République Algérienne Démocratique Et Populaire
Ministère De L'enseignement Supérieur Et De La Recherche Scientifique

Université Saad Dahlab Blida
Faculté Des Sciences
Département De Mathématiques

Master En Mathématiques
Option
Modélisation Stochastique Et Statistique

Thème

VALEURS EXTRÊMES EN PRÉSENCE DE CENSURE

SIMULATIONS ET APPLICATION

Présenté Par

- Harti Hamza
- Mounine Youcef

Promoteur

- M. Laidi Mohamed

Année Universitaire : 2018/2019

Remerciement

Tout d'abord, nous remercions Dieu le tout-puissant qui nous a donné la force et le savoir afin d'accomplir ce travail.

Un grand merci pour nos familles, surtout nos parents qui nous ont épaulés, soutenus et suivis tout au long de ce projet.

A nos chères amis qui ont toujours été présents et fidèles.

Mes sincères remerciements à mon encadreur Monsieur Dr.Laidi Mohamed pour avoir accepté d'encadrer cette mémoire.

Nous tenons aussi à remercier également tous les membres de Jury pour avoir accepté d'évaluer notre travail.

Enfin, pour toute personne qui a contribué, de près ou de loin, à l'élaboration de ce mémoire. Veuillez bien trouver ici l'expression de nos sincères remerciements.

Harti Hamza & Mounine Youcef

Dédicace

Nous dédions ce travail à :

- Nos parents

- Nos frères

- Et nos soeurs et toute la famille

- Et nos chers amis

- Nous voulons remercier tous les membres de notre promotion.

- Et à tous les professeurs

Table des matières

Notations

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

F	: Fonction de répartition (fdr).
f	: Densité de probabilité d'une va.
μ	: Espérance ou moyenne d'une va.
$\overline{F}(t)$	: Fonction de survie.
F_n	: Fonction de répartition empirique.
$v.a$	: Variable aléatoire.
TCL	: Théorème Centrale Limite.
TVE	: Théorie des valeurs extrêmes.
$i.i.d$	: Indépendantes et identiquement distribuées.
$(\Omega\,,\,F\,,\,\mathbb{P})$	: Espace probabilisé.
$\mathbb{R}$	: Ensemble des valeurs réelles.
σ^2	: Variance d'une variable aléatoire.
S_n	: Somme arithmétique.
$\overline{X}$	: Moyenne arithmétique.
$\xrightarrow{p}$	: Converge en probabilité.
$\xrightarrow{L}$	: Converge en loi.
$\xrightarrow{p.s}$	: Convergence presque sûre.
Q	: Fonction de quantile.
Q_n	: Quantile empirique.
$p.s$	: Presque sûre.
$\mathcal{N}(0,1)$	: Loi normale standard.
$X_{n:n}$	: Maximum de $X_1, ..., X_n$.
$X_{1:n}$	: Minimum de $X_1, ..., X_n$.
$X_{k:n}$	: $k^{éme}$ statistique d'ordre.
$X_1, ..., X_n$	: Echantillon de taille n de X.
x_F	: Point terminal.
$sup(A)$	: Sup de lensemble A.
t_s	: Quantile d'ordre s.
u	: Seuil.
$\mathcal{R}v_\rho$	: Variation régulière au ∞ avec l'indice ρ.

$GEVD$	: Distribution de valeurs extrêmes généralisée.
GPD	: Distribution de Paréto généralisée.
POT	: Peak Over Threshold.
$\mathcal{D}(.)$	: Domaine d'attraction du maximum.
$\mathcal{H}_\gamma$	: Famille de la lois de valeurs extrêmes généralisée.
l	: Fonction à variation lente.
$\mathbb{I}\{A\}$	: La fonction indicatrice de l'ensemble A.
MSE	: L'erreur moyenne quadratique.
BM	: Block maxima.

Table des figures

Liste des tableaux

Résumé

Introduction générale

L'analyse des données de survie est née au $XVII^e$ siècle, dans le domaine de la démographie. L'objectif des analystes de cette époque était l'estimation, à partir des registres de décès, les diverses caractéristiques de la population comme son effectif, sa longévité, etc. Ces analyse, ne sont réalisées qu'à partir du XIX^e siècle, avec l'application de catégorisation suivant des variables exogènes (sexe, nationalité, ...) . Durant ce siècle, ils apparaissent également les premières modélisations concernant la probabilité de mourir à un certain âge, probabilité qui sera par la suite désignée sous le terme de fonction de risque.

Une deuxième particularité de cette analyse, et qu'il est très commun de se trouver en face du problème de données manquantes, c'est à dire que les données de survie ne sont pas totalement observées, elles sont incomplètes. La censure et la troncature sont les deux causes les plus répandues de ce problème là. 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. Pour des détails complets sur la censure et l'analyse de survie, le lecteur est invité à se référer aux livres de Cox et Oakes [9] , Kalbeisch et Prentice [29] ,Lee et Wang [31] , Klein et Moeschberger [30] , Wienke et Hanagal [52] .

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

En 1958, Kaplan et Meier [28] présentent d'importants résultats concernant l'estimation non-paramétrique de la fonction de survie, de l'estimateur résultant, ils étudient l'espérence, la variance et les propriétés asymptotiques. Le comportement asymptotique de l'estimateur de Kaplan et Meier [28] a suscité l'intérêt d'un grand nombre d'auteurs dont Breslow et Crowley [5] qui sont les premiers qui ont traiter la convergence et la normalité asymptotique de l'estimateur de Kaplan et Meier [28].

Une étude statistique peut poser un autre problème, outre les données incomplètes est celui de la rareté des statiques et l'estimation au delà des maximas. Il y a toute une théorie, dite la théorie des valeurs extrêmes : TVE) qui traite ces problèmes là. Cette dernière est une branche de la statistique qui essaie d'amener une solution face à ces phénomènes. Elle se repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Cependant, on retrouve deux modèles : loi généralisée des extrêmes (GEV : « Generalized Extreme Value ») et loi de Paréto généralisée (GPD : « Generalized Paréto Distribution »). Ainsi, tout a commencé avec les auteurs Fisher et Tippet [19] , quand ils étudiaient la résistance. Ils ont énoncé un théorème fondamental avec la création de trois domaines d'attraction : domaine d'attraction de Fréchet, Gumbell et Weibull. Ce théorème intéressant fait référence à un paramètre appelé l'indice de queue qui donne la forme de la queue de distribution. En effet, si l'indice de queue est positif nous sommes en présence du domaine d'attraction de Fréchet ; puis si c'est négatif, domaine d'attraction de Weibull par contre si l'indice est nul alors domaine de Gumbel.Von Mises [49], puis Jenkinson [27] , ont rassemblé les distributions de ces trois domaines en une seule écriture. C'est en ce moment que plusieurs auteurs se sont focalisés aux estimations de l'indice des valeurs extrêmes. Nous pouvons citer Hill [24] , dans le cas où l'indice est positif. Puis Pickands [38] dans la même année a proposé un estimateur de l'indice des valeurs extrêmes dans le cas général. Par contre Dekkers et al [15]. ont généralisé l'estimateur de Hill, dénommé l'estimateur des Moments 1989, Beirlant et al [3]. ont présenté à leur tour, l'estimateur de l'indice des valeurs extrêmes généré à partir de l'estimateur de Hill et de la fonction quantile.

La théorie des valeurs extrêmes (TVE), qui se fait selon deux approches, la première approche appelée GEV ; elle permet de modéliser les block des maximas 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 Paréto distribution) . Pour une description détailler de la TVE, en particulier sur l'estimation de l'indice des valeurs et quantiles extrêmes, consulter les excellents bouquins comme Embrechts et al [18], Coles [11], Beirlant et al [3], Reiss et Thomas [41] et Embrechts et al [18].
généré à partir de l'estimateur de Hill et de la fonction quantile.

Ce mémoire constitue une sorte de mariage entre trois branches de la statistique : l'analyse de survie et la théorie des valeurs extrêmes et la théorie des valeurs extrêmes censurées. On va proposer une synthèse des différentes définitions et propriétés fondamentales de ces trois domaines de la statistique. Pour y faire, on a organisé ce mémoire comme suit :

Dans le premier chapitre on a fait une introduction qui est consacrée au notion de base sur l'analyse de survie, on a commencé par quelques rappels sur les les fonctions de répartitions fdr et la fonction de survie et on a discuté la relation d'équivalence entre ces fonctions, on a parlé aussi sur les lois des grands nombre et les propriétés asymptotiques de la somme des *va iid* (TCL), et sur les différentes types de convergence (forte, en probabilité, en loi, presque sûre et en moyenne quadratique).

Dans le Deuxième chapitre, on a présenté une introduction sur la théorie des valeurs extrêmes et on a définit les statistiques d'ordre, ainsi que les lois exactes des statistiques

d'ordre et les lois asymptotiques des valeurs extrêmes. Ensuite, on a donné le résultat fondamental de la distribution GEV ainsi que les caractéristiques des différents domaines d'attraction du maximum puis on a introduit la distribution GPD, et on a donné les estimateurs classiques de l'indice de queue tels l'estimateur de Hill, de Pickands et des Moments dans le cadre de sans censure.

Au troisième chapitre, on a présenté deux cas de données incomplètes (censurées et tronquées) et on a présenté les principaux estimateurs non-paramétriques qui sont l'estimateur de Kaplan-Meier de fonction de survie et l'estimateur de Nelson-Aalen pour la fonction de hasard cumulée. Enfin, on a introduit l'estimation non-paramétrique de la moyenne sous censure aléatoire.

Le quatrième chapitre, traite le comportement de l'estimateurs de Hill en fonction de la taille de l'échantillon, et en fonction du taux de censure et l'étude des données réelles (transplantation d'un rein de 863 patient). Sa performance est évaluée à travers des ensemble de données simulées, à l'aide du logiciel d'analyse statistique *R*.
Finalement, nous concluons ce travail par une conclusion générale.

1 Introduction à l'analyse de survie

Sommaire

1.1 Introduction

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié (communément appelé décès) est le passage irréversible entre deux états (communément nommé vivant et décès). L'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie (par exemple, le temps avant une rechute ou un rejet de greffe), d'une guérison (temps entre le diagnostic et la guérison), la panne d'une machine (durée de fonctionnement d'une machine, en fiabilité) ou la survenue d'un sinistre (temps entre deux sinistres, en actuariat).L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet événement. Dans

le domaine biomédical, par exemple, on étudie ces durées dans le contexte des études longitudinales comme les enquêtes de cohorte (suivi de patients dans le temps) ou les essais thérapeutiques (tester l'efficacité d'un médicament). On cherche alors à estimer la distribution des temps de survie (fonction de survie), à comparer les fonctions de survie de plusieurs groupes ou à analyser la manière dont des variables explicatives modifient les fonctions de survie.

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

1.2 Fonctions de survie

1.2.1 Fonction de survie, queue de distribution

Définition 1.2.1 *(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 :

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

Remarque 1.2.1 *La fonction de répartition F d'une variable aléatoire 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 $\overline{F}$ que est une fonction décroissante, continue à gauche telle que :

$$\lim_{t\to 0} \overline{F}(t) = 1$$

et

$$\lim_{t\to\infty} \overline{F}(t) = 0$$

1.2.2 Fonctions empiriques de survie

Définition 1.2.2 *(Fonctions empiriques de répartition et de survie)*

Soit $(X_1, \ldots, X_n)$ un échantillon de taille $n \geq 1$ d'une va positive X de fonction de répartition F et de fonction de survie $\overline{F}(t)$.Les fonctions empiriques de répartition et de survie, F_n et $\overline{F}_n$ sont respectivement définies par :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\}, \forall t \geq 0 \tag{1.1}$$

et

$$\overline{F_n}(t) = 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i > t\}, \forall t \geq 0 \tag{1.2}$$

Où $\mathbb{I}\{A\}$ est la fonction indicatrice de l'ensemble A.

1.2.3 Densité de probabilité

Définition 1.2.3 *(Densité de probabilité f)*

C'est la fonction $f(t) \geq 0$ telle que pour tout, $t \geq 0$ $F(t) = \int_0^t f(u)\,du$ et si la fonction de répartition F admet une dérivée au point t alors :

$$f(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leqslant X < t+h)}{h} = F'(t) = -S'(t) \tag{1.3}$$

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

1.2.4 Taux de hasard λ

Définition 1.2.4 *Risque instantané λ (ou taux de hasard)*

Le risque instantané (ou taux d'incidence), pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu)

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leqslant X < t+h \,|X \geqslant t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))' \tag{1.4}$$

1.2.5 Taux de hasard cumulé Λ

Définition 1.2.5 *(Taux de hasard cumulé Λ)*

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

$$\Lambda(t) = \int_0^t \lambda(u)\,du = -\ln(S(t)) \tag{1.5}$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané)

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)\,du\right)$$

Remarque 1.2.2 *La fonction (1.4), peut également être définie en termes de fonction de répartition $F(t)$ et la fonction de densité de probabilité $f(t)$*

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \tag{1.6}$$

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

$$\Lambda(t) = \int_0^t \lambda(x)\, dx = \int_0^t \frac{dF(x)}{\overline{F}(t)} \tag{1.7}$$

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 $t = 0$, $\overline{F}(t) = 1$, $\lambda\overline{F}(t) = 0$, 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 1. On note que, en vertu de (1.8), on peut écrire :

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

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 voir, par exemple, Lee et Wang [31], (Exemple 2 :2, page 17) ou Wienke [51], (Exemple 2 :1, page 17).

