République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

HOUHOU ROUMAISSA

Titre :

Statistiques des Extrêmes sous Données Censurés

Membres du Comité d'Examen :

Pr.	**MERAGHNI Djamel**	UMKB	Président
Dr.	**BENAMEUR Sana**	UMKB	Encadreur
Dr.	**SOLTANE Louiza**	UMKB	Examinateur

27 Juin 2019

Dédicace

Je dédie ce travail à

Mes chers parents.

Mes frères, ma sœur.

À tous mes collègues, mes amies et toutes les personnes qui ont contribués à la réalisation de ce travail.

Mon encadreur : Dr. Benameur Sana.

A toute ma famille.

Houhou Roumaissa

REMERCIEMENTS

Tout d'abord je glorifie "ALLAH" le tout puissant, de m'avoir donné la santé, la force, la patience et la volonté pour réaliser ce travail, dans des meilleures conditions.

Mes premiers remerciements s'adressent à mon encadreur de ce mémoire : Dr. **Benameur Sana** qui a bien voulu me proposer ce thème et m'aider au cours de sa réalisation.

Je remercie les membres de jury : Le Pr. Meraghni Djamel et le Dr. Louiza Soltane d'avoir accepté d'examiner et d'évaluer ce travail.

Je remercie également, tous nos **enseignants** du département de **Mathématiques** à l'université de **Mohamed khider,** qui ont contribués à nos formations pendant les années de Licence et de Master.

Je remercie ma famille à qui je n'ai jamais su dire toute l'affection que j'ai pour eux, mon père, ma mère, mes frères et ma *sœur qui ont été et seront toujours présents à mes côtés, merci pour votre soutient et vos encouragements.*

J'adresse mes remerciement à tous mes amies.

Merci

Table des matières

Table des figures

Liste des tableaux

Introduction

La théorie des valeurs extrêmes (EVT : Extremes Value Theory) est une branche de la statistique qui essaie d'amener une solution face à des événements rares ; c'est-à-dire des événements dont la probabilité d'apparition est très faible. Les valeurs extrêmes sont aussi rares soient-elle, puisque il s'agit des valeurs beaucoup plus grandes ou plus petites que celles observés habituellement. Cette théorie se repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Tout a commencé avec les auteurs *Ficher* et *Tippet* (1928) [11] puis plus tard avec *Gnedenko* (1943) [12].

Dans l'analyse de survie, les données sont caractérisées par l'existence d'observation incomplète. Ces données sont souvent recueillies partiellement à cause d'un processus de censure, qui empêche l'observation exacte du délai de survenue de l'évènement d'intérêt. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information.

Au cours des dernières années, la modélisation des valeurs extrêmes pour des données censurées a bénéficié d'une certaine attention. Elle voit le jour en premier fois en (1997) avec la sortie du livre Reiss et Thomas [21]. En (2007), *Beirlant et al* [2], introduit un estimateur de l'indice des valeurs extrêmes à la présence censure, l'étude a continué jusqu'en (2008), par Einmahl et al [9].

Ce sujet est appliqué à une grande variété de domaines : La médecine (malades cancéreux), biologie (la mort d'organismes biologiques), climatologie (température), hydrologie (hauteur des barrages) et physique (avec l'apparition de la théorie de la fiabilité), etc.

L'objectif principal de ce travail est l'estimation des statistiques des extrêmes : la fonction de survie, l'indice des valeurs extrêmes et les quantiles extrêmes, sous données censurées.

Ce mémoire est composé de deux chapitres.

Dans le premier, on présente tout d'abord les principaux définitions et concepts de base de l'EVT. Après avoir introduit la distribution du maximum d'un échantillon, on présente les deux principaux outils servant à modéliser la distribution des valeurs extrêmes : la distribution des valeurs extrêmes généralisées (GEV) et la distribution de Pareto généralisé (GPD). Ensuite, on rappelle la caractérisation des domaines d'attraction. La dernière section est dédiée aux fondamentales notions de l'analyse de la survie.

Dans le deuxième, on regroupe deux parties. Dans la première partie : on présente les différentes statistiques des extrêmes et leurs estimateurs dans le cas des données complètes et censurées. On commence par l'estimateur de l'indice des valeurs extrêmes : estimateur paramétrique (estimateur de maximum de vraisemblance), estimateur semi-paramétrique (estimateur de Hill et celui de Pickands) dans la cas des données complètes et l'estimateur de Hill adapté sous des données censurées. Ensuite, on présente l'estimation de la fonction de survie sous données complètes : distribution à variation finie et distribution à queue lourde et l'estimateur de Kaplan-Meier qui est le plus couramment utilisé sous des données censurés. Dans la troisième section, on présente l'estimation des quantiles extrêmes. La deuxième partie est consacrée à l'application des principaux résultats obtenus dans la première partie de ce chapitre, sur des données simulées, sous le logiciel R.

Chapitre 1

Préliminaires sur les valeurs extrêmes et censurées

Dans ce chapitre, nous rappelons des préliminaires sur la théorie des valeurs extrêmes et les différents types des données censurés.

1.1 Définitions et caractéristiques de bases

Définition 1.1.1 (Fonction de répartition)

La fonction de répartition F d'une variable aléatoire (v.a) X est définie par l'application suivante :

$$\begin{array}{rccl} F: & \mathbb{R} & \rightarrow & [0,1], \\ & x & \mapsto & F(x) := P(X \leq x). \end{array} \tag{1.1}$$

Définition 1.1.2 (Fonction de survie)

Pour t fixé, la fonction de la survie, d'une v.a positive et continue X dite "durée de survie", est définie par la probabilité de survivre jusqu'à l'instant t, c'est-à-dire :

$$S(t) := \overline{F}(t) = 1 - P(X \leq t) = P(X > t). \tag{1.2}$$

Définition 1.1.3 (Fonction de quantile)

Pour tout $0 < s < 1$*, la fonction de quantile est définie par :*

$$Q(s) =: F^{-1}(s) := \inf\{t : F(t) \geq s\}, \tag{1.3}$$

où F^{-1} *représente l'inverse généralisée de la fonction de répartition* F*, la convention que* $\inf\{\varnothing\} = +\infty$ *et* $P(X \leq Q(s)) = s$*. On l'exprime en terme de la fonction de survie par :*

$$Q(s) := \inf\left\{t : \overline{F}(t) \leq 1 - s\right\} \ , 0 < s < 1. \tag{1.4}$$

Définition 1.1.4 (Fonction quantile de queue)

La fonction quantile de queue est définie par :

$$U(x) := Q\left(1 - \frac{1}{x}\right) = F^{-1}\left(1 - \frac{1}{x}\right) = \left(1/\overline{F}\right)^{-1}(x) \ \ \textit{avec } 1 < x < \infty. \tag{1.5}$$

Définition 1.1.5 (Point terminal)

Le point terminal de la fonction de répartition F *est donné par :*

$$x_F := \sup\{x \in \mathbb{R} : F(x) < 1\} \leq \infty.$$

Définition 1.1.6 (Fonction de répartition empirique)

La fonction de répartition empirique d'un échantillon $(X_1, X_2, \ldots, X_n)$ *est définie par :*

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \ , x \in \mathbb{R}. \tag{1.6}$$

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

Définition 1.1.7 (Fonction des quantiles empirique)

La fonction des quantiles empirique d'un échantillon $(X_1, X_2, \ldots, X_n)$ *est définie par :*

$$Q_n(s) := \inf\{x \in \mathbb{R} : F_n(x) \geq s\}, 0 < s < 1.$$

Théorème 1.1.1 (Glivenko-Cantelli)

La convergence de F_n *vers* F *est presque sûrement uniforme, c'est-à-dire :*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p.s} 0 \text{ quand } n \to \infty.$$

1.1.1 Théorème central limite

Le théorème suivant connu sous le nom de théorème central limite (TCL) établit la convergence en loi vers la loi normale d'une somme de v.a's indépendantes identiquement distribuées (iid) sous des hypothèses très peu contraignantes.

Théorème 1.1.2 (TCL)

Soit $X_1, \ldots, X_n$ *une suite de v.a's réelles iid définies sur le même espace de probabilité* (Ω, A, P)*, d'une fonction de répartition commune* F*, de moyenne* μ *et de variance* σ^2 *finie. Considérons la somme et la moyenne arithmétique correspondantes respectivement :*

$$S_n := X_1 + X_2 + \ldots + X_n \quad , \ \overline{X}_n := \frac{S_n}{n} \ . \tag{1.7}$$

La variable Y_n converge en loi vers la loi normale centrée réduite

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \text{ quand } n \to +\infty. \tag{1.8}$$

1.1.2 Lois des grands nombres

Les lois des grands nombres indiquent que lorsque l'on fait un tirage aléatoire dans une série de grande taille, plus on augmente la taille de l'échantillon, plus les caractéristiques statistiques du tirage (l'échantillon) se rapprochent aux caractéristiques statistiques de la population. Elles sont deux types, lois faibles mettant en jeu la convergence en probabilité $\left(\overset{P}{\rightarrow}\right)$ et lois fortes relatives à la convergence presque sûre $\left(\overset{P.s}{\rightarrow}\right)$.

Théorème 1.1.3 (Lois des grands nombres)

Si $(X_1, \ldots, X_n)$ un échantillon provenant d'une v.a X, tel que $\mu := E(X) < \infty$, alors :

$$\begin{aligned} &\textit{Loi faible :} \quad \overline{X}_n \overset{P}{\rightarrow} \mu \ \textit{quand} \ n \rightarrow +\infty. \\ &\textit{Loi forte :} \quad \overline{X}_n \overset{P}{\rightarrow} \mu \ \textit{quand} \ n \rightarrow +\infty. \end{aligned} \tag{1.9}$$

1.1.3 Statistiques d'ordre

Définition 1.1.8 (Statistiques d'ordre)

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de v.a's iid, de fonction de distribution commune F. En rangeant ces variables aléatoires par ordre croissant, on obtient ce que l'on appelle les statistiques d'ordre notées :

$$X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{k,n} \leq \ldots \leq X_{n,n}. \tag{1.10}$$

Pour $1 \leq k \leq n$, la variable aléatoire $X_{k,n}$ s'appelle la $k^{ème}$ statistique d'ordre.

Définition 1.1.9 (Statistique d'ordre extrêmes)

Les statistiques d'ordres extrêmes $X_{1,n}$ et $X_{n,n}$ sont définies respectivement par le minimum et le maximum de l'échantillon $(X_1, \ldots, X_n)$, comme suit :

$$X_{1,n} := \min\{X_1, X_2, \ldots, X_n\}, \tag{1.11}$$

$$X_{n,n} := \max\{X_1, X_2, \ldots, X_n\}. \tag{1.12}$$

où la variable $X_{1,n}$ est la plus petite statistique d'ordre et la variable $X_{n,n}$ est la plus grande statistique d'ordre.

