République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

GUESMIA Nour El Houda

Titre :

Tests d'ajustement à une distribution basés sur la fonction de répartition empirique

Membres du Comité d'Examen :

NECIR Abdelhakim	Pr.	UMKB	Président
MERAGHNI Djamel	Pr.	UMKB	Encadreur
ROUBI Afef	M.A.A.	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce mémoire à :

Mes chers parents

Ma mère, qui a œuvré pour ma réussite, de par son soutien et ses précieux conseils, pour toute son assistance et sa présence dans ma vie.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie.

Mes sœurs, qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

Mes professeurs, qui doivent voir dans ce travail la fierté d'un savoir bien acquis.

REMERCIEMENT

Louange à Allah, Seigneur de l'univers, avant tout.

Mes plus vifs remerciements sont adressés à mon encadreur Pr MERAGHNI Djamel pour ces conseils et ses orientations, ses disponibilités, ses précieux conseils, son sens de l'écoute et pour toute l'aide qu'ils m'ont apporté qui m'ont été d'une grande utilité au cours de l'élaboration de mon mémoire.

Je remercie tous les enseignants qui ont contribué à ma formation et surtout à Pr Necir Abdelhakim et Monsieur Dr Benatia Fatah.

Je remercie tout particulièrement mes parents pour leurs encouragements et soutien sur tous les aspects et aussi toute ma famille. Je n'oubliais pas l'ensemble de mes amis proches.

Enfin, je n'oublie pas mes collègues et amis ; grâce à qui ma vie universitaire a été très agréable et joyeuse. Merci à tous ceux que j'ai oubliés.

Table des matières

Table des figures

Liste des tableaux

Introduction

Grâce aux méthodes de statistique non paramétrique, il est tout à fait possible d'extraire des informations pertinentes d'un échantillon sans connaitre la loi de probabilité dont il est issu. Cependant, si c'est possible, il est préférable d'adopter un modèle probabiliste. En effet, les estimations seront toujours plus précises dans un cadre paramétrique que dans un cadre non paramétrique. Par ailleurs, un grand nombre de procédures statistiques standard ne sont utilisables que si on fait des hypothèses particulières sur la loi de probabilité des observations (par exemple, les tests dans les modèles linéaires gaussiens).

Par conséquent, il est fondamental de disposer de méthodes permettant de déterminer s'il est vraisemblable de considérer que des observations proviennent d'un modèle probabiliste donné. Ces méthodes sont appelées les tests d'ajustement ou d'adéquation. Elles sont destinées à vérifier si un échantillon observé peut être considéré comme extrait d'une population donnée. Il existe plusieurs types de ces procédures : les graphes de probabilité, qui sont des tests d'adéquation graphiques, le test du khi-deux, le test de Kolmogorov-Smirnov, etc... Pour une description détaillée sur ces tests, on réfère le lecteur à [6], [3], [8],...

Dans le cadre de ce mémoire, on va étudier les tests d'ajustement qui sont basés sur la fonction de répartition empirique.

Ce mémoire compose de deux chapitres

– **Premier chapitre** : Fonction de répartition empirique.

Ce chapitre est dédié à la définition de la fonction de répartition empirique et à ses propriétés fondamentales.

– **Deuxième chapitre** : Tests d'ajustement.

Dans ce chapitre, après un bref rappel sur les différents tests d'ajustement, on présente, avec plus de détails, les trois principales procédures utilisant la fonction de répartition empirique, à savoir les tests de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling, puis on va donner des exemples d'applications de ces tests sur des données simulées et réelles, à l'aide du logiciel d'analyse statistique **R** version 3.5.0.

Chapitre 1

Fonction de répartition empirique

En statistique, une fonction de répartition empirique est une fonction de répartition qui attribue la probabilité $1/n$ à chacun des n nombres dans un échantillon, c'est l'estimateur non paramétrique de la fonction de répartition.

On va étudier, avec plus de détails, les propriétés fondamentales de cet estimateur, pour une description détaillée sur cette fonction, consulter, par exemple, [9], [10], [11],...

1.1 Fonction de répartition

La fonction de répartition est utilisée pour définir de façon unifiée la loi de probabilité d'une variable aléatoire (v.a) qu'elle soit discrète ou continue. Si cette fonction est connue, il est possible de calculer la probabilité de tout intervalle et donc, en pratique, de tout événement.

Soit X une v.a définie sur un espace de probabilités $(\Omega, \mathcal{T}, P)$.

Définition 1.1.1 *On appelle fonction de répartition de X, que l'on note F, la fonction définie de $\mathbb{R}$ dans $[0,1]$ par*

$$F(x) := P(X \leq x), \ x \in \mathbb{R}.$$

La valeur prise par la fonction de répartition au point x est donc la probabilité de l'événement $]-\infty, x]$.

Propriété 1.1.1 *Les propriétés principales de la fonction de répartition sont les suivantes :*

1. F est non décroissante.
2. F continue à droite en tout point x de $\mathbb{R}$.
3. $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$.

Remarque 1.1.1 *Les résultats suivants sont des conséquences directes de la définition.*

1. *Si X est discrète de valeurs $x_1, x_2, ..., x_n$, alors :*

$$F(x) = \sum_{x_k \leq x} P(X = x_k).$$

C'est une fonction en escaliers, son graphe est un diagramme en escaliers comme le montre la figure (1.1) (panneau de droite).

2. *Si X est continue de densité f, alors*

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

dans ce cas, la probabilité de tout intervalle réel de bornes a et b, avec $a \leq b$ est égale à $F(b) - F(a)$.

Exemple 1.1.1 *La figure (1.1) présente les graphes des fonctions de répartition de la loi normale centrée réduite $\mathcal{N}(0, 1)$ et de la loi de Poisson de paramètre* 5.

L'inverse de la fonction de répartition est appelée fonction des quantiles. Elle est donnée dans la définition 1.1.2.

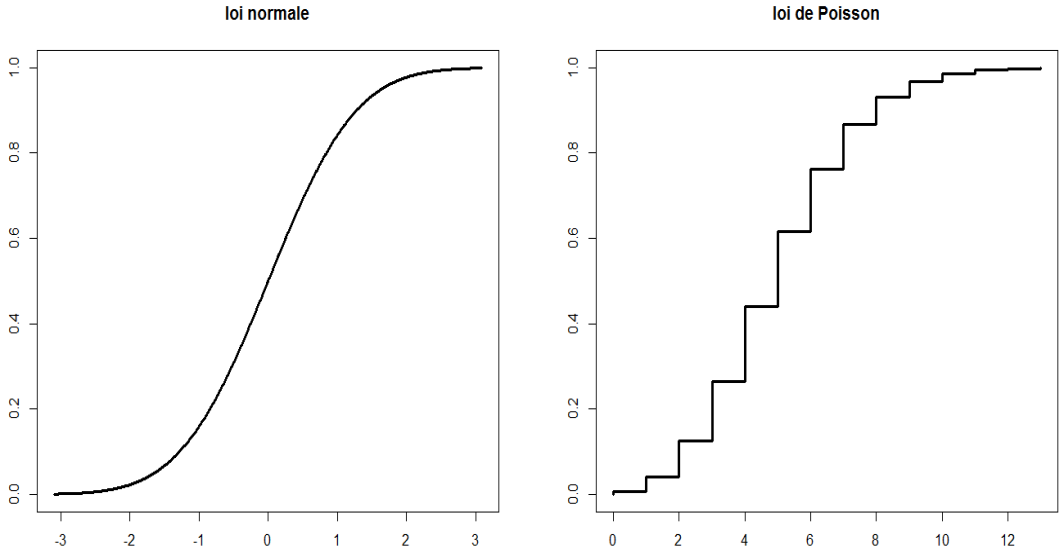

FIG. 1.1 – Fonction de répartition d'une v.a normale standard et d'une v.a de Poisson de paramètre 5.

Définition 1.1.2 (fonction des quantiles) *Soit X une v.a et F sa fonction de répartition, la fonction des quantiles, notée par $F^{\leftarrow}$, associée à F est définie par*

$$F^{\leftarrow}(q) := \inf\{x \in \mathbb{R} : F(x) \geq q\},\ 0 \leq q \leq 1.$$

On l'appelle aussi inverse généralisée de F. Il est à noter que si F est strictement croissante et continue, alors elle est bijective et dans ce cas on a $F^{\leftarrow}(q) = F^{-1}(q)$.

Exemple 1.1.2

- $F^{\leftarrow}(0.5)$ est la médiane de X.
- $F^{\leftarrow}(0.25)$ et $F^{\leftarrow}(0.75)$ sont respectivement les premier et troisième quartiles de X.
- La fonction des quantiles de la loi exponentielle $\mathcal{E}(\lambda)$, $\lambda > 0$, est définie par

$$F^{\leftarrow}(q) = -\frac{1}{\lambda}\log(1-q),\ 0 \leq q \leq 1.$$

En effet, on a $F(x) = (1 - \exp(-\lambda x))\mathbf{1}_{\mathbb{R}^+}(x)$, bijective. Alors $F^{\leftarrow}(q) = F^{-1}(q)$.

1.2 Statistiques d'ordre

Définition 1.2.1 *Soit $(X_1, X_2, ..., X_n)$ un échantillon, de taille $n \geq 1$, d'une v.a X. On appelle statistiques d'ordre associées à cet échantillon la suite classée par ordres croissants des v.a's $X_1, X_2, ..., X_n$. On les note généralement par $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ ou $X_{1,n}, X_{2,n}, ..., X_{n,n}$.*

Pour $k = 1, 2, ..., n$, la v.a $X_{(k)}$ est dite statistique d'ordre de rang k ou $k^{ème}$ statistique d'ordre. Les v.a's

$$X_{(1)} := \min_{1 \leq i \leq n} X_i \text{ et } X_{(n)} := \max_{1 \leq i \leq n} X_i,$$

sont deux statistiques d'ordre particulières, elles représentent la plus petite et la plus grande observation respectivement.

