ÉTUDE ET COMPARAISON DE DIFFÉRENTS TESTS DE NORMALITÉ

Teddy Dousset, Amélie Guého et Guillaume Morvan

Mai 2020

Table des matières

Introduction

En statistique, les lois normales sont des distributions omniprésentes. Très souvent, les méthodes d'analyse requièrent des hypothèses de normalité avant d'être appliquées. C'est le cas par exemple en régression linéaire où la normalité des résidus est nécessaire. Des outils visuels existent pour évaluer la normalité d'un échantillon comme le Q-Q plot (diagramme Quantile-Quantile), les histogrammes ou encore les boîtes à moustaches. Ces méthodes graphiques sont pratiques mais ne fournissent pas de preuves statistiques fortes. Les tests de normalité sont des tests d'adéquation qui ont été développés pour pallier ce problème. Au cours du siècle dernier, des tests basés sur des outils mathématiques différents ont été pensés.

Notre travail a pour ambition d'étudier la validité des tests les plus connus et de comparer leurs performances en se basant sur les puissances, c'est-à-dire la probabilité qu'un test rejette l'hypothèse de normalité quand l'échantillon n'est pas normalement distribué. L'analyse des puissances permettra de déterminer le test le plus adapté à chaque situation.

Pour cela nous détaillerons dans une première partie théorique les propriétés et spécificités de 5 tests de normalité. Puis dans un deuxieme temps nous présenterons la méthode de simulation qui a permis de tracer les courbes des puissances. Enfin nous discuterons les résultats.

1 Validité des tests de normalité

Soit $(X_1, ..., X_n)$ un échantillon $i.i.d$ de variables aléatoires continues de densité f inconnue. Les tests de normalité sont un cas particulier des tests d'adéquation. Ils permettent de tester (quels que soient μ et σ^2) les hypothèses :

$$H_0 : X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2) \quad \text{vs} \quad H_1 : X_1, ..., X_n \nsim \mathcal{N}(\mu, \sigma^2)$$

Il existe différentes classes pour les tests de normalité :

— les tests basés sur la régression linéaire et le coefficient de corrélation : Shapiro-Wilk, Shapiro-Francia et Ryan-Joiner en sont des exemples ;

— les tests utilisant la fonction de répartition empirique : Kolmogorov-Smirnov, Lilliefors, Anderson-Darling et Cramér-von Mises ;

— les tests basés sur la méthode des moments : test du coefficient d'asymétrie, test du kurtosis, Jarque-Bera et D'Agostino-Pearson ;

— le test du χ^2 ;

— etc...

Dans ce document, nous nous focaliserons sur les tests de Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors (qui est en fait une modification de KS) et enfin Cramér-von Mises et sa version pondérée : Anderson-Darling.

1.1 Shapiro-Wilk

Publié en 1965 par Samuel Sanford Shapiro, statisticien à l'Université de Floride en collaboration avec Martin Wilk, le test de Shapiro-Wilk découle directement du diagramme quantile-quantile. Le diagramme quantile-quantile (ou Q-Q plot) compare les quantiles des observations avec les quantiles théoriques d'une certaine distribution (loi gaussienne dans notre cas).

Soient $Z_1, ..., Z_n \sim \mathcal{N}(0,1)$, n variables aléatoires $i.i.d$ et $Z_{(1)} \leq ... \leq Z_{(n)}$ leurs statistiques d'ordre.

On note $m' = (m_1, ..., m_n)$ le vecteur des espérances et $V = (v_{ij})$ la matrice $n \times n$ de covariance de ces statistiques d'ordre. Autrement dit :

$$\begin{cases} m_i = E(Z_{(i)}) & i = 1, 2, ..., n \\ (v_{ij}) = cov(Z_{(i)}, Z_{(j)}) & i, j = 1, 2, ..., n \end{cases}$$

Les m_i sont les $\frac{i}{n}^{eme}$ quantiles théoriques de la loi $\mathcal{N}(0,1)$.

Soit l'échantillon $i.i.d$ $(X_1, .., X_n) \sim \mathcal{N}(\mu, \sigma^2)$ et $X_{(1)} \leq ... \leq X_{(n)}$ leurs statistiques d'ordre. On a les relations suivantes entre les X_i et Z_i :

$$X_i = \mu + \sigma Z_i \quad i = 1, 2, ..., n$$

Ces égalités restent vraies pour les statistiques d'ordre :

$$X_{(i)} = \mu + \sigma Z_{(i)} \quad i = 1, 2, ..., n$$

Ainsi, par passage à l'espérance

$$X_{(i)} = \mu + \sigma m_i + \epsilon_i \quad i = 1, 2, ..., n$$

où $\epsilon_i = X_{(i)} - E(X_{(i)})$

Le Q-Q plot sera donc approximativement la droite d'équation '$y = \sigma x + \mu$'

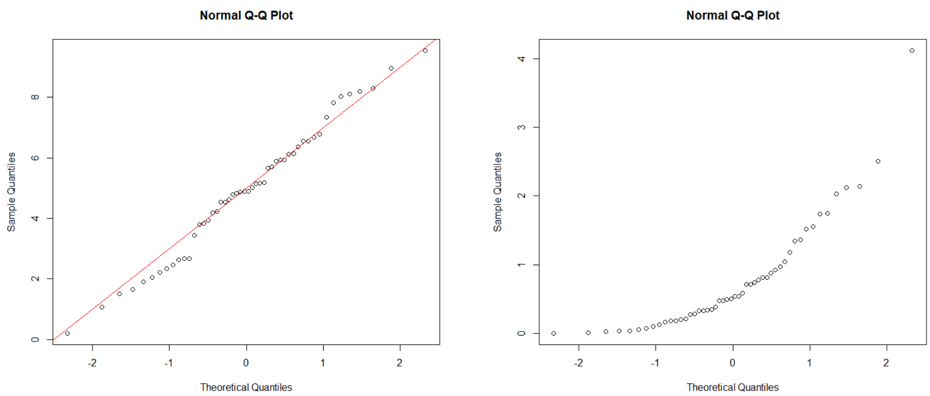


FIGURE 1 – Comparaison des Q-Q plots quand H_0 est vraie ou non

En pratique, la loi de l'échantillon $(X_1, ..., X_n)$ est inconnue. Le statistitien choisit d'accepter l'hypothèse de normalité des X_i s'il considère que les points sont suffisamment alignés. Cette méthode de validation est donc assez subjective. Une manière plus rigoureuse de procéder est d'estimer un modèle de régression linéaire par la méthode des moindres carrés, c'est l'idée derrière le test de Shapiro-Wilk.

Le modèle de régression linéaire est le suivant :

$$X_{(.)} = (\mathbb{1}, m).\begin{pmatrix}\mu \\ \sigma\end{pmatrix} + \epsilon \tag{1}$$

avec

$$X_{(.)} = \begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}, \quad (\mathbb{1}, m) = \begin{pmatrix} 1 & m_1 \\ 1 & m_2 \\ \vdots & \vdots \\ 1 & m_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Les erreurs ϵ_i de la modélisation sont centrées ($E(\epsilon_i) = 0$) mais ne sont pas indépendantes car les statistiques d'ordre $Z_{(1)} \leq ... \leq Z_{(n)}$ ne sont pas indépendantes. La matrice $Var(\varepsilon) = \sigma^2 V$ n'étant pas diagonale, il faut utiliser ici la méthode des moindres carrés généralisés. Celle-ci nous donne le meilleur estimateur linéaire sans biais de (μ, σ) :

$$(\hat{\mu}, \hat{\sigma}) = \Big(\frac{m'V^{-1}(m\mathbb{1}' - \mathbb{1}m')V^{-1}X_{(.)}}{\mathbb{1}'V^{-1}\mathbb{1}m'V^{-1}m - (\mathbb{1}'V^{-1}m)^2} \quad , \quad \frac{\mathbb{1}'V^{-1}(\mathbb{1}m' - m\mathbb{1}')V^{-1}X_{(.)}}{\mathbb{1}'V^{-1}\mathbb{1}m'V^{-1}m - (\mathbb{1}'V^{-1}m)^2} \Big) \tag{2}$$

On a le lemme suivant :

Lemme 1. *Si la loi des Z_k est symétrique alors $\mathbb{1}'V^{-1}m = 0$.*

Démonstration. Comme les Z_k sont de loi symétrique, $Z_{(1)}$ suit la même loi que $-Z_{(n)}$ et plus généralement $-Z_{(n+1-k)}$ a la même distribution que $Z_{(k)}$ pour $k = 1, ..., n$.

Ainsi on a l'égalité de leurs espérances :

$$m_j = -m_{n+1-j}, \quad j = 1, ..., n$$

En outre, la matrice V de covariance de $Z_{(1)}, ..., Z_{(n)}$ est également préservée si on échange j avec $n + 1 - j$ pour les lignes et pour les colonnes. De même pour V^{-1}.

Ainsi le vecteur $V^{-1}m$ est de somme nulle, c'est à dire $\mathbb{1}'V^{-1}m = 0$ □

Il en résulte que

$$(\hat{\mu}, \hat{\sigma}) = \Big(\frac{1}{n}\sum_{i=1}^{n} X_{(i)} \quad , \quad \frac{m'V^{-1}X_{(.)}}{m'V^{-1}m} \Big) \tag{3}$$

Ici $\hat{\sigma}$ correspond à la pente de la droite de régression du Q-Q plot. Si la loi de distribution des X_i est une loi normale (i.e sous H_0), on a donc ce nouvel estimateur sans biais $\hat{\sigma}$ de l'écart type de X_i.

La statistique de Shapiro-Wilk est définie comme le rapport entre le carré de cet estimateur $\hat{\sigma}$ et l'estimateur sans biais usuel de la variance :

$$\frac{1}{n-1}\sum_{i=1}^{n}\Big(X_{(i)} - \overline{X}_{(.)}\Big)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(X_i - \overline{X}\Big)^2$$

Ce dernier est indépendant de l'ordre. Le tout est multiplié par d'une constante de normalisation $\left(\frac{R^2}{C}\right)^2$:

$$W = \frac{R^4\hat{\sigma}^2}{C^2\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{b^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\langle a, X_{(.)}\rangle^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Et finalement :

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4}$$

où

— $R^2 = m'V^{-1}m$

— $C^2 = \langle m'V^{-1}, m'V^{-1}\rangle$

— $b = \frac{R^2\hat{\sigma}}{C} = \frac{m'V^{-1}X_{(.)}}{\langle m'V^{-1}, m'V^{-1}\rangle^{1/2}} = \langle a, X_{(.)}\rangle$

— $a' = (a_1, ..., a_n) = \frac{m'V^{-1}}{\langle m'V^{-1}, m'V^{-1}\rangle^{1/2}}$

En d'autres termes, le test de Shapiro-Wilk compare un estimateur de la variance de X_i sous H_0 avec l'estimateur usuel de la variance de X_i. Pour un échantillon non normalement distribué, ces 2 quantités n'estiment généralement pas la même chose.

On a la propriété suivante qui rend la statistique de SW appropriée à un test de normalité :

Propriété 1. *Le paramètre de position (location) et le paramètre d'échelle (scale) de la loi de X n'ont aucune influence sur la statistique de SW. En particulier si $X \sim \mathcal{N}(\mu, \sigma^2)$, la statistique de SW est invariante pour les paramètres μ et σ^2 (W est une statistique libre).*

Démonstration. Supposons qu'on ajoute une constante p à tous les X_i.

Cette constante $p \in \mathbb{R}$ s'ajoute également à la moyenne $\overline{X}$ donc le dénominateur de W reste inchangé.

Pour le numérateur, commençons par remarquer que la loi des Z_i étant $\mathcal{N}(0,1)$ (symétrique), les statistiques d'ordre $Z_{(j)}$ et leurs moyennes m_j ont des propriétés de symétrie que nous avons déjà évoqué :

$$m_j = -m_{n+1-j}, \quad j = 1, ..., n$$

ainsi

$$\sum_{j=1}^{n} m_j = 0$$

En outre, la matrice V de covariance de $Z_{(1)}, ..., Z_{(n)}$ est également préservée si on échange j avec $n+1-j$ pour les lignes et pour les colonnes. De même pour V^{-1}.

Ainsi on a une symétrie des coefficients a_i :

$$a_j = -a_{n+1-j}, \quad j = 1, ..., n$$

et

$$\sum_{j=1}^{n} a_j = 0$$

En conséquence, comme l'ajout d'une constante $p \in \mathbb{R}$ à tous les X_i entraîne également l'ajout de cette même constante p à l'échantillon ordonné $X_{(1)}, ..., X_{(n)}$, on a :

$$\sum_{i=1}^{n} a_i(X_{(i)} + p) = \sum_{i=1}^{n} a_i X_{(i)} + p \underbrace{\sum_{i=1}^{n} a_i}_{=0} = \sum_{i=1}^{n} a_i X_{(i)}$$

Le numérateur et donc W sont invariants par changement de position.

Supposons maintenant que tous les X_i sont multipliées par une constante $s > 0$. Les $X_{(i)}$ ordonnés sont également multipliées par s. Ainsi, le numérateur et le dénominateur sont tous deux multipliés par s^2, et donc W est invariant par changement d'échelle.

□

Corollaire 1. *Sous H_0, W est pivotale et sa distribution ne dépend que de la taille de l'échantillon n.*

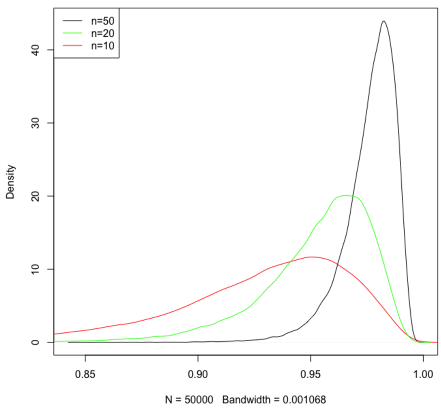


FIGURE 2 – Distribution de W sous H_0 pour différentes valeurs de n

La statistique de Shapiro-Wilk peut être également interprétée comme le carré du coefficient de corrélation de Pearson entre les poids $a_1, ..., a_n$ et l'échantillon ordonné $X_{(1)}, ..., X_{(n)}$.

En effet,

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i = 0 \quad \text{et donc} \quad \hat{\sigma}_a^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 = \frac{1}{n}$$

de plus, la covariance entre a et $X_{(.)}$ s'écrit :

$$\hat{\sigma}_{aX_{(.)}} = \frac{1}{n}\sum_{i=1}^{n} a_i X_{(i)} - \frac{1}{n^2}\sum_{i=1}^{n} X_{(i)} \underbrace{\sum_{i=1}^{n} a_i}_{=0} = \frac{1}{n}\sum_{i=1}^{n} a_i X_{(i)}$$

et comme

$$\hat{\sigma}^2_{X_{(.)}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_{(i)} - \overline{X}_{(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

on a bien :

$$W = \hat{\rho}^2_{aX_{(.)}} = \frac{\hat{\sigma}^2_{aX_{(.)}}}{\hat{\sigma}^2_X . \hat{\sigma}^2_a}$$

Propriété 2. *$W \leq 1$, et le maximum de W est atteint lorsque les vecteurs $X_{(.)}$ et a sont proportionnels.*

Démonstration. W est le carré d'un coefficient de corrélation.

□

On rappelle que les poids a_i dans la statistique de test W sont définis par :

$$a_i = \sum_{j=1}^{n} \frac{m_j v_{ij}}{C} \quad i = 1, ..., n \quad \text{où} \quad C = \langle m'V^{-1}, m'V^{-1} \rangle^{\frac{1}{2}}$$

Pour calculer les a_i il faut donc connaître les valeurs exactes des m_j et des v_{ij} de V, la matrice de covariance de taille $n \times n$ des statistiques d'ordre $Z_{(1)} \leq ... \leq Z_{(n)}$. Cependant, pour $n > 20$, V n'est pas connue et il faut passer par des approximations.

Les valeurs tabulées des a_i pour différentes tailles d'échantillon n sont disponibles en annexes. Dans chaque colonne, seulement la moitié des poids est donnée car le reste se déduit par symétrie. Les approximations des a_i par Shapiro et Wilk sont données pour $n \leq 50$. Nous ne détaillerons pas l'algorithme d'approximation des a_i proposé par Shapriro et Wilk [2] mais pour $n > 50$, J.P Royston (1982) a montré que cette approximation n'était plus possible et a proposé une nouvelle méthode d'approximation des a_i qui est valable pour des valeurs de n allant jusqu'à 5000.

Quant aux valeurs critiques, Shapiro et Wilk expliquent dans leurs travaux que celles-ci sont calculées empiriquement. La loi de W pour chaque n est estimée par Monte-Carlo et ses quantiles critiques sont répertoriés dans des tables. Nous n'avons pas trouvé de table pour $n > 50$ donc nous devrons calculer nous même les valeurs critiques.

La statistique de SW étant un coefficient de corrélation, des valeurs trop inférieures à 1 conduisent au rejet de l'hypothèse nulle.

1.2 Kolmogorov-Smirnov

En 1933, Andrey Nikolaevich Kolmogorov et Nikolai Smirnov, mathématiciens de l'URSS ont developpé un test d'adéquation basé sur la fonction de répartition empirique. De tels test sont non-paramétriques. En effet, le modèle ici est l'ensemble des fonctions CàdLàg de $\mathbb{R}$ dans $[0,1]$, espace vectoriel de dimension infini.

Soit un échantillon $X_1, ..., X_n$ $i.i.d$ d'une certaine loi continue et F sa fonction de répartition.

La fonction de répartition empirique $\hat{F_n}$ de l'échantillon est :

$$\hat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) \quad \text{avec} \quad \mathbb{1}_{]-\infty,x]}(X_i) = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{sinon} \end{cases}$$

Par la loi forte des grands nombres, $\hat{F_n}$ est un estimateur fortement consistant de F et le théorème central limite nous donne sa vitesse de convergence :

$$\forall x \in \mathbb{R}, \quad \sqrt{n}(\hat{F_n}(x) - F(x)) \xrightarrow[n\to+\infty]{Loi} \mathcal{N}\Big(0, F(x)(1-F(x))\Big)$$

Le test de Kolmogorov-Smirnov permet de tester :

$$H_0 : F = F_0 \;\text{ vs }\; H_1 : F \neq F_0$$

par la statistique de test :

$$D_n = \underset{x\in\mathbb{R}}{sup}(|\hat{F_n}(x) - F_0(x)|)$$

D_n quantifie la plus grande distance verticale entre la fonction de répartition empirique $\hat{F_n}$ d'un échantillon et la fonction de répartition F_0 d'une distribution théorique de référence. Dans le cadre des tests de normalité, cette distribution théorique est une loi normale de paramètres μ et σ connus (dans ce cas on notera $F_0 = F_{\mu,\sigma}$).

