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

FACULTÉ DE MATHÉMATIQUES ET SCIENCES DE LA MATIÈRE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté En Vue De L'obtention Du

DIPLÔME DE MASTER

EN MATHÉMATIQUES

Option : Probabilité et Statistique

Par

Hadjer DJOUHRI

Intitulé

Bootstrap de l'indice des valeurs extrêmes en présence de censure aléatoire à droite

Membres du jury

Abdelmalek BOUSAAD	M. A. A	UKMO	Président
Said ZIBAR	M. A. A	UKMO	Examinateur
Fatima MEDDI	M. C. A	UKMO	Rapporteur

1 juin 2016

Dédicace

Je dédie ce modeste travail

A ma très cher Mère et mon très cher Père
A mes chers sœurs et frères
A toutes la familles Djouhri

A ceux qui ont veillé pour mon bien être

A ceux qui m'ont toujours encouragé pour que je réussisse dans mes études
A tout ce qui m'ont encouragé lors de la réalisation de mon travail

Sans oublier de dédier ce mémoire à mes très chères amies intimes
A tous mes collègues de ma promotion de Proba-Stat 2016

Finalement à tous ceux qui m'on aider de proche ou de loin

REMERCIEMENTS

Je remercie Dieu le tout-puissant de m'avoir donné la volonté, la force et le courage pour bien mener et finir mon travail de thèse.

Je voudrais d'abord et avant tout remercier ma encadreur vertueuse MEDDI FATIMA pour tous leurs efforts en vue d'établir cette mémoire, elle a eu le rôle fondamental et essentiel et la grand mérite dans tout ce qui a été réalisé, comme cela avait toujours été, près de moi et me guider et corriger mes erreurs et me donner de précieux conseils et les conseils appropriés et les alertes considérés, je la répète mes remerciements pour tout ce qu'elle a fait pour moi à travers la mise en plan à toutes les étapes de la préparation de ce mémoire depuis le début jusqu'à la fin, étape par étape, était que le premier et dernier facteur pour le succès de ce travail, et je lui dis encore une fois merci beaucoup à pour votre appréciation profonde.

Avec un grand honneur, j'aimerais présenter mes remerciements et ma gratitude aux membres du jury, Monsieur Abdelmalek BOUSAAD, et Monsieur Said Zibar tout d'abord d'avoir accepté d'examiner mon mémoire, qui sans eux ce mémoire ne pourra jamais voir le jour pour ntérêt et apport qu'ils ont apporté à mon travail.

J'exprime ma gratitude à ma famille qui m'a toujours soutenue et encouragée dans la voie que je m'étais fixée. Je remercie particulièrement mes parents qui m'ont stimulée et encouragé pendant mes études. qui étaient toujours prêts à fournir tous les moyens physique et morale pour la réussite de ce projet.

Table des matières

Table des figures

Liste des tableaux

0.1 Abréviations et Notations

EVD	Distribution des valeurs extrêmes
EVI, γ	Indice des valeurs extrêmes
F	Fonction de répartition
F_n	Fonction de répartition empirique
$F^{\leftarrow}$	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 $X_1, ..., X_n$
N_u	Nombres des excés qui dépassent du seuil u
POT	Pics au-delà d'un seuil
$p.s$	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 $X_1, ..., X_n$
$X \wedge Y$	$\min(X, Y)$
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
$\underset{\rightarrow}{o_P}(.)$	Converge vers 0 en probabilité

$O_P(.)$	Être borné en probabilité
VR_α	Variation régulière d'indice α
$\hat{\gamma}^{(c,H)}$	Estimateur de Hill avec les données censurées
MSE	L'erreur quadratique moyenne
$Z^* = (Z_1^*, ..., Z_n^*)$	Échantillon Bootstrap
se	Erreur standard
$al.$	Autres
$(\Omega, \mathcal{A}, \mathbb{P})$	Espace probabilisé
τ_H	Point terminal
$\sup A$	Supremum de l'ensemble A
$s.o$	Statistique d'ordre
$v.a$	variable aléatoire
IC	Intervalle de confiance

0.2 Introduction

Au cours des dernières années, nous avons pu observer dans la recherche scientifique, une modélisation des événements rares. Ces événements rares sont des événements dont la probabilité d'apparition est trop faible c'est-à-dire se trouve dans les queues des distributions. Ils apparaissent en général dans les contextes physiques nombreux et variés 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 rassemblé 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 XVII^e 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 « loi de Weibull » très utilisée par les spécialistes de données de survie. En 1958, Kaplan et Meier proposent un estimateur jusque là inconnue de la fonction de répartition dans le cas où les données sont censurées. Ils en profitent pour déterminer ses propriétés asymptotiques.

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.
Nous présentons dans le *chapitre 1* quelques rappels essentielles sur la théorie des valeurs extrêmes et aussi sur la notion de censure qui permettent de faciliter la lecture de mémoire. Ainsi, 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 puis nous présentons ensuite l'estimateur classique de l'indice de queue de Hill (1975) $\hat{\gamma}^H$, ainsi que ces propriétés asymptotiques. 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, la fonction de risque. Beaucoup d'auteurs se sont intéressés sur la notion notamment Kaplan et Meier (1958), qui ont proposé un estimateur de la fonction de survie.

Dans le *deuxième chapitre* nous nous plaçons dans le cadre de données censurées aléatoirement à droite. Dans ce contexte, différents estimateurs de l'indice des valeurs extrêmes ont été proposés par Beirlant et al. Ils sont tous construits de la même façon : l'estimateur usuel (sans censure) est divisé par la proportion d'observations non censurées au-delà d'un certain seuil. Récemment, Einmahl et al.(2008) ont établi la normalité asymptotique de tels

estimateurs, en particulier l'estimateur de Hill $\hat{\gamma}_1^{(c,H)}$. Nous avons commencé par présenter l'estimateur de Hill dans le cas des données censurées et ses propriétés asymptotiques [7]. Ensuite nous présentons l'approximation gaussienne de cet estimateur dans le cas d'une distribution à queue lourde [11], et nous finissons par illustrer des nouveaux estimateurs de l'indice de queue sur la base d'intégration de Kaplan-Meier et l'approche de S.Leurgans (1987) récemment proposés par [15] .

Notre principal contribution est proposé au *troisième chapitre,* où nous appliquons la méthode du bootstrap dans le but de réduire le biais de l'estimateur de l'indice de Hill $\hat{\gamma}_1^{(c,H)}$, sur des données censurées présenté auparavant au chapitre 2. Nous présentons d'abord le principe de bootstrap, ensuite nous examinons les estimateurs de l'écart-type, du biais et de l'erreur quadratique moyenne empiriques de l' estimateur de l'indice de queue en question, nous illustrons des intervalles de confiance en effectuant quelques simulations de Bootstrap sous R. Des $QQnorm$ de la distribution limite bootsrap de $\hat{\gamma}_1^{(c,H)}$ montre une normalité asymptotique confirmée.

Chapitre 1

Quelques rappels sur la théorie des valeurs extrêmes et sur la 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 Présentation de la théorie des valeurs extrêmes

1.1.1 Définitions et théorèmes

Définition 1.1 (Statistiques d'ordre). *Soient n variables aléatoires* $(X_i)_{1\leq i\leq n}$ *indépendantes et identiquement distribuées. Les statistiques d'ordre associées à l'échantillon* $(X_1, ..., X_n)$*sont les réarrangements croissants des éléments de cet échantillon. Elle sont dénotées par* $X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$*, et* $X_{i,n}$ *est la* $i^{ième}$ *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 : $X_{1,n} = \min(X_{1,}..., X_n)$ et $X_{n,n} = \max(X_1, ..., X_n)$. [13]

Loi de $X_{i,n}$.

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

Nous en déduisons que la fonction de densité est :

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

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 $X_1, ..., X_n$, on obtient :

Loi de $X_{1,n}$.

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

d'où

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

Loi de $X_{n,n}$.

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

d'où

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

Remarque 1.1

$$\mathbb{P}\{X_{n,n} \leq x\} = (F(x))^n \to 0 \text{ ou } 1 \text{ quand } n \to \infty.$$

Les expressions de $F_{1:n}$ et $F_{n:n}$ peuvent s'obtenir trés facilement en considérant les relations [13]

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

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

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

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

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

Il existe une autre version de la définition de F_n en utilisant les $(s.o)$ comme suit :

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

Comportement asymptotique des extrêmes

On définit la variable aléatoire M_n, qui traduit le maximum d'un n-échantillon d'une variable aléatoire X, (les variables aléatoires X_i sont indépendantes et suivent la même loi que X) par :

$$M_n = \max (X_i)_{1 \leq i \leq n}$$

On pourrait aussi s'intéresser au minimum en utilisant la relation :

$$\min (X_i)_{1 \leq i \leq n} = - \max (-X_i)_{1 \leq i \leq n}$$

Le résultat central de la théorie des valeurs extrêmes concerne la distribution asymptotique du maximum (ou le minimum) en fonction de celle de la variable aléatoire X. Notons la fonction de répartition F_X de la variable aléatoire de loi de probabilité $\mathbb{P}$, à savoir $F_X(x) = \mathbb{P}(X < x)$. La fonction de répartition de M_n est alors définit par :

$$\begin{aligned} F_{M_n}(x) &= \mathbb{P}(M_n \leq x) \\ &= \mathbb{P}(X_1 \leq x, ..., X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \times ... \times \mathbb{P}(X_n \leq x) \\ &= [F_X(x)]^n \end{aligned}$$

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

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

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

1.1.2 Loi des valeurs extrêmes

Comme la fonction de répartition obtenue précédemment conduit à une loi dégénérée lorsque n tends vers l'infini, 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.

Théorème 1.1 (Fisher et Tippett, 1928, Gnedenko, 1943). Soit $X_1, ..., X_n$ une suite de n variables aléatoires réelles *iid* de loi continue P et $M_n = \max(X_i)_{1\leq i\leq n}$. S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$, et une fonction de répartition non-dégénérée G_γ telle que,

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

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

$$\text{loi de Gumbel : } \Lambda_\gamma(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty$$

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

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

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 Extreme Value Distribution) des trois distribution par,

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

Remarque 1.2. Si F vérifie le théorème 1.1. On dit alors que F appartient au domaine d'attraction de G_γ et on note $F \in D(G_\gamma)$ selon le signe de γ.(section 1.1.4)

1.1.3 Distribution coditionnelle des excés

L'idée d'utiliser un nombre croissant de statistiques d'ordre de l'échantillon a ensuite été plus largement développée dans le cadre de l'approche « Peaks Over Threshold »(POT), via l'approximation de la loi des excès au-delà d'un seuil par des GPD (Generalized Pareto Distributions). L'idée est la suivante : partant d'un échantillon $X_1, ..., X_n$, on se fixe un seuil u grand. On ne considére que les N_u observations dépassant ce seuil. On note $Y_i, i = 1, ..., N_u$. Plus précisément, soit $u < x_F$ et F_u la fonction de répartition des excès représente la probabilité que la variable aléatoire X dépasse le seuil u d'au plus une quantité y, sachant qu'elle dépasse u et définie par :

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

Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull) (section 1.1.4), alors il existe une fonction $\sigma(u)$ strictement positive et un $\gamma \in \mathbb{R}$ 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 $G_{\gamma,\sigma}$ est la fonction de répartition de la loi de Paréto Généralisée définie par :

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

1.1.4 Caractérisation des Domaines d'attraction

Selon le signe de γ on distingue trois cas de domaines d'attraction **(D.A.)** :

1. 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.
2. 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.

