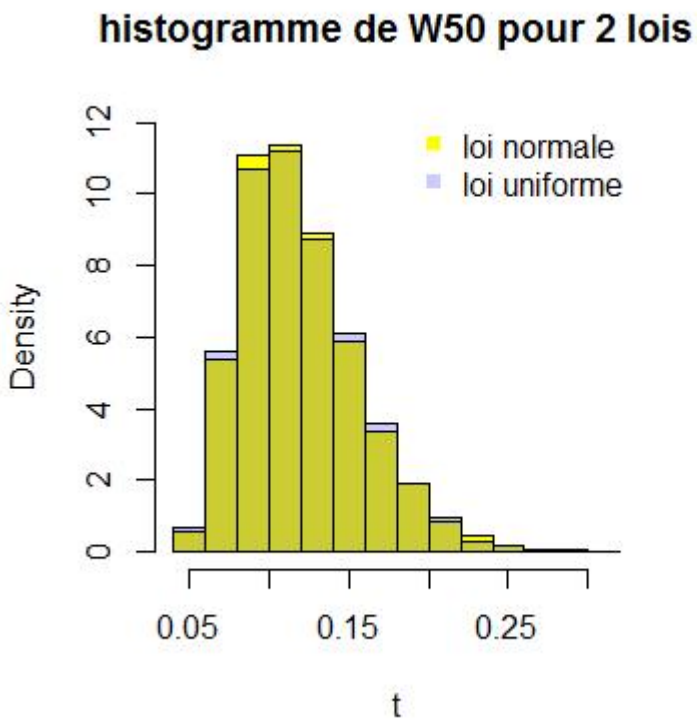
$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{u\in[0,1]} |\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\leq F^{-1}(u)} - u|$$

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{u\in[0,1]} |\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{F(X_i)\leq (u)} - u|$$

d'après le théorème 12, on a $F(X_i) \sim U_i$

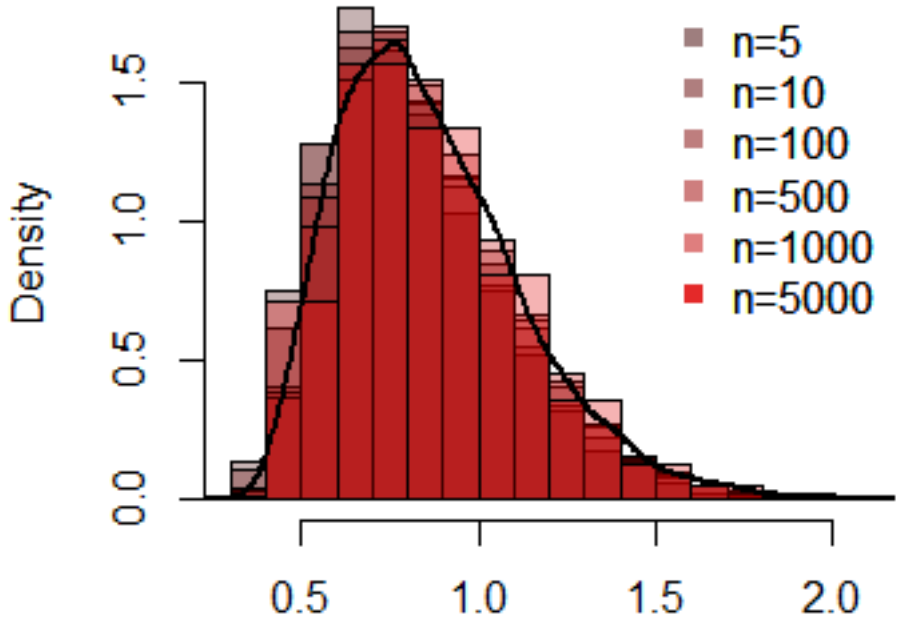
$$\text{donc } \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{u\in[0,1]} |\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq (u)} - u| = \sup_{u\in[0,1]} |U_n(u) - U(u)|$$

Exemple : histogramme de W50 pour deux lois



Théorème 14.
On a $\sqrt{n}W_n \xrightarrow{loi} K$ *où* K *est une variable aléatoire distribuée suivant la loi de Kolmogorov.*

Histogrammes de $\sqrt{n}W_n$ pour différentes valeurs de n et densité de la fonction de Kolmogorov



