Université Mohamed Khider
Biskra, Algeria
Laboratoire de Mathématiques Apliquées
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département de Mathématiques

Thèse

Présentée En Vue De l'Obtention Du

DIPLÔME DE DOCTORAT EN MATHÉMATIQUES

Option : Probabilité et Statistique

Par

Louiza SOLTANE

Titre

Analyse des Valeurs Extrêmes en Présence de Censure

Sous la Direction de Mr. ***Djamel MERAGHNI***

Membres de Comité d'Examen

Abdelhakim NECIR	Professeur	Université de Biskra	Président
Hocine FELLAG	Professeur	Université de Tizi-Ouzou	Examinateur
Fatah BENATIA	M.C.(A)	Université de Biskra	Examinateur
Brahim BRAHIMI	M.C.(A)	Université de Biskra	Examinateur
Djabrane YAHIA	M.C.(A)	Université de Biskra	Examinateur

2016

À ma Mère bien-aimée,

À l'esprit de mon Père bien-aimé,

À tous ceux qui le méritent.

Remerciements

Tout d'abord je remercie Allah qui m'a donné la volonté et le courage pour pouvoir réaliser ce travail.

Mes sincères remerciements à mon encadreur Monsieur Djamel Meraghni, Professeur à l'université Mohamed Khider de Biskra, pour avoir accepté d'encadrer cette thèse. Un grand merci pour son extréme patience au cours de ces quatre dernières années. Je suis profondément reconnaissante pour ses nombreux conseils, pour ses corrections et son soutien indéfectible tout au long de ce travail, il me faudrait des pages pour le remercier. Il m'a donné l'occasion de travailler sur un thème fascinant et intéressant, me permettant de mieux comprendre les concepts théoriques de la statistique par leur application à des cas concrets.

Mes plus vifs remerciements vont à Monsieur Abdelhakim Necir, Professeur à l'université de Biskra, pour sa disponibilité, ses précieux conseils, son sens de l'écoute et pour toute l'aide qu'il m'a apporté, malgré ses nombreuses responsabilités. Je le remercie infiniment de me faire l'honneur de présider le jury de ma thèse.

Je tiens également à exprimer ma gratitude envers le Professeur Hocine Fellag , de l'université de Tizi- Ouzou, ainsi qu'à Messieurs Fatah Benatia, Brahim Brahimi et Djabrane Yahia , de l'université de Biskra, d'avoir accepté sans hésitation de faire partie du jury d'examination. Je vous remercie énormément.

Je voudrais exprimer toute ma reconnaissance à tous les enseignants du département de mathématiques, ainsi qu'à tous les membres du Laboratoire de Mathématiques Appliquées "LMA" et aux collègues qui m'ont aidée de près ou de loin dans mon travail et en particulier mes collègues et amies Sana Benameur et Hadda Saidane, dont le soutien conseils et m'ont été très précieux.

Un merci sans limites à ma très chère et vertueuse maman, qui a sû comment me tisser et m'illuminer le chemin de la réussite. Une mention spéciale aux êtres qui me sont très chers: mes sœurs et mon frère Aissa et mon neveu Annas Abd.Arahmane Bessaker . Je leur exprime toute ma gratitude pour le soutien et l'encouragement qu'ils m'ont apporté pour mener à bien mon travail. Enfin, je ne terminerais pas sans citer mes chères amies Sara Habba et Dalal Houhou et toutes les autres amies pour les excellents moments passés ensemble.

Merci à Allah pour tout.

Table des matières

Table des figures

Liste des tableaux

Publications et Communications

Articles

1. Statistical estimate of the proportional hazard premium of loss under random censoring. Journal Afrika Statistika, 11(1), 883 – 899. (**avec. Meraghni & A. Necir,** 2016)**.**

2. Estimating the mean of a heavy-tailed distribution under random censoring, *soumis.* (**avec D. Meraghni & A. Necir,** 2016)**.**

3. EVT-based confidence intervals for the parameters of a symmetric Lévy-stable distribution, *en préparation.* (**avec D. Meraghni**)**.**

Communications

1. Sur le théorème de Fisher-Tippett. Biskra, Algeria, Séminaire Semestriel de Statistique (2012).

2. Estimating risk measures under random censoring. Biskra, Algeria, Séminaire Semestriel de Statistique (2013).

3. Estimating the mean of a heavy-tailed distribution under random censoring. Biskra, Algeria, Journées de Statistique. (2014).

4. Estimating the mean of a heavy-tailed distribution under random censoring. Biskra, Algeria, Journées de Statistique. (2016).

5. Estimating the mean of a heavy-tailed distribution under random censoring. Constantine, Algeria, 1^{er} Séminaire National de Mathématiques. (2016).

Abreviations et Notations

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

Abreviations ou Notations	Explication
$al.$	: Autres.
$\mathcal{D}(.)$	: Domaine d'attraction du maximum.
EVD	: Distribution des valeurs extrêmes.
$\mathbf{E}[X]$	: Espérance mathématique ou moyenne du va X.
F	: Fonction de répartition (fdr).
F_n	: Fonction de répartition empirique.
$\widehat{F}_n$	: Estimateur de Kaplan-Meier.
F^{-1}	: Inverse généralisé de F ou fonction des quantiles.
f	: Densité de probabilité d'une va.
$\mathcal{GEVD}$	: Distribution de valeurs extrêmes généralisée.
$\mathcal{GPD}$	: Distribution de Pareto généralisée.
$\mathcal{H}_\gamma$	: Famille de la lois de valeurs extrêmes généralisée.
IVE	: Indice des valeurs extrêmes.
iid	: Indépendantes et identiquement distribuées.
$\inf A$	: Infimum de l'ensemble A.
$\mathbb{1}_A$	: Fonction indicatrice de l'ensemble A.
ℓ	: Fonction à variation lente.
LGN	: Loi des grands nombres.
μ	: Espérance ou moyenne d'une va.
$\mathcal{N}(0,1)$	: Loi normale standard, ou distribution Gaussienne standard.
$o_{\mathbb{P}}(.)$	: Converge vers 0 en probabilité.
$O_{\mathbb{P}}(.)$	: Être borné en probabilité.
$(\Omega, \mathcal{F}, \mathbb{P})$	: Espace probabilisé.
$p.s.$	: Prèsque sûre.
Q	: Fonction de quantile.
Q_n	: Quantile empirique.
$\mathbb{R}$	: Ensemble des valeurs réelles.

σ^2	: Variance d'une variable aléatoire.
$\mathcal{RV}_\rho$	: Variation régulière au ∞ avec l'indice ρ.
$rmse$	: La racine de l'erreur moyenne quadratique.
$S = \overline{F}$	: Fonction de survie.
S_n	: Somme arithmétique.
$S_\alpha(\sigma, \beta, \mu)$	: Loi stable de paramètre $\alpha,\ \sigma,\ \beta$ et μ.
$\sup A$	: Supremum de l'ensemble A.
t_s	: Quantile d'ordre s.
TCL	: Théorème Centrale Limite.
TVE	: Théorie des valeurs extrêmes.
u	: Seuil.
va	: Variable aléatoire.
$va's$	: Variables aléatoires.
$var(X)$	: Variance mathématique du va X.
$X_{n:n}$	: Maximum de $X_1, ..., X_n$.
$X_{1:n}$	: Minimum de $X_1, ..., X_n$.
$X_{k:n}$	: $k^{ème}$ statistique d'ordre.
$X_1, ..., X_n$	: Échantillon de taille n de X.
x_F	: Point terminal.
$\Lambda_n(t)$	: Estimateur de Nelson-Aalen.
$\overline{X}$	: Moyenne arithmétique.
$\Pi_\varrho(R)$	: La prime de réassurance.
$[a]$	: Partie entière de a.
$\|.\|$	: Valeur absolue.
$:=$	: Égalité par définition.
$\stackrel{\mathcal{D}}{=}$	: Égalité en distribution.
$\stackrel{\mathcal{D}}{\longrightarrow}$	: Converge en distribution.
$\stackrel{\mathbb{P}}{\longrightarrow}$	: Converge en probabilité.
$\stackrel{p.s.}{\longrightarrow}$	: Convegence prèsque sûre.

Introduction

L'analyse de survie est une branche de statistique souvent liée à l'étude des durées de survie dans les applications médicales. En plus de la mort d'organismes biologiques, l'analyse de survie peut s'étendre et s'intéresser à l'échec de systèmes mécaniques et/ou électroniques, dans ce cas on l'appelle "analyse de fiabilité". Ce thème trouve aussi beaucoup d'applications dans les sciences sociales, économiques et actuarielles, où on l'appelle "analyse de durée". Pour des raisons de commodité, les termes propres à la survie biologique sont souvent les plus utilisés.

Dans l'analyse de survie, il est très commun de se trouver en face du problème de données manquantes. Les données de survie ne sont pas totalement observées. Il n'est pas rare, mais elles sont plutôt incomplètes. La censure et la troncature sont les deux causes de données incomplètes les plus répandues. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt. On sait bien que ce délai appartient à un certain intervalle de temps. La troncature survient qu'on ne peut pas observer les individus de l'échantillon dont le délai de survenue appartient à un certain intervalle de temps, on observe donc un sous-échantillon. Dans ce cas les techniques classiques ne s'adaptent pas correctement aux données incomplètes.

La littérature est beaucoup plus riche en censure que la troncature qui est plus récente. Dans cette thèse, on va s'intéresser particulièrement à la censure droite dans le cadre d'apporter de nouveaux résultats. Pour des détails complets sur la censure et l'analyse de survie, on réfère aux livres de Cox et Oakes [28], Kalbfleisch et Prentice [82], Lee et Wang [86], Klein et Moeschberger [84], Wienke [134] et Hanagal[75].

En 1958, Kaplan et Meier [83] ont introduit un estimateur (portant leurs noms) de la fonction de survie des données censurées. Cet estimateur possède des propriétés asymptotiques très populaire (convergence uniforme, presque sure, normalité asymptotique) similaires à celles de la fonction de répartition empirique. Le comportement asymptotique de l'estimateur de Kaplan-Meier a suscité l'intérêt d'un grand nombre d'auteurs, Breslow et Crowley [21] sont les premiers à traiter la convergence et la normalité asymptotique de l'estimateur de Kaplan-Meier. Pour plus de détails, on renvoie au livre de Shorack et Wellner [120].

Dans le cas de non censure, il y a toute une théorie (théorie des valeurs extrêmes : TVE). Pour l'analyse des extrêmes qui se fait selon deux approches. La première, qu'on appellera approche GEV ; permet de modéliser les block maxima par une distribution GEV (generalized extreme value distribution) et la seconde, appelée approche GPD consiste à ajuster les observations dépassant un certain seuil (peaks

over threshold : POT) par une GPD (generalized Pareto distribution). Pour une description détaillée de la TVE, en particulier sur l'estimation de l'indice des valeurs et quantiles extrêmes, consulter les excellents bouquins comme Embrechts et al. [46], Coles [27], Beirlant et al. [9], Reiss et Thomas [111] et Novak [105]. On essaie à travers cette thèse d'adapter les outils de la TVE sans censure au cas de données à queues lourdes censurées.

L'analyse des valeurs extrêmes de censure aléatoire est un nouveau sujet de recherche. L'objet de cette thèse est d'étendre les résultats de la théorie des valeurs extrêmes dans le cas où l'échantillon consiste en un ensemble de données censurées tout en apportant les modifications nécessaires. Des estimateurs sont proposés dans le bouquin de Reiss et Thomas [111], mais sans résultats asymptotiques. Un premier pas, dans l'analyse du comportement asymptotique, des estimateurs de l'indice des valeurs extrêmes et des quantiles extrêmes sous censure, est fait par Beirlant et al. [11]. Leurs estimateurs sont basés 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. L'année suivante Einmahl et al. [44] ont utilisé le même concept pour proposer un estimateur adapté de l'indice de queue dans le cas où les données sont censurées par un seuil aléatoire et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique. En outre, ils ont appliqué leurs résultats sur des données du SIDA. Ces données sont dans le package MASS sous R dans la base de données Aids2. Gomes et Neves [66] ont également contribué à ce domaine en fournissant une étude de simulation détaillée et en appliquant les procédures d'estimation sur certains ensembles de données de survie. À ce jour, beaucoup d'auteurs s'attardent pour améliorer l'estimation de l'indice des valeurs extrêmes (IVE) et parfois leurs quantiles extrêmes comme Ndao et al. [97], Worms et Worms [136], Brahimi et al. [18], Ndao et al. [98], Stupfler [123] et Beirlant et al. [6].

On sait que la moyenne est un paramètre de localisation d'une distribution, elle est aussi très importante dans la synthèse des valeurs de distribution. Malheureusement, elle n'est pas suffisante pour décrire les données et pour cela nous avons besoin d'un paramètre supplémentaire appelé "la variance". La variance est donc le moment centré d'ordre 2 de la distribution et une mesure de la dispersion de X autour de la moyenne. La remarque principale sur la moyenne est qu'elle ne peut pas exister toujours selon la stabilité de α, $\alpha = 1/\gamma$ où γ c'est l'IVE :

- $0 \leq \alpha \leq 1$: la moyenne et la variance sont les deux infinies, comme un exemple la distribution de la loi de Cauchy si $\alpha = 1$.
- $1 < \alpha < 2$: la moyenne finie et la variance infinie, comme les distributions à queue lourde.
- $\alpha = 2$: la moyenne et la variance sont les deux finies, comme la distribution de Gauss.

Pour un échantillon de taille $n \geq 1$, d'une variable aléatoire (va) X de fonction de répartition (fdr) F et de moyenne μ. L'estimateur naturel de μ est la moyenne empirique $\overline{X}$. Il est consistant, sans biais et sous la condition $\mathbb{E}(X^2) < \infty$ le Théorème Centrale Limite (TCL) garantit sa normalité asymptotique. Dans le cas où cette condition n'est pas vérifiée, Peng [107] a proposé un estimateur asymptotiquement normal en exploitant les outils de la théorie des valeurs extrêmes. Dans les deux cas, les estimateurs sont construits à partir de la totalité des observations. Lorsque la variable X est censurée à droite par une autre variable, de telle façon que les techniques d'estimation basées sur les données complètes deviennent inappropriées. Dans cette situation, Stute [124] a introduit un estimateur asymptotiquement normal de la moyenne d'une distribution de moment d'ordre 2 fini, en se basant sur les résultats de Kaplan et Meier [83].

Cette thèse constitue une sorte de mariage entre deux branches de la statistique : l'analyse de survie et la théorie des valeurs extrêmes. On va proposer une synthèse des différentes définitions et propriétés fondamentales de ces deux domaines de la statistique. Alors, cette thèse est organisée comme suit :

Le premier Chapitre se regroupe en cinq Sections après le menu de l'introduction, est consacré aux notions de base de l'analyse de survie. Dans la Section 1.2, on commençe par quelques rappels sur les concepts de base tels la fdr, les trois fonctions de survie et on discute la relation d'équivalence entre ces trois fonctions. Plus dans la Section 1.3, on parle sur les lois des grands nombres et les propriétés asymptotiques de la somme des va's iid (TCL). Dans la Section 1.4, on présente dans ce qui suit deux cas de données incomplètes : censurées et tronquées. La Section 1.5 présente une synthèse des principaux estimateurs des quantités pertinentes, dont les plus célèbres estimateurs non-paramétriques sont l'estimateur de Kaplan-Meier de fonction de survie et l'estimateur de Nelson-Aalen pour la fonction de hasard cumulée. Enfin, dans la dernière Section 1.6, on introduit l'estimation non-paramétrique de la moyenne sous censure aléatoire est faite par Stute [124].

Le deuxième Chapitre se compose de cinq Sections après le menu de l'introduction et se consacre à la TVE. La Section 2.2 définit les statistiques d'ordre, les extrêmes ainsi que les lois exactes des statistiques d'ordre et les lois asymptotiques des valeurs extrêmes. Ensuite, la Section 2.3 donne le résultat fondamental de la TVE et la distribution $\mathcal{GEV}$ ainsi que les caractéristiques des différents domaines d'attraction du maximum ensuite on introduit la distribution $\mathcal{GPD}$ et le théorème de Balkema et de Haan [5]. La Section 2.4 est consacrée sur la classe de distributions à queue lourde. Puis la Section 2.5 donne les estimateurs classiques de l'indice de queue tels l'estimateur de Hill, de Pickands et des Moments dans le cadre de sans censure. On va s'intéresser à la dernière Section 2.6 au problème de l'estimation de l'indice de queue en présence de données censurées aléatoirement à droite.

Le troisième Chapitre est consacré à l'estimation de la prime de réassurance en excédant des sinistres lorsque les risques sont aléatoirement censurés à droite. L'estimateur est construit et sa normalité asymptotique établie sous des conditions adéquates. Sa performance est évaluée à travers des ensembles de données simulées. Ce Chapitre correspond à l'article Statistical estimate of the proportional hazard premium of loss under random censoring, publié dans la revue Afrika Statistika, 11(1), 883 − 899 (2016).

Le quatrième Chapitre, traite l'estimation de la moyenne sous censure aléatoire. On a constaté qu'il y a des distributions à queues lourdes pour lesquelles les conditions de Stute (Stute, [124]) ne sont pas vérifiées. L'objectif principal de ce Chapitre est de proposer une méthode d'estimation de la moyenne de ce type de distributions en présence de censure aléatoire à droite, sa normalité asymptotique établie et sa performance évaluée sur des données simulées. Les résultats obtenus sont appliqués sur des données réelles médicales. Ce travail correspond à l'article Estimating the mean of a heavy-tailed distribution under random censoring, soumis pour publication.

Enfin, je voudrais mentionner que le traitement des données (calculs numériques et représentations graphiques) est réalisé à l'aide des logiciels d'analyse statistique R.

Chapitre 1

Concepts de Base en Analyse de Survie

Sommaire

1.1 Introduction

On se pose les questions suivantes : qu'est ce qui distingue l'analyse de survie des autres domaines de la statistique ? Pourquoi les données de survie ont besoin d'une théorie statistique spéciale ? La raison est que le modèle de survie est un modèle basé sur des durées de vie ou (durées de survie). Le terme de durée de vie est employé pour désigner le temps qui s'écoule jusqu'à la survenue d'un événement particulier. Cette notion de durée de vie est observée dans plusieurs domaines tels que la médecine, la fiabilité, l'assurance, l'économie, la biologie, l'astronomie et la sociologie ..., de la durée de vie d'une ampoule à la durée de vie d'un malade sous observation thérapeutique qui n'est pas forcément la mort : par exemple, le délai de rémission d'une maladie (temps écoulé entre le début du traitement et la rechute), le délai de guérison (temps qui sépare le diagnostic de la guérison).
La principale source de difficulté dans l'analyse des durées de vie et pour diverses raisons, est la présence de données incomplètes qui nécessitent un traitement statistique particulier. Les références les plus importantes dans ce domaine sont les livres de Cox et Oakes [28], Kalbfleisch et Prentice [82], Lee et Wang [86], Klein et Moeschberger [84], Wienke [134] et Hanagal [75].

1.2 Concepts et Définitions

Soit X une variable aléatoire (va), définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, représentant le temps de survie. La distribution de X peut être caractérisée par trois fonctions équivalentes. En pratique, ces trois fonctions peuvent être utilisées pour illustrer les différents aspects des données. Avant de parler de ces fonctions, on définit la fonction de répartition (fdr ou fd) de X.

Définition 1.1 (*Fonction de répartition*)**.**
La fdr d'une va X est l'application F définie de $\mathbb{R}_+$ dans $[0,1]$ par

$$F(t) := \mathbb{P}(X \leq t), \tag{1.1}$$

F : est aussi appelée fonction de distribution ou fonction de distribution cumulée.

1.2.1 Fonction de Survie

Définition 1.2 (*Fonction de survie*)**.**
La fonction de survie, aussi appelée queue de distribution, qu'on note par $S(t)$ ou $\overline{F}(t)$ est définie sur $\mathbb{R}_+$ par

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

C'est la probabilité qu'un individu vive au-delà d'une date t.

Remarque 1.1.
La fdr F d'une va X est croissante, continue à droite et vérifie

$$\lim_{t\to 0} F(t) = 0 \quad et \quad \lim_{t\to\infty} F(t) = 1,$$

alors que $\overline{F}$ est une fonction décroissante, continue à gauche telle que

$$\lim_{t\to 0} \overline{F}(t) = 1 \quad et \quad \lim_{t\to\infty} \overline{F}(t) = 0.$$

Définition 1.3 (*Fonctions empiriques de répartition et de survie*)**.**
Soit $X_1, ..., X_n$ un échantillon de taille $n \geq 1$ d'une va positive X de fdr F et de fonction de survie $\overline{F}$. Les fonctions empiriques de répartition et de survie, F_n et $\overline{F}_n$ sont respectivement définies par

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \leq t\right\}, \quad \forall\ t \geq 0, \tag{1.3}$$

et

$$\overline{F}_n(t) = 1 - F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i > t\right\}, \quad \forall\ t \geq 0, \tag{1.4}$$

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

Remarque 1.2.

1. *$F_n(t)$: c'est la proportion des n variables qui sont inférieurs ou égales à t.*
2. *$\overline{F}_n(t)$: c'est la proportion d'observations qui dépasse t.*

Pour $1 \leq i \leq n$, les fonctions (1.3) et (1.4) s'écrivent en termes des valeurs des statistiques d'ordre[1] comme suit

$$F_n(t) = \begin{cases} 0 & \text{si } t < X_{1:n}, \\ \dfrac{i}{n} & \text{si } X_{i:n} \leq t < X_{i+1:n}, \\ 1 & \text{si } t \geq X_{n:n}, \end{cases} \quad \text{et } \overline{F}_n(t) = \begin{cases} 1 & \text{si } t < X_{1:n}, \\ 1 - \dfrac{i}{n} & \text{si } X_{i:n} \leq t < X_{i+1:n}, \\ 0 & \text{si } t \geq X_{n:n}. \end{cases}$$

Pour une représentation graphique de ces deux fonctions, voir la Figure 1.1.

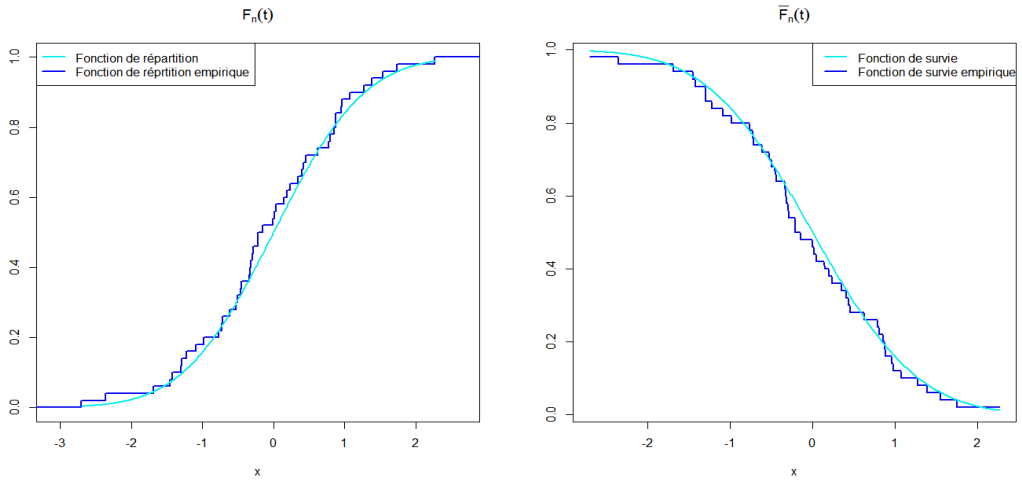

FIG. 1.1. Fonctions empiriques de répartition (gauche) et de survie (droite) d'un échantillon Gaussien standard de taille 50.

1.2.2 Fonction de Densité

Comme toute autre va continue, la durée de survie X a une fonction de densité de probabilité.

Définition 1.4 (*Fonction de densité*)**.**
Si F admet une dérivée par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$, la fonction de densité de probabilité existe, elle est définie pour tout $t \geq 0$, par

$$f(t) = \frac{dF(t)}{dt} := \lim_{dx \to \infty} \frac{\mathbb{P}\left(t < X < t + dx\right)}{dx}.$$

[1] Les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$, sont obtenues en classant ces va's par ordre croissant $X_{1:n} \leq ... \leq X_{n:n}$. Une brève étude sur ces dernières est donnée dans la Section 2.2.

1.2.3 Fonction de Hasard

Appelée selon les domaines d'application : *"taux instantané de défaillance"*, *"taux de risque"* ou encore *"quotient de mortalité"*.

Définition 1.5 (*Fonction de hasard*)**.**
Si X est une variable continue positive représentant une durée. La fonction de hasard, notée par $h(t)$, est définie par

$$h(t) := \lim_{dx \to 0} \frac{\mathbb{P}\left(t < X < t + dx / X > t\right)}{dx}. \tag{1.5}$$

Remarque 1.3.
La fonction (1.5) *, peut également être définie en termes de fdr $F(t)$ et la fonction de densité de probabilité $f(t)$*

$$h(t) := \frac{f(t)}{1 - F(t)}. \tag{1.6}$$

Parfois, il est utile de travailler avec une fonction de hasard cumulée (ou intégrée) qui est donnée par

$$\Lambda(t) := \int_0^t h(x)dx = \int_0^t \frac{dF(x)}{\overline{F}(x)}. \tag{1.7}$$

Il est facile de trouver les relations entre ces différentes notions, par exemple, (1.7) implique que

$$\Lambda(t) = -\log \overline{F}(t). \tag{1.8}$$

Ainsi, quand $t = 0$, $\overline{F}(t) = 1$, $\Lambda(t) = 0$, et quand $t = \infty$, $\overline{F}(t) = 0$, $\Lambda(t) = \infty$. La fonction de hasard cumulée peut être n'importe quelle valeur comprise entre 0 et ∞. On note que, en vertu de (1.8) , on peut écrire

$$\overline{F}(t) = \exp\left\{ -\int_0^t \frac{dF(x)}{\overline{F}(x)} \right\} = \exp\left\{-\Lambda(t)\right\}. \tag{1.9}$$

Cette égalité est la principale formule exponentielle d'analyse de survie. Elle présente une caractérisation de distribution et une fonction de survie par l'intermédiaire de fonction de hasard. Ainsi, compte tenu de l'une des trois fonctions de survie, les deux autres peuvent facilement être dérivées. L'exemple suivant illustre les relations d'équivalence entre les trois fonctions précédentes, pour plus de détails on réfère à, par exemple, (Lee et Wang [86], Exemple 2.2, page 17) ou (Wienke [134], Exemple 2.1, page 17).

Exemple 1.1 (*Distribution de Weibull*)**.**
On suppose que la va X suit une distribution de Weibull avec la fonction de densité de probabilité

$$f(t) = \lambda v t^{v-1} e^{-\lambda t^v}, \qquad t \geq 0,$$

où λ, v sont des paramètres non négatifs. Les fonctions de distribution et de survie F et $\overline{F}$ sont respectivement données par

$$F(t) = 1 - e^{-\lambda t^v} \quad et \quad \overline{F}(t) = e^{-\lambda t^v}.$$

La fonction de hasard, alors peut être obtenue de (1.6)

$$h(t) = \lambda v t^{v-1}.$$

D'après (1.8), on a la fonction de hasard cumulée

$$\Lambda(t) = \lambda t^v.$$

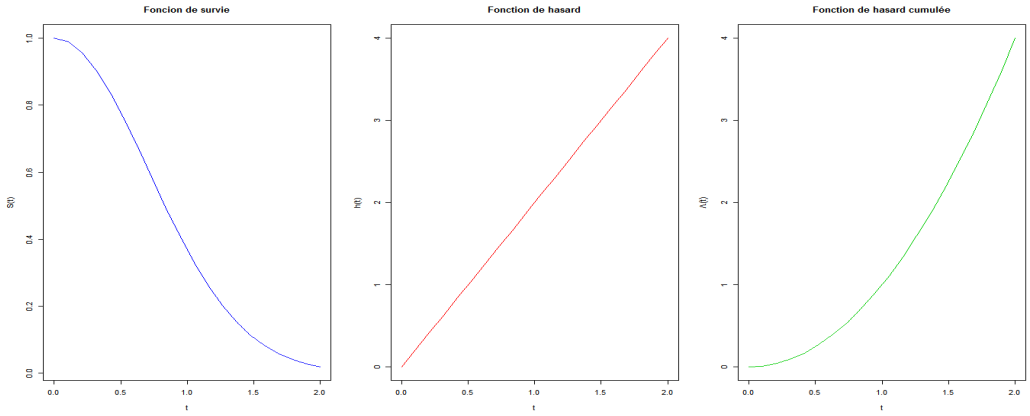

FIG. 1.2. Fonctions de survie, hasard et hasard cumulée de distribution de Weibull avec $\lambda = 1$ et $v = 2$.

1.3 Théorèmes Limites

Définition 1.6 (*Somme et moyenne arithmétique*)**.**
Soit $X_1, ..., X_n$ une suite de va's indépendantes et identiquement distribuées (iid) définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Pour tout entier $n \geq 1$, on définit la somme et la moyenne arithmétique correspondante respectivement par

$$S_n := \sum_{i=1}^{n} X_i \quad et \quad \overline{X}_n := S_n/n, \tag{1.10}$$

$\overline{X}_n$: s'appelle alors la moyenne d'échantillon ou la moyenne empirique.

1.3.1 Lois des Grands Nombres

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

Théorème 1.1 (*Lois des grands nombres*)**.**
Si $X_1, ..., X_n$ un échantillon provenant d'une va X tel que $\mu := \mathbf{E}[X] < \infty$, alors

$$\begin{aligned} &\textit{Loi faible} \quad \overline{X}_n \xrightarrow{\mathbb{P}} \mu \ \textit{quand}\ n \longrightarrow \infty, \\ &\textit{Loi forte} \quad \overline{X}_n \xrightarrow{p.s} \mu \ \textit{quand}\ n \longrightarrow \infty. \end{aligned}$$

La Figure 1.3 ci-dessous illustre la convergence de $\overline{X}_n$ vers la vraie valeur de μ. Pour plus d'information sur les lois des grands nombres, on peut consulter le Chapitre 2 de Embrechts et al. [46].

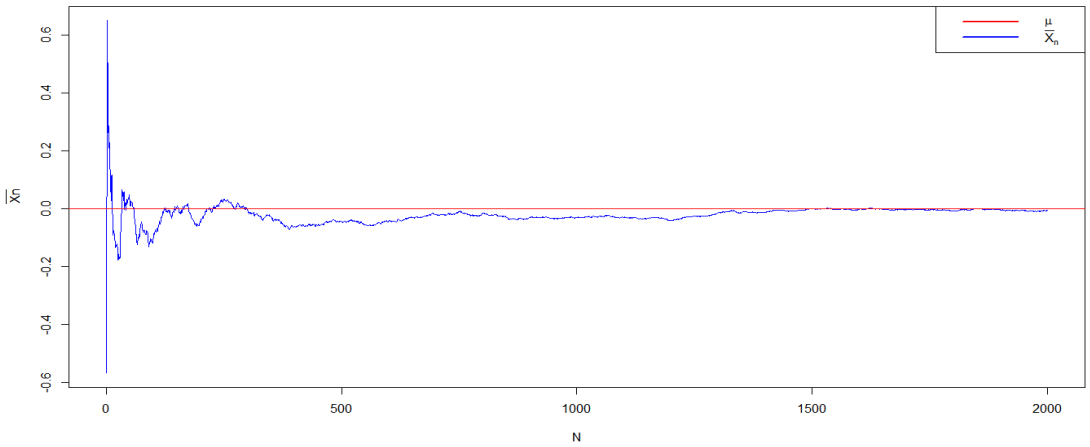

FIG. 1.3. Illustration de la loi des grands nombres : moyenne empirique d'un échantillon Gaussien standard de taille 2000.

Le théorème suivant regroupe des propriétés classiques de la fdr empirique.

Théorème 1.2 (*Propriétés asymptotiques de F_n*)**.**

(i) *Absence de biais : pour tout t, on a $F_n(t) \xrightarrow{p.s} F(t)$, c'est-à-dire que*

$$\mathbf{E}[F_n(t)] = F(t), \quad \textit{quand}\ n \longrightarrow \infty.$$

(ii) *La convergence de F_n vers F est presque sûrement uniforme, c'est-à-dire que*

$$\sup_t |F_n(t) - F(t)| \xrightarrow{p.s} 0, \quad \text{quand } n \longrightarrow \infty. \tag{1.11}$$

(iii) *Normalité asymptotique : pour un t fixé, on a*

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, F(t)(1 - F(t))).$$

La convergence (1.11) est connue sous le nom de théorème de Glivenko-Cantelli. Il est l'un des résultats fondamentaux en statistiques non-paramétriques. La preuve du Théorème 1.1 et le Théorème 1.2 peut être trouvés dans tout manuel standard de la théorie des probabilités comme (Rényi [114], Chapitre 7, page 378) ou (Billingsley [15], Chapitre 4, page 268).

Définition 1.7 (*Fonction de quantile*)**.**
Pour tout $0 < s < 1$, la fonction de quantile de la durée de survie est définie par

$$Q(s) := \inf\{t : F(t) \geq s\} =: F^{-1}(s), \tag{1.12}$$

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

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

Remarque 1.4.

1. *Lorsque la fdr F est strictement croissante et continue alors*

$$Q(s) = F^{-1}(s) = \overline{F}^{-1}(1 - s), \quad 0 < s < 1.$$

2. *Le quantile d'ordre s $(0 < s < 1)$ est défini par $t_s = F^{-1}(s)$.*

Définition 1.8 (*Quantile empirique*)**.**
La fonction de quantile empirique de l'échantillon $X_1, ..., X_n$ est définie par

$$Q_n(s) = F_n^{-1}(s) := \inf\{t : F_n(t) \geq s\}, \quad 0 < s < 1, \tag{1.13}$$

ou

$$Q_n(s) = \overline{F}_n^{-1}(1 - s) := \inf\left\{t : \overline{F}_n(t) \leq 1 - s\right\}, \quad 0 < s < 1.$$

Q_n peut être exprimée comme une fonction simple des statistiques d'ordre concernant l'échantillon $X_1, ..., X_n$. Donc, on a

$$Q_n(s) = X_{i:n} \quad \text{si} \quad \frac{i-1}{n} < s \leq \frac{i}{n}, \; i = 1, ..., n. \tag{1.14}$$

On note que pour $0 < s < 1$, $X_{[ns]+1:n}$ est le quantile empirique d'ordre s, où $[a]$ désigne la partie entière de a.

Remarque 1.5.

1. *Une fonction, notée par U et (parfois) appelée fonction de quantile de queue, elle est définie par*

$$U(t) := F^{-1}(1 - 1/t) = (1/\overline{F})^{-1}(t), \quad t \geq 1. \tag{1.15}$$

2. *La fonction empirique correspondante est*

$$U_n(t) := Q_n(1 - 1/t), \quad t \geq 1.$$

1.3.2 Théorème Central Limite

L'étude de sommes de variables indépendantes et de même loi joue un rôle capital en statistique. Le théorème suivant est établi par Saporta [118] dans le Chapitre 2 est connu sous le nom de Théorème Centrale Limite (TCL) établit la convergence en loi vers la loi normale d'une somme de va iid sous des hypothèses très peu contraignantes.

Théorème 1.3 (*TCL*)**.**
Soit $X_1, ..., X_n$ est une suite de va's iid de moyenne μ et de variance σ^2 finie, alors

$$(S_n - n\mu)/\sigma\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \text{ quand } n \longrightarrow \infty.$$

La preuve du Théorème 1.3 peut être trouvée dans n'importe quel livre standard des statistiques (voir par exemple, Saporta [118], page 66).

1.4 Données Incomplètes

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 ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information. Au lieu d'observer des réalisations iid de durée X, on observe la réalisation de la variable X soumise à diverses perturbations indépendantes ou non de l'événement étudié. Les mécanismes de censure et de troncature peuvent survenir simultanément.

1.4.1 Données Censurées

Le phénomène de censure est lié aux événements perturbateurs qui peuvent se produire dans le laps de temps nécessaire au recueil d'une donnée. Il intervient donc fréquemment lors de mesures qui portent sur les variables modélisant le temps écoulé entre deux événements : durée de vie d'un individu, durée entre le début d'une maladie et la guérison, durée d'un épisode de chômage, ... etc. Ces perturbations empêchent l'observateur d'accéder à la totalité de l'information concernant le

phénomène qu'il étudie et conduit à l'apparition d'observations incomplètes dites censurées. La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Définition 1.9 (*Variable de censure*)**.**
La variable de censure Y est définie par la non-observation de l'événement étudié. Si au lieu d'observer X, on observe Y, et que l'on sait que $X > Y$ (respectivement $X < Y$, $Y_1 < X < Y_2$), on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).

Pour un individu donné j, on va considérer

- son temps de survie X_j,
- son temps de censure Y_j,
- la durée réellement observée Z_j.

1.4.2 Types de Censures

La censure des données se fait selon plusieurs mécanismes tels la censure à droite, la censure à gauche, la censure double (ou mixte), ... On se réfère à [99] où ces types de censure sont présentés avec des exemples.

1. **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. On peut aussi trouver ces genres de phénomènes en hydrologie, en pluviométrie, ... 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.
2. **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 on veut 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. Dans la vie courante il y a plusieurs phénomènes qui présentent à la fois des données censurées à droite et à gauche.

3. **Censure double ou mixte**
   On dit qu'on a une censure double ou mixte si on a des données censurées à droite et des données censurées à gauche dans le même échantillon. Plusieurs modèles non-paramétriques ont été présentés pour l'étude de la double censure. Par exemple, le modèle de Turnbull [126] est le plus utilisé, et plusieurs travaux sont basés sur ce modèle.

Dans la littérature d'autres modèles ont été proposés notamment la censure par intervalle.

4. **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. On a 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 $[a, +\infty[$ et $[0, a]$ respectivement.

Ces quatre 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 :

- **Censure de type I : fixée**
  L'expérimentateur fixe une valeur (une date par exemple non aléatoire de fin d'expérience). Par exemple en épidémiologie on fixe la durée maximale de participation et vaut, pour chaque observation, la différence entre la date de fin d'expérience et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quant à lui, aléatoire.

  Soit Y une valeur fixée. Par exemple en censure à droite, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on observe X_j que lorsqu'elle est inférieure ou égale à une durée fixée Y. On observe donc une variable Z_j telle que $Z_j := \min(X_j, Y)$, $j = 1, ...n$.
  Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles.
- **Censure de type II : attente**

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

Par exemple en épidémiologie on décide d'observer les durées de survie des n patients jusqu'à ce que r $(1 \leq r \leq n)$ d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{j:n}$ et $Z_{j:n}$ les statistiques d'ordre des variables X_j et Z_j. La date de censure est donc $X_{r:n}$ et on observe

$$\begin{cases} Z_{j:n} = X_{j:n} & \text{si } j \leq r, \\ Z_{j:n} = X_{j:n} & \text{si } j \geq r. \end{cases}$$

- **Censure de type III : aléatoire**
  C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient).

  Soit $X_1, ..., X_n$ un échantillon d'une va positive X, on dit qu'il y a censure aléatoire de cet échantillon s'il existe une autre va positive elle aussi Y d'échantillon $Y_1, ..., Y_n$ dans ce cas au lieu d'observer les $X_j's$, on observe un couple de va's (Z_j, δ_j) avec

$$Z_j := \min(X_j, Y_j) \quad \text{et} \quad \delta_j := \mathbb{1}\{X_j \leq Y_j\} \quad \text{pour } j = 1, ...n, \tag{1.16}$$

  où δ_j l'indicateur de censure, qui détermine si X a été censuré ou non :

  - si $\delta_j = 1$, la durée d'intérêt est observée $(Z_j = X_j)$.
  - si $\delta_j = 0$, elle est censurée $(Z_j = Y_j)$. On observe des durées incomplètes.

Dans cette thèse, on s'intéresse uniquement au cas des censures à droite du type aléatoire. Celui-ci correspond à un modèle fréquemment utilisé en pratique, ce qui justifie amplement qu'on y attache à l'intérêt.

1.4.3 Données Tronquées

Les données censurées ne sont pas le type unique de données incomplètes. L'autre cas classique de données incomplètes est celui des données dites tronquées. Le phénomène de troncature est très différent de la censure.

La troncature, quant à elle, élimine de l'étude une partie des X_j. Lors d'une étude pratique sur les durées de vie, il n'est pas rare que la variable d'intérêt X ne soit pas observable quand elle est inférieure à un seuil aléatoire Y, ce qui aura pour conséquence que l'analyse ne pourra porter que sur la loi conditionnelle de X sachant $X > Y$. Il y a trois types de troncature : troncature à gauche, à droite et par intervalle.

- **Troncature à gauche :** Soit Y est une va indépendante de X, on dit qu'il y a troncature à gauche lorsque X (la durée de survie) n'est observable que si $X > Y$. On observe le couple (X, Y), avec $X > Y$.
- **Troncature à droite :** De même, il y a troncature à droite lorsque X n'est observable que si $X < Y$.
- **Troncature par intervalle :** Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.

Lyndell-Bell [90] proposa une estimation non-paramétrique de la fdr de X dans le cadre du modèle de troncature et les propriétés asymptotiques : la loi forte et la normalité asymptotique ont été étudiées par Woodroofe [137].

1.5 Estimation des $\overline{F}(t)$ et $\Lambda(t)$

Les principaux estimateurs jouant un rôle essentiel dans le cadre des données censurées sont :

- L'estimateur de Kaplan-Meier pour la fonction de survie $\overline{F}(t)$. Il est aussi appelé estimateur produit-limite.
- L'estimateur de Nelson-Aalen pour la fonction de hasard cumulée $\Lambda(t)$.