1.2.1 Distribution d'une statistique d'ordre

Proposition 1.2.1 (lois de $X_{(1)}$ et $X_{(n)}$) *Les lois de probabilité des v.a's $X_{(1)}$ et $X_{(n)}$ sont données par leurs fonctions de répartition respectives*

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n \text{ et } F_{X_{(n)}}(x) = [F(x)]^n, \ x \in \mathbb{R}.$$

Preuve. On a

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x).$$

Il est clair que l'évènement $(X_{(1)} > x)$ est équivalent à $(X_1 > x, X_2 > x, ..., X_n > x)$, Par conséquent

$$F_{X_{(1)}}(x) = 1 - P(X_1 > x, X_2 > x, ..., X_n > x),$$

qui par l'indépendance des X_i devient

$$F_{X_{(1)}}(x) = 1 - P(X_1 > x)P(X_2 > x)...P(X_n > x).$$

Finalement, on obtient le résultat en utilisant l'équidistribution des observations, alors

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - [P(X_1 > x)]^n \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

Par le même principe, on montre le résultat relatif au maximum. ■

Proposition 1.2.2 (loi de $X_{(k)}$) *De façon générale, pour $1 \leq k \leq n$, la fonction de répartition de la $k^{ème}$ statistique d'ordre est la suivante*

$$F_{X_{(k)}}(x) = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} F^i(x)[1 - F(x)]^{n-i}, \ x \in \mathbb{R}.$$

Preuve. Soit $x \in \mathbb{R}$ fixé. Dire que l'événement $\left(X_{(k)} \leq x\right)$ est réalisé est équivalent à dire que parmi les variables $X_1, X_2, ..., X_n$, au moins k sont plus petites que x. En d'autres termes, on a

$$X_{(k)} \leq x \Longleftrightarrow \sum_{j=1}^{n} \mathbf{1}_{(X_j \leq x)} \geq k,$$

ce qui implique que

$$P(X_{(k)} \leq x) = P\left(\sum_{j=1}^{n} \mathbf{1}_{(X_j \leq x)} \geq k\right).$$

Par ailleurs $\sum_{j=1}^{n} \mathbf{1}_{(X_j \leq x)}$ suit la loi binomiale de paramètres n et $F(x)$, alors on a, pour $i = 1, 2, ..., n$,

$$P\left(\sum_{j=1}^{n} \mathbf{1}_{(X_j \leq x)} = i\right) = \frac{n!}{i!(n-i)!} F^i(x)[1 - F(x)]^{n-i}.$$

En prenant la somme de $i = k$ à n, on obtient le résultat. ■

1.3 Fonction de répartition empirique

1.3.1 Distribution empirique

Soit $(X_1, X_2, ..., X_n)$ un échantillon, de taille $n \geq 1$, d'une v.a X de distribution P, on définit une distribution discrète sur $\mathbb{R}$ muni de la tribu borélienne $\mathfrak{B}$ concentrée aux réalisations $(x_1, x_2, ..., x_n)$ par $P_n(x_i) = 1/n$, $i = 1, 2, ..., n$.
La probabilité empirique d'un borélien $B \in \mathfrak{B}$ est alors définie par

$$P_n(B) := \frac{\nu(B)}{n},$$

où $0 \leq \nu(B) \leq n$ représente le nombre de réalisations x_i qui sont dans B.

Proposition 1.3.1 (convergence de P_n) *Soit $B \in \mathfrak{B}$, alors*

1. $P_n(B) \xrightarrow{p} P(B)$, *quand* $n \longrightarrow \infty$.
2. $P_n(B) \xrightarrow{p.s} P(B)$, *quand* $n \longrightarrow \infty$.

Preuve. On applique les lois des grands nombres. Pour cela, on définit les v.a's indépendantes identiquement distribuées (iid)

$$Y_i := \mathbf{1}_{X_i}(B) = \begin{cases} 1 & \text{si} \quad X_i \in B, \\ 0 & \text{si} \quad X_i \notin B, \end{cases}, \ i = 1, 2, ..., n.$$

Alors $\nu(B) = \sum_{i=1}^{n} Y_i$ et $P_n(B) = \sum_{i=1}^{n} Y_i/n$, c'est la moyenne empirique $\bar{Y}_n$ des Y_i. D'autre part, on a $E[Y_1] = P(Y_1 = 1) = P(X_1 \in B) = P(B)$. Pour le premier résultat, on applique la loi faible des grands nombres et pour le deuxième, on applique la loi forte. ■

Remarque 1.3.1 *Ce résultat implique que la distribution inconnue P (de la variable X) peut être approchée, en considérant un échantillon de taille élevée. En d'autres termes, P_n est la meilleure approximation de P.*

1.3.2 Fonction de répartition empirique

La fonction de répartition empirique, qu'on notera F_n, est la probabilité empirique du borélien particulier $B =]-\infty, x]$.

Définition 1.3.1 *La fonction de répartition empirique F_n associée à un échantillon $(X_1, X_2, ..., X_n)$ est définie par*

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}, \ x \in \mathbb{R}.$$

C'est la proportion des éléments de l'échantillon qui sont inférieurs ou égaux à x. En d'autres termes, la fonction de répartition empirique est la moyenne empirique des fonctions d'indicatrices des événements $(X_i \leq x)$.

Sa représentation par les statistiques d'ordre est donnée par

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)}, \\ \frac{i}{n} & \text{si} \quad x_{(i)} \leq x < x_{(i+1)}, \ i = 1, 2, ..., n-1, \\ 1 & \text{si} \quad x \geq x_{(n)}. \end{cases} \tag{1.1}$$

F_n est une fonction en escaliers qui fait des sauts de hauteur $1/n$ en chaque point de l'échantillon.

La figure (1.2) représente le graphe de la fonction de répartition empirique d'un échantillon, de taille 20, de la loi $\mathcal{N}(0, 1)$.

Propriété 1.3.1 *Les propriétés principales de la fonction de répartition empirique sont :*

1. *F_n est croissante.*
2. *F_n continue à droite sur $\mathbb{R}$. En fait, elle est discontinue aux points $\left(x_{(i)}\right)_{1 \leq i \leq n}$ et constante sur $[x_{(i)}, x_{(i+1)}[$ pour $i = 1, 2, ..., n-1$.*
3. $\lim\limits_{x \to -\infty} F_n(x) = 0$ *et* $\lim\limits_{x \to +\infty} F_n(x) = 1$.

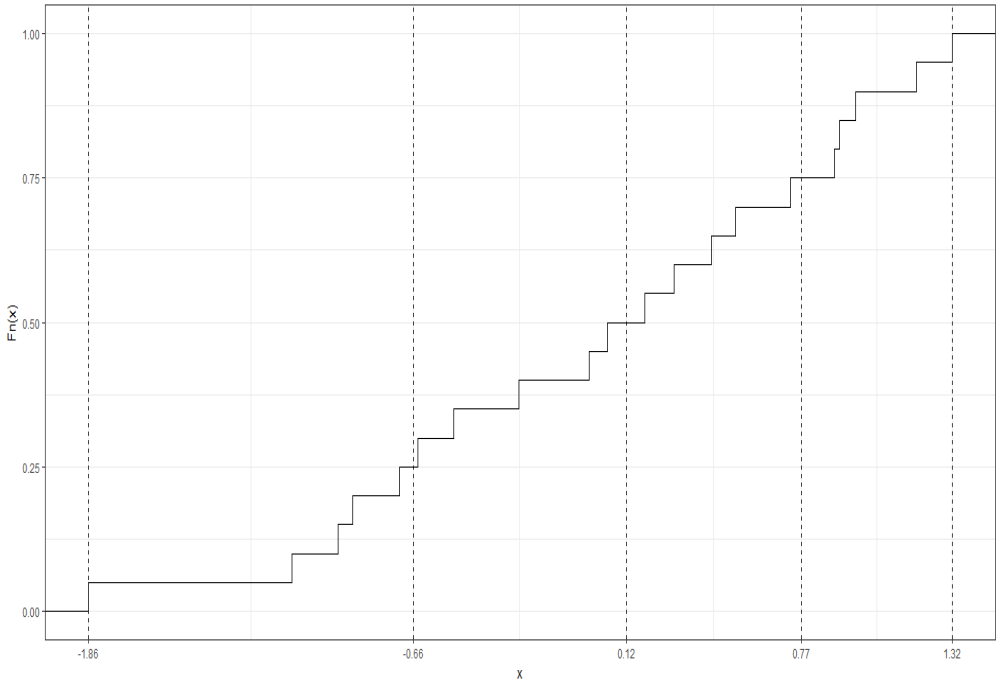

FIG. 1.2 – Fonction de répartition empirique d'un échantillon de taille 20 de loi normale standard.

Exemple 1.3.1 (fonction empirique uniforme) *Soit* $(U_n)_{n\geq 1}$ *une suite de v.a's iid uniformes sur* $[0,1]$, *la distribution empirique uniforme* G_n *associé à* $U_1, U_2, ..., U_n$ *est définie par*

$$G_n(t) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(U_i\leq t)} = \begin{cases} 0 & si \quad t < u_{(1)}, \\ \frac{i}{n} & si \quad u_{(i)} \leq t < u_{(i+1)};\ i = 1, 2, ..., n-1, \\ 1 & si \quad t \geq u_{(n)}. \end{cases} \tag{1.2}$$

L'inverse de la fonction de répartition empirique est appelée fonction des quantiles empiriques. Elle est donnée dans la définition 1.3.2.

