

République Algérienne Démocratique et Populaire  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique  
UNIVERSITÉ KASDI MERBAH OUARGLA

FACULTÉ DE MATHÉMATIQUES ET SCIENCES DE LA  
MATIÈRE

DÉPARTEMENT DE MATHÉMATIQUES



Mémoire présenté En Vue De L'obtention Du

**DIPLÔME DE MASTER**

EN MATHÉMATIQUES

**Option** : Probabilité et Statistique

Par

**REZIG Ikram**

Intitulé

Modélisation des queues de distribution par la  
loi de Pareto généralisée

Membres du jury

MANSOUL Brahim	M. A. A	UKMO	Président
BOUREDJI Hind	M. C. B	UKMO	Examineur
ARBIA Hanane	M. C. B	UKMO	Rapporteur

2023-2024

# Table des matières

Table des matières	i
Dédicace	iii
Remerciements	iv
Table des figures	v
Notations et abreviations	vi
Introduction	1
<b>1 Théorie des valeurs extrêmes</b>	<b>3</b>
1.1 Statistique d'ordre . . . . .	3
1.1.1 Définition et notation . . . . .	3
1.1.2 Densité et loi des statistiques d'ordre . . . . .	4
1.1.3 Convergence de la fonction de répartition empirique . . . . .	6
1.2 Fonction à variations régulière . . . . .	8
1.3 Autres définitions . . . . .	9
1.4 Théorème Fisher-Tippet(1928) . . . . .	11
1.5 Caractérisation des domaines d'attraction . . . . .	11
1.5.1 Domaine d'attraction de Fréchet . . . . .	12
1.5.2 Domaine d'attraction de Weibull . . . . .	12
1.5.3 Domaine d'attraction de Gumbel . . . . .	13
1.6 Distribution de Pareto généralisée . . . . .	15
1.6.1 Distribution conditionnelle des excès . . . . .	15
1.6.2 Distribution de Paréto généralisée . . . . .	17
1.6.3 Théorème de Balkema-de Haan-Pickands . . . . .	17
<b>2 Estimation dans les modèles GPD</b>	<b>22</b>
2.1 Estimation des paramètres de la GPD . . . . .	22
2.1.1 Sélection du seuil . . . . .	22

2.1.2	Estimation par le Maximum de Vraisemblance . . . . .	23
2.1.3	Méthode des moments (MOM) . . . . .	25
2.1.4	Méthode des Moments Pondérés (EMP) . . . . .	25
2.1.5	Méthode des L-moments . . . . .	27
2.2	Estimation par des méthodes semi-paramétrique . . . . .	30
2.2.1	Estimateur de Hill . . . . .	30
2.2.2	Estimateur des Moments : . . . . .	32
2.2.3	Estimateur de Pickands . . . . .	33
2.2.4	Choix du nombre optimal de statistiques d'ordre extrêmes . . . . .	34
2.3	Estimation des quantiles extrêmes et période de retour . . . . .	35
2.3.1	Estimation des quantiles extrêmes . . . . .	36
2.3.2	Estimation de période de retour . . . . .	37
<b>3</b>	<b>Application : Modélisation via la distribution de Pareto généralisée (GPD)</b>	<b>38</b>
3.1	Les données . . . . .	39
3.1.1	Analyse descriptive de la série . . . . .	39
3.1.2	L'ajustement par la loi normale . . . . .	42
3.2	Sélection de seuil . . . . .	43
3.3	Estimation des paramètres du modèle GPD . . . . .	44
3.4	Estimation par des méthodes semi-paramétrique . . . . .	44
3.5	Estimation de niveau de retour . . . . .	45
3.6	Le programme d'application sous R . . . . .	46
	<b>Conclusion générale</b>	<b>49</b>
	<b>Bibliographie</b>	<b>50</b>

# Dédicace

Je dédie ce modeste travail

A mon grand amour, ma mère qui a sacrifié sa vie pour notre bonheur et notre réussite.

A mon père à qui je témoigne de l'affection et du respect.

À mon frère Abd el Rahman.

À mes chères sœurs Inas, Ghoufrane, Mais, frdwousse.

À toute la famille Rezig.

À tous mes amis.

## Remerciements

Je tiens tout d'abord à remercier Allah le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce travail.

Au début de ce mémoire, je tiens à exprimer ma profonde gratitude et reconnaissance à tous ceux qui ont contribué à la réalisation de ce travail.

Je remercie tout particulièrement mon encadrante Mme **ARBIA Hanane**, pour son soutien constant et ses précieux conseils qui ont grandement contribué au développement de cette recherche. Ce travail n'aurait pas atteint ce niveau sans ses orientations judicieuses et sa patience face à mes questions.

Je souhaite également remercier les membres du jury pour le temps et les efforts qu'ils ont consacrés à la lecture de ce mémoire et pour leurs remarques constructives qui ont aidé à améliorer ce travail.

Je n'oublie pas de remercier profondément ma famille, mes parents, pour leur soutien et leurs encouragements inébranlables tout au long de mes études. Ils ont toujours été mon pilier et ma source de force.

Je remercie également mes collègues et amis pour leur soutien moral et leur aide dans les moments difficiles. Merci à vous tous d'avoir été à mes côtés et de m'avoir encouragé à persévérer.

Enfin, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

**Rezig Ikram**

# Table des figures

1.1	Densités standard des valeurs extrêmes. . . . .	15
1.2	Les excès . . . . .	16
1.3	Densités de probabilité de GPD avec différentes valeurs de $\xi$ . . . . .	20
1.4	Fonction de survie de GPD avec différentes valeurs de $\xi$ . . . . .	21
3.1	Pertes des compagnies d'assurance automobile du 31/1/2020 à 31/12/2023	40
3.2	Histogramme de pertes de la compagnie d'assurance automobile. . . . .	41
3.3	La distribution moyenne des exés. . . . .	43
3.4	Représentation graphique de niveau de retour dans le cas GPD. . . . .	45

## Notations et abriviations

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

$v.a$	: Variable aléatoire.
$i.i.d$	: Indépendant et identiquement distribué
$F$	: Fonction de répartition.
$F_n$	: Fonction de répartition empirique.
$f$	: fonction de densité.
$(X_{1,n}, \dots, X_{n,n})$	: Statistique d'ordre $(X_1, \dots, X_n)$ .
$E[X]$	: Espérance mathématique ou moyenne du v.a $X$ .
$Var(X)$	: Variance de la v.a $X$ .
$GEVD$	: Distribution des valeurs extrêmes généralisée.
$IVE$	: Indice des valeurs extrêmes.
$GPD$	: Distribution de Pareto Généralisée
$H_\xi$	: Distribution des valeur sextrêmes.
$D(H_\xi)$	: Domaine d'attraction de H.
$Q$	: Fonction de quantile.
$Q_n$	: Fonction de quantile empirique.
$\binom{n}{k}$	: combinaisons de k éléments parmi n.
$RV_\sigma$	: Variation régulière avec l'indice $\sigma$ .
$l(\cdot)$	: Fonction à variation lente.
$u$	: Seuil.

# Introduction

Au cours des dernières années, nous avons pu observer dans la recherche scientifique, une modélisation des événements rares. Ces derniers sont des événements dont la probabilité d'apparition est trop faible, c'est-à-dire ils se trouvent dans les queues des distributions. La modélisation des événements extrêmes tels (ouragan, tremblement de terre, finance, les crues, assurance, ou encore en météorologie etc..) est aujourd'hui un champ de recherche particulièrement actif.

Ces événements extrêmes peuvent causer des dégâts humains et matériels considérables. De telles catastrophes ne peuvent pas toujours être évitées. Cependant, la société peut prendre des actions préventives pour minimiser leurs effets. Pour cela, les spécialistes ont à leur disposition la théorie statistique des valeurs extrêmes. Elle donne un résultat très intéressant parce qu'il est d'une portée générale.

La Théorie des Valeurs Extrêmes est une branche des statistiques apparue entre 1920 et 1943 grâce à Fréchet (1927); Fisher et Tippett (1928) et Gnedenko (1943). Cette branche a connue un grand essor depuis les années cinquante bien que le théorème principal ait été énoncé par Fisher et Tippett dès 1928. Elle a pour but de modéliser et de décrire la survenue et l'intensité des événements extrêmes, c'est à dire qui présentent des variations de très grande amplitude (ayant une faible probabilité d'apparition).

Cette théorie est largement utilisée, elle a trouvé des champs d'application en :

- Hydrologie, domaine dans lequel la prévision des crues et la connaissance de leurs débits est particulièrement importante pour la conception des aménagements des cours d'eau et la protection des zones urbaines inondables, souvent théâtre d'épisodes de crues violentes causant d'importants dommages économiques.
- Climatologie : étude des événements climatiques extrêmes (précipitations, températures, chutes de neige), modélisation des grands feux de forêt.
- Assurance : survenue des sinistres d'intensité exceptionnelle (ouragan Katrina en 2005, importants incendies en risque industriels, sinistres grave en responsabilité civile automobile).
- Finance : fortes variations du cours d'actifs financiers, gestion risque opérationnel des banques (la crise fin des années 2000).

Notre travail est organisé en trois chapitres et se répartit comme suit :



- Le premier chapitre, on présente tout d’abord quelques définitions de certains outils nécessaires dans la théorie des valeurs extrêmes tels que (statistique d’ordre, fonction à variations régulière, caractérisation des domaines d’attraction, distribution de Pareto généralisée ect ...). En plus les deux théorèmes sont essentiels à la compréhension de la théorie des valeurs extrêmes : celui de Fisher-Tippett et celui de Balkema-de Haan-Pickands.
- Dans le deuxième chapitre, nous présentons différentes méthodes d’estimation de l’indice des valeurs extrêmes et des quantiles extrêmes. On distingue deux méthodes d’estimation : la première, appelée "Estimation paramétrique", est basée sur la Distribution de Pareto généralisée (GPD) utilisées dans la littérature (méthode du maximum de vraisemblance, méthode des moments, méthode des moments de probabilité pondérés, méthode des moments linéaires et la deuxième, appelée "Estimation semi-paramétrique" telle que (Hill, Moment, Pickands), tout cela après avoir détermination du seuil par méthode Mean-Excess plot.
- Le troisième chapitre est consacré à la partie pratique de la réalisation de notre application. Malgré les difficultés rencontrées dans ce domaine, nous avons appliqué la modélisation de la distribution généralisée de Pareto (GPD) à des données réelles "pertes des compagnies d’assurance automobile". Les données sont récoltées durant une période entre 2020 et 2023 par la direction du concessionnaire automobile touggourt. Dans le but d’estimer le quantile extrême et la période de retour.

# Chapitre 1

## Théorie des valeurs extrêmes

La théorie des valeurs extrêmes a été développée pour l'estimation de probabilités d'occurrences d'évènements rares. Elle permet d'extrapoler le comportement de la queue de distribution à partir des plus grandes données observées (les données extrêmes de l'échantillon).

Ce chapitre réservé a une introduction sur les statistiques d'ordre, les valeurs extrêmes et les queues de distributions. Nous énonçons les principaux résultats concernant les distributions limites des plus grandes observations d'un échantillon ainsi que les domaines d'attractions [1],[6],[10],[18],[7],[8],[20],[14],[9].

### 1.1 Statistique d'ordre

Les statistiques d'ordre jouent un rôle très important dans la théorie des valeurs extrêmes car elles (plus précisément ceux du haut) renseignent sur la répartition queue. Dans cette section, nous résumerons certaines de leurs propriétés et résultats.

Soit  $(X_1, \dots, X_n)$  une suite de variables aléatoires indépendantes et identiquement distribuées (*iid*) de fonction de répartition  $F$ , ( $F(x) = P(X_n \leq x)$  pour  $x \in \mathbb{R}$ ).

#### 1.1.1 Définition et notation

**Définition 1.1.1** *On appelle statistique d'ordre notées  $X_{(1,n)}, X_{(2,n)}, \dots, X_{(n,n)}$  les variables aléatoires ordonnées :*

$$\min(X_1, X_2, \dots, X_n) = X_{(1,n)} \leq X_{(2,n)} \leq \dots \leq X_{(n,n)} = \max(X_1, \dots, X_n).$$

- La statistique d'ordre de l'échantillon  $(X_1, \dots, X_n)$  est le réarrangement croissant (aléatoire) de  $(X_1, \dots, X_n)$  telle que  $X_1 < X_2 < \dots < X_n$  que l'on note  $(X_{(1,n)}, \dots, X_{(n,n)})$ .

Le vecteur  $(X_{1,n}, \dots, X_{n,n})$  est appelé l'échantillon ordonné associé à l'échantillon  $(X_1, \dots, X_n)$  et  $X_{k,n}$  étant la  $k^{\text{ième}}$  statistique d'ordre.