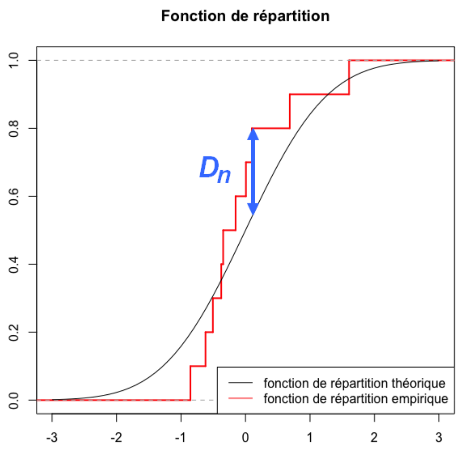


FIGURE 3 – La statistique D_n

Nous allons voir quelques propriétés de D_n et pourquoi elle est valide en tant que test d'adéquation.

Théorème 1.1. *Théorème de Glivenko-Cantelli*

Presque sûrement, la fonction de répartition empirique $\hat{F_n}$ converge uniformément vers F quand $n \to \infty$.

$$\underset{x\in\mathbb{R}}{sup}|\hat{F_n} - F| \xrightarrow{p.s} 0$$

Ainsi sous H_0, le théorème de Glivenko-Cantelli nous assure la convergence presque sûre de D_n vers 0 quand $n \to +\infty$.

On a également le théorème suivant qui rend possible l'utilisation de D_n en tant que statistique de test :

Théorème 1.2. *Sous H_0, si F est continue alors la loi de $D_n = \underset{x}{sup}|\hat{F_n}(x) - F(x)|$ ne dépend pas de F (i.e D_n est pivotale)*

Démonstration. Soit $F^{-1}(y) = \inf\{x \in \mathbb{R}, F(x) \geq y\}$ la fonction quantile. Soit F_{D_n} la fonction de répartition de la variable aléatoire D_n.

En effectuant le changement variable $y = F(x) \Longleftrightarrow x = F^{-1}(y)$, on peut écrire :

$$F_{D_n}(t) = \mathbb{P}(\underset{x\in\mathbb{R}}{sup}|\hat{F_n}(x) - F(x)| \leq t) = \mathbb{P}(\underset{0\leq y\leq 1}{sup}|\hat{F_n}(F^{-1}(y)) - y| \leq t)$$

Où

$$\hat{F_n}(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{X_i \leq F^{-1}(y)\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{F(X_i) \leq y\}$$

Finalement :

$$F_{D_n}(t) = \mathbb{P}(\underset{0\leq y\leq 1}{sup}|\hat{F_n}(F^{-1}(y)) - y| \leq t) = \mathbb{P}(\underset{0\leq y\leq 1}{sup}|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{F(X_i) \leq y\} - y| \leq t)$$

On peut prouver aisément que la loi de $F(X_i) \quad i = 1, ..., n$ est uniforme sur $[0, 1]$:

$$\mathbb{P}(F(X_i) \leq t) = \mathbb{P}(X_i \leq F^{-1}(t)) = F(F^{-1}(t)) = t \quad i = 1, ..., n$$

Dès lors on pose $U_i = F(X_i)$ $i = 1, ..., n$ des variables aléatoires $i.i.d$ suivant une loi uniforme sur $[0, 1]$, et en reprenant les calculs précédents on déduit :

$$F_{D_n}(t) = \mathbb{P}(\underset{x\in\mathbb{R}}{sup}|\hat{F_n}(x) - F(x)| \leq t) = \mathbb{P}(\underset{0\leq y\leq 1}{sup}|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{U_i \leq y\} - y| \leq t)$$

ce qui ne dépend pas de F. □

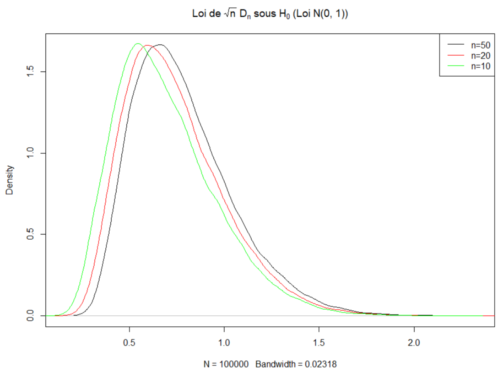


FIGURE 4 – $\sqrt{n}D_n$ sous H_0 quand F vient d'une loi $N(0,1)$

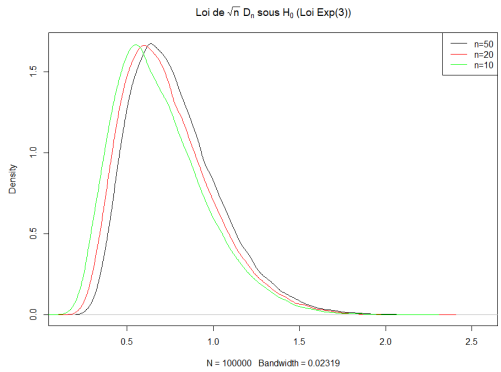


FIGURE 5 – $\sqrt{n}D_n$ sous H_0 quand F vient d'une loi $Exp(3)$

Le code qui a permis de tracer ces courbes sous R est disponible en annexe

Ainsi sous H_0 la distribution de D_n est la même pour toute fonction F continue et peut donc être utilisée pour tester l'égalité $F = F_0$. La loi de D_n sera quand même dépendante de n, mais pour n assez grand, on peut montrer la convergence en loi de $\sqrt{n}D_n$ vers une Loi limite qui ne dépend plus de n : la loi de Kolmogorov-Smirnov.

Le théorème de Glivenko-Cantelli ne dit rien sur la vitesse de convergence de D_n ni sur son comportement asymptotique. Pour approfondir cela, il nous faut d'abord définir ce qu'est un pont brownien et plus généralement, un processus stochastique :

Définition 1. *Un processus stochastique $X = (X_t)_{t \in T}$ est une famille de variables aléatoires X_t indexée par un ensemble T. L'ensemble des observations $(x_t)_{t \in T}$ constitue une réalisation du processus. L'espace T (continu ou discret) des indices peut faire référence au temps. L'indice $t \in T$ désigne alors un instant. Dans le cas continu, on peut noter le processus $X(t)$ comme une fonction du temps.*

Un pont brownien standard $B(t)$ est un processus stochastique à temps continu ($T = [0, 1]$) et dont la loi est celle d'un processus de Wiener (les incréments sont $i.i.d$ gaussiens de moyenne nulle et $Cov(B(s), B(t)) = s(1-t)$ si $0 < s < t < 1$). On parle de "pont" car il est conditionné à s'annuler en $t = 0$ et en $t = 1$.

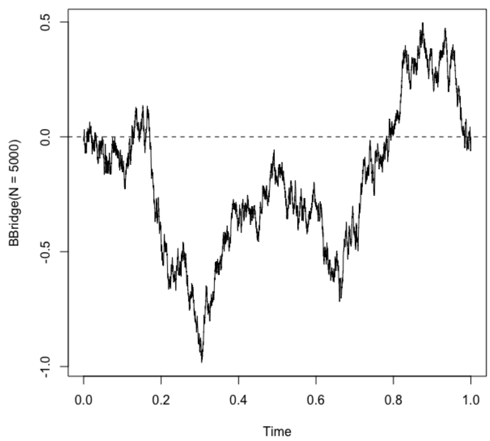


FIGURE 6 – Le pont Brownien

Le concept d'un pont Brownien entre $\hat{F_n}$ et F_0 est ici assez intuitif. Les fonctions $\hat{F_n}$ et F_0 sont "attachées" aux extrémités en $-\infty$ et $+\infty$ et entre ces deux connexions, l'écart entre $\hat{F_n}$ et F_0 fluctue aléatoirement.

On effectue le changement de variable suivant pour se ramener dans $[0,1]$:

$$\begin{array}{ccc}]-\infty,+\infty[& \longrightarrow & [0,1] \\ x & \longmapsto & F(x)=t \end{array}$$

Et on en déduit le théorème de Donsker :

Théorème 1.3. *Théorème de Donsker*
Le processus $\sqrt{n}(\hat{F_n}-F)$ indexé par x converge en Loi vers un pont brownien $B(F(x))$.

Enfin par passage au supremum, on obtient la loi limite de $\sqrt{n}D_n$:

La loi de Kolmogorov-Smirnov

Sous H_0,

$$\sqrt{n}\underset{x}{sup}|\hat{F_n}(x)-F(x)| = \sqrt{n}D_n \overset{n\to\infty}{\longrightarrow} \underset{x}{sup}|B(F(x))|$$

La loi du supremum d'un pont brownien $K = \underset{x}{sup}|B(F(x))|$ est la loi Kolmogorov-Smirnov qui ne dépend ni de F ni de n.

On admet le résultat suivant concernant le pont brownien :

Propriété 3. *Soit x un nombre réel strictement positif alors*

$$\mathbb{P}\left[\sup_{t\in[0,1]}|B(t)|\geq x\right] = 2\sum_{k\geq 1}(-1)^{k-1}e^{-2k^2x^2}$$

Cela nous permet de definir la fonction de répartition de la loi de Kolmogorov-Smirnov :

$$H(x) = \begin{cases} 0 & \text{si } x<0 \\ 1-2\sum\limits_{k\geq 1}(-1)^{k-1}e^{-2k^2x^2} & \text{sinon} \end{cases}$$

En conclusion, si l'hypothèse nulle est vraie, par le Théorème 1.2, la distribution de D_n ne dépend que de n et peut être tabulée. Et si n est assez grand alors la loi de $\sqrt{n}D_n$ est approxivement la loi de Kolmogorov-Smirnov dont les valeurs sont également tabulées.

Pour tester H_0 on considère la règle de décision suivante :

$$A = \begin{cases} H_0 & \text{si } \sqrt{n}D_n \leq c \\ H_1 & \text{si } \sqrt{n}D_n > c \end{cases}$$

Où le seuil c dépend du niveau de significativité α choisi :

$$\alpha = \mathbb{P}(A\neq H_0|H_0) = \mathbb{P}(\sqrt{n}D_n > c|H_0) \approx 1-H(c)$$

Supposons maintenant que l'hypothèse nulle est fausse, i.e. $F\neq F_0$.

Puisque F est la fonction de répartition de l'échantillon $(X_1, ..., X_n)$, par la loi forte des grands nombres : $\hat{F_n} \stackrel{p.s}{\longrightarrow} F \neq F_0$, ainsi

$$\exists \delta > 0, \exists N \in \mathbb{N}, \forall n > N, \qquad \underset{x}{sup} |\hat{F_n}(x) - F_0(x)| > \delta$$

En multiliant par $\sqrt{n}$:

$$\sqrt{n} D_n = \sqrt{n}.\underset{x}{sup} |\hat{F_n}(x) - F_0(x)| > \sqrt{n}\delta$$

Si H_0 est fausse alors $\sqrt{n} D_n > \sqrt{n}\delta \stackrel{n \to +\infty}{\longrightarrow} +\infty$.

Finalement, sous H_0 la loi de D_n peut être tabulée pour chaque n et on peut trouver le seuil $c = c_\alpha$. Quand n est assez grand, on peut utiliser H, fonction de répartition de la loi de Kolmogorov-Smirnov, pour trouver c : des valeurs trop grandes de $\sqrt{n} D_n$ (au-delà d'un certain quantile de la distribution de Kolmogorov) conduisent à rejeter l'hypothèse nulle et on conclut que la distibution de l'échantillon n'est pas F_0. La table des valeurs critiques est fournie en annexe.

Le problème d'utiliser Kolmogorov-Smirnov dans un test de normalité c'est qu'il faut spécifier $F_{\mu,\sigma}$. Cela le rend très limité car bien souvent, on ne connaît pas μ et σ. Une modification de ce test a été proposé par Lilliefors dans laquelle les paramètres inconnus μ et σ sont estimés par leurs équivalents empiriques, c'est le sujet de la section suivante.

1.3 Lilliefors

Développé en 1967 par Hubert Lilliefors, professeur de statistique à l'Université de George Washington, le test de Lilliefors est une variante du test de Kolmogorv-Smirnov. Dans le cas où les paramètres de la loi théorique ne sont pas connus, on peut les estimer.

Pour le test de normalité, cela revient à remplacer les paramètres μ et σ^2 de la loi théorique $\mathcal{N}(\mu, \sigma^2)$ par leurs équivalents empiriques :

Soit $X_1, ..., X_n$ un échantillon *i.i.d* d'une loi continue à déterminer,

— Sa moyenne empirique : $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$

— Sa variance empirique : $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$

La statistique de test de Lilliefors est :

$$\hat{D}_n = \underset{x \in \mathbb{R}}{sup} |\hat{F_n}(x) - F_{\hat{\mu},\hat{\sigma}}(x)|$$

Bien que celle-ci soit similaire à celle de Kolmogorov-Smirnov, $\hat{D}_n$ dépend des paramètres de la loi (μ et σ) en plus de dépendre de n. En effet, en estimant les paramètres de la loi de $X_1, ..., X_n$ à partir de ce même échantillon que l'on veut tester, la fonction $F_{\hat{\mu},\hat{\sigma}}$ est un peu trop "adaptée" à l'échantillon auquel on la compare. En conséquence on rejette H_0 moins souvent que l'on devrait.

$\hat{D}_n$ n'est pas *distribution-free* et les valeurs critiques C_α^{KS} de KS ne sont plus valides pour n'importe quelle distribution. Pour calculer les nouvelles valeurs critiques, on procède par Monte-Carlo (la méthode sera détaillée ultérieurement).

Comparons la table des valeurs critiques de KS avec la table des valeurs critiques obtenues par Monte-Carlo (pour une loi théorique $\mathcal{N}(-3, 2)$ par exemple) :

n	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
10	0.30347	0.26232	0.24091
20	0.22334	0.19176	0.17630
30	0.18506	0.15875	0.14588
40	0.16129	0.13853	0.12723
50	0.14539	0.12463	0.11432
80	0.11535	0.09907	0.09109

Les valeurs de critiques de KS (voir annexe) sont plus élevées que celles du Monte-Carlo et globalement :

$$C_\alpha^{KS} \approx \frac{3}{2} C_\alpha^{MC}$$

Ce ratio est d'autant plus vrai que n est grand.

Avec les valeurs critiques C_α^{KS} on accepte des valeurs un peu plus élévée de $\hat{D}_n$ donc on a tendance à moins rejeter H_0 et on augmente la probabilité β d'accepter de H_0 à tort (ou erreur de type II). Ainsi la puissance $1 - \beta$ est inférieure avec C_α^{KS}.

Un autre exemple : les $C_{1\%}^{MC}$ sont à peine plus petites que les $C_{20\%}^{KS}$: le niveau de significativité α est largement surévalué avec la table de KS.

En réalité, la statistique de LL est invariante pour une même famille de loi quand cette famille est paramétrée intégralement par une position et une échelle. C'est le cas pour les lois normales. En effet :

Soit $f(t|\mu,\sigma)$ la densité de probabilité de $\mathcal{N}(\mu,\sigma)$.
Par changement de variable :

$$F_{\mu,\sigma}(x) = \int_{-\infty}^{x} f(t|\mu,\sigma)\,\mathrm{d}t = \int_{-\infty}^{\frac{x-\mu}{\sigma}} f(t|0,1)\,\mathrm{d}t = \Phi\Big(\frac{x-\mu}{\sigma}\Big)$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Ainsi sous H_0 et quels que soient les vraies valeurs μ et σ de la loi de l'échantillon $(X_1,...,X_n)$, $\hat{D}_n$ peut s'écrire :

$$\hat{D}_n = \underset{x\in\mathbb{R}}{sup}|\hat{F_n}(x) - F_{\hat{\mu},\hat{\sigma}}(x)| = \underset{x\in\mathbb{R}}{sup}|\hat{F_n}(x) - \Phi\Big(\frac{x-\hat{\mu}}{\hat{\sigma}}\Big)|$$

de plus comme

$$\hat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,\frac{x-\mu}{\sigma}]}\Big(\frac{X_i-\mu}{\sigma}\Big)$$

Les valeurs de $\hat{F_n}(x)$ peuvent aussi bien être calculées à partir de l'échantillon initial que l'échantillon normalisé.

En conclusion, pour un test de normalité, les valeurs critiques de $\hat{D}_n$ peuvent toutes être estimées par Monte-Carlo à partir de la loi $\mathcal{N}(0,1)$.

1.4 Cramér-von Mises et Anderson-Darling

Par soucis de simplicité, les notations qui ont été utilisées jusqu'à présent resteront les mêmes dans cette partie, notamment pour les fonctions de répartition théoriques et empiriques (F, Φ, F_0 et $\hat{F_n}$) et les hypothèses (H_0 vs H_1).

Le test de Cramér-von Mises ou critère de Cramér-von Mises est un test statistique utilisé pour évaluer la qualité de l'ajustement d'une fonction de répartition notée F_0 comparée à une fonction de répartition empirique notée $\hat{F_n}$. Ce test est nommé en l'honneur de Harald Cramér et Richard von Mises. La généralisation pour deux échantillons de ce test est due à Theodore Anderson. Ce test est également une alternative au test de Kolmogorov-Smirnov.

Le test d'Anderson-Darling quant à lui porte le nom de Theodore Wilbur Anderson (1918-2016) et Donald A. Darling (1915-2014), qui l'ont inventé en 1952. Ce test d'ajustement est de la même famille que celui de Kolmogorov-Smirnov (KS) et de Cramér-von Mises (CVM).

1.4.1 Présentation générale

On teste les hypothèses :

$$H_0 : F = F_0 \quad \text{vs} \quad H_1 : F \neq F_0$$

($F_0 = F_{\mu,\sigma}$ dans le cas d'un test de normalité)

Soient les deux mesures de distance suivantes :

$$K_n = \sup_{x\in\mathbb{R}}\{\sqrt{n}\,|\hat{F_n}(x) - F_0(x)|\,\sqrt{\psi(F_0(x))}\} \tag{5}$$

$$W_n^2 = n\int_{-\infty}^{\infty}[\hat{F_n}(x) - F_0(x)]^2\,\psi(F_0(x))\,\mathrm{d}F_0(x) \tag{6}$$

où le poids $\psi(t)$ doit vérifier pour (6)

$$\int_0^1 t^2\psi(t)\,\mathrm{d}t < \infty \quad \text{et} \quad \int_0^1 (1-t)^2\psi(t)\,\mathrm{d}t < \infty\,.$$

En pratique, pour des observations ordonnées $(x_{(1)}, ..., x_{(n)})$ d'un échantillon $(X_1, ..., X_n)$ de variables aléatoires continues, on calcule W_n^2 grâce à la formule :

$$W_n^2 = 2\sum_{i=1}^{n}\left\{\phi_1(F_0(x_{(i)})) - \frac{2i-1}{2n}\phi_2(F_0(x_{(i)})\right\} + n\int_0^1(1-t)^2\psi(t)\,\mathrm{d}t \tag{7}$$

$$\text{avec} \quad \phi_1(t) = \int_0^t \psi(s)\,\mathrm{d}s \quad \text{et} \quad \phi_2(t) = \int_0^t s\psi(s)\,\mathrm{d}s.$$

W_n^2 existe si et seulement si ϕ_1 et ϕ_2 existent.