1.3 Moyenne et variance

Définition 1.3.1 *(Moyenne et variance de la durée de survie)*

Le temps moyen de survie E(X) et la variance de la durée de survie V(X) sont définis par les quantités suivantes :

$$E(X) = \int_0^\infty t dF(t) = -\int_0^\infty t d[1 - F(t)] = \int_0^\infty S(t)\, dt \quad \Leftrightarrow \quad E(X) = \int_0^\infty S(t)\, dt$$

$$V(X) = 2\int_0^\infty t\, S(t)\, dt - [E(X)]^2$$

Exemple 1.3.1 *(Distribution de* WEIBULL*)*

On suppose que la v.a X suit une distribution de WEIBULL *avec la fonction de densité de probabilité suivante :*

$$f(t) = \theta \upsilon t^{\upsilon-1} e^{-\theta t^{\upsilon}}, \qquad t \geqslant 0$$

où θ, υ sont des paramètres non négatifs.

Les fonctions de distribution et de survie F et $\overline{F}$ sont respectivement données par :

$$F(t) = 1 - e^{-\theta t^{\upsilon}}, \quad et \quad \overline{F}(t) = e^{-\theta t^{\upsilon}}$$

La fonction de hasard, alors peut être obtenue :

$$\lambda(t) = \theta \upsilon t^{\upsilon-1}$$

La fonction de hasard cumulée :

$$\Lambda(t) = \lambda t^{\upsilon}$$

Remarque 1.3.1 *Quelques définitions sont couramment utilisées dans les études de survie :*

➥ ***Date d'origine :***

Elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente (pas important car c'est la durée qui nous intéresse).

➥ ***Date de point :***

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

➥ ***Date des dernières nouvelles :***

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

1.4 Théorème central limite

L'étude de somme de variables indépendantes et de même loi joue un rôle capitale en statistique. Le théorème suivant connu sous le nom de théorème central limite (TCL) établit la convergence vers la loi de Gauss.

Théorème 1.4.1 *Soient $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires, indépendantes et identiquement distribuées définies sur un même espace probabilisé $(\Omega, A, \mathbb{P})$ à valeurs dans $\mathbb{R}$ et $S_n = \sum_{i=1}^{n} X_i$ supposons que :*

$$\forall i \in \mathbb{N}^*, E(X^2) < \infty$$

si on pose :

$$E(X_i) = \mu \quad et \quad V(X_i) = \sigma^2, avec\ (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$$

alors :

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Notation 1.4.1 *On note $\mathcal{N}(m, \sigma^2)$ la variable aléatoire qui suit une loi de probabilité normale d'espérance m et de variance σ^2.*

1.5 Lois des grands nombres

Ces lois décrivent le comportement asymptotique de la moyenne de l'échantillon. Elles sont de deux types : loi faible mettant en jeu la convergence en probabilité et loi forte relative à la convergence presque sûre.

1.5.1 Loi faible des grands nombres

Théorème 1.5.1 (Khintchine)[Foata & Fuchs (1998)]

Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes et identiquement distribuées, définies sur un même espace probabilisé $(\Omega, A, \mathbb{P})$ à valeurs dans $\mathbb{R}$ et telle que :

$$\forall i \in \mathbb{N}^*, E(|X_i|) < \infty$$

Alors :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

1.5.2 Loi forte des grands nombres

Théorème 1.5.2 [Kolmogorov & Khintchine]

Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes et identiquement distribuées, définies sur un même espace probabilité $(\Omega, A, \mathbb{P})$ à valeurs dans $\mathbb{R}$ alors on a l'équivalence suivante :

$$(\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{ps} \mu < \infty) \Leftrightarrow (E(|X_i|) < \infty \quad et \quad E(X_i) = \mu)$$

1.6 La convergence

1.6.1 Convergence en probabilité

Définition 1.6.1 *On dit que la suite de variables aléatoires $(X_n, n \in \mathbb{N}^*)$ converge en probabilité vers la variable aléatoire X si $\forall\ \varepsilon > 0$ on a :*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

Ou bien :

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

On écrit :

$$X_n \xrightarrow{p} X$$

1.6.2 Convergence en loi

Définition 1.6.2 *On dit que la suite de variables aléatoires $(X_n, n \in \mathbb{N}^*)$ de fonction de répartition F_n converge en loi vers la variable aléatoire X de fonction de répartition F si $\lim_{n\to\infty} F_n(x) = F(x)$ en tout point de continuité de F on écrit :*

$$X_n \xrightarrow{L} X$$

1.6.3 Convergence presque sûre

Définition 1.6.3 *La suite de variables aléatoires $(X_n, n \in \mathbb{N}^*)$ converge presque sûrement vers la variable aléatoire X si $E\,|X|^2 < \infty$ Alors :*

$$\mathbb{P}\{\omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1 \qquad On\ note \quad X_n \xrightarrow{p.s} X$$

1.7 Conclusion

Dans ce chapitre, nous avons fait une introduction sur l'analyse de survie, ainsi on a vu les notions de fonction de survie, taux de hasard, et en plus les théorèmes et définitions des différents types de convergences nécessaires à l'étude du comportement asymptotique d'un estimateur ... Ces notions de base vont servir à aborder les valeurs extrêmes mais en présence de censure ; ce que nous allons traiter dans le chapitre suivant.

2

Théorie des valeurs extrêmes

Sommaire

2.1 Introduction

La théorie classique des valeurs extrêmes a contribué à la solution de nombreux problèmes statistiques. La modélisation des événements extrêmes est aujourd'hui un champ de recherche particulièrement actif. Ces événements extrêmes peuvent causer des dégâts humains et matériels considérables. De telles catastrophes ne peuvent pas toujours être évitées. Cependant, l'état de la société peut prendre des actions préventives pour minimiser leurs effets. Pour cela, les spécialistes sont à leur disposition la théorie statistique des valeurs extrêmes. Elle donne un résultat très intéressant. Quelle que soit la loi de la variable parente, la loi limite des extrêmes a toujours la même forme. Il est noter avec intérêt que depuis quelques années, la théorie des valeurs extrêmes(TVE) est largement utilisée pour la modélisation de tels événements. Les domaines d'application utilisant les

modèles de la (TVE) n'ont cessé de se développer ces années en touchant des domaines variés : En hydrologie, la prévision des crues par exemple, est particulièrement importante (voir David et Nagaraja [13]), en assurance, l'une des premières préoccupations, est la prise en compte des grands sinistres (voir Embrechts et al [18]) . Leur introduction en 1997 ; Coles [11] ; Beirlant et al [3] ; Reiss et Thomas [41] ; Embrechts et al [18] , est une réponse immédiate à la remise en cause de l'hypothèse de normalité surtout avec les observations en hautes fréquences. Lorsqu'on modélise le maximum des échantillons, alors, sous certaines conditions que nous préciserons plus loin, la distribution ne peut appartenir que l'une des trois lois suivantes : Weibull négative (support borné), Gumbel ou Fréchet (support non borné). C'est le résultat d'un théorème dit Théorème des 3 types, on le détaillera ultérieurement.

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 (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. Pour des présentations plus détaillées dans ce domaine, on peut citer, par exemple, les livres de Reiss, [39] et Coles [11] , Arnold et al [2] , David et Nagaraja [13] et Castillo et al [?] . Qui ont couverts pratiquement toute la matière de statistique d'ordre.

Définition 2.2.1 *Soit $(X_1, \ldots, X_n)$ n variables aléatoires indépendantes identiquement distribuées de distribution commune F et de densité f. On appelle statistique d'ordre (croissante) la suite des variables aléatoires $(X_1, \ldots, X_n)$ qui sont rangées par ordre croissant, soit :*

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

Remarque 2.2.1 *Pour $1 \leq k \leq n$ la variable $X_{k:n}$ est connue sous le nom de la k^{eme} statistique 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 2.2.2 *(Statistiques d'ordre extrême)*

Les statistiques d'ordre extrême sont définies comme termes du maximum et du minimum des n variables aléatoires $(X_1, \ldots, X_n)$ 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) $M_n = \max(X_1, ..., X_n)$ *et* $m_n = \min(X_1, ..., X_n)$

2.2.1 Fonction de répartition empirique et statistiques d'ordre

- $F_n(t)$ est la proportion des n variables qui sont inférieurs ou égales à t.
- $\overline{F_n}(t)$ c'est la proportion d'observations qui dépasse à t.

Pour $1 \leq i \leq n$,les fonctions $F_n(t)$ et $\overline{F_n}(t)$ s'écrivent en termes des valeurs des statistiques d'ordre comme suit

$$F_n(t) = \begin{cases} 0 & \text{si} \quad \text{t} \leq X_{1:n} \\ \frac{i}{n} & \text{si} \quad X_{i:n} \leq t \leq X_{i+1:n} \\ 1 & \text{si} \quad \text{t} \geq \text{X}_{n:n} \end{cases}$$

et

$$\overline{F_n}(t) = \begin{cases} 1 & \text{si} \quad \text{t} \leq X_{1:n} \\ 1 - \frac{i}{n} & \text{si} \quad X_{i:n} \leq t \leq X_{i+1:n} \\ 0 & \text{si} \quad \text{t} \geq \text{X}_{n:n} \end{cases}$$

2.2.2 Fonction de répartition et densité du maximum

Proposition 2.2.1 [Balakrishnan & Nagaraja]

➥ *La fonction de distribution (de répartition) F_{M_n} de M_n est donnée par :*

$$\forall\, x \ \in \ \mathbb{R}, \quad F_{M_n}(x) = \mathbb{P}(M_n \leqslant x) = F^n(x) \tag{2.1}$$

➥ *Si X est absolument continue de densité f, alors la fonction de densité f_{M_n} de M_n est donnée par :*

$$\forall\, x \ \in \ \mathbb{R}, \qquad f_{M_n}(x) = nF^{n-1}(x)f(x) \tag{2.2}$$

Remarque 2.2.2 *Par analogie, on peut facilement trouver la répartition et la densité du minimum et donner par :*

$$\forall\, x \ \in \ \mathbb{R}, \quad F_{m_n}(x) = \mathbb{P}(m_n \leqslant x) = 1 - \mathbb{P}(m_n > x) = 1 - (1 - F(x))^n \tag{2.3}$$

Remarque 2.2.3

D'après la formule (2.3) ,les proposition statistiques de M_n, données par sa fonction de distribution, dépendant principalement des propriétés de F pour les grandes valeurs de x. Pour les autres valeurs de x, l'influence de F décroit rapidement avec la croissance de n. Donc, l'information la plus importante à propos des extrêmes est contenue dans la queue de la loi de probabilité de X. La loi de probabilité de M_n devait nous fournir des information au sujet des événements extrêmes.
M_n tend vers un nombre réel qui peut être infini, qu'on appelle point extrême ou point terminal de la fonction de distribution F, quand $n \ \to \ \infty$.

2.2.3 Point extrême

Définition 2.2.3 *On note par x_F (resp. x_F^*) le point extrême supérieur (resp.inférieur) de la distribution F (i.e. la plus grande valeur possible pour $X_{k:n}$ qui peut prendre la valeur $+\infty$ (resp. $-\infty$)) au sens où :*

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

respectivement

$$x_F^* = \inf\{x : F(x) > 0\}$$

Proposition 2.2.2 [Embrechts & Mikosch]

La suit des maximums $\{M_n = \max(X_1, X_2, ..., X_n), \quad n \geqslant 1\}$ *converge presque-sûrement vers* x_F *quand* $n \to \infty$, *i.e :*

$$M_n \xrightarrow{p.s} x_F \quad quand \quad n \to \infty \tag{2.4}$$

Le résultat suivant découle automatiquement de (2.3).

