UNIVERSITE KASDI MERBAH OUARGLA

Faculté des mathématiques et sciences de la matière

N° d'ordre :
N° de série :

DEPARTEMENT DE MATHEMATIQUES

MASTER

Spécialité : Mathématiques

Option : Probabilité et Statistique

Par : DEGACHI Ahmed Taki Eddine

Théme

Comparaison des estimateur de l'indice des valeurs extrêmes sous données censurées

Soutenu publiquement le : 29/05/2017

Devant le jury composé de :

Mr. AGTI Mohamed	M.A. université de KASDI Merbah - Ouargla	Président
Mr. ZIBAR Said	M.A. université de KASDI Merbah - Ouargla	Examinateur
M. MEDDI Fatima	M.C. université de KASDI Merbah - Ouargla	Rapporteur

Dédication

Je dédie ce travail à :

Mes parents

-A mes frères

et mes sœurs,et toute la famille DEGACHI

- 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

Remerciement

Je tiens tout d'abord à remercier Allah le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce modeste travail. En second lieu, je tiens à remercier mon encadreur M MEDDI Fatima, pour ses précieux conseils et son aide durant toute la période du travail. Mes vifs remerciements vont également aux membres du jury : "AGTI Mohamed", "ZIBAR Mohamed Said" pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions. Mes remerciements s'étendent également à tous mes enseignants durant les années d'études. Ma famille et mes amis qui par leurs prières et leurs encouragements, on a pu surmonter tous les obstacles. Enfin, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Table des matières

Introduction générale

L'analyse des données de survie voit le jour au XVIIe siècle, dans le domaine de la démographie. L'objectif des analystes de ce siècle est l'estimation, à partir des registres de décès, de diverses caractéristiques de la population son effectif, sa longévité, etc. Ces analyses, très générales, ne sont affinées qu'à partir du XIXe siècle, avec l'apparition de catégorisations suivant des variables exogènes (sexe, nationalité, catégories socio-professionnelles...). Durant ce siècle, apparaissent également les premières modélisations concernant la probabilité de mourir a un certain âgé, probabilité qui sera par la suite désignée sous le terme de fonction de risque . Enfin, l'analyse des données de survie commence de déborder le cadre stricte de la démographie pour investir, au xxe siècle, toutes les disciplines susceptibles d'avoir recours à de tels types de données: l'actuariat, la physique (avec l'apparition de la théorie de la fiabilité), l'industrie (pharmaceutique, biomédicale)...

Jusqu'en (1950) , *la communauté des statisticiens s'intéresse peu à l'analyse des données de survie, la principale contribution étant celle de Greenwood* (1926), *qui propose une formule pour l'erreur standard d'une table de survie.*

En 1951, Weibull conçoit un modèle paramétrique dans le domaine de la fiabilité à cet effet, il fournit une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de la survie: la loi de Weibull .

En (1958), *Kaplan et Meier présentent d'importants résultats concernant l'estimation non paramétrique de la fonction de survie: de l'estimateur résultant, ils étudient l'espérance, la variance et les propriétés asymptotiques.*

La théorie de la valeur extrême est unique comme une discipline statistique en ce qu'elle développe des techniques et des modèles pour décrire l'insolite plutôt que l'habituel. Comme une étude abstraite des phénomènes aléatoires, le sujet peut être retracé à la première partie

du 20 siècle. Ce n'est que les années (1950) que la méthodologie a été proposée de façon sérieuse pour la modélisation de véritables phénomènes physiques. Ce n'est pas un hasard si les premières applications de modèles extrêmes de valeur étaient principalement dans le domaine de l'ingénierie civile: les ingénieurs avaient toujours été tenus de concevoir leurs structures afin qu'ils résistent aux forces qui pourraient raisonnablement s'attendre à des répercussions sur eux. 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 Beirtant et al. L'étude a continué jusqu'en 2008, où Einmahl et al, il a présenté l'indice est de diviser l'indicateur de Hill* (1975) *sur* $\hat{p}$ *où* $\hat{p}$ *représent la proportion estimeé de l'estimateur des données. Warms et Warms a été atteint en* (2013) *pour créer un nouvel indice basé sur l'intégral des Kaplen -Meier.*

la comparaison des estimateur de l'indice des valeur extrêmes sous données censurées *ont un grand intérêt dans plusieurs domaines mathématiques, dans ce mémoire nous verrons quelques applications de ces derniers, Notre sujet se divise en trois chapitres:*

$1^{ère}$ *chapitre: Présentation de la théorie des valeurs extrêmes univariées, on regroupe des définitions et des résultats sur cette dernière. Après avoir introduit le comportement du maximum , on présente les deux principaux outils servant à modéliser le comportement des valeurs extrêmes d'un échantillon: la loi des valeurs extrêmes et la loi des excès . On s'intéressera ensuite à la caractérisation des domaines d'attraction. Enfin, on rappele les différentes méthodes d'estimation de quantiles extrêmes. Dans la dernière partie on présentera les données censurées.*

$2^{ème}$ *chapitre: Définitions et caractéristiques de deux estimateurs, nous présentons brièvement quelques estimateurs de l'indice des valeurs extrêmes en présence de censure aléatoire à droite . Ce chapitre n'est pas un panorama exhaustif de tous les travaux faits et publiés dans cette thématique. Nous nous focalisons essentiellement sur les estimateurs suivants : l'estimateur de Hill de l'indice des valeurs extrêmes Einmahl et al* (2008) *, et l'estimateur l'indice des valeurs extrêmes l'intégration de Kaplan-MeierWorms ,J., Worms, R., 2013.*

Notations et conventions

F	Fonction de répartition
F_n	Fonction de répartition empirique
$F^{\leftarrow}$	Inverse généralisé de F
EVD	Distribution des valeurs extrêmes
EVI, γ	Indice des valeurs extrêmes
GEV	Distribution des valeurs extrêmes généralisée
GPD	Distribution de pareto généralisée
G_γ	Famille de la loi de valeurs extrêmes généralisée
$i.i.d$	Indépendantes et identiquement distribuées.
$\mathbb{I}_{\{A\}}$	Fonction indicatrice de l'ensemble A
$\ell(x)$	Fonction à variation lente
DA	Domaine d'attraction de maxcimum
$M_n = X_{n,n}$	Maxcimum de $X_1, ..., X_n$
N_u	Nombres des excés qui dépassent du seuil u
POT	Pics au-delà d'un seuil
$p.s$	Prèsque sûre
Λ	Loi de Gumbel
Φ	Loi de frêchet
Ψ	Loi de weibull

$resp$	Respectivement
$S = \overline{F}$	$1 - F$ fonction de survie
TEV	Théorèm des valeurs extrêmes
$X_{1:n}, ..., X_{n,n}$	Statistique d'ordre associées à $X_1, ..., X_n$
$X \wedge Y$	$\min(X, Y)$
x_F	Point terminal
$\underset{=}{\mathcal{L}}$	Égalité en loi
$:=$	Égalité en définition
$\underset{\rightarrow}{D}$	Converge en distribution
$\underset{\rightarrow}{l}$	Converge en loi
$\underset{\rightarrow}{p}$	Converge en probabilité
$\underset{\rightarrow}{p.s}$	Converge prèsque sûre
$o_P(.)$	Converge vers 0 en probabilité
VR_α	Variation régulière d'indice α
τ_H	Point terminal
$\sup A$	Supremum de l'ensemble A
$Z^* = (Z_1^*, ..., Z_n^*)$	Échantillon Bootstrap
$\hat{\gamma}_X^c$	Estimateur de Hill avec les données censurées
$\hat{\gamma}_{X,Hill}^{KM}$	Estimateur de Hill par l'intégration de Kaplan-Meier
SD	Erreur standard
IC	Intervalle de confiance
MSE	L'erreur quadratique moyenne

Liste des Figures

Liste des Tables

Chapitre 1

Présentation de la théorie des valeurs extrêmes univariées

1.1 Comportement du maximum d'un échantillon

1.1.1 Densités et loi des statistiques d'ordre

Définition 1.1.1 *(les statistiques d'ordres) Soit une suite finie d'observations iid X_i, $i \in [1, n]$, classées par ordre croissant.On écrit cette suite d'observations sous la notation $X_{i,n}$ avec,*

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

$X_{i,n}$ est donc la i-i'eme statistique d'ordre (ou statistique d'ordre i) dans un échantillon de taille n. $X_{1,n}$ la plus petite valeur observée (ou statistique du minimum, $X_{1,n} = \min(X_1,..., X_n)$) et la plus grande statistique d'ordre $X_{n,n}$ (ou statistique du maximum, $X_{n,n} = \max(X_1, ..., X_n)$)

David [1970] et Balakrishnan et Clifford Cohen [1991] montrent que l'expression de la distribution de $X_{i,n}$ est

$$F_{i,n} = \mathbb{P}\{X_{i,n} \leq x\} = \sum_{r=i}^{n} \binom{n}{r} (F(x))^r (1 - F(x))^{n-r}. \tag{1.1.1}$$

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

$$f_{i,n}(x) = \frac{n!}{(i-1)!(n-i)!} [F(n)]^{i-1} [1 - F(x)]^{n-i} f(x), \tag{1.1.2}$$

Pour les statistiques d'ordre extrêmes, on obtient les expressions suivantes :

$$F_{1,n}(x) = \mathbb{P}\{X_{1,n} \leq x\} = 1 - (1 - F(x))^n, \tag{1.1.3}$$

d'où,

$$f_{1,n}(x) = nf(x)(1 - F(x))^{n-1}, \tag{1.1.4}$$

pour la statistique du minimum et

$$F_{n,n}(x) = \mathbb{P}\{X_{n,n} \leq x\} = (F(x))^n, \tag{1.1.5}$$

d'où,

$$f_{n,n}(x) = nf(x)(F(x))^{n-1}. \tag{1.1.6}$$

Les expressions de $F_{1:n}$ et $F_{n:n}$ peuvent s'obtenir trés facilement en considérant les relations

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

et,

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

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

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

et,

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

Définition 1.1.2 *(La fonction de répartition empirique).Soit* $(X_1, ..., X_n)$ *une suite des v.a.r. i.i.d. de distribution* F *inconnue,*

$$F_{i,n} = \mathbb{P}\{X_{i,n} \leq x\}$$

Pour chaque entier $n \geq 1$*, on a* $X_{1,n} \leq X_{2,n} \leq ..... \leq X_{n,n}$ *la statistique d'ordre associée à l'échantillon* $(X_1, ..., X_n)$*. On note par:*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{]-\infty,x[}(X_i), \quad x \in \mathbb{R} \tag{1.1.7}$$

il existe une autre version de la définition de F_n *en utilisant les tatistiques d'ordres comme suit:*

$$F_n(x) = \begin{cases} 0 & si, \quad x \leq X_{1,n} \\ \frac{i-1}{n} & si, \quad X_{i-1,n} < x \leq X_{i,n}, \quad 2 \leq i < n \\ 1 & si, \quad x > X_{n,n} \end{cases} \tag{1.1.8}$$

Définition 1.1.3 *On définit la fonction des quantiles* Q *par :*

$$Q(s) = F^{\leftarrow}(s) := \inf\{x \in \mathbb{R} : F(x) \geq s\}\ , \qquad 0 < s < 1 \tag{1.1.9}$$

Où $F^{\leftarrow}$ *est l'inverse généralisée de la fonction de distribution* F*.*

On définit la fonction des quantiles du queue U *par:*

$$U(t) := Q(1 - 1/t) = (1/\overline{F})^{\leftarrow}(t), \quad 1 < t < \infty \tag{1.1.10}$$

1.2 Loi des valeurs extrêmes

Théorème 1.2.1 *(Fisher-Tippet* (1928)*). Supposons n variables aléatoires* X_i , $i \in [1, n]$ *indépendantes et de même loi de distribution F et* $M_n = \max(X_i)_{1 \leq i \leq n}$. *S'il existe des constantes* a_n *et* b_n *et une distribution limite non-dégénérée* G_γ *telles que,*

$$\lim_{n \to \infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n \to \infty} F^n(a_n x + b_n) = G_\gamma(x), \qquad \forall x \in \mathbb{R}, \tag{1.2.1}$$

