الجمهورية الجزائرية الديمقراطية الشعبية

République Algerienne Démocratique et populaire

Ministère de l'ensignement supérieur et de la recherche scientifique

Université Saad Dahleb de Blida

Faculté des sciences

Département de Mathématiques

Mémoire de fin d'études en vue de l'obtention du diplome

de Master en Mathématiques

Option :

Modélisation Stochastique et Statistique

Thème

Estimation des paramètres de la loi de Pareto généralissée par la méthode de Bootstrap. Application aux données climatiques

Présenté par :

- Yahia Bouaroudj
- Mohammed Hendi

Devant le jury :

Président	Omar Tami	M.A.A	USDB
Examinateur	Abdelaziz Rassoul	M.C.A	ENSH
Promoteur	Redhouane Frihi	M.A.A	USDB

Année Universaire : 2018/2019

Remerciement

Premièrement nous remercions DIEU le tout puissant qui nous a donné la force et le courage pour mener à bien ce travail .
On désire adresser, un merci tout particulier , à notre promoteur Mr. Redhouane Frihi, pour son pertinent et judicieux conseil tout au long de notre travail .
Nous remercions chaleureusement tous les enseignants du département des mathématiques de l'université Saad Dahleb, Blida.
Enfin, nous présentons toutes notre gratitude à tous ceux ou celles qui ont contribué de près ou de loin à la réalisation de ce mémoire .

Dédicace

Je dédie ce modeste travail

A

Mon très cher père et ma très cher mère qui représentent pour moi par excellence le symbole de la bonté, la source de tendresse. Ce travail est le fruit de vos sacrifices et vos efforts fournis pour mon éducation et ma formation, vos prières et votre bénédiction m'ont été un grand soutien.

A

mon chère frère Sofiane que j'aime beaucoup .

Mes très chères sœurs : Salwa, Titiche,Fetta, Chafiaa et sa fille Tiziri, Djazira et son mari Moussa et leurs enfants :Rania, Meriem, Oussama et mon adorable Abdouch, Malgré la distance, vous etes toujours dans mon cœur,je vous remercie pour votre hospitalité sans égal et votre afection si sincère .

A

Mes amis(e)s : hassen, abd allah, Hamza, Houria et houssem qui n'ont jamais arrêté de m'encourager durant mon parcoure.

A

Mon binôme Mohammed Hendi et toute sa famille .

A vous et à ceux que j'aurais très injustement oublié ... merci .

Yahia

Dédicace

Je dédie ce mémoire de fin d'étude
A ma chère mère
Quoi que je fasse ou que je dise je ne saurai point te remercier comme il se doit. Ton affection me couvre, ta bienveillance me guide et ta présence a mes cotés a toujours été ma source de force pour effectuer les différents obstacle .
A mon chère père
Tu as toujours été a mes cotés pour me soutenir et m'encourager
Que ce travail traduit ma gratitude et mon affection
A mon chère petit frère
fouad
Mes très chères sœurs : Khoula et lamia A mes ami(e)s : houssem, abd allah, hamza et sofian merci de m'avoir soutenue et couragé toute au long de cette période.
Mon binôme :Yahia Bouaroudj que je ne cesserai jamais de la considère parmi l'un de mes frère, et à qui je souhaite une bonne continuation pour son avenir professionnel.
Pour finir, aucune dédicace pourrai exprimer mes vifs remerciements, le dévouement et le respect que j'ai toujours pour vous tous ce qui m'ont aidé de près ou de loin por réaliser ce travail.

Mohammed

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Pour un profane, la statistique est associée à la notion de moyenne ou l'écart-type. En effet dans de nombreuses applications, notamment en sciences sociales et en sciences de la physique, les statistiques se résument généralement au calcul des moyennes et à l'évaluation de dispersion d'une série de valeurs autour de leur moyenne.

Par définition, les événements rares sont des événements ayant une faible probabilité d'apparition. Lorsque le comportement de ces événements est dù au hasard on peut étudier leur loi. Ils sont dite extrêmes quand il s'agit de valeurs beaucoup plus élevées que les autres valeurs ou plus faibles que celles observées habituellement.

Les événements extrêmes peuvent être catastrophiques lorsque il s'agit (tremblements de terre , inondations, accidents nucléaires,...) dominent l'actualité quotidienne par leur caractère imprévisible compte tenu de l'importance des enjeux sociaux et scientifiques. Aucun débat sérieux sur le hasard ne serait être mené sans une réflexion sur les événements rares et extrêmes.
"La loi des grands nombres et la distribution gaussienne, fondements de l'étude statistique des grandeurs moyennes, échouent à rendre compte des événements rares ou extrêmes. Pour ce faire, des outils statistiques plus adaptés existent ... mais ne sont pas toujours utilisés !"(Rama Cont-Papers-Pour La Science -Décembre 2009).

Dés lors, la question que l'on pourrait se poser est de savoir ce que peuvent les outils statistiques face aux événements extrêmes ? Autrement dit, peut-on réellement prévoir ou quantifier le risque des événements extrêmes ?

La théorie des valeurs extrêmes (TVE) fournit une base mathématique et probabiliste rigoureuse sur laquelle il est possible de construire des modèles statistiques pour prévoir la taille et la fréquence de ces phénomènes dans les queues de distribution.

Le comportement extrême des lois appartenant aux différents domaines d'attraction maximal (DAM) est significativement différent. Les lois appartenant aux DAM de Weibull sont bornées à droite, celles appartenant aux DAM de Gumbel et Fréchet ont un support infini à droite. Mais les premières ont des queues finies alors les secondes ont des queues épaisses et exposent donc à des situations extrêmes beaucoup plus dangereuses.

Les domaines d'application sont en effet trés variés : hydrologie, météorologie, biologie, ingénierie, gestion de l'environnement, finance, assurance, sciences sociales,...etc.
Les catastrophes naturelles sont des exemples d'évènements extrêmes qui conduisent à des pertes financières importantes. Les cracks boursiers sont d'autres exemples qui conduisent à des pertes financières très importantes.

Il existe essentiellement deux approches dans la modélisation des extrêmes : la méthode des blocs de maxima et celle de dépassement des excès. L'approche des maxima uni-variés établit qu'une famille paramétrique généralisée résume le comportement asymptotique de la loi du maximum convenablement normalisé. Dans la seconde approche, il est établi que seule la distribution généralisée de Pareto qui modélise la loi de la variable excédentaire au sel d'un certain seuil fixé et assez élevé.

La modélisation la loi GPD passe par l'estimation de ces paramètres a savoir le paramètre d'échelle et l'indice de queue. Ces derniers peuvent être éstimer par différentes méthode. Dans notre mémoire nous exposons une méthode qui prend de l'ampleur a savoir la méthode de Bootstrap

Ce mémoire s'organise en deux parties. La première partie s'intitule « Théorie des valeurs extrêmes » dont l'objectif est d'exposer la théorie probabiliste des valeurs extrêmes dans le cas uni varié. Elle se compose de trois chapitres :

• Au cours de premier chapitre,on expose la théorie probabiliste des valeurs extrêmes dans le cas uni varie. D'abord, on a commencé par le théorème fondamental de la théorie des valeurs extrêmes (théorème de Fisher-Tippet) qui assure que la loi limite maximum est surement une des trois lois (Gumbel- Weibull-Fréchet). Puis on a unifié les trois lois dans une seule représentation qui est la représentation de Von-Mise (GEV) et qui sert à estimer l'indice de queue. Ensuite, on a défini les domaines d'attraction et les conditions pour que chaque distribution appartienne au max-domaine d'attraction associé.

• Dans le deuxième chapitre,Dans ce chapitre, nous définissons une autre approche des valeurs extrêmes qui est basée sur la distribution de Pareto généralisée , On constate que le deuxième théorème fondamental des V.E (théorème de Balkema- Pickands-de Haan) est considéré comme le deuxième théorème fondamental des valeurs extrêmes. Puis, on a estimé les paramètres de cette distribution mais après la détermination du seuil ,et nous estimons les paramètre par différente méthode .

• Dans le troisième chapitre on a définie la méthode de bootstrap et son algorithme de l'appliquer .

La deuxième partie est une partie de simulations et d'applications aux données climatiques de la région de Dellys . Tout d'abord nous avons fait une simulation d'estimation de GPD par maximum vraisemblance ensuite par Bootstrap , la même chose avec l'application aux données réel , ensuite par une comparaison on conclure celle de la méthode de Bootstrap est performance .

Première partie :

Partie théorique

Chapitre 1

Distribution des valeurs extrêmes généralisées

Motivation : La distribution des valeurs extrêmes généralisée est très utile en application de la théorie des valeurs extrêmes, car c'est la seule et unique loi de probabilité qui modélise le comportement du maximum d'un échantillon.

1.1 Introduction

La théorie des valeurs extrêmes a le but d'étudier la loi du maximum et du minimum d'une suite des variables aléatoires, cette théorie s'énonce de la façon suivante :
soit X_i une suite de variables aléatoires indépendantes et identiquement distribuées (iid) de fonction de répartition F qui est définie par :

$$F(X) = \Pr(X \leq x).$$

L'étude des extrêmes passe par l'analyse du maximum d'un échantillon de taille n (n assez grand), ordonné, noté $M_n = Max(X_1, X_2, ..., X_n)$. Cette analyse est appelée analyse des maxima par bloc. Grâce a la caractérisation (iid) on aura :

$$\begin{aligned} \Pr(M_n \leq x) &= \Pr(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &= \Pr(X_1 \leq x) \Pr(X_2 \leq x) ... \Pr(X_n \leq x) \\ &= \prod_{i=1}^{n} \Pr(X_i \leq x) \\ &= [F(x)]^n. \end{aligned} \tag{1.1}$$

Généralement, F est inconnue et la relation (1.1) n'est pas utilisable directement ; la théorie des valeurs extrêmes donne le comportement asymptotique de la variable M_n A partir de l'idée du théorème centrale limite (TCL), la théorie des valeurs extrêmes montre que s'il existe des suites normalisatrices $(a_n)_{n>0}$ et $(b_n)_{n>0}$ sont des suites de normalisation avec $a_n > 0$ et $b_n \in \mathbb{R}$ telle que :

$$\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F(a_n x + b_n)^n = G(x). \tag{1.2}$$

Où $G(x)$ est une variable aléatoire non dégénérer.
Remarque : $Min(x_1, x_2 ..., x_n) = -Max(-x_1, -x_2, ..., -x_n)$

Question : Quelle est la loi de $G(x)$?

Réponse : Le théorème suivant répond à cette question.

1.2 Théorème Fisher-Tippet(1928)

Théorème 1.1. *[9] Sous certaines conditions de régularité sur la fonction de répartition F, s'il existe un paramètre réel α et deux suites $(a_n)_{n>0}$ et $(b_n)_{n>0}$ tels que $a_n > 0$ et $b_n \in \mathbb{R}$ et pour tout $x \in \mathbb{R}$*

$$\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F(a_n x + b_n)^n = G_\alpha(x).$$

Alors $G_\alpha(x)$ est l'une des trois distributions suivantes :

***Gumbel** :*
pour tout $x \in \mathbb{R}$

$$\Lambda(x) = \exp(-\exp(-x))$$

***Fréchet** :*
pour tout $\alpha > 0$

$$\Phi_\alpha(x) = \begin{cases} \exp(-x^{-\alpha}) & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$$

***Weibull** :*
pour tout $\alpha > 0$

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x^{\alpha})) & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$$

$G_\alpha(x)$ est appelée fonction de répartition de la loi des valeurs extrêmes.
Pour une démonstration de ce théorème, on pourra référer aux ouvrages :Hann and Frreira ou Resnick (voir [1]ou [2])

Commentaires sur le théorème :

1. On a une seule limite dans le TCL (la loi normale) par contre dans le cas des extrêmes trois limites sont possibles (*Gumbel, Fréchet et Weibull*).
2. $(a_n)_{n\geq 1}$ est un paramètre de position,il représente $\mathbb{E}(x)$ dans le TCL.
3. $(b_n)_{n\geq 1}$ est le paramètre d'échelle, représente $\dfrac{\sigma(x)}{\sqrt{n}}$.

Le paramètre $\alpha > 0$ qui apparaisse dans les types de distribution de **Fréchet** et de **Weibull** est appelé **paramètre de forme** (indice de queue).
Plus α est petit, plus la décroissance est lente, plus des valeurs extrêmes sont susceptibles de ce produire.
Plus α est grand,plus la décroissance est forte, moins des valeurs extrêmes vont se présenter.

1.2.1 Propriété de queue des lois extrêmes Gumbel, Fréchet et Weibull

A. Distribution de Gumbel

Proposition 1.1.

$\bar{\Lambda}(x) = 1 - \Lambda(x) = 1 - \exp(-e^{-x}) \sim e^{-x}$ *quand* $x \longrightarrow \infty$. *c-à-d : pour les grandes valeurs, la queue de la distribution décroit très rapidement (de façon exponentielle).*
La queue de ***Gumbel*** *est une fonction à variation rapide à l'infini d'indice* $-\infty$. *Donc c'est une lois à queue* ***légère****.*

B. Distribution de Fréchet

Proposition 1.2.

