République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

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

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : Statistique

Par

ZERNADJI Loubna

Titre :

Estimation des queues de distributions

Membres du Comité d'Examen :

Pr.	CHERFAOUI Mouloud	UMKB	Président
Pr.	NECIR Abdelhakim	UMKB	Encadreur
Dr.	ZOUAOUI Nour Elhouda	UMKB	Examinatrice

Juin 2023

Dédicace

Je dédie ce humble travail

À mes chers parents :

Tout au long de ma vie, vous avez été le guide le plus aimant, les mentors les plus inspirants. Leurs conseils et leurs soutiens m'ont aidé à façonner qui je suis aujourd'hui, et je ne peux exprimer assez ma gratitude pour tout ce qu'ils ont fait pour moi.

À mes frères et sœurs, Oussama, Raouf, Linda, Wafa merci d'avoir été mes meilleurs compagnons, mes alliés et mes protecteurs. Vous êtes les personnes qui ont le mieux compris mes joies, mes peines et mes rêves.

À ma merveilleuse famille Zernadji, qui ont toujours été une source de soutien et générosité et gentillesse et d'amour inébranlable

À mes amies : Chaima, Imane, Khaoula qui ont partagé avec moi des moments de joie, de rire, et de défi et votre amitié est une source de soutien inestimable.

REMERCIEMENTS

En commençant, je tiens à exprimer ma gratitude envers **Allah** le Tout-Puissant pour toutes les bénédictions qu'Il m'a accordées tout au long de ma vie.

Je tiens à exprimer mes sincères remerciements au Professeur **Necir Abdelhakim**, en tant qu'encadreur de mon mémoire de Master, il a exercé un rôle essentiel dans la réussite de ce travail. Son expertise, son expérience et sa patience ont été déterminants pour m'aider à comprendre la démarche de la recherche scientifique. Ses conseils avisés m'ont poussé à fournir le meilleur de moi-même tout au long de ce travail, et pour cela je lui suis très reconnaissante.

Je tiens également à exprimer ma reconnaissance envers les membres du comité d'examen le Pr. **Cherfaoui Mouloud** et Dr. **Zouaoui Nour Elhouda** pour leur temps et leur expertise et pour avoir accepté d'évaluer ce travail.

Je suis reconnaissant aussi à tous les professeurs qui ont accompli leur devoir avec sincérité et ont soutenu les étudiants pendant leur période universitaire. Je suis particulièrement reconnaissante aux Professeurs **Yahia Djabrane** et **Benatia Fatah** pour leurs conseils et leur soutien.

Table des matières

Table des figures

Liste des tableaux

Introduction

Connaissez-vous les queues de distributions ? Ces régions souvent négligées des statistiques qui pourtant peuvent causer de grandes surprises lorsqu'on les explore ? Si vous êtes curieux de comprendre comment prédire les évènements les plus extrêmes, alors ce mémoire est pour vous. Préparez-vous à découvrir des outils statistiques avancés et des exemples concrets pour appréhender les corps gras de ces distributions. L'estimation des queues de distributions est l'un des domaines les plus importants de la statistique. Il s'agit d'un domaine qui a une longue histoire, qui a vu de nombreuses percées et qui est toujours en évolution aujourd'hui et avant de plonger dans les détails de l'estimation des queues de distributions, il convient de donner un aperçu historique de la question. Au fil des siècles, le calcul des probabilités s'est développé et a donné naissance à de nouvelles branches de la statistique. Ce n'est que dans les années 1920 que la distribution des queues a commencé à intéresser les scientifiques.

Les queues de distributions sont importantes dans de nombreuses applications pratiques, car elles peuvent concerner des phénomènes rares mais critiques tels que les crises financières, les maladies rares, etc. Où l'estimation de ces queues est une branche de la statistique qui s'intéresse aux cas où les observations extrêmes de la distribution font l'objet d'une analyse particulière.

Au terme de ce mémoire, nous pourrons répondre aux questions suivantes :

1. Comment prédire le comportement des valeurs extrêmes de la variable étudiée ?

2. Comment estimer la probabilité d'un évènement extrême ?
3. Comment est estimée les queues de distributions en cas de manque de données ?
4. Comment utiliser les résultats pour prendre des décisions éclairées et réduire les risques ?

Nous exposons notre travail en quatre chapitres. Le premier chapitre présente les statistiques d'ordre, qui sont la base de l'estimation des queues de distributions. Le deuxième chapitre traite la théorie des valeurs extrêmes, qui fournit une approche alternative pour estimer les évènements rares, ainsi que certaines méthodes courantes pour estimer l'indice de queue et pour estimer les quantiles extrêmes, et la fin de ce chapitre, nous parlerons de certaines méthodes pour déterminer le nombre de valeurs extrêmes. Le troisième chapitre traite les données incomplètes et des manières possibles d'en tenir compte lors de l'estimation des queues de distributions. Enfin, le dernier chapitre met en pratique ces méthodes sur des données réelles, permettant de mesurer leur efficacité et leur pertinence dans un contexte concret.

Chapitre 1

Statistique d'ordre

Dans certaines études, on rencontre souvent des valeurs très grandes ou très petites qui affectent l'étude, où l'échantillon est réordonné pour faciliter l'identification de ces valeurs qui est appelée les valeurs extrêmes. Dans ce chapitre, nous parlerons simplement sur la statistique d'ordre qui joue un rôle important dans la théorie des valeurs extrêmes.

1.1 Définition de la statistique d'ordre

Définition 1.1.1 *On appelle statistique d'ordre associée à l'échantillon* $(X_1, ..., X_n)$ *la suite ordonée (au sens croissant) notée* $(X_{1,n}, ..., X_{n,n})$, *tel que*

$$X_{1,n} \leq \cdots \leq X_{n,n}.$$

– Mettons un vecteur de rangs $(R(1), ..., R(n))$ avec

$$R(m) = \sum_{k=1}^{n} \mathbb{1}_{\{X_m \geq X_k\}}.$$

– Le rang de X_m égal à k c'est à dire $X_m = X_{k,n}$, cela peut être écrit en terme

mathématique comme suit :

$$\{R(m) = k\} = \{X_m = X_{k,n}\}, \text{avec } m = 1, ..., n \text{ et } k = 1, ..., n.$$

– $(r(1), ..., r(n))$ sont des permutations de valeurs $\{1, ..., n\}$ correspondant $(R(1), ..., R(n))$.

On a donc : $(X_{1,n}, ..., X_{n,n}) = (X_{\delta(1)}, ..., X_{\delta(n)})$ tel que $\delta(r(k)) = k$.

Théorème 1.1.1 *Soit $X_1, ..., X_n$ des variables aléatoires (v.a) indépendantes et de fonction de répartition F. Soit $U_1, ..., U_n$ des v.a indépendantes de loi uniforme $[0, 1]$, alors $(F^{-1}(U_{1,n}), ..., F^{-1}(U_{n,n}))$ à même loi que $(X_{1,n}, ..., X_{n,n})$ et si F continue, alors*

$$(U_{1,n}, ..., U_{n,n}) \overset{\mathcal{D}}{=} (F(X_{1,n}), ..., F(X_{n,n})) \ p.s.$$

La distribution de la statistique d'ordre permet de déterminer la probabilité q'une valeur soit valeur extrême, d'autre part sa densité permet de calculer les valeurs les plus extrêmes possibles. Dans la partie suivante, nous verrons la loi de : la $1^{ère}$ statistique d'ordre $(X_{1,n} = \min(X_1, ..., X_n))$, la dernière statistique d'ordre $(X_{n,n} = \max(X_1, ..., X_n))$ et la $i^{ème}$ statistique d'ordre $X_{i,n}$.

1.2 Loi de la statistique d'ordre

Soit $X_1, ..., X_n$ une suite des v.a' i.i.d, de fonction de répartition F_X et de densité f_X. On note par $F_{X_{1,n}}$, $F_{X_{n,n}}$, $F_{X_{i,n}}$ les distributions des v.a's $X_{1,n}, X_{n,n}$ et $X_{i,n}$ respectivement et $f_{X_{1,n}}$, $f_{X_{n,n}}$, $f_{X_{i,n}}$ les densités correspondantes.

1.2.1 Loi de $X_{1,n}$

Distribution de $X_{1,n}$

Nous avons

$$\begin{aligned} F_{X_{1,n}}(x) &= P(X_{1,n} \leq x) = P(\min(X_1, ..., X_n) \leq x) \\ &= 1 - P(\min(X_1, ..., X_n) > x) = 1 - P(X_1 > x, ..., X_n > x). \end{aligned}$$

Car les v.a sont indépandentes et de même loi F_X alors

$$F_{X_{1,n}}(x) = 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - (P(X > x))^n = 1 - (1 - F_X(x))^n.$$

De plus, on sait que la densité $f_X(x) = \frac{\partial F_X(x)}{\partial x}$ alors, la densité de $X_{1,n}$ est donnée par : $f_{X_{1,n}}(x) = n f_X(x)(1 - F_X(x))^{n-1}$.

1.2.2 Loi de $X_{n,n}$

Distribution de $X_{n,n}$

De la même façon de $X_{1,n}$ on trouve la distribution de $X_{n,n}$:

$$\begin{aligned} F_{X_{n,n}}(x) &= P(X_{n,n} \leq x) = P(\max(X_1, ..., X_n) \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) = (F_X(x))^n. \end{aligned}$$

La densité correspondante est donnée par : $f_{X_{n,n}}(x) = n f_X(x)(F_X(x))^{n-1}$.

1.2.3 Loi de $X_{i,n}$

Distribution de $X_{i,n}$

Nous avons

$$\begin{aligned} F_{X_{i,n}}(x) &= P(X_{i,n} \leq x) = P(\cup_{k=i}^{n}\{k(X_k \leq x) \cap (n-k)(X_k > x)\}) \\ &= \sum_{k=i}^{n} P(k(X_k \leq x) \cap (n-k)(X_k > x)) \\ &= \sum_{k=i}^{n} C_n^k (P(X_k \leq x))^k (P(X_k > x))^{n-k} \\ &= \sum_{k=i}^{n} C_n^k (F_X(x))^k (1 - F_X(x))^{n-k}. \end{aligned}$$

– Passons maintenant à la formule de la densité de $X_{i,n}$:

Il existe $i-1$ de X_i sont inférieurs à x et une seule X_i entre x et $x+dx$ et $n-i$ sont supérieurs à x, donc on trouve :

$$P(x \leq X_{i,n} \leq x+dx) = \frac{n!}{(i-1)!(n-i)!} P(X_i \leq x)^{i-1} P(x \leq X_i \leq x+dx) P(X_i > x)^{n-i},$$

et par conséquent on a :

$$f_{X_{i,n}}(x) = \lim_{dx \to 0} \frac{P(x \leq X_{i,n} \leq x+dx)}{dx} = n C_{n-1}^{i-1} (F_X(x))^{i-1} f_X(x) (1 - F_X(x))^{n-i}.$$

L'étude de la distribution et de la densité de la statistique d'ordre dans la théorie des valeurs extrêmes permet de développer des modèles mathématiques pour prévoir la probabilité d'occurrence de valeurs extrêmes, ces modèles peuvent être appliqués dans divers domaines, (tels que la finance et l'assurance) pour une aide précieuse dans la prise de décisions éclairées. Cependant, la distribution dont nous avons traité dans cette section est une distribution dégénérée (i.e : lorsque $n \to \infty$ la limite de F soit 0 ou 1), et cela ne nous donne pas d'informations sur la distribution des valeurs

extrêmes. La question est donc de savoir quelle est la distribution non dégénérée des valeurs extrêmes ?, sa réponse est dans le deuxième chapitre2.1.

1.3 Fonction de densité de probabilité conjointe

On définie la densité de $(X_{1,n}, ..., X_{n,n})$ par

$$f_{(X_{1,n},...,X_{n,n})}(x_{1,n}, ..., x_{n,n}) = n!\prod_{i=1}^{n} f(x_i) \text{ tel que } x_i \in \mathbb{R} \text{ pour } i = 1, \ldots, n.$$

Distribution du couple $(X_{i,n}, X_{j,n})$

Soit $X_1, ..., X_n$ une suite de v.a' i.i.d de distribution F_X et de densité $f_X(x)$.

Pour $1 \leq i < j \leq n$ on a : $F_{(X_{i,n},X_{j,n})}(x, y) = P((X_{i,n} \leq x) \cap (X_{j,n} \leq y))$.

– **1^{er} cas :** si $x \geq y$ alors, $F_{(X_{i,n},X_{j,n})}(x, y) = P(X_{j,n} \leq y) = F_{X_{j,n}}(y)$.

– **$2^{ème}$ cas :** $x < y$

C'est à dire au moint j de $X_1, ..., X_n$ sont inférieurs à y et au moint i de $X_1, ..., X_n$ sont inférieurs à x. Alors,

$$\begin{aligned} F_{(X_{i,n},X_{j,n})}(x, y) &= \sum_{k=j}^{n}\sum_{s=i}^{k} P(s \text{ de } (X_1, ..., X_n) \leq x \text{ et } k \text{ de } (X_1, ..., X_n) \leq y). \\ &= \sum_{k=j}^{n}\sum_{s=i}^{k} \frac{n!}{s!(k-s)!(n-k)!}(F_X(x))^s(F_X(y) - F_X(x))^{k-s}(1 - F_X(y))^{n-k}. \end{aligned}$$

La densité correspondante est donnée par :

$$\begin{aligned} f_{(X_{i,n},X_{j,n})}(x, y) = \tfrac{n!f_X(x)f_X(y)}{(i-1)!(j-i-1)!(n-j)!}&(F_X(y) - F_X(x))^{j-i-1} \\ &\times (F_X(x))^{i-1}(1 - F_X(y))^{n-j}. \end{aligned}$$

– Il est possible de calculer la densité de $(X_{i,n}, X_{j,n})$ à partir de la densité de $(X_{1,n}, ..., X_{n,n})$ et ça par calculer l'integral de $f_{(X_{1,n},...,X_{n,n})}$ par rapport à $(X_{1,n}, ..., X_{i,n})$,

$(X_{i+1,n}, ..., X_{j-1,n})$ et $(X_{j+1,n}, ..., X_{n,n})$ (pour plus de détailles voir la réf [1, pages [16-17]).

1.4 Densité conditionnelle

La densité conditionnelle de $X_{i,n}$ sachant que $X_{j,n} = y$ est donnée par :

$$f_{(X_{i,n}/X_{j,n})}(x/y) = \frac{f_{(X_{i,n},X_{j,n})}(x,y)}{f_{X_{j,n}}(y)}$$
$$= \begin{cases} \frac{(j-1)!}{(i-1)!(j-i-1)!}(F_X(x))^{i-1} f_X(x) & \\ \qquad \times (F_X(y))^{1-j}(F_X(y) - F_X(x))^{j-i-1} & \text{si } x < y \\ 0 & \text{sinon} \end{cases}$$

1.5 Moments de la statistique d'ordre

Soit $X_1, ..., X_n$ une suite de v.a de distribution F et $U_1, ..., U_n$ de loi uniforme $(0, 1)$.

