RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université Akli Mohand Oulhadj Bouira

Faculté des Sciences et des Sciences Appliquées
Département de Recherche Opérationnelle

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Recherche Opérationnelle

Estimation de l'Indice Extrêmal et Application

Alliche **NASSER**

Sous la direction de **Mr. Ait Yala Nabil**

Composition du jury :

Président :	Mr. Iftissen Elghani
Promoteur :	Mr. Ait Yala Nabil
Examinateur :	Mr. Demmouche Nacer
Examinateur :	Mr. Hamid Karim

June, 2024

Dedicaces

À ma chère maman, qui nous a quittés en ressentant une immense joie pour les succès que j'ai accomplis aujourd'hui, et à mon papa,

Ma chère maman, même si vous êtes maintenant dans un autre monde, je sais que vous me regardez avec fierté et bonheur. Mes rêves portaient votre silence et votre sagesse, et je vous serai toujours reconnaissant pour l'amour que vous avez semé dans mon cœur.

À mon frère et ma sœur,

Avec des mots de joie et de gratitude, je vous dédie cette réussite qui n'aurait pas été possible sans votre soutien continu et votre amour infini. Mon parcours a été rempli de rêves et d'ambitions que vous avez partagés avec moi avec goût et affection.

À mes amis,

Avec des mots reconnaissants et expressifs, je vous offre ce mémoire comme symbole de l'amitié profonde qui nous unit. Merci pour vos encouragements, votre soutien et les moments magnifiques qui sont devenus indispensable de mon parcours académique.

Enfin, à tous ceux qui ont partagé ce voyage académique avec moi de quelque manière que ce soit, que ce soit par des conseils, des encouragements ou un soutien émotionnel, je vous exprime ma profonde gratitude et mon appréciation.

Avec fierté et reconnaissance,

Remerciements

Louange à Dieu,

Je tiens à exprimer mes sincères remerciements et ma profonde gratitude à l'enseignant superviseur AIT YALLA NABIL ainsi qu'à tous les membres précieux de l'Université de Bouira pour leur soutien inestimable et leurs efforts considérables tout au long de mon parcours académique. Votre guidance précieuse et vos encouragements constants ont été une source d'inspiration constante, m'aidant à atteindre mes objectifs avec succès.

Je suis reconnaissant pour votre disponibilité et votre ouverture d'esprit, toujours prêts à répondre à mes questions et à m'offrir des conseils avisés qui ont grandement enrichi mon apprentissage et mon développement personnel.

Je tiens également à remercier le jury qui a accepté de discuter de mon mémoire de fin d'études, à leur tête le président du jury, M. Elghani Iftissen, ainsi que M. Hamid Karim et M. Demmouche Nacer.

J'aimerais également exprimer ma sincère reconnaissance à tous les étudiants qui ont partagé cette période d'apprentissage et de croissance avec moi. Cette expérience a été enrichissante grâce à votre collaboration et votre esprit de communauté, sans laquelle je n'aurais pas pu accomplir cette réussite.

Merci du fond du cœur pour votre soutien indéfectible et pour avoir contribué à rendre cette expérience universitaire si mémorable et formatrice.

Avec mes meilleurs vœux et mes salutations distinguées,

Abstract

Cette étude se concentre sur l'estimation de l'indice extrême, un outil crucial dans l'analyse des queues de distribution des données extrêmes. L'indice extrême est utilisé pour caractériser la dépendance de ces queues, ce qui est essentiel dans divers domaines comme la finance, l'assurance et l'hydrologie. Deux méthodes principales, celles de Hill et de Pickands, sont explorées dans cette recherche pour estimer cet indice.

Nous avons utilisé ces deux méthodes pour obtenir une compréhension approfondie et holistique des caractéristiques des queues de distribution extrêmes. Le choix de la méthode d'estimation optimale dépend fortement des propriétés des données et du nombre d'observations extrêmes utilisées dans le calcul. Par conséquent, une approche intégrative combinant ces deux méthodes est recommandée pour obtenir une analyse complète et précise des données extrêmes.

L'intégration de ces approches peut avoir un impact significatif sur la prise de décision éclairée dans des domaines tels que la finance, l'assurance et l'hydrologie, où les valeurs extrêmes revêtent une grande importance.

mots-clés : Indice extrême Analyse des queues de distribution, Méthode de Hill, Méthode de Pickands, Estimation statistique.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

La seconde moitié du XXe siècle a été caractérisée par un boom de l'activité économique, qui a entraîné une augmentation sans précédent de la richesse totale de la population mondiale et, par conséquent, des risques auxquels elle est exposée. Beaucoup de ces dangers proviennent d'événements politiques, économiques, naturels ou accidentels, dont il est difficile de prévoir avec précision la survenance. Afin de prévenir ou de se préparer à leurs effets négatifs potentiels, une gamme très diversifiée et sans cesse croissante d'instruments d'assurance est développée et, surtout ces dernières années, la gestion des risques a pénétré tous les aspects des processus économiques.

Un sujet important dans la gestion des risques est l'analyse, la modélisation et la prévision d'événements extrêmes rares mais dangereux, appelés « événements du pire cas ». En effet, les impacts les plus dramatiques sur un système sont généralement infligés dans des circonstances extraordinaires, lorsque l'expérience commune et les mesures de sécurité s'effondrent et que la logique des causes avalancheuses et des effets dévastateurs prend le dessus. Cela pourrait conduire à l'échec total du système, comme cela s'est produit dans un passé récent avec certains krachs boursiers et faillites spectaculaires.

De toute évidence, de tels scénarios défavorables nécessitent une analyse, de préférence avant qu'ils ne deviennent réalité, mais, de par leur nature même, les données historiques sur les événements extrêmes ont tendance à être rares. Depuis le milieu des années 70, cependant, certaines techniques innovantes basées sur la théorie stochastique des valeurs extrêmes "extreme value theory" (EVT) ont été conçues pour décrire et prédire des événements extrêmes avec plus ou moins de précision en utilisant seulement une quantité limitée de données.

En termes statistiques, le problème central de la gestion des risques est de savoir comment modéliser les queues et estimer les quantiles extrêmes de la distribution du processus à risque.

Désignons par $X_1, ..., X_n, ...$ les données sur un tel processus (par exemple, les rendements des pertes quotidiennes sur un titre particulier, les tremblements de terre ou les mesures de la vitesse du vent à un endroit particulier) et supposons que ces données soient indépendantes. et distribués de manière identique (i.i.d., une simplification assez grossière pour les rende-

ments boursiers) selon une certaine distribution de probabilité F. Pour l'analyse du pire des cas, nous nous intéressons alors aux niveaux x_p qui ne seront dépassés qu'avec une probabilité $p \in (0,1)$ proche de 0, c'est-à-dire $F(x_p) = 1-p$, proche de 1. En définissant la fonction quantile Q comme l'inverse généralisé de F, $Q(r) := inf\{x : F(x) \geq r\}$, un voit que les niveaux requis x_p correspondent aux quantiles $Q(1-p)$ avec une faible probabilité de dépassement p. L'estimation de quantiles aussi élevés est directement liée à la modélisation précise de la queue de la distribution $\overline{F}(x) := 1-F(x) = P(X_i > x)$ pour de grands seuils x.

Il est bien connu de l'EVT qu'un paramètre spécifique, à savoir l'indice des valeurs extrêmes (EVI), domine le comportement de la queue d'une distribution. Ce paramètre à valeur réelle indique la lourdeur de la queue, c'est-à-dire la fréquence et la gravité des événements extrêmes selon la distribution de probabilité donnée. Il existe un nombre important de publications sur les estimateurs de cet indice de valeur économique et nous continuons à en apprendre et à en comprendre davantage. Les différents estimateurs s'inspirent tous de diverses conditions (équivalentes) qui assurent la convergence de la distribution du maximum d'échantillon $X_{n,n} = max\{X_1, ..., X_n\}$ vers une distribution limite de type valeur extrême. Il est logique d'exiger cette convergence, sinon il n'y a aucun espoir qu'il y ait quelque chose de significatif à dire sur les quantiles extrêmes à la frontière, voire au-delà, de la plage d'échantillonnage. Cependant, un bon estimateur de quantile nécessite un estimateur d'EVI plus que bon, car certains estimateurs d'EVI renommés produisent de mauvais estimateurs de quantile. S'il est vrai que l'EVI détermine le comportement asymptotique des quantiles et des queues d'une distribution, il convient de souligner que des paramètres supplémentaires (par exemple, l'échelle et l'emplacement) ne sont pas moins importants pour une estimation précise des quantiles. La qualité de ces derniers dépend dans une large mesure du modèle dont est dérivé l'estimateur EVI, ainsi que des estimateurs correspondants pour les autres paramètres de ce modèle.

Ce project comporte alors trois chapitres ;

— Le chapitre 1 : Ce chapitre explore l'histoire et l'évolution de la théorie des valeurs extrêmes. Il couvre les contributions de chercheurs clés comme Fisher, Tippett, Gnedenko, De Haan, Von Mises, et Jenkinson. La théorie se concentre sur l'analyse des événements rares et extrêmes dans divers domaines, y compris la finance, les sciences naturelles, et la technologie. Elle inclut des méthodes pour comprendre la distribution des valeurs aberrantes et évaluer les risques associés

— Le chapitre 2 : Dans ce chapitre, les principales méthodes paramétriques et semi-paramétriques sont présentées. On y discute également de la méthode des excès de

seuil (POT) et d'autres techniques pertinentes pour l'analyse des valeurs extrêmes. Ces méthodes sont cruciales pour appliquer la théorie des valeurs extrêmes à des données réelles et pour modéliser les phénomènes rares de manière précise

— Le chapitre 3 : Ce chapitre se concentre sur les études expérimentales et la simulation des résultats théoriques présentés dans le deuxième chapitre. Il comprend la création d'ensembles de données simulées et l'application des résultats de la théorie des valeurs extrêmes pour démontrer leur utilité et leur pertinence dans la prévision des événements rares et la gestion des risques

Enfin, une conclusion générale résume les principales découvertes et souligne l'importance croissante de TVE pour expliquer et prédire les événements rares.

Chapitre 1

THEORIE DES VALEURS EXTREMES

L'histoire et l'évolution de la théorie des valeurs extrêmes remontent au début du XXe siècle, avec les travaux de plusieurs chercheurs, dont , Fisher et Tippett [1], Gnedenko [2], De Haan [3], et Les travaux de Von Mises [4] et Jenkinson [5].

Cette théorie s'intéresse aux événements rares et extrêmes dans divers domaines tels que la finance, les sciences naturelles et la technologie. Elle se base sur le concept des valeurs aberrantes, qui sont des valeurs très éloignées de la moyenne ou des valeurs typiques d'un phénomène.

Au fil du temps, la théorie des valeurs extrêmes s'est développée pour inclure un large éventail d'applications dans des domaines tels que l'évaluation des risques, la gestion des catastrophes, l'assurance, l'ingénierie, les sciences environnementales et la finance. Elle s'est avérée efficace pour comprendre et analyser des phénomènes inattendus caractérisés par leur rareté et leur impact important.

1.1 Les événements rares

Les événements rares sont caractérisés par leur faible probabilité d'occurrence. Lorsque ces événements se produisent de manière aléatoire, il est possible d'étudier leur distribution. Ils sont considérés comme extrêmes lorsqu'ils représentent des valeurs bien au-delà ou en deçà de ce qui est habituellement observé.

Ces événements rares captivent souvent l'actualité en raison de leur imprévisibilité. Étant donné l'importance sociale et scientifique de ces enjeux, aucun débat sérieux sur le hasard ne peut être mené sans une réflexion approfondie sur les événements rares et extrêmes.

1.2 Statistique d'ordre

Les méthodes statistiques commencent par la notion d'échantillon donné. Il est parfois nécessaire d'analyser la suite des valeurs observées, que ce soit par ordre croissant ou décroissant.

Définition 1.1 Soit $(X_1, X_2, ..., X_n)$ une suite de variables aléatoires réelles, (i.i.d), avec une fonction de répartition F, telle que :

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

La statistique d'ordre est le réarrangement croissant de $(X_1, ..., X_n)$, sous la notation $(X_{i,n})_{1 \leq i \leq n}$ avec

$$X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}.$$

$X_{i:n}$ est donc la i^{eme} statistique d'ordre (statistique d'ordre i) dans un échantillon de taille n. En particulier, la statistique du minimum et du maximum (statistiques d'ordre extrêmes) sont respectivement données par :

$$M_n = X_{n,n} = \max(X_1, X_2, ..., X_n)$$

et

$$m_n = X_{1,n} = \min(X_1, X_2, ..., X_n)$$

En général, les résultats concernant les valeurs minimales peuvent être obtenus à partir des résultats concernant les valeurs maximales en utilisant la relation suivante :

$$\min_{1 \leq i \leq n} (X_i) = - \max_{1 \leq i \leq n} (-X_i)$$

1.3 Lois de valeurs extrêmes

L'idée principale du problème est de déterminer toutes les lois de probabilité qui peuvent être exprimées comme des limites (en loi) d'une suite de n variables aléatoires indépendantes et identiquement distribuées, prenant leurs valeurs dans $\mathbb{R}$.

Considérons une variable aléatoire X qui suit une loi de probabilité $P(X)$. On définit M_n comme la valeur maximale dans un ensemble de n valeurs : $M_n = \max\{X_i\}_{1 \leq i \leq n}$. Nous nous intéressons à la distribution de cette nouvelle variable. La distribution de M_n est donnée par :

$$
\begin{aligned}
prob(M_n \leq x) &= P(X_1 \leq x, \ldots, X_n \leq x) \\
&= P(X_1 \leq x) \ldots P(X_n \leq x) \qquad ; \text{(iid)} \\
&= [P(X_i < x)]^n \\
&= P^n(x)
\end{aligned} \tag{1.1}
$$

Le problème est que la loi P n'est pas connue en pratique. Même si l'on peut trouver une distribution empirique $\hat{P}$ qui approximativement correspond à P, les erreurs s'accumulent, rendant généralement grande l'erreur d'estimation lorsqu'on remplace P^n par $\hat{P}^n$.

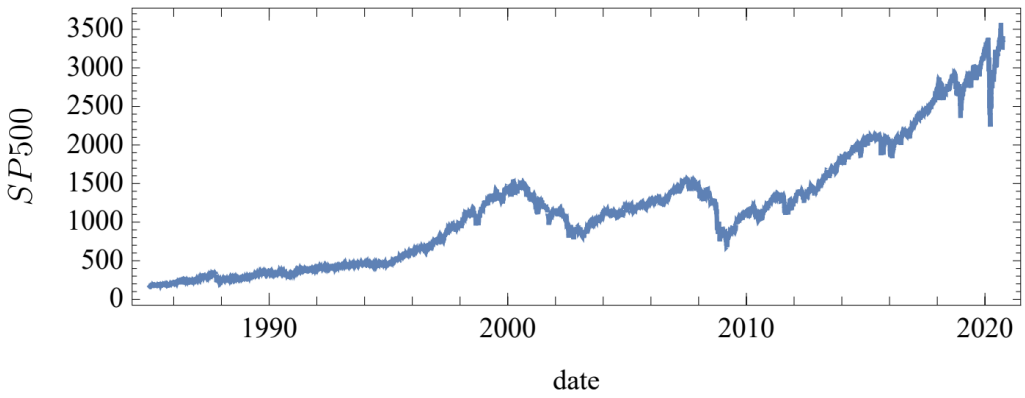


FIGURE 1.1 – Indice boursier Standard and Poor's 500 (SP500)

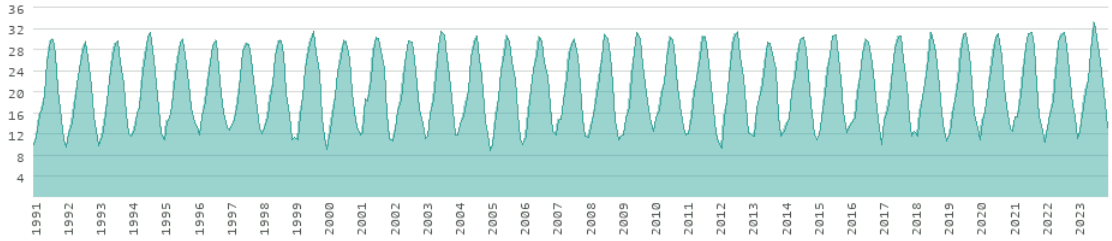


FIGURE 1.2 – Evolution des températures de 1991 à 2023 en Algerié

La théorie des valeurs extrêmes se concentre sur l'étude du comportement des valeurs extrêmes ou des valeurs maximales et minimales d'une distribution de probabilité. Son résultat principal, connu sous le nom de théorème de Fisher-Tippett-Gnedenko, établit les lois limites possibles pour le maximum (ou le minimum) d'une suite de variables aléatoires indépendantes et identiquement distribuées (v.i.i.d), même lorsque la distribution sous-jacente est inconnue. Cette théorie a des applications importantes dans des domaines tels que la gestion des risques, l'assurance, l'hydrologie, etc.

Lors de l'étude d'un phénomène aléatoire, on se concentre généralement sur la partie centrale de la distribution, qui représente le comportement typique du phénomène. Cela implique souvent le calcul de l'espérance, de la médiane, de la variance, ainsi que l'utilisation du théorème central limite et d'autres techniques statistiques. Cependant, comprendre le comportement des valeurs extrêmes est également crucial dans de nombreuses situations, telles que les risques liés aux fluctuations des marchés financiers et des actions (voir la figure 1.1 indice S&P 500), ou la prévision des sécheresses ou des inondations (la figure 1.2).

La théorie des valeurs extrêmes fournit des outils pour analyser ces valeurs et comprendre leur distribution, même lorsque les données disponibles sont limitées. En identifiant les distributions limites possibles pour les valeurs extrêmes, cette théorie permet aux chercheurs et aux praticiens d'évaluer plus précisément les risques associés aux événements extrêmes et de prendre des décisions éclairées en conséquence.

1.4 Argument asymptotique

Considérons un jeu où l'on lance un dé à six faces n fois. Si l'on obtient un 6, on gagne 2 euros, et pour tout autre résultat, on perd 1 euro. On note S_n le montant total gagné ou perdu après n lancers. Quelle est la loi de probabilité de S_n ? Intuitivement, il est clair que S_n est plus susceptible d'être proche de zéro que d'atteindre des valeurs extrêmes. Cependant, peut-on aller plus loin et déterminer la forme de la distribution de S_n ? Le théorème suivant, connu sous le nom de théorème central limite, nous permet de le faire.

1.4.1 Théorème Central limite

Soit (X_n) une suite de variables indépendantes identiquement distribuées admettant un moment d'ordre 2. On pose :

$$\mu = E[X_n], \;\; \sigma^2 = Var(X_n)$$

$$S_n = X_1 + X_2 + ... + X_n, \; Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Alors la suite (Y_n) converge en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$ (la loi normale centrée réduite). En d'autres termes, pour tout $x \in \mathbb{R}$,

$$P(Y_n \leq x) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

La force de cette théorie réside dans son exhaustivité, car elle repose sur très peu d'hy-

pothèses sur la séquence X_n. Quelle que soit la loi de probabilité de tout événement aléatoire, si cet événement se répète indépendamment et à l'infini, la moyenne finale suivra la loi naturelle. C'est ce qui fait de cette théorie une base pour comprendre que la loi naturelle régit les phénomènes naturels.

Puisque la fonction de distribution X est souvent inconnue, il devient difficile de déterminer avec précision la distribution maximale. Notre intérêt se porte donc sur la distribution approximative des valeurs extrêmes. La répartition du maximum devrait nous donner des informations sur les événements rares et extrêmes. Cette étude aide à comprendre le comportement des événements extrêmes même lorsque les détails exacts de leur distribution sont inconnus, fournissant un cadre pour travailler avec des données dans des domaines confrontés à des risques élevés et à de faibles probabilités, comme les conditions météorologiques catastrophiques ou les crises financières.