1.5.1 Estimateur de Kaplan-Meier

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Soit $X_1, ..., X_n$ une suite de va's d'intérêt iid positives, de fdr commune F et $Y_1, ..., Y_n$ une suite de va's de censure iid positives, de fdr continue G. On suppose aussi que ces variables sont indépendantes des X_j. Soit $\{(Z_j, \delta_j),\ 1 \leq j \leq n\}$ l'échantillon réellement observé défini par (1.16), dans la suite on supposera que la variable Z a comme fdr H.

Malheureusement, en présence de censure, la fonction de survie empirique (1.4) de la variable X n'est plus valable car elle dépend de va's parmi X, $X_1, ..., X_n$ qui ne sont pas observées. Afin d'estimer la loi de X, il a été donc nécessaire de construire un estimateur de fonction de survie (1.2) en présence de données censurées. En 1958, Kaplan et Meier [83] ont introduit les estimateurs non-paramétiques du maximum de vraisemblance de F et G (voir, par exemple, (1.1) et (1.2) dans Deheuvels et Einmahl [36]) sont définis par

$$\widehat{\overline{F}}_n(t) := \begin{cases} \displaystyle\prod_{Z_{j:n} \leq t} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} & \text{pour } t < Z_{n:n}, \\ 0 & \text{pour } t \geq Z_{n:n}, \end{cases} \tag{1.17}$$

et

$$\widehat{\overline{G}}_n(t) := \begin{cases} \displaystyle\prod_{Z_{j:n}\leq t} \left(\frac{n-j}{n-j+1}\right)^{1-\delta_{[j:n]}} & \text{pour } t < Z_{n:n}, \\ 0 & \text{pour } t \geq Z_{n:n}, \end{cases} \tag{1.18}$$

où $Z_{1:n} \leq ... \leq Z_{n:n}$ sont les statistiques d'ordre associées à $Z_1, ..., Z_n$, et où pour $1 \leq j \leq n$, $\delta_{[j:n]}$ le concomitant de la $j^{ème}$ statistique d'ordre, c'est-à-dire, $\delta_{[j:n]} = \delta_i$ si $Z_{j:n} = Z_i$, $1 \leq i \leq n$.

On présente ici une autre écriture de l'estimateur de Kaplan-Meier sous forme de somme. Cet écriture peut être trouvée dans le livre de Reiss et Thomas [111], page 162

$$\widehat{F}_n(t) = 1 - \widehat{\overline{F}}_n(t) := \sum_{i=2}^{n} W_{i,n} \mathbb{1}\{Z_{i:n} \leq t\}, \tag{1.19}$$

par des raisonnements combinatoires, Stute et Wang [125] obtiennent l'expression suivante des sauts de l'estimateur de Kaplan-Meier

$$W_{i,n} := \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}, \tag{1.20}$$

où $W_{i,n}$ est le saut à la $i^{ème}$ observation $Z_{i:n}$ dans l'échantillon ordonné.

Remarque 1.6.
Les estimateurs (1.17) *et* (1.18) *sont parfois écris de la manièr suivante :*

$$\widehat{\overline{F}}_n(t) = \prod_{Z_{j:n}\leq t} \left(1 - \frac{\delta_{[j:n]}}{n-j+1}\right) = \prod_{j=1}^{n} \left(1 - \frac{\delta_{[j:n]}}{n-j+1}\right)^{\mathbb{1}\{Z_{j:n}\leq t\}}, \text{ pour } t < Z_{n:n},$$

et

$$\widehat{\overline{G}}_n(t) = \prod_{Z_{j:n}\leq t} \left(1 - \frac{1-\delta_{[j:n]}}{n-j+1}\right) = \prod_{j=1}^{n} \left(1 - \frac{1-\delta_{[j:n]}}{n-j+1}\right)^{\mathbb{1}\{Z_{j:n}\leq t\}}, \text{ pour } t < Z_{n:n}.$$

Remarque 1.7.

1. *L'expression* (1.17) *définit une fonction constante par morceaux, continue à droite avec limite à gauche.*
2. *Les points de discontinuité de cette fonction correspondent aux observations non-censurées.*
3. *L'hauteur des sauts de* $\widehat{F}_n(t)$ *est aléatoire.*
4. *En l'absence de censures, on retrouve la fonction de survie empirique* (1.4).

Dans la littérature de l'analyse de survie, un grand nombre des auteurs ont été consacrés à l'étude des propriétés asymptotiques de l'estimateur de Kaplan-Meier. Par exemple. La consistance uniforme a été étudiée par Shorack et Wellner [120], Wang [127], Stute et Wang [125] et plus récemment par Gill [62]. La normalité asymptotique a été étudiée par Breslow et Crowley [21], Gill [60] et Gill [61].

Proposition 1.1 (*Propriétés asymptotiques de* $\widehat{\overline{F}}_n$)**.**

(i) *Absence de biais : pour tout* t*, on a* $\widehat{\overline{F}}_n(t) \xrightarrow{p.s} \overline{F}(t)$*, c'est-à-dire que*

$$\mathbf{E}[\widehat{\overline{F}}_n(t)] = \overline{F}(t), \quad \text{quand } n \longrightarrow \infty.$$

(ii) *Consistance uniforme : soit* x_H[2] $= H^{-1}(1) := \inf\{t : H(t) = 1\} \leq \infty$*. Alors*

$$\sup_{0 \leq t < x_H} \left| \widehat{\overline{F}}_n(t) - \overline{F}(t) \right| \xrightarrow{p.s} 0, \quad \text{quand } n \longrightarrow \infty.$$

(iii) *Normalité asymptotique : pour tout* $t \geq 0$*, on a*

$$\sqrt{n}(\widehat{\overline{F}}_n(t) - \overline{F}(t)) \xrightarrow{\mathcal{D}} \mathbb{X}_t.$$

où $\mathbb{X}_t$ *est un processus Gaussien centré de fonction de covariance*

$$cov\,(\mathbb{X}_s, \mathbb{X}_t) = \overline{F}(s)\overline{F}(t) \int_0^{\min(s,t)} \frac{dH(t)}{\left(\overline{F}(t)\right)^2}.$$

Définition 1.10 (*Fonction des quantiles empiriques sous censure aléatoire*)**.**
La fonction des quantiles empiriques sous censure aléatoire de l'échantillon $Z_1, ..., Z_n$ *est définie par*

$$\widehat{Q}_n(s) = \widehat{F}_n^{-1}(s) := \inf\left\{t : \widehat{F}_n(t) \geq s\right\} = Z_{i:n}, \tag{1.21}$$

où

$$1 - \prod_{j=2}^{i-1}\left(1 - \frac{\delta_{[j:n]}}{n-j+1}\right) < s \leq 1 - \prod_{j=2}^{i}\left(1 - \frac{\delta_{[j:n]}}{n-j+1}\right), \; 1 \leq i \leq n.$$

[2] x_H : la borne supérieure du support de H.

1.5.2 Estimateur de Nelson-Aalen

Kaplan et Meier [83] ont introduit l'estimateur de produit-limite pour la fonction de survie. L'estimateur de la fonction de hasard cumulée est l'estimateur de Nelson-Aalen introduit par Nelson [101] en 1972 et généralisé par Aalen [2] en 1978.

On observe, tout d'abord, que, sous l'hypothèse générale d'indépendance entre X et Y, on peut décomposer $H(t)$ de la manière suivante :

$$H(t) := 1 - (1 - F(t))\,(1 - G(t)) = H^{(0)}(t) + H^{(1)}(t), \tag{1.22}$$

où

$$H^{(0)}(t) := \mathbb{P}(Z \leq t,\ \delta = 0) = \int_0^t \overline{F}(x) dG(x), \tag{1.23}$$

et

$$H^{(1)}(t) := \mathbb{P}(Z \leq t,\ \delta = 1) = \int_0^t \overline{G}(x) dF(x). \tag{1.24}$$

Pour $t \geq 0$, la fonction de hasard cumulée (1.7) peut s'exprimer de la manière suivante :

$$\Lambda(t) = \int_0^t \frac{\overline{G}(x) dF(x)}{\overline{H}(x)} = \int_0^t \frac{dH^{(1)}(x)}{\overline{H}(x)}.$$

Définition 1.11 (*Estimateur de Nelson-Aalen*)**.**
L'estimateur non-paramétrique de Nelson-Aalen Λ_n de Λ basé sur l'échantillon $\{(Z_j, \delta_j)\,,\ 1 \leq j \leq n\}$ défini par

$$\Lambda_n(t) = \int_0^t \frac{dH_n^{(1)}(x)}{\overline{H}_n(x)} := \begin{cases} \displaystyle\sum_{Z_{j:n} \leq t} \frac{\delta_{[j:n]}}{n - j + 1} & si \quad t < Z_{j:n}, \\ 1 & si \quad t \geq Z_{j:n}, \end{cases}$$

où

$$H_n(t) := \frac{1}{n} \sum_{j=1}^n \mathbb{1}\{Z_j \leq t\} \quad et \quad H_n^{(1)}(t) := \frac{1}{n} \sum_{j=1}^n \delta_j \mathbb{1}\{Z_j \leq t\},$$

représentent respectivement la fdr empirique de $H(t)$ et la version empirique de $H^{(1)}(t)$ de l'échantillon $Z_1, ..., Z_n$.

Remarque 1.8.
En remplaçant $\Lambda(t)$ par $\Lambda_n(t)$ dans (1.9) on obtient un nouvel estimateur de la fonction de survie $\overline{F}$:

$$\widehat{\overline{F}}_n^{NA}(t) := \begin{cases} \displaystyle\prod_{Z_{j:n} \leq t} \exp\left\{-\frac{\delta_{[j:n]}}{n - j + 1}\right\} & pour\ t < Z_{n:n}, \\ 0 & pour\ t \geq Z_{n:n}. \end{cases}$$

connu sous le nom de l'estimateur de Breslow [20].

Fleming et Harrington [53] ont montré la relation étroite entre les estimations de Nelson-Aalen et de Kaplan-Meier, ils ont comparé numériquement pour plusieurs tailles d'échantillon et ont souligné que les deux estimateurs, sont asymptotiquement équivalents. Une belle discussion de ce point peut être trouvée dans un article plus récent de Huang et Strawderman [79].

1.6 Estimation de la Moyenne en Présence de Censure

Dans cette Section on s'intéresse à l'estimation de la moyenne d'une distribution sous censure aléatoire. La moyenne d'une va X de fdr F, elle est définie par

$$\mu = \mathbf{E}[X] := \int t dF(t).$$

Une intégration par parties, on peut réécrire la moyenne comme

$$\mu = \int_0^\infty \overline{F}(t)dt. \tag{1.25}$$

Dans le cas les données complètes, l'estimateur non-paramétrique de la moyenne est $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. Il est obtenu par substitution de $\overline{F}_n$ à la place de $\overline{F}$ dans (1.25) . Il est sans biais, consistant et sous la condition $\mathbf{E}(X^2) = \int t^2 dF(t) < \infty$ le TCL garantit sa normalité asymptotique. Lorsque la variable X est censurée l'estimateur cité ci-dessus ne fonctionne pas car il est basé sur la totalité des observations, dans ce cas, Stute [124] a introduit un estimateur qui s'appelle les intégrales de Kaplan-Meier pour des quantités plus générales que la moyenne.

Soit φ est une fonction mesurable $\varphi : \mathbb{R} \to \mathbb{R}$, l'intégrale $S^\varphi = \int \varphi(t)dF(t)$ estimée par l'intégrale de Kaplan-Meier :

$$S_n^\varphi = \int \varphi(t)\, d\widehat{F}_n(t) := \sum_{i=2}^n W_{i,n}\varphi\left(Z_{i:n}\right),$$

où $W_{i,n}$ défini par (1.20) . Pour $\varphi(t) = t$, on obtient $S^\varphi = \mu$.

Définition 1.12 (*Moyenne empirique sous censure aléatoire*)**.**
L'estimation non-paramétrique de la moyenne sous censure aléatoire est défini par

$$S_n^\varphi = \widetilde{\mu}_n := \sum_{i=2}^n W_{i,n} Z_{i:n}.$$

Stute [124] a montré que cet estimateur est asymptotiquement normal sous les deux conditions suivantes :

$$I_1 := \int_0^\infty x^2 \Gamma_0^2(x) dH^{(1)}(x) < \infty, \tag{1.26}$$

et

$$I_2 := \int_0^\infty x \left(\int_0^x \frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2} \right)^{1/2} dF(x) < \infty, \tag{1.27}$$

où $H^{(0)}$ et $H^{(1)}$ sont deux fonctions définies par (1.23) et (1.24), avec

$$\Gamma_0(x) := \exp\left\{ \int_0^x \frac{dH^{(0)}(s)}{\overline{H}(s)} \right\},$$

$$\Gamma_1(x) := \int_0^x \frac{s\Gamma_0(s)}{\overline{H}(s)} dH^{(1)}(s) \quad \text{et} \quad \Gamma_2(x) := \int_x^\infty \frac{\int_s^\infty t\Gamma_0(t)\, dH^{(1)}(t)}{\left[\overline{H}(s)\right]^2} dH^{(0)}(s).$$

Théorème 1.4 (*TCL sous censure aléatoire*)**.**
Sous (1.26) *et* (1.27), *on a*

$$\sqrt{n}\left(\widetilde{\mu}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right), \quad \textit{quand } n \to \infty,$$

où $\sigma^2 := \mathbf{Var}\left[Z_1\Gamma_0(Z_1)\delta_1 + \Gamma_1(Z_1)(1-\delta_1) - \Gamma_2(Z_1)\right]$.

Démonstration. Voir Stute [124], Corollaire 1.2. □

Chapitre 2

Théorie des valeurs extrêmes

Sommaire

2.1 Introduction

La théorie des valeurs extrêmes (TVE) 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 réponse aux intempéries, aux inondations, aux catastrophes naturelles, aux problèmes financiers, ...etc. en prédisant leurs occurrences dans les années à venir. L'étude des valeurs extrêmes revient à l'étude des queues de distributions de fonctions, ou de façon équivalente, à l'analyse de la plus grande (ou plus petite) observation d'un échantillon. En ce sens, on peut considérer la théorie des valeurs extrêmes comme la contrepartie de la théorie statistique classique, qui est principalement basée sur l'étude de la moyenne d'un échantillon plutôt que des observations extrêmes. On peut citer par exemple comme une bonne référence classique sur la théorie et les applications des valeurs extrêmes, les ouvrages de Embrechts et al. [46], Coles [27], Beirlant et al., [9], Reiss et Thomas [111] et Novak [105] qui font le point sur les différentes techniques existantes.

2.2 Statistique d'Ordre

Les statistiques d'ordre jouent un rôle de plus en plus important dans la théorie des valeurs extrêmes, parce qu'ils fournissent des informations sur la distribution de queue (à droite). On les rencontre, en effet, de façon naturelle et depuis longtemps, dans les problèmes de données censurées ou tronquées quand on étudie, par exemple, les durées de survie, mais aussi, plus récemment, dans la recherche

de méthodes robustes. On commence dans cette Section, par donner les définitions et quelques propriétés des statistiques d'ordre, puis on étudie leurs distributions exactes et asymptotiques des statistiques d'ordre. Pour des présentations plus détaillées dans ce domaine, voir, par exemple, les livres de Reiss [110], Balakrishnan et Coen [4], Arnold et al. [3], David et Nagaraja [33] et Castillo et al. [22] qui sont couverts pratiquement toute la matière de statistique d'ordre.

Définition 2.1 (*Statistique d'ordre*)**.**
Soit $X_1, ..., X_n$, n va's iid de distribution commune F et de densité f. On considère les va's $X_{1:n}, ..., X_{n:n}$ sont rangés par ordre croissant soit

$$X_{1:n} \leq ... \leq X_{n:n}. \tag{2.1}$$

Les va's (2.1) sont appelées les statistiques d'ordre de l'échantillon $X_1, ..., X_n$.

Remarque 2.1.
Pour $1 \leq k \leq n$, la variable $X_{k:n}$ est connue sous le nom de la $k^{ème}$ statistiques d'ordre ou statistique d'ordre k.

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

Définition 2.2 (*Statistiques d'ordre extrêmes*)**.**
Les statistiques d'ordre extrêmes sont définies comme termes du maximum et du minimum des n va's $X_1, ..., X_n$. La variable $X_{1:n}$ est la plus petite statistique d'ordre (ou statistique du minimum) et $X_{n:n}$ est la plus grande statistique d'ordre (ou statistique du maximum)

$$X_{n:n} := \max(X_1, ..., X_n) \quad et \quad X_{1:n} := \min(X_1, ..., X_n). \tag{2.2}$$

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

$$\min(X_1, ..., X_n) = -\max(-X_1, ..., -X_n). \tag{2.3}$$

Dans la suite de cette thèse, on se concentrera sur l'étude du maximum.

2.2.1 Distributions Exactes des Statistiques d'Ordre

Proposition 2.1 (*Distributions du maximum et du minimum*)**.**
Soient $X_1, ..., X_n$ n va's iid de fdr F, la distribution exacte du maximum $X_{n:n}$ est simplement donnée par la formule suivante

$$F_{X_{n:n}}(x) = [F(x)]^n, \quad -\infty < x < +\infty. \tag{2.4}$$

La distribution exacte du minimum $X_{1:n}$ est

$$F_{X_{1:n}}(x) = 1 - \left[\overline{F}(x)\right]^n, \quad -\infty < x < +\infty. \tag{2.5}$$

En effet, pour $x \in \mathbb{R}$

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

Et

$$\begin{aligned} F_{X_{1:n}}(x) &= P(X_{1:n} \leq x) = 1 - P(X_{1:n} \leq x) \\ &= 1 - P(X_1 > x, ..., X_n > x) \\ &= 1 - \left[\overline{F}(x)\right]^n . \end{aligned}$$

Ce sont des cas particuliers importants du résultat général de $F_{X_{k:n}}(x)$.

Remarque 2.2.
On remarque que

$$\overline{F}_{X_{1:n}}(x) = 1 - F_{X_{1:n}}(x) = \left[\overline{F}(x)\right]^n ,$$

cela implique que la fonction de survie du minimum est la fonction de survie élevée à la puissance n.

Proposition 2.2 (*Fonction de répartition de la $k^{ème}$ statistique d'ordre*)**.**
Soit $X_{k:n}$ est la $k^{ème}$ statistique d'ordre de n va's iid $X_1, ..., X_n$ avec F la fdr. Alors

$$F_{X_{k:n}}(x) = \sum_{i=k}^{n} \binom{n}{i} [F(x)]^i \left[\overline{F}(x)\right]^{n-i}, \quad -\infty < x < +\infty, \tag{2.6}$$

$F_{X_{k:n}}$: est la fonction de la $k^{ème}$ statistique d'ordre.

Démonstration. Voir Reiss [110], Lemme 1.3.1, page 20. □

Pour trouver les deux densités $f_{X_{n:n}}$ et $f_{X_{1:n}}$, il suffit de dériver les deux expressions (2.4) et (2.5) . Danc

$$f_{X_{1:n}} = n \left[\overline{F}(x)\right]^{n-1} f(x) \quad \text{et} \quad f_{X_{n:n}} = n [F(x)]^{n-1} f(x), \quad -\infty < x < +\infty.$$

Proposition 2.3 (*Densité de la $k^{ème}$ statistique d'ordre*)**.**
La densité de la $k^{ème}$ statistique d'ordre est

$$f_{X_{k,n}}(x) = \frac{n!}{(k-1)!\,(n-k)!} [F(x)]^{k-1} \left[\overline{F}(x)\right]^{n-k} f(x), \quad 1 \leq k \leq n. \tag{2.7}$$

La démonstration de la Proposition 2.3 est détaillée dans un ouvrage déjà cité [110].

Proposition 2.4 (*Densité jointe d'un couple de statistiques d'ordre*).
La densité jointe de $X_{k:n}$ *et* $X_{l:n}$ $(1 \leq l \leq k \leq n)$, *où* $x \leq y$, *est la suivante*

$$f_{X_{k:n},X_{l:n}}(x,y) = \frac{n!}{(k-1)!\,(l-k-1)!\,(n-k)!} .F^{k-1}(x)[F(y)-F(x)]^{l-k-1}[\overline{F}(x)]^{n-l}f(x)f(y). \tag{2.8}$$

autrement $f_{X_{k:n},X_{l:n}}(x) = 0$.

Corollaire 2.1 (*Densité jointe de n statistique d'ordre*).
La densité jointe de n statistique d'ordre $X_{1:n},...,X_{n:n}$ *est*

$$f_{X_{1:n},...,X_{n:n}}(x_1,...,x_n) = \begin{cases} n!f(x_1)f(x_2)...f(x_n) & si\ x_1 < ... < x_n, \\ 0 & sinon. \end{cases} \tag{2.9}$$

Corollaire 2.2 (*Distribution jointe du minimum et maximum*).
La distribution jointe du minimum et maximum est donnée par

$$f_{X_{1:n},X_{n:n}}(x) = n\,(n-1)\,[F(y)-F(x)]^{n-2}f(x)f(y), \quad -\infty < x < y < +\infty. \tag{2.10}$$

Les preuves de Proposition 2.4, Corollaire 2.1 et Corollaire 2.2 sont trouvées dans le livre de David et Nagaraja [33], pour plus de détails.

Proposition 2.5 (*Transformation de quantile*).
Soit $(U_1,...,U_n)$ *un échantillon de va's uniformes sur* $[0,1]$ *et* $U_{1:n} \leq ... \leq U_{n:n}$ *les statistiques d'ordore associées.*

(i) Pour toute fonction de distribution F, on a

$$X_{k:n} \overset{\mathcal{D}}{=} F^{-1}(U_{k:n}),\ i = 1,...,n. \tag{2.11}$$

(ii) Quand F est continue, on a

$$F(X_{k:n}) \overset{\mathcal{D}}{=} U_{k:n},\ i = 1,...,n. \tag{2.12}$$

Démonstration. Voir Reiss [110], Théorème 1.2.5, page 17. □

En pratique, la loi F inconnue, le comportement de $F_{X_{n:n}}$ sera encore plus difficile à étudier. On peut cependant remarquer que

$$\lim_{n\to\infty} F_{X_{n:n}}(x) = \lim_{n\to\infty}[F(x)]^n = \mathbb{1}\{x \geq x_F\} = \begin{cases} 1 & \text{si } x \geq x_F, \\ 0 & \text{si } x < x_F, \end{cases} \tag{2.13}$$

où

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

est le point terminal à droite (right endpoint) de F, avec la convention $\sup\{\varnothing\} = -\infty$ $\varnothing$. Le point terminal x_F représente la borne supérieure du support de la loi. Le résultat (2.13) nous indique que la distribution du maximum $X_{n:n}$ est une loi dégénérée. Ce résultat fournit très peu d'informations sur le comportement de $X_{n:n}$. On aimerait obtenir une loi non dégénérée pour le maximum.

L'idée est de procéder à une transformation. La plus connue en statistique est la normalisation illustrée à travers l'exemple du théorème central limite qui, après normalisation, donne la loi asymptotique (non dégénérée) de la moyenne de n va's.

Dans la Subsection 2.2.2 on va commencer par énoncer le résultat fondamental de la théorie des valeurs extrêmes connu sous le nom de théorème de Fisher-Tippett. Il établit la loi asymptotique du maximum de l'échantillon $X_{n:n}$ convenablement renormalisé.

2.2.2 Distributions des Valeurs Extrêmes

Historiquement, l'étude de la loi de probabilité du maximum d'un échantillon de n variables a été la première approche pour décrire les événements extrêmes. Les travaux de Fisher et Tippett [52] en 1928 ont, les premiers, déduit de manière heuristique les lois limites possibles pour le maximum d'une suite de va's iid, avant que Gnedenko [63] en 1943 n'obtienne pas rigoureusement la convergence, dont la preuve fut simplifiée par de Haan [68] en 1976.

Le théorème ci-dessous est fondamental en théorie des valeurs extrêmes car il établit la loi asymptotique du maximum $X_{n:n}$ convenablement normalisé d'un échantillon.

Théorème 2.1 (*Fisher et Tippett*)**.**
Soit $(X_i)_{i\geq 1}$ une suite de va's iid. n va's iid de fdr F. S'il existe deux suites normalisantes réelles $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ et une loi non dégénérée de distribution $\mathcal{H}$ tels que

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}(x), \ \forall x \in \mathbb{R}, \tag{2.14}$$

$\mathcal{H}$ est la distribution des valeurs extrêmes (EVD, Extrême Values Distribution). Alors à une translation et un changement d'échelle près, la fdr de la limite est du type des trois classes suivantes :

$$\begin{aligned}
&\textit{Loi de Gumbel} \quad \Lambda(x) = \exp\left(-\exp(-x)\right), \ x \in \mathbb{R} && \textit{et } \alpha = 0,\\
&\textit{Loi de Frechet} \quad \Phi_\alpha(x) = \begin{cases} 0, & x < 0 \\ \exp\left(-x^{-\alpha}\right), & x \geq 0 \end{cases} && \textit{et } \alpha > 0,\\
&\textit{Loi de Weibull} \quad \Psi_\alpha(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right), & x \leq 0 \\ 1, & x > 0 \end{cases} && \textit{et } \alpha > 0.
\end{aligned}$$

Ces trois distributions Λ, Φ_α, Ψ_α sont appelées "les distributions des valeurs extrêmes Standard" et les va's correspondantes sont "les variables aléatoires extrémales". On peut trouver une démonstration moderne de ce théorème dans la Section 0.3, page 9, de Resnick [112] ou dans la Section 3.2 de Embrechts et al. [46], page 122.

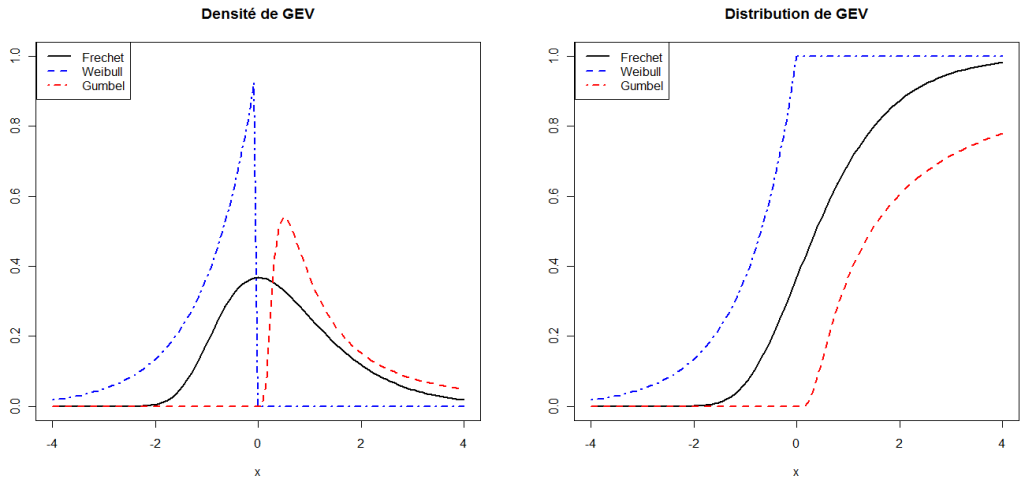

FIG. 2.1. Densités et Distributions de Lois des Valeurs Extrêmes

Le Théorème 2.1 est la base de la théorie des valeurs extrêmes. C'est l'équivalent du Théorème Centrale Limite en ce qui concerne la loi limite des maximas, où la suite $(a_n)_{n\geq 1}$ joue le rôle d'un paramètre d'échelle ou de dispersion et la suite $(b_n)_{n\geq 1}$ joue le rôle d'un paramètre de position ou de centrage. De plus, ces suites ne sont pas uniques. Pour plus d'exemples de suites de normalisation de chaque loi se référer à Embrechts et al. [46] page 145.

2.3 Distributions $\mathcal{GEV}$ et $\mathcal{GPD}$

2.3.1 Distribution $\mathcal{GEV}$

Il est difficile de travailler avec trois familles à la fois, Jenkinson [80] en 1955 montre que ces trois familles peuvent être regroupées sous une forme unique dite famille des lois des valeurs extrêmes généralisées ($\mathcal{GEV}$, Generalized Extreme Value distribution).

Définition 2.3 (Distribution $\mathcal{GEV}$)**.**
La fdr de la famille $\mathcal{H}_\gamma$ des valeurs extrêmes généralisées $\mathcal{GEV}$, est pour $\gamma \in \mathbb{R}$ et

$1+\gamma x>0$

$$\mathcal{H}_\gamma(x) := \begin{cases} \exp\left\{-(1+\gamma x)^{-1/\gamma}\right\} & si\ \gamma \neq 0, \\ \exp\left\{-\exp(-x)\right\} & si\ \gamma = 0. \end{cases} \tag{2.15}$$

Le paramètre γ qui apparaît dans la formule (2.15) est appelé "indice de queue, indice des valeurs extrêmes (IVE)" ou en anglais "extreme values index, EVI". Pour $\gamma = 0$, il faut lire $\mathcal{H}_0(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$ qui s'obtient dans la formule précédente en faisant tendre γ vers 0. Les lois de valeurs extrêmes généralisées correspondent à une translation et un changement d'échelle près aux lois de valeurs extrêmes. On a, où $\gamma := 1/\alpha$, les correspondances suivantes

$$\begin{array}{lcll} \Lambda & = & H_0(x), & x \in \mathbb{R}. \\ \Phi_{1/\gamma} & = & H_\gamma((x-1)/\gamma), & x > 0. \\ \Psi_{1/\gamma} & = & H_{-\gamma}((x+1)/\gamma), & x < 0. \end{array}$$

Pour les variables non centrées et non réduites, on peut écrire $\mathcal{H}_\gamma(x)$ sous une forme plus générale, notée par $\mathcal{H}_{\mu,\sigma,\gamma}$, dans laquelle on fait apparaître un paramètre de localisation $\mu \in \mathbb{R}$ et un paramètre d'échelle $\sigma > 0$. Pour $(1+\gamma(\frac{x-\mu}{\sigma}) > 0)$ la distribution $\mathcal{H}_{\mu,\sigma,\gamma}(x)$ s'écrit comme suit

$$\mathcal{H}_{\mu,\sigma,\gamma}(x) := \begin{cases} \exp\left\{-\left[1+\gamma(\frac{x-\mu}{\sigma})\right]^{-1/\gamma}\right\} & \text{si } \gamma \neq 0, \\ \exp\left\{\exp\left[-(\frac{x-\mu}{\sigma})\right]\right\} & \text{si } \gamma = 0. \end{cases} \tag{2.16}$$

On peut facilement montrer que la fonction de densité correspondante à $\mathcal{H}_{\mu,\sigma,\gamma}$, pour $1+\gamma(\frac{x-\mu}{\sigma}) > 0$, est

$$h_{\mu,\sigma,\gamma}(x) := \begin{cases} \frac{1}{\sigma}\left[1+\gamma(\frac{x-\mu}{\sigma})\right]^{-(\frac{1+\gamma}{\gamma})} H_{\mu,\sigma,\gamma}(x) & \text{si } \gamma \neq 0, \\ \frac{1}{\sigma}\exp\left\{-(\frac{x-\mu}{\sigma}) - \exp\left[-(\frac{x-\mu}{\sigma})\right]\right\} & \text{si } \gamma = 0. \end{cases} \tag{2.17}$$

Remarque 2.3.
Le quantile $Q(p)$ de la distribution $H_{\mu,\sigma,\gamma}$ est donné par la formule suivante

$$Q(p) = H^{-1}_{\mu,\sigma,\gamma} := \begin{cases} \mu - \sigma\gamma^{-1}\left[1-(-\log p)^{-\gamma}\right] & si\ \gamma \neq 0, \\ \mu - \sigma\log(-\log p) & si\ \gamma = 0. \end{cases}$$

Ce quantile est donc fortement influencé par les deux paramètres σ et γ. Intuitivement, on comprend que plus γ est grand, plus le quantile est élevé.

2.3.2 Domaines d'Atraction

Définition 2.4 (*Domaine d'attraction*).
On dit qu'une distribution F appartient au domaine d'attraction du maximum de la distribution $\mathcal{H}_\gamma$, et on note $F \in \mathcal{D}(\mathcal{H}_\gamma)$, s'il existe deux suites normalisantes $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ tels que la condition (2.14) soit vérifiée.

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

- Si $\gamma > 0$, on dit que $F \in \mathcal{D}(\Phi_\alpha)$, et F a un point terminal à droite infinie $(x_F = +\infty)$. Ce domaine d'attraction est celui des distributions à queues lourdes, c'est-à-dire qui ont une fonction de survie à *décroissance polynomiale*.
- Si $\gamma < 0$, on dit que $F \in \mathcal{D}(\Psi_\alpha)$, et F a un point terminal à droite finie $(x_F < +\infty)$. Ce domaine d'attraction est celui des fonctions de survie dont le support est borné supérieurement.
- Si $\gamma = 0$, on dit que $F \in \mathcal{D}(\Lambda)$, le point terminal x_F peut alors être fini ou non. Ce domaine d'attraction est celui des distributions à queues légères, c'est-à-dire qui ont une fonction de survie à décroissance exponentielle.

Les Tableaux 2.1, 2.2 et 2.3 donnent différents exemples de distributions standard dans ces trois domaines d'attraction (Embrechts et al. [46], Tableaux 3.4.2−3.4.4).

Distributions	$\overline{F}(x)$	γ
Pareto(α) $\alpha > 0$	$x^{-\alpha}$, $x > 1$	$\frac{1}{\alpha}$
Burr(β, τ, λ) , $\beta > 0$, $\tau > 0$, $\lambda > 0$	$\left(\frac{\beta}{\beta + x^\tau}\right)^\lambda$	$\frac{1}{\lambda\tau}$
Fréchet$\left(\frac{1}{\alpha}\right)$, $\alpha > 0$	$1 - \exp\left(-x^{-\alpha}\right)$	$\frac{1}{\alpha}$
Loggamma(m, λ) , $m > 0$, $\lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty (\log u)^{m-1} u^{-\lambda-1} du$	$\frac{1}{\lambda}$
Loglogistic(β, α) , $\beta > 0$, $\alpha > 1$	$\frac{1}{1 + \beta x^\alpha}$	$\frac{1}{\alpha}$

TAB. 2.1. Quelques distributions associées à un indice positif

Distributions	$\overline{F}(x)$	γ
Uniform$(0, 1)$	$1 - x$	-1
ReverseBurr$(\beta, \tau, \lambda, x_\tau)$, β, τ, $\lambda > 0$	$\left(\frac{\beta}{\beta + (x_F + x)^{-\tau}}\right)^\lambda$	$-\frac{1}{\lambda\tau}$

TAB. 2.2. Quelques distributions associées à un indice négatif

Distributions	$\overline{F}(x)$	γ
$Gamma(m,\lambda)$, $m \in \mathbb{N}$, $\lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty u^{m-1} \exp(-\lambda u)\, du$	0
$Gumbel(\mu,\beta)$, $\mu \in \mathbb{R}$, $\beta > 0$	$\exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$	0
logistic	$\frac{2}{1+\exp(x)}$	0
$Lognormale(\mu,\sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$	$\frac{1}{\sqrt{2\pi}} \int_1^\infty \frac{1}{u} \exp\left(-\frac{1}{2\sigma^2}(\log u - \mu)^2\right) du$	0
$Weibull(\lambda,\tau)$, $\lambda > 0$, $\tau > 0$	$\exp(-\lambda x^\tau)$	0

TAB. 2.3. Quelques distributions associées à un indice nul

Caractérisation des domaines d'attraction

On indique ici les critères les plus utilisés c'est-à-dire, les conditions sur la fdr F pour qu'elle appartienne à l'un des trois domaines d'attraction qui sont définis précédemment.

Théorème 2.2 (*Caractérisation du* $\mathcal{D}(\Phi_\alpha)$)**.**
La fdr F appartient au domaine d'attraction de la loi de Fréchet de paramètre $\alpha > 0$ si et seulement si

$$\overline{F}(x) = x^{-\alpha}\ell(x),$$

où la fonction ℓ est à variation lente[1]*. En particulier $x_F = +\infty$. De plus si $F \in \mathcal{D}(\Phi_\alpha)$, alors avec $a_n = U(n) = F^{-1}(1-1/n)$ et $b_n = 0$, la suite $(a_n^{-1}X_{n:n})_{n\geq 1}$ converge en loi vers une va de fdr Φ_α quand $n \to \infty$.*

Démonstration. Voir Embrechts et al. [46], Théorème 3.3.7, page 131. □

Théorème 2.3 (*Caractérisation du* $\mathcal{D}(\Psi_\alpha)$)**.**
La fdr F appartient au domaine d'attraction de la loi de Weibull de paramètre $\alpha > 0$ si et seulement si $x_F < +\infty$ et

$$\overline{F}(x) = \left(x_F - \frac{1}{x}\right) = x^{-\alpha}\ell(x),$$

où la fonction ℓ est à variation lente. De plus si $F \in \mathcal{D}(\Psi_\alpha)$, alors avec $a_n = x_F - U(n) = x_F - F^{-1}(1-1/n)$ et $b_n = x_F$, la suite $(a_n^{-1}(X_{n:n} - x_F))_{n\geq 1}$ converge en loi vers une va de fdr Ψ_α quand $n \to \infty$.

[1] Une brève étude sur les fonctions à variations régulières et à variations lentes où on va donner dans la Subsection 2.4.3.

La démonstration du Théorème 2.3 est similaire à celle du théorème précédent, voir Embrechts et al. [46] Théorème 3.3.12 pour la réciproque.

Les résultats concernant le domaine d'attraction de la loi de Gumbel sont plus délicats, puisque, il n'y a pas de représentation simple pour les lois appartenant au domaine d'attraction de Gumbel. Pour une présentation exhaustive, on renvoie à Beirlant et al. [9] Chapitre 2, ou Embrechts et al. [46], Chapitre 3.3. On rappelle tout d'abord la définition d'une fonction de Von Mises [139].

Définition 2.5 (*Fonction de von Mises*)**.**
La fdr F est dite fonction de von Mises avec la fonction auxiliaire a, s'il existe une certaine $z < x_F$ tel que

$$\overline{F}(x) = c \exp\{-\int_z^x \frac{dt}{a(t)}\}, \quad z < x < x_F \leq \infty,$$

où $c > 0$ et a est une fonction positive absolument continue (par rapport la mesure de Lebesgue) avec la densité a' vérifiant $\lim_{x \to x_F} a'(x) = 0$.

Voici deux exemples de fonctions de von Mises sont disponibles dans (Embrechts et al. [46], Exemples 3.3.19 – 3.3.23).

Exemple 2.1 (*Distribution Exponentielle*)**.**

$$\overline{F}(x) = e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0.$$

F est une fonction von Mises avec la fonction auxiliaire $a(x) = 1/\lambda$.

Exemple 2.2 (*Distribution Weibull*)**.**

$$\overline{F}(t) = e^{-\lambda x^v}, \quad x \geq 0, \quad \lambda,\ v > 0.$$

F est une fonction von Mises avec la fonction auxiliaire $a(x) = \lambda^{-1} v^{-1} x^{1-v}$, $x > 0$.

Le résultat ci-dessous est démontré notamment dans (Resnick [112], Proposition 1.4 et Corollaire 1.7).

Théorème 2.4 (*Caractérisation du $\mathcal{D}(\Lambda)$*)**.**
La fdr F appartient au domaine d'attraction de la loi de Gumbel si et seulement si

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

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

$$a_n = x_F - F^{-1}\left(1 - \frac{1}{n}\right) \quad et \quad b_n = \frac{1}{\overline{F}(a)} \int_{a_n}^{x_F} \overline{F}(y) dy.$$

2.3.3 Distribution $\mathcal{GPD}$

L'approche basée sur la distribution $\mathcal{GEV}$ peut être réductrice du fait que l'utilisation d'un seul maxima conduit à une perte d'information continue dans les autres grandes valeurs de l'échantillon. Pour pallier ce problème, Pickands [109] a introduit la méthode $\mathcal{POT}$ (Peaks-over-Threshold) encore appelée méthode des excès au-delà d'un certain seuil réel u suffisamment grand, inférieur au point terminal ($u < x_F$). Cette méthode consiste à utiliser les observations qui dépassent un certain seuil, plus particulièrement les différences entre ces observations et le seuil, appelées "excès". Il est clair que cette méthode nécessite la détermination d'un seuil ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations. On note le seuil par u.

Plus précisément, soit un échantillon de va's iid $X_1, ..., X_n$. Soit u un seuil fixé (non aléatoire) tel que $u < x_F$. On note par N_u le nombre d'exceedances $X_{i_1}, ..., X_{i_{N_u}}$ qui dépassent le seuil u. On appelle excès au-delà du seuil u les $Y_j := X_{i_j} - u$ pour $j = 1, ..., N_u$, voir Figure 2.2 ci-dessous :

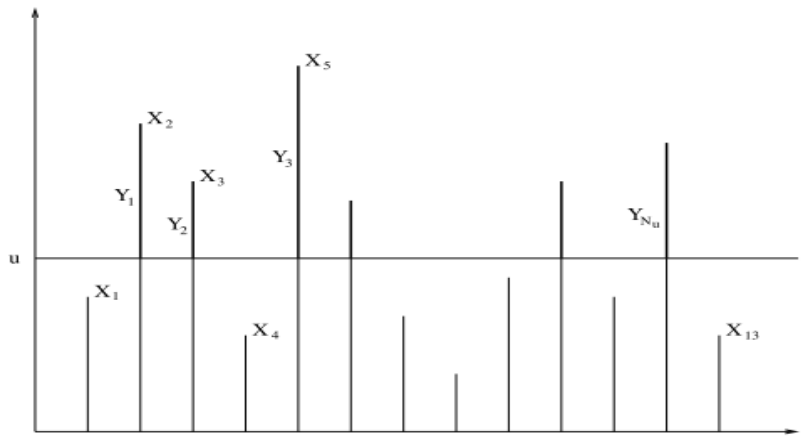

FIG. 2.2. Les données $X_1, ..., X_{13}$ et les excès correspondants $Y_1, ..., Y_{N_u}$ au-dessus du seuil u.