Corollaire 2.2.1 *La suit des maximums* $\{M_n, \ n \in \mathbb{N}^*\}$ *converge en loi vers une variable aléatoire dégénérée concentrée en* x_F, *car :*

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty} \mathbb{P}(M_n \leqslant x) = \lim_{n\to\infty} F^n(x) = \begin{cases} 0 & si \quad x < x_F \\ 1 & si \quad x \geqslant x_F \end{cases} \tag{2.5}$$

Remarque 2.2.4 *Les formules exactes de la loi du terme maximum* M_n *donnée dans* (2.3) *et de la loi limite donnée dans* (2.5) *présentent toute fois un intérêt limité, et elles ne fournissent pas beaucoup d'informations sur les extrême. En pratique, la loi de la variable aléatoire parente* X *est rarement connue avec précision, et par conséquent, il en est de même pour la loi du terme maximum* M_n.
De plus, même si la loi de X *est connue avec exactitude, la loi de* M_n *n'est pas toujours facilement calculable. Par exemple, la fonction de distribution d'une variable aléatoire normale ne possède pas d'expression analytique, puisque c'est une intégrale incalculable. Donc, sa puissance* n^{eme} *est difficile à calculer et conduit à de sérieux problèmes numériques pour de grandes valeurs de* n *et pour de grandes valeurs de* x. *Pour ces raisons, il est intéressant de considérer le comportement asymptotique du terme maximum* M_n *normalisé.*

2.2.4 Fonction quantile

Définition 2.2.4 *(Fonction quantile)*

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

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

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

$$Q(s) = \inf\{t : \overline{F}(t) \leq 1 - s\}$$

Remarque 2.2.5 *Pour tout $0 < s < 1$ lorsque la fonction de répartition F est strictement croissante et continue alors :*

$$Q(s) = F^{-1}(s) = \overline{F}^{-1}(1-s)$$

Le quantile d'ordre s est défini par :

$$t_s = F^{-1}(s)$$

Définition 2.2.5 *(Fonction Quantile empirique)*

La fonction de quantile empirique de l'échantillon $(X_1, \ldots, X_n)$ est définie par :

$$Q_n(s) = F_n^{-1}(s) = \inf\{t : \overline{F_n}(t) \geq s\}$$

Où

$$Q_n(s) = \overline{F_n}^{-1}(1-s) = \inf\{t : \overline{F_n}(t) \leq (1-s)\}$$

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

$$Q_n(s) = X_{i:n} \quad si \quad \tfrac{i-1}{n} \leq s \leq \tfrac{i}{n} \ , \quad i = 1, \ldots, n$$

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

Remarque 2.2.6

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

$$U(t) = F^{-1}\left(1 - 1/t\right) = \left(1/\bar{F}\right)^{-1}(t), \qquad t \geqslant 1$$

- *la fonction empirique correspondante est :*

$$U_n(t) = Q_n\left(1 - 1/t\right), \qquad t \geqslant 1$$

2.2.5 Distributions exactes des statistiques d'ordre

Proposition 2.2.3 *(Distributions du maximum et du minimum)*

Soient n variable aléatoire $(X_1, \ldots, X_n)$ indépendante identiquement distribuées de fonction de répartition F, la distribution exacte du maximum $X_{n:n}$ est simplement donnée par la formule suivante :

$$\forall\, x \in \mathbb{R}, \qquad F_{X_{n:n}}(x) = [F(x)]^n$$

La distribution exacte du minimum est donner par :

$$\forall\, x \in \mathbb{R}, \qquad F_{X_{1:n}}(x) = 1 - [1 - F(x)]^n$$

Preuve.

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

et

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

■

Ce sont des cas particuliers importants du résultat général de $F_{X_{k:n}}(x)$ dont il est donné par :

$$\begin{aligned} \mathbb{P}\left(X_{k:n} \leq x\right) &= F_{X_{k:n}}(x) \\ &= \quad \mathbb{P}(\text{au moins } k \text{ des } n \ X \text{ sont } \leq x) \\ &= \sum_{j=k}^{n} \binom{n}{j} \mathbb{P}(X_1 \leq x)^j (1 - \mathbb{P}(X_1 \leq x))^{n-j} \\ &= \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j}. \end{aligned}$$

2.3 Distribution des valeurs extrêmes

Maintenant, supposons qu'il existe une suite $(a_n)_{n\in\mathbb{N}^*}$ de nombre réels strictement positifs et une suite $(b_n)_{n\in\mathbb{N}^*}$ de nombres réels telles que la suite des maximums normalisée $\{a_n^{-1}(M_n - b_n),\quad n \in \mathbb{N}^*\}$ converge en distribution vers une variable aléatoire non dégénérée de fonction de distribution $\mathcal{H}$, *i.e* :

$$\forall x \in \mathbb{R}, \qquad \lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = \lim_{n\to\infty} F^n\left(a_n x + b_n\right) = \mathcal{H}(x) \tag{2.6}$$

Définition 2.3.1 *Les suites* $\{a_n > 0,\ n \geqslant 1\}$ *et* $\{b_n,\ n \geqslant 1\}$ *sont appelées suites de normalisation, les constantes* $a_n \in \mathbb{R}_+^*$ *et* $b_n \in \mathbb{R}$ *sont appelées constantes de normalisation et la variable aléatoire* $[a_n^{-1}(M_n - b_n)]$ *est appelée maximum normalisé.*

2.3.1 Loi max-stable

Comme dans la théorème de la limite centrale, nous avons défini les lois stables qui sont les seules lois limites possibles pour la suite des sommes, normalisées de n variables aléatoires indépendantes et identiquement distribuées quand n tend vers l'infini, il existe une notion similaire dans la théorie des valeurs extrêmes.

Définition 2.3.2 [Embrechts & Mikosch]

La variable aléatoire, non dégénérée, X ou la loi de probabilité de X ou, encore la fonction de distribution F de X est dite max-stable, s'il existe des constante $a_n \in \mathbb{R}_+^*$ et $b_n \in \mathbb{R}$ telles que :

$$M_n \overset{d}{=} a_n X + b_n \qquad \text{pour tout} \quad n \in \mathbb{N}^*$$

Ou, ce qui est équivalent, s'il existe des constantes $a_n \in \mathbb{R}_+^*$ et $b_n \in \mathbb{R}$ telles que :

$$F^n\left(a_n x + b_n\right) = F(x) \qquad \text{pour tout} \qquad n \in \mathbb{N}^* \ \text{ et } \ x \in \mathbb{R}$$

Exemple 2.3.1 [Coles]

Supposons que X est une variable aléatoire suivant une loi de probabilité de Fréchet *standard, i.e :*

$$\forall x \in \mathbb{R}, \quad F(x) = \mathbb{P}(X \leqslant x) = \begin{cases} \exp\left(-x^{-1}\right) & si\ x > 0 \\ 0 & si\ x \leqslant 0 \end{cases}$$

Posons $a_n = n$ et $b_n = 0$, on a alors

$$\forall x \in \mathbb{R},\ \forall n \in \mathbb{N}^*, \quad F^n(a_n x + b_n) = F^n(nx)$$

$$\text{donc} \quad F^n(a_n x + b_n) = \begin{cases} \left[\exp\left(-(nx)^{-1}\right)\right]^n & si\ nx > 0 \\ 0 & si\ nx \leqslant 0 \end{cases}$$

$$= \begin{cases} \exp\left(-x^{-1}\right) & si\ x > 0 \\ 0 & si\ x \leqslant 0 \end{cases}$$

$$= F(x)$$

Finalement

$$\forall x \in \mathbb{R},\ \forall n \in \mathbb{N}^*, \quad \exists\, a_n = n \quad et \quad b_n = 0 \quad tels\, que \quad F^n(a_n x + b_n) = F(x)$$

On déduit que la loi de probabilité de Fréchet *standard est une loi max-stable.*

Théorème 2.3.1 [Fisher & Tippett]

Soit $(X_i)_{i\geq 1}$ une suite de n variable aléatoire indépendante identiquement distribuées de fonction de répartition 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 :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}(x), \forall x \in \mathbb{R} \tag{2.7}$$

$\mathcal{H}$ est la distribution des valeurs extrêmes.
Alors à une translation et un changement d'échelle près, la fonction de répartition de la limite est du type des trois classes suivantes :

$$\textit{Loi de Gumbel :}\ \Lambda(x) = \exp(-\exp(-x)) \quad , \mathrm{x} \in \mathbb{R} \ \text{ et } \ \alpha = 0$$

$$\textit{Loi de Fréchet :}\ \Phi_\alpha(x) = \begin{cases} \exp(-x^{-\alpha}) & , \mathrm{x} \geq 0 \\ 0 & , \mathrm{x} < 0 \end{cases} \ \text{ et } \ \alpha > 0$$

$$\textit{Loi de Weibull :}\ \Psi_\alpha(x) = \begin{cases} \exp(-(-x^{-\alpha}) & , \mathrm{x} \leq 0 \\ 0 & , \mathrm{x} > 0 \end{cases} \ \text{ et } \ \alpha < 0$$

Preuve. Pour une démonstration du ce théorème, le lecteur pourra se référer aux ouvrages suivants : Resnick [40], Charpentier et Denuit [10] ■

Remarque 2.3.1 *Pour distinguer les trois distributions, on utilise généralement les notation suivantes :*
Λ *pour la distribution* Gumbel, Φ_α *pour la distribution* Fréchet *et* Ψ_α *pour la distribution* Weibull.

Remarque 2.3.2 *Le Théorème (2.3.1) n'est valable que si les suites* $\left\{a_n \in \mathbb{R}_+^*, \quad n \in \mathbb{N}^*\right\}$ *et* $\{b_n \in \mathbb{R}, \quad n \in \mathbb{N}^*\}$ *de normalisation existent et admettent une limite, et en plus, ces suites de normalisation ne sont pas uniques . Dans ce cas, il nous donne un résultat très intéressant, car si le cardinal de l'ensemble des distribution (loi de probabilité) est grand, le cardinal de l'ensemble des distribution des valeurs extrêmes est quand à lui très petit et il est vrai pour la majorité des lois usuelles. Si l'on fait un parallèle avec le théorème de la limite centrale, la constante de normalisation* $a_n \in \mathbb{R}_+^*$ *joue le rôle de* $\sigma\sqrt{n}$*, où* σ *est l'écart type de la variable aléatoire* X*, et la constante de normalisation* $b_n \in \mathbb{R}$ *joue le rôle de* $n\mu$*, où* μ *est l'espérance de la variable aléatoire* X*. La constante de normalisation* $a_n \in \mathbb{R}_+^*$ *(respectivement.* $b_n \in \mathbb{R}$ *) s'interprète comme un paramètre d'échelle (respectivement. un paramètre de position).*

On peut choisir les constantes de normalisation d'après le théorème suivante :

Théorème 2.3.2 *On a* $(a_n) \geq 0$ *et* $(b_n) \in \mathbb{R}$ *telle que :*

$$F_{M_n}(a_n x + b_n) \xrightarrow{n \longrightarrow +\infty} \mathcal{H}(x)$$

- ❖ $b_n = 0, \qquad a_n = F^{-1}(1 - \frac{1}{n}), \qquad si\ \mathcal{H} = \Phi$
- ❖ $b_n = F^{-1}(1), \qquad a_n = F^{-1}(1) - F^{-1}(1 - \frac{1}{n}), \qquad si\ \mathcal{H} = \Psi$
- ❖ $b_n = F^{-1}(1 - \frac{1}{n}), \qquad a_n = F^{-1}(\frac{1}{n}) - F^{-1}(1 - \frac{1}{n}), \qquad si\ \mathcal{H} = \Lambda$

Proposition 2.3.1 *(Fonction de densité des valeurs extrêmes)*

Les fonction de densité des distribution des valeurs extrêmes standards, des différents types de distribution extrêmes, sont les suivantes :

$$\text{Fréchet} : \quad \phi_\alpha(x) = \begin{cases} \alpha x^{-(\alpha+1)} \exp(-x^{-\alpha}) & si\ x > 0 \\ 0 & si\ x \leqslant 0 \end{cases}, \quad \alpha > 0$$

$$\text{Weibull} : \quad \psi_\alpha(x) = \begin{cases} \alpha(-x)^{\alpha-1} \exp\{-(-x)^\alpha\} & si\ x \leqslant 0 \\ 0 & si\ x > 0 \end{cases}, \quad \alpha > 0$$

$$\text{Gumbel} : \quad \lambda(x) = \exp\left\{-(x + e^{-x})\right\}, \quad x \in \mathbb{R}$$

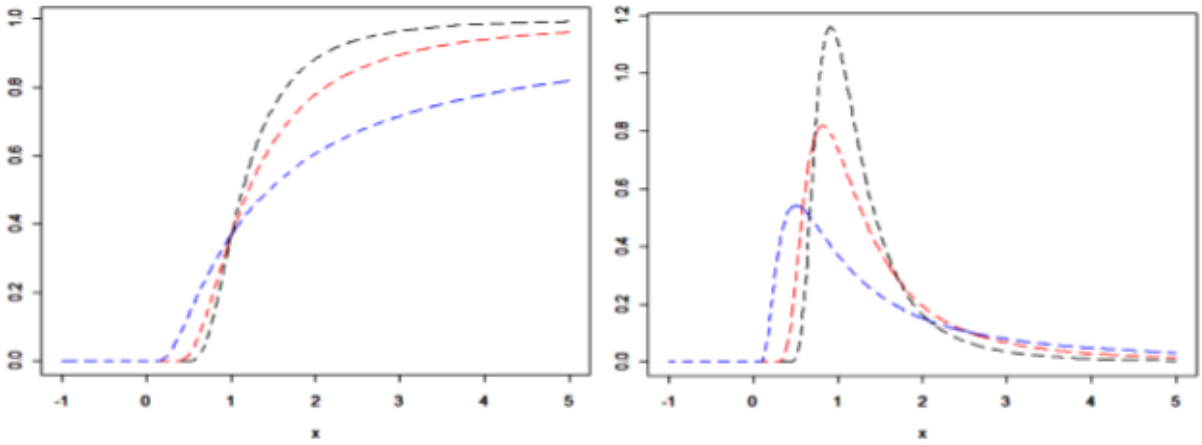

FIGURE 2.1 – Le panneau gauche représente les graphes de la distribution des valeurs extrême standard Φ_α selon les valeurs de α et le panneau droit représente les graphes de leurs densités : le bleu pour $\alpha = 1$, Le rouge pour $\alpha = 2$ et Le noir pour $\alpha = 3$.