3. Si $\gamma < 0$, 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.

Fonctions à variation régulière

Définition 1.3. *Une fonction mesurable $g : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation régulière à l'infini si et seulement si, il existe un réel α tel que, pour tout $x > 0$*

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

Et on note $g \in VR_\alpha$, α est appelé indice (ou exposant) de la fonction à variation régulière.

Dans le cas particulier où $\alpha = 0$, on dit que g est à variation lente à l'infini, c'est à dire,

$$\lim_{t\to\infty} \frac{g(tx)}{g(t)} = 1 \quad \forall x > 0$$

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

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

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

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

Définition 1.4. *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 $\ell \in RV_0$.

Théorème 1.2. Une fonction $\ell : (0,\infty) \to (0,\infty)$ est à variation lente si et seulement si elle peut être écrite comme :

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

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

Théorème 1.3. 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 :

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) \quad \text{et} \quad b_n = 0.$$

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

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

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

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

Nous en déduisons que

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

Donc $1 - F \in RV_{-1/\gamma}$

Domaine d'attraction de Weibull

Théorème 1.4. 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 $1/\gamma$ c'est-à-dire,

$$1 - F = (x_F - x)^{-1/\gamma} \ell \left[(x_F - x)^{-1} \right].$$

avec,

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

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

$$a_n = x_F - F^{-1}\left(1 - \frac{1}{n}\right) \quad \text{et} \quad b_n = x_F.$$

Domaine d'attraction de Gumbel

Définition 1.5. *Soit F une fonction de répartition de point terminal x_F fini ou infini. S'il existe $z < x$ tel que*

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

où $c > 0$ et a une fonction positive absolument continue de densité a' vérifiant $\lim_{x \uparrow x_F} a'(x) = 0$. Alors F est une fonction de Von-Mises et a est sa fonction auxiliaire.

Théorème 1.5. F appartient au domaine d'attraction de Gumbel si et seulement si il existe une fonction de Von-Mises F^* telle que pour $z < x < x_F$ on ait :

$$1 - F(x) = c(x)\left[1 - F^*(x)\right] = c(x) \exp\left\{ - \int_z^x \frac{1}{a(t)} dt \right\},$$

où $c(x) \to c > 0$ lorsque $x \to x_F$.

Les Figures 1.1 et 1.2 [10] ci-dessous illustre le comportement de différentes distributions GEV correspondant à différentes valeurs de γ. Les Tableaux (Tableau 1.1, Tableau 1.2, Tableau 1.3, [14]) donnent quelques lois et leur domaine d'attraction.

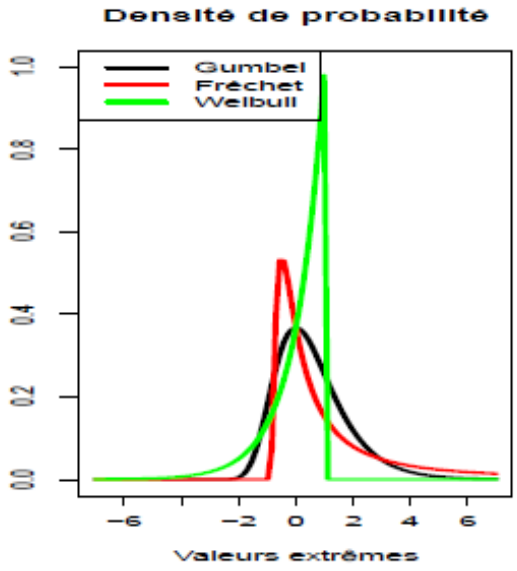

FIG. 1.1 – Représentation de la densité de probabilité : Gumbel $(\gamma = 0)$, Fréchet $(\gamma = 1)$,Weibull $(\gamma = -1)$.

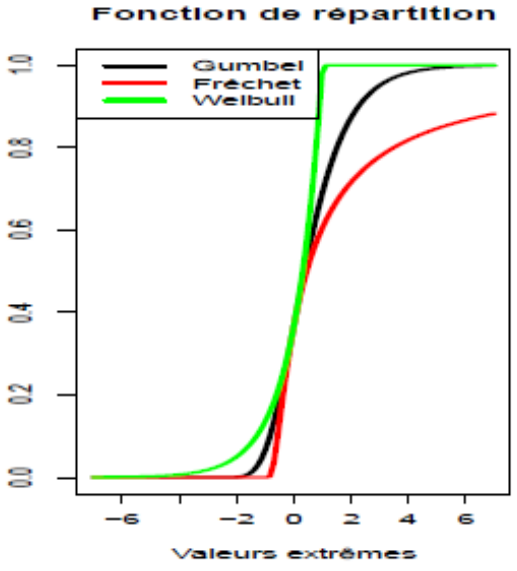

FIG. 1.2 – Représentation de la fonction de répartition : Gumbel $(\gamma = 0)$, Fréchet $(\gamma = 1)$,Weibull $(\gamma = -1)$.

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Burr$(\beta,\tau,\lambda)\,,\beta>0,\tau>0,\lambda>0$	$\left(\frac{\beta}{\beta+x^{\tau}}\right)^{\lambda}$	$\frac{1}{\lambda\tau}$
Fréchet$\left(\frac{1}{\alpha}\right),\alpha>0$	$1-\exp\left(-x^{-\alpha}\right)$	$\frac{1}{\alpha}$
Loggamma$(m;\lambda)\,,m>0,\lambda>0$	$\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}(\log(u))^{m-1}\,u^{-(\beta+1)}du$	$\frac{1}{\lambda}$
Loglogistic$(\beta,\alpha)\,,\beta>0,\alpha>0$	$\frac{1}{1+\beta x^{\alpha}}$	$\frac{1}{\alpha}$
Pareto$(\alpha)\,,\alpha>0$	$x^{-\alpha}$	$\frac{1}{\alpha}$

TAB. 1.1 – Quelques distributions appartenant au domaine d'attraction de Fréchet

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Gamma$(m,\lambda), m\in\mathbb{N},\lambda>0$	$\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}u^{m-1}\exp(-\lambda u)du$	0
Gumble$(\mu,\beta),\mu\in\mathbb{R},\beta>0$	$\exp(-\exp(-\frac{x-\mu}{\beta}))$	0
Logistic	$\frac{2}{1+\exp(x)}$	0
Log nomole$(\mu,\sigma),\mu\in\mathbb{R},\sigma>0$	$\frac{1}{\sqrt{2\pi}}\int_x^{\infty}\frac{1}{\mu}\exp(-\frac{1}{2\sigma^2}(\log u-\mu)^2du$	0
Weibull $(\lambda,\tau)\,,\lambda>0,\tau>0$	$\exp(-\lambda x^{\tau})$	0

TAB. 1.2 – Quelques distributions appartenant au domaine d'attraction de Gumble

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Uniforme $(0,1)$	$1-x$	-1
Reverse Burr$(\beta,\tau,\lambda,\tau_F),\beta>0,\tau>0,\lambda>0$	$\left(\frac{\beta}{\beta+(\tau_F-x)^{-\tau}}\right)^{\lambda}$	$-\frac{1}{\lambda\tau}$

TAB. 1.3 – Quelques distributions appartenant au domaine d'attraction de Weibull

1.1.5 Estimateur de l'indice des valeurs extrêmes γ

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

Soient $X_1, ..., X_n$ des variables aléatoires *iid* de fonction de répartition commune F et soit $X_{1,n}, ..., X_{n,n}$ les statistiques d'ordre correspondantes. On suppose que F satisfait la condition du domaine d'attraction de Fréchet, en terme de variations régulière ceci est équivalent à : pour $\gamma > 0$

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

La méthode du maximum de vraisemblance est la plus populaire et qui sous certaines conditions est la plus efficace. On se sert des statistiques d'ordre supérieurs à un certain seuil u pour ne garder que les observations les plus grandes de façon à ce quelles suivent approximativement une distribution de *Pareto.* D'autre part on peut écrire :

$$\frac{1 - F(x)}{1 - F(u)} = \left(\frac{x}{u}\right)^{-1/\gamma}, \quad x > u, \quad u \in \mathbb{R}.$$

Si $(n-k)$ désigne le nombre de statistiques d'ordre qui dépassent le seuil u, alors on estime u par $X_{n-k,n}$. En utilisant les plus grande statistiques d'ordre $X_{n-k+1,n}, ..., X_{n,n}$ la fonction de log-vraisemblance sera alors :

$$\begin{aligned} L(\gamma, X_{n-k+1}, X_{n,n}) &= -k \log(\gamma u) + k \log(1 - F(u)) \\ &\quad - \frac{\gamma}{\gamma + 1} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n}). \end{aligned}$$

En maximisant la fonction de log-vraisemblance par rapport à γ, on obtient l'estimateur de *Hill* pour $\gamma > 0$,

$$\hat{\gamma}_{k_n}^H = \frac{1}{k} \sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \tag{1.1}$$

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

plusieurs chercheurs ont essayé de déterminer les propriétés asymptotiques de l'estimateur de Hill. Mason (1982) a prouvé la consistance faible de l'estimateur de Hill pour toute suite $k = k(n)$ satisfaisant $k \to \infty$ et $k/n \to 0$ quand $n \to \infty$ appelée suite intermédiaire d'entiers. La condition $k \to \infty$ assure que la taille de statistiques d'ordre k est assez grande afin d'obtenir des estimateurs stables. Par contre, la condition $k/n \to 0$ permet de rester

dans la queue de distribution. Davis et Resnick (1984) ont proposé sa normalité asymptotique sous les conditions de Von Mises ; Csörgö et Mason (1985) ont présenté sa normalité asymptotique en introduisant l'approximation des processus empiriques par les ponts browniens. Dans cette même lancée Resnick et de Haan (1998) ont montré cette propriété asymptotique. Rappelons à présent les propriétés asymptotiques de l'estimateur. Pour cela, nous allons commencer par les conditions du premier et du second ordre avec les fonctions quantiles définies ainsi :

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

et

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

Proposition 1.2 [10] (Conditions du premier ordre, de Haan et Ferreira (2006)). Les assertions suivantes sont équivalentes :

1. F est à queue lourde

$$F \in D.A(\text{fréchet}),\ \gamma > 0$$

2. $1 - F$ est une fonction à variation régulière à l'infini d'indice $-1/\gamma$

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

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

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

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

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

Proposition 1.3 [10] (Conditions du second ordre de Haan et Ferreira (2006)). Une fonction de répartition $F(.) \in D.A(\text{fréchet}),\gamma > 0$, admet une condition du second ordre à l'infini si elle satisfait à l'une des assertions suivantes :

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

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

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

$$\lim_{s\to 0} \frac{Q(1-sx)/Q(1-s) - x^{-\gamma}}{A_2(s)} = x^{-\gamma}\frac{x^{\rho}-1}{\rho},$$

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

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

si $\rho = 0$, on remplace $(x^{\rho}-1)/\rho$ par $\log x$
Les fonctions $A(.)$, $A_1(.)$, $A_2(.)$ sont à variations régulières à l'infini d'indices respectifs ρ, ρ/γ, et $-\rho$, avec $A_1(t) = A(1/(1-F(t)))$ et $A_2(s) = A(1/s)$.

