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

KIDOUS Sihem

Titre :

Estimation de distribution,valeurs extrêmes

Membres du Comité d'Examen :

Pr.	SAYAH Abdallah	UMKB	Président
Pr.	BENATIA Fatah	UMKB	Encadreur
Dr.	KHEIREDDINE Souraya	UMKB	Examinatrice

Juin 2024

Dédicace

Je dédie cet 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. Je veux présenter mon diplôme à mon père allah yarhmo.

À mes frères et sœurs, Oussama, Ramzi, Achraf, Sara 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.

REMERCIEMENTS

Tout d'abord, je remercie **"Allah"** Le Tout-Puissant de m'avoir aidé et donné la santé et volonté pour arriver à ce stade.

Mes vifs remerciements, sont adressés à mon encadreur **Pr. BENATIA FATAH** pour ses précieux conseils, ses orientations pertinentes et sa patience tout au long de la réalisation de ce mémoire.

Je tiens à remercier : **Pr. Sayah Abdallah** et **Dr. Kheireddine Souraya** qui m'ont fait l'honneur de faire partie du jury de soutenance.

Je remercie touts les enseignants qui ont contribué à ma formation, ainsi que tous les employés du département de mathématiques.

Je remercie tout particuliérement mes parents, pour leur encouragement et soutien sur tous les aspects, ainsi que toute ma famille.

Je n'oublie pas l'ensemble de mes amis mes proches et aussi mes collégues d'étides.

À ceux qui ont contribué, de prés ou de loin, à la réalisation de ce modeste travail.

Un grand merci à vous tous.

Table des matières

Table des figures

Introduction

Nous allons dans ce mémoire présenter une contribution pour faire connaitre la théorie des valeurs extrêmes, les lois limites de variable maximale. Nous rapelons que la théorie des valeurs extrêmes permet de déterminer le comportement asymptotique des maximas des valeurs prises par les variables aléatoire indépendantes et identiquement distribuées (i.i.d). Cette distribution limite s'écrit en fonction de certains paramètres applés paramètres des distributions des queues. Ces paramètres inconnues en générale sont estimes par deux méthodes qui sont la méthode des maxima qui repose sur la valeur extrême $X_{n,n}$ a disposition et la méthode des excés au dela d'un seuil qui consiste à utilisé la distribution des données supérieures à un certain seuil applée aussi méthode P.O.T (Peaks over Threshold). Aprés une introduction générale, le premier chapitre est consacré aux statistiques d'ordre et présenté comme base de départ sur l'étude des valeurs extrêmes $X_{n-k+1,n}, ..., X_{n,n}$ les plus grandes valeurs de la distribution. Le deuxième chapitre porte sur la définition et les propriétés essentielles des valeurs extrêmes et leurs théorie. Le troisiéme chapitre comporte une étude sur l'estimateur de l'indice de queue des distributions de Paréto généralisée. L'estimateur du maximum de vraisemblence et l'éstimateur de Hill sont présentés dans cette partie. Enfin une application de ces derniérs résultats sous forme d'une simulation avec le logiciel R clôture notre travail.

Chapitre 1

Statistique d'ordre

Dans ce chapitre, nous allons présentés la définition de statistique d'ordre qui joue un rôle important dans la théorie des valeurs extrêmes, celle ci représente un point de départ pour étudier les valeurs extrêmes qui sont en réalité les statistique d'ordre extrêmes

1.1 Définition de la statistique d'ordre

Définition 1.1.1 *Un vecteur aléatoire* $(X_1, ..., X_n)$ *est une suite de v.a de fonction de répartition* $F_{X_1,...,X_n}(x_1, ..., x_n)$ *ou de densité* $f_{X_1,...,X_n}(x_1, ..., x_n)$ *si en plus les v.a sont indépendantes alors :*

$$F_{X_1,...,X_n}(x_1, ..., x_n) = F_{X_1}(x_1)...F_{X_n}(x_n).$$
$$f_{X_1,...,X_n}(x_1, ..., x_n) = f_{X_1}(x_1)...f_{X_n}(x_n).$$

Si les v.a ont la même loi on dit qu'elles sont identiquement distribuées

Définition 1.1.2 *Si les deux conditions de indépondance et de l'indetique distribution sont réalisés, ont dit que* $(X_1, ..., X_n)$ *est un n-échantillon ie que c'est une suite*

de v.a (i.i.d) de même densité $f_X(x)$

$$f_{X_1,...,X_n}(x_1,...,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = (f_X(x))^n.$$

Définition 1.1.3 *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

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

– 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 $\{1,...,n\}$ correspondantes $(R(1),...,R(n))$.

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*

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

– La $1^{ère}$ statistique d'ordre est notée par : $X_{1,n} = \min(X_1,...,X_n)$.

– la $i^{ème}$ statistique d'ordre est notée par : $X_{i,n}$.

– La dernière statistique d'ordre est notée par : $X_{n,n} = \max(X_1, ..., X_n)$.

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. Noté par $F_{X_{1,n}}$, $F_{X_{n,n}}$, $F_{X_{i,n}}$ la distribution de chaque statistique d'ordre ($X_{1,n}, X_{n,n}$ et $X_{i,n}$) et $f_{X_{1,n}}$, $f_{X_{n,n}}$, $f_{X_{i,n}}$ les densités corespondantes.

Fonction de répartition de la statistique d'ordre

Fonction de répartition 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 indépendentes et de même loi F_X alors :

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

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

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

Fonction de répartition de $X_{i,n}$ Nous avons

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

Densité de la statistique d'ordre

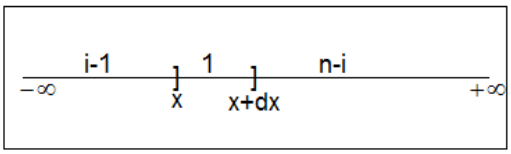
Densité de $X_{1,n}$ On sait que la densité $f_X(x) = \frac{\partial F_X(x)}{\partial x}$ alors, la densité de $X_{1,n}$ est donnée par :

$$f_{X_{1,n}}(x) = n f_X(x)(1 - F_X(x))^{n-1}.$$

Densité de $X_{n,n}$ La densité correspondante de $X_{n,n}$ est donnée par :

$$f_{X_{n,n}}(x) = n f_X(x)(F_X(x))^{n-1}.$$

Densité de $X_{i,n}$ Il existe $i-1$ de X_k 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 :

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

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 éclairée. Cependant, la distribution que nous avons traité dans cette section est une distribution dégénérée 2.1, 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 ?, la réponse est dans le deuxième chapitre2.