$\forall \alpha > 0$, $\bar{\Phi}_\alpha(x) = 1 - \Phi_\alpha(x) = 1 - \exp(-x^{-\alpha}) \sim x^{-\alpha}$ *quand* $x \longrightarrow \infty$. *Pour les grandes valeurs, la queue de cette distribution décroit d'une façon polynomiale .*
La queue de Φ_α *est une fonction à variation régulière à l'infini d'indice* $-\alpha$. *Donc on déduit que la queue de Fréchet est* ***lourde*** *à droite.*

C. Distribution de Weibull

La queue de la distribution de **Weibull** est finie, c-à-d c'est une distribution bornée à droite, donc elle a peu d'intérêt dans l'étude des événements extrêmes

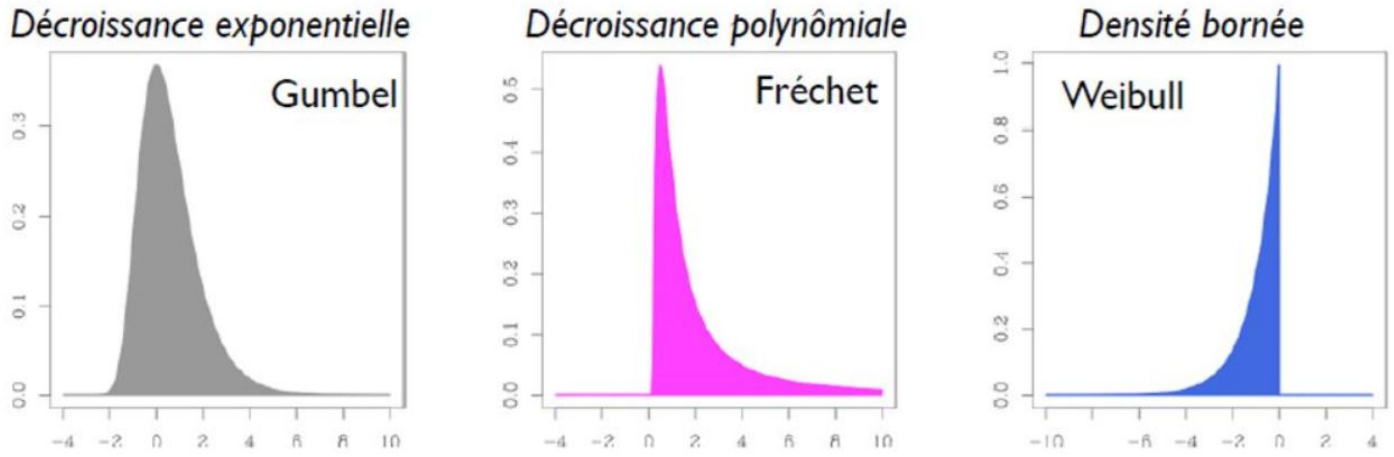

FIGURE 1.1 : *Queues des trois lois extrêmes*

Relation entre les trois lois

Proposition 1.3.

La relation entre les trois distributions des valeurs extrêmes est donnée comme suit :
$X \sim$ ***Fréchet*** $\Longrightarrow Y = \ln X \sim$ ***Gumbel****.*
$X \sim$ ***Weibull*** $\Longrightarrow Y = \dfrac{1}{X} \sim$ ***Fréchet****.*
$X \sim$ ***Weibull*** $\Longrightarrow Y = \ln\left(\dfrac{1}{X}\right) \sim$ ***Gumbel****.*

Grâce au travaux de ***Von Mises (1936)*** *et* ***Jenkinson(1955)*** *on a une forme unifiée de la fonction de répartition de la loi* ***EVD*** *à un facteur d'échelle et de position.*

1.2.2 Représentation de Jenkinson-Von Mises

Proposition 1.4. *[10] La représentation de* ***Jenkinson-Von Mises*** *de la distribution des valeurs extrêmes****EVD*** *que l'on appelle loi des valeurs extrêmes généralisée notée* ***GEV*** *à pour fonction de répartition :*

$$G_\xi(x) = \begin{cases} \exp\left(-\left(1+\xi x\right)\right)^{-1/\xi} & si\ \xi \neq 0, 1+\xi x > 0 \\ \exp\left(-\exp(-x)\right) & si\ \xi = 0, x \in \mathbb{R} \end{cases} \tag{1.3}$$

Pour obtenir une forme plus générale de la loi ***GEV*** *en introduisent les paramètres de localisation μ et d'échelle β on obtiendra :*

$$G_{\xi,\mu,\beta}(x) = \begin{cases} \exp\left(-\left(1+\xi\left(\frac{x-\mu}{\beta}\right)^{-1/\xi}\right)\right) & si\ \xi \neq 0, 1+\xi(\frac{x-\mu}{\beta}) > 0 \\ \exp\left(-\exp-\left(\frac{x-\mu}{\beta}\right)\right) & si\ \xi = 0 \end{cases} \tag{1.4}$$

Remarque 1.1.

À partir de cette écriture, on peut distinguer 3 cas :
$\xi = 0$ *correspond à la loi de* ***Gumbel****.*
$\xi > 0$ *correspond à la loi de* ***Fréchet****.*
$\xi < 0$ *correspond à la loi de* ***Weibull****.*

1. Quelle est la relation entre F et G ?
2. Comment on peut choisir les coefficients de normalisation a_n et b_n ?

Réponse : La réponse de ces deux questions est dans la partie suivante.

1.3 Domaines d'attraction et coefficients de normalisation

[10] La classe des fonctions de distribution F satisfaisant (1.2) appelé max-domaine d'attraction.
La recherche du domaine d'attraction peut être considérée comme l'étude réciproque de la recherche de la distribution des valeurs extrêmes associée à une distribution,étant donné G ,quelles sont les conditions nécessaires et/ou suffisante sur la variable aléatoire X pour que la limite $\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F(a_n x + b_n)^n = G(x)$ soit réalisée ?

Définition 1.1.

On dit qu'une distribution F appartient au max-domaine d'attraction d'une distribution des valeurs extrêmes G et on note $F \in MDA(G)$, si la distribution du maximum normalisé converge vers G i.e si F est la distribution commune de v.a X_1,...,X_n (iid) de maximum M_n, alors il

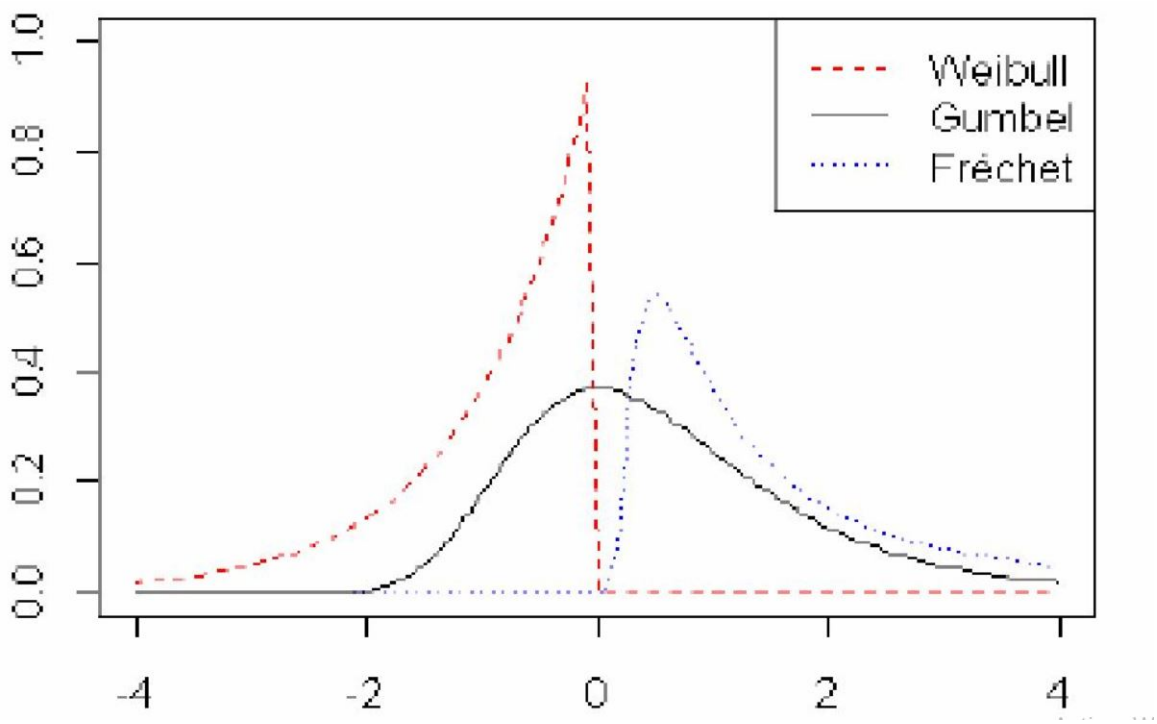

FIGURE 1.2 : *Dénsité de la loi GEV*

existe des constantes $a_n > 0$,$b_n \in \mathbb{R}$ *telles que :*
$\forall x \in \mathbb{R}$

$$\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = G.$$

Exemple 1.1.

La loi exponentielle du paramètre 1

Soit X *suit la loi* $\exp(1)$ *;* $F(x) = 1 - e^{-x}$.
On pose $a_n = 1$, $b_n = \ln n$

$$\begin{aligned}
\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \lim_{n\to\infty} \Pr\left(\frac{M_n - \ln n}{1} \leq x\right) \\
&= \lim_{n\to\infty} F\big(x + \ln n\big)^n \\
&= \lim_{n\to\infty} \left(1 - e^{-x+\ln n}\right)^n \\
&= \lim_{n\to\infty} \left(1 - \frac{e^{-x}}{n}\right)^n \\
&= e^{-e^{-x}} \\
&= \Lambda(x)
\end{aligned}$$

car $\lim\limits_{n\to\infty} \big(1 - \frac{x}{n}\big)^n = e^{-x}$.

Alors la distribution exponentielle appartient au max-domaine d'attraction de ***Gumbel****.*

Exemple 1.2.

On suppose que X suit la loi uniforme $\mathcal{U}([0,1])$
c-à-d :
$\forall x \in \mathbb{R}$

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ x & si\ 0 \leq x \leq 1 \\ 1 & si\ x > 1 \end{cases}$$

On pose $a_n = \frac{1}{n}$ *et* $b_n = 1$
alors : $\forall x \in \mathbb{R}$, $\forall n \in \mathbb{N}^*$

$$\begin{aligned} \Pr\left(\frac{M_n - b_n}{x} \leq x\right) &= F^n\big(1 + \frac{x}{n}\big) \\ &= \begin{cases} 0 & si\ 1 + \frac{x}{n} < 0 \\ \left(1 + \frac{x}{n}\right)^n & si\ 0 \leq 1 + \frac{x}{n} \leq 1 \\ 1 & si\ 1 + \frac{x}{n} > 1 \end{cases} \\ &= \begin{cases} 0 & si\ x < -n \\ \left(1 + \frac{x}{n}\right)^n & si\ -n \leq x \leq 0 \\ 1 & si\ x > 0 \end{cases} \end{aligned}$$

Donc : $\forall x \in \mathbb{R}$

$$\begin{aligned} \lim_{n \to \infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \begin{cases} e^x & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases} \\ &= \Psi_1(x) \end{aligned}$$

Alors la distribution uniforme appartient au max-domaine d'attraction de **Weibull**. $\star$ Les définitions précédentes concernant le domaine d'attraction basent sur le choix des facteurs de normalisation a_n et b_n. Mais souvent le choix n'est pas facile ; on propose d'autre théorèmes et des propositions sur le domaine d'attraction.

1.3.1 Domaine d'attraction maxima

Théorème 1.2.

Une condition nécessaire et suffisante pour q'une fonction F appartienne au domaine d'attraction maximal de $G_\alpha(x)$ *est :*

$$\lim_{\epsilon \to 0} \frac{F^{-1}(1-\epsilon) - F^{-1}(1-2\epsilon)}{F^{-1}(1-2\epsilon) - F^{-1}(1-4\epsilon)} = 2^c. \tag{1.5}$$

Cela implique que

1. *Si* $c < 0$ *alors* $F(x) \in MDA(\Psi_\alpha(x))$.
2. *Si* $c = 0$ *alors* $F(x) \in MDA(\Lambda_\alpha(x))$.
3. *Si* $c > 0$ *alors* $F(x) \in MDA(\Phi_\alpha(x))$.

Application du théorème

1. La loi uniforme
   On a $F^{-1}(\alpha) = \alpha$

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{F^{-1}(1-\epsilon)-F^{-1}(1-2\epsilon)}{F^{-1}(1-2\epsilon)-F^{-1}(1-4\epsilon)} &= \lim_{\epsilon\to 0}\frac{(1-\epsilon)-(1-2\epsilon)}{(1-2\epsilon)-(1-4\epsilon)}\\ &= \frac{1}{2}\\ &= 2^{-1}.\end{aligned}$$

   Donc $c = -1 < 0 \Longrightarrow$ la loi uniforme $\in MDA(\Psi_\alpha(x))$.