À partir de l'expression 1.1, on note a fonction de distribution de M_n par : $F_{M_n}(x) = [F(x)]^n$ et F la répartition de $X_1, X_2, ..., X_n$ iid.
Notons par $x_f = sup\{x \in \mathbb{R}; F(x) < 1\} \leq \infty$, le **point terminal** à droite (right-end point) de la fonction de répartition F. Ce point terminal peut être infini ou fini.
On s'intéresse alors à la distribution asymptotique du maximum en faisant tendre n vers l'infini, on a :

$$\lim_{x\to\infty} F_{M_n}(x) = \lim_{x\to\infty} F^n(x) = \begin{cases} 0 & \text{si } x < x_f \\ 1 & \text{si } x \geq x_f \end{cases}$$

L'idée consiste à effectuer une transformation. La plus célèbre en statistique est la normalisation, illustrée à travers l'exemple du théorème central limite. Après la normalisation, ce théorème fournit la loi asymptotique (non dégénérée) de la moyenne de n variables aléatoires.

Remareque : Pour obtenir la densité du maximum, on dérive F_{M_n}, avec les notations appropriées.

$$f_{M_n}(x) = n\, f(x)\, [F(x)]^{n-1}$$

Définition 1.2 Deux variables aléatoires réelles X et Y sont dites **de même type** si elles peuvent être transformées l'une en l'autre par une transformation **affine**. Autrement dit, il existe des constantes réelles $a > 0$ et $b \in \mathbb{R}$ telles que $Y = aX + b$.

Cela signifie que si F et H sont les lois respectives des variables Y et X, alors $F(ax+b) = H(x)$. En d'autres termes, les variables de même type partagent la même distribution, mais

elles peuvent différer par un facteur d'échelle et de translation.

De manière similaire au théorème central limite (TCL), cherchons des constantes de normalisation $a_n > 0$ et $b_n \in \mathbb{R}$, ainsi qu'une loi non dégénérée H, telles que :

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

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \left[F(a_n\, x + b_n)\right]^n \to H(x) \quad \text{lorsque} \quad n \to \infty$$

Concernant le comportement du maximum d'une suite de variables aléatoires i.i.d, le résultat le plus important est le théorème de Fisher-Tippett en 1928 donné pour des maximums normalisés.

1.4.2 Théorème $(Fisher - Tippett, Gnedenko)$

Dans la théorie classique des valeurs extrêmes (G.E.V.), l'étude se concentre sur le comportement asymptotique du maximum d'un échantillon. En analysant comment le maximum évolue à mesure que la taille de l'échantillon augmente, on peut ensuite extrapoler pour comprendre le comportement de la queue de la distribution. Cette approche est cruciale pour prévoir les événements rares et extrêmes, tels que les catastrophes naturelles, les crises financières ou d'autres phénomènes à faible probabilité mais à fort impact. Nous allons maintenant présenter le résultat principal de cette théorie, qui offre un cadre rigoureux pour ces extrapolations [6].

En approfondissant cette théorie, on découvre que la distribution des valeurs extrêmes suit souvent l'une des trois familles de distributions possibles : la distribution de Gumbel, de Fréchet ou de Weibull. Ces distributions décrivent différentes formes de queues : légère, lourde ou modérée. Par exemple, la distribution de Gumbel est souvent utilisée pour modéliser les phénomènes naturels comme les précipitations extrêmes, tandis que les distributions de Fréchet et de Weibull peuvent être appliquées à d'autres types de données extrêmes.

Le résultat principal de la théorie G.E.V. permet de déterminer laquelle de ces distributions est la plus appropriée pour modéliser les valeurs extrêmes d'un échantillon donné, en fonction de la nature des données et de leur comportement asymptotique. Ce résultat est fondamental pour les analyses de risques et pour la prise de décisions dans les domaines où les événements extrêmes ont des conséquences significatives.

Soit $(X_n)_n$ une suite de n variables aléatoires réelles i.i.d de loi F et $M_n = max(X_1, ..., Xn)$.

S'il existe deux suites réelles $a_n > 0$ et $b_n \in \mathbb{R}$ telles que ;

$$\lim_{n\to\infty} P\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n\to\infty} \left[F(a_n\, x + b_n)\right]^n = \mathrm{H}_\gamma(x), \quad \forall x \in \mathbb{R} \tag{1.2}$$

où H_γ est une fonction de distribution non dégénérée, alors H_γ est du même type que l'une des familles suivantes :

$$\text{Gumbel (type I) :} \qquad \Lambda(x) = \exp\big(-\exp(-x)\big) \qquad \text{si } x > 0 \qquad ;$$

$$\text{Fréchet (type II) :} \qquad \Phi_\gamma(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \exp\big(-(x)^{-\gamma}\big) & \text{si } x > 0, \gamma > 0 \end{cases}$$

$$\text{Weibull (type III) :} \qquad \Psi_\gamma(x) = \begin{cases} 1 & \text{si } x > 0 \\ \exp\big(-(-x)^{-\gamma}\big) & \text{si } x \leq 0, \gamma < 0 \end{cases}$$

Les trois types Λ , Φ_γ et Ψ_γ de distribution extrêmes s'appellent les **lois standards** ou traditionnelles des **des distribution des valeurs extrêmes** et γ est le paramètre de forme appelé **indice des valeurs extrêmes** ou **indice de queue**.

Ce théorème est un résultat important car il ne nécessite pas de faire d'hypothèses paramétriques sur la loi des X_i. La valeur de γ détermine le comportement de la queue de la distribution.

Relation entre Λ , Φ_γ et Ψ_γ :

Soit Y une variable aléatoire positive ($Y > 0$), alors les énoncés suivants sont équivalents :

1. $Y \sim \Phi_\gamma$;
2. $ln\big(Y^\gamma\big) \sim \Lambda$;
3. $-Y^{-1} \sim \Psi_\gamma$.

Remarques :

— Les théorèmes 1.2.1 et 1.2.2 revêtent un intérêt important car, si l'ensemble des distributions est grand, celui des distributions des valeurs extrêmes est très petit.

— La fonction de répartition H_γ est connue sous le nom de loi des valeurs extrêmes (EVD, ou "Extreme Value Distribution"). Le paramètre γ est un paramètre de forme, également désigné sous le nom d'indice des valeurs extrêmes ou indice de queue. a_n est un paramètre de localisation et b_n est un paramètre d'échelle.

— Les séquences de normalisation (a_n) et (b_n) ne sont pas uniques.

— Gnedenko a établi des conditions nécessaires et suffisantes pour l'existence de constantes de normalisation lorsque la loi de la variable X est connue. Ces conditions peuvent être utilisées pour identifier le type de la loi limite.

— Effectivement, plus l'indice γ est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On utilise alors le terme "queues épaisses" pour décrire de telles distributions.

1.4.3 Distribution des valeurs extrêmes généralisée

Pour simplifier l'utilisation des trois distributions limites des valeurs extrêmes, Jenkinson et Von Mises ont proposé une nouvelle représentation en introduisant des paramètres supplémentaires : le paramètre de localisation μ et le paramètre de dispersion σ. Ces paramètres permettent de reparamétrer les distributions des valeurs extrêmes, rendant leur application plus flexible et intuitive.

Ainsi, la forme la plus générale de la distribution des valeurs extrêmes, connue sous le nom de **GEVD** (Generalized Extreme Value Distribution), est obtenue. Cette distribution généralise et unifie les trois types de distributions limites identifiées dans le théorème 1.4.2 : les distributions de Gumbel, Fréchet et Weibull. En intégrant les paramètres μ et σ, la GEVD permet de mieux ajuster les modèles aux données empiriques, offrant ainsi une plus grande précision dans l'analyse des phénomènes extrêmes.

La paramétrisation proposée par Jenkinson et Von Mises facilite la modélisation des données dans divers domaines, tels que les sciences de l'environnement pour l'étude des événements météorologiques extrêmes, la finance pour l'évaluation des risques de marché, et l'ingénierie pour la prévision des charges maximales sur les structures. Grâce à cette approche, il devient possible d'adapter les modèles aux spécificités des données observées, améliorant ainsi la capacité à prévoir et à gérer les risques associés aux événements rares mais significatifs.

$$H_{\mu,\sigma,\gamma}(x) = \begin{cases} \exp\left\{-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right\} & \text{si } \gamma \neq 0,\ \ 1+\gamma\frac{x-\mu}{\sigma} > 0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \text{si } \gamma = 0,\ \ x \in \mathbb{R} \end{cases} \tag{1.3}$$

avec $\mu \in \mathbb{R}$ et $\sigma > 0$.
On peut facilement montrer que la fonction densité correspondante à $H_{\mu,\sigma,\gamma}$ est donnée par :

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

En remplaçant $\frac{x-\mu}{\sigma}$ par x, on obtient la forme standard de la **GEVD** :

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

où γ est le paramètre de forme.

La fonction de densité de probabilité h_γ associée et définie par :

$$h_\gamma(x) = \begin{cases} H_\gamma(x)\ \left(1+\gamma x\right)^{\frac{-1}{\gamma}-1} & \text{si } \gamma \neq 0,\ \ 1+\gamma x > 0 \\ \exp\left(x - \exp(-x)\right) & \text{si } \gamma = 0,\ \ x \in \mathbb{R} \end{cases}$$

La figure 1.3 ci-dessous illustre le comportement de **GEVD** standard, et La figure 1.4 illustre les densités standard de **GEV**.

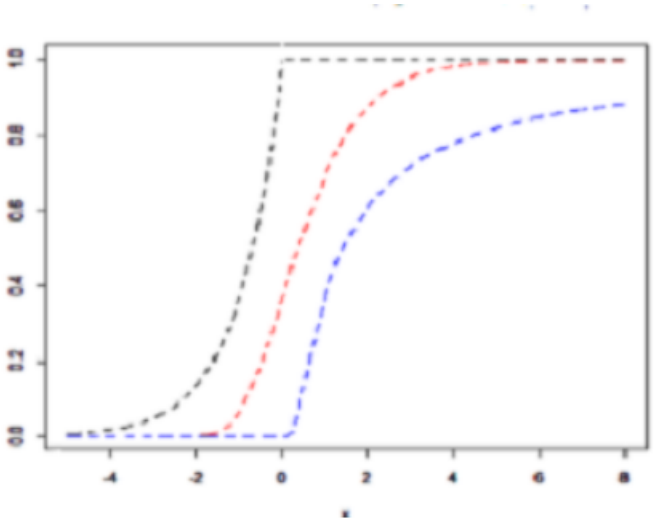


FIGURE 1.3 – Exemple : distributions standards des valeurs extrêmes (le bleu pour Frechet, le rouge pour Gumbel et le noir pour Weibull).

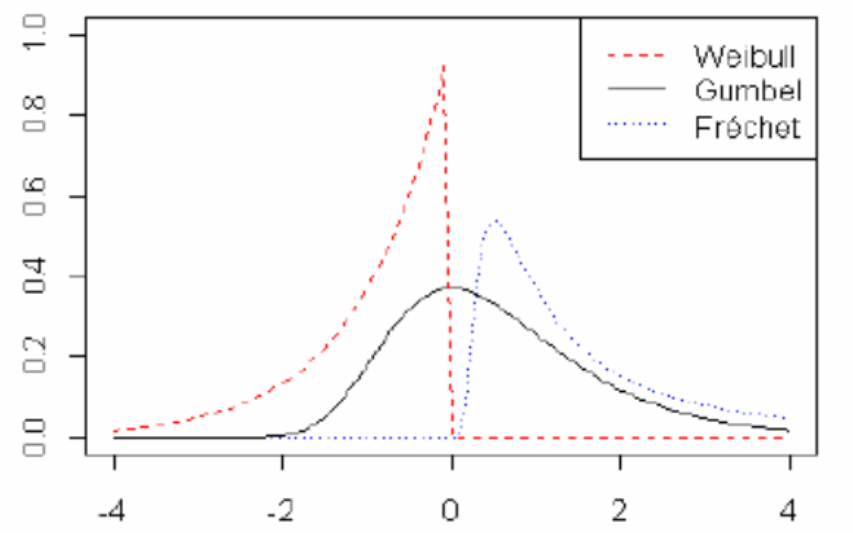


FIGURE 1.4 – Exemple : densités standard des valeurs extrêmes.

Quand $\gamma = 0$, nous avons $x \in \mathbb{R}$. Pour $\gamma = 0$, nous obtenons $H_0(x) = \exp[-\exp(-x)]$, où $x \in \mathbb{R}$. Cette expression est obtenue dans H en faisant tendre γ vers 0. Les lois des valeurs extrêmes généralisées correspondent, à une translation et à un changement d'échelle près, aux lois des valeurs.
Nous avons alors les correspondances suivantes :

$$\Lambda(x) = H_0(x) \qquad \text{si } x \in \mathbb{R},$$

$$\Phi_\gamma(x) = H_{\frac{1}{\gamma}}\big((x-1)\gamma\big) \qquad \text{si } x > 0,$$

$$\Psi_\gamma(x) = H_{\frac{-1}{\gamma}}\big((x+1)\gamma\big) \qquad \text{si } x < 0.$$

1.4.4 Domaine d'attraction

Le concept de **domaine d'attraction** est essentiel dans la théorie des valeurs extrêmes. Il se réfère à l'ensemble des lois de probabilité pour lesquelles la distribution normalisée des valeurs extrêmes converge vers une distribution limite spécifique. En d'autres termes, une distribution donnée appartient au domaine d'attraction d'une certaine loi limite des valeurs extrêmes si, après une normalisation appropriée, les valeurs extrêmes de la distribution convergent vers cette loi limite.

Dans cette section, nous examinerons les conditions nécessaires et suffisantes que doit remplir une fonction de distribution F afin que la loi du maximum converge vers H_γ.

Définition 1.3 On dit qu'une distribution appartient au **domaine d'attraction** de H_γ, et on note $F \in D(H_\gamma)$, Si deux suites normalisables $(a_n)_{n>1} > 0$ et $(b_n)_{n>1} \in \mathbb{R}$ peuvent être trouvées de manière à satisfaire la condition (1.2), Autrement dit, si

$$\exists\, a_n > 0,\ \exists\, b_n \in \mathbb{R}, \quad \lim_{n\to\infty} F(a_n\, x + b_n) = H_\gamma(x)$$

Selon le signe de γ, on distingue trois domaines d'attraction :

— Lorsque le paramètre $\gamma > 0$, la fonction appartient au domaine d'attraction Φ_γ. Dans ce cas, le point terminal $x_F = \infty$ et la fonction est de type Pareto, ce qui signifie que le coefficient de Pareto représente une valeur élevée pour les valeurs extrêmes avec une décroissance graduelle.

Exemple 1.1 Pour domaine d'attraction $\gamma = 1/\alpha > 0$:

Distribution	**Fonction de Répartition**	**Fonction de Densité**
Pareto$(\alpha), \alpha > 0$	$1 - \frac{1}{x^\alpha}$	$\frac{\alpha}{x^{\alpha+1}}$
Fréchet$(\frac{1}{\alpha}), \alpha > 0$	$\exp\left(-x^{-\alpha}\right)$	$\frac{\alpha}{x^{\alpha+1}} \exp\left(-x^{-\alpha}\right)$
Log-logistique$(\beta, \alpha), \beta > 0, \alpha > 1$	$1 - \left(1 + \beta x^\alpha\right)^{-1}$	$\alpha\beta\, \frac{x^{\alpha-1}}{\left(1+\beta x^\alpha\right)^2}$

TABLE 1.1 – Quelques distributions associées à un indice positif.

— Lorsque le paramètre $\gamma < 0$, la fonction appartient au domaine d'attraction Ψ_γ. Dans cette situation, le point terminal $x_F < \infty$ et la fonction décroît rapidement à mesure que l'on s'éloigne des valeurs extrêmes, reflétant ainsi une décroissance rapide des valeurs.

Exemple 1.2 Pour domaine d'attraction $\gamma < 0$:

— Lorsque le paramètre $\gamma = 0$, la fonction appartient au domaine d'attraction Λ_γ. Dans ce cas, le point terminal x_F peut ne pas être défini. Ces fonctions sont connues pour avoir des queues légères, ce qui signifie que la décroissance des valeurs est progressive et lente.

Distribution	Répartition	Fonction de Densité	γ
Uniforme(0, 1)	x	$\frac{\alpha}{x^{\alpha+1}}$	-1
Burr inversée(β,τ,λ,x_F), $\beta,\tau,\lambda>0$	1-$\left(\frac{\beta}{\beta+(x_F-x)^{-\tau}}\right)^{\lambda}$	$\frac{(\tau\lambda\beta^{\lambda})(x_F-x)^{-\tau-1}}{\left(\beta+(x_F-x)^{-\tau}\right)^{\lambda+1}}$	-$\frac{1}{\lambda\tau}$

TABLE 1.2 – Quelques distributions associées à un indice négatif.

Distribution	Répartition	Fonction de Densité
Gamma(m,λ), $m\in\mathbb{N},\lambda>0$	$1-\frac{\lambda^m}{\Gamma(m)}\int_0^x u^{m-1}\exp\big(-\lambda u\big)du$	$\frac{\lambda^m}{\Gamma(m)}x^{m-1}\exp(-\lambda x)$
Gumbel(μ,β), $\mu\in\mathbb{R},\beta>0$	$\exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$	$\frac{1}{\beta}\exp\left(-\frac{x-\mu}{\beta}\right)\exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$
Weibull(λ,τ), $\lambda>0,\tau>0$	$1-\exp\big(-(\lambda x)^{\tau}\big)$	$\tau\lambda(\lambda x)^{\tau-1}\exp\big(-(\lambda x)^{\tau}\big)$

TABLE 1.3 – Quelques distributions associées à un indice nul.

Exemple 1.3 Pour domaine d'attraction $\gamma=0$:

1.4.5 Les fonctions à variations régulières

Dans cette section, nous allons définir quelques notions fondamentales de la théorie des valeurs extrêmes qui permettent de caractériser les domaines d'attraction. Ces fonctions à variations régulières sont essentielles car elles fournissent une écriture unique pour chaque domaine d'attraction.

Définition 1.4 Une fonction h positive à l'infini est définie comme ayant une **variation régulière** à l'infini si elle satisfait à une condition spécifique. Cette condition stipule que lorsque le paramètre t tend vers l'infini, le rapport entre la fonction évaluée en tx et celle évaluée en t converge vers x^{γ}, pour tout $x>0$. Mathématiquement, cela se formule comme

suit :

$$\lim_{t\to\infty}\frac{h(tx)}{h(t)} = x^{\gamma} \quad \text{pour } x > 0$$

Nous désignons l'**ensemble** des fonctions à variation régulière d'indice γ par RV_{γ}.

Définition 1.5 Une fonction $L : \mathbb{R} \to [0, +\infty[$ est définie comme ayant une variation régulière d'indice 0, notée $L \in RV_0$, si elle satisfait à une condition particulière. Cette condition stipule que lorsque le paramètre t tend vers l'infini, le rapport entre la fonction évaluée en tx et celle évaluée en t converge vers 1, pour tout $x > 0$. Cela se formule comme suit :

$$\lim_{t\to\infty}\frac{L(tx)}{L(t)} = 1 \quad \text{pour } x > 0$$

Dans le contexte de la théorie des valeurs extrêmes, une fonction qui satisfait cette condition est appelée une **fonction à variation lente**.