Définition 1.3.2 *La fonction des quantiles empiriques associée à* F_n *est définie par*

$$F_n^{\leftarrow}(q) := inf\{x \in \mathbb{R} : F_n(x) \geq q\},\ 0 \leq q \leq 1.$$

Remarque 1.3.2 *L'équation (1.1) entraîne que*

$$F_n^{\leftarrow}(q) = X_{(i)},\ \textit{pour}\ \frac{i-1}{n} < q \leq \frac{i}{n},\ i = 1, 2, ..., n.$$

Alors on a pour $0 \leq q \leq 1$, $F_n^{\leftarrow}(q) = X_{([nq])}$, *où* $[nq]$ *est la partie entière de* nq.

Définition 1.3.3 (fonction des quantiles empiriques uniformes) *La fonction des quantiles uniformes associée à* $U_1, U_2, ..., U_n$ *est définie par*

$$V_n(t) := \begin{cases} 0 & si \quad t < 0, \\ G_n^{\leftarrow}(t) & si \quad 0 \leq t < 1, \\ 1 & si \quad t \geq 1. \end{cases}$$

Remarque 1.3.3 *La forme (1.2) entraîne que*

$$V_n(t) = U_{(i)},\ \textit{pour}\ \frac{i-1}{n} < t \leq \frac{i}{n},\ i = 1, 2, ..., n.$$

1.3.3 Distribution de probabilité de la v.a $F_n(x)$

Pour chaque x fixé dans $\mathbb{R}$, $F_n(x)$ est une fonction des v.a's $X_1, X_2, ..., X_n$, donc elle-même est une v.a. Les fonctions $\mathbf{1}_{(X_i \leq x)}$ sont des v.a's iid de loi de Bernoulli de paramètre $p := P\left(\mathbf{1}_{(X_i \leq x)} = 1\right) = P\left(X_i \leq x\right) = F(x)$. Par conséquent, la v.a $nF_n(x) = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}$ suit une loi binomiale de paramètres n et $F(x)$.

Paramètres de la v.a $F_n(x)$

Les moyenne et variance de la v.a $nF_n(x)$ sont $nF(x)$ et $nF(x)(1 - F(x))$ respectivement. D'où les paramètres de $F_n(x)$, pour $x \in \mathbb{R}$ fixé, sont :

Espérance-Variance

$$E[F_n(x)] = F(x) \text{ et } Var(F_n(x)) = \frac{1}{n}F(x)(1 - F(x)).$$

En effet, on a

$$E[F_n(x)] = \frac{1}{n}E[nF_n(x)] = \frac{1}{n}nF(x) = F(x),$$

et

$$Var(F_n(x)) = \frac{1}{n^2}Var\left(nF_n(x)\right) = \frac{1}{n^2}nF(x)(1-F(x)) = \frac{1}{n}F(x)(1-F(x)).$$

Biais-Erreur quadratique moyenne (ou MSE pour "mean squared error")

$$Biais(F_n(x)) = 0 \text{ et } MSE(F_n(x)) = \frac{1}{n}F(x)(1-F(x)).$$

En effet, on a

$$Biais(F_n(x)) = E[F_n(x)] - F(x) = 0,$$

et

$$\begin{aligned} MSE(F_n(x)) &= E[(F_n(x)-F(x))^2] = \left(E[F_n(x)] - F(x)\right)^2 + Var(F_n(x)) \\ &= Var(F_n(x)) = \frac{1}{n}F(x)(1-F(x)). \end{aligned}$$

Donc, pour tout $x \in \mathbb{R}$, $F_n(x)$ est un estimateur sans biais de $F(x)$.

Les résultats de convergence de la v.a $F_n(x)$ sont donnés dans la proposition 1.3.2.

Proposition 1.3.2 (convergence de $F_n(x)$)

1. *Consistance : pour tout x, on a $F_n(x) \xrightarrow{p} F(x)$, quand $n \to \infty$.*
2. *Consistance forte : pour tout x, on a $F_n(x) \xrightarrow{p.s} F(x)$, quand $n \to \infty$.*
3. *Convergence uniforme : par le théorème de Glivenko-Cantelli, la convergence de F_n vers F est presque sûrement uniforme, c'est à dire que*

$$\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid \xrightarrow{p.s} 0, \text{ quand } n \to \infty.$$

Preuve. Les deux premiers résultats, sont des cas particuliers de la proposition 1.3.1, pour $B =]-\infty, x]$. Pour le troisième résultat, se référer à [1], pages 4-6. ■

Ces trois résultats sont fondamentaux et justifient l'utilisation des échantillons en statistique.

La figure (1.3) illustre l'approximation de la fonction de répartition de $\mathcal{N}(0,1)$ par la fonction de répartition empirique correspondant à un échantillon de taille $n = 100$.

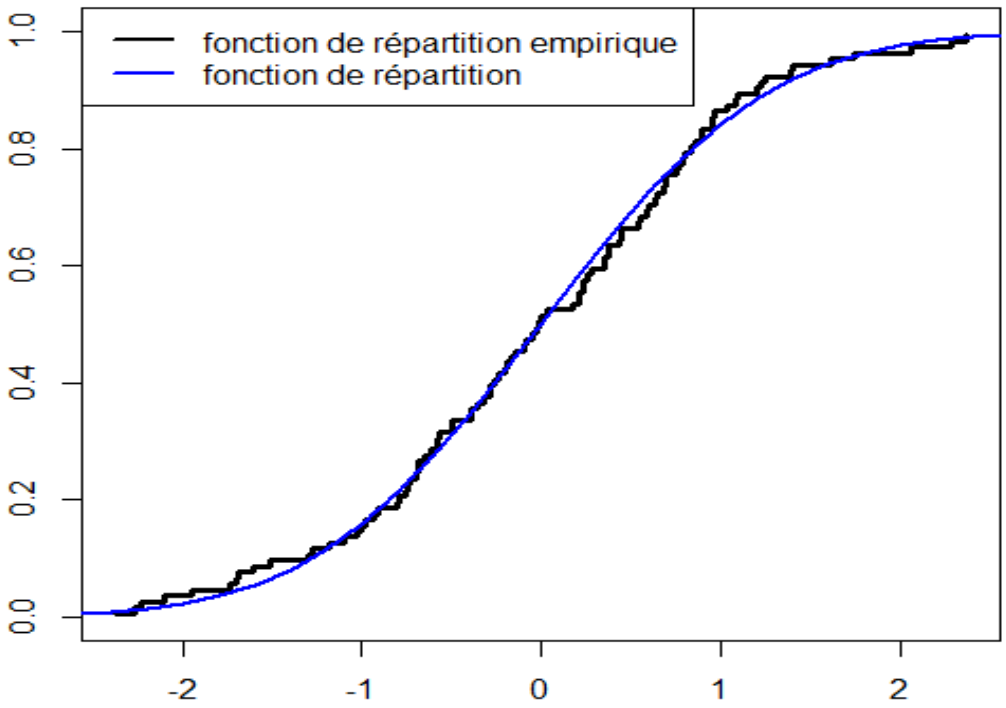

FIG. 1.3 – Approximation de la fonction de répartition de $\mathcal{N}(0,1)$ par la fonction de répartition empirique basée sur 100 observations.

Loi asymptotique de $F_n(x)$, pour $x \in \mathbb{R}$ fixé

Pour obtenir cette distribution asymptotique de $F_n(x)$, on applique le résultat suivant qui est d'une importance capitale en statistique.

Théorème 1.3.1 (théorème central limite) *Si $X_1, X_2, ..., X_n$ est une suite de v.a's iid d'espérance μ et de variance σ^2 finie, alors*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2), \text{ quand } n \to \infty,$$

où $\bar{X}_n$ désigne la moyenne empirique. Par conséquent, pour $F_n(x)$, on a

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(x)(1 - F(x))), \text{ quand } n \to \infty. \tag{1.3}$$

Le résultat (1.3) est illustré par la figure (1.4) obtenue pour des échantillons, de tailles $n = 30, ..., 200$, d'une v.a de loi uniforme sur $[0, 1]$, au point $x = 0.5$.

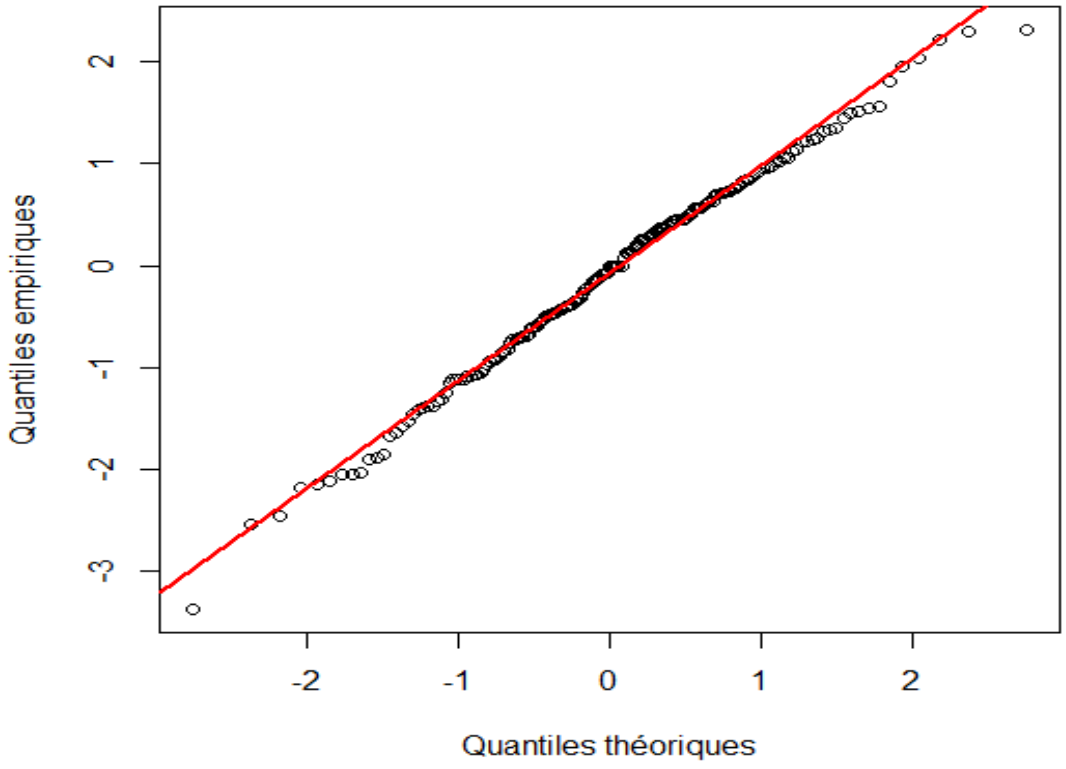

FIG. 1.4 – Illustration du théorème central limite pour la fonction de répartition empirique.