Le moment d'ordre m de la statistique d'ordre $X_{i,n}$ est donné par

$$E(X_{i,n}^m) = \int_{\mathbb{R}} x^m f_{X_{i,n}}(x)dx.$$

Il peut être écrit d'une autre manière en fonction de $U_{i,n}$, d'après le théorème 1.1.1

$$E(X_{i,n}^m) = E((F^{-1}(U_{i,n}))^m) = \int_0^1 \frac{n!}{(i-1)!(n-i)!}(F^{-1}(u))^m u^{i-1}(1-u)^{n-i}du,$$

et le moment d'ordre m de la statistique d'ordre $U_{i,n}$ est donné par

$$E(U_{i,n}^m) = \int_{\mathbb{R}} u^m f_{U_{i,n}}(u)du = \int_0^1 \frac{n!}{(i-1)!(n-i)!} u^{m+i-1}(1-u)^{n-i}du = \frac{B(i+m, n-i+1)}{B(i, n-i+1)}.$$

Après la simplification on trouve :

$$E(U_{i,n}^m) = \frac{n!(i+m-1)!}{(n+m)!(i-1)!}.$$

Existence de moment de statistique d'ordre

Théorème 1.5.1 *Soient $X_1, \ldots, X_n$ un échantillon de taille n de v.a X de loi F continue et $X_{1,n}, \ldots, X_{n,n}$ les statistiques d'ordre associées. Soit k un entier strictement positif. Si X admet un moment d'ordre k, alors pour tout $i = 1, \ldots, n$, la $i^{ème}$ statistique d'ordre $X_{i,n}$ admet aussi un moment d'ordre k. (La réciproque est fausse).*

1.6 Convergence de la statistique d'ordre

Soit $p \in]0,1[$. Supposons que F est continue et qu'il existe une seule solution x_p à l'équation $F(x) = p$. Soit $(k_n, n \geq 1)$ une suite d'entiers tel que $1 \leq k_n \leq n$ et $\lim_{n \to +\infty} k_n/n = p$. Alors

$$X_{k_n,n} \xrightarrow{p.s} x_p \text{ lorsque,} n \to +\infty.$$

– Si $p = 1 : X_{k_n,n} \xrightarrow{p.s} w_F = \inf\{x, F(x) = 1\}$.

– Si $p = 0 : X_{k_n,n} \xrightarrow{p.s} \alpha_F = \sup\{x, F(x) = 0\}$.

Preuve. Soit $x \in \mathbb{R}$ fixé. $S_n = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$. ■

On a $\{X_{k,n} \leq x\} = \{k \leq S_n\}$ alors $\{X_{k_n,n} \leq x\} = \{k_n \leq S_n\} = \{1 \leq \frac{S_n}{k_n}\}$.

D'après la loi forte des grands nombres on a $S_n \xrightarrow{p.s} np_1$ tel que $p_1 = F(x)$, alors $\frac{S_n}{n} \xrightarrow{p.s} p_1$, ce qui emplique $\frac{S_n}{k_n} \xrightarrow{p.s} \frac{p_1}{p} = \frac{F(x)}{p}$.

Donc

$$P(\frac{S_n}{k_n} \geq 1) = P(X_{k_n,n} \leq x) = \begin{cases} 0 & \text{si } F(x) \leq F(x_p) \\ 1 & \text{sinon} \end{cases} = \begin{cases} 0 & \text{si } x \leq x_p \\ 1 & \text{sinon} \end{cases}$$

Alors $X_{k_n,n} \xrightarrow{p.s} x_p$.

1.7 Représentation de Rényi

Soit $(U_1, \ldots, U_n)$ un échantillon suit la loi uniforme sur $[0,1]$ et $(U_{1,n}, \ldots, U_{n,n})$ la statistique d'ordre associée, et soit $E_1, \ldots, E_{n+1}$ un échantillon de taille $n+1$ suit la loi exponentielle de paramètre 1, alors :

$$\{U_{i,n}\}_{i=1,\ldots,n} \overset{\mathcal{D}}{=} \left\{\frac{S_i}{S_{n+1}}\right\}_{i=1,\ldots,n}, \quad \text{avec } S_i = E_1 + \ldots + E_i,\ i = 1, \ldots, n+1$$

Ce qui implique

$$\{X_{i,n}\}_{i=1,\ldots,n} \overset{\mathcal{D}}{=} \left\{F^{-1}(U_{i,n})\right\}_{i=1,\ldots,n} \overset{\mathcal{D}}{=} \left\{F^{-1}(\frac{S_i}{S_{n+1}})\right\}_{i=1,\ldots,n}.$$

1.8 Fonctions linéaires des statistiques d'ordre

Les fonctions linéaires des statistiques d'ordre (appelées aussi L-statistiques) sont écrites sous la forme de la statistique suivante : $L_n = \sum_{i=1}^{n} a_{i,n} X_{i,n}$, avec $a_{i,n}$ une suite de constantes. Ces fonctions donnent une bonne estimation des paramètres d'échelle et de position et lorsqu'elles sont utilisées comme un estimateur, on les appelle L-estimateur.

On suppose $a_{i,n} = \frac{1}{n} J(\frac{i}{n+1})$ où $J(u)$ est la fonction de poids associée tel que $u \in [0,1]$.

Donc

$$L_n = \frac{1}{n} \sum_{i=1}^{n} J(\frac{i}{n+1}) X_{i,n}.$$

Propriétés asymptotiques de L_n

Théorème 1.8.1 *Supposons que $E|X|^3 < +\infty$, tel que X représente la v.a de population avec une distribution F.*

Soit la fonction de poids $J(u)$ bornnée et continue en tout point de discontinue de $F^{-1}(u)$. De plus supposons $|J(u) - J(v)| \leq C|u - v|^{\delta + \frac{1}{2}}$ tel que C constant, $\delta > 0$, $0 < u < v < 1$, sauf pour un nombre fini de valeurs de u et v. Alors, on a les résultats suivants :

$$\begin{aligned}
&(i) \quad \lim_{n\to\infty} \sqrt{n}(E(L_n) - \mu(J,F)) = 0,\\
&(ii) \quad \lim_{n\to\infty} nVar(L_n) = \sigma^2(J,F),\\
&(iii) \quad \sqrt{n}(L_n - \mu(J,F)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(J,F)),
\end{aligned}$$

avec

$$\begin{aligned}
\mu(J,F) &= \int_{\mathbb{R}} xJ(F(x))dF = \int_0^1 J(u)F^{-1}(u)du.\\
\sigma^2(J,F) &= 2 \iint\limits_{-\infty<x<y<+\infty} J(F(x))J(F(y))F(x)(1-F(y))dxdy.
\end{aligned}$$

Distribution asymptotique de la moyenne tronquée

La moyenne tronquée est une importante L-statistiques qui est considérée comme une méthode de calcul de moyenne qui supprime un petit pourcentage des valeurs extrêmes supérieurs ou inférieurs ou les deux. Sa fonction de poids peut être exprimée par

$$J(u) = \begin{cases} (p_2 - p_1)^{-1} & 0 \leq p_1 < u < p_2 \leq 1. \\ 0 & \text{sinon.} \end{cases}$$

Si $F^{-1}(u)$ continue en p_1 et p_2, on peut appliquer le théorème 1.8.1.

Maintenant, nous commençons par quelques notations, puis énonçons un résultat plus général pour la moyenne tronquée qui est obtenue par Stigler (1973*b*).

Pour $0 \leq p_1 < p_2 \leq 1$, la L-statistique donnée par :

$$S_n = \frac{1}{[np_2] - [np_1]} \sum_{i=[np_1]+1}^{[np_2]} X_{i,n},$$

est appelée la moyenne tronquée, où les proportions p_1 et $(1 - p_2)$ représentent la proportion de l'échantillon tronqué aux deux extrémités. Soit

$$\alpha = F^{-1}(p_1) - F^{-1}(p_1-), \beta = F^{-1}(p_2) - F^{-1}(p_2-),$$

avec α et β représentent les amplitudes du saut de F^{-1} aux proportions de tronquée. Introduire une distribution H obtenue en tranquant F comme suit :

$$H(x) = \begin{cases} 0 & x \leq F^{-1}(p_1) \\ \frac{F(x)-p_1}{p_2-p_1} & F^{-1}(p_1) \leq x < F^{-1}(p_2-) \\ 1 & x \geq F^{-1}(p_2-) \end{cases}$$

Théorème 1.8.2 *Soit* $0 < p_1 < p_2 < 1$ *et* $n \to +\infty$. *Alors*

$$\sqrt{n}(S_n - \mu_H) \xrightarrow{\mathcal{D}} W, \tag{1.1}$$

où la v.a limite peut être exprimée comme

$$W = \frac{1}{p_2 - p_1} \left\{ Y + \left[F^{-1}(p_1) - \mu_H\right] Y_1 + \left[F^{-1}(p_2) - \mu_H\right] Y_2 - \alpha \max(0, Y_1) + \beta \max(0, Y_2) \right\},$$

avec

$Y \to \mathcal{N}(0, (p_2 - p_1)\sigma_H^2)$, μ_H la moyenne de H, σ_H^2 la variance de H et le vecteur

aléatoire (Y_1, Y_2) est normal bivarié tel que

$$E(Y_i) = 0; Var(Y_i) = p_i(1-p_i); i = 1, 2.$$
$$Cov(Y_1, Y_2) = -p_1(1-p_2).$$

De plus, Y et (Y_1, Y_2) sont multuellement indépendants.

1.9 L-moments

Les L-moments sont des L-statistiques, ont été introduit et extrait de Silitto (1951). Elles sont utilisées comme les méthodes d'estimation des moments habituelles tels que les estimations basées sur celle-ci sont obtenues de la même manière que dans la méthode des moments. Les L-moment permettent de distinguer le comportement des données hydrologiques asymétriques (hydrologie des eaux de surface, hydrologie des précipitations extrêmes).

Définition 1.9.1 *Soit X une v.a.de taille n d'une distribution F et soit $X_{1,n}, \ldots, X_{n,n}$ la statistique d'ordre associée de $(X_1, \ldots, X_n)$. Le $r^{ème}$ L-moment est défini par :*

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k C_{r-1}^k E(X_{r-k,r}),$$

et les quatres premiers L-moments théoriques sont

$$\lambda_1 = E(X_{1,1}).$$
$$\lambda_2 = \frac{1}{2}E(X_{2,2} - X_{1,2}).$$
$$\lambda_3 = \frac{1}{3}E(X_{3,3} - 2X_{2,3} + X_{1,3}).$$
$$\lambda_4 = \frac{1}{4}E(X_{4,4} - 3X_{3,4} + 3X_{2,4} - X_{1,4}).$$

Avec,

– λ_1 est une mesure de position.

– λ_2 est une mesure de dispersion ou d'échelle.

– λ_3 est une mesure de skewness.

– λ_4 est une mesure de kurtosis.

Les rapports théoriques des L-moments

Les rapports théoriques des L-moments sont des quantités sans dimension et écrites sous la forme suivante :

$$\tau_2 = \frac{\lambda_2}{\lambda_1} \text{ (coifficient de L-variation).}$$
$$\tau_3 = \frac{\lambda_3}{\lambda_2} \text{ (L-skewness).}$$
$$\tau_4 = \frac{\lambda_4}{\lambda_2} \text{ (L-kurtosis).}$$

1.10 Estimateur de la fonction quantile

L'estimateur de la fonction quantile qui est donnée par

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

est dépendent de l'ordre des observations dans un échantillon, en même temps il est utilisé pour estimer les quantiles et les valeurs extrêmes d'une distribution inconnue telles que la valeur à risque (VaR). L'estimateur empirique correspondant à Q est

$$Q_n(u) = \sum_{i=1}^{n} X_{i,n} \mathbb{1}_{\{\frac{i-1}{n} < u \leq \frac{i}{n}\}} = X_{1,n} + \sum_{i=2}^{n} (X_{i,n} - X_{i-1,n}) \mathbb{1}_{\{\frac{i-1}{n} < u\}}, \qquad (1.2)$$

et l'estimateur lissé de Q est donné par

$$\tilde{Q}(u) = \sum_{i=1}^{n} X_{i,n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h} k(\frac{u-y}{h}) dy,$$

avec k un noyau et h le paramètre de lissage. Falk (1984) a montré que le comportement asymptotique de $\tilde{Q}$ est plus performant que celui du quantile empirique Q_n. Sheater et Marron (1990) [15] ont donné une expression de l'erreur au moyenne quadratique asymptotique de $\tilde{Q}$ comme suit :

Théorème 1.10.1 *Supposons que $Q''(u)$ est continue au voisinage de u et k est une densité à support compact, symétrique par rapport à 0. Alors, si F n'est pas symétrique ou F symétrique mais $u \neq \frac{1}{2}$, on a*

$$MSE(\tilde{Q}(u)) = \frac{u(1-u)}{n}(Q'(u))^2 + \frac{h^4}{4}(Q''(u))^2\mu_2^2(k) - 2\frac{h}{n}(Q'(u))^2\Phi(k) + o(\frac{h}{n}) + o(h^4).$$

Si F est symétrique et $u = \frac{1}{2}$, alors

$$MSE(\tilde{Q}(u)) = \frac{1}{n}(Q'(\frac{1}{2}))^2(\frac{1}{4} - h\Phi(k) + \frac{1}{nh}R(k)) + o(\frac{h}{n}) + o(\frac{1}{(nh)^2}),$$

où

$$R(k) = \int_{\mathbb{R}} k^2(x)dx, \ \mu_2(k) = \int_{\mathbb{R}} x^2 k(x)dx.$$
$$\Phi(k) = \int_{\mathbb{R}} xk(x)k^{(-1)}(x)dx, \ \textit{avec } k^{(-1)} \textit{ la primitive de } k.$$

1.11 Estimateur de la prime de réassurance $\Pi_{\rho,R_{opt}}$

La prime est une somme d'argent paye par l'assuré à la compagnie d'assurance en contrepartie de prendre en charge des risques. Necir et al (2007) [14] proposent un estimateur semi paramétrique de $\Pi_{\rho,R}$ pour un indice d'aversion au risque fixé $\rho \geq 1$

et un niveau de rétention optimal $R_{opt} = F^{-1}(1 - \delta_{opt})$, avec δ_{opt} est un nombre réel suffisamment petit. Cet estimateur a été écrit sous la forme suivante

$$\widehat{\Pi}_{\rho,\hat{R}_{opt}} := (k/n)^{1/\rho} \frac{\rho}{1/\hat{\gamma}^{(H)} - \rho} X_{n-k,n}, \text{ où } \widehat{\gamma}^{(H)} < 1/\rho,$$

sous ces conditions : $\delta_{opt} = k/n$ et R_{opt} sera estimée par $\widehat{R}_{opt} = X_{n-k,n}$, avec $1 \leq k = k_n \leq n$, tel que $k_n \to \infty$ et $\frac{k_n}{n} \underset{n\to\infty}{\to} 0$.