Définition 2.6 (*Fonction de répartition et moyenne des excès*)**.**
Soit X une va de fdr F et de point terminal x_F, Pour tout $u < x_F$, la fonction

$$F_u(y) := P\{X - u \leq y \,|X > u\} = 1 - \frac{\overline{F}(u+y)}{\overline{F}(u)}, \quad 0 < y < x_F - u, \tag{2.18}$$

est appelée fdr des excès au-dessus du seuil u.

La fonction de moyenne des excès correspondante, notée par $e(u)$, est définie par

$$e(u) := \mathbf{E}\left(X - u \,|X > u\right), \quad u < x_F,$$

qui s'exprime également sous la forme

$$e(u) = \frac{1}{\overline{F}(u)} \int_u^{x_F} \overline{F}(t)dt, \quad u < x_F.$$

Remarque 2.4.

1. *Si $x = y + u$ pour $X > u$, on a la représentation suivante :*

$$F(x) = [1 - F(u)]\, F_u(y) + F(u). \tag{2.19}$$

2. *Pour une va qui suit une $G_{\sigma+u\gamma,\gamma}$ de paramètre $\gamma < 1$, la moyenne des excès, $e(u)$, est linéaire avec $u < x_F$*

$$e(u) = \frac{\sigma + u\gamma}{1 - \gamma}, \quad \sigma + u\gamma > 0. \tag{2.20}$$

Une fois le seuil optimal choisi, on construit une nouvelle série d'observations au dessus de ce seuil et la distribution de ces données suit une distribution généralisée de Pareto (*generalized Pareto distribution* ($\mathcal{GPD}$)).

Définition 2.7 (*Distribution de Pareto généralisée*)**.**
La fdr de Pareto généralisée standard ($\mathcal{GPD}$), notée par G_γ, est définie pour $\gamma \in \mathbb{R}$ comme suit

$$G_\gamma(y) := \begin{cases} 1 - (1 + \gamma y)^{-1/\gamma} & si\ \gamma \neq 0, \\ 1 - \exp(-y) & si\ \gamma = 0, \end{cases} \tag{2.21}$$

avec le support

$$\begin{array}{ll} y \geq 0 & si\ \gamma \geq 0, \\ 0 \leq y \leq -1/\gamma & si\ \gamma < 0. \end{array}$$

Pour les propriétés et ses preuves de cette fonction ($G_\gamma(y)$), on peut se référer par exemple à l'ouvrage de Embrechts et al. [46].

Une forme générale de $\mathcal{GPD}$, notée par $G_{\gamma,\mu,\sigma}(y) = G_\gamma(\frac{y-\mu}{\sigma})$, est obtenue en remplaçant l'argument y par $(y - \mu)/\sigma$ dans (2.21) avec un support doit être, qui ajusté en conséquence, où $\mu \in \mathbb{R}$ et $\sigma > 0$ sont respectivement les paramètres de localisation et d'échelle.

On note que la $\mathcal{GPD}$ standard est correspond au cas où $\mu = 0$ et $\sigma = 1$. Lorsque le paramètre de localisation est nul ($\mu = 0$) et le paramètre d'échelle est arbitraire

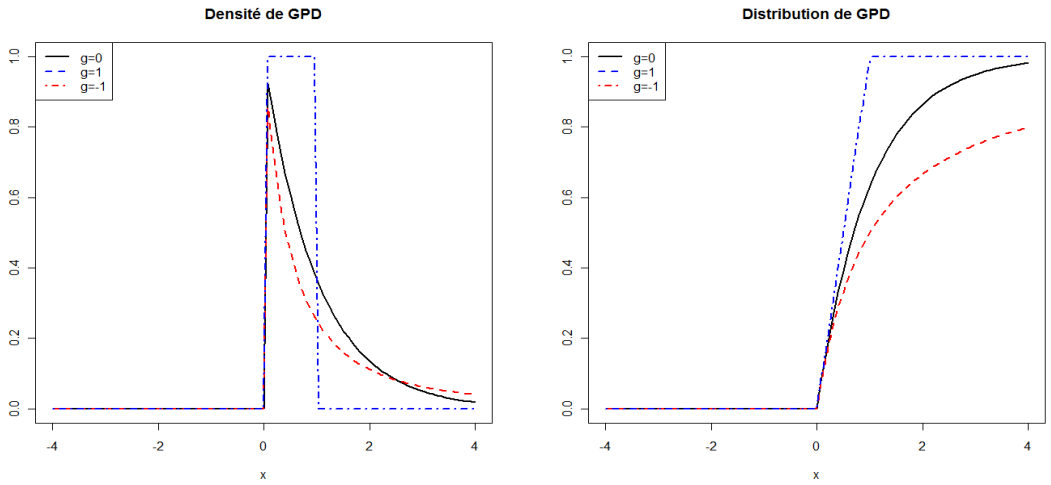

FIG. 2.3. Densités et distributions de lois de Pareto généralisée avec différentes valeurs de γ.

($\sigma > 0$), cette distribution joue un rôle important, dans l'analyse statistique des événements extrêmes, en fournissant une approximation appropriée pour l'excès au-delà d'un grand seuil. Cette famille spéciale, dénotée par $G_{\gamma,\sigma}(y)$, est définie comme suit

$$G_{\gamma,\sigma}(y) := \begin{cases} 1 - \left(1 + \gamma \dfrac{y}{\sigma}\right)^{-1/\gamma} & \text{si } \gamma \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right) & \text{si } \gamma = 0, \end{cases} \tag{2.22}$$

où

$$\begin{array}{ll} y \geq 0 & \text{si } \gamma \geq 0, \\ 0 \leq y \leq -\sigma/\gamma & \text{si } \gamma < 0. \end{array}$$

Remarque 2.5.

1. *La densité de la distribution $\mathcal{GPD}$ $(G_{\gamma,\sigma})$ s'écrit comme suit*

$$g_{\gamma,\sigma}(y) := \begin{cases} \sigma^{-1}\left(1 + \gamma \dfrac{y}{\sigma}\right)^{-\frac{1}{\gamma}-1} & si\ \gamma \neq 0, \\ \sigma^{-1} \exp\left(-\dfrac{y}{\sigma}\right) & si\ \gamma = 0. \end{cases} \tag{2.23}$$

2. *Le quantile $Q(s)$ de la distribution $G_{\gamma,\sigma}$, qui est également la VaR au niveau*

de confiance élevé s, est donné par

$$Q(s) = VaR(s) := u + \frac{\sigma}{\gamma}\left\{\left(\frac{n}{N_u}s\right)^{-\gamma} - 1\right\} \tag{2.24}$$

3. *Il y'a un rapport simple entre la $\mathcal{GPD}$ standard $G_\gamma(x)$ et la $\mathcal{GEV}$ standard $\mathcal{H}_\gamma(x)$ tels que*

$$G_\gamma(x) = 1 + \log \mathcal{H}_\gamma(x), \quad si\ \log \mathcal{H}_\gamma(x) > -1.$$

Balkema et de Haan [5] et Pickands [109] ont proposé le théorème ci-après, qui va être le résultat théorique central de la TVE. Ce théorème précise la distribution conditionnelle des excès lorsque le seuil déterministe tend vers le point terminal x_F.

Théorème 2.5 (*Balkema et de Haan, Pickands*)**.**
Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction $\sigma(u)$ strictement positive et un réel γ tels que

$$\lim_{u\uparrow x_F} \sup_{0<y<x_F-u} \left|F_u(y) - G_{\gamma,\sigma(u)}(y)\right| = 0. \tag{2.25}$$

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

Le Théorème 2.5 établit l'équivalence en loi du maximum (convenablement normalisé) d'un échantillon vers une loi des valeurs extrêmes $\mathcal{H}_\gamma$ et la convergence en loi des excès au-delà d'un seuil vers une loi de Pareto généralisée $G_{\gamma,\sigma}$, lorsque le seuil tend vers la limite supérieure du support de F.

La preuve du Théorème 2.5 doit être trouvé dans Embrechts et al. [46].

2.4 Classe de Distributions à Queue Lourde

Dans cette Section, on présente la notion de distribution à queue lourde et les différentes classes de ce type de distributions. Les distributions à queues lourdes sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines : comme nous l'avons déjà mentionné, finances, hydrologie, télécommunication, géologie ... et plus récemment en climatologie.

Il y a plusieurs définitions ont été associées à ces distributions. La caractérisation la plus simple est celle basée sur la comparaison avec la loi normale (El-Adlouni et al., [48]).

Définition 2.8. *On dit qu'une distribution de X a la queue lourde si*

$$C_k = \mathbf{E}\left[\frac{(X-\mu)^4}{\sigma}\right] = \frac{\mu_4}{\mu_2^2} > 3. \tag{2.26}$$

avec μ_i les moments centrés d'ordre i.

Ce qui est équivalent à dire qu'une distribution a une queue lourde si et seulement si son coefficient d'aplatissement C_k est supérieur à celui de la loi normale. La différence entre la loi normale et une loi avec une queue plus lourde a été illustrée par El-Adlouni et al. [49] comme dans la **Figure 2.4.** Dans cette Figure on présente les fonctions de densité de probabilité de la loi normale et d'une distribution à queue plus lourde (pour plus de détails sur la loi Halphen type B^{-1} (HIB), on peut référer à l'article de El-Adlouni et al. [49], Appendix B.4). On remarque que (voir agrandissement 1−b) la fdr de la loi normale est presque nulle au niveau des extrêmes (queue droite), alors qu'elle ne l'est pas pour la loi HIB.

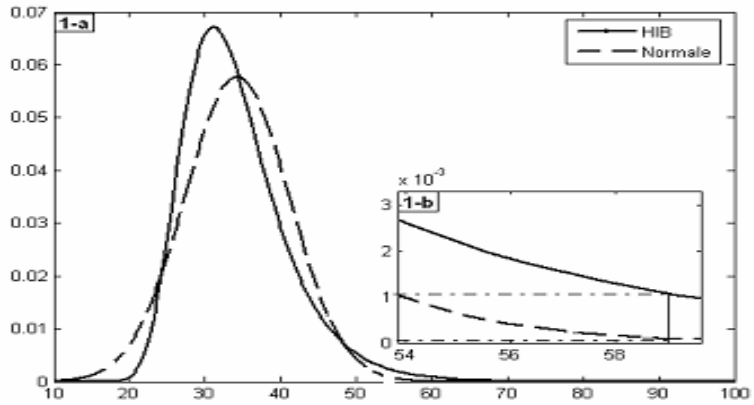

FIG. 2.4. Illustration de la différence entre la loi normale et une loi à queue lourde (HIB)

La caractérisation donnée par l'Equation 2.26 est très générale et ne peut être appliquée que si le moment d'ordre 4 existe. Par conséquent aucune discrimination, pour les distributions ayant un moment d'ordre 4 infini, ne peut être faite si on ne considère que ce critère. Malheureusement, il n'y a pas de critère pour classer toutes les distributions selon la queue droite.

Cependant, un classement de queue peut être obtenue pour des catégories particulières de distributions. Dans la suite, une discussion de cinq classes, donnée dans Werner et Upper [131]. Ces classes sont emboîtées ($A \subset B \subset C \subset D \subset E$) est illustrées à la Figure 2.5, pour plus de détails sur ces interrelations, on se réfère à Embrechts et Omey [47], Klüppelberg [85], et El-Adlouni et al. [49].

- E : Les distributions avec des moments exponentiels inexistants.
- D : Les distributions Subexponentielles.
- C : Les distributions à variations régulières.
- B : Les distributions à queue de type-Pareto.
- A : Les distributions α-Stables avec $\alpha < 2$.

Les deux classes B et C sont très importants compte tenu de leur lien à la théorie des valeurs extrêmes.

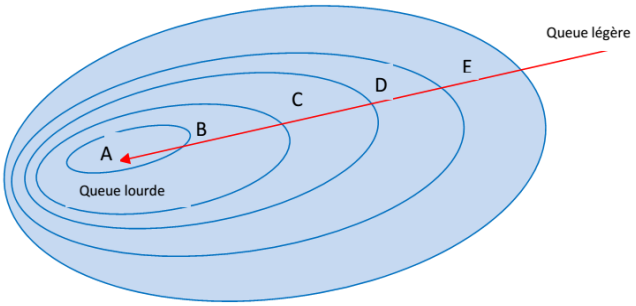

FIG. 2.5. Différentes classes de distributions à queue lourde (El-Adlouni et al., [48]).

2.4.1 Distributions avec des Moments Exponentiels Inexistants

La classe E plus large englobe toutes les distributions qui sont caractérisées par la définition suivante (Foss et al., [54])

Définition 2.9 (*Distribution à queue lourde*)**.**
Une distribution F est dite à queue lourde ou à queue épaisse si sa fdr vérifie

$$\mathbf{E}\left(e^{\lambda x}\right) = \int_{\mathbb{R}} e^{\lambda x} dF(x) = \infty, \quad \textit{pour tout } \lambda > 0. \tag{2.27}$$

On note que la distribution normale n'appartient pas à cette classe à cause de la probabilité au dépassement $\overline{F}$, pour les extrêmes de cette classe, décroît moins rapidement que celle de la loi normale. Dans ce sens, toutes les distributions de la classe E sont à queue plus lourds par rapport à la distribution normale.

Définition 2.10 (*Distribution à queue légère*)**.**
Une distribution F est dite à queue légère ou à queue fine si et seulement si

$$\int_{\mathbb{R}} e^{\lambda x} dF(x) < \infty, \quad \textit{pour tout } \lambda > 0. \tag{2.28}$$

c'est-à-dire si et seulement si elle ne parvient pas à être à queue lourde.

Théorème 2.6.
Pour toute distribution F les assertions suivantes sont équivalentes :

(i) F est une distribution à queue lourde.

(ii) La fonction $\overline{F}$ est à queue lourde.

Démonstration. Voir Foss et al. [54], Théorème 2.6, page 8. □

2.4.2 Distributions Subexponentielles

La classe D des distributions subexponentielles est caractérisée par la définition suivante (Embrechts et al., [46]).

Définition 2.11 (*Distributions subexponentielles*)**.**
Soit $X_1, ..., X_n$ une suite de va's iid, de fdr F. La loi correspondant à F est dite subexponentielle si

$$\lim_{x\to\infty} \frac{\mathbb{P}(S_n > x)}{\mathbb{P}(\max(X_1, ..., X_n) > x)} = 1, \ n \geq 1. \tag{2.29}$$

Autrement dit, le comportement dans les queues d'une somme est alors essentiellement déterminé par la loi du maximum.

On peut démontrer que (2.29) implique que

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{e^{-\varepsilon x}} = \infty, \qquad \text{pour tout } \varepsilon > 0. \tag{2.30}$$

On rappelle que $e^{-\varepsilon x}$ est la forme de la queue de la loi exponentielle. Certaines propriétés de base des distributions subexponentielles peuvent être trouvées dans le livre de Embrechts et al. [46] ou dans le livre de Foss et al. [54]. Comme son nom l'indique, la classe D contient les distributions telles que $\overline{F}$ décroissent plus lentement que n'importe quelle loi exponentielle.

Le tableau suivant donne un certain nombre de distribution subexponentielle :

Distribution	$\overline{F}(x)$ ou $f(x)$	Paramètres
Weibull	$\overline{F}(x) = e^{-cx^\tau}$	$c > 0, 0 < \tau < 1$
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma > 0$
Benktender-type I	$\overline{F}(x) = \left(1 + 2\frac{\beta}{\alpha}\ln x\right) e^{-\left(\beta(\ln x)^2 + (\alpha+1)\ln x\right)}$	$\alpha, \beta > 0$
Benktender-type II	$\overline{F}(x) = e^{\alpha/\beta} x^{-(1-\beta)} e^{-\alpha x^\beta/\beta}$	$\alpha > 0, 0 < \beta < 1$

TAB. 2.4. Quelques distributions subexponentielles

2.4.3 Distributions à Variations Régulières

Dans cette partie, on traite la classe C des fonctions qui apparaît dans un vaste nombre d'applications dans la totalité de mathématiques. Ici, on va définir quelques généralités sur ces fonctions avec certaines de leurs propriétés les plus importantes. Ceux qui sont intéressés par la théorie de variation régulière, on peut consulter par exemple Bingham et al. [16], Embrechts et al. [46], Beirlant et al. [9], de Haan et Ferreira, [69] et Resnick [113].

Définition 2.12 (*Fonctions à variation régulière et à variation lente*)**.**

- *Une fonction mesurable* $V : \mathbb{R}_+ \to \mathbb{R}_+$ *est à variation régulière à* ∞ *avec l'index* $\rho \in \mathbb{R}$ *, et on note* $V \in \mathcal{RV}_\rho$*, si*

$$\lim_{x\to\infty} \frac{V(tx)}{V(x)} = x^\rho, \quad t > 0, \tag{2.31}$$

  on appelle ρ *l'exposant de variation ou l'indice de variation régulière.*

- *Une fonction mesurable* $V : \mathbb{R}_+ \to \mathbb{R}_+$ *est à variation lente à* ∞ *avec l'index* $\rho = 0$ *et on note* $V \in \mathcal{RV}_0$*, si*

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

- *Une fonction à variation régulière d'indice* $\rho \in \mathbb{R}$ *peut toujours s'écrire sous la forme* $V(x) = x^\rho \ell(x)$*, où* $\ell \in \mathcal{RV}_0$*.*

Il est facile de donner des exemples de fonctions à variations lentes. Les exemples typiques sont les fonctions des constantes positives, fonctions convergent vers une constante positive, logarithmes et logarithmes itérés. D'autre part,

- les fonctions x^ρ, $x^\rho \log(1+x)$ et $(x\log(1+x))^\rho$: sont à variations régulières.
- les fonctions $\exp(x)$, $\sin(x+2)$ et $\exp(\log(1+x))$: ne sont pas à variations régulières.

On note que $\log(x)$ est à variation lente, mais $\exp(\log(x))$ n'est pas à variation régulière. Enfin, on donne quelques propriétés élémentaires, des fonctions à variations lentes, où les preuves peuvent être laissées au lecteur.

Proposition 2.6 (*Propriétés de fonction à variation lente*)**.**

(i) $\mathcal{RV}_0$ *est fermé sous l'addition, la multiplication et la division.*

(ii) Si ℓ *est à variation lente,* $\lim_{x\to\infty}(\log l(x))/\log x = 0$*.*

(iii) Si ℓ *est à variation lente, alors* ℓ^α *est à variation lente pour tout* $\alpha \in \mathbb{R}$*.*

(iv) Si ℓ *est à variation lente et* $\rho > 0$*,*

$$\lim_{x\to\infty} x^\rho \ell(x) = \infty, \quad \lim_{x\to\infty} x^{-\rho}\ell(x) = 0.$$

Le premier résultat utile est le théorème de convergence uniforme.

Théorème 2.7 (*Convergence uniforme*)**.**
Si $V \in \mathcal{RV}_\rho$, pour $0 < a \leq b < \infty$, alors la relation (2.31) se tient uniformément pour

(i) $x \in [a, b]$ *si* $\rho = 0$.

(ii) $x \in (0, b]$ *si* $\rho > 0$.

(iii) $x \in [a, \infty)$ *si* $\rho < 0$.

Démonstration. Voir Bingham et al. [16], Théorème 1.5.2, page 22. □

Un autre résultat important concerne la représentation des fonctions à variations régulières.

Théorème 2.8 (*Représentation de Karamata*)**.**

(i) $\ell \in \mathcal{RV}_0$, *si et seulement si peut être représentée sous la forme.*

$$\ell(x) = c(x) \exp\left\{ \int_1^x \frac{r(t)}{t} dt \right\}, \quad x > 0, \tag{2.32}$$

où c, r sont des fonctions mesurables, et

$$\lim_{x\to\infty} c(x) = c_0 \in (0, \infty), \quad \text{et } \lim_{t\to\infty} r(t) = 0. \tag{2.33}$$

(ii) Une fonction $V : \mathbb{R}_+ \to \mathbb{R}_+$ à variation régulière avec l'index ρ si et seulement si V a la représentation

$$V(x) = c(x) \exp\left\{ \int_1^x t^{-1} \rho(t) dt \right\}, \quad x > 0, \tag{2.34}$$

où c satisfait (2.33) et $\lim_{t\to\infty} \rho(t) = \rho$.

Démonstration. Voir Resnick [113], Corollaire 2.1, page 29. □

Quelques autres propriétés utiles se rassemblent dans la proposition suivante :

Proposition 2.7.

(i) Si $V \in \mathcal{RV}_\rho$, $-\infty \leq \rho \leq \infty$, alors $\lim_{x\to\infty} \log V(x)/\log x = \rho$, donc

$$\lim_{x\to\infty} V(x) \to \begin{cases} \infty & si \quad \rho > 0, \\ 0 & si \quad \rho < 0. \end{cases}$$

(ii) *On suppose que V est non décroissante, $V(\infty) = \infty$, et $V \in \mathcal{RV}_\rho$, $0 \leq \rho \leq \infty$. Alors*

$$V^{-1} \in \mathcal{RV}_{-\rho}.$$

(iii) ***Inégalité de Potter :*** *On Suppose que $V \in \mathcal{RV}_\rho$, $\rho \in \mathbb{R}$ et pour tout $\varepsilon > 0$. Alors il existe t_0 tel que pour $x \geq 1$ et $t \geq t_0$,*

$$(1-\varepsilon)x^{\rho-\varepsilon} < \frac{V(tx)}{V(x)} < (1+\varepsilon)x^{\rho+\varepsilon}. \tag{2.35}$$

La preuve de la Proposition 2.7 est détaillée dans Resnick [113], Proposition 2.6, page 32.

Définition 2.13 (*Fonction à variation rapide*)**.**
Une fonction mesurable $V : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation rapide ou variation régulière d'indice ∞ et on note $V \in \mathcal{RV}_\infty$ si pour tout $t > 0$,

$$\lim_{x\to\infty} \frac{V(tx)}{V(x)} = x^\infty := \begin{cases} 0 & si\ x < 1, \\ 1 & si\ x = 1, \\ \infty & si\ x > 1. \end{cases}$$

De même, $V \in \mathcal{RV}_{-\infty}$ si

$$\lim_{x\to\infty} \frac{V(tx)}{V(x)} = x^{-\infty} := \begin{cases} \infty & si\ x < 1, \\ 1 & si\ x = 1, \\ 0 & si\ x > 1. \end{cases}$$

Un exemple d'une fonction $V \in \mathcal{RV}_{-\infty}$ est $V(x) = \exp(-x)$. Pour une discussion sur la liste des propriétés des fonctions à variation rapide voir Resnick [113] et voir aussi Embrechts et al. [46].

Conditions du Premier et du Second ordre

Dans le contexte des modèles à queue lourd, c'est-à-dire, pour tout $x > 0$, et avec la fonction de queue de quantile U définie par (1.15) , l'une des conditions (équivalentes) suivantes sont satisfaites :

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma} \iff \lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma, \tag{2.36}$$

l'Equation 2.36 est connue comme la condition du premier ordre des fonctions à variations régulières avec l'indice $-1/\gamma$ (ou l'indice γ) et $\gamma > 0$, ce qui signifie $\overline{F} \in \mathcal{RV}_{-1/\gamma}$ ou $U \in \mathcal{RV}_\gamma$.

Cependant, une condition du premier ordre en général n'est pas suffisante pour étudier les propriétés des estimateurs des paramètres de queue, en particulier la

normalité asymptotique. Dans ce cas, une condition du second ordre des fonctions à variations régulières est nécessaire en spécifiant le taux de convergence dans l'Equation 2.36. La définition suivante de cette condition vient de de Haan et Stadtmüller [72], Geluk et al. [59], de Haan et Ferreira [69], page 48, et Resnick [113], page 67. Voir aussi Neves [102] et Neves et al. [57] pour une version légèrement différente du second ordre étendu des fonctions à variations régulières. Plus d'informations sur la condition du troisième ordre peut être trouvés dans Gomes et al. [65], Fraga Alves et al. ([56], [55]) et plus généralement dans Wang et Cheng [129].

Définition 2.14 (*Condition du second ordre*)**.**
On dit que la fonction de queue de quantile U est à variation régulière du second ordre avec le paramètre du premier ordre $\gamma > 0$ et le paramètre du second ordre $\rho \leq 0$, on écrit $U \in 2\mathcal{RV}_{\gamma,\rho}$, s'il existe une fonction $A^*(t) \to 0$ et ne change pas le signe au voisinage de ∞, telles que

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^\gamma}{A^*(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \quad x > 0, \tag{2.37}$$

où $|A^*| \in \mathcal{RV}_\rho$ est appelée la fonction auxiliaire de U.

Le corollaire suivant exprime la condition du second ordre des fonctions à variations régulières en fonction de $\overline{F}$

Corollaire 2.3.
Pour tout $x > 0$ avec $\rho \leq 0$ and $A(t) := A^*(1/(1 - F(t)))$, la relation (2.37) est équivalente à

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma} \frac{x^\rho - 1}{\gamma\rho}. \tag{2.38}$$

La preuve du Corollaire 2.3 est détaillée notamment dans une version générale du Lemme de Vervaat's (de Haan et Ferreira, [69]).

Bien que l'estimation du paramètre du second ordre ρ est difficile, tant du point de vue théorique et appliquée, plusieurs estimateurs pour ρ qui fonctionnent bien dans la pratique ont été récemment introduites, et leurs propriétés asymptotiques ont été étudiées. On se réfère ici à Gomes et al. [65], Fraga Alves et al. [56] et Caeiro et Gomes [23], entre autres. Ciuperca et Mercadier [26] ont étendu les estimateurs dans [65] et [56]. Goegebeur et al. [64] ont introduit une nouvelle classe d'estimateurs de noyau basés sur l'échelle log-excesses. D'autres estimateurs de ce paramètre ont été proposés notamment par de Wet et al. [133], Caeiro et Gomes [24] pour réduire le biais et par Worms et Worms [135] qui introduisent un estimateur de moment de probabilité pondérée ρ. La classe des estimateurs dans [56] est considérée par de nombreux auteurs, l'état de l'art sur l'estimation du paramètre

du second ordre ρ et a reçu beaucoup d'attention dans la littérature, y compris les citations dans certains livres importants dans le domaine de la statistique des extrêmes (voir Beirlant et al. [9], de Haan et Ferreira [69], Resnick [113], ...).

Une valeur de ρ proche de 0 implique une faible vitesse de convergence. On note que la distribution de Pareto, où $x^{\gamma}\left(x^{\rho}-1\right)/\rho$ dans l'Equation 2.37 serait 0 pour toute fonction $A^{*}(t)$, ne satisfait pas la condition du second ordre. Comme un exemple de distributions à queue lourde satisfaire cette condition, on a ce qu'on appelle "le modèle de Hall".

Classe de Hall

Une sous-classe simple et riche de distributions à queue lourde joue un rôle important dans la discussion des estimateurs d'un IVE positif γ. Cette classe est présentée par Hall [73] et elle est mentionnée par "le modèle de Hall". Pour lesquels la fonction de queue de quantile de cette classe est sous forme

$$U(t)=Ct^{\gamma}+Dt^{\gamma+\rho}\left(1+o(1)\right), \quad \text{quand } t\to\infty. \tag{2.39}$$

avec $C>0$, $D\neq 0$,$\gamma>0$ et $\rho<0$ satisfait la condition du second ordre avec $A^{*}(t)=\rho\gamma dc^{\rho}t^{\rho}$, pour les constantes $c=C^{1/\gamma}>0$ et $d=\gamma^{-1}DC^{\gamma/(\gamma+\rho)}\neq 0$. La relation ci-dessus peut être reformulée en termes de fonction $\overline{F}$ comme suit :

$$\overline{F}(x)=cx^{-1/\gamma}+dx^{-1/\gamma+\rho/\gamma}(1+o(1)), \quad \text{quand } x\to\infty. \tag{2.40}$$

On note que $A(t)=A^{*}(1/\overline{F}(t))$ et avec un calcul simple montre que, dans le modèle de Hall, la fonction A est équivalente à $\rho\gamma dct^{\rho/\gamma}$ quand $t\to\infty$. Cette classe est très large et contient des distributions comme Fréchet, Burr, Cauchy, Pareto généralisée ($\mathcal{GPD}$), α-stable, log-logistic et log-gamma, ...etc, on donne certains exemples et pour plus de détails on réfère à Geluk et al. [59], Exemple 1.1 et Neves et al. [57], Exemple 4.2.

Exemple 2.3 (*Distribution log-gamma*)**.**
On Suppose que X_1, X_2 sont iid avec une densité exponentielle standard. La distribution log-gamma est la distribution de $\exp\{X_1+X_2\}$. *Pour* $x>1$

$$\begin{aligned}\mathbb{P}\left[\exp\{X_1+X_2\}>x\right] &= \mathbb{P}\left[X_1+X_2>\log x\right]\\ &= \exp\{-\log x\}+\exp\{-\log x\}\log x\\ &= x^{-1}(1+\log x)=\overline{F}(x).\end{aligned}$$

Donc, pour $x>1$

$$\frac{\overline{F}(tx)}{\overline{F}(t)}=x^{-1}=x^{-1}\left(\frac{\log x}{1+\log t}\right)\sim x^{-1}\frac{\log x}{\log t},$$

et

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

et avec $A(t) = 1/\log t$ *on a* $\gamma = 1$, $\rho = 0$ *et* $\lim_{t\to\infty} \overline{F}(t)/A(t) = 0$.

Exemple 2.4 (*Distribution de Fréchet*)**.**
Pour le modèle de Fréchet, $F(x) = \exp(x^{-1/\gamma})$, $\gamma > 0$, *on a*

$$\overline{F}(x) = x^{-1/\gamma} - \frac{1}{2}x^{-1/\gamma}(1 + o(1)), \quad \textit{qaund } x \to \infty,$$

et

$$U(t) = t^{\gamma} - \frac{\gamma}{2}t^{\gamma-1}(1 + o(1)), \quad \textit{qaund } t \to \infty.$$

Par conséquent, la distribution de Fréchet fait partie de la classe de Hall. Danc, $U \in 2\mathcal{RV}_{\gamma,-1}$ *avec* $C = 1$, $D = -1/2$ *et la fonction auxiliaire* $A^*(t) = \gamma/2t$.

Exemple 2.5 (*Distribution de Burr*)**.**
On considère la distribution de $Burr(\beta, \tau, \lambda)$, *donnée par*

$$F(x) = 1 - \left(\frac{\beta}{\beta + x^{\tau}}\right)^{\lambda}, \quad x > 0, \quad \beta, \tau, \lambda > 0.$$

La fonction de queue de quantile est

$$U(t) = \beta^{1/\tau}t^{1/\tau\lambda} - \tau^{-1}\beta^{1/\tau}t^{1/(\tau\lambda)-1/\lambda}(1 + o(1)).$$

Par conséquent, la distribution de Burr fait partie de la classe de Hall. Donc, $U \in 2\mathcal{RV}_{\gamma,\rho}$ *avec* $\gamma = 1/\tau\lambda$, $\rho = -1/\lambda$ *et la fonction auxiliaire* $A^*(t) = (\lambda\tau)^{-1}\, t^{-1/\lambda}$.

2.4.4 Distributions à queue de type-Pareto

Une classe de distributions s'écrivant selon le modèle suivant

$$\overline{F}(x) = (x/C)^{-1/\gamma}, \quad x \geq C \text{ et } C > 0, \tag{2.41}$$

est dite à queue de type-Pareto (B) avec l'IVE $\gamma > 0$ et C est un paramètre d'échelle. L'Equation 2.41 est équivalente à

$$U(t) = Ct^{\gamma}, \quad C > 0,\ \gamma > 0.$$

Il existe dans la littérature de nombreux auteurs qui se sont concentrés sur cette classe.

2.4.5 Distributions α-stables

La classe des distributions stables (appelée aussi α-stables, Pareto-stables ou Lévy-stable). Cette classe a été présentée par Lévy au début des années 1920 dans son livre "Calcul des probabilités" (Lévy, [88]). Ces distributions constituent une classe très riche de lois de probabilité capables de représenter différentes asymétries et des queues très lourdes. Les définitions et les propriétés énoncées dans cette partie sont issues de Zolotarev [140] et de Samorodnitsky et Taqqu [117].

Définition 2.15 (*Variable aléatoire stable*)**.**
Une va X est dite stable ou a une distribution stable si est seulement si, pour tous nombres positifs a et b et des copies X_1 et X_2 indépendantes de X, il existe deux nombres réels $c > 0$ et d tels que :

$$aX_1 + bX_2 \stackrel{d}{=} cX + d. \tag{2.42}$$

Dans le cas $d = 0$, on dit que X est strictement stable.

Cette définition montre que la famille des lois stables est préservée par convolution. On peut utiliser une autre définition des va's stables, équivalente à la Définition 2.15.

Définition 2.16 (*Définition équivalente*)**.**
Une va X est dite stable si, pour tout entier non nul n, il existe deux constantes $a_n > 0$ et b_n telles que

$$S_n = X_1 + ... + X_n \stackrel{d}{=} a_n X + b_n, \quad n \geq 1,$$

ou de manière équivalente

$$a_n^{-1}(S_n - b_n) \stackrel{d}{=} X.$$

où $X_1, ..., X_n$ sont des va's indépendantes, ayant chacune la même distribution que X.

Les lois stables sont les seules lois qui peuvent être obtenues comme limites de sommes normalisées de va's iid. La Définition 2.16 généralise le théorème centrale limite (voir Feller, [51]).

Définition 2.17 (*Domaine d'attraction des lois $\alpha-$stables*)**.**
On dit que F appartient au domaine d'attraction d'une loi stable d'indice de stabilité $0 < \alpha \leq 2$ et on note $F \in \mathcal{D}(\alpha)$, s'il existe deux suites réelles $A_n > 0$ et B_n tels que

$$A_n^{-1}(\sum_{i=1}^{n} X_i - B_n) \xrightarrow{\mathcal{D}} S_\alpha(\sigma, \beta, \mu), \quad \text{quand } n \to \infty.$$

En raison de l'importance de cette classe de distributions, il est nécessaire de les décrire analytiquement. Leur principal inconvénient est qu'elles ne possèdent pas de formes explicites pour les densités et les fdr's, sauf dans trois cas qu'on verra par la suite. La méthode la plus commune de les décrire se fait par leurs fonctions caractéristiques.

Définition 2.18 (*Fonction caractéristique*)**.**
La fonction caractéristique d'une va stable est définie, pour $t \in \mathbb{R}$*, par*

$$\varphi(t) := \begin{cases} \exp\left\{i\mu t - \sigma^\alpha |t|^\alpha \left(1 - i\beta sign(t) \tan\left(\frac{\pi\alpha}{2}\right)\right)\right\} & si\ \alpha \neq 1, \\ \exp\left\{i\mu t - \sigma |t| \left[1 + i\beta sign(t) \frac{2}{\pi} \ln |t|\right]\right\} & si\ \alpha = 1, \end{cases}$$

où

$$i^2 = -1, \quad sign(t) := \begin{cases} 1 & quand \quad t > 0, \\ 0 & quand \quad t = 0, \\ -1 & quand \quad t < 0, \end{cases}$$

et α*,* σ*,* β *et* μ *sont des paramètres réels tels que* $0 < \alpha \leq 2$*,* $\sigma \geq 0$*,* $-1 \leq \beta \leq 1$ *et* $-\infty < \mu < +\infty$*.*

Une distribution stable est donc caractérisée par quatre paramètres :

- α: est appelé "exposant caractéristique ou indice de stabilité", décrit la forme de la distribution ou le degré d'épaisseur de ses queues.
- β : est appelé "paramètre d'asymétrie". Si β est égal à -1 (respectivement $+1$) la distribution est totalement asymétrique à gauche (respectivement à droite). Lorsque β vaut zéro alors la distribution est symétrique.
- μ: est appelé "paramètre de position ou localisation". Il correspond, pour α supérieur à 1, à l'espérance. Si $\beta = 0$ alors μ est la médiane. Dans les autres cas, le paramètre μ ne peut pas être interprété.
- σ: est appelé "paramètre d'échelle".

Remarque 2.6.

1. *Une va* X *distribuée suivant une loi stable des paramètres* α*,* σ*,* β *et* μ *sera notée par* $X \sim S_\alpha(\sigma, \beta, \mu)$ *ou* $X_\alpha(\sigma, \beta, \mu)$*.*
2. *Lorsque la variable* X *sera symétrique autour de* 0 (*si* $\beta = \mu = 0$)*, elle sera simplement notée par* $S\alpha S$*. Dans ce cas, la fonction caractéristique a la forme simple*
$$\varphi(t) = \exp\left(-\sigma^\alpha |t|^\alpha\right), \quad t \in \mathbb{R}.$$
3. *Si seulement* $\beta = 0$*, on dit que la distribution est symétrique autour du* μ*.*

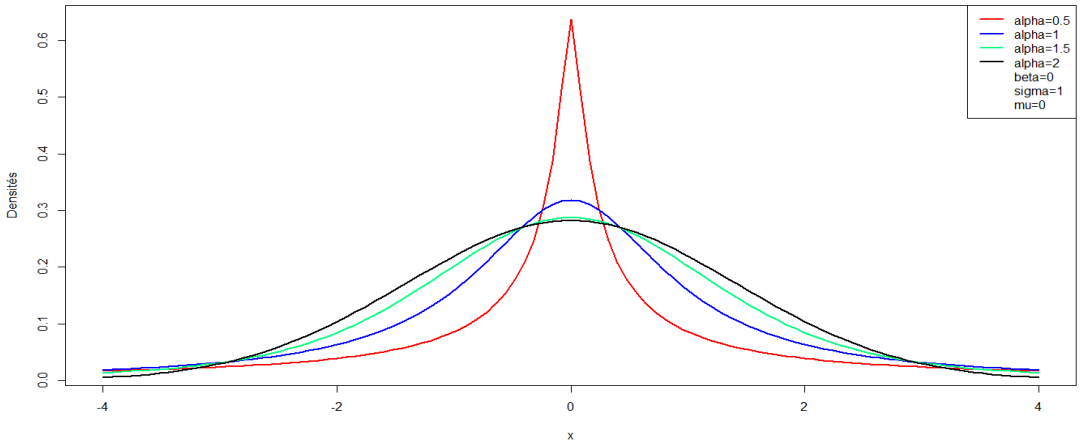

FIG. 2.6. Densités $\alpha-$stables pour différentes valeurs de α.

Exemple 2.6 (*Distributions α-stables*)**.**
Les distributions α-stables les plus connues, et les seules dont nous disposons d'une forme explicite pour les densités, sont les suivantes

- *La distribution Gaussienne $S_2(\sigma, 0, \mu)$ où*

$$f(x) = \frac{1}{2\sigma\sqrt{\pi}} \exp\left(-\frac{(x-\mu)^2}{4\sigma^2}\right).$$

- *La distribution de Cauchy $S_1(\sigma, 0, \mu)$ où*

$$f(x) = \frac{2\sigma}{\pi\left(4\sigma^2 + (x-\mu)^2\right)}.$$

- *La distribution de Lévy $S_{1/2}(\sigma, 1, \mu)$ où*

$$f(x) = \sqrt{\frac{\sigma}{2\pi}}\,(x-\mu)^{-3/2} \exp\left(-\frac{\sigma}{2\,(x-\mu)}\right).$$

On va rappeler quelques propriétés importantes des va's stables de loi $S_\alpha(\sigma, \beta, \mu)$.

Proposition 2.8 (*Propriétés arithmétiques*)**.**

(i) Soient X_1 et X_2 deux va's stables et indépendantes, avec $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$ pour $i = 1, 2$, alors $X_1 + X_2 \sim S_\alpha(\sigma, \beta, \mu)$ où

$$\sigma = (\sigma_1^\alpha + \sigma_2^\alpha)^{1/\alpha}, \quad \beta = \frac{\beta_1\sigma_1^\alpha + \beta_2\sigma_2^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}, \quad \mu = \mu_1 + \mu_2.$$

On note que si $\beta_1 = \beta_2$ alors $\beta = \beta_1 = \beta_2$.

(ii) *Soit* $X \sim S_\alpha(\sigma, \beta, \mu)$ *et* $a \in \mathbb{R}$. *Alors* $X + a \sim S_\alpha(\sigma, \beta, \mu + a)$.

(iii) *Soit* $X \sim S_\alpha(\sigma, \beta, \mu)$ *et* $a \in \mathbb{R}$. *Alors*

$$aX \sim \begin{cases} S_\alpha\left(|a|\,\sigma, sign(a)\beta, a\mu\right) & \text{pour } \alpha \neq 1, \\ S_1\left(|a|\,\sigma, sign(a)\beta, a\mu - \dfrac{2}{\pi}a\left(\log|a|\right)\sigma\beta\right) & \text{pour } \alpha = 1. \end{cases}$$

(iv) *Si* $0 < \alpha < 2$ *et* $X \sim S_\alpha(\sigma, \beta, 0)$ *alors* $-X \sim S_\alpha(\sigma, -\beta, 0)$.

La classe des distributions α-stable a des bonnes propriétés de queues lourdes.

Proposition 2.9 (*Propriétés de queues lourdes*)**.**
Soit $X \sim S_\alpha(\sigma, \beta, \mu)$ *avec* $0 < \alpha < 2$. *Alors on a, quand* $x \to \infty$

$$x^\alpha P(X > x) \longrightarrow C_\alpha \frac{1+\beta}{2}\sigma^\alpha \quad \text{et} \quad x^\alpha P(X < -x) \longrightarrow C_\alpha \frac{1-\beta}{2}\sigma^\alpha, \tag{2.43}$$

où

$$C_\alpha := \left(\int_0^\infty x^{-\alpha}\sin x\; dx\right)^{-1} = \frac{2}{\pi}\Gamma(\alpha)\sin\frac{\pi\alpha}{2},$$

et $\Gamma(\alpha)$ *la fonction gamma définie, pour* $\alpha > 0$, *par* $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$.