Intervalle de confiance asymptotique de $F(x)$

Le résultat (1.3) nous permet, après avoir estimé la variance par $F_n(x)(1-F_n(x))$, de construire des intervalles de confiance asymptotiques pour $F(x)$.

Corollaire 1.3.1 *Soit $0 \leq \alpha \leq 1$, l'intervalle de confiance asymptotique de $F(x)$, de niveau $1-\alpha$, est*

$$\left[F_n(x) - z_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1-F_n(x))}{n}}, F_n(x) + z_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1-F_n(x))}{n}}\right],$$

où $z_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. En d'autres termes, on a

$$P\left(F(x) \in \left[F_n(x) - z_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1-F_n(x))}{n}}, F_n(x) + z_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1-F_n(x))}{n}}\right]\right) = 1-\alpha.$$

Intervalle de confiance exact de $F(x)$

Pour déterminer un intervalle de confiance exact pour $F(x)$, on applique l'inégalité de Dvoretzky-Kiefer-Wolfowitze (voir [16], page 14).

Proposition 1.3.3 (inégalité de Dvoretzky-Kiefer-Wolfowitz)

Pour tout $\epsilon > 0$, on a

$$\forall n \in \mathbb{N},\ P\left(\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid > \epsilon\right) \leq 2\exp(-2n\epsilon^2).$$

Corollaire 1.3.2 *Soit $0 < \alpha \leq 1$, l'intervalle de confiance exact de $F(x)$, de niveau $1-\alpha$ (au moins), est*

$$\left[F_n(x) - \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}, F_n(x) + \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}\right].$$

En d'autres termes, on a

$$P\left(F(x) \in \left[F_n(x) - \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}, F_n(x) + \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}\right]\right) \geq 1-\alpha.$$

Preuve. Pour tout $x \in \mathbb{R}$, on a

$$\begin{aligned} P(F(x) \in \quad & [F_n(x) - \epsilon, F_n(x) + \epsilon]) = 1 - P(\mid F_n(x) - F(x) \mid > \epsilon) \\ \geq \quad & 1 - P\left(\sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid > \epsilon\right) \geq 1 - 2\exp(-2n\epsilon^2). \end{aligned}$$

En choisissant $\epsilon > 0$ tel que $2\exp(-2n\epsilon^2) = \alpha$, c'est à dire $\epsilon^2 = \frac{1}{2n}\log(2/\alpha)$, on obtient le résultat. ■

Remarque 1.3.4

– Comme $F(x) \in [0,1]$, si n est petit on peut souvent raffiner cet intervalle de confiance en prenant plutôt

$$\left[F_n(x) - \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}, F_n(x) + \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}\right] \cap [0,1].$$

– Bien qu'asymptotique le premier intervalle de confiance peut s'avérer meilleur que l'intervalle exact car ce dernier est fondé sur une borne uniforme qui peut être mauvaise pour certaines valeurs de x (voir [9]).

1.4 Fonctionnelles d'une distribution

On appelle fonctionnelle T, une application d'un espace fonctionnel $\mathcal{F}$ dans $\mathbb{R}$:

$$\begin{aligned} T: \ \mathcal{F} \quad & \rightarrow \quad \mathbb{R} \\ F \quad & \rightarrow \quad T(F). \end{aligned}$$

Exemple 1.4.1

– Moyenne : $E[F] := \int x dF(x)$.

– Variance : $Var(F) := \int (x - E[F])^2 dF(x) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$.

– Quantile : pour $0 \leq q \leq 1$, $Q_q(F) := F^{\leftarrow}(q)$. Le cas $q = 1/2$ correspond à la médiane de la distribution.

– Coefficient de dissymétrie (asymétrie) : $cd(F) := \left(\int (x - E[F])^3 dF(x)\right) / \left(Var(F)\right)^{\frac{3}{2}}$.

– Coefficient d'aplatissement : $ca(F) := \left(\int (x - E[F])^4 dF(x)\right) / \left(Var(F)\right)^2$.

1.4.1 Fonctionnelles linéaires et fonctionnelles de moment

– Une fonctionnelle T est dite linéaire s'il existe une fonction $a : \mathbb{R} \to \mathbb{R}$ telle que

$$T(F) = \int a(x) dF(x).$$

– Une fonctionnelle T est dite de moment s'il existe un entier $k \geq 1$ et une fonction $\phi : \mathbb{R}^k \to \mathbb{R}$ telle que

$$T(F) = E[\phi(X_1, ..., X_k)] = \int \phi(x_1, ..., x_k) dF(x_1)...dF(x_k).$$

Remarque 1.4.1

1. Les fonctionnelles linéaires sont des fonctionnelles de moment.
2. La moyenne est une fonctionnelle est linéaire et de moment.
3. Les variance, quantile et les coefficients d'asymétrie et d'aplatissement ne sont ni linéaires ni de moment.

1.4.2 Estimateurs par substitution ("plug in")

Un estimateur naturel (non paramétrique) de $T(F)$ est obtenu en substituant la fonction de répartition empirique F_n dans l'expression de T. En d'autres termes, $T(F_n)$ est un estimateur naturel de $T(F)$.

Exemple 1.4.2

1. La moyenne empirique $\bar{X}_n := \int x dF_n(x) = \sum_{i=1}^n X_i/n$ est l'estimateur de la moyenne théorique $E[X] := \int x dF(x)$.
2. La variance empirique S_n^2 est l'estimateur de la variance, est définie par

$$S_n^2 := \int x^2 dF_n(x) - \left(\int x dF_n(x) \right)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

3. L'estimateur par substitution d'une fonctionnelle linéaire est

$$\int a(x) dF_n(x) = \sum_{i=1}^n a(X_i)/n.$$

1.4.3 Fonction d'influence

La fonction d'influence est une dérivée de la fonctionnelle T, elle est utilisée pour approximer l'erreur standard d'un plug-in estimateur.
Pour définir une dérivée, il faut définir un taux d'accroissement. Comme une fonctionnelle T a pour argument $F \in \mathcal{F}$, il faut définir un accroissement élémentaire dans l'espace $\mathcal{F}$.

Définition 1.4.1 *Soit $T : F \to T(F)$ une fonctionnelle. La fonction d'influence de T en F en un point $x_0 \in \mathbb{R}$ est définie par la limite suivante, si elle existe*

$$L_{T,F}(x_0) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) - T(F)}{\epsilon},$$

où, pour tout $t \in \mathbb{R}$, $G_{\delta_{x_0}}(t) = \mathbf{1}_{(x_0 \leq t)}$ représente la fonction de répartition associée à la masse de Dirac δ_{x_0}.

Exemple 1.4.3 (la moyenne) *$\forall \epsilon > 0$, $\forall x_0 \in \mathbb{R}$, on a*

$$E\left[(1-\epsilon)F + \epsilon G_{\delta_{x_0}}\right] = (1-\epsilon)E\left[F\right] + \epsilon E\left[G_{\delta_{x_0}}\right],$$

car $E[.]$ est linéaire. De plus, $E[G_{\delta_{x_0}}] = x_0$, alors on a

$$\frac{E\left[(1-\epsilon)F + \epsilon G_{\delta_{x_0}}\right] - E\left[F\right]}{\epsilon} = \frac{(1-\epsilon)E\left[F\right] + \epsilon x_0 - E[F]}{\epsilon} = x_0 - E[F].$$

Ainsi, $L_{E,F}(x_0) = x_0 - E[F]$.

Plus généralement, il est facile de voir que, pour une fonctionnelle linéaire, on a

$$L_{T,F}(x_0) = a(x_0) - T(F).$$

Fonction d'influence empirique

Définition 1.4.2 *La fonction d'influence empirique de T en F au point x_0 est*

$$L_{T,F_n}(x_0) = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F_n + \epsilon G_{\delta_{x_0}}\right) - T(F_n)}{\epsilon}.$$

Exemple 1.4.4 *La fonction d'influence empirique associée à la moyenne en F au point x est*

$$L_{E,F_n}(x) = x - \bar{X}_n.$$

Dans ce cas, la quantité $L_{E,F_n}(X_i)$ mesure la contribution de l'observation X_i à la variance empirique

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(L_{E,F_n}(X_i))^2.$$

Plus généralement, la quantité $L_{E,F_n}(X_i)$ mesure la contribution de l'observation X_i à n'importe quel moment empirique.