4.3 test de Kolmogorov-Smirnov

4.3.1 Méthode du test

On considère $X_1...X_n$ des variables aléatoires i.i.d. On considère deux hypothèses :

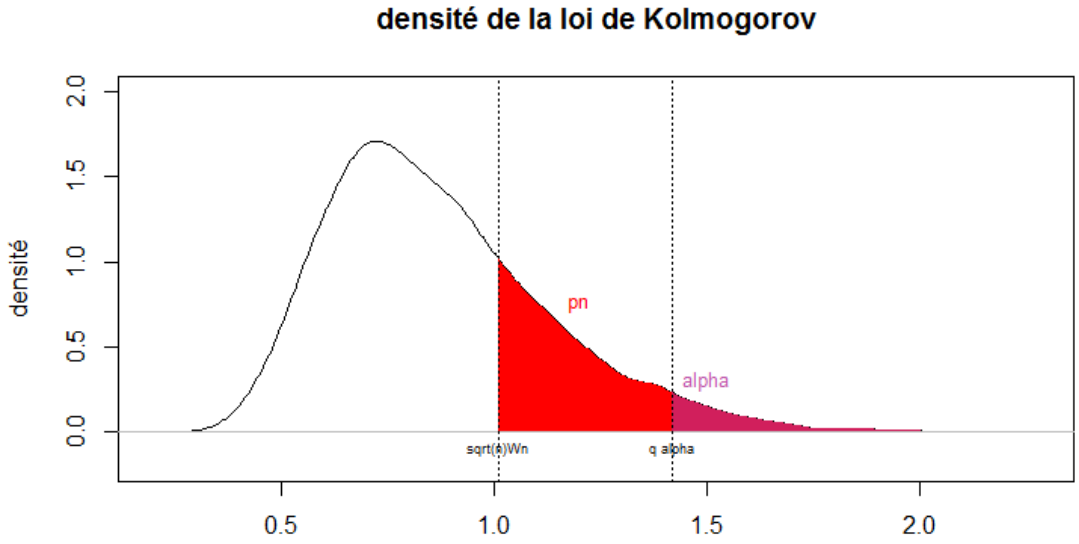
- H_0 : $X_1...X_n$ ont pour fonction de répartition F
- H_1 : $X_1...X_n$ n'ont pas pour fonction de répartition F

On appelle R la zone de rejet et α l'erreur de première espèce. Si $(X_1, ..., X_n) \in R$, on rejette l'hypothèse H_0.
On souhaite alors que $\mathbb{P}((X_1, ..., X_n) \in R \; sachant \; H_0 \; vraie) = \alpha$

$R = \{(X_1...X_n)/\sqrt{n}W_n > k_\alpha\}$
On souhaite avoir

décision / en réalité	H_0 acceptée	H_1 acceptée
H_0 vraie	1-α	α
H_1 vraie	$\underset{n\to+\infty}{\longrightarrow} 0$	$\underset{n\to+\infty}{\longrightarrow} 1$

On a par définition de la p-value, $p_n = 1 - F_K(\sqrt{n}W_n)$



Théorème 15.
Sous l'hypothèse H_0, $p_n \Longrightarrow U[0,1]$
Sous l'hypothèse H_1, $p_n \Longrightarrow 0$ (dirac en 0)

Preuve
Sous l'hypothèse H_0
On veut montrer que $\mathbb{P}(p_n \leq u) \to u = \mathbb{P}(U \leq u)$ où $U \sim \mathcal{U}[0,1]$