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

Théorème 1.6 (Propriétés asymptotiques de l'estimateur de Hill [6]). Soit k_n, $n \geq 1$ une suite d'entiers telle que $1 \leq k_n \leq n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.

1. **Consistance faible** : $\hat{\gamma}^H_{k_n}$ converge en probabilité vers γ.
2. **Consistance forte** : Si de plus $k_n/\log n \log n \to \infty$ quant $n \to \infty$, alors $\hat{\gamma}^H_{k_n}$ converge presque sûrement vers γ.
3. **Normalité asymptotique** : Si la condition

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

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

$$\sqrt{k_n}(\hat{\gamma}^H_{k_n} - \gamma) \xrightarrow{\mathcal{L}} \mathcal{N}(\lambda/(1-\rho), \gamma^2)$$

1.2 Quelques généralités 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 ([10] , [3])

Définition 1.6. *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 1.7. *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 1.8. *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 1.9. *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 1.10. *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 [10]

. 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 [10] :

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 1.11. *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 1.12. *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*

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

La censure de type III : C'est la version aléatoire du type I
Définition 1.13. *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*

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

1.2.3 Estimation de la fonction de survie

Dans la littérature plusieurs auteurs se sont intéressés pour l'estimation de la fonction de survie dans le cas où les données sont censurées. Parmi ces derniers nous pouvons citer *Kaplan* et *Meier* (1958) ont proposé un estimateur de la fonction de survie.

Estimateur de Kaplan-Meier

Soit $(Z_i,\delta_i)_{1\leq i\leq n}$ l'échantillon réellement observé et soit $(Z_{i,n},\delta_{i,n})_{1\leq i\leq n}$ sa statistique d'ordre croissant. L'estimateur de *Kaplan-Meier* est défini par :

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

Il est aussi appelé « **produit limite** » car il s'obtient comme la limite d'un produit.

. Cet estimateur de Kaplan-Meier est une fonction étagée avec des sauts seulement aux observations non-censurées.

. La hauteur des sauts de cet estimateur est aléatoire.

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

L'estimateur de Kaplan-Meier est asymptotiquement gaussien, précisément on a le résultat suivant :

Théorème 1.7 (Droesbeke et Saporta (2011)). Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sup_{t\geq 1}\left|\hat{S}_n(t)-S(t)\right| \xrightarrow{p.s} 0$$

et pour tout $t \geq 0$,

$$\sqrt{n}\left(\hat{S}_n(t)-S(t)\right) \xrightarrow{d} W_t$$

où $(W_t)_{t\geq 0}$ est un processus gaussien centré qui vérifie pour tous t et s strictement positifs,

$$cov(W_s, W_t) = S(t)S(s)\int_0^{s\wedge t}\frac{dF(u)}{(1-F(u))^2(1-G(u))}.$$

Chapitre 2

Estimation de l'indice des valeurs extrêmes en présence de censure aléatoire à droite

L'analyse des valeurs extrêmes lorsque les données sont censurées aléatoire est un nouveau sujet de recherche, il a été mentionné dans le livre de *Reiss et Thomas* (1997, [12]) dans la section 6.1 en tant que première étape, mais sans résultats asymtotiques *Beirlant et al* (2007, [2]) ont proposé des estimateurs pour les EVI et ils ont discuté leurs porpriétés asymptotiques lorsque les données sont censurées par un seuil déterministe. Plus récemment *Einmahl et al.*(2008, [7]). ont adapté divers estimateur EVI au cas des données sont censurées par un seuil aléatoire, et ont proposé une méthode unifiée pour établir leurs porpriétésasymptotiques.

2.1 Modèle et notation

Définition 2.1.[1] *On dit que la fonction de distribution F est a queue lourde si la fonction du queue $\overline{F} := 1 - F$ est à variation régulière à l'infini d'indice $-1/\gamma$,*

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

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 $X_1, ..., X_n$ 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 $Y_1,...,Y_n$, 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 : $1-G\in RV(-1/\gamma_2)$, 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

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

et d'autre part les indicateurs de censure

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

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 :

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

Alors

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

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

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

Si F et G sont supposées être dans le dommaine d'attraction maximales $D(G_{\gamma_1})$ et $D(G_{\gamma_2})$ 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 $H \in D(G_\gamma)$. *Einmahl et al.*(2008, [7]) ont examiné les trois cas les plus intéressants suivants:

$$\begin{cases} cas\ 1 & : & \gamma_1 > 0, \gamma_2 > 0 & & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \\ cas\ 2 & : & \gamma_1 < 0, \gamma_2 < 0 & \tau_F = \tau_G & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \\ cas\ 3 & : & \gamma_1 = \gamma_2 = 0 & \tau_F = \tau_G = \infty & \gamma = 0 \end{cases} \tag{2.1}$$

2.2 Estimateur de l'indice de queue $\hat{\gamma}^{(c,H)}$

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 $(Z_1, ..., Z_n)$, si l'on note l'administration concomitant de statistique d'ordre i par $\delta_{[i:n]}$, $\delta_{1:n}, ..., \delta_{n:n}$ les δ's correspondants à $Z_{1,n}, ..., Z_{n,n}$ respectivement, i.e,

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

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 γ_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 :

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

où

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

alors

$$\hat{\gamma}_1^{(c,H)} = \frac{k^{-1}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}}{k^{-1}\sum_{i=1}^{k} \delta_{[n-i+1:n]}} \tag{2.3}$$

telque $\hat{p}$ estime $p = \frac{\gamma_2}{\gamma_1+\gamma_2}$ ($\hat{p} \to p$: quand $n \to \infty$) et $\hat{\gamma}_1^{(c,H)}$ estime $\gamma_1 = \gamma/p$ avec $\delta_{[1:n]}, ..., \delta_{[n:n]}$ les indicateurs de censure retenues respectivement avec l'échantillon $Z_{1:n}, ..., Z_{n:n}$. *Einmahl et al* (2008, [7]) 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, t, est aléatoire et égal à $Z_{n-k,n}$ la $n-k-ième$ statistique d'ordre de l'échantillon $Z_1, ..., Z_n$.

2.2.1 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, [7])),

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

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 :

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

2.2.2 Application de loi à queues lourdes

Supposons X et Y sont respectivement de Pareto (γ_1) et Pareto(γ_2) , C'est-à-dire pour tout $x \geq 1$. ([9]) ,

$$\begin{aligned} F_X(x) &= 1 - x^{-1/\gamma_1} \ , \gamma_1 > 0 \\ F_Y(x) &= 1 - x^{-1/\gamma_2}, \gamma_2 > 0 \end{aligned}$$

on obtient :

$$\begin{aligned} F_Z(z) &= \mathbb{P}(\min(X, Y \leq z)) \\ &= 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= 1 - z^{-1/\gamma_1} z^{-1/\gamma_2} \\ &= 1 - z^{-\frac{\gamma_1+\gamma_2}{\gamma_1\gamma_2}}, \end{aligned}$$

ce qui implique $Z \frown \text{Pareto}(\gamma_1\gamma_2/(\gamma_1+\gamma_2))$, nous pouvons à présent calculer la fonction $p(z)$

$$\begin{aligned} p \equiv p(z) &:= \frac{(1-F_Y(z))f_X(z)}{(1-F_Y(z))f_X(z)+(1-F_X(z))f_Y(z)} \\ &= \frac{z^{-1/\gamma_2}\frac{1}{\gamma_1}z^{-1/\gamma_1}}{z^{-1/\gamma_2}\frac{1}{\gamma_1}z^{-1/\gamma_1}+z^{-1/\gamma_1}\frac{1}{\gamma_2}z^{-1/\gamma_2}} \\ &= \frac{\frac{1}{\gamma_1}z^{-1/\gamma_1-1/\gamma_2}}{(\frac{1}{\gamma_1}+\frac{1}{\gamma_2})z^{-1/\gamma_1-1/\gamma_2}} \\ &= \frac{\frac{1}{\gamma_1}}{\frac{1}{\gamma_1}+\frac{1}{\gamma_2}} \\ &= \frac{\gamma_2}{\gamma_1+\gamma_2}. \end{aligned}$$

Par conséquent, le quotient entre un estimateur de $\gamma = \gamma_Z$ et un estimateur de $p = p_z$ sera fournir une estimation $\gamma_X = \gamma_1$ du paramètre d'intérêt. En effet, beaucoup plus général, et pour tous les cas mentionnés ci-dessus (*voir* 2.1) .

$$\begin{aligned} &F_X \in D_M(EV_{\gamma_1}), \quad F_Y \in D_M(EV_{\gamma_2}) \\ &\quad \Longrightarrow F_{Z=\min(X,Y)} \in D_M(EV_\gamma), \text{tel que} : \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \end{aligned}$$

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 $u > 0$

$$\lim_{t\to\infty} \frac{H^{\leftarrow}(1-\frac{1}{tu})/H^{\leftarrow}(1-\frac{1}{t}) - u^{\gamma}}{b(t)} = u^{\gamma}\frac{u^{\rho}-1}{\rho} \tag{2.4}$$

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

$$1 < k < n;\ k \to \infty \ \text{ et } k/n \to 0, n \to \infty \tag{2.5}$$

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

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

$\mathbb{C}4$: 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.[7] Sous les condition $\mathbb{C}1 - \mathbb{C}4$ et s'il existe b_0 et σ telles que

$$\sqrt{k}(\hat{\gamma}^{(.)}_{Z,k,n} - \gamma) \xrightarrow{d} \mathcal{N}(\alpha_1 b_0, \sigma^2). \tag{2.6}$$

Alors, nous avons

$$\sqrt{k}\left(\hat{\gamma}^{(c,.)}_{Z,k,n} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\frac{1}{p}(\alpha_1 b_0 - \gamma_1\alpha_2), \frac{\sigma^2 + \gamma_1^2 p(1-p)}{p^2}\right)$$

Corollaire 2.1.[4] nous avons les résultats asymptotiques de l'estimateurs de Hill.

$$\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\mu^{(c,H)}; \frac{\gamma_1^3}{\gamma}\right)$$

où

$$\mu^{(c,H)} := -\frac{\gamma_1\alpha_2}{p} + \frac{\alpha_1}{p(1-\rho)}$$

Ce corollaire se déduit directement du théorème précédent en notant $b_0 = 1/(1-\rho)$ et $\sigma^2 = \gamma^2$.