1.12 Estimateur de Hill

L'estimation de l'indice des valeurs extrêmes γ est très important pour estimer la queue de distribution. Il y a beaucoup d'estimateurs pour γ écrits en termes de la statistique d'ordre, parmi ces estimateurs est l'estimateur de Hill qui est définie dans la définition suivante :

Définition 1.12.1 *Soit $X_1, X_2, \ldots, X_n$ une suite de v.a' i.i.d de fonction de répartition F appartenant au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $\gamma > 0$, soit une suite d'entier $k = k_n \to +\infty$ et $k/n \to 0$ quand $n \to +\infty$. L'estimateur de Hill est défini par la statistique :*

$$\widehat{\gamma}^{H} := \frac{1}{k} \sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \text{ avec } 1 \leq k \leq n. \tag{1.3}$$

Chapitre 2

Introduction sur la théorie des valeurs extrêmes

Les compagnies d'assurance et autres entreprises font face à de nombreux risques considérés comme des valeurs extrêmes qui affectent négativement ces entreprises et conduisent à des crises. Pour cela, elles étudient et déterminent ces valeurs et estiment la valeur après laquelle le risque se produit. Dans ce chapitre, nous parlerons sur les notions fondamentales de la théorie des valeurs extrêmes telles que les domaines d'attraction et les estimateurs de l'indice de queue, aussi les estimateurs des quantiles extrêmes.

2.1 Théorie des valeurs extrêmes

Soit $(X_i)_{i\geq 1}$ une suite de v.a' i.i.d de fonction de répartition F. S'il existe deux constantes de normalisation $a_n > 0$ et $b_n \in \mathbb{R}$, telle que :

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = H(x), \forall x \in \mathbb{R}, \tag{2.1}$$

la limite converge en distribution vers une loi non dégénérée $H(x)$ (appelée loi des valeurs extrêmes) et prend l'un des trois types de lois suivantes :

$$\text{Fréchet :} \quad \Phi_\gamma(x) = \begin{cases} \exp(-x^{-\frac{1}{\gamma}}) & x > 0, \gamma > 0 \\ 0 & \text{sinon.} \end{cases}$$

$$\text{Gumbel :} \quad \Lambda(x) = \exp(-\exp(-x)), \ x \in \mathbb{R}.$$

$$\text{Weibull :} \quad \Psi_\gamma(x) = \begin{cases} \exp(-(-x)^{-\frac{1}{\gamma}}) & x < 0, \gamma < 0 \\ 1 & \text{sinon.} \end{cases}$$

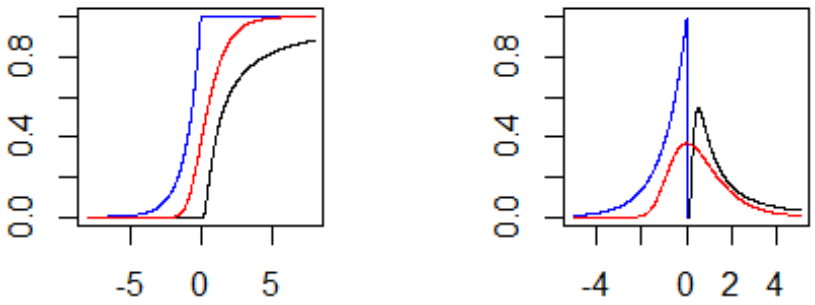

FIG. 2.1 – Lois des valeurs extrêmes(noir : Fréchet, rouge : Gumbel, bleu : Weibull)

Cela signifie que le théorème des valeurs extrêmes est un analoge du théorème central limite mais pour les phénomènes extrêmes tel que les constantes de normalisation (le paramètre d'échelle a_n et le paramètre de position b_n) jouent le rôle de l'écart-type et l'espérance respectivement et le choix de ces constantes apparaît dans le théorème suivant :

Théorème 2.1.1 *On peut choisir des constantes $a_n > 0$ et $b_n \in \mathbb{R}$ telle que* 2.1 *soit vérifiée de la manière suivante*

$$\begin{aligned}
&\textit{si } H(x) = \Phi_\gamma(x) \textit{ alors} \quad b_n = 0,\ a_n = F^{-1}(1 - \tfrac{1}{n}).\\
&\textit{si } H(x) = \Lambda(x) \textit{ alors} \quad\ b_n = F^{-1}(1),\ a_n = F^{-1}(1) - F^{-1}(1 - \tfrac{1}{n}).\\
&\textit{si } H(x) = \Psi_\gamma(x) \textit{ alors} \quad b_n = F^{-1}(1 - \tfrac{1}{n}),\ a_n = F^{-1}(1 - \tfrac{1}{ne}) - F^{-1}(1 - \tfrac{1}{n}).
\end{aligned}$$

Remarque 2.1.1 *La formule générale de $H(x)$ s'appelle distribution généralisée des valeurs extrêmes (GEVD) est définie par :*

$$H_{\mu,\sigma,\gamma}(x) = \begin{cases} \exp(-(1+\gamma(\frac{x-\mu}{\sigma}))^{-\frac{1}{\gamma}}) & \gamma \neq 0, 1+\gamma(\frac{x-\mu}{\sigma}) > 0 \\ \exp(-\exp(-(\frac{x-\mu}{\sigma}))) & \gamma = 0, x \in \mathbb{R} \end{cases}$$

avec μ : le paramètre de localisation, σ : le paramètre d'échelle, γ :l'indice de queue. Dans ce cas les trois lois sont écrites sous la forme suivante :

$$\begin{aligned}
&\Phi_\gamma(x) = H_{0,1,\frac{1}{\gamma}}(\gamma(x-1)),\\
&\Lambda(x) = H_{0,1,0}(x),\\
&\Psi_\gamma(x) = H_{0,1,-\frac{1}{\gamma}}(\gamma(x+1)), x \in \mathbb{R}.
\end{aligned}$$

Il est clair que cette distribution est écrite en termes des paramètres $(\mu, \sigma$ et $\gamma)$, ces paramètres contrôlent l'aplatissement et l'asymétrie de $H_{\mu,\sigma,\gamma}$. Le coefficient d'aplatissement K (appelé aussi le kurtosis) et le coefficient d'asymétrie S (coefficient de Skewness) s'écrivent respectivement sous la forme :

$$K = \frac{\mu_4}{\mu_2^2}, \quad S = \frac{\mu_3}{\mu_2^{3/2}},$$

avec μ_2, μ_3 et μ_4 sont les moments d'ordre respectifs $2, 3$ et 4. Si le kurtosis est supérieur à 3, on dit que la distribution à queue lourde. Mais ceci n'est valable que dans le cas où les moments d'ordre 4 et 2 existent, et si le coefficient d'asymétrie est

supérieur à 0, on dit que la distribution asymétrique à droite.

Chaque type de distribution des valeurs extrêmes que nous avons vues au début du chapitre est un domaine d'attraction pour de nombreuses distributions, si ces distributions ont réalisé ce que nous verrons dans la section suivante.

2.2 Domaine d'attraction

Définition 2.2.1 (domaine d'attraction maximum) *S'il existe des séquences réelles $a_n > 0$ et $b_n \in \mathbb{R}$ tel que*

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \forall x \in \mathbb{R},$$

pour tout point de continuité x de G_γ, on dit que la distribution F appartient au domaine d'attraction maximum de G_γ, et on écrit $F \in D(G_\gamma)$.

Dans le tableau ci-dessous, nous verrons le domaine d'attraction de quelques lois

Domaine d'attraction	Lois
Fréchet ($\gamma > 0$)	Pareto, Cauchy, Student, Log-gamma.
Gumbel ($\gamma = 0$)	Exponentielle, Normale, Log-normale, Gamma.
Weibull ($\gamma < 0$)	Beta, Uniforme.

TAB. 2.1 – Domaine d'attraction de quelques lois.

Théorème 2.2.1 *Soit F une fonction de répartition et w_F le point terminal. Suppose que $F''(x)$ existe et $F'(x)$ positive pour tout x dans un voisinage gauche de w_F. Si*

$$\lim_{t\to w_F} \left(\frac{1 - F(t)}{F'(t)}\right)' = \gamma,$$

ou

$$\lim_{t \to w_F} \frac{(1 - F(t))F''(t)}{(F'(t))^2} = -\gamma - 1,$$

alors $F \in D(H_{\mu,\sigma,\gamma})$.

Proposition 2.2.1 *$F \in D(H_{\mu,\sigma,\gamma})$ ce qui est équivalente de*

$$n\bar{F}(a_n x + b_n) \underset{n \to \infty}{\to} -\log(H_{\mu,\sigma,\gamma}), \text{ avec } a_n > 0 \text{ et } b_n \in \mathbb{R}.$$

On a alors pour $n \geq 1$ la convergence en loi de $(X_{n,n} - b_n)/a_n$ vers une v.a de fonction de répartition $H_{\mu,\sigma,\gamma}$.

Théorème 2.2.2 *La fonction de répartition F est dans le domaine d'attraction de la distrbution des valeurs extrêmes H ssi*

1. *Pour $\gamma > 0 : w_F = +\infty$ et*

$$\lim_{t \to +\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\frac{1}{\gamma}}, x > 0.$$

2. *Pour $\gamma < 0 : w_F < +\infty$ et*

$$\lim_{t \to 0} \frac{1 - F(w_F - tx)}{1 - F(w_F - t)} = x^{-\frac{1}{\gamma}}, x > 0.$$

3. *Pour $\gamma = 0 : w_F$ peut être fini ou infini et*

$$\lim_{t \to w_F} \frac{1 - F(t + xf(t))}{1 - F(t)} = e^{-x}, x \in \mathbb{R}, \text{ où } f \text{ est une fonction positive appropriée.}$$

Remarque 2.2.1 *Dans le théorème 2.2.2 la première limite signifie que $F \in D(\Phi_\gamma)$, la deuxième est équivalente à $F \in D(\Psi_\gamma)$ et la dernière est équivalente à $F \in D(\Lambda)$.*

Définition 2.2.2 *On dit qu'une fonction L est à variation lente si $L(t) > 0$ pour t assez grand et si pour tout $x > 0$, on a*

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1.$$

Théorème 2.2.3 *La fonction de répartition F appartient au domaine d'attraction de la loi de Fréchet de paramètre γ si et seulement si*

$$\overline{F}(x) = x^{-1/\gamma} L(x), \tag{2.2}$$

où $L(x)$ est une fonction à variation lente.

Remarque 2.2.2 *On dit que F est à queue lourde ssi 2.2 vérifie quand $x \to \infty$, $\gamma > 0$.*

Ce type de distribution satisfait pour tout $x > 0$, les deux conditions suivantes :

1. Condition du 1^{er} ordre

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma}. \ (\overline{F} \text{ à variation régulière à l'infini}).$$

2. Condition du $2^{ème}$ ordre : $\exists \rho > 0$ et A fonction tend vers à 0 et ne change pas le signe au voisinage de ∞ tel que

$$\lim_{t\to\infty} \frac{\bar{F}(tx)/\bar{F}(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma} \frac{x^{\rho} - 1}{\gamma\rho}. \tag{2.3}$$

Proposition 2.2.2 (Critère de Von Mises) *Soit F la fonction de répartition d'une loi de densité f*

1. Si on a

$$\lim_{x\to\infty} \frac{xf(x)}{\overline{F}(x)} = \gamma > 0,$$

alors F appartient au domaine d'attraction de Φ_γ de paramètre γ.

2. On suppose la loi de densité f strictement positive sur un intervalle (z, w_F), avec $w_F < \infty$. Si on a

$$\lim_{x \to w_F^-} \frac{(x - w_F) f(x)}{\overline{F}(x)} = \gamma > 0,$$

alors F appartient au domaine d'attraction de Ψ_γ de paramètre γ.

2.3 Distribution de Pareto généralisée

En effet, l'approche basée sur la GEVD ne prend en considération qu'une seule valeur $X_{n,n}$, ce qui conduit à une perte d'informations contenues dans les autres grandes valeurs de l'échantillon. Pour résoudre ce problème, une autre approche appelée POT (Peak Over Threshold) a été trouvée pour toutes les valeurs extrêmes qui dépassent le seuil u. Cette méthode repose sur le choix d'un seuil approprié pour étudier les excès au-delà de ce seuil (les excès sont les différences positives entre les observations et le seuil). Un résumé de ceci est dans la figure 2.2 suivante

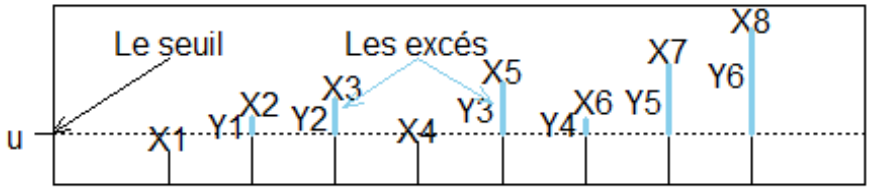

FIG. 2.2 – Représentation graphique des excès.

Théorème 2.3.1 (Balkema-de Haan-Pickands) *Si $F \in D(H_{\mu,\sigma,\gamma})$, alors il existe une fonction de répartition des excès au-delà de u noté F_u qui peut être uniformément approchée par une loi de Pareto généralisée (GPD) tel que :*

$$\lim_{x \to w_F} \sup_{0<x<w_F-u} |F_u(x) - G_{\gamma,\sigma}(x)| = 0,$$

avec $G_{\gamma,\sigma}(x)$ la GPD et

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)},\ x \in]0, w_F - u[\,.$$

Définition 2.3.1 (Distribution de Pareto généralisée) *Pour $\sigma > 0$, et $\gamma \in \mathbb{R}$, la GPD est définie par :*

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - (1 + \frac{\gamma}{\sigma}x)^{-1/\gamma} & si\ \gamma \neq 0 \\ 1 - \exp(-\frac{x}{\sigma}) & si\ \gamma = 0 \end{cases}$$

où, $x \geq 0$ si $\gamma \geq 0$ et $0 \leq x \leq \frac{\gamma}{\sigma}$ si $\gamma < 0$.

Lorsque $\gamma > 0$, c'est la loi Pareto, lorsque $\gamma < 0$, nous avons la loi Bêta et $\gamma = 0$ donne la loi exponentielle.

Les paramètres de la GEVD et la GPD sont des paramètres inconnus et pour lui donner une valeur estimée, il existe de nombreuses méthodes, et ces estimateurs peuvent être paramétriques comme l'estimateur du maximum de vraisemblance et de L-moment..., semi paramétriques (l'estimateur de Hill...) ou non paramétriques comme l'estimateur de GLW. Nous allons maintenant passer pour prendre une idée sur certaines de ces estimateurs.

