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

BENABED Rabah

Thème

Estimation de la moyenne d'une distribution à queue lourde en présence de censure

Membres du jury

BOUSAAD Abdelmalek	Maître assistant classe "A"	UKMO	Président
MANSOUL Brahim	Maître assistant classe "A"	UKMO	Examinateur
ARBIA Hanane	Maître assistant classe "A"	UKMO	Rapporteur

novembre 2020

Dédicace

Je dédie ce modeste travail

A ma tres chère mère et mon trés cher père

A mes cheres soeurs et frères

A toute la famille BENABED

A mes chers amis

Je tiens à remercier tous les membres de ma promotion

Et tous mes professeurs

Finalement à tous ceux qui m'ont aidée de proche ou de loin

Remerciements

J'exprime d'abord mon profond remerciements à Dieu qui jai donné le courage et la volonté d'achever ce travail.

Je remercier particulièrement mes parents ; notre succès demeure de loin le fruit de leurs longues années de sacrifices et d'éducation.

Mon sentiments de reconnaissance et mon remerciements vont à ma Promotrice Dr.ARBIA H, pour ses conseils, ses encouragements, sa patience, sa compétence, ses qualités humaines et scientifiques resterons pour nous un exemple, qui nous en permis de bien mener ce travail. Le suivi et l'orientation dont nous avons pu bénéficier.

Mon remerciements vont au président de jury Pr. BOUSAAD Abdelmalek, merci de nous avoir fait l'honneur d'accepter de présider ce jury.

Mon vif remerciements s'adressent aussi au Pr. MANSOUL Brahim, qui a accepté d'examiner mon travail avec la bienveillance et je suis très honoré.

Un remerciement spécial à notre responsable de spécialité Pr. BEHDDI M, pour sa disponibilité et sa sympathie.

Mon remerciements les plus chaleureux s'adressent à ma soeur MOKHTAR .H, qui a contribué à la réalisation de ce modeste travail.

Merci à tout

Table des matières

Table des figures

Notations et abriviations

EVD	:	Distribution des valeurs extrêmes
EVI, γ	:	Indice des valeurs extrêmes
F	:	Fonction de répartion $(f\ dr)$
F_n	:	Fonction de répartion empirique
$F^{\longleftarrow}$	:	Iverse généralisé de F
GEV	:	Distribution des valeurs extrêmes généralisée
GPD	:	Distribution de pareto généralisée
G_γ	:	Famille de la loi de valeurs extrêmes généralisée
iid	:	Indépendantes et identiqument distribuées
$I_{\{A\}}$	:	Fonction indicatrices de l'ensemble A
A	:	loi de Gumbel

$\ell(x)$	:	Fonction à variation lente
DA	:	Domaine d'atraction de maximum
$M_n = X_{n,n}$	:	Maxcimum de $X_1, ..., X_n$
$p.s$	:	Prèsque sûre
Φ	:	Loi de Fréchet
Ψ	:	Loi de Weibull
Λ	:	Loi de Gumbel
$resp$	:	Respectivement
$S = \overline{F}$	:	$1 - F$ fonction de survie
TEV	:	Théorème des valeurs extrêmes
$X_{1:n}, ..., X_{n:n}$	:	Statistique d'ordre asscociées à $X_1, ..., X_n$
$X \wedge Y$	:	$\min(X, Y)$
x_F	:	Point terminal

$\widehat{\gamma}^{(c,H)}$	:	Estimation de Hill avec les donnéées censurées
(Ω, A, P)	:	Espace probabilité
$N(0,1)$	:	Loi normal standard
$inf\ A$	:	Supermum de l'ensemble
$\overset{p.s}{\rightarrow}$	:	Converge prèsque sûre
$\overset{l}{\rightarrow}$	:	Converge en loi
$\overset{d}{\rightarrow}$	:	Converge en distribution
$\overset{p}{\rightarrow}$	:	Convere en probabilité
Q	:	Fonction de quantile
Q_n	:	Quantile empirique

Introduction

La théorie des valeurs extrêmes a fourni un cadre dans lequel une estimation des forces anticipées pourrait être faite à l'aide de données historiques.

Apparue au cours des dernières années la modélisation du valeur extrêmes dans le cas censure. En 1997, et en 2007 il a été étudié par *Beirlant et al* [1]. L'étude a continué jusqu'en 2008, ou *Einmahl et al* [6], il a présenté l'indice est de diviser l'indicateur de *Hill* (1975) [10]sur $\widehat{p}$.

Estimation de la moyenne d'un distribution à queue lourde en présence de censure ont un grand intérêt dans plusieurs domaines matématiques. Notre sujet se divise en trois chapitres :

Dans le premier chapitre, on a présenté une introduction sur la théorie des valeurs extrêmes et on a définit les statistiques d'ordre, ainsi que les lois exactes des statistiques d'ordre et les lois asymptotiques des valeurs extrêmes. Ensuite, on a donné le résultat fondamontal de la distribution *GEV*, ainsi que les caractéristiques des différents domaines d'attraction du maximum puis on a introduit la distribution *GPD* et le théorème de Belkema et de Haan, et on a donné les estimateurs classiques de l'indice de queue tels l'estimateur de Hill, de Pickands et des Moments.

Dans le deuxième chapitre, nous rappelons des préliminaires sur les modèles de survie. Nous introduisons les principales fonctions en analyse de survie : fonction de survie et les différentes formes du taux de risque...ect. Nous donnons aussi les différents types de censure (censure à droite, censure à gauche, censure par intervalle, ect....) et on a présenté les principaux estimateurs non-paramétriques qui sont l'estimateur de Kaplan-Meier de fonction de survie .Dans la dernière partie on présentera l'estimateur de l'indice des valeurs extrêmes avec censure.

Le troixième chapitre traite l'estimation des moyennes aléatoires. L'objectif principal de ce chapitre est de proposer une méthode pour estimer la moyenne de distribution à queue lourde en présence de censure.

Chapitre 1

THÉORIE DES VALEURS EXTRÉMES

La théorie des valeurs extrêmes(**TVE**) communément appelée **«Extrêrme Value Theory» (EVT)** en anglais, est une vaste théorie dont le but est d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'appartition est faible. Autrement dit, elle essaie d'amener des éléments de réponses aux intempéries, aux catastrophes naturelles, aux problemes financiers, ... etc. Les ouvrages de *Reiss et Thomas (2007)*[18]*, Embrechts et al.(1997)*[7]*,Beirlant et al.(2007)*[1], qui font le point sur les différentes techniques existantes.

1.1 Statistique d'Ordre

Les statistiques d'ordre fournissent des informations sur la distribution de queue, pour ça ils sont trés importantes dans la théorie des valeurs extremes. en effet, de façon naturelle et depuis long-temps, dans les problèmes de données censurées ou tronquées.

Soit $X_1, X_2, ..., X_n$ n variables aléatoires iid de distribution commune F. les va's $X_{1,n}$, $X_{2,n}$, ..., $X_{n,n}$ sont rangés par ordre croissante, soit :

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

sont appelées les statistiques d'ordre de l'échantillon X_1, X_2, ..., X_n.

Deux statistiques d'otdre sont particulièrement intéresantes pour l'étude des événements extrêmes. Ce sont les statistiques d'ordre extrêmes qui sont données par la définition suivante :

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

On note qu 'il est trés facile de passer de l'un à l'autre à l'aide de la relation :

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

Dans la suite de ce mémoire, on se concentrera sur l'étude du maximum.

1.1.1 Distribution d'une statistique d'ordre

Loi de $X_{i,n}$.

$$F_{i,n} = P\left\{X_{i,n} \leq x\right\} = \sum_{r=i}^{n} C_r^n (F(x))^r (1 - F(x))^{n-r}.$$

Nous en déduisons que la fonction de densité est :

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

où $f(x)$ est la densité de probabilité de X_i et F sa fonction de répartition associée.