Remarque 1.1.1 *On peut aussi passer de l'une à l'autre par la relation suivante :*

$$\min\{X_1,\ldots,X_n\} = -\max\{-X_1,\ldots,-X_n\}. \tag{1.13}$$

Distributions des statistiques d'ordres

Soit $X_1,\ldots,X_n$ une suite de v.a's iid de fonction de répartition commune F.

Distribution du maximum

La distribution du maximum $X_{n,n}$ est donnée par :

$$F_{X_{n,n}}(x) := [F(x)]^n. \tag{1.14}$$

En effet :

$$\begin{aligned} F_{X_{n,n}}(x) &= P(X_{n,n} \le x) = P(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \\ &= \prod_{i=1}^{n} P(X_1 \le x) = [P(X_1 \le x)]^n = [F(x)]^n. \end{aligned}$$

Distribution du minimum

La distribution du minimum $X_{1,n}$ est donnée par :

$$F_{X_{1,n}}(x) := 1 - (1 - F(x))^n. \tag{1.15}$$

En effet :

$$\begin{aligned} F_{X_{1,n}}(x) &= P(X_{1,n} \leq x) = 1 - P(X_{1,n} > x) = 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= 1 - \prod_{i=1}^{n} P(X_1 > x) = 1 - \prod_{i=1}^{n} (1 - P(X_1 \leq x)) = 1 - [1 - P(X_1 \leq x)]^n \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

Si F est continue de densité de probabilité f, alors la fonction de densité du maximum et du minimum est donnée respectivement par :

$$f_{Xn,n}(x) = nf(x)[F(x)]^{n-1}, \tag{1.16}$$

et

$$f_{X_{1,n}}(x) = nf(x)(1 - F(x)]^{n-1}. \tag{1.17}$$

Distribution de la k^{ème} statistique d'ordre

La distribution de la k^{ème} statistique d'ordre est donnée par :

$$F_{X_{k,n}}(x) := \sum_{i=k}^{n} \begin{pmatrix} n \\ i \end{pmatrix} [F(x)]^i [1 - F(x)]^{n-i} \ , \ -\infty < x < +\infty.$$

La densité de la $k^{ème}$ statistique d'ordre est :

$$f_{X_{k,n}}(x) := \frac{n!}{(k-1)!\,(n-k)!} [F(x)]^{k-1} \left[\overline{F}(x)\right]^{n-k} f(x), \ 1 \leq k \leq n. \tag{1.18}$$

La démonstration de ces formules est détaillée dans l'ouvrage de Reiss et Thomas [21].

1.2 Distributions des valeurs extrêmes

La distribution du maximum d'un échantillon $(X_1, \ldots, X_n)$ devrait nous fournir des informations sur des évènements extrêmes et comme la limite de cette distribution conduit à une loi dégénérée

$$\lim_{n\to+\infty} F_{X_{n,n}}(x) = \lim_{n\to\infty} F^n(x) = \begin{cases} 1 & \text{si} \quad F(x) = 1, \\ 0 & \text{si} \quad F(x) < 1, \end{cases} \tag{1.19}$$

on cherche une loi non dégénérée pour $X_{n,n}$. De façons analogue au TCL, Ficher et Tippett [11] en (1928), Gnedenko [12] en (1943) et de Haan [13] en (1976) ont montrés que les seules distributions limites non dégénérée sont les distributions des valeurs extrêmes.

Théorème 1.2.1 (Fisher etTippett (1928))

Soit $X_1, \ldots, X_n$ n v.a's iid de fonction de répartition F. S'il existe deux suites normalisantes réelles $(a_n)_{n\geq 0} > 0$, et $(b_n)_{n\geq 0} \in \mathbb{R}$, alors :

$$\lim_{n\to+\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to+\infty} F^n(a_n x + b_n) := H(x), \ \forall x \in \mathbb{R}, \tag{1.20}$$

où H est une fonction de distribution non dégénérée, appelée distribution des valeurs extrêmes (EVD : Extremes values Distribution).

La distribution H est du même type que l'une des trois distributions des valeurs extrêmes standard suivantes :

Distribution de Gumbel : $\Lambda(x) := \exp[-\exp(-x)]$; $x \in \mathbb{R}$.

Distribution de Fréchet : $\Phi_\alpha(x) := \begin{cases} 0 & si \quad x \leq 0, \\ \exp(-x^{-\alpha}) & si \quad x > 0. \end{cases}$; $\alpha > 0$.

Distribution de Weibull : $\Psi_\alpha(x) := \begin{cases} \exp[-(-x)^\alpha] & si \quad x \leq 0, \\ 1 & si \quad x > 0. \end{cases}$; $\alpha > 0$.

Exemple 1.2.1 (Loi exponentielle) *Soit X une v.a suit la loi exponentielle standard de fonction de répartition $F(x) = 1 - \exp(-x)$. Prenons $a_n = 1$ et $b_n = \log(n)$, alors $\frac{X_{n,n} - b_n}{a_n}$ tend asymptotiquement vers la loi de Gumbel. En effet :*

$$\begin{aligned} P(X_{n,n} - \log n \leq x) &= F^n(x + \log(n)) = (1 - \exp[-(x + \log(n))])^n = \left(1 - \frac{e^{-x}}{n}\right)^n \\ &= \exp\left[n \log\left(1 - \frac{e^{-x}}{n}\right)\right] \approx \exp[-\exp(-x)] \ \ \textit{quand } n \to +\infty \\ &= \Lambda(x). \end{aligned}$$

1.3 Distribution GEV

Définition 1.3.1 (Jenkinson (1955), [15])

La fonction de distribution H_γ de la famille des valeurs extrêmes généralisées (GEV : Generalized Extreme Value), pour $\gamma \in \mathbb{R}$, est .

$$H_\gamma(x) := \begin{cases} \exp\left[-(1+\gamma x)^{-\frac{1}{\gamma}}\right] & si \quad \gamma \neq 0, \quad 1 + \gamma x > 0, \\ \exp[-\exp(-x)] & si \quad \gamma = 0, \quad x \in \mathbb{R}. \end{cases} \tag{1.21}$$

où γ s'appelle indice des valeurs extrêmes.(EVI : Extreme Value Index).

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

$$h_\gamma(x) := \begin{cases} H_\gamma(x)(1+\gamma x)^{\frac{-1}{\gamma-1}} & si \quad \gamma \neq 0, \quad 1 + \gamma x > 0, \\ \exp[-x - \exp(-x)] & si \quad \gamma = 0, \quad x \in \mathbb{R}. \end{cases} \tag{1.22}$$

Pour $\gamma = 0$, il faut lire $H_0(x) = \exp[-\exp(-x)], x \in \mathbb{R}$, qui s'obtient dans H_γ en faisant tendre γ vers 0. Les lois des valeurs extrêmes généralisées correspondent, à une translation et changement d'échelle prés, au loi des valeurs extrêmes standard. Nous avons alors les correspondances suivantes :

$$\begin{aligned}\Lambda(x) = &\quad H_0(x),\\ \Phi_\alpha(x) = &\quad H_{\frac{1}{\alpha}}(\alpha(x-1)) \quad ,x\in\mathbb{R},\\ \Psi_\alpha(x) = &\quad H_{\frac{-1}{\alpha}}(\alpha(x+1)) \quad ,x\in\mathbb{R}.\end{aligned}$$

Pour les variables non centrées et non réduites, on peut écrire $H_\gamma(x)$ sous une forme plus générales dite forme paramétrée de von Mises, dans laquelle on fait apparaître un paramètre de localisation $\mu\in\mathbb{R}$ et un paramètre d'échelle $\sigma>0$, telle que

$$H_{\mu,\sigma,\gamma}(x) := \begin{cases} \exp\left\{-\left[1+\gamma\frac{(x-\mu)}{\sigma}\right]^{-\frac{1}{\gamma}}\right\} & si \quad \gamma\neq 0, \quad 1+\gamma\frac{x-\mu}{\sigma}>0,\\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & si \quad \gamma=0, \quad x\in\mathbb{R}.\end{cases} \tag{1.23}$$

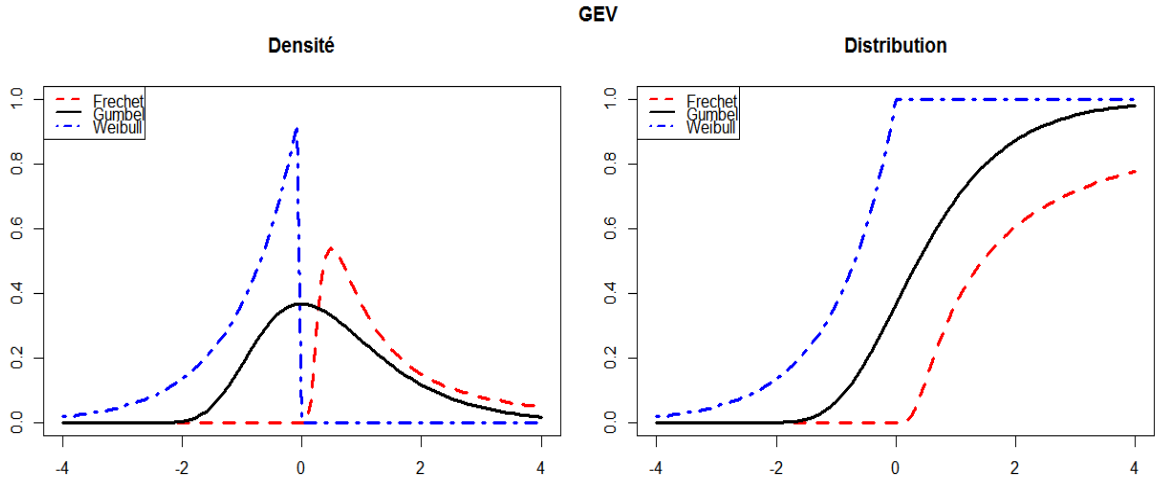

FIG. 1.1 – Densités et Distribution de Lois des Valeurs Extrêmes

Définition 1.3.2 (Fonction à variation régulière)

On dit qu'une fonction h mesurable sur $(0,\infty)$ est à variation régulière d'indice $\rho\in\mathbb{R}$ et on note $h\in RV_\rho$, si et seulement si

$$\lim_{t\to+\infty}\frac{h(tx)}{h(t)} = x^\rho, x>0. \tag{1.24}$$

où : RV_ρ l'ensemble des fonctions à variations régulières d'indice ρ.