Exemple 1.4 La fonction logarithme est connue pour être une fonction à variation lente. En effet $f(x) = \log(x), \quad x > 0$;

$$\begin{aligned}\lim_{t\to+\infty}\frac{\log(tx)}{\log(t)} &= \lim_{t\to+\infty}\frac{\log(t)+\log(x)}{\log(t)} \\ &= \lim_{t\to+\infty}\left(1+\frac{\log(x)}{\log(t)}\right) \\ &= 1\end{aligned}$$

Proposition 1.1 Si la fonction $h \in RV_{\gamma}$, alors peut toujours s'écrire sous la forme :

$$h(x) = x^{\gamma}L(x)$$

avec $L \in RV_0$.

Proposition 1.2 (Représentation de Karamata) Stipule que pour toute fonction $h \in RV_{\gamma}$, où $\gamma \in \mathbb{R}$, il existe deux fonctions mesurables $c > 0$ et ψ telles que :

$$\lim_{x\to+\infty} c(x) = c_0 \in]0, +\infty[$$

et

$$\lim_{x\to+\infty} \psi(x) = p$$

pour $x_0 > 0$ fixé, on a pour tout $x \geq x_0$:

$$L(x) = c(x) \int_{x_0}^{x} \frac{\psi(u)}{u} du \tag{1.4}$$

Dans le cas où la fonction c est constante, la fonction L correspondante est dite **normalisée**.

1.4.6 Caractérisation des domaines d'attraction

La caractérisation des domaines d'attraction est un concept utilisé en analyse et en théorie des systèmes dynamiques pour étudier le comportement asymptotique des solutions d'un système dynamique. Il existe plusieurs approches pour caractériser les domaines d'attraction en fonction des propriétés du système dynamique.

Proposition 1.3 Une distribution F appartient au domaine d'attraction de H_γ, noté $F \in D(H_\gamma)$, si et seulement si, pour une certaine suite $(a_n) > 0$ et $b_n \in \mathbb{R}$, on a

$$\lim_{n\to\infty} n\bar{F}(a_n x + b_n) = \log\left(H_\gamma(x)\right), \quad x \in \mathbb{R}$$

où $\bar{F}$ est la fonction de survie (également appelée queue de distribution), définie par :

$$\bar{F} = 1 - F(x)$$

En pratique, il est souvent plus commode de travailler avec la fonction quantile de queue plutôt qu'avec la fonction F elle-même.

Définition 1.6 On appelle l'**inverse généralisée** (fonction des quantiles) de la fonction F, l'application notée $F^{\leftarrow}$, définie par :

$$Q(p) = F^{\leftarrow}(p) = \inf\{x : F(x) \geq p\}, \quad p \in [0,1]$$

Définition 1.7 On appelle fonction **quantile de queue** de la distribution F, la fonction $U :]1, +\infty[\to \mathbb{R}$, définie par :

$$U(t) = Q\left(1 - \frac{1}{t}\right) = F^{\leftarrow}\left(1 - \frac{1}{t}\right), \quad \text{pour tout } t > 1$$

où Q est la fonction des quantiles associée à F.

Voici les critères les plus utilisés pour déterminer si une fonction de répartition F appartient à l'un des trois domaines d'attraction définis précédemment.

Théorème 1.1 $\big($**Caractérisation du** $D(H_\gamma)\big)$

Pour $\gamma \in \mathbb{R}$, les affirmations suivantes sont équivalentes :

1. F appartient à $D(H_\gamma)$
2. Pour $x, y > 0$ et $y \neq 1$, on a :

$$\lim_{s\to 1} \frac{U(sx)}{U(s)} = \begin{cases} \dfrac{x^\gamma - 1}{y^\gamma - 1} & \text{si } \gamma \neq 0 \\ \dfrac{\log\, x}{\log\, y} & \text{si } \gamma = 0 \end{cases} \tag{1.5}$$

3. Pour une certaine fonction positive b, on a :

$$\lim_{t\to x_F} \frac{\bar{F}\big(t + x\, b(t)\,\big)}{\bar{F}(t)} = \begin{cases} \Big(1 + \gamma x\Big)^{-\frac{1}{\gamma}} & \text{si } \gamma \neq 0 \\ \exp(-x) & \text{si } \gamma = 0 \end{cases}$$

avec $(1 + \gamma x) > 0$.

Démonstration Voir Embrechts al. [7].

Théorème 1.2 $\big($**Caractérisation du** $D(\Phi_\gamma)\,\big)$

La fonction de répartition F appartient au domaine d'attraction de la loi de Fréchet de paramètre $\gamma > 0$ si et seulement si $x_F = +\infty$ et

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

où L est une fonction à variation lente. Dans ce cas, on choisit $a_n = U(n) = F^{\leftarrow}\big(1 - \frac{1}{n}\big)$ et $b_n = 0$ pour tout $n > 0$. La suite $(a_n^{-1} X_{n,n})_{n\geq 1}$ converge en loi vers une variable aléatoire de fonction de répartition Φ_γ.

Exemple 1.5

Dans le cas où X suit une loi de Pareto (Pareto(γ)), les coefficients de normalisation sont définis comme suit :

$$a_n = n^\gamma \;\text{ et }\; b_n = 0.$$

Considérons $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires i.i.d. suivant une loi de Pareto de paramètre $\gamma > 0$ avec la fonction de répartition $F(x) = 1 - \frac{1}{x^{\frac{1}{\gamma}}}$.

En choisissant $a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right) = n^{\gamma}$ et $b_n = 0$, on a :

$$\begin{aligned}
\lim_{n\to\infty} F^n(a_n x + b_n) &= \lim_{n\to\infty} P\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) \\
&= \lim_{n\to\infty} \left(1 - \frac{-x^{\frac{-1}{\gamma}}}{n}\right)^n \\
&= \exp\left(-x^{\frac{-1}{\gamma}}\right) \\
&= \Phi_\gamma(x)
\end{aligned}$$

Ainsi, la loi limite est une loi de Fréchet de paramètre γ.

Toutes les distributions appartenant au domaine d'attraction, dites de type Pareto, ont une queue de distribution $\bar{F}$ qui peut être écrite, pour x très grand, comme suit :

$$\bar{F}(x) = Cx^{-\gamma}, \quad C, \gamma > 0$$

Théorème 1.3 (Caractérisation du $D(\Psi_\gamma)$)

La fonction de répartition F appartient au domaine d'attraction de la loi de Weibull de paramètre $\gamma > 0$ si et seulement si $x_F < +\infty$ et

$$\bar{F}(x_F - x^{-1}) = x^{-\gamma} L(x),$$

où L est une fonction à variation lente.
Dans ce cas, on choisit $a_n = x_F - U(n) = x_F - F^{\leftarrow}\left(1 - \frac{1}{n}\right)$ et $b_n = x_F$ pour tout $n > 0$. Alors $(a_n^{-1}(X_{n,n} - x_F))_{n\geq 1}$ converge en loi vers une variable aléatoire de fonction de répartition Ψ_γ.

Exemple 1.6

Pour une suite de variables aléatoires i.i.d. $X_1, X_2, \ldots, X_n$ suivant une loi uniforme continue sur $[0,1]$, les coefficients de normalisation sont donnés par : $a_n = \frac{1}{n}$ et $b_n = 1$. Cela signifie que pour chaque n, vous divisez par n pour standardiser la variable aléatoire, et b_n est simplement

une constante égale à 1 dans ce cas.

$$\lim_{n\to\infty} F_n(anx+bn) = \lim_{n\to\infty}\left(1+\frac{1}{n}x\right)^n = e^x$$

Donc la loi limite est une loi Weibull $F \in D(\Psi_1)$.

Définition 1.6 (Fonction de Von Mises) Soit F une fonction de répartition avec un point terminal x_F. Supposons qu'il existe $z < x < x_F$ tels que :

$$\bar{F}(x) = c\ \exp\left(-\int_z^x \frac{1}{a(t)}dt\right)$$

où $c > 0$ et a est une fonction absolument continue positive avec la densité vérifiant $\lim_{x\to x_F} a(x) = 0$. Dans ce cas, F est appelée une fonction de **von-mises** et a est une **fonction auxiliaire**.

Théorème 1.3 $\left(\text{Caractérisation du } D(\Lambda_\gamma)\right)$

La fonction de répartition F appartient au domaine d'attraction de la loi de Gumbel avec le point terminal $x_F \leq +\infty$ si et seulement s'il existe $z < x_F$ tel que :

$$\bar{F}(x) = c(x)\ \exp\left(-\int_z^{x_F} \frac{g(t)}{a(t)}dt\right)$$

où c et g sont deux fonctions mesurables telles que $c(x) \to c > 0$ et $g(x) \to 1$ quand $x \to x_F$, et a est une fonction auxiliaire. Dans ce cas, on peut choisir les constantes de normalisation $b_n = \bar{F}\left(1-\frac{1}{n}\right)$ et $a_n = a(b_n)$. Une option possible pour la fonction auxiliaire a est donnée par :

$$a(x) = \int_x^{x_F} \frac{\bar{F}(t)}{\bar{F}(x)}dt; \quad x < x_F$$

Exemple 1.7

La distribution exponentielle de paramètre λ, notée $H(x)$, est définie comme suit :

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

Dans cette définition, λ est un paramètre positif.

Puisque $H(x)$ est une fonction de Von-Mises, cela signifie que $x_F = +\infty$ et que $\forall x \in]0, +\infty[$, $\bar{H}(x) = 1 - H(x)$. En d'autres termes, $\bar{H}(x)$ est le complément de $H(x)$ et peut être exprimé comme $e^{-\lambda x}$.

De plus, il semble que vous ayez une équation pour $\bar{H}(x)$ sous forme intégrale :

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

D'après ce que vous avez dit, la fonction auxiliaire $a(x)$ est définie comme $a(x) = \lambda^{-1}$. Alors $H \in D(\Lambda)$ et

$$a_n = \frac{1}{\lambda} \quad \text{et} \quad b_n = \lambda^{-1} \ln(n).$$

Dans ce tableau (1.4) illustrant différents types de lois statistiques classées en fonction de leurs domaines d'attraction.

TABLE 1.4 – Exemple sur les lois classées selon leurs domaines d'attraction

Domainne d'attraction	**Gumbel** ($\gamma = 0$)	**Fréchet** ($\gamma > 0$)	**Weibull** ($\gamma < 0$)
Lois	Exponentielle Normale Log-normale	Gamma Pareto Student Burr Uniforme	Weibull Cauchy Log-gamma Beta Inverse de Burr Inverse de Pareto

Dans ce tableau :

— Le domaine d'attraction est classé en fonction du paramètre de forme γ de la loi de distribution.

— Les lois statistiques sont répertoriées sous chaque domaine d'attraction.

— Par exemple, dans le domaine d'attraction de Gumbel ($\gamma = 0$), les lois comme l'exponentielle, la normale et la log-normale sont répertoriées.

— De même, pour les domaines d'attraction Fréchet ($\gamma > 0$) et Weibull ($\gamma < 0$), différentes lois statistiques sont répertoriées.

1.5 Conditions de Von Mises

Les conditions de Von Mises jouent un rôle crucial dans la théorie des valeurs extrêmes, en particulier pour déterminer si une distribution donnée appartient au domaine d'attraction de la distribution de Gumbel. Essentiellement, elles stipulent que pour qu'une distribution F soit dans ce domaine d'attraction, la fonction quantile $U(y) = F^{-1}(1 - \frac{1}{y})$ doit obéir à une certaine régularité asymptotique.

En d'autres termes, lorsque la taille de l'échantillon augmente, la différence entre les valeurs quantiles normalisées doit converger de manière logarithmique. Cela permet d'assurer que les valeurs extrêmes suivent une distribution de Gumbel après une normalisation adéquate. Ces conditions sont non seulement théoriquement significatives, mais elles ont également des applications pratiques dans des domaines variés comme la météorologie pour modéliser les événements climatiques extrêmes, la finance pour évaluer les risques de marché, et l'ingénierie pour prévoir les charges maximales sur les structures.

En résumé, les conditions de Von Mises fournissent un cadre rigoureux et utile pour l'analyse des phénomènes extrêmes en garantissant la validité de l'utilisation des distributions de Gumbel dans divers contextes pratiques.

Théorème 1.4 (Conditions de Von Mises)

Les conditions suffisantes pour qu'une fonction de distribution F absolument continue dans $]x_1, x_F[$ appartienne à l'un des trois domaines d'attraction sont les suivantes :

1. Si la densité f admet une dérivée f' négative pour tout x dans l'intervalle $]x_1, x_F[$, avec $f(x) = 0$ pour $x \geq x_F$, et si

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

alors F appartient au domaine d'attraction $D(\Lambda)$.

2. Si $f(x) > 0$ pour tout x dans l'intervalle $]x_1, \infty[$, et pour $\gamma > 0$,

$$\lim_{t \to \infty} \frac{t f(t)}{1 - F(t)} = \gamma,$$

alors F appartient au domaine d'attraction $D(\Phi_\gamma)$.

3. Si $f(x) > 0$ pour tout x dans l'intervalle $]x_1, x_F[$, $f(x) = 0$ pour tout $x > x_F$, et pour $\gamma > 0$,

$$\lim_{t \to x_F^+} \frac{F(x_F - t) f(t)}{1 - F(t)} = \gamma,$$

alors F appartient au domaine d'attraction $D(\Psi_\gamma)$.

Lorsque ces conditions sont vérifiées, on peut dire que F appartient au domaine d'attraction correspondant à la loi extrême considérée. Ces conditions fournissent un moyen pratique de vérifier l'appartenance d'une fonction de répartition à un domaine d'attraction donné, mais elles ne sont applicables que dans le cas où la fonction de répartition possède une densité.

Il est important de souligner que ces conditions suffisantes simplifiées sont un outil précieux pour l'analyse statistique, mais elles sont basées sur des résultats théoriques qui peuvent ne pas être évidents à première vue. Par conséquent, leur application nécessite une compréhension appropriée de la théorie des lois de probabilité extrêmes et de leurs domaines d'attraction.

1.6 Distribution de Pareto généralisée

L'approche des maximums par blocs, bien que largement utilisée dans la modélisation des événements extrêmes, a été critiquée pour sa perte potentielle d'information. Cela est dû au fait qu'en se concentrant uniquement sur le maximum de chaque bloc, les autres valeurs extrêmes de l'échantillon peuvent être négligées, ce qui entraîne une sous-estimation de la variabilité des événements extrêmes.

Pour remédier à ce problème, Pickands (1975) [8] a introduit une nouvelle approche dans la théorie des valeurs extrêmes connue sous le nom de méthode POT (Peaks Over Threshold), également appelée approche des excès au-delà d'un seuil. Au lieu de se concentrer uniquement sur le maximum de chaque bloc, cette méthode consiste à observer toutes les valeurs qui dépassent un certain seuil préalablement fixé. En d'autres termes, elle s'intéresse aux excès par rapport à un seuil élevé.

Cette méthode a été développée par divers auteurs, notamment Smith (1987) [9] et Reiss et Thomas (2007) [10]. En utilisant la méthode POT, on peut obtenir une estimation plus précise des événements extrêmes en tenant compte de toutes les valeurs excédentaires au lieu de se limiter uniquement au maximum de chaque bloc. Cela permet une meilleure caractérisation des queues de distribution, ce qui est crucial dans de nombreux domaines tels que la modélisation des risques financiers, la prévision des crues et la gestion des catastrophes.

La GP (Generalized Pareto) est une distribution statistique couramment utilisée pour modéliser les queues de distributions, c'est-à-dire les valeurs dépassant un certain seuil. Pour

être plus précis, on définit un seuil u suffisamment élevé. On compte alors le nombre de dépassements de ce seuil parmi les observations, ce nombre étant noté $N_u = \text{card}\{i : i = 1,\ldots,n, X_i > u\}$, où X_i est la i-ème observation. Les valeurs excédant ce seuil sont notées $Y_i = X_i - u > 0$ pour $1 \leq i \leq N_u$.

En d'autres termes, N_u représente le nombre d'excès au-dessus du seuil u, et $Y_1,\ldots,Y_{N_u}$ sont les valeurs excédentaires correspondantes.

La distribution GP est souvent utilisée pour modéliser ces excès au-dessus d'un seuil car elle fournit une bonne approximation pour la distribution des valeurs extrêmes. Elle est largement utilisée dans les domaines de l'analyse des risques, de l'assurance, de l'ingénierie des fiabilités et de la finance pour modéliser les événements rares et extrêmes. En ajustant cette distribution aux données excédentaires, on peut estimer les paramètres de la distribution et ainsi caractériser les queues de la distribution de manière plus précise.

Définition 1.7 $\big($**Distribution des excès et moyenne des excès**$\big)$ La fonction de répartition des excès de X au-dessus du seuil u, notée $F_u(y)$, est définie par :

$$F_u(y) = \mathrm{P}(X - u \leq y \mid X > u)$$

De plus, cette fonction peut être réécrite en termes de la fonction de répartition cumulative de X, $F(x)$, comme suit :

$$\begin{aligned} F_u(y) &= \frac{\mathrm{P}(X \leq y+u\ ,\ X > u)}{\mathrm{P}(X > u)} \\ &= \frac{\mathrm{P}(u < X \leq y+u)}{1-\mathrm{P}(X \leq u)} \\ &= \frac{\mathrm{P}(X \leq y+u) - \mathrm{P}(u < X)}{1-\mathrm{P}(X \leq u)} \\ &= \frac{\mathrm{F}(y+u) - \mathrm{F}(u)}{1-\mathrm{F}(u)} \end{aligned}$$

En simplifiant cette expression, nous obtenons :

$$F_u(y) = \frac{\mathrm{F}(y+u) - \mathrm{F}(u)}{\bar{\mathrm{F}}(u)}$$

En effet, si X est intégrable, la fonction de moyenne des excès de X, notée $e_u(x)$, cela s'écrit comme suit :

$$e_u(x) = \mathrm{E}(X-u\,|\,X>u) = \int_u^{\infty} x\,dF_u(y), \quad x < x_F$$

où $dF_u(y)$ est la densité de probabilité des excès de X au-dessus du seuil u. En utilisant la définition de la fonction de répartition des excès $F_u(y)$, on peut réécrire $e_u(x)$ comme suit :

$$e_u(x) = \frac{1}{\bar{F}(u)} \int_u^{\infty} \bar{F}(t)\,dt, \quad x < x_F$$

1.6.1 GPD

La loi de Pareto généralisée notée **GPD** (generalized Pareto distribution), de paramètres $\sigma > 0, \gamma \in \mathbb{R}$ est défini par la fonction de répartition $G_{\gamma,\mu,\sigma}$:

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

De plus, les conditions sur les valeurs de x dépendent de la valeur de γ :

- Si $\gamma \geq 0$, alors x doit être supérieur ou égal à zéro.
- Si $\gamma < 0$, alors x doit être compris entre μ et $\mu - \sigma/\gamma$.

On note que le **GPD** standard est correspond au cas $\mu = 0$ et $\sigma = 1$.

Le **GPD** avec les paramètres $\mu = 0$ et $\sigma > 0$ joue un rôle important, dans l'analyse statistique des événements extrêmes, en fournissant une approximation appropriée pour l'excès d'un grand seuil. Cette famille spéciale sera dénotée par $G_{\gamma,\sigma}$ et définie comme suit :

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

où x appartient à $S(\gamma,\sigma)$, le support de $G_{\gamma,\sigma}$, défini comme suit :

$$S(\gamma,\sigma) = \begin{cases} [0; +\infty[& \text{si } \gamma \geq 0 \\ [0; \sigma/(-\gamma)[& \text{si } \gamma < 0 \end{cases}$$

Lorsque le paramètre de localisation est nul ($\mu = 0$) et que le paramètre d'échelle est quelconque ($\sigma > 0$), Belkema et de Haan [11], ainsi que Pickands [8], ont proposé un théorème fondamental. Ce théorème établit un lien entre le comportement asymptotique de la distribution des excès au-dessus d'un seuil élevé et la loi de Pareto généralisée (GPD). Plus précisément, il montre que, sous certaines conditions, les excès normalisés au-dessus d'un

seuil convergent vers une GPD à mesure que le seuil augmente. Cette connexion est essentielle pour l'analyse des valeurs extrêmes, car elle permet d'utiliser la GPD pour modéliser et estimer les probabilités d'événements extrêmes.