2. La loi de Cauchy
   On a $F^{-1}(p) = tg\big((p-\frac{1}{2})\Pi\big)$

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{F^{-1}\big(1-\epsilon\big)-F^{-1}\big(1-2\epsilon\big)}{F^{-1}\big(1-2\epsilon\big)-F^{-1}\big(1-4\epsilon\big)} &= \lim_{\epsilon\to 0}\frac{tg\Big(\big(\frac{1}{2}-\epsilon\big)\Pi\Big)-tg\Big(\big(\frac{1}{2}-2\epsilon\big)\Pi\Big)}{tg\Big(\big(\frac{1}{2}-2\epsilon\big)\Pi\Big)-tg\Big(\big(\frac{1}{2}-4\epsilon\big)\Pi\Big)}\\ &= 2^1.\end{aligned}$$

   Donc $c= 1 > 0 \Longrightarrow F(x) \in MDA(\Phi_\alpha(x))$.

1.3.2 Domaine d'attraction de Fréchet, Weibull et Gumbel

A. Domaine d'attraction de Fréchet

Théorème 1.3.

La fonction de répartition F appartient au domaine d'attraction de la distribution de ***Fréchet*** *si et seulement si :*

$$\bar{F}(x) = x^{-\alpha}L(x)$$

où la fonction L est à variation lente.
En particulier $x_F = +\infty$; de plus si $F(x) \in DA\ \Phi_\alpha(x)$ avec a_n=$U(n)$=$F^{-1}(1-\frac{1}{n})$ alors :
La suite $(a_n^{-1}M_n)_{n>0}$ converge en loi vers une variable aléatoire de fonction de répartition ϕ.

Demonstration On suppose que

$$\bar{F}(x) = x^{-\alpha}L(x)$$

et L est à *variation lente* $\Longrightarrow$ on peut l'écrire sous la forme :

$$L(x) = c(x)\exp\int_a^x \frac{k(u)}{u}du$$

$\Longrightarrow$

$$\bar{F}(x) \approx x^{-\alpha}c\exp\int_a^x \frac{k(u)}{u}du$$

On pose a_n=$U(n)$

$$\bar{F}(x) = U(n)^\alpha c\exp\int_a^{U_n}\frac{k(u)}{u}du = (1-\frac{1}{n})c\exp\int_a^{1-\frac{1}{n}}\frac{k(u)}{u}du = (1-\frac{1}{n})c\exp$$

$$\bar{F}(x) \leq \frac{1}{n} \leq \bar{F}(\bar{a}_n)$$

Si F est continue en a_n,alors $\bar{F}\ (a_n) = \frac{1}{n}$
sinon,Comme $\bar{F}$ est équivalente en $+\infty$ à une fonction continue, on déduit que

$$\lim_{n\to\infty} \bar{F}(a_n) = \frac{1}{n}$$

Pour $x > 0$ on a donc :

$$\lim_{n\to\infty} \bar{F}(a_n x) = \lim_{n\to\infty} \frac{\bar{F}(a_n x)}{\bar{F}(a_n)} = x^{-\alpha}$$

Proposition 1.5.

Si $F(x) \in MDA\ (\Phi_\alpha(x))$ alors les constantes de normalisations $a_n > 0$ et $b_n \in \mathbb{R}$ telles que :

$$\forall x \in \mathbb{R}\,;\ \lim_{n\to\infty} F(a_n x + b_n)^n = \Phi_\alpha(x)$$

peuvent être choisies de la manière suivante :

$$a_n = U(n) \quad et \quad b_n = 0$$

où U est la fonction quantile de queue de la variable aléatoire X.

Exemple 1.3.

Dans ce tableau on présente quelques exemples de lois de probabilité qui appartient au max-domaine d'attraction de Fréchet

Distribution de X	*Coefficients de normalisation*
$X \sim Cauchy(0,1)$	$a_n = n/\pi$ *et* $b_n = 0$
$X \sim Pareto(\alpha)$	$a_n = n^{1/\alpha}$ *et* $b_n = 0$
$X \sim log - Gamma(\alpha,\beta)(\beta > 0)$	$a_n = ((\Gamma(\beta))^{-1}(\ln n)^{\beta-1}n)^{1/\alpha}$ *et* $b_n = 0$

B.Domaine d'attraction de Weibull

Théorème 1.4.

La fonction de répartition F appartient au domaine d'attraction de la distribution de ***Weibull*** *si et seulement si :*

$x_F < +\infty$ *et* $\bar{F}\big(x_F - \frac{1}{x}\big) = x^{-\alpha}L(x)$ *où la fonction L est à variation lente.*
De plus ; si $F(x) \in DA\ (\Psi_\alpha(x))$ avec $a_n = x_F - U(n) = x_F - F^{-1}(1 - \frac{1}{n})$ alors :
La suite $(a_n^{-1}M_n)_{n>0}$ converge en loi vers une variable aléatoire de fonction de répartition ψ.

Proposition 1.6.

Si $F(x) \in (MDA\ \Psi_\alpha(x))$ alors les constantes de normalisations $a_n > 0$ et $b_n \in \mathbb{R}$ telles que :

$$\forall x \in \mathbb{R}\,;\ \lim_{n\to\infty} F(a_n x + b_n)^n = \Psi_\alpha(x)$$

peuvent être choisies de la manière suivante :

$$a_n = x_F - U(n) \quad et \quad b_n = x_F$$

où U est la fonction quantile de queue de la variable aléatoire X et x_F est le point terminal.

Exemple 1.4.

Dans ce tableau on présente quelques exemples de lois de probabilité qui appartient au max-domaine d'attraction de Weibull

Distribution de X	*Coefficients de normalisation*
$X \sim \mathcal{U}[0;1]$	$a_n = 1/n$ *et* $b_n = 1$
$X \sim \beta(\alpha, \theta), (\theta > 0)$	$a_n = \left(\dfrac{\Gamma(\alpha+\theta)}{\Gamma(\alpha+1)\Gamma(\theta)} n\right)^{-1/\alpha}$ *et* $b_n = 1$

C. Domaine d'attraction de Gumbel

Le domaine d'attraction de **Gumbel** est le plus délicat car il est difficile à énoncer du fait qu'il n'existe pas une condition nécessaire et suffisante relativement simple.

Théorème 1.5.

La fonction de répartition F de la variable aléatoire X avec le point terminal $x_F < +\infty$ appartient au domaine d'attraction de la distribution de ***Gumbel*** *si et seulement s'il existe $a < x_F$ tels que :*

$$\bar{F}(x) = 1 - F(x) = c(x) \exp\left(- \int_a^x \frac{\theta(u)}{\delta(u)} du \right)$$

où c et θ sont des fonctions mesurables sur $]a; x_F[$ vérifiant :

$$\lim_{x \to x_F} c(x) = c > 0 \ \text{ et } \lim_{x \to x_F} \theta(x) = 1$$

et δ est une fonction absolument continue sur $]a; x_F[$ de densité δ' telle que :

$$\forall x \in]a, x_F[\ : \delta > 0 \ \text{ et} \lim_{x \to x_F} \delta'(x) = 0.$$

Proposition 1.7.

Si $F(x) \in MDA\ (\Lambda_\alpha(x))$ alors :
les constantes de normalisations $a_n > 0$ et $b_n \in \mathbb{R}$ telles que :
$\forall x \in \mathbb{R}$

$$\lim_{n \to \infty} F(a_n x + b_n)^n = \Lambda_\alpha(x)$$

peuvent être choisies de la manière suivante :

$$a_n = \delta(b_n) \quad et \quad b_n = U(n)$$

où U est la fonction quantile de queue de la variable aléatoire X et δ est la fonction définie dans le théorème précédent

Proposition 1.8. *Si la fonction de répartition F de la variable aléatoire X est une fonction de* ***Von-Mises****, alors F appartient au max-domaine d'attraction de* ***Gumbel****.*

Exemple 1.5.

La distribution exponentielle du paramètre $\lambda > 0$, i.e :

$$H(x) = \begin{cases} 1 - e^{-\lambda x} & si\ x > 0 \\ 0 & si\ x \leq 0 \end{cases}, \lambda > 0$$

*H est une fonction de **Von-Mises**.*
en effet : on a $x_H = +\infty$ et
$\forall x \in]0, +\infty[$

$$\bar{H}(x) = 1 - H(x) = e^{-\lambda x} = \exp\left(- \int_0^x \frac{1}{\lambda^{-1}} du \right)$$

*On déduit que H est une fonction de **Von-Mises** de fonction auxiliaire $\delta(x) = \lambda^{-1}$.*
Alors $H \in MDA\ (\Lambda)$

Exemple 1.6.

Dans ce tableau on présente quelques exemples de lois de probabilité qui appartient au max-domaine d'attraction de Gumbel.

Distribution de X	*coefficients de normalisation*
$X \sim Rayleigh$	$a_n = a\sqrt{\log n}$ *et* $b_n = a/2\sqrt{\log n}$
$X \sim \mathcal{N}(0,1)$	$a_n = (2\ln n)^{-1/2}$ *et* $b_n = \sqrt{2\ln n} - \dfrac{\ln(4\pi) + \ln\ln n}{2\sqrt{2\ln n}}$
$X \sim \Gamma(\alpha, \beta)$	$a_n = \beta^{-1}$ *et* $b_n = \dfrac{\ln n + (\alpha - 1)\ln\ln n - \ln\Gamma(\alpha)}{\beta}$

Les conditions nécessaires et suffisantes pour q'une fonction de répartition F appartienne à un domaine d'attraction d'une distribution extrême sont pas facile à vérifier.
La partie suivante on va présenter une condition suffisante **condition de Von-Mises**, elle est simple à vérifier mais dans le cas où les fonctions de répartitions ayant une densité car cette condition est basée sur *la fonction de Hasard.*

Condition de Von-Mises(1936)

Théorème 1.6. *[8]*

Soit F une fonction de répartition absolument continue de densité f et soit la fonction de hasard

$$h(x) = \frac{f(x)}{1 - F(x)}$$

1. *Si $h(x) > 0$, x assez grand et s'il existe $\alpha \leq 0$ telle que :*

   $\lim\limits_{x \to \infty} xh(x) = \alpha$ *alors :*

   $$F(x) \in MDA(\Phi_\alpha(x))$$

2. *Si $F^{-1}(1) < 1$ et s'il existe $\alpha > 0$ telles que /*

   $\lim\limits_{x \to F^{-1}(1)} (F^{-1}(1) - x)h(x) = \alpha$ *alors :*

   $$F(x) \in MDA(\psi_\alpha(x))$$

3. Si $h(x) \neq 0$ et h dérivable au voisinage de F^{-1} (où bien $F^{-1} = \infty$) et si de plus :

$\lim_{x \to F^{-1}(1)} \frac{d}{dx} \frac{1}{h(x)} = 0$ *alors :*

$$F(x) \in MDA(\Lambda(x))$$

1.3.3 Domaine d'attraction de GEV

Théorème 1.7.

La fonction de distribution de F de la variable aléatoire X appartient au max-domaine d'attraction de la distribution des valeurs extrêmes généralisés ***GEV*** G_ξ *si et seulement si : il existe une fonction mesurable δ telle que :*

$$\lim_{t \to x_{\bar{F}}} \frac{\bar{F}(t + x\delta(t))}{\bar{F}(t)} = \begin{cases} (1+\xi x)^{-1/\xi} & si\ \xi \neq 0 \\ e^{-x} & si\ \xi = 0 \end{cases} \tag{1.6}$$

avec $1 + \xi x > 0$ et $x \in \mathbb{R}$.

Proposition 1.9.

$F \in DAG_\xi$ si et seulement si :

$$\lim_{n \to \infty} \bar{F}(a_n x + b_n) = -\log G_\xi(x).$$

Pour une certaine suite $(a_n, b_n)_{n>0}$ où $a_n > 0$ et $b_n \in \mathbb{R}$; on a alors ***la convergence en loi*** *de $a_n^{-1}(M_n - b_n)_n$ vers une variable aléatoire de la fonction de répartition G_ξ.*

Démonstration 1.3.1.

$\Longrightarrow$ *On suppose que $F \in DAG_\xi$ alors :*

$$\lim_{n \to \infty} (1 - \bar{F}(a_n x + b_n))^n = G_\xi(x) \tag{1.7}$$

En prend le log *de (1.7)*

$$\lim_{n \to \infty} n \log(1 - \bar{F}(a_n x + b_n)) = \log G_\xi(x)$$

On a : $\forall (1 + \xi x) > 0$

$$\lim_{n \to \infty} \bar{F}(a_n x + b_n)) = 0$$

$\Longrightarrow$

$$\lim_{n \to \infty} n\bar{F}(a_n x + b_n) = -\log G_\xi(x)$$

$\Longleftarrow$ *Si on a :*

$$\lim_{n \to \infty} \bar{F}(a_n x + b_n) = -\log G_\xi(x)$$

alors : $F \in DA(G_\xi)$(évident).