Définition 1.3.3 (Fonction à variation lente)

Si une fonction $L : \mathbb{R} \mapsto [0, +\infty[$ est à variation régulière d'indice 0 ($L \in RV_0$), on dit que L est à variation lente, telle que :

$$\lim_{t\to+\infty} \frac{L(tx)}{L(t)} = 1 \quad , x > 0. \tag{1.25}$$

Exemple 1.3.1 *La fonction logarithme est une fonction à variation lente. En effet, soit $t > 0$*

$$\lim_{t\to+\infty} \frac{\log(tx)}{\log x} = 1 + \lim_{t\to+\infty} \frac{\log(t)}{\log(x)} = 1.$$

Remarque 1.3.1 *Tout fonction $h \in RV_\rho$ peut s'écrire sous la forme :*

$$h(x) := x^\rho L(x) \quad ,où\ L \in RV_0.$$

Proposition 1.3.1 (Condition du second ordre)

On dit que la fonction quantile de queue U est à variation régulière du second ordre, on écrit $U \in 2RV_{\gamma,\rho}$, avec $\gamma > 0$ est le paramètre de premier ordre et $\rho \leq 0$ le paramètre du second ordre. S'il existe une fonction $A^*(t) \to 0$ et ne change pas le signe au voisinage de ∞, telles que

$$\lim_{t\to+\infty} \frac{U(tx)/U(t) - x^\gamma}{A^*(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \quad x > 0, \tag{1.26}$$

avec $|A^*| \in RV_\rho$ est appelée fonction auxiliaire.

1.4 Domaine d'attraction

Définition 1.4.1 (Domaine d'attraction)

On dit qu'une distribution F appartient au domaine d'attraction de H_γ, et on note $F \in D(H_\gamma)$, si la distribution du maximum renormalisée converge vers H. Autre-

ment dit, s'il existe des constantes réelles $a_n > 0$ *et* $b_n \in \mathbb{R}$*, tels que :*

$$\lim_{n\to+\infty} F^n\left(a_n x + b_n\right) = H_\gamma\left(x\right) = \exp\left[-\left(1+\gamma x\right)^{-\frac{1}{\gamma}}\right], avec 1+\gamma x > 0.$$

1.4.1 Caractérisation des domaines d'attraction

Théorème 1.4.1 (Caractérisation du $D\left(\Phi_\gamma\right)$**)**

La fonction da répartition F *appartient au domaine d'attraction de la loi de Fréchet de paramètre* $\gamma > 0$ *si et seulement si*

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

où la fonction $L \in RV_0$*. En particulier* $x_F = +\infty$*. De plus si* $F \in D\left(\Phi_\gamma\right)$*, alors avec* $a_n = U\left(n\right) = F^{-1}\left(1-1/n\right)$ *et* $b_n = 0$*, la suite* $\left(a_n^{-1} X_{n,n}\right)_{n\geq 1}$ *converge en loi vers une v.a de fonction de répartition* Φ_γ*, quand* $n \to \infty$*.*

Théorème 1.4.2 (Caractérisation du $D\left(\Psi_\gamma\right)$**)**

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

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

où $L \in RV_0$*. De plus si* $F \in D\left(\Psi_\gamma\right)$*, alors avec* $a_n = x_F - U\left(n\right) = x_F - Q\left(1-1/n\right)$ *et* $b_n = x_F$*, la suite* $\left(a_n^{-1}\left(X_{n,n} - x_F\right)\right)_{n\geq 1}$ *converge en loi vers une v.a de fonction de répartition* Ψ_γ*, quand* $n \to \infty$*.*

Théorème 1.4.3 (Caractérisation du $D\left(\Lambda\right)$**)**

La fonction da répartition F *appartient au domaine d'attraction de la loi de Gumbel si et seulement si*

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

où c et g sont deux fonctions mesurables satisfaites $c(x) \to c > 0$ et $g(x) \to 1$ quand $x \to x_F$ et a est une fonction positive, absolument continue (par rapport la mesure de Lebesgue) avec la densité a' ayant $\lim_{x \to x_F} a'(x) = 0$. Dans ce cas, un choix possible pour les suites de normalisation est :

$$a_n = x_F - F^{-1}(1 - 1/n) \text{ et } b_n = \frac{1}{\bar{F}(x)} \int_{a_n}^{x_F} \bar{F}(y)\, dy.$$

Propriété 1.4.1

Nous voyons que les trois distributions des valeurs extrêmes standard sont très différentes en terme de domaine d'attraction :

- *Sous le domaine d'attraction de Gumbel, nous trouvons des distributions dont la fonction de survie décroit vers 0 à une vitesse exponentielle.*
- *Sous le domaine d'attraction de Fréchet, les distributions sont à queue lourde.*
- *Sous le domaine d'attraction de Weibull, le point terminal des distributions est fini.*

Le tableau suivant résume le classement de quelques lois usuelles selon leur appartenance à l'une des domaines d'attraction.

TAB. 1.1 – Quelques loi usuelles classées en fonction de leur domaine d'attraction

Domaines d'attraction	*Gumbel* $\gamma = 0$	*Fréchet* $\gamma > 0$	*Weibull* $\gamma < 0$
	Normale	Pareto	Uniforme
	Exponentielle	Burr	Beta
Lois	Log-normale	Student	
	Gamma	Log-gamma	
	Weibull		

1.5 Distribution GPD

1.5.1 Distribution des excès

Nous supposons $X_1, X_2, \ldots, X_n$ une suite de v.a's iid, de fonction de répartition F de point terminal x_F. Alors, pour un seuil fixé $u < x_F$, on définit la variable

$$Y_i = X_i - u, \ i = 1, .., n,$$

les excès au-dessus du seuil u, dans le cadre de l'approche (POT : Peaks over Threshold ou bien piques au-delà d'un seuil).

Définition 1.5.1 (Distribution des excès)

La fonction de répartition des excès de X au-dessus du seuil u est définie par :

$$F_u(y) := P(X - u \leq y \mid X > u). \tag{1.27}$$

De plus,

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

1.5.2 Distribution de Paréto généralisée

Définition 1.5.2 (Distribution de Paréto généralisée)

La fonction de répartition de Paréto généralisée (GPD : Genereleze Pareto Distribution

), est définit, pour $\gamma \in \mathbb{R}$, $\sigma > 0$, *comme suit :*

$$G_{\gamma,\sigma,\mu}(x) := \begin{cases} 1 - \left(1 + \gamma \dfrac{x-\mu}{\sigma}\right)^{-\dfrac{1}{\gamma}} & si \quad \gamma \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right) & si \quad \gamma = 0. \end{cases} \tag{1.28}$$

où

$$\begin{aligned} x \geq 0 \qquad & si \quad \gamma \geq 0, \\ 0 \leq x \leq -\tfrac{\sigma}{\gamma} \quad & si \quad \gamma < 0. \end{aligned}$$

On note que la GPD standard est correspond au cas où $\mu = 0$ et $\sigma = 1$, telle que

$$G_{\gamma}(x) := \begin{cases} 1 - (1 - \gamma x)^{-\frac{1}{\gamma}} & si \quad \gamma \neq 0, \\ 1 - \exp(-x) & si \quad \gamma = 0. \end{cases} \tag{1.29}$$

avec

$$\begin{aligned} x \geq 0 \qquad & si \quad \gamma \geq 0, \\ 0 \leq x \leq -\tfrac{1}{\gamma} \quad & si \quad \gamma < 0. \end{aligned}$$

Lorsque le paramètre de localisation est nul ($\mu = 0$) et le paramètre d'échelle est arbitraire ($\sigma > 0$), cette distribution joue un rôle important.

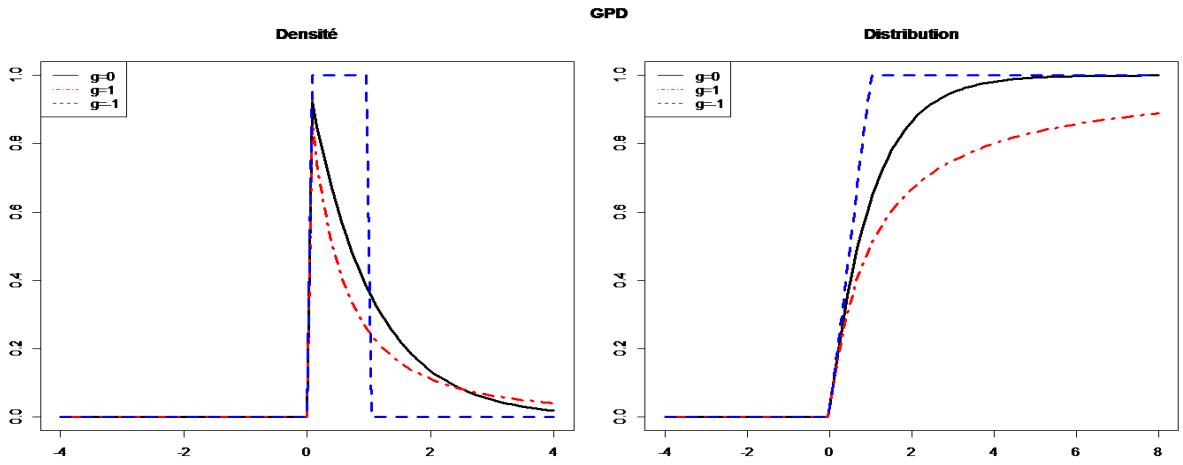

FIG. 1.2 – Densités et distributions de loi Pareto Généralisées avec différentes valeurs de γ.

Belkema et de Haan [3] et Pickands [20] ont proposés le théorème suivant qui fait le lien entre le comportement asymptotique de la distribution des excès et la loi de Paréto généralisée.

Théorème 1.5.1 (Balkema et de Haan(1974), Pickands(1975))

Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes, alors il existe une fonction $\sigma(u)$ positive et un réel γ tel que :

$$\lim_{u \to x_F} \sup_{0<y\leq x_F-u} \left|F_u(y) - G_{\gamma,\sigma(u)}(y)\right| = 0, \tag{1.30}$$

où : $G_{\gamma,\sigma(u)}$ est la fonction de répartition de la loi de Paréto généralisée et F_u est la fonction de répartition des excès au-dela du seuil u .

Remarque 1.5.1

1/ *On écrit la densité de la distribution GPD comme suit :*

$$g_{\gamma,\sigma}(x) := \begin{cases} \frac{1}{\sigma}\left(1+\gamma\frac{x}{\sigma}\right)^{-\frac{1}{\gamma}-1} & si \quad \gamma \neq 0, \\ \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right) & si \quad \gamma = 0. \end{cases} \tag{1.31}$$

2/ *Il y'a un rapport simple entre la GPD standard $G_\gamma(x)$ et la GEV standard $H\gamma(x)$, tels que :*

$$G_\gamma(x) = 1 + \log H_\gamma(x), \ si \ \log H_\gamma(x) > -1.$$

1.6 Notions de base en analyse de survie

Nous présentons dans cette section quelques définitions qui sont couramment utilisées dans l'analyse de survie.

Définition 1.6.1 (Date d'origine)

Elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque.

Définition 1.6.2 (Date des dernières nouvelles)

C'est la date la plus récente où des informations sur un sujet ont été recueillies.

Définition 1.6.3 (Date de point)

C'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

Définition 1.6.4 (Durée de survie)

La durée de survie est la durée entre la date d'origine et la survenue de l'évènement d'intérêt, c'est-à-dire du décès. Elle correspond au temps de survie lorsque les décès sont observés avant la date de point.