Le poids $\psi(t)$ ($0 \leq t \leq 1$) est une fonction continue et strictement positive. Celui-ci permet une flexibilité dans la mesure en pondérant différemment l'écart entre les fonctions de répartition théorique et empirique à certains endroits. Son choix dépendra des lois alternatives dont on veut prioriser le rejet de H_0. Lorsque $\psi(t) = 1$, on retrouve le test de CVM avec l'équation (6) et KS avec l'équation (5). Lorsque $\psi(t) = \frac{1}{t(1-t)}$ pour (6), il s'agit de la statistique d'AD.

1.4.2 Cramér-von Mises (CVM)

La statistique de CVM évalue une distance, mais contrairement à KS, celle-ci n'est pas représentée par l'écart maximal entre les deux fonctions, mais par la norme euclidienne de l'espace L^p. Cette statistique, *distribution-free* comme KS (*i.e.* la statistique de test ne dépend pas de F), est donnée par :

$$n\omega^2 = n \int_{-\infty}^{\infty} [\hat{F}_n(x) - F_0(x)]^2 \,\mathrm{d}F_0(x),$$

où F_0 est la fonction de répartition théorique de l'hypothèse H_0.

On peut exprimer cette statistique plus simplement. Soit $X_1, ...X_n$ un échantillon aléatoire de taille n, $X_{(1)}, ..., X_{(n)}$ ses statistiques d'ordre et $\hat{F}_n$ sa fonction de répartition empirique.

En notant $Z_i = F_0(X_{(i)})$, on a alors :

$$\begin{aligned}
\omega^2 &= \int_{-\infty}^{\infty} \left[\hat{F}_n(x) - F_0(x)\right]^2 \mathrm{d}F_0(x) \\
\text{(Chasles)} &= \int_{-\infty}^{X_{(1)}} [-F_0(x)]^2 \,\mathrm{d}F_0(x) + \sum_{i=1}^{n-1} \int_{X_{(i)}}^{X_{(i+1)}} \left[\frac{i}{n} - F_0(x)\right]^2 \mathrm{d}F_0(x) + \int_{X_{(n)}}^{\infty} [1 - F_0(x)]^2 \,\mathrm{d}F_0(x) \\
&= \frac{1}{3}Z_1^3 + \frac{1}{3}\sum_{i=1}^{n-1}\left[\left(Z_{i+1} - \frac{i}{n}\right)^3 - \left(Z_i - \frac{i}{n}\right)^3\right] - \frac{1}{3}(Z_n - 1)^3 \\
&= \frac{1}{3}Z_1^3 + \frac{1}{3}\sum_{i=1}^{n-1}\left[Z_{i+1}^3 - Z_i^3 + \frac{3i^2}{n^2}(Z_{i+1} - Z_i) - \frac{3i}{n}\left(Z_{i+1}^2 - Z_i^2\right)\right] - \frac{1}{3}(Z_n - 1)^3 \\
&= \frac{1}{3}Z_1^3 + \frac{1}{3}Z_n^3 - \frac{1}{3}Z_1^3 + \left(Z_n - \sum_{i=1}^{n}\frac{2i-1}{n^2}Z_i\right) - \left(Z_n^2 - \sum_{i=1}^{n}\frac{1}{n}Z_i^2\right) - \frac{1}{3}(Z_n - 1)^3 \\
&= \frac{1}{3} + \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^2 - \frac{2i-1}{n}Z_i\right) \\
\text{(Id. rem.)} &= \frac{1}{3} + \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \frac{2i-1}{2n}\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{2n}\right)^2 \\
&= \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \frac{2i-1}{2n}\right)^2. \qquad \left[\text{car } \sum_{i=1}^{n}(2i-1)^2 = \frac{1}{3}n(4n^2-1)\right]
\end{aligned}$$

Détail du passage de la ligne 4 à 5 :

$$\begin{aligned}
\sum_{i=1}^{n-1} i^2(Z_{i+1} - Z_i) &= \sum_{i=2}^{n}(i-1)^2 Z_i - \sum_{i=1}^{n-1} i^2 Z_i \\
&= (n-1)^2 Z_n - Z_1 - \sum_{i=2}^{n-1}(2i-1)Z_i \\
&= n^2 Z_n - \sum_{i=1}^{n}(2i-1)Z_i.
\end{aligned}$$

Explicitation du résultat entre crochets (dernière ligne) :

$$\begin{aligned}
\sum_{i=1}^{n}(2i-1)^2 &= 4\sum_{i=1}^{n} i^2 - 4\sum_{i=1}^{n} i + \sum_{i=1}^{n} 1 \\
&= \frac{4n(n+1)(2n+1)}{6} - \frac{4n(n+1)}{2} + n \\
&= \frac{(4n^2+4n)(2n+1) - 12n(n+1) + 6n}{6} \\
&= \frac{8n^3 + 12n^2 + 4n - 12n^2 - 12n + 6n}{6} \\
&= \frac{8n^3 - 2n}{6} \\
&= \frac{1}{3}n(4n^2-1)
\end{aligned}$$

La statistique de test est donc calculée en pratique via la formule suivante (en multipliant le résultat par n) :

$$\mathrm{CVM} = \frac{1}{12n} + \sum_{i=1}^{n}\left(F_0(x_{(i)}) - \frac{2i-1}{2n}\right)^2$$

Dans le cas d'un test de normalité, étant donné que $F_0 = F_{\mu,\sigma}$, elle peut s'écrire :

$$\mathrm{CVM} = \frac{1}{12n} + \sum_{i=1}^{n}\left(\Phi\Big(\frac{x_{(i)}-\mu}{\sigma}\Big) - \frac{2i-1}{2n}\right)^2$$

avec Φ fonction de répartition de la loi $\mathcal{N}(0,1)$.

A l'instar de KS, la statistique CVM mesure une distance. On rejetera H_0 lorsque les valeurs seront trop grandes : ce test est donc unilatéral à droite. De plus, comme CVM est fondée sur une norme euclidienne, et non un maximum comme KS, elle est moins sensible que son homologue aux *outliers*.

1.4.3 Anderson-Darling (AD)

La statistique du test d'AD est plus sensible aux extrémités des fonctions de répartition. Cette sensibilité est due à sa fonction poids. Cependant cette pondération a une conséquence lourde sur les temps de calculs puisque les valeurs critiques doivent être recalculées pour chaque loi et chaque poids. Comme les autres test basés sur la fonction de répartition empirique, AD est distribution free :

$$\mathrm{AD} = n\int_{-\infty}^{\infty} \frac{[\hat{F}_n(x) - F_0(x)]^2}{F_0(x)(1-F_0(x))}\,\mathrm{d}F_0(x) \tag{8}$$

La formule simplifié pour AD se déduit en remplaçant le poids par $\frac{1}{t(1-t)}$ dans l'équation (7) :

$$\mathrm{AD} = -n - \sum_{i=1}^{n}\frac{2i-1}{n}\Big[ln(\Phi\Big(\frac{x_{(i)}-\mu}{\sigma}\Big)) + ln(1-\Phi\Big(\frac{x_{(i)}-\mu}{\sigma}\Big))\Big] \tag{9}$$

Des valeurs d'AD et CVM au-delà d'un certain seuil z_α conduisent au rejet de l'hypothèse nulle. Il existe une loi limite à chacune de ces statistiques. Tout comme KS, déterminer ces lois limites revient à étudier les processus stochastiques.

1.4.4 Réduction à un problème stochastique

Pour la suite, nous supposons que les paramètres de F_0 sont connus. Le principe est assez similaire à celui de Kolmogorov. Sous H_0, la fonction $F_0 = F$ étant continue : $F(X_i) = U_i \sim \mathcal{U}[0,1]$ pour $i = 1, ..., n$. On note $\hat{G}_n$ la fonction empirique de $U_1, ..., U_n$ et en effectuant le changement de variable $u = F(x)$ on peut réécrire W_n^2 :

$$W_n^2 = n \int_0^1 [\hat{G}_n(u) - u]^2 \psi(u)\, \mathrm{d}u$$

On a alors le processus stochastique Y_n indépendant de F et indexé par u :

$$Y_n(u) = \sqrt{n}[\hat{G}_n(u) - u] \equiv \sqrt{n}[\hat{F_n}(x) - F(x)]$$

Soit $F_{W_n^2}$ la fonction de répartition de W_n^2 dont nous voulons connaitre la limite :

$$F_{W_n^2}(z) = \mathbb{P}(W_n^2 \leq z) = \mathbb{P}\Big\{ \int_0^1 Y_n^2(u)\psi(u)\ \mathrm{d}u \leq z \Big\}$$

$Y_n(u)$ est un processus gaussien à temps discret dont la limite $Y(u)$ ($u \in [0,1]$) est un processus de Wiener. En effet les $Y_n(u_i)$ sont des variables aléatoires gaussienne *i.i.d* de moyenne nulle et $cov(Y_n(u_i), Y_n(v_j)) = min(u_i, v_i) - u_i v_i$.

Par passage à la limite quand $n \longrightarrow \infty$ on a :

$$\mathbb{E}(Y(u)) = 0 \quad et \quad cov(Y(u), Y(v) = min(u, v) - uv \quad u, v \in [0,1]$$

Donc $Y(u)$ est bien un processus de Wiener

La propriété suivante est une conséquence du théorème de Donsker (Théorème 1.3) :

Propriété 4. *Sous certaines conditions sur $\psi(u)$ alors,*

$$F_{W_n^2}(z) \overset{n\to\infty}{\longrightarrow} \mathbb{P}\Big\{ \int_0^1 Y^2(u)\psi(u)du \leq z \Big\} = F_{W^2}(z)$$

$F_{W^2}(z)$ est la fonction de répartition de la loi limite de W_n^2

Calcul des valeurs critiques : résolution d'équations différentielles

Pour déterminer les valeurs critiques de W^2 il faut écrire sa fonction de répartition sous une forme plus explicite. Cela revient à résoudre une équation différentielle. Pour retrouver la loi limite de la statistique AD, on doit considérer le cas $\psi(t) = 1/t(1-t)$. Pour CVM $\psi(t) = 1$.

Les solutions s'expriment sous forme de série :

$$\begin{aligned} a_2(z) &= \Pr\{W^2 \leqq z\} \\ &= \sqrt{\frac{\pi}{2}}\frac{1}{z}\sum_{j=0}^{\infty}\binom{-\frac{1}{2}}{j}(4j+1)\int_0^1 e^{(rz)/8-((4j+1)^2\pi^2)/(8rz)}\frac{dr}{r^{3/2}(1-r)^{1/2}} \\ &= \frac{\sqrt{2\pi}}{z}\sum_{j=0}^{\infty}\binom{-\frac{1}{2}}{j}(4j+1)e^{-((4j+1)^2\pi^2)/(8z)}\int_0^{\infty} e^{z/(8(w^2+1))-((4j+1)^2\pi^2w^2)/(8z)}\,dw. \end{aligned}$$

FIGURE 7 – Fonction de répartition de la loi limite de AD

$$a_1(z) = \frac{1}{\pi\sqrt{z}} \sum_{j=0}^{\infty} (-1)^j \binom{-\frac{1}{2}}{j} \cdot (4j+1)^{\frac{1}{2}} e^{-(4j+1)^2/(16z)} K_{\frac{1}{4}}\left((4j+1)^2/(16z)\right).$$

where $K_{\frac{1}{4}}(x)$ is the standard Bessel function

FIGURE 8 – Fonction de répartition de la loi limite de CVM

La démonstration est disponible dans la bibliographie [11].

Ainsi quand n est petit, les valeurs critiques de CVM et AD ne dépendent que de n. Et quand n est assez grand, on peut les retrouver en utilisant les lois limites qui ne dépendent plus de n.

En pratique...

Pour les mêmes raisons que KS, la fonction $F_{\mu,\sigma}$ ne peut pas toujours être spécifiée. Si μ et σ ne sont pas connus, CVM et AD peuvent être calculés en utilisant les estimateurs empiriques $\hat{\mu}$ et $\hat{\sigma}$. Dans ce cas les statistiques AD et CVM et leurs valeurs critiques ne sont plus *distribution-free* mais restent invariantes pour la famille des loi normales : il faut les estimer par Monte-Carlo.

De plus AD doit être utilisée dans sa version corrigée :

$$\text{AD}^* = \text{AD}\left(1 - \frac{4}{n} + \frac{25}{n^2}\right)$$

Si n est assez grand, cette correction devient négligeable.

2 Simulations

Dans cette section, nous allons présenter la méthode qui a permis de tracer les courbes des puissances de chaque test de normalité. Dans un premier temps nous observerons le comportement de la puissance des test lorsque nous "déformons" la gaussienne (en altérant sa symétrie et sa courbure). Ensuite nous tracerons l'évolution de la puissance de chaque test en fonction de la taille n de l'échantillon.

2.1 Calcul des puissances empiriques

Soit $X_1, ..., X_n$ un échantillon de variables aléatoires de densité f non-normale.

Un test de normalité, par définition, permet de tester l'hypothèse (pour μ,σ^2 quelconques) :

$$H_0 : X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2) \quad \text{vs} \quad H_1 : X_1, ..., X_n \nsim \mathcal{N}(\mu, \sigma^2)$$

Ici on sait que pour l'échantillon non-normalement distribué, H_0 est fausse. Par conséquent, la probabilité de rejet de H_0 correspond bien à $1 - \beta$. Si on simule $N = 10000$ fois l'échantillon $X_1, ..., X_n$ et qu'on soumet chacun à un même test de normalité (α fixé), il devrait y avoir un nombre théorique de rejet de H_0 égal à $(1 - \beta) \times N$. Donc, la puissance du test P (à α fixé) peut être estimée par :

$$P = 1 - \beta \approx \text{Taux de rejet de } H_0 = \frac{\text{Nombre de fois que } H_0 \text{ est rejetée}}{N}$$

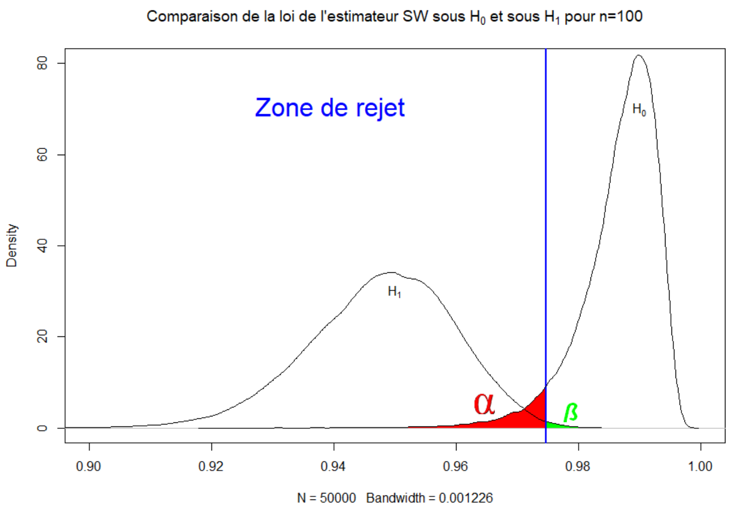


FIGURE 9 – SW sous H_0 et sous H_1(pour $n = 100$)

2.2 Régions critiques

Pour savoir si H_0 est rejetée ou non, il va falloir au préalable définir les régions critiques.

Pour cette étude nous considérerons le cas où μ et σ^2 ne sont pas spécifiés par l'utilisateur du test et sont remplacés par la moyenne et la variance empirique de l'échantillon. Cette situation est la plus courante en pratique lorsque l'on cherche à vérifier la normalité d'un échantillon. Dans ce cas, nous avons vu dans la partie "Validité des tests" que :

— La statistique de SW est pivotale
— Les statistiques basés sur la fonction de répartition empirique ($KS/LL/CVM/AD$) ne sont plus *distribution free* mais pour une même famille de loi, elles sont invariantes pour le paramètre de position μ et d'échelle σ.

Ainsi dans le cadre des tests de normalité, toutes les valeurs critique sont approximables par Monte carlo sur l'unique loi $\mathcal{N}(0,1)$.

Pour le test de KS, nous calculerons la statistique en prenant $F_0 = F_{\hat{\mu},\hat{\sigma}}$ avec les paramètres estimés. Cependant nous utiliserons la table des valeurs critiques de KS (car utiliser des valeurs critiques obtenues par Monte-Carlo reviendrait à faire le test de LL)

La Région critique $RC(\alpha)$ d'un test dépend des quantiles de sa loi (sous H_0) et de α. Donc il faut estimer les quantiles pour chaque test et pour chaque n.

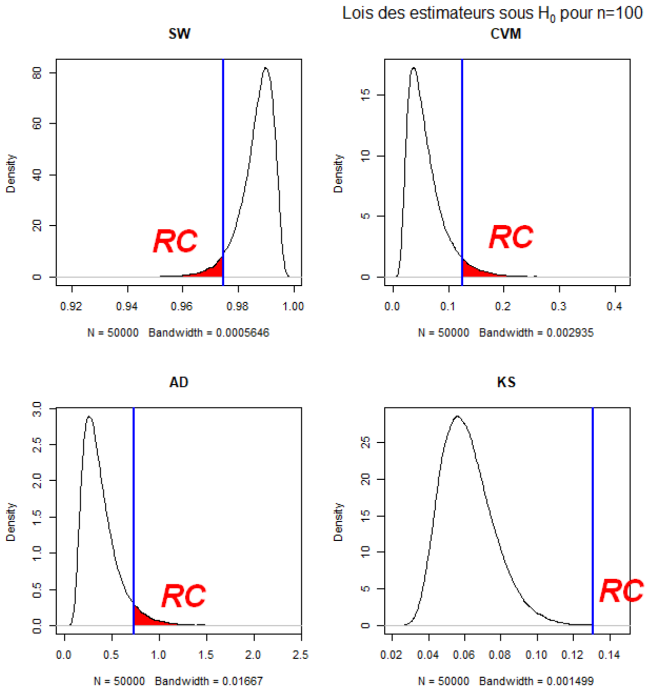


FIGURE 10 – Densité sous H_0 des estimateurs (pour $n = 100$)

Le test de SW est unilatéral gauche :

$$RC(\alpha) = \{T < q_0(\alpha)\}$$

Les tests de $KS/LL/CVM/AD$ sont unilatéraux droite :

$$RC(\alpha) = \{T > q_0(1-\alpha)\}$$

2.3 Estimation des valeurs critiques par Monte-Carlo

Comme les lois sous H_0 de ces statistiques T sont inconnues, il faut estimer les quantiles q_0 (valeurs critiques) par les quantiles empiriques. Pour ce faire, nous utilisons les statistiques d'ordres :

On fixe $\alpha = 5\%$

Pour des valeurs de n différentes :

1. Simuler $N = 50000$ n-échantillons $(Z_1, .., Z_n) \sim \mathcal{N}(0,1)$ (donc H_0 vraie)
2. Générer alors $N = 50000$ variables aléatoires $T = T(Z_1, ..., Z_n)$ (suivant bien la loi sous H_0 du Test de normalité car H_0 vraie)
3. Ordonner les N variables $(T_{(1)}, ..., T_{(N)})$
4. Estimer les quantiles q_0 par les quantiles empiriques :
   — le quantile empirique $T_{\lceil N\alpha \rceil}$ est un estimateur consistant de $q_0(\alpha)$
   — le quantile empirique $T_{\lceil N(1-\alpha) \rceil}$ est un estimateur consistant de $q_0(1-\alpha)$

3 Résultats

3.1 Courbes des puissances lors d'une déformation de la gaussienne

La GLD (Generalized Lambda Distribution) est une loi très flexible qui est paramétrée par ses moments : moyenne, variance, coefficient d'asymétrie et coefficient d'aplatissement. En spécifiant les bons paramètres à la GLD, il est possible d'approcher la plupart des lois connues

Le coefficient d'asymétrie (*skewness*) et le coefficient d'aplatissement (*kurtosis*) d'une loi normale $\mathcal{N}(0,1)$ sont respectivement 0 et 3.