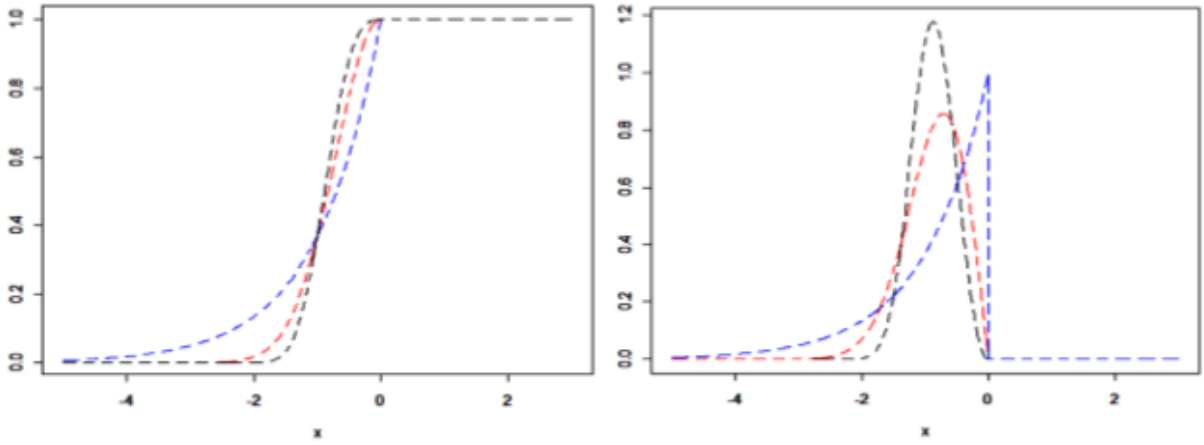

FIGURE 2.2 – Le panneau gauche représente les graphes de la distribution des valeurs extrême standard Ψ_α selon les valeurs de α et le panneau droit représente les graphes de leurs densités : le bleu pour $\alpha = 1$, Le rouge pour $\alpha = 2$ et Le noir pour $\alpha = 3$.

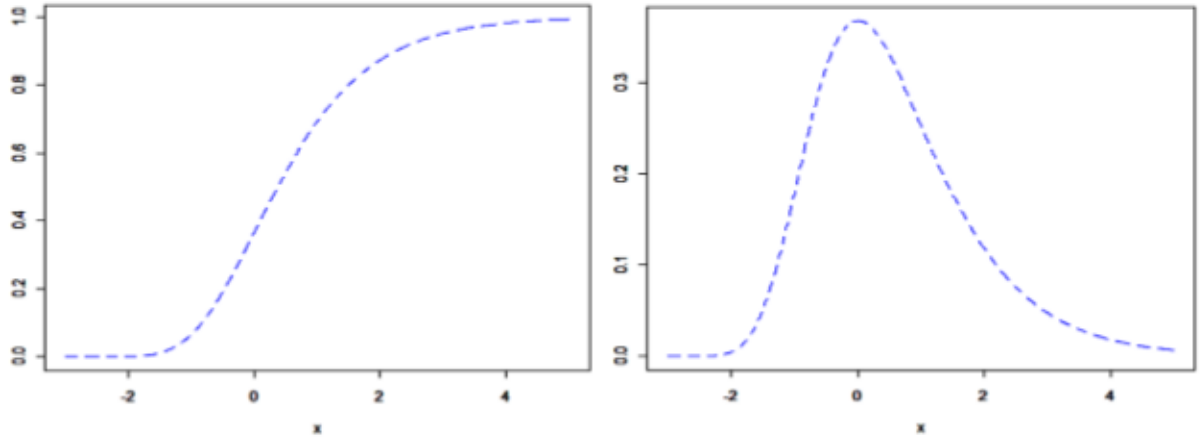

FIGURE 2.3 – Le panneau gauche représente les graphes de la distribution des valeurs extrême standard Λ et le panneau droit représente le graphe de sa densité λ .

2.4 Distributions des valeurs extrêmes généralisées

Comme on vient de le voir , les trois type de distributions extrêmes Fréchet , Weibull et Gumbel ont des comportements différents qui correspondent aux différents comportements de la fonction de queue $\overline{F}$ de la variable aléatoire X. Cela a entraîné, dans les premières applications de la théorie des valeurs extrêmes , d'adopter l'un de ces trois types pour l'analyse des données. Mais cette méthode comporte des inconvénients car, en premier lieu, on doit avoir une technique pour choisir lequel des trois de distribution extrêmes est plus approprié aux données qu'on a, et deuxièmement, une fois une telle décision est prise, les déductions subséquentes confirment que notre choix est correct, et ne tient pas compte de l'incertitude qu'une telle sélection implique, bien que cette incertitude puisse être substantielle.
Une meilleure analyse est offerte grâce aux travaux de Von Mises [49] et de Jenkinson [27] qui ont montré que les trois types de distribution extrêmes Fréchet,Weibull et Gumbel peuvent être combinées dans un seul type de distribution qu'on appelle «type de distribution des valeurs extrêmes généralisées(GEV)» ou «type de distributions des valeurs extrêmes de Von Mises-Jenkinson ».

2.4.1 Méthode des maximums par blocs

Le théorème de Fisher-Tippet donne la loi approchée du maximum d'un grand nombre d'observations *i.i.d.* En pratique, si on dispose de n observations $X_1, ..., X_n$, on commence par regrouper les données en k blocs de longueur l et on calcule le maximum sur chaque bloc

$$m_i = \max\left(x_{(i-1)l+1}, ..., x_{il}\right), \qquad pour\ i \in 1, ..., k$$

On approche ensuite la loi de la variable aléatoire M_i par une loi GEV puis on estime les paramètres de cette loi en utilisant $(m_1, ..., m_k)$. Il faut alors trouver un bon compromis entre la taille des blocs , qui doit être assez grande pour que l'approximation par la loi GEV soit réaliste, et le nombre de blocs k qui doit être assez grand pour avoir assez d'informations pour estimer les trois paramètres de la GEV.

2.4.2 Représentation de Von Mises-Jenkinson

Définition 2.4.1 [Embrechts & Mikosch]

Soient $\gamma \in \mathbb{R}$. On appelle distribution des valeurs extrêmes généralisées standard tout fonction de répartition $\mathcal{H}_\gamma$ ou toute loi de probabilité qui a $\mathcal{H}_\gamma$ comme fonction de répartition telle que pour $\gamma \in \mathbb{R}$ et $1 + \gamma x > 0$ On :

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\left\{-(1+\gamma x)^{\frac{-1}{\gamma}}\right\} & \text{si } \gamma \neq 0 \\ \exp\left\{-\exp(-x)\right\} & \text{si } \gamma = 0 \end{cases}$$

Le paramètre γ est appelé indice des valeurs extrêmes.

Proposition 2.4.1 [Ferreira (2006)]

Soient $\mathcal{H}_\gamma\,(\gamma \in \mathbb{R})$ la distribution des valeurs extrêmes généralisée et $\Phi_\alpha\,,\ \Psi_\alpha\,,\ \Lambda$ les distribution des valeurs extrêmes standards avec $\alpha > 0$, On a :

$$\forall\, x \in \mathbb{R} \ \ telque \ \ 1+\gamma x > 0, \quad \mathcal{H}_\gamma(x) = \begin{cases} \Phi_{1/\gamma}\,(1+\gamma x) & si\,\gamma > 0 \\ \Psi_{-1/\gamma}\,[-(1+\gamma x)] & si\,\gamma < 0 \\ \Lambda(x) & si\,\gamma = 0 \end{cases}$$

La proposition ci-dessus nous donne un résultat très important, dans les application de la théorie des valeurs extrêmes (Méthode des maximums par blocs),qui permette d'unifier les trois types de distributions extrêmes Fréchet, Weibull et Gumbel dans un seul type qui est le type de distribution des valeurs extrêmes généralisées. En effet, on a la proposition suivante :

Proposition 2.4.2 [Meraghni (2008)]

Si on note par $\mathcal{H}_\gamma\,(\gamma \in \mathbb{R})$ le type de distribution des valeurs extrêmes généralisées et $\Phi_\alpha\,,\ \Psi_\alpha\,,\ \Lambda$, avec $\alpha > 0$, les types de distribution des valeurs extrêmes qui sont, respectivement, de Fréchet, *de* Weibull *et de* Gumbel, *alors on a :*

$$\mathcal{H}_\gamma \ = \begin{cases} \Phi_{1/\gamma} & si\ \gamma > 0 \\ \Psi_{-1/\gamma} & si\,\gamma < 0 \\ \Lambda & si\ \ \gamma = 0 \end{cases}$$

Remarque 2.4.1 *Pour les variables non centrées et non réduites ,on peut écrire $\mathcal{H}_\gamma(x)$ sous forme plus générale notée par $\mathcal{H}_{\mu,\sigma,\gamma}(x)$ dans laquelle on fait apparaître un paramètre de localisation $\mu \in \mathbb{R}$ et un paramètre d'échelle $\sigma > 0$ pour $(1+\gamma(\frac{x-\mu}{\sigma}) > 0)$ distribution $\mathcal{H}_{\mu,\sigma,\gamma}(x)$ s'écrit comme suit :*

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

Proposition 2.4.3 *(Fonction de densité)*

Soient $\mathcal{H}_\gamma\,(\gamma \in \mathbb{R})$ *est une fonction de répartition absolument continue de densité* h_γ *définie par :*

$$\forall\, x \in \mathbb{R}, \quad h_\gamma(x) = \begin{cases} (1+\gamma x)^{\frac{-1}{\gamma}-1} \exp\left\{-(1+\gamma x)^{-1/\gamma}\right\}.\mathbb{I}_{[1+\gamma x>0]}(x) & si \quad \gamma \neq 0 \\ \exp\left\{-\left(x+e^{-x}\right)\right\} & si \quad \gamma = 0 \end{cases}$$

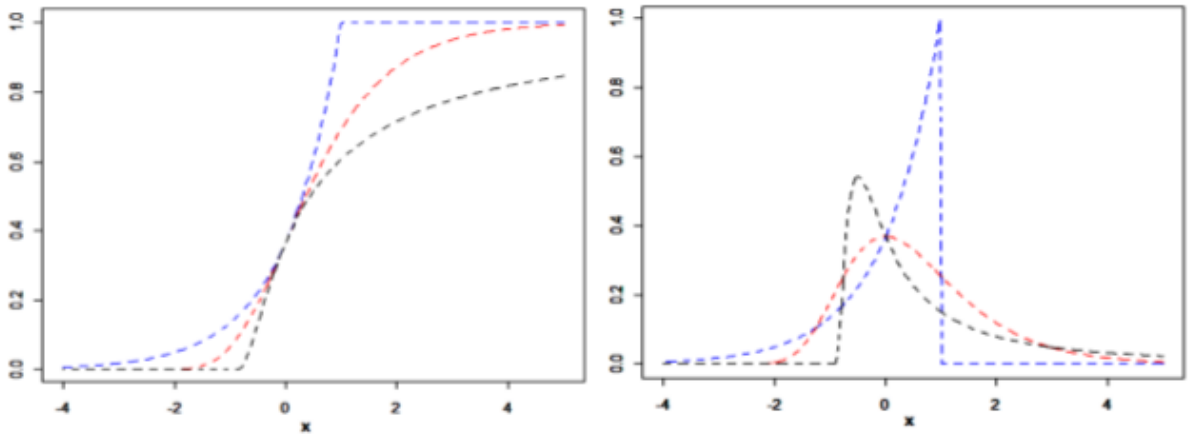

FIGURE 2.4 – Le panneau gauche représente les graphes de la distribution des valeurs extrême généralisée standard selon les valeurs de γ et le panneau droit représente les graphes de leurs densités : Le bleu pour $\gamma = -1$, Le rouge pour $\gamma = 0$ et Le noir pour $\gamma = 1$.

2.5 Distribution de Paréto généralisée (GPD)

Comme on l'a mentionné dans la section précédente, la distribution des valeurs extrêmes généralisés est très utile en application de la théorie des valeurs extrêmes, car c'est la seule et unique loi de probabilité qui modélise le comportement du maximum d'un échantillon. Pour estimer ses paramètres, les statisticiens ont souvent recours à une méthode qui s'appelle « maximums par blocs » qui consiste à construire un échantillon de maximums à partir d'un échantillon de données en format des blocs de même dimension. Cette méthode à un inconvénient majeur qui entraine une perte de certaines information, en particulier, certains blocs peuvent contenir plusieurs valeurs extrêmes, alors que d'autres blocs peuvent ne pas en contenir.
pour remédier à l'inconvénient de la méthode des maximums par blocs, les statisticiens utilisent une autre méthode qui permet de prendre en compte beaucoup plus de données pour assurer beaucoup plus de précision dans l'estimation des paramètres de la distribution des valeurs extrêmes généralisée, en particulier, l'indice des valeurs extrêmes $\gamma \in \mathbb{R}$. Cette méthode qu'on appelle « excès au-delà d'un seuil (POT) », basée sur la distribution de PARÉTO généralisée (GPD), consiste à étudier le comportement non pas du maximum des données qu'on a mais de toutes les données qui dépassent un seuil élevé u, et plus précisément, les différences entre ces données et le seuil u, appelées « excès ».

2.5.1 Méthode POT (Peak Over Threshold)

Dans la méthode des maximas par blocs, des blocs de taille identique sont constitués puis seulement le maximum de chacun d'eux est utilisé pour ajuster une loi GEV. Ce choix des blocs est généralement arbitraire, et on perd généralement de l'information sur les événements extrêmes qui par nature sont déjà peu observés. Par exemple plusieurs événements extrêmes intéressants peuvent être mis dans un même bloc alors qu'un autre bloc ne contient pas d'événement " extrême "! En pratique, cela signifie que' il faut beaucoup de données pour pouvoir mettre en place la méthode des maxima par blocs (typiquement quelques décennies si on fait des blocs annuels), ce qui n'est généralement pas le cas. Une alternative à la méthode des maxima par blocs consiste à conserver toutes les observations qui dépassent un niveau élevé puis à ajuster une loi appropriée à ces dépassements qui représente les événements " extrêmes ". Cette méthode est généralement appelée méthode des dépassements de seuil (ou " Peak Over Threshold ", POT).