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 : $F_{(X_{i,n},X_{j,n})}(x,y) = P((X_{i,n} \leq x) \bigcap (X_{j,n} \leq y))$.

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

– **2ème 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 :

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

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).*

Chapitre 2

Introduction à la théorie des valeurs extrêmes

Dans ce chapitre, nous présentons la théorie des valeurs extrêmes,qui repose sur la statistique d'ordre d'un échantillon de taille n et le comportement asymptotique du maximum. Pour cela, nous introduisons deux théorèmes essentiels bien connues, le premier concerne les lois des valeurs extrêmes (GEV :Generalized Extreme Value), qui étudie la distribution du maximum. Cette approche initiale se fonde sur le théorème de Fisher et Tippett, établi en 1928, et complété par Gnedenko en 1943. Le second théorème porte sur les lois des excés (POT :Peaks Over Threshold), qui utilisent les observations dépassant un seuil prédéterminé. Les différences entre ces observations et le seuil sont appelées les excés. Cette approche s'appuie sur la convergence vers une loi de Pareto Généralisée (GPD), établie par les théorèmes de Belkema, de Haan et Pickands en 1975, pour modéliser les excés. Enfin, nous examinons les caractérisations des domaines d'attraction du maximum et introduisons la notion de fonctions à variations réguliéres.

2.1 Distribution dégénérée

Soit $F_X(x) \in [0,1]$, on a $F_{X_{n,n}}(x) = (F_X(x))^n$, alors :

$$\lim_{n\to+\infty} F_{X_{n,n}}(x) = \begin{cases} 1 & \text{si } F_X(x) = 1 \\ 0 & \text{si } F_X(x) < 1 \end{cases}$$

Définition 2.1.1 *On dit que $G_X(x)$ est une fonction dégénérée ssi : $G_X(x)$ prends les valeurs* 0 *ou* 1.

2.2 Distribution des valeurs limites

Définition 2.2.1 *On appelle valeur minimale et maximale d'une distribution notées réspectivement α_F et ω_F est définés par :*

$$\alpha_F = \min\{x : F_X(x) > 0\}.$$
$$\omega_F = \max\{x : F_X(x) < 1\}.$$

2.2.1 Loi de la somme

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Convergence faible

$$\overline{X_n} \xrightarrow{p} \mu \text{ lorsque,} n \to +\infty.$$

Convergence forte

$$\overline{X_n} \xrightarrow{p.s} \mu \text{ lorsque,} n \to +\infty.$$

2.3 Théoréme Central limite (T.C.L)

Si $X_1, ..., X_n$ (i.i.d) de moyenne μ et de variance σ^2 alors :

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1) \text{ lorsque,} n \to +\infty.$$

2.4 Classe d'equivalence de distributions

On dit que deux distribution $F(x)$ et $G(x)$ appartiennent à la même class d'équivalence ssi : $\exists\ a, b : a > 0,\ b \in \mathbb{R}$ telle que :

$$F(ax + b) = G(x).$$

2.4.1 Distribution max-stable

On dit qu'une distribution $G(x)$ est max-stable s'il existe des suites de réeles $a_n > 0$, $b_n \in \mathbb{R}$ telles que :

$$\lim_{n \longrightarrow \infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = G(x).$$

Exemple 2.4.1 *(loi de Fréchet)*

Soit X une v.a de distribution de Fréchet ie :

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \exp(-\frac{1}{x}) & \text{si } x \geq 0. \end{cases}$$

Montrer que $F_X(x)$ est une max-stable pour $a_n = n$, $b_n = 0$:

$$\begin{aligned} \lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) &= \lim_{n\to\infty} F^n(a_n x + b_n) \\ \lim_{n\to\infty} (F_X(nx))^n &= \lim_{n\to\infty} (\exp(-\frac{1}{nx}))^n \\ &= \lim_{n\to\infty} \exp(-\frac{n}{nx}) \\ &= \exp(-\frac{1}{x}) \\ &= \begin{cases} 0 & \text{si } x \leq 0 \\ \exp(-\frac{1}{x}) & \text{si } x \geq 0. \end{cases} \end{aligned}$$

2.5 Théoréme de Fisher _Typet

Théorème 2.5.1 *Soit $X_1, ..., X_n$ une suite i.i.d de distribution $F_X(x)$. S'il existe deux constantes de normalisation $a_n > 0$ et $b_n \in \mathbb{R}$, telle que :*

$$\lim_{n\longrightarrow\infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = \lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n)$$

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

Alors la limite est une distribution non dégénérée $G(x)$ (appelée loi des valeurs extrêmes) et prend l'un des trois types de lois suivantes :

– La distribution de Fréchet ($\gamma > 0$).

– La distribution de Gumbel ($\gamma = 0$).

– La distribution de Weibull ($\gamma < 0$).

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

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

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

2.6 Domaine d'attraction

Définition 2.6.1 *On dit qu'une distribution $F_X(x)$ appartient au domaine d'attraction d'une distribution $G_X(x)$ notée $F_X(x) \in D(G_X(x))$ ssi :*

$\exists\ a_n > 0,\ b_n \in \mathbb{R}$

$$\lim_{n \longrightarrow \infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = G_X(x).$$

Exemple 2.6.1 *On a $F_X(x) \rightsquigarrow \mathcal{E}(1)$*

Montrer que $F_X(x) \in D(\Lambda(x))$

$$\exists\ a_n > 0, b_n \in \mathbb{R}\ :\ \lim_{n\to\infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = \exp(-\exp(-x))$$

$$a_n = 1\ \ ,\ \ b_n = \log(n)$$

$$\begin{aligned} \lim_{n\to\infty} P(\frac{X_{n,n} - \log(n)}{1} \leq x) &= \lim_{n\to\infty} F_{X_{n,n}}(x + \log(n)) \\ &= \lim_{n\to\infty} (F_X(x + \log(n)))^n = \lim_{n\to\infty} (1 - \exp(-(x + \log(n))))^n \\ &= \lim_{n\to\infty} (1 - \exp(-x + \log(\frac{1}{n})))^n \end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \exp(\log(1 - \frac{\exp(-x)}{n})^n) \\
&= \lim_{n\to\infty} \exp(n.\log(1 - \frac{\exp(-x)}{n})) \\
&= \exp(\lim_{n\to\infty}(-\exp(-x))\frac{\log(1-\frac{\exp(-x)}{n})}{\frac{\exp(-x)}{n}}) \\
&= \exp((-\exp(-x))\lim_{n\to\infty}(\frac{\log(1-\frac{\exp(-x)}{n})}{\frac{\exp(-x)}{n}}) \\
&= \exp(-\exp(-x))
\end{aligned}
$$

donc $F_X(x) \in D(\Lambda(x))$.