Soit $u \in [0,1]$

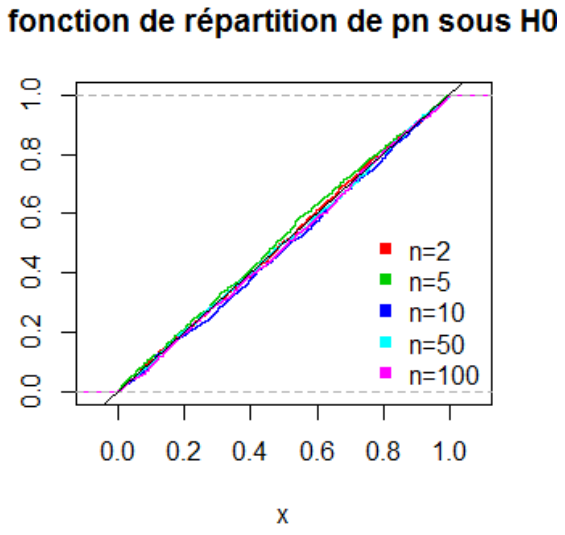
$$\begin{aligned}\mathbb{P}(p_n \leq u) &= \mathbb{P}(1 - F_K(\sqrt{n}W_n) \leq u) \\ &= \mathbb{P}(F_K(\sqrt{n}W_n) \geq 1-u) \\ &= \mathbb{P}(\sqrt{n}W_n \geq F_K^{-1}(1-u)) \text{ car } F_K^{-1} \text{ existe puisque } F_K \text{ bijection de } \mathbb{R}^+ \to [0,1] \\ &\overset{n\to+\infty}{\longrightarrow} 1 - F_K(F_K^{-1}(1-u)) = u\end{aligned}$$

car $\sqrt{n}W_n \Longrightarrow K$ et car $F_K(K) \sim U[0,1]$

Sous l'hypothèse H_1
On a $F_n(t) \overset{loi}{\to} G(t)$ tel que $\exists t_0 / G(t_0) \neq F(t_0)$
donc $W_n = \sup_x |F_n(x) - F(x)| \geq |F_n(t_0) - F(t_0)| > 0$ d'où $\sqrt{n}W_n \overset{p.s.}{\to} +\infty$
donc $p_n = 1 - F_K(\sqrt{n}W_n) \overset{n\to+\infty}{\longrightarrow} 0$

4.3.2 P_n sous H_0

On fait le test avec la loi normale.

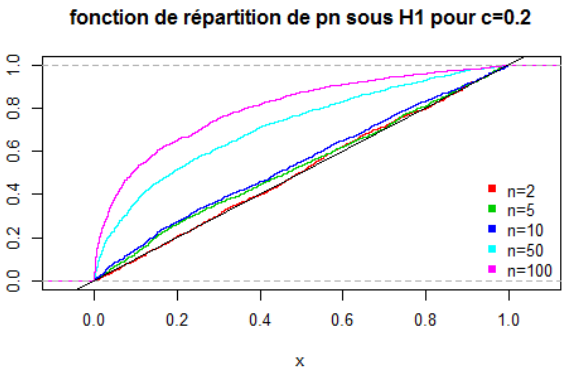


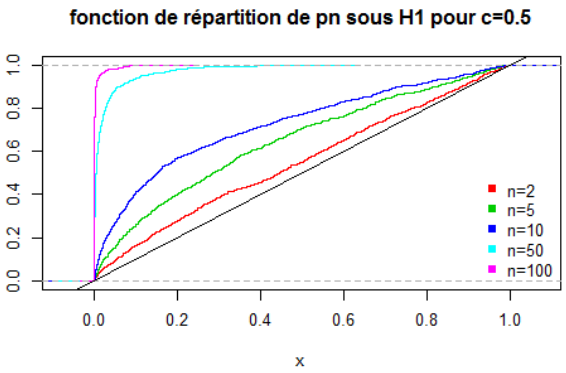
On observe pour toutes les valeurs de n une très forte proximité avec la fonction de répartition d'une loi uniforme alors qu'on devrait avoir uniquement pour des grandes valeurs de n la convergence en loi de p_n vers la loi uniforme.

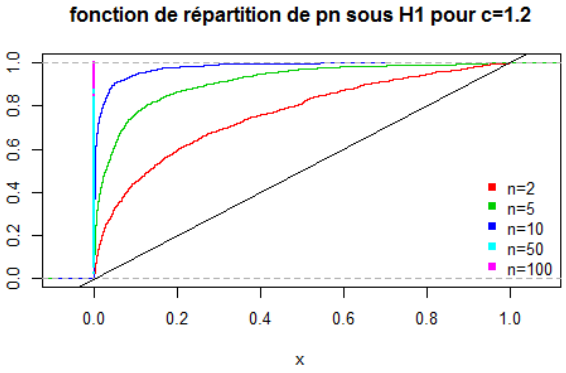
En réalité, le logiciel R choisit lors du test les valeurs de quantile $q_\alpha^{(n)}$ de la loi $\sqrt{n}W_n$ pour les petites valeurs de n et prend la valeur q_α de la loi limite de Kolmogorov lorsque n est assez grand.

