UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Mathématiques et des Sciences de la Matière

N° d'ordre :
N° de série :

DEPARTEMENT DE MATHEMATIQUES

MASTER

Spécialité : Mathématiques

Option : Probabilité et Statistique

Par :Bia Messaoud

Thème

Comparaison par simulations des estimateurs des l'indices des valeurs extrêmes conditionnel sous censure aléatoire à droite

Soutenu publiquement le : 30/09/2020

Devant le jury composé de :

Akti mouhamed	M.C.A Université KASDI Merbah-Ouargla	Président
Bousaad Abdelmalek	M.C.A Université KASDI Merbah-Ouargla	Examinateur
Fatima Meddi	M.C.B Université KASDI Merbah-Ouargla	Rapporteur

REMERCIEMENT

Tout d'abord, je remercie mon Dieu pour la santé, l'intelligence et le courage qu'Il m'accordé tout au long de ma cursus, en particulier durant la période de réalisation de ce projet.

Je voudrais dans un premier temps remercier, mon encadreur "FATIMA MEDDI" pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je remercie également toute l'équipe pédagogique de l'université Kasdi Merbah de Ouargla, pour tous leurs efforts durant ces années.

Je tiens à remercier toutes les personnes qui ont contribué au succès et m'ont aidée lors de la rédaction de ce mémoire.

Enfin, je remercie mes amis(es) qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

À tous ceux-ci, je présente mes remerciements, mon respect et ma gratitude.

Dédicace

Je dédie ce travail à :

Mes parents

-A mes frères et mes sœurs,et toute la famille "BIA"

- A mes chers amis "Aymen Bousbia" et "Slimani dia edine"

et "Said ahmed amine" ....

- Je tiens à remercier tous les membres de ma promotion

et surtout ma sœur "Djerouni Djouhaina"

-Et tous mes professeurs

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

TABLE DES MATIÈRES

ii

Abréviations et Notations **ii**

Introduction générale **ix**

1 Préliminaires mathématiques des valeurs extrêmes et censure **1**

1.1 Rappels sur la théorie des valeurs extrêmes 1

1.1.1 Statistiques d'ordre 1

1.1.2 Loi des valeurs extrêmes 3

1.1.3 Fonctions à variation régulière : 5

1.1.4 Domaines d'attraction 6

1.2 Rappel sur la censure 8

1.2.1 Les données de survie 8

1.2.2 Les données censurées 10

2 Estimation des indices des valeurs extrêmes **13**

2.1 Étude théorique des estimateurs 13

2.1.1 Estimateur de hill 15

2.1.2 Estimateur du moment 18

2.2 Rappels sur la modélisation des valeurs extrêmes conditionnelles 19

2.2.1 la fonction de répartition conditionnelle 19

2.2.2 Estimation de l'indice de queue conditionnel 20

2.2.3 Estimation quantiles extrêmes conditionnels 21

3 Comparaison des estimateurs par simulation 23

3.1 Principe du Bootstrap 23

3.2 Bootstrap des l'estimateurs 24

3.2.1 Propiétés de l'estimateur de hill $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$ 25

3.2.2 Propiétés de l'estimateur du moment $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$ 29

3.3 Simulation comparative 32

3.3.1 Échantillon initial et paramètres de simulations 32

3.3.2 Choix du nombre des valeurs extrêmes optimal k_n 34

3.3.3 Comportement de l'estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ de ses propiétés vs n 41

3.3.4 Résultats des simulations 46

Bibliographie 52

TABLE DES FIGURES

1.1 les fonctions de densité de Λ ,$\Phi_\gamma$et Ψ_γ 5

3.1 Comportement graphique $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs k issue de la distribution de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 2.5$), (10% de censure) 46

3.2 Comportement graphique $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs k issue de la distribution de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 0.5$), (40% de censure) 46

3.3 $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$) 49

3.4 Sd bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$) 49

3.5 Biais bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$) 50

3.6 MSE bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$) 50

3.7 IC bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$) 50

LISTE DES TABLEAUX

3.1 Les résultats l'estimater de hill $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ par simulation bootstrap. 48

3.2 Les résultats l'estimater du moment $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ par simulation bootstrap 48

ABRÉVIATIONS ET NOTATIONS

EVD	Distribution des valeurs extrêmes
IVE	Indice des valeurs extrêmes
F	Fonction de répartition
F_n	Fonction de répartition empirique
F^{-1}	Inverse généralisé de F
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
Λ	Loi de Gumbel
$\ell(x)$	Fonction à variation lente
DA	Domaine d'attraction de maxcimum
$M_n = X_{n,n}$	Maxcimum de $X_1, ..., X_n$
$p.s$	Prèsque sûre
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 $X_1, \ldots, X_n$

$O_P(.)$	Être borné en probabilité
VR_α	Variation régulière d'indice α
$\hat{\gamma}^{(c,H)}$	Estimateur de Hill avec les données censurées
$\hat{\gamma}^{(c,M)}$	Estimateur du moment avec les données censurées
$Z^* = (Z_1^*, ..., Z_n^*)$	Échantillon Bootstrap
MSE	L'erreur quadratique moyenne
IC	Intervalle de confiance
SD	Erreur standard
$\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$	Estimateur de hill conditionnel avec les données censurées
$\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$	Estimateur du moment conditionnel avec les données censurées
$(\Omega, \mathcal{A}, \mathbb{P})$	Espace probabilisé
τ_H	Point terminal
$\sup A$	Supremum de l'ensemble A
$s.o$	Statistique d'ordre
$v.a$	variable aléatoire
$X \wedge Y$	$\min(X, Y)$
x_F	Point terminal
$:=$	Égalité en définition
$\overset{D}{\rightarrow}$	Converge en distribution
$\overset{l}{\rightarrow}$	Converge en loi
$\overset{p}{\rightarrow}$	Converge en probabilité
$\overset{p.s}{\rightarrow}$	Converge prèsque sûre
$o_P(.)$	Converge vers 0 en probabilité

INTRODUCTION GÉNÉRALE

Au cours des dernières années, nous avons pu observer dans la recherche scientifique, une modelisation des événements rares. Ces événements rares sont des événements dont la probabilité d'apparition est trop faible c'est-àdire se trouve dans les queues des distributions. Ils apparaissent en général dans les contextes physiques nombreux et varies en particulier les catastrophes naturelles : en hydrologie crues décennales ou centennales et hauteur des barrages et digues susceptibles de les contenir, tempêtes occasionnant d'importants dommages matériels et environnementaux, dans les grands incendies, dans les tremblements de terre, dans les risques financiers (les kraks boursiers, les crises financières) dans les records sportifs, mais aussi dans l'étude de la résistance d'un matériau fibreux, etc. La théorie des valeurs extrêmes est une branche de la statistique qui essaie d'amener une solution face à ces phénomènes. Elle se repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Cependant, on y retrouve deux modèles : loi généralisée des extrêmes GEV : ≪ Generalized Extreme Value ≫ et loi de Pareto généralisée (GPD : ≪ Generalized Pareto Distribution ≫) Ainsi, tout a commencé avec les auteurs Fisher et Tippet (1928) quand ils étudiaient la résistance des fils de coton puis plus tard Gnedenko (1943) s'est intéressé à ces distributions. Ils ont énoncé un théorème fondamental avec la création de trois domaines d'attraction, domaine d'attraction de Fréchet, Gumbel et Weibull. Ce théorème intéressant fait référence à un paramètre appelé l'indice de queue qui donne la

forme de la queue de distribution. Von Mises (1954) puis Jenkinson (1955) ont rassemble les distributions de ces trois domaines en une seule écriture. C'est en ce moment que plusieurs auteurs se sont focalisés aux estimations de l'indice des valeurs extrêmes. Nous pouvons citer Hill (1975) dans le cas où l'indice est positif.

L'étude des données de survie est née au XVIIe siècle, dans le domaine de la démographie (biologie, épidémiologie, bio-statistique). L'objectif des analystes de cette époque était 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 réalisées qu'à partir du XIX ^{e} siècle, avec l'apparition de catégorisations suivant des variables exogènes (sexe, nationalité, catégories socioprofessionnelles, etc.), détermination de la probabilité de mourir à un âge donné. C'est par la suite que l'étude des données de survie commence à déborder le cadre strict de la démographie au XX ^{e} siècle au profit de toutes les disciplines susceptibles d'avoir recours à de tels types de données : finance (défaillance de crédit, intervalles entre cotations), la physique (avec l'apparition de la théorie de la fiabilité), l'industrie (pharmaceutique,biomédicale), sciences sociales (économie, sociologie, science politique), etc. C'est en ce moment que plusieurs auteurs se sont focalisés sur l'analyse des données de survie. Ainsi, en 1951, Weibull présente un modèle paramétrique dans le domaine de la fiabilité en proposant une loi dénommée

La modélisation des valeurs extrêmes censurées voit le jour en première fois en 1997 dans la littérature des extrêmes avec la sortie du livre Reiss et Thomas. Il a fallu qu'en 2007 Beirlant et al. abordent réellement la statistique non paramétrique des valeurs extrêmes avec des données censurées. Leur estimateur est basé sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dépassant un certain seuil donné. Ils ont appliqué cette théorie sur des données du SIDA. Puis, Einmahl et al. 2008 ont utilisé le même concept pour proposer un estimateur de l'indice de queue sur les k-plus grandes valeurs ensuite déterminer ses propriétés asymptotiques et enfin illustrer son comportement sur ces mêmes données du SIDA. Puis,

la recherche sur la théorie des valeurs extrêmes censurées est devenue une actualité.

Ce mémoire est réparti en trois chapitres.

1ère Chapitre : Nous présentons quelques rappels essentielles sur la théorie des valeurs extrêmes et aussi sur la notion de censure qui permettent de faciliter la lecture de mémoire.il s'agira de présenter brièvement les résultats essentiels rencontrés dans la littérature. Nous définissons rapidement les notions de domaine d'attraction, fonctions à variations régulières.Quant à la censure nous présenterons quelques définitions liées à la statistique des durées de survie. La censure est fondée avec quelques fonctions telles que la fonction de répartition, la fonction de survie

2ème Chapitre : Estimation des indices des valeurs extrêmes,nous présentons brièvement quelques estimateurs de l'indice des valeurs extrêmes en présence de censure aléatoire à droite, nous étudions propriétés asymptotiques le deux l'estimateurs hill et moment ensuite, nous présentons un rappels sur la modélisation des valeurs extrêmes conditionnelles (la fonction de répartition conditionnelle et estimation de l'indice de queue conditionnelle...), pour adaptation les deux estimateurs de l'indice conditionnel en présence de données censure aléatoire à droite dans le troisième chapitre pour comparaison

3ème Chapitre : Comparaison des estimations conditionnelles par simulation en présence de données censure aléatoire à droite,nous étudions et comparons les indicateurs de dispersion les estimateurs de l'écart-type, du biais et de l'erreur quadratique moyenne empiriques de l'estimateur de hill et l'estimateur du moment , nous parlons tout d'abord à l'application du bootstrap sur les deux estimateur, nous donnons dans un second temps les résultats de simulations et montrant une normaliée asymptotique confirmée empiriquement des estimateurs. Nous terminons en la quantile des deux estimateurs.