Loi de $X_{1,n}$
On a

$$
\begin{aligned}
\{X_{1,n} \geqslant x\} &\Longleftrightarrow \{\min(X_1, ..., X_n) \geqslant x\} \\
&\Longleftrightarrow \bigcap_{i=1}^{n} \{X_i \geqslant x\}
\end{aligned}
$$

En utilisant la propriété d'indépendance des variables aléatoires $X_1, ...,$ X_n nous en déduisons que :

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

d'où,

$$
f_{1,n}(x) = nf(x)(1 - F(x))^{n-1}
$$

Loi de $X_{n,n}$
On a

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

En utilisant la propriété d'indépendance des variables aléatoires $X_1, ...,$ X_n, nous en déduisons que :

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

d'où,

$$f_{n,n}(x) = nf(x)(F(x))^{n-1}$$

1.2 Définitions

Définition 1.1 (La fonction de répartition empirique) *La fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ notée F_n est donnée par :*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,x[}(X_i), \qquad x \in \mathbb{R}.$$

Il existe une autre version de la définition de F_n :

$$F_n(x) = \begin{cases} 0 & \text{si,} \quad x \leq X_{1,n} \\ \frac{i-1}{n} & \text{si,} \quad X_{i-1,n} < x \leq X_{i,n} \qquad 2 \leq i < n \\ 1 & \text{si,} \quad x > X_{n,n} \end{cases}$$

Définition 1.2 (Les fonctions de quantile et de quantile de queue) *On défnit la fonction des quantiles Q par :*

$$Q(s) = F^{\leftarrow}(s) = \inf\left\{x \in \mathbb{R} : F(x) \geq s\right\}, \qquad 0 < s < 1,$$

où $F^{\longleftarrow}$ est l'inverse généralisée de F. Dans la théorie des extrêmes une fonction, notée par U et (parfois) appelée la fonction quantile de queue, est utilisée assez souvent et elle est défnie comme :

$$U(t) = Q(1 - 1/t) = (1/\bar{F})^{\leftarrow}(t), \qquad 1 < t < \infty,$$

où

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

Définition 1.3 (La fonction de survie) *La fonction de survie est pour t fixé, probabitité de survivre jusqu'à l'instant t, c'est-à-dire pour $t \geq 0$*

$$\begin{aligned} S(t) &= 1 - F(x) \\ &= P(X > t) \end{aligned}$$

Définition 1.4 (Fonctions empiriques de répartition et de survie) *Soit $X_1, ..., X_n$ un échantillon de taille $n \geq 1$ d'une v.a positive X de fonction de répartition F et de fonction de survie $\overline{F}$. Les fonctions empiriques de répartition et de survie, F_n et $\overline{F}_n$ sont respectivement définies par :*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{X_i \leq t\right\}, \quad \forall t \geq 0$$

et

$$\begin{aligned} \overline{F}_n(t) &= 1 - F_n(t) \\ &= \frac{1}{n}\sum_{i=1}^{n} 1\left\{X_i \leq t\right\}, \quad \forall t \geq 0 \end{aligned}$$

où $1\{A\}$ est la fonction indicatrice de l'ensemble A.

Théorème 1.1 (Théorème Central Limite) *Soit $X_1, ..., X_n$ une suite de variables aléatoires i.i.d de moyenne μ et de variance σ^2 finie, alors :*

$$\sqrt{n}\frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0,1) \quad quand \quad n \longrightarrow \infty$$

1.3 Comportement asymptotique des extrêmes

Nous posons

$$M_n = \max(X_1, ..., X_n).$$

Nous tirons la conclusion que le maximum M_n est une variable aléatoire dont la fonction de répartition est égale à $(F_X)^n$. La fonction de répartition de X étant souvent inconnue et généralement pas possible d'être déterminée. Notons $x_F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$ le point terminal à droite de la fonctionde répartition F_X. Ce point terminal peut être infini ou fini *(Embrechts et al. 1997*[7]*)*. On s'intéresse ici au distribution asymptotique du maximum, en faisant tendre n vers l'infini,

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

On constate que la distribution asymptotique du maximum, donne une loi dégénérée, une masse de Dirac en x_F, puisque pour certaines valeurs de x, la probabilité peut être égale à 1 dans le cas où x_F est fini. Donc M_n tend vers x_F presque sûrement, Ce fait ne fournit pas assez d'informations. On s'intéresse par conséquent une loi non dégénérée pour le maximum, la théorie des valeurs extrêmes permet de donner une réponse cette problématique. Les premièrs résultats sur la caractérisation du comportement asymptotique des maxima M_n convenablement normalisés et donnés par la suite :

1.3.1 Distribution des valeurs extrêmes

Théorème 1.2 (Fisher et Tippett, 1928, Gnedenkon, 1943) [8] , [9] *Soit $X_1, ..., X_n$ une suite de n variables aléatoires réelles i.i.d de loi continue P et $M_n = \max(X_1, ..., X_n)$. S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$, avec $b_n > 0$, et une fonction de répartition non-dégénérée G_γ telle que,*

$$
\begin{aligned}
\lim_{n\to\infty} P\left[\frac{M_n - b_n}{a_n} \leq x\right] &= \lim_{n\to\infty} \left[F\left(a_n x + b_n\right)\right]^n \\
&= G_\gamma(x), \quad \forall x \in \mathbb{R}
\end{aligned}
$$

Alors G_γ est du même type qu 'une des trois lois suivantes :

loi de Gumbel :

$$\Lambda_\gamma(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty$$

loi de Frechet :

$$\Phi_\gamma(x) = \begin{cases} 0 & x < 0 \\ \\ \exp(-x^{-1/\gamma}) & x \geq 0, \gamma > 0, \end{cases}$$

loi de Weibull :

$$\Psi_\gamma(x) = \begin{cases} \exp\left(-\left(-x\right)^{-1/\gamma}\right) & x < 0,\, \gamma < 0 \ , \\ \\ 1 & x \geq 0, \end{cases}$$

avec, G_γ est la loi des valeurs extrêmes, γ est l'indice des valeurs extrêmes et a_n et b_n sont des paramètres de normalisation. Ce théorème donne la forme des lois limites de G_γ.

1.4 Distributions GEV

Il est difficile de travailler avec trois familles à la fois, *Jenkinson* en *1955*[11] montre que ces trois familles peuvent être regroupées sous une forme unique dite famille des lois des valeurs extrêmes généralisées (GEV, Generalized Extreme Value distribution).

Définition 1.5 (Distribution GEV**)** *La fonction de répartition de la famille G des valeurs extrêmes généralisées GEV, est pour $\gamma \in \mathbb{R}$ et $1+\gamma x > 0$*

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

La forme la plus générale de la distribution des valeurs extrêmes est :

$$G_{\gamma,\mu,\sigma} = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right), \gamma \neq 0, 1+\gamma\frac{x-\mu}{\sigma} > 0.$$

où G est une fonction de répartition non-dégénérée. μ est un paramètre de localisation, il est directement lié à la valeur la plus probable de la loi, il indique donc approximativement où se trouve le coeur de la distribution. σ est un paramètre de dispersion, il indique l'étalement des extrêmes. γ est l'indice de queue.

Remarque 1.1 ·*Les trois lois des extrêmes peuvent être représentés en fonction de G_γ par*

$$G_\gamma = \begin{cases} \Lambda(x) & \gamma = 0 \\ \Phi_{\frac{1}{\gamma}}(x) & \gamma > 0 \\ \Psi_{-\frac{1}{\gamma}}(x) & \gamma < 0 \end{cases}$$

La densité correspondante à $G_{\gamma,\mu,\sigma}$ est :

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

Cette densités est présentée dans la figure suivante

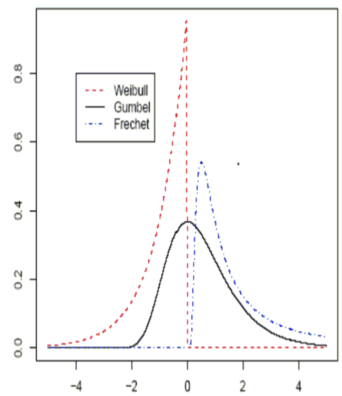

FIG. 1.1 – Fonction de distribution de Gumbel (en noire) et de Fréchet (en bleu) et de Weibull (en rouge).