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 alors* G_γ *est l'un des trois types suivants de distribution:*

$$\textit{loi de Gumbel : } G_\gamma(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty \tag{1.2.2}$$

$$\textit{loi de Fréchet : } G_\gamma(x) = \begin{cases} 0 & x < 0 \\ \exp\left(-x^{-1/\gamma}\right) & x \geq 0, \gamma > 0 \end{cases}$$

$$\textit{loi de Weibull : } G_\gamma(x) = \begin{cases} \exp\left(-(-x)^{-1/\gamma}\right) & x < 0, \gamma < 0 \\ 1 & x \geq 0. \end{cases}$$

Ce théoréme présente un intérét important, car si l'ensemble des distributions est 'grand', l'ensemble des distributions de valeurs extrêmes est lui trés petit. Stuart Coles fait le paralléle suivant entre le théoréme de limite centrale et celui de Fisher-Tippet

Remarque 1.2.1 *Pour distinguer les trois distributions, on utilise généralement les notations suivantes:* Λ *pour la distribution Gumbel,* Φ_γ *pour la distribution Frechet et* Ψ_γ *pour la distribution Weibull (Resnick [1987]).*

Remarque 1.2.2 *Les distributions* Λ *,*Φ_γ *et* Ψ_γ *sont appelées les distributions de valeurs extrêmes et les variables aléatoires correspondantes sont les variables aléatoires extrémales. Nous considérons maintenant les différences entre les trois distributions* Λ *,*Φ_γ *et* Ψ_γ. *Sur le* (*Figure* 1.2.1)*, nous représentons les fonctions de densité de* Λ *,*Φ_1 *et* Ψ_1

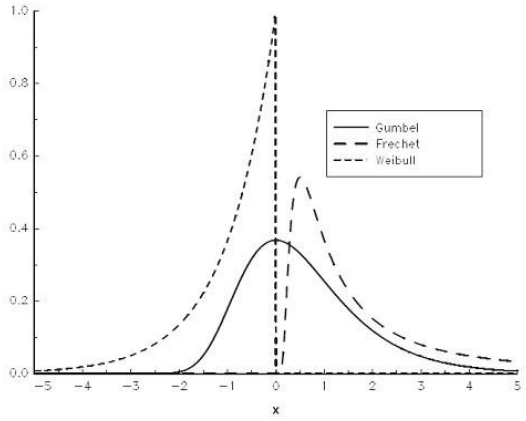

Figure 1.2.1 : les fonctions de densité de Λ ,Φ_1 et Ψ_1

Jenkinson (1955) donne l'expression générale par,

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), \forall x \in \mathbb{R}, 1+\gamma x > 0 \text{ si } \gamma \neq 0 \\ \exp(-\exp(-x)), \quad \forall x \in \mathbb{R}, \qquad \text{si } \gamma = 0 \end{cases} \tag{1.2.3}$$

s'appelle la distribution des valeurs extrêmes généralisées (distribution GEV).

1.3 La loi des excès

L'approche basée sur les distributions GEV peut être réductrice du fait que l'utilisation d'un seul maxima conduit à une perte d'information continue dans les autres grandes valeurs de l'échantillon. La solution est de considérer plusieurs grandes valeurs au lieu de la plus grande. L'approche de la théorie des valeurs extrêmes appelée POT consiste à utiliser les observations qui dépassent un certain seuil, plus particulièrement les différences entre ces observations et le seuil, appelées excès. Il est clair que cette méthode nécessite la détermination d'un seuil ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations. Notons le seuil par u.

Définition 1.3.1 *Soit X une variable aléatoire de fonction de répartition F et de point terminal x_F. Pour tout $u < x_F$, la fonction*

$$F_u(x) = \mathbb{P}(X \leq x \mid X > u) \tag{1.3.1}$$

est appelée fonction de répartition des excès au dessus du seuil u.

$$F_u(x) = \frac{F(u+y) - F(u)}{1 - F(u)} \quad \text{,si } x \geq 0 \text{ et } 0 \text{ sinon} \tag{1.3.2}$$

Nottons $Y = X - u$ pour $X > u$ et pour n v.a. observées $X_1, ..., X_n$, nous pouvons écrire $Y_j = X_i - u$ telle que i est l'indice du j^{ime} excès et $j = 1, ..., N_u$. Le $(Mean\ Excess\ Plot)$ appelé aussi le $(Mean\ Residual\ life\ Plot)$ est un outil spécifique pour retenir le seuil performant. De plus, nous approchons la loi des excès$(Y_1, ..., Y_{Nu})$ par une loi de Pareto Généralisée $GPD_{\gamma,\sigma}$ (Generalized Pareto Distribution) que nous présentons ci-dessous.

1.3.1 La loi de Pareto Généralisée

Le théorème de Pickands est très utile lorsqu'on travaille avec des observations qui dépassent un seuil fixé puisqu'il assure que la loi des excès peut être approchée par une loi de Pareto généralisée.

Définition 1.3.2 *Soient $\sigma(u)$ une fonction strictement positive et $\gamma \in R$. La loi de Pareto généralisée a pour fonction de répartition $G_{\gamma,\sigma}$:*

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

où $y \geq 0$ si $\gamma \geq 0$ et $0 \leq y \leq \frac{-\sigma}{\gamma}$ si $\gamma < 0$.

Théorème 1.3.1 $(Pickands\ 1975)$ *.Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet,Gumbel ouWeibull), alors il existe une fonction $\sigma(u)$ strictement positive et un réel γ tels que*

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

où $G_{\gamma,\sigma(u)}$ est la fonction de répartition de la loi de Pareto Généralisée et F_u est la fonction de répartition des excès au delà du seuil u.

1.4 Domaines d'attraction

1.4.1 Caractérisation des Domaines d'attraction

On définit les notions de fonctions à variations régulières et de fonctions à variations lentes qui nous seront utiles par la suite. Pour plus de détails, se référer à Bingham et al. [2] où de nombreux résultats sur les fonctions à variations régulières sont donnés.

Fonctions à variation régulière

Commençons par rappeler la définition d'une fonction à variations régulières.

Définition 1.4.1 *Une fonction G est dite à variation régulière (à l'infinie) d'indice $\alpha \in R$ si G est positive à l'infini (ie: s'il existe A tel que pour tout $x \geq A$, $G(x) > 0$) et si pour tout $x > 0$,*

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

Dans le cas particulier où $\alpha = 0$, on dit que G est une fonction à variation lente.

En remarquant que si G est à variation régulière d'indice α alors $\dfrac{G(x)}{x^{\alpha}}$ est à variation-lente, il est facile de montrer qu'une fonction à variation régulière d'indice α peut toujours s'écrire sous la forme $t^{\alpha}\ell(x)$, où ℓ est à variation lente.

Théorème 1.4.1 *ℓ est une fonction à variation lente si et seulement si pour tout $x > 0$*

$$\ell(x) = c(x) \exp\left\{\int_0^x \frac{\varepsilon(t)}{t} dt\right\}, \qquad x \geq 0,$$

où c et ϵ sont des fonctions positives telles que $\lim_{x\to\infty} c(x) = c \in]0, \infty[$ et $\lim_{x\to\infty} \varepsilon(x) = 0$

Remarque 1.4.1 *Si la fonction c est constante, on dit que ℓ est normalisée.*

Remarque 1.4.2 *Soit G une fonction à variation régulière d'indice α ($G \in VR_{\alpha}$).en utilisant le fait que $G(x) = x^{\alpha}\ell(x)$, on déduit facilement que pour tout $x > 0$,*

$$G(x) = c(x) \exp\left\{\int_0^x t^{-1}\alpha(t)\right\}, \qquad x \geq 0$$

où c et α sont des fonctions positives telles que $\lim_{x\to\infty} c(x) = c \in]0, \infty[$et $\lim_{x\to\infty} \alpha(x) = \alpha$

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

1. Si $\gamma > 0$, F appartient au domaine d'attraction de Fréchet. Il contient les lois dont la fonction de survie décroit comme une fonction puissance. On parle aussi de lois à queue lourde. Dans ce domaine d'attraction, on trouve les lois de Pareto, de Student, de Cauchy, etc...

2. Si $\gamma = 0$, F est dans le domaine d'attraction de Gumbel qui regroupe les lois ayant une fonction de survie à décroissance exponentielle. C'est le cas des lois normale, gamma, exponentielle, etc...

3. Si $\gamma < 0$, F appartient au domaine d'attraction de Weibull. Ce domaine contient les lois dont le point terminal $x_F = \inf\{x, F(x) \geq 1\}$ est fini. C'est le cas par exemple des lois uniformes, lois beta, etc...

Domaine d'attraction de Fréchet

Théorème 1.4.2 *Une fonction de répartition F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrêmes $\gamma > 0$ si et seulement si la fonction de survie $S \in RV_{-1/\gamma}$. Autrement dit, une fonction de répartition $F(.)$ appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :*

$$F(x) = 1 - x^{-1/\gamma}\ell(x), \quad \ell(x) \in RV_0. \tag{1.4.1}$$

Les suites de normalisation (a_n) et (b_n) sont données dans ce cas par $a_n = F^{-1}\left(1 - \frac{1}{n}\right)$ et $b_n = 0$.

Il faut aussi noter que toutes les fonctions de répartition du domaine d'attraction de Fréchet ont un point terminal infini. que l'équation 1.4.1 est équivalente à

$$Q(\alpha) = \alpha^{-\gamma}\ell\left(\alpha^{-1}\right), \quad \ell(x) \in RV_0, \tag{1.4.2}$$

où a $\alpha \in [0, 1]$. De nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes γ et des quantiles extrêmes $Q(\alpha)$ pour des lois à queue lourde.

Domaine d'attraction de Weibull

Théorème 1.4.3 *F appartient au domaine d'attraction de Weibul, (avec un indice des valeurs extrêmes $\gamma < 0$) si et seulement si son point terminal x_F est fini et si la fonction de répartition $F_*(x)$ définie par*

$$F_*(x) = \begin{cases} 0 & si \quad x \leq 0, \\ F\left(x_F - \frac{1}{x}\right) & si \quad x > 0. \end{cases} \tag{1.4.3}$$

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

Ainsi, une fonction de répartition F du domaine d'attraction de Weibull s'écrit pour $x < x_F$:

$$1 - F = (x_F - x)^{-1/\gamma} \ell \left[(x_F - x)^{-1} \right], \quad \ell(x) \in RV_0. \tag{1.4.4}$$

De manière équivalente, le quantile $Q(\alpha)$ associé s'écrit :

$$Q(\alpha) = x_F - (\alpha)^{-\gamma} \ell \left[\left(\frac{1}{x} \right) \right], \quad \ell(x) \in RV_0.$$

Les suites de normalisation a_n et b_n sont données par $a_n = x_F - F^{-1}\left(1 - \frac{1}{n}\right)$ et $b_n = x_F$.

Domaine d'attraction de Gumbel

Théorème 1.4.4 *Une fonction de répartition* $F(.)$ *appartient au domaine d'attraction de Gumbel si et seulement si il existe* $z < x < \infty$ *telle que,*

$$1 - F(x) = c \exp \left\{ - \int_z^x \frac{1}{a(t)} dt \right\}, \quad z < x < x_F, \tag{1.4.5}$$

où $c(x) \to c > 0$ *lorsque* $x \to x_F$ *et* $a(.)$ *est une fonction positive et dérivable de dérivée* $a'(.)$ *telle que* $a'(.) \to 0$ *lorsque* $x \to x_F$.

Le domaine d'attraction de Gumbel regroupe une grande diversité de lois comptant parmi elles la plupart des lois usuelles (loi normale, exponentielle, gamma, log-normale). Cette famille étant difficile à étudier dans toute sa généralité, de nombreux auteurs se sont concentrés sur une sous-famille : les lois à queue de type Weibull. Leur définitions est donnée dans le paragraphe suivant.