2.4 Estimation de l'indice de queue

2.4.1 Méthode du Maximum de Vraisemblance

L'estimateur du maximum de vraisemblance de θ est défini comme suit :

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} l(X_1,\ldots,X_n;\theta) = \arg\max_{\theta\in\Theta} \log L(X_1,\ldots,X_n;\theta),$$

avec

$$L(X_1,\ldots,X_n;\theta) = \prod_{i=1}^{n} h_\theta(X_i) \text{ est la fonction de vraisemblance.}$$

On obtient l'estimateur de $\hat{\theta}$ en résolvant le système suivant :

$$\begin{cases} \frac{\partial l(X_1,\ldots,X_n;\theta)}{\partial\theta} = 0. \\ \frac{\partial^2 l(X_1,\ldots,X_n;\theta)}{\partial\theta^2} < 0. \end{cases}$$

Donc pour $\gamma \neq 0$ la fonction de log vraisemblance de la GEVD est égale à :

$$l(X_1,\ldots,X_n;\theta) = -n\log\sigma - \sum_{i=1}^{n}(1+\gamma(\frac{X_i-\mu}{\sigma}))^{-1/\gamma} - (1+\frac{1}{\gamma})\sum_{i=1}^{n}\log(1+\gamma(\frac{X_i-\mu}{\sigma})).$$

On dérive $l(X_1,\ldots,X_n;\theta)$ par rapport à les paramètres μ, σ et γ, et on obtient le système suivant :

$$\begin{cases} -\frac{1}{\sigma}\sum_{i=1}^{n}(1+\gamma(\frac{X_i-\mu}{\sigma}))^{-1-1/\gamma} + (1+\gamma)\sum_{i=1}^{n}\frac{1}{\sigma+\gamma(X_i-\mu)} = 0. \\ -n - \frac{1}{\sigma}\sum_{i=1}^{n}(X_i-\mu)(1+\gamma(\frac{X_i-\mu}{\sigma}))^{-1-1/\gamma} + (1+\gamma)\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma+\gamma(X_i-\mu)} = 0. \\ \frac{1}{\gamma}\sum_{i=1}^{n}((1+\gamma(\frac{X_i-\mu}{\sigma}))^{-1/\gamma}+1)\log(1+\gamma(\frac{X_i-\mu}{\sigma})) - \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma+\gamma(X_i-\mu)}(1+\gamma+(1+\gamma(\frac{X_i-\mu}{\sigma}))^{-1/\gamma}) = 0. \end{cases}$$

Pour $\gamma = 0$ la fonction de log vraisemblance est égale à :

$$l(X_1, \ldots, X_n; \theta) = -n \log \sigma - \sum_{i=1}^{n} \exp(-\frac{X_i - \mu}{\sigma}) - \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma},$$

avec le système correspondant suivant :

$$\begin{cases} n - \sum_{i=1}^{n} \exp(-\frac{X_i - \mu}{\sigma}) = 0. \\ n + \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}(\exp(-\frac{X_i - \mu}{\sigma}) - 1) = 0. \end{cases}$$

Dans les deux cas le système est non linéaire pour lesquels aucune solution explicite existe.

2.4.2 Méthode du L-moment

La méthode de L-moment a été utilisée comme la méthode du maximum de vraisemblance pour l'estimation des paramètres de distribution. Soit $X_1, \ldots, X_n$ un échantillon de taille $n \geq 1$ d'une v.a X et $X_{1,n}, \ldots, X_{n,n}$ les statistiques d'ordre associées. L'estimateur de L-moment est donné par :

$$l_r = (C_n^r)^{-1} \sum_{1 \leq i_1 < \ldots < i_r \leq n} \cdots \sum \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k C_{r-1}^k X_{i_r - k, n}, \text{ avec } r = 1, 2, \ldots.$$

Les trois premières estimations sont écrites sous la forme suivante :

$$\begin{aligned} l_1 &= \frac{1}{n} \sum_{1 \leq i \leq n} X_{i,n}. \\ l_2 &= \frac{1}{n(n-1)} \sum \sum_{1 \leq i < j \leq n} (X_{j,n} - X_{i,n}). \\ l_3 &= \frac{2}{n(n-1)(n-2)} \sum \sum \sum_{1 \leq i < j < k \leq n} (X_{k,n} - 2X_{j,n} + X_{i,n}). \end{aligned}$$

On peut écrire les L-moments sous forme de combinaisons linéaires en fonction de moment de probabilité pondéré α_k et β_j, tout d'abord on définie par :

$$M_{i,j,k} = E(X^i F^j (1-F)^k) = \int_0^1 Q(u)(1-u)^k u^j du,$$

les moments pondéré de Weighted et α_k, β_j et $P^*_{r,k}$ sous les formes suivantes :

$$\begin{aligned}
\alpha_k &= M_{1,0,k} = \textstyle\int_0^1 Q(u)(1-u)^k du. \\
\beta_j &= M_{1,j,0} = \textstyle\int_0^1 Q(u) u^j du. \\
P^*_{r,k} &= (-1)^{r-k} C^k_r C^k_{r+k}.
\end{aligned}$$

Par conséquent les L-moments sont données par :

$$\lambda_{r+1} = (-1)^r \sum_{k=0}^{r} P^*_{r,k} \alpha_k = \sum_{j=0}^{r} P^*_{r,j} \beta_j,$$

on trouve

$$\lambda_1 = \beta_0;\ \lambda_2 = 2\beta_1 - \beta_0;\ \lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0,$$

donc l'estimateur de λ_r peut s'écrire comme suit :

$$l_r = (-1)^{r-1} \sum_{k=0}^{r-1} P^*_{r-1,k} \widehat{\alpha}_k = \sum_{j=0}^{r-1} P^*_{r-1,j} \widehat{\beta}_j,$$

avec

$$l_1 = \widehat{\beta}_0;\ l_2 = 2\widehat{\beta}_1 - \widehat{\beta}_0;\ l_3 = 6\widehat{\beta}_2 - 6\widehat{\beta}_1 + \widehat{\beta}_0.$$

Cette méthode a été utilisée pour extraire une estimation de l'indice de queue, ainsi que les paramètres de position et d'échelle. Par exemple pour trouver $\widehat{\mu}$ et $\widehat{\sigma}$ de la

GEVD lorsque $\gamma = 0$ on a

$$\begin{aligned}\beta_j &= \int_0^1 (\mu - \sigma \log(-\log u))\, u^j du \\ &= \frac{\mu}{j+1} - \sigma \int_0^1 \log(-\log u) u^j du,\end{aligned}$$

en utilisant le changement de variable on trouve

$$\beta_j = \frac{1}{j+1} \left(\mu + \sigma \left(\log\left(j+1\right) - \Gamma\right)\right),$$

avec

$$\Gamma = \int_0^{+\infty} \exp\left(-w\right) \log\left(w\right) dw = -0.57722,$$

ce qui implique

$$\lambda_1 = \beta_0 = \mu - \sigma\Gamma, \lambda_2 = 2\beta_1 - \beta_0 = \sigma \log\left(2\right),$$

donc on trouve :

$$\widehat{\sigma} = l_2 / \log\left(2\right), \ \widehat{\mu} = l_1 + \widehat{\sigma}\Gamma.$$

2.4.3 Estimateur de Pickands

Définition 2.4.1 *Soit $X_1, X_2, \ldots, X_n$ une suite de v.a'i.i.d de fonction de répartition $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$. Soit une suite d'entier $k = k_n \to \infty$ quand $n \to \infty$. L'estimateur de Pickands est donné par la statistique suivante :*

$$\widehat{\gamma}^P = \frac{1}{\log 2} \log \frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}.$$

Propriétés asymptotiques de l'estimateur de Pickands

Théorème 2.4.1 (Consistance faible) *Soit $X_1, X_2, \ldots, X_n$ une suite de v.a'i.i.d de fonction de répartition $F \in D(H_{\mu,\sigma,\gamma})$, $\gamma > 0$. Soit $(k_n)_{n\geq 1}$ une suite d'entiers tel que $1 < k_n \leq n$. Si $k_n \to \infty$ et $k_n/n \to 0$ alors on a*

$$\widehat{\gamma}^P \xrightarrow{p} \gamma, \ \textit{quand } n \to \infty.$$

Théorème 2.4.2 (Consistance forte) *Soit $X_1, X_2, \ldots, X_n$ une suite de v.a'i.i.d de fonction de répartition $F \in D(H_{\mu,\sigma,\gamma})$, $\gamma \in \mathbb{R}$. Soit $(k_n)_{n\geq 1}$ une suite d'entiers tel que $1 < k_n \leq n$. Si $k_n \to \infty$ et $\frac{k_n}{\log(\log(n))} \to \infty$ alors*

$$\widehat{\gamma}^P \xrightarrow{p.s} \gamma, \ \textit{quand } n \to \infty.$$

Normality asymptotique

Théorème 2.4.3 *Sous des conditions additionnelles sur la suite intermédiaire k_n et la fonction de répartition F, on a :*

$$\sqrt{k}(\widehat{\gamma}_n - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \eta^2), \ \textit{quand } n \to \infty,$$

où

$$\eta^2 = \frac{\gamma^2(2^{2\gamma+1}+1)}{(2(2^\gamma - 1)\log 2)^2}.$$

2.4.4 Estimateur de Hill

Pour $n \geq 1$, soit $X_1, \ldots, X_n$ suite de v.a' i.i.d définie sur un espace de probabilité $(\Omega, \mathcal{A}, P)$, de fonction de distribution F. L'estimateur de Hill est utilisé pour les distributions qui appartient au domaine d'attraction de Fréchet c'est à dire que $\overline{F} = 1 - F$ la queue de distribution (appelée aussi la fonction de survie) de ce

domaine est de forme 2.2.

L'indice des valeurs extrêmes γ régit de la queue de distribution : lorsque γ est grand, la queue est épaisse (lourde). Nous avons vu son estimation dans le premier chapitre 1.12 et dans cette partie nous allons l'examiner la consistance et l'asymptotiquement normal de cet estimateur.

Propriétés asymptotiques de l'estimateur de Hill

Théorème 2.4.4 (Cansistance faible) *Soit $X_1, X_2, \ldots, X_n$ un échantillon de v.a' i.i.d de fonction de répartition F avec $F \in D(\Phi_\gamma)$ et $\gamma > 0$. Soit $(k_n)_{n>1}$ une suite d'entiers telles que $1 \leq k_n \leq n$, donc*

$$si \begin{cases} k_n \to \infty \\ k_n/n \to 0 \end{cases} \quad alors\ \widehat{\gamma}^H \xrightarrow{p} \gamma.$$

Théorème 2.4.5 (Consistance forte) *Soit $X_1, X_2, \ldots, X_n$ un échantillon de v.a' i.i.d de fonction de répartition F avec $F \in D(\Phi_\gamma)$ et $\gamma > 0$. Soit $(k_n)_{n>1}$ une suite d'entiers telles que $1 \leq k_n \leq n$.*

$$Si \begin{cases} k_n \to \infty \\ \frac{k_n}{\log(\log(n))} \to 0 \end{cases} \quad alors\ \widehat{\gamma}^H \xrightarrow{p.s} \gamma.$$

Normalité asymptotique

Théorème 2.4.6 *Si la condition de $2^{ème}$ ordre 2.3 est satisfaite avec $\lim_{n\to\infty} \sqrt{k}A(n/k) = \lambda$, alors*

$$\sqrt{k}(\widehat{\gamma}_n - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(\frac{\lambda}{1-\rho}, \gamma^2),\ quand\ n \to \infty.$$

2.4.5 Estimateur de GLW

L'estimateur de Hill a été généralisé par Csörgö, Deheuvels et Mason (CDM) à une nouvelle classe d'estimateurs du noyau qui n'est utilisé que dans le cas où $\gamma > 0$. Groeneboom, Lopuhaä et de Wolf (2003) (GLW) ont introduit une modification sur l'estimateur CDM qui permet des valeurs négatives de $\gamma \in \mathbb{R}$.

Soit $K(.)$ une fonction noyau vérifiant les conditions suivantes :

$$
\begin{aligned}
&K(x) = 0, \text{ si } x \notin [0,1[\,.\\
&K \text{ est deux fois continument différentiable sur }]0,1]\,.\\
&K(1) = K^{'}(1) = 0.\\
&\textstyle\int_0^1 K(x)\,dx = 1.
\end{aligned}
$$

Soit

$$
\begin{aligned}
\widehat{q}_{n,h}^{(1)} &:= \sum_{i=1}^{n-1} (i/n)^{\alpha} K_h(i/n) \log(X_{n-i+1,n}/X_{n-i,n}),\\
\widehat{q}_{n,h}^{(2)} &:= \sum_{i=1}^{n-1} \Psi_h(i/n) \log(X_{n-i+1,n}/X_{n-i,n}),
\end{aligned}
$$

avec

$$
\alpha > 0, h > 0, K_h(u) := h^{-1}K(u/h), \Psi_h(u) := \frac{d}{du}(u^{\alpha+1}K_h(u)), u \in (0,1]\,.
$$

L'estimateur $\gamma_{n,h}^{(GLW)}$ est défini comme suit

$$
\gamma_{n,h}^{(GLW)} := \gamma_{n,h}^{(CDM)} - 1 + \frac{\hat{q}_{n,h}^{(2)}}{\hat{q}_{n,h}^{(1)}},
$$

où

$$
\gamma_{n,h}^{(CDM)} := \sum_{i=1}^{n-1} (i/n) K_h(i/n) \log(X_{n-i+1,n}/X_{n-i,n}).
$$

• Les propriétés de convergence forte sont traitées par Necir (2006) .

2.5 Estimation des quantiles extrêmes

On définie la fonction quantile de queue par :

$$U(t) = Q(1 - 1/t) = (1/\overline{F})^{-1}(t), t > 1,$$

donc la fonction empirique de quantile de queue correspondante est :

$$U_n(t) = Q_n(1 - 1/t),$$

avec Q_n a été exprimé dans le premier chapitre (voir 1.2).