1.7 Données censurées

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour un individu donné j, on va considérer :

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

Définition 1.7.1 (Variable de censure)

La variable de censure C est définie par la non-observation de l'événement étudié. Si l'on observe C, et non X, et que l'on sait que $X > C$ respectivement ($X < C$, $C_1 < X < C_2$), on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).

Si l'événement se produit, X est "réalisée". S'il ne se produit pas (l'individu étant perdu de vue, ou bien du vivant), c'est C qui est "réalisée".

Censure à droite

La variable d'intérêts est dite censurée à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées.

Censure à gauche

Il y a une censure à gauche lorsque l'individu a déjà subi l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue.

Censure double et mixte

On dit qu'on a une censure double ou mixte si on a des données censurées à droite et des données censurées à gauche dans le même échantillon. Plusieurs modèles non-paramétriques ont été présentés pour l'étude de la double censure.

Censure par intervalle

Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt.

1.7.1 Types de censures

Censure de type I : (fixée)

Soit C une valeur fixée, au lieu d'observer les variables $X_1, X_2, \ldots, X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq C$, sinon on sait uniquement que $X_i > C$. On observe donc une variable aléatoire Z_i telle que

$$Z_i = X_i \wedge C = \min(X_i, C),\ i = 1, \ldots, n.$$

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles.

Exemple 1.7.1 *En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps μ sont sacrifiées)*

Censure de type II : (attente)

L'expérimentateur fixe a priori le nombre d'événements à observer. La date définie d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie.

Par exemple en épidémiologie, 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 $X_{i,n}$ et $Z_{i,n}$ les statistiques d'ordre associées à la v.a X_i et Z_i respectivement La date de censure est donc $X_{k,n}$ et on observe les variables suivantes :

$$Z_{1,n} = X_{1,n}, Z_{2,n} = X_{2,n}, \ldots, Z_{k,n} = X_{k,n}, Z_{k+1,n} = X_{k,n}, \ldots, Z_{n,n} = X_{k,n}$$

Censure de type III (ou censure aléatoire de type I)

Soient $C_1, C_2, \ldots, C_n$ des v.a's iid, on observe les variables

$$Z_i = X_i \wedge C_i \text{ et } \delta_i = \mathbb{1}_{\{X_i \leq C_i\}} \text{ pour } i = \overline{1,n} \tag{1.32}$$

où Z_i la durée réellement observée, C_i est une censure aléatoire et δ sert à connaître la nature de l'observation, il indique si l'on est face à une observation réelle ($\delta = 1$) (d'où $Z_i = X_i$) ou à une censure ($\delta = 0$) (d'où $Z_i = C_i$). La censure aléatoire est la plus courante, et la plus considérée en analyse de survie.

Chapitre 2

Estimation des statistiques des extrêmes

Dans ce chapitre, on s'intéresse à l'estimation des différentes statistiques des extrêmes, telles que l'indice des valeurs extrêmes, la fonction de survie et les quantiles extrêmes, dans le cas des données complètes et censurées. Ces estimateurs jouent un rôle prémordial dans la théorie des valeurs extrêmes.

2.1 Estimation de l'indice des valeurs extrêmes

2.1.1 Estimation sous données complètes

Estimation paramétrique

Estimateur de maximum de vraisemblance

L'estimation par la méthode de maximum de vraisemblance (EMV) donne des résultats asymptotiques efficaces, et les estimateurs obtenus convergent sous certaines conditions vers les vraies valeurs des paramètres.

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de v.a's iid de densité h_θ où $\theta = (\mu, \sigma, \gamma)$, pour

lesquelles la distribution de GEV (1.23) est appropriée, la vraisemblance se factorise en

$$L\left(\theta; X_1, X_2, \ldots, X_n\right) := \prod_{i=1}^{n} h_\theta\left(X_i\right). \tag{2.1}$$

La fonction log-vraisemblance est donnée par :

$$l\left(\theta; X_1, X_2, \ldots, X_n\right) := \log L\left(\theta; X_1, X_2, \ldots, X_n\right). \tag{2.2}$$

Par conséquent :

$$\begin{aligned} l\left(\theta; X_1, X_2, \ldots, X_n\right) &= \sum_{i=1}^{n} \log h_\theta\left(X_i\right) \\ &= -n\log\sigma - \left(\frac{1}{\gamma}+1\right)\sum_{i=1}^{n}\log\left(1+\gamma\frac{X_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\left(1+\gamma\frac{X_i-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}, \end{aligned}$$

qui doit être maximale. L'EMV correspond alors :

$$\hat{\theta}_n = \hat{\theta}_n\left(X_1, X_2, \ldots, X_n\right) = \arg\max_{\theta\in\ominus} l\left(\theta; X_1, X_2, \ldots, X_n\right)$$

avec $\ominus$ est l'espace des paramètres.

Si $l\left(\theta; X_1, X_2, \ldots, X_n\right)$ admet des dérivées partielles par rapport μ, σ et γ respectivement, alors l'EMV est souvent obtenu en résolvant les équations suivantes :

$$\frac{\partial l\left(\theta; X_1, X_2, \ldots, X_n\right)}{\partial\theta} = 0.$$

Au cas où $\gamma = 0$ (loi Gumbel), la fonction log-vraisemblance est égale à

$$l\left(\theta; X_1, X_2, \ldots, X_n\right) = -n\log\sigma - \sum_{i=1}^{n}\exp\left(-\frac{X_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)$$

En dérivant cette fonction relativement aux deux paramètres μ et σ respectivement, nous

obtenons le système d'équations suivant :

$$\begin{cases} \dfrac{\partial l}{\partial \sigma} = 0 \iff n + \displaystyle\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \left[\exp\left(-\frac{X_i - \mu}{\sigma} - 1 \right) \right] = 0, \\ \dfrac{\partial l}{\partial \mu} = 0 \iff n - \displaystyle\sum_{i=1}^{n} \exp\left(-\frac{X_i - \mu}{\sigma} \right) = 0. \end{cases}$$

La résolution de ce système est relativement difficile et n'admet pas en général de solution explicite. Dans ce cas, on fait appel à des méthodes d'optimisation numériques.

Estimation semi-paramétrique

Dans cette partie, on exposera uniquement deux estimateurs de γ, tel que l'estimateur de Pickands [20] et l'estimateur de Hill [14]. On donne également certaines de leurs propriétés statistiques. Ces estimateurs sont basées fortement sur les plus grandes statistiques d'ordre $X_{k,n} \leq \ldots \leq X_{n,n}$ où la statistique $X_{n-k,n}$ est alors dite statistique d'ordre intermédiaire.

Estimateur de Pickands

L'estimateur de Pickands a été introduit en 1975 par Pickands [20] pour toute $\gamma \in \mathbb{R}$.

Définition 2.1.1 (Estimateur de Pickands)

Soit $X_1, X_2, \ldots, X_n$, n v.a's iid de fonction de répartition $F \in D(H_\gamma)$, où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Pickands est défini par :

$$\hat{\gamma}^P = \hat{\gamma}^P(k) := \frac{1}{\log 2} \log \left(\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}} \right). \tag{2.3}$$

La consistance faible et la consistance forte a été obtenue par Pickands (1975) [20] et la normalité asymptotique a été démontré par Dekkers et de Haan (1989) [8].

Théorème 2.1.1 (Propriétés asymptotiques de $\hat{\gamma}$)

Soit $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$, $k \to \infty$ et $\dfrac{k}{n} \to 0$ quand $n \to \infty$.

(i) **Consistance faible :**

$$\hat{\gamma}^P \xrightarrow{P} \gamma, \text{ quand } n \to \infty.$$

(ii) **Consistance forte :** Si $\dfrac{k}{\log\log n} \to \infty$ quand $n \to \infty$, alors

$$\hat{\gamma}^P \xrightarrow{P.s} \gamma, \text{ quand } n \to \infty.$$

(iii) **Normalité asymptotique** : Supposons que U admet des dérivés positifs U' et que $\pm t^{1-\gamma}U'(t)$ (avec l'un ou l'autre choix de signe) est à variation régulière à l'infini avec la fonction auxiliaire a. Si $k = o\left(\dfrac{n}{g^{-1}(n)}\right)$ $(n \to \infty)$, où $g(t) = t^{3-2\gamma}\left(\dfrac{U'(t)}{a(t)}\right)^2$, alors

$$\sqrt{k}\left(\hat{\gamma}^P - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \eta^2\right), \text{ quand } n \to \infty,$$

où $\eta^2 := \dfrac{\gamma^2\left(2^{2\gamma+1}+1\right)}{\left(2\left(2^\gamma-1\right)\log 2\right)^2}$.

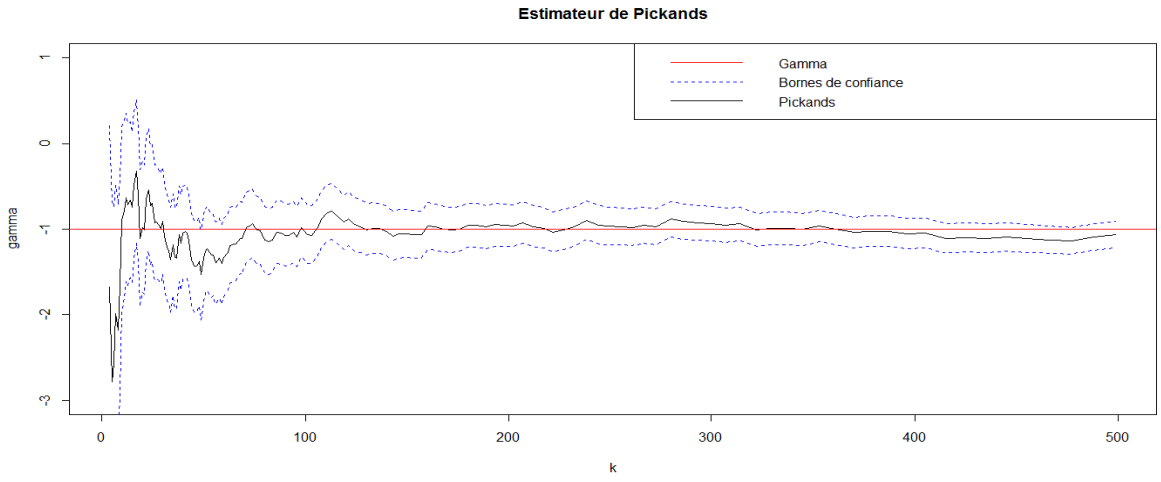

FIG. 2.1 – Estimateur de Pickands, avec un intervalle de confiance de niveau 95% pour l'EVI de la distribution de Paréto standard ($\gamma = -1$) basé sur 200 échantillons de 5000 observations.

Estimateur de Hill

L'estimateur de l'EVI le plus populaire et le plus célèbre est l'estimateur de Hill. Il a été introduit en 1975 [14], il est cependant limité au cas de Fréchet ($\gamma > 0$).