2.3 Estimateur de l'indice de queue par approximation gaussienne

(*Neci.A , Brahimi,B. et Meraghni,D.* (2014, [11])) ils ont utilisé la théorie des processus empiriques, rapprochant l'estimateur adapté Hill, pour les données censurées, en termes de processus de Gauss. Ensuite, ils ont présenté sa normalité asymptotique, que sous la condition du second ordre de variation régulière, avec le même écart que celui obtenu par *Einmahl et al.* (2008, [7]). L'approximation gaussienne nouvellement proposée est conpatible avec la représentation asymptotique de l'estimateur classique de Hill, dans le cadre de non censure. leur résultats seront d'un grand intérêt pour établir les distributions limites de nombreuses statistiques en théorie des valeurs extrêmes sous censure aléatoire tels que les estimateurs des indices de queue, les mesures de risque actuarielles et les tests d'ajustement fonctionels pour les distributions à queue lourde.
Supposons que cdf's F, G et H satisfont à la condition de second ordre de variation régulière. Autrement dit, ils supposons qu'il existe un constant $\tau_j \leq 0$ et une fonction $A_j, j = 1, 2, 3$ ce qui tend à zéro et ne pas changer le signe près de l'infini, tel que pour toute $x > 0$

$$\begin{aligned}
\lim_{t\to\infty} \frac{\bar{F}(tx)/\bar{F}(t) - x^{-1/\gamma_1}}{A_1(t)} &= x^{-1/\gamma_1} \frac{x^{\tau_1/\gamma_1} - 1}{\gamma_1 \tau_1} \\
\lim_{t\to\infty} \frac{\bar{G}(tx)/\bar{G}(t) - x^{-1/\gamma_2}}{A_2(t)} &= x^{-1/\gamma_2} \frac{x^{\tau_2/\gamma_2} - 1}{\gamma_2 \tau_2} \\
\lim_{t\to\infty} \frac{\bar{H}(tx)/\bar{H}(t) - x^{-1/\gamma}}{A_3(t)} &= x^{-1/\gamma} \frac{x^{\tau_3/\gamma} - 1}{\gamma \tau_3}
\end{aligned} \tag{2.7}$$

où $\bar{S}(x) := S(\infty) - S(x)$, pour tout S.

2.3.1 Approximation gaussienne de $\hat{\gamma}_1^H$

(*Neci.A , Brahimi,B. et Meraghni,D.* (2014, [11])) se sont intéressé à l'approximation gaussienne pour la distribution du l'estimateur adapté $\hat{\gamma}_1^{(H,c)}$, similaire à celle obtenue pour l'estimateur de Hill $\hat{\gamma}_1^H$ dans le cas de données complètes. En effet, si (2.7) maintient F, alors, pour une suite des nombres entiers k satisfaisant (2.5) avec $\sqrt{k}A_1(n/k) = O(1)$, on a lorsque $n \to \infty$,

$$\begin{aligned}
\sqrt{k}(\hat{\gamma}_1^H - \gamma_1) &= \gamma_1 \sqrt{\frac{n}{k}} \int_0^1 s^{-1} \tilde{B}_n \left(1 - \frac{k}{n} s\right) ds - \gamma_1 \sqrt{\frac{n}{k}} \tilde{B}_n \left(1 - \frac{k}{n}\right) \\
&\quad + \frac{\sqrt{k} A_1(n/k)}{1 - \tau_1} + o_P(1)
\end{aligned}$$

où $\left\{\tilde{B}_n(s); 0 \leq s \leq 1\right\}$ est une suite de ponts browniens (voir par exemple de Haan et Ferreira, (2006, [5]), page 163). En d'autres termes, si $\sqrt{k}A_1(n/k) \to \lambda_1$, puis

$$\sqrt{k}\left(\hat{\gamma}_1^H - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda_1}{1-\tau_1}, \gamma_1^2\right) \quad n \to \infty$$

L'approximation gaussienne ci-dessus permet de résoudre de nombreux problèmes en ce qui concerne le comportement asymptotique de plusieurs statistiques de distributions à queue lourde, comme les estimateurs de : la moyenne (Peng, 2001 et 2004, Brahimi et al, 2013), la prime de réassurance en excédent de pertes(Necir et al., 2007), les mesures de risque de distorsion (Necir et Meraghni, 2009 et Brahimi et al., 2011).

2.3.2 Approximation gaussienne de $\hat{\gamma}_1^{(c,H)}$

Le résultat principal de (*A.Neci ,B. Brahimi et D. Meraghni* (2014)) qui consiste en une approximation gaussienne de $\hat{\gamma}_1^{(H,c)}$ qu'en assumant les conditions de secondes d'ordre de variation régulière (2.7). Plus précisément, Ils ont démontré qu'il existe une suite de ponts browniens $\{B_n(s); 0 \leq s \leq 1\}$ définie sur $(\Omega, \mathcal{A}, \mathcal{P})$, tel que

$$\sqrt{k}(\hat{\gamma}_1^{(H,c)} - \gamma_1) = \Psi(B_n) + \frac{\sqrt{k}A_1(h)}{1-p\tau_1} + o_p(1) \quad n \to \infty$$

où Ψ est une partie fonctionnelle à définir de tel que $\Psi(B_n)$ est une v.a normale avec moyenne nulle et de variance asymptotique $\gamma_1^2/p = \gamma_1^3/\gamma$ et $h = h_n := \overleftarrow{H}(1-k/n)$, avec $\overleftarrow{K}(y) := \{x : k(x) \geq y\}, 0 < y < 1$, désignant la fonction quantile (ou l'inverse généralisée) se rapportant à une cdf K.
En plus de l'approximation gaussienne de $\sqrt{k}(\hat{\gamma}_1^{(H,c)} - \gamma_1)$, leurs résultats principal (Indiqué dans le théorème 2.2) consiste dans les représentations asymptotiques, avec processus gaussien, de deux autres statistiques utiles, à savoir $\sqrt{k}(Z_{n-k:n}/h - 1)$ et $\sqrt{k}(\hat{p} - p)$

Théorème 2.2. [11] Supposons que toutes les conditions de second ordre (2.7) détiennent. Soit $k = k_n$ être un suits des entiers satisfaisant, en plus de (2.5), $\sqrt{k}A_j(h) = O(1)$ pour $j = 1, 2, 3$ lorsque $n \to \infty$, Ensuite, il existe une suite de ponts browniens $\{B_n(s); 0 \leq s \leq 1\}$ tel que, lorsque $n \to \infty$

$$\sqrt{k}\left(\frac{Z_{n-k:n}}{h} - 1\right) = \gamma\sqrt{\frac{n}{k}}B_n^*\left(\frac{k}{n}\right) + o_p(1), \tag{2.8}$$

$$\sqrt{k}(\hat{p}-p)=\sqrt{\frac{n}{k}}\left(qB_n\left(\frac{k}{n}\right)-p\tilde{B}_n\left(\frac{k}{n}\right)\right)$$

$$-pq\left(\frac{\gamma_1^{-1}\sqrt{k}A_1(h)}{1-p\tau_1}-\frac{\gamma_2^{-1}\sqrt{k}A_2(h)}{1-q\tau_2}\right)+o_p(1), \tag{2.9}$$

et

$$\sqrt{k}\left(\hat{\gamma}_1^{(H,c)}-\gamma_1\right)=\gamma_1\sqrt{\frac{n}{k}}\int_0^1 s^{-1}B_n^*\left(\frac{k}{n}s\right)ds$$

$$-\frac{\gamma_1}{p}\sqrt{\frac{n}{k}}B_n\left(\frac{k}{n}\right)+\frac{\sqrt{k}A_1(h)}{1-p\tau_1}+o_P(1), \tag{2.10}$$

où, $B_n(s)=B_n(\theta)-B_n(\theta-ps)$, pour $0\leq s\leq\theta/p$,$\tilde{B}_n(s):=-B_n(1-qs)$, pour $0\leq s\leq 1$ et $B_n^*(s):=B_n(s)+\tilde{B}_n(s)$, sont des suites de processus gaussiens centrés, avec $\theta:=H^1(\infty)$ et $q:=1-p$

Corollaire 2.2.[11] Supposons que les conditions du théorème 2.2 sont remplis. supposons en outre que $\sqrt{k}A_1(h)\to\lambda_1$ lorsque $n\to\infty$ alors :

$$\sqrt{k}\left(\hat{\gamma}_1^{(H,c)}-\gamma_1\right)\xrightarrow{d}\mathcal{N}\left(\frac{\lambda_1}{1-p\tau_1},\frac{\gamma_1^2}{p}\right)\ \text{lorsque}\ \ n\to\infty$$

La notation $\mathcal{N}(m,v^2)$ désigne la distribution normale de moyenne m et de variance v^2.

Remarque 2.1. Il est évident que les deux premiers résultats du théorème 2.2 (*vior* 2.8 *et* 2.9) donnent respectivement les distributions asymptotiques suivantes :

$$\sqrt{k}\left(\frac{Z_{n-k:n}}{h}-1\right)\xrightarrow{d}\mathcal{N}(0,\gamma^2)\quad\text{et}\quad\sqrt{k}\,(\hat{p}-p)\ \xrightarrow{d}\mathcal{N}(\mu_p,pq)$$

où

$$\mu_p:=pq\left(\gamma_2^{-1}\frac{\lambda_2}{1-p\tau_1}-\gamma_1^{-1}\frac{\lambda_1}{1-q\tau_2}\right)$$

avec $\lambda_j:=\lim_{n\to\infty}\sqrt{k}A_j(h), j=1,2.$

Preuve (Corollaire 2.2) [11]**.**
A partir du troisième résultat du théorème(2.10), on en déduit que, $\sqrt{k}\left(\hat{\gamma}_1^{(H,c)}-\gamma_1\right)$ est asymptotiquement gaussien de moyenne

$$\frac{\lim_{n\to\infty}\sqrt{k}A_1(h)}{1-p\tau_1}=\frac{\lambda_1}{1-p\tau_1},$$

et de variance,

$$\gamma_1^2\lim\mathbb{E}\left[\sqrt{\frac{n}{k}}\int_0^1 s^{-1}B_n^*\left(\frac{k}{n}s\right)ds-\frac{1}{p}\sqrt{\frac{n}{k}}B_n\left(\frac{k}{n}\right)\right]^2.$$

Les processus $B_n(s)$, $\tilde{B}_n(s)$ et $B_n^*(s)$ satisfaisant

$$p^{-1}\mathbb{E}\left[B_n(s)B_n(t)\right]=\min(s,t)-pst$$
$$q^{-1}\mathbb{E}\left[\tilde{B}_n(s)\tilde{B}_n(t)\right]=\min(s,t)-qst$$

et

$$p^{-1}\mathbb{E}\left[B_n(s)B_n^*(t)\right]=\mathbb{E}\left[B_n^*(s)B_n^*(t)\right]=\min(s,t)-st$$

Puis, par un calcul élémentaire, on obtient $\frac{\gamma_1^2}{p}$ pour la variance asymptotique de $\sqrt{k}\left(\hat{\gamma}_1^{(H,c)}-\gamma_1\right)$.

2.4 Estimations de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier et l'approche Leurgans

Julien Worms et Rym Worms (2013, [15]) 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* et l'approche de données synthétique de S.*Leurgans* (1987). 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 conditions du premier ordre.