Dans un échantillon de taille  $n$  deux statistiques d'ordre sont particulièrement intéressantes pour l'étude des événements extrêmes ce sont :

$$X_{1,n} = \min(X_1, \dots, X_n) \quad \text{et} \quad X_{n,n} = \max(X_1, \dots, X_n).$$

En règle générale, les résultats de minima peuvent être déduits des résultats correspondant à maxima par la relation triviale :

$$\min(X_1, \dots, X_n) = -\max(X_1, \dots, X_n).$$

### 1.1.2 Densité et loi des statistiques d'ordre

*David (1970) et Balakrishnan et Clifford Cohen (1991)* montrent que l'expression de la distribution de  $X_{i,n}$  est :

$$\begin{aligned} F_{X_{i,n}}(x) &= P(X_{i,n} \leq x) \\ &= P(\text{au moins } i \text{ de } X_1, X_2, \dots, X_n \text{ sont inférieures à } x) \\ &= \sum_{r=i}^n P(\text{exactement } r \text{ de } X_1, X_2, \dots, X_n \text{ sont inférieures à } x) \\ &= \sum_{r=i}^n C_n^r [F(x)]^r [1 - F(x)]^{n-r} \end{aligned}$$

donc

$$F_{X_{i,n}}(x) = \sum_{r=i}^n C_n^r [F(x)]^r [1 - F(x)]^{n-r}.$$

Alors, on déduit que la fonction de densité est de la forme suivant :

$$f_{X_{i,n}}(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x).$$

Pour les statistique d'ordre extrême, on obtient les expressions suivantes :

$$F_{X_{1,n}}(x) = 1 - [1 - F(x)]^n.$$

et

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

Pour la statistique du minimum et

$$F_{X_{n,n}}(x) = [F(x)]^n$$

et

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

pour la statistique du maximum.

**Preuve** En utilisant la propriété d'indépendance des variables aléatoires  $X_1, X_2, \dots, X_n$  nous en déduisons que les lois de minimum  $X_{1,n}$  et de maximum  $X_{n,n}$  de la statistique d'ordre associées à l'échantillon  $X_1, X_2, \dots, X_n$  sont :

$$\begin{aligned} F_{X_{1,n}}(x) &= P(X_{1,n} \leq x) = 1 - P(X_{1,n} > x) \\ &= 1 - P(\min(X_1, \dots, X_n) > x) \\ &= 1 - P\left[\bigcap_{i=1}^n (X_i > x)\right] \\ &= 1 - \prod_{i=1}^n P(X_i > x) \\ &= 1 - \prod_{i=1}^n [1 - P(X_i \leq x)] \\ &= 1 - \prod_{i=1}^n [1 - F(x)] \end{aligned}$$

Alors

$$F_{X_{1,n}}(x) = 1 - [1 - F(x)]^n.$$

Puis

$$\begin{aligned} F_{X_{n,n}}(x) &= P(X_{n,n} \leq x) \\ &= P(\max(X_1, \dots, X_n) \leq x) \\ &= P\left[\bigcap_{i=1}^n (X_i \leq x)\right] \\ &= \prod_{i=1}^n P(X_i \leq x) \\ &= \prod_{i=1}^n F(x) \end{aligned}$$

Alors

$$F_{X_{n,n}}(x) = [F(x)]^n$$

■

De ces résultats, nous en tirons la conclusion que le maximum  $F_{X_{n,n}}$  est une variable aléatoire dont la fonction de répartition correspond à  $F^n$ .

### 1.1.3 Convergence de la fonction de répartition empirique

**Définition 1.1.2** On appelle fonction de répartition empirique associée à un échantillon  $(X_1, \dots, X_n)$ , la fonction  $F_n$  définie par :

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n I_{\{X_k \leq x\}}.$$

Il existe autre définition de  $F_n$  :

$$F_n(x) = \begin{cases} 0 & \text{si } x \leq X_{1,n} \\ \frac{k-1}{n} & \text{si } X_{k-1,n} \leq x \leq X_{k,n}, 2 \leq k \leq n \\ 1 & \text{si } x \geq X_{n,n} \end{cases}$$

**Théorème 1.1.1** (Glivenco-Cantelli). Soit  $X_1, \dots, X_n$  un échantillon i.i.d de fonction de répartition continue  $F$ . La fonction de répartition empirique  $F_n$  converge presque sûrement uniformément vers  $F$  quand  $n \rightarrow \infty$  :

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \rightarrow \infty]{p.s} 0.$$

Pour la preuve, on aura besoin du lemme suivant.

**Lemme 1.1.1** Si  $X$  est une variable aléatoire réelle de fonction de répartition  $F$ , alors pour tout  $x \in \mathbb{R}$ ,

$$1_{\{X \leq x\}} \stackrel{p.s.}{=} 1_{\{F(X) \leq F(x)\}}.$$

En particulier, si  $F$  est continue, alors  $U := F(X)$  est de loi uniforme sur  $]0, 1[$ .

**Preuve** du lemme. Comme  $F$  est croissante, on sait que

$$\{F(X) \leq F(x)\} = \{X \leq x\} \cup \{F(X) \leq F(x), x < X\}.$$

Soit  $u = \sup\{y : F(y) \leq F(x)\}$ , on distingue deux cas. Si  $F(u) = F(x)$  alors

$$P(F(X) \leq F(x), x < X) = P(F(x) \leq F(x), x < X \leq u) \leq F(u) - F(x) = 0.$$

Sinon, soit  $u_n$  une suite croissante qui tend vers  $u$  et telle que  $F(u_n) = F(x)$  pour tout  $n \in \mathbb{N}$ . On a

$$P(F(X) \leq F(x), x < X) = \lim_{n \rightarrow \infty} P(F(x) \leq F(x), x < X \leq u_n) \leq \lim_{n \rightarrow \infty} F(u_n) - F(x) = 0,$$

où la première égalité découle de l'axiome des probabilités  $\lim_{n \rightarrow \infty} P(\cup_{k=1}^n A_k) = P(\cup_{k=1}^{\infty} A_k)$ . On déduit  $1_{\{X \leq x\}} \stackrel{p.s.}{=} 1_{\{F(X) \leq F(x)\}}$ . Si  $F$  est continue, du fait que  $\lim_{x \rightarrow -\infty} F(x) = 0$  et  $\lim_{x \rightarrow \infty} F(x) = 1$ , il existe, pour tout  $u \in ]0, 1[$ , un (ou plusieurs)  $x \in \mathbb{R}$  tel que  $F(x) = u$ . Par conséquent, l'égalité  $P(F(X) \leq u) = u$  est vérifiée pour tout  $u \in ]0, 1[$ , ce qui prouve que  $F(X)$  est de loi uniforme sur  $]0, 1[$ . ■

La preuve du théorème de Glivenko-Cantelli utilise également le théorème suivant, dû à Polya (parfois appelé théorème de Dini) que l'on admettra.

**Théorème 1.1.2** Une suite  $F_n$  de fonctions croissantes définies sur un compact  $[a, b] \subset \mathbb{R}$  qui converge simplement vers une fonction  $F$  continue sur  $[a, b]$ , converge uniformément vers  $F$ .

**Preuve** du théorème de Glivenko-Cantelli. D'après le lemme

, on a

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}} - F(x) \right| \stackrel{p.s.}{=} \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n 1_{\{F(X_i) \leq F(x)\}} - F(x) \right|.$$

Par hypothèse, on sait que  $F(x)$  peut prendre toutes les valeurs de l'intervalle  $]0, 1[$ . En posant  $u = F(x)$  et  $U_i = F(X_i)$  qui suit une loi uniforme sur  $[0, 1]$ , on a donc

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \stackrel{p.s.}{=} \sup_{u \in [0, 1]} \left| \frac{1}{n} \sum_{i=1}^n 1_{\{U_i \leq u\}} - u \right|,$$

les valeurs en  $u = 0$  et  $u = 1$  étant trivialement 0. Il s'agit de montrer le résultat pour un échantillon  $U_1, \dots, U_n$  de variables aléatoires uniformes sur  $[0, 1]$ . On suppose donc sans perte de généralité que

$$F_n(u) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{U_i \leq u\}} \quad \text{et} \quad F(u) = u, \quad u \in [0, 1].$$

Soit  $q \in \mathbb{Q} \cap ]0, 1[$ , on sait que  $P(F_n(q) \rightarrow q) = 1$ . Comme une union dénombrable d'ensembles de probabilité nulle est de probabilité nulle, on déduit que l'événement  $\Omega_0 := \{\forall q \in \mathbb{Q} \cap ]0, 1[, F_n(q) \rightarrow F(q)\}$  est de probabilité 1. Soit  $u \in ]0, 1[$  et  $(u_k)_{k \in \mathbb{N}}$  et  $(v_k)_{k \in \mathbb{N}}$  deux suites rationnelles adjacentes de limite  $u$  telles que  $u_k \leq u \leq v_k$ ,  $\forall k \in \mathbb{N}$ . La croissance de  $F_n$  entraîne

$$\forall k \in \mathbb{N}, F_n(u_k) \leq F_n(u) \leq F_n(v_k).$$

En faisant tendre  $n$  vers l'infini, on déduit que sur  $\Omega_0$ ,

$$F(u_k) \leq \liminf_{n \rightarrow \infty} F_n(u) \leq \limsup_{n \rightarrow \infty} F_n(u) \leq F(v_k).$$

Comme le résultat est vrai pour tout  $k$ , il reste vrai en faisant tendre  $k$  vers l'infini. On sait de plus que  $F_n(0) = 0$  et  $F_n(1) = 1$  presque sûrement. On obtient  $\Omega_0 = \{\forall u \in [0, 1], F_n(u) \rightarrow u\}$ . Finalement, d'après le théorème 1.1.2,  $F_n$  converge uniformément vers  $F$  sur  $\Omega_0$ . ■

## 1.2 Fonction à variations régulière

**Définition 1.2.1** Une fonction  $U$  positive est à variation régulière d'indice  $\delta \in \mathbb{R}$  à l'infini, et on note  $U(\cdot) \in RV_\delta$

– si pour tout  $\lambda \neq 0$

$$\lim_{x \rightarrow \infty} \frac{U(\lambda x)}{U(x)} = \lambda^\delta.$$

Dans le cas  $\delta = 0$ , on dit que la fonction  $U(\cdot)$  est à variations lentes, et on note  $l(\cdot) \in RV_0$ .

$$\lim_{x \rightarrow \infty} \frac{l(\lambda x)}{l(x)} = 1.$$

– **Remarque 1.2.1** Une fonction à variations régulières d'indice  $\delta$  peut toujours s'écrire sous la forme  $\mathbf{x}^\delta l(x)$  où  $l$  est une fonction à variations lentes.

**Lemme 1.2.1** Soit  $U$  est une fonction à variations régulière d'indice  $\delta \in \mathbb{R}$  alors pour tout  $0 < a < b$

$$\lim_{x \rightarrow \infty} \sup_{\lambda \in [a, b]} \left| \frac{U(\lambda x)}{U(x)} - \lambda^\delta \right| = 0.$$

**Lemme 1.2.2** *Ce lemme donne un résultat sur l'inverse d'une fonction à variation régulières.*

- \* si  $U$  est une fonction à variations régulières d'indice  $\delta > 0$  alors  $U^{-1}$  est une fonction à variations régulières d'indice  $1/\delta$ .
- \* si  $U$  est une fonction à variations régulières d'indice  $\delta < 0$  alors  $U^{-1}$  est une fonction à variations régulières d'indice  $-1/\delta$ .

**Théorème 1.2.1** *Toute fonction à variation lente  $l$  s'écrit sous la forme*

$$l(x) = c(x) \exp \left( \int_1^x \frac{\varepsilon(t)}{t} dt \right),$$

où  $c > 0$  et  $\varepsilon$  sont deux fonctions mesurables telles que :

$$\lim_{x \rightarrow \infty} c(x) = c > 0 \quad \text{et} \quad \lim_{x \rightarrow \infty} \varepsilon(x) = 0.$$

## 1.3 Autres définitions

### 1) Inverse généralisé

Soit  $F : \mathbb{R} \rightarrow ]a, b[$  ( $-\infty \leq a \leq b \leq +\infty$ ) une fonction croissante.

On appelle inverse généralisé de  $F$ , l'application notée  $F^{\leftarrow}$  définie par :

$\forall p \in ]a, b[$

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

L'inverse généralisé  $F^{\leftarrow}$  coïncide avec l'inverse lorsque la fonction  $F$  est strictement croissante et continue.

### 2) Fonction survie ou fonction de queue

On appelle fonction de survie ou fonction de queue de la variable aléatoire  $X$ , la fonction  $\bar{F} : \mathbb{R} \rightarrow [0, 1]$  définie par :

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

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

### 3) Point extrême



On note par  $x_F$  (resp.  $x_F^*$ ) le point extrême supérieur (resp. inférieur) de la distribution  $F$  (i.e la plus grande valeur possible pour  $X_{k,n}$  peut prendre la valeur  $+\infty$  (resp.  $-\infty$ ) au sens où

$$x_F := \sup\{x : F(x) < 1\} \leq \infty,$$

respectivement

$$x_F^* := \inf\{x : F(x) > 0\}.$$

Le comportement asymptotique de la loi de  $X_{n,n}$  est donné, dans certaines conditions sur la queue de distribution.

Soit  $X_1, \dots, X_n$  un échantillon de  $n$  variables aléatoires indépendantes et de même fonction de répartition  $F$ . Dans la suite on note  $X_{1,n}, \dots, X_{n,n}$  échantillon ordonné associé à cette échantillon.