2.5.2 Distribution des excés

Définition 2.5.1 [LEKINA (2010)]

On appelle excès de la variable aléatoire X au-delà d'un seuil $u < x_F$ la variable aléatoire Y, qui prend ses valeurs sur $]0\ ,\ x_F - u[$, définie par :

$$Y = X - u\ |X > u, \qquad u < x_F$$

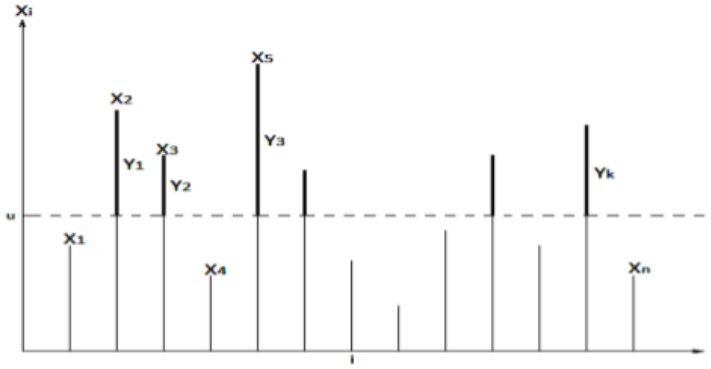

FIGURE 2.5 – Les données $X_1, X_2, ..., X_n$ et leurs k excès au-delà du seuil u correspondants $Y_1, Y_2, ..., Y_k (k \leqslant n)$.

Définition 2.5.2 *(Distribution des excès)*

On appelle distribution des excès de la variable aléatoire X par rapport à un seuil $u < x_F$ la loi de probabilité de la variable aléatoire Y excès de X au-delà du seuil $u < x_F$, donnée par sa fonction de répartition F_u, qu'on appelle fonction de distribution des excès, suivante :

$$\forall\, y \in \mathbb{R}, \qquad F_u(y) = \mathbb{P}(X - u \leqslant y \,|X > u) = \begin{cases} 0 & si\ \ y \leqslant 0 \\ 1 - \dfrac{1 - F(u+y)}{1 - F(u)} & si\ 0 < y < x_F - u \\ 1 & si\ y \geqslant x_F - u \end{cases}$$

Définition 2.5.3 [EMBRECHTS, KLUPPELBERG & MIKOSCH]

On appelle fonction moyenne des excès de la variable aléatoire X par rapport au seuil $u < x_F$, et on l'a note $e(u)$, la fonction espérance de la variable aléatoire Y excès de X au-delà du seuil $u < x_F$, définie par :

$$\forall\ u\ < x_F, \qquad e(u) = E\,(X - u \,|X > u) = \frac{1}{\bar{F}(u)} \int_u^{x_F} \bar{F}(t)\, dt$$

Notation 2.5.1 *Dans toute la suite de ce chapitre, on note F_u la fonction de distribution des excès donnée dans la définition ci-dessus.*

2.5.3 Distribution de Paréto généralisée

Définition 2.5.4 [Embrechts, Kluppelberg & Mikosch (1997)]

Soit $\gamma \in \mathbb{R}$. On appelle distribution de Paréto *généraliée standard toute fonction de répartition G_γ ou toute loi de probabilité qui a G_γ comme fonction de répartition telle que :*

$$\forall\, x > 0 \text{ tel que } \quad 1+\gamma x > 0, \qquad G_\gamma(x) = \begin{cases} 1-(1+\gamma x)^{-1/\gamma} & si\ \gamma \neq 0 \\ 1-e^{-x} & si\ \gamma = 0 \end{cases}$$

Pour les autres valeurs de x, on peut trouver G_γ puisque c'est une fonction de répartition.

Remarque 2.5.1 *On peut donner une forme plus générale à la fonction de répartition G_γ donnée dans la définition ci-dessus, qu'on note $G_{\gamma,\mu,\sigma}$, en faisant apparaître un paramètre de position $\mu \in \mathbb{R}$ et un paramètre d'échelle $\sigma > 0$ comme suit :*

$$\forall\, x > \mu \ \text{ tel que } \quad 1+\gamma\left(\frac{x-\mu}{\sigma}\right) > 0,\ G_{\gamma,\mu,\sigma}(x) = \begin{cases} 1-\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\gamma} & si\ \gamma \neq 0 \\ 1-\exp\left(\frac{x-\mu}{\sigma}\right) & si\ \gamma = 0 \end{cases}$$

On appelle $G_{\gamma,\mu,\sigma}$ « distribution de Paréto *généralisée (GPD) »*

Définition 2.5.5 [Lekina (2010)]

Le paramètre $\gamma \in \mathbb{R}$ qu'on voit donc la distribution de Paréto *généralisée est un paramètre de forme qu'on appelle « indice de queue ».*

Proposition 2.5.1 [Embrechts, Klüppelberg & Mikosch (1997)]

Si W est une variable aléatoire qui a comme fonction de répartition une distribution de Paréto *généralise $G_{\gamma,\sigma}$ $(\gamma < 1,\ \sigma > 0)$, alors sa fonction moyenne des excés $e(u)$ au-delà d'un seuil $u < w_0$ (w_0 est le point terminal de $G_{\gamma,\sigma}$) est donnée par :*

$$\forall\, u < w_0, \qquad e(u) = E(W-u\,|W>u) = \frac{\sigma+\gamma u}{1-\gamma} \qquad avec\ \ \sigma+\gamma u > 0$$

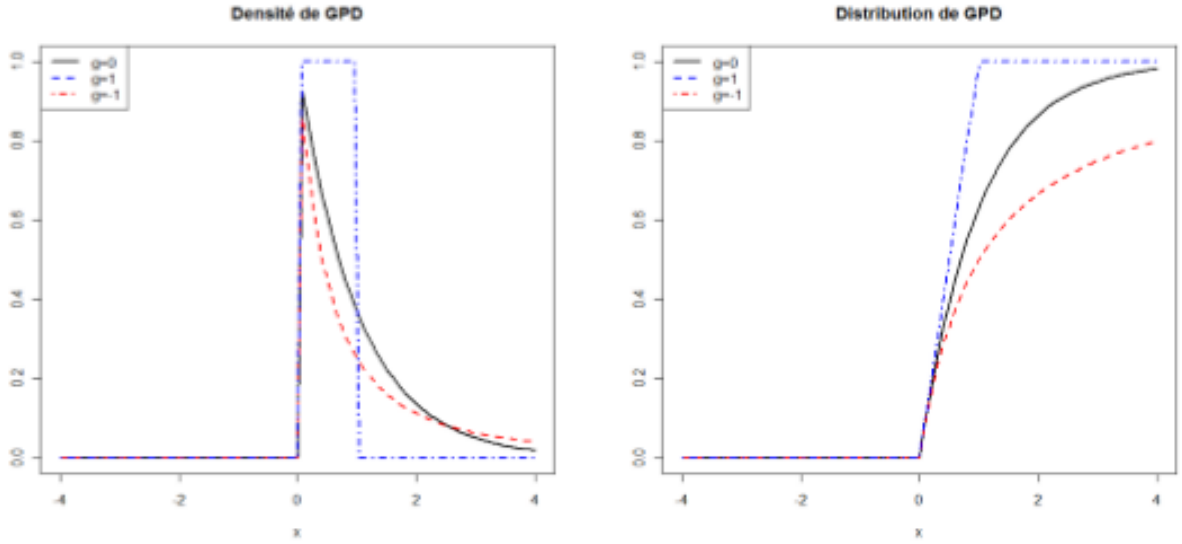

FIGURE 2.6 – Densités et distributions de loi de PARÉTO généralisée avec différentes valeurs de γ .

2.6 Distributions à variations régulières

Dans cette partie, on traite la classe C des fonctions qui apparat 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, peuvent consulter par exemple Bingham et al [6], Embrechts et al [18].

Définition 2.6.1 *(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 noté* $V \in \mathcal{R}v_\rho$, *si :*

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

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

- *Une fonction mesurable* $l :]a, +\infty[\to \mathbb{R}_+$ $(a \geqslant 0)$ *est dite à variation lente à l'infini, si :*

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

- *Une fonction à variation régulière d'indice* $\rho \in \mathbb{R}$ *peut toujours s'écrire sous la forme suivante :*

$$V(x) = x^\rho\, l(x), \qquad l \in \mathcal{R}v_0$$

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, en d'autre part :

- les fonctions x^ρ, $x^\rho \log(1+x)$, *et* $(x\log(1+x))^\rho$: sont à variation régulières .
- les fonctions $\exp(x)$, $\sin(x+2)$ *et* $\exp(\log(1+x))$ ne sont pas à variation 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 2.6.1 *(Propriétés de fonction à variation lente)*

- $\mathcal{R}\,v_0$ *est fermé sous l'addition, la multiplication et la division.*
- *Si* l *est à variation lente,* $\lim_{x\to\infty} (\log l(x)) / \log x = 0$.
- *Si* l *est à variation lente, alors* l^α *est à variation lente pour tout* $\alpha \in \mathbb{R}$.

❖ *Si l est à variation lente et $\rho > 0$,*

$$\lim_{x\to\infty} x^{\rho} l(x) = \infty, \qquad \lim_{x\to\infty} x^{-\rho} l(x) = 0$$

Théorème 2.6.1 [Représentation de Karamata]

❖ *$l \in \mathcal{R}v_0$ si et seulement si peut être représentée sous la forme :*

$$l(x) = c(x) \exp\left\{ \int_1^x \frac{r(t)}{t}\, dt \right\}, \qquad x > 0 \tag{2.8}$$

où c, r sont des fonctions mesurables, et

$$\lim_{x\to\infty} c(x) = c_0 \in [0, \infty[, \quad \text{et } \lim_{t\to\infty} r(t) = 0 \tag{2.9}$$

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

$$V(x) = c(x) \exp\left\{ \int_1^x t^{-1} \rho(t)\, dt \right\}, \qquad x > 0 \tag{2.10}$$

où c satisfait (2.9) et $\lim_{t\to\infty} \rho(t) = \rho$

Preuve. Voir Resnick [40] , Corollaire 2 :1 ; page 29 ■

Définition 2.6.2 *(Condition du second ordre)*

On dit que la fonction de queue de quantile U est à variation régulière du second ordre avec le paramètre $\gamma > 0$ et le paramètre du second ordre $\gamma \leq 0$, on écrit $U \in \mathcal{R}v_{\gamma,\rho}$, s'il existe une fonction $A^* \to 0$ et ne change pas le signe au voisinage de ∞, telles que :

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A^*(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho}, \qquad x > 0 \tag{2.11}$$

où $|A^*| \in \mathcal{R}v_{\rho}$ 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 2.6.1 *Pour tout $X > 0$ avec $\rho \leqslant 0$ et $A(t) = A^*(1/(1 - F(t)))$ la relation du second ordre est équivalente à :*

$$\lim_{t\to\infty} \frac{\overline{F}(tx)\big/\bar{F}(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma} \frac{x^{\rho} - 1}{\gamma\rho} \tag{2.12}$$

2.7 Domaine d'attraction

Définition 2.7.1 *(Domaine d'attraction)*

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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}_\gamma(x), \forall x \in \mathbb{R} \tag{2.13}$$

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

- *Si $\gamma > 0$, on dit que $F \in \mathcal{D}(\Phi_\gamma)$, 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 $F \in \mathcal{D}(\Psi_\gamma)$, 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 suivants donnent différents exemples de distributions standard dans ces trois domaines d'attraction :

Distributions	$\overline{F}(x)$	γ
$U[0\,,\,1]$	$1-x$	-1
Burr inversée $(\beta, \tau, \lambda, x_\tau)$; $\beta, \tau, \lambda > 0$	$\left(\frac{\beta}{\beta + (x_F + x)^{-\tau}}\right)^{\lambda}$	$-\frac{1}{\lambda}$

TABLE 2.1 – Tableaux de Quelques distributions associées à un indice négatif

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

TABLE 2.2 – Tableaux de Quelques distributions associées à un indice positif

Distributions	$\overline{F}(x)$	γ
Gamma (m, λ) $\lambda > 0, m \in \mathbb{N}$	$\frac{\lambda^m}{\Gamma(m)} \int_x^{\infty} u^{-m-1} Exp(-\lambda u) du$	0
Gumbel $(\mu, \beta), \beta > 0, \mu \in \mathbb{R}$	$Exp\left(-Exp\left(-\frac{x-\mu}{\beta}\right)\right)$	0
Logistic	$\frac{2}{1 + Exp(x)}$	0
Log gamma (μ, σ) $\mu \in \mathbb{R}, \sigma > 0$	$\frac{1}{\sqrt{2\pi}} \int_1^{\infty} \frac{1}{u} Exp\left(-\frac{1}{2\sigma^2} (\log u - u)^2\right) du$	0
Weibull (λ, τ) $\lambda > 0, \tau > 0$	$Exp(-\lambda x^{\tau})$	0

TABLE 2.3 – Tableaux de Quelques distributions associées à un indice nul

2.7.1 Caractérisation des domaines d'attraction

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

Théorème 2.7.1 *(Caractérisation du $\mathcal{D}(\Phi_\alpha)$)*

La fdr appartient au domaine d'attraction de la loi de Fréchet de paramètre $\alpha > 0$ si et seulement si :

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

où la fonction l est à variation lente. En particulier $x_F = +\infty$. de plus si $F \in d(\Phi_\alpha)$, avec $a_n = F^{-1}(1 - 1/n)$ et $b_n = 0$, la suite $\left(a_n^{-1} X_{n:n}\right)_{n \geqslant 1}$ converge en loi vers va de fdr Φ_α quand $n \to \infty$.

