UNIVERSITÉ KASDI MERBAH OUARGLA

Faculté des Mathématiques et des Sciences de la Matière

DÉPARTEMENT DE MATHÉMATIQUES

Master

Spécialité: Mathématiques

Option: Probabilités et Statistique

Par: ZERROUKI Ouafa

Théme

Ajustement d'un échantillon de loi inconnue Modèle "lois.test "

Soutenue publiquement le: 15/06/2023

Devant le jury composé de:

Dr. AGTI Mohammed	Université de Kasdi Merbah - Ouargla	Président
Dr. MEDDI Fatima	Université de Kasdi Merbah - Ouargla	Encadreur
Dr. ARBIA Hanane	Université de Kasdi Merbah - Ouargla	Examinateur

Dédicace

Je dédie ce mémoire

A la source de la patience, ma chère mère

A la source de ma force, mon chèr père

A mes soeurs

A mes amis et intimes.

A tous ceux qui étaient à côtés de moi et qui m'ont soutenu dans ma carrière universitaire.

Ouafa Zerrouki

REMERCIEMENTS

Mes premiers remerciments à **Dieu** tout-puissant pour la volonté et la patience qu'il ma données pour achever ce humble travail.

Mes sincères remerciements, mon respect et ma gratitude à mes chèrs parents que dieu les protégent

Je remercie mon encadreur **Dr.MEDDI Fatima,** pour sa disponabilité, sa patience, ses précieux conseils et sa confiance.

Mes vifs remerciements vont également aux membres du jury

Dr. ARBIA Hanane et **Dr. AGTI Mohammed**

Je remercie également tous ceux qui ont partagé avec moi les moments difficiles.

Ouafa Zerrouki

Annexe A : Abréviations et Notations

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

$(\Omega, \mathcal{A}, \mathbb{P})$	Espace de probabilité.
v.a	Variable aléatoire
i.i.d	Indépendantes identiquement distribuées
$f(x)$	Fonction de densité de X
$F(x)$	Fonction de répartition de X.
$F_n(x)$	Fonction de répartition empirique
$E(x)$	Espérance de la v.a X
$V(x)$	Variance de la v.a X
$F^{-1}(u)$	Fonction d'inverse de $F(x)$
$Q(u)$	Fonction de quantile
$\xrightarrow{p}$	Convergence en probabilité
$\xrightarrow{ps}$	Convergence presque sûre
$\xrightarrow{\mathcal{L}}$	Convergence en loi
R^2	Coefficient de détermination
KS	Test de kolmogorov-smirnov

Table des matières

Table des figures

Liste des tableaux

Introduction

L'ajustement d'un échantillon à une loi inconnue est une problématique fondamentale en statistique, qui se pose dans de nombreux domaines de recherche et d'application. Comprendre la distribution sous-jacente des données est crucial pour prendre des décisions éclairées et tirer des conclusions fiables. L'objectif principal de ce mémoire est d'explorer les différentes approches statistiques et les méthodes d'estimation utilisées pour déterminer la distribution théorique la mieux adaptée à un échantillon de données dont la loi est inconnue, en mettant l'accent sur l'estimation des paramètres et les tests d'ajustement.

Nous commencerons par les méthodes graphiques, telles que les QQ-plots " il remonte aux travaux de Francis Galton au 19e siècle", qui permettent une première exploration visuelle de l'ajustement. Ces approches graphiques offrent une représentation intuitive de l'échantillon par rapport à diverses distributions théoriques, facilitant ainsi l'identification des écarts et des tendances.

Ensuite, nous nous intéresserons aux techniques d'estimation des paramètres. Ces méthodes consistent à estimer les valeurs des paramètres de la distribution empirique qui correspondant le mieux à l'échantillon observé. Parmi ces techniques, l'estimation de maximum de vraisemblance (EMV) est largement utilisée. Cette méthode a été développée par le statisticien et généticien Ronald Fisher entre 1912 et 1922, nous étudierons les principes théoriques de l'EMV et discuterons de ses propriétés statistiques et de sa mise en œuvre pratique.

En complément des méthodes d'estimation, nous explorerons les tests d'ajustement sta-

tistiques. Ces tests permettent de quantifier l'adéquation entre l'échantillon de données et une distribution théorique donnée. Le test de Kolmogorov-Smirnov est le plus couramment utilisé. Il porte le nom du mathématicien russe Andréi Nikoláyevich Kolmogorov qui établit l'axiomatique des probabilités en 1933. Nous décrirons les principes de ces tests, leurs hypothèses et leur interprétation.

Ce mémoire vise à fournir une vue d'ensemble complète des approches statistiques utilisées pour l'ajustement d'un échantillon de loi inconnue. En explorant les méthodes graphiques, les techniques d'estimation et les tests d'ajustement, nous espérons fournir des outils précieux pour l'analyse statistique et la prise de décision dans des domaines aussi divers que la finance, la biologie, la psychologie et bien d'autres encore.

On donne un aspect général sur ce mémoire :

Dans le premier chapitre, on commence par donner quelques préliminaires statistiques nécessaires pour obtenir aux principaux résultats, convergence de suites de variables aléatoires avec des différents type de convergence, et des simulation des variables "méthode d'inversion".

En deuxième chapitre, on donne une vue générale sur le modèle de régression linéaire simple et de test de signification du modèle (Analyse de variance), nous explorerons les principes et les interprétations du QQ-plot et démontrerons comment il peut être utilisé pour diagnostiquer l'ajustement d'un modèle aux données et nous concentrons sur les bases de l'EMV, ses propriétés statistiques et sa mise en œuvre pratique.

Dans le troisième chapitre, nous commencerons par parler sur les principes du test de Kolmogorov-Smirnov qui occupe une place importante dans notre algorithme pour valider l'ajustement après avoir appliqué d'autres techniques d'adéquation. Notamment, nous présenterons les étapes du modèle ainsi que les codes et quelques simulations de cas pratiques pour illustrer le déroulement et la fiabilité de notre modèle lois.test. Nous finirons par une application réelle d'un jeu de données, sur le nombre de cas affecté par covid19 entre janvier 2020 et juillet 2022 et une discussion fructueuse sur les résultats.

Chapitre 1

Préliminaires probabilités statistiques

1.1 Notions sur les lois de probabilités

Dans cette section, nous concentrons sur quelques lois de probabilités usuelles. Nous donnerons les définitions de base et les propriétés les plus essentielles de chaque loi dont nous aurons besoin plus tard dans ce mémoire.

1.1.1 Définitions

Définition 1.1.1 (Tribu) *Soit A un ensemble de parties de Ω, on dit que A est une tribu si :*

1. $\{\phi\} \in A$.

2. A est stable par passage ou complémentaire $\left(\forall B \in A : B \in A \Leftrightarrow \overline{B} \in A\right)$

3. A est stable par union dénombrable $(\forall n \in N, B_n \in A \Longrightarrow \cup_{n\in N} B_n \in A)$

Le couple $(\Omega; A)$ s'appelle espace mesurable.

Définition 1.1.2 (Fonction mesurable) *Une mesure sur* $(\Omega; A)$ *est une fonction*

$$\mu : A \to \mathbb{R}_+ = [0; +\infty[$$

Telle que :

1. $\mu(\phi) = 0$
2. $\forall (A_i)_{i\in I\subseteq\mathbb{N}} \quad \mu\left(\cup_{i\in I} A_i\right) = \sum_{i\in I} \mu\left(A_i\right) \quad si \quad \forall i, j \in I \quad A_i \cap A_j = \{\phi\}$
3. $\forall i, j \in I \quad \mu\left(\cup_{i\in I} A_i\right) \leq \sum_{i\in I} \mu\left(A_i\right) \quad si \quad \forall i, j \in I \quad A_i \cap A_j \neq \{\phi\}$

Le triplet $(\Omega; A; \mu)$ *s'appelle espace mesuré .*

Remarque 1.1.1 *Si* $\mu(\Omega) = 1$ *; la mesure dite probabilité noté par* P*, l'espace* $(\Omega; A; P)$ *s'appelle espace de probabilité.*

Définition 1.1.3 (Mesurabilité) *Soient* $(\Omega; A)$ *et* $(E; \zeta)$ *deux espaces mesurables, une application* $f : \Omega \to E$ *est dite mesurable par rapport à* $(E; \zeta)$ *:*

$$\forall B \in \zeta, si\ f^{-1}(B) \in A.$$

Définition 1.1.4 (Variable aléatoire) *Soit* (Ω, A, P) *un espace de probabilité on dit une variable aléatoire* X *tout fonction sur* Ω *dans* $\mathbb{R}$*.*

$$X : \Omega \to \mathbb{R}$$

Une variable aléatoire c'est une fonction mesurable.
Il y a deux types des variables aléatoires discrètes et continues.

Définition 1.1.5 (Echantillon) *Soit* X *une v.a sur un référentiel. Un échantillon de* X *de taille* $n, (X_1; ...; X_n)$ *de v.a et i.i.d de même loi que* X*. Une réalisation de cet échantillon est un n-uplet de rèels* $(x_1; ...; x_n)$ *où* $X_i(\omega) = x_i$*.*

Définition 1.1.6 (Fonction de répartition) *La fonction définie par :*

$$F_X : \begin{array}{l} \mathbb{R} \to [0,1] \\ x \to F_X(x) = P(X \leq x) \end{array}$$

est appelée fonction de répartition de X.

Proposition 1.1.1 *Soit F_X une fonction de répartition, alors :*

- *F_X est croissante.*
- *F_X est continue à droite et admet une limite à gauche en tout point x égale à $P(X \leq x)$.*
- $\lim_{t \to -\infty} F_X(t) = 0 \quad et \quad \lim_{t \to +\infty} F_X(t) = 1$.

1.1.2 Variable aléatoire continue

Définition 1.1.7 *On dira qu'une fonction f est une densité de probabilité si :*

$$\begin{cases} - \ f_X \text{ est continue sauf en un nombre dénombrable de points,} \\ - \ f_X \text{ positive,} \\ - \ \int_{\mathbb{R}} f_X(t)dt = 1 \end{cases}$$

Définition 1.1.8 *On appellera v.a continue X toute variable aléatoire à valeurs dans $\mathbb{R}$, telle que pour tout $x \in \mathbb{R}$*

$$P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

Remarque 1.1.2 *La loi d'une v.a. X est continue donnée par (au choix) :*

- *Sa densité f_X.*
- *Les probabilités $P(a \leq X \leq b)$ pour tout $a, b \in \mathbb{R}$.*
- *Sa fonction de répartition F_X.*

Proposition 1.1.2 *La fonction de répartition F d'une v.a. X continue de f est continue, croissante. Elle est dérivable en tout point x où f est continue et $F'_X(t) = f_X(t)$.*

Espérance et variance

– L'espérance est donnée par :

$$E\left(X\right) = \int_{-\infty}^{+\infty} X f_X(x) dt$$

– Le moment d'ordre 2 est donné par :

$$E\left(X^2\right) = \int_{-\infty}^{+\infty} X^2 f_X(x) dt$$

– La variance est donné par :

$$V(X) = \int_{-\infty}^{+\infty} (X - E(X))^2 f_X(x) dt$$

et aussi par :

$$V(X) = E\left(X^2\right) - \left(E\left(X\right)\right)^2$$

1.1.3 Variable aléatoire discrète

Définitions

Définition 1.1.9 (Fonction de répartition d'une v.a discrète) *Soit X une v.a, on appelle fonction de répartition de X la fonction définie par :*

$$F_X : \begin{array}{l} \mathbb{R} \to [0,1] \\ x \to F_X(t) = P(X \leq t) = \sum_{k \leq t}^{n} P(X = k) \end{array}$$

Pour une v.a. discrète, la fonction de répartition est une fonction en escalier, avec un saut en chaque valeur k de $X(\Omega)$ et la hauteur de ces sauts est la probabilité $P(X = k)$.

Espérance et variance d'une v.a. discrète

Définition 1.1.10 *Soit Ω un espace fini ou dénombrable, P une probabilité sur Ω, et X une variable aléatoire. On appelle espérance de X (ou moyenne de X) la quantité :*

$$E(X) = \sum_{i=1}^{k} kP\left(X = k\right)$$

Définition 1.1.11 *La variance et l'écart-type d'une v.a discrète X sont les réels positifs*

$$\begin{aligned} V(X) &= E\left(\left(X - E(X)\right)^2\right) = \sum_{k\in\Omega}(k - E(X))^2 P(X = k) \\ &= E\left(X^2\right) - \left(E\left(X\right)\right)^2 \end{aligned}$$

1.1.4 Lois de probabilités usuelles

Lois usuelles continues

Loi uniforme $U([a, b])$

Définition 1.1.12 *On dit qu'une v.a X suit une loi uniforme de paramètres a et b tels que $a, b \in \mathbb{R}$ et $a < b$.si sa densité est donnée par :*

$$f(x) = \begin{cases} \frac{1}{b-a} & si\ x \in [a, b] \\ 0 & sinon \end{cases}$$

Sa fonction de répartition est :

$$F(x) = \begin{cases} 0 & si\ x < a \\ \frac{x-a}{b-a} & si\ a \leq x < b \\ 1 & si\ x \geq b \end{cases} \tag{1.1}$$

Proposition 1.1.3 *Si $X \sim U[a,b]$, on na :*

$$E(X) = \frac{a+b}{2} \quad et \quad V(X) = \frac{(b-a)^2}{12}$$

Loi exponentielle $\zeta(\lambda)$

Définition 1.1.13 *On dit qu'une v.a X suit une loi exponentielle de paramètre λ telle que $\lambda > 0$.si sa densité est donnée par :*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$$

Sa fonction de répartirion est :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases} \tag{1.2}$$

Proposition 1.1.4 *Si $X \sim \zeta(\lambda)$, on na :*

$$E(X) = \frac{1}{\lambda} \quad et \quad V(X) = \frac{1}{\lambda^2}$$

Elle est utilisée dans de nombreuses applications :

– durée de fonctionnement d'un matériel informatique avant la première panne,

– désintégration radioactive,

– temps séparant l'arrivée de deux "clients" dans un phénomène d'attente (guichet, accès à un serveur informatique, arrivée d'un accident du travail...).

Proposition 1.1.5 *La loi exponentielle vérifie la propriété d'absence de mémoire.*

Soit X un v.a. de loi exponentielle $\zeta(\lambda)$, alors pour tout

$$s, t > 0, \qquad P(X > t + s / X > t) = P(X > s)$$

Loi normale $N(\mu, \sigma^2)$ C'est la loi la plus importante. Son rôle est central dans de nombreux modèles probabilistes et dans toute la statistique. Elle possède des propriétés intéressantes qui la rendent agréable à utiliser.

Définition 1.1.14 *On dit qu'une v.a X suit une loi normale de paramètres μ, σ^2.si sa densité f est donnée par :*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

et sa fonction de répartition F_X donnée par :

$$F_X(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

On peut écrire la fonction de répartition de loi normale sous la forme suivante (évaluation de distribution normale ; voir [6]) :

$$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{1.3}$$

Telle que $\operatorname{erf}(x)$ *est la fonction d'erreur.*

Définition 1.1.15 (Fonction d'erreur) *La fonction d'erreur, notée* $\operatorname{erf}(x)$*, est une fonction mathématique qui apparaît dans diverses applications de l'analyse mathématique, de la physique et de l'ingénierie. Elle est définie par :*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$$

La fonction erf *est utilisée pour calculer les probabilités dans la distribution normale, la fonction d'erreur peut être exprimée en termes de la fonction de répartition de la loi*

normale standard $\Phi(x)$*, comme suit :*

$$\text{erf}(x) = 2\Phi(x\sqrt{2}) - 1$$

où $\Phi(x)$ *est la fonction de répartition de loi normale standard (normale centrée réduite).*

Proposition 1.1.6 *Si* $X \sim N(\mu, \sigma^2)$*, on a :*

$$E(X) = \mu \quad et \quad V(X) = \sigma^2$$

Loi normale centrée réduite Si $\mu = 0$ et $\sigma^2 = 1$, on dit que X suit une loi normale centrée réduite $N(0, 1)$, si sa fonction de densité donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

et :

$$E(X) = 0 \quad \text{et} \quad V(X) = 1$$

Loi de type Pareto

Définition 1.1.16 *On dit qu'une v.a* X *suit une loi de type Pareto de paramètre* γ*, telle que* $\gamma > 0$ *si sa densité* f *est donnée par :*

$$f(x) = \frac{1}{\gamma} x^{-\frac{1}{\gamma}-1}$$

et sa fonction de répartition F_X *donnée par :*

$$F_X(x) = 1 - x^{-1/\gamma} \tag{1.4}$$

Proposition 1.1.7 *Si X suit une loi de pareto, on na :*

$$E(X) = \frac{1}{1-\gamma} \quad et \quad V(X) = \frac{1}{(1-\gamma)(1-2\gamma)}$$

Loi de cauchy $C(\varkappa_0, \alpha)$

Définition 1.1.17 *On dit qu'une v.a X suit une loi Cauchy de paramètres $\varkappa_0, \alpha$, telle que $\varkappa_0 \in \mathbb{R}, \alpha > 0, \forall x \in \mathbb{R}$ si sa densité f est donnée par :*

$$\begin{aligned} f(x) &= \frac{1}{\pi\alpha\left[1 + \left(\frac{x-\varkappa_0}{\alpha}\right)^2\right]} \\ &= \frac{1}{\pi\alpha}\left[\frac{\alpha^2}{(x-\varkappa_0)^2 + \alpha^2}\right] \end{aligned}$$

et sa fonction de répartition F_X donnée par :

$$F_X(t) = \frac{1}{\pi}\arctan\left(\frac{x-\varkappa_0}{\alpha}\right) + \frac{1}{2} \tag{1.5}$$

Remarque 1.1.3 *Si $X \sim C(\varkappa_0, \alpha), X$ n'a pas d'espérance ni de variance.car la fonction de densité n'étant pas intégrable sur $\mathbb{R}$.*

Lois usuelles discrètes Les lois discrètes sont caractérisées par leur fonction de probabilité, qui attribue des probabilités spécifiques à chaque valeur possible de la variable aléatoire. Ces lois permettent de modéliser différentes situations où les résultats sont déterminés de manière discrète et non continue.

Le tableau suivant 1.1.4 résumé quelque lois usuelles discrètes et leurs propriétés.

Distributions	Loi de probabilité	$E(X)$	$V(X)$	Fonction de réparition $F(x)$
Uniforme $U(\frac{1}{n})$	$p(X=k)=\frac{1}{n}$ si $k\in\{1,...,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$F(x)=\frac{K}{n}1_{[1;n[}+1_{[n;+\infty[}$
Bernoulli $B(p)$	$p(X=k)=p^k(1-p)^{1-k}$ $k\in[0,1]$	p	$p(1-p)$	$F(x)=(1-p)\,1_{[0;1[}+1_{[1;+\infty[}$
Binomiale $B(n;p)$	$p(X=k)=C_n^k p^k(1-p)^{n-k}$ $k\in\{0,...,n\}$	np	$np(1-p)$	$F(x)=\sum_{k=1}^{n} C_n^k p^k(1-p)^{n-k}$
Poisson $p(\lambda)$	$p(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ $k\in\{0,1,...\}$	λ	λ	$F(x)=\sum_{k=1}^{n}\frac{\lambda^k}{k!}e^{-\lambda}$
Géométrique $G(p)$	$p(X=k)=p(1-p)^{k-1}$; $k=1,2,..$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$	$F\,(X)=1-(1-p)^k$

TAB. 1.1 – Tableaux de lois usuelles discrètes

1.2 La transformée inverse

Dans cette section, nous présenterons la méthode d'inversion pour calculer la fonction d'inverse généralisée F^{-1} où la fonction de quantile Q.