Exemple 1.7.

Soit X ***v.a*** *suit la loi de Pareto standard*

$$F(x) = 1 - x^{-\theta} \text{ avec } \theta \in \mathbb{R}_+^*$$

Ce qui implique que :

$$1 - F(x) = \begin{cases} x^{-\theta} & si\ x > 1 \\ 1 & si\ x \geq 0 \end{cases}$$

On pose : $a_n = n^{1/\theta}$ *;* $b_n = 0$
alors :

$$n[1 - F(a_n x + b_n)] = n[1 - F(n^{\frac{1}{\theta}} x)] = \begin{cases} x^{-\theta} & si\ x > n^{-1/\theta} \\ n & si\ x < n^{-1/\theta} \end{cases}$$

passant à la limite

$$\begin{aligned} \lim_{n \to \infty} n[1 - F(a_n x + b_n)] &= \begin{cases} x^{-\theta} & si\ x > 0 \\ \infty & si\ x \leq 0 \end{cases} \\ &= \begin{cases} -\ln \exp(-x^{\theta}) & si\ x > 0 \\ -\ln \theta & si\ x \leq 0 \end{cases} \\ &= -\ln \Phi_{\alpha}(x) \end{aligned}$$

Donc $F \in MDA\ (\Phi_{\alpha}(x))$.

1.3.4 Résultats obtenus

Considérons les 2 lois suivantes :

1. La loi x-exponentielle
   $\forall \lambda, \alpha > 0$
   $$F(x) = (1 - (1 - \lambda x)e^{-\lambda x})^{\alpha}.$$
   ,
2. La loi exponentielle généralisée
   $\forall \alpha > 0, \forall \lambda > 0$
   $$G(x) = (1 - e^{-\lambda x})^{\alpha}.$$

La loi x-exponentielle

On a

$$F(x) = (1 - (1 - \lambda x)e^{-\lambda x})^{\alpha}.$$

Pour trouver le domaine d'attraction maxima, on va utilisé le Théorème (1.2)D'abord :

$$x = F^{-1}(y) = -\frac{1}{\lambda}(1 + ((y^{1/\alpha})e^{-1}))$$

où est la fonction de Lambert.

$$\lim_{\epsilon \to 0} \frac{G^{-1}(1-\epsilon) - G^{-1}(1-2\epsilon)}{G^{-1}(1-2\epsilon) - G^{-1}(1-4\epsilon)} = 1/2 = 2^{-1}$$

$$2^c = 1/2 \Rightarrow c = -1.$$

Donc la distribution maximum de la loi exponentielle généralisée appartient au max domaine d'attraction de Weibull ($c < 0$).
On peut calculer les coefficients de normalisation à l'aide de la proposition(1.6)

$$a_n = -\frac{1}{\lambda}\left(1 - \left(1 + \big(\frac{n-1}{n}\big)^{1/\alpha} - 1\big)e^{-1}\right)$$
$$b_n = -\frac{1}{\lambda}$$

La loi exponentielle généralisée

On a

$$F(x) = ((1 - e^{-\lambda x})^{\alpha}).$$

Pour trouver le domaine d'attraction maxima, on va utilisé le théorème (1.2) D'abord :

$$x = -\frac{1}{\lambda}\log(1 - F^{1/\alpha}).$$

On applique le théorème (1.2)

$$\lim_{\epsilon \to 0} \frac{F^{-1}(1-\epsilon) - F^{-1}(1-2\epsilon)}{F^{-1}(1-2\epsilon) - F^{-1}(1-4\epsilon)} = 1.$$

$$2^c = 1 \Rightarrow c = 0.$$

Donc la distribution maximum de la loi exponentielle généralisée appartient au max domaine d'attraction de Gumbel (c=0).
On peut calculer les coefficients de normalisation à l'aide de la proposition(1.7)

$$a_n = -\frac{1}{\lambda}\log\big(1 - \big(\frac{n-1}{n}\big)^{1/\alpha}\big)$$
$$b_n = \frac{1}{\lambda}\big(\log\big(1 - (\frac{n-1}{n})^{1/\alpha}\big)\big) - \log\big(1 - \big(\frac{e^{-1}}{n}\big)^{1/\alpha}\big)$$

Remarque 1.2.

La fonction de Lambert, c'est la fonction inverse de la fonction f définie par $f(x) = xe^x$.
Quelques propriétés de la fonction :
$y = xe^x \Rightarrow x = (y)$.
$(xe^x) = x$.

Remarque importante

La loi de Bernoulli de paramètre $p \in]0, 1[$ n'admet pas une convergence du maximum normalisé.
En effet :
Soit X_i une variable aléatoire de loi de Bernoulli de paramètre $p \in]0, 1[$ alors :

$$\Pr(X_i = 1) = p = 1 - \Pr(X_i = 0).$$

On pose

$$T = \inf\{k \geq 1; X_k = 1\}.$$

Si $n > T$ on aura $M_n = 1$. La loi de T est une loi géométrique de paramètre p. Donc T est finis $p.s$ et M_n est constante égal à 1 $p.s$ alors, à partir d'un certain rang il n'existe pas de suite $(a_n, b_n)_{n\geq 1}$, avec $a_n > 0$ telle que la suite $a_n^{-1}(M_n - b_n)$ converge en loi vers une limite non dégénéré.
Donc il n'existe pas une limite non dégenerée pour la limite de maximum normalisé de la loi de Bernoulli.
De façon analogue, on peut démontrer aussi la même chose pour la loi géométrique et la loi de poisson.

1.4 Conclusion

Dans ce chapitre, nous avons donner quelque fondement de la théorie probabiliste des valeurs extrêmes. Au début, on a présenté le théorème fondamental de la théorie des valeurs extrêmes (théorème de Fisher-Tippet) et les différents domaines d'attraction ainsi que le coefficient de normalisation avec des résultats obtenus dans ce cadre.

Chapitre 2

Distribution de Pareto généralisée

Motivation : L'approche basée sur la GEV a été critiqué dans la mesure où l'utilisation d'un seul maxima conduit à une perte d'information contenue dans les autres grandes valeurs de l'échantillon.
Pour pallier ce problème,la méthode POT (Peaks-Over-Threshold) où méthode des excès au-delà d'un seuil élevé a été introduit par Pickands(1975),basé sur la distribution de pareto généralisées GPD.

2.1 Introduction

La méthode des excès au-delà d'un seuil repose sur le comportement des données qui dépassent un seuil donnée.Autrement dit,elle consiste à étudier le comportement non pas du maximum des données qu'on a en main,mais toutes les données qui dépassent un seuil élevé u.Plus précisément,les différences entre ces données et le seuil u appelées *excès*.

Définition 2.1.

On appelle ***excès*** *de la variable aléatoire X au-delà d'un seuil $u < x_F$ la variable aléatoire Y qui prend ses valeurs sur $]0, x_F - u[$ définie par :*

$$Y = X - u | X > u \tag{2.1}$$

avec $u < x_F$.

Définition 2.2.

On appelle ***distribution des excès*** *de la variable aléatoire X par rapport à un seuil $u < x_F$ la loi de probabilité de la variable alétoire Y excès de X au-delà du seuil $u < x_F$ donnée par sa fonction de distribution de répartition F_u,qu'on appelle fonction de distribution des excès suivante :*

$\forall y \in \mathbb{R}$

$$F_u(y) = \begin{cases} 0 & y \leq 0 \\ 1 - \dfrac{1 - F(u+y)}{1 - F(u)} & 0 < y < x_F - u \\ 1 & y \geq x_F - u \end{cases} \tag{2.2}$$

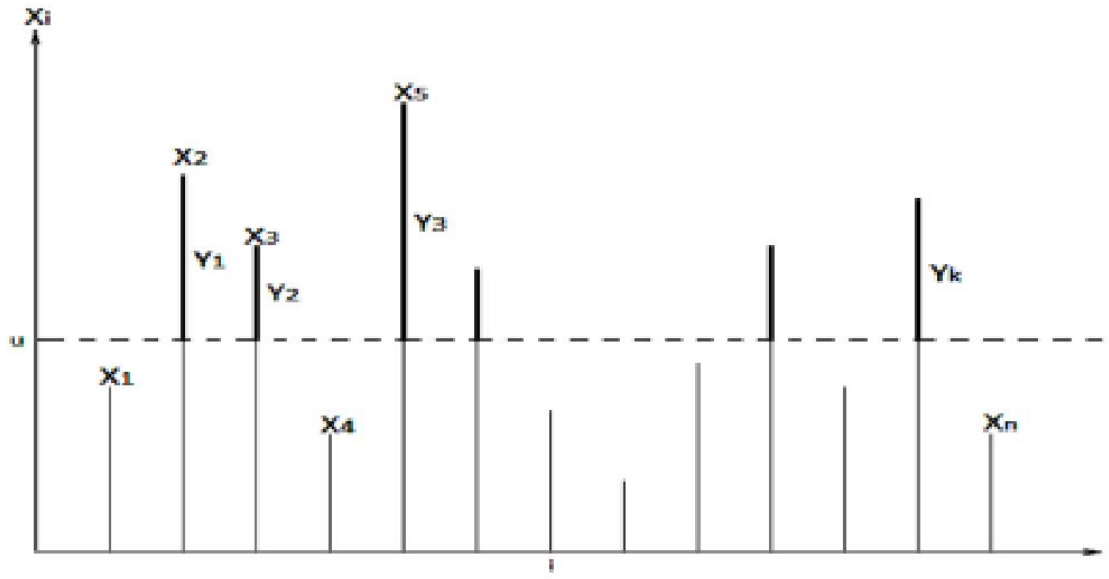

FIGURE 2.1 : La loi des excès .

2.1.1 Distribution de Pareto généralisé(GPD)

La distribution de **Pareto généralisé** joue un rôle très important à la modélisation des excès.

Définition 2.3.

*Soit $\xi \in \mathbb{R}$,On appelle **distribution de Pareto généralisé standard** toute fonction de répartition H_ξ où toute loi de probabilité qui a H_ξ comme fonction de répartition telle que :* $\forall x \in \mathbb{R},\ 1+\xi x > 0$

$$H_\xi(x) = \begin{cases} 1-(1+\xi x)^{-1/\xi} & \xi \neq 0 \\ 1-e^{-x} & \xi = 0 \end{cases} \tag{2.3}$$

Définition 2.4.

*Une distribution $H_{\xi,\beta}$ est dite de **Pareto généralisée** de paramètre $\xi \in \mathbb{R}$ et $\beta > 0$ si :*

$$H_{\xi,\beta}(x) = \begin{cases} 1-\left(1+\frac{\xi}{\beta}x\right)^{-1/\xi} & \xi \neq 0 \\ 1-\exp\left(-\frac{x}{\beta}\right) & \xi = 0 \end{cases} \tag{2.4}$$

Cette distribution est définie pour :

$$\begin{cases} x \geq 0 & \xi \geq 0 \\ 0 < x \leq -\beta/\xi & \xi < 0 \end{cases}$$

β représente le paramètre d'échelle et ξ le paramètre de forme.