1.5 Domaines d'attraction

Le théorème précédant découle immédiatement que le comportement de la queue de distribution d'une fonction est complètement caractérisé par un unique paramètre noté γ, et appelé indice des valeurs extrêmes. Le signe de ce paramètre est un indicateur essentiel sur la forme de la queue de distribution. il faut donc distinguer les troix cas possibles :

- Si $\gamma > 0$, F appartient au domaine d'attraction de Fréchet,et l'on note $F \in D(Fréchet)$, Il contient toutes les lois dont la fonction de survie décroitcomme une fonction puissance. Ce sont les lois à «queue lourde» ou lois de type Pareto, ces lois ont un point terminal x_F infini, ex : Burr, Pareto strict,...ect.
- Si $\gamma = 0$, F appartient au domaine d'attraction de Gumbel, et l'on note $F \in D(Gumbel)$, les queues des lois appartenant à ce domaine décroissance de manière exponentielle (i.e. les lois à queue légères) et le point terminal x_F peut être fini ou non, ex : Gamma, Logestic,...ect.
- Si $\gamma < 0$, F appartient au domaine d'attraction de Weibull, et l'on note $F \in D(Weibll)$, les lois de ce domaine d'attraction ont un point terminal x_F fini, ex : uniforme,...ect.

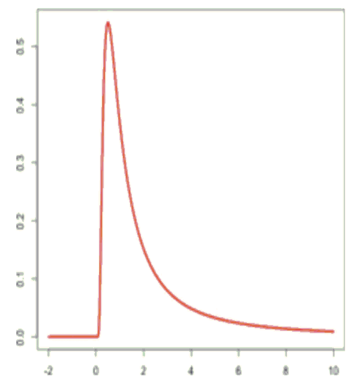

FIG. 1.2 – Queue lourde.

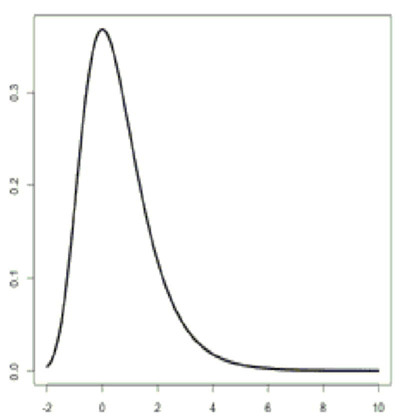

FIG. 1.3 – Queue légère.

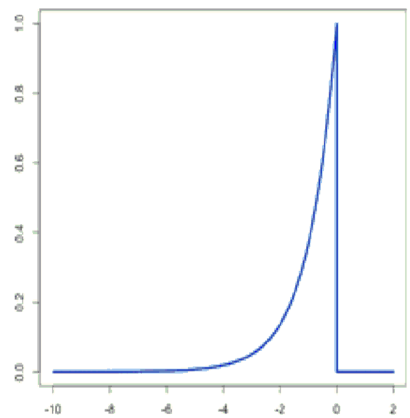

FIG. 1.4 – Queue finie.

1.5.1 Caractérisations des domaines d'attraction

Nous donnons dans cette partie la caractérisation des trois domaines dattractions, Fréchet, Weibull et Gumbel.

D'abord, nous donnons la définition d'une fonction variation régulière parce quelle est présente dans la caractérisation des domaines dattraction.

Définition 1.6 (Fonctions à variation régulière) *Une fonction mesurable $f : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ est à variation régulière l'infini si et seulement si, il existe un réel α tel que, pour tout $x > 0$,*

$$\lim_{x\to\infty} \frac{f(tx)}{f(t)} = x^{\alpha}$$

Et on note $f \in VR_\alpha, \alpha$ est appelé indice de la fonction variation régulière.

Proposition 1.1 *Soient $\alpha \in \mathbb{R}$ et $f \in VR_\alpha$ alors il existe une fonction à variation lente l à l'infini telle que :*

$$\forall x > 0;\ f(x) = x^{\alpha} l(x).$$

Définition 1.7 *Une fonction de répartition F sur $\mathbb{R}$ appartient une classe à variation régulière VR s'il existe $\alpha \geq 0$ tel que $1 - F \in VR_{-\alpha}$ sur $\mathbb{R}$, ou d'une manière équivalente :*

$$1 - F(x)\tilde{}x^{-\alpha}l(x),\ quand\ x \to \infty,$$

pour certaines $\ell \in RV_0$.

Domaine d'attraction de Fréchet

Théorème 1.3 *Une fonction de répartition F appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $\gamma > 0$ si et seulement si*

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

et sa fonction de survie $\overline{F} \in RV_{-1/\gamma}$:

$$1 - F(x) = x^{-1/\gamma}l(x)$$

Dans ce cas un choix possible des suites de normalisation a_n et b_n est :

$$a_n = \overline{F}(1 - \frac{1}{n}) = \overline{F}^{\longleftarrow}(\frac{1}{n})\ et\ b_n = 0$$

Domaine d'attraction de Gumbel

Théorème 1.4 *Une fonction de répartition F appartient au domaine d'attraction de Gumbel si et seulement si*

$$x_F \leq \infty\ et\ \overline{F}(x) = c(x)\exp\{-\int_y^x \frac{g\,(t)}{a\,(t)}dt\},\ y < x < x_F,$$

où $c(x) \to c > 0, g(x) \to 1$ et $a'(x) \to 0$ si $x \to x_F$.

un choix possible pour la fonction a est la fonction moyenne des excès définie par :

$$a(x) = \frac{1}{F\,(x)} \int_{x}^{x_F} F(t)dt,\ x < x_F.$$

Dans ce cas les suites a_n et b_n sont ainsi définies :

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

Domaine d'attraction de Weibull

Théorème 1.5 *Une fonction de répartition F appartient au domaine d'attraction de Weibull avec $\gamma < 0$ si et seulement si $x_F < +\infty$ et en plus $1 - F^*$ est une fonction à variation régulière d'indice $-1/\gamma$ c'est-à-dire*

$$F^* = 1 - (x_F - x)^{-1/\gamma} l((x_F - x)^{-1})$$

avec

$$F^*(x) = \begin{cases} 0 & si\ x \leq 0, \\ \\ F\left(x_F - x^{^{-1}}\right) & sinon. \end{cases}$$

où l est une fonction à variation lente à l'infini ($l \in RV_0$).

Dans ce domaine d'attraction les suites de normalisation sont déterminées comme suit :

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

1.6 Distribution GPD

Pickands a introduit la méthode POT (**P**eaks-**o**ver-**T**hreshold) encore appelée méthode des excés au-delà d'un certain seuil réel u suffisamment grand, infèrieur au point terminal ($u < x_F$). Cette méthode consiste utiliser les observations qui dépassent un certain seuil, plus particulièrement les diffirences entre ces observations et le seuil, appelées "excés".

Soient un échantillon de va's i.i.d $X_1, ..., X_n$ et u un seuil fixételque $u < x_F$. On note par N_u le nombre d'exceedances $X_1, ..., X_{N_u}$ qui dépassent le seuil u. On appelle excés au-delà du seuil u les $Y_j = X_i - u$, pour $j = 1, ..., N_u$.

Le principe de cette méthode est répresenté dans la figure suivante

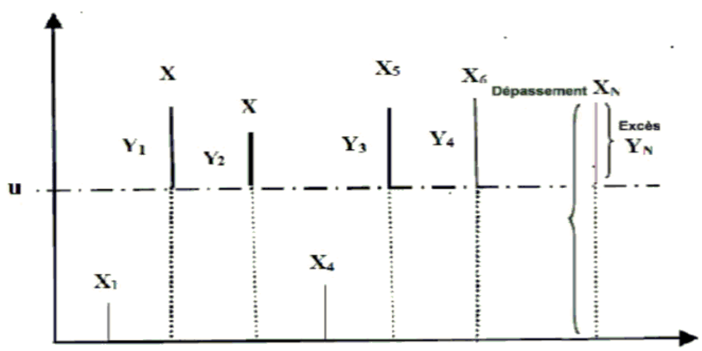

FIG. 1.5 – Représentation des excés Y issus des dépassements Y au-delà d'un seuil u.