En simulant des variables aléatoires de loi GLD(0,1,0,3) on obtient la densité suivante :

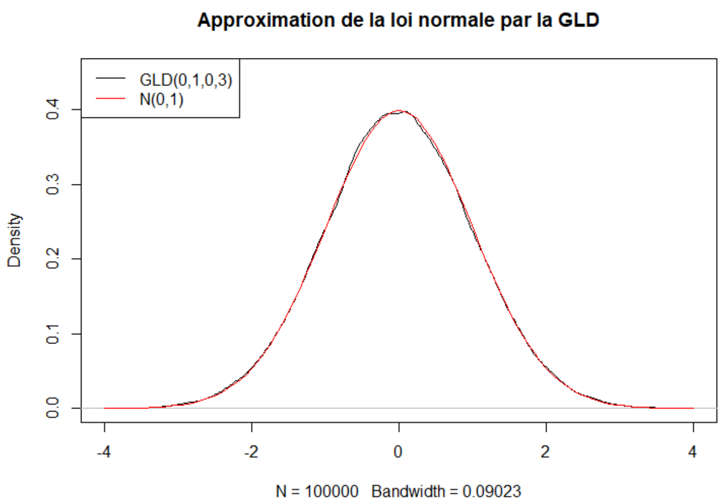


FIGURE 11 – Approximation de la loi normale par la GLD

Ainsi la GLD(0,1,0,3) est une assez bonne approximation de la loi $\mathcal{N}(0,1)$ et en faisant varier le coefficient d'asymétrie et le coefficient d'aplatissement on peut alors simuler une déformation progressive de la gaussienne. De cette façon nous pouvons apprécier l'évolution de la puissance des tests de normalité quand on modifie les paramètres de forme. L'objectif est de tracer les courbes des puissances pour différentes tailles d'échantillon n de la GLD quand elle s'éloigne de la forme d'une loi normale.

La moyenne et la variance de la GLD seront toujours fixé à 0 et 1 respectivement. Lorsque nous varierons le coefficient d'asymétrie, le coefficient d'aplatissement sera fixé à 3 et quand ce sera le coefficient d'aplatissement qui varie, l'asymétrie sera fixée à 0.

3.1.1 Altération de la symétrie

Pour $n = 20$, $n = 50$, $n = 100$ et n=2000, on obtient les courbes suivantes :

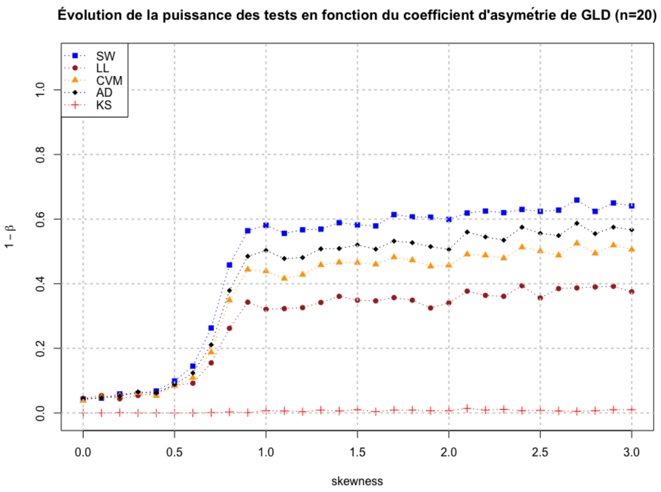


FIGURE 12 – $n = 20$

Quand nous déformons le caractère symétrique de la GLD, la puissance des tests augmente. Cependant elle n'atteint pas 1. On notera que pour le test de KS elle reste quasi nulle.

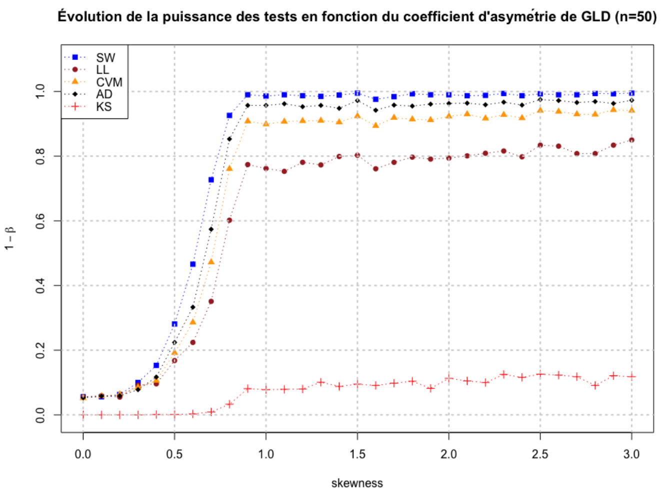


FIGURE 13 – $n = 50$

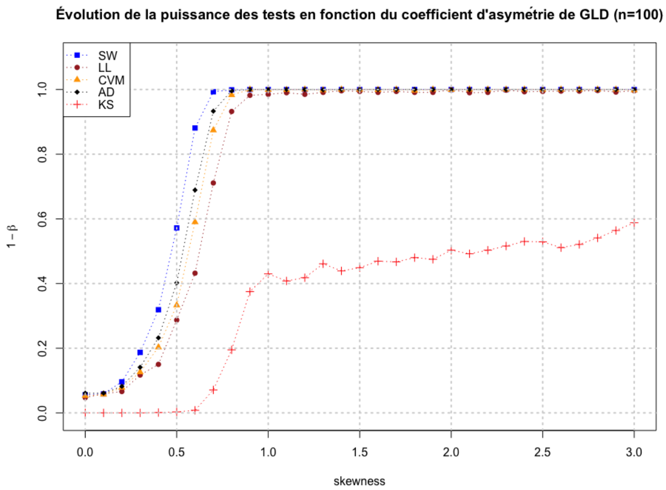


FIGURE 14 – $n = 100$

Quand $n = 50$ et $n = 100$, les figures ci-dessus nous montrent une amélioration globale de la qualité des tests. En effet, lorsque l'on s'éloigne de 0, la puissance des tests monte rapidement vers 1. Mais la performance du test de KS reste discutable.

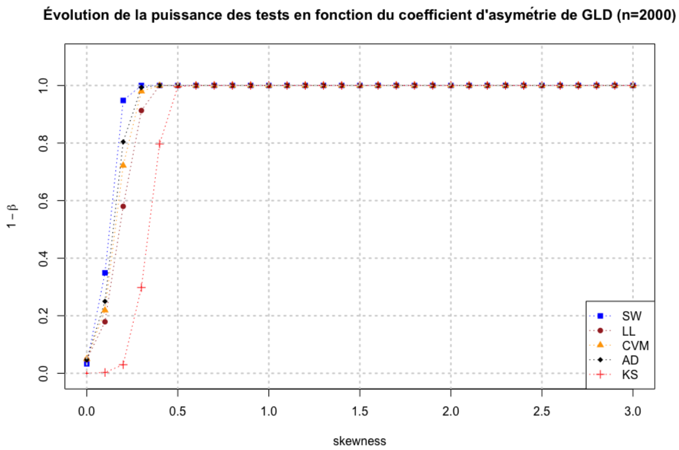


FIGURE 15 – $n = 2000$

Enfin, lorsque $n = 2000$ la quantité d'information est tellement importante que même KS parvient à rejeter H_0 dès que l'asymétrie est suffisante.

En faisant varier le coefficient d'asymétrie, on évalue la capacité d'un test à détecter un défaut de symétrie. Pour toutes les taille d'échantillon n, c'est le test de SW suivi du test d'AD qui ont montré les meilleurs résultats. Le test Kolmogorov-Smirnov est beaucoup trop conservatif à cause d'un problème dans le choix de la table des valeurs critiques (voir la partie sur Lilliefors).

3.1.2 Variation du kurtosis

Ici, le coefficient d'asymétrie est fixé à 0 et le coefficient d'aplatissement varie.

Pour $n = 20$, $n = 50$, $n = 100$ et n=2000 on a les courbes suivantes :

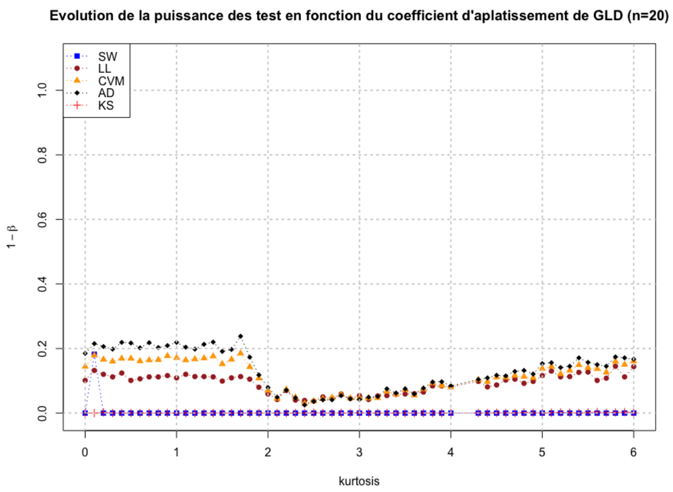


FIGURE 16 – $n = 20$

Pour un échantillon de taille 20, les test ne sont pas puissants. C'est d'autant plus vrai pour SW et KS qui ne détectent pas du tout la variation du coefficient d'aplatissement.

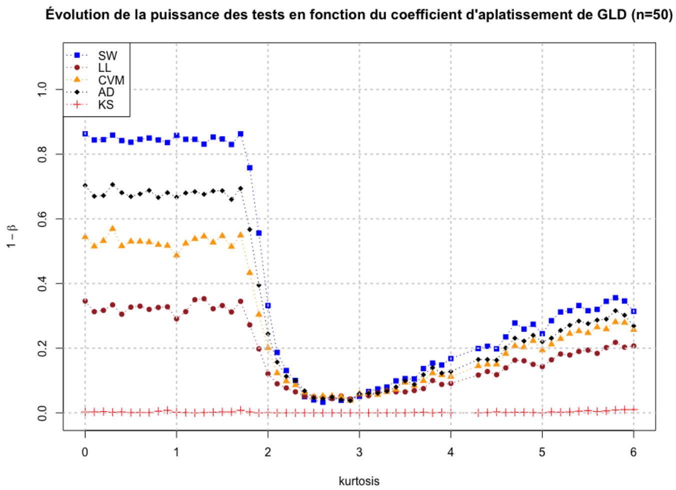


FIGURE 17 – $n = 50$

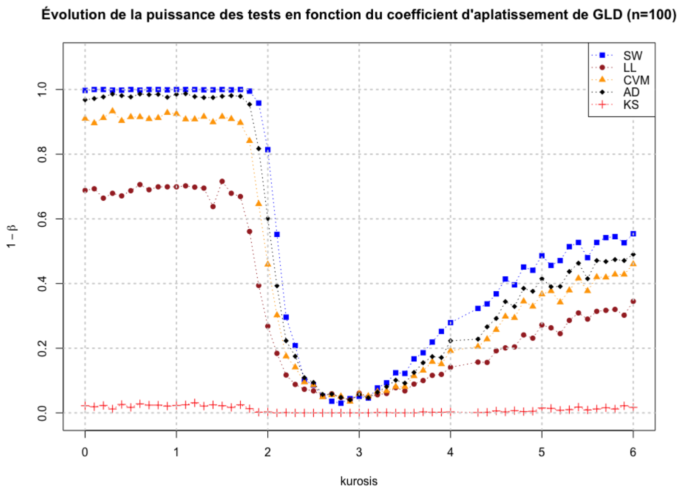


FIGURE 18 – $n = 100$

La puissance des tests s'améliore lorsque $n = 50$ ou $n = 100$. De plus, les courbes des puissances ne sont pas symétrique par rapport à l'axe $x = 3$, ceci montre que les tests ont plus de difficulté à détecter le caractère trop "pointu" de la distribution que trop "plat"

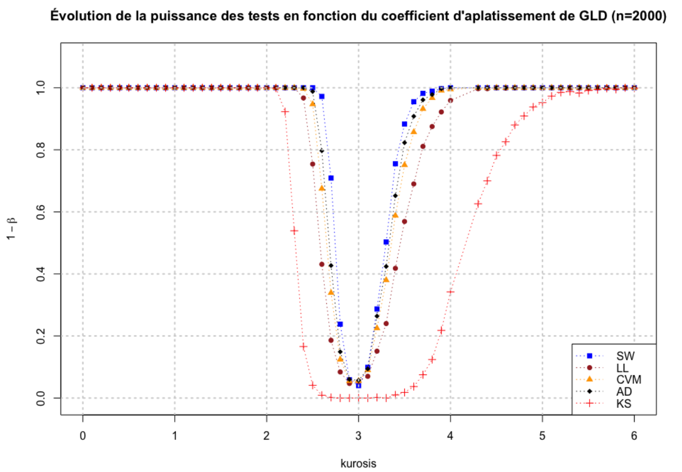


FIGURE 19 – $n = 2000$

En conclusion, pour détecter une asymétrie dans la gaussienne, les tests de SW et AD seront les plus fiables et pour détecter un problème de courbure, les tests seront moins performant sur des distributions trop 'pointues'. De plus quand l'échantillon est petit le test de SW se montre inefficace.

Ici nous pouvons voir que la taille de l'échantillon a une influence sur la puissance des tests, la suite logique est donc de tracer les courbes des puissance en fonction de n afin de trouver le test de normalité le mieux adapté à chaque taille d'échantillon. Nous effectuerons ce travail sur plusieurs familles de loi.

3.2 Courbes des puissances en fonction de n

Cette section a pour but de présenter, d'évaluer et de comparer les puissances de 5 tests : Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, Cramér-von Mises et Anderson-Darling selon les lois alternatives évoquées en §3.2.1.

Les valeurs critiques utilisées pour calculer la puissance du test de Shapiro-Wilk sont composées des 50 premières valeurs de sa table (Annexe B) puis à partir de 55 jusqu'à 2000 des valeurs critiques empiriques. Celles du test de Kolmogorov-Smirnov sont toutes issues de sa table disponible en Annexe D. Quant aux tests de Cramér-Von Mises, Anderson-Darling et Lilliefors, leurs valeurs critiques sont les valeurs critiques empiriques dont la démarche de calcul est détaillée au §2.3.

Le code utilisé pour calculer la puissance emprique des test est disponible en Annexe.

3.2.1 Lois alternatives testées

Nous avons judicieusement choisi nos lois alternatives afin de tester les puissances sur des distributions de forme variées :

— La loi Logistique(0,1)
— La loi Uniforme sur $[0, 1]$: $\mathcal{U}[0, 1]$
— La loi Bêta : $\beta(2, 1)$
— La loi Bêta : $\beta(3, 2)$
— La loi du Chi-2 à 4 degrés de liberté : $\chi^2(4)$
— La loi du Chi-2 à 10 degrés de liberté : $\chi^2(10)$
— La loi du Chi-2 à 20 degrés de liberté : $\chi^2(20)$
— La loi des lambdas généralisés (GLD) : $GLD(0, 1, \frac{3}{4}, \frac{3}{4})$
— La loi Gamma : $\Gamma(4, \frac{1}{5})$
— La loi de Laplace : $\mathcal{L}(0, 1)$
— La loi Normale "Location-Contaminated" (à interférences de position) : $LoConN(0.2, 3)$ et $LoConN(0.05, 3)$
— La loi Normale "Scale-Contaminated" (à interférences d'échelle) : $ScConN(0.2, 3)$ et $ScConN(0.05, 3)$
— La loi de Student à 10 degrés de liberté : $St(10)$
— La loi de Student à 15 degrés de liberté : $St(15)$
— La loi Normale Tronquée en -2 et 2 : $TruncN(-2, 2)$
— La loi de Weibull : $W(3, 1)$
— La loi Log-normale(0,1)

3.2.2 Loi Logistique

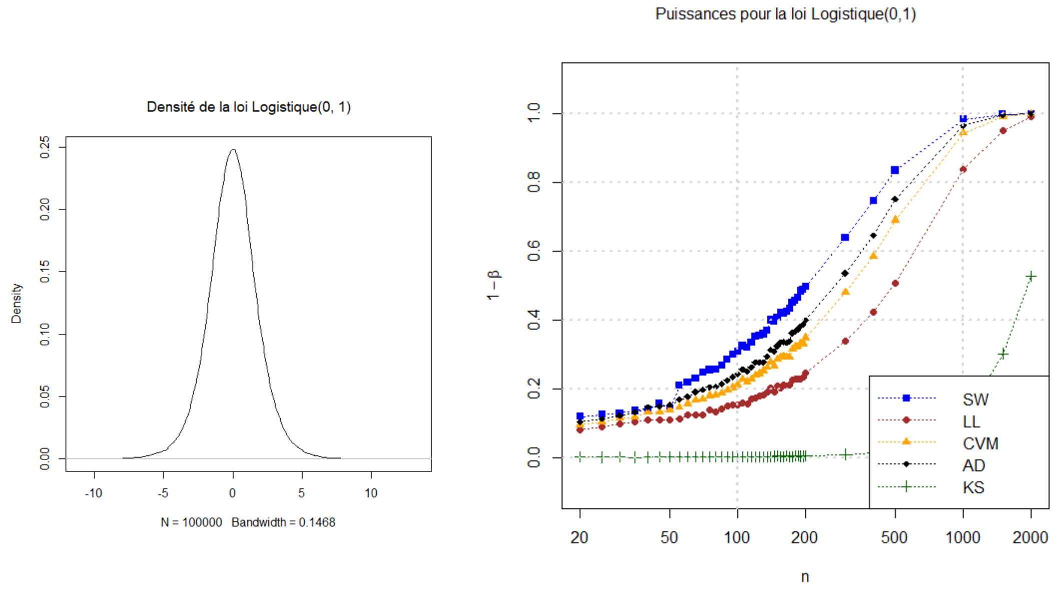


FIGURE 20 – Densité de la loi Logistique(0,1) et puissance des tests

3.2.3 Loi Uniforme

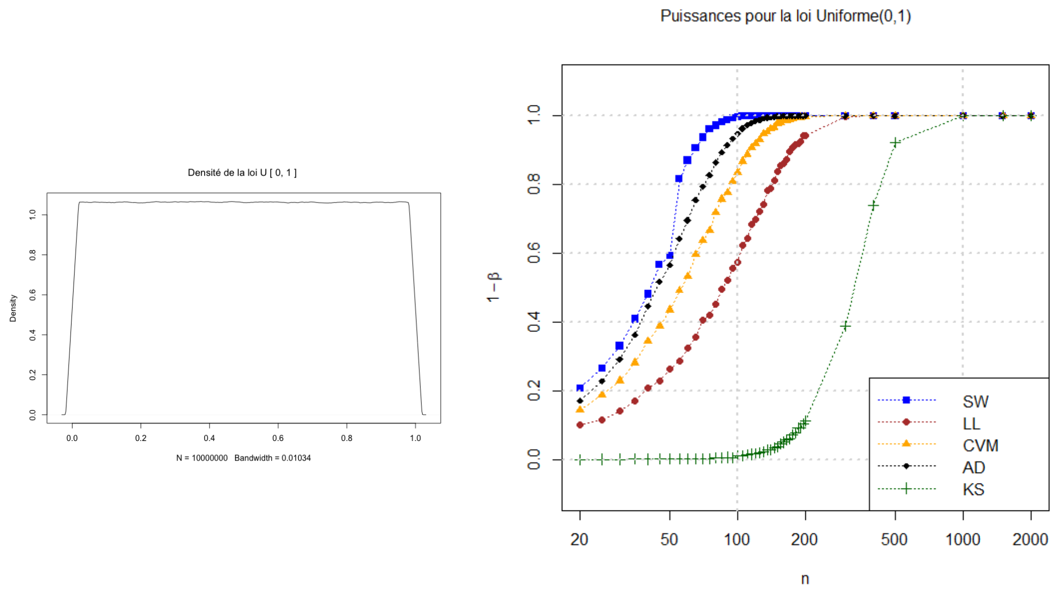


FIGURE 21 – Densité de la loi Uniforme $\mathcal{U}[0,1]$ et puissances des tests