CHAPITRE 1

PRÉLIMINAIRES MATHÉMATIQUES DES VALEURS EXTRÊMES ET CENSURE

La théorie des valeurs extrêmes communément appelée "**Extreme Value Theory**" (EVT) en anglais, est une vaste théorie dont le but est d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'apparition est faible. Autrement dit elle essaie d'amener des éléments de réponses aux intempéries, aux inondations, aux catastrophes naturelles, aux problèmes financiers, etc en prédisant leurs occurrences dans les années à venir. En d'autres termes on veut estimer des petites probabilités ou des quantités dont la probabilité d'observation est très faible c'est-à-dire proche de zéro.

1.1 Rappels sur la théorie des valeurs extrêmes

1.1.1 Statistiques d'ordre

Définition 1.1.1 *Soient n variables aléatoires $(X_i)_{1\leq i\leq n}$ indépendantes et identiquement distribuées. Les statistiques d'ordre associées à l'échantillon $(X_1, \ldots, X_n)$ sont les réarrangements croissants des éléments de cet échantillon. Elle sont*

dénotées par $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ *et* $X_{i,n}$ *est la* $i^{\text{ième}}$ *statistique d'ordre (ou statistique d'ordre i).*

Dans un échantillon de taille n, deux statistiques d'ordre sont particulièrement intéressantes pour l'étude des événements extrêmes, le minimum et le maximum [1] :

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

• **Loi de** $X_{i,n}$ **:**

$$F_{X_{i,n}}(x) = \mathbb{P}\{X_{i,n} \leq x\} = \sum_{r=i}^{n} \begin{pmatrix} n \\ r \end{pmatrix} (F(x))^r (1 - F(x))^{n-r}$$

Nous en déduisons que la fonction de densité est :

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

où $f(x)$ est la densité de probabilité de X_i et F sa fonction de répartition associée. En utilisant la propriété d'indépendance des variables aléatoires $X_1, \ldots, X_n$ on obtient :

• **Loi de** $X_{1,n}$ **:**

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

d'où

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

• **Loi de** $X_{n,n}$ **:**

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

d'où

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

Remarque 1.1.2 *:*

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

quand $n \to \infty$

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

Soit $X_1, ..., X_n$ un n-échantillon issu d'une fonction de répartition commune F telle que,
$F(x) = \mathbb{P}(X_i < x)$, $i = 1, .., n$ et on note la fonction inverse généralisée de F par Q, telle que

$$Q(t) = F^{-1}(t) = \inf\{s, F(s) \geq t\} \quad 0 < t < 1$$

Définition 1.1.3 *(La fonction de répartition empirique) :*
La fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ notée F_n est donnée par :

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

Il existe une autre version de la défnition de F_n en utilisant les statistiques d'ordres comme suit :

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

1.1.2 Loi des valeurs extrêmes

Comme la fonction de répartition obtenue précédemment conduit à une loi dégénérée lorsque n tends vers l'infni, on recherche une loi non dégénérée pour le maximum de X. Cette loi limite non dégénérée est fournie par le "théorème des distributions extrêmes" qui donne une condition nécessaire et suffisante pour l'existence d'une loi limite non dégénérée pour le maximum. Ce théorème est proposé par Gnedenko (1943) qui donne la forme des lois limites et Jenkinson (1955) qui en donne l'expression générale.

Théorem 1.1.4 ***(Fisher et Tippett, 1928, Gnedenko, 1943)*** *:*
Soit $X_1, ..., X_n$ une suite de n variables aléatoires réelles indépendantes et identiquement distribuées de loi continue P et $M_n = \max(X_i)$. s'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$ et une fonction de répartition non-dégénérée G_γ telle que

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

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

$$\text{loi de Gumbel : } G_\gamma = \Lambda(x) = \exp(-\exp(-x)) \quad -\infty < x < +\infty$$

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

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

avec G_γ est la loi des valeurs extrêmes et γ est l'indice des valeurs extrêmes. a_n et b_n sont des paramètres de normalisation. Ce théorème est proposé par Gnedenko (1943) qui donne la forme des lois limites. Jenkinson (1955) donne l'expression générale notée GEV (Generalized Exêtreme Value Distribution) des trois distribution par

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

Remarque 1.1.5 *Si F vérife le théorème 1.1.4. On dit alors que F appartient au domaine d'attraction de G_γ et on note $F \in D(G_\gamma)$*

Remarque 1.1.6 *Les distributions Λ ,Φ_γ et Ψ_γ sont appelées les distributions de valeurs extrêmes et les variables aléatoires correspondantes sont les variables aléatoires extrêmales. Nous considérons maintenant les différences entre les trois distributions Λ ,$\Phi_\gamma$et Ψ_γ. Sur le (Figure 1.1), nous représentons les fonctions de densité de Λ ,$\Phi_\gamma$et Ψ_γ*

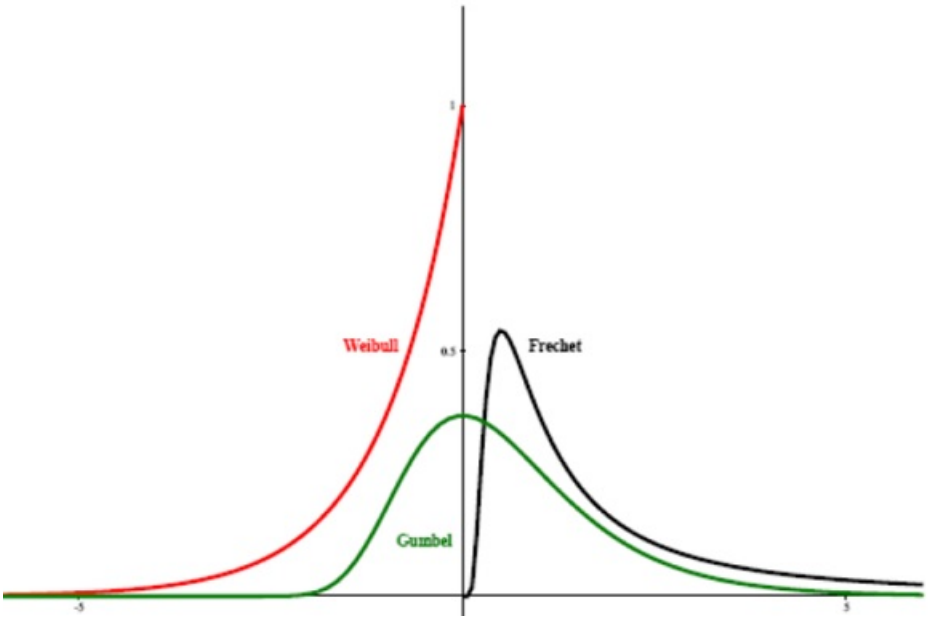

FIGURE 1.1 – les fonctions de densité de Λ ,$\Phi_\gamma$et Ψ_γ

1.1.3 Fonctions à variation régulière :

Définition 1.1.7 *Une fonction mesurable $f : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation régulière à l'infni si et seulement si, il existe un réel α tel que pour tout $x > 0$*

$$\lim_{u\to\infty} \frac{f(ux)}{f(u)} = x^\alpha$$

Et on note $f \in VR_\alpha$, α est appelé indice (ou exposant) de la fonction à variation régulière.Dans le cas particulier où $\alpha = 0$,on dit que f est à variation lente à l'infni, c'est à dire,

$$\lim_{u\to\infty} \frac{f(ux)}{f(u)} = 1 \quad \forall x > 0$$

Les fonctions à variation lente sont génériquement notées $\ell(x)$. Pour toute fonction à variation lente ℓ à l'infni, on a :

$$\lim_{x\to\infty} \frac{\log(\ell(x))}{\log(x)} = 0$$

Proposition 1.1.8 *soient $\alpha \in \mathbb{R}$ et $f \in VR_\alpha$: alors il existe une fonction à variation lente ℓ à l'infni telle que :*

$$\forall x > 0, f(x) = x^\alpha \ell(x)$$

Définition 1.1.9 *Une fonction de répartition F sur $\mathbb{R}$ appartient à une classe à variation régulière VR s'il existe $\alpha \geq 0$ tel que $1 - F \in VR_{-\alpha}$ sur $\mathbb{R}$, ou d'une manière équivalente :*

$$1 - F(x) \sim x^{-\alpha}\ell(x), \text{ quand } x \to \infty$$

pour certaines $\ell \in VR_0$

Théorem 1.1.10 *Une fonction $\ell : (0, \infty) \to (0, \infty)$ est à variation lente si et seulement si elle peut être écrite comme :*

$$\ell(x) = c(x) \exp\left\{\int_{x_0}^{x} \frac{\varepsilon(u)}{u} du\right\}, \quad x \geq x_0$$

pour certain $x_0 > 0$, on $\lim_{x\to\infty} c(x) = c \in (0, \infty)$ et $\lim_{x\to\infty} \varepsilon(x) = 0$. Des exemples typiques de fonctions à variation lente sont des constantes positives ou des fonctions convergeant vers une constante positive, logarithmes, puissances des logarithmes et logarithmes itérés.

1.1.4 Domaines d'attraction

Selon le signe de γ, on distingue trois cas de domaines d'attraction $(D.A)$[2] :

— Si $\gamma > 0$, F appartient au $D.A$ **de Fréchet**, et l'on note $F \in D.A$ **(Fréchet)**. Il contient toutes les lois dont la fonction de survie décroit comme une fonction puissance. Ce sont les lois à « queue lourde ». Les distributions du domaine de Fréchet sont beaucoup utilisées en fiabilité mécanique, dans les phénomènes climatiques tels que la météorologie, l'hydrologie, la vitesse du vent enregistrée en continu dans les aéroports et en finance dans les études de risque.

— Si $\gamma = 0$, F appartient au $D.A$ **de Gumbel**, et l'on note $F \in D.A.$ **(Gumbel)**. Ce sont les lois dont la fonction de survie décroit vers zéro à une vitesse exponentielle. Ces distributions sont souvent utilisées pour faire des prévisions dans les événements environnementaux tels que le séisme (le tremblement de terre), l'hydrologie (les inondations, la destruction des barrages), etc.