#### 4) Quantile d'ordre $p$

On appelle quantile ou fractile d'ordre  $p$ , le nombre définie par :

$$x_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Si  $F$  est strictement croissante et continue, alors  $x_p$  est l'unique nombre réel tel que :

$$F(x_p) = p.$$

#### 5) Fonction quantile

On appelle fonction quantile de la variable aléatoire  $X$  la fonction  $Q$  définie par :

$$\forall p \in ]0, 1[$$

$$Q(p) = F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

#### 6) Fonction quantile de queue

On appelle fonction quantile de queue de la variable aléatoire  $X$  la fonction  $U$  :

$$]1, \infty[ \longrightarrow \mathbb{R} \text{ définie par :}$$

$$\forall x \in ]1, \infty[$$

$$U(x) = Q\left(1 - \frac{1}{x}\right) = \left(\frac{1}{1 - F}\right)^{\leftarrow}(x).$$

Où  $Q$  est la fonction quantile de  $X$ .

#### 7) Quantile extrême

On appelle quantile extrême , le quantile d'ordre  $(1 - p)$  , définie par :

$$\begin{aligned} x_p &= \inf\{x \in \mathbb{R} : F(x) \geq 1 - p\} \\ &= F^{\leftarrow}(1 - p). \end{aligned}$$

où  $p$  proche de zéro.

## 1.4 Théorème Fisher-Tippet(1928)

**Théorème 1.4.1** *Sous certaines conditions de régularité sur la fonction de répartition  $F$  s'il existe un paramètre réel  $\alpha$  et deux suites  $(a_n)_{n>0}$  et  $(b_n)_{n>0}$  tels que  $a_n > 0$  et  $b_n \in \mathbb{R}$  et pour tout  $x \in \mathbb{R}$*

$$\lim_{n \rightarrow \infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n \rightarrow \infty} F(a_n x + b_n)^n = H_\alpha(x)$$

alors  $H_\alpha(x)$  est l'une des trois distributions suivantes :

*Gumbel :*

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

*Fréchet :*

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

*Weibull :*

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x^\alpha)) & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases} \alpha < 0$$

$H_\alpha$  est appelée fonction de répartition de la loi des valeurs extrêmes.

où  $H$  est non-dégénérée, alors  $H$  est un membre de la famille DGVE (Distribution Généralisée des Valeurs Extrêmes) définie par :

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

## 1.5 Caractérisation des domaines d'attraction

On va pouvoir caractériser les différents domaines d'attraction. Sachant la distribution  $F$ , on voudrait connaître son domaine d'attraction et ses constantes de normalisation. On rappelle les conditions nécessaires et suffisantes sur une fonction de répartition pour qu'elle appartienne à un domaine d'attraction.

### 1.5.1 Domaine d'attraction de Fréchet

Rappelons que le domaine d'attraction de Fréchet contient les lois dont la fonction de survie est à décroissance polynomiale, i.e. les lois à queues lourdes ou lois de type Pareto. Les lois de ce domaine ont un point terminal  $x_F$  infini. toute fonction appartenant au domaine d'attraction de Fréchet est une fonction à variations régulières et inversement.

**Théorème 1.5.1** *Une fonction de répartition  $F$  appartient au domaine d'attraction de Fréchet (avec un indice des valeurs extrêmes  $\gamma > 0$ ) si et seulement si  $x_F = \infty$  et si la fonction de survie  $\bar{F}$  est à variations régulières d'indice  $-1/\gamma$  qui s'écrit  $\bar{F} \in RV_{-1/\gamma}$*

i.e

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

Autrement dit, une fonction de répartition  $F$  appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :

$$F(x) = 1 - x^{-1/\gamma}l(y) \quad \text{avec } l \in RV_0. \quad (1.1)$$

Les suites de normalisation  $(a_n)$  et  $(b_n)$  sont données dans ce cas pour tout  $n > 0$  par :

$$a_n = \bar{F}^{\leftarrow}(1/n) \quad \text{et} \quad b_n = 0.$$

L'expression (1.1) est équivalente à :

$$Q(\alpha) = \alpha^{-\gamma}l(\alpha^{-1}).$$

ou  $l \in RV_0$ ,  $\alpha \in [0, 1]$  et  $Q(\alpha)$  est la fonction quantile définie par :

$$Q(\alpha) = \inf\{x, F(x) \geq \alpha\}.$$

Le domaine d'attraction de Fréchet regroupe une grande diversité de lois comptant parmi elles des lois usuelles (loi de Student, loi du Chi-deux, loi Log-gamma, loi de Fréchet).

### 1.5.2 Domaine d'attraction de Weibull

Toutes les lois appartenant au domaine d'attraction de Weibull ont un point terminal  $x_F$  fini, l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.

**Théorème 1.5.2** Une fonction de répartition  $F$  appartient au domaine d'attraction de Weibull (avec un indice des valeurs extrêmes  $\gamma < 0$ ) si et seulement si  $x_F < \infty$  et si la fonction de répartition  $F_*$  définie par :

$$F_*(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ F(x_F - 1/x) & \text{si } x > 0 \end{cases}$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes  $-\gamma > 0$  i.e.

$$\forall x > 0, \quad \lim_{t \rightarrow \infty} \frac{\overline{F}(x_F - 1/tx)}{\overline{F}(x_F - 1/t)} = \lim_{t \rightarrow \infty} \frac{\overline{F}_*(tx)}{\overline{F}_*(t)} = x^{1/\gamma}.$$

Ainsi, une fonction de répartition  $F$  du domaine d'attraction de Weibull s'écrit pour  $x \leq x_F$  sous la forme :

$$F(x) = 1 - (x_F - x)^{-1/\gamma l} ((x_F - x)^{-1}) \quad \text{avec } l \in RV_0.$$

Comme on peut facilement passer du domaine d'attraction de Fréchet à celui de Weibull on a les suites de normalisation  $(a_n)$  et  $(b_n)$  données pour tout  $n > 0$  par :

$$a_n = x_F - \overline{F}^{\leftarrow}(1/n) \quad \text{et} \quad b_n = x_F.$$

De même, il est facile d'obtenir une caractérisation de la fonction quantile associée

$$Q(\alpha) = x_F - \alpha^{-\gamma l} (1/\alpha), \quad \text{avec } l \in RV_0.$$

où  $\alpha \in ]0, 1[$ .

### 1.5.3 Domaine d'attraction de Gumbel

Le domaine d'attraction de Gumbel contient les lois dont la fonction de survie est à décroissance exponentielle, i.e. les lois à queues légères. A contrario des deux autres domaines, il n'y a pas de représentation simple pour les lois appartenant au domaine d'attraction de Gumbel .

**Théorème 1.5.3** Une fonction de répartition  $F$  appartient au domaine d'attraction de Gumbel si et seulement si il existe  $t < x_F \leq \infty$  tel que :

$$\overline{F}(x) = c(x) \exp \left\{ - \int_t^x \frac{1}{\delta(u)} du \right\} \quad \text{où } t < x < x_F.$$

$c(x) \rightarrow c > 0$  lorsque  $y \rightarrow x_F$  et  $\delta$  est une fonction positive et dérivable de dérivée  $\delta'$  telle que  $\delta'(x) \rightarrow 0$  lorsque  $y \rightarrow x_F$ .

Dans ce cas, un choix possible pour les suites  $(a_n)$  et  $(b_n)$  pour tout  $n > 0$  est :

$$a_n = Q(1/n) \quad \text{et} \quad b_n = \frac{1}{F(a_n)} \int_{a_n}^{x_F} \overline{F}(z) dz.$$

Il est difficile de caractériser la fonction quantile de lois appartenant au domaine d'attraction de Gumbel. Du fait que ce domaine d'attraction regroupe une grande diversité de lois, comptant parmi elles la plupart des lois usuelles.

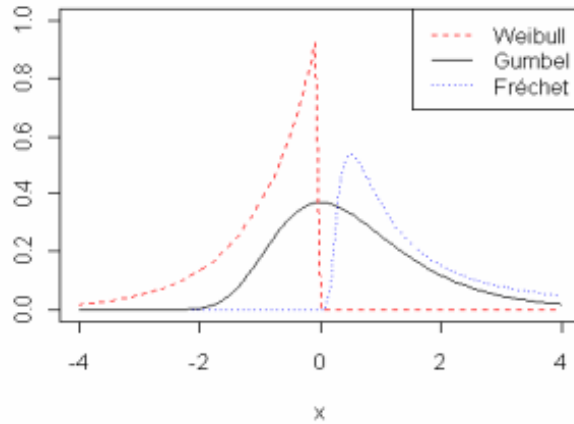


FIG. 1.1 – Densités standard des valeurs extrêmes.

## 1.6 Distribution de Pareto généralisée

### 1.6.1 Distribution conditionnelle des excès

On va introduire dans la suite, la notion d'un excès au dessus d'un seuil  $u$ . Ce dernier est suffisamment élevé pour que ces excès soient considérés comme des extrêmes.

**Définition 1.6.1** On appelle excès de la variable aléatoire  $X$  au dessus d'un seuil  $u \leq x_F$  où  $x_F$  est le point terminal de la distribution i.e  $x_F = \sup\{x : F(x) < 1\}$ , la variable aléatoire  $Y$  qui prend ses valeurs dans  $[0, x_F - u]$  définie par :

$$Y = \{X - u/X > u\}, \quad u \leq x_F.$$

**Définition 1.6.2** On appelle distribution des excès de la variable aléatoire  $X$  au dessus d'un seuil  $u \leq x_F$  la distribution de probabilité de la variable aléatoire  $Y = \{X - u/X > u\}$  donnée par sa fonction de répartition  $F_u$  suivante :

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

Alors

$$\forall y \in \mathbb{R} \quad F_u(y) = P(X - u \leq y / X > u) = \begin{cases} 0 & \text{si } y \leq 0 \\ \frac{F(y+u)-F(u)}{1-F(u)} & \text{si } 0 \leq y \leq x_F - u \\ 1 & \text{si } y \geq x_F - u \end{cases}$$

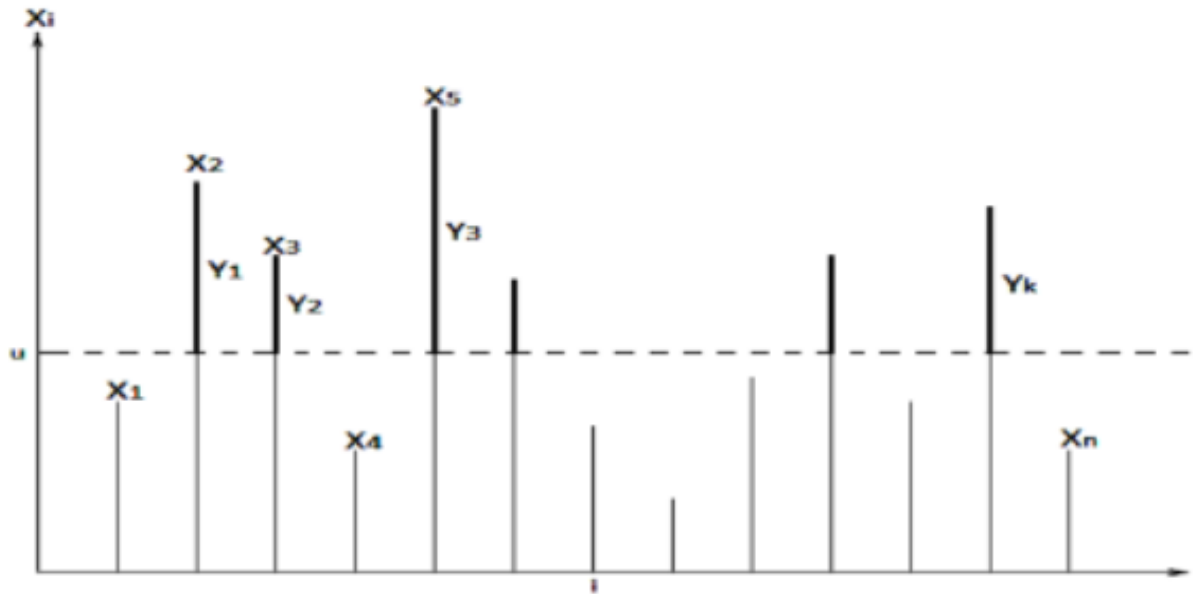


FIG. 1.2 – Les excès

La fonction de survie de  $Y$  au dessus d'un seuil  $u$  est définie par :

$$P(Y > y) = \bar{F}_u(y) = 1 - F_u(y) = \frac{1 - F(y + u)}{1 - F(u)}.$$

## 1.6.2 Distribution de Paréto généralisée

La distribution de Paréto généralisée joue un rôle essentiel dans la modélisation des excès.

**Définition 1.6.3** Soit  $X$  une variable aléatoire. On dit que  $X$  suit une DGP de paramètres  $\xi \in \mathbb{R}$  et  $\sigma > 0$ , on note  $X \sim DGP(\xi, \sigma)$  si sa fonction de répartition notée  $G$  est définie, pour  $x > 0$  et  $1 + \xi \frac{x}{\sigma} > 0$ , par :

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

1) L'ensemble de définition d'une GPD est :

$$\mathbb{R}_+ \quad \text{si } \xi \geq 0. \quad ; \quad [0, -\frac{\sigma}{\xi}] \quad \text{si } \xi < 0.$$

2) La distribution GPD dépend de deux paramètres :

- \*  $\sigma > 0$  est une paramètre d'échelle.
- \*  $\xi \in \mathbb{R}$  est un paramètre de forme.

3) La DGP a les propriétés suivantes :

$$E(X) = \frac{\sigma}{1-\xi}, \quad \xi < 1 \quad \text{et} \quad Var(X) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}, \quad \xi < \frac{1}{2}$$

Lorsque  $\xi \geq 1$ , la moyenne n'est pas définie, et lorsque  $\xi \geq \frac{1}{2}$ , la variance n'est pas définie.