Définition 2.6.2 *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*

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

Définition 2.6.3 *On dit que la fonction $H(x)$ est à variation réguliére d'ordre α ssi :*

$$\exists\ \alpha > 0 :\ H(x) = x^{\alpha}.L(x),$$

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

Proposition 2.6.1

$$\bar{F}_X(x) = 1 - F_X(x) = x^{-\alpha}.L(x), \Leftrightarrow\ F_X(x) \in D(\Phi_\gamma).$$

Théoréme de Bingham et al

Si $h \in [0,a] \to \mathbb{R}_+$ est mesurable , h est à variation réguliére d'ordre α ssi :

$$\lim_{t\to\infty} \frac{h(tx)}{h(t)} = x^\alpha \ , RV_\alpha.$$

Théorème 2.6.1 *(Mario 2000)*

$h \in [0,a] \to \mathbb{R}_+$ une fonction mesurable est à variation réguliére à droite de 0 ssi :

$$\lim_{t\to 0^+} \frac{h(tx)}{h(t)} = x^\alpha \ , \ \alpha > 0.$$

Remarque 2.6.1 *Si h est à variation réguliére à droite de 0 RV_α^0 alors :*

$g(x) = h(\frac{1}{x})$ est à variation réguliére à l'infini RV_α.

Remarque 2.6.2 *Les trois formes de distribution limit de maximum (Gumbel,Fréchet,weibulle) sont régroupées en une seule distribution dite forme de Jenkison-yon-Mizes(1954).*

Théorème 2.6.2 *(G.E.V.D)*

Si $X_1, ..., X_n$ une suite i.i.d de loi parente $F_X(x)$ et de statistique d'ordre associée $X_{1,n}, ..., X_{n,n}$ s'il existe deux suites $a_n > 0$,$b_n \in \mathbb{R}$ telles que :

$$\lim_{n\longrightarrow\infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = \lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = H_\gamma(x)$$

Alors :

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

et $H_{\mu,\sigma,\gamma}(x)$ s'appelle (GEVD) distribution généralisée des valeurs extrêmes.

avec μ : le paramètre de localisation, σ : le paramètre d'échelle, γ :l'indice de queue.

– Si la limite existe alors :

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

– a_n joue le rôle du paramétre d'échelle (T.C.L ($\frac{\sigma}{\sqrt{n}}$)).

– b_n joue le rôle du paramétre de position (T.C.L (μ)).

– γ l'indice de queue,il donne la forme de queue de distribution.

– Si la convergence est rapide (Distribution à queue légère).

– Si la convergence est lente (Distribution à queue lourde) .

1) Si $\gamma > 0$, $F_X(x)$ appartient au domaine d'attraction de Fréchet

$$F_X(x) \in D(\Phi_\gamma(x)) \{Pareto, Cauchy, Student, Log-gamma.\}.$$

Elle est surtout utilisée en actuariat (Assurrance ,Reassurrance).Cette distribution a une décroissance polynomiâle elle est non bornée et dont les moments peuvent ne pas existés.

2) Si $\gamma = 0$, $F_X(x)$ appartient au domaine d'attraction de Gumbel

$$F_X(x) \in D(\Lambda(x)) \{Exponentielle, Normale, Log-normale, Gamma.\}.$$

– Elle converge rapidement .

– Distribution à queue légère .

– Sa décroissance est exponentielle ,elle est non bornée,mais ses moments existent .

– Utilisée surtout en hydraulogie (étude des crues).

3) Si $\gamma < 0$, $F_X(x)$ appartient au domaine d'attraction de Weibul

$$F_X(x) \in D(\Psi_\gamma(x)) \{Beta, Uniforme, l'inverse\ de\ Paréto.\}.$$

– Sa queue est bornée .

– Utilisée surtout en fiabilité (étude des pannes).

Remarque 2.6.3 *Il ya des distributions qui n'appartiennent à aucun des trois domaines d'attraction.*

Exemple 2.6.2

$$F_X(x) = 1 - \frac{1}{\log(x)}, x \in \mathbb{R}_+^*.$$

$$F_X(x) = 1 - \frac{1}{\log(\log(x))}, x \in \mathbb{R}_+^*.$$

Proposition 2.6.2 *Les trois proposition suivants sont équivalentes :*

1. $\Phi_\gamma(x) = H_{\frac{1}{\gamma}}(\gamma(x-1)), \gamma > 0.$
2. $\Psi_\gamma(x) = H_{-\frac{1}{\gamma}}(\gamma(x+1)), \gamma < 0.$
3. $\Lambda(x) = H_0(x).$

Proposition 2.6.3 *Les proposition suivants sont équivalent*

1. $X \to \Phi_\gamma(x).$
2. $-\frac{1}{X} \to \Psi_\gamma(x).$
3. $\log(X^2) \to \Lambda(x).$

– Comment déterminer les constantes de normalisation $a_n > 0, b_n \in \mathbb{R}$?

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = H(x).$$

$$\lim_{n\longrightarrow\infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = H(x).$$

Le théorème suivant répond à cette question.