— $\gamma < 0$, F appartient au $D.A$ **de Weibull**, et l'on note $F \in D.A.$**(Weibull)**. Ce domaine regroupe toutes les lois dont le point terminal, $x_F = \inf\{x, F(x) \geq 1\}$ est fini. Les distributions de type de Weibull sont souvent utilisées pour décrire la résistance mécanique d'un matériau ou encore le temps de fonctionnement d'un appareil électronique ou mécanique.

pour certain $x_0 > 0$, où $\lim_{x\to\infty} c(x) = c \in (0, \infty)$ et $\lim_{x\to\infty} \varepsilon(x) = 0$. Des exemples typiques de fonctions à variation lente sont des constantes positives ou des fonctions convergeant vers une constante positive, logarithmes, puissances des logarithmes et logarithmes itérés.

Domaine d'attraction de Fréchet :

Théorem 1.1.11 *F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrêmes $\gamma > 0$ si et seulement si $x_F = +\infty$ et $1 - F$ est une fonction à variation régulière d'indice $1/\gamma$ c'est-à-dire :*

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

où ℓ est une fonction à variation lente. Dans ce cas, un choix possible pour les suites a_n et b_n est :

$$a_n = F^{-1}(1 - 1/n) \quad et \quad b_n = 0$$

Ce théorème permet de caractériser très simplement les distributions appartient au domaine d'attraction de Fréchet. En effet, elles doivent vérifer

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}$$

Prenons par exemple le cas de la distribution Pareto. Nous avons

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

Nous en déduisons que

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

donc $1 - F \in VR_{-1/\gamma}$

Domaine d'attraction de Weibull :

Théorem 1.1.12 *F appartient au domaine d'attraction de Weibul avec un indice de valeur extrême $\gamma < 0$ si et seulement si $x_F < +\infty$ et $1 - F^*$ est une fonction à variation régulière d'indice $1/\gamma$ avec :*

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

Dans ce cas, un choix possible pour les suites a_n et b_n est :

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

Domaine d'attraction de Gumbel :

Théorem 1.1.13 *Une fonction de répartition F appartient au domaine d'attraction de Gumbel si et seulement si :*

$$x_F \leq +\infty \quad et \quad \bar{F}(x) = c(x) \exp\left\{-\int_z^x \frac{g(t)}{a(t)} dt\right\}, z < x < x_F$$

où $c(x) \to c > 0$, $g(x) \to 1$ et $a^{'}(x) \to 0 si\ x \to x_F$.et un choix possible pour la fonction $a(.)$ est la fonction moyenne des excès définie par

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

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

$$b_n = F^{-1}(1 - 1/n) \quad et \quad a_n = a(b_n)$$

1.2 Rappel sur la censure

1.2.1 Les données de survie

L'analyse de survie, autrement dit la modélisation du temps de survenue d'un événement, apporte un outil principal d'évaluation théorique et pratique. L'analyse de ce type de données possède deux particularités intrinsèques, d'une part, celle ci ne concerne que des variables aléatoires positives et d'autre part, la présence de données non complètement observées comme nous l'expliquons ci dessous.

Désignons par X une variable d'intérêt, c'est à dire une variable aléatoire positive décrivant le temps qui s'écoule entre deux evénements par exemple [3]

- *En fiabilité* : durée de vie d'une lampe, durée de vie d'un matériel...
- *En biologie* : en culture de cellules les durées d'apparition de parasites...
- *En médecine* : durée de guérison d'un patient, durée de rémission d'un malade...
- *En economie* : durée de chômage...
- *En education* : durée d'apprentissage d'une langue...
- *En assurance* : durée de vie d'un contrat qui peut être défnie comme la diérence entre la date de résiliation et la date de création du contrat.

Nous donnons ci-dessous les définitions des fonctions utilisées habituellement en analyse des données de survie ([3],[4])

Définition 1.2.1 La « **durée de vie** » d'un individu est une variable aléatoire (v.a.) X positive et continue. Sa fonction de répartition

$$F(t) = \mathbb{P}(X \leq t)$$

est la probabilite que l'événement se produise entre 0 et t

Définition 1.2.2 La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire *:*

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \mathbb{P}(X > t) \quad t \geq 0. \end{aligned}$$

Définition 1.2.3 La fonction de risque instantané, pour t tixé représente 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) :

$$\begin{aligned} \lambda(t) &= \lim_{h \to 0} \frac{\mathbb{P}(t \leq X < t + h \mid X \geq t)}{h} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

où f est la densité de probabilité de X

Définition 1.2.4 La fonction de risque cumulé est l'intégrale du risque instantané définie par :

$$\Lambda(t) = \int_0^t \lambda(u)du$$

1.2.2 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 et de troncature. 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 de durée X, on observe la réalisation de la variable X soumise à diverses perturbations indépendantes ou non de l'événement étudié.

Définition 1.2.5 La variable de censure C est définie par la non-observation de l'événement étudié. Si au lieu d'observer X , on observe C, 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).

Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique. Pour un individu donné i, nous allons considérer [4]

- son temps de survie X_i de fonction de répartition F
- sa variable de censure C_i de fonction de répartition G
- sa variable réellement observée Z_i de fonction de répartition H.

Dans la littérature on distingue trois types de censure [4]

Censure à droite : La variable d'intérêt 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. Un exemple typique est celui où l'événement considéré est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation. On trouve aussi ce genre de phénomène dans les études de fiabilité quand la

panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant..., L'expérimentateur peut fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

Censure à gauche : Il y a 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. Par exemple si nous voulons étudier en fiabilité un certain composant électronique qui est branché en parallèle avec un ou plusieurs autres composants : le système peut continuer à fonctionner, quoique de façon aberrante, jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). Ainsi donc, la durée observée pour ce composant est censurée à gauche.

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. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[c, +\infty[$ et $[0, c]$ respectivement.

Ces trois catégories de censure décrites ci-dessus peuvent se présenter en fonction du mode ou mécanisme de censure. Ainsi, dans la littérature on retrouve les types suivants :

La censure de type I :

Définition 1.2.6 La censure est dite non-aléatoire du type I si, étant donné un nombre positif fixé C et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où

$$\begin{cases} Z_i = X_i \wedge C \\ \delta_i = \mathbb{I}_{\{X_i \leq C\}} \end{cases}$$

La censure de type II :

Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que r 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 $X_{(r)}$ et on observe les variables suivantes

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

Définition 1.2.7 La censure est dite non-aléatoire du type II si, étant donné un nombre positif fixé r et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où

$$\begin{cases} Z_i = X_i \wedge X_{(r)} \\ \delta_i = \mathbb{I}_{\left\{X_i \leq X_{(r)}\right\}} \end{cases}$$

La censure de type III : C'est la version aléatoire du type I

Définition 1.2.8 La censure est dite aléatoire du type I si, étant donné un n-échantillons $X_1, ..., X_n$, il existe un v.a. n-dimensionnelle $(C_1, ..., C_n)$ de $(\mathbb{R}^+)^n$ telle que les observations consistent en (Z_i, δ_i) où

$$\begin{cases} Z_i = X_i \wedge C_i \\ \delta_i = \mathbb{I}_{(X_i \leq C_i\}} \end{cases}$$

CHAPITRE 2

ESTIMATION DES INDICES DES VALEURS EXTRÊMES

2.1 Étude théorique des estimateurs

Définition 2.1.1 *[2] On dit que la fonction de distribution F est a queue lourde si la fonction du queue $\bar{F} := 1 - F$ est à variation régulière à l'infini d'indice $-1/\gamma$*

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

où γ est l'indice des valeurs extrêmes et ℓ est une fonction à variation lente. On considère l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et, soit l'échantillon $X_1, \ldots, X_n$ de variables aléatoires non-négatives définie sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ avec sa fonction de répartition F, supposons que la queue de distribution $1 - F$ est à variations réguliéres à l'infini d'indice $(-1/\gamma_1)$ noté $(1 - F \in RV(-1/\gamma_1))$, tel que :

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

L'échantillon $(X_i)_{1<i\leq n}$ n'est pas observé mais il est censuré par un deuxiéme échantillon $Y_1, \ldots, Y_n$, iid et sont supposés indépendants de l'échantillon X_i de fonction de répartition G, la queue de la distribution est à variations réguliéres

aussi : $1-G \in RV\left(-1/\gamma_2\right)$, *tel que :*

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

Les variables que nous observons sont d'une part les Z_i *définies par*

$$Z_i = X_i \wedge Y_i, \quad i=1,\ldots,n$$

et d'autre part les indicateurs de censure

$$\delta_i = \mathbb{I}_{\{X_i \leq Y_i\}}, \quad {}_i = 1,\ldots,n$$

Autrement dit, nous savons si la donnée observée a été censurée ou non, il est claire que les Z_i *sont des variables indépendantes de loi* H *liée à* F *et* G*, par la relation*

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

Le point terminal de H *est* $\tau_H=\sup\{x, H(x)<1\}$*. On a* F *et* G *satisfaisont la condition du domaine d'attraction de Fréchet :*

$$1-F(x)=x^{-1/\gamma_1}\ell_1(x) \text{ et } 1-G(x)=x^{-1/\gamma_2}\ell_2(x)$$

Alors

$$\begin{aligned}
1-H(x) &= (1-F(x))(1-G(x)) \\
&= x^{-1/\gamma_1}\ell_1(x)x^{-1/\gamma_2}\ell_2(x) \\
&= x^{-\left(\frac{1}{\gamma_1}+\frac{1}{\gamma_2}\right)}\ell_1(x)\ell_2(x) \\
&= x^{-\left(\frac{\gamma_1+\gamma_2}{\gamma_1\gamma_2}\right)}\ell(x) \\
&= x^{-1/\gamma}\ell(x)
\end{aligned}$$

où $(\ell(x)=\ell_1(x)\ell_2(x))$. *Donc* H *est une fonction de répartition appartenant au domaine d'attraction de Fréchet*

$$1-H(x)\in RV(-1/\gamma), \qquad \text{avec} \qquad \gamma=\gamma_1\gamma_2/\left(\gamma_1+\gamma_2\right)$$

Si F *et* G *sont supposées être dans le dommaine d'attraction maximales* $D\left(G_{\gamma_1}\right)$ *et* $D\left(G_{\gamma_2}\right)$ *respectivement où* $\gamma_1,\gamma_2\in\mathbb{R}$ *avec points terminales* τ_F *et* τ_G*, où* $\tau_F=\sup\{x, F(x)<1\}$*, alors cela signifie que* $H\in D\left(G_\gamma\right)$*. Einmahl et al. (2008,[6]) ont examiné les trois cas les plus intéressants suivants :*

$$\begin{cases}
\text{cas}\,1 : \gamma_1>0, \gamma_2>0 \\
\text{cas}\,2 : \gamma_1<0, \gamma_2<0, x_F=x_G \\
\text{cas}\,3 : \gamma_1=\gamma_2=0, x_F=x_G=\infty
\end{cases} \tag{2.1}$$