1.2.1 Introduction

On suppose que l'on dispose d'un bon générateur de nombres pseudo-aléatoires et on se demande comment à partir d'une suite $(U_i)_{i\geq 1}$ de variables aléatoires i.i.d suivant la loi uniforme sur $[0,1]$ construire une variable aléatoire de loi donnée, avec une attention particulière pour les lois usuelles continues et discrètes.

1.2.2 Méthode d'inversion

Nous avons alors le lemme suivant, parfois connu sous le nom de transformée intégrale de probabilité, qui nous donne une représentation de toute variable aléatoire comme un transformée d'une variable aléatoire uniforme.

Fonction de répartition d'inverse généralisée

Définition 1.2.1 *Pour une fonction non décroissante F sur $\mathbb{R}$, l'inverse généralisé de F, F^{-1}, est la fonction définie par :*

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}; \forall x \in \mathbb{R}; \forall u \in]0; 1[$$

F^{-1} est aussi la fonction de quantile.$Q(u)$

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\}$$

1.2.3 Simulation des variables aléatoires continues

Lemme 1.2.1 *Supposons que la v.a. X a pour fonction de répartition F continue et strictement croissante, avec $0 < F(x) < 1$.*
Soit U une v.a.tel que $U \rightarrow U[0, 1]$. Alors, la v.a .$F^{-1}(U)$ a pour fonction de répartition F.

Preuve. Soit G la fonction de répartition de $F^{-1}(U)$.
Alors :

$$\begin{aligned} G(x) &= P(F^{-1}(U) \leq x) \\ &= P(F(F^{-1}(U)) \leq F(x)).(\text{Par la monotonie de } F) \\ &= P(U \leq F(x)) = F(x) \ (\text{Car } U \rightarrow U(0, 1)) \end{aligned}$$

■

Algorithme Cette méthode suggère que pour générer des échantillons d'une v.a. X pour laquelle F^{-1} est connue, on peut générer des nombres aléatoires U uniformes sur $[0, 1]$ et faire $X = F^{-1}(U)$. Nous avons alors l'algorithme général d'inversion suivant :

Générer $U \rightarrow U[0,1]$
Faire $X = F^{-1}(U)$
Sortir X

TAB. 1.2 – Algorithme général d'inversion d'une v.a continue

Remarque 1.2.1 *Une condition minime pour l'application de cette méthode est de connaître la forme explicite de F^{-1}. Cela est vérifié pour plusieurs lois de probabilités, comme l'uniforme, l'exponentielle, de Weibull, de Cauchy, ...*
Remarquons qu'une telle condition n'est pas suffisante, par exemple, pour la loi beta, il est possible théoriquement de la simuler par inversion, mais elle peut résulter très couteuse. Parfois, nous disposons d'une bonne approximation de F^{-1}, d'où on peut utiliser la méthode par approximation.

1.2.4 Simulation des variables aléatoires discrètes

Fonction d'inverse d'une loi de probabilité discrète :

Définition 1.2.2 *On définit la fonction F^{-1}*

$$F^{-1}(u) = \begin{cases} x_1 \quad si \quad 0 < u \leq F(x_1) = p_1 \\ x_2 \quad si \quad F(x_1) < u \leq F(x_2) = p_1 + p_2 \\ x_3 \quad si \quad F(x_2) < u \leq F(x3) = p_1 + p_2 + p_3 \\ \\ \\ \\ x_n \quad si \quad F(x_{n-1}) < u \leq F(x_n) \end{cases}$$

$$F^{-1}(u) = \begin{cases} x_1 & si \quad 0 < u \leq p_1 \\ x_2 & si \quad p_1 < u \leq p_1 + p_2 \\ x_3 & si \quad p_1 + p_2 < u \leq p_1 + p_2 + p_3 \\ & \\ & \\ & \\ x_n & si \quad p_1 + p_2 + ... + p_{n-1} < u \leq 1 \end{cases}$$

Théorème 1.2.1 *Soit :* $F^{-1}(u) = \inf\{x : u \leq F(x)\}$

Si $U \rightarrow U([0,1])$*, alors la v.a* $.X = F^{-1}(u)$ *a pour fonction de répartition* F*.*

Preuve. Remarquons le minimum est atteint parce que F est continue à droite, alors F^{-1} est bien définie.

En plus, $F\left(F^{-1}(u)\right) \geq u$

et $F^{-1}\left(F(x)\right) = \inf\{y : F(y) \geq F(x)\} \leq x$

D'où l'égalité des ensembles : $\{(x;u) : F^{-1}(u) \leq x\} = \{(u;x) : u \leq F(x)\}$

et de probabilités : $P(X \leq x) = P\left(F^{-1}(u) \leq x\right) = P\left(U \leq F(x)\right) = F(x)$ ■

Algorithme Pour une loi discrète générale, on a $F^{-1}(u) = i$ avec $F_{i-1} < u \leq F_i$, donc la méthode d'inversion est équivalente à chercher l'indice i convenable dans la liste des F_i.

En général :

$$F^{-1}(u) = x_i \quad \text{si} \quad F^{-1}(x_{i-1}) = F_{i-1} \leq u < F_i = F(x_i)$$

Nous avons alors l'algorithme général d'inversion suivant :

Générer $U \rightarrow U(0,1)$
Tant que $F_i \leq U$, faire $i = i + 1$
Sortir $X = i$

TAB. 1.3 – Algorithme général d'inversion d'une v.a discrète

1.2.5 Simulation de quelque lois

Nous allons présenter la simulation de quelques lois dans le but de les mettre en ouvre au chapitre deux et trois dans l'intégralité de notre nouvel algorithme d'ajustement de loi inconnue.

Simulation d'une v.a. $U(a, b)$

La fonction de répartition d'une v.a de loi uniforme 1.1.

Selon l'algorithme général, précité, il suffit de faire

$$\begin{aligned} Q(u) &= F^{-1}(U) \\ &= a + (b - a)U \end{aligned}$$

Simulation d'une v.a $\zeta(\lambda)$

La fonction de répartition d'une v.a de loi exponentielle 1.2.

Selon l'algorithme général, il suffit de faire :

$$\begin{aligned} Q(u) &= F^{-1}(U) \\ &= -\frac{1}{\lambda}\log(1 - U) \end{aligned}$$

Simulation d'une v.a $N(\mu; \sigma^2)$

Selon fonction de répartition d'une v.a de loi normale 1.3.

Selon l'algorithme général, il suffit de faire :

$$\begin{aligned} Q(u) &= F^{-1}(U) \\ &= \mu + \sigma\sqrt{2}\,\mathrm{erf}^{-1}(2U - 1) \end{aligned}$$

Simulation d'une v.a $P(\gamma)$

La fonction de répartition d'une v.a de loi Pareto 1.4.

Selon l'algorithme général, il suffit de faire :

$$\begin{aligned} Q(u) &= F^{-1}(U) \\ &= -\gamma \log(1-U) \end{aligned}$$

Simulation d'une v.a $C(\varkappa_0, \alpha)$

La fonction de répartition d'une v.a de loi Cauchy 1.5.

Selon l'algorithme général, il suffit de faire :

$$\begin{aligned} Q(u) &= F^{-1}(U) \\ &= \varkappa_0 + \alpha \tan\left(\pi\left(U - \frac{1}{2}\right)\right) \end{aligned}$$

1.3 Convergence de suites de variables aléatoires

Dans cette section, nous rappelons les convergences de lois de probabilités et différentes types de convergence avec des exemples.

1.3.1 Les différents types de convergence

Une suite (X_n) de variables aléatoires étant une suite de fonctions de $\mathbb{R}$ dans Ω il existe diverses façons de définir la convergence de (X_n) dont certaines jouent un grand rôle en calcul des probabilités.

La convergence en probabilité

Définition 1.3.1 *La suite* (X_n) *converge en probabilité vers la constante* a *si,* $\forall\ \varepsilon$ *et* η *(arbitrairement petits), il existe* n_0 *tel que* $n > n_0$ *entraîne :*

$$P(|X_n - a| > \varepsilon) < \eta$$

On note alors : $(X_n) \xrightarrow{p} a$

On définit alors la convergence en probabilité vers une v.a X comme la convergence vers 0 de la suite $X_n - X$.

Lorsque $E(X_n) \to a$, il suffit de montrer que $V(X_n) \to 0$ pour établir la convergence en probabilité de X_n vers a. En effet, d'après l'inégalité de **Bienaymé-Tchebycheff.**

$$P(|X_n - E(X_n)| > \varepsilon) < \frac{V(X_n)}{\varepsilon^2}$$

On en déduit donc sans difficulté que $X_n - E(X_n) \xrightarrow{p} 0$, ce qui établit le résultat.

La convergence presque sûre ou convergence forte

Définition 1.3.2 X *et* Y *sont égales presque sûrement si* $P(\{\omega : X_n(\omega) \neq y(\omega)\}) = 0$.

C'est l'égalité presque partout des fonctions mesurables. On définit donc ainsi des classes de v.a presque sûrement égales.

Définition 1.3.3 (La convergence presque sûre) *La suite* (X_n) *converge presque sûrement vers* X *si :* $P(\{\omega : X_n(\omega) \neq X(\omega)\}) = 0$

et on note : $X_n \xrightarrow{ps} X$.

En d'autres termes, l'ensemble des points de divergence est de probabilité nulle.

Remarque 1.3.1 *La convergence presque sûremant implique la convergence en probabilité.*

La convergence en moyenne d'ordre p

Définition 1.3.4 *$(X_n) \to X$ en moyenne d'ordre p si $E\left[|X_n - X|^p\right] \to 0$.*

Remarque 1.3.2 *La plus utilisée est la convergence en moyenne quadratique si $p = 2$. La convergence en moyenne d'ordre p implique la convergence en probabilité.*

La convergence en loi

Bien quel est plus la faible, elle est très utilisée en pratique car elle permet d'approximer la fonction de répartition de X_n par celle de X.

Définition 1.3.5 *La suite (X_n) converge vers la variable X de fonction de répartition F si, en tout point de continuité de F, la suite (F_n) des fonctions de répartition des X_n converge vers F.*
on note : $X_n \xrightarrow{\mathcal{L}} X$.

Un théorème dû à Polya établit que si F est continue alors la convergence est uniforme.
Pour des variables discrètes, la convergence en loi vers une v.a discrète s'exprime par

$$P(X_n = x) \to P(X = x)$$

C'est ainsi qu'on a établi la convergence de la loi binomiale vers la loi de Poisson.
Une suite de variables discrètes peut cependant converger en loi vers une variable continue.
On a également que, Si x_n est une suite de variables de densités. F_n et X une variable de densité f, alors :

$$X_n \xrightarrow{\mathcal{L}} X \Longrightarrow f_n(x) \to f(x) \quad \forall x \in \mathbb{R}$$

La convergence en loi est intimement liée à la convergence des fonctions caractéristiques comme le précise le résultat fondamental suivant, que nous énoncerons sans démonstration :

Théorème 1.3.1 (Levy-cramer-dugué) *Si $X_n \xrightarrow{\mathcal{L}} X$ alors $\varphi_{X_n}(t) \to \varphi_X(t)$ uniformément dans tout intervalle fini. Si la suite des fonctions caractéristiques $\varphi_{X_n}(t)$, converge*

vers une fonction φ dont la partie réelle est continue à l'origine, alors φ est une fonction caractéristique et la suite X_n converge en loi vers une v.a X dont φ est la fonction caractéristique.

La convergence en probabilité entraîne la convergence en loi.
Nous présentons dans ce qui suit la convergence de la loi poisson et la loi binomiale vers la loi normale en détail dans le but de tester la stabilité de notre nouvel algorithme.

1.3.2 Convergence en loi de la loi Binomiale vers la loi normale

Théorème 1.3.2 (théorème de Moivre-Laplace) *X_n étant une suite de variables Binomiales $B(n,p)$, alors $\dfrac{X_n - np}{\sqrt{npq}} \xrightarrow{\mathcal{L}} N(0,1)$ et $q = 1-p$.*

Preuve. *La fonction caractéristique de X_n vaut $(p\exp(it)+1-p)^n$ donc celle de $\frac{X_n-np}{\sqrt{npq}}$ vaut :*

$$\varphi(t) = \left(p\exp\left(\frac{it}{\sqrt{npq}}\right)+1-p\right)^n \exp\left(-\frac{itnp}{\sqrt{npq}}\right)$$

$$\ln\varphi(t) = n\ln\left(p\left(\exp\left(\frac{it}{\sqrt{npq}}\right)-1\right)\right) - \frac{itnp}{\sqrt{npq}}$$

Développons au deuxième ordre l'exponentielle, il vient :

$$\ln\varphi \simeq n\ln\left(1+p\left(\frac{it}{\sqrt{npq}} - \frac{t^2}{2npq}\right)\right) - \frac{itnp}{\sqrt{npq}}$$

Puis le logarithme :

$$\ln\varphi \simeq n\left[\frac{pit}{\sqrt{npq}} - \frac{pt^2}{2npq} + \frac{p^2t2}{2nppq}\right] - \frac{itnp}{\sqrt{npq}}$$

Soit :

$$\ln\varphi \simeq -\frac{t^2}{2q} + \frac{pt^2}{2q} = \frac{t^2}{2q}(p-1) = -\frac{t^2}{2}$$

$\varphi(t) \to \exp(-t^2/2)$ *qui est la fonction caractéristique de loi normale cntrée réduite.* ■

Application :Lorsque n est assez grand, on peut donc approximer la loi binomiale par la loi de Gauss. On donne généralement comme condition np et $nq > 5$.

Il convient cependant d'effectuer ce que l'on appelle la correction de continuité : la convergence de la loi binomiale vers la loi de Gauss se traduit par le fait que les extrémités des bâtons du diagramme de la loi binomiale $B(n,p)$. sont voisines de la courbe de densité de la loi $N(np, \sqrt{npq})$.

On obtient donc une valeur approchée de $P(X = x)$ par la surface sous la courbe de densité comprise entre les droites d'abscisse $x - \frac{1}{2}$ et $x + \frac{1}{2}$ (Figure 1.1)

$$p(X = x) \simeq p\left(\frac{x - \frac{1}{2} - np}{\sqrt{npq}} < U < \frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right)$$

On aura alors :

$$P(X \leq x) \simeq P\left(U < \frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right)$$

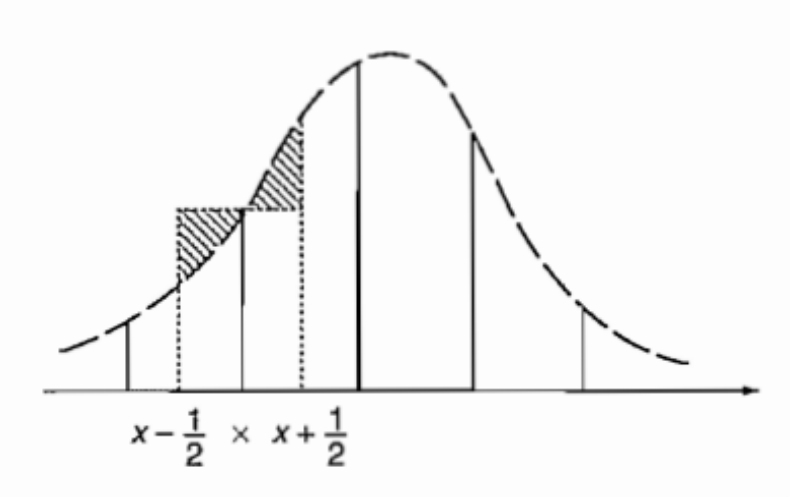

FIG. 1.1 – L'approximation la loi de binomiale par la loi de Gauss

1.3.3 Convergence de la loi de Poisson vers la loi de loi normale

Théorème 1.3.3 *Soit* (X_λ) *une famille de variables* $p(\lambda)$ *alors si* $\lambda \to \infty$, $\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \xrightarrow{\mathcal{L}} N(0,1)$

Preuve.

$$\varphi(t) = \exp(\lambda)\left(\exp(it-1)\right)$$

d'où :

$$\begin{aligned}\frac{\varphi_{X-\lambda}(t)}{\sqrt{\lambda}} &= \exp\left(\lambda(\exp\left(\frac{it}{\sqrt{\lambda}}\right)-1)\right)\exp\left(-\frac{it\lambda}{\sqrt{\lambda}}\right)\\ &= \exp\left(\lambda\exp\left(\frac{it}{\sqrt{\lambda}}\right)-\lambda-it\sqrt{\lambda}\right)\end{aligned}$$

comme :

$$\exp\left(\frac{it}{\sqrt{\lambda}}\right) \simeq 1 + \frac{it}{\sqrt{\lambda}} - \frac{t^2}{2\lambda}$$

il vient :

$$\frac{\varphi_{X-\lambda}(t)}{\sqrt{\lambda}} \simeq \exp\left(\lambda + it\sqrt{\lambda} - \frac{t^2}{2} - \lambda - it\sqrt{\lambda}\right) = \exp(-\frac{t^2}{2})$$

■

La figure 1.2 illustre l'approximation de la loi de Poisson $p(\lambda)$ par la loi de Gauss de même espérance λ et de même écart-type $\sqrt{\lambda}$.

L'approximation est très satisfaisante pour $\lambda > 18$. On trouvera en [10] d'autres formules d'approximation plus précises. On a, ici encore, intérêt à effectuer la correction de continuité.