Théorème 2.6.3 *Selon la loi parente $F_X(x)$ on détermine les suit $a_n > 0$ et $b_n \in \mathbb{R}$ de la façon suivante :*

si $F_X(x) \in D(\Phi_\gamma(x))$ *alors* $b_n = 0,\ a_n = F^{-1}(1 - \frac{1}{n})$.
si$F_X(x) \in D(\Psi_\gamma(x))$ *alors* $b_n = F^{-1}(1),\ a_n = F^{-1}(1) - F^{-1}(1 - \frac{1}{n})$.
si $F_X(x) \in D(\Lambda(x))$ *alors* $b_n = F^{-1}(1 - \frac{1}{n}),\ a_n = F^{-1}(1 - \frac{1}{ne}) - F^{-1}(1 - \frac{1}{n})$.

Exemple 2.6.3 *Soit $X_1, ..., X_n$ i.i.d $\rightsquigarrow$ exponeille $\mathcal{E}(1)$*

$$F_X(x) = \begin{cases} 1 - \exp(-x) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$\exists\ b_n \in \mathbb{R}, a_n > 0$

On a : $F_X(x) \in D(\Lambda(x))$

donc $b_n = F^{-1}(1 - \frac{1}{n})$ et $a_n = F^{-1}(1 - \frac{1}{ne}) - F^{-1}(1 - \frac{1}{n})$.

$$y = F_X(x) = 1 - \exp(-x)$$
$$x = F_X^{-1}(y) = -\log(1 - y)$$
$$F_X^{-1}(1 - \frac{1}{n}) = -\log(1 - (1 - \frac{1}{n}))$$

$$-\log(\frac{1}{n}) = \log(n)$$
$$F_X^{-1}(1 - \frac{1}{n}) = \log(n) \in \mathbb{R}$$
$$a_n = -\log(1 - (1 - \frac{1}{ne})) - \log(n)$$
$$a_n = 1 + \log(n) - \log(n) = 1$$

alors $a_n = 1$ et $b_n = \log(n)$

$$\lim_{n\to\infty} F_{X_{n,n}}(x+\log(n)) = \lim_{n\to\infty}(F_X(x+\log(n))^n$$
$$\lim_{n\to\infty}(F_X(1-\exp-(x+\log(n)))^n = \lim_{n\to\infty}(1-\frac{\exp(-x)}{n})^n$$
$$= \exp(-\exp(-x))$$

Donc $\exists\ a_n > 0, b_n \in \mathbb{R}\ \ ,\ \ F_X(x) \in D(\Lambda(x))$.

Exemple 2.6.4 *Soit $X_1, ..., X_n$ i.i.d $X_i \rightsquigarrow \mathcal{U}_{[0\ 1]}$*

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \in [0\ 1] \\ 1 & \text{si } x > 0 \end{cases}$$

$\exists\ a_n > 0, b_n \in \mathbb{R}$,telle que :

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \begin{cases} \exp(-(-x)^{-\frac{1}{\gamma}}) & x \leq 0 \\ 1 & \text{sinon.} \end{cases}$$

$b_n = F^{-1}(1)$, $a_n = F^{-1}(1) - F^{-1}(1-\frac{1}{n})$.

$$y = F_X(x) = x$$
$$\Rightarrow x = F_X^{-1}(y) = y$$

$b_n = 1$

$$a_n = 1 - (1 - \frac{1}{n})$$
$$= \frac{1}{n}$$

$a_n = \frac{1}{n}$

$$\begin{aligned}\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) &= \lim_{n\to\infty} [F_X(\frac{x}{n}+1)]^n \\ \lim_{n\to\infty} (\frac{x}{n}+1)^n &= \exp(x) \\ &= \exp -(-x)^{-\frac{1}{(-1)}}\end{aligned}$$

Donc $F_X(x) \in D(\Psi_{-1}(x))$.

Exemple 2.6.5 *Soit* $X_1, ..., X_n$ *i.i.d* $X_i \rightsquigarrow \mathcal{C}_{[0\ 1]}$

$$F_X(x) = 1 - \frac{1}{\pi x}.$$

$\exists\ a_n > 0, b_n \in \mathbb{R}$,telle que :

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \begin{cases} \exp(-x^{-\frac{1}{\gamma}}) & x > 0 \\ 0 & \text{sinon.} \end{cases}$$

$b_n = 0$, $a_n = F_X^{-1}(1 - \frac{1}{n})$.

$$\begin{aligned} y &= F_X(x) = 1 - \frac{1}{\pi x} \\ x &= \frac{1}{\pi}\frac{1}{(1-y)} = F_X^{-1}(y)\end{aligned}$$

$$a_n = \frac{1}{\pi}\frac{1}{\frac{1}{n}} = \frac{n}{\pi} > 0$$

$a_n = \frac{n}{\pi}$

$$\lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \lim_{n\to\infty} [F_X(\frac{nx}{\pi})]^n = \lim_{n\to\infty} (1 - \frac{1}{\pi(\frac{nx}{\pi})})^n$$
$$= \lim_{n\to\infty} (1 - \frac{1}{nx})^n = \lim_{n\to\infty} (1 - \frac{x^{-1}}{n})^n = \exp(-x^{-1}).$$

donc $F_X(x) \in D(\Phi_1(x))$

Le point faible de la (T.V.E) est qu'on n étudé que la valeure extrême $X_{n,n}$,Il est plus intervessant d'étudier l'échantillon des plus grandes valeur.

$$(X_{n-k+1,n})_{1\leq k\leq p}$$
$$(X_{n-p+1,n}, ..., X_{n-1,n}, X_{n,n})$$

On-s'interesse donc à l'échantillon des valeur au-delà d'un seuil P.O.T

$$Y_1, ..., Y_{N_u} \qquad N_u < n$$

N_u :nombre des valeur P.O.T

$X_1, ..., X_n$ i.i.d , $F_X(x)$.

$Y_1, ..., Y_{N_u}$ i.i.d ,$F_u(y) = P(Y \leq y/X > u)$.

$$y = X - u$$

Distribution des valeur (P.O.T)

$$F_u(y) = P(Y \leq y / X > u) = \frac{P(Y \leq y, X > u)}{P(X > u)}$$
$$\frac{P(X - u \leq y, X > u)}{1 - P(X \leq u)} = \frac{P(X \leq y + u, X > u)}{\bar{F}_X(u)}$$
$$\frac{P(u < X \leq y + u)}{\bar{F}_X(u)} = \frac{F(y + u) - F_X(u)}{\bar{F}_X(u)}$$
$$F_u(y) = \frac{F(y + u) - F_X(u)}{\bar{F}_X(u)}.$$

$F_u(y)$: Distribution des excés.