2.5.1 Estimateur des quantiles extrêmes d'une Pareto généralisée

On suppose qu'on ait trouvé un seuil u (assez élevé), la quantité $F_u(y) = \frac{F(y+u)-F(u)}{1-F(u)}$ est une approche de $G_{\gamma,\sigma}(y)$ d'après le théorème 2.3.1. On pose $x = y + u$ alors on a

$$\frac{F(x) - F(u)}{1 - F(u)} \simeq G_{\gamma,\sigma}(x - u),$$

d'après la simplification on trouve $\overline{F}(x) = \overline{F}(u)\overline{G}_{\gamma,\sigma}(x - u)$, donc pour estimer le qunatile extrême x_p, il faut estimé la queue de la distribution F. Comme $\overline{F}(u) = \frac{N_u}{n}$, $\overline{F}(x_p) = p$ et après un petit calcul, on trouve

$$\widehat{x}_p = \begin{cases} \frac{\widehat{\sigma}}{\widehat{\gamma}}((\frac{n}{N_u}p)^{-\widehat{\gamma}} - 1) + u & \text{si } \gamma \neq 0. \\ \widehat{\sigma}\log(\frac{N_u}{np}) + u & \text{si } \gamma = 0. \end{cases}$$

2.5.2 Estimateur des quantiles extrêmes d'une GEV

1. Le cas où $F = H_{\mu,\sigma,\gamma}$, l'estimateur de quantile extrême x_p est défini par :

$$\widehat{x}_p = H^{-1}_{\widehat{\mu},\widehat{\sigma},\widehat{\gamma}}(1-p) = \begin{cases} \widehat{\mu} - \frac{\widehat{\sigma}}{\widehat{\gamma}}[1 - (-\log(1-p))^{-\widehat{\gamma}}] & \text{si } \gamma \neq 0. \\ \widehat{\mu} - \widehat{\sigma}\log(-\log(1-p)) & \text{si } \gamma = 0. \end{cases}$$

2. Le cas où $F \in D(H_{\mu,\sigma,\gamma})$, il peut être estimée le quantile extrême $(x_p)_{p\geq\frac{1}{n}}$ en utilisant la proposition 2.2.1 comme suite :

$$\overline{F}(a_n x + b_n) = -\frac{1}{n}\log(H_{\mu,\sigma,\gamma}),$$

ce qui implique

$$F(a_n x + b_n) = 1 - \frac{1}{n}(1 + \gamma(\frac{x-\mu}{\sigma}))^{-\frac{1}{\gamma}}.$$

On pose $p = \frac{1}{n}(1 + \gamma(\frac{x-\mu}{\sigma}))^{-\frac{1}{\gamma}}$, alors

$$F(a_n x + b_n) = 1 - p,$$

donc

$$a_n x + b_n = Q(1-p) = x_p, \text{avec } x = \frac{\sigma}{\gamma}((np)^{-\gamma} - 1) + \mu,$$

et par conséquence

$$\widehat{x}_p := \widehat{a}_n(\frac{\hat{\sigma}}{\widehat{\gamma}}((np)^{-\widehat{\gamma}} - 1) + \hat{\mu}) + \widehat{b}_n.$$

Maintenant pour l'estimation de quantile extrême $(x_p)_{p<\frac{1}{n}}$ nous utilisons un sous séquance (n/k), où $k = k_n \to \infty, n/k \underset{n\to\infty}{\to} \infty$ on obtient

$$\widehat{x}_p := \widehat{a}_{n/k}(\frac{\hat{\sigma}}{\widehat{\gamma}}((np/k)^{-\widehat{\gamma}} - 1) + \hat{\mu}) + \widehat{b}_{n/k}.$$

2.5.3 Estimateur de Weissman

Pour la classe de Fréchet ($\gamma > 0$), l'estimateur de quantile d'ordre $(1-p)$ du type Weissman prend la forme suivante :

$$\widehat{x}_p^W := X_{n-k,n}(k/np)^{\widehat{\gamma}},$$

et l'estimateur de quantile extrême de Weissman pour l'estimateur de Hill est défini par :

$$\widehat{x}_p^H := X_{n-k,n}(k/np)^{\widehat{\gamma}^H}.$$

L'estimateur $\widehat{x}_p^P$ du quantile d'ordre $(1-p)$ lié au l'estimateur de Pickands est de la forme suivante :

$$\widehat{x}_p^P := X_{n-k+1,n} + \frac{(k/np)^{\widehat{\gamma}^P} - 1}{1 - 2^{\widehat{\gamma}^P}}(X_{n-k+1,n} - X_{n-2k+1,n}).$$

2.6 Choix du nombre optimal de statistiques d'ordre extrêmes

On note que ce soit dans une estimation de l'indice de queue ou dans une estimation de quantile extrême, on a besoin de connaître le nombre de statistiques d'ordre extrêmes (noté k) pour pouvoir trouver une estimation de les deux. La méthode de choisir cette dernière est toujours un problème difficile même si la forme de l'estimation est déterminée. Quand k est petit, la variance est grand et lorsque k est grand, le biais petit, dans cette partie nous parlerons de quelques méthodes pour obtenir k_{opt} (qui détermine le point de départ de la queue de la distribution) afin d'équilibrer la variance et le biais.

2.6.1 Méthode Graphique

Dans cette méthode consiste à tracer les points de coordonnées $\{(k, \widehat{\gamma}(k)) : k = 1, \ldots, n\}$, où k représente un certain nombre de statistiques d'ordre (k a été sélectionnée) et $\widehat{\gamma}$ désigne n'importe quel estimateur introduit dans la section 2.4. La valeur de k devrait être prise, où le graphe est stable, nous allons expliquer cela dans la figure 2.3 :

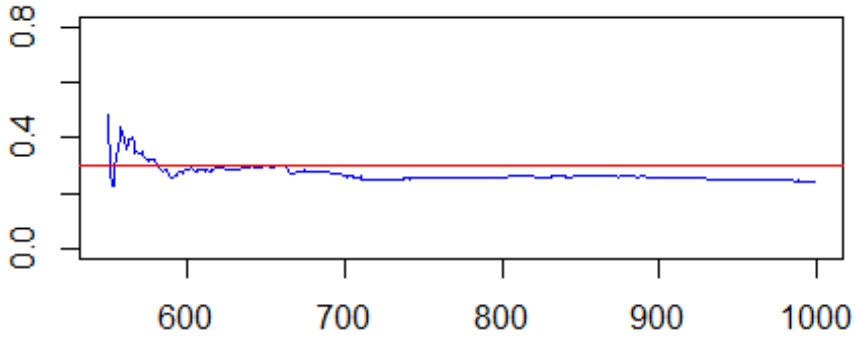

FIG. 2.3 – Choix de k optimal pour l'estimateur de Hill.

Apparemment dans cette Figure, l'estimateur semble stable un k autour de 650.

2.6.2 Erreur moyenne quadratique

Dans cette méthode pour trouver k_{opt}, l'erreur moyenne quadratique doit être minimisée. Donc le choix optimal de k se fait en résolvant le problème suivant :

$$k_{opt} := \arg\min_{k} MSE(\widehat{\gamma}),$$

avec

$$MSE(\widehat{\gamma}) := E_\infty(\widehat{\gamma} - \gamma)^2 = Biais(\widehat{\gamma})^2 + Var(\widehat{\gamma}),$$

où E_∞ dénote l'espérance en ce qui concerne la distribution de la limite.

2.6.3 Procédures adaptatives

Approche de Hall et Welsh

Le processus de sélection k n'est pas facile pour cela, les procédures adaptatives ont proposé de calculer $\widehat{k}_{opt}$ pour k_{opt} dans le sens

$$\frac{\widehat{k}_{opt}}{k_{opt}} \xrightarrow{p} 1, \text{ quand } n \to \infty.$$

Hall et Welsh (1985) [10] ont prouvé que l'erreur moyenne quadratique asymptotique de l'estimateur de Hill est minimale pour

$$k_{opt} \sim \left(\frac{c^{2\rho}\,(\rho+1)^2}{2d^2\rho^3} \right)^{1/(2\rho+1)} n^{2\rho/(2\rho+1)}, \text{ avec } \rho \text{ le paramètre de } 2^{ème} \text{ ordre.}$$

Si la fonction de répartition F satisfaite la classe de Hall (i.e : F à queue lourde qui satisfaite la condition du second ordre où

$$F(x) = 1 - cx^{-1/\gamma}\left(1 - dx^{-\rho/\gamma} + o\left(x^{-\rho/\gamma}\right)\right), \text{ quand } x \to \infty.$$

pour $\gamma > 0$, $\rho \leq 0$, $c > 0$, $d \in \mathbb{R} \diagdown \{0\}$).

Mais ce résultat ne peut pas être utilisé directement pour déterminer k_{opt} car ρ, c et d sont inconnus, pour cela Hall et Welsh ont construit une estimation de k_{opt} définie par :

$$\widehat{k}_{opt} := \widehat{\lambda}_0 n^{2\widehat{\rho}/(2\widehat{\rho}+1)},$$

où

$$\widehat{\lambda}_0 := \left| (2\widehat{\rho})^{-1/2} \left(\frac{n}{t_1}\right)^{\widehat{\rho}} \frac{\left(\widehat{\gamma}^H(t_1)\right)^{-1} - \left(\widehat{\gamma}^H(s)\right)^{-1}}{\left(\widehat{\gamma}^{(H)}(s)\right)^{-1}} \right|^{2/(2\widehat{\rho}+1)},$$

et

$$\widehat{\rho} := \left| \log \left| \frac{\left(\widehat{\gamma}^H(t_1)\right)^{-1} - \left(\widehat{\gamma}^H(s)\right)^{-1}}{\left(\widehat{\gamma}^H(s)\right)^{-1}} \right| / \log \frac{t_1}{t_2} \right|,$$

avec $\frac{\widehat{k}_{opt}}{k_{opt}} \xrightarrow{p} 1$ si $t_i = [n^{\tau_i}], i = 1,2$ et $s = [n^{\sigma}]$ pour $0 < 2\rho(1-\tau_1) < \sigma < 2\rho/(2\rho+1) < \tau_1 < \tau_2 < 1$.

Approche de Reiss et Thomas

Soit $\widehat{\gamma}(i)$ des estimations du paramètre de forme γ basées sur les i extrêmes supérieurs. Choisir k_{opt} la fraction d'échantillon optimal comme valeur qui minimise

$$\frac{1}{k}\sum_{i=1}^{k} i^{\beta} \left|\widehat{\gamma}(i) - médiane\left(\widehat{\gamma}(1), \ldots, \widehat{\gamma}(k)\right)\right|,$$

avec $0 \leq \beta < \frac{1}{2}$, un léger lissage de la série d'estimations améliore les performances de la procédure pour les tailles d'échantillon petites et moyennes.

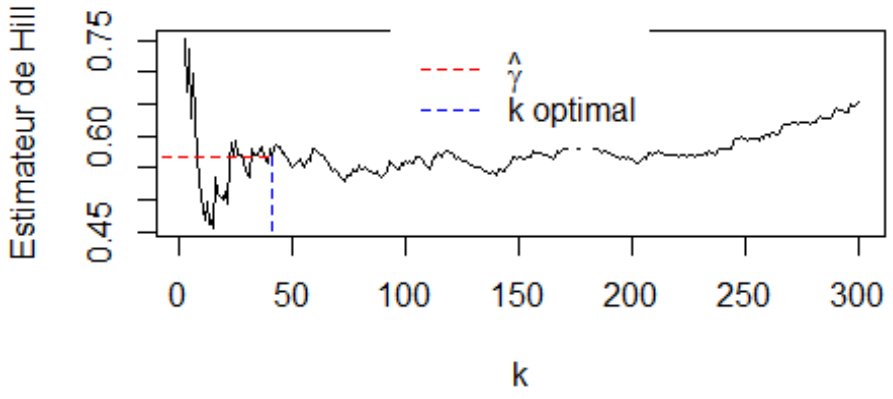

FIG. 2.4 – Choix de k avec l'approche de Reiss et Thomas (kopt=41).

Chapitre 3

Données incomplètes

On rencontre souvent des données incomplètes en assurance (comme l'assurance-vie) ou autre, et l'étude du comportement de la queue dans ce cas nécessite d'autres méthodes que celles utilisées dans le cas des données complètes, dont nous avons parlé dans le deuxième chapitre. La question est donc de savoir quels sont les types de ces données et comment trouver dans ce cas une estimation de l'indice de queue γ?

3.1 Données censurées

On dit qu'une donnée est "censurée" lorsqu'on ne peut pas l'observer complètement la variable d'intérêt, on a seulement l'information que la variable est supérieure ou inférieure à une certaine valeur, ou qu'elle est entre deux valeurs connues, mais on ne connaît pas sa valeur exacte. Le premier cas est connu sous le nom de censure à droite qui elle survient lorsque la valeur d'une observation est supérieure à un seuil de censure connu, rendant l'observation encore plus extrême que ce que les données peuvent révéler. Par exemple, dans une étude de survie des patients atteints de cancer, cette censure survient lorsqu'un patient survit au-delà de la période de

suivi de l'étude. Le second cas est la censure à gauche c'est l'inverse de censure à droite. Par exemple, dans une étude de temps de réaction de conducteurs, la censure à gauche peut survenir si le conducteur réagit trop rapidement pour être détecté par le système de mesure. Le dernier cas est connu sous le nom de censure par intervalle qui consiste à connaître les bornes d'intervalle dans lesquelles se trouve la valeur observée, sans connaître cette valeur précise. Par exemple, si l'on étudie la distribution des précipitations annuelles dans une ville, on peut avoir des observations censurées en raison de l'incapacité à mesurer les précipitations en dehors de certains intervalles de temps. Dans ce cas, la censure par intervalle se produirait lorsque l'on connaît les bornes supérieures et inférieures des intervalles au sein desquels les précipitations ont été mesurées. On peut résumé cela ci-dessous :

Censure à droite (C censure X à droite)

La durrée de survie X est inconnue et C v.a de censure avec $C < X$. Autrement dit, soit $X_1, \ldots, X_n$ une suite de taille $n \geq 1$ d'une v.a X continue et soit C la v.a de cencure, on observe alors :

$$Z_i = \min(X_i, C) = X_i \wedge C, \tag{3.1}$$

et

$$\delta_i = \begin{cases} 1 & X \leq C \ \ (X \text{ n'est pas censurée par } C)\,. \\ 0 & X > C \ \ (X \text{ est censurée par } C)\,. \end{cases}$$

Censure à gauche (C censure X à gauche)

Dans ce cas on observe :

$$Z_i = \max(X_i, C) = X_i \vee C,$$

et

$$\delta_i = \begin{cases} 1 & X > C \ \ (X \text{ n'est pas censurée par } C)\,. \\ 0 & X \leq C \ \ (X \text{ est censurée par } C)\,. \end{cases}$$

Censure par intervalle

Pour $C_1 < C_2$ on observe que $C_1 < X < C_2$ avec X inconnue et C_1, C_2 l'une est une borne inférieure et l'autre est une borne supérieure.

3.2 Données tronquées

Il existe une autre donnée incomplète appelée donnée tronquée qui correspond à un échantillonnage biaisé, où seules des données partielles ou incomplètes sont disponibles sur la variable d'intérêt. Comme la censure, même en troncature, il y a trois types, comme suit :

1. **Troncature à droite :** seuls les individus dont la durée de l'évènement est inférieure à un certain seuil sont inclus dans l'étude (X la variable intérêt n'est observée que si $X > Y$ avec Y la variable aléatoire de troncature droite).
2. **Troncature à gauche :** en raison de la structure de la conception de l'étude, nous ne pouvons observer que les individus dont la durée de l'évènement est supérieure à un certain seuil de troncature.(X la variable intérêt n'est observée que si $X < Y$ avec Y la variable aléatoire de troncature gauche).
3. **Troncature par intervalle :** quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.

3.3 Estimation de la fonction de survie

3.3.1 Estimation sous données censurées

La fonction de survie peut être estimée dans le cas de données censurées par plusieurs méthodes non paramétriques dont la plus courante est celle de Kaplan-Meier (1958) [18]. L'idée survivre après le temps t c'est être en vie juste avant t et ne pas mourir au temps t, était le début de l'estimateur de Kaplan-Meier qui est s'écrit sous la forme

$$F_n^{KM}(t) := \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i,n]}}{n-i+1}\right)^{\mathbf{1}_{\{Z_{i,n} \leq t\}}}, \text{ pour } t < Z_{n,n}, \tag{3.2}$$

où $\left(Z_{i,n}, \delta_{[i,n]}\right)_{1 \leq i \leq n}$ est les statistiques d'ordre associées à l'échantillon réellement observé $(Z_i, \delta_i)_{1 \leq i \leq n}$ défini par 3.1.