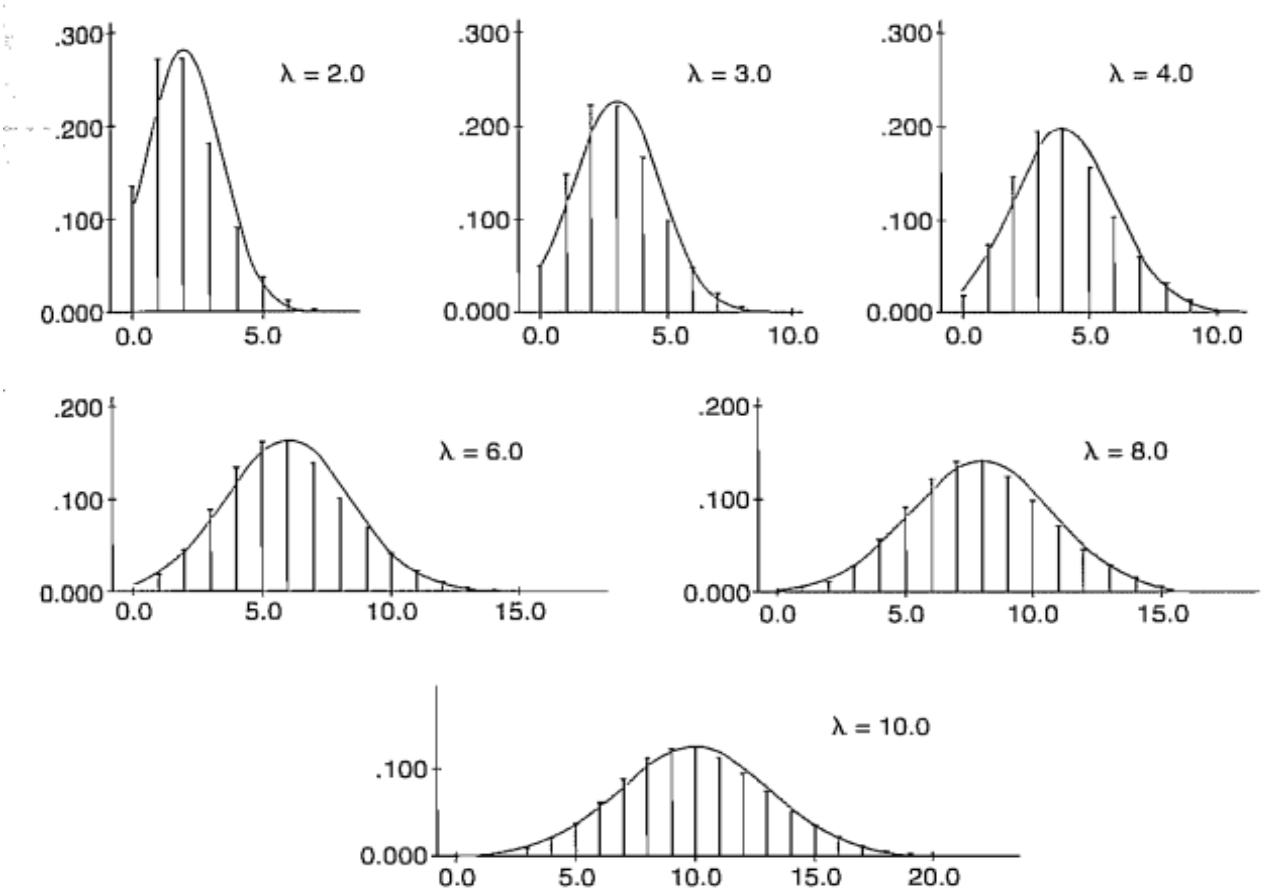

FIG. 1.2 – L'approximation de la loi de Poisson par la loi de Gauss

1.3.4 Théorème de central-limite (TCL)

L'étude de sommes de variables indépendantes et de même loi joue un rôle capital en statistique.

Le théorème suivant connu sous le nom de théorème central-limite (il vaudrait mieux dire théorème de la limite centrée) établit la convergence vers la loi de Gauss sous des hypothèses peu contraignantes.

Théorème 1.3.4 *Soit* (X_n) *une suite de variables aléatoires indépendantes de même loi*

d'espérance μ et d'écart-type σ Alors :

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma}\right) \xrightarrow{\mathcal{L}} N(0,1)$$

Preuve.

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma}\right) = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$$

Soit $\varphi_X(t)$ la fonction caractéristique de X ; la fonction caractéristique de $\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$ est donc égale à $\left[\varphi_{\frac{X-\mu}{\sigma\sqrt{n}}}(t)\right]^n$.

Le développement en série de la fonction caractéristique de $\frac{X-\mu}{\sigma\sqrt{n}}$ commence par $1 - \frac{t^2}{2n}$ les termes suivants sont des infiniment petits d'ordre $1/n^2$.

Donc, en élevant à la puissance n, la fonction caractéristique de $\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$ est équivalente à $(1 - \frac{t^2}{2n})^n$ et tend si $n \to \infty$ vers $\exp(-\frac{t^2}{2})$ selon un résultat classique.

■

On remarque que, si les variables X_i sont des variables de Bernoulli, on retrouve comme cas particulier la convergence de la loi binomiale vers la loi de Gauss.

Chapitre 2

Outils statistiques de modélisation des lois de probabilités usuelles

L'un des concepts fondamentaux en statistique est la loi de probabilités. Une loi de probabilités décrit la distribution des valeurs possibles d'une variable aléatoire. La modélisation des lois de probabilités usuelles vise à trouver des distributions statistiques qui correspondent le mieux aux données observées. Dans ce contexte, nous nous intéressons dans ce chapitre à présenter quelques outils statistiques qui seront utiles pour le développement de notre nouvel algorithme d'ajustement de données observées par des lois usuelles continues, respectivement, loi Exponentielle, loi Normale, loi Uniforme continue, loi de type Pareto et loi de Cauchy.

Nous avons utilisé dans un premier temps le QQ-plot qui est un outil d'ajustement graphique pour les lois précitées en calculant le coefficient de détermination de la régression R^2 basé sur la droite de régression entre les quantiles théoriques et les quantiles empiriques ce qui nous permet un choix meilleur parmi les lois candidates, et nous finissons par estimer les paramètres de cette dernière par la méthode de maximum de vraisemblance.

Ces trois outils, nous ont offert un ajustement quasi-complet des données observées par une loi reste à confirmer au chapitre trois par notre nouvel algorithme.

2.1 Régression linéaire simple

Dans cette section, nous concentrons sur les modèles de régression linéaire simple et leurs propriétés avec le test de signification du modèle.

2.1.1 Introduction

Le terme « régression » et les méthodes pour étudier les relations entre deux variables peut remonter à environ 100 ans. C'était d'abord introduit par Francis Galton en 1908, le célèbre biologiste britannique, quand il était engagé dans l'étude de l'hérédité. Une de ses observations était que les enfants de parents de grande taille sont plus grands que la moyenne mais pas aussi grands que leurs parents. Cette « régression vers la médiocrité » a donné leur nom à ces méthodes statistiques. Le terme régression et son évolution décrivent principalement relations statistiques entre les variables. En particulier, la régression simple est la méthode de régression pour discuter de la relation entre un dépendant variable (y) et une variable indépendante (x).

Illustration 2.1.1 Les données classiques suivantes contient les informations sur la taille des parents et la taille des enfants :

parents	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5
enfants	65.8	66.7	67.2	67.6	68.2	68.9	69.5	69.9	72.2

TAB. 2.1 – Taille des parents et des enfants

La taille moyenne est de $68,44$ pour les enfants et de $68,5$ pour les parents. La régression ligne pour les données des parents et des enfants peut être décrite comme.

$$\text{Taille des enfants} = 21.52 + 0.69 * \text{Taille des parents}$$

Pour une exécution illustrative de cette dernière sous R voir (page 30).

2.1.2 Modèle de régression linéaire simple

Définition 2.1.1 (Régression linéaire simple) *Le modèle de régression linéaire simple est généralement énoncé sous la forme :*

$$y = \beta_0 + \beta_1 x + \varepsilon$$

où

Y est la variable dépendante (variable de réponse), β_0 et β_1 sont les coefficients (ordonnée à l'origine et pente), x est la variable indépendante (variable explicative) et ε est l'erreur aléatoire.

Définition 2.1.2 *L'expérience typique pour la régression linéaire simple est que nous observons n paires de données $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ à partir d'une expérience scientifique, et modèle en termes de n paires de données peut être écrit comme :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \forall i \in [1; n] \tag{2.1}$$

où $E(\varepsilon_i) = 0$ et $var(\varepsilon_i) = \sigma^2$, et tout ε_i sont indépendantes.

Maintenant, nous trouvons à bon estimation de β_0 et β_1 pour le modèle de régression linéaire simple.

2.1.3 Estimation des paramètres de régression

Les paramètres β_0 et β_1 sont inconnus, donc on a utilisé la méthode des moindres carrés pour estimer β_0 et β_1. Autrement dit, nous estimons β_0 et β_1 de sorte que la somme des carrés des différences entre les observations y_i et la droite est un minimum. Le critère des moindres carrés est donnée par :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

On résoud le système à deux inconnues $\nabla S(\beta_0, \beta_1)$:

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0}\Big|_{\widehat{\beta}_0,\widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_{i-}\widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \\ \dfrac{\partial S}{\partial \beta_1}\Big|_{\widehat{\beta}_0,\widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_{i-}\widehat{\beta}_0 - \widehat{\beta}_1 x_i)x_i = 0 \end{cases}$$

Ce qui est équivalent à :

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \tag{2.2}$$

Les équations 2.2 sont appelées les équations normales **des moindres carrés.** La solution de ces équations est :

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

et

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$= \frac{n\overline{XY} - \sum_{i=1}^{n} y_i x_i}{n\overline{X}^2 - \sum_{i=1}^{n} x_i^2}$$

Où :

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{et} \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sont les moyennes de y_i et x_i. Donc, $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont les estimateurs de β_0 et β_1 par la méthode des moindre carrées. Alors :

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Théorème 2.1.1 – $E(\widehat{\beta}_0) = \beta_0$ *et* $E(\widehat{\beta}_1) = \beta_1$ *(Estimateurs sans biais)*

– $var(\widehat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n x_i^2 - n\overline{X}^2} \right]$ *et* $var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\overline{X}^2}$

– $cov[\widehat{\beta}_0, \widehat{\beta}_1] = -\frac{\sigma^2 \overline{X}}{\sum_{i=1}^n x_i^2 - n\overline{X}^2}$

2.1.4 Test de signification du modèle (analyse de la variance)

Pour évaluer la qualité d'ajustement du modèle par le coefficient de détermination R^2 nous utilisons l'équation d'**analyse de la variance** donnée par :

$$ScT = ScE + ScR$$

D'où :

ScT : somme carrée totale

$$\sum_{i=1}^{n} (y_i - \overline{y})^2$$

ScE : somme carrée expliquée

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

ScR : somme carrée résiduelle

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Donc la qualité du modèle est jugée par le coefficient de détermination de la régression R^2.

Définition 2.1.3 (Coefficient de détermination) *Le coefficient de détermination R^2 est défini par :*

$$R^2 = \frac{ScE}{ScT} = 1 - \frac{ScR}{ScT}$$

Cette quantité implique les pourcentages de Y expliquée par X, par exemple, si $R^2 = 0,4$ cela veut dire 40% des Y seulement sont expliqués par X et 60% sont expliqués par d'autre variables aléatoire. Si R^2 est proche de 1 cela veut dire un ajustement meilleur.

Remarque 2.1.1 *Comme nous allons se limiter dans notre nouvel algorithme à calculer le coefficient de détermination R^2 pour un meilleur ajustement parmi les lois intégrées dans notre programme d'exécution, sans prendre en considération l'estimation des coefficients de la droite de la régression (β_0, β_1) par la méthode des moindre carrées (dans notre cas ça serait une estimation empirique de β_0 et β_1).*
Pour cela nous exécutons uniquement le code qui donne R^2.

Code illustratif sous R

D'aprés l'illustration 2.1.1

Supposons que vous ayez deux vecteurs de données x et y, telle que :

x : Taille des parents.

y :Taille des enfants.

x<-c(64.5, 65.5, 66.5, 67.5, 68.5, 69.5, 70.5, 71.5, 72.5) # Créer des vecteurs de données x.

y<-c(65.8, 66.7, 67.2, 67.6, 68.2, 68.9, 69.5, 69.9, 72.2) # Créer des vecteurs de données y.

model <- lm(y ~x) # Ajuster un modèle de régression linéaire.

plot(x,y,xlab="Taille les parents",ylab="Taille les enfants") # Tracer le graphe entre de x et y

abline(coef(model), col="red", lwd="3") # Tracer la droite de régression

Rcarré<-summary(model)$r.squared # Extraire le coefficient de détermination (R^2) du modèle

Rcarré

[1] 0.9415187

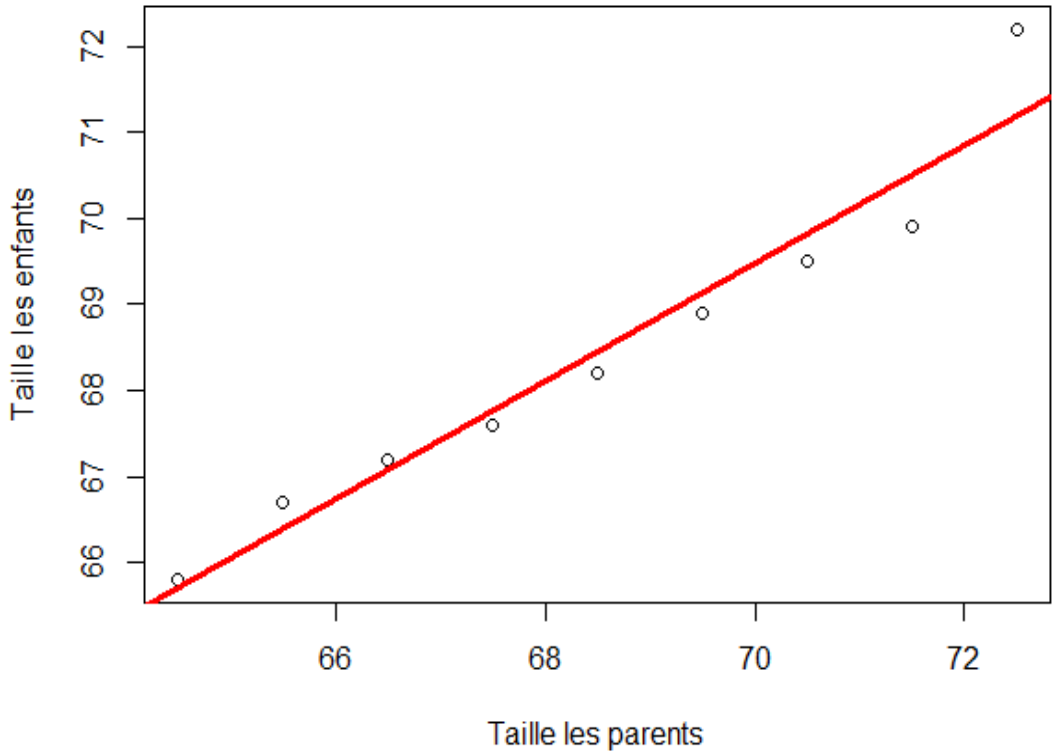

FIG. 2.1 – Graphe expliquons la relation linéaire entre la taille des enfants et leurs parents

Remarque 2.1.2 $R^2 = 0.94$ *cela veut dire* 94% *des* Y *sont expliquées par* X.

La droite de régression est proche de la plupart des points de nuage.

2.2 Quantile-quantile plot (QQ-plot)

2.2.1 Introduction

Le QQ-plot (Quantile-Quantile plot) est une technique graphique apparue dans les années 1960. Elle est couramment utilisée de manière informelle pour déterminer si un échantillon aléatoire univarié de taille n provient d'une distribution spécifiée F. La méthode consiste à tracer les quantiles empirique de l'échantillon par rapport aux quantiles théoriques de

la distribution F, puis à effectuer une vérification visuelle pour déterminer si les points présentent une allure linéaire. La linéarité dans la courbe peut être facilement vérifiée à l'œil et peut également être quantifiée à l'aide d'un coefficient de détermination.

2.2.2 Ajustement par QQ plot

Cette méthode consiste en la comparaison des quantiles empiriques et des quantiles théoriques. Soient $F(x) = P(X \leq x)$ la fonction de répartition de X et x_p le quantile d'ordre p définie par

$$x_p = \inf\{x \in \mathbb{R}; F(x) \geq p\} \quad \text{et} \quad p \in [0,1]$$

Soient $x_{(1)}, x_{(2)}, ..., x_{(n)}$ les données rangées par ordre croissant. Alors on peut écrire la fonction de répartition empirique comme

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}} \\ &= \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ \frac{i}{n} & \text{si} \quad x_{(i)} \leq x < x_{(i+1)}, \ i \in \{1, ..., n-1\} \\ 1 & \text{si} \quad x \geq x_{(n)} \end{cases} \end{aligned}$$

Soit $(p_1, p_2, ..., p_n)$ une suite strictement croissante de n réels vérifiant :

$$\frac{i-1}{n} < p_i < \frac{i}{n}$$

On appelle QQ plot le nuage de points N dans le repère orthonormé (O, I, J) défini par :

$$N = \{(x_{p_1}, x_{(1)}), (x_{p_2}, x_{(2)}), ..., (x_{p_n}, x_{(n)})\}.$$

Si X suit une loi F, les données font que F_n est une bonne estimation de F et, a fortiori, $x_{(i)}$ doit bien estimer x_{p_i} .$x_{(i)} \simeq x_{p_i}$ pour tout $i \in \{1, ..., n\}$; les points du nuage N

doivent être proche de la "**droite diagonale**" d'équation : $y = x$.

2.2.3 Construction de QQ-plot

Dans ce qui suit, un choix très pratique de valeurs de p est donnée par :

$$p_{i,n} \in \left\{ \frac{1}{n+1}, \frac{2}{n+1}, ..., \frac{n}{n+1}, 1 \right\}.$$

QQ-plot d'une distribution exponentielle $\xi(\lambda)$

La fonction de répartition générale est :

$$F_\lambda(x) = 1 - e^{-\lambda x}, x > 0.$$

la distribution exponentielle standard :

$$F_1(x) = 1 - e^{-x}, x > 0.$$

La fonction quantile de la distribution exponentielle a la forme simple :

$$Q_\lambda(p) = -\frac{1}{\lambda} \log(1-p) \quad \text{pour} \quad p \in [0,1].$$

On a de la même façon.

$$Q_1(p) = -\log(1-p) \quad \text{pour} \quad p \in [0,1].$$

En conséquence, il y a une relation linéaire simple entre les quantiles théoriques d'une distribution exponentielle et les quantiles d'une distribution exponentielle standard.

$$Q_1(p) = \lambda Q_\lambda(p) \quad \text{pour} \quad p \in [0,1].$$

À partir de notre échantillon $X_1, X_2, ..., X_n$, nous estimons la fonction quantile Q en uti-

lisant le quantile empirique Q_n. Dans un système de coordonnées orthogonales, les points représentent les valeurs des quantiles empiriques (Q_n) en fonction des quantiles théoriques correspondants (Q).

$$(Q_n(p), -\log(1-p))$$

Nous traçons le graphique pour différentes valeurs de $p \in [0,1]$. Nous nous attendons à ce qu'un motif de ligne droite apparaisse dans le nuage de points si le modèle exponentiel permet un ajustement statistique plausible pour la population donnée. Lorsqu'un motif de ligne droite est observé. En effet, si le modèle est correct, alors l'équation :

$$(-\log(1-p)) = \lambda Q_\lambda(p) \tag{2.3}$$

est valable. Remarquons que le point d'intersection pour le modèle donné doit être égal à 0 quand $Q_\lambda(0) = 0$.