Définition 2.1.2 (Estimateur de Hill)

Soit $X_1, X_2, \ldots, X_n$, n v.a's iid de fonction de répartition $F \in D\left(\Phi_{1/\gamma}\right)$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Hill est défini par la statistique

$$\hat{\gamma}^H = \hat{\gamma}^H(k) := \frac{1}{k-1} \sum_{i=n-k+2}^{n} \log X_{i,n} - \log X_{n-k+1,n}, \tag{2.4}$$

ou encore par :

$$\hat{\gamma}_k^H := \frac{1}{k} \sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}. \tag{2.5}$$

Cet estimateur est consistant au sens [17] faible et fort [7], de plus il est asymptotiquement normal de variance [6].

Théorème 2.1.2 (Propriétés asymptotiques $\hat{\gamma}^H$)

On suppose que $F \in D\left(\Phi_{\frac{1}{\gamma}}\right), \gamma > 0, k \to \infty$ et $\dfrac{k}{n} \to 0$ quand $n \to \infty$.

(a) **Consistance faible :**

$$\hat{\gamma}^H \xrightarrow{P} \gamma, \text{ quand } n \to \infty.$$

(b) **Consistance forte :** Si $\dfrac{k}{\log\log n} \to \infty$ quand $n \to \infty$, alors

$$\hat{\gamma}^H \xrightarrow{P.s} \gamma, \text{ quand } n \to \infty.$$

(c) **Normalité asymptotique :** Supposons que F satisfaisant (1.26) si $\sqrt{k}A\left(\dfrac{n}{k}\right) \to \lambda$ quand $n \to \infty$, alors

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

Ce dernier résultat permet de calculer des intervalles de confiance pour γ. Par exemple, à un niveau de confiance de $(1-\alpha)\,\%$; on a

$$\gamma \in \left]\hat{\gamma}^H - q_{1-\alpha/2}\frac{\hat{\gamma}^H}{\sqrt{k}}, \hat{\gamma}^H + q_{1-\alpha/2}\frac{\hat{\gamma}^H}{\sqrt{k}}\right[.$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ d'une loi normale centrée réduite.

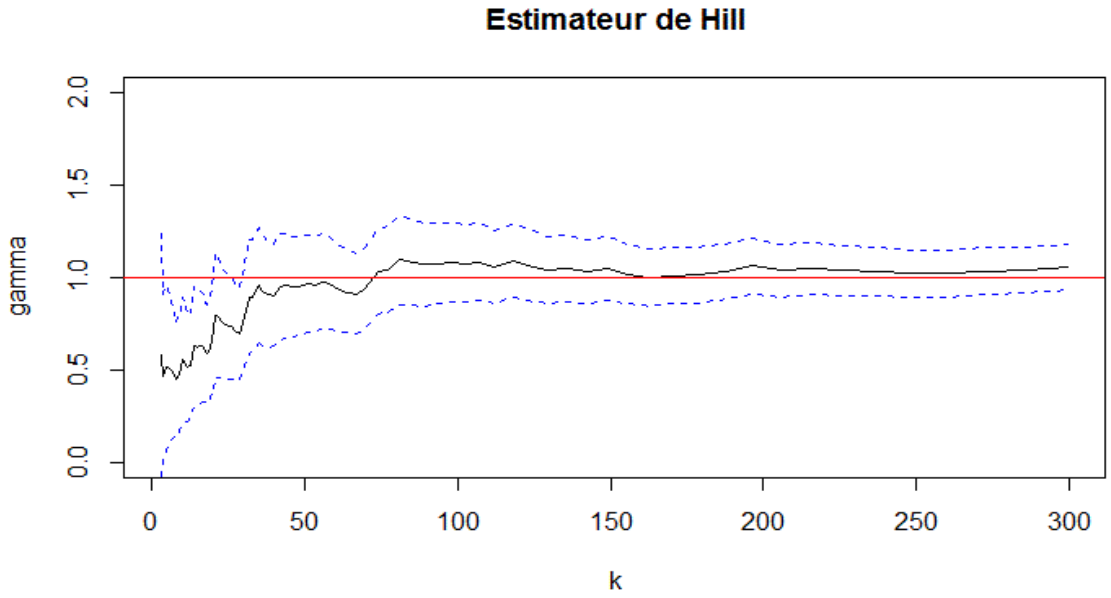

FIG. 2.2 – Estimateur de Hill, avec un intervalle de confiance de niveau 95% pour l'EVI de la distribution de pareto standard ($\gamma = 1$) basé sur 100 échantillons de 3000 observations.

2.1.2 Estimation sous données censurés

Estimateur de Hill adapté

Soient $(X_i)_{i\leq n}$ et $(C_i)_{i\leq n}$ deux échantillons de v.a iid où F et G sont respectivement leur fonctions de répartition absolument continues (avec x_F et x_G sont les point terminaux respectifs). Dans le cas des données censurées aléatoirement à droite, on observe

$\{(Z_i, \delta_i), 1 \leq i \leq n\}$ un échantillon de couple de v.a's, tel que

$$Z_i = X_i \wedge C_i \text{ et } \delta_i = \mathrm{I}_{\{X_i \leq C_i\}}.$$

Nous définissons H la distribution de l'échantillon $(Z_i)_{i \leq n}$, satisfaisant :

$$1 - H = (1 - F)(1 - G),$$

et $Z_{1,n} \leq Z_{2,n} \leq \cdots \leq Z_{n,n}$ les statistiques d'ordre lui associées. Avec $\delta_{[1:n]}, \ldots, \delta_{[n:n]}$ sont les indicateurs de censure retenues avec ces dernières $\left(\delta_{[i:n]} = \delta_j \text{ si } Z_{i,n} = Z_j\right)$. Si F et G sont absolument continue et que $F \in D(H_{\gamma_1})$, $G \in D(H_{\gamma_2})$ respectivement, pour certains γ_1, $\gamma_2 \in \mathbb{R}$. Pour tout $\gamma \in \mathbb{R}$, Einmahl et al (2008) ont proposés les trois cas les plus intéressants suivants :

$$\begin{cases} \text{cas } \mathbf{1}: & \gamma_1 > 0, \gamma_2 > 0 & & \gamma = \dfrac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2} \\ \text{cas } \mathbf{2}: & \gamma_1 < 0, \gamma_2 < 0 & x_F = x_G & \gamma = \dfrac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2} \\ \text{cas } \mathbf{3}: & \gamma_1 = \gamma_2 = 0 & x_F = x_G = \infty & \gamma = 0 \end{cases} \tag{2.6}$$

Dans le $3^{ème}$ cas, nous définissons également, pour une présentation commode, $\gamma = \dfrac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2} = 0$. Les autres possibilités ne sont pas très intéressantes. Pratiquement, ils sont très proches du cas non censuré, qui a été étudié en détail dans la littérature (cela arrive, en particulier, quand $\gamma_1 > 0$ et $\gamma_2 < 0$) ou la situation complètement censurée, où l'estimation est impossible

Beirlant et al (2007) [2] ont proposés différents estimateurs de l'indice des valeurs extrêmes γ_1 associé à F dans le cas des données censurées, ces derniers sont tous construits de façon similaire, à partir d'un estimateur non adapté à la censure, par exemple l'estimateur de Hill. Ces estimateurs basés sur les observations Z_i, estiment par conséquent l'indice γ de H. Il s'agit alors de les modifier de façon à estimer γ_1 et non γ. Une façon de procéder

consiste à diviser ces estimateurs usuels (non adaptés à la censure) par la proportion de données non censurées au-delà d'un seuil u, c'est-à-dire à utiliser

$$\hat{\gamma}_1^{(c,.)} = \frac{\hat{\gamma}^{(.)}}{\hat{p}}, \tag{2.7}$$

où

$$\hat{p} = \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]},$$

avec k est le nombre des excès au-delà de u et $\hat{p}$ estime $p = \dfrac{\gamma_2}{\gamma_1+\gamma_2}$ ($\hat{p} \to p$ quand $n \to \infty$), par conséquent $\hat{\gamma}^{(.)}$ l'estimateur de γ divisé par $\dfrac{\gamma_2}{\gamma_1+\gamma_2}$ qui est égal à γ_1.

Pour adapter l'estimateur de Hill dans le cas de censure nous allons diviser cet estimateur $\hat{\gamma}^H$ par la proportion de données non censurées des k plus grandes valeurs de Z, alors l'estimateur de Hill adapté à l'indice de queue $\hat{\gamma}_1^c$ est défini par :

$$\hat{\gamma}_1^c = \frac{\hat{\gamma}^H}{\hat{p}}, \tag{2.8}$$

où

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

alors

$$\hat{\gamma}_1^c(k) = \frac{\dfrac{1}{k}\displaystyle\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}}{\dfrac{1}{k}\displaystyle\sum_{i=1}^{k} \delta_{[n-i+1:n]}}. \tag{2.9}$$

Einmahl et al (2008) [9] ont établis de façon unifiée, la normalité asymptotique de tout estimateur de l'indice des valeurs extrêmes écrit sous la forme (2.7) dans le cas où le seuil choisi u est aléatoire et égal à $Z_{n-k,n}$ la $(n-k)$-ième statistique d'ordre de l'échantillon $Z_1, \ldots, Z_n$.

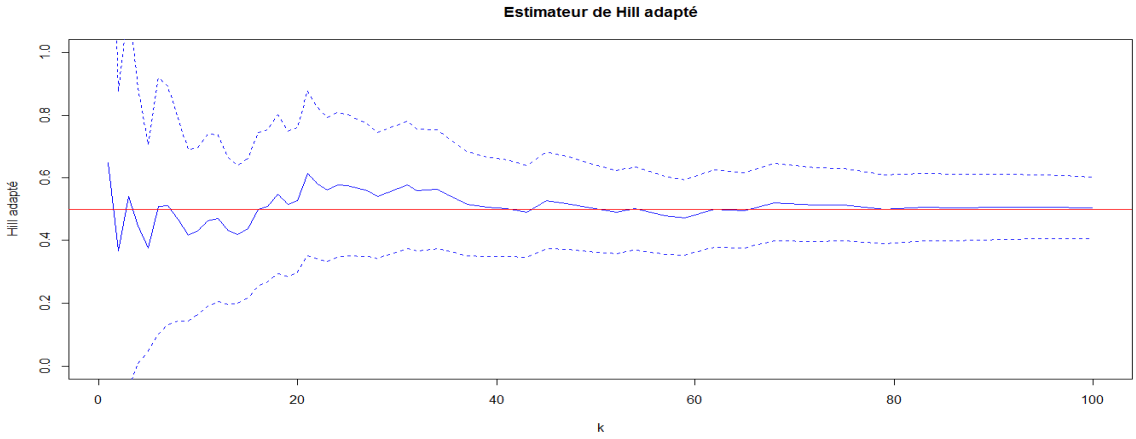

FIG. 2.3 – Estimateur de Hill adapté pour 100 échantillons de 1000 observations.