2.1.1 Estimateur de hill

Soit $\{(Z_i, \delta_i), 1 \leq i \leq n\}$ un échantillon de couple de v.a's (Z, δ). Soient $Z_{1,n} \leq \ldots \leq Z_{n,n}$ représente les statistiques d'ordre associées à l'échantillon $(Z_1, \ldots, Z_n)$, si l'on note l'administration concomitant de statistique d'ordre i par $\delta_{[i:n]}, \delta_{1:n}, \ldots, \delta_{n:n}$ les δ 's correspondants à $Z_{1,n}, \ldots, Z_{n,n}$ respectivement, i.e,

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

Beirlant et al. ont proposé différents estimateurs de γ_1, 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_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 t, c'est-à-dire à utiliser

$$\hat{\gamma}^{(c,)}_{X,k,n} = \frac{\hat{\gamma}^{(.)}_{Z,k,n}}{\hat{p}} \tag{2.2}$$

où

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

avec k le nombre d'excés au-delà de t. Et $\hat{p}$ estime $p = \frac{\gamma_2}{\gamma_1+\gamma_2}$ par conséquent $\hat{\gamma}^{(.)}_{Z,k,n}$ estimations γ divisé par $\frac{\gamma_2}{\gamma_1+\gamma_2}$ qui est égal à $\gamma_1.\hat{\gamma}^{(.)}_{Z,k,n}$ peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill $\hat{\gamma}^{(H)}_{Z,k,n}$. 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 adaptée du indice de queue

γ_1 est difini par :

$$\hat{\gamma}^{(c,H)}_1 = \frac{\hat{\gamma}^H}{\hat{p}}$$

où

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

alors

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

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 (2.2) 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 $Z_1, \ldots, Z_n$

Propriétés asymptotiques de l'estimateur de l'indice $\hat{\gamma}^{(c,H)}$

Pour déterminer les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes nous avons besoin de la fonction suivante comme difinie dans (Einmahl et al. (2008,[6])),

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

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 densites de F et G et on a :

$$\lim_{z \to \tau_H} p(z) = \frac{\gamma_2}{\gamma_1 + \gamma_2} := p \tag{2.3}$$

Pour expliquer (2.3) : Supposons X et Y suivent respectivement une loi de Pareto (γ_X) et Pareto(γ_Y) respectivement, C'est-à-dire pour tout $x \geq 1$.

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

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 $Z \sim \text{Pareto}\left(\gamma_X \gamma_Y / \left(\gamma_X + \gamma_Y\right)\right)$, nous pouvons à présent

calculer la fonction $p(z)$

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

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)).

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

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 $u > 0$

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

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

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

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

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

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

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

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

Théorem 2.1.2 *Sous les condition* $\mathbb{C}1 - \mathbb{C}4$ *et s 'il existe* b_0 *et* σ *telles que*

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

Alors, nous avons

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

telle que $b_0 = 1/(1-\rho)$ *et* $\sigma^2 = \gamma^2$ *nous avons les résultats asymptotiques de l'estimateurs de Hill.*

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

On a :

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

2.1.2 Estimateur du moment

Cet estimateur introduit par Dekkers et al (1989) [5] est une generalisation de l'estimateur de Hill, valable pour tout $\gamma \in \mathbb{R}$:

$$\hat{\gamma}_n^{(M)} := M_1 + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}$$

où

$$M_n^{(r)} := \frac{1}{k}\sum_{i=0}^{k}\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)^r ; r = 1, 2$$

γ_2 est difini par :

$$\hat{\gamma}_2^{(c,M)} = \frac{\hat{\gamma}^M}{\hat{p}}$$

où

$$\hat{\gamma}_n^{(M)} := M_1 + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1} \quad et \quad \hat{p} := \frac{1}{k}\sum_{i=1}^{k}\delta_{[n-i+1:n]}$$

alors

$$\hat{\gamma}_2^{(c,M)} = \frac{M_1 + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}}{k^{-1}\sum_{i=1}^{k}\delta_{[n-i+1:n]}}$$

Propriétés asymptotiques de $\hat{\gamma}_n^{(c,M)}$

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

(a) Consistance faible :

$$\hat{\gamma}_n^{(c,M)} \xrightarrow{p} \gamma \quad \text{quand } n \to \infty$$

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

$$\hat{\gamma}_n^{(c,M)} \xrightarrow{p.s} \gamma \quad \text{quand } n \to \infty$$

(c) Normalité asymptotique :

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

où

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

2.2 Rappels sur la modélisation des valeurs extrêmes conditionnelles

2.2.1 la fonction de répartition conditionnelle

Soient $(x_i)_{1\leq i\leq n}$, un échantillon déterministe et $(Y_i)_{1\leq i\leq n}$ et $(C_i)_{1\leq i\leq n}$ des réalisations respectives des variables aléatoires X, Y et C dans l'espace $(\Omega, \mathcal{A}, \mathbb{P})$. Nous supposons que Y et C ont respectivement pour fonctions de répartition F et G. Soient $(Z_i)_{1\leq i\leq n} = (Y_i \wedge C_i)_{1\leq i\leq n}$ et $\delta_i = \mathbb{I}_{\{Y_i \leq C_i\}}$, où $\mathbb{I}_{\{.\}}$ est une fonction indicatrice. Nous supposons aussi Y et C sont indépendantes conditionnellement à X. Ainsi, dire que nous sommes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $(Y_i)_{1\leq i\leq n}$ n'est pas observé mais est censuré par $(C_i)_{1\leq i\leq n}$. Nous observons uniquement $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ des copies indéplendantes de (Z, δ). Dans la suite nous supposerons que la variable Z a pour fonction de répartition H. Dans tout ce chapitre nous considérons que les fonctions de répartition conditionnelles de Y et C pour tout $x \in \mathcal{X}$ sont ainsi définies :

$$F(u \mid x) = 1 - u^{-1/\gamma_1(x)} L_1(u, x)$$
$$G(u \mid x) = 1 - u^{-1/\gamma_2(x)} L_2(u, x)$$

où $\gamma_1(.)$ et $\gamma_2(.)$ sont des paramètres fonctionnels positifs de la covariable x fixé, $L_1(.,x)$ et $L_2(.,x)$ sont des fonctions à variations lentes, i.e. pour tout $\lambda > 0$

$$\lim_{u\to\infty} \frac{L_i(\lambda u, x)}{L_i(u,x)} = 1, i = 1, 2$$

Si Y et C sont indépendantes, alors pour tout x, nous avons

$$\bar{H}(. \mid x) = \bar{F}(. \mid x)\bar{G}(. \mid x)$$

Alors

$$\bar{H}(u \mid x) = u^{-1/\gamma_1(x)} L_1(u,x) u^{-1/\gamma_2(x)} L_2(u,x) = u^{-\left(\frac{\gamma_1(x)+\gamma_2(x)}{\gamma_1(x)\gamma_2(x)}\right)} L_1(u,x) L_2(u,x)$$

Nous allons considérer

$$\gamma(x) = \frac{\gamma_1(x)\gamma_2(x)}{\gamma_1(x)+\gamma_2(x)} > 0 \text{ et } L(u,x) = L_1(u,x)L_2(u,x)$$

et

$$\lim_{u\to\infty} \frac{L(\lambda u, x)}{L(u,x)} = 1, \forall \quad \lambda > 0$$

car le produit de fonctions à variations lentes donne une fonction à variations lentes. Finalement,

$$H(u \mid x) = 1 - u^{-1/\gamma(x)} L(u,x)$$

donc H est une fonction de répartition conditionnelle appartenant au domaine d'attraction de Fréchet

2.2.2 Estimation de l'indice de queue conditionnel

si $x \in \mathcal{X}$ et $r > 0$, soit $B(x,r)$ la boule dans $\mathbb{R}^p$ de center x et de rayon r c'et à dire que, $B(x,r) = \{t \in \mathbb{R}^p, d(x,t) \leq r\}$. Soit $h_{n,x}$ une suite positive tendant à 0 lorsque n tend vers l'infini, et définir $m_{n,x} := \sum_{i=1}^{n} 1_{\{x_i \in B(x,h_{n,x})\}}$ (respectivement $\phi(h_{n,x}) := n^{-1} m_{n,x}$) comme le nombre (respectivement la proportion) des observations (Z_i, x_i) se trouvant dans $[0,\infty) \times B(x, h_{n,x})$. soit $Z^x_{(1)} \leq \ldots \leq Z^x_{(m_{n,x})}$ les valeur ordonnées de Z pour ces observations, et soit $\delta^x_{(1)}, \ldots, \delta^x_{(m_{n,x})}$ soit le correspondant δ (c'est-à-dire, $\delta^x_{(i)} = \delta_j$ if $Z^x_{(i)} = Z_j\big)$. Notez que puisque X est contrôlé, $m_{n,x}$ et $\phi(h_{n,x})$ sont des nombres non aléatoires. Si le $Z^x_{(i)}$ n'étaient pas censurés, Gardes and Girard (see [10]) proposer

d'estimer $\gamma_1(x)$ par un estimateur de type Hill du forme

$$\widehat{\gamma}^{(H)}_{k_x,m_{n,x}}(x) := M^{(1)}_{k_x,m_{n,x}} := \frac{1}{k_x}\sum_{i=1}^{k_x} i \log\left(\frac{Z^x_{(m_{n,x}-i+1)}}{Z^x_{(m_{n,x}-i)}}\right) \tag{2.4}$$

où k_x est un suite d'entires tels que $1 \leq k_x \leq m_{n,x}$. plusieurs autres les estimateurs de l'indices des valeurs extrême peuvent être adaptés à la situation et γ_1 est une fonction d'une covariable. par exemple, on peut considérer ce qui suit adapté version de l'estimateur du moment :

$$\widehat{\gamma}^{(M)}_{k_x,m_{n,x}}(x) := M^{(1)}_{k_x,m_{n,x}} + 1 - \frac{1}{2}\left(1 - \frac{\left(M^{(1)}_{k_x,m_{n,x}}\right)^2}{M^{(2)}_{k_x,m_{n,x}}}\right)^{-1} \tag{2.5}$$

et

$$M^{(2)}_{k_x,m_{n,x}} := \frac{1}{k_x}\sum_{i=1}^{k_x}\left(i \log\left(\frac{Z^x_{(m_{n,x}-i+1)}}{Z^x_{(m_{n,x}-i)}}\right)\right)^2$$

nous adaptons les estimateurs (2.4) et (2.5) dans le cas censure se produit (notez que ces estimateurs ne sont pas cohérents pour $\gamma_1(x)$ s'ils sont directement appliqués à l'échantillon $(Z_i, \delta_i, x_i)\,, i = 1, \ldots, n$. En effet, ils convergeront vers l'extrême indice de valeur $\gamma(x)$ de la distribution conditionnelle de Z). Pour tenir compte de la censure, nous divisons les estimateurs (2.4) and (2.5) par la proportion

$$\widehat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x} \delta^x_{(m_{n,x}-i+1)}$$

d'observations non censurées parmi les k_x les plus grands Z dans un voisinage de x (une idée similaire a été utilisée, par exemple, dans [11] et [6] pour estimer l'extrême indice de valeur sans covariable). Pour tout $x \in \mathcal{X}$, notre proposition est donc d'estimer $\gamma_1(x)$ par

$$\widehat{\gamma}^{(c,.)}_{k_x,m_{n,x}}(x) := \frac{\widehat{\gamma}^{(.)}_{k_x,m_{n,x}}(x)}{\widehat{p}_x}$$