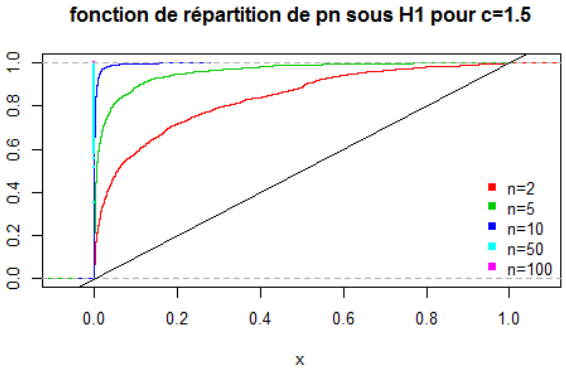
4.3.3 P_n sous H_1

On effectue plusieurs tests de Kolomogorov pour comparer N(0,1) avec un échantillon légèrement modifié en choisissant une loi N(c,1) avec $c \neq 0$. On choisit uniquement des valeurs positives de c car on a une symétrie pour c négatif. Pour chaque test, on calcule le p-value p_n. On obtient ainsi un échantillon de valeurs de p_n et on trace la fonction de répartition empirique de cet échantillon. Sous H_0, on doit être proche de la droite y=x (fonction de répartition de U[0,1]). Sous H_1, le test est performnt lorsque la courbe est proche du dirac en 0. On observe pour c proche de 0 qu'on reste proche de la fonction de répartition de la loi uniforme alors que pour c au dessus de 1 on a un écart avec la loi normale dès n=2.