2.1.2 Particularités de la GPD

1. Si $\xi > 0$, la distribution $H_{\xi,\beta}$ est la loi de Pareto usuelle avec $\alpha = \frac{1}{\xi}$ et $K = \frac{\beta}{\xi}$.
   En effet :

$$\begin{aligned} H_{\xi,\beta}(x) &= 1-\left(1+\frac{\xi}{\beta}x\right)^{-1/\xi} \\ &= 1-\left(\frac{1}{1+\frac{\xi}{\beta}x}\right)^{1/\xi} \\ &= 1-\left(\frac{\beta}{\beta+\xi x}\right)^{1/\xi} \\ &= 1-\left(\frac{\frac{\beta}{\xi}}{\frac{\beta}{\xi}+y}\right)^{1/\xi}. \end{aligned}$$

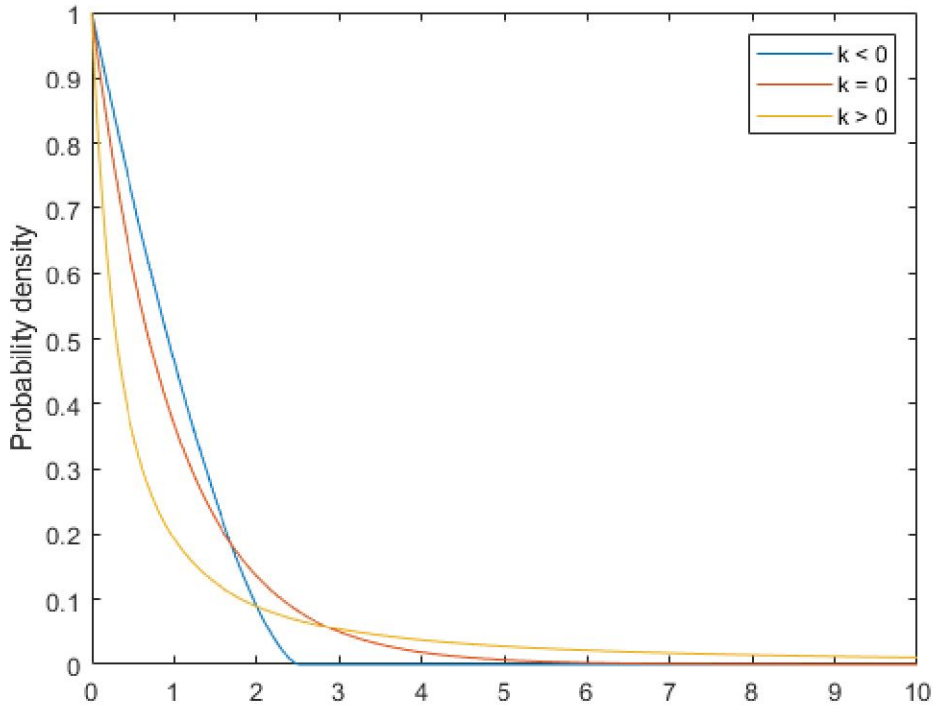

FIGURE 2.2 : Densité de pareto

2. Si $\xi = 0$, la distribution $H_{0,\beta}(x)$ est une distribution exponentielle d'espérance β.

$$\begin{aligned}
\lim_{\xi\to 0} H_{\xi,\beta}(x) &= \lim_{\xi\to 0} 1 - \left(1 + \frac{\xi}{\beta}x\right)^{-1/\xi} \\
&= 1 - \exp\left(- \lim_{\xi\to 0}\left(\frac{\frac{x}{\beta}}{1+\frac{\xi x}{\beta}}\right)\right) \\
&= 1 - \lim_{\xi\to 0}\exp\left(-\frac{1}{\xi}\right)\log\left(1+\frac{\xi}{\beta}x\right) \\
&= 1 - \exp\left(- \lim_{\xi\to 0}\left(\frac{\log\left(1+\frac{\xi}{\beta}x\right)}{\xi}\right)\right) \\
&= 1 - \exp\left(\frac{-x}{\beta}\right) \\
&= H_{0,\beta}(x).
\end{aligned}$$

3. $H_{\xi,\beta}(x) \in MDA(G_\xi)\,;\, \forall \xi \in \mathbb{R}$.
   En effet :
   Soit $\xi > 0$, on sait que si $F \in MDA(G_\xi)$ alors : $\bar{F}(x) = x^{-1/\xi}L(x)$ où $L(x)$ est une fonction à variation régulière (*théorème de Karamata*).

On montre que : $\bar{H}_{\xi,\beta}(x) = x^{-1/\xi}L(x)$.

$$\begin{aligned}
\bar{H}_{\xi,\beta}(x) &= 1 - H_{\xi,\beta}(x) \\
&= 1 - \left(1 - \left(1 + \frac{\xi}{\beta}x\right)^{-1/\xi}\right) \\
&= \left(1 + \frac{\xi}{\beta}x\right)^{-1/\xi} \\
&= x^{-1/\xi}\left(\frac{x^{-1}\beta + \xi}{\beta}\right)^{-1/\xi} \\
&= x^{-1/\xi}\left(\frac{\xi}{\beta} + \frac{1}{x}\right) \\
&= x^{-1/\xi}L(x).
\end{aligned}$$

Où : $L(x) = \dfrac{\xi}{\beta} + \dfrac{1}{x}$.
On montre que $L(x)$ est à variation régulière.
L est à variation régulière $\Longleftrightarrow \lim\limits_{x\to\infty} \dfrac{L(tx)}{L(x)} = 1$.

$$\begin{aligned}
\lim_{x\to\propto} \frac{L(tx)}{L(x)} &= \lim_{x\to\propto} \frac{\left(\dfrac{(tx)^{-1}\beta + \xi}{\beta}\right)^{-1/\xi}}{\left(\dfrac{x^{-1}\beta + \xi}{\beta}\right)^{-1/\xi}} \\
&= \lim_{x\to\propto} \left(\frac{\dfrac{(tx)^{-1}\beta + \xi}{\beta}}{\dfrac{x^{-1}\beta + \xi}{\beta}}\right)^{-1/\xi} \\
&= \lim_{x\to\propto} \left(\frac{(tx)^{-1}\beta + \xi}{x^{-1}\beta + \xi}\right)^{-1/\xi} \\
&= 1.
\end{aligned}$$

Donc $L(x)$ est à variation régulière ; alors $H_{\xi,\beta}(x) \in MDA(G_\xi)$.

4. Si $\xi = -1$, elle correspond à une loi uniforme sur $[0, \beta]$.

5. Si $\xi > 0$, on retrouve la loi de Pareto décentrée.

Question : Quel est le lien entre la loi de Pareto généralisée et la distribution des excès ?

Réponse : Le théorème suivant fait le lien entre le comportement asymptotique de la distribution des excès et la loi de Pareto généralisée.

2.2 Théorème de Pickands-Balkema-de Haan

Théorème 2.1. *[3] et [4] La fonction de distribution F de la variable aléatoire X appartient au max-domaine d'attraction de la distribution des valeurs extrêmes généralisée standard G_ξ ($\xi \in \mathbb{R}$) si et seulement s'il existe une fonction strictement positive β telle que la fonction de distribution des excès F_u de X par rapport au seuil $u < x_F$ converge uniformément vers une distribution de Pareto généralisée $H_{\xi,\beta(u)}$ lorsque u tend vers x_F.*

$$F \in MDA(G_\xi) \Longleftrightarrow \lim_{u \to x_F} \sup_{0<y<x_F-u} |F_u(y) - H_{\xi,\beta(u)}| = 0. \tag{2.5}$$

Démonstration 2.2.1.

Pour une démonstration de ce théorème, on pourra se référer au [3] ou [4]

Remarque 2.1.

*Le théorème de **Pickands-Balkema-de Haan** est considéré aussi comme le deuxième théorème fondamental de la théorie des valeurs extrêmes,ce qui a donné une importance à la distribution de Pareto généralisée dans cette théorie.*

Dans cette approche on ne retient que les observations dépassant un seuil fixé $u < x_F$ .On a définit l'excès Y de la variable aléatoire X au dessus du seuil u par $X - u|X > u$.Si l'on note par F_u la fonction de répartition d'un excès au dessus du seuil u,on a pour tout $y > 0$

$$\begin{aligned} 1 - F_u(y) &= \Pr(Y > y) \\ &= \Pr(X - u|X > u) \\ &= \frac{\Pr(X > u + y, X > u)}{\Pr(X > u)} \\ &= \frac{1 - F(u + y)}{1 - F(u)}. \end{aligned} \tag{2.6}$$

Lorsque le seuil u est grand, on peut approcher cette quantité par la fonction de survie d'une loi **GPD**.Afin d'approcher le quantile,il suffit d'utiliser le résultat de **Pickands-Balkema-de Haan** qui établit l'équivalence entre la convergence en loi du maximum vers une loi des valeurs extrêmes G_ξ et la convergence en loi d'un excès vers une **GPD**.

Exemple 2.1.

1. *La loi exponentielle du paramètre 1 : $F(x) = 1 - e^{-x}$*
   on prend $\beta_u = 1$

$$\begin{aligned} F_u(y) = \Pr(X - u \leq x|X > u) &= \frac{F(u + y) - F(u)}{1 - F(u)} \\ &= \frac{e^{-u} - e^{-u-y}}{e^{-u}} \\ &= 1 - e^{-y}. \end{aligned}$$

   *Pour tout $y > 0$,la loi limite est la loi **GPD** de paramètre $\xi =$ et $\beta_u = 1$.*
   *Donc dans ce cas, la loi **GPD** n'est pas simplement la loi limite, mais il s'agit de la loi exacte pour tout u.*

2. *La loi de Pareto :* $F(x) = 1 - cx^{-\alpha}$, $(c > 0, \alpha > 0)$
On pose : $\beta_u = ub$, $b > 0$
$\forall y > 0$

$$\begin{aligned} F_u(y) &= \frac{F(u + uby) - F(u)}{1 - F(u)} \\ &= \frac{cu^{-\alpha} - c(u + uby)}{cu^{-\alpha}} \\ &= 1 - (1 + by)^{-\alpha} \end{aligned}$$

Si on pose $\xi = \frac{1}{\alpha}$ *et* $b = \xi$, *la limite est alors la loi* ***GPD*** *de paramètre* ξ.

Petite synthèse :(Relation entre **GEV** et **GPD**)

Si pour une variable aléatoire X de fonction de répartition F inconnue,l'échantillon des maximums normalisés *converge en distribution* vers une loi de probabilité non dégénérer,alors il est équivalent de dire que F est dans le max-domaine d'attraction d'une distribution **GEV** d'indice de queue $\xi \in \mathbb{R}$ ***(Théorème de Fisher-Tippet)***.
Dans ce cas, il s'en déduit que la distribution des excès de X au-delà d'un seuil u *converge uniformément* vers une distribution **GPD**,de même indice de queue que celui de la distribution **GEV**,lorsque ce seuil tend vers le point terminal x_F de la fonction de répartition F ***(Théorème Pickands-Balkema de Haan)***.

2.3 Estimation des paramètres de la GPD

L'estimation des paramètres d'une distribution **GPD** pose le problème de la détermination du seuil u,car il doit être suffisamment grand pour que l'on puisse appliquer le théorème précédent,mais ne doit pas être trop grand afin d'avoir suffisamment de données pour obtenir des estimateurs de bonne qualité.
Donc, tout d'abord avant d'estimer les paramètre de cette distribution on doit choisir le bon seuil.

2.3.1 Choix du seuil

Le choix du seuil reste toujours délicat mais il y'a une méthode graphique nous aide à déterminer le bon seuil u, dans cette méthode en utilise une fonction qui est ***la fonction moyenne des excès***.

Définition 2.5.
On appelle ***fonction moyenne des excès*** *de la variable alétoire* X *par rapport au seuil* $u < x_F$, *et on l'a note* $e(u)$,*la fonction espérance de la variable aléatoire* Y *excès de* X *au-delà du seuil* $u < x_F$ *définie par :*
$\forall u < x_F$

$$e(u) = \mathbb{E}(X - u | X > u) = \frac{1}{\bar{F}(u)} \int_u^{x_F} \bar{F}(t)dt. \tag{2.7}$$

Définition 2.6.
Soient $(X_1, X_2, ..., X_n)$ *l'échantillon de taille* $n \in \mathbb{N}^*$ *de la variable aléatoire* X *et* F_n *sa fonction de répartition empirique.*
On appelle ***fonction moyenne des excès empirique*** *de la variable alétoire* X *par rapport*

au seuil $u < x_F$ la fonction $e_n(u)$ définie par :
$\forall u < x_F$

$$\begin{aligned} e_n(u) &= \frac{1}{\bar{F}_n(u)} \int_u^{x_F} \bar{F}_n(t)dt \\ &= \frac{1}{card\{\Delta_n(u)\}} \sum_{i \in \Delta(u)} \big(X_i - u\big) \end{aligned} \tag{2.8}$$

avec : $\Delta_n(u)$=$\{i = 1, ..., n\}$ tel que $X_i > u$ et $\frac{0}{0} = 0$ (conventionnellement).

Proposition 2.1.

Si W est une variable aléatoire qui a comme fonction de répartition une distribution de Pareto généralisée $H_{\xi,\beta}$ avec $(\xi < 1, \beta > 0)$,alors sa fonction moyenne des excès $e(u)$ au-delà d'un seuil $u < w_0$ (w_0 est le point terminal de $H_{\xi,\beta}$) est donnée par :
$\forall u < w_0$

$$e(u) = \mathbb{E}(W - u|W > u) = \frac{\beta + \xi u}{1 - \xi} \tag{2.9}$$

avec : $\beta + \xi u > 0$.

• On peut résumé l'idées principale pour choisir le bon seuil on utilisant la méthode ***graphique*** comme suit :

1. **Cadre théorique :**
   D'après le théorème de **Pickands-Belkema-de Haan**, si à partir d'un certain seuil $u_0 < x_F$ l'excès de la variable aléatoire X au-delà du seuil u_0 suit une loi de Pareto généralisée, c-à-d : $\forall u_0 < x_F$

$$\begin{aligned} e(u_0) &= \mathbb{E}(X - u_0|X > u_0) \\ &= \frac{\beta(u_0)}{1 - \xi} \end{aligned}$$

   • Si cette approximation est vraie pour le seuil u_0,elle sera vraie pour un autre seuil u tel que : $\forall u_0 < u < x_F$

$$\begin{aligned} e(u) &= \mathbb{E}(X - u_0|X > u_0) \\ &= \frac{\beta(u_0) + \xi u}{1 - \xi}. \end{aligned} \tag{2.10}$$

   Avec : $\beta(u_0) + \xi u > 0$.
   • Et pour déterminer le seuil u,on exploite la linéarité en u de la fonction moyenne des excès $e(u)$.