1.6.2 Théorème $\left(\text{Balkema et de Haan(1974), Pickands(1975)}\right)$

Théorème 1.5 Si la fonction de répartition F appartient au domaine d'attraction de la distribution de Pareto généralisée G_γ, alors il existe une fonction mesurable et positive $\sigma(u)$, telle que :

$$\lim_{u \to x_F} \sup_{0<y<x_F} \left| F_u(y) - G_{\gamma,\sigma(u)}(y) \right| = 0 \tag{1.6}$$

L'interprétation de ce théorème est la suivante : lorsqu'une fonction de répartition F appartient au domaine d'attraction de la distribution de Pareto généralisée G, cela signifie que les dépassements au-dessus d'un certain seuil peuvent être modélisés asymptotiquement par la distribution de Pareto généralisée.

Préuve. La preuve de ce théorème doit être trouvée dans Embrecht et al [7]

Remarque :

La densité de la distribution de **GPD** est souvent exprimée comme suit :

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

où γ est le paramètre de forme, σ est le paramètre d'échelle, et μ est le paramètre de localisation, mais comme mentionné précédemment, $\mu = 0$ pour la densité standard.

Quant à la relation entre la distribution de Pareto généralisée standard G_γ et la distribution GEV (Generalized Extreme Value) standard H_γ, elle peut être exprimée comme suit :

$$G_\gamma(x) = 1 + \log\left(H_\gamma(x)\right), \quad \log\left(H_\gamma(x)\right) > -1$$

où γ est le paramètre de forme de la distribution GEV.

Cette relation montre que $G_\gamma(x)$ et $H_\gamma(x)$ sont liés de manière directe et simple.

Exemple 1.8

Pour une loi exponentielle standard, avec une fonction de répartition $F(x) = 1 - e^{-x}$ pour $x > 0$, supposons $u > 0$:

$$F_u(y) = \frac{e^{-(u)} - e^{-(u+y)}}{e^{-(u)}}$$

$$= 1 - e^{-y}$$

Ceci correspond effectivement à $\gamma = 0$ et $\sigma = 1$, car la distribution exponentielle standard est un cas spécial de la distribution de Pareto généralisée.

Exemple 1.9

Pour la loi de Fréchet standard, avec une fonction de répartition $F(x) = \exp\left(\frac{1}{x}\right)$ pour $x > 0$, supposons $u > 0$:

$$F_u(y) = \frac{1 - \exp\left(\frac{-1}{u+y}\right)}{1 - \exp\left(\frac{-1}{u}\right)}$$

$$= 1 - \left(1 + \frac{x}{u}\right)^{-1}$$

Ceci correspond effectivement à $\gamma = 1$ et $\sigma(u) = u$.

La figure 1.5 illustre la distribution et la densité de la loi de GPD standard.

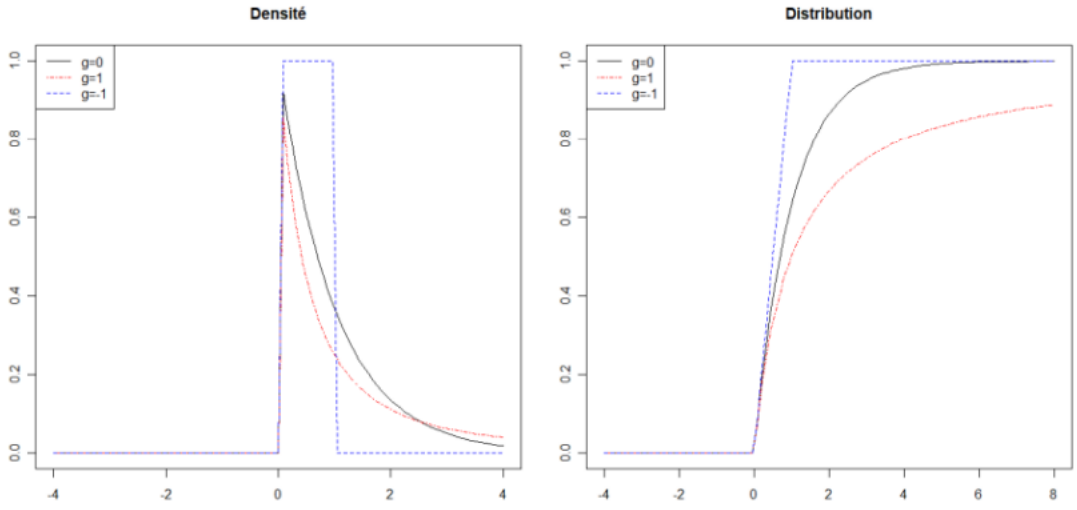


FIGURE 1.5 – Densité et distribution de GPD standard.

Dans la Figure 1.6 ci-dessous, nous illustrons le comportement de différentes GPD pour

$\sigma = 1$ et différentes valeurs de γ :

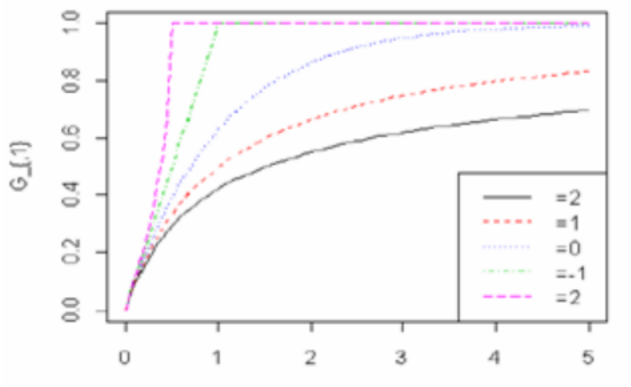


FIGURE 1.6 – Distributions de Pareto généralisées $G_{\gamma,1}$.

1.7 Conclusion

En conclusion, la théorie des valeurs extrêmes (TVE) représente bien plus qu'un simple cadre théorique : elle constitue un ensemble d'outils méthodologiques essentiels pour comprendre et gérer les événements extrêmes. En combinant des concepts statistiques avancés avec des applications concrètes dans divers domaines tels que l'hydrologie, la finance, l'ingénierie et la santé publique, la TVE joue un rôle crucial dans la prévision et la gestion des risques.

Grâce à la TVE, nous sommes en mesure d'identifier les distributions adaptées pour modéliser les queues de distribution, où se concentrent les événements rares mais potentiellement dévastateurs. Cela permet non seulement d'évaluer les probabilités de ces événements, mais aussi de prendre des décisions éclairées en matière de politique publique, de planification urbaine, de conception d'infrastructures et de stratégies financières.

En continuant à développer et à raffiner les méthodes de la TVE, nous renforçons notre capacité à anticiper les crises, à minimiser les pertes et à promouvoir la résilience face aux chocs naturels et industriels. Ainsi, la TVE contribue de manière significative à la sécurité, à la durabilité et à la prospérité à long terme des sociétés et des environnements dans lesquels nous vivons.

Chapitre 2

ESTIMATION DE L'INDICE DES VALEURS EXTREMES

Introduction

La loi des valeurs extrêmes, lorsqu'elle existe, est caractérisée par un paramètre essentiel appelé l'indice des valeurs extrêmes. Cet indice contrôle l'épaisseur de la queue de distribution, ce qui en fait un paramètre crucial à estimer à partir d'un échantillon fini (X1, ..., Xn). Deux approches principales sont couramment utilisées pour cette estimation : l'approche basée sur la distribution généralisée des valeurs extrêmes (GEVD) et l'approche utilisant la distribution de Pareto généralisée (GPD), également connue sous le nom de méthode du premier excès (POT).

De nombreux chercheurs ont contribué à ce domaine, en proposant différentes méthodes d'estimation, telles que celles développées par Hill (1975), Pickands (1975), Mason (1982), Csörgő et al. (1985), Dekkers et al. (1989), et Davis et Resnick (1984). Ces méthodes sont utilisées pour estimer l'indice des valeurs extrêmes, ainsi que d'autres paramètres associés, en se concentrant sur les propriétés asymptotiques des estimateurs.

Dans ce chapitre, nous examinons de près certains des estimateurs les plus utilisés, en mettant en évidence leurs propriétés asymptotiques. Nous explorons également différentes méthodes pour construire des estimateurs pour les quantiles extrêmes, les queues de distribution et les périodes de retour, afin de mieux comprendre et modéliser les événements extrêmes.

2.1 Estimation

On cherche à estimer les paramètres de la loi des valeurs extrêmes définie par la fonction de répartition GEV. Cette fonction, $H_\theta(x)$, est déterminée comme suit :

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

Les paramètres $\theta = (\gamma, \mu, \sigma)$ sont constitués d'un paramètre de forme, d'un paramètre de localisation et d'un paramètre d'échelle, et doivent être estimés à partir d'un échantillon $(X_1, ..., X_n)$ de n V.A indépendantes suivant la même fonction de répartition F.

Deux situations se présentent : lorsque F correspond exactement à H_θ, ou lorsque F est dans $D(H_\gamma)$.

Dans cette section, nous examinons la première situation où nous pouvons estimer les paramètres de la GEVD. Deux méthodes d'estimation sont envisageables : l'estimation par Maximum de Vraisemblance (EMV), les Moments Pondérés (EMP).

2.1.1 Méthode du Maximum de Vraisemblance (EMV)

La méthode du maximum de vraisemblance (EMV) est une approche largement utilisée pour estimer les paramètres d'un modèle statistique. Elle repose sur le principe de sélection des paramètres qui maximisent la fonction de vraisemblance, ou plus fréquemment, son logarithme, en fonction des paramètres spécifiés par la famille de lois choisie pour ajustement.

En théorie de l'estimation paramétrique, l'EMV est particulièrement prisée en raison de sa simplicité et de ses bonnes propriétés statistiques. Elle fournit des estimateurs qui sont efficaces, c'est-à-dire qu'ils exploitent au mieux l'information contenue dans les données disponibles. De plus, ces estimateurs sont consistants, ce qui signifie qu'ils convergent vers les vraies valeurs des paramètres lorsque la taille de l'échantillon augmente. Enfin, ils sont asymptotiquement normaux, ce qui implique que leur distribution converge vers une loi normale lorsque la taille de l'échantillon devient grande, facilitant ainsi l'analyse statistique et l'inférence.

En résumé, l'EMV représente un outil puissant et robuste pour estimer les paramètres des modèles statistiques, offrant une base solide pour la prise de décision dans divers domaines allant de la recherche scientifique à l'analyse des données économiques et sociales.

La fonction de vraisemblance $L(\theta; x_1, ..., x_n)$ basée sur les données $(X_1, ..., X_n)$, où H_θ a

pour fonction de densité h_θ, est définie comme suit :

$$L(\theta;x_1,...,x_n) = \prod_{i=1}^{n} h_\theta(x_i)\ \mathbf{1}_{\{1+\gamma(x_i-\mu)/\sigma>0\}}$$

Puis, la fonction log-vraisemblance est :

$$l(\theta;x_1,...,x_n) = \log L(\theta;x_1,...,x_n)$$

L'estimateur de maximum de vraisemblance $\hat{\theta}_n$ est défini comme suit :

$$\hat{\theta}_n = \hat{\theta}_n(x_1,...,x_n) = \arg\max_{\theta\in\Theta} l(\theta;x_1,...,x_n)$$

Cela signifie que $\hat{\theta}_n$ maximise $l(\theta;x_1,...,x_n)$ dans l'espace paramétrique Θ. Une approche courante pour trouver $\hat{\theta}_n$ est de résoudre le système d'équations obtenues en égalant à zéro les dérivées partielles de $l(\theta;x_1,...,x_n)$ par rapport à γ, μ et σ.

Dans le cas où $\gamma = 0$, la log-vraisemblance est :

$$l((0,\mu,\sigma);x_1,...,x_n) = -n\log\sigma - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma}$$

En dérivant cette fonction par rapport à μ et σ, nous obtenons le système d'équations suivant :

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

C'est un système d'équations non linéaires pour lesquelles aucune solution explicite n'existe.

Remarque : Lorsque $\gamma \neq 0$, dans ce cas, la maximisation de la log-vraisemblance implique la résolution d'un système d'équations non linéaires qui peuvent être difficiles à résoudre analytiquement. Cela rend l'estimation des paramètres plus complexe et nécessite souvent l'utilisation de méthodes numériques d'optimisation pour obtenir une solution.

2.1.2 Estimateurs des Moments Pondérés (EMP)

C'est une méthode basée sur les moments pondérés des données. Cette méthode, introduite par Hosking, Wallis et Wood en 1985, utilise la quantité suivante :

$$\omega_r(\theta) = \mathbb{E}[XH_\theta^r(X)]$$

où $r \in \mathbb{N}$. L'analogue empirique de cette quantité est donné par :

$$\hat{\omega}_r(\theta) = \int_{-\infty}^{+\infty} xH_\theta^r(x)\, dF_n(x) \quad r \in \mathbb{N}$$

où F_n est la fonction de répartition empirique de l'échantillon $(X_1, \ldots, X_n)$. Il est important de noter que $\omega_r(\theta)$ est la moyenne théorique de la distribution et $\hat{\omega}_r(\theta)$ est la moyenne empirique de l'échantillon.

En observant que, par un changement de variable, on peut exprimer $\hat{\omega}_r(\theta)$ comme suit :

$$\hat{\omega}_r(\theta) = \int_0^1 H_\theta^{\leftarrow}(y) y^r\, dy$$

où, pour $0 < y < 1$, la fonction $H_\theta^{\leftarrow}(y)$ est définie comme suit :

$$H_\theta^{\leftarrow}(y) = \begin{cases} \mu - \frac{\sigma}{\gamma}(1 - (1 - \log y)^{-\gamma}) & \text{si } \gamma \neq 0 \\ \mu - \sigma \log(-\log y) & \text{si } \gamma = 0 \end{cases}$$

D'autre part, l'estimateur empirique des moments pondérés $\hat{\omega}_r(\theta)$ peut également être exprimé en termes des statistiques d'ordre. Soit $X_{1,n}, \ldots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $(X_1, \ldots, X_n)$. Alors, nous avons :

$$\hat{\omega}_r(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_{i,n} H_\theta^r(X_{i,n}) \tag{2.1}$$

En utilisant le lemme de la transformation de quantile, nous pouvons réécrire l'équation (2.1) en termes des statistiques d'ordre associées à l'échantillon uniforme standard $(U_1, \ldots, U_n)$:

$$(H(X_{1,n}), \ldots, H(X_{n,n})) = (U_{1,n}, \ldots, U_{n,n}) \tag{2.2}$$

où $U_{1,n}, \ldots, U_{n,n}$ sont les statistiques d'ordre associées à l'échantillon uniforme standard $(U_1, \ldots, U_n)$. Avec cette interprétation, l'équation (2.1) peut être écrite comme :

$$\hat{\omega}_r(\theta) = \hat{\omega}_r = \frac{1}{n} \sum_{i=1}^{n} X_{i,n} U_{i,n}^r \quad \text{pour } r = 0, 1, 2$$

Dans la pratique, l'espérance $E(U_{i,n}^r)$ est souvent approximée par :

$$E(U_{i,n}^r) = \frac{(n-i)(n-i-1)\ldots(n-i-r+1)}{(n-1)(n-2)\ldots(n-r)} \quad \text{pour } r = 1, 2$$

L'estimateur des moments pondérés (EMP) de θ est obtenu en résolvant le système de

trois équations :

$$\omega_r(\theta) = \hat{\omega}_r(\theta) \quad \text{pour } r = 0,1,2$$

Pour $\gamma < 1$ et $\gamma \neq 0$, nous avons :

$$\omega_r(\theta) = \frac{1}{r+1}\left(\mu - \frac{\sigma}{\gamma}\left(1 - \Gamma(1-\gamma)(1+r)^{\gamma}\right)\right) \tag{2.3}$$

La fonction Gamma, est définie comme suit :

$$\Gamma(u) = \int_0^{\infty} e^{-t} t^{u-1} dt, \quad u > 0$$

Dans ce cas, le système ci-dessus est équivalent à :

$$\begin{cases} \hat{\omega}_0(\theta) = \mu - \frac{\sigma}{\gamma}(1 - \Gamma(1-\gamma)) \\ 2\hat{\omega}_1(\theta) - \hat{\omega}_0(\theta) = \frac{\sigma}{\gamma}\Gamma(1-\gamma)(2^{\gamma}-1) \\ 3\hat{\omega}_2(\theta) - \hat{\omega}_0(\theta) = \frac{\sigma}{\gamma}\Gamma(1-\gamma)(3^{\gamma}-1) \end{cases}$$

Par conséquent, nous avons :

$$\frac{3\hat{\omega}_2(\theta) - \hat{\omega}_0(\theta)}{2\hat{\omega}_1(\theta) - \hat{\omega}_0(\theta)} = \frac{3^{\gamma}-1}{2^{\gamma}-1}$$

La solution de cette équation est l'estimateur de EMP $\hat{\gamma}$ de γ. Les autres paramètres σ et μ sont estimés respectivement par :

$$\hat{\sigma} = \frac{(2\hat{\omega}_1 - \hat{\omega}_0)\hat{\gamma}}{\Gamma(1-\hat{\gamma})(2\hat{\gamma}-1)}$$

et

$$\hat{\mu} = \hat{\omega}_0 + \frac{\hat{\sigma}}{\hat{\gamma}}(1 - \Gamma(1-\hat{\gamma}))$$

2.2 Estimation semi - paramétrique

L'estimation semi-paramétrique est une méthode statistique qui combine les avantages des modèles paramétriques et non paramétriques pour l'analyse des données. Dans un modèle semi-paramétrique, on utilise un mélange de composantes paramétriques et non paramétriques, permettant aux chercheurs de tirer parti des structures préétablies des modèles paramétriques ainsi que de la grande flexibilité offerte par les méthodes non paramétriques.

Les modèles paramétriques reposent sur des hypothèses spécifiques concernant la distribution des données et la forme mathématique de la relation entre les variables, ce qui les rend puissants et efficaces pour estimer les paramètres lorsque ces hypothèses sont correctes. En revanche, les modèles non paramétriques ne nécessitent pas d'hypothèses préalables sur la forme de la distribution ou la relation entre les variables, ce qui leur confère une plus grande flexibilité pour représenter des schémas complexes et inattendus dans les données.

L'estimation semi-paramétrique combine ces deux approches en permettant au modèle d'utiliser une partie paramétrique et une partie non paramétrique. Par exemple, un modèle semi-paramétrique peut inclure des paramètres fixes pour décrire certains aspects fondamentaux de la relation entre les variables, tout en estimant d'autres aspects de cette relation de manière non paramétrique à l'aide de techniques telles que la régression spline ou les méthodes à noyau.

Les méthodes semi-paramétriques sont largement utilisées dans de nombreux domaines scientifiques tels que l'économie, la médecine, la biologie et les sciences sociales, où les relations entre les variables sont trop complexes pour être représentées avec précision à l'aide de modèles purement paramétriques. Par exemple, ces méthodes peuvent être utilisées dans l'analyse des données de survie, où un modèle semi-paramétrique peut traiter les effets paramétriques de certaines variables démographiques tout en permettant à la distribution des temps de survie d'être non paramétrique.