2.4.1 Première approche ($intégrale\ de\ Kaplan-Meier$)

Le premiére 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,

$$\mathbb{E}\left[\frac{\delta}{1-G(Z)}\phi(Z)\right]=\mathbb{E}\left[\phi(X)\right] \tag{2.11}$$

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 (*voir* 2.1)

$$F \in D(G_{\gamma_X}), G \in D(G_{\gamma_Y})\ \gamma_X, > 0\ , \gamma_Y > 0 \tag{2.12}$$

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

$$\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$$

Dans ce cas, il est bien connu que (voir remarque 1.2.3 dans [Haan et Ferreira (2006, [5])])

$$\lim_{t\to\infty}\mathbb{E}\left[\log(X/t\ ) \mid X > t\right] = \gamma_X \tag{2.13}$$

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

$$k_n \to +\infty\ k_n = o(n). \tag{2.14}$$

Alors nous pouvons définir une version aléatoire de ϕ

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

avec un seuil aléatoire $t = Z_{n-k_n,n}$,

$$\hat{\phi}_n(x) := \frac{1}{1-\hat{F}_n(Z_{n-k_n,n})}\log\left(\frac{x}{Z_{n-k_n,n}}\right)\mathbb{I}_{x>Z_{n-k_n,n}}. \tag{2.15}$$

Par conséquent, en combinant (2.11) et (2.13) 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 (2.12),

$$\hat{\gamma}_{X,Hill}^{KM} := \frac{1}{n(1-\hat{F}_n(Z_{n-k_n,n}))}\sum_{i=1}^{k_n}\frac{\delta_{n-i+1,n}}{1-\hat{G}_n(Z_{n-i+1,n}^{-})}\log\left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}\right) \quad (2.16)$$

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 $poH^{\leftarrow}$, qui sont similaires à celles utilisées dans [*Einmahl et al.* (2008)] : si $p(z) = \mathbb{P}(\delta = 1, Z = z)$

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(\overleftarrow{H}(1-\frac{i}{n})) - p\right] \xrightarrow{P} c \in \mathbb{R} \quad (2.17)$$

$$\sqrt{k}\sup_{(s,t)\in C_n}\left|p(\overleftarrow{H}(t)) - p(\overleftarrow{H}(s))\right| \to 0 \quad C > 0 \quad (2.18)$$

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

Théorème 2.3.[15] Sous les hypothèses $(2.12), (2.14), (2.17), (2.18)$ si l'on suppose en outre que, pour $\delta > 0$

$$-\log(k_n/n)/k_n = O(n^{-\delta}) \quad (2.19)$$

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

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

2.4.2 Deuxième approche ($Leurgans$)

La deuxième approche alternative à l'approche intégrale de Kaplan-Meier présentée dans le paragraphe précédent, est basé sur les idées de [*Leurgans* (1987)], qui ont développé une stratégie " données synthétiques" dans les problèmes de régression censurés (voir [Delacroix et al. (2008)] pour une référence plus récente à cette méthode). Le point de cette seconde approche de départ est le résultat suivant :

Si ϕ et ψ sont deux fonctions non négatif $\mathbb{R}_+ \to \mathbb{R}$ tel que $\int_0^x \psi(t)dt = \phi(x)$,

$$\mathbb{E}\left[\int_0^Z \frac{\psi(t)}{1-G(t)}dt\right] = \mathbb{E}\left[\phi(X)\right] \quad (2.20)$$

Effectivement,

$$\begin{aligned}
\mathbb{E}\left[\int_0^Z \frac{\psi(t)}{1-G(t)}dt\right] &= \int_0^{+\infty}\left(\int_t^{+\infty} dH(z)\right)\frac{\psi(t)}{1-G(t)}dt \\
&= \int_0^{+\infty}(1-F(t))\psi(t)dt \\
&= \int_0^{+\infty}\left(\int_t^{+\infty} dF(x)\right)\psi(t)dt \\
&= \int_0^{+\infty}\left(\int_0^{x}\psi(t)dt\right)dF(x) \\
&= \mathbb{E}\left[\phi(X)\right]
\end{aligned}$$

Par conséquent, si $\phi(x) = \int_0^x \psi(z)dz$, puis

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^{Z_i}\frac{\psi(z)}{1-\hat{G}_n(z^-)}dz$$

devrait correctement estimer $\mathbb{E}\left[\phi(X)\right]$. Cet estimateur peut être réécrite en utilisant la forme particulière de la fonction ψ et la forme constante de l'estimateur de Kaplan-Meier : $\hat{G}_n$ notant, $Z_{0,n} = 0$ et $rk(Z_i)$ du (ordre croissant) rang de Z_i observation dans le Z-échantillon, nous avons en effet,

$$\begin{aligned}
\int_0^{Z_i}\frac{\psi(z)}{1-\hat{G}_n(z^-)}dz &= \sum_{j=1}^{rk(Z_i)}\int_{]Z_{j-1,n},Z_{j,n}]}\frac{\psi(z)}{1-\hat{G}_n(z^-)}dz \\
&= \sum_{j=1}^{rk(Z_i)}\frac{1}{1-\hat{G}_n(z_{j-1,n})}\int_{z_{j-1,n}}^{z_{j,n}}\psi(z)dz \\
&= \sum_{j=1}^{rk(Z_i)}\frac{\phi(z_{j,n})-\phi(z_{j-1,n})}{1-\hat{G}_n(z_{j-1,n})}
\end{aligned}$$

Considérant, encore une fois, la fonction $\hat{\phi}_n$ introduit dans (2.15), nous pouvons maintenant définir la deuxième nouvelle adaptation de l'estimateur de Hill, valable dans le cas (2.12)

$$\hat{\gamma}_{X,Hill}^{Leurg} := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{i}\frac{\hat{\phi}_n(z_{j,n})-\hat{\phi}_n(z_{j-1,n})}{1-\hat{G}_n(z_{j-1,n})}$$

qui se révèle être, après quelques calculs simples,

$$\hat{\gamma}_{X,Hill}^{Leurg} = \frac{1}{n(1-\hat{F}_n(Z_{n-k_n,n}))}\sum_{j=1}^{k_n}\frac{1}{1-\hat{G}_n(Z^-_{n-k_n,n})} i\log\left(\frac{Z_{n-i+1,n}}{Z_{n-i,n}}\right) \quad (2.21)$$

On remarque que, tandis que $\hat{\gamma}_{X,Hill}^{KM}$ apparu comme une version pondérée de la forme classique de l'estimateur de Hill (moyenne des excés des log relatifs $\log(z_{n-i+1,n}/z_{n-k_n,n})$) le deuxième estimateur $\hat{\gamma}_{X,Hill}^{Leurg}$ est une version pondérée (mais toujours avec un poids positif) de la moyenne des espacements $i\log(Z_{n-i+1,n}/Z_{n-i,n})$,
Le théorème suivant fournit la consistance de cet estimateur, sous conditions moins restrictives que le théorème (2.3).

Théorème 2.4.[15] Selon des hypothèses (2.12) , (2.14) et (2.19), si nous supposons $\gamma_X < \gamma_Y$ puis, quand n tend vers ∞

$$\hat{\gamma}_{X,Hill}^{Leurg} \quad \xrightarrow{P} \gamma_X.$$

Chapitre 3

Application du bootstrap sur l'estimateur de Hill de l'indice des valeurs extrêmes sur des données censurées

La méthode du bootstrap a été proposée par *Bradley Efron* (1979) comme une alternative aux modèles mathématiques traditionnels dans des problèmes d'inférence compliqués où une modélisation mathématique de la distribution des erreurs est difficile.

Le mot bootstrap provient de l'expression anglaise "to pull oneself up by one's bootstrap" (Efron, Tibshirani, 1993), qui signifie littéralement "se soulever en tirant sur les languettes de ses bottes".

Le bootstrap est une technique de rééchantillonnage permettant de simuler la distribution d'un estimateur quelconque pour en apprécier le biais, la variance donc le risque quadratique ou encore pour en estimer un intervalle de confiance même si la loi théorique est inconnue.

Le bootstrap est une méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles.

Dans ce chapitre nous appliquons la méthode du bootstrap sur l'estimateur de l'indice de queue de valeurs extrêmes dans le cas des données censurées.

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 $\frac{1}{n}$ 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 $\hat{F}_n$ est proche de F, on aura donc une bonne approximation de la loi de X en utilisant $\hat{F}_n$ 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.

3.2 Bootstrap de l'estimateur de l'indice de queue $\hat{\gamma}_1^{(c,H)}$

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 rééchantillonnage suivant : On génère un échantillon bootstrapé,

$$(Z_1^*, \delta_1^*), ..., (Z_n^*, \delta_n^*) \tag{3.1}$$

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

$$(Z_1, \delta_1), ..., (Z_n, \delta_n) \tag{3.2}$$

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 $D(G_{\gamma_1})$ et $D(G_{\gamma_2})$ respectivement où $\gamma_1 > 0, \gamma_2 > 0$, alors cela signifie que $H \in D(G_\gamma)$, avec $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$. L'estimateur de Hill bootstrapé de l'indice de valeurs extrêmes γ_1 construit avec les données $(Z^*_{i:n}, \delta^*_{i:n})_{i=1,...,n}$ s'écrit :

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

où

$$\hat{\gamma}^{*(H)} = \frac{1}{k}\sum_{i=1}^{k} \log Z^*_{n-i+1,n} - \log Z^*_{n-k,n} \tag{3.3}$$

et

$$\hat{p}^* := \frac{1}{k}\sum_{i=1}^{k} \delta^*_{[n-i+1:n]} \tag{3.4}$$

3.3 Propiétés de l'estimateur de l'indice de queue $\hat{\gamma}_1^{(c,H)}$

Soit l'indice des valeurs extrêmes γ_1 associé à l'échantillon $(X_i)_{1\leq i\leq n}$, et soit $\hat{\gamma}_1^{(c,H)}$: 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 $\hat{\gamma}_1^{(c,H)}$.

$$\hat{\gamma}_1^{*(c,H)}(b) \ , \qquad b = 1, ..., B$$

3.3.1 Estimation Bootstrap de l'erreur standard de $\hat{\gamma}_1^{(c,H)}$

On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs $\hat{\gamma}_1^{*(c,H)}(b)$, la moyenne est :

$$\hat{\gamma}_1^{*(c,H)}(.) = \frac{1}{B}\sum_{b=1}^{B} \hat{\gamma}_1^{*(c,H)}(b) \tag{3.5}$$

L'écart type est aussi une caractéristique importante de chaque distribution. Pour un ensemble d'estimateurs $\hat{\gamma}_1^{*(c,H)}(b)$ l'écart type estimé est calculé par la formule :

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_1^{*(c,H)}(b) - \hat{\gamma}_1^{*(c,H)}(.)\right)^2} \tag{3.6}$$

où B est le nombre total d'échantillons bootstrap .