Définition 1.8 (La fonction de distribution des excés) *Nous définissons la fonction de distribution des excés au dessus du seuil* u *par :*

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

Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction $\sigma(u)$ strictement positive et un $\gamma \in \mathbb{R}$ tels que :

$$G_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 - \gamma \frac{y}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{si} \quad \gamma \neq 0 \ , \sigma > 0, \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & \text{si} \quad \gamma = 0 \ , \sigma > 0, \end{cases}$$

1.7 Théorème de Balkema-de Haan-Pickands

Le théorème suivant fait le lien entre le comportement asymptotique de la distribution des excés et la loi de Pareto généralisé.

Théorème 1.6 (de Balkema-de Haan-Pickands) *Soit F_u la distribution des excés. Si $F \in DA(H_\gamma)$,la GPD est la distribution limite de la distribution des excés lorsque le seuil tend vers x_F :*

$$\lim_{u \to x_F} \sup_{0 \leq x \leq x_F - u} \left| F_u(x) - G_{\gamma,\beta(u)}(x) \right| = 0$$

où $\beta(u)$ une fonction positive mesurable.

1.8 Estimateur de l'indice des valeurs extrêmes

1.8.1 L'estimateur de Pickands

Cet estimateur a été intoduit par *Pickands (1975)*[17] pour $\gamma \in \mathbb{R}$. Il est le premier estimateur de γ

Définition 1.9 *L'estimateur de Pickands(1975) combine 4 statistique d'ordre. Il est calculé pour un ensemble de rang k. Sa formule est donnée par :*

$$\widehat{\gamma}_{n,k}^{(p)} = \frac{1}{\log 2} \log \left(\frac{X_{(n-k+1,n)} - X_{(n-2k+1,n)}}{X_{(n-2k+1,n)} - X_{(n-4k+1,n)}} \right)$$

où $X_1, ..., X_n$ un échantillon de variables aléatoires i.i.d et n est la taille de l'échantillon observé.

Pickands (1975)[17] a démontré la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par *Dekkers et De Haan*. Des améliorations de cet estimateur ont été introduites notamment par *Drees et Segers*.

Théorème 1.7 *(propriétés de $\widehat{\gamma}_{(n,k)}^{(p)}$)*

Supposons que $F \in DA(H_\gamma)$, $\gamma \in \mathbb{R}$, $k = k(n) \to \infty$ *et* $k/n \to 0$ quand$\to \infty$

a . consistance faible :

$$\widehat{\gamma}_{n,k}^{(p)} \xrightarrow{p} \gamma \quad \text{quand} \quad n \to \infty$$

b . consistance fote : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alos

$$\widehat{\gamma}_{n,k}^{(p)} \xrightarrow{p.s} \gamma \quad \text{quand} \quad n \to \infty$$

c . Normalité asymptotique : Sous des conditions additionnelles sur la suite$k(n)$ et la distribution F

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

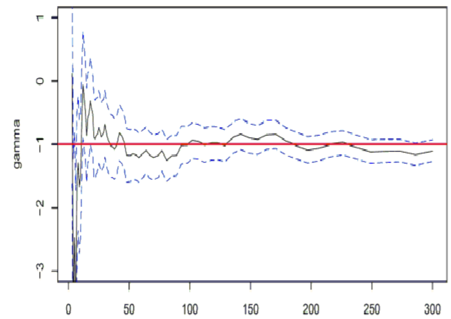

FIG. 1.6 – Estimateur de Pickands avec un intervalle de confiance 95% de γ basés sur 100 échantillons de taille 3000 pour la loi uniforme.

1.8.2 Estimateur de Hill (1975) $\widehat{\gamma}^{H}_{k_n}$

L'estimateur de Hill de l'indice de queue (également connu sous le nom indice des valeurs extrêmes), uniquement défini pour les indices positifs $\gamma > 0$. La construction de l'estimateur de Hill est basée sur la méthode du Maximum de Vraisemblance où on se sert des statistiques d'ordre supérieur à un certain seuil u, pour ne garder que les observations les plus grandes, de façon à ce quelles suivent approximativement une distribution Pareto. Il est défini par la statistique suivante :

$$\begin{aligned}\widehat{\gamma}^{H}_{n,k} &= \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \\ &= \frac{1}{k}\sum_{i=1}^{k} i\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)\end{aligned}$$

Si on choisit k, $n \to 0$ de sorte que $\frac{k}{n} \to 0$ alors on peut montrer que

$$\lim_{k\to\infty} \widehat{\gamma}^{H}_{n,k} = \gamma$$

et l'estimateur de Hill est de plus asymptotiquement normal :

$$\sqrt{k}\frac{\widehat{\gamma}_{n,k}^{H} - \gamma}{\gamma} \xrightarrow{d} N(0,1)$$

Dans le cas général du domaine de Fréchet, la fonction de survie est de la forme $1 - F(x) = x^{-1/\gamma}l(x)$ avec ℓ une fonction à variation lente, Cela induit un biais important sur l'estimateur de Hill, qui est donc en pratique d'un maniement délicat. Dans le cas général, la fonction ℓ apparaît comme un paramètre de nuisance de dimension infinie, qui complique l'estimation *BERTAIL, (2002)*. Pour plus détails sur la consistance de $\widehat{\gamma}^{H}$ *Beirlant et al.(2007)*[1]. Pour cela, nous allons commencer par les conditions du premières et du seconds ordre :

Proposition 1.2 (*Condition du première ordre, de Haan et Ferreira (2006)*) [3]*Les assertions suivantes sont équivalentes :*

a. F est à queue lourde

$$F \in D(Fréchet),\ \gamma > 0$$

b. $1 - F$ est une fonction variation régulière à l'infini d'indice $-1/\gamma$

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

c. $Q(1 - s)$ est une fonction à variations réguliéres à zéro d'indice $-\gamma$

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

d. U est une fonction à variation régulière l'infini d'indice γ

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

Proposition 1.3 (Condition du seconde ordre de Haan et Ferreira (2006)) *Une fonction de répartition $F \in D(Fréchet)$, $\gamma > 0$, admet une condition du seconde ordre à l'infini si elle satisfait l'une des assertions suivantes :*

a. Il existe un paramètre $\rho \leq 0$, et une fonction $A_1(\cdot)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall x > 0$

$$\lim_{t\to\infty} \frac{(1-F(tx))/(1-F(t)) - x^{-1/\gamma}}{A_1(t)} = x^{-1/\gamma}\frac{x^{\rho}-1}{\rho}.$$

b. S'il existe un paramètre $\rho \leq 0$ et une fonction $A_2(\cdot)$ qui tend vers 0 (ne change pas de signe à zéro) définie par, $\forall x > 0$

$$\lim_{t\to\infty} \frac{Q(1-sx)/Q(1-s) - -x^{-\gamma}}{A_1(t)} = x^{-\gamma}\frac{x^{\rho}-1}{\rho},$$

c. S'il existe un paramètre $\rho \leq 0$,et une fonction $A(\cdot)$qui tend vers 0(ne change pas de signe à l'infini) définie par,$\forall x > 0$

$$\lim_{t\to 0} = \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{\gamma}\frac{x^{\rho}-1}{\rho},$$

si $\rho = 0$, on remplace $(x^{\rho}-1)/\rho$ par $\log x$.

Les fonctions $A(\cdot)$, $A_1(\cdot)$, $A_2(\cdot)$ sont à variations régulières l'infini d'indices respectifs ρ, ρ/γ, et $-\rho$, avec $A_1(t) = A(1/(1-F(t)))$ *et* $A_2(s) = A(1/s)$.

Ces deux conditions ont permis de déterminer les propriétés asymptotiques de certains estimateurs de l'indice des valeurs extêrmes.