## 1.6.3 Théorème de Balkema-de Haan-Pickands

La distribution limite d'un excès est donnée par le théorème de Balkema-de Haan-Pickands. Soit  $X$  une variable aléatoire de fonction de distribution inconnue  $F$ . Soit  $x_F$  la limite supérieure finie ou infinie du support de la distribution  $F$  et soit  $Y = X - u$  un excès au dessus d'un seuil élevé  $u$ . La fonction de distribution de l'excès  $Y$  est définie telle que :

$$F_u(x) = P\{X - u \leq x / X > u\} = P\{Y \leq x / X > u\} = \frac{F(x+u) - F(u)}{1 - F(u)},$$

où  $0 \leq x < x_F - u$ .



**Théorème 1.6.1** *Le théorème énonce que si  $F$  appartient à l'un des trois domaines d'attraction de la loi limite des extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction de répartition des excès au-delà de  $u$ , noté  $F_u$  qui peut être approchée par une loi de Pareto généralisée (GPD) telle que :*

$$\lim_{u \rightarrow x_F} \sup_{0 \leq x < x_F - u} |F_u(x) - G_{\xi, \sigma(u)}(x)| = 0,$$

où  $\sigma$  est une fonction strictement positive.

**Preuve** En effet, d'après le théorème de Fisher-Tippett, il est naturel de supposer que pour  $u \rightarrow x_F$  on a :

$$F^n(x) \approx \exp \left( - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right) \quad (1.2)$$

De plus, comme on s'intéresse aux valeurs de la queue de la distribution, alors  $x > u$  pour  $u$  suffisamment grand implique  $F(x) \approx 1$  et donc

$$\log F(x) \approx -(1 - F(x)).$$

En passant au logarithme dans l'équation (1.2), on aura

$$n \log F(x) \simeq - \left( 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right)^{-\frac{1}{\xi}}.$$

ou bien

$$\log F(x) \simeq -\frac{1}{n} \left( 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right)^{-\frac{1}{\xi}}.$$

D'où

$$1 - F(x) = \frac{1}{n} \left( 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right)^{-\frac{1}{\xi}}. \quad (1.3)$$

Par ailleurs, la fonction de survie des excès  $Y = (X - u | X > u)$  est telle que

$$P(X - u > y | X > u) = \frac{1 - F(y + u)}{1 - F(u)}. \quad (1.4)$$

Par conséquent en utilisant la relation (1.3) dans la relation(1.4) on aura

$$\begin{aligned}
P(X - u > y/X > u) &\approx \frac{\frac{1}{n} \left(1 + \xi \left(\frac{y+u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}}{\frac{1}{n} \left(1 + \xi \left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}} \\
&= \left(\frac{1 + \xi \left(\frac{y+u-\mu}{\sigma}\right)}{1 + \xi \left(\frac{u-\mu}{\sigma}\right)}\right)^{\frac{-1}{\xi}} \\
&= \left(\frac{\sigma + \xi (y + u - \mu)}{\sigma + \xi (u - \mu)}\right)^{\frac{-1}{\xi}} \\
&= \left(1 + \xi \frac{y}{\tilde{\sigma}}\right)^{\frac{-1}{\xi}}
\end{aligned}$$

où  $\tilde{\sigma} = \sigma + \xi (u - \mu)$

Finalement, la fonction de survie des dépassements au dessus d'un seuil  $u$  est telle que :

$$P(X > u + y/X > u) = \frac{1 - F(y + u)}{1 - F(u)} \approx \left[1 + \xi \frac{y}{\tilde{\sigma}}\right]^{\frac{-1}{\xi}},$$

où  $\tilde{\sigma} = \sigma + \xi (u - \mu)$

La distribution GPD regroupe trois distributions selon les valeurs du paramètre de forme.

Lorsque  $\xi = 0$ , c'est la distribution exponentielle :

$$G_{0,\tilde{\sigma}}(x) = 1 - \exp\left[-\frac{x}{\tilde{\sigma}}\right].$$

Lorsque  $\xi > 0$ , c'est la loi de Pareto usuelle :

$$G_{\xi,\tilde{\sigma}}(x) = 1 - \left[\frac{\tilde{\sigma}}{\xi x}\right]^{-\frac{1}{\xi}}.$$

Lorsque  $\xi < 0$ , c'est la loi est de pareto de type II :

$$G_{\xi,\tilde{\sigma}}(x) = \left[1 + \frac{x}{\tilde{\sigma}}\right]^{-\frac{1}{\xi}}.$$

■

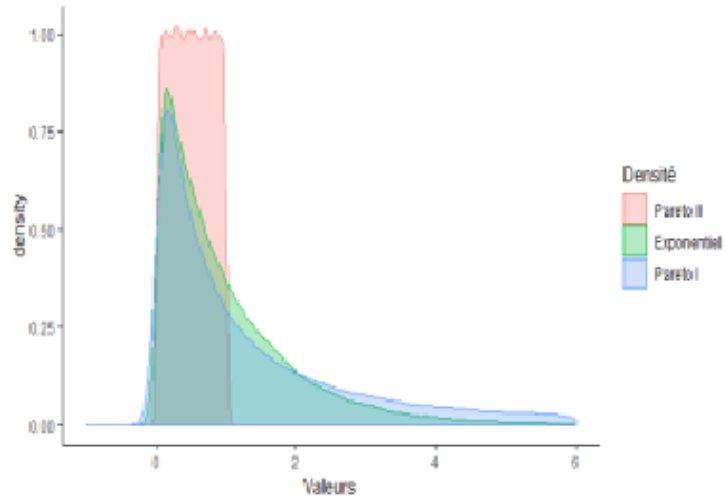


FIG. 1.3 – Densités de probabilité de GPD avec différentes valeurs de  $\xi$ .

Les fonctions de densité sont bornées à gauche et ils sont décalés à droite de la médiane. Aussi que seulement la densité de Pareto usuelle est bornée à droite.

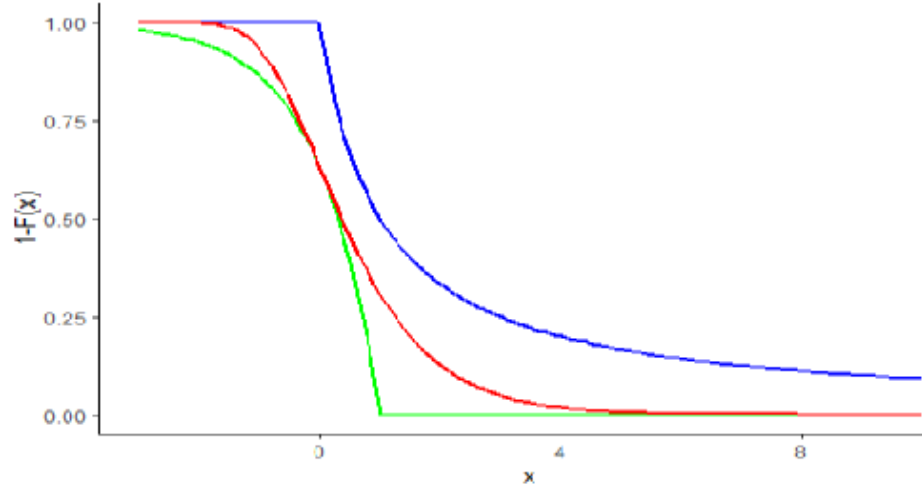


FIG. 1.4 – Fonction de survie de GPD avec différentes valeurs de  $\xi$ .

Les fonctions de survie que la distribution de Pareto I décroît moins vite que les distributions exponentielle et Pareto II. Ainsi, on dit que la distribution de Pareto I est à queue lourde.

# Chapitre 2

## Estimation dans les modèles GPD

Dans ce chapitre nous intéressons aux différentes méthodes d'estimation des paramètres  $\xi$ ,  $\sigma$  d'un modèle statistique gouverné par une distribution  $GPD_{\xi,\sigma}$ . On trouve dans la littérature plusieurs méthodes d'estimation. Parmi ces méthodes, on peut citer la méthode du maximum de vraisemblance (*Reiss and Thomas (2007)*), la méthode des moments pondérés (*Hosking and Wallis (1987)*), Méthode des L-moments (*Hosking (1990)*), l'estimateur de Hill (*Hill (1975)*), l'estimateur de Pickands (*Pickands III et al. (1975)*), et d'autres techniques d'estimation telles que, l'estimateur des moments Dekkers et al. (*Dekkers et al. (1989)*). Les méthodes les plus utilisées sont celles du maximum de vraisemblance, moments, moments pondérés, L-moments, Hill, et des Pickands [17],[12],[11],[16],[15],[3],[4],[19],[13],[5],[2].

### 2.1 Estimation des paramètres de la GPD

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

#### 2.1.1 Sélection du seuil

Une bonne détermination de  $u$  est primordiale si nous voulons appliquer la théorie des valeurs extrêmes. En effet, un seuil trop faible conduit à une mauvaise adéquation entre la fonction de répartition et la loi GPD et les estimateurs sont biaisés. Inversement un seuil trop élevé entraîne une faible taille d'échantillon de dépassements et donc des estimateurs volatiles. Nous allons voir dans la suite la méthode Mean-Excess plot pour le choix de seuil  $u$ .

## Mean-Excess plot

Etant donnée un échantillon aléatoire  $X_1, X_2, \dots, X_n$  d'une distribution  $F$  et soit  $u$  un seuil donné. Le Mean Excess Plot est le graphe des points  $(u, e(u))$  où  $e(u)$  est la moyenne des excès au dessus du seuil  $u$ , appelée aussi la Mean Excess Function définie par :

$$e(u) = E(X - u \mid X > u).$$

En pratique, la fonction des excès moyens  $e(u)$  est estimée par  $\hat{e}_n(u)$  :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^n (x_i - u) \mathbf{1}_{(x_i > u)}(x_i)}{\sum_{i=1}^n \mathbf{1}_{(x_i > u)}(x_i)} = \frac{1}{N_u} \sum_{i=1}^n (x_i - u) \mathbf{1}_{(x_i > u)}.$$

Un des principaux outils du choix du seuil est le graphe de la sample mean excess function  $\hat{e}_n(u)$  ou le mean excess plot.

**Définition 2.1.1** *Le mean excess plot ou ME-plot est défini par le nuage de points suivant*

$$\{(u, \hat{e}_n(u)), X_{1,n} < u < X_{n,n}\},$$

où  $X_{n,n}$  et  $X_{1,n}$  sont respectivement le maximum et le minimum de l'échantillon.

Le seuil  $u$  à retenir, est celui pour lequel la moyenne des excès est approximativement linéaire. Plus les queues des distributions sont épaisses, plus la fonction  $e(u)$  tend rapidement vers l'infini.

### 2.1.2 Estimation par le Maximum de Vraisemblance

Considérons  $y_1, \dots, y_{N_u}$  comme étant uniquement les valeurs de  $x_1, \dots, x_n$  qui dépassent un seuil  $u$ . Supposons que l'échantillon des excès  $(y_1, \dots, y_{N_u})$  soit *i.i.d* de densité donnée par :

$$g_{\xi, \sigma}(y) = \begin{cases} \frac{1}{\sigma} (1 + \xi \frac{y}{\sigma})^{-\frac{1}{\xi} - 1} & \text{si } \xi \neq 0 \\ \frac{1}{\sigma} \exp(-\frac{y}{\sigma}) & \text{si } \xi = 0 \end{cases}.$$

L'expression de la fonction de vraisemblance est donnée par :

$$L(\xi, \sigma, Y) = \prod_{i=1}^{N_u} g_{\xi, \sigma}(y_i).$$

Pour maximiser la log-vraisemblance

$$l(\xi, \sigma, Y) = \log L(\xi, \sigma, Y).$$

– Maximiser la log-vraisemblance revient à résoudre les deux systèmes suivants :

$$\begin{cases} \frac{\partial \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma)}{\partial \sigma} = 0; & \frac{\partial^2 \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma)}{\partial \sigma^2} < 0. \\ \\ \frac{\partial \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma)}{\partial \xi} = 0; & \frac{\partial^2 \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma)}{\partial \xi^2} < 0. \end{cases} \quad (2.1)$$

On distingue deux cas selon les valeurs de  $\xi$  :

1. Si  $\xi \neq 0$  à condition que  $1 + \xi \frac{y_i}{\sigma} > 0$  :

$$\begin{aligned} \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma) &= \sum_{i=1}^{N_u} \log g_{\xi, \sigma}(y_i) \\ &= \sum_{i=1}^{N_u} \log \left( \frac{1}{\sigma} \left(1 + \xi \frac{y_i}{\sigma}\right)^{-\frac{1}{\xi} - 1} \right) \\ &= \sum_{i=1}^{N_u} \left( -\log \sigma - \left(\frac{1}{\xi} + 1\right) \log \left(1 + \xi \frac{y_i}{\sigma}\right) \right) \\ &= -N_u \log \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \log \left(1 + \xi \frac{y_i}{\sigma}\right) \end{aligned}$$

Résoudre les deux systèmes (2.1) revient à résoudre les deux équation

$$\text{suivants : } \begin{cases} -\frac{N_u}{\sigma} - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \left( \frac{-\xi y_i}{\sigma(\sigma + \xi y_i)} \right) = 0, \\ \\ \frac{1}{\xi^2} \sum_{i=1}^{N_u} \log \left(1 + \xi \frac{y_i}{\sigma}\right) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \left( \frac{y_i}{\sigma + \xi y_i} \right) = 0. \end{cases}$$

On a deux systèmes des équations non linéaires n'admettent pas des solutions explicites. On fait appel alors à des méthodes numériques de type algorithmes de Newton-Raphson.