2.2.3 Estimation quantiles extrêmes conditionnels

Soit Y une variable aléatoire réelle mesurée conjointement avec une covariable non-aléatoire $X \in \mathcal{X}$, où $\mathcal{X}$ est un espace métrique muni d'une distance d. Soit $\{(x_i, Y_i)\,, i = 1, \ldots, n\}$ un échantillon d'observations indépendantes et identiquement distribuées du couple $(X, Y) \in \mathcal{X} \times \mathbb{R}$. On suppose que la distribution conditionnelle de Y, F est à a queue lourde " c'est-à-dire

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

où $l \in \mathcal{RV}_0$ et $\gamma(\cdot)$ est une fonction positive et inconnue de la covariable x que l'on appelle "indice de queue conditionnel" ou "indice des valeurs extrêmes" on obtient l'échantillon retenu par la procédure :

$$\{Z_i^x, i = 1, \ldots, m_{n,x}\}$$

et son échantillon ordonné associé

$$Z_{(1)}^x \leq \ldots \leq Z_{(m_{n,x})}^x$$

Le quantile extrême conditionnel d'ordre $\alpha_{m_{n,x}}$ est ainsi défini :

$$\widehat{q}_1\left(\alpha_{m_{n,x}}, x\right) = Z^x\left(m_{n,x} - \left\lfloor m_{n,x}\alpha_{m_n,x|}\right\rfloor\right)$$

si $\alpha_{m_{n,x}}$ converge lentement vers 0, ie $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to \infty$ lorsque $m_{n,x} \to \infty$ où (a) désigne la partie entière de a. Par contre si $\alpha_{m_{n,x}}$ converge rapidement vers 0, ie $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to c \in [0, 1[$ lorsque $m_{n,x} \to \infty$, le quantile extrême conditionnel est ainsi défini :

$$\widehat{q}_2\left(\alpha_{m_{n,x}}, x\right) = \widehat{q}_1\left(\beta_{m_{n,x}}, x\right)\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)^{\widehat{\gamma}_n(x)}$$

où $\beta_{m_{n,x}}$ converge lentement vers 0 et $\hat{\gamma}_n(x)$ est un estimateur de l'indice des valeurs extrêmes conditionnel.

CHAPITRE 3

COMPARAISON DES ESTIMATEURS PAR SIMULATION

3.1 Principe du Bootstrap

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 F_n qui donne un poids $1/n$ à chaque réalisation . Ainsi on obtient un échantillon de taille n dit échantillon bootstrap selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales. Si n est grand, la distribution empirique F_n est proche de F, on aura donc une bonne approximation de la loi de X en utilisant F_n à la place de F. La méthode de bootstrap consiste à construire un nombre B(B entier) d'échantillons bootstrap notée X^* (images de l'échantillon initial), afin de les utiliser pour faire des inférences, et plus le nombre d'images simulées est grand, plus la statistique est précise, pour chaque nouvel échantillon, on calcule de la même façon un nouvel estimateur (image simulée de l'estimateur initial). L'ensemble des images simulées de l'estimateur initial est considéré comme un modèle de sa distribution sur la population de l'échantillon initial.

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.

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([7],[8],[9])

3.2 Bootstrap des l'estimateurs

Soit $X_1, X_2, \ldots, X_n, n$ 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 $Y_1, \ldots, Y_n$, les instants de censures associés, positives, de fonction de répartition G. On note $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$ 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 $\{(Z_{1:n}, \delta_{1:n}), \ldots, (Z_{n:n}, \delta_{n:n})\}$ l'échantillon ordonné suivant les valeurs de Z_i. Efron (1981) suggère le plan du re-échantillonnage suivant : On génère un échantillon bootstrap

$$(Z_1^*, \delta_1^*), \ldots, (Z_n^*, \delta_n^*)$$

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

$$(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$$

et soit $(Z^*_{i:n}, \delta^*_{i:n})_{i=1,\ldots,n}$ l'échantillon ordonné suivant les valeurs de Z^*_i. Si F et G sont sup- posées 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 $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$ et l'estimateur du moment $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$ construit avec les données $(Z^*_{i:n}, \delta^*_{i:n})_{i=1,\ldots,n}$ s'écrit :*

$$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x) := \frac{\frac{1}{k_x}\sum_{i=1}^{k_x} i \log\left(\frac{Z^{*x}_{(m_{n,x}-i+1)}}{Z^{*x}_{(m_{n,x}-i)}}\right)}{\frac{1}{k_x}\sum_{i=1}^{k_x} \delta^x_{(m_{n,x}-i+1)}} \quad et \quad \widehat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x} \delta^{*x}_{(m_{n,x}-i+1)}$$

et

$$\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x) = \frac{M^{*(1)}_{k_x,m_{n,x}} + 1 - \frac{1}{2}\left(1 - \frac{\left(M^{*(1)}_{k_x,m_{n,x}}\right)^2}{M^{*(2)}_{k_x,m_{n,x}}}\right)^{-1}}{\frac{1}{k_x}\sum_{i=1}^{k_x} \delta^x_{(m_{n,x}-i+1)}}$$

et

$$M^{*(2)}_{k_x,m_{n,x}} := \frac{1}{k_x}\sum_{i=1}^{k_x}\left(i \log\left(\frac{Z^{*x}_{(m_{n,x}-i+1)}}{Z^{*x}_{(m_{n,x}-i)}}\right)\right)^2$$

3.2.1 Propiétés de l'estimateur de hill $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$

Soit l'indice des valeurs extrêmes $\gamma_{(x)}$ associé à l'échantillon $(X_i)_{1\le i\le n}$, et soit $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ une estimation de ce indice, obtenue à partir des données de l'échantillon initial

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

Chaque échantillon

$$Z^{*b} = \left\{\left(Z^{*b}_1, \delta^{*b}_1\right), \ldots, \left(Z^{*b}_n, \delta^{*b}_n\right)\right\}$$

obtenu par rééchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$

$$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b), \quad b = 1, \ldots, B$$

Estimation Bootstrap de l'erreur standard :

On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ la moyenne est :

$$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$$

L'écart type est aussi une caractéristique importante de chaque distribution. Pour un ensemble d'estimateurs $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ l'ecart type estimé est calculé par la formule :

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(.)\right)^2}$$

où B est le nombre total d'échantillons bootstrap.

Algorithme 3.2.1 *estimation bootstrap de l'erreur standard*

La procédure bootstrap pour estimer l'erreur standard d'un estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ *est la suivante :*

1. On crée B échantillons indépendants $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ où chaque échantillon Z^{*b} est obtenu en tirant n observations avec remise dans l'échantillon $Z = \{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$ de départ.

2. Pour chaque échantillon b tel que, $1 \leq b \leq B$, on calcule l'estimateur de queue :

$$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$$

On obtient alors un échantillon de B valeurs

$$\left\{\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(1), \widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(2), \ldots, \widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(B)\right\}$$

3. On estime alors l'erreur standard $\mathrm{se}_F\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right)$ *par l'erreur standard de cet échantillon de* $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$ *i.e*

$$\widehat{se}_F\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(.)\right)^2}$$

où

$$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$$

Estimation bootstrap du biais

Définition 3.2.2 *Monde réel. le biais de l'estimateur de l'indice de queue s'exprime comme*

$$\mathrm{Biais}_F\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right) = E_F\left[\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right] - \gamma^{(c,H)}_{k_x,m_{n,x}}(x)$$

Monde bootstrap. l'estimateur bootstrap du biais de l'estimateur de l'indice de queue, est défini par

$$\widehat{\text{Biais}}_{\hat{F}}\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right) = E_{\hat{F}}\left[\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)\right] - \widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$$

Comme pour l'écart-type, il n'existe généralement pas d'expression analytique et il faut avoir recours à une approximation par simulation.

Algorithme 3.2.3 *Estimation bootstrap du biais*

Variable B : *entier assez grand.*

Début

Pour $b = 1$ *à* B ***faire***

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

Calculer $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$

Calculer $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) = \#\left\{j, \left(Z_i^{*b}, \delta_i^{*b}\right) = (Z_i, \delta_i)\right\}/n$ *pour tout* i

Fin pour

Calculer $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(i) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ *pour tout* i.

Retourner

$$\widehat{Biaisboot}\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(.)$$

Fin.

Estimation Bootstrap de l'erreur quadratique moyenne

Définition 3.2.4 *Monde réel : l'erreur quadratique moyenne* (MSE) *de* $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ *est égale à*

$$MSE_F = \mathbb{E}_F\left[\left(\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x) - \gamma_{(X)}\right)^2\right]$$

Monde bootstrap : l'estimateur bootstrap de l'erreur quadratique moyenne de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ *est defini par :*

$$\widehat{MSE}_{\hat{F}} = \mathbb{E}_{\hat{F}}\left[\left(\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right)^2\right]$$

Algorithme 3.2.5 *estimation bootstrap de la* MSE

Variable

B : *entier assez*

Début

Pour *b variant de 1 à B*

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

Calculer $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ *réplication bootstrap de* $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$

FinPour

Retourner

$$\widehat{MSE}_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)\right)^2$$

Fin

Estimation des Intervalles de confiance

Méthode des percentiles simples. les limites de confiance sont données par les pour- centiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ L'algorithme est le suivant :

Algorithme 3.2.6 *Estimation Bootstrap de l'intervalle de confiance*

Variable

B : entier assez grand.