Proposition 1.4.1 *Si $T(F)$ est une fonctionnelle linéaire, alors on a*

1. La fonction d'influence empirique

$$L_{T,F_n}(x) = a(x) - T(F_n) = a(x) - \frac{1}{n}\sum_{i=1}^{n} a(X_i).$$

2. $E[L_{T,F}(X)] = \int L_{T,F}(x)dF(x) = 0.$

3. $Var(L_{T,F}(X)) = \int (a(x) - T(F))^2 \, dF(x) = \int a^2(x)dF(x) - T^2(F).$

Preuve. Ces résultats sont directs car on considère une fonctionnelle linéaire. ■

Théorème 1.4.1 *Si* $\tau^2 := Var(L_{T,F}(X)) < \infty$*, alors*

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2), \ \textit{quand } n \to \infty.$$

En estimant τ^2 *par*

$$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n} L^2_{T,F_n}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\left[a(X_i) - \frac{1}{n}\sum_{i=1}^{n} a(X_i)\right]^2,$$

on a

$$\frac{\sqrt{n}(T(F_n) - T(F))}{\hat{\tau}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \ \textit{quand } n \to \infty.$$

Preuve. Voir [16], pages 18-19 (étapes 4, 5 et 6). ■

Chapitre 2

Tests d'ajustement

Les tests d'ajustement sont des procédures non paramétriques permettant de juger l'adéquation entre une situation réelle et un modèle théorique. Ils ont pour but de vérifier si un échantillon, d'une population X de fonction de répartition F inconnue, provient ou non d'une v.a de distribution connue F_0. Il s'agit donc de tester les hypothèses suivantes :

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases} \text{ ou bien } \begin{cases} H_0 : & X \sim & \text{une certaine loi spécifiée,} \\ H_1 : & X \nsim & \text{cette loi.} \end{cases}$$

Il existe une grande variété de tests d'ajustement, parmi lesquels on peut citer :

- Le test du khi-deux, basés sur les effectifs.
- Les tests qui reposent sur la fonction de répartition empirique : tests de Kolmogorov-Smirnov, d'Anderson-Darling et de Cramer-von Mises.
- Les tests basés sur les moments, comme celui de d'Agostino's K-squared.
- Les tests d'ajustement appliqués à la loi normale, appelés tests de normalité, comme les tests de Lilliefors (fonction de répartition empirique), de Jarque-Bera (moments) et de Shapiro-Wilk (L-statistiques). Pour une description détaillée de ces tests, on recommande de voir [15], [2], [4],...

On note qu'il est conseillé de faire une étude descriptive préliminaire des données dans

le but d'avoir une idée sur la distribution adéquate à ces données. Les éléments les plus pertinents à la modélisation sont :

– Discussion de la forme de l'histogramme.
– Vérification sommaire de certaines propriétés des paramètres statistiques.
– Ajustement graphique : quantile-quantile plot (Q-Q plot), probability-probability plot (P-P plot),...

La conclusion à laquelle on arrive sera ensuite confirmée ou infirmée par les tests statistiques d'ajustement.

Dans ce chapitre, on s'intéresse aux tests construits sur la base de la fonction de répartition empirique, cités ci-dessus. L'idée générale de ces tests est la suivante : si l'hypothèse H_0 est vraie, alors la fonction de répartition empirique F_n doit être proche, par rapport à une certaine distance, de la fonction de répartition hypothétique F_0. Par conséquent, différents choix de la distance engendrent différents types de procédures.

2.1 Test de Kolmogorov-Smirnov

C'est le plus populaire parmi les tests d'adéquation qui sont basés sur la fonction de répartition empirique. Il a été proposé par Andreï N. Kolmogorov en 1933 et étendu par Vladimir I. Smirnov en 1939.

2.1.1 Statistique du test

La distance utilisée pour définir la statistique D_n de ce test est celle de la norme uniforme. La statistique de Kolmogorov-Smirnov est alors définie par

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| .$$

Remarque 2.1.1 *Puisque* $0 \leq F_0(x) \leq 1$ *et* $0 \leq F_n(x) \leq 1$, *alors* $0 \leq D_n \leq 1$.

Pour calculer les valeurs de la statistique D_n, il suffit d'évaluer la différence entre F_n et F_0 aux points $x_{(i)}$ comme l'indique la proposition 2.1.1.

Proposition 2.1.1 *La statistique de Kolmogorov-Smirnov s'écrit comme suit :*

$$D_n = \max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - F_0(x_{(i)})\right], \max_{1\leq i\leq n}\left[F_0(x_{(i)}) - \frac{i-1}{n}\right], 0\right\}. \tag{2.1}$$

Preuve. La statistique D_n peut s'écrire $D_n = \max(D_n^+, D_n^-)$, avec

$$D_n^+ := \sup_{x\in\mathbb{R}}\left(F_n(x) - F_0(x)\right) \text{ et } D_n^- := \sup_{x\in\mathbb{R}}\left(F_0(x) - F_n(x)\right).$$

On a

$$D_n^+ = \sup_{x\in\mathbb{R}}\left(F_n(x) - F_0(x)\right) = \max_{0\leq i\leq n}\sup_{x_{(i)}\leq x<x_{(i+1)}}\left(F_n(x) - F_0(x)\right).$$

On définit $x_{(0)} = -\infty$ et $x_{(n+1)} = +\infty$, on peut écrire $F_n(x) = i/n$ pour $x_{(i)} \leq x < x_{(i+1)}$, $i = 0, 1, ..., n$, alors on a

$$D_n^+ = \max_{0\leq i\leq n}\sup_{x_{(i)}\leq x<x_{(i+1)}}\left(\frac{i}{n} - F_0(x)\right) = \max_{0\leq i\leq n}\left(\frac{i}{n} - \inf_{x_{(i)}\leq x<x_{(i+1)}} F_0(x)\right).$$

Or, F_0 est une fonction croissante, d'où

$$D_n^+ = \max_{0\leq i\leq n}\left(\frac{i}{n} - F_0(x_{(i)})\right) = \max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - F_0(x_{(i)})\right], 0\right\}.$$

Par le même principe, on montre le résultat relatif à D_n^-. Pour une description détaillée de ce résultat, on recommande de voir [5], pages 109-110. ■

Remarque 2.1.2 *Dans le cas où F_0 est continue, les lois de D_n, D_n^+ et D_n^- sont indépendantes de F_0. En effet, si F_0 est continue, les v.a's $F_0\left(X_{(i)}\right)$, $i = 1, 2, ..., n$, sont uniformes sur $[0, 1]$, c'est-à-dire $F_0\left(X_{(i)}\right) \overset{\mathcal{L}}{=} U_{(i)}$, pour $i = 1, 2, ..., n$, indépendamment de F_0. Par conséquent, D_n, D_n^+ et D_n^- ont des distributions indépendantes de la F_0.*

Par le changement de variable $y = F_0(x)$, on peut écrire

$$D_n = \sup_{0 \leq y \leq 1} |G_n(y) - y| ,$$

où G_n est la fonction de répartition empirique uniforme.

En 1933, Kolmogorov a proposé une approximation à la loi (d'une fonction) de la statistique D_n et Smirnov a donné, en 1939, une démonstration plus simple du résultat. Cette distribution asymptotique est présentée, sans démonstration, dans le théorème 2.1.1.

Théorème 2.1.1 *Pour tout $t > 0$, on a*

$$\lim_{n \to \infty} P(\sqrt{n} D_n < t) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2 t^2).$$

La distribution exacte de D_n est présentée dans le théorème 2.1.2.

Théorème 2.1.2 *Dans le cas où F_0 est continue, on a, pour tout réel t et $n \geq 1$,*

$$P(D_n < t) = \begin{cases} 0 & si \quad t \leq \frac{1}{2n}, \\ \int_{1/n-t}^{t} \int_{5/6n-t}^{t-1/6n} \cdots \int_{1-t}^{(n-1)/n+t} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n & si \quad \frac{1}{2n} < t < 1, \\ 1 & si \quad t \geq 1, \end{cases}$$

où

$$f(u_1, u_2, ..., u_n) = n! \mathbf{1}_{(0 < u_1 < u_2 < ... < u_n < 1)}.$$

Preuve. Voir [5], pages 111-112. ■

Pour une description détaillée de la distribution commune aux deux statistiques D_n^+ et D_n^-, voir [5], pages 115-116.

2.1.2 Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation (2.1) puis on décide du rejet ou non du modèle proposé. L'exécution du test de Kolmogorov-Smirnov est donnée par

les étapes suivantes :

1. classer les valeurs observées par ordre croissant ;
2. calculer, pour $i = 1$, les valeurs absolues des écarts

$$\left|F_0(x_{(i)}) - i/n\right| \text{ et } \left|F_0(x_{(i)}) - (i-1)/n\right| \ ;$$

3. prendre le plus grand des deux écarts absolus ;
4. répéter les étapes 2 et 3 pour $i = 2, ..., n$;
5. la valeur de la distance de Kolmogorov-Smirnov est égale au maximum des plus grands écarts.

La région critique du test est de la forme $\{D_n > D_{crit}\}$, où D_{crit} est une certaine valeur critique vérifiant, $P(D_n > D_{crit}/H_0 \text{ est vraie}) = \alpha$, $0 \leq \alpha \leq 1$. On conclut le test en acceptant, au seuil de signification, l'hypothèse H_0 si la distance D_n calculée est inférieure D_{crit}. La distribution théorique spécifiée est alors acceptée, c'est à dire : $F = F_0$.
La valeur critique D_{crit} est lue dans la table de Kolmogorov-Smirnov (voir, par exemple, [12], pages 585-586). Pour les petites tailles d'échantillon, il y a, dans [5], pages 113-114, un exemple de calcul de D_{crit} (pour $n = 2$) à partir de la loi exacte de D_n. Pour $n \geq 50$, les valeurs critiques sont données, selon quelques valeurs de α, dans le tableau 2.1.