2.2 Estimation de la fonction de survie

2.2.1 Estimation de la queue sous données complètes

Distribution à variation finie

Soit $(X_1, \ldots, X_n)$ un échantillon de n v.a's iid de fonction de répartition commune F, d'espérance μ et de variance fini σ^2. L'estimateur de $\bar{F}$ est donné par la fonction de queue empirique :

$$\widehat{\overline{F}}(x) := \bar{F}_n(x) = 1 - F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}}.$$

Propriété 2.2.1

Sur la base des propriétés de la fonction de répartition empirique F_n, on a :

Biais : *$\bar{F}_n(x)$ est un estimateur sans biais se $\overline{F}$, c'est-à-dire*

$$E\left[\bar{F}_n(x)\right] = \overline{F}(x).$$

Variance : *La variance de $\bar{F}_n(x)$ est donnée par :*

$$Var\left[\bar{F}_n(x)\right] = F(x)\,\overline{F}(x).$$

Consistance faible et forte : *A travers la loi des grands nombres, on a :*

$$\bar{F}_n(x) \xrightarrow{P} \overline{F}(x), \text{ quand } n \to \infty.$$

et

$$\bar{F}_n(x) \xrightarrow{P.s} \overline{F}(x), \text{ quand } n \to \infty.$$

Normalité asymptotique : *L'application du théorème central limite sur cet estimateur nous donne :*

$$\sqrt{n}\frac{\bar{F}_n(x) - \overline{F}(x)}{\sqrt{F(x)\overline{F}(x)}} \xrightarrow{d} \mathcal{N}(0,1), \text{ quand } n \to \infty.$$

Distribution à queue lourde

Définition 2.2.1 (Distribution à queue lourde)

Soit F une fonction de répartition appartient au domaine d'attraction de Fréchet d'indice $1/\gamma$, $F \in D\left(\Phi_{1/\gamma}\right)$. On dit que F est à queue si la queue de distribution $\overline{F}$ est une fonction à variation réguliére d'indice $-1/\gamma$, c'est-à-dire :

$$\overline{F}(x) = x^{-1/\gamma} L(x), x \to \infty, L \in RV_0.$$

Proposition 2.2.1 (Fonction à variation régulière du premier ordre)

Les assertions suivantes sont équivalentes :

a) F à queue lourde.

b) $\bar{F}$ est à variation régulière à l'infini d'indice $-1/\gamma$

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

c) $Q(1-s)$ est à variation régulière à 0 d'indice $-\gamma$

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

d) U est à variation régilière à l'infini d'indice γ

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

Estimation de la queue

Soit $X_1,\ldots,X_n$ une suite de v.a's iid de fonction de répartition commune F, on suppose que $F\in D\left(\Phi_{1/\gamma}\right)$ (F est à queue lourde). L'estimateur de $\bar{F}$ est donné par :

$$\widehat{\overline{F}}(x)=1-\hat{F}(x)=1-F\left(\hat{x}_p\right), x\to\infty,$$

où $\hat{x}_p$ est un estimateur du quantile extrême $x_p:=F^{-1}(1-p), p\to 0$. Cet estimateur est donnée dans la section 2.3 pour l'estimateur de Hill par la statistique (2.13).

Par conséquent

$$\widehat{\overline{F}}(x)=p, x\to\infty.$$

Et comme

$$\hat{x}_p^H:=X_{n-k,n}\left(\frac{k}{np}\right)^{\hat{\gamma}_k^H}. \tag{2.10}$$

En exposant (2.10) par $1/\hat{\gamma}_k^H$, on trouve

$$\hat{x}_p^{1/\hat{\gamma}_k^H}=\left(X_{n-k,n}\right)^{1/\hat{\gamma}_k^H}(k/np),\ p\to 0,$$

ce qui nous donne :

$$p=\frac{k}{n}\left(X_{n-k,n}\right)^{1/\hat{\gamma}_k^H}\hat{x}_p^{-1/\hat{\gamma}_k^H},\ p\to 0,$$

alors

$$p = \frac{k}{n} \left(X_{n-k,n}\right)^{1/\hat{\gamma}_k^H} x^{-1/\hat{\gamma}_k^H}, \ x \to \infty.$$

Finalement, on déduit l'estimateur de la queue de distribution pour les distributions à queue lourde :

$$\widehat{\overline{F}}\left(x\right) = \frac{k}{n} \left(X_{n-k,n}\right)^{1/\hat{\gamma}_k^H} x^{-1/\hat{\gamma}_k^H}, x \to \infty.$$

2.2.2 Estimateur de la fonction de survie sous données censurées

Estimateur de Kaplan-Meier

Dans la littérature plusieurs auteurs se sont intéressés par l'estimation de la fonction de survie dans le cas des données censurées. Cette fonction peut être estimé grâce à plusieurs méthodes non paramétriques dont la plus intéressante est celle de Kaplan-Meier (1958) [16]. Cet estimateur est aussi appelé un estimateur à limite produit (P-L estimateur), car il s'obtient comme un produit.

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste en avant t et ne pas mourir au temps t. Si $0 < t_1 < \cdots < t_k$, où $1 \leq k \leq n$.

$$\begin{aligned} P\left(X > t_k\right) &= P\left(X > t_k, X > t_{k-1}\right) \\ &= P\left(X > t_k \mid X > t_{k-1}\right) \times P\left(X > t_{k-1}\right) \\ &= \vdots \\ &= P\left(X > t_k \mid X > t_{k-1}\right) \times \ldots \times P\left(X > t_2 \mid X > t_1\right) \times P\left(X > t_1\right). \end{aligned}$$

On considère les temps d'événements (décès et censure) distincts $Z_{i,n}$ $(i = 1, \ldots, n)$ ordonnés par ordre croissant, on obtient :

$$P\left(X > Z_{i,n}\right) = \prod_{k=1}^{i} P\left(X > Z_{k,n} \mid X > Z_{k-1,n}\right), \ i = 1, \ldots, n,$$

avec $Z_{0,n} = 0$. Considérons les notations suivantes :

- n_i le nombre d'individus à risque de subir l'événement juste avant le temps $Z_{i,n}$,

- d_i le nombre de décès en $Z_{i,n}$.

Alors la probabilité p_i de mourir dans l'intervalle $]Z_{i-1,n}, Z_{i,n}[$ sachant que l'on était vivant en $Z_{i-1,n}$, (c'est à dire $p_i = P\left(X \leq Z_{i,n} \mid X > Z_{i-1,n}\right)$, peut être estimé par

$$\hat{p}_i := \frac{d_i}{n_i}.$$

Comme les temps d'événements sont supposés distincts, on a :

- $d_i = 0$ en cas de censure en $Z_{i,n}$, c'est-à-dire quand $\delta_i = 0$,

- $d_i = 0$ en cas de décès en $Z_{i,n}$, c'est-à-dire quand $\delta_i = 1$.

Si on désigne par $\delta_{[i,n]}$ le concomitant de $Z_{i,n}$ ou les indicateurs de censure (c'est-à-dire, $\delta_{[i,n]} = \delta_j$ si $Z_{i,n} = Z_j$ avec $j = \overline{1,n}$), alors :

$$\begin{aligned} \delta_{[i,n]} &= 0 \quad \text{en cas de censure en } Z_{i,n}, \\ \delta_{[i,n]} &= 1 \quad \text{en cas de décès en } Z_{i,n}. \end{aligned}$$

On obtient alors l'estimateur de Kaplan-Meier :

$$\bar{F}^{KM}(t) := \prod_{\substack{i=1,\ldots,n \\ Z_{i,n} \leq t}} \left(1 - \frac{\delta_{[i,n]}}{n_i}\right) = \prod_{i:Z_{i,n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n-i+1}\right) = \prod_{i:Z_{i,n} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{[i,n]}},$$

Soit $(Z_i, \delta_i)_{1 \leq i \leq n}$ l'échantillon réellement observé défini par (1.32) et soit $\left(Z_{(i,n)}, \delta_{(i,n)}\right)_{1 \leq i \leq n}$ les statistiques d'ordre lui associées. L'estimateur Kaplan-Meier est défini par :

$$\begin{aligned} \bar{F}^{KM}(t) := 1 - \hat{F}_n(x) &= \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)} \mathrm{I}\left\{Z_{(i)} \leq t\right\}} \\ &= \prod_{i=1}^{n} \left[1 - \frac{\delta_{(i)} \mathrm{I}\left\{Z_{(i)} \leq t\right\}}{n-i+1}\right]. \end{aligned}$$

Remarque 2.2.1

· *L'estimateur de Kaplan-Meier est une fonction étagé avec des sauts seulement aux observations non-censurées.*

· *La hauteur des sauts de cet estimateur est aléatoire.*

· *Quand toutes les observations sont non-censurées alors on obtient la fonction de répartition empirique.*

Proposition 2.2.2 (Propriétés asymptotiques de $\bar{F}^{KM}(t)$) .

(i) ***Biais*** *: L'estimateur de Kaplan-Meier de la fonction de survie a un biais positif :*

$$\mathbb{E}\left[\overline{F}^{KM}(t) - \overline{F}(t)\right] \geq 0.$$

Il est asymptotiquement non biaisé :

$$\mathbb{E}\left[\overline{F}^{KM}(t)\right] \underset{n\to\infty}{\to} \overline{F}(t).$$

(ii) ***Convergence uniforme :*** *Soit* $x_H = H^{-1}(1) := \inf\{t : H(t) = 1\} \leq \infty$, *($x_H$ la borne supérieure du support de H). Alors*

$$\sup_{0\leq t<x_H} \left|\bar{F}^{KM}(t) - \bar{F}(t)\right| \overset{P.s}{\to} 0, \text{ quand } n \to \infty.$$

(iii) ***Estimation de la variance :*** *Cet estimateur donné par Greenwood (*1926*), de la forme suivante :*

$$\widehat{Var}\left(\overline{F}^{KM}(t)\right) = \left(\overline{F}^{KM}(t)\right)^2 \sum_{\substack{i=\overline{1,n} \\ Z_{i,n}\leq t}} \frac{d_i}{n_i(n_i - d_i)}.$$

(iv) ***Normalité asymptotique :*** *Si la fonction de répartition, de la survie et de la*

censure n'ont aucune discontinuité commune, alors pour tout $t \geq 0$*, on a :*

$$\sqrt{n}\left(\overline{F}^{KM}(t) - \overline{F}(t)\right) \xrightarrow{d} \mathcal{N}(0, \nu^2(t)),$$

où :

$$\nu^2(t) := -\overline{F}(t)^2 \int_0^t \frac{\overline{F}(ds)}{\overline{F}^2(s)\overline{G}(s)}.$$

(v) ***Intervalle de confiance :*** *L'intervalle de confiance de* $\overline{F}^{KM}(t)$*, pour un niveau de signification* $\alpha \in]0,1[$*, est donné comme suit :*

$$IC(\alpha) := \overline{F}^{KM}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\overline{F}^{KM}(t)\right)},$$

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

2.3 Estimation des quantiles extrêmes

2.3.1 Estimation des quantiles extrêmes sous données complètes

On désire estimer x_p le quantile d'ordre $(1-p)$ quand p est petit. Si la fonction de répartition F est continue strictement croissante, cela revient à résoudre l'équation $F(x_p) = 1-p$.