Théorème 1.8 (Propriétés asymptotiques de l'estimateur de Hill) *Soit k_n, $n \geq 1$ une suite d'entiers telle que $1 \leq k_n \leq n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.*

a. Consistance faible **:** $\widehat{\gamma}_{k_n}^H$converge en probabilité vers γ

b. Consistance forte : Si de plus $k_n/\log n \log n \to \infty$ quant $n \to \infty$, alors $\widehat{\gamma}_{k_n}^H$ converge presque sûrement vers γ.

c. Normalité asymptotique **:** Si la condition

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

est satisfaite avec $\sqrt{k_n} A(n/k_n) \to \lambda \in \mathbb{R}$ quand $n \to \infty$, alors

$$\sqrt{k_n} \left(\widehat{\gamma}_{k_n}^H - \gamma \right) \xrightarrow{\mathcal{L}} N\left(\lambda/(1-\rho),\ \gamma^2\right)$$

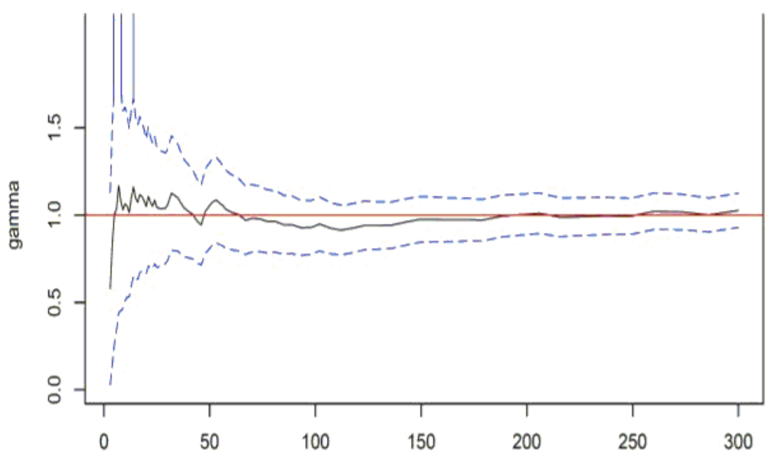

FIG. 1.7 – L'estimateur de Hill avec un intervalle de conance 95% de γ basés sur 100 échantillons de taille 3000 pour la loi de Pareto standard.

1.8.3 L'estimateur des moments de Dekkers-Einmahl De Haan

En 1989, *Dekkers-Einmahl-DeHaan*[4] ont proposé un estimateur de $\gamma \in \mathbb{R}$ qui présente une extension de l'estimateur de Hill, appelé un estimateur de moment pour $\gamma \in \mathbb{R}$.

Définition 1.10 *L'estimateur des moments est défni comme suit :*

$$\widehat{\gamma}_{n,k}^{(M)} = 1 + M_n^1 - \frac{1}{2}\left(\frac{\left(M^{(1)}\right)}{\left(M_n^{(2)}\right)} - 1\right)^{-1}$$

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

Sous certaines conditions sur k, l'estimateur de *Dekkers-Einmahl-DeHaan* converge asymptotiquement vers la loi normale.

Dekkers-Einmahl-DeHaan ont démontré la consistance faible, la consistance forte et la normalité asymptotique de leur estimateur.

Théorème 1.9 (Propriétés de $\widehat{\gamma}_{n,k}^{(M)}$)

Supposons que $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$, $k = k(n) \to \infty$ et $k/n \to 0$ quand $n \to \infty$

a . Consistance faible :

$$\widehat{\gamma}_{n,k}^{(M)} \xrightarrow{p} \gamma \text{ quand } n \to \infty$$

b . Consistance forte : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors

$$\widehat{\gamma}_{n,k}^{(M)} \xrightarrow{p.s} \gamma \text{ qauand } n \to \infty$$

c . Normalité asymptotique : Sous des conditions de régularité convenables

$$\sqrt{k}\left(\widehat{\gamma}_{n,k}^{(M)}-\gamma\right)\xrightarrow{\mathcal{L}}N\left(0,\eta^{2}\right)\text{ quand }n\rightarrow\infty,$$

ou

$$\eta^{2}=\begin{cases}1+\gamma^{2} & \text{si }\gamma\geq0\\ (1-\gamma)^{2}(1-2\gamma)\left[4-8\frac{1-2\gamma}{1-3\gamma}+\frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right] & \text{si }\gamma<0\end{cases}$$

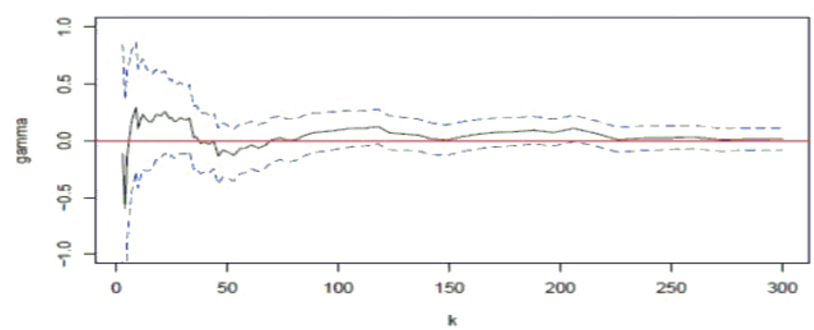

Fig. 1.8 – Estimateur des moments avec un intervalle de confance 95% de γ basésur 100 échantillons de taille 3000 pour la loi de Gumbel standard.

Chapitre 2

Introduction à l'analyse de survie

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées ($i.i.d$) de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

2.1 Définitions

Définition 2.1 (Fonction de survie S) *Soit X une v.a. positive et continue dite "durée de vie". Pour t fixé, la fonction de survie est la probabilité de survivre jusqu'à l'instant t, c'est-à-dire*

$$\begin{aligned} S(t) &= \overline{F}(t) \\ &= 1 - F(t) \\ &= 1 - P(X \leq t) \\ &= P(X > t) \end{aligned}$$

Définition 2.2 (Risque instantané λ) *Le risque instantané (ou taux d'incidence), pour t fixé caractérise la probabilité de mourir dans un petit inter-*

valle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu)

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \le X < t + h \mid X \ge t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))'$$

Définition 2.3 (Taux de hasard cumulé Λ) *Le taux de hasard cumulé est l'intégrale du risque instantané λ*

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\ln(S(t))$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé :

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)\, du\right)$$

On en déduit que

$$f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(u)\, du\right)$$

2.2 Données de censure

Définition 2.4 (variable de censure) *La variable de censure Y est définie par la non-observation de l'événement étudié. Si au lieu d'obsever X , on observe Y, on a :*

1. $X > Y$ est censure à droite.
2. $X < Y$ est censure à gauche.
3. $Y_1 < X < Y_2$ est censure par intervalle.

2.2.1 Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique.

Pour un individu donné j, on va considérer :

- Temps de survie X_j.
- Son temps de censure Y_j.
- La durée réellement observée Z_j.

2.2.2 Types de censures

Dans la littérature on distingue trois types de censure :

Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

Censure à gauche

Une durée de survie est dite censurée à gauche si l'individu a déjà subi l'évènement d'intérèt avant lentrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple (Z, δ) :

$$Z = X \vee Y = \max(X, Y) \text{ et } \delta = 1_{\{X \geq Y\}}$$

Notons X l'âge à laquelle une certaine maladie apparait pour la première fois chez un individu. Aprés un examen médical on a reçu deux types de réponses :

1. l'individu a déjà été malade mais l'age exact de la première apparition n'a pas été retenu :

Dans ce cas on n'a pas observé X mais on sait que X est inférieur à l'âge de l'individu lors de l'examen Y. Il s'agit d'une observation censurée à gauche.

2. l'individu n'a jamais eu de maladie : Dans ce cas on sait seulement que X est supérieur l'âge de l'individu, donc on a une observation censurée à droite.

Remarque 2.1 *Si les variables de censures sont dégénérée (c'est-à-dire constantes), alors on dit que la censure est fixée.*

Censure par intervalle

Dans ce cas, comme son nom l'indique, on observe la fois une borne inférieure et une borne supérieure de la variable dintérêt. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite aprés que l'évènement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[y, +\infty[$ *et* $[0, y]$ respectivement.