3.2.4 Loi Bêta

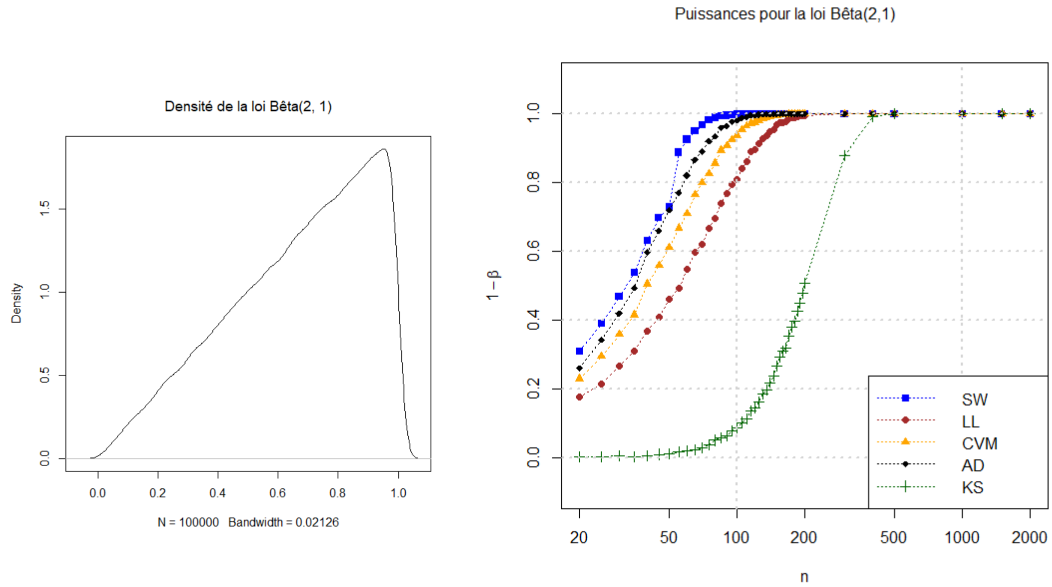


FIGURE 22 – Densité de la loi $\beta(2, 1)$ et puissances des tests

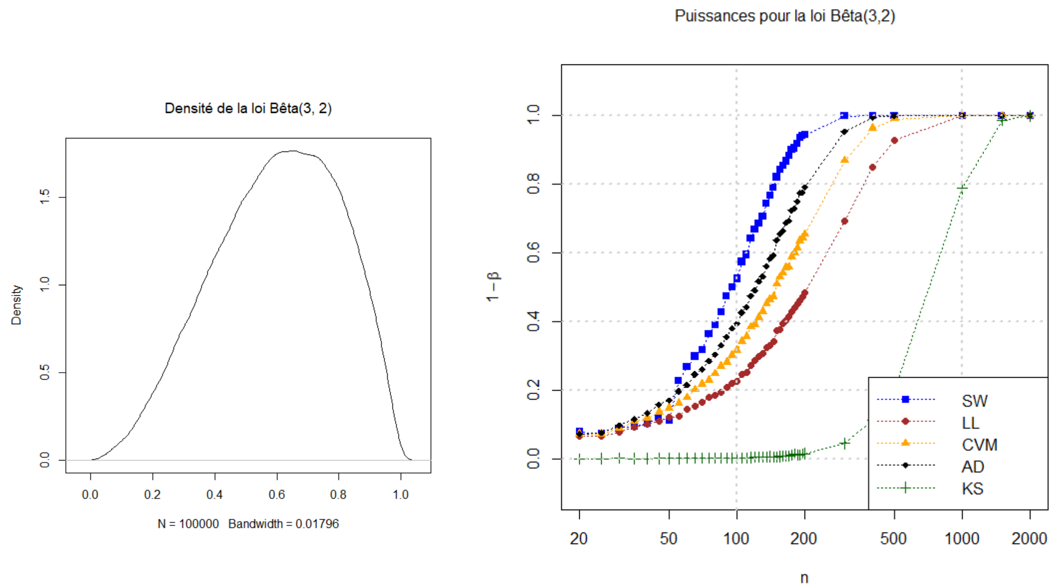


FIGURE 23 – Densité de la loi $\beta(3, 2)$ et puissances des tests

3.2.5 Loi du χ^2

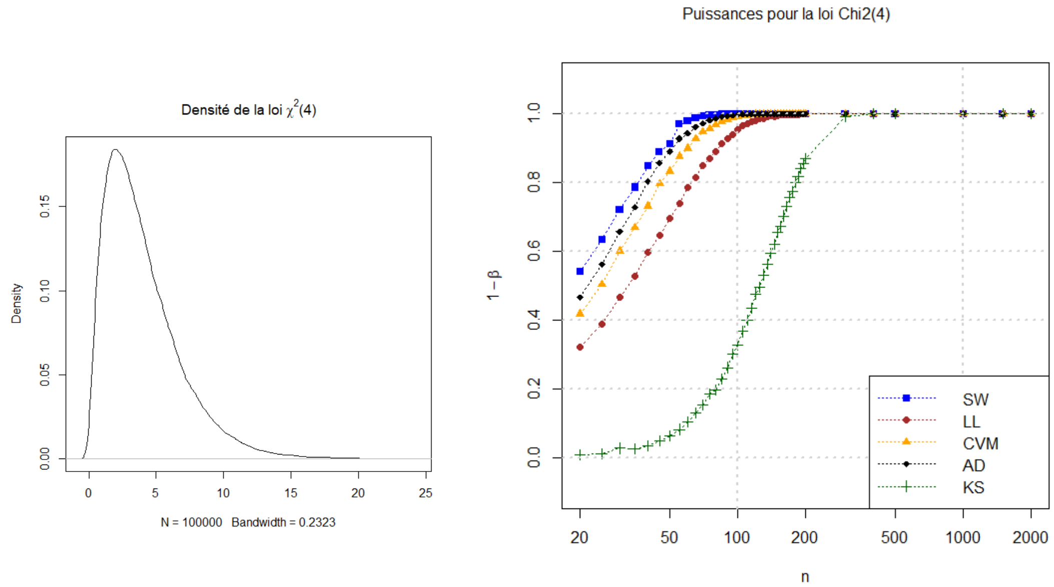


FIGURE 24 – Densité de la loi $\chi^2(4)$ et puissances des tests

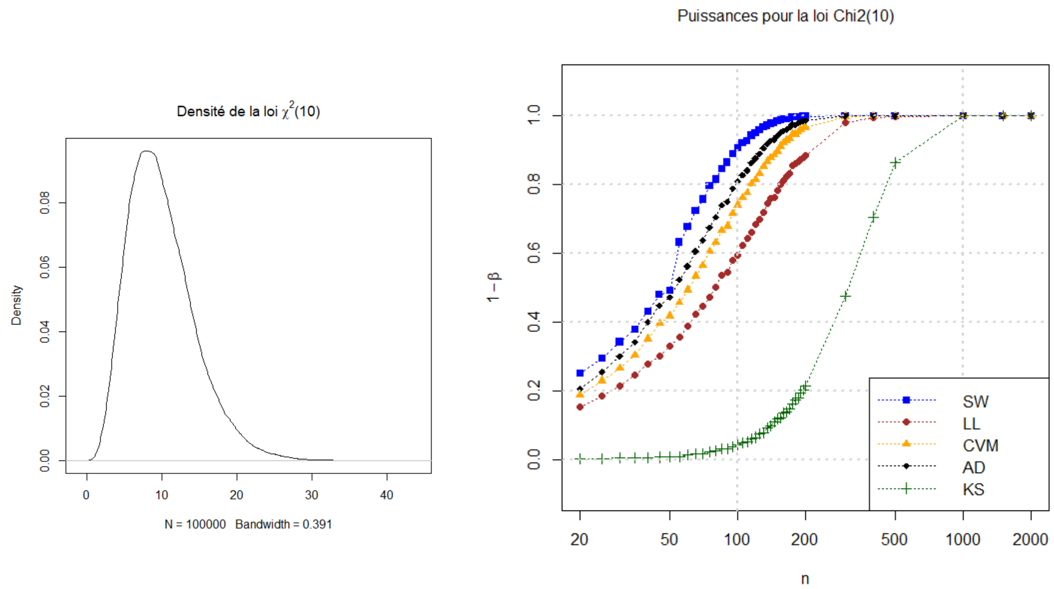


FIGURE 25 – Densité de la loi $\chi^2(10)$ et puissances des tests

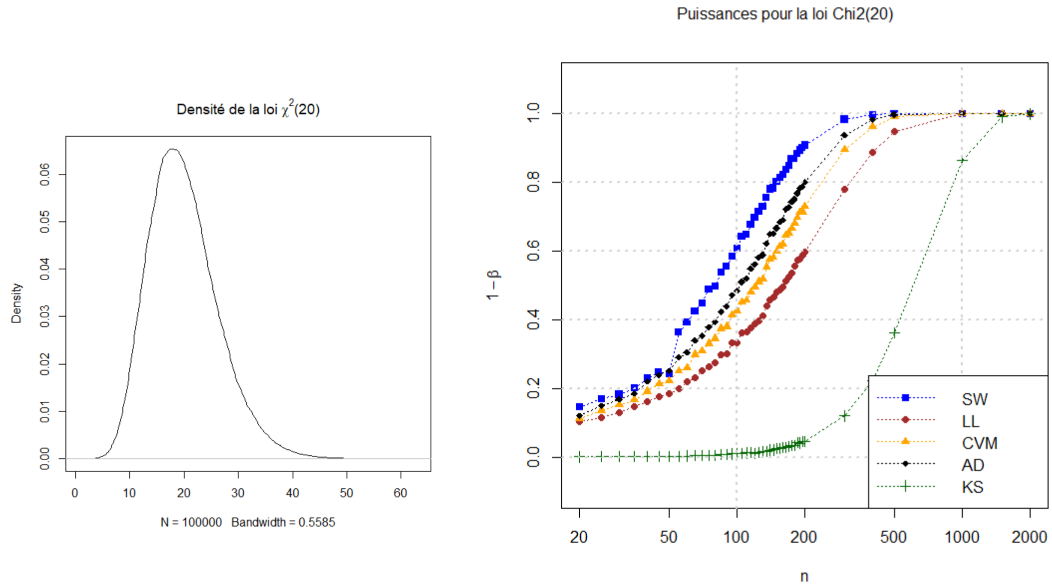


FIGURE 26 – Densité de la loi $\chi^2(20)$ et puissances des tests

3.2.6 La GLD

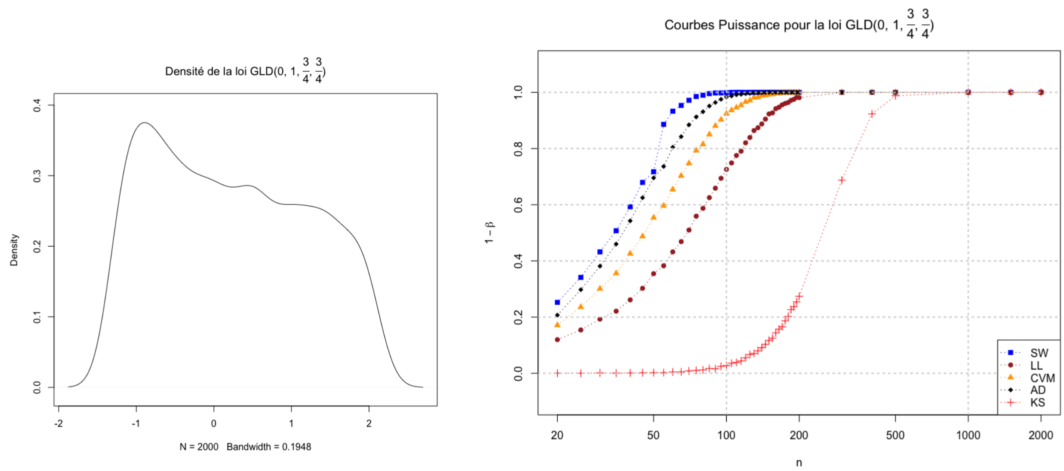


FIGURE 27 – Densité de la GLD(0,1,$\frac{3}{4}$,$\frac{3}{4}$) et puissances des tests

3.2.7 Loi Gamma

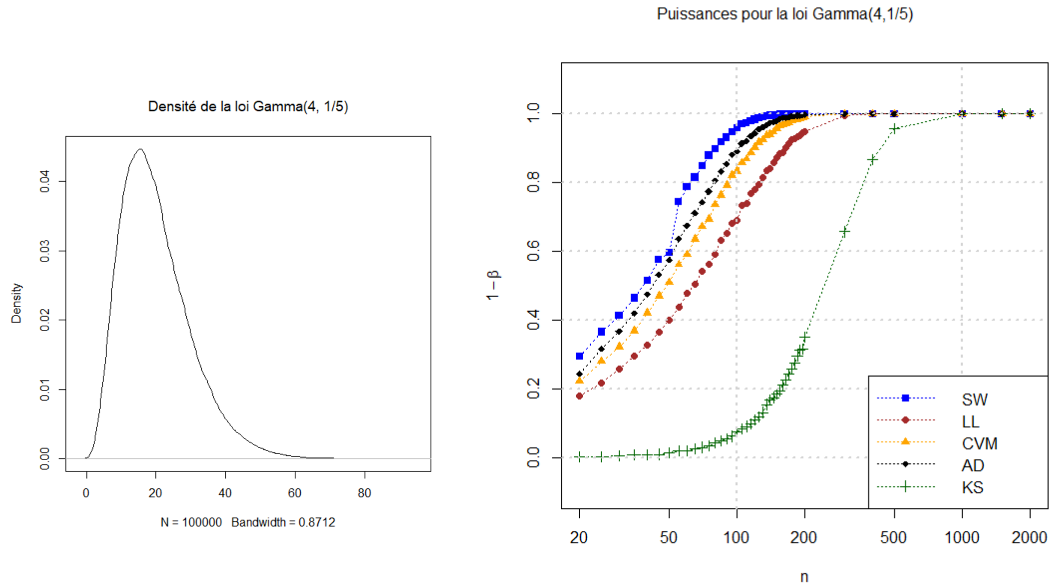


FIGURE 28 – Densité de la loi $\Gamma(4, \frac{1}{5})$ et puissances des tests

3.2.8 Loi de Laplace

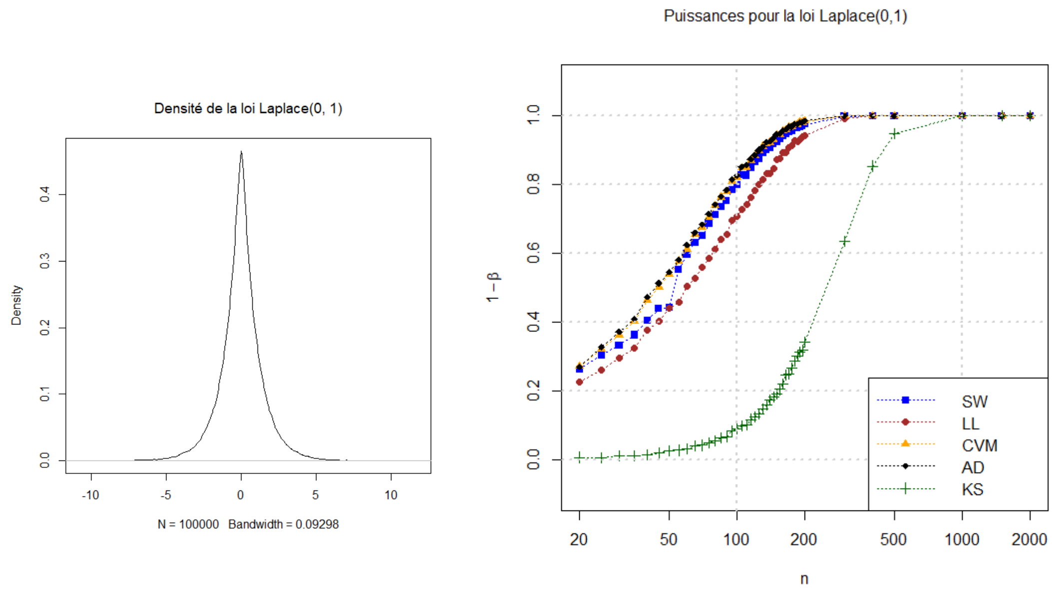
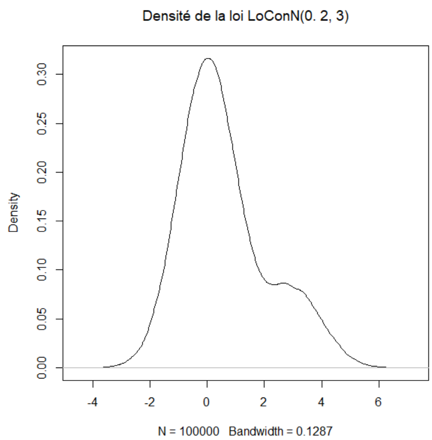


FIGURE 29 – Densité de la loi $\mathcal{L}(0, 1)$ et puissances des tests

3.2.9 Loi Normale "Location-Contaminated" (à interférences de position)

Soit $Z \sim Ber(p)$, alors :

$$X \sim LoConN(p,a) \Longleftrightarrow X \sim \begin{cases} \mathcal{N}(a,1) & \text{si } Z=1 \\ \mathcal{N}(0,1) & \text{si } Z=0 \end{cases}$$