L'estimation semi-paramétrique est un outil puissant qui permet aux chercheurs de bénéficier des avantages des modèles paramétriques et non paramétriques, offrant un équilibre idéal entre précision et flexibilité dans l'analyse des données complexes.

Les estimateurs classiques sont basés sur les plus grandes statistiques d'ordre $X_{n-k,n} \leq \ldots \leq X_{n,n}$, où k est une suite intermédiaire d'entiers liés à la taille de l'échantillon n de la manière suivante :

$$\text{quand} \quad n \to \infty, \text{ alors} \quad k = k_n \to \infty$$

2.2.1 Estimateur de Pickands

L'estimateur de Pickands a été introduit par James Pickands III dans les années 1970 dans le cadre de la théorie des valeurs extrêmes, une branche des statistiques qui étudie les comportements des événements rares et extrêmes dans les jeux de données. Cette théorie est essentielle pour des domaines tels que la finance, l'hydrologie et la météorologie, où la modélisation et la prévision des événements extrêmes, comme les crues, les vagues de chaleur ou les effondrements de marché, sont cruciales.

James Pickands III a développé cet estimateur pour estimer l'indice de forme d'une dis-

tribution de valeurs extrêmes, aussi connu sous le nom d'indice de queue. L'indice de queue est un paramètre fondamental qui décrit la probabilité et la gravité des événements extrêmes. L'estimateur de Pickands utilise les valeurs ordonnées d'un échantillon et repose sur des quantiles spécifiques pour fournir une estimation robuste de cet indice.

Formellement, si $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ sont les valeurs ordonnées d'un échantillon de taille n, l'estimateur de Pickands est donné par :

$$\hat{\gamma}_k = \frac{1}{\ln(2)} \ln\left(\frac{X_{k:n} - X_{k/2:n}}{X_{2k:n} - X_{k:n}}\right),$$

où k est un paramètre choisi en fonction de la taille de l'échantillon.

Avantages de l'estimateur de Pickands L'un des principaux avantages de l'estimateur de Pickands est sa simplicité. Il est relativement facile à calculer et ne nécessite pas de méthodes complexes. De plus, il est robuste, fournissant des estimations fiables de l'indice de forme même pour des échantillons de taille modeste. Cette robustesse le rend particulièrement utile dans des applications pratiques où les tailles d'échantillons peuvent être limitées. Enfin, il est largement utilisé pour modéliser des événements extrêmes, ce qui en fait un outil précieux dans de nombreux domaines.

Inconvénients de l'estimateur de Pickands Malgré ses avantages, l'estimateur de Pickands présente également certains inconvénients. Le choix du paramètre k est crucial pour la précision de l'estimation et peut fortement influencer les résultats. Un mauvais choix de k peut conduire à des estimations biaisées. De plus, comme l'estimateur se concentre sur les valeurs extrêmes, il peut être sensible aux anomalies ou aux erreurs dans les données. Enfin, les propriétés théoriques de l'estimateur sont asymptotiques, ce qui signifie qu'elles s'appliquent principalement lorsque la taille de l'échantillon est grande, limitant ainsi son efficacité pour les petits échantillons.

En résumé, l'estimateur de Pickands, bien que puissant et simple à utiliser, nécessite une attention particulière au choix du paramètre et à la qualité des données pour fournir des estimations précises et fiables des événements extrêmes.

Construction de l'estimateur de Pickands

On déduit de la relation 1.5 que pour $\gamma \in \mathbb{R}$ et on a avec le choix $t = 2s$, $x = 2$ et $y = \frac{1}{2}$,

$$\lim_{t\to\infty} \frac{U(t) - U(t/2)}{U(t/2) - U(t/4)} = 2^{\gamma}$$

En fait, en utilisant la croissance de U qui se déduit de la croissance de F, on obtient

$$\lim_{t\to\infty}\frac{U(t)-U(tc_1(t))}{U(tc_1(t))-U(tc_2(t))}=2^{\gamma}$$

dès que $\lim_{t\to\infty} c_1(t)=\frac{1}{2}$ et $\lim_{t\to\infty} c_2(t)=\frac{1}{4}$. Il reste donc à trouver des estimateurs pour $U(t)$.
Soit $(k(n),n\geq 1)$ une suite d'entiers telle que $1\leq k(n)\leq \frac{n}{4}$, $\lim_{n\to\infty}\frac{k(n)}{n}=0$ et $\lim_{n\to\infty} k(n)=\infty$.

Soit $(V_{1,n},\ldots,V_{n,n})$ la statistique d'ordre d'un échantillon de variables aléatoires indépendantes suivant une loi de Pareto, notée $F_V(x)=1-x^{-1}$, pour $x\geq 1$.

On déduit, avec certains résultats de base liés à $(V_{1,n},\ldots,V_{n,n})$, que les suites $\left(\frac{k}{n}V_{n-k+1,n},n\geq 1\right)$, $\left(\frac{2k}{n}V_{n-2k+1,n},n\geq 1\right)$ et $\left(\frac{4k}{n}V_{n-4k+1,n},n\geq 1\right)$ convergent en probabilité vers 1.

On en déduit en particulier que les convergences en probabilité suivantes ont lieu :

$$V_{n-k+1,n}\xrightarrow[n\to\infty]{}\infty,$$

$$\frac{V_{n-2k+1,n}}{V_{n-k+1,n}}\xrightarrow[n\to\infty]{}\frac{1}{2},$$

et

$$\frac{V_{n-4k+1,n}}{V_{n-k+1,n}}\xrightarrow[n\to\infty]{}\frac{1}{4}.$$

Donc, la convergence suivante a lieu en probabilité :

$$\frac{U(V_{n-k+1,n})-U(V_{n-2k+1,n})}{U(V_{n-2k+1,n})-U(V_{n-4k+1,n})}\xrightarrow[n\to\infty]{}2^{\gamma}.$$

Remarquons que si $x\geq 1$, alors $U(x)=F^{-1}(F_V(x))$. Ainsi,

$$(U(V_{1,n}),\ldots,U(V_{n,n}))=\left(F^{-1}(F_V(V_{1,n})),\ldots,F^{-1}(F_V(V_{n,n}))\right).$$

où F_V est la fonction de répartition de la loi de Pareto.

On déduit de la croissance de F_V que $(F^{-1}(F_V(V_{1:n})),\ldots,(F^{-1}(F_V(V_{n:n}))$ a la même loi que la statistique d'ordre de n variables aléatoires uniformes sur $[0,1]$ indépendantes. Donc on conclut que le vecteur aléatoire $(F^{-1}(F_V(V_{1,n})),\ldots,F^{-1}(F_V(V_{n,n})))$ a la même loi que $X_{1,n},\ldots,X_{n,n}$. Donc, la variable aléatoire

$$\frac{U(V_{n-k+1,n})-U(V_{n-2k+1,n})}{U(V_{n-k+1,n})-U(V_{n-4k+1,n})}$$

a la même loi que

$$\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-k+1,n} - X_{n-4k+1,n}}.$$

alors, cette quantité converge en loi vers 2^{γ} lorsque n tend vers l'infini.

Les propriétés asymptotiques de γ_n^p

Supposons que F appartient à l'espace de domaines d'attraction $D(H_\gamma)$, où H_γ est une fonction de répartition généralisée de Pareto, et γ est un paramètre réel. De plus, considérons y comme un réel.

(a) La consistance faible de l'estimateur $\hat{\gamma}_n^{(p)}$ signifie que lorsque la taille de l'échantillon tend vers l'infini ($n \to \infty$), l'estimateur converge en probabilité vers la vraie valeur du paramètre γ.

(b) La consistance forte de $\hat{\gamma}_n^{(p)}$ signifie que si le rapport $k/\log\log n$ tend vers l'infini lorsque n tend vers l'infini, alors $\hat{\gamma}_n^{(p)}$ converge presque sûrement vers γ. Cela signifie que la probabilité de l'écart entre $\hat{\gamma}_n^{(p)}$ et γ tend vers zéro lorsque la taille de l'échantillon augmente.

(c) La normalité asymptotique de $\hat{\gamma}_n^{(p)}$ sous certaines conditions pour k et F signifie que lorsque la taille de l'échantillon tend vers l'infini, la distribution de $\sqrt{k}(\hat{\gamma}_n^{(p)} - \gamma)$ converge en loi vers une distribution normale avec une moyenne de zéro et une variance η^2. La variance η^2 est déterminée par l'expression donnée et dépend du paramètre γ.

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

La figure (2.1) Estimateur de Pickands, en fonction du nombre des extrêmes (en trait plein) avec l'intervalle de confiance 95(lignes tirées), pour l'IVE de la distribution uniforme(γ = 1) basé sur 100 échantillons de 3000 observations.

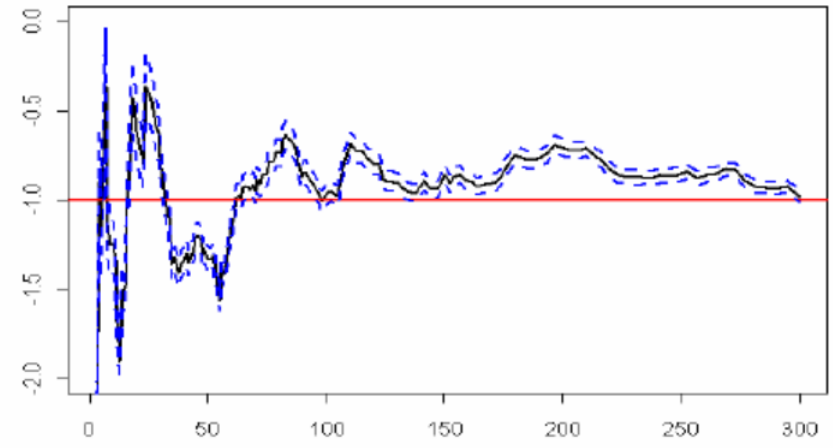


FIGURE 2.1 – Estimateur de Pickands

2.2.2 Estimateur de Hill

L'estimateur de Hill est une méthode statistique utilisée dans la théorie des valeurs extrêmes pour estimer l'indice de forme d'une distribution de valeurs extrêmes, également connu sous le nom d'indice de queue. Cet indice est crucial pour évaluer la probabilité et la gravité des événements rares et extrêmes, ce qui est essentiel dans des domaines comme la finance, l'hydrologie, et les sciences de l'atmosphère.

Historique et Développement

L'estimateur de Hill a été introduit par le statisticien néerlandais Bruce M. Hill en 1975. Il a développé cette méthode pour fournir une estimation robuste et efficace de l'indice de queue, particulièrement pour les distributions de type Pareto, qui sont souvent utilisées pour modéliser les extrêmes.

Définition de l'estimateur de Hill

L'estimateur de Hill est basé sur les valeurs ordonnées d'un échantillon. Si $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ sont les valeurs ordonnées d'un échantillon de taille n, alors l'estimateur de Hill pour les k plus grandes valeurs est défini par :

$$\hat{\gamma}_k = \frac{1}{k} \sum_{i=1}^{k} \left(\ln X_{n-i+1:n} - \ln X_{n-k:n} \right),$$

où k est le nombre de plus grandes valeurs utilisées pour l'estimation, et $X_{n-i+1:n}$ représente la $(n-i+1)$-ième valeur ordonnée (c'est-à-dire, l'une des k plus grandes valeurs).

Avantages de l'estimateur de Hill

- Simplicité : L'estimateur est facile à calculer et ne nécessite pas de techniques complexes.
- Efficacité : Il est particulièrement efficace pour les distributions de type Pareto et est largement utilisé pour modéliser les queues épaisses des distributions.
- Robustesse : Il fournit des estimations fiables de l'indice de queue, même avec des échantillons de taille modeste.

Inconvénients de l'estimateur de Hill

- Choix du paramètre k : Le choix de k est crucial et peut grandement affecter les résultats. Un choix inapproprié peut entraîner des estimations biaisées.

— Sensibilité aux valeurs extrêmes : Bien que l'estimateur de Hill soit conçu pour les extrêmes, il peut être sensible aux anomalies ou aux erreurs dans les données.

— Asymptotique : Les propriétés théoriques de l'estimateur sont asymptotiques, ce qui signifie qu'elles sont plus précises pour de grandes tailles d'échantillon.

Applications Pratiques

L'estimateur de Hill est largement utilisé dans des domaines où la modélisation des événements extrêmes est cruciale. Par exemple :

— Finance : Pour évaluer les risques extrêmes tels que les krachs boursiers.

— Hydrologie : Pour prédire les crues extrêmes et autres événements hydrologiques rares.

— Météorologie : Pour modéliser les phénomènes météorologiques extrêmes comme les ouragans et les vagues de chaleur.

En résumé, l'estimateur de Hill est un outil statistique essentiel pour l'analyse des valeurs extrêmes. Sa simplicité et son efficacité en font un choix privilégié pour estimer l'indice de queue des distributions, malgré la nécessité de choisir avec soin le paramètre k pour garantir la précision des estimations.

À l'origine, l'estimateur de Hill a été introduit comme un MLE (en fait, c'est un quasi ou pseudo MLE). Soit $F \in D(\Phi_{1/\gamma})$ et soit $Y_u := (X \mid X > u)$ désigne les excès relatifs au-dessus d'un seuil u. La distribution conditionnelle de Y_u satisfait

$$\overline{F}_{Y_u}(y) = y^{-1/\gamma} L(uy)/L(u), \quad y \geq 1,$$

où $L \in R_0$. Il s'ensuit immédiatement que

$$F_{Y_u}(y) \to 1 - y^{-1/\gamma} \text{ lorsque } u \to \infty,$$

c'est-à-dire que Y_u est asymptotiquement distribué selon Pareto$(1/\gamma)$. En supposant que cette approximation soit valable pour les k excès relatifs $X_{n,n}/u, \ldots, X_{n-k+1,n}/u$ au-dessus d'un seuil élevé $u = X_{n-k,n}$, on obtient le MLE

$$\frac{1}{k} \sum_{i=1}^{k} \log(X_{n-i+1,n}/X_{n-k,n}),$$

ce qui est exactement $\hat{\gamma}_n^{(H)}$.

Dans son article original en 1977, Hill n'a pas étudié le comportement asymptotique de l'estimateur. C'est Mason qui a prouvé la consistance faible en 1993. La consistance forte a été prouvée en 1988 par Deheuvels, Häsler et Mason qui ont donné un taux optimal de

convergence pour une séquence k_n appropriée. La normalité asymptotique a été établie, sous certaines conditions supplémentaires sur F, dans plusieurs articles tels que, par exemple, ceux de 1984, 1987, 1989, 1991 et 1993. Récemment, Beirlant, Bouquiaux et Werker en 2006 ont dérivé un résultat de normalité asymptotique locale montrant que la variance asymptotique de l'estimateur de Hill atteint une borne inférieure.

Proposition 2.1. (Condition de Variation Régulière de Premier Ordre) Les assertions suivantes sont équivalentes :

(a) F a une queue lourde $F \in D(\Phi_{1/\gamma}), > 0$.

(b) $\overline{F}$ varie régulièrement à l'infini avec l'indice $-1/\gamma$:

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

(c) $Q(1-s)$ varie régulièrement à 0 avec l'indice $-\gamma$:

$$\lim_{s\to 0} \frac{Q(1-sx)}{Q(1-s)} = x^{-\gamma}, \quad x > 0.$$

(d) U varie régulièrement à l'infini avec l'indice :

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}, \quad x > 0.$$

Définition 2.7. (Hypothèse de Variation Régulière de Second Ordre) Nous disons que (la queue de) $F \in D(\Phi_{1/\gamma})$, $\gamma > 0$, varie régulièrement au second ordre à l'infini si elle satisfait l'une des conditions (équivalentes) suivantes :

(a) Il existe un paramètre $\rho \leq 0$ et une fonction A^*, tendant vers 0 et ne changeant pas de signe près de l'infini, telle que pour tout $x > 0$,

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

(b) Il existe un paramètre 0 et une fonction A^{**}, tendant vers 0 et ne changeant pas de signe près de 0, telle que pour tout $x > 0$,

$$\lim_{s\to 0} \frac{Q(1-sx)/Q(1-s) - x^{-\gamma}}{A^{**}(s)} = x^{-\gamma} \cdot \frac{x^{\rho}-1}{\rho}.$$

(c) Il existe un paramètre $\rho 0$ et une fonction A, tendant vers 0 et ne changeant pas de

signe près de l'infini, telle que pour tout $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{\gamma} \cdot \frac{x^{\rho} - 1}{\rho}.$$

Si $\rho = 0$, interpréter $\frac{x^{\rho}-1}{\rho}$ comme $\log x$.

Les fonctions A, A^* et A^{**} sont des fonctions variant régulièrement, où $A^*(t) = A(1/(1 - F(t)))$ et $A^{**}(s) = A(1/s)$. Leur rôle est de contrôler la vitesse de convergence dans les équations présedents respectivement. Plus précisément, nous avons $A \in R_{\rho}$, $A^* \in R_{\rho/\gamma}$ et $A^{**} \in R_{-\rho}$.

Les relations ci-dessus peuvent être reformulées respectivement comme suit :

$$\lim_{t\to\infty} \frac{\log(1 - F(tx)) - \log(1 - F(t)) + (1/\gamma)\log x}{A^*(t)} = \frac{x^{\rho} - 1}{\rho},$$

$$\lim_{s\to 0} \frac{\log Q(1 - sx) - \log Q(1 - s) + \gamma \log x}{A^{**}(s)} = \frac{x^{\rho} - 1}{\rho},$$

et

$$\lim_{t\to\infty} \frac{\log U(tx) - \log U(t) - \gamma \log x}{A(t)} = \frac{x^{-1}}{\rho}.$$

Ces hypothèses équivalentes spécifient les taux de convergence nécessaires (pour dériver la normalité asymptotique des estimateurs d'indice de queue) dans la Proposition 2.1. Bien qu'elles soient rarement vérifiables en pratique, les conditions de second ordre sont principalement utiles d'un point de vue méthodologique.

La Classe de Distributions de Hall À titre d'exemple de distributions à queue lourde satisfaisant l'hypothèse de second ordre, nous avons le modèle de Hall souvent utilisé (introduit en 1972) qui est une classe de fonctions de répartition (df's) :

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

où $\gamma > 0$, $\rho 0$, $c > 0$, et $d \in \mathbb{R} \setminus \{0\}$. Cette sous-classe de distributions à queue lourde contient les distributions de Pareto, Burr, Fréchet et t-Student généralement utilisées, en mathématiques de l'assurance, comme modèles pour les risques dangereux. La relation (2.11) peut être reformulée en termes des fonctions Q et U comme suit :

$$Q(1 - s) = \frac{c}{\gamma} s^{-\gamma}\left(1 + c^{\rho} s^{-\rho} + o(s^{-\rho})\right) \text{ comme } s \to 0,$$

et

$$U(t) = \frac{c}{\rho} t^{\rho} \left(1 + \gamma d c^{\rho} t^{\rho} + o(t^{\rho})\right) \text{ comme } t \to .$$

Des calculs simples montrent que, dans le cas du modèle de Hall, les fonctions $A(t)$ et $A^*(t)$ sont respectivement équivalentes à $d\gamma c^{\rho} t^{\rho}$ et $d\gamma c^{\rho} t^{\rho/\gamma}$ lorsque $t \to$, tandis que la fonction $A^{**}(s)$ est équivalente à $d\gamma c^{\rho} s^{-\rho}$ lorsque $s \to 0$. Passons maintenant au comportement asymptotique de l'estimateur de Hill.