La démonstration des deux Propositions 2.8 et 2.9 sont données dans Samorodnitsky et Taqqu.

Si on note par $G(x) := \mathbb{P}(|X| \leq x) = F(x) - F(-x)$, $x > 0$, la fdr de $|X|$ et F est la fdr de X, alors les relations de (2.43) rapportent que les queues de distribution de $F \in \mathcal{D}(\alpha)$, pour $0 < \alpha < 2$, satisfant les deux conditions suivantes :

1. Condition de variation régulière (voir Geluk et al., [59]),

$$\lim_{t\to\infty}\frac{\overline{G}(tx)}{\overline{G}(t)} = x^{-\alpha}, \quad x > 0.$$

2. Condition d'équilibre de queue, pour $0 \leq p \leq 1$, on a

$$\frac{\overline{F}(t)}{\overline{G}(t)} \to p := \frac{1+\beta}{2} \quad \text{et} \quad \frac{F(-t)}{\overline{G}(t)} \to q := \frac{1-\beta}{2}. \tag{2.44}$$

Mandelbrot [91] et Fama [50] ont montré, dans les premiers travaux, que la loi α-stable est une bonne candidate pour adapter les séries financières à queues lourdes et ils ont été proposés également comme un modèle pour beaucoup de types de systèmes économiques physiques. Pour plus de détails sur ce sujet voir aussi Weron [132].

La loi normale est la seule loi α-stable qui possède des moments finis de tous ordres. Pour toutes les autres lois α-stables, on a les résultats suivants.

Corollaire 2.4 (*Moments*).
Soit $X \sim S_\alpha(\sigma, \beta, \mu)$ *avec* $\alpha \in]0, 2[$. *Alors*

$$\mathbf{E}\,|X|^\delta < \infty \quad si \quad \delta \in]0, \alpha[\,,$$
$$\mathbf{E}\,|X|^\delta = \infty \quad si \quad \delta \in [\alpha, \infty[\,.$$

Considérant que pour $\alpha = 2$, *on a* $\mathbf{E}\,|X|^\delta < \infty$, $\forall \delta$.

Les résultats concernant la moyenne $\mathbf{E}(X)$ et la variance $Var(X)$ d'un va stable X sont résumées dans la Table 2.5.

Indice de stabilité	$0 < \alpha \leq 1$	$1 < \alpha < 2$	$\alpha = 2$
$\mathbf{E}(X)$	∞	μ	μ
$Var(X)$	∞	∞	$2\sigma^2$

TAB. 2.5. Moments d'une va suivant une loi stable selon α.

Remarque 2.7.

1. *L'existence d'une variance finie pour la loi normale est simplement liée à une plus grande décroissance de queue par rapport aux autres lois stables.*
2. *Le fait que lorsque* $\alpha < 2$, *la variance de la distribution est infinie. Dans ce cas le TCL n'est pas applicable.*
3. *La loi de Cauchy* ($\alpha = 1$) *et celle de Lévy* ($\alpha = 1/2$) *ont chacune une espérance mathématique infinie. Dans ces deux cas, la variance n'est pas définie.*

Après avoir étudié la classe de $\alpha-$stable, dans les paragraphes suivants, on s'intéresse à présenter les estimateurs des trois paramètres de loi α-stable.

Estimation de l'Indice de Stabilité

L'exposant caractéristique α est le paramètre principal, il régit le comportement des queues de distribution. Beaucoup d'estimateurs sont proposés pour α via l'approche des valeurs extrêmes. Le plus célèbre, mais pas nécessairement le meilleur, de ces estimateurs est celui défini par Hill [76], comme suit :

$$\widehat{\alpha}_n = \widehat{\alpha}_n(k) := \left(\frac{1}{k} \sum_{i=1}^{k} \log |X_{n-i+1:n}| - \log |X_{n-k:n}| \right)^{-1}, \tag{2.45}$$

où $|X_{1:n}| \leq ... \leq |X_{n:n}|$ sont les statistiques d'ordre associées à $|X_1|, ..., |X_n|$, $(n \geq 1)$, de la va $|X|$, avec X une va stable et $k = k_n$ est une suite d'entiers liées à la taille de l'échantillon n telle que :

$$\lim_{n\to\infty} k_n = \infty \quad \text{et} \quad \lim_{n\to\infty} k_n/n = 0. \tag{2.46}$$

Une suite vérifiant les deux conditions de (2.46) sera appelée suite intermédiaire d'entiers, où la première condition $k \to \infty$ nous assure que le nombre de statistiques d'ordre k est assez grand afin d'obtenir des estimateurs stables. Par contre, la deuxième condition $k/n \to 0$ permet de rester dans la queue de distribution (Gardes et Girard, [58]).

Sous la condition du second ordre (2.37) avec $\gamma = 1/\alpha$, $k = k_n$ vérifiant les deux conditions de (2.46) et $\lim_{n\to\infty} \sqrt{k}A(n/k) = \lambda$, avec λ fini, Peng [106] a montré que

$$\sqrt{k}\left(\widehat{\alpha}_n^{-1} - \alpha^{-1}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\lambda/(1-\rho), \alpha^{-2}\right), \text{ quand } n \to \infty. \tag{2.47}$$

Weron [132] a discuté la performance de $\widehat{\alpha}_n$ en cas de distributions de Lévy-stable et a noté que pour $\alpha \leq 1,5$ l'estimation est tout à fait raisonnable mais qaund α approche 2, il y a une surestimation significative lors de l'examen des échantillons de taille typique. On trace l'estimateur de α en fonction de k basés sur 1000 échantillons de taille 5000 pour la loi $\alpha-$stable, voir la Figure 2.7.

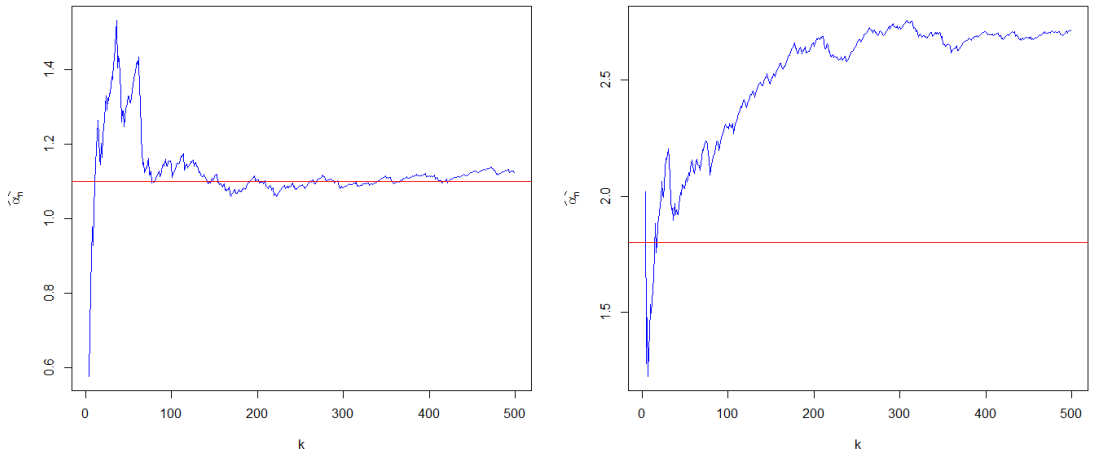

FIG. 2.7. Estimateur de Hill pour α de $S_\alpha(1,0,0)$ avec $\alpha = 1.1$ (gauche) et $\alpha = 1.8$ (droite). La ligne horizontale représente la vraie valeur de α.

Estimation du Paramètre de Localisation

Pour $1 < \alpha \leq 2$ la moyenne de X existe et elle est égale au paramètre de localisation μ, la variance de X est finie et elle est égale à $2\sigma^2$. On a vu dans le premier chapitre l'estimateur naturel de μ est la moyenne empirique $\overline{X}$, et en vertu au Théorème 1.3 l'estimation de X est asymptotiquement normale. Mais, dans le cas $1 < \alpha < 2$, le Théorème 1.3 n'est pas applicable parce que la variance de X est

infinie. Par conséquent, la normalité asymptotique de la moyenne de l'échantillon $\overline{X}$ n'est pas établie. Pour pallier ce problème, Peng [107] a proposé un estimateur asymptotiquement normal basé sur la théorie des valeurs extrêmes, comme suit :

$$\widehat{\mu}_n^P = \widehat{\mu}_n^P(k) := \widehat{\mu}_n^{(1)} + \widehat{\mu}_n^2 + \widehat{\mu}_n^{(3)},$$

où

$$\widehat{\mu}_n^2 = \widehat{\mu}_n^2(k) := \frac{1}{n}\sum_{i=k+1}^{n-k} X_{i:n}, \tag{2.48}$$

$$\begin{aligned} \widehat{\mu}_n^{(1)} = \widehat{\mu}_n^{(1)}(k) &:= \frac{k}{n} X_{k:n} \frac{\widehat{\alpha}_n^{(1)}}{\widehat{\alpha}_n^{(1)} - 1}, \\ \widehat{\mu}_n^{(3)} = \widehat{\mu}_n^{(3)}(k) &:= \frac{k}{n} X_{n-k+1:n} \frac{\widehat{\alpha}_n^{(3)}}{\widehat{\alpha}_n^{(3)} - 1}, \end{aligned} \tag{2.49}$$

avec

$$\begin{aligned} \widehat{\alpha}_n^{(1)} &= \widehat{\alpha}_n^{(1)}(k) := \left(\frac{1}{k}\sum_{i=1}^{k} \log\left(-X_{i:n}\right) - \log\left(-X_{k:n}\right) \right)^{-1}, \\ \widehat{\alpha}_n^{(3)} &= \widehat{\alpha}_n^{(3)}(k) := \left(\frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n} \right)^{-1}. \end{aligned}$$

On note que $\widehat{\alpha}_n^{(1)}$ et $\widehat{\alpha}_n^{(3)}$ sont également des estimateurs convergents en probabilité vers α (see Mason, [92]) et la convergence presque sûre est établie dans Necir [100] en 2006.

Sous la condition du second ordre sur la fdr G et avec

$$\lim_{t\to\infty} \frac{\overline{F}(t)/\overline{G}(t) - p}{A(t)} = r \in \mathbb{R},$$

où p est défini dans (2.44) , Peng [107] a montré que, quand $k = o(n^{-2\rho(\alpha-2\rho)})$, avec le paramètre du second ordre $\rho < 0$, alors

$$\frac{\sqrt{n}}{\sigma(k/n)}\left(\widehat{\mu}_n^P - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^2) \quad \text{quand } n \to \infty, \tag{2.50}$$

ou de façon équivalente

$$\frac{\sqrt{n}}{\delta\sigma(k/n)}\left(\widehat{\mu}_n^P - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad \text{quand } n \to \infty, \tag{2.51}$$

où

$$\delta^2 := 1 + \left(\frac{(2-\alpha)\,(2\alpha^2 - 2\alpha + 1)}{2\,(\alpha-1)^4} + \frac{(2-\alpha)}{(\alpha-1)} \right),$$

et

$$\sigma^2(s) := \int_s^{1-s} \int_s^{1-s} (u \wedge v - uv)\, dF^{-1}(u) dF^{-1}(v), \quad 0 < s < 1.$$

Pour calculer les bornes de confiance de μ, Meraghni et Necir [94] ont utilisé l'approximation suivante prouvée dans Peng [107]

$$\sqrt{k/n} F^{-1}(k/n)\sigma(k/n) \xrightarrow{\mathbb{P}} - \left(\frac{2-\alpha}{2\,(p^{2/\alpha} + (1-p)^{2/\alpha})} \right)^{1/2} (1-p)^{1/\alpha}, \text{ quand } n \to \infty. \tag{2.52}$$

Dans le cas les distributions symétriques ($\beta = 0$), la relation (2.52) peut être réécrite comme

$$\sigma(k/n) \sim -\frac{2\sqrt{k/n} F^{-1}(k/n)}{\sqrt{2-\alpha}}, \quad \text{quand} \quad n \to \infty. \tag{2.53}$$

Estimation du Paramètre d'Échelle

Meraghni et Necir [95] ont introduit un estimateur asymptotiquement normal pour le paramètre d'échelle σ d'une distribution Lévy-stable par l'approche de la TVE, ils ont illustré leurs résultats sur des ensembles d'observations simulées stables. La forme de l'estimateur de σ est

$$\widehat{\sigma}_n := |X_{n-k:n}| \left(\frac{k\pi}{2n\Gamma\left(\widehat{\alpha}_n\right) \sin \frac{\pi \widehat{\alpha}_n}{2}} \right)^{1/\widehat{\alpha}_n}.$$

La consistance et la normalité asymptotique de cet estimateur ont été établies par Meraghni et Necir [95], sous la condition de G satisfaisant (2.40) avec $k = k_n$ est une suite d'entiers satisfaisant (2.46).

Théorème 2.9 (*Propriétés asymptotiques de* $\widehat{\sigma}_n$)**.**

(i) *Consistance :*

$$\widehat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma, \quad \textit{quand} \quad n \to \infty.$$

(ii) *Normalité asymptotique : pour* $k \sim \left(\frac{\rho^{-3}(1-\rho)^2}{(-2d^2c^2)} \right)^{1/(1-2\rho)} n^{-2\rho/(1-2\rho)}$, *on a*

$$\frac{\sqrt{k}}{\log(k/n)} \left(\log \widehat{\sigma}_n - \log \sigma \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\lambda/(1-\rho), \alpha^{-2} \right), \quad \textit{quand} \quad n \to \infty.$$

Démonstration. Voir Meraghni et Necir [95], Théoème 4.1 et Théoème 4.2. □

2.5 Estimation de l'IVE sans Censure

On a vu dans ce Chapitre que pour la majorité des fd's F la loi asymptotique du maximum $X_{n:n}$ (convenablement normalisé) est une loi des valeurs extrêmes qui étant indexée par le paramètre de queue γ, ce paramètre apporte une information sur la forme de la queue de distribution de F. Notamment, selon que $\gamma > 0$, $\gamma < 0$ ou $\gamma = 0$, on distingue trois domaines d'attraction : Frechet, Weibull et Gumbel. Il existe dans la littérature de la TVE de nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes γ et des quantiles extrêmes. Dans cette Section, on exposera uniquement trois estimateurs de γ, on cite l'estimateur de Pickands [109], l'estimateur de Hill [76] et l'estimateur des moments (**Dekkers et al.,** [39]). On donne également certaines de leurs propriétés statistiques. Ces estimateurs sont basées fortement sur les plus grandes statistiques d'ordre $X_{n-k:n} \leq ... \leq X_{n:n}$, où la statistique $X_{n-k:n}$ est alors dite statistique d'ordre intermédiaire.

2.5.1 Estimateur de Pickands

L'estimateur de Pickands a été introduit en 1975 par James Pickands dans [109] pour toute $\gamma \in \mathbb{R}$.

Définition 2.19 (*Estimateur de Pickands*)**.**
Soit $X_1, ..., X_n$, n va's iid de fdr $F \in \mathcal{D}(\mathcal{H}_\gamma)$, où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, l'estimateur de Pickand est défini par

$$\widehat{\gamma}^P = \widehat{\gamma}^P(k) := \frac{1}{\log 2} \log \left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}} \right). \tag{2.54}$$

L'auteur démontre la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan [38]. Des améliorations de cet estimateur ont été introduites notamment par Drees [42] et Segers [119].

Sous certaines conditions sur la suite entière k et la fdr F, l'estimateur de γ a des bonnes propriétés asymptotiques, ils sont regroupés dans le théorème suivant :

Théorème 2.10 (*Propriétés asymptotiques de $\widehat{\gamma}^P$*)**.**
Soit $F \in \mathcal{D}(H_\gamma), \gamma \in \mathbb{R}$, $k \to \infty$ et $k/n \to 0$ quand $n \to \infty$

(i) *Consistance faible :*

$$\widehat{\gamma}^P \xrightarrow{\mathbb{P}} \gamma, \quad \text{quand } n \to \infty.$$

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

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

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

$$\sqrt{k}\left(\widehat{\gamma}^P - \gamma\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{V}^2), \quad \text{quand } n \to \infty,$$

où

$$\mathcal{V}^2 := \frac{\gamma^2\left(2^{2\gamma+1}+1\right)}{\left(2\left(2^\gamma - 1\right)\log 2\right)^2}.$$

Ce dernier résultat *(iii)* permet donc de donner un intervalle de confiance pour l'estimation. Mais attention, l'estimateur de Pickand est biaisé. Pour un échantillon de taille n fixée, on trace le graphe de l'estimateur de Pickand : $\widehat{\gamma}^P$ en fonction de k, voir la Figure 2.8. On est alors confronté au dilemme suivant :

- Pour k petit, il y a de grandes oscillations avec un intervalle de confiance large.
- Pour k grand, on aura un intervalle de confiance plus étroit mais pas centrée sur la vraie valeur.

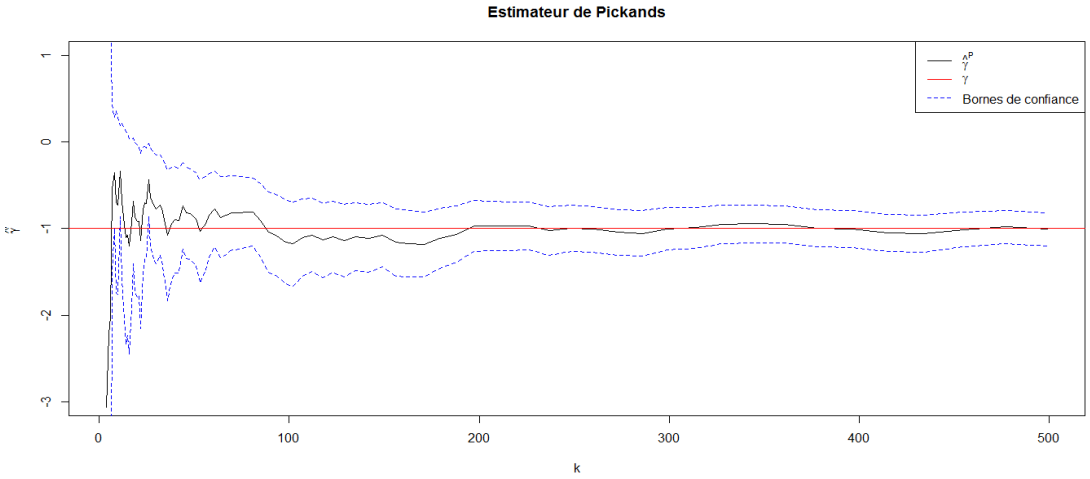

FIG. 2.8. Estimateur de Pickands avec l'intervalle de confiance au niveau 95% pour γ basés sur 1000 échantillons de taille 5000 pour la loi uniform standard ($\gamma = 1$).

Définition 2.20 (*Estimateur* $\widehat{Q}^P$)**.**
L'estimateur $\widehat{Q}^P$ de quantile $Q(1-s)$ associé à l'estimateur de Pickands est

$$\widehat{Q}^P := X_{n-k+1:n} + \frac{(k/ns)^{\widehat{\gamma}^P} - 1}{1 - 2^{-\widehat{\gamma}^P}}\left(X_{n-k+1:n} - X_{n-2k+1:n}\right). \tag{2.55}$$

Les propriétés asymptotiques de l'estimateur (2.55), sont discutées par Dekkers et de Haan [38], voir aussi Matthys et Beirlant [93].

2.5.2 Estimateur de Hill

Les recherches se sont principalement concentrées sur le cas où l'IVE est positif ($\gamma = \alpha^{-1} > 0$) parce que les ensembles de données dans la plupart des applications réelles, qui correspond aux distributions appartenant au domaine d'attraction de Fréchet $\mathcal{D}\left(\Phi_{1/\gamma}\right)$, c'est-à-dire, quand la queue de distribution a une forme de Pareto. L'estimateur le plus connu de γ est l'estimateur proposé par Hill [76] et donné par la définition suivante :

Définition 2.21 (*Estimateur de Hill*)**.**
Soit $X_1, ..., X_n$, n va's iid de fdr $F \in \mathcal{D}\left(\Phi_{1/\gamma}\right)$, où $\gamma < 0$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, l'estimateur de Hill est défini par

$$\widehat{\gamma}^H = \widehat{\gamma}^H(k) := \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n}. \tag{2.56}$$

la construction de cet estimateur est donnée dans le livre de de Haan et Ferreira [69] et dans le livre de Beirlant et al. [9]. D'autres estimateurs de l'IVE ont été proposés notamment par Beirlant et al. ([7], [8]) qui utilisent un modèle de régression exponentiel pour débiaiser l'estimateur de Hill et par Csörgő et al. [31]. qui utilisent un noyau dans l'estimateur de Hill. Un grand nombre de travaux théoriques ont été consacrés à l'étude des propriétés de l'estimateur de Hill. La consistance faible a été établie par Mason [92], La consistance forte fut établie en 1988 par Deheuvels et al [37] et plus récemment par Necir [100]. La normalité asymptotique est due entre autres à Davis et Resnick [34], Csörgő et Mason [32] et Häusler et Teugels [74].

Théorème 2.11 (*Propriétés asymptotiques $\widehat{\gamma}^H$*)**.**
On suppose que $F \in \mathcal{D}\left(\Phi_{1/\gamma}\right)$, $\gamma > 0$, $k \to \infty$ et $k/n \to 0$ quand $n \to \infty$

(i) *Consistance faible :*

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

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

$$\widehat{\gamma}^H \xrightarrow{p.s.} \gamma, \quad \text{quand } n \to \infty.$$

(iii) *Normalité asymptotique : On Suppose que F satisfaisant (2.37). Si $\sqrt{k}A(n/k) \to \lambda$ quand $n \to \infty$, alors*

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

Comme l'estimateur de Pickands, cet estimateur est biaisé et le résultat sur sa normalité asymptotique permet de donner un intervalle de confiance pour γ. Concernant l'étude du quantile extrême $Q(1-\varepsilon)$, l'estimateur le plus fréquemment utilisé a été proposé par Weissman [130]. Il est donné par la définition suivante :

Définition 2.22 (*Estimateur de Weissman* $\widehat{Q}^H$)**.**
L'estimateur de Weissman est défini par

$$\widehat{Q}^H := X_{n-k+1:n} + \left(\frac{k}{ns}\right)^{\widehat{\gamma}^H}.$$

Plus de détails sur les propriétés de cet estimateur sont disponibles dans Weissman [130], Embrechts et al. [46] et de Haan et Ferreira [69].

La Figure 2.9 illustre le graphe de $\widehat{\gamma}^H$ et l'intervalle de confiance en fonction de k. On observe que pour k petit, il y a de grandes oscillations avec un intervalle de confiance large et pour k grand, l'intervalle de confiance devient plus étroit.

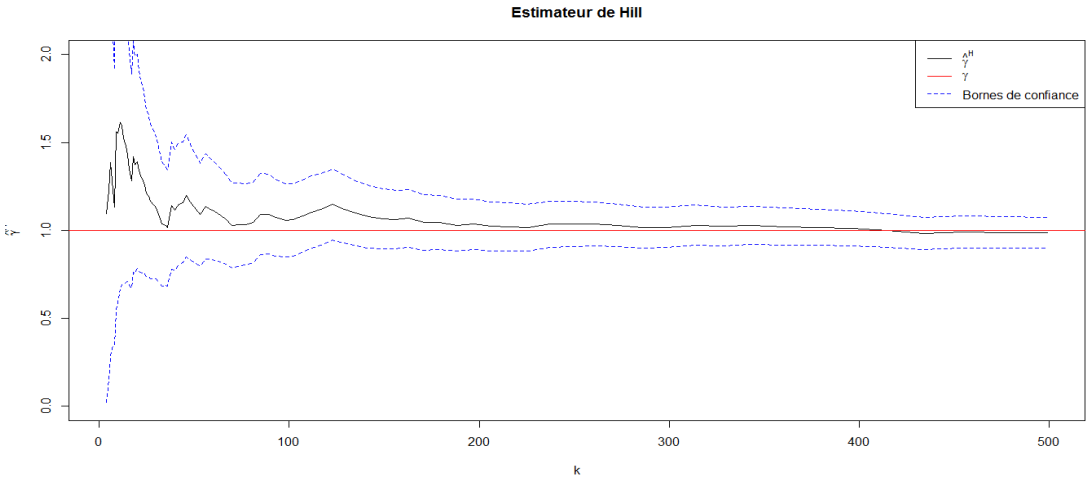

FIG. 2.9. Estimateur de Hill et l'intervalle de confiance de niveau 95%, pour l'IVE de la loi de Pareto standard ($\gamma = 1$) basé sur 1000 échantillons de 5000 observations.

Exemple 2.7. *(Loi de Pareto)*
Pour voir la performance de l'estimateur de Hill, on a effectué une étude de simulation basée sur 1000 *échantillons de la loi de Pareto avec* $\gamma \in \{0.6, 1\}$. *Les résultats de* $\widehat{\gamma}^H$, *biais abs., la racine de l'erreur moyenne quadratique (root of the*

mean squared error : rmse) et les bornes de confiance au niveau 95% *de* γ *respectivement définis par*

$$biais := \frac{1}{r}\sum_{i=1}^{r}\left(\widehat{\gamma}^H - \gamma\right), \quad rmse := \sqrt{\frac{1}{r}\sum_{i=1}^{r}\left(\widehat{\gamma}^H - \gamma\right)^2},$$

et

$$\gamma \in \left[\widehat{\gamma}^H - z_{1-\alpha/2}\frac{\widehat{\gamma}^H}{\sqrt{k}}; \widehat{\gamma}^H + z_{1-\alpha/2}\frac{\widehat{\gamma}^H}{\sqrt{k}}\right],$$

où $z_{1-\alpha/2}$ *est le quantile d'ordre* $(1-\alpha/2)$ *d'une loi normale centrée réduite, sont résumés dans la Table 2.6.*

γ	n	$\widehat{\gamma}^H$	biais abs.	mse	interv.conf.	prob.couv	longueur
0.6	1000	0.579	0.0201	0.010	0.463 – 0.789	0.91	0.326
	3000	0.594	0.006	0.003	0.520 – 0.696	0.92	0.176
	5000	0.594	0.006	0.002	0.536 – 0.668	0.93	0.133
1	1000	0.967	0.033	0.030	0.771 – 1.320	0.92	0.549
	3000	0.992	0.008	0.009	0.868 – 1.162	0.92	0.294
	5000	0.992	0.008	0.006	0.891 – 1.122	0.92	0.231

TAB. 2.6. Résultats de simulation de l'estimation de γ d'une distribution Pareto basé sur 1000 échantillons.

2.5.3 Estimateur des Moments

Un inconvénient de l'estimateur de Hill est qu'il est conçu seulement pour l'IVE des distributions à queues lourdes. En 1989, Dekkers et al. ont proposé dans [39] une extension de tout type de distribution, appelée estimateur de moment.

Définition 2.23 (*Estimateur des Moments*)**.**
Pour $\gamma \in \mathbb{R}$*, l'estimateur des moment est*

$$\widehat{\gamma}^M = \widehat{\gamma}^M(k) := M_n^{(1)} + T_n := M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}, \tag{2.57}$$

avec

$$M_n^{(r)} = M_n^{(r)}(k) := \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1:n} - \log X_{n-k:n})^r, \quad r = 1, 2, \tag{2.58}$$

où $M_n^{(1)}$ *est l'estimateur de Hill* $\widehat{\gamma}_n^H$*.*

Les propriétés asymptotiques de cet estimateur ont été étudiées dans Dekkers et al. [39].

Théorème 2.12 (*Propriétés asymptotiques de* $\widehat{\gamma}^M$)**.**
Soit $F \in \mathcal{D}(\mathcal{H}_\gamma)$, $\gamma \in \mathbb{R}$, $k \to \infty$ *et* $k/n \to 0$ *quand* $n \to \infty$.

(i) *Consistance faible :*

$$\widehat{\gamma}^M \xrightarrow{\mathbb{P}} \gamma, \quad \text{quand } n \to \infty.$$

(ii) *Consistance forte : si* $k/(\log n)^\delta \to \infty$ *quand* $n \to \infty$, *pour* $\delta > 0$, *alors*

$$\widehat{\gamma}^M \xrightarrow{p.s.} \gamma, \quad \text{quand } n \to \infty.$$

(iii) *Normalité asymptotique : si les conditions du Théorème* 3.1 *de Dekkers et al. [39] sont satisfaisants et si* $k = o(n/g_1^{-1}(n))$ *où* $g_1(t) := t\,(U(t)/a(t))^2$, *alors*

$$\sqrt{k}\left(\widehat{\gamma}^M - \gamma\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{V}^2), \quad \text{quand } n \to \infty,$$

avec

$$\mathcal{V}^2 := \begin{cases} 1+\gamma^2 & \text{si } \gamma \geq 0, \\ (1-\gamma)^2(1-2\gamma)\left[4 - 8\dfrac{1-2\gamma}{1-3\gamma} + \dfrac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right] & \text{si } \gamma < 0. \end{cases}$$

Définition 2.24 (*Estimateur* $\widehat{Q}^M$)**.**
Estimation de quantile extrême sur la base de l'estimateur du moment est

$$\widehat{Q}^M := X_{n-k+1:n} + \widehat{a}_n^M \frac{(k/ns)^{\widehat{\gamma}^M} - 1}{\widehat{\gamma}^M}, \quad \text{pour } k < n, \tag{2.59}$$

avec le choix

$$\widehat{a}_n^M = \frac{M_n^{(r)}}{\rho_1(\widehat{\gamma}^M)} X_{n-k:n}, \quad \rho_1(\gamma) = \begin{cases} 1, & \text{pour } \gamma \geq 0, \\ \dfrac{1}{1-\gamma}, & \text{pour } \gamma < 0. \end{cases}$$

La normalité asymptotique de l'estimateur (2.59) pour différentes conditions sur la queue de distribution et sur l'ordre de limitation de $s = s_n$ pour $n \to \infty$ est prouvée par Dekkers et al. [39] et de Haan et Rootzén [71].

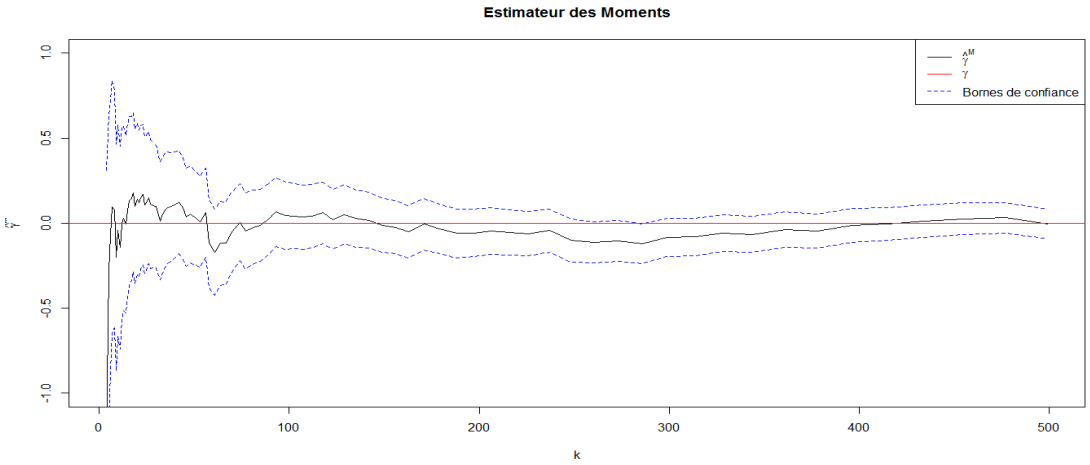

FIG. 2.10. Estimateur des Moments et l'intervalle de confiance de niveau 95%, pour l'IVE de la loi de Gumbel ($\gamma = 0$) basés sur 1000 échantillons de taille 5000.

2.5.4 Choix du Nombre k de Statistiques d'Ordre

Il est évident que les estimateurs de γ sont basées essentiellement sur le nombre de k plus grandes observations qui sont utilisées dans l'estimation où leurs comportements sont affectés par ce nombre crucial k. Cependant, le choix de ce nombre crucial k n'est pas une tâche facile, où le choix de k plus grandes observations, on se trouve en face à l'un des deux cas ou problèmes :

- Si k est très grand (autrement dit, l'utilisation un trop grand nombre d'observations dans la procédure d'estimation), entraîne à un biais grand.
- Si k est trop petit la variance d'estimation devient trop important.

Pour le choix le plus approprié de k qui permet de donner un équilibre entre les deux cas c'est-à-dire, qui permet d'obtenir une bonne estimation de γ, il y a plusieurs méthodes graphiques et numériques pour déterminer la valeur optimale de k, parmi lesquelles on peut citer l'algorithme de Reiss et Thomas [111], qui consiste à minimiser la quantité

$$\frac{1}{k}\sum_{i=1}^{k} i^{\beta}\left|\widehat{\gamma}^{H}(i) - med(\widehat{\gamma}^{H}(1),...,\widehat{\gamma}^{H}(k)\right|,\ 0 \leq \beta < 1/2.$$

2.6 Estimation de l'IVE avec Censure

Des techniques statistiques pour analyser les ensembles de données censurées sont étudiées très bien maintenant, mais elles concernent principalement des caractéristiques centrales de la distribution sous-jacente. On va s'intéresser dans cette

Section au problème de l'estimation de l'IVE et cela en présence de données censurées aléatoirement à droite. Ce problème est très récent dans la littérature, au meilleur de notre connaissance, les premiers qui ont mentionné le sujet sont Beirlant et al. [13] au Section 2.7 et Reiss et Thomas [111] au Section 6.1, mais sans résultats asymptotiques. Puis certains estimateurs des paramètres de la queue ont été proposées par Beirlant et Guillou [10] pour les données tronquées et étendu à la censure aléatoire à droite par Beirlant et al. [11] et l'année suivante par Einmahl et al. [44].

Dans le cas de censure, on suppose disposer de deux échantillons $X_1, ..., X_n$ et $Y_1, ..., Y_n$, ces deux échantillons sont formés de va's iid de loi F et G respectivement et que $F \in \mathcal{D}(\mathcal{H}_{\gamma_1})$ et $G \in \mathcal{D}(\mathcal{H}_{\gamma_2})$ pour certains $\gamma_1,\ \gamma_2 \in \mathbb{R}$. Soit $\{(Z_j, \delta_j)\,,\ 1 \leq j \leq n\}$ l'échantillon réellement observé défini par (1.16) . Il est clair que les $Z'_j s$ sont des variables indépendantes de loi H liée à F et G par la relation (1.22) . L'IVE de H la fdr de Z, existe et il est notée par γ où $\gamma := \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$. Soit x_F, x_G et x_H les points terminaux du support de F, G et H respectivement. Einmahl et al. [44] ont fourni une adaptation générale des estimateurs existants de l'IVE dans les cas suivants :

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

Le premier point important qui devrait être mentionné est le fait que tous les estimateurs précédents (Hill, Moment, ...) ne sont pas évidemment cohérentes si elles sont basées sur l'échantillon $Z_1, ..., Z_n$, autrement dit, si la censure n'est pas pris en compte. Leurs estimateurs sont basés sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dans le plus grand k de $Z's$:

$$\widehat{\gamma}_1^{(\bullet,c)} = \widehat{\gamma}_1^{(\bullet,c)}(k) := \frac{\widehat{\gamma}^{\bullet}}{\widehat{p}}, \quad \text{où} \quad \widehat{p} = \widehat{p}(k) := \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]}, \tag{2.60}$$

$\widehat{\gamma}^{\bullet}$ peut être n'importe quel estimateur non adapté à la censure, en particulier, $\widehat{\gamma}^{H}$, $\widehat{\gamma}^{M}$, ... et $\widehat{p}$ l'estimateur de la proportion de données observées dans la queue à droit de distribution, avec $k = k_n$ satisfaisant (2.46). Beirlant et al. [11] sont les premiers qui ont introduit cette méthodologie dans le cas d'estimateurs de Hill et de Moment. En plus, ils ont proposé les estimateurs des quantiles extrêmes et ont discuté leurs propriétés asymptotiques lorsque les données sont censurées par un seuil déterministe. Einmahl et al. [44] ont adapté différents estimateurs de l'IVE au cas où les données sont censurées par un seuil aléatoire et proposé une méthode unifiée pour établir leur normalité asymptotique.

Einmahl et al. [44] ont prouvé que, si $\widehat{\gamma}^{\bullet}$ un estimateur consistant et asymptotiquement normal de γ et $\widehat{p}$ un estimateur consistant et asymptotiquement normal de $p := \gamma/\gamma_1$, alors $\widehat{\gamma}_1^{(\bullet,c)}$ est un estimateur consistant et asymptotiquement normal de γ_1. Plus récemment, Brahimi et al. [18] ont également établi la consistance de $\widehat{p}$ sous la condition du premier ordre sur les fdr's F et G. Ils ont aussi conclu, que l'estimateur $\widehat{\gamma}_1^{(\bullet,c)}$ consistant de γ_1 pour l'estimateur de Hill. En outre, Brahimi et al. [18] ont utilisé la théorie des processus empiriques pour approcher l'estimateur de Hill adaptée en termes de processus Gaussiens. Ndao et al. [97, 98] ont adressée l'estimation non paramétrique de l'IVE conditionnel et son quantile pour les distributions à queue lourde qui ont été récemment généralisées par Stupfler [123] pour les trois domaines d'attraction des extrêmes. Dans le même contexte, Worms et Worms [136] ont présenté une nouvelle approche, basée sur l'intégration de Kaplan-Meier. Cette approche de définir un estimateur pour l'indice de queue positif et prouver sa consistance. Récemment, Beirlant et al. [6] ont utilisé un modèle de type Pareto censuré pour débiaiser l'estimateur de l'IVE et la queue des probabilités.

Le principal estimateur de quantile extrême $Q(1-s)$ sous censure aléatoire disponible dans la littérature a été proposé par Beirlant et al. en 2007 et par Einmahl et al. en 2008. Il est donné par la définition suivante :

Définition 2.25 (*Estimation du quantile extrême sous censure aléatoire*)**.**
L'estimation du quantile extrême sous censure aléatoire est

$$\widehat{Q}^{(\bullet,c)} := Z_{n-k:n} + \widehat{a}^{(\bullet,c)} \frac{\left(\left(1-\widehat{F}_n(Z_{n-k:n})\right)/s\right)^{\widehat{\gamma}_1^{(\bullet,c)}} - 1}{\widehat{\gamma}_1^{(\bullet,c)}},$$

où $\widehat{a}^{(\bullet,c)} = Z_{n-k:n} M_n^{(1)} (1-T_n)/\widehat{p}$, *avec* $M_n^{(1)}$ *et* T_n *sont défini dans* (2.57).

Exemple 2.8 (*Hill adapté*)**.**
Pour voir la performance de l'estimateur de Hill adapté $\widehat{\gamma}_1^{(H,c)} = \widehat{\gamma}^H/\widehat{p}$, *on a réalisé une étude de simulation basée sur* 1000 *échantillons de la loi de Burr de paramètre* γ_1 *censurées par une autre variable de Burr de paramètre* $\gamma_2 = p\gamma_1/(1-p)$, *où* p *représente la proportion de données observées dans la queue à droit de distribution. Les résultats numériques et graphiques finaux sont obtenus en faisant les moyennes sur les* 1000 *réplications. Les Figures 2.11 et 2.12 représentent l'estimateur de Hill adapté de* γ_1 *et l'estimateur de* p *en fonctions de* k. *Pour les résultats numériques, on commence par déterminer le nombre optimal d'observations extrêmes utilisées dans le calcul de* $\widehat{\gamma}_1^{(H,c)}$. *Pour cela, on applique l'algorithme de Reiss et Thomas (voir la Subsection 2.5.4). Les résultats obtenus de* $\widehat{\gamma}_1^{(H,c)}$, *biais abs. et mse (sont déjà définis dans l'Exemple 2.7) sont résumés dans la Table 2.7 avec différents*

choix de l'indice γ_1 et du pourcentage p. Pour faciliter la lecture de ce tableau, on remarque que l'estimation de $\widehat{\gamma}_1^{(H,c)}$ est meilleure pour un grande valeur de p.