En général

$$Q_n(p) = X_{i,n} \quad \text{pour} \quad \frac{i-1}{n} < p < \frac{i}{n}$$

Pour obtenir finalement, les coordonnées

$$(X_{i,n}, -\log(1-p_{i,n}))$$

Procédure sous R et illustration graphique

Pour avoir le graphique QQ-plot des données $x_1, x_2, ..., x_n$ contre les quantiles exponentielles, on exécute le programme suivant sous R :

x<- .# Entrer l'échantillon à étudier des observations x_i

n<-length(x)

x<-sort(x) # On ordonne l'échantillon des observations x_i

p<-(1 :n)/(n+1) # Vecteur des valeurs i/(n+1)

Droite<-lm((-log(1-p))~x) # Ajuster un modèle de régression linéaire

coef(Droite) # Estimation du paramètres

qqplot(x,-log(1-p), xlab="Quantiles empiriques ", ylab="Quantiles thèoriques", main = "QQ-plot de la loi Exponentielle") # Tracer le qqplot

abline(coef(Droite), col="3", lwd="3") # Tracer la droite de régression

RcExp<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de loi Exponentielle

Cas pratique 2.2.1

Dans une station-service, un jour donné, les durées du passage en caisse de 50 clients ont été mesurés. Les résultats en secondes sont :

0.96	1.45	0.42	3.69	2.58	1.95	1.74	0.01	1.02	1.12
0.17	3.19	0.85	1.27	0.68	3.60	1.23	0.34	0.31	0.16
0.07	0.79	0.02	1.20	0.05	2.09	0.24	5.46	2.57	0.89
0.74	1.67	0.88	2.27	0.22	3.39	0.12	0.06	0.78	0.32
5.79	2.09	0.39	1.82	2.96	0.20	0.08	0.37	2.58	0.30

Soit X la v.a égale à la durée de passage d'un client. On s'interroge sur le fait que X suit ou non une loi exponentielle.

On trace le graphe de qq-plot de x selon le programme de qq-plot de loi exponentielle :

x<-c(0.96, 1.45, 0.42, 3.69, 2.58, 1.95, 1.74, 0.01, 1.02, 1.12, 0.17, 3.19, 0.85, 1.27, 0.68, 3.60, 1.23, 0.34, 0.31, 0.16, 0.07, 0.79, 0.02, 1.20, 0.05, 2.09, 0.24, 5.46, 2.57, 0.89, 0.74, 1.67, 0.88, 2.27, 0.22, 3.39, 0.12, 0.06, 0.78, 0.32, 5.79, 2.09, 0.39, 1.82, 2.96, 0.20, 0.08, 0.37, 2.58, 0.30)

Le résultat est donnée comme suivante :

> RcExp

[1] 0.9892635

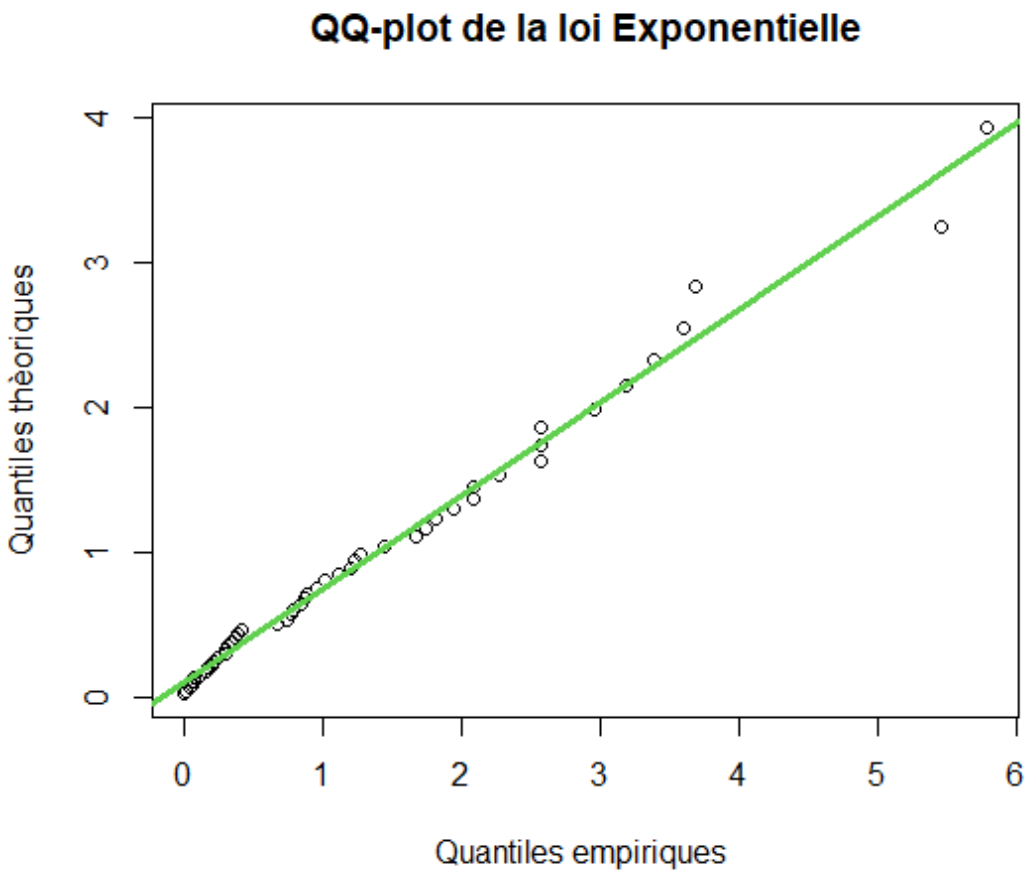

FIG. 2.2 – QQ-plot de la loi exponentielle

La droite diagonale $y = x$ ajuste bien le nuage de points ; on peut envisager que X suit une loi exponentielle.

QQ-plot d'une distribution uniforme continue $U[a, b]$

Nous présentons maintenant le QQ-plot d'une distribution uniforme continue de paramètres a et b..

$$F_{a,b}(x) = \frac{x - a}{b - a}$$

Ainsi pour $a = 0$ et $b = 1$:

$$F_{0,1}(x) = x$$

La fonction quantile de la distribution uniforme a la forme simple :

$$Q_{a,b}(p) = a + (b-a)p \quad \text{pour } p \in [0,1],$$

et de la même manière on a :

$$Q_{0,1}(p) = p \quad \text{pour } p \in [0,1].$$

Ainsi, il existe une relation linéaire directe entre les quantiles théoriques d'une distribution uniforme et les quantiles empiriques d'une distribution uniforme.

$$Q_{0,1}(p) = \frac{-a}{(b-a)} + \frac{1}{(b-a)} Q_{a,b}(p) \quad \text{pour } p \in [0,1].$$

De la même façon, on estime Q par Q_n. Dans un système de coordonnées orthogonale, les points de valeurs

$$(Q_n(p), p),$$

sont tracer pour différentes valeurs de p dans l'intervalle $[0,1]$, nous traçons le graphe de qq-plot. Nous anticipons l'apparition d'un schéma linéaire dans le nuage de points, indiquant un ajustement statistique plausible pour la population statistique considérée. Lorsqu'un schéma linéaire est observé, cela confirme que l'équation

$$Q_{0,1}(p) = \frac{-a}{(b-a)} + \frac{1}{(b-a)} p,$$

est valable.

En général,

$$Q_n(p) = X_{i,n} \quad \text{pour} \quad \frac{i-1}{n} < p < \frac{i}{n}$$

Pour obtenir finalement, les coordonnées

$$(X_{i,n}, p_{i,n})$$

Procédure sous R et illustration graphique

Pour avoir le graphique QQ-plot des données $x_1, x_2, ..., x_n$ contre les quantiles uniforme continue, on exécute le programme suivant sous R :

x<- . # Entrer l'échantillon à étudier des observations x_i

n<-length(x)

x<-sort(x) # On ordonne l'échantillon des observations x_i

p<-(1 :n)/(n+1) # vecteur des valeurs i/(n+1)

Droite<-lm(p~x) # Ajuster un modèle de régression linéaire

coef(Droite) # Estimation du paramètres

qqplot(x,p, xlab="Quantiles empiriques ", ylab="Quantiles thèoriques", main = " Q-Q plot de la loi Uniforme continue")

abline(coef(Droite), col="3", lwd="3") # Tracer la droite de regression.

RcUnif<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de loi Uniforme continue.

Cas pratique 2.2.2

L'échantillon suivant représente une série de 20 valeurs numériques réelles générées de manière aléatoire.

0.562	0.206	0.781	0.912	0.395	0.649	0.845	0.107	0.423	0.319
0.697	0.581	0.924	0.044	0.713	0.866	0.238	0.578	0.961	0.137

Soit X une v.a qui représente les valeurs de l'échantillon précédent. On cherche à déterminer si la v.a X suit une distribution uniforme continue ou non.

Selon, le programme de qq-plot de loi Uniforme, on a les résultats suivants :

x<-c(0.044,0.107,0.137,0.206,0.238,0.319,0.395,0.423,0.562,0.578,0.581,0.649,0.697,0.713,0.781,0.845,

0.866, 0.912,0.924,0.961)

> RcUnif

[1] 0.9858377

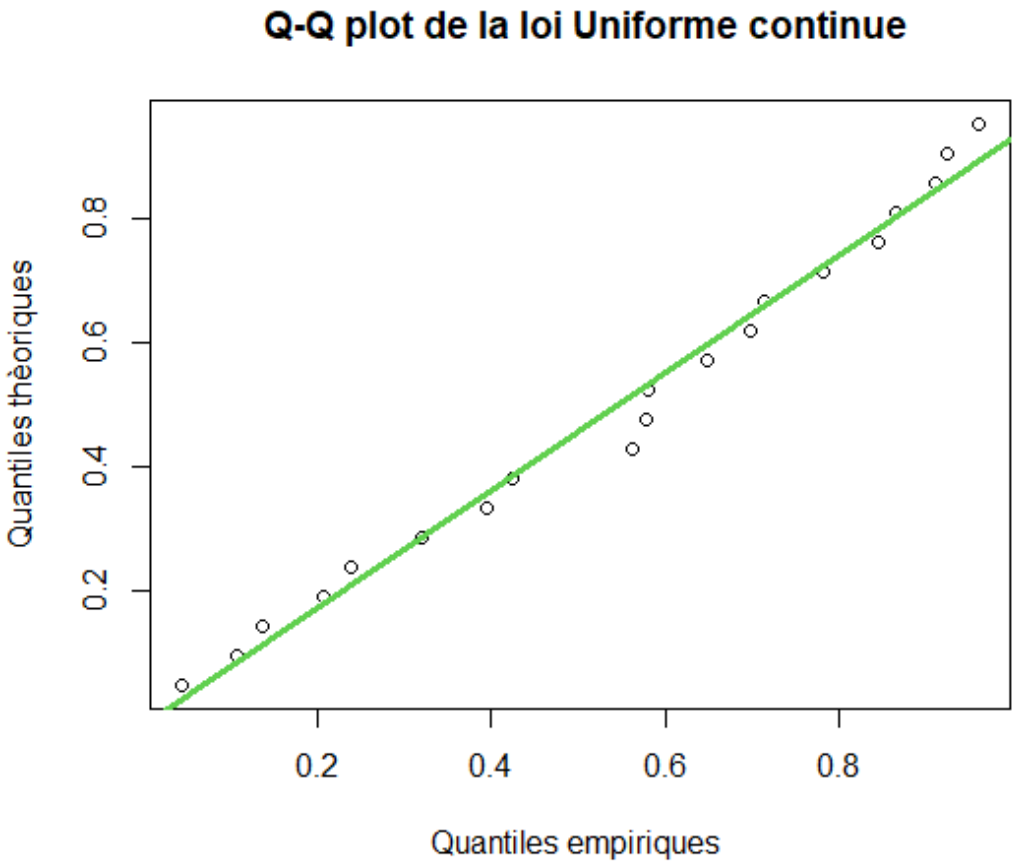

FIG. 2.3 – QQ-plot d'une loi uniforme continue

Donc, un motif de ligne droite apparaîtra dans le nuage de points. C'est-à-dire cet échantillon suit une loi uniforme.

QQ-plot d'une distribution de type Pareto

Le QQ-plot d'une distribution Pareto de paramètre $1/\gamma$, telle que pour tout $\gamma > 0$, ainsi pour $\gamma = 1$

$$F_\gamma(x) = 1 - x^{-1/\gamma} \qquad F_1(x) = 1 - x^{-1}$$

Les fonctions quantiles de la distribution de pareto à la forme simple :

$$\log(Q_\gamma(p)) = -\gamma \log(1-p) \quad \text{pour } p \in [0,1].$$

$$\log(Q_1(p)) = -\log(1-p) \quad \text{pour } p \in [0,1].$$

Il existe une relation linéaire simple entre les quantiles théoriques et les quantiles standard.

$$\log(Q_1(p)) = \frac{1}{\gamma} \log(Q_\gamma(p)) \quad \text{pour } p \in [0,1].$$

De la même façon qu'on a fait avec la distribution pareto on estime Q par Q_n. Dans un système de coordonnées orthogonales, les points de valeurs

$$(\log(Q_n(p)), -\log(1-p))$$

sont tracées pour déférentes valeurs de $p \in [0,1]$.Nous attendons alors qu'un motif de ligne droite apparaîtra dans le nuage de points si le modèle permet un ajustement statistique plausible pour la population statistique donnée. En effet, si le modèle est correct, alors l'équation

$$(-\log(1-p)) = \frac{1}{\gamma} \log(Q_\gamma(p))$$

est valable.

En général,

$$Q_n(p) = X_{i,n} \quad \text{pour} \quad \frac{i-1}{n} < p < \frac{i}{n}$$

Pour obtenir finalement, les coordonnées

$$(\log(X_{i,n}), -\log(1-p_{i,n}))$$

Procédure sous R et illustration graphique

Pour avoir le graphique QQ-plot des données $x_1, x_2, ..., x_n$ contre les quantiles de Pareto, on exécute le programme suivant sous R :

x<-. # Entrer l'échantillon à étudier des observations x_i

n<-length(x)

x<-sort(x) # On ordonne l'échantillon des observations x_i

p<-(1 :n)/(n+1) # Vecteur des valeurs i/(n+1)

Droite<-lm((-log(1-p))~ log(x)) # Ajuster un modèle de régression linéaire.

coef(Droite) # Estimation du paramètres.

qqplot(log(x),-log(1-p), xlab="Quantiles empiriques ", ylab="Quantiles théoriques", main = " Q-Q plot de la loi type Pareto")

abline(coef(Droite), col="3", lwd="3")# Tracer la droite de regression.

RcPareto<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de loi de type pareto.

Cas pratique 2.2.3

Soit X un v.a d'un échantillon aléatoire de taille $n = 20$.

1.198885	14.421026	1.156537	19.540713	1.205317
125.452346	5.279182	2.949933	1.422006	35.687986
1.082133	19.056664	1.150247	2.156494,	7.690356
2.106200	3.935910	3.425991	1.463136	4.079289

On s'interroge pour savoir si la v.a X suit ou non une loi de type pareto.

Selon, le programme de qq-plot de la loi de type pareto, on a les résulte suivante :

x<-c(1.198885, 14.421026, 1.156537, 19.540713, 1.205317, 125.452346, 5.279182, 2.949933, 1.422006, 35.687986,1.082133, 19.056664, 1.150247, 2.156494, 7.690356, 2.1062, 3.935910, 3.425991, 1.463136, 4.079289)

> RcPareto

[1] 0.9863724

FIG. 2.4 – Q-Q plot de la loi type Pareto

Donc, un motif de ligne droite apparaîtra dans le nuage de points. C'est-à-dire cet échantillon suit une loi de type pareto.

QQ-plot d'une distribution Normale $N(\mu, \sigma^2)$

Nous présentons le QQ-plot d'une distribution Normale de paramètres μ, σ^2.

Soit la fonction de répartition de loi Normale :

$$F_{\mu,\sigma^2}(x) = \frac{1}{2}(1 + \text{erf}(\frac{x-\mu}{\sigma\sqrt{2}}))$$

Ainsi pour $\mu = 0$, $\sigma^2 = 1$

$$F_{0,1}(x) = \frac{1}{2}(1 + \text{erf}(\frac{x}{\sqrt{2}}))$$

Les fonctions de quantiles de la distribution normale à la forme simple :

$$Q_{\mu,\sigma^2}(p) = \mu + \sigma\sqrt{2}\,\text{erf}^{-1}(2p-1) \quad \text{pour } p \in [0,1].$$

$$Q_{0,1}(p) = \sqrt{2}\,\text{erf}^{-1}(2p-1) \quad \text{pour } p \in [0,1].$$

Il existe une relation linéaire simple entre les quantiles théoriques d'une distribution normale et les quantiles d'une distribution normale standard.

$$Q_{0,1}(p) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}Q_{\mu,\sigma^2}(p) \quad \text{pour } p \in [0,1].$$

On estime Q par Q_n. Dans un système de coordonnées orthogonales, les points de valeurs,

$$(Q_n(p), \sqrt{2}\,\text{erf}^{-1}(2p-1)),$$

sont tracées pour déférentes valeurs de $p \in [0,1]$. Nous attendons alors qu'un motif de ligne droite apparaîtra dans le nuage de points si le modèle permet un ajustement statistique plausible pour la population statistique donnée.

En général,

$$Q_n(p) = X_{i,n} \quad \text{pour} \quad \frac{i-1}{n} < p < \frac{i}{n}$$

Pour obtenir finalement, les coordonnées :

$$(X_{i,n}, \sqrt{2}\,\text{erf}^{-1}(2p_{i,n}-1))$$

Procédure sous R et illustration graphique

Pour avoir le graphique QQ-plot des données $x_1, x_2, ..., x_n$ contre les quantiles normale, on exécute le programme suivant sous R :

x<-. # Entrer l'échantillon à étudier des observations x_i

n=length(x)

erf <- function(x) {

2 * pnorm(x * sqrt(2)) - 1} # Fonction d'erreur

inverf<-function(u){

qnorm((u+1)/2)/sqrt(2)} # Fonction d'inverse d'erreur

x<-sort(x) # On ordonne l'échantillon des observations x_i

p<-(1 :n)/(n+1) # Vecteur des valeurs i/(n+1)

Droite<-lm(sqrt(2)*inverf(2*p-1)~x) # Ajuster un modèle de régression linéaire

coef(Droite)

qqplot(x,sqrt(2)*inverf(2*p-1),xlab="Quantiles empiriques ", ylab="Quantiles théoriques" , main = "QQ-plot de la loi Normale")

abline(coef(Droite), col="3", lwd="3") #Tracer le droite de regression

RcNorm<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de loi Normale.