2. Dans le cas ou  $\xi = 0$

$$\begin{aligned} \log L(y_1, y_2, \dots, y_{N_u}; \xi, \sigma) &= \sum_{i=1}^{N_u} \log g_{\xi, \sigma}(y_i) \\ &= \sum_{i=1}^{N_u} \left( -\log \sigma - \frac{y_i}{\sigma} \right) \\ &= -N_u \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{N_u} y_i. \end{aligned}$$

Résoudre les deux systèmes (2.1) revient à résoudre l'équation suivant :

$$-\frac{N_u}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{N_u} y_i = 0,$$

donc

$$\hat{\sigma} = \frac{1}{N_u} \sum_{i=1}^{N_u} y_i = \bar{Y}.$$

### 2.1.3 Méthode des moments (MOM)

*Hosking et Wallis (1987)* ont été les premiers à proposer la méthode des moments pour estimer les deux paramètres de la GPD. Pour  $\xi < \frac{1}{2}$ , on sait que :

$$E(X) = \frac{\sigma}{1-\xi} \quad \text{et} \quad \text{Var}(X) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}.$$

On peut alors facilement exprimer les paramètres de la loi GPD  $\xi$  et  $\sigma$  en fonction de l'espérance et de la variance de  $X$ , soit :

$$\xi = \frac{1}{2} \left( 1 - \frac{E(X)^2}{\text{Var}(X)} \right) \quad \text{et} \quad \sigma = \frac{E(X)}{2} \left( 1 + \frac{E(X)^2}{\text{Var}(X)} \right).$$

Ainsi en remplaçant  $E(X)$  et  $\text{Var}(X)$  par leurs moments empiriques :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad S^2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

On obtient les estimateurs des moments de  $\xi$  et  $\sigma$  (respectivement,  $\xi^{MOM}$  et  $\sigma^{MOM}$ ), soit :

$$\hat{\xi}^{MOM} = \frac{1}{2} \left( 1 - \frac{\bar{X}^2}{S^2(X)} \right) \quad \text{et} \quad \hat{\sigma}^{MOM} = \frac{\bar{X}}{2} \left( 1 + \frac{\bar{X}^2}{S^2(X)} \right).$$

### 2.1.4 Méthode des Moments Pondérés (EMP)

Il peut arriver que certains moments n'existent pas, ne soient pas finis. Au lieu de la Méthode des Moments, nous utilisons alors la Méthode des Moments Pondérés. Définissons, avec  $r$  l'ordre du moment :

$$\omega_r(\xi, \sigma) = E[X(\bar{G}_{\xi, \sigma}(X))^r] = \int_{-\infty}^{+\infty} x \bar{G}_{\xi, \sigma}^r d\bar{G}_{\xi, \sigma}(x) \quad \text{avec } r \in \mathbb{N}.$$



où  $\overline{G}_{\xi, \sigma} = 1 - G_{\xi, \sigma}$  avec  $G$  la distribution GPD.

Alors

$$\begin{aligned}\omega_r(\xi, \sigma) &= E \left[ X \left( 1 + \frac{\xi}{\sigma} X \right)^{\frac{-r}{\xi}} \right] \\ &= \int_0^{\frac{\sigma}{\xi}} x \left( 1 + \frac{\xi}{\sigma} x \right)^{\frac{-r}{\xi}} \left( \frac{1}{\xi} \right) \left( \frac{\xi}{\sigma} \right) \left( 1 + \frac{\xi}{\sigma} x \right)^{\frac{-1}{\xi} - 1} dx \\ &= \frac{1}{\sigma} \int_0^{\frac{\sigma}{\xi}} x \left( 1 + \frac{\xi}{\sigma} x \right)^{-\frac{r+1}{\xi} - 1} dx\end{aligned}$$

on pose  $y = 1 + \frac{\xi}{\sigma} x$  alors  $x = \frac{\sigma}{\xi}(y - 1)$

$$\begin{aligned}\omega_r(\xi, \sigma) &= \frac{1}{\sigma} \int_1^0 \frac{\sigma}{\xi} (y - 1) y^{-\frac{r+1}{\xi} - 1} \left( \frac{\sigma}{\xi} \right) dy \\ &= \frac{\sigma}{\xi^2} \int_1^0 (y - 1) y^{-\frac{r+1}{\xi} - 1} dy \\ &= \frac{\sigma}{\xi^2} \left[ \int_1^0 y^{-\frac{r+1}{\xi}} dy - \int_1^0 y^{-\frac{r+1}{\xi} - 1} dy \right]\end{aligned}$$

pour  $r > \xi - 1$  on a

$$\begin{aligned}\omega_r(\xi, \sigma) &= \frac{\sigma}{\xi} \left[ \frac{1}{r+1-\xi} - \frac{1}{r+1} \right] \\ &= \frac{\sigma}{(r+1-\xi)(r+1)}\end{aligned}$$

ce qui nous permet d'obtenir avec  $r = 0$  et  $r = 1$

$$\omega_0(\xi, \sigma) = \frac{\sigma}{1-\xi} \quad \text{et} \quad \omega_1(\xi, \sigma) = \frac{\sigma}{2(2-\xi)}.$$

alors

$$\begin{aligned}\sigma &= \frac{2\omega_0\omega_1}{\omega_0 - 2\omega_1}. \\ \xi &= 2 - \frac{\omega_0}{\omega_0 - 2\omega_1}.\end{aligned}$$

Pour calculer  $\widehat{\sigma}$  et  $\widehat{\xi}$ , nous remplaçons le moment pondéré  $\omega_r$  par son estimateur empirique  $\widehat{\omega}_r$  défini tel que :

$$\widehat{\omega}_r(\xi, \sigma) = \frac{1}{n} \sum_{i=1}^n X_i \widehat{F}^r(X_i), \quad r = 0, 1.$$

où  $\widehat{F}$  est la fonction de répartition empirique de l'échantillon  $X_1, X_2, \dots, X_{N_u}$ , alors pour  $r = 0, 1$ ,

$$\widehat{\sigma} = \frac{2\widehat{\omega}_0\widehat{\omega}_1}{\widehat{\omega}_0 - 2\widehat{\omega}_1},$$

$$\widehat{\xi} = 2 - \frac{\widehat{\omega}_0}{\widehat{\omega}_0 - 2\widehat{\omega}_1}.$$

### 2.1.5 Méthode des L-moments

Les L-moments sont des combinaisons linéaires des statistiques d'ordres. Ont été introduites et extrait de *Silitto (1951)* et revu de manière exhaustive par *Hosking (1990)*.

#### Principe de la méthode des L-moments

**Définition 2.1.2** Soient  $X_1, X_2, \dots, X_n$  un échantillon de taille  $n$  d'une distribution continue  $F_X(x)$  avec la fonction de quantile  $Q(u) = F_X^{-1}(u)$ , et soit  $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$  les statistiques d'ordres associées à cet échantillon, pour  $r \geq 1$  le  $r$ -ième L-moments  $\lambda_r$  est donné par :

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(X_{r-k,r}), \quad \forall r \geq 1.$$

telle que  $E(X_{r-k,r})$  présente l'espérance de la statistique d'ordre.

En particulier, pour  $r = 1, 2, 3, 4$ , nous obtenons les premiers L-moments qui sont donnés par :

- Le premier L-moment  $\lambda_1$  est utilisé pour calculer la moyenne (mesure de position) est défini par :

$$\lambda_1 = E(X_{1,1}).$$

- Le deuxième L-moment  $\lambda_2$  est utilisé pour calculer (mesure de dispersion ou d'échelle) est donné par :

$$\lambda_2 = \frac{1}{2}E(X_{2,2} - X_{1,2}).$$

— Le troisième L-moment  $\lambda_3$  pour étudier la symétrie (mesure de skewnes) est donné par :

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

— Le quatrième L-moment  $\lambda_4$  pour étudier l'aplatissement (mesure de kurtosis) est défini par :

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

**Remarque 2.1.1** *Les quatre premiers L-moments précédents sont analogues aux moments classiques.*

### Représentation de L-moment en terme de polynômes orthogonaux

Les L-moments peuvent être écrite en fonction de polynômes de Legendre déplacé  $P_r^*$  qui sont définies par :

$$P_r^*(u) = \sum_{k=0}^r P_{r,k}^* u^k = \sum_{k=0}^r (-1)^{r+k} \binom{r}{k} \binom{r+k}{k} u^k.$$

tel que  $r$  un entier positif, pour  $r = 0, 1, 2$  et  $4$ , nous avons :

$$P_0^*(u) = 1.$$

$$P_1^*(u) = (2u - 1).$$

$$P_2^*(u) = (6u^2 - 6u + 1).$$

$$P_3^*(u) = (20u^3 - 30u^2 + 12u - 1).$$

Alors les  $\lambda_r$  réécrire comme suit :

$$\lambda_r = \int_0^1 Q(u) P_{r-1}^*(u) du, \quad r = 1, 2, \dots$$

et les premiers L-moments donc données comme ci :

$$\begin{aligned}\lambda_1 &= \int_0^1 Q(u)du. \\ \lambda_2 &= \int_0^1 (2u - 1)Q(u)du. \\ \lambda_3 &= \int_0^1 (6u^2 - 6u + 1)Q(u)du. \\ \lambda_4 &= \int_0^1 (20u^3 - 30u^2 + 12u - 1)Q(u)du.\end{aligned}$$

### Application de la méthode des L-moments à la GPD

Soit  $X$  une variable aléatoire qui suit une loi GPD dont les paramètres  $\xi$  et  $\sigma$  sont inconnus. Les estimateurs de ces paramètres sont la solution d'un système d'équations obtenues en égalant les L-moments empirique et les L-moments de la distribution. Les 4 premiers moments d'une GPD :

$$\begin{aligned}\lambda_1 &= \int_0^1 Q(u)du = \int_0^1 \frac{\sigma}{\xi}((1-u)^{-\xi} - 1)du = \frac{\sigma}{1-\xi}. \\ \lambda_2 &= \int_0^1 (2u-1)\frac{\sigma}{\xi}((1-u)^{-\xi} - 1)du = \frac{\sigma}{(-\xi+1)(-\xi+2)}. \\ \lambda_3 &= \int_0^1 (6u^2 - 6u + 1)\frac{\sigma}{\xi}((1-u)^{-\xi} - 1)du = \frac{\sigma(\xi+1)}{(-\xi+1)(-\xi+2)(-\xi+3)}. \\ \lambda_4 &= \int_0^1 (20u^3 - 30u^2 + 12u - 1)\frac{\sigma}{\xi}((1-u)^{-\xi} - 1)du = \frac{\sigma(\xi+1)(\xi+2)}{(-\xi+1)(-\xi+2)(-\xi+3)(-\xi+4)}.\end{aligned}$$

Et on déduit les rapports de L-moments

L Skewness est donnée par :

$$\begin{aligned}\tau_3 = \lambda_3/\lambda_2 &= \frac{(\sigma(\xi+1))/(-\xi+1)(-\xi+2)(-\xi+3)}{(\sigma/(-\xi+1)(-\xi+2))} \\ &= \frac{1+\xi}{3-\xi}.\end{aligned}$$

L Kurtosis est donnée par :

$$\begin{aligned}\tau_4 = \lambda_4/\lambda_2 &= \frac{(\sigma(\xi+1)(\xi+2))/(-\xi+1)(-\xi+2)(-\xi+3)(-\xi+4)}{(\sigma/(-\xi+1)(-\xi+2))} \\ &= \frac{(\xi+1)(\xi+2)}{(-\xi+3)(-\xi+4)}.\end{aligned}$$

Les estimateurs des paramètres de la GPD sont donnés par :

$$\widehat{\xi}^{LM} = \frac{3t_3 - 1}{1 + t_3} \quad \text{et} \quad \widehat{\sigma}^{LM} = l_2(-\widehat{\xi}^{LM} + 1)(-\widehat{\xi}^{LM} + 2),$$

où  $l_2$  est l'estimateur de  $\lambda_2$  et  $t_3$  est l'estimateur de  $\tau_3$  ( $t_3 = \frac{1+\xi}{3-\xi}$ ).

## 2.2 Estimation par des méthodes semi-paramétrique

Nous donnons quelques estimateurs de l'IVE $_{\xi}$  construits sous les conditions de domaine d'attraction. Par opposition aux méthodes paramétriques de la section précédente, des procédures statistiques semi-paramétrique appropriées à cette situation, n'assument pas la connaissance de la distribution entière mais seulement sur les queues de distribution.

### 2.2.1 Estimateur de Hill

Dans le domaine des distributions à queue épaisse, la statistique de la théorie des valeurs extrêmes se traduit par un problème d'estimation semi-paramétrique. En effet, si  $F$  appartient au domaine de Fréchet, alors sa fonction de survie est de la forme  $x^{-\frac{1}{\xi}}l(x)$ . Le paramètre d'intérêt est  $\xi > 0$  et  $l$  fonction à variation lente.