2. **Cadre empirique :**

   Si on suppose que nos données $x_1, ..., x_n$ $(n \in \mathbb{N})$ sont une réalisation de l'échantillon $(X_1, ..., X_n)$ de la variable aléatoire X,alors on procède de la manière suivante pour déterminer le seuil $u < x_F$:
   On calcule l'estimateur $\hat{e}_n(u)$ de la fonction moyenne des excès $e(u)$ de X au-delà du seuil u,en utilisant sa version empirique $e_n(u)$ par :

$$\hat{e}_n(u) = \frac{1}{card\{i; x_i > u\}} \sum_{i; x_i > u} \big(x_i - u\big). \tag{2.11}$$

Avec :

$$\min_{1\leq i\leq n} x_i \leq u < \max_{1\leq i\leq n} x_i.$$

• On trace le graphe

$$\kappa_u = [u, \hat{e}_n(u)]; \min_{1\leq i\leq n} x_i \leq u < \max_{1\leq i\leq n} x_i. \tag{2.12}$$

• Une fois le graphe est tracé,on exploite la linèarité en u de la fonction moyenne des excès d'une distribution de Pareto-généralisée $\kappa_{\xi,\beta}$ au-delà du seuil choisissant $u \leq x$ où x est le point à partir duquel le graphe $\kappa_{\xi,\beta}$ est approximativement une droite.

2.3.2 Estimation des paramètres de la GPD

Une fois le seuil optimal choisi,on construire une nouvelle série d'observations au dessus de ce seuil et la distribution de ces données suit approximativement une distribution **GPD**.
Dans cette partie, on va estimer les paramètres de la **GPD** on utilisons la méthode du maximum de vraisemblance

. Méthode du maximum de vraisemblance

[11] La densité de la distribution **GPD** est :

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

La fonction de vraisemblance est donnée par :

$$l(\xi, \beta, x_1, ..., x_n) = \prod_{i=1}^{n} H_{\xi,\beta}(x_i)$$

Ce qui implique

$$\begin{aligned} \log l(\xi, \beta, x_1, ..., x_n) &= \sum_{i=1}^{n} H_{\xi,\beta}(x_i) \\ &= -n\log\beta - \left(1 - \frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1 + \frac{\xi}{\beta}x_i\right). \end{aligned}$$

On pose que $\tau = \dfrac{\xi}{\beta}$, l'annulation des dérivées partielles des logarithmes de la fonction de vraisemblance conduit au système :

$$\begin{cases} \hat{\xi} & = \dfrac{1}{n}\sum_{i=1}^{n}\log\big(1 + \tau X_i\big) = \hat{\tau} \\ \dfrac{1}{\tau} & = \dfrac{1}{n}\big(1 + \dfrac{1}{\xi}\big)\sum_{i=1}^{n}\dfrac{X_i}{1 + \tau X_i} \end{cases}$$

L'estimateur de **MV** (ξ, τ) est $(\hat{\xi} = \hat{\xi}(\hat{\tau}), \tau)$ où τ est la solution de :

$$\frac{1}{\tau} = \frac{1}{n}\big(1 + \frac{1}{\xi}\big)\sum_{i=1}^{n}\frac{X_i}{1 + \tau X_i}.$$

Cette équation se résous numériquement de manière itérative.
$\star$ Cet estimateur est *asymptotiquement normale*

$$\sqrt{n}\left(\hat{\xi}-\xi;\frac{\hat{\beta}}{\beta}-1\right)\xrightarrow{L}\mathcal{N}(0,M^{-1}).$$

Où

$$M^{-1}=(1+\xi)\begin{pmatrix}1+\xi & -1\\ -1 & 2\end{pmatrix}.$$

La normalité asymptotique des estimateurs $\hat{\xi}$ et $\hat{\beta}$ est prouvé pour $\xi>-\frac{1}{2}$
Qaund $\sqrt{n}\mu_n^{\sigma}L(\mu)\longrightarrow 0$ quand $n\longrightarrow+\infty$ et la fonction $x^{-\sigma}L(x)$ non décroissante au voisinage de l'infini

Ce résultat a été démontrer par ***Smith (1987)*** [13].

2.4 Conclusion

Dans ce chapitre on a exposé une autre méthode pour étudier les valeurs extrêmes, la méthode (***Peaks-Over-Tershold***) où ***méthode des excès au-delà d'un seuil***. Cette dernière est basée sur la distribution de Paréto généralisée, elle a l'avantage de diminuer la perte d'information par rapport à la méthode précédente.
Ce chapitre souligne l'essentiel du fondement théorique de cette méthode et le théorème de ***Pickands-Belkama-de Haan*** qui est considéré comme le deuxième théorème fondamental de la théorie des valeurs extrêmes. On a estimé les paramètres avec trois méthodes différentes mais après la détermination du seuil .

Chapitre 3

Méthode de Bootstrap

Motivation : La motivation du bootstrap (Efron, 1982 ; Efron et Tibshirani, 1993) est d'approcher par simulation la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent lorsque l'on ne peut pas supposer qu'elle est gaussienne. L'objectif est de remplacer des hypothèse probabilistes pas toujours vérifiées ou même invérifiables par des simulations et donc beaucoup de calcul .

3.1 Introduction

La méthode de Bootstrap a été proposée par Bradley Efron (1979) [6] comme une alternative aux modèles mathématiques traditionnels dans des problèmes d'inférence compliqué où une modélisation mathématique de la distribution des erreurs est difficile. Ce dernier a écrit beaucoup au sujet de la méthode et de ses générations (Efron 1982 ,Efron 1985,...).Des milliers de papiers ont été rédigés sur le Bootstrap dans les derniers décennies. La méthode et ses modifications ont trouvé une utilisation très large dans des problèmes appliqués en remplaçant les méthodes asymptotiques usuelles. Le Bootstrap est une méthode de type Monte-Carlo basée sur des données observées (Efron et Tibshirani 1993)[7].Dépendant de l'information disponible au sujet de la loi de population ou du paramètre d'intérêt, il existe la méthode de Bootstrap paramétrique et non paramétrique. Le nom Bootstrap provient de l'expression "to pull oneself up by one's bootstrap".("Les aventures du baron Munchausen" par Rudolph Erich Raspe).Dans ce livre, le baron est tombé au fond d'un lac profond. Quand tout semblait perdu, une idée géniale lui est venue à l'esprit : se soulever en tirant sur les languettes de ces bottes. Le Bootstrap fournit plusieurs avantages par rapport à l'approche paramétrique traditionnelle.Il est facile de décrire la méthode et celle-ci s'applique aus situations compliquée , car ses supposition ayant à la distribution des données, telle que la normalité ou d'autre, ne sont pas nécessaires (la distribution a priori est inconnue).

3.2 Définition de la méthode Bootstrap

Schématiquement, le Bootstrap est une méthode qui consiste à réaliser par le biais d'un ordinateur ce que l'expérimentateur pourrait faire en pratique si cela était possible, c'est a dire répéter l'expérimentation un grand nombre de fois. Ainsi, avec le Bootstrap de nouvelle expérimentation ne sont plus nécessaire . En effet, les données disponibles réutilisées.Pour être plus précis, les observations originales sont réaffectées de façons aléatoire, puis l'estimateur est recalculé. Ces réaffections et recalcules d'estimateur sont réalisés un grand nombre de fois et sont considères comme des répétitions de l'expérimentateur à première vue, cela ressemble à une simulation Monte-Carlo, mais ce n'est pas le ca. En effet, le plus grand avantage du Bootstrap comparé avec la méthode de Monte-Carlo réside dans le fait que le Bootstrap ne nécessite pas la répétition de l'expérimentation, contrairement aux simulation Monte-Carlo.

D'un autre point de vue, l'idée principale de Bootstrap est de simuler autant que possible la probabilité du monde réel en substituant les inconnues par des estimés provenant des données observées. Ainsi, à partir des simulations du monde Bootstrap , des entités d'intéret inconnues dans le monde réel peuvent être estimées comme il est montré dans la figure .

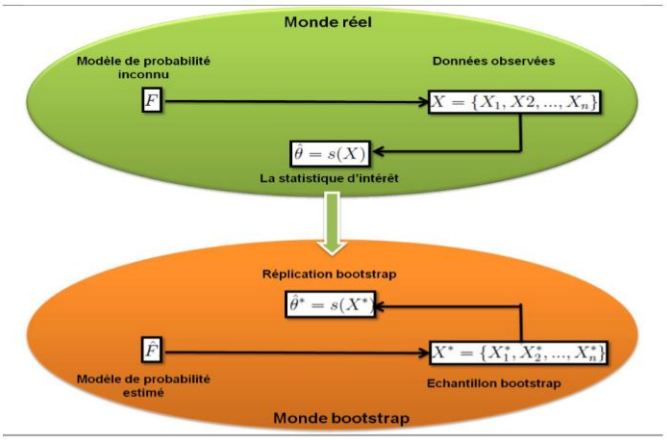

FIGURE 3.1 : Approche de bootstrap

3.3 Principe de substitution

[11] Le principe de Substitution est une méthode simple et intuitive d'estimation de paramètre à partir d'un échantillon. Ce principe préconise la Substitution de F par $\widehat{F}$ dans tout fonctionnelle définissant le paramètre à estimer pour obtenir l'estimateur. On estime donc $\theta = \theta(F)$ par $\widehat{\theta} = \theta(\widehat{F})$, les exemples sont nombreux ainsi si

$$\theta(F) = \int (XdF) = E(X)$$

. Où X est une variable aléatoire de fonction de répartition F, on a

$$\theta(\widehat{F}) = \int X d\widehat{F} = \sum_{i=1}^{n} X_i \frac{1}{n} = \overline{X}.$$

De la même façons on peut établir que si :

$$\theta(F) = \int (X - E(X(X))^2 dF = Var(X)$$

alors

$$\theta(\widehat{F}) = \int (X - E_{\widehat{F}}(x))^2 d\widehat{F} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \widehat{Var}(X).$$

Ou si bien si $\theta(F) = F^{-1}(\frac{1}{2})$ alors $\widehat{\theta} = \theta(\widehat{F}) = \widehat{F}^{-1}(\frac{1}{2})$ et donc on estime la médiane dans la population par la médiane dans l'échantillon. L'exemple ci-après illustré de principe de Substitution pour l'estimation de paramètre d'une distribution bidimensionnelle.

Exemple 3.1.

On considère une Population de 45 patients ayant subi de diabète où l'on a entre autre mesuré les niveaux de glucose et d'insuline

Patient	*Glucose*	*Insuline*	*Patient*	*Glucose*	*Insuline*	*Patient*	*Glucose*	*Insuline*
1	*393*	*202*	*16*	*360*	*134*	*31*	*356*	*124*
2	*364*	*152*	*17*	*336*	*134*	*32*	*289*	*117*
3	*359*	*185*	*18*	*352*	*169*	*33*	*319*	*143*
4	*296*	*116*	*19*	*353*	*263*	*34*	*356*	*199*
5	*345*	*123*	*20*	*373*	*174*	*35*	*323*	*240*
6	*378*	*136*	*21*	*376*	*134*	*36*	*381*	*157*
7	*304*	*134*	*22*	*367*	*182*	*37*	*350*	*221*
8	*347*	*184*	*23*	*335*	*241*	*38*	*301*	*186*
9	*327*	*192*	*24*	*396*	*138*	*39*	*379*	*142*
10	*386*	*279*	*25*	*277*	*222*	*40*	*296*	*131*
11	*365*	*228*	*26*	*378*	*165*	*41*	*353*	*221*
12	*365*	*145*	*27*	*360*	*282*	*42*	*306*	*178*
13	*352*	*172*	*28*	*291*	*94*	*43*	*290*	*136*
14	*325*	*179*	*29*	*269*	*121*	*44*	*371*	*200*
15	*321*	*222*	*30*	*318*	*73*	*45*	*312*	*208*

Avec les données de tableau sur les patients ayant subi eu test de diabète et en posant x le niveau de glucose et y le niveau d'insuline on calcule

$$\mu_x = \frac{1}{45}\sum_{i=1}^{45} X_i$$

et

$$\mu_y = \frac{1}{45}\sum_{i=1}^{45} Y_i$$

on obtient $\mu_x = \frac{15350}{45} = 341,111$ *et* $\mu_y = \frac{7777}{45} = 172,822$ *les médiane de x et y nous donnent* $F_x^{-1}(\frac{1}{2}) = 353$ *et* $F_y^{-1}(\frac{1}{2}) = 172$ *on peut aussi calculer le coefficient de corrélation*

$$r(x,y) = \frac{\sum_{i=1}^{45}(X_i - \mu_x)(Y_i - \mu_y)}{[\sum_{i=1}^{45}(X_i - \mu_x)^2 \sum_{i=1}^{45}(Y_i - \mu_y)^2]^{\frac{1}{2}}}$$

ce qui nous donnent $r(x,y) = 0,238$. Toutes ces statistiques ont été calculées avec toute l'information contenue dans F. Si on tire l'échantillon suivant de taille 10 :

(289, 117)(335, 241)(381, 157)(356, 124)(365, 228)(345, 123)(277, 222)
(366, 279)(378, 136)(359, 185)

qui sont les patients 32,23 36,31 11,5,25,10,6 et 3 respectivement, on calcule par le principe de substitution :$\widehat{\mu}_x = \overline{x} = \frac{1}{45}\sum_{i=1}^{10} x_i = \frac{3590}{10} = 359,0$ et $\widehat{\mu}_y = \overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1718}{10} = 171,8$

ainsi que la médiane :$\widehat{F}_x^{-1}(\frac{1}{2}) = 362$ et $\widehat{F}_y^{-1}(\frac{1}{2}) = 126$

et finalement le coefficient de corrélation :

$$\widehat{r}(x,y) = \frac{\sum_{i=1}^{10}(x_i - \widehat{\mu}_x)(y_i - \widehat{\mu}_y)}{[\sum_{i=1}^{10}(x_i - \widehat{\mu}_x)^2 \sum_{i=1}^{10}(y_i - \widehat{\mu}y)^2]^{\frac{1}{2}}} = 0,221.$$

Lorsque la seul information disponible sur F est celle donnée par l'échantillon x, le principe de substitution donne généralement le bons résultats.