2.2 Test de Cramer-von Mises

Le test était développé par Harald Cramer et Richard E. von Mises (1928-1930).

2.2.1 Statistique du test

Ce test est basé sur la différence quadratique entre la fonction de répartition empirique et la fonction de répartition théorique. La statistique du test d'ajustement de Cramer-von

Mises $\bar{w}_n^2$ est définie par

$$\bar{w}_n^2 := n \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 dF_0(x). \tag{2.2}$$

En pratique, son calcul est simplifié comme l'indique la proposition 2.2.1.

Proposition 2.2.1

$$\bar{w}_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} (\frac{2i-1}{2n} - F_0(x_i))^2. \tag{2.3}$$

Preuve. On découpe l'intégrale de la formule (2.2) sur les intervalles de la forme $\left[x_{(i)}, x_{(i+1)}\right[$, pour écrire

$$\begin{aligned}\bar{w}_n^2 = n \Bigg\{ &\int_{-\infty}^{x_{(1)}} (F_n(x) - F_0(x))^2 dF_0(x) + \int_{x_{(1)}}^{x_{(2)}} (F_n(x) - F_0(x))^2 dF_0(x) + ... \\ &+ \int_{x_{(n-1)}}^{x_{(n)}} (F_n(x) - F_0(x))^2 dF_0(x) + \int_{x_{(n)}}^{+\infty} (F_n(x) - F_0(x))^2 dF_0(x) \Bigg\}.\end{aligned}$$

On utilise maintenant l'équation (1.1) , pour avoir

$$\begin{aligned}\bar{w}_n^2 = n \Bigg\{ &\int_{-\infty}^{x_{(1)}} (0 - F_0(x))^2 dF_0(x) + \int_{x_{(1)}}^{x_{(2)}} (\frac{1}{n} - F_0(x))^2 dF_0(x) + ... \\ &+ \int_{x_{(n-1)}}^{x_{(n)}} (\frac{n-1}{n} - F_0(x))^2 dF_0(x) + \int_{x_{(n)}}^{+\infty} (1 - F_0(x))^2 dF_0(x) \Bigg\}.\end{aligned}$$

Par le changement de variable $y = F_0(x)$, ceci devient

$$\begin{aligned}\bar{w}_n^2 &= n \Bigg\{ \int_0^{F_0(x_{(1)})} y^2 dy + \int_{F_0(x_{(1)})}^{F_0(x_{(2)})} \left(\frac{1}{n} - y\right)^2 dy + ... + \int_{F_0(x_{(n-1)})}^{F_0(x_{(n)})} \left(\frac{n-1}{n} - y\right)^2 dy + \\ &\int_{F_0(x_{(n)})}^{1} (1-y)^2 dy \Bigg\} \\ &= \frac{n}{3} \Bigg\{ \left(F_0(x_{(1)})\right)^3 + \left[\left(\frac{1}{n} - F_0(x_{(1)})\right)^3 - \left(\frac{1}{n} - F_0(x_{(2)})\right)^3 \right] + ... \\ &+ \left[\left(\frac{n-1}{n} - F_0(x_{(n-1)})\right)^3 - \left(\frac{n-1}{n} - F_0(x_{(n)})\right)^3 \right] + \left(1 - F_0(x_{(n)})\right)^3 \Bigg\}.\end{aligned}$$

On ordonne les termes de façon à écrire

$$\bar{w}_n^2 = \frac{n}{3}\left\{\left(\frac{1}{n} - F_0(x_{(1)})\right)^3 - \left(\frac{0}{n} - F_0(x_{(1)})\right)^3 + \left(\frac{2}{n} - F_0(x_{(2)})\right)^3 - \left(\frac{1}{n} - F_0(x_{(2)})\right)^3 + ... + \left(\frac{n}{n} - F_0(x_{(n)})\right)^3 - \left(\frac{n-1}{n} - F_0(x_{(n)})\right)^3\right\}. \tag{2.4}$$

On remarque que ces termes sont du type $(a^3 - b^3)$ avec $b = a - 1/n$ et $a = i/n - F_0(x_{(i)})$, pour $i = 1, 2, ..., n$. On a

$$\begin{aligned} a^3 - b^3 &= \frac{1}{n}\left[3a^2 - \frac{3a}{n} + \frac{1}{n^2}\right] = \frac{1}{n}\left[3\left(a^2 - \frac{a}{n} + \frac{1}{4n^2}\right) - \frac{3}{4n^2} + \frac{1}{n^2}\right] \\ &= \frac{1}{n}\left[3\left(a - \frac{1}{2n}\right)^2 + \frac{1}{4n^2}\right] = \frac{3}{n}\left[\left(\frac{2i-1}{2n} - F_0(x_{(i)})\right)^2\right] + \frac{1}{4n^3}. \end{aligned}$$

En remplaçant dans l'équation (2.4), on obtient

$$\bar{w}_n^2 = \frac{n}{3}\left[\frac{3}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F_0(x_{(i)})\right)^2\right] + \sum_{i=1}^{n}\frac{1}{12n^2} = \frac{1}{12n} + \sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F_0(x_i)\right)^2.$$

C'est ce qu'il fallait trouver. ■

Remarque 2.2.1 *Comme pour le test de Kolmogorov-Smirnov, la loi de $\bar{w}^2$ est indépendante de F_0, si elle est continue. Par le changement de variable $y = F_0(x)$; dans la relation* (2.2) *on peut écrire*

$$\bar{w}_n^2 = n\int_0^1 (G_n(y) - y)^2 dy.$$

La distribution asymptotique de la statistique $\bar{w}^2$ est présentée dans le théorème 2.2.1.

Théorème 2.2.1

$$\bar{w}_n^2 \xrightarrow{\mathcal{L}} \sum_{j=1}^{+\infty}\frac{1}{j^2\pi^2}Z_j^2, \text{ quand } n \to \infty,$$

où $Z_1, Z_2, ...$ sont des v.a's de la loi normale standard.

Preuve. Voir [15], pages 131-132. ■

2.2.2 Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation (2.3) , puis on décide du rejet ou non du modèle proposé. L'exécution du test de Cramer-von Mises est donnée par les étapes suivantes :

1. classer les valeurs observées par ordre croissant ;
2. utiliser la fonction de répartition de la loi pour obtenir les valeurs de $F_0(x_i)$, pour $i = 1, 2, ..., n$;
3. calculer $\sum_{i=1}^{n}(\frac{2i-1}{2n} - F_0(x_i))^2$, puis la valeur de la statistique $\bar{w}_n^2$.

On rejette l'hypothèse H_0 si cette dernière est supérieure à une certaine valeur critique n'ayant qu'une probabilité α d'être dépassée, sous l'hypothèse H_0. Il existe une table statistique, connue sous le nom de table de Cramer-von Mises, dans laquelle sont résumées les valeurs critiques pour les niveaux de signification usuelles avec différentes tailles d'échantillon (voir, par exemple, [12], page 584). Pour $n \geq 50$, les valeurs critiques sont données, selon quelques valeurs de α, dans le tableau 2.1.

2.3 Test d'Anderson-Darling

Construit en 1954 par Theordore W. Anderson et Donald A. Darling dans une première version, puis généralisé par Michael A. Stephens en 1974. Il s'agit d'une modification du test de Cramer-von Mises, il donne plus d'importance aux queues de distribution.

2.3.1 Statistique du test

La statistique du test d'Anderson-Darling A_n^2 est définie par

$$A_n^2 := n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

Remarque 2.3.1 *Une simplification de cette statistique est donnée par*

$$A_n^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)\left\{\log(F_0(x_i)) + \log(1-F_0(x_{n+1-i}))\right\}\right) - n,$$

ou

$$A_n^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)\log(F_0(x_i)) + (2n+1-2i)\log(1-F_0(x_i))\right) - n. \tag{2.5}$$

La loi de A_n^2 est, comme pour les deux tests précédents, indépendante de F_0 dans le cas où cette dernière est continue.

La distribution asymptotique de la statistique A_n^2 est présentée dans le théorème 2.3.1.

Théorème 2.3.1

$$A_n^2 \xrightarrow{\mathcal{L}} \sum_{j=1}^{+\infty} \frac{1}{j(j+1)} Z_j^2, \text{ quand } n \to \infty,$$

où $Z_1, Z_2, ...$ sont des v.a's de la loi normale standard.

Preuve. Voir [15], page 133. ■

2.3.2 Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation (2.5) puis on décide du rejet ou non du modèle proposé. L'exécution du test d'Anderson-Darling est donnée par les étapes suivantes :

1. ordonner les observations de manière croissante ;
2. obtenir les fréquences théoriques $F_0(x_i)$, puis déduire $\log(F_0(x_i))$ et $\log(1-F_0(x_i))$;
3. calculer $\sum_{i=1}^{n}(2i-1)(\log(F_i) + (2n+1-2i)\log(1-F_0(x_i))$, puis la valeur de la statistique A_n^2 ;

Pour faire la décision, on compare la valeur A_n^2 avec une certaine valeur critique A_{crit}^2 au certain seuil de signification α, l'hypothèse H_0 est rejetée lorsque la statistique A_n^2 prend des valeurs trop élevées, c'est à dire que : $A_n^2 > A_{crit}^2$.

Les valeurs critiques de A_n^2 ont été tabulées (voir, par exemple, [4], page 112 et [14]).