1.5 Estimateur de l'indice des valeurs extrêmes γ

1.5.1 Représentation graphique

Le Pareto quantile plot. Le domaine de Fréchet ($\gamma > 0$) a été le plus largement étudié dans la littérature dans la mesure ou il englobe un grand nombre d'applications pratiques. Dans

ce domaine, les distributions F ont la propriété suivante

$$1 - F(x) = x^{-1/\gamma}\ell(x), \quad x > 0, \tag{1.5.1}$$

avec F une fonction à variations lentes à l'infini. Elle satisfait donc la convergence suivante

$$\forall \alpha > 0, \ \lim_{x\to\infty} \frac{\ell(\alpha x)}{\ell(x)} = 1. \tag{1.5.2}$$

En pratique, il est souvent plus commode, non pas de travailler sur la fonction F elle-même, mais sur la fonction queue définie par:

$$U(t) = \inf\left\{x \in \mathbb{R} : F(x) \geq 1 - \frac{1}{x}\right\}. \tag{1.5.3}$$

Dans ce cas, supposer (1.5.1) est équivalent à supposer que

$$\forall x > 0, \ U(x) = x^{\gamma}\ell_U(x), \tag{1.5.4}$$

avec ℓ_U également une fonction à variations lentes à l'infini. Notons que cette fonction queue est directement liée à la notion de période de retour, qui sera ultérieurement définie.Si nous considérons la statistique d'ordre $X_{1,n} \leq X_{2,n} \leq ..... \leq X_{n,n}$ associée à notre échantillon initial, le Pareto quantile plot, correspondant au graphe de

$$(\log\frac{n+1}{i}, \log X_{n-i+1,n}),$$

est une représentation trés utile pour visualiser graphiquement si des données sont distribuées selon une loi du domaine de Fréchet ou non. En effet, de (1.5.4), il découle que,

$$\begin{aligned}\log U(x) &= \gamma \log x + \log \ell_U(x) \\ &= \gamma \log x \left(1 + \frac{\log \ell_U(x)}{\gamma \log x}\right).\end{aligned} \tag{1.5.5}$$

En utilisant les propriétés des fonctions à variations lentes, il est immédiat que $\dfrac{\log \ell_U(x)}{\gamma \log x} \to 0, (x \to \infty)$ ce qui implique que

$$\log U(x) \sim \gamma \log x, (x \to \infty).$$

En remplaçant la fonction queue par sa version empirique $\hat{U}_n$ et en remarquant que $\hat{U}_n = X_{n-i+1,n}$, nous obtenons finalement l'équivalence suivante:

$$\log X_{n-i+1,n} \sim \gamma \log\frac{n+1}{i}, \text{quand}\left(\frac{n+1}{i}\right) \to \infty.$$

En d'autres termes, le Pareto quantile plot sera approximativement linéaire, avec une pente γ, pour les petites valeurs de i, c'est-à-dire les points extrêmes.

Estimateur de Hill $\hat{\gamma}_{k_n}^H$

Il existe beaucoup d'estimateurs de l'indice proposés dans la littérature. Les plus utilisés en hydrologie sont sans aucun doute les estimateurs des moments, du maximum de vraisemblance. Dans le domaine de Fréchet. Comme nous venons de le signaler, le comportement linéaire dans les points extrêmes a lieu avec une pente γ. Autrement dit, on peut facilement construire des estimateurs de l'indice à partir de ce graphe. Cette linéarité apparaît au-del à d'un point $(\log \frac{n+1}{k}, \log X_{n-k+1,n})$. Deux approches sont donc possibles pour la construction de tels estimateurs: soit en forçant la droite à passer par ce point, ce que l'on appelera par la suite avec contrainte; soit simplement par moindres carrés, donc sans contrainte. Dans le cas avec contrainte, Csörgő et al. (1985) ont proposé les estimateurs à noyau $K_{k,n}$ définis de la façon suivante :

$$K_{k,n} = \frac{\sum_{i=1}^{k} \frac{i}{k} K\left(\frac{j}{k}\right) \left(\log_{n-i+1,n} - \log X_{n-i,n}\right)}{\sum_{i=1}^{k} \frac{1}{k} K\left(\frac{i}{k}\right)} \tag{1.5.6}$$

ou K représente un noyau d'intégrale égale à 1. Suivant le choix de ce noyau, différents estimateurs peuvent en résulter, le plus connu étant l'estimateur de $Hill(1975)$, correspondant au cas particulier $K(x) = \mathbb{I}_{]0,1[}(x)$, qui peut donc se réécrire simplement comme

$$\hat{\gamma}_{k_n}^H = \frac{1}{k} \sum_{i=1}^{k} i \left(\log X_{n-i+1,n} - \log X_{n-i,n}\right) \tag{1.5.7}$$

$$= \frac{1}{k} \sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \tag{1.5.8}$$

1.6 Quelques généralités sur la censure

1.6.1 Distributions de la durée de survie

Supposons que la durée de survie X soit une variable positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions):

Fonction de survie S

La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = P(X > t) \ = 1 - P(X \leq t) = 1 - F(t), \quad t > 0. \tag{1.6.1}$$

Densité de probabilité f

C'est la fonction $f(t) \geq 0$ telle que pour tout $t > 0$

$$F(t) = \int_0^t f(u)du \tag{1.6.2}$$

Si la fonction de répartition F admet une dérivée au point t alors,

$$F(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X < t + h)}{h} = F^{'}(t) = -S'(t) \tag{1.6.3}$$

Pour t fixé, la densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

Risque instantané (ou taux de hasard)

Le risque instantané (ou taux d'incidence), pour t fixé caractérise la probabilité de mourir dans un petit intervalle 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) :

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X < t + h \mid X \geq t)}{h} = \frac{f(t)}{S(t)} = -(\ln S(t))' \tag{1.6.4}$$

Taux de hasard cumulé

Le taux de hasard cumulé est l'intégrale du risque instantané λ :

$$C(t) = \int_0^t \lambda(u)du = -(\ln S(t)).$$

1.6.2 Les données censurées

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.

Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. pour l'individu i; considérons

son temps de survie X_i,

son temps de censure Y_i,

la durée réellement observée Z_i.

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.

1 **La censure de type I.** Soit Y une valeur fixée, au lieu d'observer les variables $(X_1, ..., X_n)$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq Y$, sinon on sait uniquement que $X_i > Y$. on utilise la notation suivante :

$$Z_i = X_i \wedge Y = min(X_i, Y).$$

ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Par exemple, on peut tester la durée de vie de n objet identiques (ampoules) sur un intervalle d'observation fixé $[0, u]$. En biologie, on peut tester l'eficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps u sont sacrifiées).

2 **La censure de type II.** Elle est 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 $X_{(i)}$ et $Z_{(i)}$ les statistiques d'ordre des variables X_i et Z_i. La date de censure est donc $X_{(k)}$ et on observe les variables suivantes :

$$\begin{aligned} Z_{(1)} &= X_{(1)} \\ Z_{(2)} &= X_{(2)} \\ &\vdots \\ Z_{(r)} &= X_{(r)} \\ Z_{(r+1)} &= X_{(r)} \\ &\vdots \\ Z_{(n)} &= X_{(r)} \end{aligned}$$

3 **La censure de type III.** (ou censure aléatoire de type I). Soient $(Y_1, ..., Y_n)$ des variables aléatoires i.i.d. On observe les variables

$$Z_i = X_i \wedge Y$$

L'information disponible peut être résumée par :

1. la durée réellement observée Z_i
2. un indicateur $\delta_i = \mathbb{I}_{\{X_i \leq Y\}}$

- $\delta_i = 1$ si l'événement est observé (d'où $Z_i = X_i$). On observe les vraies durées ou les durées complètes.
- $\delta_i = 0$ si l'individu est censuré (d'où $Z_i = Y_i$). On observe des durées incomplètes (censurées).

La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par:

(a) la perte de vue: le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients perdus de vue.

(b) l'arrêt ou le changement du traitement : les effets secondaires ou l'ineficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.

(c) lafin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients exclus-vivants. Les perdus de vue (et les exclusions) et les exclus-vivants correspondent à des observations censurées mais les deux mécanismes sont de nature différente (la censure peut être informative chez les perdus de vue).

Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (Z, δ):

$$Z_i = X_i \wedge Y_i \quad \text{et} \quad \delta_i = \mathbb{I}_{\{X_i \geq C_i\}}$$

Comme pour la censure à droite, on suppose que la censure Y est indépendante de X. Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs: dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues.

Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'événement s'est produit entre

ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à de la censure à droite.

1.6.3 Estimateur de Kaplan-Meier

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 $t^{''} < t^{'} < t$

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

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

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

Y_i le nombre d'individus à risque de subir l'événement juste avant le temps $Z_{(i)}$, d_i le nombre de décès en $Z_{(i)}$. Alors la probabilité p_i est de mourir dans l'intervalle $]Z_{(i-1),}Z_{(i)}]$ sachant que l'on était vivant en $Z_{(i-1)}$, i.e.

$$p_i = \mathbb{P}(X \leq Z_{(i)} \mid X > Z_{(i)}),$$

peut être estimée par $p_i = \dfrac{d_i}{Y_i}$. Comme les temps d'événements sont supposés distincts, on a

- $d_i = 0$ en cas de censure en $Z_{(i)}$, i,e. quand $\delta_i = 0$.

- $d_i = 1$ en cas de décès en $Z_{(i)}$, i,e. quand $\delta_i = 1$.

On obtient alors l'estimateur de Kaplan-Meier:

$$\begin{aligned}\hat{S}_n(t) &= 1 - \hat{F}_n(t) \\ &= \prod_{i:Z_{i,\leq t}}^{n} \left(1 - \frac{\delta_i}{n - (i-1)}\right) \\ &= \prod_{i=1}^{n} \left[\frac{n-1}{n-i+1}\right]^{\delta_i}\end{aligned}$$

L'estimateur $S(t)$ 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. $S(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier 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).

Chapitre 2

Définitions et caractéristiques de deux estimateurs

2.1 Estimateur de l'indice de queue $\hat{\gamma}_H^c$

Soient $(X_i)_{i\leq n}$ et $(Y_i)_{i\leq n}$deux échantilons de v.a. indépendantes et identiquement distribuées où F et G sont respectivement leur fonctions de répartition absolument continues (avec τ_F et τ_G sont les points terminaux respectifs où $\tau_F = \sup\{x, F(x) < 1\}$. Dans le cas des données censurées à droite, on observe pour $i \leq n$,

$$Z_i = X_i \wedge Y_i \quad \text{et} \quad \delta_i = \mathbb{I}_{\{X_i \geq Y_i\}}.$$

Nous définissons H la distribution de l 'échantillon Z, satisfaisant:

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

et par $Z_{1,n} \leq Z_{2,n} \leq ..... \leq Z_{n,n}$ les statistiques d'ordre associées. $\delta_{[1:n]}, ..., \delta_{[n:n]}$ sont les indicateurs de censure associés à ces dérnières. Si F et G sont absolument continues et que $F \in D(G\gamma_X)$, $G \in D(G\gamma_Y)$ respecivement, pour certains $\gamma_X, \gamma_Y \in \mathbb{R}$. Pour tout $\gamma \in \mathbb{R}$, *Einmahl et al* (2008) ont proposé les trois cas les plus intéressents suivants:

$$\begin{cases} cas\ 1 & : & \gamma_X > 0, \gamma_Y > 0 & & \gamma = \frac{\gamma_X \gamma_Y}{\gamma_X + \gamma_Y} \\ cas\ 2 & : & \gamma_X < 0, \gamma_Y < 0 & \tau_F = \tau_G & \gamma = \frac{\gamma_X \gamma_Y}{\gamma_X + \gamma_Y} \\ cas\ 3 & : & \gamma_X = \gamma_Y = 0 & \tau_F = \tau_G = \infty & \gamma = 0 \end{cases} \tag{2.1.1}$$

Dans le cas 3, nous définissons également, pour une présentation commode, $\gamma = \frac{\gamma_X \gamma_Y}{\gamma_X + \gamma_Y} = 0$. 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 $\gamma_X > 0$ et $\gamma_Y < 0$) ou la situation complètement censurée, où l'estimation est impossible (cela arrive, en particulier, quand $\gamma_X > 0$ et $\gamma_Y < 0$).