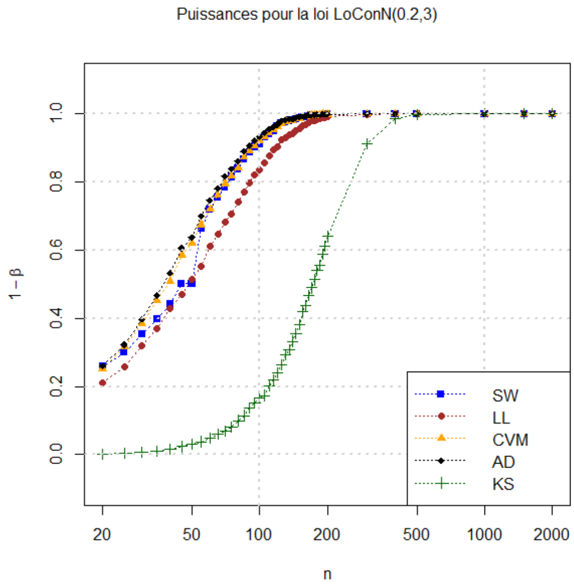
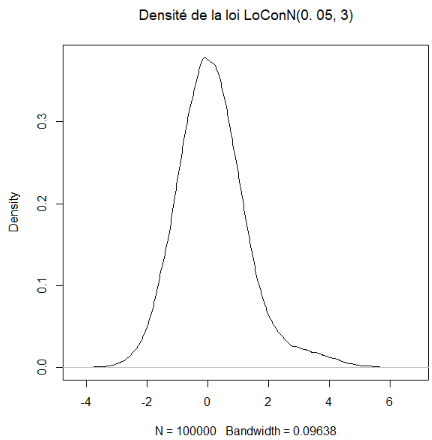


FIGURE 30 – Densité de la loi $LoConN(0.2,3)$ et puissances des tests



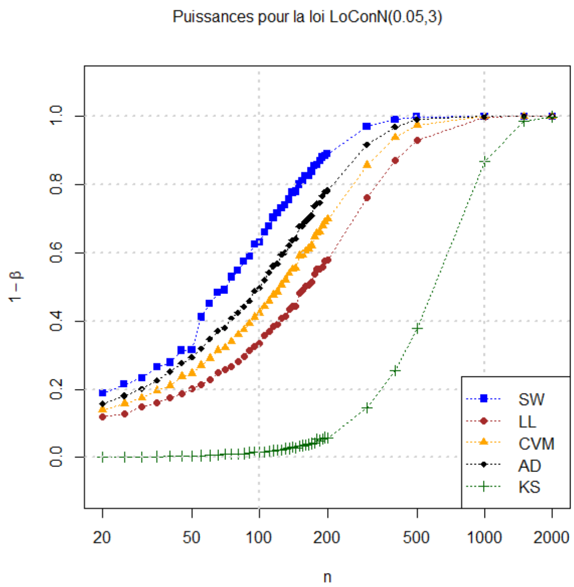


FIGURE 31 – Densité de la loi $LoConN(0.05,3)$ et puissances des tests

3.2.10 Loi Normale "Scale-Contaminated" (à interférences d'échelle)

Soit $Z \sim Ber(p)$, alors :

$$X \sim ScConN(p,b) \iff X \sim \begin{cases} \mathcal{N}(0,b^2) & \text{si } Z=1 \\ \mathcal{N}(0,1) & \text{si } Z=0 \end{cases}$$

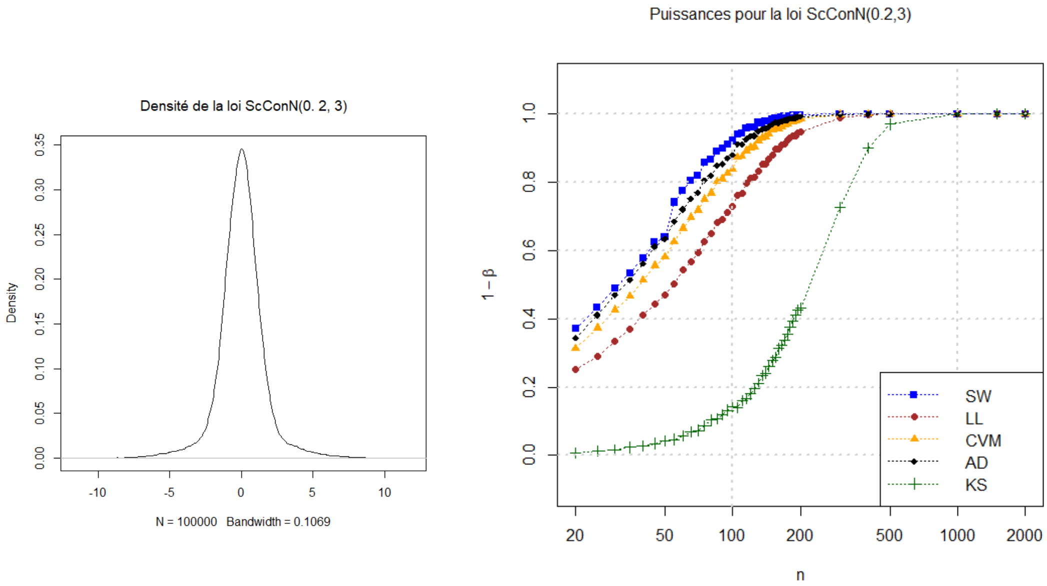


FIGURE 32 – Densité de la loi $ScConN(0.2, 3)$ et puissances des tests

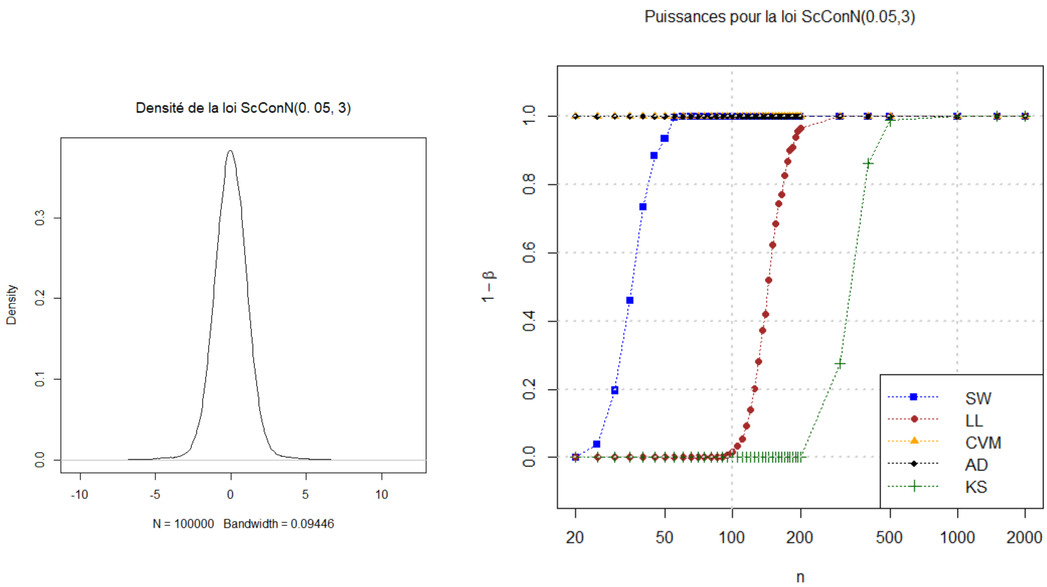


FIGURE 33 – Densité de la loi $ScConN(0.05, 3)$ et puissances des tests

3.2.11 Loi de Student

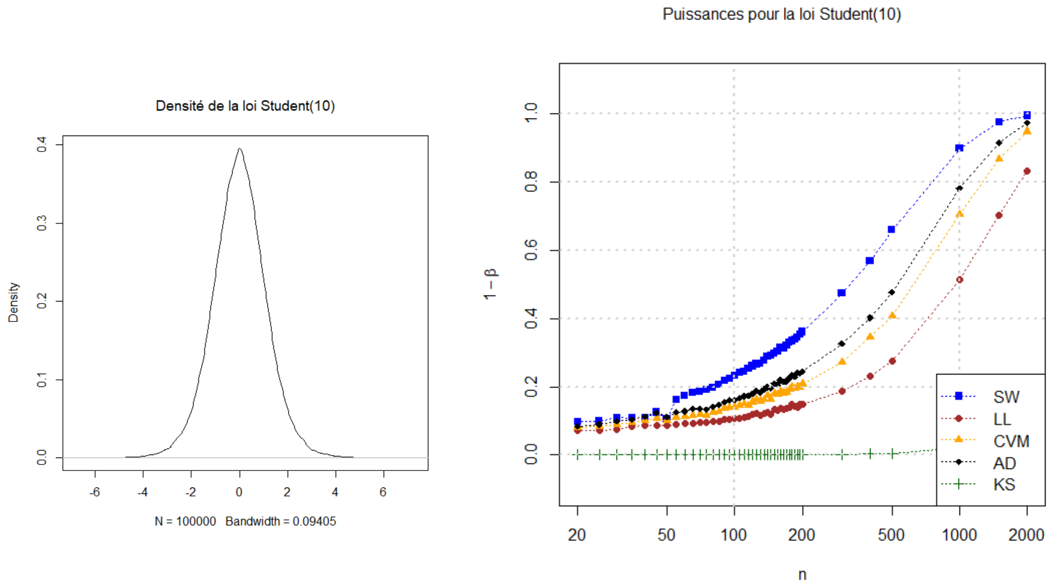


FIGURE 34 – Densité de la loi $St(10)$ et puissances des tests

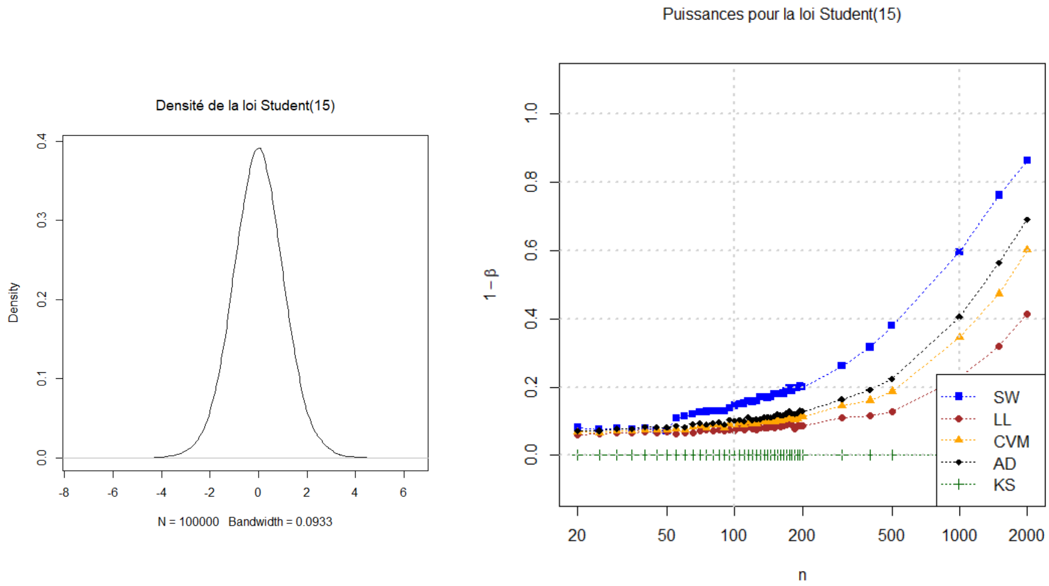


FIGURE 35 – Densité de la loi $St(15)$ et puissances des tests

3.2.12 Loi Normale Tronquée

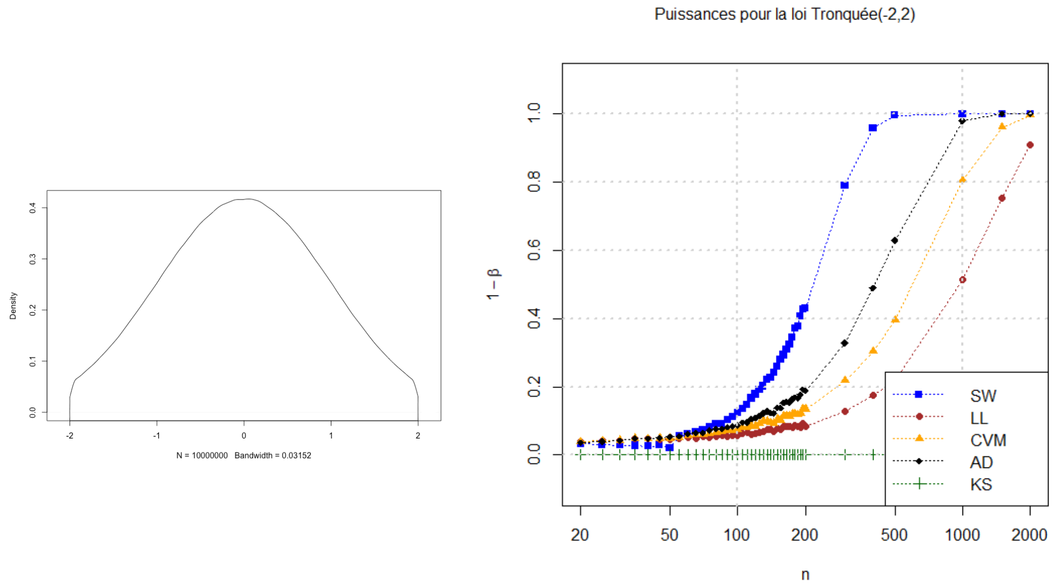


FIGURE 36 – Densité de la loi $TruncN[-2,2]$ et puissances des tests

3.2.13 Loi Weibull

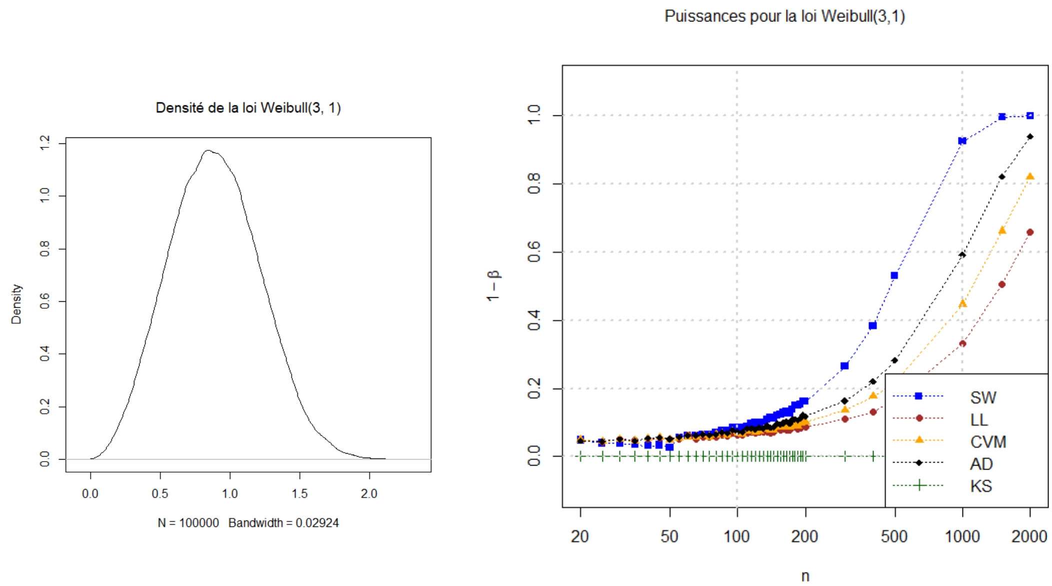


FIGURE 37 – Densité de la loi $W(3,1)$ et puissances des tests

3.2.14 Loi Log-normale

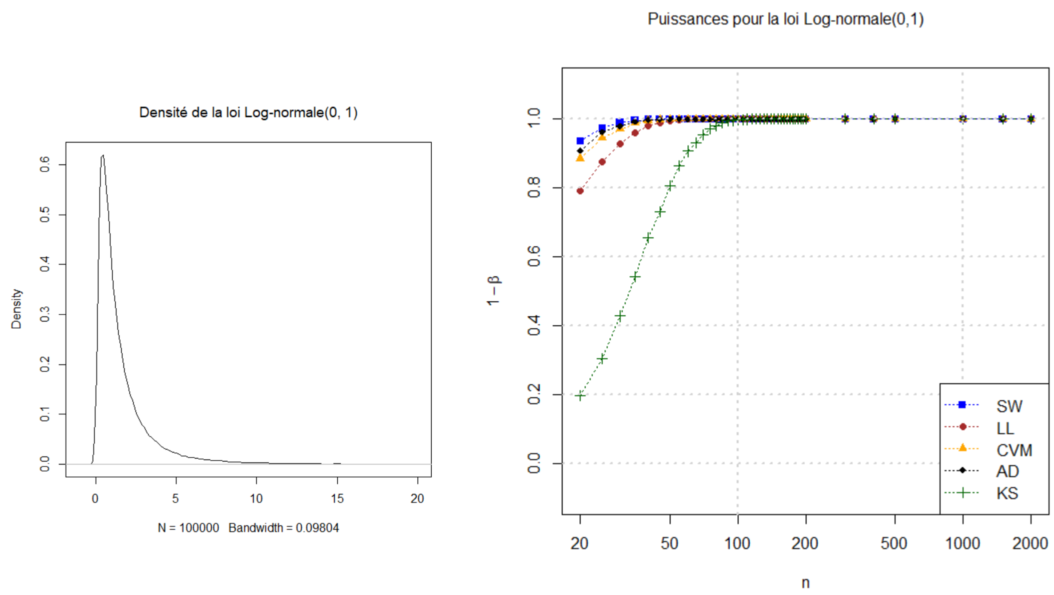


FIGURE 38 – Densité de la loi Log-normale(0,1) et puissances des tests

3.2.15 Interprétations

Dans cette partie, nous tenterons de décrire et interpréter les courbes précédentes et d'en déduire les tests les plus adaptés en différentes situations.

Les lois alternatives étudiées peuvent être divisées en deux catégories :

La première catégorie est celle des lois asymétriques, elle regroupe les lois β (figure 22 et 23), χ^2 (figure 24, 25 et 26), la GLD$(0, 1, \frac{3}{4}, \frac{3}{4})$ (figure 27), la loi Γ (figure 28), LoConN (figure 30 et 31) et Weibull (figure 37). On observe pour ces lois un schéma récurrent : *SW* est le meilleur test suivi de *AD* puis *CVM*. Cependant, dans certains cas *AD* puis *CVM* sont les meilleurs tests, notamment lorsque $n \leq 50$ pour les loi $\beta(3, 2)$ et Weibull$(3, 1)$ ou pour LoConN$(0.2, 3)$, pour toute taille d'échantillon.