3.4 L'erreur type par la méthode de Bootstrap

après avoir calculé $\widehat{\theta} = \theta(\widehat{F})$; l'estimateur par substitution de θ, on s'intéresse alors à la précision de $\widehat{\theta}$. L'erreur type est une mesure nous permettant d'avoir une idée de cette précision, mais lorsque l'estimateur est complexe, il arrive parfois que la formule théorique pour son erreur type soit très difficile à dériver. Le Bootstrap a été introduit à cette fin.

Pour bien comprendre le principe de substitution et la méthode de Bootstrap, nous voyons leur application sur l'erreur type d'une statistique bien connue, la moyenne .

Définition 3.1.

Soit une variable aléatoire X à valeur réel issue d'une fonction de répartition F. Les symboles $\mu_F = E_F(X)$ *et* $\sigma_F^2 = Var_F(X)$ *représentent respectivement l'espérance et la variance de F. si on tire un échantillon aléatoire* $X = (x_1, x_2, ..., x_n)$ *de taille n , alors la moyenne*

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

a respectivement comme espérance et variance μ_F *et* $\frac{\sigma_F^2}{n}$*. L'erreur type de* $\overline{x}$ *est notée :*

$$se_F(\overline{x}) = [Var_F(\overline{x})]^{\frac{1}{2}} = \frac{\sigma}{\sqrt{n}}$$

Exemple 3.2.

Avec la population donnée dans le tableau 1, on calcule :

$$\sigma_x^2 = \frac{1}{45} \sum_{i=1}^{45} (X_i - \mu_x)^2 = 1102,68$$

et on calcule donc l'erreur type de la moyenne $se(\overline{x}) = \frac{\sigma_x}{\sqrt{10}} = 10,5$

Dans cette exemple le calcule de $se(\overline{x})$ est assez simple, mais ce n'est pas toujours le cas. Dans plans d'échantillonnage plus complexe , par grappes ou stratifiés, et avec des statistiques plus complexe de $\overline{x}$, le calcul de l'erreur type peut s'avérer très difficile . De plus , dans l'exemple on avait des valeurs exacte de μ_F et $\frac{\sigma_F^2}{n}$, mais dans la majorité des cas on doit utiliser des estimations de ces paramètre. On doit alors utiliser le théorème central limite pour procéder à des inférence .

La méthode de Bootstrap, qui repose principalement sur des simulations et des calculs fait à l'ordinateur, est simple et représente une application direct du principe de substitution. La méthode de Bootstrap fait partie de la famille des méthodes dites "ré-échantillonnage" car on utilise successivement appelés "échantillon Bootstrap" qui sont tirés de l'échantillon étudie.

Définition 3.2.

Soit $X = (x_1, x_2, ..., x_n)'$ *un échantillon issue d'une loi F. À partir de X on peut construire la fonction de répartition empirique* $\widehat{F}$ *. Un échantillon Bootstrap* $x^* = (x_1^*, x_2^*, ..., x_n^*)'$ *est un échantillon aléatoire tiré avec remise parmi les observations* $x_1, x_2, ..., x_n$*. Il s'agit donc d'observations indépendantes et identiquement distribuées issues de* $\widehat{F}$*. On notera occasionnellement* $\widehat{F}$*.*

puisque $x^* = (x_1^*, x_2^*, ..., x_n^*)'$ *est un échantillon avec remise , nous pouvons par exemple , avoir tiré* $x_1^* = x_2, x_2^* = x_2, ..., x_n^* = x_n$*, avec un ou plusieurs des* x_i *apparaissant plusieurs fois*

et d'autres apparaissant aucune fois. Sur cette échantillon Bootstrap, nous pouvons calculer $\widehat{\theta}$. *Nous aurons alors par exemple*

$$\widehat{\theta}^* = \overline{x}^* = \sum_{i=1}^{n} \frac{x_i^*}{n}$$

qui est la moyenne de l'échantillon Bootstrap pour la variable X.

L'estimation Bootstrap de $se_F(\overline{x}^*)$ *est calculé par le principe de substitution , utilisant la fonction de répartition empirique* $\widehat{F}$. *En fait ,*$se_F(\overline{x}^*)$ *est l'erreur type de* $\overline{x}$ *pour les échantillon aléatoire de taille n tirés de* $\widehat{F}$.

Illustrons maintenant, dans un exemple, la création d'un échantillon Bootstrap ainsi que le calcul de l'erreur type de la moyenne par le principe de substitution

Exemple 3.3.
Soit

(335, 241)(289, 117)(359, 185)(345, 123)(396, 138)(359, 185)(289, 117)
(345, 123)(356, 124)(378, 136)

un échantillon Bootstrap tiré de l'échantillon de l'exemple précédent, on a tiré les patients 23,32,3,5,24,3,32,5,31 et 6 respectivement. on note que l'on a observé deux fois les patients 32 ,3 , et 5 et aucune fois les patients 10,11 et 36.on calcule $\overline{x}^* = 345,1$ *et* $se_F(\overline{x}^*)$ *par :*

$$se_F(\overline{x}^*) = \frac{\sigma_F}{\sqrt{n}} = \left[\frac{\sum_{i=1}^{n}(x_i^* - \overline{x}^*)^2}{n^2}\right]^{\frac{1}{2}}.$$

Ce qui nous donne $se_F(\overline{x}^*) = 10,28$.

On a utilisé le fait qu'il existe une formule déjà établie pour calculer l'erreur type de la moyenne .Généralement , pour un paramètre θ quelconque, il n'y a pas de formule . On doit alors utilisér l'algorithme de Bootstrap, expliqué ici en 5 étapes simple.

1. On part de l'échantillon étudie $X = (x_1, x_2, ..., x_n)'$ sur lequel on a estimer le paramètre θ par $\widehat{\theta}$
2. Tirer aléatoirement B échantillon Bootstrap $x^{*1}, x^{*2}, ..., x^{*B}$ de $\widehat{F}$. Ces échantillon sont en fait des sous-échantillon tirés avec remise de X .
3. Èvaluer $\widehat{\theta}^*$ pour chacun des échantillon Bootstrap tirés , $\widehat{\theta}^{*b} = \theta(x^{*b})$ pour b=1,2,...,B. On appelle cette évaluation de $\widehat{\theta}^*$ une réplique Bootstrap .
4. Estimer l'erreur type $se_F(\widehat{\theta})$ par l'écart type empirique des B répliques Bootstrap

$$\widehat{se}_B = \left[\frac{\sum_{b=1}^{B}[\widehat{\theta}^{*b} - \widehat{\theta}^*(.)]^2}{B-1}\right]^{\frac{1}{2}}$$

   avec

$$\widehat{\theta}^*(.) = \frac{\sum_{b=1}^{B}\widehat{\theta}^{*b}}{B}$$

5. On obtient une estimation de l'erreur type se_F pour $\widehat{\theta}$.

La figure (3.2) explique plus clairement une implémentation pratique du principe du bootstrap non-paramétrique pour l'estimation de la fonction de distribution d'un estimateur donné.

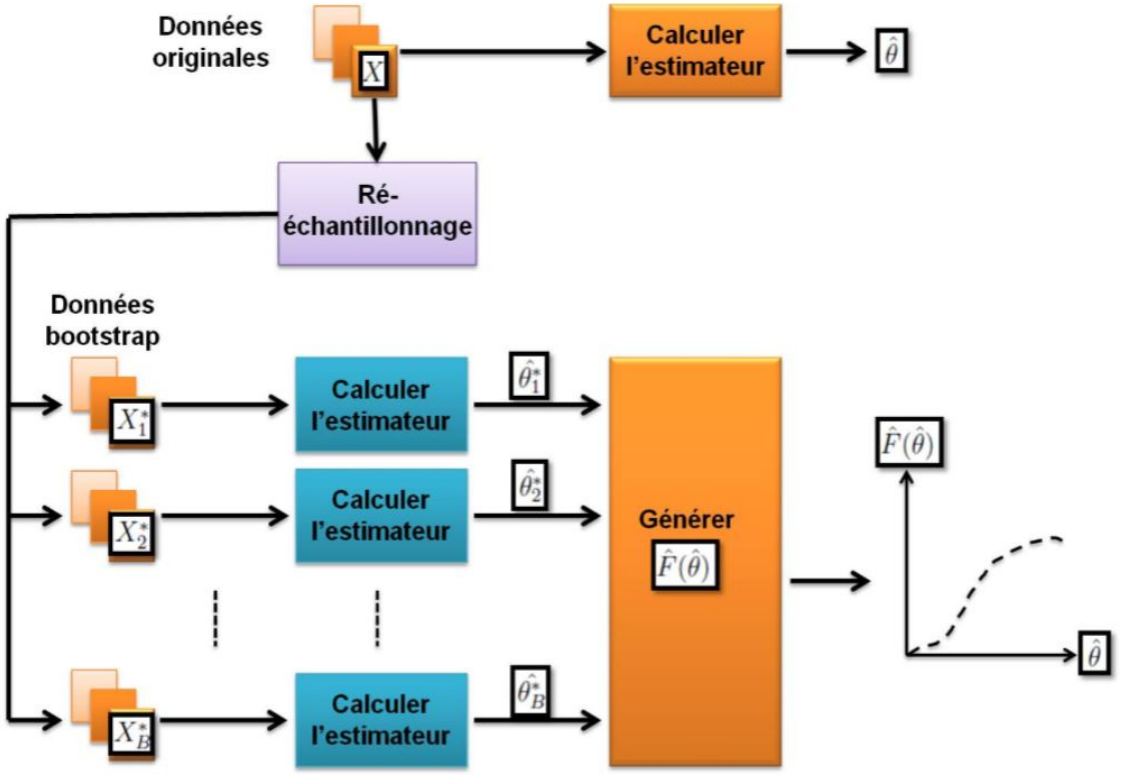

FIGURE 3.2 : *Principe du bootstrap non-paramétrique pour l'estimation de la fonction de* $\widehat{F}(\widehat{\theta})$ *du paramètre* $\widehat{\theta}$

On utilisant les données sur le glucose et l'insuline de tableau 1, on calcule, par la méthode de Bootstrap, l'erreur type de la moyenne empirique de la mesure du glucose .

Exemple 3.4.

Après avoir appliqué l'algorithme de Bootstrap décrit plus haut et avec B=50, on calcule l'erreur de la moyenne de la mesure du glucose $\widehat{se}_{50} = 11,23$*, avec B=200 on obtient* $\widehat{se}_{200} = 10,30$

3.5 Conclusion

Le bootstrap est une méthode récente qui consiste à développer les techniques de construction de certaines catégorie statistique , elle est basé sur la puissance de l'ordinateur pour facilité les calculs .

Dans ce chapitre, la méthode du bootstrap est utilisé dans l'étude statistique pour estimer la moyenne et l'écart-type (erreur standard) du bootstrap.

Deuxième partie :

Partie d'application

Chapitre 4

Application de la méthode de bootstrap

4.1 Introduction

Dans ce chapitre, nous réalisons une étude de cas réelle pour modéliser la distribution des valeurs extrêmes généralisées, les données sont obtenues suite à une collecte des données de Direction de la météo (Dellys - boumerdes).
L'objectif principal de ce chapitre est l'estimation des paramètres de la loi (GPD) par la méthode de bootstrap et les comparer avec la méthode maximum vraisemblance .