Dans la littérature, il existe d'autre types de censure :

- **censure de type 1 : fixée**

Soit Y une valeur fixée, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on n'observe X_j uniquement lorsque $X_j \leq Y$ si non on sait uniquement que $X_j > Y$. on utilise la notation suivante :

$$Z_j = X_j \wedge Y = \min(X_j, Y)$$

Exemple 2.1

Dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note X la durée d'apprentissage de cette langue. Pour certains étudiants nous allons observer leurs durées X_j d'apprentissage de la langue par contre pour d'autres leurs X_j ne seront pas observées car le stage est limité dans le temps.

- **Censure de type 2 : attente**

Elle présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là.

Soient $X_{j,n}$ et Z_j les statistiques d'ordre des variables X_j et Z_j. La date de censure est donc $X_{k:n}$ et on observe les variables suivantes :

$$
\begin{aligned}
Z_{1,n} &= X_{1,n} \\
&\vdots \\
Z_{k,n} &= X_{k,n} \\
Z_{k+1,n} &= X_{k,n} \\
&\vdots \\
Z_{n,n} &= X_{k,n}
\end{aligned}
$$

- **Censure de type 3 : aléatoire**

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expérience, la date d'inclusion du patient dans l'étude est fixé, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, la durée d'hospitalisation du patient). Ici, le nombre d'évènement observés et la durée totale de l'expérience sont aléatoires.

Soient $Y_1, ..., Y_n$ des variables aléatoires $i.i.d.$ On observe les variables

$$Z_j = X_j \wedge Y.$$

L'information disponible peut être résumée par :

1- la durée réellement observée Z_j,

2- un indicateur $\delta_j = 1_{\{X_j \leq Y_j\}}$

(a) $\delta_j = 1$ si l'événement est observé (d'où $Z_j = X_j$). On observe les "vraies" durées ou les durées complètes.

(b) $\delta_j = 0$ si l'individu est censuré (d'où $Z_j = Y_j$). On observe des durées incomplètes (censurées).

Exemple 2.2 *Lors d'un essai thérapeutique, on peut citer certaines causes entranant la censure aléatoire :*

1. *Perdu de vue : le patient peut décider de se faire soigner ailleurs cause d'un déménagement et on le revoit plus.*
2. *Arrêt du traitement : suite à des effets secondaire le traitement est arrêté.*
3. *Fin de l'étude : l'étude se termine et certains patients soit toujours vivants (exclus-vivants).*

2.3 Estimateur de Kaplan-Meier $\widehat{S}(t)$

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t, c'est-à-dire, si $t^{"} < t^{'} < t$

$$\begin{aligned} P\left(X > t\right) &= P\left(X > t^{'}, X > t\right) \\ &= P\left(X > t \mid X > t\right) \\ &= P(X > t \mid X > t^{'}) \times P(X > t^{'} \mid X > t^{''}) \times P(X > t^{''}) \end{aligned}$$

En considérant les temps d'événements (décès et censure) distincts $Z_{(j)}(i = 1, ..., n)$ rangés par ordre croissant, $Z_{(0)} = 0$ on obtient

$$P(X > Z_{(j)}) = \prod_{k=1}^{j} P(X > Z_{(k)} \mid X > Z_{(k-1)})$$

Considérons les notations suivantes :

Y_j le nombre d'individus à risque de subir l'événement juste avant le temps $Z_{(j)}$, d_j le nombre de décès en $Z_{(j)}$. Alors la probabilité p_j de mourir dans l'intervalle$\left] Z_{(j-1)}, Z_{(j)} \right]$ sachant que l'on était vivant en $Z_{(j-1)}$, *i.e.*

$$p_j = P(X \leq Z_{(j)} \mid X > Z_{(j-1)}),$$

peut être estimée par $p_j = \frac{d_j}{Y_j}$. Comme les temps d'événements sont supposés distincts, on a

- $d_j = 0$ en cas de censure en $Z_{(j)}$, *i.e.* quand $\delta_j = 0$,

- $d_j = 1$ en cas de décès en $Z_{(j)}$, $i.e.$quand $\delta_j = 1$.

On obtient alors l'estimateur de Kaplan-Meier :

$$\begin{aligned}
\widehat{S}_n(t) &= 1 - \widehat{F}_n(t) \\
&= \prod_{j=1:n,\ Z_{(j)} \leq t} \left(1 - \frac{\delta_j}{Y_j}\right) \\
&= \prod_{j:Z_j \leq t}^{n} \left(1 - \frac{\delta_j}{n-(j-1)}\right) \\
&= \prod_{j=1}^{n} \left[\frac{n-1}{n-j+1}\right]^{\delta_j}
\end{aligned}$$

L'estimateur $\widehat{S}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance. $\widehat{S}(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de $Kaplan - Meier$ [12] dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car les temps de décès ne sont pas connus).

2.4 Estimation de l'indice des valeurs extrême avec censure

Nous travaillons dans l'espace de probabilité (Ω, A, P) et soit l'échantillon $(X_1, ..., X_n)$ variable aléatoire définit sur (Ω, A, P) sa fonction répartition F et sa queue de distribution :

$$1 - F \in RV(-1/\gamma_1)$$

Soit le deuxième échantillon $(Y_1, ..., Y_n)$, des variables aléatoires $i.d.d$, de fonction de répartition G et de queue de distribution :

$$1 - G \in RV(-1/\gamma_2)$$

Alors les variables Z_j définies par :

$$Z_j = X_j \wedge Y_j,\ j = 1, ..., n$$

avec, Z_j sont des variables indépendantes de loi H liées F et G par la relation :

$$1 - H(x) = (1 - F(x))(1 - G(x)).$$

Le points terminal de H est $x_H = sup\{x,\, H(x) < 1\}$. On a F et G satisfaisant la condition de domaine d'atraction de Féchet, $F \in D\left(\Phi_{1/\gamma_1}\right)$ et $G \in D\left(\Phi_{1/\gamma_2}\right)$, telles que :

$$1 - F\left(x\right) = x^{-1/\gamma_1} l_1(x) \ et \ 1 - G(x) = x^{-1/\gamma_2} l_2(x),$$

avec $l_1\left(.\right)$ et $l_2\left(.\right)$ sont des fonctions à variations lentes. Alors :

$$\begin{aligned} 1 - H\left(x\right) &= \left(1 - F\left(x\right)\right)\left(1 - G\left(x\right)\right) \\ &= x^{-1/\gamma_1} l_1(x) x^{-1/\gamma_2} l_2(x) \\ &= x^{-(\gamma_1+\gamma_2)/\gamma_1\gamma_2} \widetilde{l}\left(x\right) \\ &= x^{-1/\gamma} \widetilde{l}\left(x\right), \end{aligned}$$

où $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$, avec $\widetilde{l}\left(x\right) = l_1(x) l_2(x)$. Par conséquent, H appartenant au domaine d'attraction de Fréchet :

$$1 - H\left(x\right) \in RV\left(-1/\gamma\right)$$

Si F et G appartiennent au domaine d'attraction du maximum $F \in D(G_{\gamma_1})$ et $G \in (G\gamma_2)$ respectivement, pour certain $\gamma_1,\ \gamma_2 \in \mathbb{R}$ avec points terminales x_F et x_G, où $x_F = \sup\left\{x, F\left(x\right) < 1\right\}$, alors cela signifie que $H \in D\left(G_\gamma\right)$. *Einmahl et al.(2008)*[6], ont proposés les trois cas les plus intéressant suivants :

$$\begin{cases} cas\ 1:\ \gamma_1 > 0,\ \gamma_2 > 0,\ x_F = x_G = +\infty, & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}, \\ cas\ 2:\ \gamma_1 < 0,\ \gamma_2 < 0,\ x_F = x_G < \infty\ , & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \\ cas\ 3:\ \gamma_1 = 0,\ \gamma_2 = 0,\ x_F = x_G = \infty, & \gamma = 0. \end{cases}$$

La méthode générale existante pour l'estimation de l'indice de queue en présence de censure à droite aléatoire, apparue d'abord dans *Beirlant et al.(2007)*[1] et développée dans *Einmahl et al.(2008)*[6], est considérer tout estimateur consistant $\widehat{\gamma}$ de **l'EVI** γ appliqué l'échantillon $(Z_1, ..., Z_n)$ et deviser par la proportion $\widehat{p}$ d'observation non censurées dans les plus grandes k valeurs de Z :

$$\widehat{\gamma}^{(c,.)}_{Z,n,k} = \frac{\widehat{\gamma}^{(.)}_{Z,n,k}}{\widehat{p}},$$

où

$$\widehat{p} = \frac{1}{k}\sum_{j=1}^{k}\delta_{[n-j+1,n]},$$

avec $\delta_{[1,n]}, ..., \delta_{[n,n]}$ les indicateurs de censure retenues correspendant à la statistique d'ordre $(Z_{1,n}, ..., Z_{n,n})$, respectivement. Il sera suivre que $\widehat{p}$ estime $p = \frac{\gamma_2}{\gamma_1+\gamma_2}$.