Définition 2.1.1 *Soit* $\{(Z_i, \delta_i), 1 \leq i \leq n\}$ *un échantillon de couple de v.a's* (Z, δ). *Soient* $Z_{1,n} \leq ... \leq Z_{n,n}$ *représente les statistiques d'ordre associées à l'échantillon* $(Z_1, ..., Z_n)$. *Avec les* $\delta_{[1:n]}, ..., \delta_{[n:n]}$ *sont les indicateurs de censure retenues respectivement avec l'échantillon* $Z_{1:n}, ..., Z_{n:n}$,

$$\delta_{[i:,n]} = \delta_j \ si \ Z_{i,n} = Z_j.$$

Beirlant et al. *ont proposé différents estimateurs de* γ_X, *l'indice des valeurs extrêmes 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_X *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* t, *c'est-à-dire à utiliser,*

$$\hat{\gamma}_{X,k,n}^{(c,.)} = \frac{\hat{\gamma}_{Z,k,n}^{(.)}}{\hat{p}} \tag{2.1.2}$$

où,

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

avec, k *est le nombre d'excés au-delà de* u. $\hat{p}$ *estime* $p = \dfrac{\gamma_Y}{\gamma_X + \gamma_Y}$ *(*$\hat{p} \to p$: *quand* $n \to \infty$*), par conséquent* $\hat{\gamma}_{Z,k,n}^{(.)}$ *estimateur d* γ *divisé par* $\dfrac{\gamma_Y}{\gamma_X + \gamma_Y}$ *qui est égal à* γ_X. $\hat{\gamma}_{Z,k,n}^{(.)}$ *peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill* $\hat{\gamma}_{Z,k,n}^{(H)}$. *Pour adapter l'estimateur de Hill dans le cas censuré nous allons diviser cet estimateur 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* $\hat{\gamma}_X^c$ *est difini par*

$$\hat{\gamma}_X^c = \frac{\hat{\gamma}^H}{\hat{p}},$$

où,

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

alors,

$$\hat{\gamma}_X^c = \frac{k^{-1}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}}{k^{-1}\sum_{i=1}^{k}\delta_{[n-i+1:n]}} \tag{2.1.3}$$

Einmahl et al (2008) *ont établi de façon unifiée, la normalité asymptotique de tout estimateur de l'indice des valeurs extrêmes écrit sous la forme (*2.1.2*) 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 $Z_1, ..., Z_n$.*

2.1.1 Propriétés asymptotiques de l'estimateur de l'indice $\hat{\gamma}_X^c$

Pour déterminer les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes nous avons besoin de la fonction suivante:

$$p(z) = \mathbb{P}(\delta = 1, Z = z).$$

Nous pouvons l'écrire d'une autre manière,

$$p(z) = \frac{(1-G(z))f(z)}{(1-G(z))f(z) + (1-F(z))g(z)},$$

où f et g désignent respectivement les densités de F et G et on a:

$$\lim_{z\to\tau_H} p(z) = \frac{\gamma_Y}{\gamma_X + \gamma_Y} := p \tag{2.1.4}$$

Pour expliquer (2.1.4): Supposons X et Y suivent respectivement une loi de Pareto (γ_X) et Pareto(γ_Y) respectivement, C'est-à-dire pour tout $x \geq 1$.

$$F_X(x) = 1 - x^{-1/\gamma_X}, \qquad \gamma_X > 0$$
$$F_Y(x) = 1 - x^{-1/\gamma_Y}, \qquad \gamma_Y > 0.$$

On obtient:

$$\begin{aligned} F_Z(z) &= \mathbb{P}(\min(X, Y \leq z)) \\ &= 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= 1 - z^{-1/\gamma_X} z^{-1/\gamma_Y} \\ &= 1 - z^{-\frac{\gamma_X+\gamma_Y}{\gamma_X\gamma_Y}}, \end{aligned}$$

ce qui implique que $Z \sim \text{Pareto}(\gamma_X\gamma_Y/(\gamma_X+\gamma_Y))$, nous pouvons à présent calculer la fonction $p(z)$,

$$\begin{aligned} p\equiv p(z) &= \frac{(1-F_Y(z))f_X(z)}{(1-F_Y(z))f_X(z)+(1-F_X(z))f_Y(z)} \\ &= \frac{z^{-1/\gamma_Y}\frac{1}{\gamma_X}z^{-1/\gamma_X}}{z^{-1/\gamma_Y}\frac{1}{\gamma_X}z^{-1/\gamma_X}+z^{-1/\gamma_X}\frac{1}{\gamma_Y}z^{-1/\gamma_Y}} \\ &= \frac{\frac{1}{\gamma_X}z^{-1/\gamma_X-1/\gamma_Y}}{(\frac{1}{\gamma_X}+\frac{1}{\gamma_Y})z^{-1/\gamma_X-1/\gamma_Y}} = \frac{\frac{1}{\gamma_X}}{\frac{1}{\gamma_X}+\frac{1}{\gamma_Y}} \\ &= \frac{\gamma_Y}{\gamma_X+\gamma_Y}. \end{aligned}$$

Par conséquent, le quotient entre un estimateur de $\gamma=\gamma_Z$ et un estimateur de $p=p_z$. En effet, beaucoup plus général, et pour tous les cas mentionnés ci-dessus (*voir* 2.1.1) .

$$\begin{aligned} &F_X \in D_M(EV_{\gamma_X}), \quad F_Y \in D_M(EV_{\gamma_Y}) \\ &\quad\Longrightarrow F_{Z=\min(X,Y)} \in D_M(EV_\gamma), \text{tel que} : \gamma = \frac{\gamma_X\gamma_Y}{\gamma_X+\gamma_Y} \end{aligned}$$

Pour déterminer les propriétés asymptotiques de l'estimateur nous avons besoin de quelques hypothèses de régularité, nous supposons les assertions suivantes :

$\mathbb{C}1$: Il existe $\rho<0$ et une fonction à variation réguliéres $b(.)$ d'indice ρ telle que pour tout $u>0$

$$\lim_{t\to\infty}\frac{H^{\leftarrow}(1-\frac{1}{tu})/H^{\leftarrow}(1-\frac{1}{t})-u^\gamma}{b(t)} = u^\gamma\frac{u^\rho-1}{\rho} \tag{2.1.5}$$

si la suite $k=k_n$ est une suite intermédiaire, telle que :

$$1<k<n;\ k\to\infty \ \text{ et } k/n\to 0, n\to\infty \tag{2.1.6}$$

$\mathbb{C}2$: $\sqrt{k}b(\frac{n}{k})\to\alpha_1\in\mathbb{R}$

$\mathbb{C}3$: $\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\to\alpha_2\in\mathbb{R}$

$\mathbb{C}4$: Soit $c>0$ et $\mathcal{A}(s,t) := \left\{1-k/n\le t<1, |t-s|\le C\sqrt{k}/n, s<1\right\}$ si $n\to\infty$,

$$\sqrt{k}\sup_{\mathcal{A}(s,t)}|p(H^{\leftarrow}(t))-p(H^{\leftarrow}(s))|\to 0$$

Sous ces conditions, nous avons les résultats asymptotiques des estimateurs.

Théorème 2.1.1 *Sous les condition* $\mathbb{C}1 - \mathbb{C}4$ *et s'il existe* b_0 *et* σ *telles que*

$$\sqrt{k}(\hat{\gamma}^{(.)}_{Z,k,n} - \gamma) \xrightarrow{d} \mathcal{N}(\alpha_1 b_0, \sigma^2). \tag{2.1.7}$$

Alors, nous avons

$$\sqrt{k}\left(\hat{\gamma}^{(c,.)}_{Z,k,n} - \gamma_X\right) \xrightarrow{d} \mathcal{N}\left(\frac{1}{p}(\alpha_1 b_0 - \gamma_X \alpha_2), \frac{\sigma^2 + \gamma_X^2 p(1-p)}{p^2}\right)$$

telle que $b_0 = 1/(1-\rho)$ *et* $\sigma^2 = \gamma^2$.

nous avons les résultats asymptotiques de l'estimateurs de Hill.

$$\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n} - \gamma_X\right) \xrightarrow{d} \mathcal{N}\left(\mu^{(c,H)}; \frac{\gamma_X^3}{\gamma}\right)$$

On a :

$$\mathbb{E}\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n} - \gamma_X\right)\right) = \mu^{(c,H)} := -\frac{\gamma_X \alpha_2}{p} + \frac{\alpha_1}{p}\frac{\gamma}{\tilde{\rho} + \gamma(1-\tilde{\rho})}$$
$$\mathbb{V}\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n} - \gamma_X\right)\right) = \frac{\gamma_X^3}{\gamma}$$

2.2 Estimations de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier

Julien Worms et Rym Worms (2013) ont proposé deux nouvelles approches pour l'estimation de l'indice des valeurs extrêmes dans le cadre de la censure aléatoire (à droite) des échantillons, sur la base des idées d'intégration de Kaplan-Meier. Ces idées sont développées dans le cas des distributions à queue lourde, et pour l'adaptation de l'estimateur de Hill, dont la consistance est prouvée sous la conditions du premier ordre. Le premier point de la nouvelle approche de départ est le résultat bien connu suivant, qui est la base des méthodes de régression censurés: Si ϕ est une fonction réel positif,

$$\mathbb{E}\left[\frac{\delta}{1 - G(Z)}\phi(Z)\right] = \mathbb{E}\left[\phi(X)\right]. \tag{2.2.1}$$

Il est prouvé: depuis $Z = X$ quand $\delta = 1$,

$$\begin{aligned}\mathbb{E}\left[\frac{\delta}{1-G(Z)}\phi(Z)\right] &= \int \mathbb{I}_{x<y}\frac{\delta}{1-G(x)}dF(x)dG(y)\\ &= \int \phi(x)\left(1-G(x)\right)^{-1}\left(\int \mathbb{I}_{y>x}dG(y)\right)dF(x)\\ &= \int \phi(x)dF(x)\\ &= \mathbb{E}\left[\phi(X)\right].\end{aligned}$$

Dans le contexte des statistiques de valeurs extrêmes, l'idée est de tirer parti de cette propriété et du fait que certains paramètres de queue de la distribution de X peuvent être approchés par l'espérance d'une fonction de X, permttant leur estimation. Nous allons l'illustré dans le cadre des distributions à queue lourde, et pour l'estimation de l'indice des valeurs extrêmes, en supposant que nous sommes dans la première des trois cas,

$$F \in D(G_{\gamma_X}), G \in D(G_{\gamma_Y})\ \gamma_X > 0 \text{ et} \gamma_Y > 0, \tag{2.2.2}$$

qui, comme indiqué plus haut, implique que $H \in D(G_\gamma)$ avec,

$$\gamma = \frac{\gamma_X\gamma_Y}{\gamma_X + \gamma_Y}.$$

Dans ce cas, il est bien connu que (voir remarque 1.2.3 dans [Haan et Ferreira (2006,)])

$$\lim_{u\to\infty}\mathbb{E}\left[\log(\tfrac{X}{u}) \mid X > u\right] = \gamma_X, \tag{2.2.3}$$

Si k_n est une suite des nombres entiers satisfaisant, quand n tend vers $+\infty$,

$$k_n \to +\infty \qquad k_n = o(n). \tag{2.2.4}$$

Alors nous pouvons définir une version aléatoire de ϕ

$$\phi(x) = \left(\mathbb{P}(X > u)\right)^{-1}\log(\tfrac{x}{u})\mathbb{I}_{x>u},$$

avec un seuil aléatoire $u = Z_{n-k,n}$,

$$\hat{\phi}_n(x) := \frac{1}{1-\hat{F}_n(Z_{n-k,n})}\log\left(\frac{x}{Z_{n-k,n}}\right)\mathbb{I}_{x>Z_{n-k,n}}. \tag{2.2.5}$$

Par conséquent, en combinant (2.2.1) et (2.2.3) avec cette fonction $\hat{\phi}_n$,

$$\int \hat{\phi}_n(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_{in}\hat{\phi}_n(Z_i) \text{ où } w_{in} = \frac{\delta_i}{1-\hat{G}_n(Z_i)}.$$

L'adaptation première de l'estimateur de Hill est valable dans le cas (2.2.2),

$$\hat{\gamma}_{X,Hill}^{KM} := \frac{1}{n(1-\hat{F}_n(Z_{n-k,n}))}\sum_{i=1}^{k_n}\frac{\delta_{n-i+1,n}}{1-\hat{G}_n(Z^-_{n-i+1,n})}\log\left(\frac{Z_{n-i+1,n}}{Z_{n-k,n}}\right) \tag{2.2.6}$$

où $\hat{F}_n$ et $\hat{G}_n$ représentent les estimations de Kaplan-Meier de F et G, respectivement. Notez que nous prenons $\hat{G}_n(Z^-_{n-i+1,n})$ au lieu de $\hat{G}_n(Z_{n-i+1,n})$, dans la défnition de $\hat{\gamma}_{X,Hill}^{KM}$, parce que $1-\hat{G}_n(Z_{n,n})$ peut être nul. Le théorème suivant fournit la consistance de cet estimateur. A cet effet, il faut deux hypothèses supplémentaires sur le comportement de la fonction $p\circ H^{\leftarrow}$, qui sont similaires à celles utilisées dans [*Einmahl et al.* (2008)]: si $p(z) = \mathbb{P}(\delta = 1, Z = z)$