$p = 0.4$						
$\gamma_1 = 0.3$				$\gamma_1 = 0.8$		
n	$\widehat{\gamma}_1^{(H,c)}$	*biais abs.*	*mse*	$\widehat{\gamma}_1^{(H,c)}$	*biais abs.*	*mse*
1000	0.3677	0.0677	0.0261	0.8501	0.0501	0.1178
3000	0.3438	0.0438	0.0064	0.8164	0.0164	0.0245
5000	0.3440	0.0440	0.0044	0.8150	0.0150	0.0137
$p = 0.9$						
$\gamma_1 = 0.3$				$\gamma_1 = 0.8$		
n	$\widehat{\gamma}_1^{(H,c)}$	*biais abs.*	*mse*	$\widehat{\gamma}_1^{(H,c)}$	*biais abs.*	*mse*
1000	0.3015	0.0015	0.0033	0.7846	0.0154	0.0247
3000	0.3024	0.0024	0.0011	0.7878	0.0122	0.0071
5000	0.3034	0.0034	0.0006	0.7967	0.0033	0.0041

TAB. 2.7. Biais et mse de l'estimation de γ_1, basée sur 1000 échantillons de la loi de Burr de paramètre γ_1 censurée par une variable de Burr de paramètre γ_2

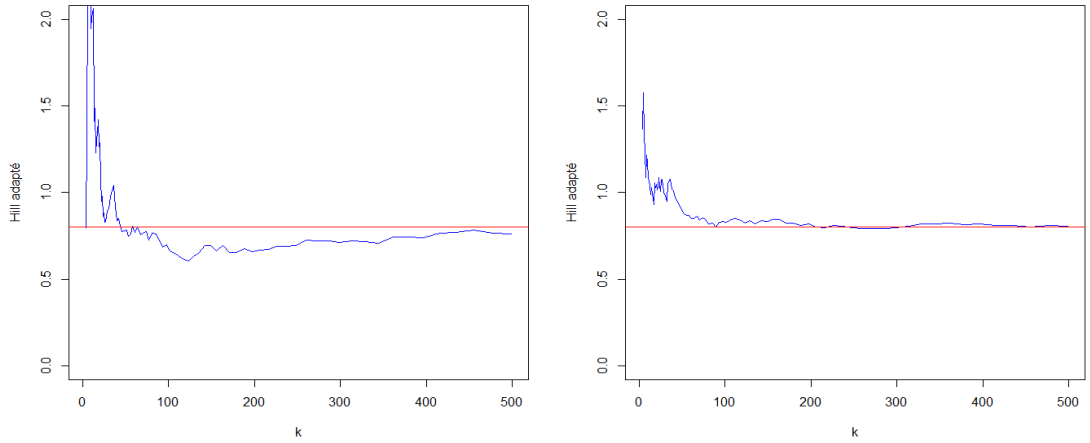

FIG. 2.11. Estimateur de Hill adapté de l'IVE, basé sur 1000 échantillons de taille 5000 de loi de Burr$(\gamma_1, 1/4)$ censurée par une autre variable de Burr$(\gamma_2, 1/4)$ avec $p = 0.4$ (gauche) et $p = 0.9$ (droite). La ligne horizontale représente la vraie valeur de $\gamma_1 = 0.8$.

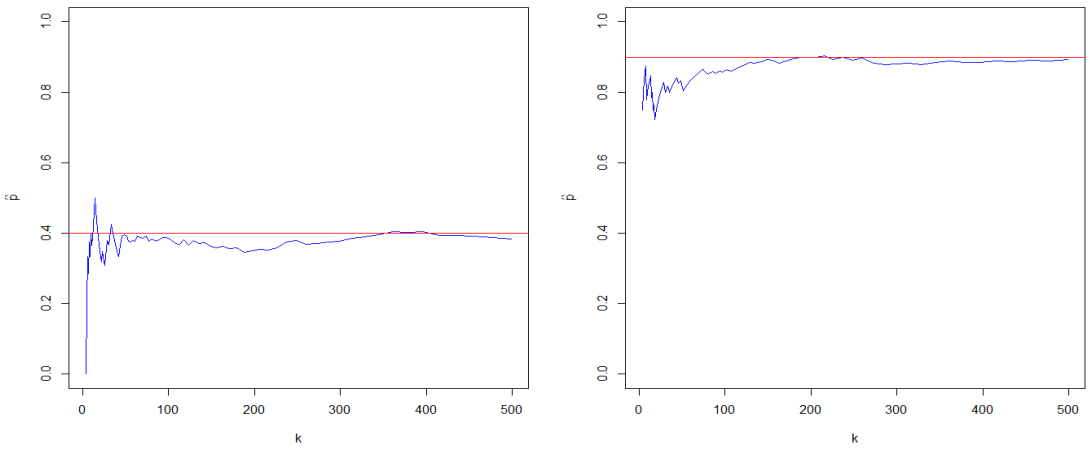

FIG. 2.12. observées dans la queue à droite de distribution, basé sur 1000 échantillons de taille 5000 de loi de Burr de paramètre $\gamma_1 = 0.8$ censurée par une variable de Burr de paramètre γ_2 avec $p = 0.4$ (gauche) et $p = 0.9$ (droite). La ligne horizontale représente la vraie valeur de p.

Chapitre 3

Statistical estimate of the proportional hazard premium of loss under random censoring

Louiza Soltane[1], Djamel Meraghni, Abdelhakim Necir

Contents

Plusieurs principes de calcul de prime d'assurance sont définis et diverses procédures d'estimation sont introduites dans la littérature. Dans ce Chapitre, on se concentre sur l'estimation de la prime de réassurance en excédent de sinistres lorsque les risques sont aléatoirement censurés à droite. La normalité asymptotique de l'estimateur proposé est établie sous des conditions adéquates et sa performance évaluée à travers des ensembles de données simulées. Les résultats présentés ici sont disponibles dans un article *publié*.

Abstract

Many insurance premium principles are defined and various estimation procedures introduced in the literature. In this Chapter, we focus on the estimation of the excess-of-loss reinsurance premium when the risks are randomly right-censored. The asymptotic normality of the proposed estimator is established under suitable conditions and its performance evaluated through sets of simulated data.

Keywords : Heavy tails ; Hill estimator ; Kaplan-Meier estimator ; Proportional hazard premium ; Random censoring ; Reinsurance treaty.

AMS 2010 Subject Classification 62G32, 62N01, 91B30, 62P05.

[1]Soltane, L., Meraghni, J. , and Necir, A., (2016). Statistical estimate of the proportional hazard premium of loss under random censoring. *Journal Afrika Statistika,* **11** (1), 883-899.

3.1 Introduction

Let $X_1, ..., X_n$ be $n \geq 1$ independent copies of a non-negative random variable (rv) X, defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with continuous cumulative distribution function (cdf) F. An independent sequence of independent rv's $Y_1, ..., Y_n$ with continuous cdf G censor them to the right, so that at each stage j we only can observe $Z_j := \min(X_j, Y_j)$ and the variable $\delta_j := \mathbb{1}\{X_j \leq Y_j\}$ (with $\mathbb{1}\{\cdot\}$ denoting the indicator function) informing whether or not there has been censorship. This model is very useful in a variety of areas where random censoring is very likely to occur such as in biostatistics, medical research, reliability analysis, actuarial science,... For more general censoring schemes and other issues involving censored data, we refer, for instance, to Cox and Oakes [28], Kalbfleisch and Prentice [82] and Gill [60].

In insurance, the worst scenarios are those caused by extreme events such as natural catastrophes, human-made disasters and financial crashes. These events increase the bill of insurance and reinsurance companies. A typical requirement for actuaries is the determination of adequate premiums for such risks. Usually, the insurer's claims data do not correspond to the underlying losses, because they are censored from above, since the insurer stipulates an upper limit to the amount to be paid out and the reinsurer covers the excess over this fixed threshold. This kind of reinsurance is called excess-of-loss reinsurance (see, e.g., Rolski et al. [116] and Embrechts et al. [46]) and the upper limit has distinct designations that are specific to each insurance type. For instance, in life insurance, it is called the cedent's company retention level while in non-life insurance, it is called the deductible, where the losses should be treated separately. For a discussion on the occurrence of right-random censorship in the area of insurance, one refers to Denuit et al. [41] in which a study on the allocated loss adjustment expenses (ALAE's) is given.

Let us assume that both F and G are heavy-tailed, that is there exist two constants $\gamma_1 > 0$ and $\gamma_2 > 0$, called tail indices or extreme value indices (EVI's), such that

$$\overline{F}(z) \sim z^{-1/\gamma_1}\ell_1(z) \text{ and } \overline{G}(z) \sim z^{-1/\gamma_2}\ell_2(z), \text{ as } z \to \infty, \tag{3.1}$$

where ℓ_1 and ℓ_2 are slowly varying functions at infinity, i.e. $\lim_{z\to\infty} \ell_i(xz)/\ell_i(z) = 1$ for every $x > 0$, $i = 1, 2$. Throughout the paper, we use the notation $\overline{\mathcal{S}}(x) := \mathcal{S}(\infty) - \mathcal{S}(x)$, for any function $\mathcal{S}(x)$ of $x > 0$. If relations (3.1) hold, then we have, for any $x > 0$

$$\lim_{z\to\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma_1} \text{ and } \lim_{z\to\infty} \frac{\overline{G}(xz)}{\overline{G}(z)} = x^{-1/\gamma_2}, \tag{3.2}$$

and we say that $\overline{F}$ and $\overline{G}$ are regularly varying at infinity as well, with respective tail indices $-1/\gamma_1$ and $-1/\gamma_2$, which we denote by $\overline{F} \in \mathcal{RV}_{-1/\gamma_1}$ and $\overline{G} \in \mathcal{RV}_{-1/\gamma_2}$. Note that, in virtue of the independence of X and Y, the cdf of the observed $Z's$,

that we denote by H, is also heavy-tailed and we have $H \in \mathcal{RV}_{-1/\gamma}$ with $\gamma := \gamma_1\gamma_2/(\gamma_1+\gamma_2)$. This class of distributions, which includes models such as Pareto, Burr, Fréchet, Lévy-stable and log-gamma, plays a prominent role in extreme value theory. Also known as Pareto-type or Pareto-like distributions, these models have important practical applications and are used rather systematically in certain branches of non-life insurance as well as in finance, telecommunications, geology and many other fields (see e.g. Resnick [113]). The analysis of extreme values of randomly censored data is a new research topic to which Reiss and Thomas [111] made a very brief reference, in Section 6.1, as a first step but with no asymptotic results. In the last decade, several authors started to be interested in the estimation of the tail index along with large quantiles under random censoring as one can see in Gomes and Oliveira [67], Beirlant et al. [11], Einmahl et al. [44] and Worms and Worms [136]. Gomes and Neves [66] also made a contribution to this field by providing a detailed simulation study and applying the estimation procedures on some survival data sets. Let $\{(Z_i, \delta_i)\,,\ 1 \le i \le n\}$ be a sample from the couple of rv's (Z, δ) and $Z_{1:n} \le ... \le Z_{n:n}$ the order statistics pertaining to $(Z_1, ..., Z_n)$. If we denote the concomitant of the ith order statistic by $\delta_{[i:n]}$ (i.e. $\delta_{[i:n]} = \delta_j$ if $Z_{i:n} = Z_j$), then Hill's estimator of γ_1 adapted to censored data is defined as $\widehat{\gamma}_1^{(H,c)} := \widehat{\gamma}^H/\widehat{p}$, where $\widehat{\gamma}^H := k^{-1}\sum_{i=1}^{k} \log\left(Z_{n-i+1:n}/Z_{n-k:n}\right)$ represents Hill's estimator (Hill, [76]) of γ, with $k = k_n$ being an integer sequence satisfying

$$1 < k < n,\ k \to \infty \text{ and } k/n \to 0 \text{ as } n \to \infty, \tag{3.3}$$

and $\widehat{p} := k^{-1}\sum_{i=1}^{k} \delta_{[n-i+1:n]}$ being the proportion of upper non-censored observations. Einmahl et al. [44] established the asymptotic normality of $\widehat{\gamma}_1^{(H,c)}$ by assuming that cdf's are absolutely continuous. Recently, Brahimi et al. [18] proved that $\widehat{p}$ consistently estimates $p := \gamma_2/(\gamma_1+\gamma_2)$ leading to the consistency of $\widehat{\gamma}_1^{(H,c)}$. They also established the asymptotic normality of $\widehat{\gamma}_1^{(H,c)}$ by adopting an approach that is different from that of Einmahl et al. [44].

In the excess-of-loss reinsurance treaty, the ceding company covers claims which do not exceed a (high) number $R \ge 0$ (called retention level), while the reinsurer pays the part $(X_i - R)_+ := \max(0, X_i - R)$ of each claim beyond R. Applying Wang's premium calculation principle (Wang [128]), with a distortion function equal to $x^{1/\varrho}$, one defines what is called the proportional hazard premium (PHP), where $\varrho \ge 1$ represents the distortion parameter or the risk aversion index. Then, the PHP of loss for the layer from R to infinity is defined as follows :

$$\Pi_\varrho(R) := \int_R^\infty \left(\overline{F}(x)\right)^{1/\varrho} dx,$$

which may be rewritten into

$$\Pi_{\varrho}(R) = R(\overline{F}(R))^{1/\varrho} \int_1^{\infty} \left(\frac{\overline{F}(Rx)}{\overline{F}(R)} \right)^{1/\varrho} dx.$$

By using the well-known Karamata theorem (see, for instance, de Haan and Ferreira, [69], page 363), we get

$$\Pi_{\varrho}(R) \sim \frac{\varrho}{1/\gamma_1 - \varrho} R \left(\overline{F}(R) \right)^{1/\varrho}, \ 0 < \gamma_1 < 1/\varrho,$$

for large R. Since $\overline{F} \in \mathcal{RV}_{-1/\gamma_1}$, then $\overline{F}(x) \sim \overline{F}(h)(x/h)^{-1/\gamma_1}$ as $x \to \infty$, where $h = h_n := H^{-1}(1 - k/n)$ with $H^{-1}(y) := \inf\{x : H(x) \geq y\}, 0 < y < 1$, denoting the quantile function pertaining to H. This leads us to derive a Weissman-type estimator (see Weissman, [130]) for the distribution tail $\overline{F}$ for censored data as follows :

$$\widehat{\overline{F}}(x) = \left(\frac{x}{Z_{n-k:n}} \right)^{-1/\widehat{\gamma}_1^{(H,c)}} \widehat{\overline{F}}_n(Z_{n-k:n}).$$

In the context of randomly right censored observations, the nonparametric maximum likelihood estimator of F is given by Kaplan and Meier [83] as the product limit estimator

$$\widehat{\overline{F}}_n(x) := \prod_{Z_{i:n} \leq x} \left(1 - \frac{\delta_{[i:n]}}{n-i+1} \right) = \prod_{Z_{i:n} \leq x} \left(\frac{n-i}{n-i+1} \right)^{\delta_{[i:n]}}, \text{ for } x < Z_{n:n},$$

which gives $\widehat{\overline{F}}_n(Z_{n-k:n}) = \prod_{i=1}^{n-k} \left(1 - \frac{\delta_{[i:n]}}{n-i+1} \right)$. Thus, the distribution tail estimator is of the form

$$\widehat{\overline{F}}(x) := \left(\frac{x}{Z_{n-k:n}} \right)^{-1/\widehat{\gamma}_1^{(H,c)}} \prod_{i=1}^{n-k} \left(1 - \frac{\delta_{[i:n]}}{n-i+1} \right),$$

and consequently, we define the PHP estimator as follows :

$$\widehat{\Pi}_{\varrho}(R) := \frac{\varrho R}{1/\widehat{\gamma}_1^{(H,c)} - \varrho} \left(\frac{R}{Z_{n-k:n}} \right)^{-1/\left(\varrho \widehat{\gamma}_1^{(H,c)}\right)} \prod_{i=1}^{n-k} \left(1 - \frac{\delta_{[i:n]}}{n-i+1} \right)^{1/\varrho}.$$

The outline of this chapter is as follows. In Section 3.2, we state our main result that consists in the asymptotic normality of the newly proposed estimator $\widehat{\Pi}_{\varrho}(R)$, which we prove in Section 3.4. In Section 3.3, we carry out a simulation study to illustrate its finite sample behavior. Finally, some results, that are instrumental to our needs, are gathered in the Appendix.

3.2 Main Results

It is well-known that the asymptotic normality of extreme value theory based estimators is adequately achieved within the second-order framework (see de Haan and Stadtmüller, [72]). Thus, it seems quite natural to suppose that cdf's F and G satisfy the well-known second-order condition of regular variation. That is, we assume that there exist two constants $\rho_j \leq 0$ (called second-order parameters) and two functions A_j, $j = 1, 2$, tending to zero and not changing sign near infinity, such that for any $x > 0$

$$\begin{aligned} \lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma_1}}{A_1(t)} &= x^{-1/\gamma_1} \frac{x^{\rho_1/\gamma_1} - 1}{\gamma_1\rho_1}, \\ \lim_{t\to\infty} \frac{\overline{G}(tx)/\overline{G}(t) - x^{-1/\gamma_2}}{A_2(t)} &= x^{-1/\gamma_2} \frac{x^{\rho_2/\gamma_2} - 1}{\gamma_2\rho_2}. \end{aligned} \tag{3.4}$$

Theorem 3.1. *Assume that the second-order conditions of regular variation* (3.4) *hold, with* $0 < \gamma_1 < 1/\varrho$ *and let* $k = k_n$ *be an integer sequence satisfying, in addition to* (3.3), $\sqrt{k}A_1(h) \to \lambda_1$. *Assume further that* $R/h \to 1$. *Then*

$$\sqrt{k}\frac{\widehat{\Pi}_\varrho(R) - \Pi_\varrho(R)}{(R/h)^{-1/\varrho\gamma_1} R\left(\overline{F}(h)\right)^{1/\varrho}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu, \mathcal{V}^2\right), \text{ as } n \to \infty,$$

where

$$\mu := \frac{\varrho\lambda_1}{(1 - p\rho_1)(1 - \varrho\gamma_1)^2} + \frac{\lambda_1}{\varrho(\gamma_1 + \rho_1 + \varrho - 2)(2 - \varrho - \gamma_1)},$$

and

$$\mathcal{V}^2 := \frac{\gamma_1^2}{(1 - \varrho\gamma_1)^2}\left(p(2 - p) + \frac{\varrho(p - 1)}{(1 - \varrho\gamma_1)} + \frac{\varrho^2(1 - 2p)}{p(1 - \varrho\gamma_1)^2}\right).$$

3.3 Simulation Study

We carry out a simulation study to illustrate the performance of our estimator, through two sets of censored and censoring data, both drawn from the following Burr model. That is

$$\overline{F}(x) = \left(1 + x^{\eta/\gamma_1}\right)^{-1/\eta} \text{ and } \overline{G}(x) = \left(1 + x^{\eta/\gamma_2}\right)^{-1/\eta}, \ x \geq 0,$$

where $\gamma_1, \gamma_2 > 0$. We fix $\eta = 1/4$, we choose the values 0.10 and 0.25 for γ_1 and two distinct aversion index values $\varrho = 1.00$ and $\varrho = 1.10$. For the proportion of the really observed extreme values, we take $p = 0.40$, 0.60 and 0.80, that is, we allow the percentage of censoring in the right tail of X to be 60%, 40% and 20%. For each couple (γ_1, p), we solve the equation $p = \gamma_2/(\gamma_1 + \gamma_2)$ to get the pertaining

γ_2-value. We vary the common size n of both samples $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$, then for each size, we generate 1000 independent replicates. Our overall results are taken as the empirical means of the results obtained through the 1000 repetitions. To determine the optimal number of upper order statistics (that we denote by k^*) used in the computation of $\widehat{\gamma}_1^{(H,c)}$, we apply the algorithm of Reiss and Thomas [111], page 137. The retention level R is taken as the value of the intermediate order statistic $Z_{n-k^*:n}$. The simulation results are summarized in Table 3.1 for $\gamma_1 = 0.10$ and in Table 3.2 for $\gamma_1 = 0.25$ (where abs bias and rmse respectively stand for the absolute value of the bias and the root of the mean squared error of the estimation). On the light of these results we see that, from the point of view of the rmse, the estimation accuracy increases when the censoring percentage decreases, which seems logical. On the other hand, we note that the sample size does not have a significant effect on the estimation when the percentage of observed data is high. Moreover, the estimator performs better for the smaller value of the distortion parameter ϱ.

$p = 0.40$								
ϱ	1.00				1.10			
n	$\Pi_\varrho(R)$	$\widehat{\Pi}_\varrho(R)$	abs bias	rmse	$\Pi_\varrho(R)$	$\widehat{\Pi}_\varrho(R)$	abs bias	rmse
500	0.0175	0.0274	0.0099	0.1372	0.0237	0.0429	0.0192	0.2012
1000	0.0174	0.0197	0.0023	0.0616	0.0236	0.0339	0.0102	0.0863
1500	0.0170	0.0158	0.0012	0.0146	0.0233	0.0233	0.0000	0.0203
$p = 0.60$								
500	0.0097	0.0066	0.0032	0.0135	0.0142	0.0103	0.0039	0.0145
1000	0.0095	0.0037	0.0058	0.0065	0.0138	0.0063	0.0076	0.0091
1500	0.0095	0.0029	0.0066	0.0069	0.0137	0.0045	0.0092	0.0098
$p = 0.80$								
500	0.0062	0.0014	0.0048	0.0049	0.0093	0.0026	0.0067	0.0074
1000	0.0062	0.0008	0.0054	0.0055	0.0092	0.0014	0.0077	0.0078
1500	0.0060	0.0006	0.0054	0.0054	0.0090	0.0010	0.0079	0.0079

TAB. 3.1. PHP estimates based on 1000 right-censored samples of size n from Burr model with tail index $\gamma_1 = 0.10$.

3.4 Proof

Before we start the proof of the theorem, let us give a brief introduction on some uniform empirical processes under random censoring. To this end, we define the

$p = 0.40$								
ϱ	1.00				1.10			
n	$\Pi_\varrho(R)$	$\widehat{\Pi}_\varrho(R)$	abs bias	rmse	$\Pi_\varrho(R)$	$\widehat{\Pi}_\varrho(R)$	abs bias	rmse
500	0.0265	0.0767	0.0501	0.3604	0.0410	0.1148	0.0738	0.8875
1000	0.0266	0.0633	0.0368	0.1602	0.0411	0.1134	0.0723	0.4842
1500	0.0266	0.0462	0.0196	0.0664	0.0409	0.0632	0.0223	0.0941
$p = 0.60$								
500	0.0196	0.0229	0.0034	0.0965	0.0310	0.0222	0.0088	0.4596
1000	0.0197	0.0119	0.0078	0.0133	0.0317	0.0203	0.0114	0.0236
1500	0.0199	0.0093	0.0106	0.0140	0.0316	0.0152	0.0164	0.0196
$p = 0.80$								
500	0.0153	0.0056	0.0097	0.0118	0.0251	0.0091	0.0160	0.0178
1000	0.0154	0.0030	0.0125	0.0127	0.0254	0.0054	0.0200	0.0204
1500	0.0157	0.0020	0.0136	0.0137	0.0254	0.0040	0.0214	0.0215

TAB. 3.2. PHP estimates based on 1000 right-censored samples of size n from Burr model with tail index $\gamma_1 = 0.25$.

functions

$$H^{(j)}(v) := \mathbb{P}(Z \leq v,\ \delta = j),\ j = 0, 1;\ v \geq 0,$$

which have a prominent role to play in the random censorship setting. Their empirical counterparts are defined by

$$H_n^{(j)}(v) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i \leq v, \delta_i = j),\ j = 0, 1;\ v \geq 0.$$

In the sequel, we will use the following two empirical processes

$$\sqrt{n}\left(\overline{H}_n^{(j)}(v) - \overline{H}^{(j)}(v)\right),\ j = 0, 1;\ v \geq 0,$$

which may be represented, almost surely, by a uniform empirical process. Indeed, let us define, for each $i = 1, ..., n$ with $\theta := H^{(1)}(\infty)$, the following rv

$$U_i := \delta_i H^{(1)}(Z_i) + (1 - \delta_i)(\theta + H^{(0)}(Z_i)).$$

From Einmahl and Koning [45], the rv's $U_1, ..., U_n$ are independent and identically distributed according to the $(0, 1)$-uniform law. The empirical cdf and the uniform empirical process based upon $U_1, ..., U_n$ are respectively denoted by

$$\mathbb{U}_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(U_i \leq s) \text{ and } \alpha_n(s) := \sqrt{n}(\mathbb{U}_n(s) - s),\ 0 \leq s \leq 1.$$

Deheuvels and Einmahl [35] state that almost surely

$$H_n^{(0)}(v) = \mathbb{U}_n(H^{(0)}(v) + \theta) - \mathbb{U}_n(\theta), \text{ for } 0 < H^{(0)}(v) < 1 - \theta,$$

and

$$H_n^{(1)}(v) = \mathbb{U}_n(H^{(1)}(v)), \text{ for } 0 < H^{(1)}(v) < \theta.$$

It is easy to verify that we almost surely have

$$\sqrt{n}\left(\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)\right) = \alpha_n(\theta) - \alpha_n\left(\theta - \overline{H}^{(1)}(v)\right), \text{ for } 0 < \overline{H}^{(1)}(v) < \theta, \tag{3.5}$$

and

$$\sqrt{n}\left(\overline{H}_n^{(0)}(v) - \overline{H}^{(0)}(v)\right) = -\alpha_n\left(1 - \overline{H}^{(0)}(v)\right), \text{ for } 0 < \overline{H}^{(0)}(v) < 1 - \theta. \tag{3.6}$$

Our methodology strongly relies on the well-known Gaussian approximation given in Proposition 3.1. For our needs, we use the following form :

$$\sup_{1/n \le s \le 1} \frac{n^{\zeta}\left|\alpha_n(1-s) - B_n(1-s)\right|}{s^{1/2-\zeta}} = O_{\mathbb{P}}(1). \tag{3.7}$$

For the increments $\alpha_n(\theta) - \alpha_n(\theta - s)$, we will need an approximation of the same type as (3.7). Following similar arguments, mutatis mutandis, as those used to in the proof of assertions (2.2) of Theorem 2.1 and (2.8) of Theorem 2.2 in Csörgő et al. [30], we may show that, for every $0 < \theta < 1$ and $0 \le \zeta < 1/4$, we have

$$\sup_{1/n \le s \le \theta} \frac{n^{\zeta}\left|\{\alpha_n(\theta) - \alpha_n(\theta - s)\} - \{B_n(\theta) - B_n(\theta - s)\}\right|}{s^{1/2-\zeta}} = O_{\mathbb{P}}(1). \tag{3.8}$$

The following Gaussian processes will be crucial to our needs :

$$\mathbf{B}_n(v) := B_n(\theta) - B_n\left(\theta - \overline{H}^{(1)}(v)\right), \text{ for } 0 < \overline{H}^{(1)}(v) < \theta, \tag{3.9}$$

and

$$\mathbf{B}_n^*(v) := \mathbf{B}_n(v) - B_n\left(1 - \overline{H}^{(0)}(v)\right), \text{ for } 0 < \overline{H}^{(0)}(v) < 1 - \theta. \tag{3.10}$$

Proof of Theorem 3.1. In the sequel, for two sequences of rv's, we write $V_n^{(1)} = o_{\mathbb{P}}\left(V_n^{(2)}\right)$ and $V_n^{(1)} \approx V_n^{(2)}$, as $n \to \infty$, to say that $V_n^{(1)}/V_n^{(2)} \to 0$ in probability and $V_n^{(1)} = V_n^{(2)}(1 + o_{\mathbb{P}}(1))$ respectively. With the premium

$$\Pi_{\varrho}(R) = R(\overline{F}(R))^{1/\varrho} \int_1^{\infty} \left(\frac{\overline{F}(Rx)}{\overline{F}(R)}\right)^{1/\varrho} dx,$$

and its estimator

$$\widehat{\Pi}_{\varrho}(R)=\frac{\varrho R}{1/\widehat{\gamma}_1^{(H,c)}-\varrho}\left(\frac{R}{Z_{n-k:n}}\right)^{-1/\left(\varrho\widehat{\gamma}_1^{(H,c)}\right)}\left(\widehat{\overline{F}}_n(Z_{n-k:n})\right)^{1/\varrho},$$

it is easy to verify that

$$\sqrt{k}\frac{\widehat{\Pi}_{\varrho}(R)-\Pi_{\varrho}(R)}{(R/h)^{-1/(\varrho\gamma_1)}\,R\left(\overline{F}\,(h)\right)^{1/\varrho}}=\sum_{i=1}^{5}S_{ni},$$

where

$$S_{n1}:=\frac{\varrho}{1/\widehat{\gamma}_1^{(H,c)}-\varrho}\left(\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\right)^{1/\varrho}\left(\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})}\right)^{1/\varrho}$$
$$\times\sqrt{k}\left\{\left(\frac{(R/Z_{n-k:n})^{-1/\widehat{\gamma}_1^{(H,c)}}}{(R/h)^{-1/\gamma_1}}\right)^{1/\varrho}-1\right\},$$

$$S_{n2}:=\left(\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\right)^{1/\varrho}\left(\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})}\right)^{1/\varrho}\sqrt{k}\left\{\frac{\varrho}{1/\widehat{\gamma}_1^{(H,c)}-\varrho}-\frac{\varrho}{1/\gamma_1-\varrho}\right\},$$

$$S_{n3}:=\frac{\varrho}{1/\gamma_1-\varrho}\left(\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\right)^{1/\varrho}\sqrt{k}\left\{\left(\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})}\right)^{1/\varrho}-1\right\},$$

$$S_{n4}:=\frac{\varrho}{1/\gamma_1-\varrho}\sqrt{k}\left\{\left(\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\right)^{1/\varrho}-1\right\},$$

and

$$S_{n5}:=\sqrt{k}\left\{\frac{\varrho}{1/\gamma_1-\varrho}-\frac{\left(\overline{F}\,(R)/\overline{F}(h)\right)^{1/\varrho}}{(R/h)^{-1/(\varrho\gamma_1)}}\int_1^{\infty}\left(\frac{\overline{F}(Rx)}{\overline{F}(R)}\right)^{1/\varrho}dx\right\}.$$

We will represent the first three terms S_{ni}, $i=1,2,3$, in terms of the Gaussian processes $\mathbf{B}_n$ and $\mathbf{B}_n^*$ and we will show that $S_{n4}\xrightarrow{\mathbb{P}}0$ while S_{n5} converges to a deterministic limit. For the first term S_{n1}, we have $\widehat{\gamma}_1^{(H,c)}\xrightarrow{\mathbb{P}}\gamma_1$ (see Brahimi et al., [18]) and $Z_{n-k:n}/h\xrightarrow{\mathbb{P}}1$, which, in view of the regular variation of $\overline{F}$, implies that $\overline{F}\,(Z_{n-k:n})/\overline{F}\,(h)\xrightarrow{\mathbb{P}}1$. Moreover, from (3.19) we have $\widehat{\overline{F}}_n\,(Z_{n-k:n})/\overline{F}\,(Z_{n-k:n})\xrightarrow{\mathbb{P}}$

1. It follows that $S_{n1} = S_{n1}^{(1)} + S_{n1}^{(2)}$, where

$$S_{n1}^{(1)} := (1 + o_{\mathbb{P}}(1)) \frac{\varrho\gamma_1}{1 - \varrho\gamma_1} \times \sqrt{k} \left\{ \left(\frac{Z_{n-k:n}}{h} \right)^{1/\left(\varrho\widehat{\gamma}_1^{(H,c)}\right)} - 1 \right\} \left(\left(\frac{R}{h} \right)^{1/\gamma_1 - 1/\widehat{\gamma}_1^{(H,c)}} \right)^{1/\varrho},$$

and

$$S_{n1}^{(2)} := (1 + o_{\mathbb{P}}(1)) \frac{\varrho\gamma_1}{1 - \varrho\gamma_1} \sqrt{k} \left\{ \left(\left(\frac{R}{h} \right)^{1/\gamma_1 - 1/\widehat{\gamma}_1^{(H,c)}} \right)^{1/\varrho} - 1 \right\}.$$

For $S_{n1}^{(1)}$, we use the mean value theorem, the consistency of $\widehat{\gamma}_1^{(H,c)}$ and the fact that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$, to have

$$S_{n1}^{(1)} = (1 + o_{\mathbb{P}}(1)) \frac{1}{1 - \varrho\gamma_1} \sqrt{k} \left(\frac{Z_{n-k:n}}{h} - 1 \right).$$

Next, we apply result (2.7) of Theorem 2.1 in Brahimi et al. [18] to get

$$S_{n1}^{(1)} = (1 + o_{\mathbb{P}}(1)) \frac{\gamma}{1 - \varrho\gamma_1} \sqrt{\frac{n}{k}} \mathbf{B}_n^*(h).$$

In view of the consistency and asymptotic normality of $\widehat{\gamma}_1^{(H,c)}$ and the assumption $R/h \to 1$, we show, by applying the mean value theorem twice, that $S_{n1}^{(2)} = o_{\mathbb{P}}(1)$. Thus, we end up with

$$S_{n1} = (1 + o_{\mathbb{P}}(1)) \frac{\gamma}{1 - \varrho\gamma_1} \sqrt{\frac{n}{k}} \mathbf{B}_n^*(h) + o_{\mathbb{P}}(1). \tag{3.11}$$

By similar arguments and using the mean value theorem once again, we easily show that

$$S_{n2} = (1 + o_{\mathbb{P}}(1)) \frac{\varrho}{(1 - \varrho\gamma_1)^2} \sqrt{k} \left(\widehat{\gamma}_1^{(H,c)} - \gamma_1 \right),$$

$$S_{n3} = (1 + o_{\mathbb{P}}(1)) \frac{\gamma_1}{1 - \varrho\gamma_1} \sqrt{k} \left(\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})} - 1 \right),$$

and

$$S_{n4} = (1 + o_{\mathbb{P}}(1)) \frac{\gamma_1}{1 - \varrho\gamma_1} \sqrt{k} \left\{ \frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)} - 1 \right\}.$$

By applying result (2.9) of Theorem 2.1 in Brahimi et al. [18] we get, after a change of variables, that

$$S_{n2} = (1+o_{\mathbb{P}}(1))\frac{\varrho}{(1-\varrho\gamma_1)^2}\left\{\frac{1}{p}\sqrt{\frac{n}{k}}\int_1^{\infty} v^{-1}\mathbf{B}_n^*(hv)\,dv - \frac{\gamma_1}{p}\sqrt{\frac{n}{k}}\mathbf{B}_n(h)\right\} \quad (3.12)$$
$$+ (1+o_{\mathbb{P}}(1))\frac{\varrho\sqrt{k}A_1(h)}{(1-p\rho_1)(1-\varrho\gamma_1)^2},$$

From Proposition 3.2, we infer that

$$S_{n3} = (1+o_{\mathbb{P}}(1))\frac{\gamma_1}{1-\varrho\gamma_1}\left(\sqrt{\frac{n}{k}}\mathbf{B}_n(h) + \sqrt{\frac{k}{n}}\Delta_n\right) + o_{\mathbb{P}}(1). \quad (3.13)$$

Now, we decompose S_{n4} into the sum of two terms

$$S_{n4}^{(1)} := (1+o_{\mathbb{P}}(1))\frac{\gamma_1}{1-\varrho\gamma_1}\sqrt{k}\left\{\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)} - \left(\frac{Z_{n-k:n}}{h}\right)^{-1/\gamma_1}\right\},$$

and

$$S_{n4}^{(2)} := (1+o_{\mathbb{P}}(1))\frac{\gamma_1}{1-\varrho\gamma_1}\sqrt{k}\left\{\left(\frac{Z_{n-k:n}}{h}\right)^{-1/\gamma_1} - 1\right\}.$$

The second-order condition (3.4) of $\overline{F}$ and the fact that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$ yield that

$$S_{n4}^{(1)} = o_{\mathbb{P}}\left(\sqrt{k}A_1(h)\right) = o_{\mathbb{P}}(1).$$

For $S_{n4}^{(2)}$, we, once again, apply the mean value theorem (with $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$) then we use result (2.7) of Theorem 2.1 in Brahimi et al. [18] to get

$$S_{n4}^{(2)} = -(1+o_{\mathbb{P}}(1))\frac{\gamma}{1-\varrho\gamma_1}\sqrt{\frac{n}{k}}\mathbf{B}_n^*(h).$$

Consequently, we have

$$S_{n4} = -(1+o_{\mathbb{P}}(1))\frac{\gamma}{1-\varrho\gamma_1}\sqrt{\frac{n}{k}}\mathbf{B}_n^*(h) + o_{\mathbb{P}}(1). \quad (3.14)$$

For the last term S_{n5}, we start by decomposing it into the sum of

$$S_{n5}^{(1)} := -\frac{\varrho\gamma_1}{1-\varrho\gamma_1}\frac{1}{(R/h)^{-1/(\varrho\gamma_1)}}\sqrt{k}\left\{\left(\frac{\overline{F}(R)}{\overline{F}(h)}\right)^{1/\varrho} - \left(\left(\frac{R}{h}\right)^{-1/\gamma_1}\right)^{1/\varrho}\right\},$$

and

$$S_{n5}^{(2)} := -\left(\frac{\left(\overline{F}\left(R\right)/\overline{F}(h)\right)}{(R/h)^{-1/\gamma_1}}\right)^{1/\varrho}\sqrt{k}\int_1^{\infty}\left(\left(\frac{\overline{F}(Rx)}{\overline{F}(R)}\right)^{1/\varrho}-\left(x^{-1/\gamma_1}\right)^{1/\varrho}\right)dx.$$

By similar arguments as those used for $S_{n4}^{(1)}$, we show that (here we use the assumption that $R/h \to 1$)

$$S_{n5}^{(1)} = o_{\mathbb{P}}\left(\sqrt{k}A_1\left(h\right)\right) = o_{\mathbb{P}}\left(1\right).$$

For $S_{n5}^{(2)}$, we first apply the mean value theorem to have

$$S_{n5}^{(2)} = -\frac{1}{\varrho}\sqrt{k}\int_1^{\infty}\left(\frac{\overline{F}(Rx)}{\overline{F}(R)}-x^{-1/\gamma_1}\right)\zeta^{1/\varrho-1}(x)dx,$$

where ζ lies between $\overline{F}(Rx)/\overline{F}(R)$ and x^{-1/γ_1}. Then we use Potter's inequalities, given in assertion 5 of Proposition B.1.9 in de Haan and Ferreira [69], to get

$$S_{n5}^{(2)} = \left(1+o\left(1\right)\right)\frac{\sqrt{k}A_1\left(h\right)}{\varrho\left(\gamma_1+\rho_1+\varrho-2\right)\left(2-\varrho-\gamma_1\right)}.$$

Thereforec

$$S_{n5} = \left(1+o\left(1\right)\right)\frac{\sqrt{k}A_1\left(h\right)}{\varrho\left(\gamma_1+\rho_1+\varrho-2\right)\left(2-\varrho-\gamma_1\right)}+o_{\mathbb{P}}\left(1\right). \tag{3.15}$$

Finally, by gathering results (3.11), (3.12), (3.13), (3.14) and (3.15), we obtain the following asymptotic representation to the premium estimator :

$$\begin{aligned}\sqrt{k}\frac{\widehat{\Pi}_{\varrho}(R)-\Pi_{\varrho}(R)}{(R/h)^{-1/\varrho\gamma_1}R\left(\overline{F}\left(h\right)\right)^{1/\varrho}} &= o_{\mathbb{P}}\left(1\right)+\frac{\gamma_1}{1-\varrho\gamma_1}\sqrt{\frac{k}{n}}\Delta_n+\frac{1}{1-\varrho\gamma_1}\sqrt{\frac{n}{k}}\Gamma_n \\ &+\left\{\frac{\varrho\sqrt{k}A_1\left(h\right)}{\left(1-p\rho_1\right)\left(1-\varrho\gamma_1\right)^2}+\frac{\sqrt{k}A_1\left(h\right)}{\varrho\left(\gamma_1+\rho_1+\varrho-2\right)\left(2-\varrho-\gamma_1\right)}\right\},\end{aligned} \tag{3.16}$$

where Δ_n is as defined in 3.18 and

$$\Gamma_n := \gamma_1\left(1-\frac{\varrho}{p\left(1-\varrho\gamma_1\right)}\right)\mathbf{B}_n\left(h\right)+\frac{\varrho}{p\left(1-\varrho\gamma_1\right)}\int_1^{\infty}v^{-1}\mathbf{B}_n^{*}\left(hv\right)dv.$$

From (3.16), we deduce that $\sqrt{k}\left(\widehat{\Pi}_{\varrho}(R)-\Pi_{\varrho}(R)\right)/\left((R/h)^{-1/\varrho\gamma_1}R\left(\overline{F}\left(h\right)\right)^{1/\varrho}\right)$ is asymptotically Gaussian with mean

$$\left\{\frac{\varrho}{\left(1-p\rho_1\right)\left(1-\varrho\gamma_1\right)^2}+\frac{1}{\rho\left(\gamma_1+\rho_1+\varrho-2\right)\left(2-\varrho-\gamma_1\right)}\right\}\lim_{n\to\infty}\sqrt{k}A_1\left(h\right)=\mu,$$

and variance

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{\gamma_1}{1-\varrho\gamma_1}\sqrt{\frac{k}{n}}\Delta_n + \frac{1}{1-\varrho\gamma_1}\sqrt{\frac{n}{k}}\Gamma_n\right]^2.$$

Note that from the covariance structure in Csörgő [29] , page 2768, we have the following useful formulas :

$$\begin{cases} \mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n(v)\right] = \min\left(\overline{H}^{(1)}(u), \overline{H}^{(1)}(v)\right) - \overline{H}^{(1)}(u)\,\overline{H}^{(1)}(v), \\ \mathbf{E}\left[\mathbf{B}_n^*(u)\,\mathbf{B}_n^*(v)\right] = \min\left(\overline{H}(u), \overline{H}(v)\right) - \overline{H}(u)\,\overline{H}(v), \\ \mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n^*(v)\right] = \min\left(\overline{H}^{(1)}(u), \overline{H}^{(1)}(v)\right) - \overline{H}^{(1)}(u)\,\overline{H}(v). \end{cases} \tag{3.17}$$

After elementary but very tedious computations, using these formulas with l'Hôpital's rule, we get as $n \to \infty$,

$$\int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n(h)\right]}{\overline{H}^2(u)} d\overline{H}(u) \to -p, \quad \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(h)\,\mathbf{B}_n^*(u)\right]}{\overline{H}^2(u)} d\overline{H}^{(1)}(u) \to -p^2,$$

$$\int_0^h \int_1^\infty \frac{\mathbf{E}\left[\mathbf{B}_n(v)\,\mathbf{B}_n^*(hu)\right]}{u\overline{H}^2(v)} du d\overline{H}(v) \to -p\gamma,$$

$$\int_0^h \int_1^\infty \frac{\mathbf{E}\left[\mathbf{B}_n^*(v)\,\mathbf{B}_n^*(hu)\right]}{u\overline{H}^2(v)} du d\overline{H}^{(1)}(v) \to -p\gamma,$$

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n(v)\right]}{\overline{H}^2(u)\overline{H}^2(v)} d\overline{H}(u) d\overline{H}(v) \to 2p,$$

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n^*(u)\,\mathbf{B}_n^*(v)\right]}{\overline{H}^2(u)\overline{H}^2(v)} d\overline{H}^{(1)}(u) d\overline{H}^{(1)}(v) \to 2p^2,$$

and

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n^*(v)\right]}{\overline{H}^2(u)\overline{H}^2(v)} d\overline{H}(u) d\overline{H}^{(1)}(v) \to 2p^2,$$

Using the results above with some further calculations leads to σ^2. □

3.5 Appendix

The following proposition consists in Corollary 2.1 of Csörgő et al. [30]

Proposition 3.1. *There exists a probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ *with independent* $(0,1)$*-uniform rv's* U_1, $U_2, ...$ *and a sequence of Brownian bridges* $\{B_i(s);\ 0 \le s \le 1\}$ $(i = 1, 2, ...)$ *such that, for every* $0 < \lambda < \infty$, *we have as* $n \to \infty$

$$\sup_{\lambda/n \le s \le 1} \frac{n^{\zeta} \left|\alpha_n(s) - B_n(s)\right|}{s^{1/2-\zeta}} = \begin{cases} O_{\mathbb{P}}(\log n) & \text{when } \zeta = \frac{1}{4}, \\ O_{\mathbb{P}}(1) & \text{when } 0 \le \zeta < \frac{1}{4}, \end{cases}$$

$$\sup_{0 \le s \le 1-\lambda/n} \frac{n^{\zeta} \left|\alpha_n(s) - B_n(s)\right|}{(1-s)^{1/2-\zeta}} = \begin{cases} O_{\mathbb{P}}(\log n) & \text{when } \zeta = \frac{1}{4}, \\ O_{\mathbb{P}}(1) & \text{when } 0 \le \zeta < \frac{1}{4}, \end{cases}$$

and

$$\sup_{\lambda/n \le s \le 1-\lambda/n} \frac{n^{\zeta} \left|\alpha_n(s) - B_n(s)\right|}{(s(1-s))^{1/2-\zeta}} = \begin{cases} O_{\mathbb{P}}(\log n) & \text{when } \zeta = \frac{1}{4}, \\ O_{\mathbb{P}}(1) & \text{when } 0 \le \zeta < \frac{1}{4}. \end{cases}$$

Proof. See Csörgő et al. [30], page 48.