Preuve. Voir Embrechts et al [18], Théorème 3 :3 :7, page 131 ■

Théorème 2.7.2 *(Caractérisation du $\mathcal{D}(\Psi_\alpha)$)*

La fdr appartient au domaine d'attraction de la loi de Weibull de paramètre $\alpha > 0$ si et seulement si $x_F < \infty$ et :

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

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

Preuve. La démonstration du ce théorème est similaire à celui du théorème précédent, et Voir Embrechts et al [18], Théorème 3 :3 :7, page 131. 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.

Définition 2.7.2 *(Fonction de* VON MISES*)*

La fdr est dite fonction de VON MISES *avec la fonction auxiliaire a, s'il existe une certaine $z < x_F$ tel que :*

$$\overline{F}(x) = c \exp\left\{-\int_z^x \frac{dt}{a(t)}\right\}, \qquad z < x < x_F \leqslant \infty \tag{2.14}$$

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

Exemple 2.7.1 *(Distribution Exponentielle)*

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

F est une fonction VON MISES *avec la fonction auxiliaire $a(x) = \frac{1}{\lambda}$.*

Exemple 2.7.2 *(Distribution de* WEIBULL*)*

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

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

Théorème 2.7.3 *(Caractérisation du $\mathcal{D}(\Lambda)$)*

La fdr appartient au domaine d'attraction de la loi de GUMBEL *si et seulement si :*

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

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 suite de normalisation est :

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

2.8 Estimation de l'indice des valeurs extrêmes

Les deux estimateurs sans doute les plus populaires dans la littérature sont les estimateurs de Hill 1975 et de Pickands 1975.
On note $X_{1,n}, ..., X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$, c'est à dire que l'on classe $X_1, ..., X_n$ par ordre croissant de sorte que :

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

On considère les k valeurs les plus grandes (ou les plus petites). k dépend a priori de n , même si on ne le mentionnera pas dans la notation : l'idée est d'avoir $k \to \infty$ lorsque $n \to \infty$, mais sans prendre " trop " de valeurs de l'échantillon , ce qui conduit à imposer $\frac{k}{n} \to 0$. Incidemment , cela implique qu'on se posera la question du choix optimal de k.
En effet , il est indispensable de calculer cet estimateurs sur les queues de distribution .Choisir un k trop élevé engendre le risque de prendre en compte des valeurs qui ne sont pas extrêmes, à l'inverse, un sous-échantillon trop petit ne permet pas aux estimateurs d'atteindre leur niveau de stabilité.
Enfin, on retiendra que l'approche non paramétrique n'est envisageable que si l'on dispose d'un nombre important d'observations : dans le cas où les échantillons sont de petite taille, on se tournera vers l'approche paramétrique.

2.8.1 Estimation par des méthodes paramétriques

❶ Estimateur du maximum de vraisemblance

Cette méthode notée E.M.V est fréquemment utilisée en statistique. L'estimation par le maximum de vraisemblance donne des résultats asymptotiques efficaces, et les estimateurs obtenus convergent sous certaines conditions vers les vraies valeurs des paramètres.

Définition 2.8.1 *Soit $(Y_1, ..., Y_n)$ un n-échantillon, les Y_i sont supposées indpendantes et identiquement distribuées, de densité h_θ où $\theta = (\mu, \sigma, \gamma)$. L'expression de la fonction de vraisemblance est donnée par :*

$$L\left(\theta = (\mu, \sigma, \gamma); (Y_1, ..., Y_n)\right) = \prod_{i=1}^{n} h_\theta\left(y_i\right)$$

L'estimateur $\widehat{\theta}$ est donné par la résolution du système suivant :

$$\begin{cases} \dfrac{\partial \log L}{\partial \theta} = 0 \\ \dfrac{\partial^2 \log L}{\partial^2 \theta} < 0 \end{cases}$$

Exemple 2.8.1 *Dans le cas où $\gamma = 0$ (loi de* Gumbel *), la fonction log de la vraisemblance égale à :*

$$\log\ L\left(\theta = (\mu, \sigma, \sigma); (Y_1, ..., Y_n)\right) = -n \log \sigma - \sum_{i=1}^{n} \exp\left(-\frac{y_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma}$$

En dérivant cette fonction relativement aux deux paramètres, nous obtenons le système d'équations à résoudre suivant :

$$\begin{cases} \dfrac{\partial \log L}{\partial \sigma} = 0 \ \Leftrightarrow\ n + \displaystyle\sum_{i=1}^{n} \frac{y_i - \mu}{\sigma} \left[\exp\left(-\frac{y_i - \mu}{\sigma} - 1\right)\right] = 0 \\ \dfrac{\partial \log L}{\partial \mu} = 0 \ \Leftrightarrow\ n - \displaystyle\sum_{i=1}^{n} \exp\left(-\frac{y_i - \mu}{\sigma}\right) = 0 \end{cases}$$

La résolution de ce système est relativement difficile et n'admet pas en général de solutions explicites.

2.8.2 Estimation par des méthodes semi-paramétriques

❶ L'estimateur de Pickands

L'estimateur de Pickands a été introduit en 1975 par James Pickands [38] pour toute $\gamma \in \mathbb{R}$.

Définition 2.8.2 *(Estimateur de* Pickands*)*

Soit $X_1, ..., X_n$, n variable aléatoire i.i.d de $F \in \mathcal{D}(\Phi_{1/\gamma})$,où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, l'estimateur de Pickand *est défini par :*

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

L'auteur démontre la Convergence faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan [14]. Des améliorations de cet estimateur ont été introduites notamment par Drees [12].
Sous certaines conditions sur la suite entière k et la fdr F, l'estimateur de γ a des bonnes propriétés asymptotiques, elles sont regroupés dans le théorème suivant :

Théorème 2.8.1 *(Propriétés asymptotiques de $\hat{\gamma}^P$)*

Soit $F \in \mathcal{D}(\mathcal{H}_\gamma)$,$\gamma \in \mathbb{R}$, $k \to \infty$ et $\frac{k}{n} \to 0$, quand $n \to \infty$

- ❖ *Convergence en probabilité :*

$$\hat{\gamma}^P \xrightarrow{p} \gamma, \quad quand \ n \to \infty$$

- ❖ *Convergence forte (presque sûre) : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors*

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

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

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

où

$$v^2 = \frac{\gamma^2\left(2^{2\gamma+1}+1\right)}{\left(2(2^\gamma - 1)\log 2\right)^2}$$

Ce dernier (résultat de la normalité asymptotique) 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 : $\hat{\gamma}^P$ en fonction de k ; voir la Figure (2.7), 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.

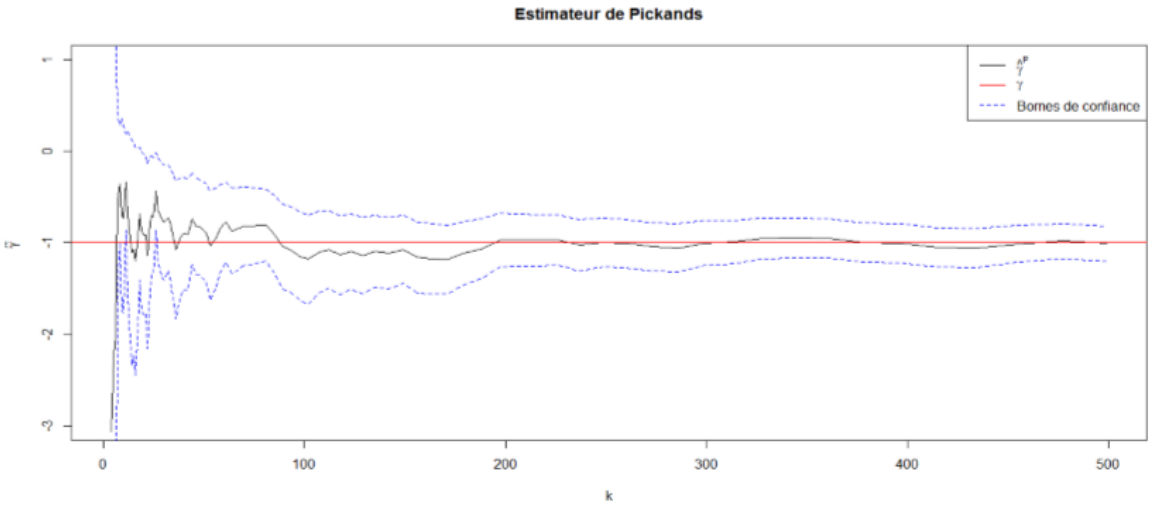

Figure 2.7 – Estimateur de Pickands avec intervalle de confiance au niveau 95% pour γ basés sur 1000 échantillons de taille 500 pour la loi uniform standard $\gamma = -1$.

❷ L'estimateur de Hill

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

Définition 2.8.3 *(Estimateur de* Hill*)*

Soit $X_1, ..., X_n$, n variable aléatoire i.i.d de $F \in \mathcal{D}(\Phi_{1/\gamma})$,où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, l'estimateur de Hill *est défini par :*

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

La construction de cet estimateur est donnée dans le livre de de Haan et Ferreira [23] et dans le livre de Beirlant et al [3]. D'autres estimateurs de l'IVE ont été proposés notamment par Beirlant et al [3] qui utilisent un modèle de régression exponentiel pour débiaiser l'estimateur de Hill et par Csorgo et al [8] 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 [32], et la consistance forte fut établie en 1988 par Deheuvels et al [16] et plus récemment par Necir [36].

Théorème 2.8.2 *(Propriétés asymptotiques de $\hat{\gamma}^H$)*
Soit $F \in \mathcal{D}(\mathcal{H}_\gamma)$, $\gamma \in \mathbb{R}$, $k \to \infty$ et $\frac{k}{n} \to 0$, quand $n \to \infty$

- ❖ *Convergence en probabilité :*

$$\hat{\gamma}^H \xrightarrow{p} \gamma, \quad quand \ n \to \infty$$

- ❖ *Convergence forte (presque sûre) : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors*

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

- ❖ *Normalité asymptotique : On Suppose que F satisfaisant la Condition du second ordre . Si $\sqrt{k}\, A\,(n/k) \to \lambda$, quand $n \to \infty$, alors :*

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

Dans le cas général du domaine de FRÉCHET , la fonction de survie est de forme $S(x) = 1 - F(x) = x^{-1/\gamma} l(x)$ avec l une fonction à variation lente . Cela induit un biais important sur l'estimateur de HILL, qui est donc en pratique d'un maniement délicat dans le cas général.
La Figure (2.8) illustre le graphe de $\hat{\gamma}^H$ 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.

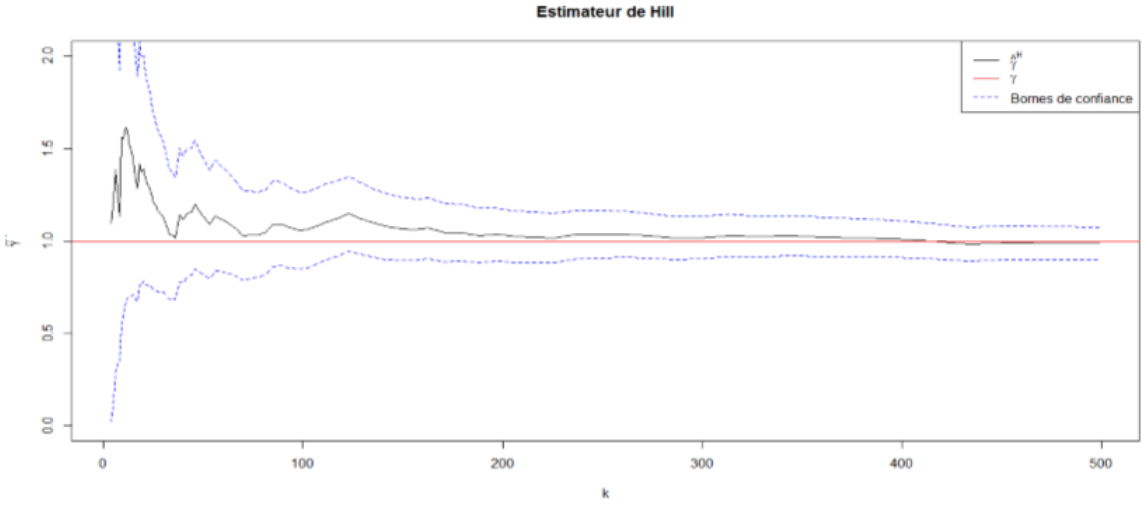

FIGURE 2.8 – Estimateur de HILL avec l'intervalle de confiance au niveau 95% pour l'IVE de la loi de Paréto standard $\gamma = 1$ basés sur 1000 échantillons de taille 500 observation .

❸ L'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 [15] ont donné une extension de tout type de distribution, appelée estimateur des moments.