L'estimateur de Hill n'est utilisable que pour les distributions appartenant au domaine de Fréchet. Une fonction de répartition  $F$  appartient au domaine d'attraction de Fréchet si et seulement si sa fonction de survie  $F \in RV_{\xi}$ , la fonction de survie de la distribution de ce domaine est la forme :

$$\overline{F}(x) = 1 - F(x) = x^{-\frac{1}{\xi}}l(x).$$

avec  $l$  une fonction à variation lente. La fonction de survie  $\overline{F}$  comporte :

- Une partie paramétrique  $x^{-\frac{1}{\xi}}$  ne dépendant que d'un paramètre réel.
- Une partie non-paramétrique  $l$ .

Pour un réel  $t > 1$  il est clair que

$$\lim_{u \rightarrow \infty} \frac{\overline{F}(tu)}{\overline{F}(u)} = t^{-\frac{1}{\xi}}; \quad \lim_{u \rightarrow \infty} \frac{l(tu)}{l(u)} = 1,$$

on en déduit l'approximation

$$\overline{F}(tu) \simeq \overline{F}(u)t^{-\frac{1}{\xi}}. \tag{2.2}$$

Dans l'approximation (2.2) nous remplaçons  $tu$  par  $x$  on trouve

$$\bar{F}(x) = \bar{F}(u) \left(\frac{x}{u}\right)^{-\frac{1}{\xi}}.$$

La densité  $f(x)$  est comme suit :

$$f(x) = \frac{1}{\xi} \frac{\bar{F}(u)}{u} \left(\frac{x}{u}\right)^{-\frac{1}{\xi}-1}.$$

L'expression de la fonction de vraisemblance est :

$$L(x_1, x_2, \dots, x_k; \xi) = \prod_{i=1}^k f(x_i)$$

Maximiser la vraisemblance  $L$  revient à résoudre le système suivant :

$$\left\{ \begin{array}{l} \frac{\partial \log L(x_1, x_2, \dots, x_k; \xi)}{\partial \xi} = 0; \\ \frac{\partial^2 \log L(x_1, x_2, \dots, x_k; \xi)}{\partial \xi^2} < 0. \end{array} \right. \quad (2.3)$$

en effet

$$\begin{aligned} \log L(x_1, x_2, \dots, x_k; \xi) &= \log \left( \prod_{i=1}^k f(x_i) \right) \\ &= \sum_{i=1}^k \log f(x_i) \\ &= \sum_{i=1}^k \log \left( \frac{1}{\xi} \frac{\bar{F}(u)}{u} \left(\frac{x_i}{u}\right)^{-\frac{1}{\xi}-1} \right) \\ &= \sum_{i=1}^k \left[ \log \bar{F}(u) - \log(\xi u) - \left(\frac{1}{\xi} + 1\right) \log \left(\frac{x_i}{u}\right) \right] \end{aligned}$$

Résoudre le système (2.3) revient à résoudre l'équation suivant :

$$\begin{aligned} \sum_{i=1}^k \left[ \frac{-u}{\xi u} + \frac{1}{\xi^2} \log \left(\frac{x_i}{u}\right) \right] &= 0 \\ -\xi k + \sum_{i=1}^k \log \frac{x_i}{u} &= 0, \end{aligned}$$

la solution de cette équation est de la forme

$$\hat{\xi}^H = \frac{1}{k} \sum_{i=1}^k \log \frac{x_i}{u} = \frac{1}{k} \sum_{i=1}^k [\log x_i - \log u].$$

On remplace  $x_i$  et  $u$  par leurs équivalents dans la statistique d'ordre on trouve la formule de l'estimateur de Hill

$$\widehat{\xi}^H = \frac{1}{k} \sum_{i=1}^k [\log X_{n-i+1,n} - \log X_{n-k+1,n}].$$

**Théorème 2.2.1** Soit  $k_n, n \geq 1$  une suite d'entiers telle que  $1 \leq k_n < n, k_n \rightarrow \infty$  et  $\frac{k_n}{n} \rightarrow 0$  quand  $n \rightarrow \infty$ .

- (a) Consistance faible : Alors,  $\widehat{\xi}_{k_n}^H$  converge en probabilité vers  $\xi$ .
- (b) Consistance forte : Si de plus  $\frac{k_n}{\log \log n} \rightarrow \infty$  quand  $n \rightarrow \infty$ , alors,  $\widehat{\xi}_{k_n}^H$  converge presque sûrement vers  $\xi$ .
- (c) Normalité asymptotique :

$$\sqrt{k_n} \frac{\widehat{\xi}_{k_n}^H}{\xi - 1} \rightarrow^L N(0, 1).$$

On peut associer à l'estimateur de Hill un intervalle de confiance asymptotique de niveau  $(1 - \alpha)$  donné par :

$$\left( \widehat{\xi}_{k_n}^H \left( 1 + z_{1-\alpha/2} \frac{1}{\sqrt{k_n}} \right)^{-1}, \widehat{\xi}_{k_n}^H \left( 1 - z_{1-\alpha/2} \frac{1}{\sqrt{k_n}} \right)^{-1} \right),$$

où  $z_\alpha$  est le quantile d'une distribution normale standard au niveau  $\alpha$

## 2.2.2 Estimateur des Moments :

Un inconvénient de l'estimateur de Hill est qu'il est conçu seulement pour l'IVE des distributions à queues lourdes. En 1989, Dekkers et al ont proposé une extension de tous types de distribution, appelé estimateur des moments.

**Définition 2.2.1** (Estimateur des Moments), Pour  $\xi \in \mathbb{R}$ , l'estimateur des moments est

$$\widehat{\xi}^M = M_n^{(1)} + 1 - \frac{1}{2} \left( 1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}} \right)^{-1}$$

Où  $M_n^{(1)}$  est l'estimateur de Hill  $\widehat{\xi}^H$ .

$$\widehat{\xi}^M = \widehat{\xi}^H + 1 - \frac{1}{2} \left( 1 - \frac{(\widehat{\xi}^H)^2}{M_n^{(2)}} \right)^{-1}$$

Avec :

$$M_n^{(r)} = \frac{1}{k} \sum_{i=1}^k (\log(X_{n-i+1:n}) - \log(X_{n-k:n}))^r \quad , \quad r = 1, 2$$

**Théorème 2.2.2** (*Propriétés asymptotique de  $\widehat{\xi}^M$* ) Supposons que  $F \in D(H_\xi)$ ,  $\xi \in \mathbb{R}$ ,  $k \rightarrow \infty$  et  $k/n \rightarrow 0$  quand  $n \rightarrow \infty$  :

(a) *Consistance faible* :

$$\widehat{\xi}^M \xrightarrow{P} \xi \quad \text{quand } n \rightarrow \infty.$$

(b) *Consistance forte* : Si  $k/(\log n)^\sigma \rightarrow \infty$  quand  $n \rightarrow \infty$  pour certaine  $\sigma > 0$ , alors

$$\widehat{\xi}^M \xrightarrow{P.s} \xi \quad \text{quand } n \rightarrow \infty.$$

(c) *Normalité asymptotique* :

$$\sqrt{k_n} (\widehat{\xi}^M - \xi) \xrightarrow{\mathcal{L}} N(0, \eta^2) \quad \text{quand } n \rightarrow \infty.$$

où

$$\eta^2 = \begin{cases} 1 + \xi^2 & , \xi \geq 0 \\ (1 - \xi^2)(1 - 2\xi) \left( 4 - 8\frac{1-2\xi}{1-3\xi} + \frac{(5-11\xi)(1-2\xi)}{(1-3\xi)(1-4\xi)} \right) & , \xi < 0 \end{cases}$$

### 2.2.3 Estimateur de Pickands

Cet estimateur a été introduit par (*Pickands III et al. (1975)*). Cet estimateur est construit en utilisant trois statistiques d'ordres et a l'avantage d'être valable quelque soit le domaine d'attraction de la distribution. Il est défini de la manière suivante :

$$\widehat{\xi}_{K_n}^P = \frac{1}{2} \log \frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}, k_n = 1, \dots, [n/4].$$

Notons que l'estimateur de Pickands  $\widehat{\xi}_{K_n}^P$ , qui utilise l'information apportée par des distances entre deux statistiques ordonnées n'utilise pas le maximum de l'échantillon  $X_{n,n}$ , ce qui implique une perte d'information sur la queue de distribution. Une justification plus formelle de l'estimateur de Pickands a été fournie dans *Embrechts et al. (2013)*. Les propriétés de l'estimateur de Pickands ont été principalement explorées dans *Dekkers et al. (1989)*. Plus précisément, ils ont prouvé les résultats suivants :



**Théorème 2.2.3** *On suppose que  $F \in D(H_\xi)$ ,  $\xi \in \mathbb{R}$ , la condition du second ordre soit vérifiée. Soit  $k_n, n \geq 1$  une suite d'entiers telle que  $1 \leq k_n < [n/4], k_n \rightarrow \infty$  et  $k_n/n \rightarrow 0$  quand  $n \rightarrow \infty$ .*

(a) *Consistance faible :*

$$\widehat{\xi}_{K_n}^P \xrightarrow{P} \xi \quad \text{quand } n \rightarrow \infty$$

(b) *Consistance forte : Si de plus  $k_n/\log \log n \rightarrow \infty$  quand  $n \rightarrow \infty$ , alors*

$$\widehat{\xi}_{K_n}^P \xrightarrow{P,s} \xi \quad \text{quand } n \rightarrow \infty$$

(c) *Normalité asymptotique : On suppose que la fonction de queue  $U$  associée à  $F$  satisfait la condition du second ordre ci-dessous : ils existent une constante réelle  $\rho \leq 0$  et une fonction  $A(t) \rightarrow 0$  quand  $t \rightarrow \infty$  telles que pour tout  $x > 0$ ,*

$$\lim_{t \rightarrow \infty} \frac{\frac{U(tx) - U(t)}{A(t)} - D_\xi(x)}{A(t)} = \int_1^x s^{\xi-1} \int_1^s u^{\rho-1} du ds$$

*Alors pour  $\sqrt{k_n}A(n/k_n) \rightarrow \xi \in \mathbb{R}$  quand  $n \rightarrow \infty$ ,*

$$\sqrt{k}(\widehat{\xi}_{K_n}^P - \xi) \xrightarrow{\mathcal{L}} N\left(0, \frac{\gamma^2(2^{2\gamma+1} + 1)}{4(\log 2)^2(2\gamma - 1)^2}\right) \quad \text{quand } n \rightarrow \infty$$

## 2.2.4 Choix du nombre optimal de statistiques d'ordre extrêmes

Les résultats concernant les estimateurs de l'indice des valeurs extrêmes sont asymptotiques : ils sont obtenus lorsque  $k_n \rightarrow \infty$  et  $\frac{k_n}{n} \rightarrow 0$ . Choisir un nombre optimal de statistiques d'ordre extrêmes  $k_{opt}$  n'est pas facile à manipuler. Lorsque  $k$  est petit la variance de l'estimateur est grande et l'utilisation de  $k$  introduit un grand biais. L'équilibrage de ces composants (la variance et le biais) est important dans les applications de la théorie des valeurs extrêmes, parce que cela réduit l'erreur quadratique moyenne.

### Erreur quadratique moyenne

L'erreur quadratique moyenne est très utile pour comparer plusieurs estimateurs, notamment lorsque l'un d'eux est biaisé. Si les deux estimateurs à comparer sont sans biais, l'estimateur le plus efficace est simplement celui qui a la variance la plus petite. On peut effectivement exprimer l'erreur quadratique moyenne en fonction du biais de l'estimateur  $E(\widehat{\xi}_n - \xi)$  ainsi que sa variance :

$$MSE = \text{Biais}(\widehat{\xi}_n)^2 + \text{Var}(\widehat{\xi}_n).$$

Le  $MSE$  d'un estimateur de l'indice de queue  $\widehat{\xi}_n$  est défini par :

$$MSE(\widehat{\xi}_n) := E_\infty(\widehat{\xi}_n - \xi)^2,$$

où  $E_\infty$  dénote l'espérance en ce qui concerne la distribution de la limite. On le vérifie facilement que  $MSE(\widehat{\xi}_n)$ , qui est réellement une fonction de  $k$ .

Pour une estimation exacte de l'indice de queue, il est nécessaire pour n'importe quel estimateur classique à faire un compromis entre le biais et la variance. Il semble raisonnable que le  $MSE$  permet de réduire au minimum un compromis entre le biais et la variance rendent l'estimation la plus précise possible. C'est le choix optimal de  $k$ ; noté  $k_{opt}$ , correspond à la plus petite  $MSE$ , i.e.

$$k_{opt} := \arg \min MSE(\widehat{\xi}_n).$$

### Méthode Graphique

C'est la méthode la plus simple pour la détermination de  $k_n$ . Elle consiste à tracer le graphe

$$\left\{ \left( k_n, \widehat{\xi}_n \right) : 1 \leq k < n \right\},$$

dans le but de trouver une valeur optimale de  $k_n$ . Ainsi le  $k_n$  optimal (notée  $k_{opt}$ ) est choisi dans la première région où l'estimateur  $\widehat{\xi}_n$  devient stable c'est à dire le graphe devient horizontal. Cet estimateur est valable seulement dans le domaine d'attraction de Fréchet c'est à dire si  $\xi > 0$ . Pour généraliser aux autres domaines d'attractions, différents estimateurs ont été proposés, entre autres l'estimateur de Dekkers-Einmahl-de Hann et l'estimateur de Pickands.