Cas pratique 2.2.4

On fait passer à 50 adolescents le test psychologique de Rorsçhach. Les temps de passation en minutes du test sont :

43	48	65	55	51	51	44	51	59	62
45	53	55	55	49	34	52	69	45	54
59	36	36	29	52	59	41	58	54	55
72	53	52	49	57	42	70	58	42	53
57	68	40	65	54	49	32	56	50	59

On s'interroge pour savoir si la v.a X qui à un adolescent associe son temps de passation

au test suit ou non une loi normale.

x<-c(43, 48, 65, 55, 51, 51, 44, 51, 59, 62, 45, 53, 55, 55, 49, 34, 52,69, 45, 54, 59, 36, 36, 29, 52, 59, 41, 58, 54, 55, 72, 53, 52, 49, 57, 42,70, 58, 42, 53, 57, 68, 40, 65, 54, 49, 32, 56, 50, 59)

Selon, le programme de qq-plot de la loi Normale. Le résultat est donnée comme suivant :

> RcNorm

[1] 0.9790692

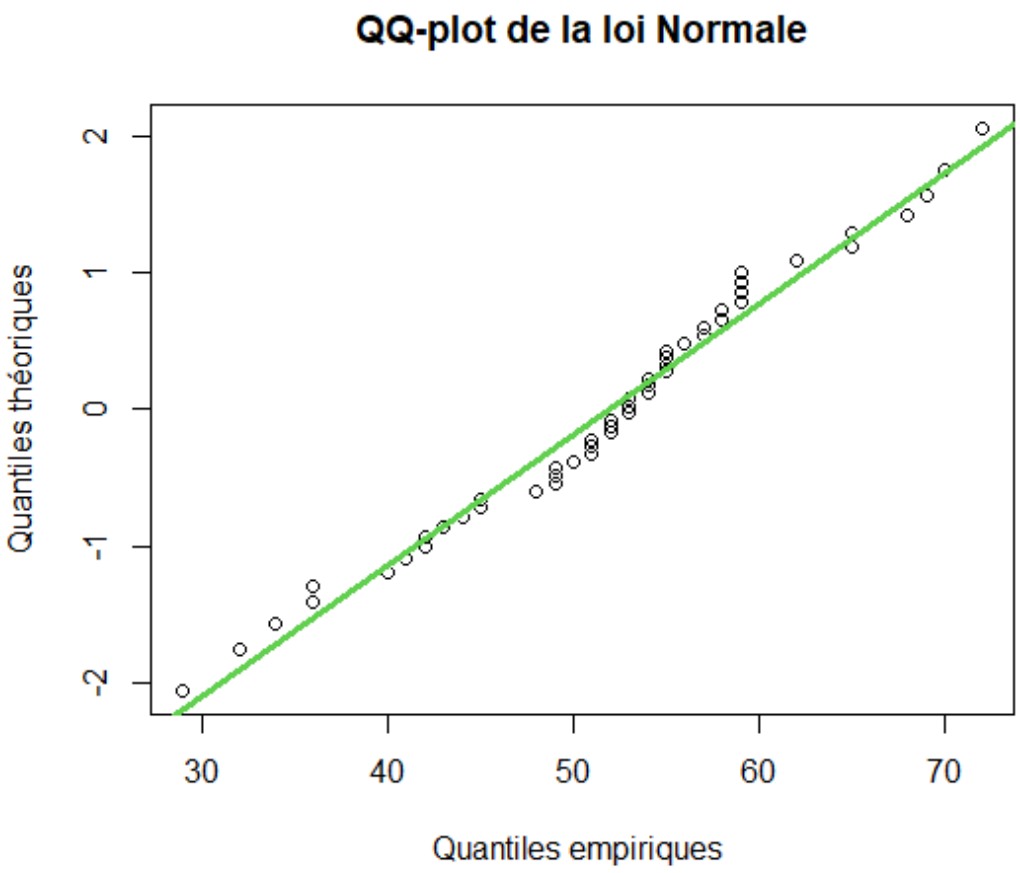

FIG. 2.5 – QQ-plot de la loi Normale

La droite ajuste bien le nuage de points. Donc, on peut dire que cet échantillon suit une loi normale.

Remarque 2.2.1 *Nous avons une instruction intégrée dans R, qui donne directement*

le qq-plot de la loi normale. Ainsi, repronons l'échantillon de dernier cas pratique, en exécutant le code qqnorm(x).

Cette instruction a été codé selon la référence [1]

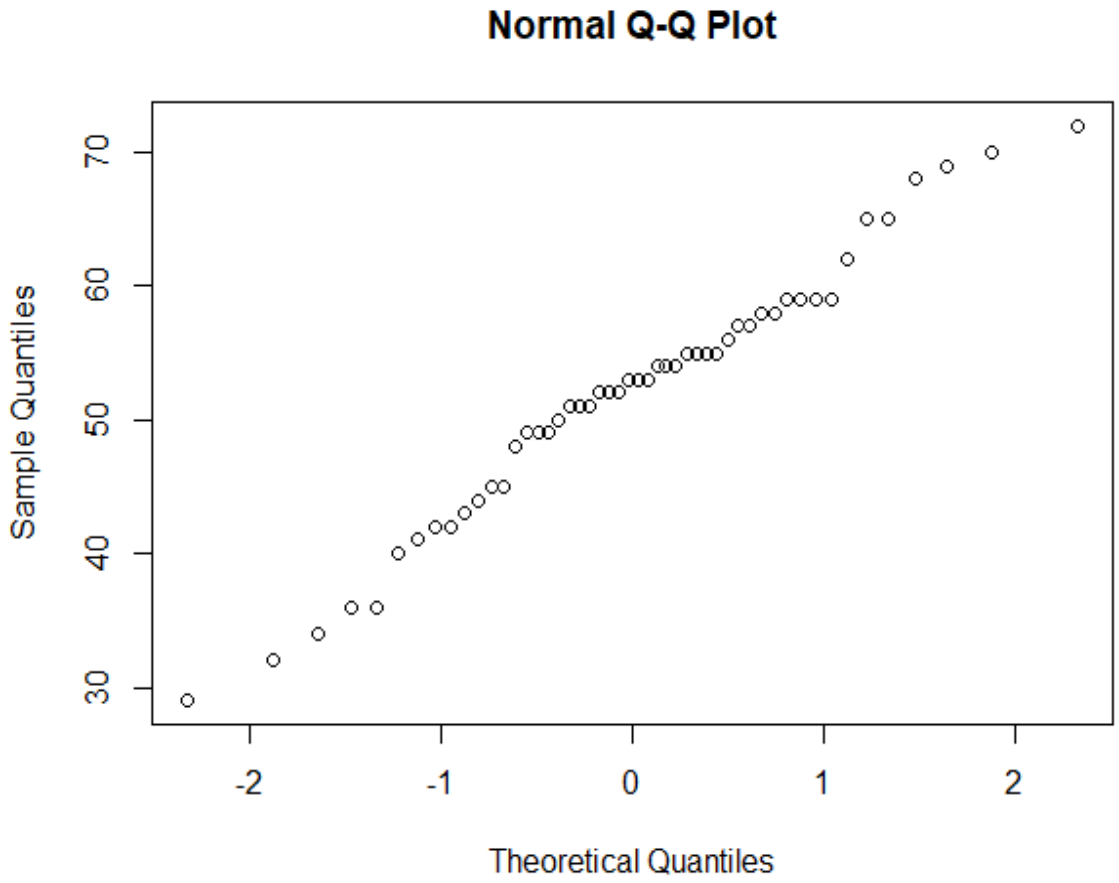

FIG. 2.6 – graphe de QQnorm

QQ-plot d'une distribution de Cauchy

Nous présentons le QQ-plot d'une distribution cauchy de paramètres $\alpha, \varkappa_0$ telle que ($\alpha \in \mathbb{R}^+, \varkappa_0 \in \mathbb{R}$)

$$F_{\alpha,\varkappa_0}(x) = \frac{1}{\pi}\arctan(\frac{x-\varkappa_0}{\alpha}) + \frac{1}{2},$$

et quand $\alpha = 1$ et $\varkappa_0 = 0$:

$$F_{1,0}(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}$$

Les fonctions de quantiles de la loi de Cauchy :

$$Q_{\alpha,\varkappa_0}(p) = \varkappa_0 + \alpha\tan(\pi(p - \frac{1}{2})) \quad \text{pour } p \in [0,1].$$

$$Q_{1,0}(p) = \tan(\pi(p - \frac{1}{2})) \quad \text{pour } p \in [0,1].$$

Remarqons qu'il existe une relation linéaire simple entre les quantiles théoriques d'une distribution Cauchy et les quantiles standard d'une distribution Cauchy :

$$Q_{1,0}(p) = \frac{\varkappa_0}{\alpha} + \frac{1}{\alpha}Q_{\alpha,\varkappa_0}(p) \quad \text{pour } p \in [0,1].$$

On estime Q par Q_n. Dans un système de coordonnées orthogonales, les points de valeurs

$$(Q_n(p), \tan(\pi(p - \frac{1}{2}))).$$

sont tracées pour déférentes valeurs de $p \in [0,1]$. Nous attendons alors qu'un motif de ligne droite apparaîtra dans le nuage de points si le modèle permet un ajustement statistique plausible pour la population statistique.

En général,

$$Q_n(p) = X_{i,n} \quad \text{pour} \quad \frac{i-1}{n} < p < \frac{i}{n}$$

Pour obtenir finalement, les coordonnées

$$(X_{i,n}, \tan(\pi(p_{i,n} - \frac{1}{2})))$$

Procédure sous R et illustration graphique

Pour avoir le graphique QQ-plot des données $x_1, x_2, ..., x_n$ contre les quantiles de Cauchy,

on exécute le programme suivant sous R :

x<-# Entrer l'échantillon à étudier des observations x_i

n<-length(x)

x<-sort(x) # On ordonne l'échantillon des observations x_i

p<-(1 :n)/(n+1) # vecteur des valeurs i/(n+1)

Droite<-lm(tan(pi*(p-(1/2)))~x) # Ajuster un modèle de régression linéaire

coef(Droite) #Estimation du paramètres

qqplot(x,tan(pi*(p-(1/2))), xlab="Quantiles empiriques ", ylab="Quantiles théoriques",

xlim=c(min(x),max(x)), main = " Q-Q plot de la loi Cauchy")

abline(coef(Droite), col="3", lwd="3") # Tracer le droite de regression

RcCauchy<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2)

Cas pratique 2.2.5

Soit l'échantillon constitué d'une série de 40 valeurs numériques réelles générées de manière aléatoire.

−0.139	0.218	0.308	0.425	0.499	0.537	0.632	0.667
0.785	0.805	0.816	0.932	1.052	1.054	1.191	1.236
1.302	1.417	1.603	1.668	1.716	1.727	1.882	1.889
1.972	1.982	2.094	2.102	2.171	2.334	2.352	2.396
2.451	2.490	2.515	2.617	2.785	2.865	2.911	3.354

On veut savoir si la v.a X suit ou non une loi de Cauchy.

Selon, le programme de qq-plot de loi de Cauchy, on a le résultat suivant :

x<–c(-0.139, 0.218, 0.308, 0.425, 0.499, 0.537, 0.632, 0.667,0.785, 0.805, 0.816, 0.932, 1.052, 1.054, 1.191, 1.236,1.302, 1.417, 1.603, 1.668, 1.716, 1.727, 1.882, 1.889,1.972, 1.982, 2.094, 2.102, 2.171, 2.334, 2.352, 2.396,2.451, 2.490, 2.515, 2.617, 2.785, 2.865, 2.911, 3.354)

>RcCauchy

[1] 0.7024112

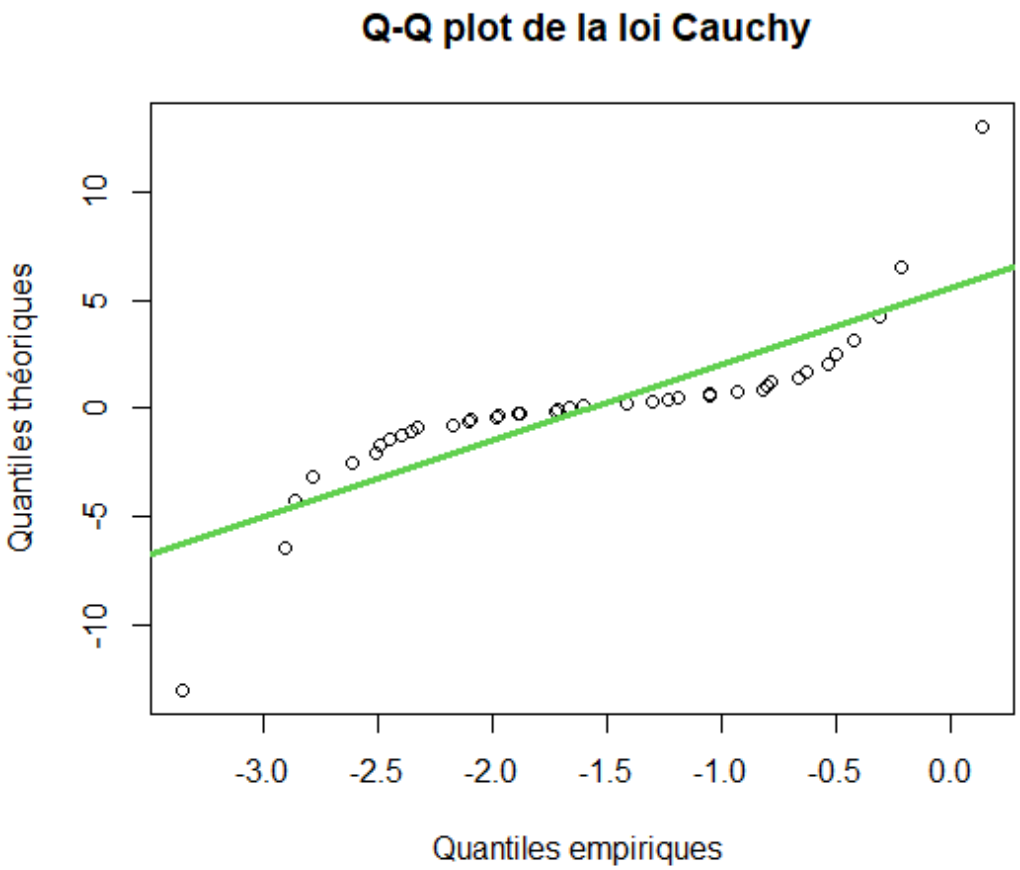

FIG. 2.7 – QQ-plot de la loi de Cauchy

Donc, un motif de ligne droite apparaîtra dans le nuage de points. C'est-à-dire cet échantillon suit une loi de Cauchy.

2.3 Estimation par maximum de vraisemblance

Dans cette section, nous rappelons les méthodes de l'estimation paramétrique et reconcentrons sur l'estimation de maximum de vraisemblance.

2.3.1 Estimation paramétrique

L'estimation paramétrique implique l'estimation ou l'évaluation d'un ou plusieurs paramètres inconnus à partir des données observées $x_1, ..., x_n$ de la variable X, tels que la moyenne, la variabilité, etc. Il est naturel de vouloir obtenir une estimation aussi précise que possible. Les trois principaux types d'estimation paramétrique sont les suivants :

- **L'estimation ponctuelle** : on estime directement le ou les paramètres inconnus par des valeurs réelles.
- **L'estimation par intervalles de confiance** : on détermine des intervalles réels, aussi étroits que possible, qui ont de fortes chances de contenir un paramètre inconnu.
- **Les tests statistiques** : il s'agit de démarches visant à accepter ou rejeter une hypothèse impliquant un ou plusieurs paramètres inconnus, tout en minimisant le risque d'erreur.

Dans notre nouvel algorithme, nous utilisons l'estimation du maximum de vraisemblance (EMV) comme méthode d'estimation ponctuelle.

2.3.2 Estimateurs du maximum de vraisemblance

Fonction de vraisemblance

On appelle fonction de vraisemblance pour $(x_1, ..., x_n)$ la fonction de θ :

$$L_n(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

On rappelle que :

$$f(x;\theta) = \begin{cases} P_\theta(X = x) & \text{Si } X \text{ une v.a discrète} \\ f_\theta(x) & \text{Si } X \text{ une v.a continue} \end{cases}$$

La fonction de vraisemblance n'est intéressante que si θ et x_i vérifient :

$$f(x_i;\theta) \neq 0 \text{ pour tout } i \in \{1,...,n\},$$

sinon on peut d'ores et déjà remettre en cause l'hypothèse que X suit la loi F_θ.

Fonction de vraisemblance : contexte théorique

Dans un contexte théorique, il est parfois nécessaire de décrire uniquement la modélisation de X, sans faire référence aux données spécifiques $x_1, ..., x_n$. Dans ce cas, on utilise le terme "fonction de vraisemblance" pour désigner la fonction de vraisemblance d'une réalisation donnée $(x_1, ..., x_n)$ de $(X_1, \ldots, X_n)$.

Estimateurs du maximum de vraisemblance pour $(x_1, ..., x_n)$

L'estimateur du maximum de vraisemblance (EMV) de θ pour $(x_1, ..., x_n)$ est défini comme un réel $\widehat{\theta}$ qui maximise la fonction de vraisemblance $L_n(x_1, ..., x_n; \theta)$ en θ, i.e. pour tout θ.

$$L_n(x_1, ..., x_n; \theta) \leq L_n(x_1, ..., x_n; \widehat{\theta})$$

Une expression alternative est :

$$\widehat{\theta} \in \arg\max_\theta L_n(x_1, ..., x_n; \theta)$$

Où arg max désigne l'argument du maximum qui est l'ensemble des points en lesquels une expression atteint sa valeur maximale. Puisqu'il dépend de $x_1, ..., x_n$, θ est une estimation ponctuelle de θ. Un tel estimateur n'existe pas toujours et peut ne pas être unique.

Fonction de Log-vraisemblance

On appelle fonction de log-vraisemblance pour $(x_1, ..., x_n)$ la fonction de θ définie par :

$$\ell_n(x_1, ..., x_n; \theta) = \log(L_n(x_1, ..., x_n; \theta))$$

Elle n'a de sens que si θ vérifie : $L_n(x_1, ..., x_n; \theta) > 0$.