Théorème 2.2. (Propriétés Asymptotiques de $\hat{\gamma}_n^{(H)}$ Supposons que $F \in D(\Phi_{1/\gamma})$, $\gamma > 0$, $k \to \infty$ et $k/n \to 0$ lorsque $n \to \infty$.

(a) **Consistance Faible** :

$$\hat{\gamma}_n^{(H)} \xrightarrow{p} \gamma \text{ lorsque } n \to \infty.$$

(b) **Consistance Forte** : Si $k/\log\log n \to \infty$ lorsque $n \to \infty$, alors

$$\hat{\gamma}_n^{(H)} \xrightarrow{a.s.} \gamma \text{ lorsque } n \to \infty.$$

(c) **Normalité Asymptotique** : Supposons que F satisfait la condition (2.8). Si $\sqrt{k}A(n/k) \to \lambda$ lorsque $n \to \infty$, alors

$$\sqrt{k}\left(\hat{\gamma}_n^{(H)} - \gamma\right) \xrightarrow{d} N\left(\frac{\lambda}{1-\rho}, \gamma^2\right) \text{ lorsque } n \to \infty.$$

Remarquez qu'en général, l'estimateur de Hill n'est pas (asymptotiquement) sans biais. Sa variance asymptotique γ^2/k diminue pour des valeurs croissantes de k, ce qui peut malheureusement introduire un biais dans l'estimation. Par conséquent, il est nécessaire de faire un choix approprié du nombre k de statistiques d'ordre supérieur impliquées dans l'estimation de γ (voir Section 2.4). Par exemple, nous pouvons choisir k de telle manière que $\sqrt{k}A(n/k)$ ait une limite nulle, ce qui donne un estimateur sans biais pour γ. Mais cela peut se faire au détriment de la variance.

Comme il mesure l'augmentation moyenne de la courbe quantile de Pareto au-dessus d'un point d'ancrage fixe $(\log(n+1)/(k+1), \log X_{n-k,n})$, l'estimateur de Hill peut être interprété graphiquement comme l'estimateur de la pente de la courbe quantile de Pareto dans le cas où l'on considère des lignes de régression passant par ce point fixe.

En effectuant une régression linéaire non contrainte pour les k points les plus élevés (au lieu de la régression contrainte ci-dessus), Kratz et Resnick, ont proposé un estimateur cohérent et asymptotiquement normal pour l'indice de queue qu'ils appellent l'estimateur

QQ. La variance asymptotique de cet estimateur est deux fois celle de l'estimateur de Hill, mais ce n'est pas un critère de supériorité car ce dernier présente un biais plus considérable.

2.3 Modèle POT

La méthode des excès, également connue sous le nom de POT (Peaks Over Threshold), est une méthode d'estimation de la queue de distribution largement utilisée en statistique. Elle a été initialement développée dans les années 1970 en hydrologie avant d'être étudiée et appliquée dans divers domaines statistiques.

Cette méthode consiste à utiliser les observations qui dépassent un seuil prédéterminé. Plutôt que de modéliser directement toute la distribution des données, on se concentre uniquement sur les valeurs excédentaires au-dessus du seuil. Ces excès sont ensuite modélisés à l'aide d'une distribution appropriée, généralement une distribution généralisée de valeurs extrêmes (GEV) ou une distribution de Pareto généralisée (GPD).

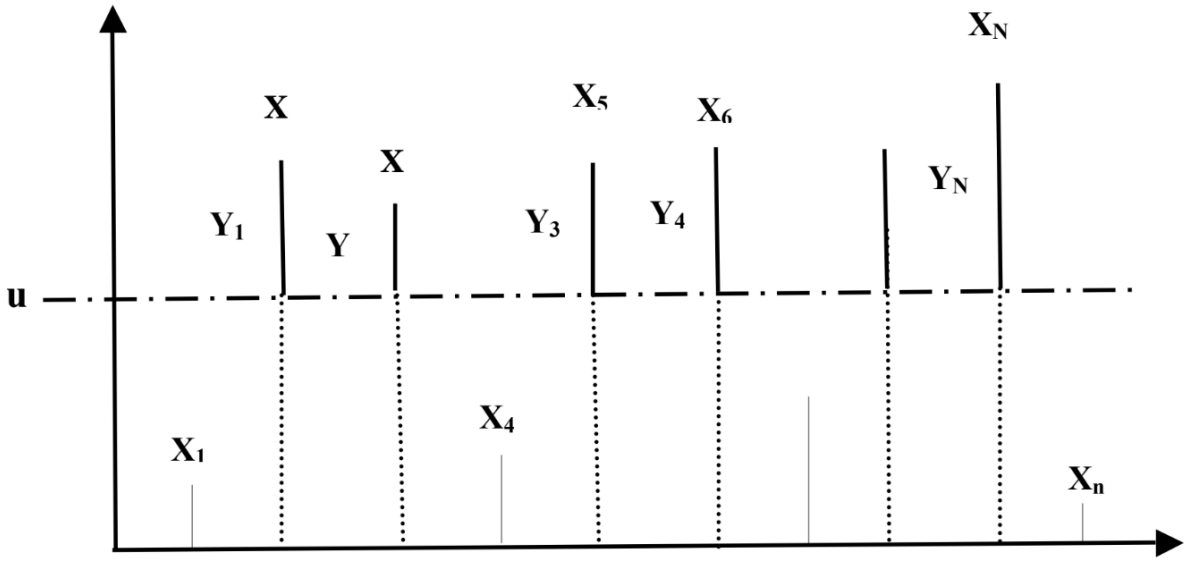


FIGURE 2.2 – Méthode des excès : u réel suffisamment élevé appelé seuil,Y :excès de X au-delà de u.

La méthode des excès repose sur une approximation de la loi des excès au-dessus d'un seuil donné u pour une variable aléatoire réelle X. Cette loi des excès est une approximation de la loi conditionnelle de la variable aléatoire réelle positive $X - u$ sachant que $X > u$. En d'autres termes, on s'intéresse uniquement aux valeurs excédentaires de X au-dessus du seuil u, et l'approximation vise à modéliser la distribution de ces excès.

2.3.1 Loi des excès

Pour définir la distribution conditionnelle F_u par rapport au seuil u pour les variables aléatoires dépassant ce seuil, on utilise les excès au-delà du seuil u, définis comme suit :

Soit X une variable aléatoire avec une fonction de répartition F et u un réel suffisamment grand mais inférieur au point terminal x_F (appelé seuil). Les excès au-delà du seuil u sont définis comme l'ensemble des variables aléatoires Y telles que $y_i = x_i - u$, où $x_i > u$.

En d'autres termes, si une observation x_i dépasse le seuil u, alors y_i est l'excès correspondant, c'est-à-dire la différence entre x_i et u.

L'objectif est alors de définir la distribution conditionnelle F_u basée sur les excès, ce qui permet de modéliser la distribution des valeurs excédentaires au-dessus du seuil u. Cette distribution conditionnelle F_u est essentielle dans la méthode des excès pour estimer les queues de distribution des variables aléatoires.

L'objectif de la méthode POT est de déterminer par quelle loi de probabilité on peut approcher cette distribution conditionnelle. Balkema et de Haan (1974), Pickands (1975) voir 1.6.2. Ce Théorème est très utile lorsque on travaille avec des observations qui dépassent un seuil fixé puisqu'il assure que la loi des excès peut-être approchée par une loi de Pareto généralisée.

Motivation de la Pareto généralisée

La distribution de Pareto généralisée (GPD) est souvent utilisée pour modéliser les observations qui dépassent un seuil u dans la méthode des excès. L'avantage principal de l'utilisation de la GPD par rapport aux distributions de valeurs extrêmes est qu'elle permet d'inclure plusieurs observations qui dépassent le seuil dans la modélisation, plutôt que de se limiter à une seule observation par période (comme c'est souvent le cas avec les distributions de valeurs extrêmes).

Cependant, un défi important lors de l'utilisation de la GPD est de déterminer un seuil approprié u. Ce problème est similaire à celui de déterminer le paramètre k pour les estimateurs de la distribution de valeurs extrêmes généralisée. Le choix du seuil u est crucial car il influence directement la qualité de l'ajustement du modèle. Un seuil mal choisi peut conduire à des estimations biaisées ou peu fiables.

Ainsi, bien que l'utilisation de la GPD offre des avantages en termes de modélisation des excès au-dessus d'un seuil, la sélection d'un seuil approprié reste une étape critique dans le processus d'estimation des queues de distribution des variables aléatoires.

2.3.2 Stabilité du seuil

Dans le contexte de la distribution de Pareto généralisée (GPD), une variable aléatoire est dite stable par rapport au seuil si certaines propriétés sont vérifiées. Si $u > 0$ est le seuil, alors pour $x > 0$, $P(X - u \leq x | X > u) = G_{\gamma,\sigma+\gamma(u-\mu)}(x)$ notée $G_{\gamma,\sigma(u)}(x)$, où :
- Si $\mu = 0$, alors $\sigma(u) = \sigma + \gamma u$.
- Si $\mu \neq 0$, alors $\sigma(u) = \sigma + \gamma(u - \mu)$.

Ces relations expriment comment le paramètre de dispersion σ de la distribution de Pareto généralisée évolue en fonction du seuil u, en tenant compte du paramètre de forme γ et du paramètre de position μ, le cas échéant. Cela permet de caractériser la stabilité de la distribution par rapport au seuil.

2.3.3 Détermination du seuil

Le choix du seuil doit être équilibré pour trouver un compromis. Plus le seuil est élevé, moins le biais de l'estimateur est important, ce qui permet d'obtenir un modèle plus précis. En revanche, un seuil plus bas réduit la variance de l'estimateur car davantage de données sont prises en compte dans l'estimation. Voici quelques-unes des méthodes suggérées dans la littérature :

Méthode graphique

La première méthode, complémentaire à la première, implique d'ajuster plusieurs modèles de distribution de Pareto généralisée (GPD) aux données en utilisant différents seuils. Cela permet d'obtenir différentes estimations, et l'objectif est d'analyser la stabilité des paramètres pour sélectionner le seuil "optimal".

Le raisonnement sous-jacent repose sur le Théorème de Balkema-de Haan-Pickands, qui stipule que si une distribution généralisée de Pareto (GPD) est un modèle raisonnable pour les excès d'un certain seuil u_0, alors les excès d'un seuil supérieur à u_0 suivent également une distribution généralisée de Pareto. Il est important de noter que l'estimation du paramètre de forme γ ne dépend pas du choix du seuil u, et donc sa valeur devrait rester constante quel que soit le seuil choisi.

Par contre, l'estimation du paramètre d'échelle $\sigma(u)$ est influencée par le choix du seuil. Comme nous l'avons vu précédemment, $\sigma(u)$ et u sont liés par la relation suivante :

$$\sigma(u) = \sigma(u_0) + \gamma(u - u_0)$$

Ainsi, $\sigma(u)$ change de manière linéaire en fonction du seuil u. Pour remédier à cela, nous pou-

vons définir un paramètre d'échelle reparamétrisé σ^* (échelle modifiée), qui reste constant avec u, défini comme suit :

$$\sigma^* = \sigma(u) - \gamma u$$

Avec cette nouvelle définition, et sachant que γ est constant en fonction de u, l'estimation de σ^* devrait également être constante.

La méthode consiste donc à estimer les paramètres γ et σ^* pour différents seuils dans les modèles GPD correspondants, avec leurs intervalles de confiance, puis de tracer les graphiques de ces paramètres en fonction du seuil u. Ensuite, il s'agit de rechercher le plus petit seuil pour lequel l'estimation des paramètres reste constante au-delà du seuil, ce qui permet de limiter la variance de l'estimateur tout en maintenant la stabilité des paramètres.

La deuxième méthode,D'accord, cette méthode expérimentale pour le choix du seuil repose sur la moyenne de la distribution de Pareto. Soit Y une variable aléatoire suivant une distribution de Pareto avec les paramètres $\sigma(u)$ et γ. La moyenne de Y, notée $E[Y]$, est donnée par l'expression suivante :

$$E[Y] = \begin{cases} \frac{\sigma(u)}{1-\gamma} & \text{si } \gamma < 1 \\ +\infty & \text{si } \gamma \geq 1 \end{cases}$$

Maintenant, supposons que la distribution de Pareto soit un modèle valide pour les observations qui excèdent un certain seuil u_0, provenant d'une suite de variables aléatoires $X_1, \ldots, X_n$ (c'est-à-dire que la distribution limite est une bonne approximation), alors la moyenne conditionnelle $e(u_0) = E[X - u_0 \,|\, X > u_0]$ est donnée par :

$$e(u_0) = \begin{cases} \frac{\sigma(u_0)}{1-\gamma} & \text{si } \gamma < 1 \\ +\infty & \text{si } \gamma \geq 1 \end{cases}$$

où $\sigma(u_0) = \sigma + \gamma(u_0 - \mu)$.

Cette méthode utilise la moyenne conditionnelle des excès au-delà du seuil u_0 pour guider le choix du seuil optimal. Si la moyenne conditionnelle est stable et finie, cela indique que la distribution de Pareto est un bon modèle pour les excès au-dessus de ce seuil.

Mais alors, si la distribution de Pareto est une bonne approximation en choisissant le seuil u_0, elle le sera également en choisissant tout seuil u supérieur à u_0 ($u > u_0$). Ainsi,

$$e(u) = E[X - u \,|\, X > u_0] = \frac{\sigma(u_0) + \gamma(u - u_0)}{1-\gamma} = \frac{\sigma + \gamma(u - \mu)}{1-\gamma}$$

pour autant que $\gamma < 1$, $u > u_0$.

Nous observons donc que $E[X-u\,|\,X>u]$ est une fonction linéaire de u, pour $u>u_0$. Étant donné que $E[X-u\,|\,X>u]$ n'est rien d'autre que la moyenne des observations extrêmes dépassant le seuil u, nous pouvons l'approximer par la valeur suivante (en vertu de la loi forte des grands nombres) :

$$\hat{e}_n(u) = E[X-u\,|\,X>u] \approx \frac{1}{N_u}\sum_{i=1}^{N_u}(x_i-u)$$

où $x_1,\ldots,x_{N_u}$ représentent simplement les N_u observations qui excèdent le seuil u.

Cette procédure consiste à calculer la moyenne conditionnelle des excès $e(u)$ pour chaque seuil u plus petit que la valeur maximale des données. Ensuite, nous construisons un graphique où nous avons les points suivants :

$$\{(u,\hat{e}_n(u)) : u < x_{\max}\}$$

En effectuant cette démarche, nous nous attendons à observer un graphique approximativement linéaire en fonction de u. À partir de la valeur de u qui fournit un modèle adéquat pour les données, nous devrions voir une stabilisation des valeurs de $\hat{e}_n(u)$, indiquant ainsi que ce seuil fournit une approximation fiable de la moyenne conditionnelle des excès.

2.3.4 Estimation des paramètres de la GPD

Dans la méthode du Maximum de Vraisemblance (EMV), supposons que notre échantillon des excès $X=(X_1,\ldots,X_{N_u})$ est i.i.d. (indépendantes et identiquement distribuées) avec une distribution de Pareto généralisée (GPD) $G_{\gamma,\sigma}$. La fonction de densité $g_{\gamma,\sigma}$ de la GPD est définie comme suit :

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

La log-vraisemblance est donc donnée par :

$$l((\gamma,\sigma);X) = -N_u \ln\sigma - \left(\frac{1}{\gamma}+1\right)\sum_{i=1}^{N_u}\ln\left(1+\gamma\frac{x_i}{\sigma}\right)$$

En dérivant cette expression par rapport aux deux paramètres d'intérêt, γ et σ, nous obtenons un système de deux équations à deux inconnues. En résolvant ces équations, généralement à l'aide de méthodes numériques, nous obtenons les estimateurs du maximum de vraisemblance $(\hat{\gamma}_{\mathrm{Nu}},\hat{\sigma}_{\mathrm{Nu}})$.

Lorsque $\gamma=0$, la fonction de densité de la GPD devient simplement :

$$g_{0,\sigma}(x) = \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right)$$

La log-vraisemblance correspondante est donnée par :

$$l((0,\sigma);X) = -N_u \ln\sigma - \frac{1}{\sigma}\sum_{i=1}^{N_u} x_i$$

Dans ce cas, l'estimateur du maximum de vraisemblance pour σ, noté $\hat{\sigma}_{\text{Nu}}$, est simplement la moyenne empirique des excès, car pour $\gamma = 0$, la GPD se réduit à la loi exponentielle. Ainsi, $\hat{\sigma}_{\text{Nu}} = \frac{1}{N_u}\sum_{i=1}^{N_u} x_i$.

Dans la méthode des Moments pour une distribution GPD(γ,σ), l'estimateur des moments (proposé par Hosking et Wallis) est basé sur la relation suivante :

$$E\left[\left(1+\frac{\gamma X}{\sigma}\right)^r\right] = \frac{1}{1-r\gamma}$$

Si $1-r\gamma > 0$. À partir de cette relation, nous pouvons exprimer les paramètres σ et γ en fonction des deux premiers moments l_1 et l_2 :

$$\sigma = \frac{1}{2}l_1\left(1+\frac{l_1^2}{l_2-l_1^2}\right)$$

$$\gamma = \frac{1}{2}\left(1-\frac{l_1^2}{l_2-l_1^2}\right)$$

En substituant les expressions des moments empiriques l_1 et l_2, nous obtenons l'estimateur des moments pour le couple $(\hat{\sigma}_{\text{MOM}}, \hat{\gamma}_{\text{MOM}})$ comme suit :

$$\hat{\sigma}_{\text{MOM}} = \frac{1}{2}\rho_1\left(1+\frac{\rho_1^2}{\rho_2-\rho_1^2}\right)$$

$$\hat{\gamma}_{\text{MOM}} = \frac{1}{2}\left(1-\frac{\rho_1^2}{\rho_2-\rho_1^2}\right)$$

Où ρ_1 et ρ_2 sont les estimateurs empiriques des moments d'ordre 1 et 2 de l'échantillon.
La normalité asymptotique du couple $(\hat{\sigma}_{\text{MOM}}, \hat{\gamma}_{\text{MOM}})$ peut être établie sous la condition $\gamma < \frac{1}{4}$.

Dans la méthode des Moments Pondérés (EMP), nous utilisons les moments pondérés

$w_r(\gamma,\sigma)$ au lieu des moments classiques, surtout lorsque certains moments n'existent pas ou ne sont pas finis.

Définissons $w_r(\gamma,\sigma)$ comme l'espérance du r-ème moment sous la distribution $G_{\gamma,\sigma}$, où r est un nombre réel :

$$w_r(\gamma,\sigma) = E[X(\overline{G}^r_{\bar{\gamma},\sigma})]$$

Où $\overline{G}_{\bar{\gamma},\sigma} = 1 - G_{\gamma,\sigma}$.

Ensuite, nous pouvons exprimer $w_r(\gamma,\sigma)$ comme suit :

$$\begin{aligned} w_r(\gamma,\sigma) &= \int_{-\infty}^{\infty} x\,(\overline{G}^{-1}_{\bar{\gamma},\sigma}(x))^r\, d\overline{G}^{-1}_{\bar{\gamma},\sigma}(x) \\ &= \int_0^1 x\,\overline{G}^{-1}_{\bar{\gamma},\sigma}(y)\, y^r\, dy \\ &= \int_0^1 \frac{\sigma}{\gamma}\left(y^{-\gamma}-1\right) y^r\, dy \end{aligned}$$

Grâce à la dernière formulation et après quelques calculs, nous obtenons les estimations des paramètres σ et γ comme suit :

$$\sigma = \frac{2w_0 \cdot w_1}{w_0 - 2w_1}$$
$$\gamma = 2 - \frac{w_0}{w_0 - 2w_1}$$

Nous pouvons également estimer les moments pondérés empiriques $\hat{w}_r(\gamma,\sigma)$ pour $r = 0,1$ comme suit :

$$\hat{w}_r(\gamma,\sigma) = \frac{1}{N_u}\sum_{i=1}^{N_u} X_i \hat{F}^r(X_i) \quad \text{pour } r = 0,1$$

Où $\hat{F}$ est la fonction de répartition empirique de l'échantillon $X_1,\ldots,X_{N_u}$. Pour estimer γ et σ, nous remplaçons w_r par $\hat{w}_r$ pour $r = 0,1$.