Debut

Pour $b = 1$ *à* ***faire***

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

Calculer pour chacun les réplications bootstrap $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$

Fin pour

Retourner

les stat. d'ordre $B(\alpha/2)$ *et* $B(1-\alpha/2)$ *percentile de* $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ *dans la liste ordonnée des B réplications de* $\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$

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

Fin.

3.2.2 Propiétés de l'estimateur du moment $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$

Soit l'indice des valeurs extrêmes γ_2 associé à l'échantillon $(X_i)_{1\leq i\leq n}$, et soit $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$ une estimation de ce indice, obtenue à partir des données de l'échantillon initial

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

Chaque échantillon

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

obtenu par réchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$

$$\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b), \quad b = 1, \ldots, B$$

Estimation Bootstrap de l'erreur standard

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

$$\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(.) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$$

L'écart type est aussi une caractéristique importante de chaque distribution. Pour un ensemble d'estimateurs $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$ l'écart type estimé est calculé par la formule :

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(.)\right)^2}$$

où B est le nombre total d'échantillons bootstrap.

Algorithme 3.2.7 *estimation bootstrap de l'erreur standard*

Variable

B :entier assez grand

Début

pour *b = 1 á B* ***faire***

on calcule l'estimateur de queue :

$$\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$$

On obtient alors un échantillon de B valeurs

$$\left\{\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(1), \widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(2), \ldots, \widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(B)\right\}$$

On estime alors l'erreur standard $se_F\left(\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)\right)$ *par l'erreur standard de cet échantillon* de $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$, *i.e*

$$\widehat{se}_F\left(\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(.)\right)^2}$$

fin pour

Fin.

Estimation bootstrap du biais

Le biais d'un estimateur s'exprime comme

$$\widehat{\text{Biais}}_F\left(\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)\right) = \mathbb{E}_{\hat{F}}\left[\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)\right] - \widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$$

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

$$\widehat{\text{Biais}_{\text{boot}}}\left(\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)\right) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(.)$$

Algorithme 3.2.8 *Estimation bootstrap du biais*

Variable

B :entier assez grand.

Début

Pour $b = 1$ *à* B ***faire***

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

Calculer $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$

Calculer $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) = \#\left\{j, \left(Z_i^{*b}, \delta_i^{*b}\right) = (Z_i, \delta_i)\right\}/n$ *pour tout i.*

Fin pour

Calculer $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(i) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(b)$ *pour tout i.*

Retourner

$$\widehat{Biais_{boot}}\left(\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)\right) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(.)$$

Fin.

Estimation Bootstrap de l'erreur quadratique moyenne

Monde réel. l'erreur quadratique moyenne (MSE) de $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ est égale $\grave{a}$

$$MSE_F = \mathbb{E}_F\left[\left(\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x) - \gamma_{(X)}\right)^2\right]$$

Monde bootstrap. l'estimateur bootstrap de l'erreur quadratique moyenne de $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ est défini par :

$$\widehat{MSE}_{\hat{F}} = \mathbb{E}_{\hat{F}}\left[\left(\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(.)\right)^2\right]$$

Algorithme 3.2.9 *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 $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$ réplication bootstrap de $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(.)$

FinPour

Retourner

$$\widehat{MSE}_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b) - \widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(.)\right)^2$$

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 $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$

Algorithme 3.2.10 *Estimation Bootstrap de l'intervalle de confiance*

Variable

B :entier assez grand.

Début

Pour *$b = 1$ à B* ***faire***

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

Calculer pour chacun les réplications bootstrap $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$

Fin pour

Retourner

les stat. d'ordre $B(\alpha/2)$ *et* $B(1-\alpha/2)$ *percentile de* $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(b)$ *dans la liste ordonnée des* B *réplications de* $\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$

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

Fin.

3.3 Simulation comparative

3.3.1 Échantillon initial et paramètres de simulations

La loi de simulation utilisée dans ce cas est une loi de Pareto de paramètre $\gamma_{(X)}$ de fonction de répartition :

$$F(y/X) = 1 - y^{-1/\gamma_{(X)}} L(y, X)$$

Nous avons généré un échantillon $(X_i)_{1\leq i\leq n} \sim$ Pareto $\left(\gamma_{(X)}\right)$ 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 deuxieme échantillon $(Y_i)_{1\leq i\leq n} \sim$ Pareto $\left(\gamma_{(Y)}\right)$ 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 $Z_i \sim$ Pareto(γ) 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= ?
alpha= ?
# 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
r<-rank(z)
#ordonnéeslesdonnées
Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for(i in 1 :length(z))
{
L[i]<-(delta[i]+length(z))*r[i]
}
L
A<-sort(L)
A
for(i in 1 :length(z))
{
B[i]<-A[i]/r1[i]
}
```

B

Delta<-B-length(z)

Delta

Z<-log(Y)

3.3.2 Choix du nombre des valeurs extrêmes optimal k_n

Tous les estimateurs basés sur des valeurs extrêmes reposent essentiellement sur le nombre k des statistiques d'ordre supérieurs impliqués dans le calcul de l'estimation le fait que ce nombre localise où la queue de la distribution commence.

Les résultats asymptotiques des estimateurs de l'indice de queue sont en général obtenus lorsque $k \to \infty$ et $k/n \to 0$. 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. Ce probème a été longuement abordé dans la littérature. Plusieurs méthodes ont été développées, pour la choise de k, mais aucune n'est adoptée d'une manière générale.

Nous allons utiliser dans nos simulations la méthode de minimisation de l'erreur quadratique moyenne (MSE) pour déterminer la valeur optimale de k correspendante à l'estimateurs $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$

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 :

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

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 compro-

mis 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 $\{(k_n, MSE\,(k_n)\,, \quad k_n = 1, \ldots, n-1)\}$. La valeur optimale de k_n est donnée par

$$k_n^{opt} := \underset{1 \leq k \leq n-1}{\arg\min} \left\{ \frac{1}{R} \sum_{j=1}^{R} \left(\hat{\gamma}_{k_n}^{(c,H),j} - \gamma_1 \right)^2 \right\}$$

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 $(j = 1, \ldots, R)$ avec $(k_n = 1, \ldots, n-1)$. Il semble donc naturel de trouver une valeur k_{nopt} qui minimise les valeurs de l'erreur quadratique moyenne par rapport a k_n

Calcul de k^{opt} par minimisat MSE sous R

Le programme suivant sous R calcul directement la valeur de k^{opt}

```
1ère calcul kopt1
rm(list=ls())
R=300
a<-flooor(n/5)
b<-floor(n-(n/5))
EQM<-seq(1 :b)
for (k in 1 :b)
{
hills<-seq(1,R)
for(j in 1 :R){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
zboot
deltaboot
r<-rank(zboot)
Y<-sort(zboot)
r1<-sort(r)
```

```
L<-seq(1,length(zboot))
B<-seq(1,length(zboot))
for (i in 1 :length(zboot))
{
L[i]<-(deltaboot[i]+length(zboot))*r[i]
}
L
A<-sort(L)
for (i in 1 :length(zboot))
{
A
B[i]<-A[i]/r1[i]
}
B
Deltaboot<-B-length(zboot)
Deltaboot
Zboot<-log10(Y)
l<-seq(1,k)
hills[j]<-(((1/k)*sum(l*(Zboot[n-l+1]))-Zboot[n-
l]))/((1/k)*(sum(Deltaboot[n-l+1])))
hills
}
hills
EQM[k]<-(1/R)*sum((hills-gammaX)^2)
}
EQM
kopt1<-which.min(EQM[a :b])+a-1
kopt1
```

$2^{éme}$ pour le calcul directement la valeur de kopt nous avons recueilli programme dans l'instruction suivant :

```
kopt1<- minAMSE1(z)
kopt2<- minAMSE2(z)
```

Estimateur bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$

Pour construire des échantillons bootstrapés, on utilise la commande :

```
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
# estimateur initial du kopt
l<-seq(1,kopt1)
hill1<-(((1/kopt1)*sum(l*(Z[n-l+1]))-Z[n-l]))/((1/kopt1)*sum(Delta[n-l+1]))
hill1
```

```
#gamma hills conditionnels censurées
R1=1000
hillls<-seq(1,R1)
for(s in 1 : R1){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
zboot
deltaboot
r<-rank(zboot)
Y<-sort(zboot)
r1<-sort(r)
L<-seq(1,length(zboot))
B<-seq(1,length(zboot))
for (i in 1 :length(zboot))
{
L[i]<-(deltaboot[i]+length(zboot))*r[i]
}
L
A<-sort(L)
for (i in 1 :length(zboot))
{
A
B[i]<-A[i]/r1[i]
```

```
}
B
Deltaboot<-B-length(zboot)
Deltaboot
Zboot<-log10(Y)
l<-seq(1,kopt1)
hillls[s]<-((1/kopt1)*sum(l*(Zboot[n-l+1])-Zboot[n-l]))/((1/kopt1)*sum(Deltaboot[n-
l+1]))
}
rééé<-cbind(hillls)
rééé
hilllboot<-mean(hillls)
# Estimation Bootstrap de l'erreur standard
Sd<-sd(hillls)
Sd
# Estimation bootstrap de biais
biais<-mean(hillls)-hill1
biais
# Estimation bootstrap de l'EQMp
EQMboot<-(1/R1)*sum((hillls-hill1)^2)
EQMboot
# L'intervalle de confiancep
ICbootquantile<-function(alpha,R1,f){
vectcroiss<-sort(f)
icinf<-vectcroiss[R1*(alpha/2)]
icsup<-vectcroiss[R1*(1-alpha/2)]
cbind(icinf,icsup) }
ICbootquantile(0.05,1000,hillls)
```

Estimateur bootstrap de $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$

```
# estimateur initial du kopt2
u<-seq(1,kopt2)
moment1<-((((1/kopt2)*sum(u*(Z[n-u+1]))-Z[n-u]))+1-0.5*(1-((1/kopt2)*sum(
u*(Z[n-u+1]))-Z[n-u])))^2/((1/kopt2)*sum(u*(Z[n-u+1]))-Z[n-u]))^2) )^-1 )/((
1/kopt2)*sum(Delta[n-u+1]))
moment1
#gamma moment conditionnels censurées
R1=1000
momentt<-seq(1,R1)
for(t in 1 : R1){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
zboot
deltaboot
r1<-rank(zboot)
Y1<-sort(zboot)
r2<-sort(r1)
L1<-seq(1,length(zboot))
B1<-seq(1,length(zboot))
for (i in 1 :length(zboot))
{
L1[i]<-(deltaboot[i]+length(zboot))*r1[i]
}
L
A<-sort(L)
for (i in 1 :length(zboot))
{
A1
B1[i]<-A1[i]/r2[i]
}
B1
```