2.7 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 excés au-delà de ce seuil (les excés sont les différences positives entre les observations et le seuil). Un résumé de ceci est dans la figure 2.1 suivante

Théorème 2.7.1 (Balkema-de Haan-Pickands) *Si $F_X(x) \in D(H_\gamma(x))$, alors il existe une fonction de répartition des excés au-delà de u noté F_u qui peut être uniformément approchée par une loi de Pareto généralisée (GPD) $G_{\gamma,\beta}(x)$ tel que :*

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

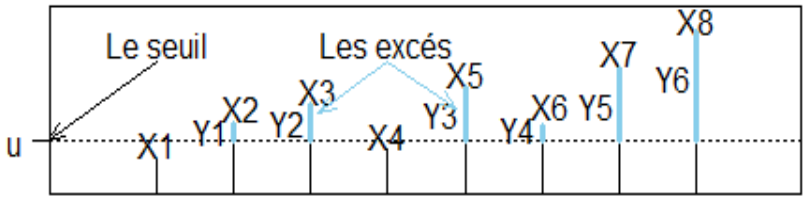


FIG. 2.1 – Les excés

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

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

Définition 2.7.1 (Distribution de Pareto généralisée) *Pour* $\beta > 0$, *et* $\gamma \in \mathbb{R}$, *la GPD est définie par :*

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

où, $x \geq 0$ *si* $\gamma \geq 0$ *et* $0 \leq x \leq \frac{\gamma}{\beta}$ *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.

Distribution de Pareto

On dit que X est une v.a de loi de pareto si sa fonction de répartition est $G_{\gamma,\beta}$ définée par :

$$G_{\gamma,\beta}(x) = 1 - (\frac{\gamma}{x})^{\beta}, x > 0.$$

On rajoute la condition $\beta \neq 1$ pour que le moment d'ordre 1 existe.

$$g_{\gamma,\beta}(x) = -\beta(-\frac{\gamma}{x^2}(\frac{\gamma}{x})^{\beta-1})$$
$$g_{\gamma,\beta}(x) = \frac{\beta\gamma}{x^2}(\frac{\gamma}{x})^{\beta-1}$$
$$g_{\gamma,\beta}(x) = \beta\gamma^{\beta}x^{-\beta-1}$$

$$E(X) = \int_{\gamma}^{+\infty} x.g_{\gamma,\beta}(x)dx$$

$$= \beta\gamma^{\beta}\int_{\gamma}^{+\infty} x^{-\beta}dx$$
$$= [\beta\gamma^{\beta}\frac{x^{-\beta+1}}{(-\beta+1)}]_{\gamma}^{+\infty}$$
$$E(X) = \frac{\gamma\beta}{\beta-1} \qquad \beta \neq 1$$

Exemple 2.7.1 *Montrer que si* $X_1, ..., X_n$ *,i.i.d* $\rightsquigarrow \mathcal{E}(1)$ *alors de distribution des excés est aussi une* $\mathcal{E}(1)$

$$F_X(x) = 1 - \exp(-x), x \geq 0 \ \in D(\Lambda(x))$$

$$\lim_{x \to w_F} \sup_x |F_u(x) - G_{\gamma,\beta}(x)| = 0,$$

alors :

$$F_u(x) = G_{0,\beta} = 1 - \exp(-\frac{x}{\beta}), x > 0 \ , \quad \beta = 1$$

$$\begin{aligned} F_u(x) &= \frac{F_X(x+u) - F_X(u)}{1 - F_X(u)} \\ &= \frac{1 - \exp(-x-u) - (1 - \exp(-u))}{1 - (1 - \exp(-u))} \\ &= \frac{-\exp(-x-u) + \exp(-u)}{\exp(-u)} \\ &= \frac{\exp(-u)(1 - \exp(-x))}{\exp(-u)} \\ &= (1 - \exp(-x)) \rightsquigarrow \mathcal{E}(1). \end{aligned}$$

Exemple 2.7.2 *Montrer que $G_{-1,\beta}$ est la distribution Uniforme $\mathcal{U}_{[0,\beta]}$*

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

$$G_{-1,\beta}(x) = 1 - (1 - \frac{x}{\beta}), \gamma = -1, 0 < x \leq \beta$$

$$= \begin{cases} 0 \ \text{ si } x \leq 0 \\ \frac{x}{\beta} \quad \text{si } x \in [0, \beta] \\ 1 \quad \text{si } x \geq \beta \end{cases}$$

Exemple 2.7.3 *Montrer que*

$$\lim_{\gamma \to 0} G_{\gamma,\beta}(x) = G_{0,\beta}(x)$$

On a :

$$\begin{aligned}\lim_{\gamma\to 0} 1-(1+\frac{\gamma}{\beta}x)^{-1/\gamma} &= 1-\exp(-\frac{x}{\beta}) \\ &= G_{0,\beta}(x).\end{aligned}$$

Chapitre 3

Estimateurs de l'indice de queue et application

Dans ce chapitre, on s'intéresse à l'estimation de l'indice de queue en théorie des valeurs extrêmes. Il existe différentes approches pour estimer ces quantités. Nous présentons par la suite deux approches differentes : l'une basée sur l'estimateur de Hill et l'autre par la méthode du maximum de vraisemblance. Et on va terminer ce chapitre par une application réelle sur les données de Covid-19.

3.1 Estimateur de Hill

Une grande partie de la théorie de l'estimation de l'indice de valeur extrême est développée pour des valeurs extrêmes des indices positifs. Le meilleur estimateur connu d'un indice des valeurs extrêmes positif est l'estimateur de Hill (Hill, 1975) défini de la façon suivante :

$$\hat{\gamma}_n^{(H)} = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}$$

avec $X_{1,n}, ..., X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$.

En premier lieu, si on note $E_{j,u}$ les excés relatifs au-dela de u i.e. $E_{j,u} = X_j/u$ avec $X_j > u$, on peut facilement vérifier que

$$P(E_{j,u} > x/E_{j,u} > 1) \to x^{-1/\gamma} \qquad \text{quand } u \to \infty,\ x > 1. \tag{3.1}$$

En formant la vraisemblance basée sur cette distribution limite, on vérifie facilement que l'estimateur de Hill n'est rien d'autre que l'estimateur du maximum de vraisemblance dans le cas où le seuil $u = X_{n-k,n}$ et en utilisant les statistiques d'ordre $X_{n-k+1,n}, ..., X_{n,n}$.