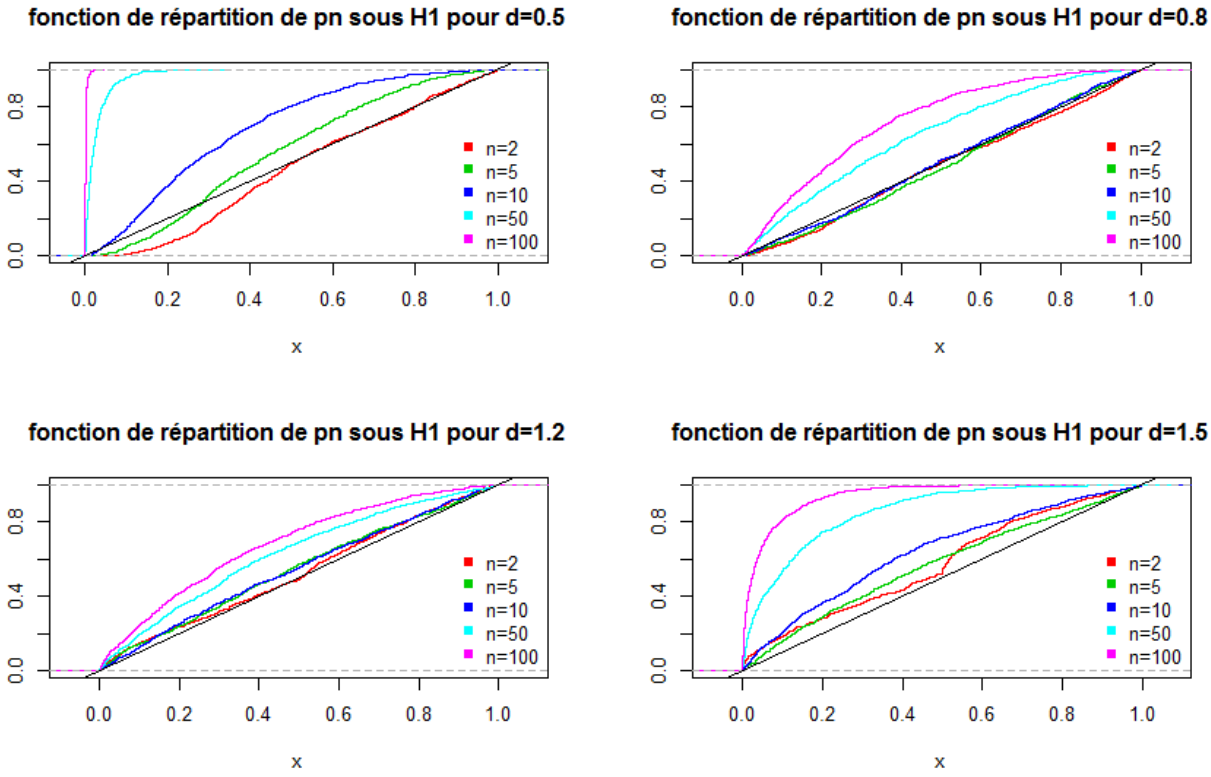








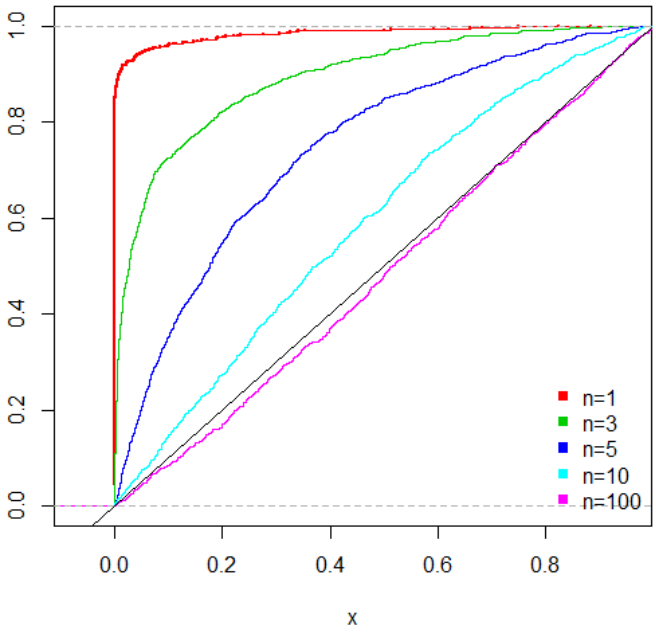
Désormais, on garde la moyenne à 0 et on fait varier l'écart-type d autour de 1. On observe une meilleure réactivité du test pour des petites valeurs de n lorsque l'écart-type s'éloigne de plus en plus de 1.



4.3.4 Application du test à la loi de Student

On considère $X \sim N(0,1)$ et $(Z_i)_{i\geq 1}$ i.i.d de loi χ^2_1 avec X indépendant des Z_i. On appelle loi de Student à n degré de liberté la loi T_n où $T_n = \frac{X}{\sqrt{\frac{\sum_{i=1}^{n} Z_i}{n}}}$.

Comme X$\overset{loi}{\Longrightarrow} X$ et $\sqrt{\frac{\sum_{i=1}^{n} Z_i}{n}} \overset{p.s.}{\to} \sqrt{E[Z_1]} = 1$.
On est donc censé avoir $T_n \overset{loi}{\Longrightarrow} X \sim N(0,1)$
Testons cette convergence grâce au test de Kolmogorov


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
x
n=1
n=3
n=5
n=10
n=100

Conclusion

Tout d'abord, le théorème central limite m'a permis de démontrer des résultats de convergence en loi dans le cas de variables aléatoires indépendantes et identiquement distribuées. Il est donc possible d'estimer si un échantillon suit une loi grâce au test de Kolmogorov. Il existe des résultats similaires lorsque les variables aléatoires sont indépendantes mais non-identiques avec une condition d'ordre 3 sur l'espérance avec le théorème de Liapounov.

Le théorème central limite est donc crucial dans les probabilités car il permet d'estimer de nombreuses lois. A partir de ce théorème central limite, la delta méthode permet d'estimer la moyenne lorsque la moyenne et la variance sont reliées. Enfin, on peut constater à travers les simulations réalisées que les différentes méthodes ne convergent pas aussi bien suivant le paramètre à estimer.

Ensuite, j'ai pu progresser dans le domaine de la programmation. J'ai en effet découvert le logiciel R avec lequel je peux désormais faire des simulations et tester le logiciel Latex pour écrire un document mathématique.

J'aimerai enfin remercier Mme Anne PHILIPPE qui m'a fait découvrir le domaine de la recherche et qui a su m'aider dans mon stage. Je suis désormais davantage intéressée par la recherche dans le domaine des probabilités.

Références

– Lehmann, *Elements of large sample theory*

– J.Jiang, *Large Sample Techniques for Statistics*

– Charles Suquet, *Théorème central limite*

– Charles Suquet, *Cours I.S.*