Proposition 3.2. *Assume that all second-order conditions* (3.4) *hold.Let* $k = k_n$ *bbe an integer sequence satisfying, in addition to* (3.3) $\sqrt{k} A_j(h) = O(1)$, *for* $j = 1, 2$, *as* $n \to \infty$. *Then there exists a sequence of Brownian bridges* $\{B_n(s);\ 0 \le s \le 1\}$ *such that*

$$\sqrt{k}\left\{\frac{\widehat{\overline{F}}_n\left(Z_{n-k:n}\right)}{\overline{F}\left(Z_{n-k:n}\right)} - 1\right\} = \sqrt{\frac{n}{k}}\mathbf{B}_n(h) + \sqrt{\frac{k}{n}}\Delta_n + o_{\mathbb{P}}(1),$$

where

$$\Delta_n := \int_0^h \frac{\mathbf{B}_n(v)}{\overline{H}^2(v)} d\overline{H}(v) - \int_0^h \frac{\mathbf{B}_n^*(v)}{\overline{H}^2(v)} d\overline{H}^{(1)}(v), \tag{3.18}$$

with $\mathbf{B}_n(v)$ *and* $\mathbf{B}_n^*(v)$ *respectively defined in* (3.9) *and* (3.10). *Consequently,*

$$\sqrt{k}\left\{\frac{\widehat{\overline{F}}_n\left(Z_{n-k:n}\right)}{\overline{F}\left(Z_{n-k:n}\right)} - 1\right\} \xrightarrow{d} \mathcal{N}(0, p(1-p)), \text{ as } n \to \infty, \tag{3.19}$$

Proof. In view of Proposition 5 of Csörgő [29], combined with equation (4.9) in the same reference, we have for any $x \le Z_{n-k:n}$,

$$\begin{aligned} &\frac{\widehat{\overline{F}}_n(x) - \overline{F}(x)}{\overline{F}(x)} \\ &= \int_0^x \frac{d\left(\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)\right)}{\overline{H}(v)} - \int_0^x \frac{\overline{H}_n(v) - \overline{H}(v)}{\overline{H}^2(v)} d\overline{H}^{(1)}(v) + O_{\mathbb{P}}\left(\frac{1}{k}\right). \end{aligned}$$

Upon integrating the first integral by parts, we get

$$\begin{aligned}
&\frac{\widehat{\overline{F}}_n(x)-\overline{F}(x)}{\overline{F}(x)} \\
&=-\left(\overline{H}_n^{(1)}(0)-\overline{H}^{(1)}(0)\right)+\frac{\overline{H}_n^{(1)}(x)-\overline{H}^{(1)}(x)}{\overline{H}(x)} \\
&+\int_0^x \frac{\overline{H}_n^{(1)}(v)-\overline{H}^{(1)}(v)}{\overline{H}^2(v)}d\overline{H}(v)-\int_0^x \frac{\overline{H}_n(v)-\overline{H}(v)}{\overline{H}^2(v)}d\overline{H}^{(1)}(v)+O_{\mathbb{P}}\left(\frac{1}{k}\right).
\end{aligned} \tag{3.20}$$

Recall that

$$\sqrt{n}\left(\overline{H}_n(v)-\overline{H}(v)\right)=\sqrt{n}\left(\overline{H}_n^1(v)-\overline{H}^1(v)\right)+\sqrt{n}\left(\overline{H}_n^0(v)-\overline{H}^0(v)\right),$$

which by representations (3.5) and (3.6) becomes

$$\sqrt{n}\left(\overline{H}_n(v)-\overline{H}(v)\right)=\left(\alpha_n(\theta)-\alpha_n\left(\theta-\overline{H}^{(1)}(v)\right)\right)-\alpha_n\left(1-\overline{H}^{(0)}(v)\right).$$

On the other hand, by the classical central limit theorem, we have $\overline{H}_n^{(1)}(0)-\overline{H}^{(1)}(0)=O_{\mathbb{P}}\left(n^{-1/2}\right)$. Using these results in (3.20) and then multiplying by $\sqrt{k}$, we get

$$\begin{aligned}
&\sqrt{k}\frac{\widehat{\overline{F}}_n(x)-\overline{F}(x)}{\overline{F}(x)} \\
&=O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right)+O_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right)+\sqrt{\frac{k}{n}}\frac{\alpha_n(\theta)-\alpha_n\left(\theta-\overline{H}^{(1)}(x)\right)}{\overline{H}(x)} \\
&+\sqrt{\frac{k}{n}}\int_0^x \frac{\alpha_n(\theta)-\alpha_n\left(\theta-\overline{H}^{(1)}(v)\right)}{\overline{H}^2(v)}d\overline{H}(v) \\
&-\sqrt{\frac{k}{n}}\int_0^x \frac{\alpha_n(\theta)-\alpha_n\left(\theta-\overline{H}^{(1)}(v)\right)-\alpha_n\left(1-\overline{H}^{(0)}(v)\right)}{\overline{H}^2(v)}d\overline{H}^{(1)}(v).
\end{aligned}$$

The Gaussian approximations (3.7) and (3.8), in $x=Z_{n-k:n}$, and the facts that $\sqrt{k/n}$ and $1/\sqrt{k}$ tend to zero as $n\to\infty$, lead to

$$\begin{aligned}
&\sqrt{k}\frac{\widehat{\overline{F}}_n(Z_{n-k:n})-\overline{F}(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})} \\
&=\sqrt{\frac{n}{k}}\mathbf{B}_n(Z_{n-k:n})+\sqrt{\frac{k}{n}}\int_0^{Z_{n-k:n}}\frac{\mathbf{B}_n(v)}{\overline{H}^2(v)}d\overline{H}(v)-\sqrt{\frac{k}{n}}\int_0^{Z_{n-k:n}}\frac{\mathbf{B}_n^*(v)}{\overline{H}^2(v)}d\overline{H}^{(1)}(v)+o_{\mathbb{P}}(1).
\end{aligned}$$

Applying Lemma 3.1 completes the proof. The asymptotic normality property is straightforward. For the variance computation, we use the covariance formulas (3.17) and the results at the end of Section 3.4. □

Lemma 3.1. *Assume that the second-order conditions of regular variation* (3.4) *and let* $k := k_n$ *be an integer sequence satisfying* (3.3). *Then*

$$(i)\ \sqrt{\frac{k}{n}}\int_h^{Z_{n-k:n}} \frac{\mathbf{B}_n(v)}{\overline{H}^2(v)} d\overline{H}(v) = o_{\mathbb{P}}(1).$$

$$(ii)\ \sqrt{\frac{k}{n}}\int_h^{Z_{n-k:n}} \frac{\mathbf{B}_n^*(v)}{\overline{H}^2(v)} d\overline{H}^{(1)}(v) = o_{\mathbb{P}}(1).$$

$$(iii)\ \sqrt{\frac{n}{k}}\left\{\mathbf{B}_n(Z_{n-k:n}) - \mathbf{B}_n(h)\right\} = o_{\mathbb{P}}(1)$$

$$(iv)\ \sqrt{\frac{n}{k}}\left\{\mathbf{B}_n^*(Z_{n-k:n}) - \mathbf{B}_n^*(h)\right\} = o_{\mathbb{P}}(1).$$

Proof. We begin by proving the first assertion. For fixed $0 < \eta, \varepsilon < 1$, we have

$$\mathbb{P}\left(\left|\sqrt{\frac{k}{n}}\int_h^{Z_{n-k:n}} \mathbf{B}_n(v)\frac{d\overline{H}(v)}{\overline{H}^2(v)}\right| > \eta\right)$$
$$\leq \mathbb{P}\left(\left|\frac{Z_{n-k:n}}{h} - 1\right| > \varepsilon\right) + \mathbb{P}\left(\left|\sqrt{\frac{k}{n}}\int_h^{(1+\varepsilon)h} \mathbf{B}_n(v)\frac{d\overline{H}(v)}{\overline{H}^2(v)}\right| > \eta\right).$$

It is clear that the first term the right-hand side tends to zero as $n \to \infty$. Then, it remains to show that the second one goes to zero as well. Indeed, observe that

$$\mathbf{E}\left|\sqrt{\frac{k}{n}}\int_h^{(1+\varepsilon)h} \mathbf{B}_n(v)\frac{d\overline{H}(v)}{\overline{H}^2(v)}\right| \leq -\sqrt{\frac{k}{n}}\int_h^{(1+\varepsilon)h} \mathbf{E}\left|\mathbf{B}_n(v)\right|\frac{d\overline{H}(v)}{\overline{H}^2(v)}.$$

From the first result of (3.17), we have $\mathbf{E}\left|\mathbf{B}_n(v)\right| \leq \sqrt{\overline{H}^1(v)}$. Then

$$\mathbf{E}\left|\sqrt{\frac{k}{n}}\int_h^{(1+\varepsilon)h} \mathbf{B}_n(v)\frac{d\overline{H}(v)}{\overline{H}^2(v)}\right| \leq -\sqrt{\frac{k}{n}}\int_h^{(1+\varepsilon)h} \sqrt{\overline{H}^1(v)}\frac{d\overline{H}(v)}{\overline{H}^2(v)},$$

which, in turn, is less than or equal to

$$\sqrt{\frac{k}{n}}\sqrt{\overline{H}^{(1)}(h)}\left(\frac{1}{\overline{H}((1+\varepsilon)h)} - \frac{1}{\overline{H}(h)}\right).$$

Since $\overline{H}(h) = k/n$, then this may be rewritten into

$$\sqrt{\frac{\overline{H}^{(1)}(h)}{\overline{H}(h)}}\left(\frac{\overline{H}(h)}{\overline{H}((1+\varepsilon)h)} - 1\right).$$

Since $\overline{H}^{(1)}(h) \sim p\overline{H}(h)$ and $\overline{H} \in \mathcal{RV}_{(-1/\gamma)}$, then the previous quantity tends to $p^{1/2}\left((1+\varepsilon)^{1/\gamma} - 1\right)$ as $n \to \infty$. Being arbitrary, ε may be chosen small enough so that this limit be zero. By similar arguments, we also show assertion (ii), therefore we omit the details. The last two assertions are shown following the same technique, that we use to prove (iv). Notice that, from the definition of $\mathbf{B}_n^*(v)$ and the second covariance formula in (3.17),

$$\left\{\mathbf{B}_n^*(v);\ v \geq 0\right\} \overset{d}{=} \left\{\mathcal{B}_n\left(\overline{H}(v)\right);\ v \geq 0\right\},$$

where $\{\mathcal{B}_n(s);\ 0 \leq s \leq 1\}$ is a sequence of standard Brownian bridges. Hence

$$\sqrt{\frac{n}{k}}\left\{\mathbf{B}_n^*(Z_{n-k:n}) - \mathbf{B}_n^*(h)\right\} \overset{d}{=} \sqrt{\frac{n}{k}}\left\{\mathcal{B}_n\left(\overline{H}(Z_{n-k:n})\right) - \mathcal{B}_n\left(\overline{H}(h)\right)\right\}.$$

Let $\{\mathcal{W}_n(t);\ 0 \leq s \leq 1\}$ be a sequence of standard Wiener processes such that $\mathcal{B}_n(t) = \mathcal{W}_n(t) - t\mathcal{W}_n(1)$. Then $\sqrt{n/k}\left\{\mathbf{B}_n^*(Z_{n-k:n}) - \mathbf{B}_n^*(h)\right\}$ equals in distribution to

$$\sqrt{\frac{n}{k}}\left(\left\{\mathcal{W}_n\left(\overline{H}(Z_{n-k:n})\right) - \mathcal{W}_n\left(\overline{H}(h)\right)\right\} - \left\{\overline{H}(Z_{n-k:n}) - \overline{H}(h)\right\}\mathcal{W}_n(1)\right).$$

By using the facts that $\overline{H}(h) = k/n$ and $\overline{H}(Z_{n-k:n})/\overline{H}(h) \approx 1$, we get

$$\sqrt{\frac{n}{k}}\left(\overline{H}(Z_{n-k:n}) - \overline{H}(h)\right) = \sqrt{\frac{k}{n}}\left(\frac{\overline{H}(Z_{n-k:n})}{\overline{H}(h)} - 1\right) = o_{\mathbb{P}}(1).$$

Now, we prove that

$$\vartheta_n := \sqrt{\frac{n}{k}}\left\{\mathcal{W}_n\left(\overline{H}(Z_{n-k:n})\right) - \mathcal{W}_n\left(\overline{H}(h)\right)\right\} = o_{\mathbb{P}}(1).$$

To this end, we show that $\mathbb{P}(|\vartheta_n| > \eta) \to 0$, for any fixed real number $\eta > 0$. Since $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$, then for arbitrary $\epsilon > 0$ and sufficiently large n, the probability of $A_n(\epsilon) := \{|Z_{n-k:n}/h - 1| \leq \epsilon\}$ is close to 1. Next, we use the following useful inequality : $\mathbb{P}(|\vartheta_n| > \eta) \leq \mathbb{P}\{|\vartheta_n| > \eta, A_n(\epsilon)\} + \mathbb{P}\{A_n^c(\epsilon)\}$, with $A_n^c(\epsilon)$ denoting the complement set of $A_n(\epsilon)$. It is easy to verify that ϑ_n may be rewritten into

$$\frac{\mathcal{W}_n\left(\overline{H}(h)\,\xi_n + \overline{H}(h)\right) - \mathcal{W}_n\left(\overline{H}(h)\right)}{\sqrt{\overline{H}(h)}},$$

where $\xi_n := \overline{H}\left(Z_{n-k:n}\right)/\overline{H}\left(h\right) - 1$. Thereby $\mathbb{P}\left(|\vartheta_n| > \eta\right) \leq I_n + \mathbb{P}\left\{A_n^c\left(\epsilon\right)\right\}$, where

$$I_n := \mathbb{P}\left(\sup_{0\leq t\leq \overline{H}(h)\xi_n} \left|\mathcal{W}_n\left(t + \overline{H}\left(h\right)\right) - \mathcal{W}_n\left(\overline{H}\left(h\right)\right)\right| > \eta\sqrt{\overline{H}\left(h\right)}, A_n\left(\epsilon\right)\right).$$

Since $\overline{H}$ is regularly varying at infinity, then we readily show that, in the set $A_n\left(\epsilon\right)$, we have $|\xi_n| \leq \epsilon$. Note that, for a fixed $0 \leq s \leq 1$, we have

$$\left\{\mathcal{W}_n\left(t+s\right) - \mathcal{W}_n\left(s\right);\ 0 \leq t \leq 1-s\right\} \overset{d}{=} \left\{\mathcal{W}_n\left(t\right);\ 0 \leq t \leq 1-s\right\}.$$

It follows that $I_n = \mathbb{P}\left(\sup_{0\leq t\leq \epsilon\overline{H}(h)} \left|\mathcal{W}_n\left(t\right)\right| > \eta\sqrt{\overline{H}\left(h\right)}\right)$ which, by Doob's martingale inequality, satisfies

$$I_n \leq \frac{\mathbf{E}\left|\mathcal{W}_n\left(\epsilon\overline{H}\left(h\right)\right)\right|}{\eta\sqrt{\overline{H}\left(h\right)}}.$$

Since $\mathbf{E}\left|\mathcal{W}_n\left(\epsilon\overline{H}\left(h\right)\right)\right| \leq \sqrt{\epsilon\overline{H}\left(h\right)}$ and $\mathbb{P}\left\{A_n^c\left(\epsilon\right)\right\} < \epsilon$, then $\mathbb{P}\left(|\vartheta_n| > \eta\right) \leq \eta^{-1}\epsilon^{1/2} + \epsilon$ which tends to zero as $\epsilon \downarrow 0$, as sought. □

Chapitre 4

Estimating the mean of a heavy-tailed distribution under random censoring

Louiza Soltane[1], Djamel Meraghni, Abdelhakim Necir

Contents

L'objectif principal de ce Chapitre est de proposer une méthodologie d'estimation de la moyenne d'une distribution à queue lourde en présence de censure aléatoire à droite. Les résultats présentés ici sont disponibles dans un article *soumis pour publication*.

Abstract

The central limit theorem introduced by Stute [The central limit theorem under random censorship. Ann. Statist. 1995 ; 23 : 422-439] does not hold for some class of heavy-tailed distributions. In this paper, we make use of the extreme value theory to propose an alternative estimating approach of the mean ensuring the asymptotic normality property. A simulation study is carried out to evaluate the performance of this estimation procedure and, as an application, confidence bounds to the mean of the survival time of Australian male Aids patients are provided.

Keywords : Extreme values ; Hill estimator ; Kaplan-Meier estimator ; Random censoring.

AMS 2010 Subject Classification : 62P05 ; 62H20 ; 91B26 ; 91B30.

[1]Soltane, L., Meraghni, J., and Necir, A. (2016). Estimating the mean of a heavy-tailed distribution under random censoring. *Submitted.*

4.1 Introduction

Let $X_1, ..., X_n$ be $n \geq 1$ independent copies of a non-negative random variable (rv) X, defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with cumulative distribution function (cdf) F. These rv's are censored to the right by a sequence of independent copies $Y_1, ..., Y_n$ of a non-negative rv Y, independent of X, with cdf G. At each stage $1 \leq j \leq n$, we can only observe the rv's $Z_j := \min(X_j, Y_j)$ and $\delta_j := \mathbb{1}\{X_j \leq Y_j\}$, with $\mathbb{1}\{\cdot\}$ denoting the indicator function. The latter rv indicates whether there has been censorship or not. If we denote by H the cdf of the observed $Z's$, then, by the independence of X and Y, we have $1 - H = (1 - F)(1 - G)$. Throughout this Chapter, we will use the notation $\overline{\mathcal{S}}(x) := \mathcal{S}(\infty) - \mathcal{S}(x)$, for any function $\mathcal{S}$. Assume further that F and G are heavy-tailed or, in other words, that $\overline{F}$ and $\overline{G}$ are regularly varying at infinity with negative indices $-1/\gamma_1$ and $-1/\gamma_2$ respectively. That is

$$\lim_{z\to\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma_1} \text{ and } \lim_{z\to\infty} \frac{\overline{G}(xz)}{\overline{G}(z)} = x^{-1/\gamma_2}, \tag{4.1}$$

for any $x > 0$. Consequently, H is heavy-tailed too, with tail index $\gamma := \dfrac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}$. Examples of censored data with apparent heavy tails can be found in Gomes and Neves [66]. The convergence rates of the limits (4.1) are formulated by the well-known second-order condition of regularly varying functions. In other words, there exist constants $\rho_j < 0$ and functions A_j, $j = 1, 2$ tending to zero, not changing sign near infinity and having regularly varying absolute values with indices ρ_j, such that for any $x > 0$

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma_1}}{A_1(t)} = x^{-1/\gamma_1} \frac{x^{\rho_1/\gamma_1} - 1}{\gamma_1\rho_1}, \tag{4.2}$$

and

$$\lim_{t\to\infty} \frac{\overline{G}(tx)/\overline{G}(t) - x^{-1/\gamma_2}}{A_2(t)} = x^{-1/\gamma_2} \frac{x^{\rho_2/\gamma_2} - 1}{\gamma_2\rho_2}. \tag{4.3}$$

The class of heavy-tailed distributions, satifying the second-order condition, takes a significant role in extreme value theory. It includes distributions such as Burr, Fréchet, Benktander, generalised Pareto, the log-logistic, log-gamma and α-stable $(0 < \alpha < 2)$, known to be appropriate models for fitting large insurance claims, log-returns, large fluctuations of prices, etc ... (see, e.g., Resnick, [113]).

The nonparametric maximum likelihood estimator of cdf F is given by Kaplan and Meier [83] as the product limit estimator

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

where $Z_{1:n} \leq ... \leq Z_{n:n}$ denote the order statistics pertaining to the sample $Z_1, ..., Z_n$ with the corresponding concomitants $\delta_{[1:n]}, ..., \delta_{[n:n]}$ satisfying $\delta_{[j:n]} = \delta_i$ if $Z_{j:n} = Z_i$. This estimator, known as Kaplan-Meier estimator of F, may be expressed as follows

$$\widehat{F}_n(x) := \sum_{i=2}^{n} W_{i,n} \mathbb{1}\{Z_{i:n} \leq x\}, \tag{4.4}$$

where $W_{i,n} := \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}$ (see, e.g., Reiss and Thomas, [111], page 162). The aim of this paper is to propose an asymptotically normal estimator for the mean $\mu = \mathbf{E}[X] := \int_0^\infty \overline{F}(x)dx$, whose existence requires that $\gamma_1 < 1$. By substituting $\widehat{F}_n$ for F in the previous equation, Stute [124] defined the empirical mean for censored data by

$$\widetilde{\mu}_n := \sum_{i=2}^{n} \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{i:n}. \tag{4.5}$$

and established, in Corollary 1.2, its asymptotic normality. Explicitly, the author showed that

$$\sqrt{n}\left(\widetilde{\mu} - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right) \text{ as } n \to \infty,$$

where $\sigma^2 := \mathbf{Var}\left[Z_1 \Gamma_0\left(Z_1\right) \delta_1 + \Gamma_1\left(Z_1\right)\left(1-\delta_1\right) - \Gamma_2\left(Z_1\right)\right]$, with

$$\Gamma_0(x) := \exp\left\{\int_0^x \frac{dH^{(0)}(s)}{\overline{H}(s)}\right\}, \tag{4.6}$$

$$\Gamma_1(x) := \int_0^x \frac{s\Gamma_0(s)}{\overline{H}(s)} dH^{(1)}(s) \text{ and } \Gamma_2(x) := \int_x^\infty \frac{\int_s^\infty t\Gamma_0(t)\, dH^{(1)}(t)}{\left[\overline{H}(s)\right]^2} dH^{(0)}(s),$$

provided that

$$I_1 := \int_0^\infty x^2 \Gamma_0^2(x) dH^{(1)}(x) \text{ and } I_2 := \int_0^\infty x \left(\int_0^x \frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2}\right)^{1/2} dF(x), \tag{4.7}$$

be finite, where $H^{(j)}(v) := \mathbb{P}\left(Z_1 \leq v,\ \delta_1 = j\right)$, $j = 0, 1$, are two functions defined on $\mathbb{R}_+$, that will play a prominent role in this work. However, assumptions (4.7) may be violated by a class of heavy-tailed distributions. Indeed, in Lemma 4.3 we show that when F and G satisfy the second order conditions (4.2)-(4.3) with $\gamma_1 > \gamma_2/\left(1+2\gamma_2\right)$, then both I_1 and I_2 are infinite. In other words, the range

$$\mathcal{R} := \left\{\gamma_1, \gamma_2 > 0 : \frac{\gamma_2}{1+2\gamma_2} < \gamma_1 < 1\right\}, \tag{4.8}$$

is not covered by the central limit theorem established by Stute [124]. As an example of censored real datasets with indices belonging to $\mathcal{R}$, we may cite the Australian Aids data that will be described and analyzed in Section 4.4. After noting that these medical observations exhibit a heavy right tail (see Einmahl et al., [44]), we estimate, in Section 4.4, the corresponding extreme value index (EVI) γ_1 and the proportion $p := \gamma_2/(\gamma_1+\gamma_2)$ by 0.29 and 0.90, respectively, leading to a γ_2 estimate equal to 0.37. These values of (γ_1, γ_2) clearly lie in the range $\mathcal{R}$ where Stute's central limit theorem is not valid and thus no confidence interval could be constructed for the mean of this dataset. Consequently, we need to handle this situation by adopting an approach that is different from that of Stute [124]. This problem has already been addressed by Peng [107] and Johansson [81] for sets of complete data from heavy-tailed distributions with tail indices lying between 1/2 and 1. A bias reduced version of Peng's estimator is provided in Brahimi et al. [19]. Note that in the non censoring case, we have $\gamma_1 = \gamma$ meaning that $\gamma_2 = \infty$, consequently $\mathcal{R}$ reduces to Peng's range. To define our estimator, we introduce an integer sequence $k = k_n$, representing a fraction of extreme order statistics, satisfying

$$1 < k < n,\ k \to \infty \text{ and } k/n \to 0 \text{ as } n \to \infty, \tag{4.9}$$

and we set $h = h_n := H^{-1}(1-k/n)$, where $K^{-1}(y) := \inf\{x : K(x) \geq y\}$, $0 < y < 1$, denotes the quantile function of a cdf K. We start by decomposing μ into the sum of two terms as follows : $\mu = \int_0^h \overline{F}(x)dx + \int_h^\infty \overline{F}(x)dx =: \mu_1 + \mu_2$, then we estimate each term separately. Integrating the first integral by parts and changing variables in the second respectively yield

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

By replacing h and $F(x)$ by $Z_{n-k:n}$ and $\widehat{F}_n(x)$ of formula (4.4) respectively, we get

$$\widehat{\mu}_1 = \prod_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{n-k:n} + \sum_{i=2}^{n-k}\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{i:n}, \tag{4.10}$$

as an estimator to μ_1. Regarding μ_2, we apply the well-known Karamata theorem (see, for instance, de Haan and Ferreira, [69], page 363), to write

$$\mu_2 \sim \frac{\gamma_1}{1-\gamma_1} h\overline{F}(h), \text{ as } n \to \infty,\ 0 < \gamma_1 < 1. \tag{4.11}$$

The quantities h and $\overline{F}(h)$ are, as above, naturally estimated by $Z_{n-k:n}$ and

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

respectively. Now, it is clear that to derive an estimator to μ_2, one needs to estimate the tail index γ_1. The general existing method, which first appeared in Beirlant et al. [11] and then developed in Einmahl et al. [44], is to consider any consistent estimator of the extremal index γ based on the Z-sample and divide it by the proportion of observed observations in the tail. For instance, Einmahl et al. [44] adapted Hill's estimator to introduce an estimator $\widehat{\gamma}_1^{(H,c)} := \widehat{\gamma}^H/\widehat{p}$ to the tail index $\gamma_1 = \gamma/p$ under random right censorship, where

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

are the classical Hill estimator and the proportion of upper non-censored observations respectively. Note that $\widehat{\gamma}^H$ is a consistent estimator of γ (Mason, [92]) and that it is proved in Einmahl et al. [44] that $\widehat{p}$ consistently estimates p through its asymptotic normality. More recently, Brahimi et al. [18] also established the consistency of $\widehat{p}$ only by assuming the first-order condition of regular variation of cdf's F and G. In conclusion, the adapted estimator $\widehat{\gamma}_1^{(H,c)}$ is consistent for the tail index γ_1. Moreover, the authors of Brahimi et al. [18] provide a Gaussian approximation, that yields the asymptotic normality of $\widehat{\gamma}_1^{(H,c)}$, by adopting an alternative approach (based on the theory of empirical processes) which will be the key to the main result of the present work. By following similar procedures as Einmahl et al. [44], the authors of Ndao et al. [97] addressed the nonparametric estimation of the conditional EVI and large quantiles for heavy-tailed distributions which was recently generalized by Stupfler [123] to the three extreme domains of attraction namely, Féchet, Gumbel and Weibull. Based on Kaplan-Meier integration, Worms and Worms [136] introduced two new estimators for γ_1 and proved their consistency. They showed, by simulation, that they perform better, in terms of bias and root of the mean squared error (RMSE) than the adapted Hill estimator in the weak censoring case $(\gamma_1 < \gamma_2)$. Recently, Beirlant et al. [6] developed improved estimators for the EVI and tail probabilities by reducing their biases which can be quite substantial. Let us now continue with the construction our new estimator. By replacing, in (4.11), F and γ_1 by their respective empirical counterparts $\widehat{F}_n$ and $\widehat{\gamma}_1^{(H,c)}$, we obtain

$$\widehat{\mu}_2 := \frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}} Z_{n-k:n} \prod_{j=1}^{n-k} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}, \quad \text{for } \widehat{\gamma}_1^{(H,c)} < 1, \tag{4.12}$$

as an estimator for μ_2. Finally, with (4.10) and (4.12), we construct our estimator

$\widehat{\mu}$ of the mean μ as follows :

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

The rest of this Chapter is organized as follows. In Section 4.2, we state our main result which we prove in Section 4.5. We devote Section 4.3 to a simulation study in which we investigate the finite sample behavior of the newly proposed estimator $\widehat{\mu}$. In Section 4.4, we apply our estimation procedure to build a 95%-confidence interval for the mean of the survival time of Australian male Aids patients. Finally, some results, that are instrumental to our needs, are gathered in the Appendix.

4.2 Main Results

Our main result consists in the asymptotic normality of the newly introduced estimator $\widehat{\mu}$. It is stated in the following theorem which results in a corollary that is very useful in the practical construction of an asymptotic confidence interval for the expected value μ.

Theorem 4.1. *Assume that both second-order conditions of regular variation (4.2) and (4.3) hold with $(\gamma_1, \gamma_2) \in \mathcal{R}$. Let $k = k_n$ be an integer sequence satisfying (4.9) and $h = h_n := H^{-1}(1-k/n)$ such that $\sqrt{k}A_1(h) \to \lambda$, $\sqrt{k}A_2(h) = O(1)$ and $\sqrt{k}h\overline{F}(h) \to \infty$. Then*

$$\frac{\sqrt{k}\left(\widehat{\mu}-\mu\right)}{h\overline{F}(h)} \xrightarrow{\mathcal{D}} \mathcal{N}\left(m, \mathcal{V}^2\right), \text{ as } n \to \infty,$$

where

$$m := \frac{\lambda}{(1-p\rho_1)(1-\gamma_1)^2} + \frac{\lambda}{(\gamma_1+\rho_1-1)(1-\gamma_1)},$$

and

$$\mathcal{V}^2 = \mathcal{V}^2(p,\gamma_1) := \frac{2p\gamma_1\left(\gamma_1 - p^2\gamma_1^2 + p^2 + 2p\gamma_1^2 - 3p\gamma_1\right)}{(\gamma_1-1)^2(1-2p+2p\gamma_1)(1-p+p\gamma_1)} - \frac{4\gamma_1^2}{(1-\gamma_1)^3(1-2p+2p\gamma_1)} + \frac{2(1+2p)\gamma_1^2}{p(1-\gamma_1)^4}.$$

Corollary 4.1. *Under the assumptions of Theorem 4.1, with $\lambda = 0$, we have*

$$\sqrt{k}\left(\widehat{\mu}-\mu\right)/\sigma_{n,k} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \text{ as } n \to \infty,$$

where

$$\sigma_{n,k} := Z_{n-k:n} \prod_{j=1}^{n-k} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} \mathcal{V}\left(\widehat{p}, \widehat{\gamma}_1^{(H,c)}\right).$$

Remark 4.1. *It is worth mentioning, that, from a practical point of view, the assumption* $\sqrt{k}A_1(h) \to \lambda$ *is usually a bothering one. In the case of complete data, Cheng and Peng [25] solved this problem by taking* $\lambda = 0$ *to build confidence bounds for the tail index and so did Drees et al. [43] to perform tests for extremes values. Of course, this is somewhat restrictive but it allows us to avoid the estimation of the second-order parameter* ρ_1 *appearing in the asymptotic bias of* $\widehat{\mu}$*, which has not been addressed yet the extreme value analysis of censored data.*

Remark 4.2. *In this remark, we give some justification to the hypotheses of the theorem. As the development of our main result is based on the two quantities* $\sqrt{k}\left(Z_{n-k:n}/h - 1\right)$ *and* $\sqrt{k}\left(\widehat{\gamma}_1 - \gamma_1\right)$*, the assumptions* $\sqrt{k}A_i(h) = O\left(1\right)$*,* $i = 1, 2$*, are needed for their Gaussian approximations (see Theorem 2.1 in Brahimi et al., [18]). The condition* $\sqrt{k}A_1(h) \to \lambda$ *is a standard requirement, often encountered in extreme value theory, for the asymptotic bias. As for the last assumption* $\sqrt{k}h\overline{F}\left(h\right) \to \infty$*, it comes from the remainder term* $O_{\mathbb{P}}\left(1/k\right)$ *appearing in the asymptotic representation (4.18) of the Kaplan-Meier estimator on* $(-\infty, Z_{n-k:n}]$ *as a sum of independent identically distributed (iid) rv's. In Lemma 4.1, we show that the latter assumption is equivalent to*

$$k/n^{\nu} \to \infty \text{ with } 0 < \nu := \frac{2\gamma_2\left(1-\gamma_1\right)}{\gamma_1 + 3\gamma_2 - 2\gamma_1\gamma_2} < 1. \tag{4.13}$$

An example of such a sample fraction, we have $k = [n^{\alpha}]$*, with* $0 < \nu < \alpha < 1$*, where* $[a]$ *stands for the integer part of* a*.*

Remark 4.3. *Suppose that both* F *and* G *belong Hall's class (Hall, [73]), which contains the most popular heavy-tailed cdf's, such as Burr, Fréchet, Generalized Pareto, and Generalized Extreme Value (see, e.g., Beirlant et al., [6]). That is, there exist* $c_1, c_2 > 0$ *and* $d_1, d_2 \neq 0$ *such that, as* $x \to \infty$*, we have*

$$\overline{F}\left(x\right) = c_1 x^{-1/\gamma_1} + d_1 x^{-1/\gamma_1 + \rho_1/\gamma_1}\left(1 + o\left(1\right)\right),$$

and

$$\overline{G}\left(x\right) = c_2 x^{-1/\gamma_2} + d_2 x^{-1/\gamma_2 + \rho_2/\gamma_2}\left(1 + o\left(1\right)\right).$$

The corresponding cdf H *is a Hall model as well and more precisely*

$$\overline{H}\left(z\right) = cx^{-1/\gamma} + dx^{-1/\gamma + \rho/\gamma}\left(1 + o\left(1\right)\right),$$

where $\rho := \max\left(\rho_1, \rho_2\right)$*,* $c := c_1 c_2$ *and*

$$d := \begin{cases} d_1, & \rho_1 > \rho_2, \\ d_2, & \rho_1 < \rho_2, \\ d_1 + d_2, & \rho_1 = \rho_2. \end{cases}$$

It is easy to check that $\overline{F}$, $\overline{G}$ *and* $\overline{H}$ *satisfy the second-order condition of regular variation with respective convergence rates* $A_i(t) = \rho_i \gamma_i^{-1} d_i c_i^{-1} t^{\rho_i/\gamma_i}$, $i = 1, 2$, *and* $A(t) = \rho\gamma^{-1} dc^{-1} t^{\rho/\gamma}$. *We may readily show that the two conditions on the sample fraction* k, *namely* $\sqrt{k} A_1(h) \to \lambda$ *and* $\sqrt{k} A_2(h) = O(1)$, *may be expressed into* $k/n^{\nu_1} \to \lambda_*$ *and* $k/n^{\nu_2} = O(1)$, *where* $\nu_i := 2\rho_i\gamma/(2\rho_i\gamma - \gamma_i)$, $i = 1, 2$, *with* $\lambda_* := \lambda^{2\gamma_1/(2\rho_1\gamma - \gamma_1)}$. *Note that whenever* $k/n^{\nu_2} \to 0$ *and* $\lambda = 0$, *all the assumptions on* k, *including* (4.9) *and* (4.13), *may be summarized into* $1 < k < n$, $k \to \infty$, $k/n^{\nu_*} \to 0$, *where* $\nu_* := \min\{\nu_1, \nu_2, \nu, 1\}$.

4.3 Simulation Study

We carry out a simulation study to illustrate the performance of our estimator, through two sets of censored and censoring data, both drawn, in the first part, from Fréchet model $F(x) = \exp\{-x^{-1/\gamma_1}\}$, $G(x) = \exp\{-x^{-1/\gamma_2}\}$, $x \geq 0$, and, in the second part, from Burr model

$$F(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_1}, \; G(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_2}, \; x \geq 0,$$

where $\eta, \gamma_1, \gamma_2 > 0$. We fix $\eta = 1/4$ and choose the values 0.3, 0.4 and 0.5 for γ_1. For the proportion of the really observed extreme values, we take $p = 0.40$, 0.50, 0.60 and 0.70. For each couple (γ_1, p), we solve the equation $p = \gamma_2/(\gamma_1 + \gamma_2)$ to get the pertaining γ_2-value. We vary the common size n of both samples $X_1, ..., X_n$ and $Y_1, ..., Y_n$, then for each size, we generate 1000 independent replicates to take our overall results as the empirical means of the results obtained through all the repetitions. To determine the optimal number (that we denote by k^*) of upper order statistics used in the computation of $\widehat{\gamma}_1^{(H,c)}$, we apply the algorithm of automatic selection given in page 137 of Reiss and Thomas [111], detailed and evaluated in Neves and Fraga Alves [103]. The performance of the newly defined estimator $\widehat{\mu}$ is evaluated in terms of absolute bias (abs bias), mean squared error (mse) and confidence interval (conf int) accuracy via length and coverage probability (cov prob). The results, summarized in Table 4.1, Table 4.2 and Table 4.3 for Fréchet model and Table 4.4, Table 4.5 and Table 4.6 for Burr distribution, show that the same conclusions might be drawn in both cases. As expected, the sample size influences the estimation in the sense that the larger n gets, the better the estimation is. On the other hand, it is clear that the estimation accuracy increases when the censoring percentage decreases, which seems logical. Moreover, the estimator performs best for the smaller value of the tail index, as we can see from Table 4.1 and Table 4.4. Finally, many simulations realized with extreme value indices larger than 0.5, but whose results are not reported here, show that the estimator behaves poorly especially when the censorship proportion is high.