Définition 2.8.4 *(Estimateur des* Moments*)*

Pour $\gamma \in \mathbb{R}$*, l'estimateurs des* Moments *est donner par :*

$$\hat{\gamma}^M = \hat{\gamma}^M(k) = M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1} \tag{2.15}$$

avec

$$M_n^{(r)} = M_n^{(r)}(k) = \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{n-i+1:n} - \log X_{n-k:n}\right)^r, \quad r = 1, 2, \tag{2.16}$$

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

Les propriétés asymptotiques de cet estimateur ont été étudiées dans Dekkers et al [15].

Théorème 2.8.3 *(Propriétés asymptotiques de* $\hat{\gamma}^M$*)*

Soit $F \in \mathcal{D}(\mathcal{H}_\gamma)$, $\gamma \in \mathbb{R}$, $k \to \infty$ *et* $\frac{k}{n} \to 0$, *quand* $n \to \infty$

- ❖ *Convergence en probabilité :*

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

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

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

- ❖ *Normalité asymptotique : si* $k = o\left(n \big/ g_1^{-1}(n)\right)$, $g_1(t) = t(U(t)/a(t))^2$, *alors*

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

avec

$$v^2 = \begin{cases} 1 + \gamma^2 & si\ \gamma \geqslant 0 \\ (1-\gamma)^2(1-2\gamma)\left[4 - 8\dfrac{1-2\gamma}{1-3\gamma} + \dfrac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right] & si\ \gamma < 0 \end{cases}$$

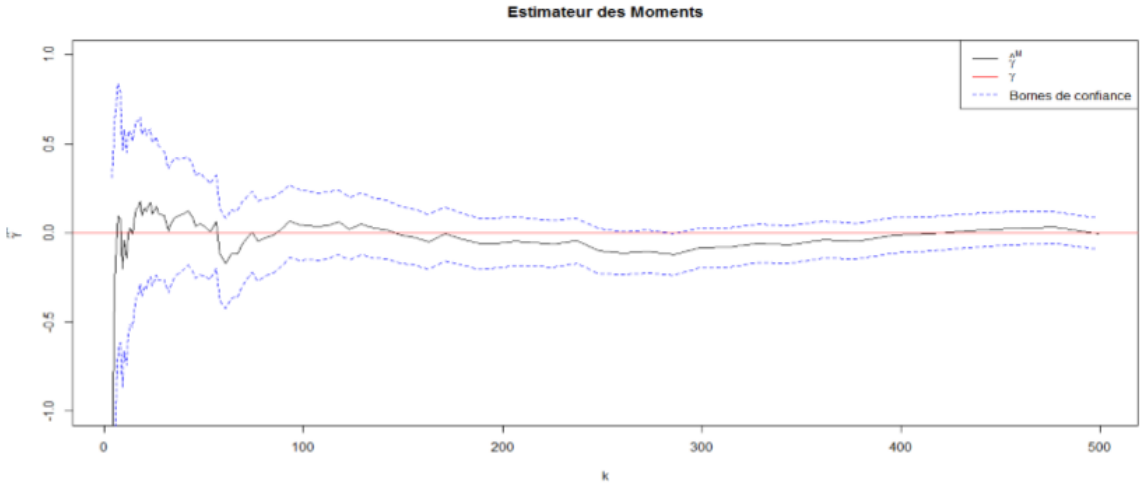

FIGURE 2.9 – Estimateur de MOMENTS avec l'intervalle de confiance au niveau 95% pour l'IVE de la loi de Gumbel $\gamma = 0$ basés sur 1000 échantillons de taille 500 observation .

❹ Choix du nombre k

Le nombre k de la statistique d'ordre est difficile à choisir. Les résultats concernant les estimateurs de l'indice des valeurs extrêmes sont asymptotique lorsque $k \to \infty$ et $k/n \to 0$. Comme en pratique, on ne dispose que d'un nombre d'observation n fini, il s'agit de choisir k de manière à ce que l'on dispose de suffisamment de données statistiques tout en restant dans la queue de distribution.

◆ Méthode graphique

C'est une méthode la plus simple pour la détermination de k. Elle consiste à tracer le graphe $\left(k\,,\,\widehat{\gamma}^{H}_{k_n\,,n}\right)$ avec $k_n = k$ une suite d'entiers et $1 < k < n$.
et de prendre la valeur où $\left(k\,,\,\widehat{\gamma}^{H}_{k_n\,,n}\right)$ devient horizontal. cet estimateur est valable seulement dans le domaine d'attraction de FRÉCHET c'est à dire $\gamma > 0$. pour généraliser aux autres domaines d'attraction, différents estimateurs ont été proposés, entre autres l'estimateur de PICKANDS.

◆ Méthode analytique

Il est nécessaire pour donner une précision à l'estimateur $\left(k\,,\,\widehat{\gamma}^{H}_{k_n\,,n}\right)$, de calculer l'erreur moyenne quadratique (MSE), elle est en fonction de k

$$\begin{aligned} MSE\,(\widehat{\gamma}_{k_n\,,n}) &= MSE(\widehat{\gamma}_{k_n\,,n} - \gamma)^2 \\ &= biais^2\,(\widehat{\gamma}_{k_n\,,n}) + \text{var}\,(\widehat{\gamma}_{k_n\,,n}) \end{aligned}$$

Le choix optimal de k, correspond à minimiser MSE. Concernant l'estimateur de Hill pour des fonctions appartenant au domaine d'attraction maximale de FRÉCHET, de

HAAN ET PENG en 1998 ont proposé de retenir le nombre d'observation k_{opt} qui minimise l'erreur moyenne quadratique de l'estimateur de HILL qui est :

$$k_{opt} = \begin{cases} 1 + 2^{2\gamma/(\frac{2\gamma+1}{2\gamma})} \left(\dfrac{(\gamma+1)^2}{2\gamma} \right)^{1/(2\gamma+1)} & si \;\; 0 < \gamma < 1 \\ 2n^{2/3} & si \;\; \gamma > 1 \end{cases}$$

◆ **Méthode numérique**

Il existe plusieurs algorithmes pour trouver un estimateur $\widehat{k}_{opt}$ de k_{opt}

$$\frac{\widehat{k}_{opt}}{k_{opt}} \xrightarrow{p} 1 \qquad quand \;\; n \to \infty$$

alors $\widehat{\gamma}_{\left(\widehat{k}_{opt\,,\,n}\right)}$ converge asymptotiquement vers $\gamma_{(k_{opt\,,\,n})}$

Remarque 2.8.1

➥ *Si k est petit, $\widehat{\gamma}_{(k\,,\,n)}$ utilise peu d'observation et on a une grande variance.*

➥ *Si k est grand, le biais est grand, la variance est petite.*

2.9 Conclusion

Dans ce chapitre, nous avons fait un aperçu général sur la théorie des valeurs extrêmes, en mentionnant les différentes caractéristiques et les notions de base qui sont très utiles pour l'estimation des quantiles extrêmes et les données censurées que nous allons les aborder dans le chapitre suivant.

3

Données censurées

Sommaire

3.1 Introduction

Lorsque l'on s'intéresse par exemple à l'étude de la survenue au cours du temps d'un événement « en tout ou rien » comme le décès (mais aussi, la récidive tumorale ou l'apparition de métastases, etc.), on désigne souvent ces données sous le terme générique de « données de survie ». La particularité de ces données, c'est qu'à la fin de la période d'observation, l'événement d'intérêt (ici le décès) ne sera probablement pas survenu pour tous les patients. Pour ces patients, le temps de survie est dit « censuré » (à droite), indiquant que le délai exact de décès du sujet (non observé) est supérieur ou égal « à droite » à son délai de suivi. Nous ne savons pas quand et si l'événement se produira mais à la date où sont analysées les données, le patient est toujours vivant.

On dit parfois que ces sujets sont des « exclus vivants ». L'autre mécanisme principal de censure concerne les patients dits « perdus de vue », c'est-à-dire ceux dont le suivi s'interrompt avant la date de point, de manière inopinée (du fait d'un déménagement ou de changement de filière médicale, par exemple). Pour ces derniers sujets, le temps de

survie sera également censuré puisque la période d'observation s'est arrêtée à la dernière date où l'on savait que le patient était vivant. Dans ce cas une hypothèse importante est que la raison du départ des patients de l'étude doit être indépendante de leur risque de décès. C'est-à-dire qu'à chaque temps, les patients censurés ont la même perspective de survie que ceux qui continuent d'être suivis. En d'autres termes, si le patient est perdu de vue du fait d'une altération de son état de santé, l'indépendance entre la cause de censure et le décès ne peut plus être assurée. On parle de censure « informative » (ou dépendante du décès). Ceci est important à vérifier car les méthodes qui seront exposées plus loin ne sont valides qu'en cas de censure dite « non-informative » (ou indépendante du décès).

Enfin la dernière notion associée aux données censurées est celle de censure à droite ou à gauche. Dans toutes les situations identiques à l'exemple précédent, où l'on sait seulement que l'événement ne s'est pas produit à une certaine date, on a vu que le délai de survie constitue une observation dite censurée à droite. En effet la durée de survie est supérieure à un délai donné. Mais il peut arriver que l'événement se soit produit avant la date de point sans qu'il soit possible d'en connaître la date exacte. L'observation est dite censurée à gauche. C'est-à-dire que le véritable délai de survie du patient est inférieur au délai d'observation. Le plus souvent on se trouve dans les conditions de censure à droite et c'est ce cas qui sera traité dans ce chapitre.

3.2 Censure et troncature

En général, pour une suite de données $x_1, x_2, ..., x_n$ réalisation de la suite de *va* $X_1, X_2, ..., X_n$ (suites de durées dans un état du système), l'observateur n'accède pas à tous les X_i, $i = 1, ..., n$ à cause d'un empêchement, d'un arrêt des observations, d'une complication à l'accès à l'information. Le statisticien utilise un modèle pour chaque situation. Pour les modèles de durée, les plus utilisés sont généralement ceux de censure et ceux de troncatures. La censure correspond à l'introduction d'une variable concurrente Y qu'on appelle variable de troncature. Cette variable peut-être fixé ou aléatoire. L'analyse des données censurées consiste à étudier la loi de la durée X et Y en tenant compte des données non observées. La troncature agit comme la censure, mais ne prend pas en considération les données non observées (c'est-a'-dire ayant subi la troncature) ; la loi de la durée n'est, dans ce dernier cas, qu'une loi conditionnelle sur la partie non tronquée. Une donnée est dite censurée si elle n'est pas observée au delà d'un certain seuil (prédéfini ou aléatoire), mais que l'information que le phénomène n'a pas été observé au delà du seuil est prise en considération. Si l'information que le phénomène n'a pas été observé au delà du seuil n'est pas prise en considération, on est devant le cas de données tronquées.

3.2.1 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 l'apparition d'observations incomplètes dites censurées. La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Définition 3.2.1 *(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 :

- *Temps de survie X_j*
- *Son temps de censure Y_j*
- *La durée réellement observée Z_j*

3.2.2 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).

❶ **Censure à droite** $Z_j = \min(X_j, Y_j)$

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 considère 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 d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date qui seront censurées à droite.

❷ **Censure à gauche** $Z_j = \max(X_j, Y_j)$

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.

❸ **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 modeles non-paramétriques ont été présentés pour l'étude de la double censure. Par exemple, le modèle de Turnbull [48] 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.

❹ **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 à 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 $]-\infty\,,\,a]$ 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 l'expérience). Par exemple en épidémiologie on fixe la durée maximale de participation et vaut, pour chaque observation, la différence entre la date de l'expérience et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quand à lui, aléatoire.
Soit Y une valeur fixée. Par exemple en censure à droite, au lieu d'observer les variable $X_1, X_2, ..., X_n$ qui nous intéressent, on observe X_j que lorsqu'elle est inférieure ou égale une durée fixée Y , On observe donc une variable Z_j telle que $Z_j = \min(X_j\,,\, Y)\,, \quad j = 1, ...n$

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

◆ **Censure de type II : attente**

L'expérimentateur fixe à priori le nombre d'événements à observer. La date 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 $(1 \leqslant r \leqslant n)$ 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.

$$\begin{cases} Z_{j:n} = X_{j:n} & si\ j \leqslant r \\ Z_{j:n} = X_{r:n} & si\ j \geqslant r \end{cases}$$

◆ **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 des n d'observations est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient).
Soit $X_1, X_2, ..., X_n$ un échantillon d'une *va* positive X, on dit qu'il y a censure aléatoire de cette échantillon s'il existe une autre *va* positive elle aussi Y d'échantillon $Y_1, Y_2, ..., Y_n$ dans ce cas au lieu d'observer les X_j ; on observe un couple de *va* (Z_j, δ_j) avec

$$Z_j = \min(X_j, Y_j) \quad et\ \ \delta_j = \mathbb{I}\{X_j \leqslant Y_j\} \qquad pour\ j = 1, ..., n \tag{3.1}$$

où j est l'indicateur de censure, qui détermine si X a été censurée 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.

3.2.3 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, quand 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 .

❶ **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 donc le couple (X, Y), avec $X > Y$.

❷ **Troncature à droite**

De même, il y a troncature à droite lorsque X n'est observable que si $X > Y$.

❸ **Troncature par intervalle**

Quand une durée est tronquée à droite et à gauche, on dit quelle est tronquée par intervalle.