La fonction de log-vraisemblance sera alors :

$$L(\gamma; X_{n-k+1,n}, ..., X_{n,n}) = -k \log(\gamma u) + \log(1 - F(u)) - \frac{\gamma + 1}{\gamma} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log u.$$

En maximisant la fonction log-vraisemblance par rapport à γ,on obtient l'estimateur de Hill pour $\gamma > 0$.

En second lieu, un côté trés attrayant de l'estimateur de Hill est qu'il est facile à interpréter graphiquement. Ceci est particuliérement important pour les partieiens, qui préférent souvent des interprétations graphiques à des formules mathématiques. Plus précisément, si on utilise le graphe "Pareto quantile plot" dans le cas de distributions de type Pareto, ce graphe sera approximativement linéaire, dans les points extrêmes, avec une pente γ.

3.1.1 Comportement de l'estimateur de Hill

(a) Les propriétés asymptotiques de l'estimateur de Hill ont été établies par Mason (1982) qui a prouvé la consistance faible de l'estimateur de Hill $\gamma_n^{(H)}$ pour

toute suite k_n vérifiant :

$$k = k_n \to \infty \text{ et } k_n/n \to 0 \text{ quand } n \to \infty$$

(b) La consistance forte a été prouvée dans Deheuvels, Haeusler et Mason (1985) sous les conditions suivantes :

$$k/\log\log n \to \infty \text{ et } k_n/n \to 0 \text{ quand } n \to \infty$$

(c) Sous certaines conditions du second ordre, la normalité asymptotique de l'estimateur de Hill a été démontréé entre autre par Hall (1982), Davis et Resnick (1984), Haeusler et Teugels (1985), Goldie et Smith (1987) et Dekkers et al. (1989), à savoir

$$\sqrt{k}(\gamma_n^{(H)} - \gamma) \rightsquigarrow \mathcal{N}(0, \gamma^2).$$

On peut associer à l'estimateur de Hill un intervalle de confiance asymptotique $I_N(\alpha)$ de niveau α.

$$I_N(\alpha) = [\hat{\gamma} - z_{\alpha/2}\hat{\gamma}\frac{1}{\sqrt{k}}, \hat{\gamma} + z_{\alpha/2}\hat{\gamma}\frac{1}{\sqrt{k}}],$$

où $z_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

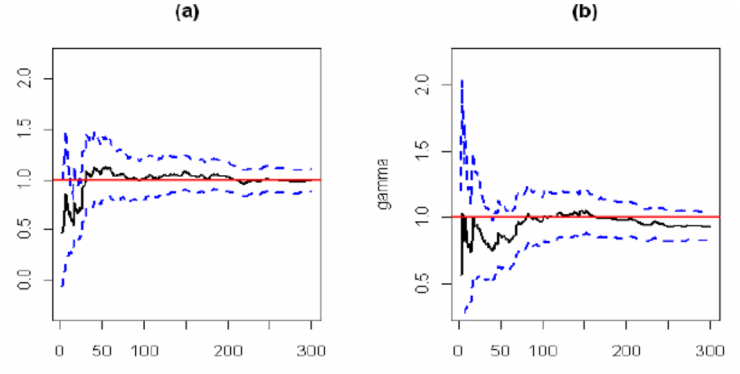


FIG. 3.1 – Estimateur de Hill, en fonction du nombre des extrêmes (en trait plein) avec l'intervalle de confiance 95%, pour (a) la distribution de Pareto standard et pour (b) la distribution de Fréchet(1) basées sur 100 échantillons de 3000 observations.

3.1.2 Classe de Hall de la fonction de distribution

Hall (1982) a considéré les fonctions de distribution F qui satisfont :

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

pour $\gamma > 0$, $\rho \leq 0$, $c > 0$, $d \in \mathbb{R} \setminus \{0\}$. Il a prouvé la normalité asymptotique pour l'estimateur de Hill.

Cette sous-classe des distributions à queue lourde contient les distributions Pareto, Burr, Fréchet et t-student.

3.2 Consistance de l'estimateur de Hill

3.2.1 Convergence faible

Théorème 3.2.1 *(Mason 1982) Soit $X_1, X_2, ..., X_n$ des variables aléatoires i.i.d de fonction de répartition commune F, avec $F \in D(H_\gamma), \gamma > 0$ et k_n suite d'entier tq $1 < k_n \leq n$*

$$\text{si}\begin{cases} k_n \to \infty \\ \quad et \qquad \text{alors } H_{k,n} \underrightarrow{P} \gamma, \quad \text{lorsque } n \to \infty \\ \frac{k_n}{n} \to 0 \end{cases}$$

3.2.2 Convergence forte

Théorème 3.2.2 *(Deheuvels,Hausler et Mason 1988) Soient $X_1, X_2, ..., X_n$ des variables aléatoirs i.i.d de fonction de répartition commune F,avec $F \in D(H_\gamma)$.*

$$\text{si}\begin{cases} k_n \to \infty \\ \quad et \qquad \text{alors } H_{k,n} \underrightarrow{p.s} \gamma, \quad \text{lorsque } n \to \infty. \\ \frac{k_n}{\log(\log(n))} \to \infty \end{cases}$$

3.3 Méthode du Maximum de vraisemblance

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

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

avec

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

On obtient l'estimateur de θ en résolvant le systéme suivant :

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

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

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

On dérive $l(X_1, ..., X_n; \theta)$ par rapport aux paramètres μ, σ et γ, et on obtient le systéme suivant :

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

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

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

avec le systéme 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 systéme est non linéaire pour lesquels aucune solution analytique n'existe. Par conséquence on utilise les méthodes numérique pour trouver la solution du systéme d'équations ainsi obtenu. Telle que la méthode de Monte Carlo.

3.4 Application

Nous allons modéliser les données Covid-19 en utilisant la méthode des blocs maxima, cette méthode dépend de la division des données en bloc ensuite nous prenons le maximum de chaque bloc puis nous les modélisons en utilisant la loi de Fréchet, mais avant cela, il faut voir s'il ya des valeurs extrêmes, pour cela nous utiliserons une ou deux instruction grâce auxquelles nous saurons que les données contiennent des valeurs extrêmes ce qui indique que la queue des données est lourde.