$\gamma_1 = 0.3 \rightarrow \mu = 1.298$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.247	0.052	0.021	1.043 – 1.450	0.88	0.407
1000	1.244	0.054	0.020	1.099 – 1.389	0.88	0.291
1500	1.233	0.065	0.005	1.119 – 1.346	0.80	0.227
2000	1.231	0.067	0.005	1.135 – 1.328	0.74	0.193
$p = 0.50$						
500	1.248	0.050	0.008	1.049 – 1.447	0.96	0.399
1000	1.247	0.051	0.004	1.107 – 1.387	0.90	0.280
1500	1.250	0.048	0.003	1.134 – 1.365	0.90	0.231
2000	1.248	0.050	0.003	1.146 – 1.350	0.86	0.204
$p = 0.60$						
500	1.254	0.044	0.009	1.050 – 1.458	0.90	0.408
1000	1.257	0.041	0.003	1.119 – 1.395	0.94	0.275
1500	1.266	0.032	0.002	1.153 – 1.379	0.96	0.226
2000	1.264	0.034	0.002	1.164 – 1.364	0.92	0.200
$p = 0.70$						
500	1.265	0.033	0.003	1.069 – 1.460	0.97	0.391
1000	1.269	0.029	0.002	1.123 – 1.415	0.96	0.291
1500	1.279	0.019	0.001	1.162 – 1.395	0.98	0.233
2000	1.278	0.020	0.001	1.178 – 1.377	0.96	0.199

TAB. 4.1. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Fréchet model with shape parameter 0.3

4.4 Application to AIDS Survival Data

In this Section, we apply our estimation procedure to the dataset known as Australian Aids data and provided by Dr P.J. Solomon and the Australian National Centre in HIV Epidemiology and Clinical Research. It consists in medical observations on 2843 patients (among whom 2754 are male) diagnosed with Aids in Australia before July 1^{st}, 1991. The datafile is available under the name "Aids" in the package MASS of the statistical software R. In the literature, these data were analyzed with different prospects by several authors like, for instance, Ripley and Solomon [115] and Venables and Ripley [138] (pages 379 – 385), Einmahl et al. [44], Ndao et al. [97] and Stupfler [123]. The graphical representations of the estimators of the tail index and the proportion of observed upper observations are

$\gamma_1 = 0.4 \rightarrow \mu = 1.489$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.370	0.120	0.074	$1.147 - 1.593$	0.71	0.446
1000	1.377	0.112	0.048	$1.217 - 1.536$	0.57	0.319
1500	1.367	0.122	0.019	$1.241 - 1.493$	0.48	0.252
2000	1.363	0.126	0.018	$1.256 - 1.470$	0.36	0.214
$p = 0.50$						
500	1.396	0.093	0.027	$1.169 - 1.624$	0.81	0.455
1000	1.394	0.095	0.018	$1.237 - 1.551$	0.66	0.313
1500	1.392	0.097	0.012	$1.264 - 1.521$	0.65	0.257
2000	1.389	0.101	0.012	$1.275 - 1.502$	0.55	0.227
$p = 0.60$						
500	1.407	0.082	0.013	$1.189 - 1.625$	0.89	0.436
1000	1.405	0.084	0.010	$1.251 - 1.559$	0.77	0.308
1500	1.419	0.070	0.007	$1.292 - 1.546$	0.84	0.254
2000	1.418	0.071	0.007	$1.308 - 1.529$	0.71	0.222
$p = 0.70$						
500	1.420	0.069	0.010	$1.199 - 1.641$	0.92	0.442
1000	1.433	0.056	0.006	$1.273 - 1.593$	0.86	0.320
1500	1.443	0.046	0.004	$1.312 - 1.575$	0.90	0.263
2000	1.442	0.047	0.004	$1.329 - 1.554$	0.89	0.226

TAB. 4.2. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Fréchet model with shape parameter 0.4

given in Figure 4.1. The vertical line corresponds to the optimal k-value, $k^* = 162$, while the horizontal lines stand for the pertaining estimate values $\widehat{\gamma}_1^{(H,c)} = 0.90$ and $\widehat{p} = 0.29$. The mean survival time of male patients is estimated to be 1083.61 days with a 95%-confidence interval of $1082.58 - 1084.64$.

4.5 Proofs

We begin by a brief introduction on some uniform empirical processes under random censoring. The empirical counterparts of $H^{(j)}$ $(j = 0, 1)$ are defined, for $v \geq 0$, by $H_n^{(j)}(v) := n^{-1} \sum_{i=1}^{n} \mathbb{1}\{Z_i \leq v, \delta_i = j\}$, $j = 0, 1$. In the sequel, we will use the following two empirical processes $\sqrt{n}\left(\overline{H}_n^{(j)}(v) - \overline{H}^{(j)}(v)\right)$, $j = 0, 1$; $v \geq 0$, which may be represented, almost surely, by a uniform empirical process. Indeed, let

$\gamma_1 = 0.5 \rightarrow \mu = 1.772$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.566	0.206	0.398	1.262 – 1.870	0.52	0.608
1000	1.550	0.223	0.176	1.372 – 1.727	0.28	0.355
1500	1.559	0.214	0.064	1.415 – 1.703	0.20	0.289
2000	1.549	0.224	0.061	1.426 – 1.671	0.13	0.245
$p = 0.50$						
500	1.577	0.195	0.180	1.309 – 1.846	0.53	0.537
1000	1.573	0.199	0.139	1.386 – 1.761	0.37	0.375
1500	1.578	0.195	0.051	1.430 – 1.725	0.20	0.294
2000	1.576	0.196	0.044	1.447 – 1.706	0.22	0.259
$p = 0.60$						
500	1.626	0.147	0.128	1.362 – 1.889	0.65	0.527
1000	1.617	0.155	0.034	1.430 – 1.805	0.56	0.375
1500	1.606	0.166	0.033	1.465 – 1.747	0.34	0.282
2000	1.622	0.150	0.029	1.494 – 1.751	0.34	0.258
$p = 0.70$						
500	1.632	0.141	0.046	1.375 – 1.888	0.72	0.513
1000	1.646	0.126	0.024	1.459 – 1.833	0.70	0.370
1500	1.668	0.104	0.017	1.516 – 1.821	0.68	0.305
2000	1.666	0.107	0.016	1.535 – 1.797	0.57	0.262

TAB. 4.3. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Fréchet model with shape parameter 0.5

us define, for each $i = 1, ..., n$, the rv $U_i := \delta_i H^{(1)}(Z_i) + (1 - \delta_i)(\theta + H^{(0)}(Z_i))$, where $\theta := H^{(1)}(\infty)$. From Einmahl and Koning [45], the rv's $U_1, ..., U_n$ are iid $(0, 1)$-uniform. The empirical cdf and the uniform empirical process based upon $U_1, ..., U_n$ are respectively denoted by

$$\mathbb{U}_n(s) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{U_i \leq s\} \text{ and } \alpha_n(s) := \sqrt{n}(\mathbb{U}_n(s) - s), \ 0 \leq s \leq 1.$$

Deheuvels and Einmahl [35] state that almost surely

$$H_n^{(0)}(v) = \mathbb{U}_n(H^{(0)}(v) + \theta) - \mathbb{U}_n(\theta), \text{ for } 0 < H^{(0)}(v) < 1 - \theta,$$

and

$$H_n^{(1)}(v) = \mathbb{U}_n(H^{(1)}(v)), for 0 < H^{(1)}(v) < \theta.$$

$\gamma_1 = 0.3 \rightarrow \mu = 1.228$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.186	0.042	0.077	0.972 – 1.399	0.90	0.428
1000	1.179	0.049	0.019	1.038 – 1.32	0.80	0.282
1500	1.163	0.064	0.005	1.053 – 1.273	0.80	0.220
2000	1.164	0.063	0.005	1.068 – 1.261	0.72	0.193
$p = 0.50$						
500	1.186	0.042	0.009	0.991 – 1.380	0.94	0.388
1000	1.173	0.054	0.004	1.039 – 1.308	0.93	0.269
1500	1.068	0.047	0.003	1.180 – 1.292	0.88	0.224
2000	1.181	0.046	0.003	1.086 – 1.276	0.86	0.190
$p = 0.60$						
500	1.184	0.043	0.004	0.997 – 1.371	0.95	0.374
1000	1.192	0.036	0.002	1.058 – 1.326	0.96	0.268
1500	1.196	0.031	0.002	1.088 – 1.305	0.96	0.217
2000	1.194	0.034	0.002	1.099 – 1.288	0.92	0.190
$p = 0.70$						
500	1.198	0.029	0.003	1.012 – 1.384	0.97	0.373
1000	1.200	0.028	0.001	1.066 – 1.334	0.98	0.269
1500	1.208	0.020	0.001	1.098 – 1.317	0.98	0.219
2000	1.207	0.021	0.001	1.113 – 1.301	0.98	0.188

TAB. 4.4. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Burr model with shape parameter 0.3

It is easy to verify that almost surely

$$\beta_n(v) := \sqrt{n}\left(\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)\right) = \alpha_n(\theta) - \alpha_n\left(\theta - \overline{H}^{(1)}(v)\right), \text{ for } 0 < \overline{H}^{(1)}(v) < \theta, \tag{4.14}$$

and

$$\widetilde{\beta}_n(v) := \sqrt{n}\left(\overline{H}_n^{(0)}(v) - \overline{H}^{(0)}(v)\right) = -\alpha_n\left(1 - \overline{H}^{(0)}(v)\right), \text{ for } 0 < \overline{H}^{(0)}(v) < 1 - \theta. \tag{4.15}$$

Our methodology strongly relies on the well-known Gaussian approximation given in Corollary 2.1 by Csörgő et al. [30]. It says that : on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, there exists a sequence of Brownian bridges $\{B_n(s);\ 0 \leq s \leq 1\}$ such

$\gamma_1 = 0.4 \rightarrow \mu = 1.498$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.426	0.071	0.093	1.193 – 1.660	0.76	0.466
1000	1.388	0.110	0.033	1.224 – 1.551	0.58	0.327
1500	1.374	0.124	0.020	1.248 – 1.499	0.44	0.252
2000	1.374	0.123	0.019	1.268 – 1.480	0.29	0.212
$p = 0.50$						
500	1.402	0.096	0.047	1.176 – 1.627	0.80	0.451
1000	1.389	0.109	0.017	1.231 – 1.546	0.64	0.316
1500	1.401	0.097	0.012	1.272 – 1.530	0.66	0.258
2000	1.402	0.096	0.011	1.292 – 1.511	0.53	0.219
$p = 0.60$						
500	1.422	0.076	0.043	1.186 – 1.657	0.85	0.471
1000	1.421	0.077	0.009	1.261 – 1.581	0.86	0.320
1500	1.429	0.069	0.007	1.302 – 1.556	0.80	0.254
2000	1.427	0.071	0.006	1.316 – 1.538	0.76	0.223
$p = 0.70$						
500	1.436	0.061	0.009	1.214 – 1.658	0.94	0.444
1000	1.441	0.057	0.006	1.285 – 1.597	0.92	0.312
1500	1.451	0.047	0.004	1.322 – 1.580	0.91	0.259
2000	1.449	0.049	0.004	1.340 – 1.558	0.88	0.218

TAB. 4.5. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Burr model with shape parameter 0.4

that for every $0 \leq \zeta < 1/4$,

$$\sup_{1/n \leq s \leq 1} \frac{n^{\zeta} \left| \alpha_n(1-s) - B_n(1-s) \right|}{s^{1/2-\zeta}} = O_{\mathbb{P}}(1). \tag{4.16}$$

For the increments $\alpha_n(\theta) - \alpha_n(\theta - s)$, we will need an approximation of the same type as (4.16) . Following similar arguments, mutatis mutandis, as those used in the proofs of assertions (2.2) of Theorem 2.1 and (2.8) of Theorem 2.2 in Csörgő et al. [30], we may show that, for every $0 < \theta < 1$ and $0 \leq \zeta < 1/4$, we have

$$\sup_{1/n \leq s \leq \theta} \frac{n^{\zeta} \left| \{\alpha_n(\theta) - \alpha_n(\theta - s)\} - \{B_n(\theta) - B_n(\theta - s)\} \right|}{s^{1/2-\zeta}} = O_{\mathbb{P}}(1). \tag{4.17}$$

$\gamma_1 = 0.5 \rightarrow \mu = 1.854$						
$p = 0.40$						
n	$\widehat{\mu}$	abs bias	mse	conf int	cov prob	length
500	1.654	0.200	0.760	1.330 – 1.978	0.50	0.649
1000	1.648	0.206	0.114	1.460 – 1.836	0.26	0.375
1500	1.630	0.224	0.098	1.478 – 1.782	0.14	0.304
2000	1.621	0.233	0.090	1.491 – 1.752	0.14	0.260
$p = 0.50$						
500	1.603	0.252	0.554	1.253 – 1.952	0.67	0.700
1000	1.658	0.196	0.090	1.470 – 1.847	0.34	0.378
1500	1.653	0.202	0.049	1.501 – 1.804	0.25	0.303
2000	1.656	0.198	0.045	1.530 – 1.782	0.22	0.252
$p = 0.60$						
500	1.688	0.166	0.066	1.417 – 1.959	0.67	0.542
1000	1.693	0.161	0.036	1.508 – 1.879	0.54	0.371
1500	1.695	0.159	0.031	1.544 – 1.846	0.39	0.301
2000	1.705	0.149	0.027	1.576 – 1.834	0.34	0.258
$p = 0.70$						
500	1.737	0.117	0.060	1.462 – 2.012	0.77	0.550
1000	1.737	0.117	0.036	1.547 – 1.927	0.74	0.380
1500	1.749	0.105	0.016	1.593 – 1.904	0.70	0.311
2000	1.753	0.101	0.014	1.621 – 1.885	0.60	0.264

TAB. 4.6. Absolute bias, mean squared error and 95%-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Burr model with shape parameter 0.5

Proof of Theorem 4.1. Observe that $\widehat{\mu} - \mu = (\widehat{\mu}_1 - \mu_1) + (\widehat{\mu}_2 - \mu_2)$, where

$$\widehat{\mu}_1 - \mu_1 = \int_0^{Z_{n-k:n}} \widehat{\overline{F}}_n(x)dx - \int_0^h \overline{F}(x)dx$$

and

$$\widehat{\mu}_2 - \mu_2 = \prod_{j=1}^{n-k} \left(1 - \frac{\delta_{[j:n]}}{n-j+1}\right) \frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}} Z_{n-k:n} - \int_h^\infty \overline{F}(x)dx.$$

It is clear that $\widehat{\mu}_1 - \mu_1 = \int_0^{Z_{n-k:n}} (\widehat{\overline{F}}_n(x) - \overline{F}(x))dx + \int_{Z_{n-k:n}}^h \overline{F}(x)dx$. In view of Proposition 5 combined with equation (4.9) in Csörgő [29], we have for any

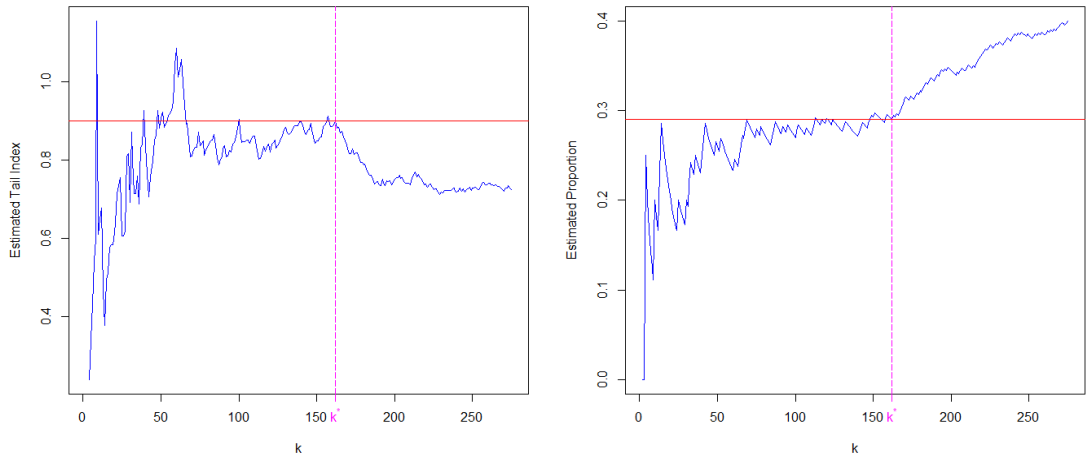

FIG. 4.1. Estimators of the tail index (left) of the survival time of Australian males diagnosed with aids, and the proportion of observed top obdervations (right) as functions of the number k of upper order statistics.

$x \leq Z_{n-k:n}$,

$$\frac{\widehat{\overline{F}}_n(x) - \overline{F}(x)}{\overline{F}(x)} = \int_0^x \frac{d\left(\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)\right)}{\overline{H}(v)} - \int_0^x \frac{\overline{H}_n(v) - \overline{H}(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) + O_{\mathbb{P}}(1/k). \tag{4.18}$$

Integrating the first integral by parts yields

$$\frac{\widehat{\overline{F}}_n(x) - \overline{F}(x)}{\overline{F}(x)} = \frac{\overline{H}_n^{(1)}(x) - \overline{H}^{(1)}(x)}{\overline{H}(x)} - \left(\overline{H}_n^{(1)}(0) - \overline{H}^{(1)}(0)\right) + \int_0^x \frac{\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}(v) - \int_0^x \frac{\overline{H}_n(v) - \overline{H}(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) + O_{\mathbb{P}}(1/k).$$

Recall that

$$\sqrt{n}\left(\overline{H}_n(v) - \overline{H}(v)\right) = \sqrt{n}\left(\overline{H}_n^{(1)}(v) - \overline{H}^{(1)}(v)\right) + \sqrt{n}\left(\overline{H}_n^{(0)}(v) - \overline{H}^{(0)}(v)\right),$$

which by representations (4.14) and (4.15) becomes

$$\sqrt{n}\left(\overline{H}_n(v) - \overline{H}(v)\right) = \alpha_n(\theta) - \alpha_n\left(\theta - \overline{H}^{(1)}(v)\right) - \alpha_n\left(1 - \overline{H}^{(0)}(v)\right).$$

From the classical central limit theorem, we have $\overline{H}_n^{(1)}(0) - \overline{H}^{(1)}(0) = O_{\mathbb{P}}\left(n^{-1/2}\right)$. Therefore, we have

$$\begin{aligned}\frac{\widehat{\overline{F}}_n(x) - \overline{F}(x)}{\overline{F}(x)} &= \frac{1}{\sqrt{n}}\frac{\beta_n(x)}{\overline{H}(x)} + \frac{1}{\sqrt{n}}\int_0^x \frac{\beta_n(v)}{\overline{H}^2(v)} d\overline{H}(v) \\ &\quad - \frac{1}{\sqrt{n}}\int_0^x \frac{\beta_n(v) + \widetilde{\beta}_n(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) + O_{\mathbb{P}}(1/k) + O_{\mathbb{P}}\left(1/\sqrt{n}\right).\end{aligned} \tag{4.19}$$

By letting $a_n := (k/n)^{1/2}/(h\overline{F}(h))$, it is easy to verify that

$$\frac{\sqrt{k}\left(\widehat{\mu}_1 - \mu_1\right)}{h\overline{F}(h)} = \sum_{i=1}^{6} T_{ni},$$

where

$$T_{n1} := a_n \int_0^{Z_{n-k:n}} \frac{\beta_n(x)}{\overline{H}(x)} \overline{F}(x) dx,$$

$$T_{n2} := a_n \int_0^{Z_{n-k:n}} \left\{ \int_0^x \frac{\beta_n(v)}{\overline{H}^2(v)} d\overline{H}(v) \right\} \overline{F}(x)\, dx,$$

$$T_{n3} := -a_n \int_0^{Z_{n-k:n}} \left\{ \int_0^x \frac{\beta_n(v) + \widetilde{\beta}_n(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) \right\} \overline{F}(x)\, dx,$$

$$T_{n4} := a_n O_{\mathbb{P}}\left(\sqrt{n}/k\right) \int_0^{Z_{n-k:n}} \overline{F}(x)\, dx,$$

$$T_{n5} := -a_n\sqrt{n} \int_{Z_{n-k:n}}^{h} \overline{F}(x)\, dx \text{ and } T_{n6} := O_{\mathbb{P}}(a_n).$$

We are now in position to apply the Gaussian approximations (4.16) and (4.17). To this end, we first have to check that for all large n and $v \in [0, Z_{n-k:n}]$, we have $\overline{H}^{(1)}(v) \in [n^{-1}, \theta]$. Note that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$ as $n \to \infty$ (see for instance, Theorem 2.1 in Brahimi et al., [18]), then for a fixed $0 < \epsilon < 1$, the probability of $A_n(\epsilon) := \{|Z_{n-k:n}/h - 1| \le \epsilon\}$ is close to 1, for all sufficiently large n. It follows that, in the set $A_n(\epsilon)$, we have

$$\overline{H}^{(1)}((1+\epsilon)h) \le \overline{H}^{(1)}(Z_{n-k:n}) \le \overline{H}^{(1)}((1-\epsilon)h),$$

because $\overline{H}^{(1)}$ is non-increasing function. On the other hand, from Lemma 4.1 in Brahimi et al. [18], we infer that $\overline{H}^{(1)}(x) \sim p\overline{H}(x)$, as $x \to \infty$, therefore $\overline{H}^{(1)}((1 \pm \epsilon)h) \sim p\overline{H}((1 \pm \epsilon)h)$. Since $\overline{H}$ is regularly varying at infinity with

index $(-1/\gamma)$ and $\overline{H}(h) = k/n$, then $\overline{H}^{(1)}((1\pm\epsilon)\,h) \sim p\,(1\pm\epsilon)^{-1/\gamma}\,k/n$, as $n\to\infty$. Hence $\overline{H}^{(1)}((1\pm\epsilon)\,h)$ are less than to $1/n$, for all large n. Since $\overline{H}^{(1)}(0) = H^{(1)}(\infty) = \theta$, then $\overline{H}^{(1)}(v) \in [n^{-1}, \theta]$ for any $v \in [0, Z_{n-k:n}]$ and all large n. Thus, the Gaussian approximation (4.17) may be applied, in T_{n1}, to obtain

$$T_{n1} = a_n \int_0^{Z_{n-k:n}} \frac{\overline{F}(x)}{\overline{H}(x)} \mathbf{B}_n\,(x)\,dx + O_{\mathbb{P}}\left(n^{-\zeta}\right) a_n \int_0^{Z_{n-k:n}} \frac{\left[\overline{H}^{(1)}(x)\right]^{1/2-\zeta}}{\overline{H}(x)} \overline{F}(x) dx,$$

for a fixed real number $0 \le \zeta < 1/4$, where

$$\mathbf{B}_n\,(x) := B_n\,(\theta) - B_n\left(\theta - \overline{H}^{(1)}\,(x)\right), \text{ for } 0 < \overline{H}^{(1)}\,(x) < \theta, \tag{4.20}$$

Next we show that

$$T_{n1} = a_n \int_0^{Z_{n-k:n}} \frac{\overline{F}(x)}{\overline{H}\,(x)} \mathbf{B}_n\,(x)\,dx + o_{\mathbb{P}}\,(1)\,.$$

Note that $\overline{H} = \overline{H}^{(0)} + \overline{H}^{(1)}$ implies that $\overline{H}^{(1)} \le \overline{H}$. Since $0 \le \zeta < 1/4$, then

$$\int_0^{Z_{n-k:n}} \frac{\left[\overline{H}^{(1)}(x)\right]^{1/2-\zeta}}{\overline{H}(x)} \overline{F}(x) dx \le \int_0^{Z_{n-k:n}} \frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}} dx.$$

which, in the set $A_n\,(\epsilon)$, is less than or equal to $\int_0^{(1+\epsilon)h} \left[\overline{H}(x)\right]^{1/2+\zeta} \overline{F}(x) dx$. By Lemma 4.2, we conclude that $n^{-\zeta} a_n \int_0^{Z_{n-k:n}} \left[\overline{H}^{(1)}(x)\right]^{1/2-\zeta} \overline{F}(x)/\overline{H}(x) dx \xrightarrow{\mathbb{P}} 0$, as sought. Next, we also show that T_{n1} may be rewritten into

$$T_{n1} = a_n \int_0^{h} \frac{\overline{F}(x)}{\overline{H}(x)} \mathbf{B}_n\,(x)\,dx + o_{\mathbb{P}}\,(1)\,. \tag{4.21}$$

To this end let us write

$$T_{n1} = a_n \int_0^{h} \frac{\overline{F}(x)}{\overline{H}(x)} \mathbf{B}_n\,(x)\,dx + a_n \int_h^{Z_{n-k:n}} \frac{\overline{F}(x)}{\overline{H}(x)} \mathbf{B}_n\,(x)\,dx + o_{\mathbb{P}}\,(1)\,,$$

with the second term in the right-hand side tending to zero in probability. Indeed, for a fixed $\varrho > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left| a_n \int_h^{Z_{n-k:n}} \frac{\overline{F}(x)}{\overline{H}(x)} \mathbf{B}_n\,(x)\,dx \right| > \varrho\right) \\
&\le \mathbb{P}\left(A_n^c\,(\epsilon)\right) + \mathbb{P}\left(a_n \int_h^{Z_{n-k:n}} \frac{\overline{F}(x)}{\overline{H}(x)} \left|\mathbf{B}_n\,(x)\right| dv > \varrho,\ A_n\,(\epsilon)\right),
\end{aligned}$$

where $A_n^c(\epsilon)$ is the complement set of $A_n(\epsilon)$. This implies that

$$\mathbb{P}\left(\left|a_n\int_h^{Z_{n-k:n}}\frac{\overline{F}(x)}{\overline{H}(x)}\mathbf{B}_n(x)\,dx\right|>\varrho\right)$$
$$\leq \mathbb{P}\left(A_n^c(\epsilon)\right)+\mathbb{P}\left(a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\overline{H}(x)}\left|\mathbf{B}_n(x)\right|dx>\varrho\right).$$

Since $\mathbb{P}\left(A_n^c(\epsilon)\right)\to 0$, then it remains to show that the second probability is also asymptotically negligible. Observe that

$$\mathbf{E}\left[a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\overline{H}(x)}\left|\mathbf{B}_n(x)\right|dx\right]\leq a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\overline{H}(x)}\mathbf{E}\left|\mathbf{B}_n(x)\right|dx,$$

with $\mathbf{E}\left|\mathbf{B}_n(x)\right|\leq\sqrt{\overline{H}^{(1)}(x)}$ and $\overline{H}^{(1)}\leq\overline{H}$. It follows that

$$\mathbf{E}\left|a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\overline{H}(x)}\mathbf{B}_n(x)\,dx\right|\leq a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\sqrt{\overline{H}(x)}}dx.$$

From now on, a key result related to the regular variation concept, namely Potter's inequalities (see, e.g., Proposition B.1.9, assertion 5 in de Haan and Ferreira [69]), will be applied quite frequently. For this reason, we need to recall this very useful tool here. Suppose that ℓ is a regularly varying function at infinity with index κ, then there exists $t_0=t_0(\epsilon)$ such that for $t\geq t_0$ and $x\geq 1$,

$$(1-\epsilon)\,x^{\kappa-\epsilon}<\ell(tx)/\ell(t)<(1+\epsilon)\,x^{\kappa+\epsilon}. \tag{4.22}$$

Since $x\to\overline{F}(x)/\sqrt{\overline{H}(x)}$ is regularly varying with index $-1/\gamma_1+1/(2\gamma)$, then by using the right inequality in (4.22), we get $\overline{F}(hx)/\sqrt{\overline{H}(hx)}\leq(1+\epsilon)\,x^{-1/\gamma_1+1/(2\gamma)+\epsilon}$, for all large n, $\epsilon>0$ and $x\geq 1$. Therefore

$$\mathbf{E}\left|a_n\int_h^{(1+\epsilon)h}\frac{\overline{F}(x)}{\overline{H}(x)}\mathbf{B}_n(x)\,dx\right|\leq a_n\frac{h\overline{F}(h)}{\sqrt{\overline{H}(h)}}(1+\epsilon)\int_1^{1+\epsilon}x^{-1/\gamma_1+1/(2\gamma)+\epsilon}dx,$$

which equals $(1+\epsilon)\int_1^{1+\epsilon}x^{-1/\gamma_1+1/(2\gamma)+\epsilon}dx$. The latter integral is clearly finite and tends to zero as $\epsilon\downarrow 0$. By similar arguments using approximations (4.16) and (4.17), we also show that

$$T_{n2}=a_n\int_0^h\left\{\int_0^x\frac{\mathbf{B}_n(v)}{\left[\overline{H}(v)\right]^2}d\overline{H}(v)\right\}\overline{F}(x)dx+o_{\mathbb{P}}(1) \tag{4.23}$$

and

$$T_{n3} = -a_n \int_0^h \left\{ \int_0^x \frac{\mathbf{B}_n^*(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) \right\} \overline{F}(x)dx + o_{\mathbb{P}}(1), \tag{4.24}$$

where

$$\mathbf{B}_n^*(v) := \mathbf{B}_n(v) - B_n\left(1 - \overline{H}^0(v)\right), \text{ for } 0 < \overline{H}^0(v) < 1 - \theta. \tag{4.25}$$

Before we examine T_{n4}, we provide an approximation to T_{n5}, for which a change of variables yields yields $T_{n5} = -\sqrt{k}\int_{Z_{n-k:n}/h}^1 \overline{F}(hx)/\overline{F}(h)dx$ which may be rewritten into

$$T_{n5} = -\sqrt{k}\int_{Z_{n-k:n}/h}^1 \left(\frac{\overline{F}(hx)}{\overline{F}(h)} - x^{-1/\gamma_1} \right) dx - \sqrt{k}\int_{Z_{n-k:n}/h}^1 x^{-1/\gamma_1} dx. \tag{4.26}$$

By the uniform inequality of the second-order regularly varying functions, given in Proposition 4 of Hua and Joe [78], we write : for possibly different function $\widetilde{A}_1$, with $\widetilde{A}_1(t) \sim A_1(t)$, as $t \to \infty$, any $0 < \epsilon < 1$ and large n, we have

$$\left| \frac{\overline{F}(hx)/\overline{F}(h) - x^{-1/\gamma_1}}{\widetilde{A}_{\mathbf{1}}(h)} - x^{-1/\gamma_1} \frac{x^{\rho_1/\gamma_1} - 1}{\gamma_1 \rho_1} \right| \leq \epsilon x^{-1/\gamma_1 + \epsilon}, \text{ for any } x \geq 1. \tag{4.27}$$

By using the inequalities (4.27), in the first integral in (4.26), we get

$$\sqrt{k}\widetilde{A}_{\mathbf{1}}(h)\int_{Z_{n-k:n}/h}^1 x^{-1/\gamma_1} \frac{x^{\rho_1/\gamma_1} - 1}{\rho_1 \gamma_1} dx + o_{\mathbb{P}}(1),$$

wich by an elementary integration, equals

$$\sqrt{k}\widetilde{A}_{\mathbf{1}}(h)\left(g(1) - g\left(Z_{n-k:n}/h\right)\right) + o_{\mathbb{P}}(1),$$

where

$$g(x) := \frac{\left(\rho_1 + \gamma_1 + x^{\rho_1/\gamma_1} - x^{\rho_1/\gamma_1}\gamma_1 - 1\right) x^{(\gamma_1 - 1)/\gamma_1}}{\rho_1(1 - \gamma_1)(\rho_1 + \gamma_1 - 1)}.$$

By the continuity of g with the fact that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$, $\widetilde{A}_{\mathbf{1}}(h) \sim A_1(h)$ and $\sqrt{k}A_1(h) \to \lambda < \infty$, we get $\sqrt{k}A_1(h)\left(g(1) - g\left(Z_{n-k:n}/h\right)\right) = o_{\mathbb{P}}(1)$ which implies that the first term in (4.26) tends to zero in probability. We develop the second integral and make a Taylor's expansion. Knowing, once again, that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$ ultimately yields that $T_{n5} = (1 + o_{\mathbb{P}}(1))\sqrt{k}\left(Z_{n-k:n}/h - 1\right)$. By using result (2.7) of Theorem 2.1 in Brahimi et al. [18], we get

$$T_{n5} = \gamma\sqrt{\frac{n}{k}}\mathbf{B}_n^*(h) + o_{\mathbb{P}}(1). \tag{4.28}$$

Next, we readily check that the fourth term T_{n4} tends to zero in probability. Indeed, we have $\int_0^{Z_{n-k:n}} \overline{F}(x)dx < \mu$ and by assumption $\sqrt{k}h\overline{F}(x) \to \infty$. Finally, for the last term T_{n6} we use the second-order regular variation of the tails $\overline{F}$ and $\overline{G}$. Indeed, from Lemma 3 in Hua and Joe [78] , the second-order conditions (4.2)−(4.3) , imply that

$$\overline{F}(u) \sim c_1 u^{-1/\gamma_1} \text{ and } \overline{G}(u) \sim c_2 u^{-1/\gamma_2}, \text{ as } u \to \infty, \tag{4.29}$$

for some positive constants c_1 and c_2. Therefore $\overline{H}(u) \sim c_1c_2u^{-1/\gamma}$ it follows that $h \sim (c_1c_2)^{\gamma}(k/n)^{-\gamma}$ and thus $a_n \sim c_1^{-1}(c_1c_2)^{\gamma/\gamma_1+\gamma}(k/n)^{1/2+\gamma-\gamma/\gamma_1}$. But the indices γ_1 and γ_2 belong to $\mathcal{R}$, hence $1/2+\gamma-\gamma/\gamma_1$ are positive, therefore, $a_n \to 0$ and $T_{n6} = o_{\mathbb{P}}(1)$. The four approximations (4.21) , (4.23) , (4.24) and (4.28) together with the asymptotic negligibility of both T_{n4} and T_{n6} give

$$\begin{aligned}
&\frac{\sqrt{k}(\widehat{\mu}_1 - \mu_1)}{h\overline{F}(h)} \\
&= a_n\int_0^h \frac{\mathbf{B}_n(x)}{\overline{H}(x)}\overline{F}(x)dx + a_n\int_0^h\left\{\int_0^x \frac{\mathbf{B}_n(v)}{\left[\overline{H}(v)\right]^2}d\overline{H}(v)\right\}\overline{F}(x)\,dx \\
&-a_n\int_0^h\left\{\int_0^x \frac{\mathbf{B}_n^*(v)}{\left[\overline{H}(v)\right]^2}d\overline{H}^{(1)}(v)\right\}\overline{F}(x)\,dx + \gamma\sqrt{\frac{n}{k}}\mathbf{B}_n^*(h) + o_{\mathbb{P}}(1).
\end{aligned} \tag{4.30}$$

Let us now treat the term $\sqrt{k}(\widehat{\mu}_2 - \mu_2)/h\overline{F}(h)$. Consider the following forms of μ_2 and $\widehat{\mu}_2$:

$$\mu_2 = h\overline{F}(h)\int_1^\infty \frac{\overline{F}(hx)}{\overline{F}(h)}dx \text{ and } \widehat{\mu}_2 = \frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}}Z_{n-k:n}\overline{F}(Z_{n-k:n})\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})},$$

and decompose $\sqrt{k}(\widehat{\mu}_2 - \mu_2)/h\overline{F}(h)$ into the sum of

$$S_{n1} := \sqrt{k}\frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}}\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})}\left\{\frac{Z_{n-k:n}}{h} - 1\right\},$$

$$S_{n2} := \sqrt{k}\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})}\left\{\frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}} - \frac{\gamma_1}{1-\gamma_1}\right\},$$

$$S_{n3} := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)}\left\{\frac{\widehat{\overline{F}}_n(Z_{n-k:n})}{\overline{F}(Z_{n-k:n})} - 1\right\},$$

$$S_{n4} := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\left\{\frac{\overline{F}(Z_{n-k:n})}{\overline{F}(h)} - \left(\frac{Z_{n-k:n}}{h}\right)^{-1/\gamma_1}\right\},$$

$$S_{n5} := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\left\{\left(\frac{Z_{n-k:n}}{h}\right)^{-1/\gamma_1} - 1\right\},$$

and

$$S_{n6} := \sqrt{k}\left\{\frac{\gamma_1}{1-\gamma_1} - \int_1^{\infty}\frac{\overline{F}(hx)}{\overline{F}(h)}dx\right\},$$

For the first term, we have $\widehat{\gamma}_1^{(H,c)} \xrightarrow{\mathbb{P}} \gamma_1$ and $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$, which, in view of the regular variation of $\overline{F}$ and its corresponding Potter's inequalities (2.35), imply that $\overline{F}(Z_{n-k:n}) = (1+o_{\mathbb{P}}(1))\,\overline{F}(h)$. Moreover, from assertion 5.19 of Proposition 5.2 in Soltane et al. [121], we infer that $\widehat{\overline{F}}_n(Z_{n-k:n}) = (1+o_{\mathbb{P}}(1))\,\overline{F}(Z_{n-k:n})$. It follows that $S_{n1} = (1+o_{\mathbb{P}}(1))\,\frac{\gamma_1}{1-\gamma_1}\sqrt{k}\,(Z_{n-k:n}/h-1)$, which, by applying result (2.7) of Theorem 2.1 in Brahimi et al. [18], is approximated as follows :

$$S_{n1} = (1+o_{\mathbb{P}}(1))\,\frac{\gamma_1\gamma}{1-\gamma_1}\sqrt{\frac{n}{k}}\mathbf{B}_n^*(h). \tag{4.31}$$

By using similar arguments, we easily show that

$$S_{n2} = (1+o_{\mathbb{P}}(1))\,\frac{1}{(1-\gamma_1)^2}\sqrt{k}\left(\widehat{\gamma}_1^{(H,c)} - \gamma_1\right),$$

which, by applying result (2.9) of Theorem 2.1 in Brahimi et al. [18] (after a change of variables), becomes

$$S_{n2} = \frac{(1+o_{\mathbb{P}}(1))}{(1-\gamma_1)^2}\left\{\frac{1}{p}\sqrt{\frac{n}{k}}\int_1^{\infty}x^{-1}\mathbf{B}_n^*(hx)\,dx - \frac{\gamma_1}{p}\sqrt{\frac{n}{k}}\mathbf{B}_n(h) + \frac{\sqrt{k}A_1(h)}{(1-p\tau_1)}\right\}. \tag{4.32}$$

For For the third term, we have $S_{n3} = (1+o_{\mathbb{P}}(1))\,\frac{\gamma_1}{1-\gamma_1}\sqrt{k}\left(\widehat{\overline{F}}_n(Z_{n-k:n})/\overline{F}(Z_{n-k:n})-1\right)$. By using the Gaussian approximation given in Proposition 5.2 of Soltane et al. [121], we write

$$\begin{aligned} S_{n3} &= (1+o_{\mathbb{P}}(1))\sqrt{\frac{k}{n}}\frac{\gamma_1}{1-\gamma_1}\left(\int_0^h\frac{\mathbf{B}_n(x)}{\left[\overline{H}(x)\right]^2}d\overline{H}(x) - \int_0^h\frac{\mathbf{B}_n^*(x)}{\left[\overline{H}(x)\right]^2}d\overline{H}^{(1)}(x)\right) \\ &\quad + (1+o_{\mathbb{P}}(1))\,\frac{\gamma_1}{1-\gamma_1}\sqrt{\frac{n}{k}}\mathbf{B}_n(h) + o_{\mathbb{P}}(1). \end{aligned} \tag{4.33}$$

For S_{n4}, we use the second-order condition (4.2) of $\overline{F}$ and the fact that $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$ to get

$$S_{n4} = o_{\mathbb{P}}\left(\sqrt{k}\left|A_1\left(h\right)\right|\right) = o_{\mathbb{P}}\left(1\right), \text{ as } n \to \infty. \tag{4.34}$$

For S_{n5}, we apply the mean value theorem with the fact $Z_{n-k:n}/h \xrightarrow{\mathbb{P}} 1$ to have

$$S_{n5} = -\left(1 + o_{\mathbb{P}}\left(1\right)\right)\frac{1}{1-\gamma_1}\sqrt{k}\left(\frac{Z_{n-k:n}}{h} - 1\right).$$

Using, once again, result (2.7) of Theorem 2.1 in Brahimi et al. [18] yields

$$S_{n5} = -\left(1 + o_{\mathbb{P}}\left(1\right)\right)\frac{\gamma}{1-\gamma_1}\sqrt{\frac{n}{k}}\mathbf{B}_n^*\left(h\right). \tag{4.35}$$

For the last term, we first note that $k^{-1/2}S_{n6} = \int_1^\infty x^{-1/\gamma_1}dx - \int_1^\infty \overline{F}\left(hx\right)/\overline{F}\left(h\right)dx$. Once again, by applying (4.27) with the fact that $\sqrt{k}A_{\mathbf{1}}\left(h\right) = O\left(1\right)$, we end up with

$$S_{n6} = \frac{\sqrt{k}A_{\mathbf{1}}\left(h\right)}{\left(1-\gamma_1\right)\left(\rho_1+\gamma_1-1\right)} + o\left(1\right). \tag{4.36}$$

By gathering (4.31), (4.32), (4.33), (4.34), (4.35) and (4.36) we end up with

$$\begin{aligned}
&\frac{\sqrt{k}\left(\widehat{\mu}_2 - \mu_2\right)}{h\overline{F}(h)} \\
&= \frac{\gamma_1}{1-\gamma_1}\sqrt{\frac{k}{n}}\left\{\int_0^h \frac{\mathbf{B}_n(x)}{\left[\overline{H}\left(x\right)\right]^2}d\overline{H}(x) - \int_0^h \frac{\mathbf{B}_n^*(x)}{\left[\overline{H}\left(x\right)\right]^2}d\overline{H}^{(1)}(x)\right\} \\
&+ \sqrt{\frac{n}{k}}\left\{-\frac{\gamma_1\mathbf{B}_n\left(h\right)}{p\left(1-\gamma_1\right)^2} - \gamma\mathbf{B}_n^*\left(h\right) + \frac{\int_1^\infty x^{-1}\mathbf{B}_n^*\left(hx\right)dx}{p\left(1-\gamma_1\right)^2}\right\} + R_{n1} + o_{\mathbb{P}}(1),
\end{aligned} \tag{4.37}$$

where

$$R_{n1} := \frac{\sqrt{k}A_1\left(h\right)}{\left(1-\gamma_1\right)}\left\{\frac{1}{\left(1-p\rho_1\right)\left(1-\gamma_1\right)} + \frac{1}{\left(\gamma_1+\rho_1-1\right)}\right\}.$$

Finally, by summing up equations (4.30) and (4.37) we obtain

$$\frac{\sqrt{k}\left(\widehat{\mu}-\mu\right)}{h\overline{F}(h)} = \sum_{i=1}^{5} D_{ni} + R_{n1} + o_{\mathbb{P}}(1),$$

where

$$D_{n1} := a_n\int_0^h \frac{\mathbf{B}_n\left(x\right)}{\overline{H}\left(x\right)}\overline{F}(x)dx, \quad D_{n2} := a_n\int_0^h\left\{\int_0^x \frac{\mathbf{B}_n\left(v\right)}{\left[\overline{H}\left(v\right)\right]^2}d\overline{H}\left(v\right)\right\}\overline{F}(x)dx,$$

$$D_{n3} := -a_n \int_0^h \left\{ \int_0^x \frac{\mathbf{B}_n^*(v)}{\left[\overline{H}(v)\right]^2} d\overline{H}^{(1)}(v) \right\} \overline{F}(x)dx,$$

$$D_{n4} := \frac{\gamma_1}{1-\gamma_1}\sqrt{\frac{k}{n}} \left(\int_0^h \frac{\mathbf{B}_n(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}(x) - \int_0^h \frac{\mathbf{B}_n^*(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}^{(1)}(x) \right),$$

$$D_{n5} := \sqrt{\frac{n}{k}} \left(-\frac{\gamma_1}{p(1-\gamma_1)^2} \mathbf{B}_n(h) + \frac{1}{p(1-\gamma_1)^2} \int_1^\infty v^{-1} \mathbf{B}_n^*(hx)\, dx \right).$$