Algorithme 1 : estimation bootstrap de l'erreur standard

La procédure bootstrap pour estimer l'erreur standard d'un estimateur $\hat{\gamma}_1^{(c,H)}$ est la suivante :

1. On crée B échantillons indépendants $Z^{*1}, Z^{*2}, ..., Z^{*B}$ où chaque échantillon Z^{*b} est obtenu en tirant n observations avec remise dans l'échantillon $Z = \{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$ de départ.
2. Pour chaque échantillon b tel que, $1 \leq b \leq B$, on calcule l'estimateur de queue :
$$\hat{\gamma}_1^{*(c,H)}(b).$$
On obtient alors un échantillon de B valeurs
$$\left\{\hat{\gamma}_1^{*(c,H)}(1), \hat{\gamma}_1^{*(c,H)}(2), ..., \hat{\gamma}_1^{*(c,H)}(B)\right\}$$
3. On estime alors l'erreur standard $se_F\left(\hat{\gamma}_1^{(c,H)}\right)$ par l'erreur standard de cet échantillon de $\hat{\gamma}_1^{*(c,H)}$, i.e
$$\widehat{se}_F\left(\hat{\gamma}_1^{(c,H)}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_1^{*(c,H)}(b) - \hat{\gamma}_1^{*(c,H)}(.)\right)^2}$$
où $\hat{\gamma}_1^{*(c,H)}(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_1^{*(c,H)}(b)$

3.3.2 Réduction du biais de l'estimateur de l'indice de queue $\hat{\gamma}_1^{(c,H)}$

Estimation bootstrap du biais

Définition 3.1
Monde réel. le biais de l'estimateur de l'indice de queue s'exprime comme

$$Biais_F\left(\hat{\gamma}_1^{(c,H)}\right) = E_F\left[\hat{\gamma}_1^{(c,H)}\right] - \gamma_1$$

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

$$\widehat{Biais}_{\hat{F}}\left(\hat{\gamma}_1^{(c,H)}\right) = E_{\hat{F}}\left[\hat{\gamma}_1^{*(c,H)}\right] - \hat{\gamma}_1^{(c,H)}$$

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

Algorithme 2 : estimation bootstrap du biais

Soit $Z = \{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$ l'échantillon observé et $\hat{\gamma}_1^{(c,H)}$, l'estimateur de l'indice de queue, La méthode du bootstrap permet d'estimer le biais de cet estimateur.

1. pour $b = 1 : B$
   - sélectionner l'échantillon bootstrap $Z^{*b} = \left\{(Z_1^{*b}, \delta_1^{*b}), ..., (Z_n^{*b}, \delta_n^{*b})\right\}$, par tirage avec remise dans Z
   - Estimer sur cet échantillon la réplication bootstrap de $\hat{\gamma}_1^{(c,H)}$ par $\hat{\gamma}_1^{*(c,H)}(b)$.
2. Approcher $\mathbb{E}_{\hat{F}}\left[\hat{\gamma}_1^{*(c,H)}\right]$ par $\hat{\gamma}_1^{*(c,H)}(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_1^{*(c,H)}(b)$.
   - L'approximation bootstrap du biais est la différence entre la moyenne de la distribution bootstrap et l'estimateur de l'indice de queue pour les données observées originales :

$$\widehat{Biais}_{boot}\left(\hat{\gamma}_1^{(c,H)}\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_1^{*(c,H)}(b) - \hat{\gamma}_1^{(c,H)} \tag{3.7}$$

Réduction du biais de $\hat{\gamma}_1^{(c,H)}$

Nous allons utiliser la méthode de bootstrap pour la réduction du biais de l'estimateur de l'indice de queue $\hat{\gamma}_1^{(c,H)}$. Corriger le biais de l'estimateur $\hat{\gamma}_1^{(c,H)}$ par la méthode de bootstrap revient à considérer :

$$\hat{\gamma}_{1_{corr}}^{(c,H)} = \hat{\gamma}_1^{(c,H)} - \widehat{Biais}_{boot}\left(\hat{\gamma}_1^{(c,H)}\right).$$

En remplaçant $\widehat{Biais}_{boot}\left(\hat{\gamma}_1^{(c,H)}\right)$ par (3.7), on obtient

$$\hat{\gamma}_{1_{corr}}^{(c,H)} = 2\hat{\gamma}_1^{(c,H)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_1^{*(c,H)}(b) \tag{3.8}$$

3.3.3 Estimation Bootstrap de l'erreur quadratique moyenne de $\hat{\gamma}_1^{(c,H)}$

Définition 3.2 [8]
Monde réel. l'erreur quadratique moyenne (MSE) de $\hat{\gamma}_1^{(c,H)}$ est égale à

$$MSE_F = \mathbb{E}_F\left[\left(\hat{\gamma}_1^{(c,H)} - \gamma_1\right)^2\right]$$

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

$$\widehat{MSE}_{\hat{F}} = \mathbb{E}_{\hat{F}}\left[\left(\hat{\gamma}_1^{*(c,H)}(b) - \hat{\gamma}_1^{(c,H)}\right)^2\right]$$

Algorithme 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 $\hat{\gamma}_1^{*(c,H)}(b)$ réplication bootstrap de $\hat{\gamma}_1^{(c,H)}$
FinPour
Retourner

$$\widehat{MSE}_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\gamma}_1^{*(c,H)}(b) - \hat{\gamma}_1^{(c,H)}\right)^2 \tag{3.9}$$

Fin

3.4 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 $\hat{\gamma}_1^{*(c,H)}(b)$. L'algorithme est le suivant [16] :

Algorithme 4 : Estimation Bootstrap de l'intervalle de confiance

1. Générer B échantillons bootstrap $\left(Z^{*1}, ..., Z^{*B}\right)$
2. Calculer pour chacun les réplications bootstrap $\hat{\gamma}_1^{*(c,H)}(b)$

3. Soient $\hat{\gamma}^{*(c,H)}_{1,B(\alpha/2)}$ et $\hat{\gamma}^{*(c,H)}_{1,B(1-\alpha/2)}$ respectivement $B\alpha/2$-ième et $B(1-\alpha/2)$-ième percentile de $\hat{\gamma}^{*(c,H)}_{1}(b)$ dans la liste ordonnée des B réplications de $\hat{\gamma}^{*(c,H)}_{1}$

   - L'intervalle approximé est alors

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

3.5 Simulations

3.5.1 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 $(X_i)_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_1)$ de taille $n = 1000$, à partir d'une variable u de $U([0,1])$, le modèle ajusté sera :

$$F^{-1}(u) = (1-u)^{-\gamma_1}$$

L'échantillon $(X_i)_{1\leq i\leq n}$est censuré par un deuxième échantillon $(Y_i)_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_2)$ à partir d'une variable v de $U([0,1])$:

$$G^{-1}(v) = (1-v)^{-\gamma_2}$$

Les variables que nous observons sont d'une part les $Z_i \backsim \text{Pareto}(\gamma)$ 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
gamma1<- ?
gamma2<- ?

# échantillon initial
```

```
u<-runif(n,0,1)
x<-(1-u)^(-gamma1)
v<-runif(n,0,1)
y<-(1-v)^(-gamma2)
z<-pmin(x,y)
delta=as.numeric(x<=y)
z
delta
plot(z)
r<-rank(z)
#ordonnées les données

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)

# graphe hills vs k.
hills<-seq(1,n-1)
for(k in 1 :n-1)
{
l<-seq(1,k)
hills[k]<-(((1/k)*sum(Z[n-l+1]))-Z[n-k])/((1/k)*sum(Delta[n-l+1]))
}
k<-seq(1,n-1)
plot(k,hills)
```

```
abline(h=gamma1, col="red", lwd="2").

# k optimal graphique
kgr<-k[which.min(abs(gamma1-hills))]
kgr
hillgrph<-hills[kgr]
hillgrph
```

3.5.2 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'estimateur $\hat{\gamma}_1^{(c,H)}$.

Méthode basée sur l'erreur quadratique moyenne

La valeur optimale de k peut être obtenue par la minimisation de *l'erreur quadratique moyenne* de l'estimateur. On peut se baser sur des méthodes de *Bootstrap* pour calculer (MSE).
Pour toute réplication R nous estimons γ_1 et soit $\hat{\gamma}_{k_n}^{(c,H),j}$ l'estimateur de γ_1 obtenu à la $j - iéme$ réplication $(j = 1, ..., R)$ avec $(k_n = 1, ..., n-1)$. Il semble donc naturel de trouver une valeur k^{opt} qui minimise les valeurs de l'erreur quadratique moyenne $\{(k_n, MSE(k_n), \quad k_n = 1, ..., n-1)\}$ par rapport à k. La valeur optimale de k_n est donnée par :

$$k_n^{opt} := \underset{1 \le k \le n-1}{\arg\min} \left\{ \frac{1}{R} \sum_{j=1}^{R} \left(\hat{\gamma}_{k_n}^{(c,H),j} - \gamma_1 \right)^2 \right\}. \tag{3.10}$$

Il est donc facile de voir que la MSE de $\hat{\gamma}_{k_n}^{(c,H)}$, qui est en fonction de k_n n'est rien d'autre que le carré du biais plus la variance de l'estimateur, ils est nécessaire 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.

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

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

```
R=300
a<-floor(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(Zboot[n-l+1]))-Zboot[n-k])/((1/k)*(sum(Deltaboot[n-
l+1])))
hills
}
hills
EQM[k]<-(1/R)*sum((hills-gamma1)^2)
}
EQM
kopt<-which.min(EQM[a :b])+a-1
kopt
```

3.5.3 Estimateur bootstrap de $\hat{\gamma}_1^{(c,H)}$

Pour construire des échantillons bootstrapés, on utilise la commande :

```
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice] ;deltaboot<-delta[indice]
```

Procédure sous R

```
# estimateur initial du kopt
l<-seq(1,kopt)
hill1<-(((1/kopt)*sum(Z[n-l+1]))-Z[n-kopt])/((1/kopt)*sum(Delta[n-l+1]))
hill1

# hills censurées Bootstrap
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,kopt)
hillls[s]<-((1/kopt)*sum(Zboot[n-l+1])-Zboot[n-kopt])/((1/kopt)*sum(Deltaboot[n-
l+1]))
}
réé<-cbind(hillls)
réé
hilllboot<-mean(hillls)
qqnorm(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)
```

3.5.4 Comportement de l'estimateur $\hat{\gamma}_1^{(c,H)}$ et de ses propiétés vs n

Dans le but d'observer la stabilité de l'estimateur $\hat{\gamma}_1^{(c,H)}$, 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.