$\widehat{\gamma}^{(c,.)}_{Z,n,k}$ peut être n'importe quel estimateur non adapté la censure. En particulier, une adaptation de l'estimateur de Hill $\widehat{\gamma}^{(H)}_{Z,n,k}$ de l'indice γ_1 dans le cas de censure est défini par :

$$\widehat{\gamma}^{(c,H)}_{Z,n,k} = \frac{\widehat{\gamma}^{H}_{n}}{\widehat{p}}$$

où

$$\widehat{\gamma}^{H}_{n} := \frac{1}{k}\sum_{j=1}^{k}\log\left(Z_{n-i+1,n}\right) - \log\left(Z_{n-k,n}\right),\ 1 \le k \le n$$

Alors

$$\widehat{\gamma}_1^{(c,H)} = \frac{\sum_{j=1}^{k} \log\left(Z_{n-i+1,n}\right) - \log\left(Z_{n-k,n}\right)}{\sum_{j=1}^{k} \delta_{[n-j+1,n]}}$$

Chapitre 3

ESTIMATION DE LA MOYENNE AVEC CENSURE

Soit $X_1, ..., X_n$, $n \geq 1$ sont des des variable aléatoire positives, indépendantes et de fonction de répartition F, et indépendent des variables aléatoire $Y_1, ..., Y_n$, les instants de censure associés, positives, de fonction de répartition G. On note $\{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$ l'échantillon réellement observé, où, pour $1 \leq j \leq n$

$$Z_j = \min(X_j, Y_j) \ \ et \ \delta_j = 1\{X_j, Y_j\}$$

avec $1\{.\}$ indiquant la fonction de l'indicateur. Ce dernier indique s'il y a censure ou non. Si l'on note H le *cdf* des Z observés , par l'indépendance de X et Y, nous avons $1 - H = (1 - F)(1 - G)$. Dans ce chapitre, nous utiliserons la notation $\overline{S}(x) = S(\infty) - S(x)$, pour toute fonction S supposons en outre que F et G sont à queue lourde ou, en d'autres termes, que F et G varient régulièrement à l'infinité avec des indices négatifs $-1/\gamma_1$ et $-1/\gamma_2$ respectivement.

$$\lim_{x\to\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma_1} \text{ et } \lim_{x\to\infty} \frac{\overline{G}(xz)}{\overline{G}(z)} = x^{-1/\gamma_2}$$

pour tout $x > 0$. Par conséquent, H est aussi à queue lourde, avec un indice de queue $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

Il existe des constantes $\rho_j < 0$ et des fonctions A_j, $j = 1, 2$ tendant vers zéro, ne changeant pas de signe près de l'infinité tel que pour tout $x > 0$

$$\lim_{t\to\infty}\frac{\overline{F}(tx)/\overline{F}(t)-x^{-1/\gamma_1}}{A_1(t)}=x^{-1/\gamma_1}\frac{x^{\rho_1/\gamma_1-1}}{\gamma_1\rho_1}$$

et

$$\lim_{t\to\infty}\frac{\overline{G}(tx)/\overline{G}(t)-x^{-1/\gamma_2}}{A_2(t)}=x^{-1/\gamma_2}\frac{x^{\rho_2/\gamma_2-1}}{\gamma_2\rho_2}$$

L'estimateur du maximum de vraisemblance de *cdf* F est donné par *Kaplan et Meier* [12] :

$$\widehat{F}_n(x)=\begin{cases}1-\prod\limits_{Z_{j:n}\leq x}^{n}\left(\frac{n-j}{n-(j-1)}\right)^{\delta_{[j:n]}} & \text{pour } x<Z_{n:n},\\ \\ 1 & \text{pour} \quad x\geq Z_{n:n},\end{cases}$$

$\widehat{F}_n$ est donné aussi par :

$$\widehat{F}_n(t)=\sum_{i=1}^{n}W_{i,n}1\{Z_{i:n}\leq x\}\ \textit{où}\ W_{i,n}=\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j:n]}}$$

L'estimateur asymptotiquement normal pour la moyenne est $\mu=E[X]=\int\limits_0^{\infty}\overline{F}(x)\,dx$, en substituant $\widehat{F}_n$ à F dans l'équation précédente, *Stute* [22] a réduit la moyenne empirique des données censurées de

$$\widehat{\mu}_n:=\sum_{i=1}^{n}\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j:n]}}$$

μ est un somme de deux termes : $\mu=\int\limits_0^{h}\overline{F}(x)\,dx+\int\limits_h^{\infty}\overline{F}(x)\,dx=\mu_1+\mu_2$,

$$\mu_1=h\overline{F}(h)+\int_0^h xF(x)\,dx \text{ et } \mu_2=h\overline{F}(h)\int_1^{\infty}\frac{\overline{F}(hx)}{\overline{F}(h)}.$$

L'estimation de μ_1 et μ_2 sont (remplaçant h et $F(x)$ par Z_{n-kn}et $\widehat{F}_n(x)$)[20] :

$$\widehat{\mu}_1=\sum_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}Z_{n-k}+\sum^{n-k}\frac{\delta_{[i:n]}}{n-j+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}Z_{i:n}, \tag{3.1}$$

$$\mu_2 \tilde{} \frac{\gamma_1}{1-\gamma_1} h\overline{F}(h), \text{ comme } n \to \infty, \; 0 < \gamma_1 < 1. \tag{3.2}$$

Les quantités h et $\overline{F}(h)$ sont, comme ci-dessus, naturellement estimées par $Z_{n-k:n}$ et

$$\widehat{\overline{F}}_n(Z_{n-k}) = \prod_{j=1}^{i-k} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}}$$

pour dériver un estimateur de μ_2 , il faut estimer l'indice γ_1. La méthode générale existante, apparue pour la première fois dans *Beirlant et al*, puis développé dans *Einmahl et al*, cette méthode est basé sur l'évaluation Z et divisez-le par le proportion observées. Par exemple, *Einmahl et al*, adapté l'estimateur de Hill par un estimateur $\widehat{\gamma}_1^{(H,c)} = \widehat{\gamma}^H / \widehat{p}$

$$\widehat{\gamma}^H = \frac{1}{k} \sum_{i=1}^{k} \log \frac{Z_{n-i+1:n}}{Z_{n-k}} \quad \text{et} \quad \widehat{p} = \frac{1}{k} \sum_{i=1}^{k} \delta_{[n-i+1:n]},$$

En remplaçant, dans (3.2), F et γ_1 par $\widehat{F}_n$ et $\widehat{\gamma}_1^{(H,c)}$

$$\widehat{\mu}_2 := \frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}} Z_{n-k} \prod_{j=1}^{n-k} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}}, \widehat{\gamma}_1^{(H,c)} < 1 \tag{3.3}$$