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n})) - p\right] \xrightarrow{P} c \in \mathbb{R} \tag{2.2.7}$$

$$\sqrt{k}\sup_{(s,t)\in C_n}|p(H^{\leftarrow}(t)) - p(H^{\leftarrow}(s))| \to 0 \quad C > 0 \tag{2.2.8}$$

où $C_n = \left\{(s,t) \text{ tel que } s < 1, 1-k_n/n \le u < 1, |u-s| \le C\sqrt{k_n}/n\right\}$

Théorème 2.2.1 *Sous les hypothèses* $(2.2.2), (2.2.4), (2.2.7), (2.2.8)$ *si l'on suppose en outre que, pour* $\delta > 0$,

$$-\log(k_n/n)/k_n = O(n^{-\delta}) \tag{2.2.9}$$

et que $\gamma_X < \gamma_Y$*, puis, lorsque* n *tend vers* $+\infty$,

$$\hat{\gamma}_{X,Hill}^{KM} \xrightarrow{\mathcal{P}} \gamma_X.$$

Remarque 2.2.1 *Ce théorème n'a été prouvé que pour* $\gamma_X < \gamma_Y$ *(une condition utilisée à plusieurs reprises dans la preuve;voir remarque suivante), qui peut être interprété comme une légère censure dans la queue (finalement, pas plus de* 50% *des observations dans la queue sont censurées). en fait, si nous utilisons notre estimateur dans le cas de censure forte (*$\gamma_X \ge \gamma_Y$*), les simulations semblent montrer que la performance (en termes de MSE) est, étonnamment, pire que celle de l'affaire* $\gamma_X < \gamma_Y$ *(voir chapitre 3). Cependant, le même phénomène est observé (et même fortement) pour la version* $(\hat{\gamma}_Z/p)$*de l'estimateur de la colline.*

Remarque 2.2.2 *La condition* $\gamma_X < \gamma_Y$ *provient essentiellement du fait que l'etimateur* $\hat{\gamma}^{KM}_{X,Hill}$ *est convergent vers la puissance* α *d'une lois i.i.d. standard de Pareto, où* α *est proche de* γ/γ_Y. *cet exposant* $\gamma/\gamma_Y = \frac{\gamma_X}{\gamma_X+\gamma_Y}$ *est toujours plus petit que* 1, *mais (pour les conditions de moment) nous avons été amenés à supposer qu'il soit en fait plus petit que* $1/2$, *c.-à-d.* $\gamma_X < \gamma_Y$.

Chapitre 3

Comparaison des estimateurs par simulation

3.1 Principe du Bootstrap

Les techniques de "reechantillonnage", appeleès aussi en anglais methodes du "bootstrap" a été proposée par *Bradley Efron* (1979), cette méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles.

Le principe fondamental de cette technique de ré-échantillonnage est de substituer à la distribution de probabilité inconnue F, dont est issu l'échantillon d 'apprentissage, la distribution empirique $\hat{F}$ qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit échantillon bootstrap selon la distribution empirique $\hat{F}$ par n tirages aléatoires avec remise parmi les n observations initiales.

Il est facile de construire un grand nombre d'échantillons bootstrap sur lesquels calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables de la loi de l'estimateur. Cette approximation fournit ainsi des estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance de l'estimateur sans hypothèse (normalité) sur la vraie loi.

3.1.1 Choix du B, le nombre de Bootstraps

L'importance des valeurs extrêmes de la statistique γ etudiée est un facteur important dans la détermination du choix de B: plus ces valeurs sont fréquentes, plus B devrait tre grand. On notera cependant que certaines autres applications du bootstrap exigent un B beaucoup plus grand ; ce sera en particulier le cas pour l'application à la construction d'intervalles de confiance. Selon *B.Efron* , il est rarement nécessaire d'utiliser plus de $B = 200$ échantillons bootstrap pour estimer une variance ; dans bien des cas, $B = 50$ ou 100 sont suffisants.

3.1.2 Application de bootstrap sur des données censurées

Le bootstrap est bien validée par de nombreuses études statistiques et une des premières applications du bootstrap a été faite dans le contexte d'analyse de la survie (*Bradley Efron*, 1981) pour répondre à certaines questions notamment pour construire les bandes de confiance

3.2 Choix du nombre des valeurs extrêmes optimal k_n

Les résultats concernant les estimateurs de l'indice des valeurs extrêmes sont asymptotiques : ils sont obtenus lorsque $k_n \to \infty$ et $k_n/n \to 0$. La difficulté en pratique consiste à choisir le nombre d'extrêmes k_n utilisé dans les estimations. L'issue est importante: l'extrême volatilité du graphe $\{(k_n, \hat{\gamma}_{k_n}) : 1 \leq k_n < n\}$, où $\hat{\gamma}_{k_n}$ représente n'importe quel estimateur introduit précédement, rend difficile l'utilisation de l'estimateur en pratique si aucune indication sur le choix de k_n n'est donnée. Des travaux ont montré qu'en utilisant trop d'observations, dans la procédure d'estimation de γ, on observe un biais substantiel tandis que l'utilisation de peu d'observations conduit à une variance considérable.

3.2.1 Méthode basée sur l'erreur quadratique moyenne

Dans un autre point de vue, une minimisation de l'erreur moyenne quadratique asymptotique (en anglais mean squared error: $AMSE$) est souvent donnée comme critère. L'erreur

quadratique moyenne de l'estimateur $\hat{\gamma}_{k_n}$ de l'indice de queue γ est définie par:

$$MSE = E_\infty((\hat{\gamma}_{k_n} - \gamma)^2),$$

où E_∞ est l'espérance mathématique suivant la distribution asymptotique. Il est donc facile de voir que l'erreur quadratique moyenne de $\hat{\gamma}_{k_n}$, qui est en fonction de k_n, n'est rien d'autre que le carré du biais plus la variance de l'estimateur considéré. Par conséquent pour une estimation précise de l'indice de queue γ, il est nécessaire pour un estimateur classique de trouver un compromis entre le biais et la variance. Il semble raisonnable qu'une minimisation du MSE permet de trouver une valeur intermédiaire entre les composantes du biais et de la variance pour ce compromis.

Il semble donc naturel de trouver une valeur $\hat{\gamma}_{k_n}$ qui minimise le graphe de l'erreur quadratique moyenne estimée $\{(k_n, MSE(k_n), \quad k_n = 1, ..., n-1)\}$ .La valeur optimale de k_n est donnée par

$$k_{nopt} = \arg\min_{k_n} MSE(k_n) = \arg\min_{k_n} \left\{ \frac{1}{R} \sum_{j=1}^{R} \left(\hat{\gamma}_{k_n}^{j} - \gamma \right)^2 \right\}. \tag{3.2.1}$$

par sumilation. On peut se baser sur des méthodes de "bootstrap" pour calculer (MSE). Pour toute réplication R nous estimons γ_X et soit $\hat{\gamma}_{k_n}^{c,j}$ l'estimateur de γ_X obtenu à la $j-iéme$ réplication $(j = 1, ..., R)$ avec $(k_n = 1, ..., n-1)$. Il semble donc naturel de trouver une valeur k_{nopt} qui minimise les valeurs de l'erreur quadratique moyenne par rapport à k_n.

3.3 Bootstrap des l'estimateurs

Soit $X_1, X_2, ..., X_n$, n variables représentant les durées de vie de n sujets, sont des variables aléatoires positives, indépendantes et de fonction de répartition F, et indépendamment des variables aléatoires $Y_1, ..., Y_n$, les instants de censures associés, positives, de fonction de répartition G. On note $\{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$ l'échantillon réellement observé, où, pour $i \leq n$,

$$Z_i = X_i \wedge Y_i \quad \text{et} \quad \delta_i = \mathbb{I}_{X_i \leq Y_i},$$

avec H la fonction de distribution de $Z-$échantillons. Soit $\{(Z_{1:n}, \delta_{1:n}), ..., (Z_{n:n}, \delta_{n:n})\}$ l'échantillon ordonné suivant les valeurs de Z_i. *Efron (1981)* suggère le plan du ré-échantillonnage

suivant : On génère un échantillon bootstrapé,

$$(Z_1^*, \delta_1^*), ..., (Z_n^*, \delta_n^*), \tag{3.3.1}$$

en tirant chaque couple aléatoirement et avec remise dans l'échantillon observé,

$$(Z_1, \delta_1), ..., (Z_n, \delta_n), \tag{3.3.2}$$

et soit $(Z_{i:n}^*, \delta_{i:n}^*)_{i=1,...,n}$l'échantillon ordonné suivant les valeurs de Z_i^*. Si F et G sont supposées être dans le dommaine d'attraction maximales teleque $F \in D(G_{\gamma_X}), G \in D(G_{\gamma_Y})$ $\gamma_X, > 0$,$\gamma_Y > 0$, comme indiqué plus haut, implique que $H \in D(G_\gamma)$ avec $\gamma = \gamma_X\gamma_Y/(\gamma_X + \gamma_Y)$. L'estimateur de Hill bootstrapé de l'indice de valeurs extrêmes $\hat{\gamma}_X^{*c}$ et l'estimateur de Kaplan-Meier $\hat{\gamma}_{X,Hill}^{*KM}$ construit avec les données $(Z_{i:n}^*, \delta_{i:n}^*)_{i=1,...,n}$ s'écrit:*

$$\hat{\gamma}_X^{*c} = \frac{\frac{1}{k}\sum_{i=1}^{k} \log Z_{n-i+1,n}^* - \log Z_{n-k,n}^*}{\frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]}^*}, \tag{3.3.3}$$

et,

$$\hat{\gamma}_{X,Hill}^{*KM} := \frac{1}{n(1-\hat{F}_n(Z_{n-k_n,n}^*))} \sum_{i=1}^{k_n} \frac{\delta_{n-i+1,n}^*}{1-\hat{G}_n(Z_{n-i+1,n}^{*-})} \log\left(\frac{Z_{n-i+1,n}^*}{Z_{n-k_n,n}^*}\right). \tag{3.3.4}$$

3.3.1 Propriétés de l'estimateur de l'indice de queue $\hat{\gamma}_X^{*c}$

Soit l'indice des valeurs extrêmes γ_X associé à l'échantillon $(X_i)_{1\leq i\leq n}$, et soit $\hat{\gamma}_X^{*c}$: une estimation de cet indice, obtenue à partir des données de l'échantillon initial

$$Z = \{(Z_1, \delta_1), ..., (Z_n, \delta_n)\},$$

Chaque échantillion

$$Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\},$$

obtenu par rééchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\hat{\gamma}_{X,Hill}^{KM}$,

$$\hat{\gamma}_X^{*c}(b) \ , \qquad b = 1, ..., B.$$

Estimation Bootstrap de l'erreur standard

Définition 3.3.1 *On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs* $\hat{\gamma}_X^{*c}(b)$ *, la moyenne est:*

$$\hat{\gamma}_X^{*c}(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_X^{*c}(b) \tag{3.3.5}$$

L'écart type est aussi une caractéristique importante de chaque distribution. Pour un ensemble d'estimateurs $\hat{\gamma}_X^{*c}(b)$ *l'écart type estimé est calculé par la formule:*

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_X^{*c}(b) - \hat{\gamma}_X^{*c}(.)\right)^2} \tag{3.3.6}$$

où B *est le nombre total d'échantillons bootstrap .*

Algorithme 3.3.1 *estimation bootstrap de l'erreur standard*

Variable

B*:entier assez grand*

Début

pour $b = 1$ ***à*** B ***faire***

on calcule l'estimateur de queue:

$$\hat{\gamma}_X^{*c}(b).$$

On obtient alors un échantillon de B *valeurs*

$$\{\hat{\gamma}_X^{*c}(1), \hat{\gamma}_X^{*c}(2), ..., \hat{\gamma}_X^{*c}(B)\}$$

On estime alors l'erreur standard $se_F(\hat{\gamma}_X^c)$ *par l'erreur standard de cet échantillon de* $\hat{\gamma}_X^{*c}$*, i.e*

$$\widehat{se}_F(\hat{\gamma}_X^c) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_X^{*c}(b) - \hat{\gamma}_X^{*c}(.)\right)^2}$$

fin pour

Fin.