```
Deltaboot<-B1-length(zboot)
Deltaboot
Zboot<-log10(Y1)
u<-seq(1,kopt2)
momentt[t]<-((((1/kopt2)*sum(u*(Zboot[n-u+1]))-Zboot[n-u]))+1-0.5*(1-((1/kopt2)*sum(
u*(Zboot[n-u+1]))-Zboot[n-u])))^2/((1/kopt2)*sum(u*(Zboot[n-u+1]))-Zboot[n-
u]))^2) )^-1 )/((1/kopt2)*sum(Deltaboot[n-u+1]))
}
réé1<-cbind(momentt)
réé1
momenttboot<-mean(momentt)
# Estimation Bootstrap de l'erreur standard
Sd<-sd(momentt)
Sd
# Estimation bootstrap de biais
biais<-mean(momentt)-moment1
biais
# Estimation bootstrap de l'EQMp
EQMboot<-(1/R1)*sum((momentt-moment1)^2)
EQMboot
# L'intervalle de confiancep
ICbootquantile<-function(alpha,R1,f){
vectcroiss<-sort(f)
icinf<-vectcroiss[R1*(alpha/2)]
icsup<-vectcroiss[R1*(1-alpha/2)]
cbind(icinf,icsup) }
ICbootquantile(0.05,1000,momentt)
```

3.3.3 Comportement de l'estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ de ses propiétés vs n

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

```
N=1000
alpha=0.05
R1=1000
hill1n<-seq(1 :N)
hilllboot<-seq(1 :N)
moment1n<-seq(1 :N)
momentboot<-seq(1 :N)
Sd<-seq(1 :N)
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)
r<-rank(z)
Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for (i in 1 :length(z))
```

```
{
L[i]<-(delta[i]+length(z))*r[i]
}
A<-sort(L)
for (i in 1 :length(z))
{
B[i]<-A[i]/r1[i]
}
Delta<-B-length(z)
Z<-log(Y)
# calcule k^opt le deux estimateur par l'instruction suivant :
kopt1<- minAMSE1(z)
kopt2<- minAMSE2(z)
l<-seq(1,kopt1)
hill1<-(((1/kopt1)*sum(l*(Z[n-l+1]))-Z[n-l]))/((1/kopt1)*sum(Delta[n-l+1]))
hillls<-seq(1,R1)
for(N in 1 : R1){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
r<-rank(zboot)
Y<-sort(zboot)
r1<-sort(r)
L<-seq(1,length(zboot))
B<-seq(1,length(zboot))
for (i in 1 :length(zboot))
{
L[i]<-(deltaboot[i]+length(zboot))*r[i]
}
A<-sort(L)
for (i in 1 :length(zboot))
{
B[i]<-A[i]/r1[i]
```

```
} Deltaboot<-B-length(zboot)
Zboot<-log(Y)
hills[s]<-((1/kopt1)*sum(l*(Zboot[n-l+1])-Zboot[n-l]))/((1/kopt1)*sum(Deltaboot[n-
l+1]))
}
hilllboot[n]<-mean(hills)
Sd1[n]<-sd(hills)
biais1[n]<-mean(hillls)-hill1
EQMboot1[n]<-(1/R1)*sum((hillls-hill1)^2) Sor<-sort(hillls)
icinf1[n]<-Sor[R1*(alpha/2)]
icsup1[n]<-Sor[R1*(1-alpha/2)]
hill1n[n]<-hill1
}
u<-seq(1,kopt2)
moment1<-((((1/kopt2)*sum(u*(Z[n-lu+1]))-Z[n-u]))+1-0.5*(1-((1/kopt)*sum(
u*(Z[n-u+1]))-Z[n-kopt2])))^2/((1/kopt2)*sum(u*(Z[n-u+1]))-
Z[n-u]))^2) )^-1 )/((1/kopt2)*sum(Delta[n-u+1]))
momentt<-seq(1,R1)
for(N in 1 : R1){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
r1<-rank(zboot)
Y1<-sort(zboot)
r2<-sort(r1)
L1<-seq(1,length(zboot))
B1<-seq(1,length(zboot))
for (i in 1 :length(zboot))
{
L1[i]<-(deltaboot[i]+length(zboot))*r1[i]
}
A1<-sort(L1)
for (i in 1 :length(zboot))
```

```
{
B1[i]<-A1[i]/r2[i]
} Deltaboott<-B1-length(zboot)
Zboott<-log(Y1)
{
momentt[N]<-((((1/kopt2)*sum(u*(Zboott[n-u+1]))-Zboott[n-u]))+1-0.5*(1-
((1/kopt2)*sum(u*(Zboott[n-u+1]))-Zboott[n-u])))^2/((1/kopt2)*sum(u*(Zboott[n-
u+1]))-Zboott[n-u]))^2) )^-1 )/((1/kopt2)*sum(Deltaboott[n-u+1])) }
{
momentboot[n]<-mean(momentt)
Sd2[n]<-sd(momentt)
biais2[n]<-mean(momentt)-moment1
EQMboot2[n]<-(1/R1)*sum((momentt-moment1)^2) Sor<-sort(momentt)
icinf2[n]<-Sor[R1*(alpha/2)]
icsup2[n]<-Sor[R1*(1-alpha/2)]
moment1n[n]<-moment1
}
hill1n
hilllboot
Sd1
biais1
EQMboot1
icinf1
icsup1
moment1n
momentboot
Sd2
biais2
EQMboot2
icinf2
icsup2
# estimateur Boot et intial
```