Au cours de ce travail nous réaliserons les étapes suivantes :

1 Calculerons le kurtosis et verrons plot d'excés moyennes.

2 Estimer le paramètre γ par la méthode de Hill et maximum de vraisemblance.

3 La modélisation de la queue avec la loi de Fréchet.

3.4.1 Les packages utilisés

Parmi les packages que nous avons utilisés dans cette application est :

• "readxl" : ce package est utilisé pour l'instruction "read_excel".

• "evir" : ce package est utilisé pour l'instruction"$meplot$".

• "moments" : ce package est utilisé pour l'instruction"$kurtosis$".

• "DescTools" : ce package est utilisé pour l'instruction"$Gini$".

• "evd" : ce package est utilisé pour l'instruction"$dfrechet$".

Au cas où le package est n'existe pas dans la portefeuille de packages vous téléchargez comme suit :

install.packages("nom de package "), par exemple pour télécharger le package evir entrer install.packages("evir").

• Aprés le téléchargement , nous utilisons l'instruction Library pour lire le package, puis utilisons les instructions qu'il contient.

• Nous lisons toutes les données Covid-19 à l'aide de la directive "read_excel" comme indiqué dans l'encadré suivant :

$$\gg alle = read\_excel("C : /Users/Pc/Downloads/jhudata\_2020\_06 - 30.xls")$$

3.4.2 Le coefficient de gini avec langage R

• Parmi les instructions qui nous montrent si les données sont à queue lourde se trouve l'instruction "$Gini$", qui calcule le coefficient de gini et l'écrit par la formule

suivante :

$$\gg giniest = Gini(alle\$cumdeaths, conf.level = 0.95)$$

• Nous affichons la valeur de ce paramètre :

$$\gg giniest$$

• Le résultat obtenu :0,9078399, puisque la valeur de coefficient de gini est proche de 1,cela indique que la queue est lourd.

• On séléctionner les données supérieures à 200 pour le calcul du kurtosis et le plot d'excées moyenne :

$$\gg jhudata = alle[which(alle\$cumdeaths > 200),]$$

$$\gg est = jhudata\$cumdeaths$$

3.4.3 Le kurtosis avec langage R

$$\gg K = kurtosis(est)$$

• Le resultats obtenue avec R est K=49 >3, ce qui signifie que la queue est lourde.

• Nous avons ordonés l'échantillon en utilisant "*order*" :

$$\gg data = est[order(est)]$$

• Nous mettons les données dans un tableau :

$$\gg table(data)$$

• On utilise l'instruction "meplot" pour représenté le plot d'excés moyenne et on trouve :

$$\gg me = meplot(data = data, omit = 3)$$

3.4.4 Résultats et discussion de plot d'excés moyenne

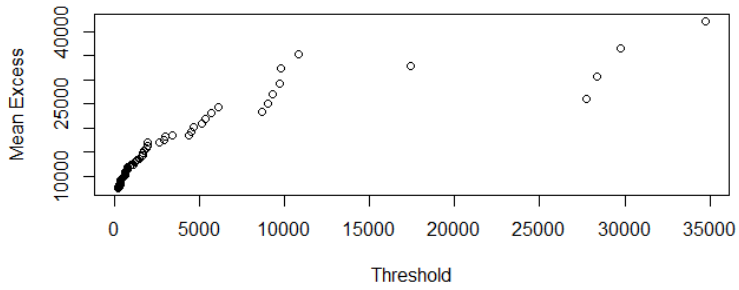


FIG. 3.2 – Plot d'excés moyenne

– Le graphique démontre une augmentation significative de la moyenne des excés du seuil 0 jusqu'à environ 12000, suivie d'une diminution progressive jusqu'à près de 30000. Malgré cette baisse, la moyenne des excés reste élevée.
– Les résultats indiquent également une importante disparité entre les pays, avec certains affichant des taux de mortalité bien plus élevés que d'autres. Parmi les pays les plus touchés par la pandémie figurent les États-Unis, le Brésil et le Royaume-Uni, avec des taux de mortalité cumulés élevés.
– On peut affirmer que la distribution des données présente une queue lourde.

3.4.5 Estimateur de Hill avec le langage R

• Pour calculer l'estimateur de Hill on détermine l'échantillon X d'une taille N.

$\gg X = alle\$cumdeaths$

$\gg N = length(X)$

• On choisit le nombre des valeurs extrêmes k comme le nombre de valeurs supérieures à 500 :

$$\gg k = length(X[which(X > 500)])$$

$$\gg Y = sort(X)$$

$$\gg t = N - k + 1$$

$$\gg Z = log(Y[t : N])$$

$$\gg H = sum(Z)/k - log(Y[N - k])$$

• Déclaration d'estimation de Hill :

$$\gg H$$

• Puisque la valeur que nous obtenons γ =1.903775, est supérieure à 0 donc on peut dire que la loi des valeurs extrêmes est la loi de Fréchet.

3.4.6 Méthode des blocs maxima

Nous savons que la Méthode des blocs maxima dépend de la division des données en blocs, afin que nous divisions les données en 33 bloc. Puis nous prenons les maximum de chaque bloc.