Estimation bootstrap du biais

Le biais d'un estimateur s'exprime comme

$$\widehat{Biais}\left(\hat{\gamma}_X^c\right) = \mathbb{E}_{\hat{F}}\left[\hat{\gamma}_X^c\right] - \hat{\gamma}_X^c$$

Définition 3.3.2 *On appelle estimateur bootstrap du biais, l'estimateur de l'indice de queue pour les données observées*

$$\widehat{Biais}_{boot}\left(\hat{\gamma}_X^c\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_X^{*c}\left(b\right) - \hat{\gamma}_X^{*c} \tag{3.3.7}$$

Algorithme 3.3.2 *Estimation bootstrap du biais*

Variable

B: entier assez grand.

Début

Pour $b = 1$ *à* B ***faire***

Générer $Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$

Calculer $\hat{\gamma}_X^{*c}(b)$.

Calculer $\hat{\gamma}_X^{*c}\left(b\right) = \#\left\{j, (Z_i^{*b}, \delta_i^{*b}) = (Z_i, \delta_i)\right\} / n$ *pour tout* i.

Fin pour

Calculer $\hat{\gamma}_X^{*c}\left(i\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_X^{*c}\left(b\right)$ *pour tout* i.

Retourner

$$\widehat{Biais}_{boot}\left(\hat{\gamma}_X^c\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_X^{*c}\left(b\right) - \hat{\gamma}_X^c \tag{3.3.8}$$

Fin.

Estimation Bootstrap de l'erreur quadratique moyenne

Définition 3.3.3 *Monde réel: l'erreur quadratique moyenne* (MSE) *de* $\hat{\gamma}_X^c$ *est égale à*

$$MSE_F = \mathbb{E}_F\left[\left(\hat{\gamma}_X^c - \gamma_X\right)^2\right]$$

*Monde bootstrap**:** l'estimateur bootstrap de l'erreur quadratique moyenne de* $\hat{\gamma}_1^{(c,H)}$ *est défini par:*

$$\widehat{MSE}_{\hat{F}} = \mathbb{E}_{\hat{F}}\left[\left(\hat{\gamma}_X^{*c}(b) - \hat{\gamma}_X^c\right)^2\right]$$

Algorithme 3.3.3 *estimation bootstrap de la MSE*

Variable

B: entier assez grand

Début

Pour *b* ***variant de*** 1 ***à*** *B*

Générer Z^{*b} *réalisation d'un échantillon bootstrap*

Calculer $\hat{\gamma}_X^{*c}(b)$ *réplication bootstrap de* $\hat{\gamma}_X^c$

FinPour

Retourner

$$\widehat{MSE}_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\gamma}_X^{*c}(b) - \hat{\gamma}_X^c\right)^2 \tag{3.3.9}$$

Fin.

Estimation des Intervalles de confiance

Méthode des percentiles simples. les limites de confiance sont données par les pourcentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\hat{\gamma}_X^{*c}(b)$. L'algorithme est le suivant :

Algorithme 3.3.4 *Estimation Bootstrap de l'intervalle de confiance*

Variable

B: entier assez grand.

Début

Pour $b = 1$ *à* B ***faire***

Générer $Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$ *réalisation d'un echantillon bootstrap*

Calculer pour chacun les réplications bootstrap $\hat{\gamma}_X^{*c}(b)$.

Fin pour

Retourner

les stat. d'ordre $B(\alpha/2)$ *et* $B(1-\alpha/2)$ *percentile de* $\hat{\gamma}_X^{*c}(b)$ *dans la liste ordonnée des B réplications de* $\hat{\gamma}_X^{*c}$

$$\left[\hat{\gamma}_{X,B(\alpha/2)}^{*c} \quad ; \quad \hat{\gamma}_{X,B(1-\alpha/2)}^{*c}\right]$$

Fin.

3.3.2 Propriétés de l'estimateur de Kaplan-Meier $\hat{\gamma}^{*KM}_{X,Hill}$

Soit l'indice des valeurs extrêmes γ_1 associé à l'échantillon $(X_i)_{1\leq i\leq n}$, et soit $\hat{\gamma}^{*KM}_{X,Hill}$: une estimation de ce indice, obtenue à partir des données de l'échantillon initial

$$Z = \{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$$

Chaque échantillon

$$Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$$

obtenu par rééchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\hat{\gamma}^{KM}_{X,Hill}$.

$$\hat{\gamma}^{*KM}_{X,Hill}(b) \ , \qquad b = 1, ..., B$$

Estimation Bootstrap de l'erreur standard

Définition 3.3.4 *On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs* $\hat{\gamma}^{*KM}_{X,Hill}(b)$ *, la moyenne est:*

$$\hat{\gamma}^{*KM}_{X,Hill}(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}^{*KM}_{X,Hill}(b) \tag{3.3.10}$$

L'écart type est aussi une caractéristique importante de chaque distribution. Pour un ensemble d'estimateurs $\hat{\gamma}^{*KM}_{X,Hill}(b)$ *l'écart type estimé est calculé par la formule:*

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}^{*KM}_{X,Hill}(b) - \hat{\gamma}^{*KM}_{X,Hill}(.)\right)^2} \tag{3.3.11}$$

où B est le nombre total d'échantillons bootstrap .

Algorithme 3.3.5 *estimation bootstrap de l'erreur standard*

Variable

B:entier assez grand

Début

pour $b = 1$ ***à*** B ***faire***

on calcule l'estimateur de queue:

$$\hat{\gamma}^{*KM}_{X,Hill}(b).$$

On obtient alors un échantillon de B valeurs

$$\left\{\hat{\gamma}_{X,Hill}^{*KM}(1), \hat{\gamma}_{X,Hill}^{*KM}(2), ..., \hat{\gamma}_{X,Hill}^{*KM}(B)\right\}$$

On estime alors l'erreur standard $se_F\left(\hat{\gamma}_{X,Hill}^{KM}\right)$ par l'erreur standard de cet échantillon de $\hat{\gamma}_{X,Hill}^{*KM}$, i.e

$$\widehat{se}_F\left(\hat{\gamma}_{X,Hill}^{*KM}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_{X,Hill}^{*KM}(b) - \hat{\gamma}_{X,Hill}^{*KM}(.)\right)^2}$$

fin pour

Fin.

Estimation bootstrap du biais

Le biais d'un estimateur s'exprime comme

$$\widehat{Biais}_F\left(\hat{\gamma}_{X,Hill}^{KM}\right) = \mathbb{E}_{\hat{F}}\left[\hat{\gamma}_{X,Hill}^{KM}\right] - \hat{\gamma}_{X,Hill}^{KM}$$

Définition 3.3.5 *On appelle estimateur bootstrap du biais, l'estimateur de l'indice de queue pour les données observées*

$$\widehat{Biais}_{boot}\left(\hat{\gamma}_{X,Hill}^{KM}\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_{X,Hill}^{*KM}(b) - \hat{\gamma}_{X,Hill}^{KM} \tag{3.3.12}$$

Algorithme 3.3.6 *Estimation bootstrap du biais*

Variable

B:entier assez grand.

Début

Pour $b = 1$ *à* B ***faire***

Générer $Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$

Calculer $\hat{\gamma}_{X,Hill}^{*KM}(b)$.

Calculer $\hat{\gamma}_{X,Hill}^{*KM}(b) = \#\left\{j, (Z_i^{*b}, \delta_i^{*b}) = (Z_i, \delta_i)\right\} / n$ *pour tout* i.

Fin pour

Calculer $\hat{\gamma}_{X,Hill}^{*KM}(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_1^{*(c,H)}(b)$ *pour tout* i.

Retourner

$$\widehat{Biais}_{boot}\left(\hat{\gamma}_{X,Hill}^{*KM}\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_{X,Hill}^{*KM}(b) - \hat{\gamma}_{X,Hill}^{KM} \tag{3.3.13}$$

Fin.

Estimation Bootstrap de l'erreur quadratique moyenne

Définition 3.3.6 ***Monde réel.*** *l'erreur quadratique moyenne* (MSE) *de* $\hat{\gamma}_{X,Hill}^{KM}$ *est égale à*

$$MSE_F = \mathbb{E}_F\left[\left(\hat{\gamma}_{X,Hill}^{KM} - \gamma_X\right)^2\right]$$

Monde bootstrap. *l'estimateur bootstrap de l'erreur quadratique moyenne de* $\hat{\gamma}_{X,Hill}^{KM}$ *est défini par:*

$$\widehat{MSE}_{\hat{F}} = \mathbb{E}_{\hat{F}}\left[\left(\hat{\gamma}_{X,Hill}^{KM}(b) - \hat{\gamma}_{X,Hill}^{KM}\right)^2\right]$$

Algorithme 3.3.7 *estimation bootstrap de la MSE*

Variable

B *:entier assez grand*

Début

Pour b ***variant de*** 1 ***à*** B

Générer Z^{*b} *réalisation d'un échantillon bootstrap*

Calculer$\hat{\gamma}_{X,Hill}^{*KM}(b)$ *réplication bootstrap de* $\hat{\gamma}_{X,Hill}^{KM}$

FinPour

Retourner

$$\widehat{MSE}_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\gamma}_{X,Hill}^{*KM}(b) - \hat{\gamma}_{X,Hill}^{KM}\right)^2 \tag{3.3.14}$$

Fin.

Estimation des Intervalles de confiance

Méthode des percentiles simples. dans la méthode des percentiles simples, les limites de confiance sont données par les pourcentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\hat{\gamma}_{X,Hill}^{*KM}(b)$.

Algorithme 3.3.8 *Estimation Bootstrap de l'intervalle de confiance*

Variable

B:entier assez grand.

Début

Pour $b = 1$ *à* B ***faire***

Générer $Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$ *réalisation d'un echantillon bootstrap*

Calculer pour chacun les réplications bootstrap $\hat{\gamma}_{X,Hill}^{*KM}(b)$.

Fin pour

Retourner

les stat. d'ordre $B(\alpha/2)$ *et* $B(1-\alpha/2)$ *percentile de* $\hat{\gamma}_{X,Hill}^{*KM}(b)$ *dans la liste ordonnée des* B *réplications de* $\hat{\gamma}_{X,Hill}^{*KM}$

$$\left[\hat{\gamma}_{X,Hill,B(\alpha/2)}^{*KM} \quad ; \quad \hat{\gamma}_{X,Hill,B(1-\alpha/2)}^{*KM}\right]$$

Fin.

3.4 Simulations

3.4.1 Échantillon initial et paramètres de simulations

La loi de simulation utilisée dans ce cas est une loi de Pareto de paramètre γ de fonction de répartition,

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

Nous avons généré un échantillon $(X_i)_{1\leq i\leq n} \sim \text{Pareto}(\gamma_X)$ de taille $n = 1000$, à partir d'une variable u de $U([0,1])$, le modèle ajusté sera:

$$F^{-1}(u) = (1-u)^{-\gamma_X}.$$

L'échantillon $(X_i)_{1\leq i\leq n}$est censuré par un deuxième échantillon $(Y_i)_{1\leq i\leq n} \sim \text{Pareto}(\gamma_Y)$ à partir d'une variable v de $U([0,1])$:

$$G^{-1}(v) = (1-v)^{-\gamma_Y}.$$

Les variables que nous observons sont d'une part les $Z_i \sim \text{Pareto}(\gamma)$ définies par:

$$Z_i = X_i \wedge Y_i$$

les indicateurs de censure sont,

$$\delta_i = \mathbb{I}_{\{X_i \leq Y_i\}}$$

Programme sous R.

```
# paramètre de simulation
   n=1000
   gammaX=?
   gammaY=?
   alfa=?
   B=100
   # échantillon initial
   u<-runif(n,0,1)
   x<-(1-u)^(-gammaX)
   v<-runif(n,0,1)
   y<-(1-v)^(-gammaY)
   z<-pmin(x,y)
   delta=as.numeric(x<=y)
   z
   delta
   plot(z)
```

Calcul de k_{opt} par minimisation du MSE sous R.

pour le calcul directement la valeur de k_{opt} nous avons recueilli programme dans l'instruction suivant:

```
   minAMSE1(z)
   minAMSE2(z)
```