## 2.3 Estimation des quantiles extrêmes et période de retour

Le choix de la loi pour les valeurs extrêmes et l'estimation de la fonction de répartition en déterminant l'indice de queue de distribution ne sont souvent qu'un objectif intermédiaire, l'objectif réel étant plutôt l'estimation d'un quantile extrême ou d'un niveau de retour ainsi que l'estimation d'une période de retour.

### 2.3.1 Estimation des quantiles extrêmes

Dans cette approche, on va utiliser une méthode nommée Méthode POT (PeakOver-Threshold) pour estimer les quantiles extrêmes. Cette méthode est basée sur l'approximation de la distribution des excès pour la loi de Pareto généralisée en ce sens qu'il est plus facile d'avoir un échantillon d'excès que de max. Dans la pratique, on remplace  $u$  par  $X_{n-k+1,n}$  qui représente la  $k$  plus grande observation de l'échantillon. Pour réaliser cette méthode, on va suivre les étapes suivantes :

- 1 Soit  $X_1, \dots, X_{N_u}$  un échantillon, à partir d'un certain seuil  $u$ , on note  $N_u$  le nombre d'observation qui dépassent ce seuil.
- 2 Soit  $Y_1, \dots, Y_{N_u}$  un échantillon des excès au dessus du seuil  $u$  de distribution conditionnelle

$$F_u(x) = P(X - u \leq x \mid X > u) = \frac{F(x + u) - F(u)}{1 - F(u)}.$$

D'après le théorème de Belkema-Pickands De Haan une approximation pour cette distribution on peut approximer cette distribution par :

$x > u$  et  $u \rightarrow \infty$

$$F_u(x) \approx G_{\xi, \sigma}(x)$$

Et de plus :

$$\bar{F}_u(x) \approx \bar{G}_{\xi, \sigma}(x) \tag{2.4}$$

- 3 La décomposition de  $F$  peut s'écrire de la manière suivante :

Pour  $x > u$

$$F(x) = P(X > u)F_u(x - u) + P(X \leq u).$$

Cette distribution est estimée par :

$$\hat{F}(x) = (1 - F_n(u))G_{\hat{\xi}, \hat{\sigma}}(x) + F_n(u).$$

Où  $F_n(u)$  est la fonction de répartition empirique au point  $u$ , tel que :

$$F_n(u) = 1 - \frac{N_u}{n}$$

L'estimateur pour la queue  $\bar{F}(u + x)$  prend la forme :

$$\hat{\bar{F}}(u + x) = \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x}{\hat{\sigma}}\right)^{-1/\hat{\xi}},$$

et

$$\widehat{F}(u+x) = 1 - \frac{N_u}{n} \left(1 + \widehat{\xi} \frac{x}{\widehat{\sigma}}\right)^{-1/\widehat{\xi}},$$

d'où

$$\widehat{F}(x) = \frac{N_u}{n} \left(1 + \widehat{\xi} \frac{x-u}{\widehat{\sigma}}\right)^{-1/\widehat{\xi}}.$$

4 On inverse la fonction  $\widehat{F}$  définie dans (2.4), on trouve l'estimateur du quantile extrême qui est donnée par :

$$\begin{aligned} \widehat{x}_p &= G_{\widehat{\xi}, \widehat{\sigma}}^{-1} \left( \frac{1-p-F_n(u)}{1-F_n(u)} \right) \\ &= u + \frac{\widehat{\sigma}}{\widehat{\xi}} \left( \left( \frac{n}{N_u} (1-p) \right)^{-\widehat{\xi}} - 1 \right). \end{aligned}$$

Donc :

$$\widehat{x}_p^{POT} = u + \frac{\widehat{\sigma}}{\widehat{\xi}} \left( \left( \frac{n}{N_u} (1-p) \right)^{-\widehat{\xi}} - 1 \right). \quad (2.5)$$

### 2.3.2 Estimation de période de retour

On sait que l'estimation de la période de retour  $T$  d'un quantile  $x_{POT}$  revient à estimer l'ordre de ce quantile  $\widehat{p}_T = 1 - \frac{1}{T}$ . Selon (2.5) nous estimons la période de retour pour la loi des valeurs extrêmes GPD par :

$$\widehat{T}_p^{POT} = \frac{n}{N_u} \frac{1}{1 - G_{\widehat{\xi}, \widehat{\sigma}}(x)}, \quad \forall \widehat{\xi} \in \mathbb{R} \text{ et } \sigma > 0.$$

## Chapitre 3

# Application : Modélisation via la distribution de Pareto généralisée (GPD)

On s'intéresse dans notre application à la modélisation de la distribution de Pareto généralisée (GPD) sur des données réelles telles que Pertes de compagnie d'assurance automobile. Notre base de données s'étale du 31 janvier 2020 à 31 décembre 2023 fournie par la direction Agence CIAR Touggourt. On vise au cours de notre étude pratique à :

- Faire une analyse descriptive des données.
- Sélection de seuil.
- Estimation des paramètres du modèle GPD.
- Estimation par des méthodes semi-paramétrique.
- Estimation de niveau de retour.

La réalisation de cette application nécessite de faire appel à des commandes et des fonctions prédéfinies sous le logiciel R.

**R** est un logiciel de traitement statistique des données et un langage de programmation. Il fonctionne sous la forme d'un interpréteur de commandes. Il dispose d'une bibliothèque très large de fonctions statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des "packages", des modules externes compilés (sous forme de DLL sous Windows) que l'on peut télécharger gratuitement sur internet. R propose également une palette étendue de fonctionnalités graphiques. Il est possible d'utiliser R en mode interactif sans jamais avoir à programmer.

De plus il est libre, donc gratuit. R fonctionne sur tout ordinateur (Mac, PC, Windows, Linux), R est un logiciel d'excellente qualité, utilisé notamment par de nombreux chercheurs en statistique.

## 3.1 Les données

Les données sont des pertes de la compagnie d'assurance automobile pendant quatre ans (en mois).

	<b>2020</b>	<b>2021</b>	<b>2022</b>	<b>2023</b>
<b>janvier</b>	1156586	926489.87	1120510.38	1281206.15
<b>février</b>	1495093.41	737763.02	700368.12	1335906.80
<b>mars</b>	2097321.6	954744.55	747471.23	1575796.23
<b>avril</b>	1539917.21	838977.25	1174436.79	1662007.01
<b>mai</b>	1723533.78	1944874.14	617774.77	2332900.29
<b>juin</b>	1610588.83	1563328.6	1822986.5	1717210.52
<b>juillet</b>	1115547.73	620923.36	728408.96	803291.33
<b>aout</b>	1041389.19	1241216.74	841380	1081437.14
<b>septembre</b>	3466933.28	910029.65	875110.79	971893.05
<b>octobre</b>	704179.43	335188.72	2042784.1	1852697.14
<b>novembre</b>	1182532.79	523794.51	1508634.51	1636586.36
<b>décembre</b>	1086667.05	2419634.58	1359084.5	362250.65

TAB. 3.1 – Pertes de la compagnie d'assurance automobile.

### 3.1.1 Analyse descriptive de la série

D'abord, nous faisons une simple analyse statistique sur notre série des données, les résultats selon le langage R sont les suivants :

Min	1st Qu	Median	Mean	3rd Qu	Max	skewness	kurtosis
335119	840779	1165511	1278944	1617088	3466933	1.121125	5.107274

TAB. 3.2 – Statistique des données.

On remarque que la plage des variations des données est entre 335119 et 3466933 et la moyenne est 1278944 (DA). La kurtosis est supérieure à 3, donc la distribution à queue lourde.

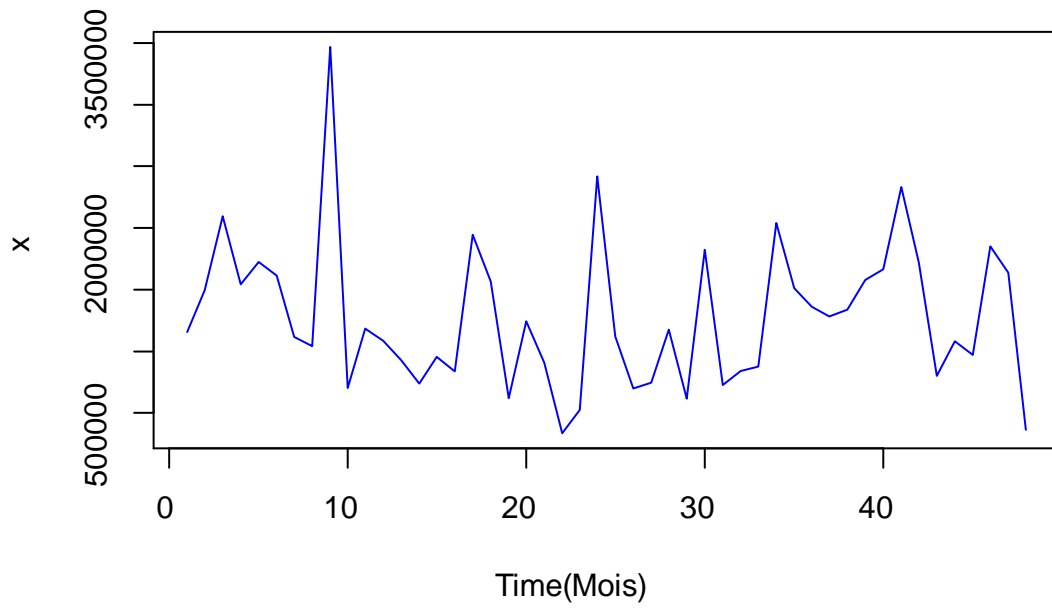


FIG. 3.1 – Pertes des compagnies d'assurance automobile du 31/1/2020 à 31/12/2023

Histogramma des données est :

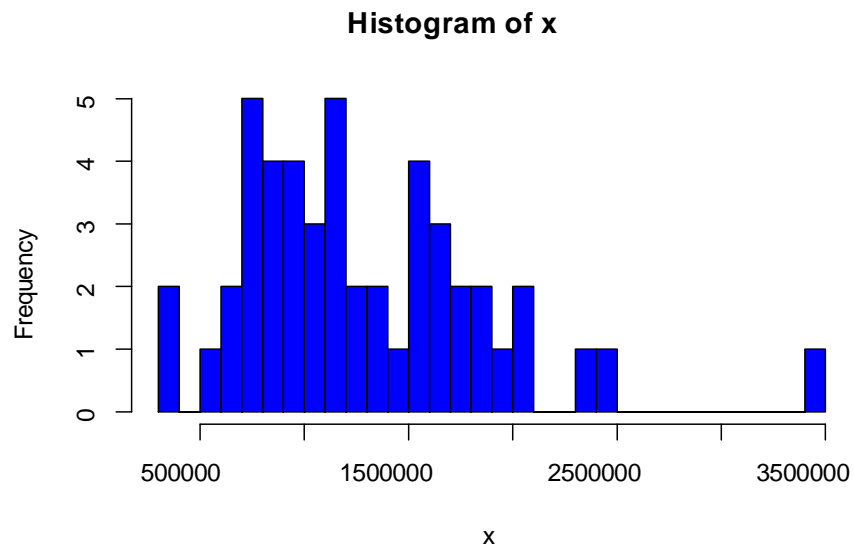


FIG. 3.2 – Histogramme de pertes de la compagnie d'assurance automobile.



### 3.1.2 L'ajustement par la loi normale

Pour l'ajustement par la loi normale en utilisant le test de shapiro wilk.

On teste les hypothèses suivantes :

$$\begin{cases} H_0 : \text{"L'échantillon suit une loi normale"}. \\ H_1 : \text{"L'échantillon ne suit pas une loi normale"}. \end{cases}$$

On fixe un seuil de signification  $\alpha = 0.05$ .

le résultat :

p-value = 0.008266.

#### Décision du test :

L'ajustement des données de notre échantillon par la loi normale donne une p-valeur (valeur renvoyée par logiciel R) égale à 0.008266, cette valeur est inférieure au seuil de signification  $\alpha = 0.05$ , donc l'échantillon ne suit pas une loi normale.

## 3.2 Sélection de seuil

Dans un cadre statistique, le choix du seuil  $u$  est très important car il induit une grande variabilité dans l'estimation des quantiles extrêmes et des paramètres de la loi des excès. Il existe différentes approches pour le choix du seuil  $u$  de la méthode POT. En effet, le seuil doit être suffisamment grand pour satisfaire la caractéristique asymptotique du modèle, mais pas trop élevé non plus, afin de garder un nombre d'excès suffisant pour estimer convenablement les paramètres du modèle. Généralement,  $u$  est déterminé graphiquement par le graphe de la fonction moyenne des excès (Mean Excess Plot), voir la Figure 3.3 .