La fonction logarithme népérien étant croissante, l'emv $\widehat{\theta}$ de θ pour $(x_1, ..., x_n)$ vérifie :

$$\widehat{\theta} \in \arg\max_{\theta} L_n(x_1, ..., x_n; \theta) = \arg\max_{\theta} \ell_n(x_1, ..., x_n; \theta)$$

Équation de vraisemblance

On appelle équation de vraisemblance l'équation en θ :

$$\frac{\partial}{\partial\theta}\ell_n(x_1, ..., x_n; \theta) = 0$$

Expression analytique de l'EMV

Pour envisager d'avoir une expression analytique de l'emv $\widehat{\theta}$ de θ pour $(x_1, ..., x_n)$, une approche consiste à exprimer $L_n(x_1, ..., x_n; \theta)$ en fonction de produits de termes exponentiels/puissances, puis de considérer la fonction de log-vraisemblance $\ell_n(x_1, ..., x_n; \theta)$. Si cette dernière est dérivable en θ, une condition nécessaire que doit vérifier $\widehat{\theta}$ est d'être solution de l'équation de vraisemblance. Il faut ensuite vérifier que $\hat{\theta}$ est bien un maximum pour $\ell_n(x_1, ..., x_n; \theta)$:

– soit en étudiant les variations de $\ell_n(x_1, ..., x_n; \theta)$,

– soit en montrant que

$$\frac{\partial^2}{\partial\theta^2}\ell_n(x_1, ..., x_n; \hat{\theta}) < 0$$

2.3.3 Estimation des paramètres de lois

Loi Exponentielle Soit $x_1, ..., x_n \sim \xi(\lambda)$

$$f_{x_i}(x_1, ..., x_n; \lambda) = \lambda e^{-\lambda x_i}$$

La fonction de vraisemblance est la suivante :

$$\begin{aligned} L_n(x_1, ..., x_n; \lambda) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

La fonction de log-vraisemblance est la suivante :

$$\ell_n(x_1, ..., x_n; \lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

Calculons la dérivée première par rapport à λ :

$$\frac{\partial}{\partial \lambda} \ell_n(x_1, ..., x_n; \lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \lambda} \ell_n(x_1, ..., x_n; \lambda) = 0 \Leftrightarrow \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Donc la fonction de vraisemblance est maximale en :

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X}} \tag{2.4}$$

Loi Normale Soit $x_1, ..., x_n \sim N(\mu, \sigma^2)$

$$f_{x_i}(x_1, ..., x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right)$$

La fonction de vraisemblance est la suivante :

$$\begin{aligned} L_n(x_1,...,x_n;\mu,\sigma^2) &= \prod_{i=1}^{n} f(x_i;\mu,\sigma^2) \\ &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

La fonction de log-vraisemblance est la suivante :

$$\ell_n(x_1,...,x_n;\mu,\sigma^2) = -n\log(\sqrt{2\pi}) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Calculons la dérivée première par rapport à μ et σ^2 :

$$\begin{left} \left\{ \begin{array}{l} \frac{\partial}{\partial\mu}\ell_n(x_1,...,x_n;\mu) = \frac{1}{2\sigma^2}2\sum_{i=1}^{n}(x_i-\mu) \\ \frac{\partial}{\partial\sigma^2}\ell_n(x_1,...,x_n;\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \end{array} \right. \\ \iff \left\{ \begin{array}{l} \frac{\partial}{\partial\mu}\ell_n(x_1,...,x_n;\mu) = \sum_{i=1}^{n} x_i - n\mu \\ \frac{\partial}{\partial\sigma^2}\ell_n(x_1,...,x_n;\sigma^2) = \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} \end{array} \right. \end{left}$$

$$\left\{ \begin{array}{l} \frac{\partial}{\partial\mu}\ell_n(x_1,...,x_n;\mu) = 0 \\ \frac{\partial}{\partial\sigma^2}\ell_n(x_1,...,x_n;\sigma^2) = 0 \end{array} \right. \iff \left\{ \begin{array}{l} \sum_{i=1}^{n} x_i - n\mu = 0 \\ \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0 \end{array} \right.$$

Donc la fonction de vraisemblance est maximale en :

$$\left\{ \begin{array}{l} \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X} \\ \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = S_n^2 \end{array} \right. \tag{2.5}$$

Loi de type Pareto Soit $x_1,...,x_n \sim Pr(\gamma)$

$$f_{x_i}(x_1,...,x_n;\gamma) = \frac{1}{\gamma}x_i^{-\frac{1}{\gamma}-1}$$

La fonction de vraisemblance est la suivante :

$$\begin{aligned} L_n(x_1, ..., x_n; \gamma) &= \prod_{i=1}^{n} f(x_i; \gamma) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\gamma} x_i^{-\frac{1}{\gamma}-1} \right) \\ &= \frac{1}{\gamma^n} \prod_{i=1}^{n} x_i^{-\frac{1}{\gamma}-1} \end{aligned}$$

La fonction de log-vraisemblance est la suivante :

$$\ell_n(x_1, ..., x_n; \gamma) = -n \log(\gamma) - \left(\frac{1}{\gamma} + 1 \right) \sum_{i=1}^{n} \log(x_i)$$

Calculons la dérivée première par rapport à γ :

$$\begin{aligned} \frac{\partial}{\partial \gamma} \ell_n(x_1, ..., x_n; \gamma) &= -\frac{n}{\gamma} + \frac{1}{\gamma^2} \sum_{i=1}^{n} \log(x_i) \\ &= \frac{-n\gamma + \sum_{i=1}^{n} \log(x_i)}{\gamma^2} \end{aligned}$$

$$\frac{\partial}{\partial \gamma} \ell_n(x_1, ..., x_n; \gamma) = 0 \Leftrightarrow \frac{-n\gamma + \sum_{i=1}^{n} \log(x_i)}{\gamma^2} = 0$$

Donc la fonction de vraisemblance est maximale en :

$$\widehat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \log(x_i) \tag{2.6}$$

Loi uniforme continue Soit $x_1, ..., x_n \sim U[a, b]$

$$f_{x_i}(x_1, ..., x_n; a, b) = \frac{1}{b-a} 1_{[a;b]}(x_i)$$

La fonction de vraisemblance est la suivante :

$$\begin{aligned} L_n(x_1,...,x_n;a,b) &= \prod_{i=1}^{n} f(x_i;a,b) \\ &= \prod_{i=1}^{n} \frac{1}{b-a} 1_{[a;b]}(x_i) \\ &= \left(\frac{1}{b-a}\right)^n \prod_{i=1}^{n} 1_{[a;b]}(x_i);\ \ \forall x_i \quad a \leq x_i \leq b \end{aligned}$$

Ce que implique :

$$\begin{cases} a \in]-\infty; x_i] \\ b \in [x_i; +\infty[\end{cases} \Longrightarrow \begin{cases} a \in (-\infty; \min\limits_{1<i<n}\{x_i\}) \\ b \in (\max\limits_{1<i<n}\{x_i\}; +\infty) \end{cases}$$

Si les X_i sont ordonné ($x_1 < x_2 < ... < x_n$), alors :

$$L_n(x_1,...,x_n;a,b) = \frac{1}{(b-a)^n} 1_{\{a \leq x_{(1)}; b \geq x_{(n)}\}}(a,b)$$

Donc,

$$\begin{cases} \widehat{a} = \min\limits_{1<i<n}\{x_i\} \\ \widehat{b} = \max\limits_{1<i<n}\{x_i\} \end{cases}$$

Loi de Cauchy Soit $x_1,...,x_n \sim C(\varkappa_0, \alpha)$

$$f_{x_i}(x_1,...,x_n;\varkappa_0,\alpha) = \frac{1}{\pi\alpha} \frac{1}{1+\left(\frac{x_i-\varkappa_0}{\alpha}\right)^2}$$

La fonction de vraisemblance est la suivante :

$$\begin{aligned} L_n(x_1,...,x_n;\varkappa_0,\alpha) &= \prod_{i=1}^{n} f(x_i;\varkappa_0,\alpha) \\ &= \prod_{i=1}^{n}\left(\frac{1}{\pi a}\frac{1}{1+\left(\frac{x_i-\varkappa_0}{\alpha}\right)^2}\right) \\ &= \frac{1}{(\pi\alpha)^n}\prod_{i=1}^{n}\left(\frac{1}{1+\left(\frac{x_i-\varkappa_0}{\alpha}\right)^2}\right) \end{aligned}$$

La fonction de log-vraisemblance est la suivante :

$$\ell_n(x_1,...,x_n;\varkappa_0,\alpha) = -n\log(\pi\alpha) - \sum_{i=1}^{n}\log\left(1+\left(\frac{x_i-\varkappa_0}{\alpha}\right)^2\right)$$

Calculons la dérivée première par rapport à $\varkappa_0$ et α :

$$\begin{cases} \dfrac{\partial}{\partial\varkappa_0}\ell_n(x_1,...,x_n;\varkappa_0) = 2\displaystyle\sum_{i=1}^{n}\frac{(x_i-\varkappa_0)}{\alpha^2+(x_i-\varkappa_0)^2} \\ \dfrac{\partial}{\partial\alpha}\ell_n(x_1,...,x_n;\alpha) = -\dfrac{n}{\alpha} + 2\displaystyle\sum_{i=1}^{n}\frac{(x_i-\varkappa_0)^2}{\alpha[\alpha^2+(x_i-\varkappa_0)^2]} \end{cases}$$

$$\begin{cases} \dfrac{\partial}{\partial\varkappa_0}\ell_n(x_1,...,x_n;\varkappa_0) = 0 \\ \dfrac{\partial}{\partial\alpha}\ell_n(x_1,...,x_n;\alpha) = 0 \end{cases}$$
$$\iff \begin{cases} 2\displaystyle\sum_{i=1}^{n}\frac{(x_i-\varkappa_0)}{\alpha^2+(x_i-\varkappa_0)^2} = 0 \\ -\dfrac{n}{\alpha} + 2\displaystyle\sum_{i=1}^{n}\frac{(x_i-\varkappa_0)^2}{\alpha[\alpha^2+(x_i-\varkappa_0)^2]} = 0 \end{cases}$$

Donc la fonction de vraisemblance est maximale en :

$$\begin{cases} \widehat{\varkappa}_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i = \overline{X} \\ \widehat{\alpha} = \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\varkappa_0)^2} \end{cases}$$

2.3.4 Simulation des paramètres

Soit $x_1, x_2, ..., x_n$

#L'EMV pour le paramètre de loi Exponentielle.

EMV.lambda<-1/mean(x)

#L'EMV pour le paramètre de loi Normale.

EMV.mu<-mean(x)

EMV.sigma<-sd(x)^2

#L'EMV pour le paramètre de loi Pareto.

EMV.gamma<-(1/n)*sum(log(x))

#L'EMV pour le paramètre de loi Uniforme Continue.

EMV.a<-min(x)

EMV.b<-max(x)

#L'EMV pour le paramètre de loi Cauchy.

EMV.x0<-mean(x)

EMV.alpha<-sqrt((1/n)*sum(x-x0)^2)

Chapitre 3

Algorithme d'ajustement d'observation par une loi de probabilité

Nous arrivons finalement au chapitre trois par la présentation de notre nouveau modèle « **lois.test** » qui utilise les outils statistiques d'exploration de données illustrés au chapitre deux. Notamment les résultats théoriques des lois de probabilité et de convergence mentionnés au chapitre un.

Dans cette partie de notre mémoire, nous présenterons une brève introduction sur le test d'hypothèse de Kolmogorov-Smirnov, qui va servir à confirmer par notre modèle d'ajustement l'hypothèse, $H_0 : F = F_0$, qui stipule que l'échantillon d'observations est approché par une loi F_0 qu'on a déjà retrouvée par la technique du QQplot et estimer ses paramètres par la méthode du maximum de vraisemblance.

Dans ce contexte, nous présenterons des simulations illustratives pour montrer le déroulement de notre algorithme. Ainsi nous finissons par une application réelle sur le nombre de cas affecté par covid19 entre janvier 2020 et juillet 2022.

3.1 Test d'ajustement de Kolmogorov-Smirnov

3.1.1 Introduction

Le test d'ajustement est une méthode statistique utilisée pour évaluer si un échantillon de données suit une distribution théorique donnée. Il permet de déterminer si les données observées diffèrent de manière significative de la distribution théorique spécifiée.

Soit $x_1, x_2, ..., x_n$ une suite des v.a qui suit une loi F telle que F est continue et inconnue.

- **Formulation des hypothèses**

  Hypothèse nulle (H_0) : suppose que l'échantillon de données suit la distribution théorique spécifiée F_0

  Hypothèse alternative (H_1) : suppose que les données ne suivent pas cette distribution

  On note :

$$\begin{cases} H_0 : & F = F_0 \\ H_1 : & F \neq F_0 \end{cases}$$

- **Statistique du test :** Le test d'ajustement calcule une statistique de test qui mesure l'écart entre la distribution théorique et les données observées. Cette statistique est ensuite comparée à une valeur critique ou à une distribution de référence pour évaluer la significativité de l'écart.
- **Règle de décision :** En fonction de la statistique de test et du seuil de signification choisi, on peut rejeter ou ne pas rejeter l'hypothèse nulle. Si l'hypothèse nulle est rejetée, cela suggère que les données ne suivent pas la distribution théorique spécifiée.

Remarque 3.1.1 *Il existe plusieurs tests d'ajustement, tels que le test de Kolmogorov-Smirnov, le test de khi-deux, le test de Lilliefors, etc. Le choix du test dépend de la nature des données et de la distribution théorique supposée. Dans ce mémoire, nous concentrons sur le test de Kolmogorov-Smirnov, et nous nous limitions à l'ajustement des lois usuelles continue.*

Le test de Kolmogorov-Smirnov (KS) c'est le plus populaire parmi les tests d'adéquation. Il mesure la différence maximale entre la fonction de répartition empirique de l'échantillon et la fonction de répartition théorique, souvent notée F_0. L'hypothèse nulle du test est que l'échantillon provient de la distribution théorique spécifiée. Il a été proposé par Andreï N. Kolmogorov en 1933 et étendu par Vladimir I. Smirnov en 1939.

3.1.2 Statistique de 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 :

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

Où F_0 est la fonction de répartition théorique et

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

Désigne la fonction de répartition empirique définie pour l'échantillon $(X_1, ..., X_n)$.
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 suivante.

Proposition 3.1.1 *La statistique de Kolmogorov-Smirnov s'écrit comme suit :*

$$D_n := \max_{1\leq i\leq n}\left(\max\left(\left|F_0(X_{i,n}) - \frac{i}{n}\right|; \left|F_0(X_{i,n}) - \frac{i-1}{n}\right|\right)\right) \tag{3.1}$$

Remarque 3.1.2 *En pratique, le test de Kolmogorov-Smirnov est souvent utilisé pour évaluer si une v.a X suit une distribution spécifique avec une fonction de densité de probabilité continue.*

3.1.3 Région critique

La région critique du test est de la forme $\{D_n > D_{crit}\}$ où D_{crit} est une certaine valeur critique vérifiant : $P\left(D_n > D_{crit}/H_0 \text{ est vrai}\right) = \alpha; \quad 0 \leq \alpha \leq 1$.

Tell que : $D_{crit} = \frac{c}{\sqrt{n}}$

On conclut le test en acceptant, au seuil de signification, l'hypothèse H_0 si la distance D_n calculée est inférieure à D_{crit}. C-à-d, on accepte H_0 si : $D_n < D_{crit}$

Les valeurs de c en fonction de valeurs usuelles sont données selon la tableau de Kolmogorov-Smirnov (Tableau 3.2)

α	0.20	0.10	0.05	0.02	0.01
c	1.073	1.223	1.358	1.518	1.629

(3.2)

TAB. 3.1 – Valeurs de c pour calculer la valeur critique du test

3.1.4 P-valeur

En statistique, la $P - valeur$ est le plus petit seuil de significativité pour lequel H_0 est accepté. On va comparer un seuil de signification α et P-valeur pour accepter ou rejeter H_0 comme suit :

– Si $P - valeur \leq \alpha$ on va rejeter l'hypothèse H_0.

– Si $P - valeur > \alpha$ on va accepter l'hypothèse H_0.

Cas pratique 3.1.1 On mesure les durées de vie de 10 ampoules d'un même type. Les résultats sont en heures :

673 389 1832 570 522 2694 3683 644 1531 2916

Si la v.a X égale à la durée de vie en heures d'une ampoule de ce type, est-ce que l'on peut affirmer, au risque 5% que X suit ou pas la loi exponentielle $\zeta(1/1545)$.

Par l'énoncé, on observe la valeur de X sur chacun des n individus (ampoules) d'un échantillon avec $n = 10$: $(x_1, ..., x_n)$ (avec $x_i \in \mathbb{R}$).

On considère les hypothèses :

$$\begin{cases} H_0 : & X \text{ suit la loi exponentielle } \zeta(1/1545) \\ H_1 : & X \text{ ne suit pas la loi exponentielle } \zeta(1/1545) \end{cases}$$

On utilise le tableau suivant pour simplifier les calculs :

i	x_i	$F_0(x_i)$	$\frac{i}{n}$	$\frac{i-1}{n}$	$\left\|F_0(X_{i,n}) - \frac{i}{n}\right\|$	$\left\|F_0(X_{i,n}) - \frac{i-1}{n}\right\|$
1	389	0.2225842	0.1	0.0	0.12258	0.22258
2	522	0.2867078	0.2	0.1	0.08670	0.18670
3	570	0.3085276	0.3	0.2	0.00852	0.10852
4	644	0.3408660	0.4	0.3	0.05913	0.04086
5	673	0.3531227	0.5	0.4	0.14687	0.04687
6	1531	0.6287719	0.6	0.5	0.02877	0.12877
7	1832	0.6944863	0.7	0.6	0.00551	0.09448
8	2694	0.8251260	0.8	0.7	0.02512	0.12512
9	2916	0.8485317	0.9	0.8	0.12512	0.04853
10	3683	0.9078022	1	0.9	0.09219	0.00780

donc,

$$\begin{aligned} D_n &:= \max_{1\leq i\leq n}\left(\max\left(\left|F_0(X_{i,n}) - \frac{i}{n}\right|; \left|F_0(X_{i,n}) - \frac{i-1}{n}\right|\right)\right) \\ &= \max_{1\leq i\leq n}(0.22258, 0.18670, 0.10852, 0.05913, 0.14687, 0.12877, 0.09448, 0.12512, 0.12512, 0.09219) \\ &= 0.22258 \end{aligned}$$

Alors,

$$P(\text{accepte } H_1/H_0\text{est vrai}) = P\left(D_n > D_{crit}/H_0 \text{ est vrai}\right) \tag{3.3}$$
$$= \alpha, \quad \alpha = 0.05$$

D'aprés le Tableau 3.2, on trouve $c = 1.358$.

$$D_{crit} = \frac{c}{\sqrt{n}} = \frac{1.358}{\sqrt{10}} = 0.4294$$

Donc,

$$D_n = 0.22258 < 0.4294 = D_{crit}$$

Donc, $D_n < D_{crit}$, qui ce implique une contraduction par rapport à 3.3, au lieu d'accepter H_1, on accepte H_0, X suit la loi Exponentielle au risque $\alpha = 5\%$

Code R :

x = c(673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916)

x<-sort(x)

lambda<-1/1545

ks.test(x,pexp,lambda)

Exact one-sample Kolmogorov-Smirnov test

data : x

D = 0.22258, p-value = 0.6286

alternative hypothesis : two-sided

Commentaire : On remarque que $p-value > \alpha$ $(0.6286 > 0.05)$. Alors au risque $\alpha = 0,05$ on accepte l'hypothèse H_0 (la variable X suit une loi Exponentielle).

3.2 Nouvau modèle d'ajustement de jeu de données

3.2.1 Etapes d'algorithme

Dans cette partie, nous présenterons les étapes de notre nouvel algorithme dans le but de faciliter la compréhension du code R (voir page 66-72). Ainsi, notre modèle chargera un échantillon d'observations, son exécution fournira des résultats très précis il choisira parmi les cinq lois programmées la loi qui ajuste le mieux l'échantillon de données, en estimant les paramètres de cette loi avec une exactitude fiable, car nous avons utilisé des outils et des méthodes statistiques puissantes, d'exploration de données.

Chargement du jeu de données

- Soit $x_1, x_2, ..., x_n$ une suite de v.a de loi inconnue.
- Charger les observations dans un vecteur, noté x.
- Ordonner l'échantillon x
- Extraire la taille de l'échantillon, notée n.

Application de QQ-plot

- Tracer les QQ-plot des lois, respectivement (Loi Exponentielle, loi Uniforme continue, loi de type Pareto, loi Normale et loi Cauchy).
- Extraire les coefficients de déterminations de chaque loi, respectivement (RcExp, RcUnif, RcPareto, RcNorm, RcCauchy).
- Choisir la loi qui présente une allure linéaire des nuage de points du qq-plot. Ainsi qui correspond au plus grande valeur du coefficient de détermination appellée : « Rcarré ».

Estimation des paramétres de la loi "choisi" par maximum de vraisemblance

- Dans l'étape précédente, nous avons choisi la meilleure loi qui ajuste le mieux le jeu de donnée mais sans trouver ses paramètres.

– Ainsi nous estimons ces paramètres par la méthode de maximum de vraisemblance.