3.3.2 Estimation sous données tronquées

Dans cette partie, nous présenterons l'estimateur de Woodroofe (1985) de la fonction de queue qui seront présentés dans le cas de données tronquées. L'estimateur est défini par

$$\overline{F}_n^{(W)}(x) := 1 - \prod_{i:X_i > x} \left\{ \exp\left(-\frac{1}{nC_n(X_i)}\right) \right\}, \tag{3.3}$$

et

$$C_n(x) := n^{-1} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x \leq Y_i\}}, \text{ tel que } Y \text{ tronqué } X \text{ à droite.}$$

3.4 Estimation de l'indice des queues sous données incomplètes

3.4.1 Estimation sous censure aléatoire à droite

La queue de la distribution de censure est supposée également varier régulièrement, c'est-à-dire $\overline{F} \in RV(-1/\gamma_1)$ et $\overline{G} \in RV(-1/\gamma_2)$ alors pour tout $x > 0$

$$\lim_{t\to\infty} \frac{\overline{F}(tx)\overline{G}(tx)}{\overline{F}(t)\overline{G}(t)} = x^{-1/\gamma_1}x^{-1/\gamma_2} = x^{-1/\gamma} = \lim_{t\to\infty} \frac{\overline{H}(tx)}{\overline{H}(t)}, \text{ avec } \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2},$$

et donc $\overline{H} \in RV(-1/\gamma)$. La méthode d'estimation de γ a été développée par Einmahl et al (2008) [10] et l'estimateur donné :

$$\widehat{\gamma}^{(c,.)}_{Z,k,n} := \frac{\widehat{\gamma}^{(.)}_{Z,k,n}}{\widehat{p}}, \text{ où } \widehat{p} := k^{-1}\sum_{i=1}^{k} \delta_{[n-i+1,n]} \text{ l'estimateur de } p = \gamma/\gamma_1,$$

avec $\delta_{[1,n]}, ..., \delta_{[n,n]}$ étant les δ correspondant respectivement à $Z_{1,n}, ..., Z_{n,n}$ et $\widehat{\gamma}^{(.)}_{Z,k,n}$ pourrait être n'importe quel estimateur non adapté à la censure.

3.4.2 Estimation sous troncature aléatoire à droite

Soit (X_i, Y_i), $1 \leq i \leq N$, soit $N \geq 1$ copies indépendantes d'un couple (X, Y) de v.a positives indépendantes définies sur un espace de probabilité $(\Omega, \mathcal{A}, P)$, avec les fonctions de distribution marginale continue F et G respectivement. On Suppose que X est tronqué à droite par Y, $\overline{F} \in RV(-1/\gamma_1)$ et $\overline{G} \in RV(-1/\gamma_2)$. Les fonctions de distributions marginaux F^* et G^* correspondant au la distribution conjoint de couple (X, Y) sont donnés par :

$$F^*(x) := \rho^{-1}\int_0^x \overline{G}(w)\, dF(w), G^*(y) := \rho^{-1}\int_0^y F(w)\, dG(w),$$

avec $\rho := P(X \leq Y) = \int_0^\infty F(w)\, dG(w)$ la constante correspond à la probabilité de l'échantillon observé qui est supposé non nul, sinon rien n'est observé.

Gardes et Stupfler (2015) ont utilisé la définition de γ pour construire l'estimateur suivant :

$$\widehat{\gamma}_1^{(GS)}\left(k, k^{'}\right) := \frac{\widehat{\gamma}(k)\,\widehat{\gamma}_2\left(k^{'}\right)}{\widehat{\gamma}_2(k') - \widehat{\gamma}(k)},$$

où

$$\widehat{\gamma}(k) := \frac{1}{k}\sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{X_{n-k,n}}, \widehat{\gamma}_2\left(k^{'}\right) := \frac{1}{k^{'}}\sum_{i=1}^{k^{'}} \log \frac{Y_{n-i+1,n}}{Y_{n-k,n}}.$$

Mais Benchaira et al (2015) [11] ont vu dans l'analyse de valeur extrême qu'il est inhabituel de traiter deux parties distinctes de l'échantillon simultanément $\left(k \neq k^{'}\right)$ et pour cela ils ont proposé un estimateur où $k = k^{'}$ $\left(\text{au lieu de } k/k^{'} \to 1\right)$ donné par :

$$\widehat{\gamma}_1^{(BMN)}(k) := \frac{1}{k}\frac{\sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{X_{n-k,n}} \sum_{i=1}^{k} \log \frac{Y_{n-i+1,n}}{Y_{n-k,n}}}{\sum_{i=1}^{k} \log \frac{X_{n-k,n}Y_{n-i+1,n}}{X_{n-i+1,n}Y_{n-k,n}}}.$$

Aussi en 2022 Mancer et al [12] ont proposé un estimateur semi-paramétrique de l'indice de queue des données tronquées aléatoires défini comme suit

$$\widehat{\gamma}_1^{(MNB)} := \frac{\sum_{i=1}^{k}\left(\overline{G}_{\widehat{\theta}_n}\left(X_{n-i+1,n}\right)\right)^{-1} \log\left(X_{n-i+1,n}/X_{n-k,n}\right)}{\sum_{i=1}^{k}\left(\overline{G}_{\widehat{\theta}_n}\left(X_{n-i+1,n}\right)\right)^{-1}},$$

où G_θ est un modèle connu paramétré la distribution G avec une densité $g_\theta, \theta \in \Theta \subset \mathbb{R}^d, d \geq 1$ et

$$\widehat{\theta}_n := \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} g_\theta(Y_i) / \overline{G}_\theta(X_i).$$

Nous terminons cette partie avec la normalité asymptotique de $\widehat{\gamma}_1^{(MNB)}$, sous la condi-

tion $\lim_{n\to\infty} \sqrt{k}A(a_k) = \lambda < \infty$, tel que $a_k = \left(F^*\left(1-k/n\right)\right)^{-1}$ on a

$$\sqrt{k}(\widehat{\gamma}_1 - \gamma_1) \xrightarrow{\mathcal{D}} \mathcal{N}(\frac{\lambda}{1-\rho_1}, \sigma^2),$$

avec $\sigma^2 := \gamma^2 \left(1 + \frac{\gamma_1}{\gamma_2}\right) \left(1 + \left(\frac{\gamma_1}{\gamma_2}\right)^2\right) \left(1 - \frac{\gamma_1}{\gamma_2}\right)^3$.

3.5 Estimation des quantiles extrêmes sous données incomplètes

3.5.1 Estimation des quantiles extrêmes sous censure aléatoire à droite

Soit un échantillon aléatoire $(Z_i, \delta_i), 1 \leq i \leq n$, de copies indépendantes de (Z, δ), soit les notation suivants :

$$\widehat{a}^{(c,.)} := Z_{n-k,n} M^{(1)} \left(1 - S_{Z,k,n}\right) / \widehat{p},$$
$$S_{Z,k,n} := 1 - \frac{1}{2}\left(1 - \frac{\left(M^{(1)}\right)^2}{M^{(2)}}\right)^{-1} \text{ et } M^{(r)} := \frac{1}{k}\sum_{i=1}^{k}\left(\log \frac{Z_{n-i+1,n}}{Z_{n-k,n}}\right)^r.$$

Alors l'estimateur du quantile extrême $q_v := F^{-1}\left(1-v\right)$ sous censure aléatoire à droite qui a été proposé par Einmahl et al en 2008 [8] est défini par :

$$\widehat{q}_v := Z_{n-k,n} + \widehat{a}^{(c,.)} \frac{\left[\left(1 - F_n^{KM}\left(Z_{n-k,n}\right)\right)/v\right]^{\widehat{\gamma}^{(c,.)}_{Z,k,n}} - 1}{\widehat{\gamma}^{(c,.)}_{Z,k,n}},$$

avec F_n^{KM} l'estimateur de Kaplan-Meier 3.2.

3.5.2 Estimation des quantiles extrêmes sous troncature aléatoire à droite

Benchaira et al (2016) [4] ont proposé un estimateur du quantile extrême d'ordre $(1-v)$ comme estimateur de type Weissman sous troncature aléatoire à droite donné comme suit :

$$\widehat{q}_v := X_{n-k,n}\left(\frac{v}{\overline{F}_n^{(W)}\left(X_{n-k,n}\right)}\right)^{-\widehat{\gamma}_1},$$

avec $\overline{F}_n^{(W)}$ est l'estimateur non paramétrique de Woodroofe 3.3 pour la distribution de queue $\bar{F}$ de la variable intérêt X.

3.6 Exemple sous données tronquées

L'application au jeu de données AIDS. Les données présentent les temps d'infection et d'induction pour $n = 5258$ adultes qui ont été infectés par le virus VIH et ont développé le SIDA au 30 juin 1986. La variable d'intérêt est le temps d'induction T, et la date d'infection M et $M+T$ la date de déclaration de la maladie. L'échantillon $(T_1, M_1), \ldots, (T_n, M_n)$ sont prise entre 0 et 8 (entre 01/04/1978 et 30/06/1986 avec

0 représente les trois premiers mois). On observe (T, M) si $0 \leq T+M \leq 8$ on d'autre terme si $0 \leq M \leq S = 8 - T$, donc on a troncature à gauche. On pose

$$X = \frac{1}{S+\varepsilon},\ Y = \frac{1}{M+\varepsilon},$$

pour que nous ayons une troncature à droite, avec $\varepsilon = 0.05$. On peut ajuster F et G par la loi de Fréchet de deux paramètres (i.e : $H_{(a,r)}(x) = \exp(-a^r x^{-r})$, $a > 0, r > 0, x > 0$), donc F et G à queue lourde, et l'estimateur de a et r donne par : $\widehat{a} = 0.004, \widehat{r} = 2.1$. Ainsi, on peut considérer que G est connu et vaut $G_\theta = H_{(\widehat{a},\widehat{r})}$, en utilisant l'algorithme de Thomas et Reiss on trouve : $k = 19, X_{n-k,n} = 0.356, \widehat{\gamma}_1^{(MNB)} = 0.917$. et la valeur de l'estimateur de quantile extrême pour $v = 1/2n$ est : $\widehat{q}_v = 0.061$. Donc pour trouver la fin de la date d'indiction t_{end} compensons dans ce qui suit : $P(X \geq \widehat{q}_v) = v$ et qui nous donne : $P(T \geq 1/\widehat{q}_v - 8 + \varepsilon)$ ce qui implique que $t_{end} = 1/\widehat{q}_v - 8 + \varepsilon = 8.4$. Donc le temps de fin d'induction du SIDA est : 8 ans, 4 mois. [12]

Chapitre 4

Simulation et applications

4.1 Simulation

En se basant sur des échantillons de tailles finies, nous étudions ci-dessous la performance de l'estimateur de la prime de réassurance et l'estimateur des quantiles extrêmes de Weissman.

A cet effet :

• Nous générons un échantillon pour deux distributions ayant des queues lourdes, à savoir les distributions de $burr(\gamma, \delta)$ et $Fréchet(\gamma)$.

• Nous choisissons une taille de l'échantillon $N = 1000$ et le nombre de valeurs extrêmes ($k = 100$).

• Nous évaluons la performance des deux estimateurs à travers les courbes associés.

4.1.1 Estimateur de la prime

En utilisant le logiciel R, nous obtenons les courbes suivants :

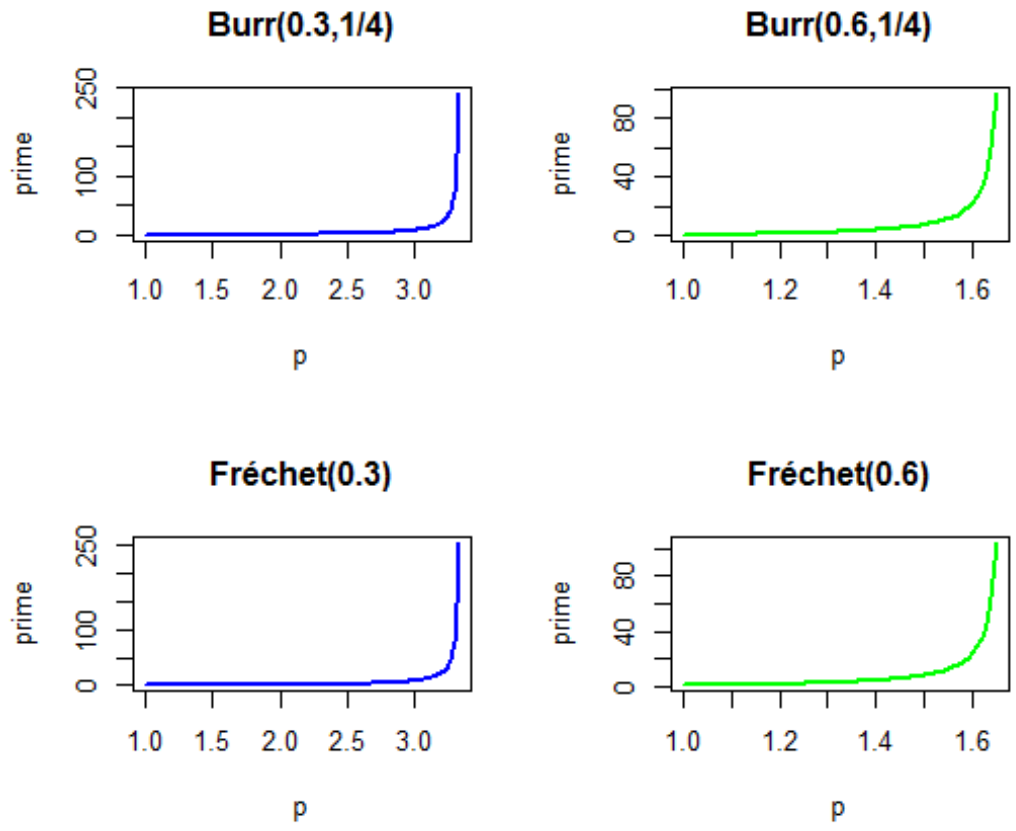

FIG. 4.1 – Le graphe de la prime pour les distributions de Burr et Fréchet.

Dans la figure 4.1, la prime est tracée en termes de l'indice d'aversion au risque ρ, aussi il est à noter que ce soit dans le cas de la loi de Burr ou la loi de Fréchet, plus le ρ proche de l'inverse de l'indice de queue, plus la valeur de la prime est élevée. Une autre remarque, plus la valeur de γ est élevée, plus la prime augmente.

4.1.2 Estimateur de Weissman

D'après le logiciel R, nous obtenons la figure 4.2. Dans ce graphe le quantile extrême est tracé en termes de la probabilité p, et on remarque que lorsque p augmente, la valeur de x_p diminue, cela correspond au fait que lorsque le quantile tend vers à l'infini, la probabilité d'un v.a plus grand que ce quantile, tend vers à 0.