Pour $n \geq 50$, les valeurs critiques sont données, selon quelques valeurs de α, dans le tableau 2.1.

n	Kolmogorov-Smirnov				Cramer-von Mises				Anderson-Darling			
	Niveau de signification											
	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
50	0.15	0.16	0.17	0.19	0.24	0.28	0.34	0.45	1.43	1.62	1.90	2.42
100	0.11	0.11	0.12	0.14	0.24	0.29	0.36	0.47	1.39	1.60	1.91	2.41
200	0.08	0.08	0.09	0.10	0.24	0.24	0.28	0.45	1.39	1.59	1.92	2.49
500	0.05	0.05	0.05	0.06	0.23	0.28	0.33	0.44	1.41	1.61	1.93	2.50
800	0.04	0.04	0.04	0.05	0.24	0.29	0.35	0.47	1.41	1.62	1.90	2.40
1000	0.03	0.04	0.04	0.04	0.25	0.29	0.35	0.45	1.42	1.64	1.95	2.51
2000	0.02	0.03	0.03	0.03	0.24	0.29	0.35	0.48	1.39	1.59	1.91	2.44

TAB. 2.1 – Quelques valeurs critiques des tests de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling (source [13]).

2.4 Test de normalité de Lilliefors

Les tests précédents sont des tests généraux qui s'appliquent à n'importe quelle distribution F_0 de l'hypothèse nulle. Lorsque cette dernière est la loi normale, on parle de test de normalité. Il s'agit donc de vérifier l'ajustement d'un ensemble d'observations à un modèle Gaussien. Les hypothèses suivantes à tester sont donc

$$\begin{cases} H_0 : & \text{les données suivent une loi normale,} \\ H_1 : & \text{les données ne suivent pas une loi normale.} \end{cases}$$

Il existe plusieurs procédures de ce type, parmi lesquelles le test de Lilliefors qui a été introduit en 1967 par Hubert Lilliefors. C'est une approche non paramétrique visant à tester si une variable continue X suit une loi normale de paramètres μ et σ^2 inconnus et

qui sont alors estimés par leurs contre parties empiriques $\bar{x}_n$ et s_n^2 respectivement.

La statistique de Lilliefors L est définie par :

$$L := \max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - \Phi\left(z_{(i)}\right)\right], \max_{1\leq i\leq n}\left[\Phi\left(z_{(i)}\right) - \frac{i-1}{n}\right], 0\right\}, \tag{2.6}$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite et $z_{(i)}$ est la valeur ordonnée de z_i, où $z_i := (x_i - \bar{x}_n)/s_n$, $i = 1, 2, ..., n$.

2.4.1 Principe du test

Le principe de calcul est très similaire au test de Kolmogorov-Smirnov, à la différence que les paramètres de la loi sont estimés et que les valeurs critiques sont modifiées. L'exécution du test de Lilliefors est donnée par les étapes suivantes :

1. ordonner les observations de manière croissante ;
2. calculer les paramètres $\bar{x}_n$ et s_n^2 ;
3. calculer alors les données centrées et réduites z_i ;
4. obtenir les valeurs $\Phi\left(z_{(i)}\right)$;
5. calculer la valeur de la statistique L.

Pour faire la décision, on compare la valeur L avec une certaine valeur critique L_{crit} correspondant à un seuil de signification α fixé. Si $L > L_{crit}$, l'hypothèse H_0 est rejetée avec un risque maximum de se tromper égal à α, sinon elle est acceptée. Les valeurs critiques L_{crit} sont tabulées (voir, par exemple, [17]).

2.5 Application sous R

On illustre les résultats théoriques de ce chapitre sur des exemples de données simulées et réelles. Les résultats numériques sont obtenus à l'aide du logiciel d'analyse statistique **R**.

2.5.1 Données simulées

Le but est d'appliquer les procédures ci-dessus pour vérifier si un ensemble de données s'ajustent à un modèle de probabilité proposé. Pour cela, on génère 1000 échantillons, de 100 observations chacun, d'une population X de distribution exponentielle de paramètre 1. La statistique et la p-valeur de chaque test sont prises comme les moyennes sur les 1000 réplications des quantités correspondantes. On considère les deux tests suivants :

$$\begin{cases} H_0 : & X \sim & \mathcal{E}(1), \\ H_1 : & X \nsim & \mathcal{E}(1), \end{cases} \quad \text{et} \quad \begin{cases} H_0 : & X \sim & \mathcal{U}([0,10]), \\ H_1 : & X \nsim & \mathcal{U}([0,10]). \end{cases}$$

Les résultats de cette étude de simulation sont présentés dans le tableau 2.2. Mais, tout d'abord on commence par une investigation graphique des Q-Q plots illustrée par la figure (2.1). Comme attendu, on constate une allure linéaire dans le panneau de gauche et une forme non linéaire dans le panneau de droite.

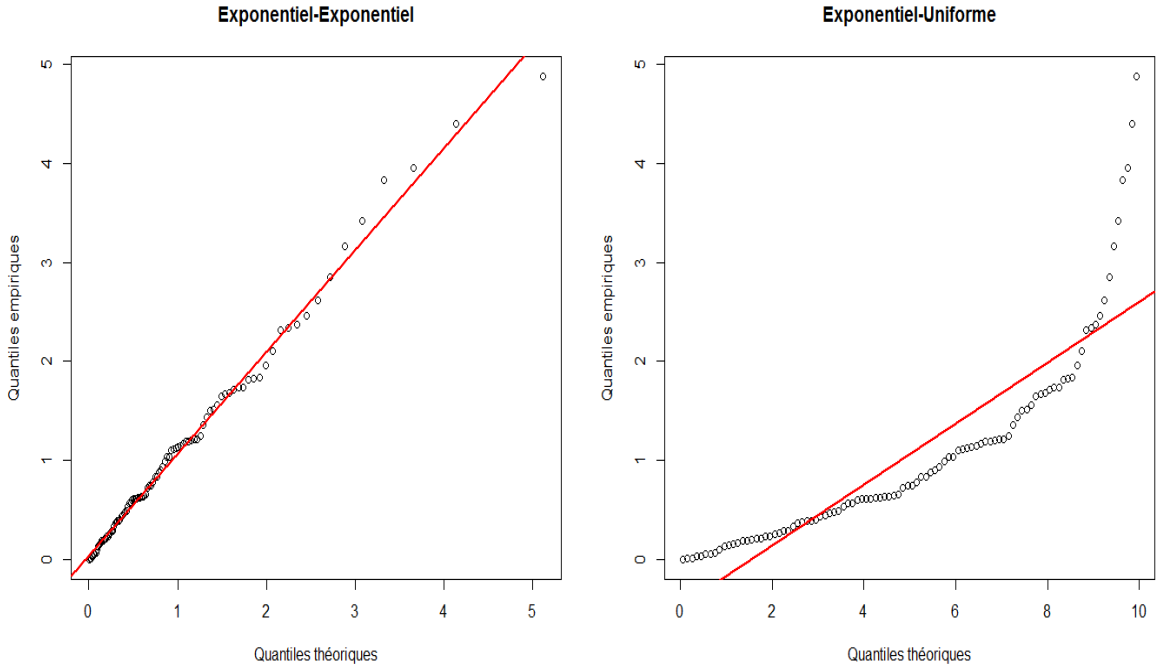

FIG. 2.1 – Q-Q plots des hypothèses d'exponentialité de paramètre 1 (à gauche) et d'uniformité sur $[0, 10]$ (à droite) d'un échantillon exponentiel $\mathcal{E}(1)$.

	Exponentiel-Exponentiel		Exponentiel-Uniforme	
Test	statistique	p-valeur	statistique	p-valeur
Kolmogorov-Smirnov	0.086	0.521	0.689	0
Cramer-von Mises	0.165	0.499	20.391	0
Anderson-Darling	0.996	0.500	∞	0

TAB. 2.2 – Résultats des tests d'exponentialité et d'uniformité d'un échantillon exponentiel.

La troisième colonne du tableau 2.2 indique que tous les tests ont accepté l'hypothèse d'exponentialité des observations pour chacun des niveaux de signification usuels. En effet, les p-valeurs des trois tests sont supérieures aux seuils 1%, 5% et 10%. D'autre part, la dernière colonne montre que l'hypothèse d'uniformité est rejetée par les trois tests pour chacun des niveaux de signification usuels. Les p-valeurs étant nulles, elles sont donc inférieures aux seuils cités ci-dessus.

On note que ces conclusions ne contredisent pas la figure (2.1) où les points du Q-Q plot sont approximativement alignés dans le graphe de gauche et ne le sont pas dans celui de droite.

2.5.2 Données réelles

L'objectif de cette étude est de vérifier la normalité d'un ensemble de données réelles. Ces dernières représentent les résultats (en mètres) de l'épreuve du "lancer de poids", de 33 joueurs au concours de décathlon des jeux olympiques de Séoul (Corée du Sud) en 1988 (voir [7], page 293). C'est un exemple qui est adopté par le logiciel **R,** sous le nom "olympic" dans le package ade4.

Les paramètres statistiques de cette série sont résumés dans le tableau 2.3.

minimum	1^{er} quartile	mediane	$3^{ème}$ quartile	maximum
10.27	13.15	14.12	14.97	16.60
moyenne	variance	mode	asymétrie	aplatissement
13.98	1.72	14.05	-0.41	3.23

TAB. 2.3 – Paramètres statistiques des données réelles.

On voit sur ce dernier que les moyenne, médiane et mode sont approximativement égaux, ce qui implique une symétrie de la distribution. De plus, la valeur du coefficient d'aplatissement favorise le modèle gaussien pour les observations. D'autre part, l'examen du Q-Q plot et de l'histogramme des fréquences, donné par la figure (2.2), permet de conclure que les données suivent loi normale. En effet, on remarque que l'histogramme a une forme plus ou moins symétrique et que le Q-Q plot présente une allure linéaire.