Test de kolmogorov-Smirnov

– La technique du qqplot et l'estimation de maximum de vraisemblance nous ont fourni une loi paramétrique « candidate » d'ajustement de jeu de donnée de paramètres « Estimés», cette loi est notée F_0. Nous allons confirmer ce choix par un test d'hypothèse d'ajustement de Kolmogorov-Sminrov, soit :

$$\begin{cases} H_0 : & F = F_0 \\ H_1 : & F \neq F_0 \end{cases}$$

Programmation des résulatats

– Si l'échantillon suit l'une des cinq lois précitées alors le résultat sera :

"x suit la loi « choisie » des paramétres « estimés » pour un seuil critique α représenté par un intervalle de [0,1]"

– Si l'échantillon ne suit aucunes des cinq lois, Alors le résultat sera :

"x n'appartient à aucune loi du programme."

– Stabilité par convergence :

Si l'échantillon suit une loi qui converge théoriquement vers l'une des cinq lois alors notre modèle arrivera à ajuster cet échantillon convenablement pour confirmer cette convergence.

3.2.2 Code R

Cette partie et consacré à notre véritable contribution ; le code du nouveau modèle que nous avons inséré dans d'une fonction qu'on a appelée **« lois.test »**. Ce modèle d'ajustement facilitera l'utilisation pour tout praticien à la recherche d'une technique fiable et rapide qui ne demande pas la connaissance des outils statistiques.

```
lois.test<-function(x){
n<-length(x)
x<-sort(x) # On ordonne l'échantillon des observations X_{i}
p<-(1 :n)/(n+1) # Vecteur des valeurs i/(n+1)
alpha<-(1 :1000)/1000
par(mfrow=c(3,3))
#Loi Exponentielle
Droite<-lm((-log(1-p))~x) # Ajuster un modèle de régression linéaire
coef(Droite) # Estimation du paramètres
qqplot(x,-log(1-p), xlab="Quantiles empiriques ", ylab="Quantiles thèoriques", main =
"QQ-plot de loi Exponentielle") # Tracer le qqplot
abline(coef(Droite), col="3", lwd="3") # Tracer la droite de régression
RcExp<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de
loi Exponentielle.
#Loi uniforme Continue :
Droite<-lm(p~x) # Ajuster un modèle de régression linéaire
coef(Droite) # Estimation des paramètres
qqplot(x,p, xlab="Quantiles empiriques ", ylab="Quantiles thèoriques", main = " Q-Q
plot de loi Uniforme C")
abline(coef(Droite), col="3", lwd="3")# Tracer la droite de regression
RcUnif<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de
loi Uniforme continue
#Loi de type Pareto :
if(any (x<=0)){
RcPareto<-0
}else {
Droite<-lm((-log(1-p))~log(x)) # Ajuster un modèle de régression linéaire
```

```
coef(Droite) # Estimation des paramètres
qqplot(log(x),-log(1-p), xlab="Quantiles empiriques ", ylab="Quantiles thèoriques", main
= " Q-Q plot de loi type Pareto")
abline(coef(Droite), col="3", lwd="3")# Tracer la droite de regression
RcPareto<-summary(Droite)$r.squared }# Extraire le coefficient de détermination (R^2)
de loi de type pareto
#Loi Normale :
erf <- function(x) {
2 * pnorm(x * sqrt(2)) - 1} # Fonction d'erreur
inverf<-function(u){
qnorm((u+1)/2)/sqrt(2)} # Fonction d'inverse d'erreur
Droite<-lm(sqrt(2)*inverf(2*p-1)~x) # Ajuster un modèle de régression linéaire
coef(Droite) # Estimation des paramètres
qqplot(x,sqrt(2)*inverf(2*p-1),xlab="Quantiles empiriques ", ylab="Quantiles thèoriques"
, main = "QQ-plot de loi Normale")
abline(coef(Droite), col="3", lwd="3") #Tracer le droite de regression
RcNorm<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2) de
loi normale.
#Loi Cauchy :
Droite<-lm(tan(pi*(p-(1/2)))~x) # Ajuster un modèle de régression linéaire
coef(Droite) #Estimation des paramètres
qqplot(x,tan(pi*(p-(1/2))),xlab="Quantiles empiriques ", ylab="Quantiles thèoriques",
xlim=c(min(x),max(x)), main = " Q-Q plot de loi Cauchy")
abline(coef(Droite), col="3", lwd="3") # Tracer le droite de régression
RcCauchy<-summary(Droite)$r.squared # Extraire le coefficient de détermination (R^2)
Rcarré<-max(c(RcExp,RcUnif,RcPareto,RcNorm,RcCauchy))
if(Rcarré==RcExp){
```

```
EMV.lambda<-(1/mean(x)) # EMV de lambda
ks.test(x, "pexp",EMV.lambda)# Test de Kolmogrov-Smirnov
p.valeur<-ks.test(x, "pexp",EMV.lambda)$p.value #P-valeur
if(all(alpha>p.valeur)){
print(paste("X n'appartient à aucune loi du programme."))
} else {
seuil <- c()
i <- 1
while (i <= length(alpha)) {
if (alpha[i] < p.valeur) {
seuil <- c(seuil, alpha[i])
}
i <- i + 1
}
print(paste("x suit la loi exponentielle avec de paramètre lambda=" ,EMV.lambda, " avec
R^2=" ,Rcarré, "Pour un seuil critique α ∈[",min(seuil),",",max(seuil),"]"))
}
}
if(Rcarré==RcUnif){
EMV.a<-min(x) #EMV de (a)
EMV.b<-max(x)# EMV de (b)
ks.test(x, "punif",EMV.a,EMV.b)# Test de Kolmogrov-Smirnov
p.valeur<-ks.test(x, "punif",EMV.a,EMV.b)$p.value
if(all(alpha>p.valeur)){
print(paste("x n'appartient à aucune loi du programme."))
} else {
seuil <- c()
```

```
i <- 1
while (i <= length(alpha)) {
if (alpha[i] < p.valeur) {
seuil <- c(seuil, alpha[i])}
i <- i + 1
}
print(paste("x suit la loi Uniforme continue de paramètre a=" ,EMV.a,"b=",EMV.b, "
avec R²=" ,Rcarré, "Pour un seuil critique α ∈[",min(seuil),",",max(seuil),"]"))
}
}
if(Rcarré==RcPareto){
EMV.gamma<-(1/n)*sum(log(x))# EMV de gamma.
ppareto<-function(x,EMV.gamma){
1-x^(-1/EMV.gamma)} #Fonction de répartition de loi de type Pareto
ks.test(x, "ppareto",EMV.gamma)# Test de Kolmogrov-Smirnov
p.valeur<-ks.test(x, "ppareto",EMV.gamma)$p.value
if(all(alpha>p.valeur)){
print(paste("x n'appartient à aucune loi du programme."))
} else {
seuil <- c()
i <- 1
while (i <= length(alpha)) {
if (alpha[i] < p.valeur) {
seuil <- c(seuil, alpha[i])
}
i <- i + 1
}
```

```
print(paste("x est suit la loi de type Pareto avec de paramètre gamma=" ,EMV.gamma,
" avec R²=" ,Rcarré, "Pour un seuil critique α ∈[",min(seuil),",",max(seuil),"]"))
}
}
if(Rcarré==RcNorm){
EMV.mu<-mean(x)# EMV de mu
EMV.sigma<-sd(x)# EMV de sigme
ks.test(x, "pnorm",EMV.mu,EMV.sigma)# Test de Kolmogrov-Smirnov
p.valeur<-ks.test(x, "pnorm",EMV.mu,EMV.sigma)$p.value
if(all(alpha>p.valeur)){
print(paste("X n'appartient à aucune loi du programme."))
} else {
seuil <- c()
i <- 1
while (i <= length(alpha)) {
if (alpha[i] < p.valeur) {
seuil <- c(seuil, alpha[i])
}
i <- i + 1
}
print(paste("x suit la loi Normale avec de paramètres mu=" ,EMV.mu,"sigma=",EMV.sigma,
" avec R²=" ,Rcarré, "Pour un seuil critique α ∈[",min(seuil),",",max(seuil),"]"))
}
}
if(Rcarré==RcCauchy){
EMV.x0<-mean(x)# EMV de x0
EMV.alpha<-sqrt((1/n)*sum(x-EMV.x0)^2) # EMV de alpha
```

```
ks.test(x, "pcauchy",EMV.x0,EMV.alpha)# Test de Kolmogrov-Smirnov
p.valeur<-ks.test(x, "pcauchy",EMV.x0,EMV.alpha)$p.value
if(all(alpha>p.valeur)){
print(paste("x n'appartient à aucune loi du programme."))
} else {
seuil <- c()
i <- 1
while (i <= length(alpha)) {
if (alpha[i] < p.valeur) {
seuil <- c(seuil, alpha[i])}
i <- i + 1
}
print(paste("x suit la loi Cauchy avec des paramètres x0=" ,EMV.x0,"a=",EMV.alpha, "
avec R²=" ,Rcarré, "Pour un seuil critique α ∈[",min(seuil),",",max(seuil),"]"))
}}
}
```

3.2.3 Simulation et stabilitée

Simulation d'une v.a suivant une loi de Poisson

Soit X un échantillon de données constitué d'une série de 50 valeurs numériques réelles simulées avec R de manière aléatoire qui suit à une loi de poisson de paramètre $\lambda = 20$.

Simuler une v.a x sous R :

```
n=50
lambda=20
x<-rpois(n,lambda)
```

Exécuter notre modèle « lois.test » :

```
lois.test(x)
```

Résultats illustratifs obtenus :

"x suit la loi Normale de paramètres mu= 19.4 sigma= 4.56249126362037 avec R^2= 0.969603064157088 pour un seuil critique $\alpha \in$[0.001 , 0.461]"

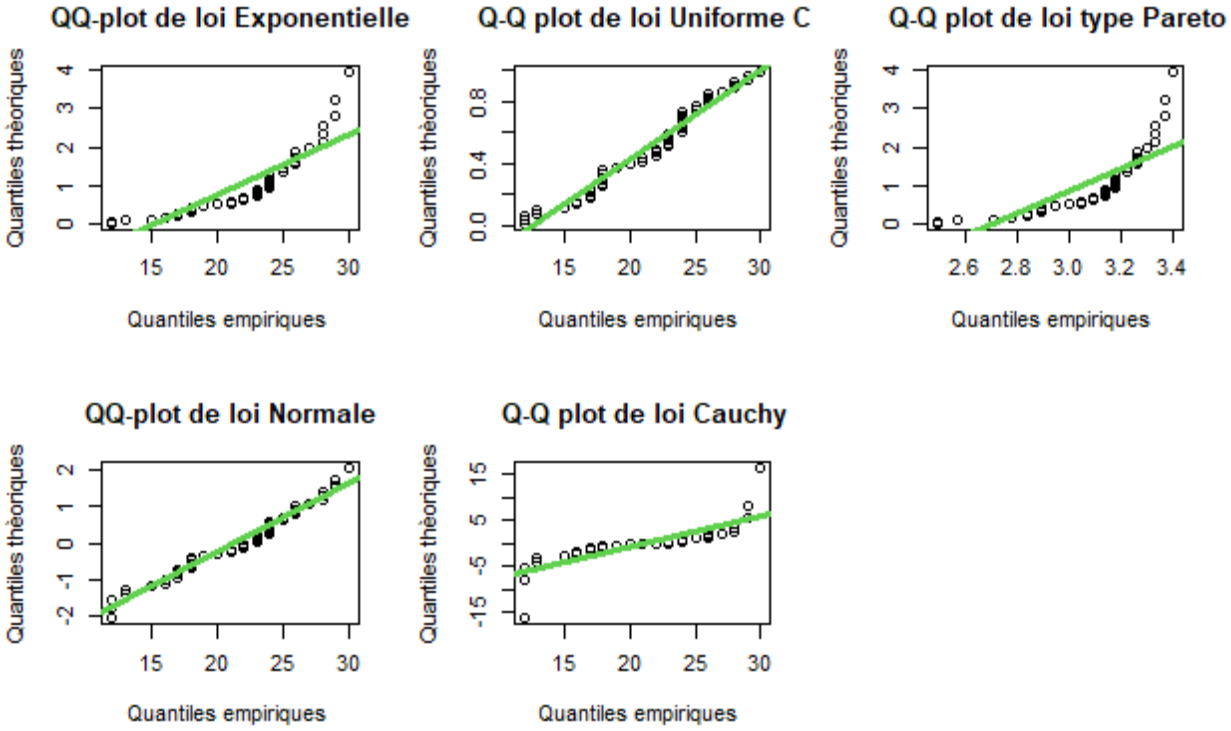

FIG. 3.1 – QQ-plot du modèle « **lois.test** » pour un échantillon de loi de Poisson de $\lambda = 20$

Commentaire : Nous constatons clairement d'après l'exécution du modèle d'ajustement que l'échantillon généré à partir d'une loi de Poisson de paramètre $\lambda > 18$, suit une loi Normale $N(19.4, 4.56)$, ce qui explique la convergence de la loi Poisson vers la loi normale (pour $\lambda > 18$, voir théorème 1.3.3). Ceci rend notre modèle stable par convergence. On voit clairement sur la figure 3.1 que le qqplot de la loi Normale présente une allure linéaire parfaite.On voit clairement que l'ajustement est accepté pour $\alpha \in [0.001, 0.461]$.

Simulation d'une v.a suivant une loi de type Pareto

On reprend l'échantillon d'observations donné dans le cas pratique 2.2.3 qui suit une loi de type Pareto (de paramètre γ inconnue).

Exécuter notre modèle « lois.test » :

lois.test(x)

Résultats illustratifs obtenus :

"x est suit la loi de type Pareto de paramètre gamma= 1.43834079761937 avec R^2= 0.986372430035738 pour un seuil critique $\alpha \in$[0.001 , 0.856]"

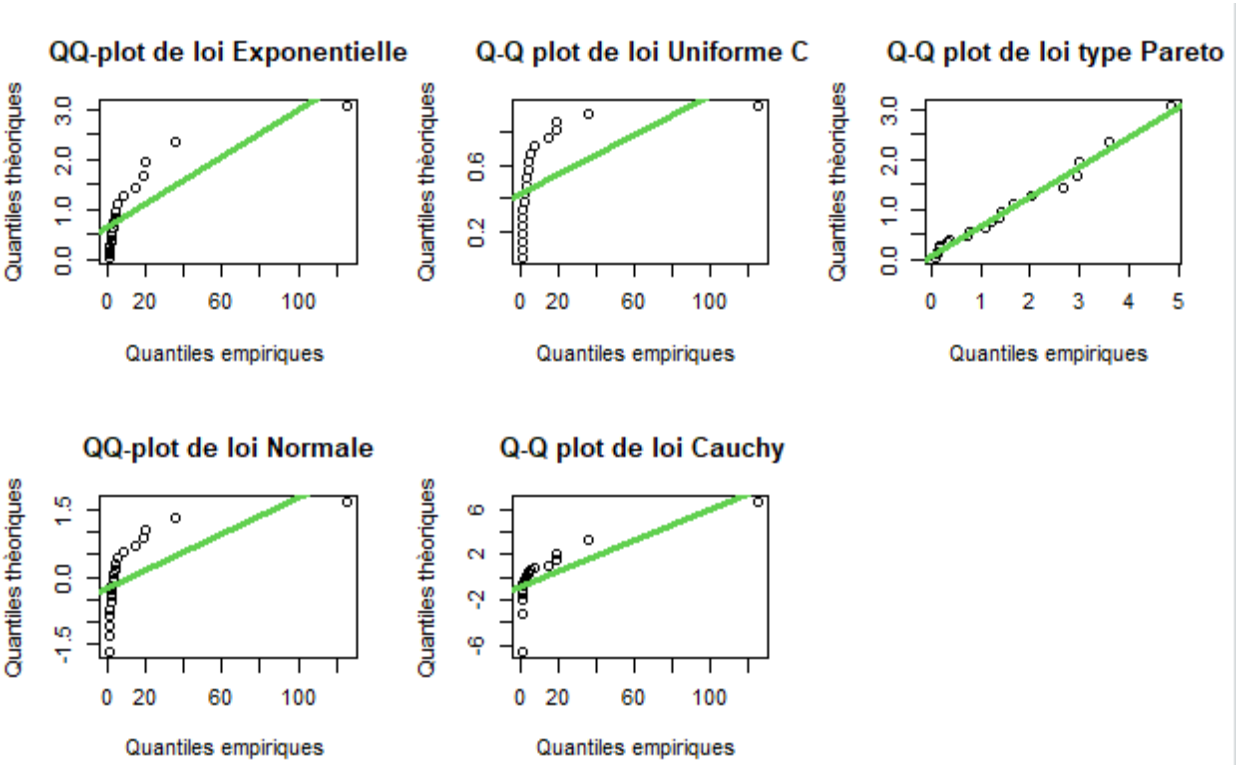

FIG. 3.2 – QQ-plot du modèle « **lois.test** ».pour un échantillon de loi de type Pareto avec γ inconnue

Commentaire : Après l'exécution du modèle, nous constatons la confirmation d'ajustement des observations pour une loi de type Pareto. Ainsi un estimateur du paramètre γ est donné ($\gamma = 1.43$). Nous avons remarqué que notre modèle fonctionne bel est bien même

dans le cas de taille d'échantillon réduit selon la linéarité des nuages des points aperçus sur la figure 3.2 de la loi de type Pareto.

On voit clairement que l'ajustement est accepté pour $\alpha \in [0.001, 0.856]$.

Simulation d'une v.a suivant une loi Uniforme continue.

Soit un échantillon de données constitué d'une série de 5000 valeurs numériques réelles simulées avec R de manière aléatoire suitvant une loi Uniforme continue de paramètres a et b.