```
s=1000
alpha=0.05
R1=1000
hill1n<-seq(1 :s)
hilllboot<-seq(1 :s)
Sd<-seq(1 :s)
biais<-seq(1 :s)
EQMboot<-seq(1 :s)
icinf<-seq(1 :s)
icsup<-seq(1 :s)
for (n in 90 :s)
{
gamma1<-0.35
gamma2<-1
u<-runif(n,0,1)
x<-(1-u)^(-gamma1)
v<-runif(n,0,1)
y<-(1-v)^(-gamma2)
z<-pmin(x,y)
delta=as.numeric(x<=y)
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)
R=300
a<-floor(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]
}
A<-sort(L)
for (i in 1 :length(zboot))
{
B[i]<-A[i]/r1[i]
}
Deltaboot<-B-length(zboot)
Deltaboot
Zboot<-log(Y)
l<-seq(1,k)
hills[j]<-(((1/k)*sum(Zboot[n-l+1]))-Zboot[n-k])/((1/k)*(sum(Deltaboot[n-
l+1])))
}
EQM[k]<-(1/R)*sum((hills-gamma1)^2)
}
```

```
kopt<-which.min(EQM[a :b])+a-1
l<-seq(1,kopt)
hill1<-(((1/kopt)*sum(Z[n-l+1]))-Z[n-kopt])/((1/kopt)*sum(Delta[n-l+1]))
hillls<-seq(1,R1)
for(s 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)
l<-seq(1,kopt)
hillls[s]<-((1/kopt)*sum(Zboot[n-l+1])-Zboot[n-kopt])/((1/kopt)*sum(Deltaboot[n-
l+1]))
}
hilllboot[n]<-mean(hillls)
Sd[n]<-sd(hillls)
biais[n]<-mean(hillls)-hill1
EQMboot[n]<-(1/R1)*sum((hillls-hill1)^2)
Sor<-sort(hillls)
icinf[n]<-Sor[R1*(alpha/2)]
icsup[n]<-Sor[R1*(1-alpha/2)]
hill1n[n]<-hill1
}
hill1n
hilllboot
Sd
biais
EQMboot
```

```
    icinf
    icsup

    # estimateur Boot et intial
    plot( hilllboot, xlim = c(130,1000), ylim = c(0.2,1),xlab=expression("n"),
type ="l", col = "blue")
    lines (hill1n, type ="l", col = "2" )
    abline(h=gamma1, col="3", lwd="1")

    # Sd, Biais, EQMboot
    plot(Sd, xlim = c(125,1000), ylim = c(0,0.2),xlab=expression("n"), type
="l",col = "3")
    plot(biais, xlim = c(130,1000), ylim = c(-0.01,0.2),xlab=expression("n"),
type ="l",col = "4")
    plot(EQMboot, xlim = c(130,1000),ylim = c(0,0.02),xlab=expression("n"),
type ="l",col = "11")