Hosking et Wallis ont montré que lorsque $0 \leq \gamma \leq 0.4$ et pour des échantillons de petite taille, la méthode des Moments Pondérés (EMP) produit des estimateurs plus précis que la méthode du Maximum de Vraisemblance (EMV), avec des écarts-types plus faibles. Cependant, cette différence diminue avec l'augmentation de la taille de l'échantillon. De plus, Rootzén et Tajvidi ont révélé que pour $\gamma \geq 1/2$, l'EMP calcule des estimateurs fortement biaisés, contrairement aux estimateurs de l'EMV qui sont efficaces. Enfin, pour $\gamma \geq -1/2$, les conditions de régularité de l'EMV sont satisfaites et les estimateurs du Maximum de Vraisemblance $(\hat{\gamma}_{\mathrm{Nu}}, \hat{\sigma}_{\mathrm{Nu}})$ calculés sur l'échantillon des N_u excès sont asymptotiquement normaux.

Chapitre 3

Simulation

Introduction

Les valeurs extrêmes sont des événements rares mais souvent critiques dans de nombreux domaines tels que la finance, la météorologie, et la gestion des risques. La théorie des valeurs extrêmes (Extreme Value Theory, EVT) fournit un cadre statistique pour modéliser et analyser ces événements rares. En particulier, elle permet de comprendre et prédire les occurrences des valeurs extrêmes au-delà de ce qui est généralement observé.

La distribution de Cauchy est un exemple classique d'une distribution avec des queues lourdes, ce qui signifie qu'elle a une probabilité plus élevée de produire des valeurs extrêmes par rapport à des distributions comme la normale. La distribution de Cauchy, caractérisée par une densité de probabilité qui décroît lentement, est souvent utilisée dans des contextes où des valeurs très grandes ou très petites peuvent survenir.

Dans ce contexte, l'objectif de ce projet est de simuler des valeurs extrêmes issues d'une distribution de Cauchy et d'estimer ces valeurs en utilisant des méthodes statistiques appropriées en langage R. Le processus comprend la génération de données à partir de la distribution de Cauchy, puis l'application de techniques de la théorie des valeurs extrêmes, telles que l'analyse des maxima de blocs et l'estimation par dépassement de seuil.

3.1 Simulation de Données de Cauchy

Tout d'abord, nous allons générer des données à partir d'une distribution de Cauchy.

Pour simuler un échantillon de la loi de Cauchy en utilisant le langage R, vous pouvez utiliser la fonction **rcauchy**, qui est conçue pour générer des nombres aléatoires suivant une

distribution de Cauchy. Voici comment simuler un échantillon de 9000 valeurs de la loi de Cauchy :

```
# Charger les packages nécessaires
library(extRemes)
library(evir)
library(ismev)

# Définir les paramètres de la simulation
n <- 9000  # Nombre d'échantillons
location <- 0  # Paramètre de position
scale <- 1  # Paramètre d'échelle

# Simuler les données de la distribution de Cauchy
set.seed(123)  # Pour reproductibilité
data <- rcauchy(n, location, scale)
```

Avant d'appliquer les estimateurs présentés au chapitre deux, il est crucial de comprendre les caractéristiques des données générées. Nous débuterons par examiner les statistiques descriptives.

Les statistiques descriptives nous aident à saisir la distribution des données et à repérer d'éventuelles valeurs aberrantes. Une fois cette analyse préliminaire achevée, nous sommes prêts à mettre en œuvre les estimateurs.

```
> # Afficher un résumé des données simulées
> summary(data)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1618.113    -1.014     0.018     1.026     1.051  3905.195
```

Ces statistiques indiquent que les données couvrent une large plage de valeurs avec des valeurs extrêmes significatives. La présence de valeurs entre (-1618.113) et (3905.195) indique une distribution avec des valeurs aberrantes potentielles.

3.2 Estimation des valeurs extrêmes par méthode GEV

La méthode du maximum par blocs, utilisée pour estimer les paramètres de la distribution des valeurs extrêmes généralisée (GEV), est une approche couramment utilisée en statistique pour modéliser le comportement des valeurs extrêmes dans un ensemble de données. Voici comment elle fonctionne :

1. Division en blocs : Tout d'abord, l'ensemble de données est divisé en blocs de taille égale. Dans ce cas, la taille de bloc de n = 30, ce qui correspond à environ un mois de données. Cette taille de bloc est choisie pour capturer les variations mensuelles dans les données.

2. Sélection des maximums : Pour chaque bloc, le maximum est sélectionné. Cela crée un nouvel échantillon constitué des maximums de chaque bloc.
3. Estimation des paramètres : Une fois que nous avons cet échantillon de maximums de blocs, nous pouvons utiliser des méthodes statistiques pour estimer les paramètres de la distribution des valeurs extrêmes généralisée (GEV). Ces paramètres sont généralement la localisation, l'échelle et la forme de la distribution GEV.
4. Validation du modèle : Après avoir estimé les paramètres, il est important de valider le modèle pour vérifier s'il convient bien aux données observées. Cela peut impliquer des tests statistiques pour évaluer l'adéquation du modèle aux données.

```
# Diviser les données en blocs et trouver les maxima de chaque bloc
block_size <- 30  # Taille de chaque bloc
num_blocks <- n/block_size
block_maxima <- sapply(1:num_blocks, function(i) {
  max(data[((i-1) * block_size + 1):(i * block_size)])
})

summary(block_maxima)
```

```
> summary(block_maxima)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   1.318    7.221   15.430   83.760   40.948 3905.195
```

En R, on peut utiliser la fonction '**fevd**' pour estimer les paramètres du modèle GEV par la méthode du maximum de vraisemblance.

```
# Ajuster une distribution GEV aux maxima des blocs
gev_fit <- fevd(block_maxima, type = "GEV", method = "MLE")

# l'ajustement GEV
 gev_fit$results$par
```

— location : 10.575678

— scale : 10.919345

— shape : 1.028561

La validation du modèle GEV dans le contexte de l'estimation des valeurs extrêmes implique généralement plusieurs étapes,

Ce code appliquera la distribution GEV aux données et générera un graphique Q-Q pour montrer à quel point les données correspondent à la distribution supposée. Si les points dans le graphique suivent approximativement une ligne droite, cela signifie que les données suivent bien la distribution générale des valeurs extrêmes (GEV).

```r
# Afficher les résultats de l'ajustement GEV
plot(gev_fit,"qq")
```

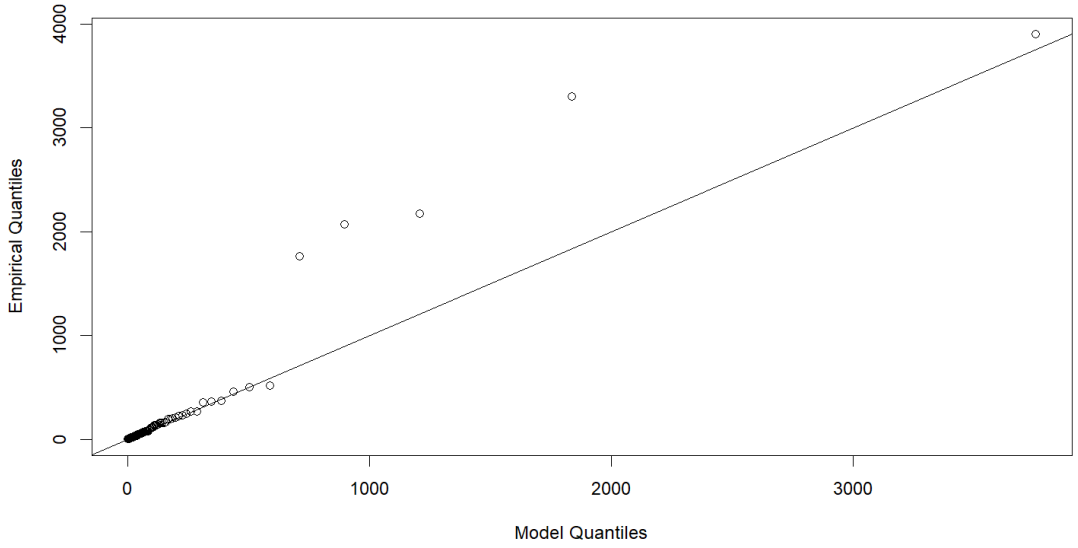


FIGURE 3.1 – Le graphique Q-Q

```r
# Afficher les résultats de l'ajustement GEV
plot(gev_fit,"density")
```

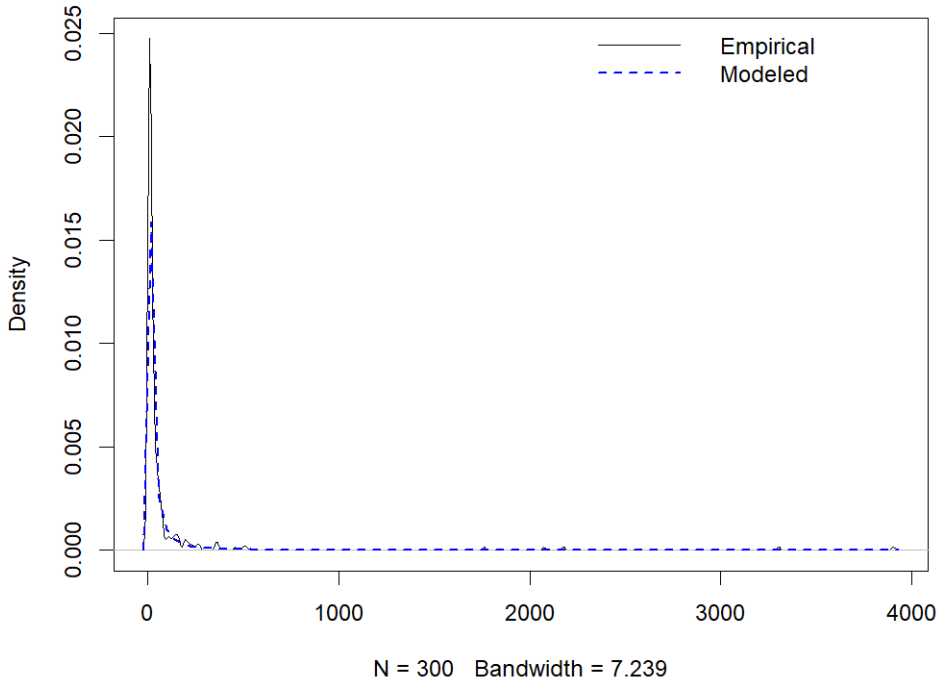


FIGURE 3.2 – La densité

D'après les figures (3.1) et (3.2), on peut dire que l'échantillon suit la loi de Fréchette

Niveau de retour

La figure suivant regroupe les quantiles extrêmes (niveau de retour) calculés pour quelques périodes de retour par les méthodes d'estimation MV (Maximum de Vraisemblance) :

```
# Afficher les résultats de l'ajustement GEV
plot(gev_fit,"rl")
```

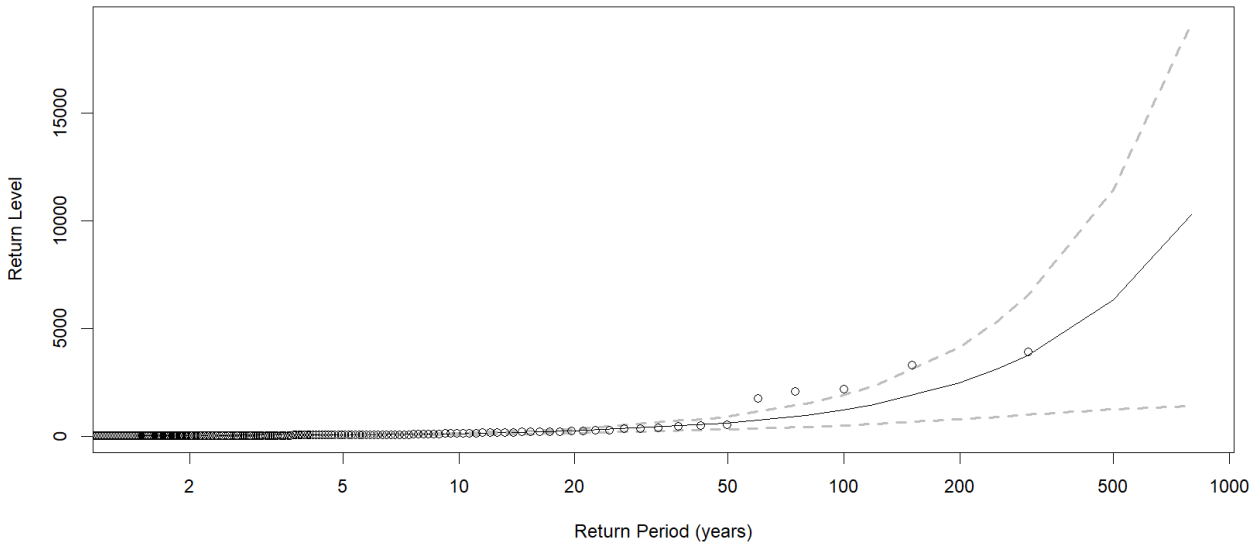


FIGURE 3.3 – La niveau de retour

Il ressort clairement du graphique que les niveaux de retour augmentent de manière significative avec l'augmentation de la période de retour. Cela signifie que les événements environnementaux extrêmes deviennent plus intenses à mesure qu'ils sont plus rares.

3.3 La méthode des Excès de Seuil POT

La détermination du seuil pour le modèle POT (Peaks Over Threshold) est une étape cruciale dans l'analyse des valeurs extrêmes. Le seuil choisi doit être suffisamment élevé pour que les valeurs au-dessus de ce seuil suivent une distribution de Pareto généralisée (GPD), tout en assurant qu'il y a suffisamment de données pour obtenir des estimations fiables.

Le choix du seuil u pour des estimateurs precis doit étre grand ; On va calculer u par des differentes méthodes.

En utilisant le logiciel **R**, on va tracer le graphe de la fonction moyenne des exces notée $e(u)$, on obtient le graphe suivant :

```
#creation graphique
meplot(data)
```

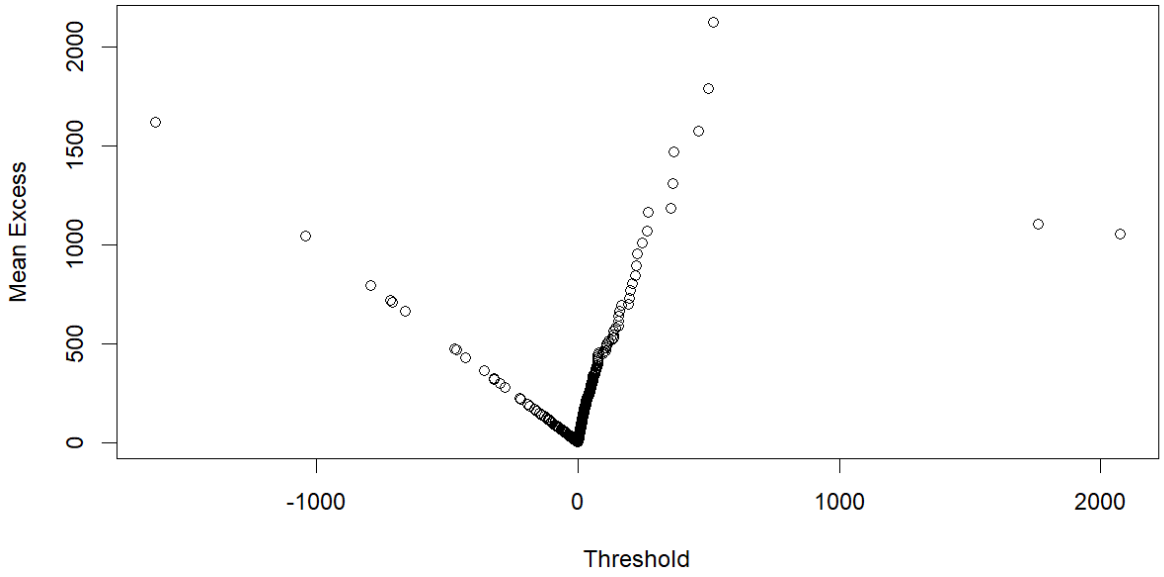


FIGURE 3.4 – La distribution moyenne des exces.

D'apres le graphe, le seuil 65 tel que le nuage de points $(u, e(u))$ soit approximativement lineaire pour $u > u_0$; Nous estimons la GPD par la méthode de moments Pondérés et nous obtenons le resultat suivant :

```
# Ajuster une distribution GPD aux excédents au-dessus du seuil
gpd_fit<-gpd(data,65,method = "pwm") # Estimation des Moments Pondérés (EMP)

# L'ajustement GPD
gpd_fit$par.ests
```

— xi : 0.7223586

— beta : 95.5039280

```
# Afficher les résultats de l'ajustement GPD
plot(gpd_fit)
```

Validation du modèle

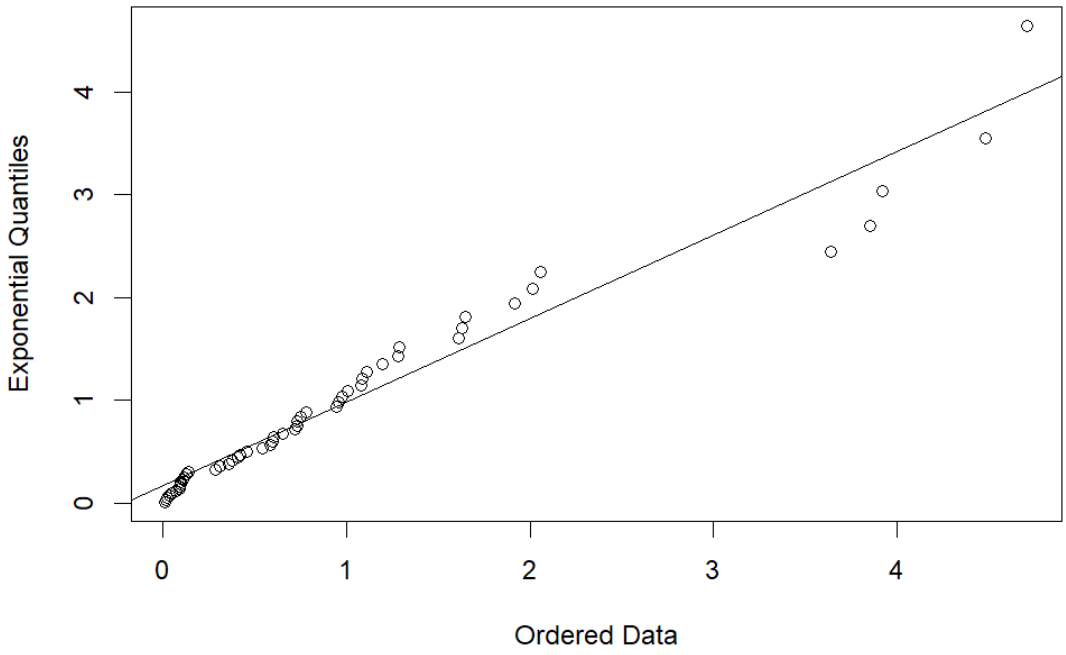


FIGURE 3.5 – Le QQ plot

Observez si les points suivent une ligne droite. Cela suggère une bonne concordance avec la distribution théorique.