Définition 2.3.1

Soit $X_1, \ldots, X_n$*,* n *v.a's iid de fonction de répartition commune* F*. On définit le quantile extrême par :*

$$x_p := F^{-1}(1-p).$$

avec $p \to 0$*,* $np \to m$ *pour* $n \to \infty$ *et* $m > 0$*.*

Estimation basée sur la GEV

Pour obtenir les estimateurs des quantiles extrêmes du GEV, il suffit d'inverser la fonction $H_{\mu,\sigma,\gamma}$ donnée par 1.23. Ils se présentent comme suit :

$$\begin{aligned}\hat{x}_p &= H^{-1}_{\hat{\mu},\hat{\sigma},\hat{\gamma}}(1-p) \\ &= \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\gamma}}\left\{1 - [-\log(1-p)]^{-\hat{\gamma}}\right\} & \text{si} \quad \gamma \neq 0. \\ \hat{\mu} - \hat{\sigma}\log\{-\log(1-p)\} & \text{si} \quad \gamma = 0. \end{cases}\end{aligned}$$

Supposons que $F \in D(H_\gamma)$. Dans le cas où le quantile extrême est à l'intérieur des données (i.e. $p \geq 1/n$). Il peut être estimé par

$$\hat{x}_p := \hat{a}_n \frac{(np)^{-\hat{\gamma}} - 1}{\hat{\gamma}} + \hat{b}_n. \tag{2.11}$$

où $\hat{\gamma}$, $\hat{a}_n$ et $\hat{b}_n$ sont des estimateurs appropriés (basés sur les k plus grande statistiques d'ordre) de l'indice de la queue de distribution, et de constantes de normalisation a_n et b_n (resp.)

Pour x suffisamment grand, le cas le plus typique où le quantile extrême est hors des données (i.e. $p < 1/n$). Il est estimé par

$$\hat{x}_p := \hat{a}_{n/k} \frac{(np/k)^{-\hat{\gamma}} - 1}{\hat{\gamma}} + \hat{b}_{n/k}. \tag{2.12}$$

On suppose que $\gamma > 0$. L'estimateur du quantile extrême, associé à l'estimateur de Hill (2.5), est donnée par :

$$\hat{x}_p^H := X_{n-k,n}\left(\frac{k}{np}\right)^{\hat{\gamma}_k^H}, \tag{2.13}$$

où $\hat{b}_{n/k} = \hat{a}_{n/k}/\hat{\gamma}_k^H = X_{n-k,n}$.

Estimation basée sur la GPD

Les quantiles aux ordres élevés au-dessus du seuil u ($x_p > u$) sont estimés en inversant l'expression de l'estimateur du GPD donnée dans (1.28)

$$\hat{x}_p := u + \frac{\hat{\sigma}_u}{\hat{\gamma}_u}\left(\left(\frac{N_u}{np}\right)^{\hat{\gamma}_u} - 1\right),\ p < \frac{N_u}{n}, \tag{2.14}$$

avec $\hat{\sigma}_u$ et $\hat{\gamma}_u$, les estimateurs des paramètres de la loi GPD et N_u, le nombre d'excès.

Cette expression figure dans Embrechts et al. (1997, p.354) [10].

Le seuil u est souvent choisi égal à une des statistiques d'ordre $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$. Si l'on choisit comme seuil $u = X_{n-k,n}$ la $(k+1)$-ième plus grande observation, alors $N_u = k$ et l'estimateur des quantiles aux ordres élevés se réécrit de la manière suivante

$$\hat{x}_p^{POT} := X_{n-k,n} + \frac{\hat{\sigma}^{POT}}{\hat{\gamma}^{POT}}\left(\left(\frac{k}{np}\right)^{\hat{\gamma}^{POT}} - 1\right),\ p < \frac{k}{n}, \tag{2.15}$$

où $\hat{\gamma}^{POT}$, $\hat{\sigma}^{POT}$ sont les estimateurs résultants de γ et σ respectivement.

2.3.2 Estimation des quantiles extrêmes sous données censurées

Le principe de l'estimation des quantiles extrêmes x_p d'ordre $(1-s)$ en présence de la censure aléatoire est disponible dans la littérature. Il a été proposé par Beirlant et al [2] en (2007) et par Einmahl et al [9] en (2008). Il est définit par l'expression suivante :

$$\hat{x}_s^c := Z_{n-k,n} + \hat{a}^{(c,.)}\frac{\left(\left(1 - \bar{F}^{KM}\left(Z_{n-k,n}\right)\right)/s\right)^{\hat{\gamma}_1^c} - 1}{\hat{\gamma}_1^{(c,.)}}.$$

où $\hat{a}^c := Z_{n-k,n}\hat{\gamma}^H\left(1 - D_n\right)/\hat{p}$, avec

$$D_n := 1 - \frac{1}{2}\left(1 - \frac{\left(\hat{\gamma}^H\right)^2}{H_k^{(2)}}\right)^{-1},$$

et

$$H_k^{(2)} := \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)^2 .$$

2.4 Application sous R

2.4.1 Estimateur de Kaplan-Meier sous des données simulées

On génère un échantillon de v.a X, qui représente la durée de survie, issu d'une loi uniforme standard, censuré à droite par un autre échantillon de v.a C uniformément distribué sur $[0, 0.8]$, qu'il s'agit de la durée de censure. Les deux échantillons ont la même taille $n = 25$.

Les résultats numériques de l'estimation de la fonction de survie sont résumés dans le tableau 2.1, où la $1^{ère}$ colonne contient les six durées réellement observées, parmi les 25 individus qui sont analysés, ce qui est équivalent à 76% de censure. La $2^{ème}$ colonne indique le nombre d'individu n_i à risque sur l'intervalle de temps écoulé. Le nombre d'évènements observé d_i est indiqué dans la $3^{ème}$ colonne. Dans la $5^{ème}$ et $6^{ème}$ colonne on donne l'erreur standard $\hat{\sigma}$ et l'intervalle de confiance à 95% respectivement, associée à l'estimateur de la fonction de survie de Kaplan-Meier pour chaque t_i.

t_i	n_i	d_i	$\bar{F}^{KM}$	$\hat{\sigma}$	IC (95%)
0.0289	24	1	0.958	0.0408	$0.882 - 1.00$
0.1293	19	1	0.908	0.0625	$0.793 - 1.00$
0.2824	13	1	0.838	0.0885	$0.681 - 1.00$
0.3482	11	1	0.762	0.1084	$0.576 - 1.00$
0.4306	10	1	0.686	0.1214	$0.485 - 0.97$
0.7521	3	1	0.457	0.2034	$0.191 - 1.00$

TAB. 2.1 – Résultats relatifes aux données simulées de 25 observations uniformes standards censurées par une variable uniforme sur [0,0.8]

Les résultats obtenus de l'estimateur de Kaplan-Meier en fonction de la durée de survie sont illustrés dans la figure 2.4. Il est clair que la courbe de cet estimateur est en escalier décroisant. L'intervalle de confiance au niveau 95% de la survie est également représenté sur cette courbe. On peut aussi observer 19 valeurs, à travers les petits tirets verticaux, qui représentent les durées censurées.

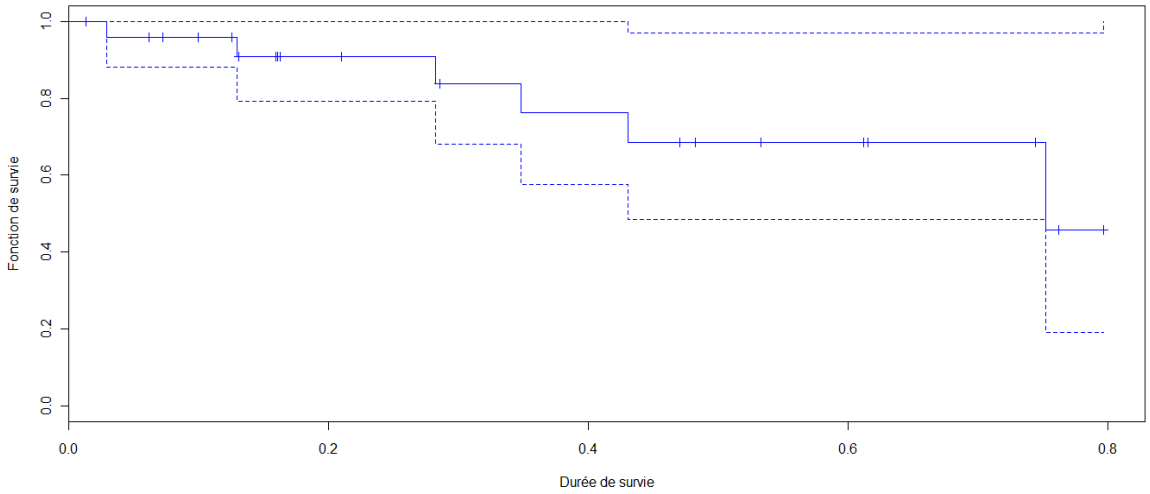

FIG. 2.4 – Estimateur de Kaplan-Meier (ligne continue) et bornes de confiance (lignes en tirets) de la fonction de survie sous données simulées.

2.4.2 Estimation de la fonction de survie d'une distribution à queue lourde

Soit X une v.a de la loi de $Burr$ de paramètre $\alpha, \rho, \tau > 0$, notée $Burr(\alpha, \rho, \tau)$ et de fonction de répartition :

$$F(x) = 1 - \left(\frac{\alpha}{\alpha + x^{\rho}}\right)^{\tau}, \text{ avec } \gamma = \frac{1}{\tau\rho}.$$

On génère dans ce cas $r = 200$ échantillons de v.a X de taille $n = 2000$ observations, issus d'une loi de $Burr(1, 4, 1/4\gamma_1)$ pour $\gamma_1 = 0.7$, censurées par une autre v.a C de

$Burr\,(1, 4, 1/4\gamma_2)$ où $\gamma_2 = \dfrac{p\gamma_1}{(1-p)}$ avec $p = 0.5$ est la proportion de données observées dans la queue de distribution. Les résultats graphiques et numériques finaux sont obtenus en faisant les moyennes sur les 200 réplications.

La figure 2.5 représente l'estimateur de Hill adapté de γ_1, définit par (2.8) en fonctions de statistiques d'ordre k. Cette figure, montre que l'estimateur de Hill adapté présente une allure très proche de la valeur réelle $\hat{\gamma}_1^c$. Cet estimateur est très stable à partir d'une certaine valeur qui représente le nombre optimal de statistiques d'ordre extrêmes.