    # intervalle de confiance
    plot(icinf, xlim = c(110,1000), ylim=c(0.2,0.6),xlab=expression("n"), type
="l", col = "6")
    lines (hill1n,xlab=expression("n"), type ="l", col = "2" )
    lines (icsup,xlab=expression("n"), type ="l", col = "6" )
```

3.6 Résultats des simulations

3.6.1 Simulation bootstrap de l'estimateur $\hat{\gamma}_1^{(c,H)}$ vs k

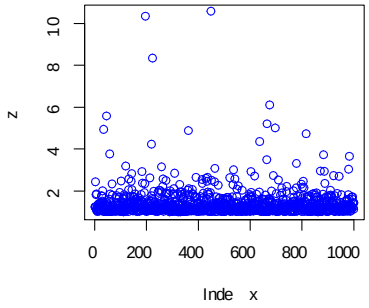

FIG. 3.1 – $\gamma = \gamma_1\gamma_2/(\gamma_1+\gamma_2)$, pour $\gamma_1 = 0.35$ et $\gamma_2 = 2.5$, (10% de censure)

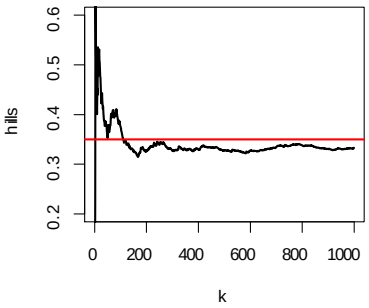

FIG. 3.2 – Comportement graphique de $\hat{\gamma}_1^{(c,H)}$ vs k issue de la distribution de Pareto ($\gamma_1 = 0.35$) censurées par Pareto ($\gamma_2 = 2.5$) , (10% de censure)

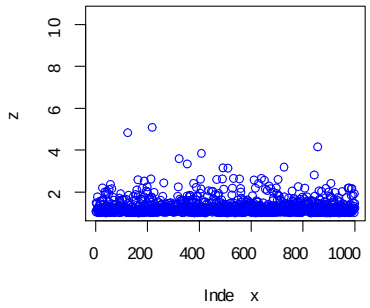

FIG. 3.3 – $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$, pour $\gamma_1 = 0.35$ et $\gamma_2 = 1$, (25% de censure)

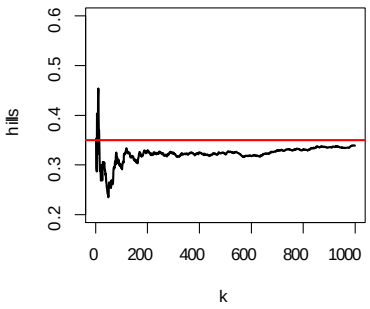

FIG. 3.4 – Comportement graphique de $\hat{\gamma}_1^{(c,H)}$ vs k issue de la distribution de Pareto ($\gamma_1 = 0.35$) censurées par Pareto ($\gamma_2 = 1$), (25% de censure)

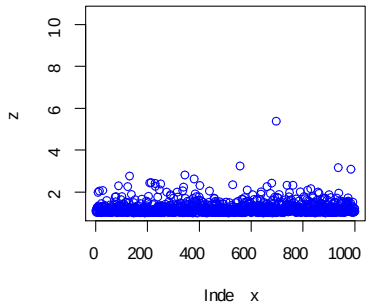

FIG. 3.5 – $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$, $\gamma_1 = 0.35$ et $\gamma_2 = 0.5$, (40% de censure)

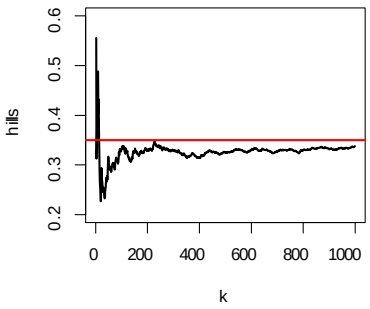

FIG. 3.6 – Comportement graphique de $\hat{\gamma}_1^{(c,H)}$ vs k issue de la distribution de Pareto ($\gamma_1 = 0.35$) censurées par Pareto ($\gamma_2 = 0.5$), (40% de censure)

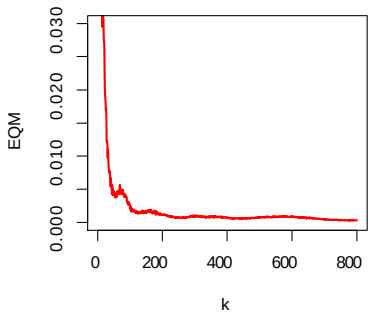

FIG. 3.7 – MSE de $\hat{\gamma}_1^{(c,H)}$ bootstrap de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 2.5$) , (10% de censure), $n = 1000$, $k_{opt} = 786$

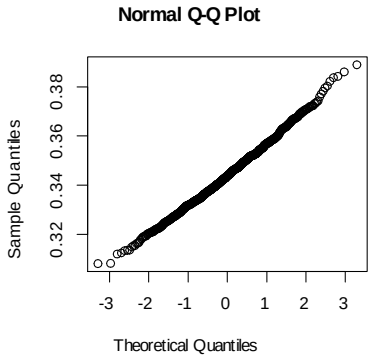

FIG. 3.8 – QQ-norm de la distribution limite bootstrap de 1000 répétition de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 2.5$) , (10% de censure), $n = 1000$

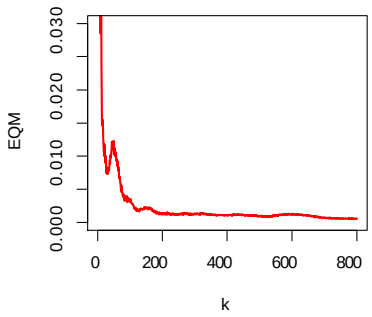

FIG. 3.9 – MSE de $\hat{\gamma}_1^{(c,H)}$ bootstrap de 1000 répétitions de Pareto $(\gamma_1 = 0.35)$ censurée par Pareto $(\gamma_2 = 1)$, (25% de censure), $n = 1000$, $k_{opt} = 772$

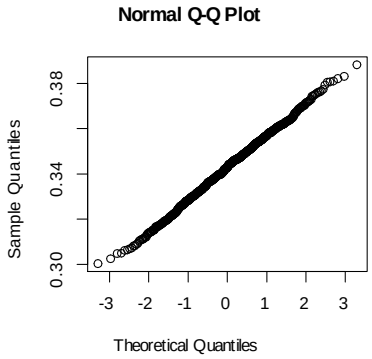

FIG. 3.10 – QQ-norm de la distribution limite bootstrap de 1000 répétitions de Pareto $(\gamma_1 = 0.35)$ censurée par Pareto $(\gamma_2 = 1)$, (25% de censure), $n = 1000$

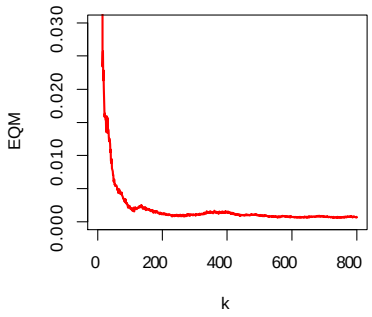

FIG. 3.11 – MSE de $\hat{\gamma}_1^{(c,H)}$ bootstrap de 1000 répétitions de Pareto $(\gamma_1 = 0.35)$ censurée par Pareto $(\gamma_2 = 0.5)$, (40% de censure), $n = 1000$, $k_{opt} = 747$

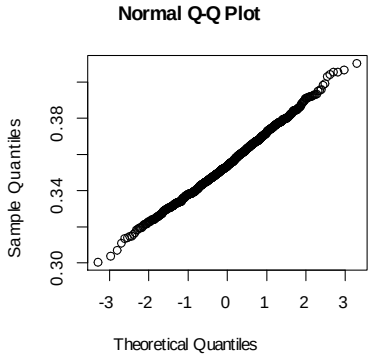

FIG. 3.12 – QQ-norm de la distribution limite bootstrap de 1000 répétitions de Pareto $(\gamma_1 = 0.35)$ censurée par Pareto $(\gamma_2 = 0.5)$, (40% de censure), $n = 1000$

Commentaires.

Sur les figures (*Fig.3.1, Fig.3.3, Fig.3.5*) nous avons ploté l'échantillon Z (γ) qui représente l'échantillon X (γ_1) censurée par l'échantillon Y (γ_2) où nous avons conclut que plus le censure augmente plus, les valeurs extrêmes diminuent.

Les figures (*Fig.3.2, Fig.3.4, Fig.3.6*), montrent que l'estimateur $\hat{\gamma}_1^{(c,H)}$ présente une allure très proche de la valeur réelle de γ_1. En particulier, nous avons constaté que l'estimateur devient très stable à partir de $k/n = 0.2$.

Après plusieurs simulations en variant la valeur de γ_2, nous avons constaté que l'augmentation du censure influence sur la stabilité de l'estimateur, c.-à-d., plus le pourcentage du censure $p = (40\%, 25\%, 10\%)$ décroit plus l'estimateur devient encore plus stable.

Les graphiques (*Fig.3.7, Fig.3.9, Fig.3.11*) montrent que l'erreur quadratique moyenne Bootstrap est très petite, sa stabilité dépend directement de la stabilité de son estimateur $\hat{\gamma}_1^{(c,H)}$.

Bien qu'en minimisant l'erreur quadratique moyenne Bootstrap par rapport à k , nous obtenons notre k_{opt} optimal un peu plus élevé pour la simulation de l'estimateur..

L'erreur quadratique moyenne est stable à partir de $k/n = 0.2$, ce qui nous permettra un large choix du nombre de valeurs extrêmes k.

Les graphiques (*Fig.3.8, Fig.3.10, Fig.3.12*) des QQ-norm qui correspondent aux distributions limites empiriques de l'estimateur $\hat{\gamma}_1^{(c,H)}$ Bootstrap pour $\gamma_2 = (2.5, 1, 0.5)$ respectivement illustrent clairement par sa linéarité que la normalité asymptotique est fortement confirmé et théoriquement et empiriquement.

3.6.2 Simulation bootstrap de l'estimateur $\hat{\gamma}_1^{(c,H)}$ vs n

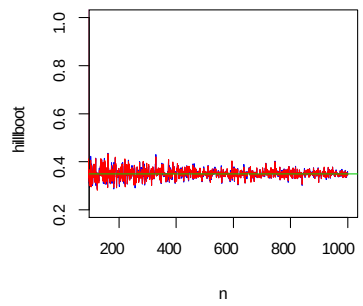

FIG. 3.13 – $\hat{\gamma}_1^{(c,H)}$et $\hat{\gamma}_{boot}^{(c,H)}$ bootstrap de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 1$).

On observe que $\hat{\gamma}_{boot}^{(c,H)}$ converge vers $\hat{\gamma}_1^{(c,H)}$ Pour toutes les valeurs de n, de 100 jusqu'à 1000, et tout les deux sont très proche de γ_1. Pour des échantillons de petites taille ($n \leq 400$) à peu prés, l'estimateur est relativement perturbé, car sa consistance demande un nombre suffisamment grand d'observation pour converger vers la valeur théorique γ_1.

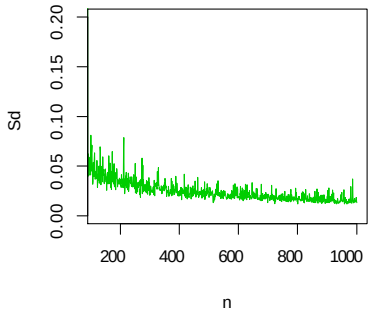

FIG. 3.14 – *Ecart − type* bootstrap de $\hat{\gamma}_1^{(c,H)}$ de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 1$).

L'écart-type prend en générale des valeurs très petites. Plus n est grand plus L'écart-type décroit vers 0.

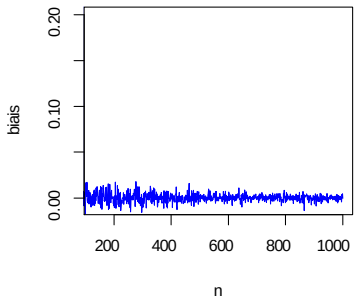

FIG. 3.15 – *Biais* bootstrap de $\hat{\gamma}_1^{(c,H)}$ de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 1$) .

Le biais est significativement petit, plus n est grand plus il est stable.

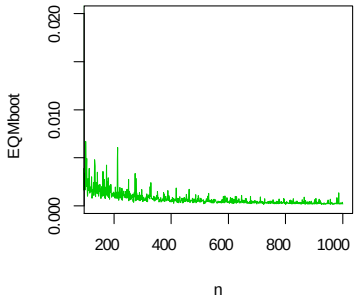

FIG. 3.16 – *MSE* bootstrap de $\hat{\gamma}_1^{(c,H)}$ de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 1$) .

On remarque que L'erreur quadratique moyenne est stable et petit à partir de $n \geq 400$.

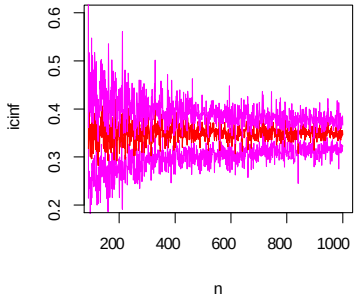

FIG. 3.17 – IC bootstrap de $\hat{\gamma}_1^{(c,H)}$ de 1000 répétitions de Pareto ($\gamma_1 = 0.35$) censurée par Pareto ($\gamma_2 = 1$).

Le graphe en rouge représente l'estimateur $\hat{\gamma}_1^{(c,H)}$ Les graphes en rose représentent l'estimation bootstrap des intervalles confiances sup et inf. ils deviennent plus exacte lorsque n est plus grand, en se rapprochant de plus en plus vers $\hat{\gamma}_1^{(c,H)}$.

Nous avons simuler trois échantillons de taille $n = 1000$ de pareto ($\gamma_1 = 0.35$), censuré par un échantillon de taille $n = 1000$ de pareto.(γ_2) pour différentes valeurs du paramètre ($\gamma_2 = 0.5, 1, 2.5$).

Le caractère c représente le pourcentage de censure ($10\%, 25\%, 40\%$).

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

	$c = 10\%$		$c = 25\%$		$c = 40\%$	
k_{opt}	786		772		747	
$\hat{\gamma}_1^{(c,H)}$	0.340511		0.3316888		0.3303923	
$\hat{\gamma}_{1_{boot}}^{(c,H)}$	0.3396358		0.3301014		0.3304379	
$\hat{\gamma}_{corr}^{(c,H)}$	0.3413862		0.3332762		0.3303467	
sd	0.01331123		0.01363844		0.01534961	
$biais$	-0.0008752431		-0.001587428		$4.558391e-05$	
MSE	0.00017777		0.0001883411		0.0002353769	
IC	$ic\,\inf$	$ic\,\sup$	$ic\,\inf$	$ic\,\sup$	$ic\,\inf$	$ic\,\sup$
	0.3151654	0.3648449	0.3035091	0.3586323	0.3007389	0.3606425

TAB. 3.1 – Résultats de simulation pour n=1000

Conclusion.

Nous avons utilisé la méthode du bootsratp dans le but d'étudier les indicateurs de dispersion $(sd, Biais, MSE)$ de l'estimateur de l'indice de queue dans le cas du domaine d'attraction de Fréchet pour des données censurées à droite noté $\hat{\gamma}_1^{(c,H)}$. Notamment notre but essentiel était de réduire son Biais par un estimateur corrigé qu'on a noté $\hat{\gamma}_{corr}^{(c,H)}$, ainsi confirmer son comportement asymptotique est proposer des intervalles de confiances bootstrap très significatives.

Nous avons constaté la puissance de la méthode du bootstrap vu son application qui ne nécessite pas la vérification de toutes les conditions théoriques pour la convergence, cette dernière joue le double rôle, en service pour vérifier la qualité de l'estimation avant de passer aux étapes théoriques et en même temps pour confirmer des résultats auparavant posées. Un rôle primordial dans les applications réelles vu qu'un praticien veut obtenir des résultats directs de la machine sans passer par la vérification théorique. Cette méthode aussi peut facilement nous alerter quand il s'agit des perturbations de stabilités, et tester la taille optimal de l'échantillon observées vue sa capacité de génération.

Nous avons pu sortir avec des perspectif que nous souhaiterons continuer à travailler d'avantage.

. Faire une comparaison entre les propriétes théoriques et empiriques bootstrap.

. Effectuer une comparaison entre les estimateurs de l'indice proposés dans la littérature.

. Faire une étude plus approfondi sur le choix du nombre de valeurs extrêmes vu que l'erreur quadratique moyenne présente des petites valeurs tôt avant d'arriver a sa valeur minimale, en prenant en compte la vraie valeur du nombre des extrêmes de l'échantillon.

. Nous allons être un peu plus ambitieuse et penser un proposer un nouveau estimateur de l'indice de queue dans le même cas étudié.

. De nombreuses applications sont permises par l'utilisation de cet estimateur pour d'autre statistique comme les mesures de risque et les quantiles extrêmes ainsi que des tests de normalité.

Bibliographie

[1] Berkane, H., 2005, Méthodes du Bootstrap pour les queues de distributions, université Mohamed Khider, Biskra, Algeria.

[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] Brahimi, B., Meraghni, D., Necir, A.,2013, On the asymptotic normality of hill's estimator of the tail index under random censoring. Preprint : arXiv-1302.1666,

[5] de Haan, L., and Ferreira, A., 2006, Extreme Values Theory : An introduction. New York,Springer.

[6] Deme, E.H., 2013, Quelques contributions à la Théorie univariée des Valeurs Extrêmes et Estimation des mesures de risque actuariel pour des pertes à queues lourdes,Université Gaston Berger.

[7] Einmahl, J.H.J.,Fils-Villetard, A., Guillou, A,2008, Statistics of extremes under random censoring. Bernoulli, 14 :207–227.

[8] Fromont, M. et Vimond,M., 2012, Bootstrap et rééchantillonnage,Université Européenne de Bretagne.

[9] Ivette Gomes,M.,and Manuela Neves,M.,2010,Estimation of the Extreme Value Index for Randomly Censored Data.

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

[11] Necir,A.,Brahimi.B.,Meraghni,D.,2014,Approximations to the tail index estimator of a heavy-tailed distribution under random censoring and application, Mohamed Khider University, Biskra, Algeria.

[12] Reiss,R.-D.,Thomas,M.S.,2007, Statistical analysis of extreme values. From insurance,finance, hydrology and other fields. Birkhäuser Verlag, Basel., Boston, Berlin.

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

[14] Toulemonde,G.,2008, Estimation et tests en théorie des valeurs extrêmes,thése de doctorat de l'université Paris VI.

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

[16] Inférence Statistique Assistée par Ordinateur - 2A, 2006-2007

Résumé. Nous considérons le problème de l'estimation de l'indice des valeurs extrêmes censurées aléatoirement à droite. Nous nous concentrons sur le domaine d'attraction de Fréchet dans le cas des distributions à queues lourdes. Lorsque les données sont complètes, l'indice des valeurs extrêmes le plus fameux est l'estimateur de Hill (1975), Einmahl et al. (2008) l'ont ajusté au cas où les données sont censurées à droite, Ils ont aussi établi sa normalité asymptotique. Nous avons appliqué la technique du ré-échantillonnage pour vérifier les indicateurs de dispersion de l'estimateur en question. Notamment, la réduction du biais et l'illustration des intervalles de confiance sont effectuées.

Mots clés : théorie des valeurs extrêmes, distribution à queue lourde, censure aléatoire à droite, indice des valeurs extrêmes, estimateur de Hill, méthode de Ré-échantillonnage, intervalles de confiance.

Abstract.
We consider the estimation problem of the extreme value index in the presence of censoring we focus on the Fréchet domains of attraction, in the case of no censoring data. The most famous estimator of the Pareto index in the classical hill estimator (1975), Einmahl et al. (2008) adjusted him for randomly censored data. They also established its asymptotic normality. We applied the bootstrap to check its dispersion indicators. In particular, the reduction in bias and illustration of confidence intervals are performed.

Key words : extreme value theory, heavy-tailed distribution, random censoring, extreme value index, Hill estimator, bootstrap, confidence intervals.