• Cette explication est présentée dans la figure suivante :

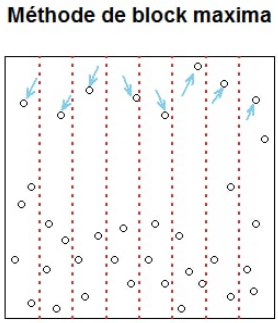


FIG. 3.3 – Méthode de block maxima

$$\gg maxima = tapply(X, gl(length(X)/5, 5), max)$$

• Nous avons arrangé les maximums :

$$\gg maax = sort(maxima)$$

• Convertir les données en valeurs numériques :

$$\gg x = as.numeric(maax)$$

3.4.7 Estimation des paramètres par la méthode MV.

• Certaines instructions permettent de trouver des estimations de paramètres à l'aide de la méthode MV. Parmi ces instructions se trouve l'instruction"$fevd$".

$$\gg es\_MV = fevd(x, method = "MLE")$$

• Puis les résultats des estimations obtenues selon cette méthode :

$$\gg es\_MV$$

estimateur de μ	estimateur de σ	estimateur de γ
175.348087	519.210645	2.978077

• Enfin, nous modélisons les maximum (les valeurs extrêmes) en utilisant la loi de Fréchet.

$$\gg plot(x, xlim = c(0, 15000), dfrechet(x, 175.348087, 519.210645, 2.978077), type = "l")$$

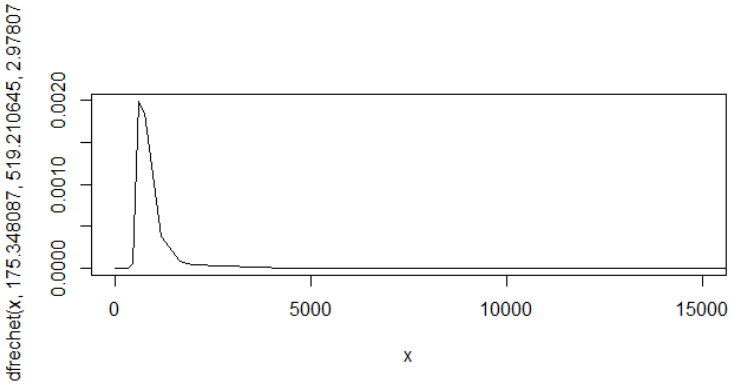


FIG. 3.4 – Les maximum

Conclusion

Nous avons dans ce mémoire essayé de présenter d'une façon très simple et accessible l'estimation de distribution.valeurs des extrêmes, vu lours importance et leurs utilisations dans beaucoups de domaines sensibles les finances, l'actuariat , l'assurance, l'hydrologie...etc.

Dans ce mémoire, nous avons étudié deux méthodes pour estimer les distribution des queues : l'estimateur de Hill et la méthode MV. De plus, nous avons réalisés une application sur des données réelles.

On peut dire qu'il n'existe pas de méthodes universelles pour estimer les queues de distribution dans toutes les situations. Chaque méthode présente ses avantages et ses limites, donc il est préférable d'utiliser plusieurs méthodes pour obtenir une estimation à la fois robuste et précise. Ces résultats peuvent servir de base à des décisions éclairées en matière de gestion des risques. Par exemple, dans le domaine de l'assurance, estimer la probabilité d'événements extrêmes permet de fixer les primes. De plus, les erreurs d'estimation peuvent avoir des conséquences graves, surtout en cas d'effets sévères d'événements rares.

Enfin, il faut mentionner qui une nouvelle branche et prometteuse de l'analyse des valeurs extrêmes est celle de méthodes des valeurs extrêmes multivariées.

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] Bateka, S.,(2010). Determination du Noumbre de Statistiques D'ordre Extrêmes. Mémoire de magister d'université de Mohammad khider, Biskra, Algeria.

[4] Benatia, F., (2023). Cours de Proucessus Empirique et Statistique d'ordre. Université de Mohammad khider, Biskra, Algeria.

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

[6] De Haan L., Ferreira A. (2006). Extreme Value Theory : An Introduction Springer Verlag.

[7] Lachecheb, S.,(2019). Théorie des valeurs extrêmes et Applications. Mémoire de master d'université de Ghardaia, Algeria.

[8] Medjder, M.,Moussaoui, S.,(2019). Estimation de l'indice des valeurs extrêmes-Application en hydrologie-Mémoire de master d'université de Mohammad Seddik Ben Yahia, Jijel, Algeria.

[9] Zernadji, L., (2023). Estimation des queues de distributions. Mémoire de master d'université de Mohammad khider, Biskra, Algeria.

Annexe A : Logiciel R

3.5 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.
$Var\,(.)$	:	Variance.
$Cov(.,.)$	:	Covariance.
$v.a$	:	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.
$\xrightarrow{p.s}$	:	Convergence presque sur.
$\xrightarrow{\mathcal{D}}$	:	Convergence en distribution.
GPD	:	Distribution Généralisée de Pareto.
MV	:	Maximum de Vraisemblance.
POT	:	Peak Over Threshold.
$\xrightarrow{\mathcal{P}}$	:	Convergence en probabilité
$GVED$	:	Distribution Généralisée des Valeurs Extrêmes

RÉSUMÉ

Dans ce travail, nous avons présentés une étude des maxima de distribution,dans leurs deux présentation (GVED) et (GPD) avec l'estimation de leurs distributions respectives et qui sont la méthode des maxima (au bien loi max-stable) et la méthode P.O.T (Peaks over Threshold) ou méthode des excés au dela d'un seuil. Nous avons aussi consacré une partie à l'estimateur trés repondu qui est l'estimateur de Hill pour l'estimation de l'indice de queue des distribution. Nous terminons notre travail par une application aux données réelles(Covid-19), en utilisant le langage R.

Mots-clés : maxima de distribution,excés,l'estimation de l'indice de queue.

ملخص

قدمنا في هذا العمل دراسة الحد الأقصى للتوزيع في عرضيهما(GVED) و (GPD) مع تقدير التوزيعات الخاصة بكل منهما وهما طريقة الحد الأقصى (قانون الحد الأقصى المستقر) وطريقة P.O.T(القمم فوق العتبة) أو طريقة التجاوزات التي تتجاوز العتبة. لقد خصصنا أيضًا قسمًا للمقدر الشائع جدًا وهو مقدر Hill لتقدير مؤشر ذيل التوزيعات. لقد أنهينا عملنا بتطبيق على البيانات الحقيقية (Covid-19)، وذلك باستخدام لغة R.

كلمات مفتاحية : اقصى توزيع ,افراط ,تقدير مؤشر الذيل .

ABSTRACT

In this work, we presented the study of the distribution maxima, in their two presentations (GEVD) and (GPD) with the estimation of their respective distributions and which are the method of maxima (instead max-stable law) and the P.O.T method (Peaks over Threshold) or method of excesses beyond a threshold. We also devoted a section to the very common estimator which is the Hill estimator for estimating the tail index of the distributions. We finish our work with an application to real data(Covid-19), using the R language.

KEYWORDS: distribution maxima, excesses, estimating the tail index.