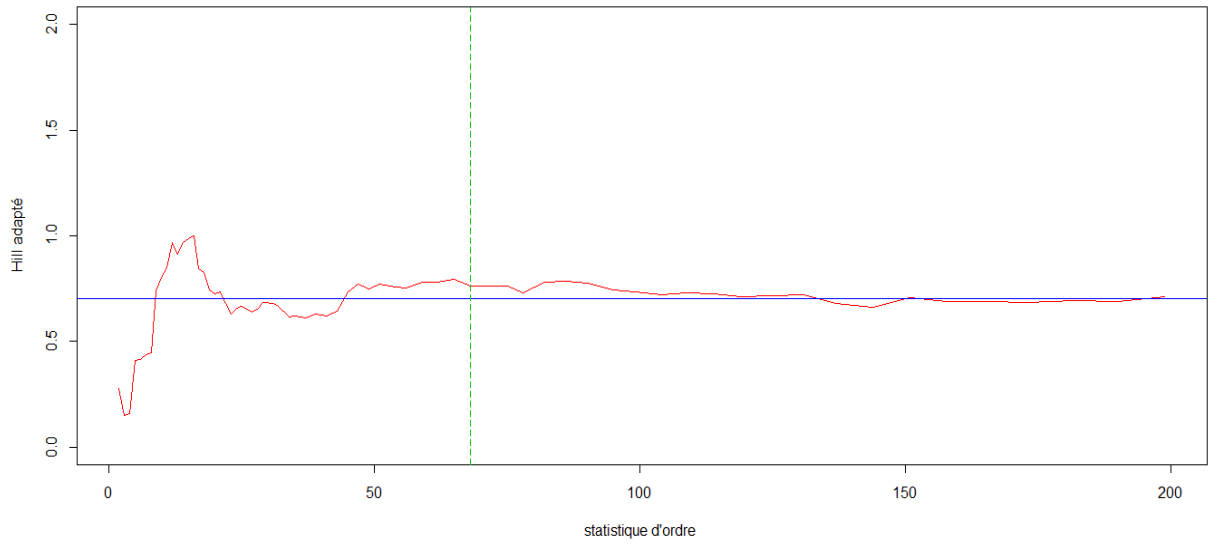

FIG. 2.5 – Estimateur de Hill adapté issue de 200 échantillons de taille 2000 de la loi de $Burr$ pour $(\gamma_1 = 0.7)$ censurée par une variable de $Burr$ pour γ_2, avec $p = 0.5$. La ligne horizontale représente la vraie valeur de γ_1 et la ligne verticale représente le nombre optimal de la statistique d'ordre supérieure.

Pour les résultats numériques, on commence par la détermination du nombre de statistique d'ordre extrême utilisé dans le calcul de l'estimateur de Hill adapté au cas de censure $\hat{\gamma}_1^c$, on applique l'algorithme de Reiss et Thomas, qui consiste à minimiser la quatité

$$\frac{1}{k}\sum_{i\leq k} i^{\beta} \mid \hat{\gamma}_1^c(i) - med(\hat{\gamma}_1^c(1), ..., \hat{\gamma}_1^c(k)) \mid, 0 \leq \beta \leq 1/2,$$

où *med* dénote la médiane.

Les valeurs du biais et la racine de l'erreur moyenne quadratique (rmse : root of the mean squared error) sont calculées respectivement par :

$$biais := \frac{1}{r}\sum_{i=1}^{k}\left(\widehat{\bar{F}}_i(t) - \bar{F}(t)\right) \quad \text{et} \quad rmse := \sqrt{\frac{1}{r}\sum_{i=1}^{k}\left(\widehat{\bar{F}}_i(t) - \bar{F}(t)\right)^2}.$$

Les résultats obtenus pour l'estimateur de la fonction de survie de Kaplan-Meier à des instants distincts t sont résumés dans le tableau 2.2.

	$p =$	0.5			
	t	$\bar{F}$	$\widehat{\bar{F}}$	$biais$	$rmse$
$\gamma_1 = 0.7$	10	0.0373	0.0696	0.0323	0.0392
	20	0.0138	0.0119	−0.0019	0.0112
	35	0.0062	0.0063	0.0001	0.0060
	60	0.0029	0.0034	0.0005	0.0033
	95	0.0015	0.0021	0.0006	0.0020
	140	0.0009	0.0013	0.0004	0.0014

TAB. 2.2 – Résultats de simulation obtenues sur la base de 200 échantillons de taille2000 de la loi de Burr pour γ_1 censurée par une variable de Burr pour γ_2

Conclusion

L'estimation de l'indice des valeurs extrême, la fonction de survie et les quantiles extrêmes pour des données incomplètes (ou bien qui ne sont pas totalement observées) a fait l'objet de plusieurs publications ces dernières année, du fait que beaucoup de quantités statistiques sont exprimées en terme de ces statistique.

Ce travail est consacré à la présentation des différentes méthodes existent dans la littérature pour estimer ces statistique, en particulier à la présence de données censurées.

Dans un premier temps, nous avons énoncés les principaux résultats de la théorie des valeurs extrêmes, et les fondamentales notions de censure, avec un accent particulier sur la censure aléatoire à droite.

Nous avons ensuite intéressés à, une adaptation de l'estimateur de Hill en présence de censure proposé par Beirlant et al (2007) [2], l'estimateur de Kaplan-Meier pour estimer la fonction de survie et une estimation des quantiles extrêmes basée sur l'estimateur de Hill.

A la fin, nous avons appliqué les principaux résultats des estimateurs sur des données simulés, issues des lois usuelles et à queue lourde.

Bibliographie

[1] Beirlant, J. , Vynckier, P., and Teugels, J. (1996). Tail index estimation, pareto quantile plots regression diagnostics. Journal of the American Statistical Association. **91** (436), 1659 – 1667.

[2] Beirlant, J., Guillou, A. Dierckx,G.,and Fils-Villetard, A. (2007) . Estimation of the extreme value index and extreme quantiles under random censoring. Extremes, **10**(3), 151 – 174.

[3] Belkema, A. A. ,and De Haan, L. (1974). Residual life time at great age. Ann. Probab,792 – 804.

[4] Benameur, S. (2010). Sur l'estimateur de l'indice des valeurs extrêmes. Mémoire de Magistère, de l'université Mohamed Khider,Biskra.

[5] Boualam,K (2017). Etude de l'estimateur de Hill sous dépendance faible. Thése de doctorat, de l'université Mouloud Mammeri, Tizi-ouzou.

[6] Davis, R., and Resnick, S. (1984) . Tail estimes motivated by extreme value theory. The Annals of Statistics, 1467 – 1487.

[7] Deheuvels, P., Haeusler, E., and Mason, D. M. (1988). Almost sure convergence of the Hill estimator. Math. Proc. Cambridge Philos. Soc, **104** (02), 371 – 381.

[8] Dekkers, A. L., and De Haan, L. (1989). On the estimation of the extreme value index and large quantile estimation. Ann. Statist, 1795 – 1832.

[9] Einmahl, J.H.J., Fils-Villetard, A., Guillou. (2008) Statistics of extremes under random censoring ; Bernoulli.

[10] Embrechts, P. , Kluppelberg, C. , and Mokosch, T. (1997). Modelling extremal avents for insurance and finance. Springer-Verlag. Berlin. .

[11] Ficher, R.T., and Tippett, L.H.C.(1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc.Cambridge Philos. Soc, **24** (02),180 − 190.

[12] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math, 423 − 453.

[13] de Haan,L. (1976). Sample extremes : an Introduction. Springer-Verlag,New York.

[14] Hill, B. M. (1975) . A simple general approach to inference about the tail of a distribution. Ann. Statist, **3**(5), 1163 − 1174.

[15] Jankinson, A. F.(1955). The Frequency Distribution of the Annual Maximum (or Minimum) of the Meteorological Elements Quarterly Jornal ofthe Royel Meteorological society **81**, 185 − 171.

[16] Kaplan, E.L, Meier, P. (1958). Nonparamertic estimation from incomplete observations. Jornal of American Statistical Association and, **53 :**457 − 481.

[17] Mason, D. M. (1982). Laws of large numbers for sums of extreme values. Ann. Proba, **754 − 764**.

[18] Meraghani, D (2016) .Modelling Distribution tails. Thèse de doctorat de l'université Mohamed Khider, Biskra.

[19] Pathé, N. Modélisation de valeurs extrêmes conditionnelles en présence de censure.Thèse de doctorat, de université Gaston Berger de Saint-Louis du Senegal

[20] .Pickands III, J.(1975). Statistical inference using extreme order statistics. Ann. Statist, 119 − 131.

[21] Reiss,R.D, and Thomas,M. (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhauser, Basel.

[22] Soltane, L (2016). Analyse des valeurs extrêmes en présence de censure. Thèse de doctorat, de l'université Mohamed Khider, Biskra, Algéria.

Annexe : Abréviations et Notations

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

S	Fonction de survie $\left(\overline{F}\right)$.
Q	Fonction de quantile.
Q_n	Fonction de quantile empirique.
F^{-1}	Inverse généralisé de F.
$\mathbb{1}_A$	Fonction indicatrice de l'ensemble A.
$v.a$	variable aléatoire.
iid	Indépendantes et identiquements distribuées.
$\underrightarrow{P}$	convergence en probabilité.
$\underrightarrow{P.s}$	Convergence presque sur.
$\binom{n}{i}$	Combinaison.
H_γ	Famille de laloi de valeurs extrêmes généralisées.
RV_ρ	Variation régulière au ∞ avec l'indice ρ.
L	Fonction à variation lente.
POT	Pearks over Threshold.
TCL	Théorème Centrale Limite.
F	Fonction de répartition.

F_n	Fonction de répartition empirique.
$E[X]$	Espérence mathématique de X.
σ^2	Varience d'une variable aléatoire.
EMV	Estimateur de Maximum de Vraisemblance.
$D(.)$	Domaine d'attraction.
GEV	Distribution des valeurs extrêmes généralisées.
GPD	Distribution de Paréto généralisée.
EVI	Indice des valeurs extrêmes.
μ	Espérence ou moyenne d'une v.a.
$\mathbb{R}$	Ensemble des valeurs réelles.
TEV	Théorie des valeurs extrêmes.
u	Seuil.
$X_{n,n}$	Maximum de $(X_1, \ldots, X_n)$.
$X_{1,n}$	Minimum de $(X_1, \ldots, X_n)$.
x_F	Point terminal.
$:=$	Egalité par définition.
$\hat{\gamma}^H$	Estimateur deHill.
$\hat{\gamma}^p$	Estimateur de Pickands.
$\hat{\gamma}^{adH}$	Estimateur de Hill adapté.
$\mathcal{N}(0,1)$	Loi normale standard.
(Ω, A, P)	Espace probabilité.
$(X_1, \ldots, X_n)$	Un échantillon.
$X_1, \ldots, X_n$	Une suite de n variable aléatoire.
$L(\theta; X_1, X_2, \ldots, X_n)$	Fonction de vraisemblance.