On notera que les courbes de puissance ne sont pas les mêmes pour une même famille de loi selon le paramètre choisi. Par exemple pour la loi χ^2 : la puissance des tests diminue en même temps que le degré de liberté augmente. En effet les χ^2 à faible degré de liberté ont une asymétrie plus prononcées, ainsi la puissance des tests est meilleure.

La deuxième catégorie est représentée par les lois symétriques.

Le test de *SW* montre les meilleurs performances pour la plupart des lois symétriques. D'autres lois préfèrent d'autres tests pour une taille d'échantillon spécifique ou pour toutes tailles d'échantillon. La loi Normale Tronquée, pour des échantillons inférieur à 50, préfère le test de *AD* (ou à performance quasi équivalente *CVM*), on notera que pour cette même taille d'échantillon et pour certaines lois (logistique par exemple) *SW* se démarque difficilement.

La loi Scale-Contaminated peut-être considérée à part. Le test de *SW* est le meilleur test pour la loi ScConN$(0.2, 3)$ (suivi de près de *AD* et *CVM*). Tandis que pour la loi ScConN$(0.05, 3)$ ce sont les tests de *AD* suivi de *CVM* qui montrent les meilleurs résultats.

On pourra noter deux choses. Premièrement, comme pour les lois asymétriques, les paramètres de la loi influencent la puissance des tests, c'est le cas pour la loi Scale-Contaminated avec le paramètre p et de la loi de Student qui lorsque son degré de liberté augmente la puissance des tests diminue. Deuxièmement l'évolution de la puissance des tests progresse lentement vers 1 pour les lois Logistique et de Student.

Finalement, pour toutes les lois on préférera utiliser le test de SW pour tester la normalité, quelques écarts entre les tests peuvent se faire sentir lorsque $n \leq 50$. Pour la loi ScConN où l'on préférera utiliser AD ou CVM qui donne de meilleurs résultats, pour les différents paramètres.

4 Conclusion

Nous avons étudié la validité de plusieurs tests basés sur différents outils mathématiques : SW utilise un coefficient de corrélation et LL, KS et CVM quant à eux évaluent une distance entre les fonctions de répartition empirique et théorique. Certaines de ces statistiques de test sont dites *distribution-free*, *i.e.* elles ne dépendent pas des paramètres de la loi testée. Certaines en revanche dépendent de ces paramètres : il est donc nécessaire de calculer les valeurs critiques empiriquement. En particulier, si l'on procède ainsi pour KS au lieu d'utiliser la table théorique, on obtient le test de LL, ce qui le rend bien plus performant.

On note d'une manière générale une faible performance de Kolmogorov-Smirnov et de très bons résultats pour le test de Shapiro-Wilk. Il est cependant important de garder à l'esprit qu'il faut s'adapter à la taille de l'échantillon testé et ne pas mettre de côté AD notamment. Le test de Shapiro-Wilk semble être, dans la majeure partie des cas, le plus performant des cinq pour différentes lois non Normales mais il possède aussi ses inconvénients, il est par exemple connu que celui ci ne fonctionne pas bien si l'échantillon contient trop de valeurs identiques.

Notre comparaison des test de normalité reste cependant incomplète, il existe encore de nombreux tests différents à étudier. Il faudrait également considérer la situation où les valeurs exactes de μ et σ sont connues pour voir ce que valent KS, CVM et AD dans ce cas là.

Références

[1] B. W. Yap and C. H. Sim, *Comparisons of various types of normality tests*, Journal of Statistical Computation and Simulation, Vol. 81, No. 12, December 2011, 2141-2155.

[2] S. S. Shapiro and M. B. Wilk, *An Analysis of Variance Test for Normality*, Biometrika, Vol. 52, No. 3/4 (Dec., 1965), pp. 591-611.

[3] *The Shapiro-Wilk and related tests for normality.*
`https://math.mit.edu/%7Ermd/465/shapiro.pdf`

[4] Patrick Royston, *Remark AS R94 : A Remark on Algorithm AS 181 : The W-test for Normality*, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 44, No. 4 (1995), pp. 547-551.

[5] *Tests non paramétriques*, Cours de Statistique Inférentielle, Université de Toulouse.

[6] A. Kolmogorov, *Confidence limits for an unknown distibution function*, Moscou, URSS.

[7] W. Feller, *On the Kolmogorov-Smirnov limit theorems for empirical distribution*, Cornell University.

[8] David, F. N., and N. L. Johnson. *"The Probability Integral Transformation When Parameters Are Estimated from the Sample."* Biometrika, vol. 35, no. 1/2, 1948, pp. 182–190. JSTOR, www.jstor.org/stable/2332638. Accessed 10 May 2020.

[9] *Statistics for Application - Section 13 : Kolmogorov-Smirnov test*, 2006.
`https://ocw.mit.edu/index.html`

[10] *On the Kolmogorov–Smirnov test for normality with mean and variance unknown*, Journal of the American Statistical Association Vol. 62, No. 318 (Jun., 1967), pp. 399-402.

[11] T. W. Anderson and D. A. Darling, *Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes*, Columbia University and University of Michigan.

[12] Cramér-von Mises test. Encyclopedia of Mathematics.
`http://www.encyclopediaofmath.org/index.php?title=Cram%C3%A9r-von_Mises_test&oldid=44377`

[13] Gilbert Colletaz, *Statistique non paramétrique*, Master 2 Économétrie et Statistique Appliquée, 7 décembre 2017.
`https://www.univ-orleans.fr/deg/masters/ESA/GC/sources/CoursNP.pdf`

[14] `https://perso.univ-rennes2.fr/system/files/users/fromont_m/PolyTests.pdf`

ANNEXES

Table des Annexes

A.1 Coefficients a_i du test de Shapiro-Wilk

$i \backslash n$	2	3	4	5	6	7	8	9	10
1	0·7071	0·7071	0·6872	0·6646	0·6431	0·6233	0·6052	0·5888	0·5739
2	—	·0000	·1677	·2413	·2806	·3031	·3164	·3244	·3291
3	—	—	—	·0000	·0875	·1401	·1743	·1976	·2141
4	—	—	—	—	—	·0000	·0561	·0947	·1224
5	—	—	—	—	—	—	—	·0000	·0399

$i \backslash n$	11	12	13	14	15	16	17	18	19	20
1	0·5601	0·5475	0·5359	0·5251	0·5150	0·5056	0·4968	0·4886	0·4808	0·4734
2	·3315	·3325	·3325	·3318	·3306	·3290	·3273	·3253	·3232	·3211
3	·2260	·2347	·2412	·2460	·2495	·2521	·2540	·2553	·2561	·2565
4	·1429	·1586	·1707	·1802	·1878	·1939	·1988	·2027	·2059	·2085
5	·0695	·0922	·1099	·1240	·1353	·1447	·1524	·1587	·1641	·1686
6	0·0000	0·0303	0·0539	0·0727	0·0880	0·1005	0·1109	0·1197	0·1271	0·1334
7	—	—	·0000	·0240	·0433	·0593	·0725	·0837	·0932	·1013
8	—	—	—	—	·0000	·0196	·0359	·0496	·0612	·0711
9	—	—	—	—	—	—	·0000	·0163	·0303	·0422
10	—	—	—	—	—	—	—	—	·0000	·0140

$i \backslash n$	21	22	23	24	25	26	27	28	29	30
1	0·4643	0·4590	0·4542	0·4493	0·4450	0·4407	0·4366	0·4328	0·4291	0·4254
2	·3185	·3156	·3126	·3098	·3069	·3043	·3018	·2992	·2968	·2944
3	·2578	·2571	·2563	·2554	·2543	·2533	·2522	·2510	·2499	·2487
4	·2119	·2131	·2139	·2145	·2148	·2151	·2152	·2151	·2150	·2148
5	·1736	·1764	·1787	·1807	·1822	·1836	·1848	·1857	·1864	·1870
6	0·1399	0·1443	0·1480	0·1512	0·1539	0·1563	0·1584	0·1601	0·1616	0·1630
7	·1092	·1150	·1201	·1245	·1283	·1316	·1346	·1372	·1395	·1415
8	·0804	·0878	·0941	·0997	·1046	·1089	·1128	·1162	·1192	·1219
9	·0530	·0618	·0696	·0764	·0823	·0876	·0923	·0965	·1002	·1036
10	·0263	·0368	·0459	·0539	·0610	·0672	·0728	·0778	·0822	·0862
11	0·0000	0·0122	0·0228	0·0321	0·0403	0·0476	0·0540	0·0598	0·0650	0·0697
12	—	—	·0000	·0107	·0200	·0284	·0358	·0424	·0483	·0537
13	—	—	—	—	·0000	·0094	·0178	·0253	·0320	·0381
14	—	—	—	—	—	—	·0000	·0084	·0159	·0227
15	—	—	—	—	—	—	—	—	·0000	·0076

$i \backslash n$	31	32	33	34	35	36	37	38	39	40
1	0·4220	0·4188	0·4156	0·4127	0·4096	0·4068	0·4040	0·4015	0·3989	0·3964
2	·2921	·2898	·2876	·2854	·2834	·2813	·2794	·2774	·2755	·2737
3	·2475	·2463	·2451	·2439	·2427	·2415	·2403	·2391	·2380	·2368
4	·2145	·2141	·2137	·2132	·2127	·2121	·2116	·2110	·2104	·2098
5	·1874	·1878	·1880	·1882	·1883	·1883	·1883	·1881	·1880	·1878
6	0·1641	0·1651	0·1660	0·1667	0·1673	0·1678	0·1683	0·1686	0·1689	0·1691
7	·1433	·1449	·1463	·1475	·1487	·1496	·1505	·1513	·1520	·1526
8	·1243	·1265	·1284	·1301	·1317	·1331	·1344	·1356	·1366	·1376
9	·1066	·1093	·1118	·1140	·1160	·1179	·1196	·1211	·1225	·1237
10	·0899	·0931	·0961	·0988	·1013	·1036	·1056	·1075	·1092	·1108
11	0·0739	0·0777	0·0812	0·0844	0·0873	0·0900	0·0924	0·0947	0·0967	0·0986
12	·0585	·0629	·0669	·0706	·0739	·0770	·0798	·0824	·0848	·0870
13	·0435	·0485	·0530	·0572	·0610	·0645	·0677	·0706	·0733	·0759
14	·0289	·0344	·0395	·0441	·0484	·0523	·0559	·0592	·0622	·0651
15	·0144	·0206	·0262	·0314	·0361	·0404	·0444	·0481	·0515	·0546
16	0·0000	0·0068	0·0131	0·0187	0·0239	0·0287	0·0331	0·0372	0·0409	0·0444
17	—	—	·0000	·0062	·0119	·0172	·0220	·0264	·0305	·0343
18	—	—	—	—	·0000	·0057	·0110	·0158	·0203	·0244
19	—	—	—	—	—	—	·0000	·0053	·0101	·0146
20	—	—	—	—	—	—	—	—	·0000	·0049

$i \backslash n$	41	42	43	44	45	46	47	48	49	50
1	0·3940	0·3917	0·3894	0·3872	0·3850	0·3830	0·3808	0·3789	0·3770	0·3751
2	·2719	·2701	·2684	·2667	·2651	·2635	·2620	·2604	·2589	·2574
3	·2357	·2345	·2334	·2323	·2313	·2302	·2291	·2281	·2271	·2260
4	·2091	·2085	·2078	·2072	·2065	·2058	·2052	·2045	·2038	·2032
5	·1876	·1874	·1871	·1868	·1865	·1862	·1859	·1855	·1851	·1847
6	0·1693	0·1694	0·1695	0·1695	0·1695	0·1695	0·1695	0·1693	0·1692	0·1691
7	·1531	·1535	·1539	·1542	·1545	·1548	·1550	·1551	·1553	·1554
8	·1384	·1392	·1398	·1405	·1410	·1415	·1420	·1423	·1427	·1430
9	·1249	·1259	·1269	·1278	·1286	·1293	·1300	·1306	·1312	·1317
10	·1123	·1136	·1149	·1160	·1170	·1180	·1189	·1197	·1205	·1212
11	0·1004	0·1020	0·1035	0·1049	0·1062	0·1073	0·1085	0·1095	0·1105	0·1113
12	·0891	·0909	·0927	·0943	·0959	·0972	·0986	·0998	·1010	·1020
13	·0782	·0804	·0824	·0842	·0860	·0876	·0892	·0906	·0919	·0932
14	·0677	·0701	·0724	·0745	·0765	·0783	·0801	·0817	·0832	·0846
15	·0575	·0602	·0628	·0651	·0673	·0694	·0713	·0731	·0748	·0764
16	0·0476	0·0506	0·0534	0·0560	0·0584	0·0607	0·0628	0·0648	0·0667	0·0685
17	·0379	·0411	·0442	·0471	·0497	·0522	·0546	·0568	·0588	·0608
18	·0283	·0318	·0352	·0383	·0412	·0439	·0465	·0489	·0511	·0532
19	·0188	·0227	·0263	·0296	·0328	·0357	·0385	·0411	·0436	·0459
20	·0094	·0136	·0175	·0211	·0245	·0277	·0307	·0335	·0361	·0386
21	0·0000	0·0045	0·0087	0·0126	0·0163	0·0197	0·0229	0·0259	0·0288	0·0314
22		—	·0000	·0042	·0081	·0118	·0153	·0185	·0215	·0244
23	—	—	—	—	·0000	·0039	·0076	·0111	·0143	·0174
24	—	—	—	—	—	—	·0000	·0037	·0071	·0104
25	—	—	—	—	—	—	—	—	·0000	·0035

FIGURE 39 – Table des a_i

A.2 Table des valeurs critiques du test de Shapiro Wilk

	Level								
n	**0·01**	**0·02**	**0·05**	**0·10**	**0·50**	**0·90**	**0·95**	**0·98**	**0·99**
3	0·753	0·756	0·767	0·789	0·959	0·998	0·999	1·000	1·000
4	·687	·707	·748	·792	·935	·987	·992	·996	·997
5	·686	·715	·762	·806	·927	·979	·986	·991	·993
6	0·713	0·743	0·788	0·826	0·927	0·974	0·981	0·986	0·989
7	·730	·760	·803	·838	·928	·972	·979	·985	·988
8	·749	·778	·818	·851	·932	·972	·978	·984	·987
9	·764	·791	·829	·859	·935	·972	·978	·984	·986
10	·781	·806	·842	·869	·938	·972	·978	·983	·986
11	0·792	0·817	0·850	0·876	0·940	0·973	0·979	0·984	0·986
12	·805	·828	·859	·883	·943	·973	·979	·984	·986
13	·814	·837	·866	·889	·945	·974	·979	·984	·986
14	·825	·846	·874	·895	·947	·975	·980	·984	·986
15	·835	·855	·881	·901	·950	·975	·980	·984	·987
16	0·844	0·863	0·887	0·906	0·952	0·976	0·981	0·985	0·987
17	·851	·869	·892	·910	·954	·977	·981	·985	·987
18	·858	·874	·897	·914	·956	·978	·982	·986	·988
19	·863	·879	·901	·917	·957	·978	·982	·986	·988
20	·868	·884	·905	·920	·959	·979	·983	·986	·988
21	0·873	0·888	0·908	0·923	0·960	0·980	0·983	0·987	0·989
22	·878	·892	·911	·926	·961	·980	·984	·987	·989
23	·881	·895	·914	·928	·962	·981	·984	·987	·989
24	·884	·898	·916	·930	·963	·981	·984	·987	·989
25	·888	·901	·918	·931	·964	·981	·985	·988	·989
26	0·891	0·904	0·920	0·933	0·965	0·982	0·985	0·988	0·989
27	·894	·906	·923	·935	·965	·982	·985	·988	·990
28	·896	·908	·924	·936	·966	·982	·985	·988	·990
29	·898	·910	·926	·937	·966	·982	·985	·988	·990
30	·900	·912	·927	·939	·967	·983	·985	·988	·900
31	0·902	0·914	0·929	0·940	0·967	0·983	0·986	0·988	0·990
32	·904	·915	·930	·941	·968	·983	·986	·988	·990
33	·906	·917	·931	·942	·968	·983	·986	·989	·990
34	·908	·919	·933	·943	·969	·983	·986	·989	·990
35	·910	·920	·934	·944	·969	·984	·986	·989	·990
36	0·912	0·922	0·935	0·945	0·970	0·984	0·986	0·989	0·990
37	·914	·924	·936	·946	·970	·984	·987	·989	·990
38	·916	·925	·938	·947	·971	·984	·987	·989	·990
39	·917	·927	·939	·948	·971	·984	·987	·989	·991
40	·919	·928	·940	·949	·972	·985	·987	·989	·991
41	0·920	0·929	0·941	0·950	0·972	0·985	0·987	0·989	0·991
42	·922	·930	·942	·951	·972	·985	·987	·989	·991
43	·923	·932	·943	·951	·973	·985	·987	·990	·991
44	·924	·933	·944	·952	·973	·985	·987	·990	·991
45	·926	·934	·945	·953	·973	·985	·988	·990	·991
46	0·927	0·935	0·945	0·953	0·974	0·985	0·988	0·990	0·991
47	·928	·936	·946	·954	·974	·985	·988	·990	·991
48	·929	·937	·947	·954	·974	·985	·988	·990	·991
49	·929	·937	·947	·955	·974	·985	·988	·990	·991
50	·930	·938	·947	·955	·974	·985	·988	·990	·991