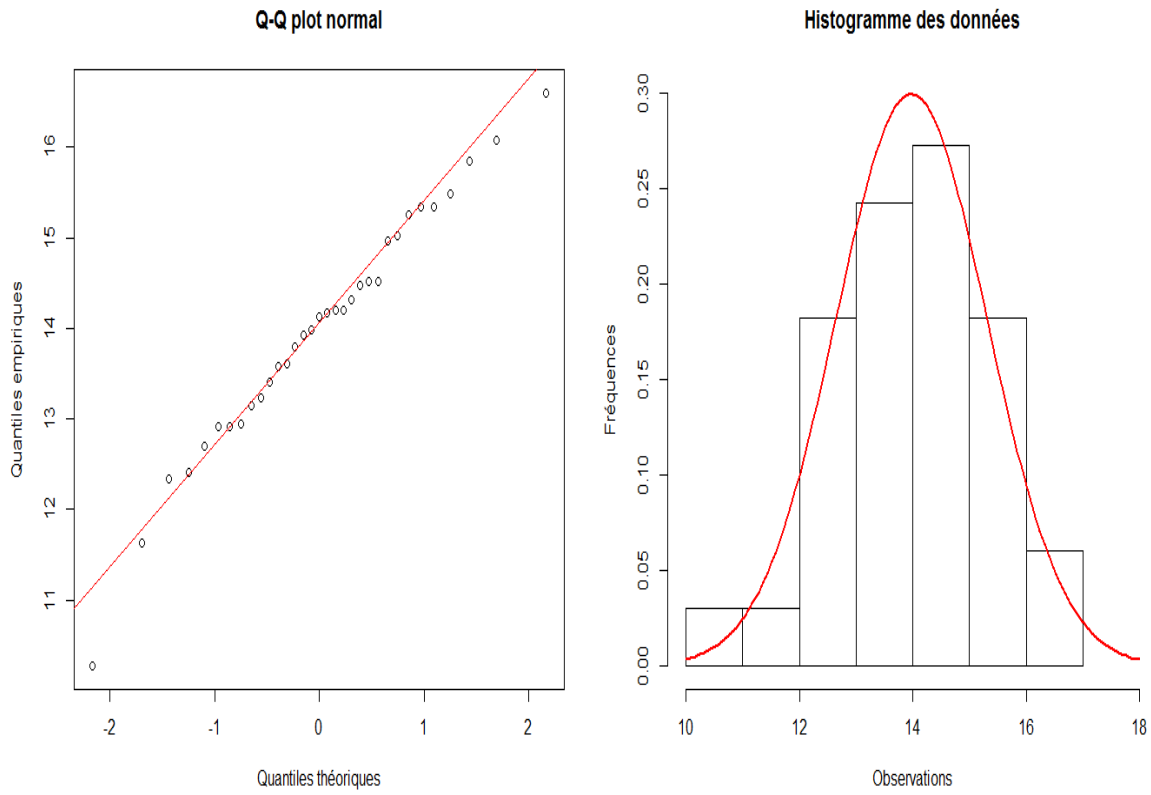

FIG. 2.2 – Q-Q plot et Histogramme de fréquences des données réelles.

Test	Statistique	p-valeur
Kolmogorov-Smirnov	0.069	0.998
Cramer-von Mises	0.024	0.992
Anderson-Darling	0.180	0.995
Lilliefors	0.069	0.957

TAB. 2.4 – Résultats des tests avec des données réelles d'une hypothèse de normalité.

Enfin, les tests statistiques, dont les résultats sont résumés dans le tableau 2.4, permettant

d'ajuster les observations à une distribution normale et confirment ainsi les conclusions obtenues par les considérations numériques et graphiques ci-dessus. En effet, on voit clairement que les p-valeurs des quatre tests sont largement supérieures à tous les niveaux de signification usuels, ce qui implique que l'hypothèse de normalité ne peut être rejetée.

Conclusion

L'objectif principal de ce travail est de passer en revue les tests d'ajustement, d'un ensemble d'observations à un modèle de probabilité, basés sur la fonction de répartition empirique. Pour ce faire, ce mémoire est consacré premièrement aux définitions et propriétés fondamentales de cette dernière et deuxièmement aux tests d'ajustement dont les statistiques représentent des distances entre les fonctions de répartition théorique et empirique. L'intérêt a surtout porté sur les tests de Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling ainsi que celui de Lilliefors relatif à la normalité. Appliquées à des distributions continues, ces procédures sont plus fiables que le test d'ajustement du khi-deux (basé sur les effectifs) qui repose sur le regroupement des observations en classes, entraînant ainsi une perte d'information sur les données.

Le calcul des puissances, qui reste à faire, permettra de comparer ces tests entre eux ainsi qu'avec d'autres basés sur des outils statistiques autres que la fonction de répartition empirique.

Bibliographie

[1] Borovkov, A A. (1999). Mathematical Statistics. CRC Press.

[2] Boulay, J-P. (2010). Statistique mathématique : Applications commentées. Ellipses.

[3] Colletaz, G. (2017). Statistique non paramétrique : Économétrie et statistique appliquée. Lien : *www.univ-orleans.fr/deg/masters/ESA/GC/sources/CoursNP.pdf.*

[4] D'Agostino, R. B. & Stephens, M. A. (1986). Goodness-of-fit techniques. Marcel Dekker, Inc.

[5] Gibbons, J. D. & Chakraborti, S. (2010). Nonparametric statistical inference. CRC Press.

[6] Gaudoin, O. (2011). Statistique inférentielle avancée. Ensimag-2ème année. INP Grenoble. Lien : *https ://www-ljk.imag.fr/membres/Olivier.Gaudoin/SIA.pdf.*

[7] Hand, D. J. Daly, F. McConway, K. J. Lunn, A. D. & Ostrowski, E. (1994). A handbook of small data sets. Springer.

[8] Lejeune, M. (2010). Statistique : la théorie et ses applications. Springer.

[9] Matias, C. & Atelier, S. F. D. S. (2013). Introduction à la statistique non paramétrique. Laboratoire Statistique & Génome, Évry. Lien : *http ://cmatias.perso.math.cnrs.fr/Docs/atelier_stat_np_1_intro.pdf.*

[10] Meraghni, Dj. (2017). Cours de première MASTER. Université Mohamed Khider de Biskra.

[11] Necir, A. (2016). Cours de troisième année. Université Mohamed Khider de Biskra.

[12] Saporta, G. (2006). Probabilité, analyse de données et statistique. Technip.

[13] Singla, N. Jain, K. & Sharma, S. K. (2016). Goodness of fit tests and power comparisons for weighted gamma distribution. REVSTAT–Statistical Journal, 14(1), 29-48.

[14] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association, 69(347), 730-737.

[15] Thas, O. (2010). Comparing distributions. Springer.

[16] Wasserman, L. (2006). All of nonparametric statistics : with 52 illustrations. Springer.

[17] http ://courses.wcupa.edu/rbove/eco252/252KStest.doc.

Annexe : Abréviations et Notations

Les différentes abrévariations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

A_n^2	Statistique d'Anderson-Darling.
A_{crit}^2	Valeur critique d'Anderson-Darling.
a	Fonctionnelle linéaire.
α	Risque de premier espèce.
D_n	Statistique de Kolmogorov-Smirnov.
D_{crit}	Valeur critique de Kolmogorov-Smirnov.
$\mathcal{E}(\lambda)$	Loi exponentielle de paramètre λ.
$E[X]$	Espérance mathématique de X.
exp	Fonction exponentielle.
F	Fonction de répartition.
F_0	Distribution hypothétique.
F_n	Fonction de répartition empirique.
$F^{\leftarrow}$	Fonction des quantilles.
$F_n^{\leftarrow}$	Fonction des quantilles empiriques.
$\mathcal{F}$	Espace fonctionnel.
f	Densité de probabilité.
Φ	Fonction de répartition de la loi normale centrée réduite.
G_n	Fonction de répartition empirique uniforme.

iid	Indépendantes identiquement distribuées.
$\inf(A)$	Borne inférieure de A.
L	Statistique de Lilliefors.
L_{crit}	Valeur critique de Lilliefors.
$L_{T,F}$	Fonction d'influence de la fonctionnelle T en F.
L_{T,F_n}	Fonction d'influence empirique de la fonctionnelle T en F.
log	Fonction logarithme.
$\max(A)$ (ou $\min(A)$)	Maximum de A (ou minimum de A).
MSE	Erreur quadratique moyenne.
$\mathcal{N}(0,1)$	Loi normale centrée réduite.
$\sup(A)$	Borne supérieure de A.
S_n^2	Variance empirique.
T	Fonctionnelle.
$\mathcal{U}([a,b])$	Loi uniforme sur l'intervalle $[a,b]$.
v.a	Variable aléatoire.
$Var(X)$	Variance mathématique de X.
V_n	Fonction des quantiles empiriques uniformes.
$\bar{w}_n^2$	Statistique de Cramer-von Mises.
X	Population.
$\bar{X}_n$	Moyenne empirique.
$(X_1, X_2, ..., X_n)$	Échantillon de taille n de X.
$(X_{(1)}, X_{(2)}, ..., X_{(n)})$	Statistiques d'ordre associées à $(X_1, X_2, ..., X_n)$.
$\mathbf{1}_A$	Fonction indicatrice de l'ensemble A.
$\xrightarrow{\mathcal{L}}$	Convergence en loi.
$\overset{\mathcal{L}}{=}$	Égalité en loi.
$\xrightarrow{p}$	Convergence en probabilité.
$\xrightarrow{p.s}$	Convergence presque sûrement.
$:=$	Égalité par définition.