Estimateur bootstrap de $\hat{\gamma}_X^c$

```
#gamma hills censurées
```

```
gammaHC<-function (data,delta,kopt){
E<-sort(z)
Z<-log(E)
Delta<-ord(delta)
l<-seq(1,kopt)
hill1<-(((1/kopt)*sum(Z[n-l+1]))-Z[n-kopt])/((1/kopt)*sum(Delta[n-l+1]))
hill1
}
gammaHC(data,delta,kopt)
#gamma hills censurées Bootstrap
FgammaHCB<-function (data,delta,kopt,B){
hil<-seq(1,B)
for(s in 1:  B){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice];deltaboot<-delta[indice]
zboot
deltaboot
Deltaboot<- ord(deltaboot)
Zboot<-log(sort(zboot))
l<-seq(1,kopt)
hil[s]<-((1/kopt)*sum(Zboot[n-l+1])-Zboot[n-kopt])/((1/kopt)*sum(Deltaboot[n-l+1]))
}
hil
}
gammaHCB<-FgammaHCB(z,delta,kopt1,B)
hill.cen.boot<-mean(gammaHCB)

# Estimation Bootstrap de l'erreur standard
Sd<-sd(gammaHCB)
Sd

# Estimation bootstrap de biais
```

```
biais<-mean(gammaHCB)-gammaHC
biais

# Estimation bootstrap de l'EQMp
EQMboot<-(1/B)*sum((gammaHCB-gammaHC)^2)
EQMboot

# L'intervalle de confiancep
ICbootquantile<-function(alpha,B,f){
vectcroiss<-sort(f)
icinf<-vectcroiss[B*(alpha/2)]
icsup<-vectcroiss[B*(1-alpha/2)]
cbind(icinf,icsup) }
ICbootquantile(alfa,B,gammaHCB)
```

Estimateur bootstrap de $\hat{\gamma}^{KM}_{X,Hill}$

```
#G représentent les estimations de Kaplan-Meier
G<-function (t){
Delta<-ord(delta);z<-sort(z)
ind=function(z,t){ifelse(z<=t,1,0)}
f<-seq(1,n)
for(i in 1:n){f[i]<-(1-((ind(z,t)[i]*Delta[i])/(n-i+1)))}
f=prod(f)
f}
Fgammakm<-function(z,delta,kopt){
hil<-seq(1,n)
E<-sort(z)
Delta<-ord(delta)
v<-seq(1,kopt)
Z<-log(E)
u<-(1-G(E[n-kopt]))/kopt
```

```
for(i in 1:kopt){v[i]<-(Delta[n-i+1])*((1-F(E[n-i+1]))^-1)}
v
i<-seq(1,kopt)
hil<-u*sum((v[i])*(Z[n-i+1]/Z[n-kopt]))
hil
}
gammakm(z,delta,kopt)
#hills censurées Bootstrap
FHCKM<-function (data,delta,kopt,B){
hil<-seq(1,B)
for(s in 1:  B){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice];deltaboot<-delta[indice]
zboot
deltaboot
E<-sort(zboot)
Deltaboot<-ord(deltaboot)
Zboot<-log(E)
u<-(n*(1-G(E[n-kopt])))^-1
for(i in 1:kopt){v[i]<-(Delta[n-i+1])*((1-G(E[n-i+1]))^-1)}
v
i<-seq(1,kopt)
hil<-u*sum((v[i])*(Zboot[n-i+1]/Zboot[n-kopt]))
hil
}
gammakmb<-FHCKM(z,delta,kopt,B)
hill.cen.boot<-mean(gammakm)

#Estimation Bootstrap de l'erreur standard
Sd<-sd(gammakmb)
Sd
```

```
#Estimation bootstrap de biais
biais<-mean(gammakmb)-gammakm
biais

#Estimation bootstrap de l'EQMp
EQMboot<-(1/B)*sum((gammakmb-gammakm)^2)
EQMboot

# L'intervalle de confiancep
ICbootquantile<-function(alpha,B,f){
vectcroiss<-sort(f)
icinf<-vectcroiss[B*(alpha/2)]
icsup<-vectcroiss[B*(1-alpha/2)]
cbind(icinf,icsup) }
ICbootquantile(alfa,B,gammakmb)
```

3.4.2 Comportement de l'estimateur $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ de ses propiétés vs n

,nous traçons les valeurs de sd, $Biais$, MSE , ainsi que les intervalles de confiance IC versus n, en variant sa valeure de 100 jusqu'à 1000.

```
N=1000
alpha=0.05
B=1000
hill1n<-seq(1:N)
hilllboot<-seq(1:N)
Sd<-seq(1:s)
biais<-seq(1:N)
EQMboot<-seq(1:N)
icinf<-seq(1:N)
icsup<-seq(1:N)
for (n in 90:N)
```

```
{
gammaX=0.35
gammaY=1
u<-runif(n,0,1)
x<-(1-u)^(-gammaX)
v<-runif(n,0,1)
y<-(1-v)^(-gammaY)
z<-pmin(x,y)
delta=as.numeric(x<=y)
Delta<-ord(delta)
kopt1<-minAMSE1(z)
kopt2<-minAMSE2(z)
gammaHC(data,delta,kopt1)
FgammaHCB<-FgammaHCB(z,delta,kopt1,B)
gammakmb<-FHCKM(z,delta,kopt1,B)
gammakm<-Fgammakm(z,delta,kopt2)
gammakmb<-FHCKM(z,delta,kopt2,B)
gammaHCB[n]<-mean(FgammaHCB)
Sd1[n]<-sd(gammaHCB)
biais1[n]<-mean(gammaHCB)-gammaHC
EQMboot1[n]<-(1/B)*sum((gammaHCB-gammaHC)^2)
Sor<-sort(gammaHCB)
icinf1[n]<-Sor[B*(alpha/2)]
icsup1[n]<-Sor[B*(1-alpha/2)]
gammakmn[n]<-gammakm
Sd2[n]<-sd(gammakmb)
biais2[n]<-mean(gammakmb)-gammakm
EQMboot2[n]<-(1/B)*sum((gammakmb-gammakm)^2)
Sor<-sort(gammakmb)
icinf2[n]<-Sor[B*(alpha/2)]
```

```
   icsup2[n]<-Sor[B*(1-alpha/2)]
   gammakmn[n]<-gammakm
    }
   gammaHCn
   gammaHCB
   Sd1
   biais1
   EQMboot1
   icinf1
   icsup1
   gammakmn
   gammakmb
   Sd2
   biais2
   EQMboot2
   icinf2
   icsup2
   # estimateur Boot et intial
   plot( gammaHCB, xlim = c(130,1000), ylim = c(0.2,1),xlab=expression("n"),
type ="l", col = "blue")
   lines (gammaHCn, type ="l", col = "2" )
   abline(h=gammaX, col="3", lwd="1")
   # Sd, Biais, EQMboot
   plot(Sd1, xlim = c(125,1000), ylim = c(0,0.2),xlab=expression("n"), type ="l",col
= "3")
   plot(biais1, xlim = c(130,1000), ylim = c(-0.01,0.2),xlab=expression("n"),
type ="l",col = "4")
   plot(EQMboot1, xlim = c(130,1000),ylim = c(0,0.02),xlab=expression("n"), type
="l",col = "11")
   # intervalle de confiance
```

```
   plot(icinf1, xlim = c(110,1000), ylim=c(0.2,0.6),xlab=expression("n"), type
="l", col = "6")
   lines (gammaHCn,xlab=expression("n"), type ="l", col = "2" )
   lines (icsup1,xlab=expression("n"), type ="l", col = "6"
   # estimateur Boot et intial
   plot( gammakmb, xlim = c(130,1000), ylim = c(0.2,1),xlab=expression("n"),
type ="l", col = "blue")
   lines (gammakmn, type ="l", col = "2" )
   abline(h=gammaX, col="3", lwd="1")
   # Sd, Biais, EQMboot
   plot(Sd2, xlim = c(125,1000), ylim = c(0,0.2),xlab=expression("n"), type ="l",col
= "3")
   plot(biais2, xlim = c(130,1000), ylim = c(-0.01,0.2),xlab=expression("n"),
type ="l",col = "4")
   plot(EQMboot2, xlim = c(130,1000),ylim = c(0,0.02),xlab=expression("n"), type
="l",col = "11")
   # intervalle de confiance
   plot(icinf2, xlim = c(110,1000), ylim=c(0.2,0.6),xlab=expression("n"), type
="l", col = "6")
   lines (gammakmn,xlab=expression("n"), type ="l", col = "2" )
   lines (icsup2,xlab=expression("n"), type ="l", col = "6"
```

3.5 Résultats des simulations

3.5.1 Simulation bootstrap de l'estimateur $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ vs k

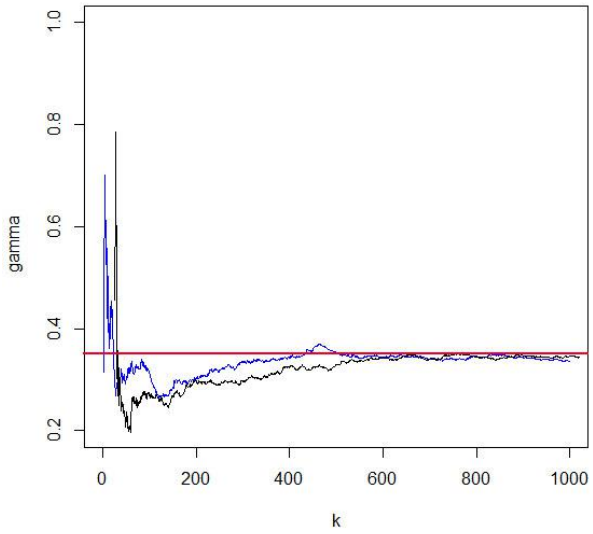

Figure 3.5.1 : Comportement graphique de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ vs k issue de la distribution de Pareto $(\gamma_X = 0.35)$ censurées par Pareto $(\gamma_Y = 2.5)$, (10% de censure)

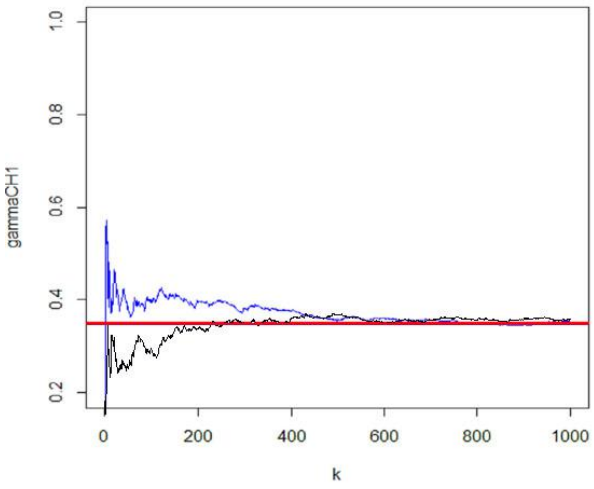

Figure 3.5.2 : Comportement graphique de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ vs k issue de la distribution de Pareto $(\gamma_X = 0.35)$ censurées par Pareto $(\gamma_Y = 0.5)$, (40% de censure)

Pour un pourcentage réduit de censure: on voit que l'estimateur $\hat{\gamma}_X^c$ et meiller que $\hat{\gamma}_{X,Hill}^{KM}$ jusqu'a $\left(k \leq \frac{n}{2}\right)$ et inversement (figue 3.5.1), pour un pourcentage grand de censure on doit que l'estimateur $\hat{\gamma}_{X,Hill}^{KM}$ et mailler que $\hat{\gamma}_X^c$ sue jus qu'a $\left(k \leq \frac{n}{2}\right)$ et inversement (figue 3.5.2), en général les deux estimateurs sont stables à partir de $\left(k \geq \frac{n}{2}\right)$

Les résultats de notre simulation bootstrap sont illustrés dans le tableau 3.1 et 3.2

Nous avons simuler deux échantillons de taille $n = 1000$ de pareto ($\gamma_X = 0.35$), censuré par un échantillon de taille $n = 1000$ de pareto. pour différentes valeurs du paramètre ($\gamma_Y = 0.5, 2.5$). Le caractère c représente le pourcentage de censure ($10\%, 40\%$).

	$c = 10\%$		$c = 40\%$	
k_{opt}	732		789	
$\hat{\gamma}_X^c$	0.3312207		0.3397039	
$\hat{\gamma}_{X_{boot}}^c$	0.3383147		0.3489045	
sd	0.01648889		0.01216952	
$biais$	0.007093974		0.0005311879	
MSE	0.000321936		0.0001482313	
IC	$ic\,\inf$	$ic\,\sup$	$ic\,\inf$	$ic\,\sup$
	0.3044285	0.3708987	0.3157617	0.3634473

Table 3.1: Les résultats l'estimater de l'indice de queue $\hat{\gamma}_X^{*c}$ par simulation bootstrap.