4.2 Simulation

D'abord on fait une simple simulation pour l'estimation des paramètre de la loi GPD (ξ et β), nous avons regroupé les résultats obtenu sur un tableaux , les résultats a été exécuter avec logiciel $\mathcal{R}$ sont :

Taille d'échantillon	Nombre de réplication	ξ	Estimation du ξ par (MV)	Estimation du ξ après avoir échantillon bootstrap	β	Estimation de β par (MV)	Estiamtion de β après avoir échantillon bootstrap
100	100	0,3	0,3760854	0,3463019	1	1,3052792	1,2828
100	100	0,5	0,3479909	0,5394003	1	1,1895337	1,120738
100	100	0,7	0,507491	0,6771015	1	1,0670205	0,9527311

TABLE 4.1 : Estimation du ξ et β pour N = 100 avec B = 100

Taille d'échantillon	Nombre de réplication	ξ	Estimation du ξ par (MV)	Estimation du ξ après avoir échantillon bootstrap	β	Estimation de β par (MV)	Estimation de β après avoir échantillon bootstrap
100	500	0,3	0,2352065	0,3242528	1	1,29569990	1,252199
100	500	0,5	0,4150281	0,4722365	1	1,12016110	0,9591528
100	500	0,7	0,6110486	0,697437	1	0,8395827	0,9587812

TABLE 4.2 : Estimation de ξ et β pour N =100 avec B= 500

Taille d'échantillon	Nombre de réplication	ξ	Estimation du ξ par (MV)	Estimation du ξ après avoir échantillon bootstrap	β	Estimation de β par(MV)	Estiamtion de β après avoir échantillon bootstrap
500	1000	0,3	0,313268	0,310592	1	0,8891872	0,9656958
500	1000	0,5	0,4477305	0,5243101	1	0,9304253	1,046106
500	1000	0,7	0,644573	0,6994459	1	1,0656652	0,9904037

TABLE 4.3 : Estimation de ξ et β pour N =500 avec B= 1000

On remarque d'après les trois tableaux que l'estimation des paramètres de la loi GPD par la méthode de bootstrap est meilleur que la méthode du maximum de vraisemblance, nous avons remarqué que plus le nombre de réplication augmente plus l'estimation des paramètres est meilleure.

4.3 Application de la méthode de bootstrap aux données climatiques

Nous avons appliqué la méthode de bootstrap pour des données climatiques de la région de Dellys ces données mensuel sont établi sur 13 ans (156 mois).

nous avons fait une analyse statistique descriptive, et les résultats sont représenter dans le tableaux suivant.

	Min	Mean	Max
Température	10,1	18,22244	27,6
Humidité	9	69,60256	86
Vent	1,6	3,098718	5,9
Précipitation	0	57,77564	308,7

TABLE 4.4 : Analyse statistique des données climatiques

Nous avons estimé les paramètres de la loi de (GPD) après le choix de seuil par la méthode de (MV) et bootstrap. les résultats a été exécuter avec logiciel $\mathcal{R}$

	threshold	Estimation de ξ par (MV)	Estimation de β par (MV)	Estimation de ξ par (bootstrap)	Estimation de β par (bootstrap)
Température	15	-0,794323	10,04919	-0,7730243	9,650251
Humidité	70	-0,4522117	7,824814	-0,4815472	8,06588
Vent	3	-0,1890025	0,8808928	-0,1669094	0,9042077
Précipitation	100	-0,794323	68,735779	-0,1240874	65,84011

TABLE 4.5 : Estimation des paramètres des données climatiques pour B=1000 et N=156

	threshold	Estimation de ξ par (MV)	Estimation de β par (MV)	Estimation de ξ par (bootstrap)	Estimation de β par (bootstrap)
Température	15	-0,794323	10,04919	-0,7825336	11,69575
Humidité	70	-0,4522117	7,824814	-0,4808428	8,09715
Vent	3	-0,1890025	0,8808928	-0,1990736	0,8848719
Précipitation	100	-0,794323	68,735779	-0,1253985	67,51786

TABLE 4.6 : Estimation des paramètres des données climatiques pour B=10000 et N=156

	threshold	Estimation de ξ par (MV)	Estimation de β par (MV)	Estimation de ξ par (boots-trap)	Estimation de β par (boots-trap)
Température	15	-0,794323	10,04919	-0,794323	9,855963
Humidité	70	-0,4522117	7,824814	-0,474641	7,935745
Vent	3	-0,1890025	0,8808928	-0,1701292	0,8874125
Précipitation	100	-0,794323	68,735779	-0,1209151	70,88683

TABLE 4.7 : Estimation des paramètres des données climatiques pour B=100000 et N=156

On remarque d'après les trois tableaux que l'estimation des paramètres de la loi (GPD) par la méthode de bootstrap et la méthode de maximum du vraisemblance sont très proche sauf pour la précipitation .

Conclusion générale

La méthode de bootstrap est une méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles.

L'utilisation des techniques de ré-échantillonnage a été rendue possible grâce à la généralisation des moyens de calculs performants, ces techniques reposent, au départ, sur des idées simples. Toutefois, il faut bien admettre que les développements apportés aux méthodes de base leur on fait perdre une partie de cette simplicité .

Dans ce mémoire, nous nous sommes limité au problème de l'estimation des paramètres de la loi (GPD). Il s'agit cependant d'un seul application de méthode de ré-échantillonnage. Le bootstrap est ainsi particulièrement utile lorsque les échantillons de données sont de petits tailles .

L'étude de simulation nous a montré que l'estimation des paramètres de loi GPD par la méthode du bootstrap est généralement meilleure que la méthode du maximum de vraisemblance. Nous avons achevé ce travail par une application aux données climatiques.

Nous avons remarqué que l'estimation des paramètres de la loi (GPD) par le bootstrap et la méthode du maximum de vraisemblance sont très proches sauf pour la précipitation.

Bibliographie

[1] **de Haan,A.Ferreira,A**, *Extreme value theory . An introduction*, (2006)

[2] **Resnick,S**, *Extreme values, Regular variation, and Point Processes.*, Applied Probability Trust, Springer-Verlag. New York, (1987).

[3] **Berger,J**, Lecture Notes in statistics.

[4] **Borchani,A**, *Statistiques des valeurs extrêmes dans le cas discrètes*, Research center. ESSEC working paper 10009, Décembre 2009.

[5] **Christan,Y.Robert**, *Cour Théorie des valeurs extrêmes* , ISFA-Université de Lyon 1, 2016.

[6] , **Efron, B. (1979)** *Bootstrap methods : another look at the jackknifie*, annals of statistical association 82 :171-185

[7] **Efron, B. & R.J.Tibshirani .(1993)** *An Introduction to the Bootstrap*, chapman & Hall, New York

[8] **Raotke, M. (1988).** *Rates of convergence and asymptotic expansions under von Mises conditions*Ph.D. dissertation, Univ. Siegen (in German).

[9] **Fisher,R et Tippett,L .(1928).**, *Limiting forms of the frequency loi of the largest or smallest member of a sample*, Proceedings of the Cambridge Philosophical Society 24, 1928, p. 180-190.

[10] **Embrechts,P et Klüppelberg et C and Mikosch,T (1997)**, *Modelling extremal events For insurance* , Applications of Mathematics (New York),33. Springer-Verlag, Berlin.

[11] **Resnick,S**, *extrême values,regular variation and point processes* SpringerVerlag,1987.[cité en page 11].

[12] **Palm,R**, *Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres.*Biotechnol. Agron. Soc. Environ. 2002, 6(3), [143–153]

[13] **Smith,R.(1987)**, *Estimating tails of probability distributions*The annals of statistics,15,1174-1207.

Annexe

• Les sorties du langage $\mathcal{R}$ sont comme suit :

estimation de paramètre ξ du partie de simulation

```
> h5<-seq(1:50)
> h6<-seq(1:50)
> for(j in 1:50)
+ {
+ data<-rgpd(500,0.3,10,1)
+ data
+ h1<-gpd(data, threshold = 10, method = c("ml"))
+ h1
+ B<- replicate(50, sample(data, 500, TRUE), simplify = FALSE)
+ B
+ for(i in 1:50){
+ h3[[i]]<-gpd(B[[i]], threshold = 10, method = c("ml"))
+ h4[i]<-h3[[i]][7]
+ }
+  for(i in 1:50)
+ {
+ + print(h4[[i]][1])
+ h5[i]<-h4[[i]][1]
+ }
+ h6[j]<-mean(h5)
+ }
> mean(h6[j]
```

estimation de paramètre β du partie de simulation

```
> h5<-seq(1:50)
> h6<-seq(1:50)
> for(j in 1:50)
+ {
+ data<-rgpd(500,0.3,10,1)
+ data
+ h1<-gpd(data, threshold = 10, method = c("ml"))
+ h1
+ B<- replicate(50, sample(data, 500, TRUE), simplify = FALSE)
+ B
+ for(i in 1:50){
+ h3[[i]]<-gpd(B[[i]], threshold = 10, method = c("ml"))
+ h4[i]<-h3[[i]][7]
+ }
+  for(i in 1:50)
+ {
+ + print(h4[[i]][2])
+ h5[i]<-h4[[i]][2]
+ }
+ h6[j]<-mean(h5)
+ }
> mean(h6[j]
```

estimation le paramètre ξ du partie d'application

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dv<-read.csv2("vent.csv",header=TRUE)
>dv

>ev<-gpd(dv[,1], threshold=3, method = c("ml"))
>ev

>B<- replicate(1000, sample(dv[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 3, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][1])
+h5[i]<-h4[[i]][1]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-ev[[7]][1]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dp<-read.csv2("pre.csv",header=TRUE)
>dp

>ep<-gpd(dp[,1], threshold=100, method = c("ml"))
>ep

>B<- replicate(1000, sample(dp[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 100, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][1])
+h5[i]<-h4[[i]][1]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-ep[[7]][1]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dt<-read.csv2("tem.csv",header=TRUE)
>dt

>et<-gpd(dt[,1], threshold=15, method = c("ml"))
>et

>B<- replicate(1000, sample(dt[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 15, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][1])
+h5[i]<-h4[[i]][1]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-et[[7]][1]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dh<-read.csv2("hum.csv",header=TRUE)
>dh

>eh<-gpd(dh[,1], threshold=70, method = c("ml"))
>eh

>B<- replicate(1000, sample(dh[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 70, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][1])
+h5[i]<-h4[[i]][1]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-eh[[7]][1]
>h1
```

estimation le paramètre β du partie d'application

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dv<-read.csv2("vent.csv",header=TRUE)
>dv

>ev<-gpd(dv[,1], threshold=3, method = c("ml"))
>ev

>B<- replicate(1000, sample(dv[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 3, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][2])
+h5[i]<-h4[[i]][2]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-ev[[7]][2]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dp<-read.csv2("pre.csv",header=TRUE)
>dp

>ep<-gpd(dp[,1], threshold=100, method = c("ml"))
>ep

>B<- replicate(1000, sample(dp[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 100, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][2])
+h5[i]<-h4[[i]][2]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-ep[[7]][2]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dt<-read.csv2("tem.csv",header=TRUE)
>dt

>et<-gpd(dt[,1], threshold=15, method = c("ml"))
>et

>B<- replicate(1000, sample(dt[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 15, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][2])
+h5[i]<-h4[[i]][2]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-et[[7]][2]
>h1
```

```
>h4<-seq(1:1)
>h5<-seq(1:1)
>h6<-seq(1:1)

>dh<-read.csv2("hum.csv",header=TRUE)
>dh

>eh<-gpd(dh[,1], threshold=70, method = c("ml"))
>eh

>B<- replicate(1000, sample(dh[,1], 156, TRUE), simplify = FALSE)
>B
>for(i in 1:1000){
+h3[[i]]<-gpd(B[[i]], threshold = 70, method = c("ml"))
+h4[i]<-h3[[i]][7]
+}
>for(i in 1:1000){
+print(h4[[i]][2])
+h5[i]<-h4[[i]][2]
+}
>h6<-mean(h5)

>mean(h6)

>h1<-eh[[7]][2]
>h1
```

Le package qui on a utilisé c'est le package "evir"

ملخص

إن تقدير القيم القصوى بموجب القانون ***(GPD)*** *له أهمية كبيرة في التنبؤ بالأحداث المتطرفة التي يمكن أن تسبب كوارث النمذجة حسب القانون****(GPD)*** *يمر حسب تقدير هذه المعلمات.*

*لقد حاولنا تقديرها بواسطة طريقة****(Bootstrap)****من خلال مقارنتها بالطريقة الكلاسيكية التي تمثل الحد الاقصى لطريقة الاحتمالية.*

Résumé

L'estimation des valeurs extrêmes par la loi (GPD) à une grande importance dans la prévision des phénomènes extrêmes qui peuvent engendrer des catastrophes.la modélisation par la loi GPD passe par l'estimation de ces paramètres.

Nous avons essayé d'estimer ces derniers par la méthode de bootstrap en le comparant avec la méthode classique qui est la méthode du maximum de vraisemblance.

Abstract

The estimation of extreme values by law (GPD) is of great importance in the forecastingextremephenomenathatcan lead to disasters. Mdeling by the law (GPD) passes by the estimate of theseparameters.

Wetried to estimatethese last by the bootstrapmethod by comparingitwith the classicalmethodwhichis the maximum likelihoodmethod.