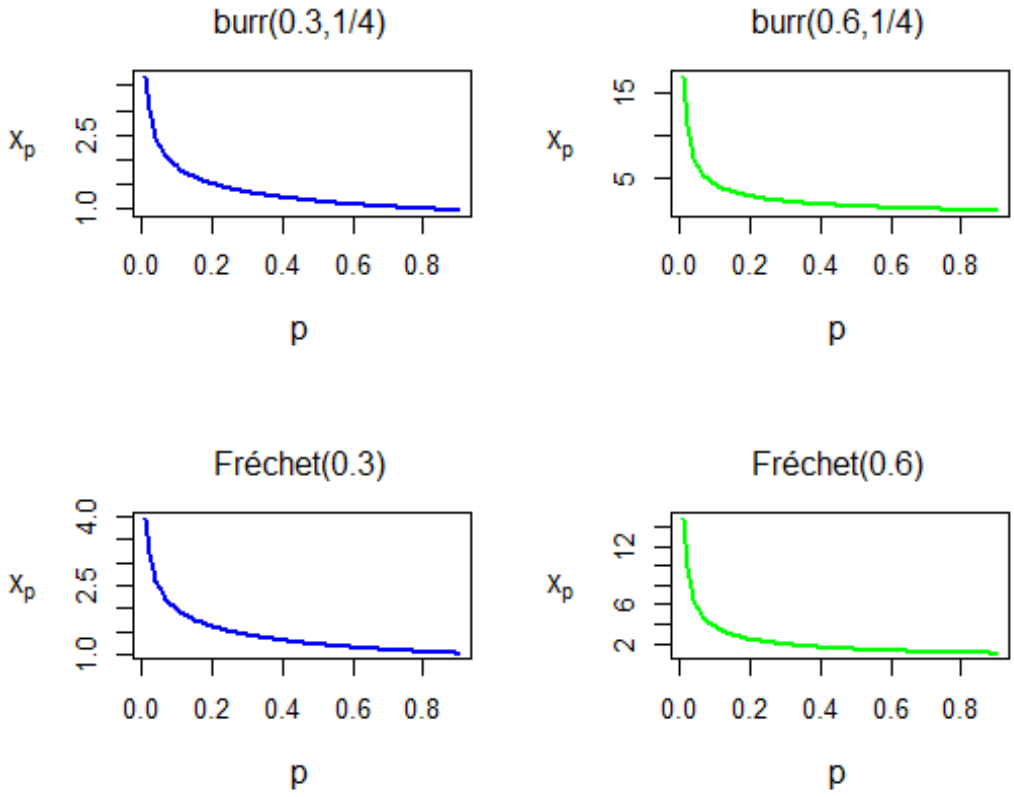

FIG. 4.2 – L'estimateur de Weissman pour les distributions de Burr et Fréchet.

4.2 Applications

Nous allons étudié deux bases de données :

- Pertes des incendies danoises (danish fire) dans la période de 03 janvier 1980 à 31 décembre 1990.
- Nombre de décès de Covid-19.

Danish fire : Dans l'histoire de l'assurance, les incendies danoises dans la période de dix ans durant a marqué un tournant, tel que ces incendies ont causé d'énormes pertes financières aux compagnies d'assurance, ce qui a conduit à remettre en cause les techniques d'évaluation des risques utilisées par ces compagnies. La théorie des valeurs extrêmes, qui vise à étudier les évènements extrêmes tels que les incendies majeurs dans l'évaluation des risques et la tarification des assurances a été développée

pour résoudre ces problèmes. Ainsi, les incendies danoises ont été la source qui a poussé les compagnies d'assurance à développer de nouvelles stratégies pour mieux comprendre, gérer et estimer les risques. Dans cette partie, nous allons modéliser la distribution des grandes pertes on utilise la GEVD pour les données des incendies danoises qui consistant en 2167 observations et ça sera après l'analyse de ces données.

Covid-19 : L'ensemble de données de Covid-19 comprend des données du 3 Juin 2020 sont extraites de la référence [6]. Dans cette partie, nous allons modéliser la distribution de la queue des données des décès de Covid-19 on utilise la GPD.

Le but : Par le biais de la théorie des valeurs extrêmes, nous allons estimer la queue de distribution de chacune de ces deux ensembles de données. Ceci nécessite l'utilisation de certaines commandes et fonctions spécifiques du langage R.

4.2.1 Détection de queue lourde pour les pertes des incendies danoises

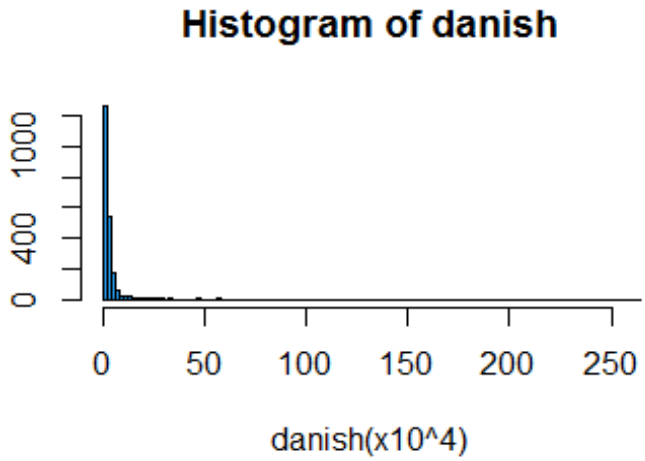

FIG. 4.3 – Histogramme des données des incendies danoises

D'après la figure 4.3, nous remarquons que les données sont asymétrique à droite (positive skewness), et leurs coefficient d'asymétrie est égal à $S = 18.74983$ (Il a été calculé en utilisant le logiciel R). Cette asymétrique à droite est une indiquation de la queue lourde, et dans ce cas la moyenne est supérieure à la médiane, cela est illustré dans le tableau suivant :

Min	1st Qu	Median	Mean	3rd Qu	Max
1.000	1.321	1.778	3.385	2.967	263.250

TAB. 4.1 – Statistiques des données des incendies danoises.

En plus, le tableau 4.1 représente la valeur minimale des données ($Min = 1.000$), le 1^{er} quartile : 1.321, le $3^{ème}$ quartile : 2.967, et la valeur maximale 263.250.

Q-Q Plot :

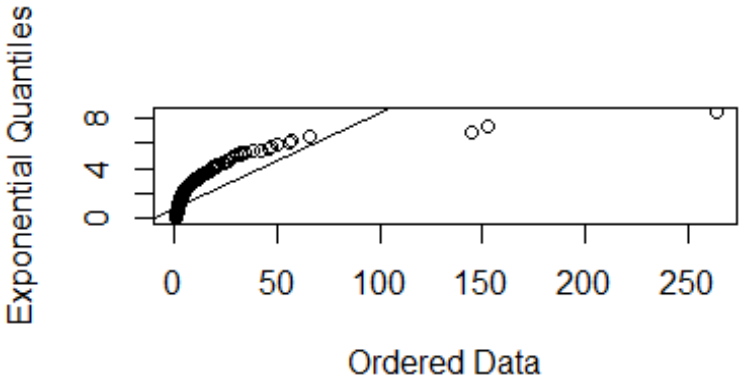

FIG. 4.4 – Représentation graphique quantiles-quantiles pour les données danoises

Puisque la longueur des données est de 2167, il semble que l'on puisse en déduire que la queue des données est plus lourde que l'exponentielle et la concavité dans la figure 4.4 est un signe de l'existance de la queue lourde.

Le Kurtosis : A l'aide du langage R, nous avons trouvé le coefficient d'aplatissement $K = 485.6461 > 3$, donc la queue est lourde, et ceci d'après ce que l'on trouve dans la partie théorique 2.1.

4.2.2 Modélisation de la distribution des grandes pertes

Nous allons modéliser la distribution des données "danish" à l'aide de la méthode des blocs maxima. Les données sont divisées en 217 blocs, puis prend la plus grande valeur de chaque bloc et étudions ensuite le comportement des valeurs obtenues.

Estimation des paramètres de la GEVD

1. Estimations par la méthode de maximum de vraisemblance (E.MV) et la méthode L-moments (E.LM) :

	P. de position $\widehat{\mu}$	P. d'échelle $\widehat{\sigma}$	P. de forme $\widehat{\gamma}$
MV	5.8088457	3.9712825	0.6501871
LM	5.9632150	4.3019387	0.5555336

TAB. 4.2 – E.MV et E.LM pour les données des incendies danoises.

2. Estimations de Hill et de Pickands :

	Estimateur de Hill $\widehat{\gamma}^H$	Estimateur de Pickands $\widehat{\gamma}^P$
P. de forme $\widehat{\gamma}$	0.5953429	0.7223427

TAB. 4.3 – Estimateurs de Hill et Pickands pour les données des incendies danoises.

3. Estimateur de GLW : $\widehat{\gamma}^{(GLW)} = 0.6320656$.

D'après le tableau 4.2 et 4.3 et l'estimateur $\widehat{\gamma}^{(GLW)}$ on remarque que le paramètre de la forme est de signe positif ce qui signifie que la GEVD est de type Fréchet. Dans l'étape suivante on va tester l'ajustement par la loi de Fréchet.

Test de Kolmogrov-Smirnov

Les hypothèses de test sont données par :

$$\begin{cases} H_0 : \text{ La queue suit la loi de Fréchet.} \\ H_1 : \text{ La queue ne suit pas la loi de Fréchet.} \end{cases}$$

A l'aide du langage R, on trouve $p - value = 0.3127$, cette valeur est supérieure au seuil de signification $\alpha = 0.05$, donc on accepte l'hypothèse H_0 (c'est à dire la queue suit la loi de Fréchet).

Représentaion graphique du modèle choisi

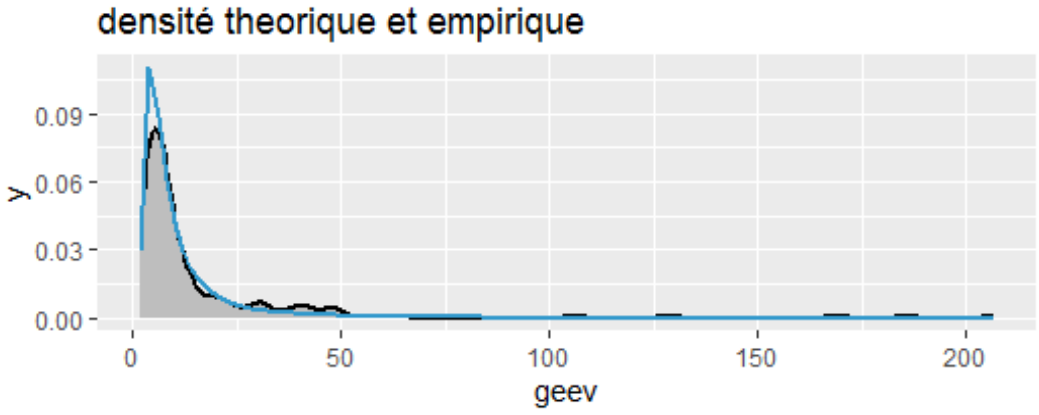

FIG. 4.5 – Représentation graphique de la densité théorique et la densité empirique.

On remarque que la figure 4.5 montre que la densité empirique (ligne noir) correspond approximativement à la densité théorique (ligne bleu), donc on accepte que la queue

est bien ajustée par la loi de Fréchet.

Une autre méthode est utilisée pour modéliser les queues lourdes de distributions est appelée la méthode de POT. Ci-dessous, nous allons modéliser les données de Covid-19 en utilisant cette méthode, après avoir vérifié que les données de Covid-19 ont une queue lourde.

4.2.3 Les propriétés typiques de la distribution à queue lourde des données de Covid-19

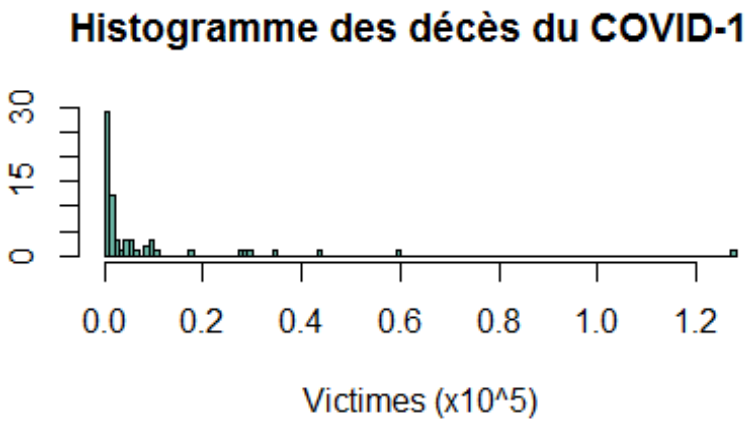

FIG. 4.6 – Histogramme des décès du Covid-19.

Dans ce cas, nous remarquons la même remarque que nous avons remarquée pour l'histogramme des données danoises 4.2.1. Dans le tableau suivant on va voir quelques statistiques des données :

Moyenne	Medianne	Variance	Skewness	Kurtosis
$Moy = 7670.121$	$Med = 1385$	$V = 347720455$	$S = 4.635261$	$K = 27.88347$

La moyenne des données autour 7670.121 et la medianne égal à 1385 avec un grand variance qui mesure la dispertion des données. Le coefficient de Skewness est positif

ce qui signifie que la queue est étendue du côté des grandes valeurs (c'est ce que nous remarquons dans la figure 4.6), avec un kurtosis supérieure à 3 (indique que la queue est lourde)

Le rapport du maximum sur la somme

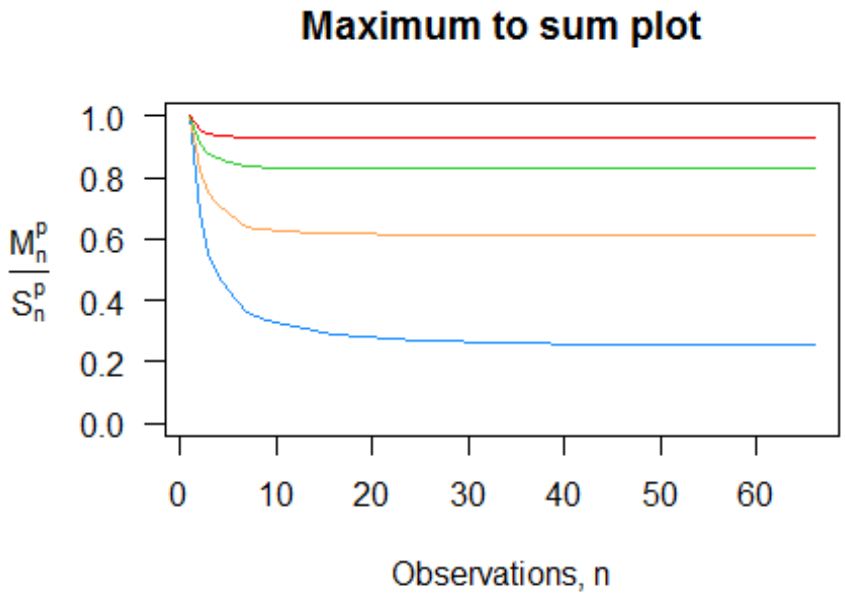

FIG. 4.7 – Taux de moralité entre les différents pays.

En augmentant la valeur de p, on réduit l'impact des pays ayant des nombres de décès cumulés faibles et on donne plus de poids aux pays ayant des nombres de décès cumulés élevés. Cela réduit la variabilité des taux de mortalité entre les différents pays. D'après la figure 4.7 pour différentes valeurs de p $(1, 2, 3$ et $4)$, nous pouvons conclure que la queue lourde est une caractéristique de la distribution des taux de mortalité.

Coefficient de Gini

En calculant le coefficient de Gini, nous pouvons quantifier l'inégalité de la répartition des données. Dans ce cas le coefficient de Gini égal à 0.9078399 proche de 1, cela signifie que les décès Covid-19 ne sont pas répartis uniformément dans la population. On peut donc conclure que la queue de la distribution des données est à queue lourde.