3.4 Estimateur de Hill

Nous expliquerons comment appliquer la fonction de l'estimateur de Hill aux données simulées que nous avons générées dans la section précédente. Rappelons que la formule de l'estimateur de Hill pour un indice des valeurs extrêmes positif γ, est donné par :

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

Nous allons maintenant appliquer l'estimateur de Hill aux données simulées que nous avons générées. Nous devrons d'abord trier les données par ordre décroissant, puis sélectionner le nombre de valeurs extrêmes (k) à utiliser dans l'estimation.

```
hill_estimator <- function(data, k) {
  n <- length(data)
  sorted_data <- sort(data)
  log_ratios <- log(sorted_data[(n - k + 1):n] / sorted_data[n - k + 1])
  hill_general <- mean(log_ratios)
  return(hill_general)
}
```

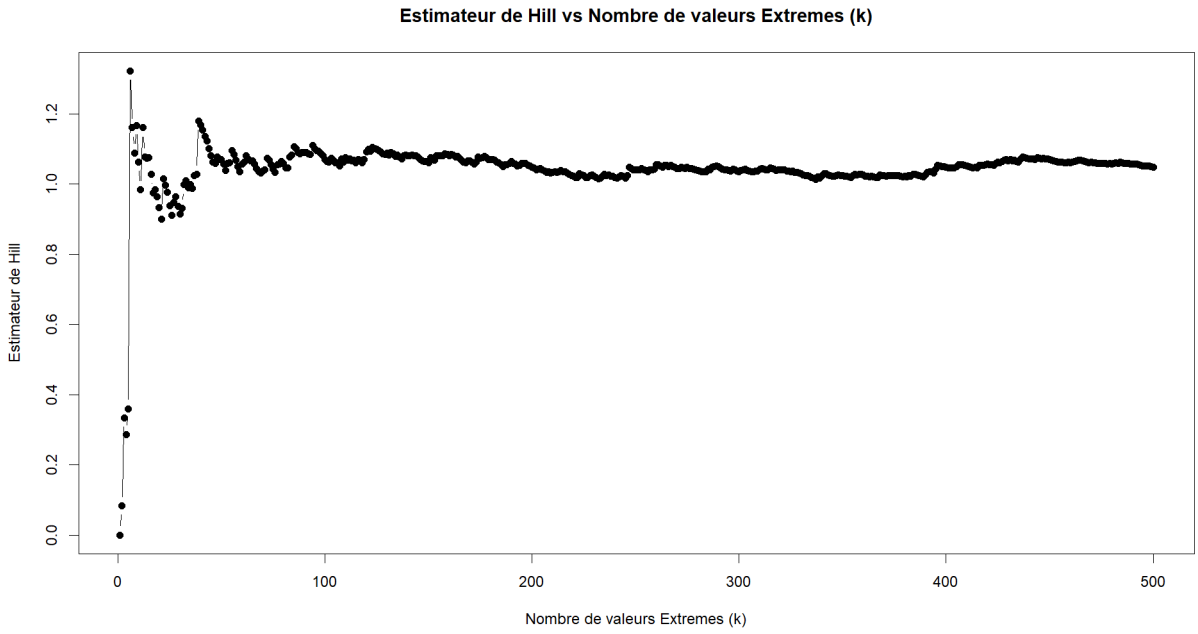
Pour analyser les résultats de l'estimateur de Hill en fonction du nombre de valeurs extrêmes (k) utilisées dans l'estimation, nous allons créer deux graphiques : l'un montrant l'évolution des valeurs de l'estimateur de Hill en fonction de (k) et l'autre montrant l'effet de (k) sur les statistiques d'ordre.

3.4.1 Graphique de l'Estimateur de Hill vs k

Nous allons d'abord créer un graphique qui montre comment l'estimateur de Hill varie en fonction du nombre de valeurs extrêmes utilisées (k)

```
# Calculer les estimations de l'indice de queue pour différentes valeurs de k
k_values <- seq(1, 100, by = 1)
hill_estimates <- sapply(k_values, function(k) hill_estimator(data, k))

# Créer le graphique
plot(k_values, hill_estimates, type = "b", pch = 19, col = "blue",
     xlab = "Nombre de Valeurs Extrêmes (k)", ylab = "Estimateur de Hill",
     main = "Estimateur de Hill vs Nombre de Valeurs Extrêmes (k)")
```



Le graphique montre comment l'estimateur de Hill varie en fonction du nombre de valeurs extrêmes (k). Nous pouvons observer une certaine stabilité de l'estimation lorsque k est dans une plage optimale. Si k est trop petit, l'estimation peut être très volatile car elle est basée sur un nombre limité de valeurs. Si k est trop grand, l'estimation peut devenir biaisée car elle inclut des valeurs qui ne sont pas réellement des extrêmes. voir 3.6

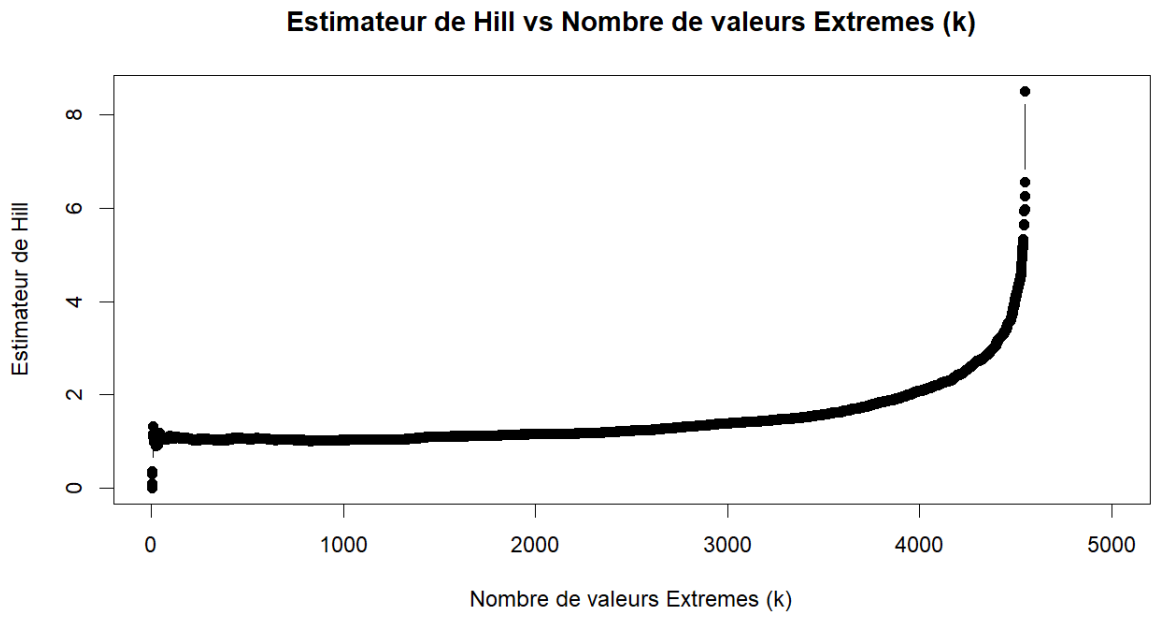


FIGURE 3.6 – estimateur de hill vs k

Nous avons maintenant estimé l'indice de queue ($\hat{\gamma}_n^{(H)}$) à l'aide de l'estimateur de Hill. Nous pouvons utiliser cette estimation pour comprendre la queue de la distribution et prendre des décisions éclairées sur les valeurs extrêmes.

```
# Sélectionner le nombre de valeurs extemes (k)
k<-100
xi_hat <- hill_estimator(data,k)
print(xi_hat)
```

Dans le cas $k = 100$ ona : $\hat{\gamma}_n^{(H)} \approx 1.069$.

3.5 Application de l'Estimateur de Pickands

L'estimateur de Pickands est une autre méthode pour estimer l'indice de queue, couramment utilisée en complément de l'estimateur de Hill. Dans cette section, nous appliquerons l'estimateur de Pickands aux données simulées et comparerons les résultats avec ceux de l'estimateur de Hill.

3.5.1 Application de l'Estimateur de Pickands aux Données Simulées

Nous allons maintenant appliquer l'estimateur de Pickands aux données simulées dans R.

```r
# Fonction de l'estimateur de pickands
pickands_estimator <- function(data, k){
  sorted_data <- sort(data, decreasing = T)
  xi_hat <- (1/log(2)) * log((sorted_data[k] -
            sorted_data[2*k])/ (sorted_data[2*k] - sorted_data[4*k]))
return(xi_hat)
}
# Calculer les estimations de l'indice de queue pour diff valeurs de k
k_values <- seq(1,1000, by=1)
# Créer les graphiques
pickands_estimates <- sapply(k_values, function(k) pickands_estimator(data, k))

pickands_plot <- ggplot(data.frame(k=k_values,
                        Pickands_Estimate = pickands_estimates),
                        aes(x = k, y = Pickands_Estimate)) +
  geom_line() +
  geom_point() +
  labs(title = "Estimateur de Pickands vs nbr de valeurs Extremes (k)",
       x = "nbr de valeurs Extremes (k)" , y="Estimateurs de Pickands")
# Afficher les graphiques
print(pickands_plot)
```

Le graphique montre comment l'estimateur de Pickands varie avec le nombre de valeurs extrêmes (k). Nous observons que l'estimateur de Pickands peut présenter des fluctuations similaires à celles de l'estimateur de Hill lorsque k est trop petit ou trop grand voir la fig 3.7

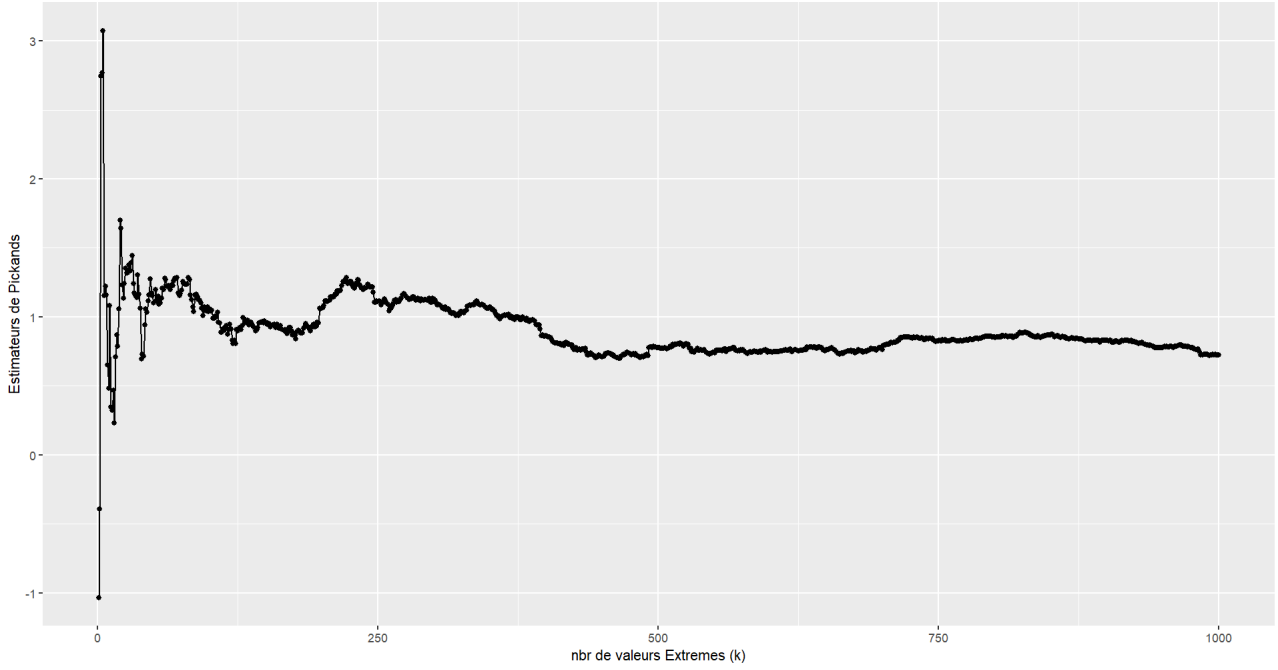


FIGURE 3.7 – Estimateur de Pickands vs Nombre de Valeurs Extrêmes (k)

3.6 Comparaison des Estimateurs de Hill et de Pickands

Nous allons maintenant combiner les estimations de Hill et de Pickands dans un seul graphique pour comparer les résultats.

```
# Combiner les estimations de Hill et de Pickands dans un seul data frame
combined_estimates <- data.frame(
  k = rep(k_values, 2),
  Estimate = c(hill_estimates, pickands_estimates),
  Method = rep(c("Hill", "Pickands"), each = length(k_values))
)

# Créer le graphique de comparaison
comparison_plot <- ggplot(combined_estimates, aes(x = k, y = Estimate, color = Method)) +
  geom_line() +
  geom_point() +
  labs(title = "Comparaison des Estimateurs de Hill et de Pickands",
  x = "Nombre de Valeurs Extrêmes (k)", y = "Estimation de l'Indice de Queue")

# Afficher le graphique de comparaison
print(comparison_plot)
```

Le graphique comparatif suivant montre comment les estimations de Hill et de Pickands diffèrent en fonction du nombre de valeurs extrêmes (k). Nous pouvons voir que parfois les estimations sont similaires, tandis que dans d'autres cas, elles diffèrent en fonction de la nature des données et des valeurs extrêmes.

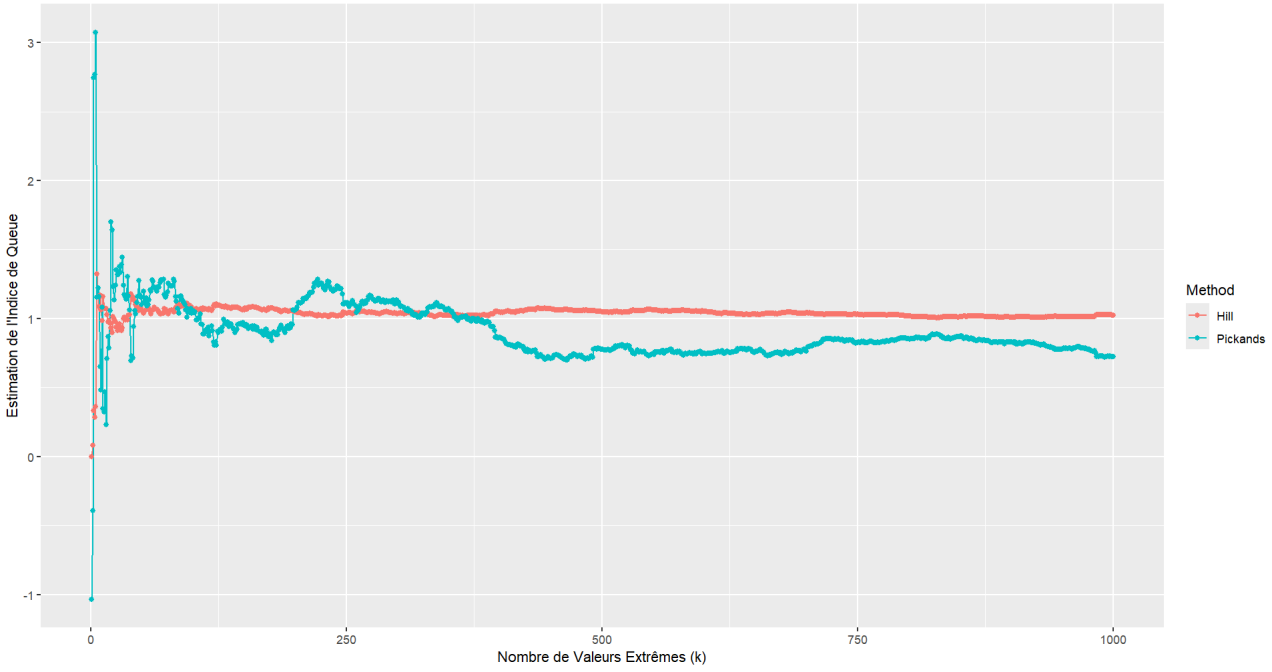


FIGURE 3.8 – Comparaison des Estimateurs de Hill et de Pickands

L'analyse visuelle des résultats montre clairement l'importance de choisir un k approprié pour l'estimateur de Hill. Un k trop petit ou trop grand peut conduire à des estimations imprécises de l'indice de queue. L'objectif est de trouver une plage de k où l'estimation est stable et représentative des extrêmes réels de la distribution.

3.7 Conclusion

À travers notre analyse, nous avons conclu que les estimations de Hill et de Pickands fournissent des perspectives précieuses sur les caractéristiques de la queue dans la distribution des données extrêmes. Nous avons également utilisé ces deux estimations pour obtenir une compréhension plus approfondie et plus exhaustive. Le choix optimal d'estimation dépend largement des caractéristiques des données et du nombre de valeurs extrêmes utilisées dans le calcul. Par conséquent, il est crucial d'adopter une approche intégrée qui combine ces deux méthodes pour obtenir une analyse complète et précise des données extrêmes. Cette intégration peut avoir un impact significatif sur la prise de décisions éclairées dans des domaines tels que la finance, l'assurance et l'hydrologie, où les valeurs extrêmes jouent un rôle crucial.

Conclusion Générale

Au cours des dernières années, la théorie des valeurs extrêmes a suscité un intérêt croissant tant sur le plan théorique que pratique, en raison de l'importance découverte par les chercheurs dans l'explication des événements rares et leur caractère prédictif. Dans ce mémoire, nous avons détaillé les résultats et définitions les plus importants de la théorie des valeurs extrêmes.

Dans le deuxième chapitre, nous avons expliqué les principales méthodes paramétriques et semi-paramétriques ainsi que la méthode des excès de seuil (POT) et d'autres techniques. Enfin, dans le dernier chapitre, nous avons effectué des études expérimentales des résultats présentés dans le deuxième chapitre en simulant un ensemble de données et en appliquant les résultats de la théorie des valeurs extrêmes.

Grâce à cette approche, nous visons à démontrer l'utilité et la pertinence des méthodes d'analyse des valeurs extrêmes dans la prévision des événements rares et la gestion des risques dans divers domaines d'application.

Bibliographie

[1] R. Fisher and L. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 180–190, 1928.

[2] B. Gnedenko, "Sur la distribution limite du terme maximum d'une série aléatoire," *Annals of Mathematics*, vol. 44, no. 3, pp. 423–453, 1943.

[3] L. De Haan and A. Ferreira, *Extreme Value Theory : An Introduction*. Springer Series in Operations Research and Financial Engineering, Springer Science and Business Media LLC, 2006.

[4] R. Von Mises, "La distribution de la plus grande de n valeurs," *Revue de Mathématique Union Interbalcanique*, vol. 1, pp. 141–160, 1936.

[5] A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *The Quarterly Journal of the Royal Meteorological Society*, vol. 81, no. 348, pp. 158–171, 1955.

[6] Y. Ziane, *Sur l'estimation non paramétrique de l'indice de variabilité et la distribution des densités à queue lourde*. PhD thesis, Université A. Mira, 2016.

[7] P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*. Berlin : Springer, 1997.

[8] J. Pickands III, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, p. 119, 1975.

[9] R. Smith, "Estimating tails of probability laws," *The Annals of Statistics*, vol. 3, pp. 1174–1207, 1987.

[10] R. Reiss and M. Thomas, *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Bâle : Birkhäuser, 2007.

[11] A. Belkema and L. De Haan, "Residual lifetime at great age," *Annals of Probability*, pp. 792–80, 1974.