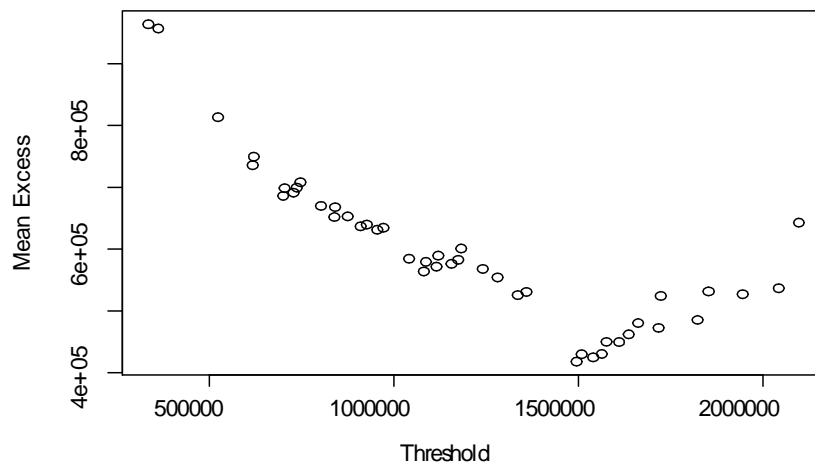


FIG. 3.3 – La distribution moyenne des excès.

le graphique semble linéaire à partir de seuil 1495093 environs, ça nous a permis de considérer la valeur  $u = 1495093$  comme seuil optimale.

### 3.3 Estimation des paramètres du modèle GPD

Dans cette partie on va estimer les paramètres  $\xi$  et  $\sigma$  de la GPD par les méthodes suivantes :

- Méthode de maximum de vraisemblance (MV).
- Méthode des moments pondérés (MP).
- Méthode L-moments (LM).

Nous obtenons les résultats suivants :

	$\hat{\xi}$	$\hat{\sigma}$
<b>MV</b>	2.963979e-01	2.717449e+05
<b>MP</b>	2.635886e-01	2.905206e+05
<b>LM</b>	3.111811e-01	2.717449e+05

TAB. 3.3 – Estimation des paramètres de la GPD.

- \* On remarque que le paramètre de la forme  $\xi$  estimé par les méthode (MV, MP, LM) est strictement supérieur à 0 , donc notre série suit une loi de Pareto usuelle.

### 3.4 Estimation par des méthodes semi-paramétrique

On utilisons quelque méthodes semi-paramétrique pour estimer l'indice de valeur extrême  $\xi$  (Hill, Moments, Pickands). Selon le langage R les résultats sont les suivants :

$\hat{\xi}^H$	$\hat{\xi}^M$	$\hat{\xi}_{K_n}^P$
0.1798269	0.3343327	0.3674706

TAB. 3.4 – Estimation semi-paramétrique.

### 3.5 Estimation de niveau de retour

Le niveau de retour pour la loi des valeurs extrêmes GPD dans logiciel R en années est :

Périodes	MV	LM
2	5417775	5630228
20	10154596	10875521
40	12338710	13343822
60	13840503	15054658

TAB. 3.5 – Niveau de retour pour GPD.

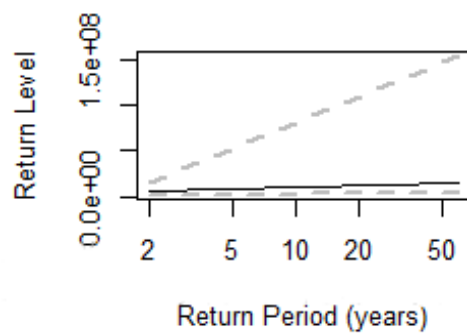


FIG. 3.4 – Représentation graphique de niveau de retour dans le cas GPD.

## 3.6 Le programme d'application sous R

### \* Les données

```
x<-c(1156586,1495093.41,2097321.6,1539917.21,1723533.78,1610588.83,1115547.73,
1041389.19,3466933.28,704179.43,1182532.79,1086667.05,926489.87,737763.02,954744.55,
838977.25,1944874.14,1563328.6,620923.36,1241216.74,910029.65,335188.72,523794.51,
2419634.58,1120510.38,700368.12,747471.23,1174436.79,617774.77,1822986.5,728408.96,
841380,875110.79,2042784.1,1508634.51,1359084.5,1281206.15,1335906.80,1575796.23,
1662007.01,2332900.29,1717210.52,803291.33,1081437.14,971893.05,1852697.14,
1636586.36,362250.65)
```

### \* Analyse descriptive

```
>summary(x)
>skewness(x)
>kurtosis(x)

#tracer le graphe et histogramme de la série
>plot(x,type='l',xlab="Time(Mois)",col="Blue")
> hist(x,breaks = 24,col="Blue")
```

### \* Détermination du seuil.

```
>meplot(x)
```

### \* Estimation des paramètres de la GPD

```
#Pour méthodes(maximum de vraisemblance, moments Pondérés et
L-moments).
```

```
a<- fevd(x,threshold=1495093,method="MLE",type="GP")
b<- gpdFit(x,u=1495093,type="pwm")
c<- fevd(x,threshold=1495093,method="Lmoment",type="GP")
```

### \* Estimation par des méthodes semi-paramétrique

```
#Le choix du nombre k
# Définition de la fonction pour calculer MSE
calculate_mse <- fonction(data, k) {
n <- length(data)
sorted_data <- sort(data)
mean_excess <- numeric(length = n - k)
for (i in 1 :(n - k)) {
mean_excess[i] <- mean(sorted_data[(i + 1) :n] - sorted_data[i])
}
```

```

}
pickands_estimator <- sum(mean_excess) / (n - k)
mse <- sum((mean_excess - pickands_estimator)^2) / (n - k)
return(mse)
}
# Définition de la fonction pour choisir la valeur optimale de k en utilisant
MSE
choose_k_mse <- fonction(data, max_k) {
mse_values <- numeric(max_k)
for (k in 1 :max_k) {
mse_values[k] <- calculate_mse(data, k)
}
return(which.min(mse_values))
}
data <- x
# Choisir la valeur optimale de k
optimal_k <- choose_k_mse(data, max_k = 12)
# Affichage du résultat
cat("Nombre optimal de blocs (k) :", optimal_k, "\n")
#Estimateur de Hill
# Fonction pour calculer l'estimateur de Hill
hill_estimator <- fonction(data, k) {
# Trier les données dans l'ordre croissant
sorted_data <- sort(data)
# Calcul de l'estimateur de Hill
hill <- (1/k) * sum(log(sorted_data[(length(sorted_data) - (1 :k) + 1)]) -
log(sorted_data[length(sorted_data) - k+1]))
return(hill)}
data <- x
k <- 2
# Appel de la fonction
hill <- hill_estimator(data, k)
# Affichage de l'estimation
print(hill)
#Estimateur des Moments
# Fonction pour calculer l'estimateur de moments
moments_estimator <- fonction(data, k) {
# Trier les données dans l'ordre croissant
sorted_data <- sort(data)
#calcul M2

```

```

M2<- (1/k) * sum((log(sorted_data[(length(sorted_data) - (1 :k) + 1)]) -
log(sorted_data[length(sorted_data) - k]))^2)
return(M2)
}
data <- x
k <- 2
# Appel de la fonction
M2<- moments_estimator(data, k)
hill<- 0.1798269
# Calcul de l'estimateur de moments
moments<-hill+1-(1/2)*((1-((hill^2)/M2))^(-1))
print(moments)
#Estimateur de Pickands
# Fonction pour calculer l'estimateur de pickands
pickands_estimator <- function(data, k) {
# Trier les données dans l'ordre croissant
sorted_data <- sort(data)
# Calcul de l'estimateur de pickands
pickands <- (1/2) * log(((sorted_data[length(sorted_data) - k+1])
-(sorted_data[length(sorted_data)-((2 * k)+1)])) / (((sorted_data[length(sorted_data)
-((2*k)+1)])-(sorted_data[length(sorted_data) - ((4 * k)+1)]))))))
return(pickands)
}
data <- x
k <- 2
# Appel de la fonction
pickands <- pickands_estimator(data, k)
# Affichage de l'estimation
print(pickands)

```

**\* calculée le niveau de ratour**

**#Pour méthodes MV et LM.**

```
>return.level(a, return.period = c(2, 20,40, 60),do.u = T)
```

```
>return.level(c, return.period = c(2, 20,40, 60),do.u = T)
```

**\* les packages utilisé est :**

moments,POT,evd,extRemes,fExtremes,evir,mgcv,Lmoments,nlme,ismev,evmix.

# Conclusion

L'utilisation du modèle de distribution généralisée de Pareto (GPD) dans l'analyse et la modélisation des queues de distribution est extrêmement utile. Ce modèle nous permet de mieux comprendre la zone de la queue des distributions, qui contient souvent des valeurs aberrantes et peut être très importante dans de nombreux domaines tels que l'assurance, la finance, l'ingénierie et l'environnement.

Dans ce travail, nous avons introduit des définitions et des propriétés des valeurs extrêmes et présentons différentes méthodes d'estimation de l'indice des valeurs extrêmes paramétrique et semi-paramétrique de loi GPD.

Nous modélisons les données réelles "pertes des compagnies d'assurance automobile" durant une période 2020 et 2023 en mois par la direction du concessionnaire Agence CIAR Touggourt.

Dans ce cadre on a arrivé à l'objectif de notre application qui est l'estimation la période de retour et le niveau de retour.

Nous concluons du modèle GPD qu'il importance permettant responsables de mettre en place une politique pour éviter toute catastrophe en cas de période de retour pertes grand.



# Bibliographie

- [1] Balkema, A. A., De Haan, L., (1974). Residual life time at great age. *The Annals of probability*, 792–804.
- [2] BI'LKOVA', D , L-moments and tl-moments as an alternative tool of statistical data analysis. *Journal of Applied Mathematics and Physics* (2014), 919–929.
- [3] Dekkers, A. L., De Haan, L., et al., 1989a. On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics* 17 (4), 1795–1832.
- [4] Dekkers, A. L., Einmahl, J. H., De Haan, L., 1989b. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 1833–1855.
- [5] Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. *Modelling extremal events : for insurance and finance*. Vol. 33. Springer Science & Business Media.
- [6] Fisher, R. A., Tippett, L. H. C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In : *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 24. Cambridge University Press, pp. 180–190.
- [7] Girard, S., Guillou, A. et Stupřer, G. (2012). Estimating an endpoint with high order moments in the Weibull domain of attraction. *Statistics and Probability Letters*, 82 :2136–2144.
- [8] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics*, 44(3) :423–453.
- [9] Gumbel, E. (1954). *Statistical theory of extreme values and some practical applications : a series of lectures*. Numéro Applied Mathematics Series, 33. National Bureau of Standards, Washington.
- [10] Gumbel, E. (1958). *Statistics of extremes*. Columbia University Press, Columbia.
- [11] Hill, B. M., (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163–1174.

- [12] Hosking, J. et Wallis, J.,(1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics* 29, (3), 339–349.
- [13] Hosking, J. R. M. et Wallis. J .R. (1990). L-moments : analysis and estimation of distributions using linear combinations of statistics, Royal Statistical Society.
- [14] Mason, D. M., (1982). Laws of large numbers for sums of extreme values. *The Annals of Probability*, 754–764.
- [15] Peng, L., (1998). Asymptotically unbiased estimators for the extreme-value index. *Statistics & Probability Letters* 38 (2), 107–115.
- [16] Pickands III, J., et al., 1975. Statistical inference using extreme order statistics. *the Annals of Statistics* 3 (1), 119–131.
- [17] Reiss, R.-D., Thomas, M., 2007. *Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields*, 3rd Edition. Birkhkauser Verlag, Basel, with 1 CD-ROM (Windows).
- [18] Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer Verlag.
- [19] Sillitto, G.P. (1951). Interrelations between certain linear systematic statistics of samples from any continuous population. *Biometrika* 38, 377-382.
- [20] Von Mises, R. (1954). La distribution de la plus grande de n valeurs. *American Mathematical Society, Providence, RI, II(In Selected Papers)* :271–294.

## Résumé

La modélisation des queues de distribution par la loi de pareto généralisée (GPD) à une grande importance dans la prévention des phénomènes extrêmes qui peuvent engendrer des catastrophes. La modélisation par la loi de GPD passe par l'estimation de ces paramètres.

Nous avons modélisé par la distribution de GPD les données réelles de CIAR Touggourt management en assurance (automobile) entre le 31 janvier 2020 et le 31 décembre 2023 afin d'estimer le niveau de retour et de déterminer les niveaux de réserves financières nécessaires pour couvrir les pertes dans les cas extrêmes.

### **Mots-clés**

Distribution de Pareto généralisée (GPD), valeurs extrêmes, niveau de retour.

## Abstract

The modeling of distribution tails by the generalized pareto law (GPD) is of great importance in the prevention of extreme phenomena that can generate catastrophes. The modeling by GPD's law involves estimating these parameters.

We modeled by the distribution of GPD the real data of CIAR Touggourt management in insurance (automobile) between January 31, 2020 and December 31, 2023 in order to estimate the level of return and to determine the levels of financial reserves necessary to cover the losses in extreme cases.

### **Keywords**

Generalized Pareto distribution (GPD), extreme values, return level.

## المخلص

نمذجة ذيول التوزيع من قبل قانون باريتو المعمم (GPD) له أهمية كبيرة في الوقاية من الظواهر المتطرفة التي يمكن أن تسبب الكوارث. النمذجة من قبل قانون GPD ينطوي على تقدير هذه المعلمات.

لقد قمنا بنمذجة من خلال توزيع GPD البيانات الحقيقية لإدارة CIAR تفرقت في التأمين (السيارات) في الفترة ما بين 31 يناير 2020 و 31 ديسمبر 2023 من أجل تقدير مستوى العائد وتحديد مستويات الاحتياطيات المالية اللازمة لتغطية الخسائر في الحالات القصوى.

### **الكلمات المفتاحية**

توزيع باريتو المعمم (GPD), القيم المتطرفة, مستوى العودة.