Simuler une v.a x sous R :

a=40

b=70

n=5000

x<-runif(n,a,b)

Exécuter notre modèle « lois.test » :

lois.test(x)

Résultats illustratifs obtenus :

"x suit la loi Uniforme continue de paramètres a= 40.0005461648107 b= 69.9979029875249 avec R^2= 0.999263496127875 pour un seuil critique $\alpha \in$ [0.001 , 0.624]"

Commentaire : Après l'exécution du modèle, nous constatons la confirmation d'ajustement des observations pour une loi Uniforme continue. Ainsi un estimateur des paramètres a et b est donné ($a = 40.00, b = 69.99$). Selon la linéarité des nuages des points aperçus sur la figure 3.3 du qq-plot de la loi de Uniforme continue. Nous confirmons que notre modèle fonctionne bel est bien même dans le cas d'un d'échantillon de taille trés grande

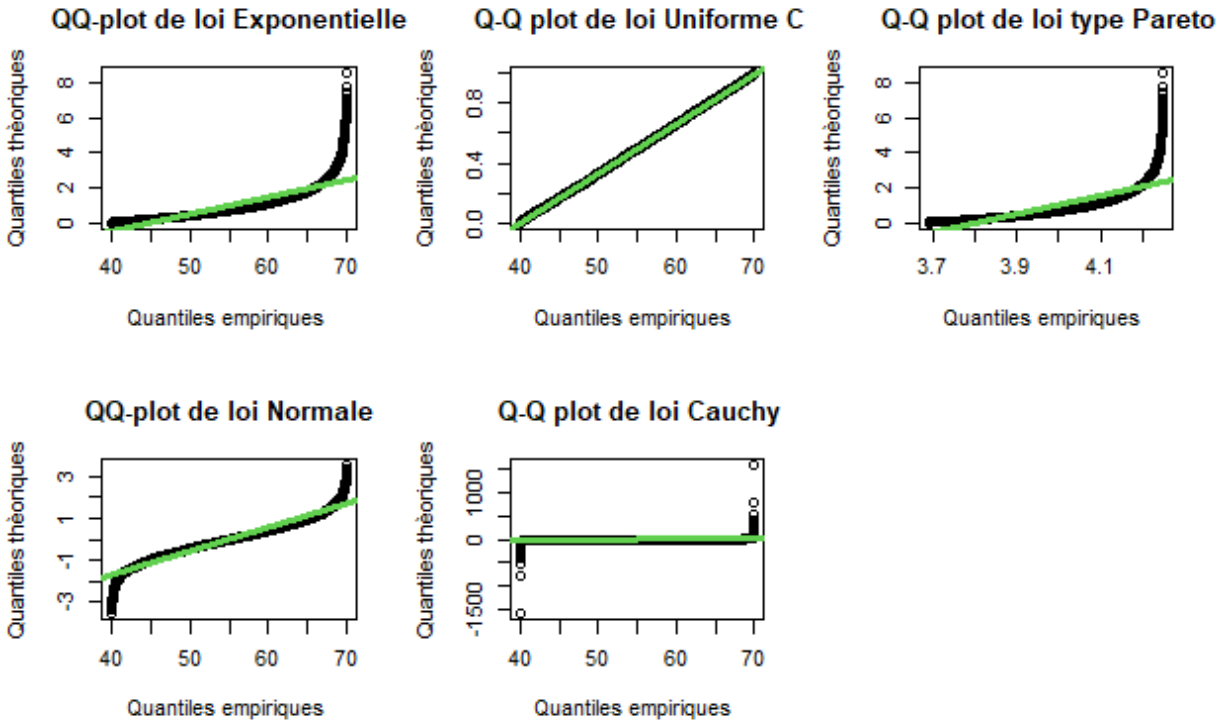

FIG. 3.3 – QQ-plot du modèle « **lois.test** » pour un échantillon de loi de Uniforme continue de paramètres a et b inconnue

3.3 Application rèelle

3.3.1 Représentation du jeu de données

Nous exécutons le code suivant pour charger le jeu de donnés de Covid19 du mois de janvier 2020 au mois de juin 2022 dans le monde et on a téléchargé sur le net :

>covid =read.table(file.choose(),header=FALSE,sep=",") # chargement de données

>View(covid) # Affichage des données (voir la figure 3.4)

Nous affichons la colonne **V6** qui représente l'echantillon à modéliser qui est le nombre de cas affecté dans le monde par jour de 2020-01-22 à 2022-07-07. (Voir le figure 3.5)

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	OWID_WRL	NA	World	2020-01-22	557	0	NA	17	0	NA	0.071	0.000	
2	OWID_WRL	NA	World	2020-01-23	657	100	NA	18	1	NA	0.083	0.013	
3	OWID_WRL	NA	World	2020-01-24	944	287	NA	26	8	NA	0.120	0.036	
4	OWID_WRL	NA	World	2020-01-25	1437	493	NA	42	16	NA	0.182	0.063	
5	OWID_WRL	NA	World	2020-01-26	2120	683	NA	56	14	NA	0.269	0.087	
6	OWID_WRL	NA	World	2020-01-27	2929	809	338.857	82	26	9.286	0.372	0.103	
7	OWID_WRL	NA	World	2020-01-28	5580	2651	717.571	131	49	16.286	0.709	0.337	
8	OWID_WRL	NA	World	2020-01-29	6169	589	801.714	133	2	16.571	0.783	0.075	
9	OWID_WRL	NA	World	2020-01-30	8237	2068	1082.857	171	38	21.857	1.046	0.263	

Showing 1 to 9 of 898 entries, 67 total columns

FIG. 3.4 – Tableau des données de covid 19

```
> x<-covid[2:898,c("V6")]
> x
  [1]     100     287     493     683     809    2651     589    2068    1690    2111
 [11]    4749    3100    4012    3745    3162    3594    2731    3031    2609    2043
 [21]     418   15152    6528    2143    2183    2035    1882     500     561     630
 [31]    1762     382     568     854     973    1344    1424    1871    2379    1980
 [41]    2612    2322    2711    3930    4131    3854    4323    4786    7504    6754
 [51]   13196   10888   11233   14567   15174   17560   27087   29532   32433   34213
 [61]   42555   41900   51479   60815   63930   69583   56518   65009   78441   81708
 [71]   82445   84164   77333   72178   74515   72322   81995   86765   86524   75453
 [81]  119052   71732   84449   77790   94424   87830   78233   76713   75923   76215
 [91]   81952   84313   94071   83493   70870   70941   75801   78826   83847   89016
[101]   78546   74176   77394   79623   90531   90250   91939   84550   75267   76053
[111]   85303   84363   95656   96354   94072   78449   89056   95772  105411  106117
```

FIG. 3.5 – Echantillon de nombre de cas effectés par jours, taille n=898(nombre de jours)

3.3.2 Exécution du modèle

Nous exécutons notre modèle, ensuite on l'appliquant sur l'échantillon x dans le but de trouver la loi qui modélise le mieux ce dernier parmi les cinq lois intégrées dans le modèle sachant que nous n'avons aucune informations statistiques sur cette distribution d'observations (ni la loi ni ses paramètres).

> lois.test(x)

[1] "X n'appartient à aucune loi du programme."

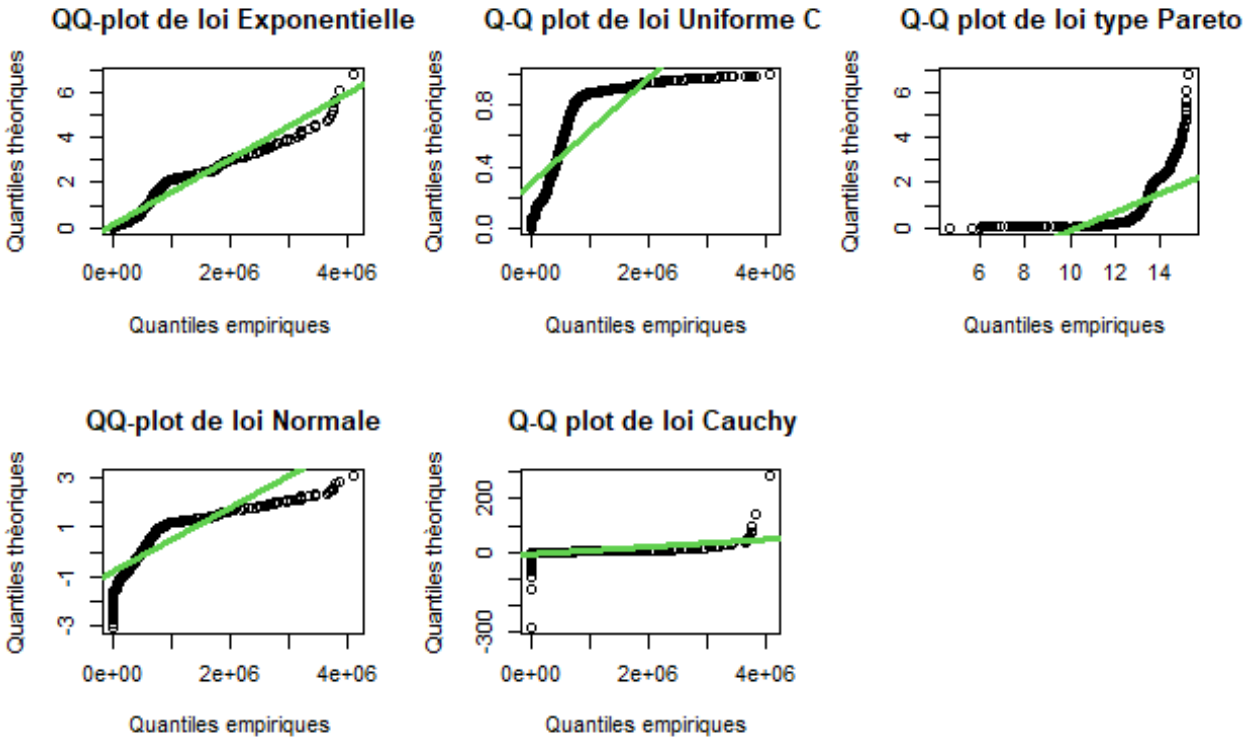

FIG. 3.6 – QQplot du modèle sur l'echantillon x de nombre de cas effectée par jours

3.3.3 Discussions des résultats

L'exécution du modèle entraîne le résultat : "X n'appartient à aucune loi du programme."

- Nous avons remarqué d'après le graphique du qqplot du modèle que l'échantillon réelle présente une allure linéaire remarquable avec quelques valeurs aberrantes bien claires selon le qq-plot de la loi exponentielle.(voir figure 3.7).
- Le modèle a donnée un coefficient de détermination $R^2 = 0.93$ qui est une valeur parfaite pour que l'échantillon x soit expliqué par la loi exponentielle.
- La détermination de paramètre de la loi exponentielle a été donné par le modèle ; $\lambda_{EMV} = 1.63e - 06$.

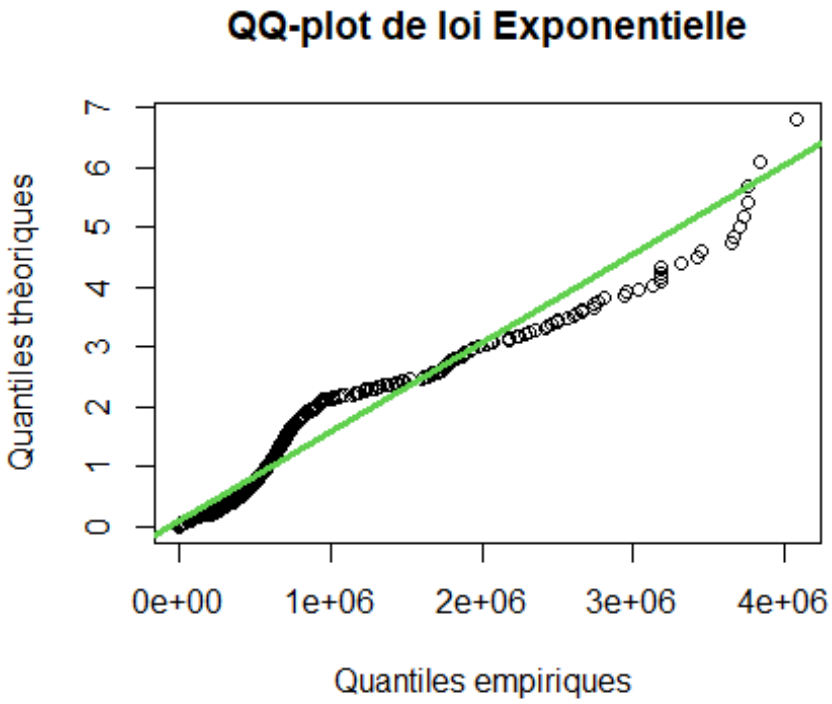

FIG. 3.7 – QQplot de l'echantillon x de nombre de cas effectée par jours

– L'exactitude de l'estimation de λ par notre modèle a été conclue empiriquement par le coefficient de la droite de régression $\lambda_{reg} = 1.48e-06$ qui est très proche de λ_{EMV}.
– Relativement à la sensibilité du test de Kolmogorov-Smirnov, le modèle ne valide pas l'ajustement de l'échantillon par la loi exponentielle $\zeta(898, 1.63e-06)$ car la distance de Kolmogorov-Smirnov ($Dn = 0.11$) a pris une valeur relativement grande par rapport aux valeurs aberrantes de l'échantillon d'observations.

Conclusion

Présentation du modèle

. Les méthodes étudiées dans ce mémoire fournissent des outils solides pour aborder la problématique d'ajustement de données par une loi de probabilité permettant de créer un algorithme sous forme de fonction sous R contenant cinq lois d'ajustement,

. Notre nouvel modèle a été nommée **lois.test**

Importance du modèle

. L'ajustement d'un échantillon de données réelles est une étape cruciale de l'analyse statistique.

. Ce mémoire a contribué à la compréhension et à la comparaison des méthodes d'ajustement d'échantillon de loi inconnue, en mettant en évidence l'utilité du QQ-plot, de regression linéaire simple, de l'EMV et du test de KS.

. L'utilisation conjointe de ces méthodes ont permet une évaluation complète de l'ajustement de l'échantion de données par l'une des cinq lois du modèle.

Déroulement du modèle

. La fonction lois.test(x) charge l'échantillon d'observations x

. Un qq-plot entre l'échantillon théorique et empirique est tracé par l'utilisation de l'instruction de la régression linéaire simple pour chaque loi contenant dans le modèle (Exp, Norm, Pareto, Unif,Cauchy)

. Le coefficient de détermination de validation pour chaque loi est calculé

. Un chois de loi est effectué selon le coefficient de validation le plus élevé notée F_0

. Les paramètres de loi choisie sont calculés par méthode de maximum de vraisemblance

. Un test d'hypothèse de KS est effectué pour accepter ou rejeter la loi F_0, c'est-à-dire accepter d'ajuster l'échantillon de données par cette loi F_0 ou non, cela en parcourant un niveau de confiance $\alpha \in [0, 1]$.

Avantages du modèle

. Le modèle **lois.test** est un modèle fiable et donne un ajustement parfait

. Il est stable par convergence de loi et par rapport à la taille de l'échantillon

. Il présente une estimation exacte des paramètres des lois selon les trois méthodes utilisées

. Il est sévère et sensible aux valeurs aberrantes et extrêmes à cause du test de KS

. Il valide la loi que lorsque l'ajustement et vrais pour chaque observations.

Lacunes du modèle

. Le modèle **lois.test** contient que cinq loi continues, ce qui le rend non exhaustive

. Concernant la loi Pareto, lorsque les valeurs sont négatives ou nulles le QQplot ne s'affiche pas.

. Le modèle lois.test n'ajuste pas les lois discrètes. Nous avons essayé quelques un parmi eux et avons trouvé que le test de KS n'est pas vraiment le meilleur test à utiliser car il affiche 'warning message'

. Le test de Kolmogorov-Smirnov est un test très sensible et sévère et rejette souvent la loi à cause des valeurs aberrantes ou des valeurs extrêmes qui rendent la distance maximale entre F_0 et toutes observations "grande" et non acceptables même si la autres obsevations sont proche de F_0

. Le modèle lois.test ajuste que des lois usuelles

Perspectifs

. On propose de rajouter plus de lois continues.

. On propose aussi d'intégrer les lois discrètes. Par contre c'est le test de khi-deux qui sera le plus compatible.

. On propose d'intégrer plusieurs tests en même temps et générer plus de décisions pour avoir plus de souplesse dans nos choix par exemple le test de khi-deux, etc

. On propose d'élargir ce travail pour donner des ajustements pour des lois de probabilités quelconques et pas uniquement usuelles

. Possibilité d'intégrer des outils qui détectent et traitent les valeurs aberrante sans perdre de l'information

. Possibilité d'intégrer des outils qui examine l'existence des valeurs extrêmes pour mieux choisir le test à considérer.

. Question ouverte : est ce qu il est possible d'élargir ce modèle au cas multidimensionnelle, où le QQplot n'est pas pratique ? ce qui nous emmène à trouver d'autres techniques plus commodes.

Annexe B : Logiciel R

R est un langage de programmation interactif interprété et orienté objet contenant une très large collection de méthodes statistiques et des facilités graphiques importantes.
C'est un clône gratuit du logiciel S-Plus commercialisé par Math Soft et d´enveloppé par Statistical Sciences autour du langage S (conçu par les laboratoires Bell).
Initié dans les années 90 par Robert Gentleman et Ross Hakka (Département de Statistique, Université d'Auckland, Nouvelle-Zélande), auxquels sont venus depuis s'ajouter de nombreux chercheurs, le logiciel R constitue aujourd'hui un langage de programmation intégré d'analyse statistique. Le site Internet de la "R core-development Team", http ://www.r-project.org , est la meilleure source d'informations sur le logiciel R. Vous pourrez y trouver les différentes distributions du logiciel, de nombreuses bibliothèques de fonctions et des documents d'aide. Des bibliothèques supplémentaires sont aussi disponibles sur le « compréhensive R archive network » (CRAN) http ://lib.stat.cmu.edu/R/CRAN/.

Bibliographie

[1] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

[2] Chesneau, C. (2016). Sur l'adéquation à une loi de probabilité avec R.

[3] Chesneau, C. (2017). Sur l'Estimateur du Maximum de Vraisemblance (emv).

[4] Ferignac, P. (1962). Test de Kolmogorov-Smirnov sur la validité d'une fonction de distribution. Revue de statistique appliquée, 10(4), 13-32.

[5] Marin, J. M. (2005). Initiation au logiciel R. Université Paris Dauphine.

[6] Marsaglia, G. (2004). Evaluating the normal distribution. Journal of Statistical Software, 11, 1-11.

[7] Meddi, F. (2014). Estimation des measures de risqué pour les distributions à queue lourde (Doctoral dissertation, Université Mohamed Khider Biskra).

[8] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

[9] Robert, C. P., Casella, G., & Casella, G. (1999). Monte Carlo statistical methods (Vol. 2). New York : Springer.

[10] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions technip.

[11] Yan, X., & Su, X. (2009). Linear regression analysis : theory and computing. world scientific.

Résumé

L'objet de ce mémoire est la création et la généralisation d'un algorithme sous R permettant d'ajuster n'importe quel échantillon provenant d'une distribution inconnue F. En utilisant des techniques statistiques avancées, tels que le qqplot, la régression linéaire simple, l'estimation par maximum de vraisemblance et le test de Kolmogorov–Smirnov, notre modèle «lois.test» vise à ajuster des données observées par l'une des cinq lois intégrées, estimer ses paramètres et fournir une meilleure compréhension de la distribution sous-jacente.

Mots clés : lois.test, qqplot, régression linéaire simple, estimation par maximum de vraisemblance, test de Kolmogorov–Smirnov.

Abstract

The object of this memory is the creation and the generalization of an algorithm under R allowing to adjust any sample coming from an unknown distribution F. By using advanced statistical techniques, such as qqplot, linear regression simple, maximum likelihood estimation and the Kolmogorov–Smirnov test, our model" lois.test" aims to fit observed data to one of the five built-in laws, estimate its parameters and provide a better understanding of the distribution underlying.

Keywords: lois.test, qqplot, simple linear regression, maximum likelihood estimation, Kolmogorov–Smirnov test.

الملخص

الهدف من هذه المذكرة هو إنشاء وتعميم خوارزمية باستعمال برنامج R تسمح بنمذجة أي عينة قادمة من توزيع غير معروف F. باستخدام تقنيات إحصائية متقدمة، مثل qqplot ، الانحدار الخطي البسيط، تقدير الاحتمالية القصوى و اختبار كولموغوروف سميرنوف، يهدف نموذجنا "lois.test"إلى ملاءمة البيانات المرصودة مع أحد القوانين الخمسة المتضمنة ، وتقدير معلماته وتوفير فهم أفضل للتوزيع الأساسي.

الكلمات المفتاحية: lois.test، qqplot، الانحدار الخطي البسيط، تقدير الاحتمالية القصوى، اختبار كولموغوروف سميرنوف.