Enfin, avec (3.1) et (3,3), l'estimateur $\widehat{\mu}$ de la moyenne μ est :

$$\widehat{\mu} = \frac{1}{k} \sum_{j=2}^{n-k} \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}} Z_{i:n} + \prod_{j=1}^{n-k} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}} \frac{Z_{n-k}}{1-\widehat{\gamma}_1^{(H,c)}}$$

3.1 Simulation

On a réalisé une étude de simulation basée sur 3000 échantillons de loi de Pareto de paramètre γ_1 censurées par une autre variable de Pareto de paramètre γ_2.

$$F^{-1}(u) = (1-u)^{-\gamma_1}, \qquad G^{-1}(v) = (1-v)^{-\gamma_2}.$$

$\{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$ l'échantillon réellement observé, où, pour $1 \leq j \leq n$

$$Z_j = \min(X_j, Y_j)$$

Le programme ci-dessous calcule l'estimateur $\widehat{\mu}$ de la moyenne μ sous R :

```
gamma1<-1
gamma2<-0.5
u<-runif(n,0,1)
X<-(1-u)^gamma1
V<-runif(n,0,1)
Y<-(1-V)^gamma2
Z<-pmin(X,Y)
delta=as.numeric(X<=Y)
k<-100
for(i in 1 :k)
{
i<-seq(1,k)
gammah<-(1/k)*sum(log(Z[n-i+1]/Z[n-k]))
}
i<-seq(1,k)
p<-(1/k)*sum(delta[n-i+1])
gammach<-gammah/p
i<-seq(1,k)
for(i in 2 :n-k)
for(j in 1 :i-1)
{
S<-prod((n-j)/(n-j+1))^delta[j]
}
U1<-(1/k)*sum(delta[i]/n-i+1)*S*Z[i]
for(j in 1 :n-k)
{
U2<-(prod((n-j)/(n-j+1))^delta[j])*(Z[n-k]/(1-gammach))
}
U<-U1+U2
```

Conclusion

Dans notre étude , nous avons abordé différents aspects de la théorie des valeurs etrêmes. Aprés avoir le premier chapitre la théorie des valeurs extrêmes en mentionnent les différentes caractéristiques et les notions de base qui sont trés utile pour l'estimation des quantiles extrêmes et les données censurées et aussi nous nous sommes intéréssés à la famille de lois à que de type Fréchet, on peut citer la loi de Pareto. On a présenté trois estimateurs (estimateur de Pickands, estimateur Hill (1975) et estimateur des moments), nous nous sommes intéréssés dans le deuxième chapitre au certaines des notions clés en analyse de survie et quelques estimations (estimation de la fonction de survie et taux hasard ...ect) et les estimateurs de l'IVE avec censure. Dans le dernier chapitre, nous avons fait un rappel général sur l'estimation de la moyenne et de traiter une méthode pour estimer la moyenne d'un distribution à queue lourde en présence de censure.

Bibliographie

[1] Beirlant, J., Guillou, A., Dierckx, G. and Fils-Viletard, A., 2007. Estimation of the extreme value index and extreme
quantiles unr random censoring. Extremes 10 :3, 151-17

[2] Brahimi, B., Meraghni, D. et Necir, A. (2015). Approximation gaussienne de l'estimateur de l'indice de valeur extrême d'une distribution à queue lourde sous censure aléatoire. Math. Méthodes Statist., 24 (4), 266279.

[3] De Haan, L. et Ferreira, A. (2006). Théorie des valeurs extrêmes : une introduction. SpringerVerlag, New York.

[4] Dekkers, A., Einmahl, J. H. J. et de Haan, L., (1989). Estimateur de moment pour l'indice d'une distribution de valeurs extrêmes. Annals of Statistics, 17 (4) : 1833 -1855.

[5] Drees, H., (1996). Estimateur Rened Pickands de l'indice des valeurs extrêmes. Annals of Statistics, 23 (6) : 2059-2080.

[6] Einmahl, J. H ; Fils-Villetard, A et Guillon, A. (2008) Statistiques des extrêmes sous censure aléatoire, Bernoulli, 14(1) ; 207-227.

[7] Embrechts, P., Klppelberg, C. et Mikosch, T. (1997). Modélisation d'événements extrêmes pour l'assurance et la finance, Springer-Verlag, Berlin.

[8] Fisher, R., Tippet, L. (1928). La limitation forme la distribution de fréquence du membre le plus grand ou le plus petit d'un échantillon. Actes de la Cambridge Philosophical Society, 24 : 180-190.

[9] Gnedenko, B. (1943). Sur la distribution limite du terme maximum dune serie aléatoire. Ann. Math., 423- 453.

[10] Hill, B.M. (1975) .Une approche générale simple à consulter sur le détail de la distribution, Ann. Statist., 3, 1163-1174.

[11] Jenkinson, A. F. (1955). La distribution de fréquence des valeurs annuelles maximales (ou minimales) des éléments météorologiques. Quarterly J.R. Methodol.Soc., 81 (348), 158-171

[12] Kaplan, E. L et Meier, P. (1958). Estimation nano paramétrique à partir d'observations incomplètes. J. Amer. Statist Assoc ; 53(282) ; 457-481.

[13] Klein, J. P ; Moeschberger, M. L. (1997) Analyse de survie : techniques pour les données censurées et tronquées, Springer. Verlag Inc, Berlin, New York.

[14] Lee, E.T ; et Wang, J. (2003) Méthodes statistiques pour l'analyse de données de survie. John. Wiley.

[15] Mason, D. M. (1982). Lois des grands nombres pour des sommes de valeurs extrêmes. Ann. Probab., 754-764.

[16] Ndao, P., Diop, A. et Dupuy, J. F. (2014). Estimation non paramétrique de l'indice de queue conditionnel et des quantiles extrêmes sous censure aléatoire. Comput. Statiste. Data Anal., 79, 6379.

[17] Pickands III, J. (1975). Inférence statistique à l'aide de statistiques d'ordre extrême. Statist., 119- 131.

[18] Reiss,R.-D.,Thomas,M.S.,2007, Analyse statistique des valeurs extrêmes. De l'assurance, de la finance, de l'hydrologie et d'autres domaines. Birkhäuser Verlag, Basel., Boston, Berlin.

[19] Segars. J., (2001). Estimateurs de résidus. Journal de StSmith, R. L., (1985). Estimation du maximum de vraisemblance dans une classe de cas non réguliers. Biometrika, 72 (1) : 6792

[20] Soltane, Louiza, (2016). Analyse des valeurs extrêmes en présence de censure. Thèse de doctorat de université Mohamed khider, Biskra, Algérie.

[21] Stuper, G. (2016). Estimation de l'indice conditionnel de valeur extrême sous censure aléatoire à droite. J. Multivariate Anal., 144, 1-24.

[22] Stute, W. (1995). Le théorème de la limite centrale sous censure aléatoire. Ann. Statist., 422-439

Résumé

Ce mémoire étude l'estimation de la moyenne d'une distribution à queue lourde en présence de censure, d'un part est constitué de la théorie des valeurs extrêmes et d'autre part de l'analyse de survie, l'intérêt principal est de proposer une méthodologie d'estimation de la moyenne avec censure aléatoire à droite . Mots clés : Analyse de survie, Censure aléatoire, Estimateur de Hill, Estimateur de Kaplan-Meier, Estimation de la moyenne.

Abstract

This is a study of the estimation of the mean of a heavy-tailed distribution in the presence of censorship, on the one hand is made up of the theory of extreme values and on the other hand of the survival analysis. The main thing is to propose a method of estimating the mean with random right-censoring. Key words: Survival analysis, Random censuring, Hill estimator; Kaplan-Meier estimator, Estimating the mean.

ملخص

هذه الأطروحة عبارة عن دراسة لتقدير متوسط توزيع الذيل الثقيل في وجود الرقابة ، من ناحية تتكون من نظرية القيم المتطرفة ومن ناحية أخرى تحليل البقاء ، الهدف الرئيسي هو اقتراح منهجية لتقدير المتوسط مع وجود الرقابة

الكلمات المفتاحية : تحليل البقاء ، الرقابة العشوائية ، مقدر هيل ، مقدر كابلان ماير ، تقدير المتوسط