FIGURE 40 – Table des valeurs critiques du test de Shapiro-Wilk

A.3 Code R : Kolmogorov-Smirnov : $\sqrt{n}D_n$ est pivotale

```
#------ KS sous H0 est pivotale ------

N=100000
n1=10
n2=20
n3=50

KS1=rep(0,N)
KS2=rep(0,N)
KS3=rep(0,N)

#Generation de VA N(0,1)
NORMALE1=replicate(N,rnorm(n=n1,mean=0,sd =1))
NORMALE2=replicate(N,rnorm(n=n2,mean=0,sd =1))
NORMALE3=replicate(N,rnorm(n=n3,mean=0,sd =1))

#Generation de VA Exp(3)
EXP1=replicate(N,rexp(n=n1,3))
EXP2=replicate(N,rexp(n=n2,3))
EXP3=replicate(N,rexp(n=n3,3))


#Calcule N fois sqrt(n)*sup(Fn-F0) quand F0 est la FdR de N(0,1)
for (i in 1:N){
  FDR_emp1=function(t){return(sum(NORMALE1[1:n1,i]<t)/n1)}
  KS1[i]=max(abs(Vectorize(FDR_emp1)(NORMALE1[1:n1,i])-pnorm(
     NORMALE1[1:n1,i],mean = 0, sd = 1))*sqrt(n1))

  FDR_emp2=function(t){return(sum(NORMALE2[1:n2,i]<t)/n2)}
  KS2[i]=max(abs(Vectorize(FDR_emp2)(NORMALE2[1:n2,i])-pnorm(
     NORMALE2[1:n2,i],mean = 0, sd = 1))*sqrt(n2))

  FDR_emp3=function(t){return(sum(NORMALE3[1:n3,i]<t)/n3)}
  KS3[i]=max(abs(Vectorize(FDR_emp3)(NORMALE3[1:n3,i])-pnorm(
     NORMALE3[1:n3,i],mean = 0, sd = 1))*sqrt(n3))

}

plot(density(KS3), main='')
title(TeX('Loi␣de␣\\sqrt{n}␣D_n␣sous␣H_0␣(Loi␣N(0,1))'))

lines(density(KS2),col='red')
lines(density(KS1),col='green')


legend('topright', legend = c('n=50','n=20','n=10'), col=c('black','
   red','green'),lty=1)

#Calcule N fois sqrt(n)*sup(Fn-F0) quand F0 est la FdR de Exp(3)
for (i in 1:N){
  FDR_emp1=function(t){return(sum(EXP1[1:n1,i]<t)/n1)}
  KS1[i]=max(abs(Vectorize(FDR_emp1)(EXP1[1:n1,i])-pexp(EXP1[1:n1,i
     ],3))*sqrt(n1))

  FDR_emp2=function(t){return(sum(EXP2[1:n2,i]<t)/n2)}
  KS2[i]=max(abs(Vectorize(FDR_emp2)(EXP2[1:n2,i])-pexp(EXP2[1:n2,i
     ],3))*sqrt(n2))

  FDR_emp3=function(t){return(sum(EXP3[1:n3,i]<t)/n3)}
  KS3[i]=max(abs(Vectorize(FDR_emp3)(EXP3[1:n3,i])-pexp(EXP3[1:n3,i
```

```
    ],3))*sqrt(n3))
}

plot(density(KS3), main= '' )
title(TeX("Loi␣de␣\\sqrt{n}␣D_n␣sous␣H_0␣(Loi␣Exp(3))"))

lines(density(KS2),col='red')
lines(density(KS1),col='green')


legend('topright', legend = c('n=50','n=20','n=10'), col=c('black','
  red','green'),lty=1)
```

A.4 Les valeurs critiques du test de Kolmogorov-Smirnov

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35831	.40962	.45427	.50654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897
36	.17418	.19910	.22119	.24732	.26532
37	.17188	.19646	.21826	.24404	.26180
38	.16966	.19392	.21544	.24089	.25843
39	.16753	.19148	.21273	.23786	.25518
40	.16547	.18913	.21012	.23494	.25205
41	.16349	.18687	.20760	.23213	.24904
42	.16158	.18468	.20517	.22941	.24613
43	.15974	.18257	.20283	.22679	.24332
44	.15796	.18053	.20056	.22426	.24060
45	.15623	.17856	.19837	.22181	.23798
46	.15457	.17665	.19625	.21944	.23544
47	.15295	.17481	.19420	.21715	.23298
48	.15139	.17302	.19221	.21493	.23059
49	.14987	.17128	.19028	.21277	.22828
50	.14840	.16959	.18841	.21068	.22604
51	.14697	.16796	.18659	.20864	.22386
52	.14558	.16637	.18482	.20667	.22174
53	.14423	.16483	.18311	.20475	.21968
54	.14292	.16332	.18144	.20289	.21768
55	.14164	.16186	.17981	.20107	.21574
56	.14040	.16044	.17823	.19930	.21384
57	.13919	.15906	.17669	.19758	.21199
58	.13801	.15771	.17519	.19590	.21019
59	.13686	.15639	.17373	.19427	.20844
60	.13573	.15511	.17231	.19267	.20673
61	.13464	.15385	.17091	.19112	.20506
62	.13357	.15263	.16956	.18960	.20343
63	.13253	.15144	.16823	.18812	.20184
64	.13151	.15027	.16693	.18667	.20029
65	.13052	.14913	.16567	.18525	.19877
66	.12954	.14802	.16443	.18387	.19729
67	.12859	.14693	.16322	.18252	.19584
68	.12766	.14587	.16204	.18119	.19442
69	.12675	.14483	.16088	.17990	.19303
70	.12586	.14381	.15975	.17863	.19167
71	.12499	.14281	.15864	.17739	.19034
72	.12413	.14183	.15755	.17618	.18903
73	.12329	.14087	.15649	.17498	.18776
74	.12247	.13993	.15544	.17382	.18650
75	.12167	.13901	.15442	.17268	.18528
76	.12088	.13811	.15342	.17155	.18408
77	.12011	.13723	.15244	.17045	.18290
78	.11935	.13636	.15147	.16938	.18174
79	.11860	.13551	.15052	.16832	.18060
80	.11787	.13467	.14960	.16728	.17949
81	.11716	.13385	.14868	.16626	.17840
82	.11645	.13305	.14779	.16526	.17732
83	.11576	.13226	.14691	.16428	.17627
84	.11508	.13148	.14605	.16331	.17523
85	.11442	.13072	.14520	.16236	.17421
86	.11376	.12997	.14437	.16143	.17321
87	.11311	.12923	.14355	.16051	.17223
88	.11248	.12850	.14274	.15961	.17126
89	.11186	.12779	.14195	.15873	.17031
90	.11125	.12709	.14117	.15786	.16938
91	.11064	.12640	.14040	.15700	.16846
92	.11005	.12572	.13965	.15616	.16755
93	.10947	.12506	.13891	.15533	.16666
94	.10889	.12440	.13818	.15451	.16579
95	.10833	.12375	.13746	.15371	.16493
96	.10777	.12312	.13675	.15291	.16408
97	.10722	.12249	.13606	.15214	.16324
98	.10668	.12187	.13537	.15137	.16242
99	.10615	.12126	.13469	.15061	.16161
100	.10563	.12067	.13403	.14987	.16081
$n > 100$	$1.073/\sqrt{n}$	$1.223/\sqrt{n}$	$1.358/\sqrt{n}$	$1.518/\sqrt{n}$	$1.629/\sqrt{n}$

FIGURE 41 – Table des c_α pour le test de Kolmogorov-Smirnov

A.5 Code R : Initialisation de n et des statistiques de test

```r
library("nortest")
library("fBasics")

alphalist=c(0.05,0.1)
#vecteur contenant les valeurs de alpha

nlist=c(seq(20,200,5),seq(300,500,100),seq(1000,2000,500))
#vecteur contenant les 43 valeurs de n , tailles des samples

#Nos 5 tests de normalite
#Puisque les tests retournent une liste, on cree ces fonctions
   permettant de garder uniquement la valeur de la statistique de
   test:

shapiro.statistic=function(X){
  return(shapiro.test(X)$statistic)}

lillie.statistic=function(X){
  return(lillie.test(X)$statistic)}

ks.statistic=function(X){
  return(ks.test(X,'pnorm',mean(X),sd(X))$statistic)}

cvm.statistic=function(X){
  return(cvm.test(X)$statistic)}

ad.statistic=function(X){
  return(ad.test(X)$statistic)}
```

A.6 Code R : Génération par Monte-Carlo des valeurs critiques, exemple du test de Lilliefors

```
# Le code R suivant permet de calculer les valeurs critiques d'un
   test pour differents echantillons n. Ici il est execute avec le
   test Lilliefors, mais reste valable pour les autres tests.

RC=rep(0,length(nlist))
names(RC)=nlist

RC_LL=RC


N=50000
c=1
a=0.05

for (n in nlist){
  print(c) #permet de suivre l'avancee de la boucle en temps reel
  Z=replicate(N, rnorm(n=n,mean=0,sd=1)) #chaque colonne de Z
     correspond a 1 echantillon de taille n. Z est de taille (n,N)

  #Vecteur contenant chacun N VA simulees selon la loi sous H0 de l'
     estimateur

  LL=apply(Z,2,lillie.statistic) #Choix du test


  ###### STATISTIQUE D'ORDES ######

  LL_ordre=sort(LL, decreasing = FALSE)


  ###### QUANTILE EMPIRIQUE ######


  #right-tailed tests

  q_LL=LL_ordre[ceiling(N*(1-a))]

  ### Valeurs critiques pour n ###
  RC_LL[c]=q_LL

  c=c+1
}
```

A.7 Code R : Évolution de la puissance des tests en fonction du coefficient d'asymétrie et du coefficient d'aplatissement de la GLD

```
########### DEVIER DE LA LOI NORMALE EN UTILISANT GLD #############

library('gld')
library('nortest')

# Simulation de la GLD
Qgld=function(n,l1,l2,l3,l4){
  U=runif(n,0,1)
  return(l1+(1/l2)*((U**l3-1)/l3-((1-U)**l4-1)/l4))
}

#pour une loi normale: skewness=0  et kurtosis =3

skewness=0 #faire varier le skewness autour de 0 pour tracer les
   courbes des puissances
kurtosis=3 #faire varier le kurtosis autour de 3 pour tracer les
   courbes des puissances

# Approximation des parametres
n=100
norm.approx <- fit.fkml.moments.val(c(0,1,skewness,kurtosis))
l2=norm.approx$lambda[2]
l3=norm.approx$lambda[3]
l4=norm.approx$lambda[4]
GLD=Qgld(199,0,l2,l3,l4)



######### Approximation de la loi N(0,1) avec la GLD ##########
plot(density(GLD), xlim=c(-4,4),ylim=c(0,0.45) )
curve(dnorm(x), add = TRUE, col = "red")


######### Evolution de la puissance des tests en fonction du
   coefficient d'asymetrie #########

#pour n choisi
n=50
intervalle=seq(0,3,0.1) #intervalle autour de skewness=0 (par
   sym trie on peut prendre simplement l'intervalle [0,3])

#initialisation des variables
P=rep(0,length(intervalle))
P_SW_GLD=P
P_LL_GLD=P
P_CVM_GLD=P
P_AD_GLD=P
P_KS_GLD=P
N2=1000
c=1


#boucle qui calcule la puissance pour chaque valeur de skewness
   entre 0 et 3
for (i in intervalle){
  print(c)

  #Approximation des parametres
  approx <- fit.fkml.moments.val(c(0,1,i,kurtosis)) #kurtosis fixe
```

```
        3
  l2=approx$lambda[2]
  l3=approx$lambda[3]
  l4=approx$lambda[4]

  #Calcul des stats
  GLD=replicate(N2,Qgld(n,0,l2,l3,l4))
  SW_GLD=apply(GLD,2,shapiro.statistic)
  CVM_GLD=apply(GLD,2,cvm.statistic)
  LL_GLD=apply(GLD,2,lillie.statistic)
  AD_GLD=apply(GLD,2,ad.statistic)
  KS_GLD=apply(GLD,2,ks.statistic)

  #Calcul de la puissance
  ##left-tailed tests
  P_SW_GLD[c]=sum(SW_GLD<RC_SW[7])/N2
  #right-tailed tests
  P_LL_GLD[c]=sum(LL_GLD>RC_LL[7])/N2
  P_CVM_GLD[c]=sum(CVM_GLD>RC_CVM[7])/N2
  P_AD_GLD[c]=sum(AD_GLD>RC_AD[7])/N2
  P_KS_GLD[c]=sum(KS_GLD>0.188)/N2
  c=c+1
}


# Graphique puissance
plot(intervalle,P_SW_GLD, type="o", col="blue", pch=15, lty=3 , xlab
   ="skewness", ylab=expression(1-beta),ylim=c(-0.01,1.1))
points(intervalle,P_LL_GLD, col="brown",pch=16)
lines(intervalle,P_LL_GLD,col="brown", lty=3)
points(intervalle,P_CVM_GLD, col="orange",pch=17)
lines(intervalle,P_CVM_GLD,col="orange", lty=3)
points(intervalle,P_AD_GLD, col="black",pch=18)
lines(intervalle,P_AD_GLD,col="black", lty=3)
points(intervalle,P_KS_GLD, col="red",pch=3)
lines(intervalle,P_KS_GLD,col="red", lty=3)
grid(lwd = 2)
legend(x="topleft",legend=c("SW","LL","CVM","AD","KS"), col=c("blue"
   ,"brown","orange", "black",'red'),
       pch=c(15:18,3), ncol=1, lty=3)
title("Evolution de la puissance des tests en fonction du 
   coefficient d'asymtrie de GLD (n=50)")



################ Evolution de la puissance des tests en fonction du
   coefficient d'aplatissement pour n choisi ################
n=50
intervalle=c(seq(0,4,0.1),seq(4.3,6,0.1)) #Intervalle au tout de
   kurtosis = 3
P=rep(0,length(intervalle))
P_SW_GLD=P
P_LL_GLD=P
P_CVM_GLD=P
P_AD_GLD=P
P_KS_GLD=P
N2=1000
c=1


#boucle qui calcule la puissance pour chaque valeur de kurtosis
   entre 0 et 6
for (i in intervalle){
```

```
  print(c)

  #Approximation des parametres
  approx <- fit.fkml.moments.val(c(0,1,skewness,i)) #pour skewness
    fixe    0
  l2=approx$lambda[2]
  l3=approx$lambda[3]
  l4=approx$lambda[4]

  #Calcul des stats
  GLD=replicate(N2,Qgld(n,0,l2,l3,l4))
  SW_GLD=apply(GLD,2,shapiro.statistic)
  CVM_GLD=apply(GLD,2,cvm.statistic)
  LL_GLD=apply(GLD,2,lillie.statistic)
  AD_GLD=apply(GLD,2,ad.statistic)
  KS_GLD=apply(GLD,2,ks.statistic)

  #Calcul de la puissance
  ##left-tailed tests
  P_SW_GLD[c]=sum(SW_GLD<RC_SW[7])/N2
  #right-tailed tests
  P_LL_GLD[c]=sum(LL_GLD>RC_LL[7])/N2
  P_CVM_GLD[c]=sum(CVM_GLD>RC_CVM[7])/N2
  P_AD_GLD[c]=sum(AD_GLD>RC_AD[7])/N2
  P_KS_GLD[c]=sum(KS_GLD>0.188)/N2
  c=c+1
}


# Graphique puissance
plot(intervalle,P_SW_GLD, type="o", col="blue", pch=15, lty=3 , xlab
   ="kurtosis", ylab=expression(1-beta),ylim=c(-0.01,1.1))
points(intervalle,P_LL_GLD, col="brown",pch=16)
lines(intervalle,P_LL_GLD,col="brown", lty=3)
points(intervalle,P_CVM_GLD, col="orange",pch=17)
lines(intervalle,P_CVM_GLD,col="orange", lty=3)
points(intervalle,P_AD_GLD, col="black",pch=18)
lines(intervalle,P_AD_GLD,col="black", lty=3)
points(intervalle,P_KS_GLD, col="red",pch=3)
lines(intervalle,P_KS_GLD,col="red", lty=3)
grid(lwd = 2)
legend(x="topleft",legend=c("SW","LL","CVM","AD","KS"), col=c("blue"
   ,"brown","orange", "black",'red'),
       pch=c(15:18,3), ncol=1, lty=3)
title("Evolution␣de␣la␣puissance␣des␣tests␣en␣fonction␣du␣
   coefficient␣d'aplatissement␣de␣GLD␣(n=50)")
```

A.8 Code R : Puissance des tests en fonction de n, exemple de la loi Uniforme

```
# Le code R suivant permet de calculer la puissance des tests pour
   differents echantillons n. Ici il est execute avec une loi
   Uniforme, mais on peut tres bien l'utiliser pour d'autres lois.

################## Valeurs critiques tabulees: ##################

#Shapiro-Wilk
rc_sw=c(0.905,0.918,0.927,0.934,0.94,0.945,0.947) #les 50 premieres
   valeurs de shapiro sont tabul es
RC_SW=c(rc_sw,RC_SW[8:43]) # avec RC_SW valeurs critiques empiriques

#Kolmogorov-Smirnov
rc_ks = c(0.29408, 0.26404, 0.24170, 0.22425, 0.21012, 0.19837,
   0.18841, 0.17981, 0.17231, 0.16567, 0.15975, 0.15442, 0.14960,
   0.14520, 0.14117, 0.13746, 0.13403, 1.358/sqrt(105), 1.358/sqrt
   (110), 1.358/sqrt(115), 1.358/sqrt(120), 1.358/sqrt(125), 1.358/
   sqrt(130), 1.358/sqrt(135), 1.358/sqrt(140), 1.358/sqrt(145),
   1.358/sqrt(150), 1.358/sqrt(155), 1.358/sqrt(160), 1.358/sqrt
   (165), 1.358/sqrt(170), 1.358/sqrt(175), 1.358/sqrt(180), 1.358/
   sqrt(185), 1.358/sqrt(190), 1.358/sqrt(195), 1.358/sqrt(200),
   1.358/sqrt(300), 1.358/sqrt(400), 1.358/sqrt(500), 1.358/sqrt
   (1000), 1.358/sqrt(1500), 1.358/sqrt(2000))


############# Puissance des tests : Loi Uniforme(0,1) #############
#Initialisation

P = rep(0,length(nlist))
names(P) = nlist

P_SW_Unif = P
P_LL_Unif = P
P_CVM_Unif = P
P_AD_Unif = P
P_KS_Unif = P

c = 1
N2 = 10000

for (n in nlist) {
  print(c)

  #Generation de la loi
  Unif = replicate(N2, runif(n,0,1)) #choix de la loi alternative

  #Calcul de la statistique de test
  SW_Unif = apply(Unif,2,shapiro.statistic)
  CVM_Unif = apply(Unif,2,cvm.statistic)
  LL_Unif = apply(Unif,2,lillie.statistic)
  AD_Unif = apply(Unif,2,ad.statistic)
  KS_Unif = apply(Unif,2,ks.statistic)

  #Calcul de la puissance
  ##left-tailed tests
  P_SW_Unif[c] = sum(SW_Unif<RC_SW[c])/N2

  ##right-tailed tests
  P_LL_Unif[c] = sum(LL_Unif>RC_LL[c])/N2
  P_CVM_Unif[c] = sum(CVM_Unif>RC_CVM[c])/N2
  P_AD_Unif[c] = sum(AD_Unif>RC_AD[c])/N2
```

```
  P_KS_Unif[c] = sum(KS_Unif>rc_ks[c])/N2

  c = c + 1

}

# GRAPHIQUE PUISSANCE
par(mfrow = c(1,1))

plot(nlist, P_SW_Unif, log = "x", type = "o", col = "blue", pch =
   15, lty = 3, xlab = "n", ylab = expression(1-beta), ylim = c
   (-0.1, 1.1))

points(nlist,P_LL_Unif, col="brown",pch=16)
lines(nlist,P_LL_Unif ,col="brown", lty=3)

points(nlist,P_CVM_Unif, col="orange",pch=17)
lines(nlist,P_CVM_Unif ,col="orange", lty=3)

points(nlist,P_AD_Unif, col="black",pch=18)
lines(nlist,P_AD_Unif ,col="black", lty=3)

points(nlist,P_KS_Unif, col="dark␣green",pch=3)
lines(nlist,P_KS_Unif ,col="dark␣green", lty=3)

grid(lwd = 2)
legend(x = "bottomright", legend = c("SW","LL","CVM","AD","KS"), col
    = c("blue","brown","orange", "black","dark␣green"),
       pch=c(15:18,3), ncol=1, lty=3)
title(TeX("Courbe␣puissance␣de␣la␣loi␣\\mathcal{U}(0,1)"))
```