	$c = 10\%$		$c = 40\%$	
k_{opt}	786		538	
$\hat{\gamma}^{KM}_{X,Hill}$	0.3359939		0.3725373	
$\hat{\gamma}^{KM}_{X,Hill_{boot}}$	0.3418095		0.3615116	
sd	0.01948867		0.0360337	
$biais$	0.005815626		-0.06757429	
MSE	0.0004132501		0.005863413	
IC	$ic\,\inf$	$ic\,\sup$	$ic\,\inf$	$ic\,\sup$
	0.3251654	0.3724978	0.314904	0.3725373

Table 3.2: Les résultats l'estimater de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier $\hat{\gamma}^{KM}_{X,Hill}$ par simulation bootstrap.

3.5.2 Simulation bootstrap de l'estimateur $\hat{\gamma}^{c}_{X}$ vs n

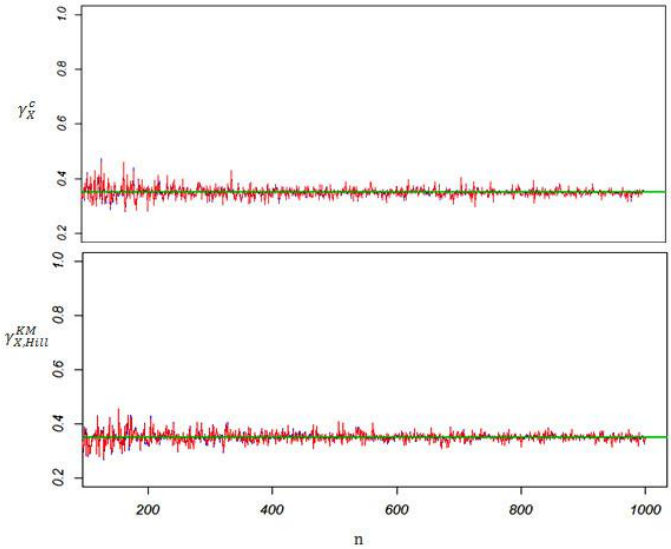

Figure 3.5.3 : $\hat{\gamma}^{c}_{X}$ et $\hat{\gamma}^{KM}_{X,Hill}$ bootstrap de 100 répétitions de Pareto ($\gamma_X = 0.35$) censurée par Pareto ($\gamma_Y = 1$) .

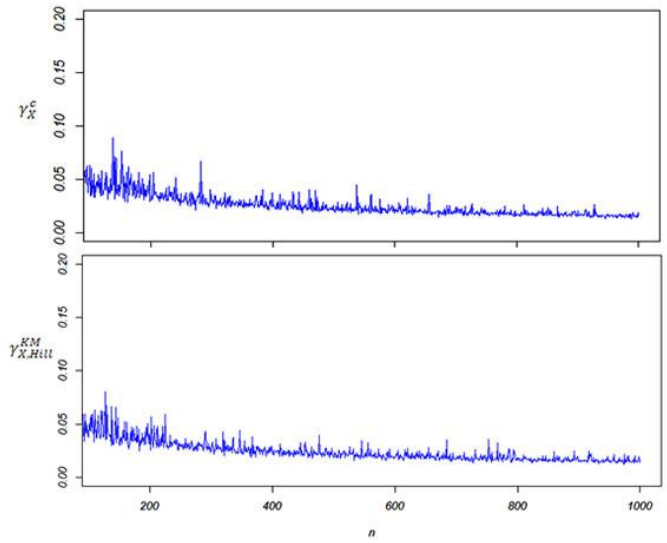

Figure 3.5.4 : *Biais* bootstrap de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ de 100 répétitions de Pareto ($\gamma_X = 0.35$) censurée par Pareto ($\gamma_Y = 1$).

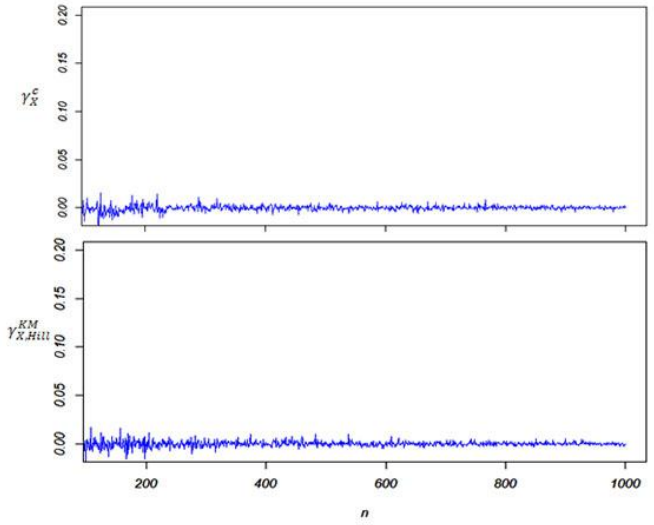

Figure 3.5.5 : *Biais* bootstrap de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ de 100 répétitions de Pareto ($\gamma_X = 0.35$) censurée par Pareto ($\gamma_Y = 1$).

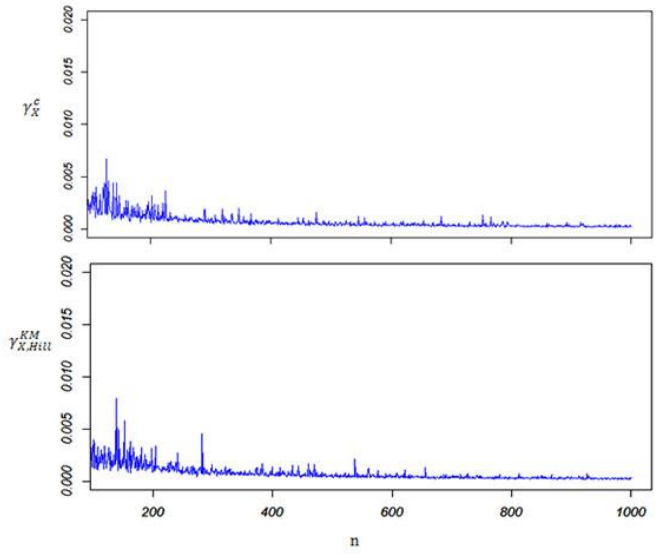

Figure 3.5.6 : MSE bootstrap de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ de 100 répétitions de Pareto ($\gamma_X = 0.35$) censurée par Pareto ($\gamma_Y = 1$) .

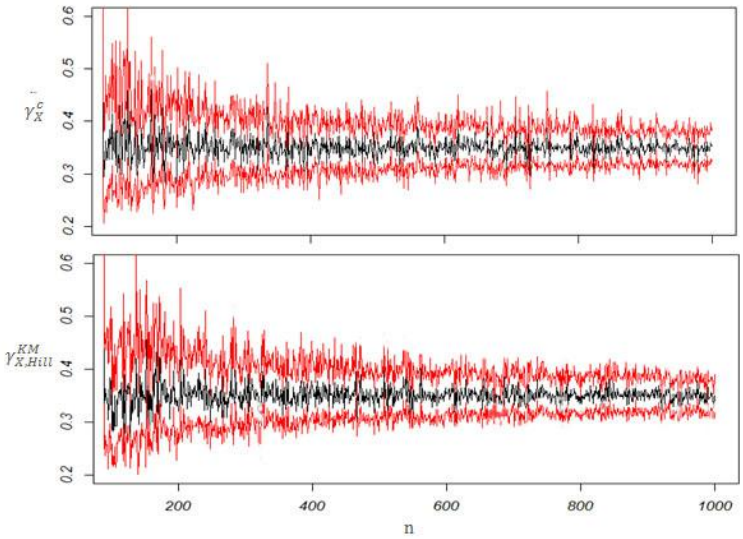

Figure 3.5.7 : IC bootstrap de $\hat{\gamma}_X^c$ et $\hat{\gamma}_{X,Hill}^{KM}$ de 100 répétitions de Pareto ($\gamma_X = 0.35$) censurée par Pareto ($\gamma_Y = 1$) .

à partir de la distribution limite Bootstrap nous avons tracé des intervalles de confiance en fonction de n qui est stable être très satisfisant

Conclusion.

Nous avons utilisé la méthode du Bootsratp sur deux estimateur de l'indice de queue $\hat{\gamma}_X^c$, et l'estimateur de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier $\hat{\gamma}_{X,Hill}^{KM}$, dans le cas du domaine d'attraction de Fréchet pour des données censurées à droite pour étudier les indicateurs de dispersion $(sd, Biais, MSE)$. Notre objectif était d'observer le comportement de chacun des deux estimateurs.

Pour le moments, nous n'avons pas consacré notre travaille pour le comparaison théorique exhaustive mais les simulations montrent que nouvelle version de Worms et Worms (2013) et meilleur en termes de biais et erreur quadratique moyenne.

Bibliographie

[1] **A. Guillou, P. Willems**,*Application de la Théorie des valeurs extrêmes en hydrologie,*2006

[2] **Anis Borchani**, *Statistiques des valeurs extrmes dans le cas de lois discretes,* December 2010.

[3] **Beirlant, J., Guillou, A., Dierckx, G. and Fils-Viletard, A**. *Estimation of the extreme value index and extreme quantiles under random censoring*, 2007.

[4] **Bradley Efron, R.J. Tibshirani**, *An Introduction to the Bootstrap*, 1993.

[5] **El Hadji Deme,***Quelques contributions a la Théorie univari ee des Valeurs Extrêmes et Estimation des mesures de risque actuariel pour des pertes a queues lourdes,* 2014.

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

[7] **Haan,L.,FerreiraA**, *Extreme Values Theory : An introduction.* New York,Springer, 2006.

[8] **Laurent Gardes et Stéphane Girard,** *Estimation de quantiles extrêmes pour les lois à queue de type Weibull*, 2013.

[9] **Magalie F., Myriam V.**, *Bootstrap et rééchantillonnage*, Université Européenne de Bretagne, 2012.

[10] **Philippe Saint Pierre**, *Introduction à l'analyse des durées de survie*,Université Pierre et Marie Curie, Février 2015.

[11] **Thierry Roncalli,** *Théorie des valeurs extrêmes*, Université Paris,Janvier 2002.

[12] **Worms ,J., Worms, R.**, *New estimators of the extreme value index under random right censoring, for heavy-tailed distributions*, 2013.

Résumé

Dans ce travail, nous nous concentrons sur le domaine d'attraction de Fréchet dans le cas des distributions à queues lourdes lorsque les données sont incomplètes. Nous nous intéressons à effectuer une comparaison de l'indice des valeurs extrêmes dans le cas des données censurées à droite de Einmahl et al, (2008) et celui de Worms et Worms (2013), basé sur les idées de l'intégration de Kaplan-Meier. Par des simulations effectués, nous avons appliqué la méthode du Re-échantillonnage pour les deux estimateurs dans le but de calculer les paramètres de dispersion, notamment pour vérifier les performances et l'efficacité de chacun d'entre eux.

Mots clés: indice des valeurs extrêmes, censure aléatoire à droite, distribution à queue lourde, domaine d'attraction de Fréchet, estimateur de Hill, méthode de Ré-échantillonnage.

Abstract

In this work, we focus on the domain of attraction of Fréchet in the case of heavy-tailed distributions when the data are incomplete. We are interested in comparing the index of extreme values in the case of data censored to the right of Einmahl et al, (2008) and that of Worms and Worms (2013), based on the ideas of the integration of Kaplan-Meier. By simulations carried out, we applied the method of Bootstrap for the two estimators in order to calculate the dispersion parameters, in particular to verify the performance and efficiency of each of them.

Key words: heavy-tailed distribution, random censoring, extreme value index, domain of attraction of Frechet, Hill estimator, bootstrap.

ملخص

في هذا العمل ركزنا على مجال جاذبية فريشي في حالة التوزيعات ذات الذيول الثقيلة في حالة البيانات غير المكتملة. نهتم بالمقارنة بين مؤشر القيم القصوى في حالة البيانات غير المكتملة المعدل من قبل انمهال وآخرون (2008)، وبين المؤشر المعدل من قبل ورمزو ورمز (2013) استنادا على افكار تكامل كبلان ماير. من خلال عملية المحاكات قمنا بتطبيق طريقة تكرار العينة على المقدرين لحساب مقاييس التشتت على غرار أداء وفاعلية كل مقدر.

الكلمات المفتاحية: مؤشر القيم القصوة، مجال جاذبية فريشي، التوزيعات ذات الذيول الثقيلة، تكرار العينة.