```
plot( hilllboot, xlim = c(200,1000), ylim = c(0.2,1),xlab=expression("n"), type
="l", col = "blue")
lines (hill1n, type ="l", col = "2" )
abline(h=gammaX, col="3", lwd="1")
# Sd, Biais, EQMboot
plot(Sd1, xlim = c(200,1000), ylim = c(0,0.2),xlab=expression("n"), type ="l",col
= "4") plot(biais1, xlim = c(200,1000), ylim = c(0,0.2),xlab=expression("n"),
type ="l",col = "4")
plot(EQMboot1, xlim = c(200,1000),ylim = c(0,0.02),xlab=expression("n"),
type ="l",col = "4")
# intervalle de confiance
plot(icinf1, xlim = c(200,1000), ylim=c(0.2,0.6),xlab=expression("n"), type ="l",
col = "2")
lines (hill1n,xlab=expression("n"), type ="l", col = "black" )
lines (icsup1,xlab=expression("n"), type ="l", col = "2")
# estimateur Boot et intial
plot( momentbot, xlim = c(200,1000), ylim = c(0.2,1),xlab=expression("n"),
type ="l", col = "blue")
lines (moment1n, type ="l", col = "2" )
abline(h=gammaX, col="3", lwd="1")
# Sd, Biais, EQMboot
plot(Sd2, xlim = c(200,1000), ylim = c(0,0.2),xlab=expression("n"), type ="l",col
= "4")
plot(biais2, xlim = c(200,1000), ylim = c(0,0.2),xlab=expression("n"), type
="l",col = "4")
plot(EQMboot2, xlim = c(200,1000),ylim = c(0,0.02),xlab=expression("n"),
type ="l",col = "4")
# intervalle de confiance
plot(icinf2, xlim = c(200,1000), ylim=c(0.2,0.6),xlab=expression("n"), type ="l",
col = "2")
lines (hill1n,xlab=expression("n"), type ="l", col = "6" )
lines (icsup2,xlab=expression("n"), type ="l", col = "2")
```

3.3.4 Résultats des simulations

Simulation bootstrap de l'estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs k

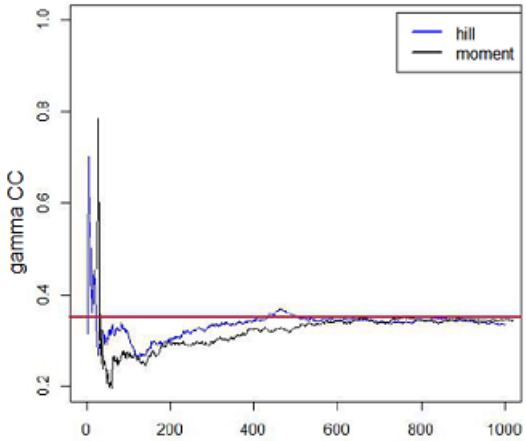

FIGURE 3.1 – Comportement graphique $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs k issue de la distribution de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 2.5$), (10% de censure)

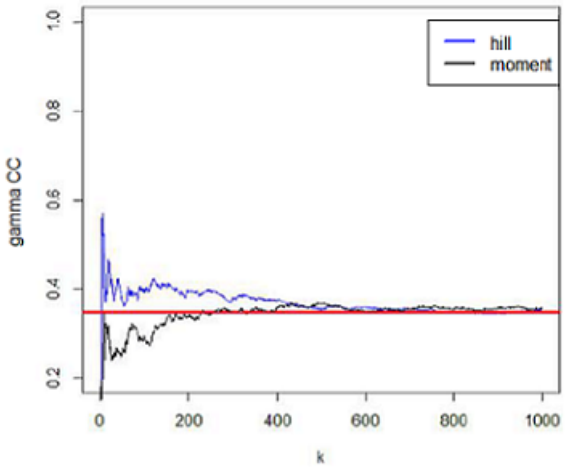

FIGURE 3.2 – Comportement graphique $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs k issue de la distribution de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 0.5$), (40% de censure)

Pour un pourcentage réduit de censure : on voit que l'estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et meiller que $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ jusqu'a $\left(k \leq \frac{n}{2}\right)$ et inversement (figue1 3.3.4) pour un pourcentage grand de censure on doit que l'estimateur $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ et mailler que $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ sue jus qu'a $\left(k \leq \frac{n}{2}\right)$ et inversement (figue2 3.3.4), 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 $(\gamma_X = 0.35)$, censurée par un échantillon de taille $n = 1000$ de paréto.et C est la variable de censure de loi Paréto de paramètre fonctionnel γ_Y défini selon le pourcentage de censure

$$\gamma_Y = \begin{cases} 0.5, & \text{si} \quad \% \text{ de censure } \simeq 40\% \\ 1, & \text{si} \quad \% \text{ de censure } \simeq 25\% \\ 2.5, & \text{si} \quad \% \text{ de censure } \simeq 10\% \end{cases}$$

	c=10%		c=25%		c=40%	
k_{opt}	645		633		682	
$\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$	0.3499868		0.3499991		0.3398567	
$\widehat{\gamma}^{*(c,H)}_{k_x,m_{n,x}}(x)$	0.3446431		0.3419632		0.3402936	
Sd	0.01579789		0.01346219		0.01511182	
biais	0.01406515		0.01484466		0.01205823	
MSE	0.0003920529		0.0003942478		0.0004312249	
IC	$ic\inf$	$ic\sup$	$ic\inf$	$ic\sup$	$ic\inf$	$ic\sup$
	0.3371476	0.3596124	0.336105	0.3551854	0.3347744	0.3659637

TABLE 3.1 – Les résultats l'estimater de hill $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ par simulation bootstrap.

	c=10%		c=25%		c=40%	
k_{opt}	667		586		612	
$\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$	0.3499866		0.3600052		0.3701057	
$\widehat{\gamma}^{*(c,M)}_{k_x,m_{n,x}}(x)$	0.3391044		0.3615116		0.3690599	
Sd	0.01439509		0.01650337		0.01501585	
biais	0.00360337		0.00474870		0.004041339	
MSE	0.0005997496		0.0005863413		0.0004304698	
IC	$ic\inf$	$ic\sup$	$ic\inf$	$ic\sup$	$ic\inf$	$ic\sup$
	0.3200737	0.3629636	0.3276725	0.3658161	0.3292755	0.3643591

TABLE 3.2 – Les résultats l'estimater du moment $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ par simulation bootstrap

Simulation bootstrap de l'estimateur $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ vs n

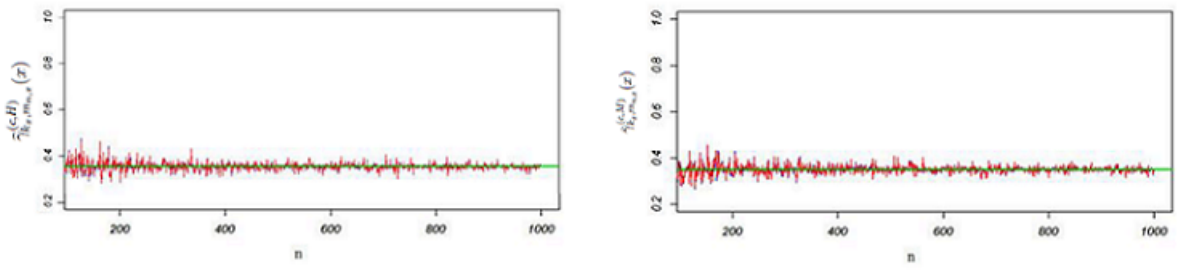

FIGURE 3.3 – $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$)

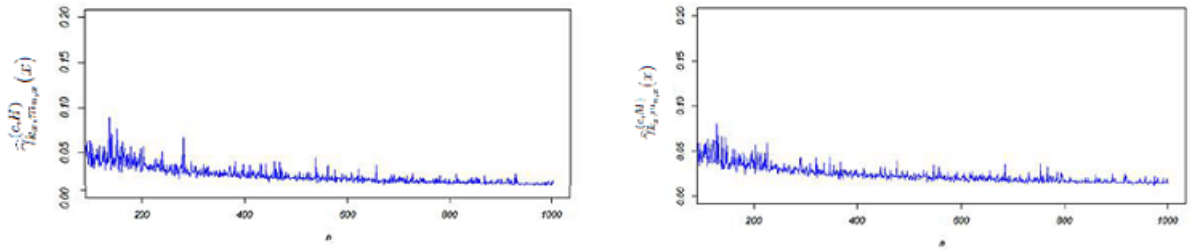

FIGURE 3.4 – Sd bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$)

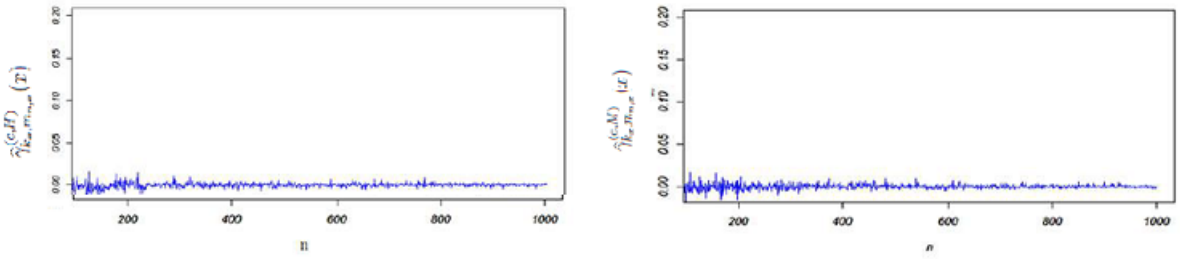

FIGURE 3.5 – Biais bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$)

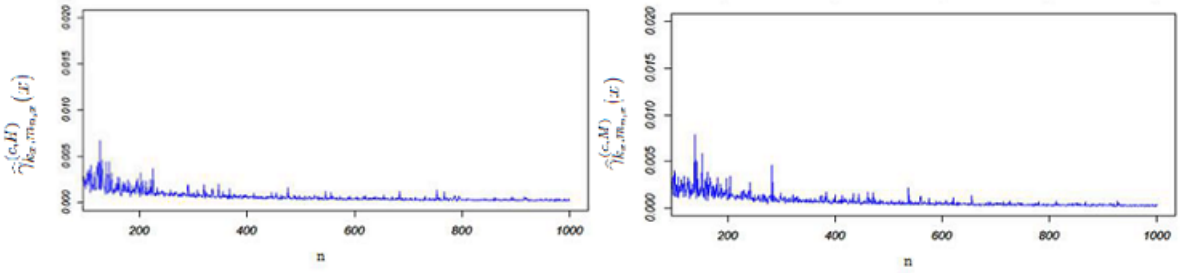

FIGURE 3.6 – MSE bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$)

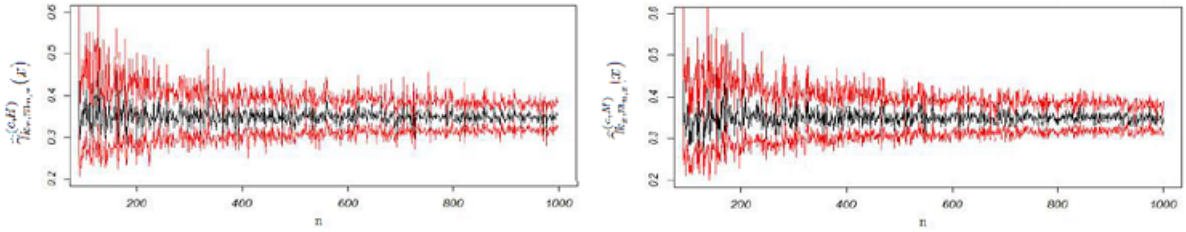

FIGURE 3.7 – IC bootstrap de $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ 100 répétitions de Pareto ($\gamma_{(X)} = 0.35$) censurées par Pareto ($\gamma_{(Y)} = 1$)

Conclusion :

Nous avons utilisé la méthode du Bootsratp sur deux estimateur de l'indice de queue $\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)$ et $\widehat{\gamma}^{(c,M)}_{k_x,m_{n,x}}(x)$ dans le cas conditionnel du domaine d'attraction de Fréchet pour des données censurées á droite pour étudier les indicateurs de dispersion (Sd,Biais,MSE,IC). Notre objectif était d'observer le comportement de chacun des deux estimateurs.
Les résultats des simulatons montrent que la performance des estimateurs dépend le pourcentage de censure, la taille de l'indice des valeurs extrêmes et le nombre de statistiques ordre... , cette étude est considérée comme la continuation d'une précédente étude de "Richard Minkah 2017" à travers laquelle nous avons atteint les mêmes résultats.

BIBLIOGRAPHIE

[1] *Roncalli, T., 2002, Théorie des Valeurs Extrêmes ou Modélisation des événements rares pour la Gestion des Risques, DESS 203 de l'Université Paris IX Dauphine Marchés Financiers, Marchés des Matières Premières et Gestion des Risques.*

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

[3] *Belkadi, S.,analyse par bootstrap des données censurées, université Houari Boumediéne.*

[4] *Ndao,P.,Modélisation de valeurs extrêmes conditionnelles en présence de censure,thése de doctorat,université Gaston berger de Saint-Louis.*

[5] *Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, Annals of Statistics 17(4) : 1833-1855.*

[6] *Einmahl, J. H., Fils-Villetard, A., and Guillou, A. (2008). Statistics of extremes under random censoring. Bernoulli, 14(1), 207 - 227.*

[7] *T. Duchesne. Analyse des durées de vie, New York (2004).*

[8] *B. Efron. Censored data and the bootstrap. Technical report. (February 1980). Public health service grant 5 R01 GM21215-05 Stanford, California.*

[9] *F. Utzet. Á. Sánchez. Some Applications of the Bootstrap to Survival Analysis. Anuario de Psicologia , no 55, 155-167 (1992).*

[10] *L. Gardes and S. Girard. A moving window approach for non-parametric estimation of the conditional tail index. J. Multivariate Anal.99(10) :2368{2388, 2008.}*

[11] *B. Brahimi, D. Meraghni, and A. Necir. On the asymptotic normality of Hill's estimator of the tail index under random censoring. ArXiv e-prints, February 2013.*

[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 concentrons sur le domaine d'attraction de Fréchet dans le cas des distributions à queues lourdes lorsque les données sont incomplètes. Nous intéressons à effectuer une comparaison de l'indice des valeurs extrêmes conditionnels dans le cas des données censurées, entrée deux estimateurs hill (Einmahl et al, 2008) et moment (Beirlant et al., 2004). Par des simulations effectués, nous avons appliqué la méthode du Re-échantillonnage"bootstrap", 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 conditionnels , censure aléatoire à droite , distribution à queue lourde , méthode de bootstrap

Abstract :

In this work, we focus on Fréchet's domain of attraction in the case of heavy-tailed distributions when the data are incomplete. we are interested in making a comparison of the index of conditional extreme values in the case of censored data, between two estimators hill(Einmahl and al, 2008) and moment(Beirlant and al., 2004) . By simulations carried out, we applied the "bootstrap" resampling method, to the two estimators in order to calculate the dispersion parameters, in particular to check the performance and efficiency of each of them.

Key words : conditional extreme value index, right random censoring, heavy tailed distribution, bootstrap method

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 الشرطية هيل (انمهال وآخرون ، ٢٠٠٨)وبين مؤشرالقيم القصوى الشرطية مومونه (بيرلانت وآخرون ، ٢٠٠٤) من خلال عملية المحاكات قمنا بتطبيق طريقة تكرار العينة على المقدرين لحساب مقاييس التشتت على غرار أداء وفاعلية كل مقدر.

الكلمات المفتاحية:القيم القصوى الشرطية، التوزيعات ذات الذيول الثقيلة ، تكرار العينة ، الرقابة العشوائية على اليمين