Note that D_{n2} is of the form $-a_n \int_0^h \psi(x) d\varphi(x)$, where $\varphi(x) := \int_x^\infty \overline{F}(u)du$ and $\psi(x) := \int_0^x \mathbf{B}_n(v) / \overline{H}^2(v)\, d\overline{H}(v)$. Integrating by parts yields

$$D_{n2} = a_n \int_0^h \varphi(x) \frac{\mathbf{B}_n(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}(x) - \sqrt{\frac{k}{n}} \frac{\int_h^\infty \overline{F}(x)dx}{h\overline{F}(h)} \int_0^h \frac{\mathbf{B}_n(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}(x).$$

Equation $(B.1.9)$ in Theorem B.1.5 (Karamata's theorem) in de Haan and Ferreira [69] yields that $\int_h^\infty \overline{F}(x)dx/(h\overline{F}(h)) \to \gamma_1/(1-\gamma_1)$. We apply the same technique to D_{n3} and get

$$D_{n2} + D_{n3} + D_{n4} = L_{n2} + L_{n3} + R_{n2},$$

where $R_{n2} = o_{\mathbb{P}}(D_{n4})$ and

$$L_{n2} := a_n \int_0^h \varphi(x) \frac{\mathbf{B}_n(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}(x) \text{ and } L_{n3} := -a_n \int_0^h \varphi(x) \frac{\mathbf{B}_n^*(x)}{\left[\overline{H}(x)\right]^2} d\overline{H}^{(1)}(x).$$

This yields the following new decomposition :

$$\frac{\sqrt{k}(\widehat{\mu}-\mu)}{h\overline{F}(h)} = \sum_{i=1}^4 L_{ni} + R_{n1} + R_{n2} + o_{\mathbb{P}}(1),$$

with $L_{n1} := D_{n1}$ and $L_{n4} := D_{n5}$. The four L_{ni} are centred Gaussian rv's whose asymptotic second moments are finite, as we will see thereafter. From the covariance structure in Csörgő [29], page 2768, we have the following useful formulas :

$$\begin{cases} \mathbf{E}\left[\mathbf{B}_n(u)\mathbf{B}_n(v)\right] = \min\left(\overline{H}^{(1)}(u), \overline{H}^{(1)}(v)\right) - \overline{H}^{(1)}(u)\overline{H}^{(1)}(v), \\ \mathbf{E}\left[\mathbf{B}_n^*(u)\mathbf{B}_n^*(v)\right] = \min\left(\overline{H}(u), \overline{H}(v)\right) - \overline{H}(u)\overline{H}(v), \\ \mathbf{E}\left[\mathbf{B}_n(u)\mathbf{B}_n^*(v)\right] = \min\left(\overline{H}^{(1)}(u), \overline{H}^{(1)}(v)\right) - \overline{H}^{(1)}(u)\overline{H}(v). \end{cases} \tag{4.38}$$

Note in view (4.29), we may easily show that $k/n = \overline{H}(h) \sim c_1 c_2 h^{-1/\gamma}$. This allows us with formulas (4.38) and L'Hôpital's rule, after elementary but very tedious calculations, to get as $n \to \infty$,

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n(v)\right]}{\left[\overline{H}(u)\,\overline{H}(v)\right]^2} d\overline{H}(u)\,d\overline{H}(v) \to 2p,$$

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n^*(u)\,\mathbf{B}_n^*(v)\right]}{\left[\overline{H}(u)\,\overline{H}(v)\right]^2} d\overline{H}^{(1)}(u)\,d\overline{H}^{(1)}(v) \to 2p^2,$$

$$\frac{k}{n}\int_0^h \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n^*(v)\right]}{\left[\overline{H}(u)\,\overline{H}(v)\right]^2} d\overline{H}(u)\,d\overline{H}^{(1)}(v) \to 2p^2,$$

$$\int_0^h \int_1^\infty \frac{\mathbf{E}\left[\mathbf{B}_n^*(v)\,\mathbf{B}_n^*(hu)\right]}{u\left[\overline{H}(v)\right]^2} du d\overline{H}^{(1)}(v) \to -p\gamma,$$

$$\int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(u)\,\mathbf{B}_n(h)\right]}{\left[\overline{H}(u)\right]^2} d\overline{H}(u) \to -p \text{ and } \int_0^h \frac{\mathbf{E}\left[\mathbf{B}_n(h)\,\mathbf{B}_n^*(u)\right]}{\left[\overline{H}(u)\right]^2} d\overline{H}^{(1)}(u) \to -p^2.$$

By using the results above, we obtain

$$\mathbf{E}\left[L_{n1}\right]^2 \to \frac{2\gamma_1 p^3}{(1-p+p\gamma_1)(1-2p+2p\gamma_1)},$$

$$\mathbf{E}\left[L_{n2}\right]^2 \to \frac{2p\gamma_1^2}{(1-\gamma_1)^2(1-p+p\gamma_1)(1-2p+2p\gamma_1)},$$

$$\mathbf{E}\left[L_{n3}\right]^2 \to \frac{2p^2\gamma_1^2}{(1-\gamma_1)^2(1-p+p\gamma_1)(1-2p+2p\gamma_1)},$$

$$\mathbf{E}\left[L_{n4}\right]^2 \to \frac{2\gamma_1^2+4p\gamma_1^2}{p(1-\gamma_1)^4},\ 2\mathbf{E}\left[L_{n1}L_{n4}\right] \to 0,\ 2\mathbf{E}\left[L_{n3}L_{n4}\right] \to 0,$$

$$2\mathbf{E}\left[L_{n1}L_{n2}\right] \to -\frac{4p^2\gamma_1^2}{(1-\gamma_1)(1-p+p\gamma_1)(1-2p+2p\gamma_1)},$$

$$2\mathbf{E}\left[L_{n1}L_{n3}\right] \to \frac{4p^3\gamma_1^2}{(1-\gamma_1)(1-p+p\gamma_1)(1-2p+2p\gamma_1)},$$

$$2\mathbf{E}\left[L_{n2}L_{n3}\right] \to -\frac{4p^2\gamma_1^2}{(1-\gamma_1)^2(1-p+p\gamma_1)(1-2p+2p\gamma_1)}$$

and

$$2\mathbf{E}\left[L_{n2}L_{n4}\right] \to -\frac{4p\gamma_1^2}{p(1-\gamma_1)^3(1-2p+2p\gamma_1)}.$$

As a consequence, we conclude that $\sqrt{k}\left(\widehat{\mu}-\mu\right)/\left(h\overline{F}(h)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(m,\mathcal{V}^2\right)$, as $n\to\infty$, where $m:=\lim_{n\to\infty}R_{n1}$ and $\mathcal{V}^2:=\lim_{n\to\infty}\mathbf{E}\left[\sum_{i=1}^{4}L_{ni}\right]^2$. After gathering all the previous limits of $\mathbf{E}\left[L_{ni}L_{nj}\right]$, we end up with the asymptotic variance given in the Theorem. The expression of m is simple and easily obtainable since $\sqrt{k}A_1(h)\to\lambda$, which achieves the proof. □

4.6 Appendix

Lemma 4.1. *Assume that both the two second-order conditions of regular variation* (4.2) − (4.3) *hold with* $\gamma_2/\left(1+2\gamma_2\right)<\gamma_1<1$. *Then the following two assertions are equivalent :* (*i*) $\sqrt{k}h\overline{F}\left(h\right)\to\infty$; (*ii*) $k/n^{\nu}\to\infty$, *as* $n\to\infty$, *where* ν *is as in* (4.13).

Proof. From (4.29), we infer that $H^{-1}(1-s)\sim\left(c_1c_2\right)^{\gamma}s^{-\gamma}$ as $s\downarrow 0$. Recall that $h=H^{-1}\left(1-k/n\right)$, then $\sqrt{k}h\overline{F}\left(h\right)\sim c_1\left(c_1c_2\right)^{\gamma}k^{1/2-\gamma+\gamma/\gamma_1}/n^{-\gamma+\gamma/\gamma_1}$, for all large n. It is easy to verify that $\sqrt{k}h\overline{F}\left(h\right)\sim c_1\left(c_1c_2\right)^{\gamma}\left(k/n^{v_*}\right)^{1/2-\gamma+\gamma/\gamma_1}$, where $\nu_*=\left(-\gamma+\gamma/\gamma_1\right)/\left(1/2-\gamma+\gamma/\gamma_1\right)$. Since we have $\gamma=\gamma_1\gamma_2/\left(\gamma_1+\gamma_2\right)$, then ν_* becomes $2\gamma_2\left(1-\gamma_1\right)/\left(\gamma_1+3\gamma_2-2\gamma_1\gamma_2\right)$ which meets v. By assumption we have $\gamma_2/\left(1+2\gamma_2\right)<\gamma_1<1$, then we may easily check that indeed $0<v<1$, which completes the proof. □

Lemma 4.2. *Under the assumptions of Lemma 4.1, we have*

$$n^{-\zeta}a_n\int_0^{(1+\epsilon)h}\frac{\overline{F}(x)}{\left(\overline{H}(x)\right)^{1/2+\zeta}}dx\to 0,\ \ \textit{as } n\to\infty,$$

for any fixed $0\le\zeta<1/4$ *and* $0<\epsilon<1$.

Proof. Recall that $a_n=(k/n)^{1/2}/\left(h\overline{F}(h)\right)$ and $\overline{H}(h)=k/n$. It is clear that

$$n^{-\zeta}a_n\int_0^{(1+\epsilon)h}\frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}}dx=k^{-\zeta}\frac{\left[\overline{H}(h))\right]^{1/2+\zeta}}{h\overline{F}(h)}\int_0^{(1+\epsilon)h}\frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}}dx.$$

Since $k^{-\zeta}\to 0$, as as $n\to\infty$, then it suffices to show the limit of

$$\phi\left(h\right):=\frac{\left[\overline{H}(h))\right]^{1/2+\zeta}}{h\overline{F}(h)}\int_0^{(1+\epsilon)h}\frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}}dx$$

is finite. For the purpose of using L'Hôpital's rule, we first have to verify that both $\int_0^{(1+\epsilon)h}\frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}}dx$ and $h\overline{F}(h)/\left[\overline{H}(h))\right]^{1/2+\zeta}$ tend to infinity as $n\to\infty$. It is

obvious that

$$\int_0^{(1+\epsilon)h} \frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}} dx \geq \int_h^{(1+\epsilon)h} \frac{\overline{F}(x)}{\left[\overline{H}(x)\right]^{1/2+\zeta}} dx,$$

which, by a change of variables, equals $h\int_1^{1+\epsilon} \frac{\overline{F}(hx)}{\left[\overline{H}(hx)\right]^{1/2+\zeta}} dx$. Since $h \to \infty$ and $x \to \overline{F}(x)/\left[\overline{H}(x)\right]^{1/2+\zeta}$ is regularly varying at infinity with index $(1/2+\zeta)/\gamma - 1/\gamma_1$, then by using Potter's inequalities (4.22), we get

$$\int_1^{1+\epsilon} \frac{\overline{F}(hx)}{\left[\overline{H}(hx)\right]^{1/2+\zeta}} dx \geq (1-\epsilon) \int_1^{1+\epsilon} x^{-\frac{1}{\gamma_1}+\frac{1/2+\zeta}{\gamma}-\epsilon} dx,$$

which equals $(1-\epsilon)(1+\epsilon)^{1-\frac{1}{\gamma_1}+\frac{1/2+\zeta}{\gamma}-\epsilon} / \left(1-\frac{1}{\gamma_1}+\frac{1/2+\zeta}{\gamma}-\epsilon\right) =: c(\epsilon)$. Thus we have $\int_0^{(1+\epsilon)h} \overline{F}(x)/\left[\overline{H}(x)\right]^{1/2+\zeta} dx \geq c(\epsilon)\, h$, which tends to ∞ as $n \to \infty$. For the quantity $h\overline{F}(h)/\left[\overline{H}(h)\right]^{1/2+\zeta}$, we use (4.29) to show that

$$\left[\overline{H}(h)\right]^{1/2+\zeta} / \left(h\overline{F}(h)\right) \sim \eta_1 (k/n)^{\gamma-\gamma/\gamma_1+1/2+\zeta},$$

for some positive $\eta_1 = \eta_1(c_1, c_2, \gamma, \gamma_1, \zeta)$. Note that

$$\gamma - \gamma/\gamma_1 + 1/2 = \frac{2\gamma_1\gamma_2 - \gamma_2 + \gamma_1}{2\gamma_1(\gamma_1+\gamma_2)},$$

and by assumption $\gamma_2/(1+2\gamma_2) < \gamma_1$ which implies that $2\gamma_1\gamma_2 - \gamma_2 + \gamma_1 < 0$, it follows that $\gamma - \gamma/\gamma_1 + 1/2 < 0$. Then we can choose $0 \leq \zeta < 1/4$, such that $\gamma - \gamma/\gamma_1 + 1/2 + \zeta < 0$, to get $(k/n)^{\gamma-\gamma/\gamma_1+1/2+\zeta} \to \infty$, as $n \to \infty$, because $k/n \to 0$. Next we compute the $\lim_{n\to\infty} \phi(h)$. By using (4.29) (for $\overline{F}$), we write

$$\frac{\left[\overline{H}(h)\right]^{1/2+\zeta}}{h\overline{F}(h)} \sim \eta_2 h^{1-1/\gamma_1+(1/2+\zeta)/\gamma}.$$

or some positive $\eta_2 = \eta_2(c_1, c_2, \gamma, \gamma_1, \zeta)$, , it follows that

$$\phi(h) \sim \frac{\int_0^{(1+\epsilon)h} \overline{F}(x)/\left(\overline{H}(x)\right)^{1/2+\zeta} dx}{\eta_2 h^{1-1/\gamma_1+(1/2+\zeta)/\gamma}}, \text{ as } n \to \infty,$$

By using L'Hôpital's rule, we have

$$\lim_{n\to\infty} \phi(h) = \frac{1}{1-1/\gamma_1+(1/2+\zeta)/\gamma} \lim_{n\to\infty} \frac{\overline{F}((1+\epsilon)h)/\left(\overline{H}((1+\epsilon)h)\right)^{1/2+\zeta}}{\eta_2 h^{-1/\gamma_1+(1/2+\zeta)/\gamma}}.$$

Once again, by using (4.29), we write

$$\frac{\overline{F}((1+\epsilon)\,h)}{\left(\overline{H}((1+\epsilon)\,h)\right)^{1/2+\zeta}} \sim \eta_2\left((1+\epsilon)\,h\right)^{-1/\gamma_1+(1/2+\zeta)/\gamma}, \text{ as } n\to\infty,$$

it follows that

$$\lim_{n\to\infty}\frac{\overline{F}((1+\epsilon)\,h)/\left(\overline{H}((1+\epsilon)\,h)\right)^{1/2+\zeta}}{\eta_2 h^{-1/\gamma_1+(1/2+\zeta)/\gamma}} = (1+\epsilon)^{-1/\gamma_1+(1/2+\zeta)/\gamma},$$

which is indeed finite, as sought. □

Lemma 4.3. *Under the assumptions of Lemma 4.1, both the integrals I_1 and I_2, given in (4.7), are infinite.*

Proof. Let $u>0$ be a large real number, then

$$I_1=\int_0^\infty x^2\Gamma_0^2(x)dH^{(1)}(x)\geq\int_u^\infty x^2\Gamma_0^2(x)dH^{(1)}(x).$$

Note that $dH^{(1)}(x)=\overline{G}\,(x)\,dF\,(x)$ and $dH^{(0)}(x)=\overline{F}\,(x)\,dG\,(x)$, therefore

$$\Gamma_0(x)=\exp\left\{\int_0^x dH^{(0)}(z)/\overline{H}(z)\right\}=\frac{1}{\overline{G}\,(x)}.$$

It follows that $\int_u^\infty x^2\Gamma_0^2(x)dH^{(1)}(x)=\int_u^\infty x^2dF(x)/\overline{G}\,(x)$, which may be rewritten into

$$u^2\frac{\overline{F}(u)}{\overline{G}\,(u)}\int_1^\infty x^2\frac{\overline{G}\,(u)}{\overline{G}\,(ux)}\frac{dF(ux)}{\overline{F}(u)}.$$

Making use of Potters inequalities (4.22), applied to the regularly varying functions $\overline{F}$ and $\overline{G}$, we write, for any small $\epsilon>0$

$$\int_1^\infty x^2\frac{\overline{G}\,(u)}{\overline{G}\,(ux)}\frac{dF(ux)}{\overline{F}(u)}\geq\frac{1-\epsilon}{1+\epsilon}\left(\epsilon+1/\gamma_1\right)\int_1^\infty x^{1/\gamma_2-1/\gamma_1+1}dx.$$

Note that $\gamma_2/\left(1+2\gamma_2\right)<\gamma_1<1$, then $1/\gamma_2-1/\gamma_1+2>0$, and therefore $\int_1^\infty x^{1/\gamma_2-1/\gamma_1+1}dx=\infty$. On the other hand, by using (4.29) we show that $u^2\overline{F}(u)/\overline{G}\,(u)\sim$ $(c_1/c_2)\,u^{1/\gamma_2-1/\gamma_1+2}$, which tends to ∞ as $u\to\infty$, hence I_1 is infinite. Let us now focus on I_2 which is great than or equal to

$$\int_u^\infty x\left(\int_0^x\frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2}\right)^{1/2}dF(x).$$

Since $dH^{(0)}(x) = \overline{F}(x)\, dG(x)$ and $\overline{H}(x) = \overline{F}(x)\,\overline{G}(x)$, then the previous integral becomes $\int_u^{\infty} xdF(x)/\sqrt{\overline{G}(x)}$. Let us write

$$\int_u^{\infty} \frac{xdF(x)}{\sqrt{\overline{G}(x)}} = u\frac{\overline{F}(u)}{\sqrt{\overline{G}(u)}} \int_1^{\infty} x\sqrt{\frac{\overline{G}(u)}{\overline{G}(ux)}} xd\frac{F(ux)}{\overline{F}(u)}.$$

On again, by using (4.29), we may write $u\overline{F}(u)/\sqrt{\overline{G}(u)} \sim u^{1/(2\gamma_2)-1/\gamma_1+1}$. It is clear, since $\gamma_1 < 1$, that $1/(2\gamma_2) - 1/\gamma_1 + 1 > 0$, then by similar arguments as those used for I_1, we also show $I_2 = \infty$, therefore the details are omitted. This completes the proof. □

Conclusion et perspective

Dans cette thèse, on s'est intéressé à un problème récent en théorie des valeurs extrêmes, à savoir la présence de censure aléatoire. Ce problème est très fréquent dans plusieurs domaines de la vie socio-économique où les données sont souvent censurées aléatoirement à droite, tels la médecine, la finance, l'assurance, la fiabilité,...

Cette thèse se décompose en quatre parties distinctes auxquelles s'ajoutent une introduction. Dans l'introduction on a rappelé les domaines où l'on rencontre les données incomplètes (censurées-tronquées) avec une attention particulière sur les données censurées. Pour faciliter la lecture du document, on a rappelé dans le premier Chapitre quelques notions fondamentales de l'analyse de survie.

Le deuxième Chapitre est une sorte de rappel des concepts de base et des résultats essentiels de la théorie des valeurs extrêmes. Parmi ces résultats, on a décrit les limites possibles de la loi du maximum d'un échantillon. Ces lois sont indexées par un paramètre appelé indice des valeurs extrêmes ou indice de queue. Des estimateurs de cet indice sont présentés dans les cas de données complètes et censurées.

Au troisième Chapitre, on a considéré l'estimation de la prime de réassurance en excédent de sinistres qui suscite un très grand intérêt de la part des gens travaillant dans le domaine de la gestion des risques en général et du risque actuariel en particulier. En effet, il existe dans la littérature une grande variété d'estimateurs asymptotiquement normaux sur la base des données complètes. L'objet de ce Chapitre est de proposer un estimateur de la prime de réassurance lorsque les risques sont aléatoirement censurés à droite.

Le théorème central limite dans le cas de censure, introduit par Stute en 1995 ne s'applique pas pour une certaine classe de distributions à queues lourdes. Dans le dernier Chapitre, on utilise de la théorie des valeurs extrêmes pour proposer une approche alternative d'estimation de moyenne qui garantit la propriété de normalité asymptotique. Une étude de simulation est réalisée pour évaluer la performance de cette procédure d'estimation.

Cette thèse offre des perspectives intéressantes d'un point de vue aussi bien théorique que pratique. En effet, en plus des nouveaux axes de recherche qu'il ouvre, ce travail peut contribuer dans de nombreuses situations réelles à résoudre certains problèmes statistiques comme en recherche médicale et en assurance par exemple. Parmi les sujets succeptibles de faire l'objet de travaux de recherche scientifique, on peut citer la construction d'estimateurs asymptotiquements normaux pour :

- quelques unes des mesures de risque les plus connues, telles la valeur-en-risque (Value-at-Risque : VaR), l'espérance conditionelle de queue (conditional tail expectation), la déviation bilatérale (two-sided deviation), ... dans le cas de sinistres à queues lourdes aléatoirement censurés à droite.
- le paramètre de distorsion de la prime d'assurance du risque proportionnel sur la base de données censurées.
- le paramètre du second ordre de variation régulière sous censure aléatoire.

Bibliographie

[1] Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes : a large sample study. *Ann. Statist.*, 1100 – 1120.

[2] Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.*, 534 – 545.

[3] Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1992). A first course in order statistics. *John Wiley & Sons, New York.*

[4] Balakrishnan, N. and Cohen, A. C. (1991). Order Statistics and Inference : Estimation Methods. *Statist. Model. Decis. Sci. Academic Press.*

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

[6] Beirlant, J., Bardoutsos, A., de Wet, T., and Gijbels, I. (2016). Bias reduced tail estimation for censored Pareto type distributions. *Statist. Probab. Lett.,* **109,** 78 – 88.

[7] Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, **2**(2), 177 – 200.

[8] Beirlant, J., Dierckx, G., Guillou, A., and Staăricaă, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, **5**(2), 157 – 180.

[9] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006). Statistics of extremes : theory and applications. *John Wiley.*

[10] Beirlant, J., and Guillou, A. (2001). Pareto index estimation under moderate right censoring. *Scand. Actuar. J.*, 111 – 125.

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

[12] Beirlant, J., Matthys, G., and Dierckx, G. (2001). Heavy-tailed distributions and rating. *Astin Bull.*, **31**(01), 37 – 58.

[13] Beirlant, J., Teugels, J. L., and Vynckier, P. (1996). Practical analysis of extreme values. Leuven University Press.

[14] Benlagha, N., Grun-Réhomme, M., and Vasechko, O. (2009). Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes. *Revue MODULAD*, **47**(39).

[15] Billingsley, P. (1995) . Probability and Measure, 3rd edition. *Wiley, New York.*

[16] Bingham, N. H., Goldie, C. M., and Teugels, J. L.(1987). *Cambridge University Press.*

[17] Borchani, A. (2010). Statistiques des valeurs extrêmes dans le cas de lois discretes.

[18] Brahimi, B., Meraghni, D., and Necir, A. (2015). Gaussian approximation to the extreme value index estimator of a heavy-tailed distribution under random censoring. *Math. Methods Statist.,* **24**(4), 266 − 279.

[19] Brahimi, B., Meraghni, D., Necir, A., and Yahia, D. (2013). A bias-reduced estimator for the mean of a heavy-tailed distribution with an infinite second moment. *J. Statist. Plann. Inference*, **143**(6), 1064 − 1081.

[20] Breslow, N. E. (1972). Contribution to the discussion of the paper by DR Cox. *J. R. Stat. Soc. Ser. B*, **34**(2), 216 − 217.

[21] Breslow, N., and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, **2**(3), 437 − 453.

[22] Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J. M. (2005). Extreme value and related models with applications in engineering and science. *Wiley, Hoboken, NJ.*

[23] Caeiro, F., and Gomes, M. I. (2006). A new class of estimators of a "scale" second order parameter. *Extremes*, **9**(3-4), 193 − 211.

[24] Caeiro, F., and Gomes, M. I. (2015). Bias reduction in the estimation of a shape second-order parameter of a heavy-tailed model. *J. Stat. Comput. Simul.*, **85**(17), 3405 − 3419.

[25] Cheng, S., and Peng, L. (2001). Confidence intervals for the tail index. *Bernoulli*, 751 − 760.

[26] Ciuperca, G., and Mercadier, C. (2010). Semi-parametric estimation for heavy tailed distributions. *Extremes*, **13**(1), 55 − 87.

[27] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. *Springer, London.*

[28] Cox, D. R., and Oakes, D. (1984). Analysis of survival data. *CRC Press.*

[29] Csörgő, S. (1996). Universal Gaussian approximations under random censorship. *Ann. Statist.*, 2744 − 2778.

[30] Csörgő, M., Csorgo, S., Horváth, L., and Mason, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.*, 31 − 85.

[31] Csörgő, S., Deheuvels, P., and Mason, D. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.*, 1050 − 1077.

[32] Csörgő, S. and Mason, D.M. (1985). Central limit theorems for sums of extreme values. *Math. Proc. Cambridge Philos. Soc.,* **98**, 547 − 558.

[33] David, H. A., and Nagaraja, H. N. (2003). Order Statistics, Third Edition. *John Wiley.*

[34] Davis, R., and Resnick, S. (1984). Tail estimates motivated by extreme value theory. *Ann. Statist.*, 1467 − 1487.

[35] Deheuvels, P., and Einmahl, J. H. (1996). On the strong limiting behavior of local functionals of empirical processes based upon censored data. *Ann. Probab.*, 504 − 525.

[36] Deheuvels, P., and Einmahl, J. H. (2000). Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications. *Ann. Probab.*, 1301 − 1335.

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

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

[39] Dekkers, A. L., Einmahl, J. H., and De Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 1833 − 1855.

[40] Delmas, J. F., and Jourdain, B. (2006). Modèles aléatoires : applications aux sciences de l'ingénieur et du vivant. *Springer Sci. and Bus. Media.*

[41] Denuit, M., Purcaruff, O., and Van Keilegorni, I. (2006). Bivariate archimedean copula models for censored data in non-life insurance. *J. Actuar. Practice.* **13**, 5 − 32.

[42] Drees, H. (1995). Refined Pickands estimators of the extreme value index. *Ann. Statist.*, 2059 − 2080.

[43] Drees, H., de Haan, L., and Li, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions. *J. Statist. Plann. Inference*, **136**(10), 3498 − 3538.

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

[45] Einmahl, J.H.J. and Koning, A.J. (1992). Limit theorems for a general weighted process under random censoring. *Canad. J. Statist.,* **20**, 77 − 89.

[46] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance, *Springer-Verlag, Berlin.*

[47] Embrechts, P., and Omey, E. (1984). A property of longtailed distributions. *J. Appl. Probab.*, 80 − 87.

[48] El Adlouni, S., Bobée, B., and Ouarda, T. B. (2007). Caractérisation des distributions à queue lourde pour l'analyse des crues. Rapport de recherche no R-929, INRS-ETE, Université du Québec.

[49] El Adlouni, S., Bobée, B., and Ouarda, T. B. M. J. (2008). On the tails of extreme event distributions in hydrology. *J. Hydrology*, **355**(1), 16 – 33.

[50] Fama, E. F.(1965). The behavior of stock-market prices. *J. Business*, **38**(1), 34 – 105.

[51] Feller, W. (1971). An Introduction to Probability Theory and its Applications, 2nd edition. *Wiley, New York.*

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

[53] Fleming, T. R., and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Comm. Statist. Theory Methods*, **13**(20), 2469 – 2486.

[54] Foss, S., Korshunov, D., and Zachary, S. (2011). An introduction to heavy-tailed and subexponential distributions. *New York : Springer.*

[55] Fraga Alves, M. I., De Haan, L., and Lin, T. (2006). Third order extended regular variation. *Publ. Inst. Math.*, **80**(94), 109 – 120.

[56] Fraga Alves, M. I., Gomes, M. I., and De Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter. *Port. Math.*, **60**(2), 193 – 214.

[57] Fraga Alves, M. I., Gomes, M. I., De Haan, L., and Neves, C. (2007). A note on second order conditions in extreme value theory : linking general and heavy tail conditions. *REVSTAT*, **5**(3), 285 – 304.

[58] Gardes, L., and Girard, S. (2013). Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique. *J. de la Société Française de Statistique*, **154**(2), 98 – 118.

[59] Geluk, J., De Haan, L., Resnick, S., and Stărică, C. (1997). Second-order regular variation, convolution and the central limit theorem. *Stochastic Process. Appl.*, **69**(2), 139 – 159.

[60] Gill, R. D. (1980). Censoring and Stochastic Integrals," Mathematical Centre Tracts **124**. *Mathematisch Centrum, Amsterdam*

[61] Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, 49 – 58.

[62] Gill, R. D. (1994). Glivenko-Cantelli for Kaplan-Meier. *Math. Methods Statist.*, 3(1), 76.

[63] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.*, 423 – 453.

[64] Goegebeur, Y., Beirlant, J., and de Wet, T. (2010). Kernel estimators for the second order parameter in extreme value statistics. *J. Statist. Plann. Inference.*, **140**(9), 2632 – 2652.

[65] Gomes, M. I., De Haan, L., and Peng, L. (2002). Semi-parametric estimation of the second order parameter in statistics of extremes. *Extremes*, **5**(4), 387–414.

[66] Gomes, M. I., and Neves, M. M. (2011). Estimation of the extreme value index for randomly censored data. *Biometrical Lett.*, **48**(1), 1 – 22.

[67] Gomes, M. I., and Oliveira, O. (2003). Censoring estimators of a positive tail index. *Statist. Probab. Lett.*, **65**(3), 147 – 159.

[68] de Haan, L. (1976). Sample extremes : an elementary introduction. *Stat. Neerl.*, **30**(4), 161 – 172.

[69] de Haan, L. and Ferreira, A. (2006). Extreme Value Theory : An Introduction. *Springer-Verlag, New York.*

[70] de Haan, L., and Pereira, T. T. (1999). Estimating the index of a stable distribution. *Statist. Probab. Lett.*, **41**(1), 39 – 55.

[71] de Haan, L., and Rootzén, H. (1993). On the estimation of high quantiles. *J. Statist. Plann. Inference.*, **35**(1), 1 – 13.

[72] de Haan, L., and Stadtmüller, U. (1996). Generalized regular variation of second order. *J. Aust. Math. Soc.*, **61**(03), 381 – 395.

[73] Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. R. Stat. Soc.*, 37 – 42.

[74] Häeusler, E., and Teugels, J. L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Ann. Statist.*, 743 – 756.

[75] Hanagal, D. D. (2011). Modeling survival data using frailty models. *CRC Press.*

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

[77] Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, **27**(3), 251 – 261.

[78] Hua, L., and Joe, H. (2011). Second order regular variation and conditional tail expectation of multiple risks. *Insurance Math. Econom.*, **49**(3), 537–546.

[79] Huang, X., and Strawderman, R. L. (2006). A note on the Breslow survival estimator. *J. Nonparametr. Stat.*, **18**(1), 45 – 56.

[80] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly J. R. Methodol. Soc.*, **81**(348), 158 − 171.

[81] Johansson, J. (2003). Estimating the mean of heavy-tailed distributions. *Extremes*, **6**(2), 91 − 109.

[82] Kalbfleisch, J. D., and Prentice, R. L. (2011). The statistical analysis of failure time data. *Wiley, New York*.

[83] Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**(282), 457 − 481.

[84] Klein, J. P., and Moeschberger, M. L. (2005). Survival analysis : techniques for censored and truncated data. *Springer*.

[85] Klüppelberg, C. (1988). Subexponential distributions and integrated tails. *J. Appl. Probab.*, 132 − 141.

[86] Lee, E. T., and Wang, J. (2003). Statistical methods for survival data analysis. *John Wiley*.

[87] Lehmann, E. L., and Casella, G. (1998). Theory of Point Estimation. *Springer*

[88] Lévy, P. (1925). Calcul des probabilités. Paris : Gauthier-Villars.

[89] Lindskog, F. (2004). The Mathematics and Fundamental Ideas of Extreme Value Theory. *Stockholm : Royal Institute of Technology in Stockholm.*

[90] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Not. R. Astronomical Soc.*, **155**(1), 95 − 118.

[91] Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *J. Business* **36**, 394 − 419.

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

[93] Matthys, G., and Beirlant, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models. *Statist. Sinica*, 853 − 880.

[94] Meraghni, D. and Necir, A. (2006). Computing Confidence Bounds for the Mean of a Lévy-Stable Distribution. Proceeding in Computational Statistics (Edited by Alfredo Rizzi and Maurizio Vichi), 1285 − 1291. Physica-Verlag, *Springer,* ISBN : 3-7908-1708-2.

[95] Meraghni, D., and Necir, A. (2007). Estimating the Scale Parameter of a Lévy-stable Distribution via the Extreme Value Approach. Methodol. *Comput. Appl. Probab.*, **9**(4), 557 − 572.

[96] Meraghni, D. (2008). Modelling Distribution Tails. Thèse de Doctorat, Université de Biskra, Algérie.

[97] Ndao, P., Diop, A., and Dupuy, J. F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Comput. Statist. Data Anal.*, **79**, 63 – 79.

[98] Ndao, P., Diop, A., and Dupuy, J. F. (2016). Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. *J. Statist. Plann. Inference*, **168**, 20 – 37.

[99] Ndao, P. (2015) Modélisation de valeurs extrêmes modélisation de valeurs extrêmes conditionnelles en présence de censure. Thèse de Doctorat, Université de Gaston Berger.

[100] Necir, A. (2006). A functional law of the iterated logarithm for kernel-type estimators of the tail index. *J. Statist. Plann. Inference*, **136**(3), 780 – 802.

[101] Nelson, W. (1972). A short life test for comparing a sample with previous accelerated test results. *Technometrics*, **14**(1), 175 – 185.

[102] Neves, C. (2009). From extended regular variation to regular variation with application in extreme value statistics. *J. Math. Anal. Appl.*, **355**(1), 216 – 230.

[103] Neves, C., and Alves, M. F. (2004). Reiss and Thomas' automatic selection of the number of extremes. *Comput. Statist. Data Anal.*, **47**(4), 689 – 704.

[104] Nolan, J. P. (2009). Stable Distributions-Models for heavy tailed data. Department of Mathematics and Statistics, American University.

[105] Novak, S. Y. (2011). Extreme value methods with applications to finance. *CRC Press.*

[106] Peng, L. (1998). Asymptotically unbiased estimators for the extreme-value index. *Statist. Probab. Lett.*, **38**(2), 107 – 115.

[107] Peng, L. (2001). Estimating the mean of a heavy tailed distribution. *Statist. Probab. Lett.*, **52**(3), 255 – 264.

[108] Peterson Jr, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.*, **72**, 854 – 858.

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

[110] Reiss, R.D. (1989). Approximate distributions of order statistics. *Springer, New York.*

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

[112] Resnick, S.I. (1987). Extreme values, regular variation, and point processes. *Springer, New York.*

[113] Resnick, S.I. (2007). Heavy-Tail Phenomena, probabilistic and statistical modeling. *Springer.*

[114] Rényi, A. (1966). Calcul des Probabilités. *Dunod, Paris.*

[115] Ripley, B. D., and Solomon, P. J. (1994). A note on Australian AIDS survival. University of Adelaide, Department of Statistics.

[116] Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (2009). Stochastic processes for insurance and finance. *John Wiley & Sons.*

[117] Samoradnitsky, G., and Taqqu, M. S. (1994). Stable non-Gaussian random processes : Stochastic models with infinite variance. *CRC Press.*

[118] Saporta, G. (2006). Probabilités, analyse des données et statistique. *Editions Technip.*

[119] Segers, J. (2001). Residual estimators. *J. Statist. Plann. Inference*, **98**(1), $15 - 27$.

[120] Shorack, G.R., and Wellner, J.A. (1986). Empirical Processes with Applications to Statistics. *John Wiley & Sons.*

[121] Soltane, L., Meraghni, D., and Necir, A. (2016). Statistical estimate of the proportional hazard premium of loss under random censoring. *Afr. Stat.*, **11**(1), $883 - 899$.

[122] Soltane, L., Meraghni, D., and Necir, A. (2016). Estimating the mean of a heavy-tailed distribution under random censoring. *Submitted.*

[123] Stupfler, G. (2016). Estimating the conditional extreme-value index under random right-censoring. *J. Multivariate Anal.*, **144**, 1-24.

[124] Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.*, $422 - 439$.

[125] Stute, W., and Wang, J. L. (1993). A strong law under random censorship. *Ann. Statist.*, $1591 - 1607$.

[126] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. J. *Amer. Statist. Assoc.*, **69**(345), $169 - 173$.

[127] Wang, J. G. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *Ann. Statist.*, $1313 - 1316$.

[128] Wang, S. (1996). Premium calculation by transforming the layer premium density. *Astin Bull.*, **26**(01), $71 - 92$.

[129] Wang, X. Q., and Cheng, S. H. (2005). General regular variation of n–th order and the 2nd order edgeworth expansion of the extreme value distribution (i). *Acta Math. Sin.*, **21**(5), $1121 - 1130$.

[130] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.*, **73**(364), 812 − 815.

[131] Werner, T., and Upper, C. (2002). Time variation in the tail behaviour of bunds futures returns.

[132] Weron, R. (2001). Levy-stable distributions revisited : tail index > 2 does not exclude the Levy-stable regime. *Internat. J. Modern Phys. C*, **12**(02),209 − 223.

[133] de Wet, T., Goegebeur, Y., and Munch, M. R. (2012). Asymptotically unbiased estimation of the second order tail parameter. *Statist. Probab. Lett.*, **82**(3), 565 − 573.

[134] Wienke, A. (2010). Frailty models in survival analysis. *CRC Press.*

[135] Worms, J., and Worms, R. (2012). Estimation of second order parameters using probability weighted moments. *ESAIM Probab. Stat.*, **16**, 97 − 113.

[136] Worms, J. and Worms, R. (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes,* **17**, 337 − 358.

[137] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann.Statist.*, 163 − 177.

[138] Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S, 4th edition. *Springer.*

[139] Von Mises, R. (1936). La Distribution de la plus grande des n valeurs. Selected papers, *Amer. Math. Soc.*, 271 − 294.

[140] Zolotarev, V.M. (1986). One-dimensional Stable Distributions. *Amer. Math. Soc., Providence, RI.*

مُلَخَّص

هذه الرسالة تمثل نوعا من التزاوج بين فرعين من الإحصاء: تحليل البقاء على قيد الحياة و نظرية القيم المتطرفة. الاهتمام الرئيسي هو تمديد نتائج نظرية القيم المتطرفة إلى حالة البيانات الخاضعة للرقابة.

يبنى مقدر مقارب طبيعي لإعادة تأمين قسط في زيادة المطالبات في حالة كون هذه الأخيرة تخضع للرقابة من اليمين و يتم تقييم أدائه من خلال مجموعة بيانات محاكاة. نعتبر أيضا تقدير متوسط لتوزيع ذي ذيل ثقيل على أساس بيانات مراقبة. في هذا السياق، يقترح مقدر مقارب طبيعي يدرس سلوكه على بيانات محاكاة. يتم تطبيق النتائج المحصل عليها على بيانات طبية حقيقية لمرض الإيدز.

الكلمات المفتاحية: االرقابة العشوائية؛ التقارب الطبيعي؛ الذيول الثقيلة؛ القيم المتطرفة؛ تقدير هيل؛ تقدير كابلان-ميار؛ قسط الخطر النسبي؛ مؤشر القيم القصوى.

Résumé

Cette thèse constitue une sorte de mariage entre deux branches de la statistique : l'analyse de survie et la théorie des valeurs extrêmes. L'intérêt principal est d'étendre les résultats de la théorie des valeurs extrêmes au cas où les données sont censurées.

Un estimateur asymptotiquement normal pour la prime de réassurance en excédent des sinistres est construit dans le cas où ces derniers sont censurés à droite et sa performance évaluée à travers des ensembles de données simulées. L'estimation de la moyenne d'une distribution à queue lourde sur la base de données censurées est aussi considérée. Dans ce cadre, un estimateur est proposé, sa normalité asymptotique établie et son comportement examiné sur des données simulées. Les résultats obtenus sont appliqués sur des données réelles médicales du SIDA.

Mots Clés : Censure aléatoire ; Estimateur de Hill ; Estimateur de Kaplan-Meier ; Indice des valeurs extrêmes ; Normalité asymptotique ; Prime de risque proportionnel ; Queues lourdes ; Valeurs extrêmes.

Abstract

This thesis represents a kind of marriage between two branches of statistics: survival analysis and the theory of extreme values. The main interest is to extend the results of the theory of extreme values to the case where the data are censored.

An Asymptotically normal estimator to the excess-of-loss reinsurance premium is built when the risks are right-censored and its performance evaluated through sets of simulated data. The estimation of the mean of a heavy-tailed distribution on the basis of censored data is considered as well. In this context, an estimator is proposed, its asymptotic normality established and his behavior examined on simulated data. In this context, an estimator is proposed, its asymptotic normality established and its behavior examined on simulated data. The results obtained are applied to medical real data of AIDS.

Keywords: Asymptotic normality; Extreme values; Extreme value index; Hill estimator; Heavy tails; Kaplan-Meier estimator; Proportional hazard premium; Random censoring.