Plot d'excès moyenne

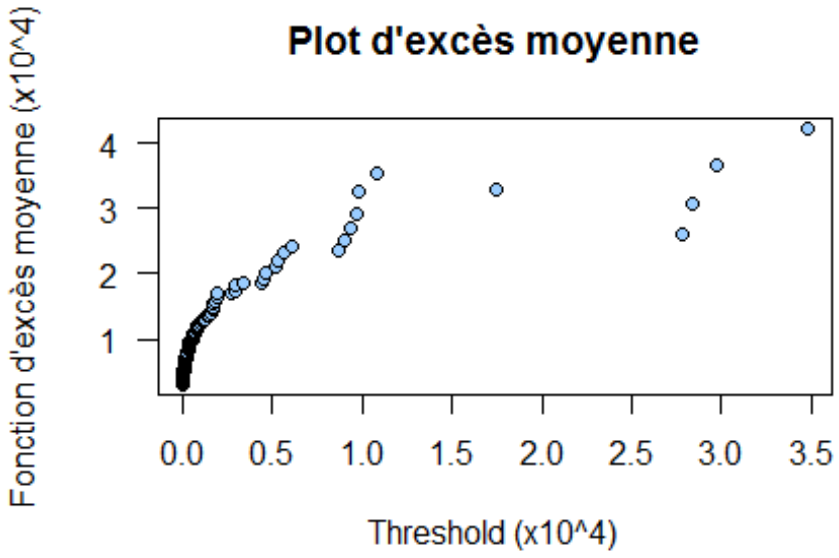

FIG. 4.8 – Plot d'excès moyenne pour les données des décès de Covid-19.

▶ Le graphique 4.8 montre une forte augmentation de la moyenne d'excès du seuil 0 à environ 12000, qui a ensuite commencé à diminuer jusqu'à un seuil proche de 30000. Cependant, malgré cette diminution, la moyenne d'excès reste élevée.

▶ Les résultats montrent également une grande variation entre les pays, avec de nombreux pays ayant des taux de moralité beaucoup plus élevés que d'autres. (Les pays les plus touchés par la pandémie sont les EU, le Brésil et le Royaume-Uni, avec des taux de moralité cumulés élevés).

▶ On peut donc dire que la queue de la distribution des données est lourde.

Courbe de Lorenz

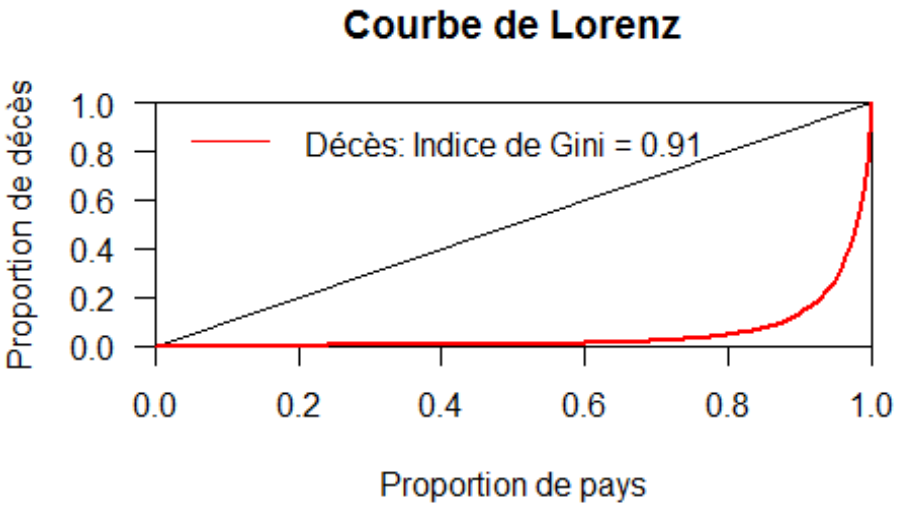

FIG. 4.9 – Courbe de Lorenz des décès cumulés de Covid-19.

On remarque dans la figure 4.9 une forte déviation de la courbe par rapport à la droite, ce qui signifie que la répartition est inégale, c'est à dire que certains pays ont été beaucoup plus touchés que d'autres par l'épidémie. Cette distribution inégale indique qu'il existe une queue lourde.

4.2.4 Modélisation de la queue des décès cumulés de Covid-19

Estimation des paramètres de la GPD

1. Estimations par les méthodes de (MV) et (LM) :

	P. d'échelle $\widehat{\sigma}$	P. de forme $\widehat{\gamma}$
MV	1752.153205	1.098218
LM	1752.1803805	0.7654415

TAB. 4.4 – E.MV et E.LM pour les données de Covid-19.

2. Estimations de Hill et Pickands :

	Estimateur de Hill $\widehat{\gamma}^H$	Estimateur de Pickands $\widehat{\gamma}^P$
P. de forme $\widehat{\gamma}$	0.6725703	0.5590987

TAB. 4.5 – Estimateurs de Hill et Pickands pour les données de Covid-19.

3. Estimateur de GLW : $\widehat{\gamma}^{(GLW)} = 0.8330848$.

Représentation graphique du modèle choisi

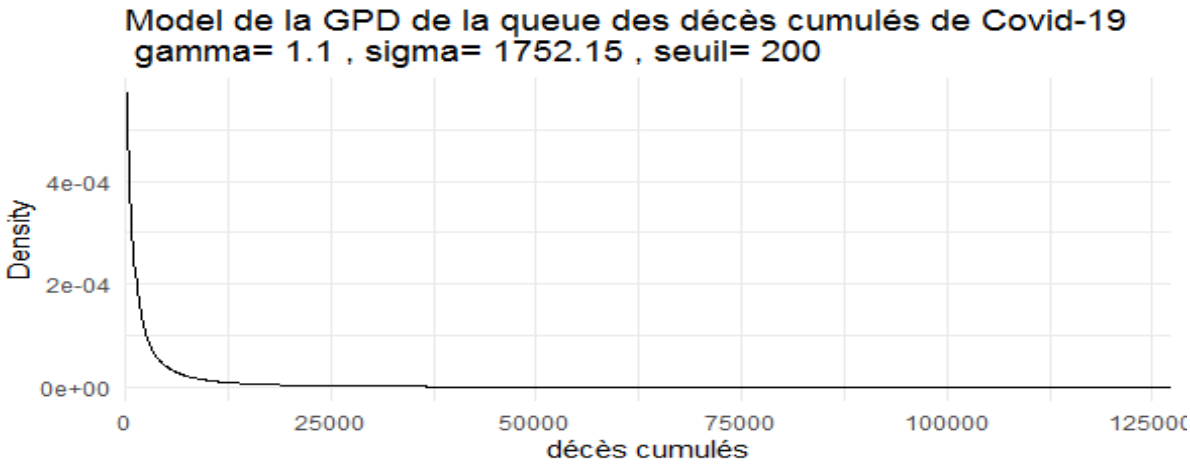

FIG. 4.10 – Modélisation de la queue des décès cumulés de Covid-19.

Conclusion

L'estimation des queues de distributions est un domaine important et nécessite la sélection de méthodes appropriées basées sur les caractéristiques de la distribution, y compris la forme de la queue et l'échantillon disponible.

Dans le cadre de ce mémoire, nous avons exploré plusieurs méthodes d'estimation des queues de distributions, notamment dans le cas des données complètes : la méthode de Hill, la méthode MV, la méthode GLW, etc et nous avons pris un petit aperçu du cas des données incomplètes. Nous avons également examiné des applications concrètes de ces méthodes à des données réelles, telles que des données d'assurance et des données épidémiques.

On peut dire qu'il n'y a pas de méthode universelle pour estimer les queues de distributions dans toutes les situations et chaque méthode a ses avantages et ses limites, ainsi qu'il est préférable d'utiliser plusieurs méthodes pour obtenir une estimation fiable et précise. Les résultats obtenus peuvent également être utilisés pour prendre des décisions éclairées sur la gestion des risques. Par exemple, en assurance, l'estimation de la probabilité d'évènements extrêmes peut être utilisée pour déterminer les primes, et aussi pour la pandémie de Covid-19, la méthode de modélisation des valeurs extrêmes peut être utilisée pour étudier les conséquences financières et évaluer les stratégies de prévention, à cela s'ajoutent les erreurs d'estimation entraînant des conséquences graves, notamment en cas d'effets graves d'évènements rares.

En résumé, la modélisation des queues de distributions est une étape critique dans

la compréhension des risques associés à la variable étudiée. Bien que les méthodes GEVD et GPD présentent certains inconvénients (par exemple, la méthode GEVD nécessite une grande quantité de données pour être fiable, de plus, les données doivent être de bonne qualité et doivent être distribuées de manière aléatoire et les résultats obtenus par la méthode GEVD peuvent également varier fortement selon les hypothèses de départ (distribution), quant à la méthode GPD elle peut être sensible aux valeurs aberrantes, qui peuvent avoir un impact significatif sur les estimations. La mauvaise sélection du seuil peut conduire à une erreur dans la modélisation des valeurs extrêmes, et donc cela conduit à un impact sérieux sur les anticipations futures de la variable étudiée), le développement se poursuit dans ce domaine pour permettre une meilleure compréhension et modélisation d'évènements extrêmes.

Bibliographie

[1] Arnold, B.C., Balakrishnan, N. et Nagaraja, H.N. (1992). A First Course in Order Statistics. Wiley, New York.

[2] Balakrishnan N., Rao C. (1998). Handbook of Statistics 16_ Order Statistics_ Theory and Methods.

[3] Benchaira, S., Meraghni, D., and Necir, A. (2015). On the asymptotic normality of the extreme value index for right-truncated data. Statist. Probab. Lett, 378 − 384.

[4] Benchaira, S., Meraghni, D., and Necir, A. (2016). Tail product-limit process for truncated data with application to extreme value index estimation. Extremes, 19(2) : 219 − 251.

[5] Bobée, B. et Robitaille, R. (1975). Etude sur les coefficients d'asymétrie et d'aplatissement d'un échantillon. INRS-Eau, rapport scientifique no 49, 22 p.

[6] Covid-19 : https ://www.en.ibe.med.uni-muenchen.de/research/heavy-tail-issues/index.html.

[7] De Haan L., Ferreira A. (2006). Extreme Value Theory : An Introduction SpringerVerlag.

[8] Einmahl, J.H.J., Fils-Villetard, A. and Guillou, A., (2008). Statistics of extremes under random censoring. Bernoulli. 14, no.1, 207 − 227.

[9] F.Delmas et B. Jourdain. (2006). Modèles aléatoires. Applications aux sciences.

[10] Hall, P. et Welsh, A.H. (1985). Adaptive Estimates of Parameters of Regular Variation. Annals of Statistics 13, 331 − 341.

[11] Hosking, J.R.M , Wallis.J.R. (1997). Regional frequency analysis_ An approach based on L-Moments-Cambridge University Press.

[12] Mancer, S., Necir, A., Benchaira, S. (2022) . Semiparametric tail-index estimation for randomly right-truncated heavy-tailed data. Arab Journal of Mathematical Sciences

[13] M. C. JONES.(1991). Estimating densities, quantiles, quantile densities and density quantiles,Volume(44),page 721 − 727.

[14] Necir, A., Meraghni, D., Meddi, F. (2007) . Statistical estimate of the proportional hazard premium of loss. Scandinavian Actuarial Journal, 1 − 15.

[15] Sheater, S. J. and Marron, J. S. (1990). kernel quatile estimtors. journal of the American Statistical Association.

[16] William H. Asquith. (2012) . Distributional analysis with L-moment statistics using the R environment for statistical computing.

[17] Yahia, D. (2010) . Conditional Quantile for Truncated Dependent data. Thése de doctorat d'université Mohamed khider, Biskra, Algeria.

[18] Kaplan, E.L, Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association, 53 :457 − 481.

Annexe A : Logiciel R

4.3 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

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

$E(.)$	:	Espérence mathématique.
$Cov(.,.)$	:	Covariance.
$Var\,(.)$	:	Variance.
$v.a$	:	Variable aléatoire.
$i.i.d$	:	Indépendantes et identiquement distribuées.
$R\,(m)$	:	Rang de X_m.
$F\,(x-)$	:	$P\,(X < x)\,.$
$\mathbb{1}_a$	:	Indicatrice de a.
$X_{i,n}$	:	$i^{ème}$ statistique d'ordre dans un échantillon de taille n.
$\xrightarrow{p}$	:	Convergence en probabilité.
$\xrightarrow{p.s}$	:	Convergence presque sur.
POT	:	Peak Over Threshold.
$B\,(a,b)$	:	Fonction bêta complète.
$\Gamma\,(a)$	:	Fonction gamma complète.

$I_\alpha(a,b)$	:	fonction bêta incomplète.
$J(.)$	:	Fonction de poids.
$[np]$	:	Partie entier de np.
$N(\mu,\sigma^2)$	:	Loi normal de moyenne μ et de variance σ^2.
MSE	:	L'erreur au moyenne quadratique.
$GEVD$	:	Distribution Généralisée des Valeurs Extrêmes.
$D(G_\gamma)$	:	Domaine d'attraction maximum de G_γ.
w_F	:	Point terminal.
GPD	:	Distribution Généralisée de Pareto.
$\xrightarrow{\mathcal{D}}$	:	Convergence en distribution.
EMV	:	Estimateur de Maximum de Vraisemblance.
$U(.)$	:	Fonction quantile de queue.
x_p	:	Quantile d'ordre $1-p$ sous données complètes.
q_v	:	Quantile d'ordre $1-v$ sous données incomplètes.
LM	:	L-Moment.
MV	:	Maximum de Vraisemblance.
E.MV	:	Estimateur de Maximum de Vraisemblance.
E.LM	:	Estimateur de L-Moment.

RÉSUMÉ :

Ce mémoire porte sur l'estimation de l'indice des valeurs extrêmes qui correspond aux trois domaines d'attraction, à savoir Gumble, Fréchet et Weibull. Nous présentons aussi quelques estimateurs de ce crucial paramètre aux cas des données incomplètes. En outre nous exposons des méthodes d'estimation de la queue de distribution et les quantiles extrêmes associés. Nous terminons notre travail par des simulations, en utilisant le langage R, avec des applications aux données réelles.

Mots-clés : **Estimation des queues, les valeurs extr*êmes, quantiles* extr*êmes.***

ملخص

تتناول هذه الأطروحة تقدير مؤشر القيم المتطرفة الذي يتوافق مع مجالات الجذب الثلاثة، وهي Gumble و Fréchet و Weibull. نقدم أيضًا بعض تقديرات هذه المعلمة المهمة في حالة البيانات غير الكاملة. بالإضافة إلى ذلك ، نكشف عن طرق لتقدير ذيل التوزيع والكميات المتطرفة المرتبطة به. ننهي عملنا بالمحاكاة ، باستخدام لغة R ، مع تطبيقات لبيانات حقيقية.

كلمات مفتاحية : تقدير الذيول ، القيم المتطرفة، الكميات المتطرفة.

ABSTRACT

We deal with the estimation of the extreme values index pertaining to the three domains of attraction, namely Gumble, Fréchet and Weibull. We also present some estimators of this crucial parameter in the case of incomplete data. Moreover, we expose some estimation methods for the distribution tail and its corresponding high quantiles. Finally, by using the R language, we provide a simulation study with applications to real data.

KEYWORDS: **Estimation of tails, extreme values, extreme quantiles.**