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

3.3.1 Estimateur de Kaplan-Meier

Soit $(\Omega, \mathrm{A}, \mathbb{P})$ un espace probabilisé. Soit $X_1, ..., X_n$ une suite de variable aléatoire $i.i.d$ positives, de fdr commune F et $Y_1, ..., Y_n$ une suite de *va* de censure iid positives, de fdr continue G. On suppose aussi que ces variables sont indépendantes des X_j, Soit $\{(Z_j, \delta_j), \quad 1 \leqslant j \leqslant n\}$ l'échantillon réellement observé défini par (3.1), dans la suite on supposera que la variable Z a comme fonction de répartition H.
Malheureusement, en présence de censure, la fonction de survie empirique de la variable X n'est plus valable car elle dépend d'une *va* 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 en présence de données censurées. En 1958, Kaplan et Meier [28] ont introduit les estimateurs non-paramétriques du maximum de vraisemblance de F et G (voir, par exemple, dans Deheuvels et al [16] et Einmahl et al [17]).

Définition 3.3.1 *(L'estimateur de* KAPLAN-MEIER*)*

$\{(Z_j, \delta_j)\,,\ \ 1 \leqslant j \leqslant n\}$ *l'échantillon réellement observé défini par* (3.1)*. L'estimateur de* KAPLAN-MEIER *est défini par :*

$$\widehat{\overline{F_n}}(t) = \begin{cases} \prod\limits_{Z_{j:n} \leqslant t} \left(\dfrac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} & \text{pour } t < Z_{n:n} \\ 0 & \text{pour } t \geqslant Z_{n:n} \end{cases} \tag{3.2}$$

et

$$\widehat{\overline{G_n}}(t) = \begin{cases} \prod\limits_{Z_{j:n} \leqslant t} \left(\dfrac{n-j}{n-j+1}\right)^{1-\delta_{[j:n]}} & \text{pour } t < Z_{n:n} \\ 0 & \text{pour } t \geqslant Z_{n:n} \end{cases} \tag{3.3}$$

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

On présente ici une autre écriture de l'estimateur de KAPLAN ET MEIER [28] sous forme de somme. Cet écriture peut être trouvée dans le livre de REISS ET THOMAS [41] , page 16.

$$\widehat{F}_n(t) = 1 - \widehat{\overline{F}}_n(t) = \sum_{i=2}^{n} W_{i:n} \mathbb{I}\{Z_{i:n} \leqslant t\} \tag{3.4}$$

Par des raisonnements combinatoires, STUTE [43] et WANG [53] obtiennent l'expression suivante des sauts de l'estimateur de KAPLAN ET MEIER [28].

$$W_{i:n} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} \tag{3.5}$$

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

Remarque 3.3.1 *Les estimateurs* (3.2) *et* (3.3) *sont parfois écris de la manière suivante :*

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

et

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

Remarque 3.3.2

- *L'expression (3.2) définit une fonction constante par morceaux, continue à droite avec limite à gauche.*
- *Les points de discontinuité de cette fonction correspondent aux observations non-censurées.*
- *L'auteur des sauts de de $\widehat{F}_n(t)$ est aléatoire.*
- *En l'absence de censures, on retrouve la fonction de survie empirique (1.2).*

Dans la littérature de l'analyse de survie, un grand nombre d'auteurs ce sont consacrés à l'étude des propriétés asymptotiques de l'estimateur de Kaplan et Meier [28]. Par exemple, la consistance uniforme a été étudiée par Shorack et Wellner [47] , Wang [53] , Stute et Wang [46] et plus récemment par Gill [22], et la normalité asymptotique a été étudiée par Breslow et Crowley [5].

Proposition 3.3.1 *(Propriétés asymptotiques de $\widehat{\overline{F}}_n$)*

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

$$E\left[\widehat{\overline{F}}_n(t)\right] = \overline{F}(t), \qquad \text{quand } n \to \infty$$

- *Consistance uniforme : soit $x_H = H^{-1}(1) = \inf\{t : H(t) = 1\} \leqslant \infty$ Alors :*

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

- *Normalité asymptotique : pour tout $t \geq 0$, on a :*

$$\sqrt{n}\left|\widehat{\overline{F}}_n(t) - \overline{F}(t)\right| \xrightarrow{L} \mathbb{X}_t$$

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

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

Définition 3.3.2 *(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 :

$$\widehat{Q}_n(s) = \widehat{F}_n^{-1}(s) = \inf\left\{t : \widehat{F}_n^t(t) \geqslant s\right\} = Z_{i:n} \tag{3.6}$$

où

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

3.3.2 Estimateur de Nelson-Aalen

Kaplan et Meier 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 [35] et généralisé par Aalen [1]. On observe, tout d'abord, que sous l'hypothèse générale d'indépendance entre X et Y, on peut décomposer H de la manière suivante :

$$H(t) = 1 - (1 - F(t))(1 - G(t)) = H^{(0)}(t) + H^{(1)}(t) \tag{3.7}$$

où

$$H^0(t) = P(Z \leqslant t,\ \delta = 0) = \int_0^t \overline{F}(x)\, dG(x) \tag{3.8}$$

et

$$H^1(t) = P(Z \leqslant t,\ \delta = 1) = \int_0^t \overline{G}(x)\, dF(x) \tag{3.9}$$

$H(t) = P(T > t)$ *et* $H^1(t) = P(T > t, \delta = 1)$. Pour $t \geq 0$; la fonction de hasard cumulée peut s'exprimer de la manière suivante :

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

Définition 3.3.3 *(Estimateur de* NELSON-AALEN*)*

L'estimateur non-paramétrique de NELSON-AALEN Λ_n *de* Λ *basé sur l'échantillon* $\{(Z_j, \delta_j), \quad 1 \leqslant j \leqslant n\}$ *est défini par :*

$$\Lambda_n(t) = \int_0^t \frac{dH_n^{(1)}(x)}{\overline{H}_n(x)} = \begin{cases} \sum\limits_{Z_{j:n} \leqslant t} \dfrac{\delta_{[j:n]}}{n-j+1} & si \ \ t < Z_{j:n} \\ 1 & si \ \ t \geqslant Z_{j:n} \end{cases} \tag{3.10}$$

où

$$H_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{Z_j \leqslant t\} \ \text{et} \ H_n^{(1)}(t) = \frac{1}{n}\sum_{j=1}^{n} \delta_j \mathbb{I}\{Z_j \leqslant t\}$$

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

Remarque 3.3.3 *En remplaçant* $\Lambda(t)$ *par* $\Lambda_n(t)$ *dans (1.9) on obtient un nouveau estimateur de la fonction de survie F*

$$\widehat{\overline{F}}_n^{NA}(t) = \begin{cases} \prod\limits_{Z_{j:n} \leqslant t} \exp\left\{-\dfrac{\delta_{[j:n]}}{n-j+1}\right\} & pour \ \ t < Z_{n:n} \\ 0 & pour \ \ t \geqslant Z_{n:n} \end{cases} \tag{3.11}$$

connu sous le nom de l'estimateur de BRESLOW *[5].*

FLEMING ET HARRINGTON [20] ont montré la relation étroite entre les estimateurs de NELSON [35] , AALEN [1] et de KAPLAN-MEIER [28], ils ont comparé numériquement pour plusieurs tailles d'échantillons et ont souligné que les deux estimateurs, sont asymptotiquement équivalents. Une très instructive discussion sur ce point peut être trouvée dans un article plus récent de HUANG ET STRAWDERMAN [26].

3.4 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 $v.a$ X de fdr est définie par :

$$\mu = E[X] = \int t\, dF(t).$$

Si la $v.a$ X est positive, une intégration par parties, on peut réécrire la moyenne comme :

$$\mu = \int \overline{F}(t)\, dt \tag{3.12}$$

Dans le cas les données complétés, l'estimateur non-paramétrique de la moyenne $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$. Il est obtenu par substitution de $\overline{F_n}$ à la place de $\overline{F}$ dans (3.7) Il est sans biais, consistant et sous la condition $E(X^2) = \int t^2\, dF(t) < \infty$ 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 [43] a introduit un estimateur qui s'appelle l'intégrale 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

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

où $W_{i:n}$ défini par (3.5). Pour $\varphi(t) = t$ on obtient $S^{\varphi} = \mu$.

Définition 3.4.1 *(Moyenne empirique sous censure aléatoire)*

L'estimation non-paramétrique de la moyenne sous censure aléatoire est défini par :

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

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

$$I_1 = \int_0^{\infty} x^2\, \Gamma_0^2(x)\, dH^{(1)}(x) < \infty \tag{3.13}$$

et

$$I_2 = \int_0^{\infty} x \left(\int_0^{x} \frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2} \right)^{1/2} dF(x) < \infty \tag{3.14}$$

Où $H^{(0)}$ *et* $H^{(1)}$ sont deux fonctions définies par (3.8) et (3.9) avec :

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

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

Théorème 3.4.1 *(TCL sous censure aléatoire)*

Sous (3.13) et (3.14), on a :

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

Où $\sigma^2 = \text{var}\left[Z_1\Gamma_0(Z_1)\,\delta_1 + \Gamma_1(Z_1)\,(1-\delta_1) - \Gamma_2(Z_1)\right]$

Preuve. Voir Stute [43] , Corollaire (1.2) ■

3.5 Estimation de l'IVE avec censure

Des techniques statistiques pour analyser les ensembles de données censurées sont maintenant très bien étudiées, 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, les premiers qui ont mentionné le sujet sont Beirlant et al [3]. et Reiss et Thomas [41] , mais sans résultats asymptotiques. Puis certains estimateurs des paramètres de la queue ont été proposées par Beirlant et Guillou [7], pour les données tronquées et étendu la censure aléatoire à droite par Beirlant et al [3] . et l'année suivante par Einmahl et al [17].
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 variable aléatoire $i.i.d$ 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 $\{(Z_j\,,\,\delta_j)\,,\ 1 \leqslant j \leqslant n\}$ l'échantillon réellement observé défini par (3.1). Il est clair que les Z_j sont des variables indépendantes de loi H liée à F et G par la relation (3.7). 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, etH respectivement. Einmahl et al [17]. ont fourni une adaptation générale des estimateurs existants de L'IVE dans les cas suivants :

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

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

$$\widehat{\gamma}_1^{(\bullet,c)} = \widehat{\gamma}_1^{(\bullet,c)}(k) = \frac{\widehat{\gamma}^{\bullet}}{\widehat{p}}\ , \quad et \quad \widehat{p} = \widehat{p}\,(k) = \frac{1}{k}\sum_{i=1}^{k}\delta_{[n-i+1\,:\,n]} \tag{3.15}$$

$\widehat{\gamma}^{\bullet}$ 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 droite de distribution, avec $k = k_n$ satisfaisant $\lim\limits_{n\to\infty} k_n = \infty$ et $\lim\limits_{n\to\infty} k_n/n = 0$. Beirlant et al [3] . 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 [17] . ont adapté différents estimateurs de l'IVE au cas où les données sont censurées par un seuil aléatoire et ont proposé une méthode unifiée pour établir leur normalité asymptotique.

Einmahl et al [17].ont prouvé que, si $\widehat{\gamma}^{\bullet}$ est 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 [17]. ont également établi la consistance de $\widehat{p}$ sous la condition du premier ordre sur les fonction F et G. Ils ont aussi conclu, que l'estimateur $\widehat{\gamma}_1^{(\bullet,c)}$ de γ_1 est consistant pour l'estimateur de Hill. En outre, Brahimi et al [4] ont utilisé la théorie des processus empiriques pour approcher l'estimateur de Hill adaptée en termes de processus Gaussiens. Ndao et al [34]. ont adressée l'estimation non paramétrique de l'IVE conditionnel et son quantile pour les distributions queue lourde qui ont été récemment généralisées par Stupfler [45] pour les trois domaines d'attraction des extrêmes. Dans le même contexte, Worms et Worms [50], ont présenté une nouvelle approche, basée sur l'intégration de Kaplan-Meier. Cette dernière approche est de définir un estimateur pour l'indice de queue positif et prouver sa consistance. Récemment, Beirlant et al [3] . ont utilisé un modèle de type Paréto 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 [3]. en 2007 et par Einmahl et al [17]. en 2008. Il est donné par la définition suivante :

Définition 3.5.1 *(Estimation du quantile extrême sous censure aléatoire)*

L'estimation du quantile extrême sous censure aléatoire est donner par :

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

où $\widehat{a}^{(\bullet,c)} = Z_{n-k:n} M_n^{(1)}\left(1-T_n\right)\Big/\widehat{p}$, *avec* $M_n^{(1)}$ *est l'estimateur de* Hill $\widehat{\gamma}_n^H$

3.6 Conclusion

Dans ce chapitre, nous avons fait un rappel les valeurs extrêmes en présence de censure, et quelques estimations (estimation de la fonction de survie et taux de hasard,...) et les estimateurs de l'IVE avec censure.
Dans le chapitre suivant nous allons faire quelques simulations puis traiter une application sur des données réelles.

4 Simulations et étude de données réelles

Sommaire

4.1 Introduction

Il est étudié dans cette partie du mémoire le comportement de l'estimateur de Hill, en premier lieu, en fonction de la taille de l'échantillon générée ($n = 200, 500, 850, 1000$) par une loi de Paréto de paramètre de forme égale à 5 et une loi de censure de Paréto de paramètre de forme égale à 1, et en deuxième lieu, cette étude se focalisera sur le comportement de l'estimateur de Hill suivant le taux de censure noté T_c dans ce document ($T_c = 50\%, 25\%, 10\%, 5\%$).

4.2 Simulations

4.2.1 Fonction de survie par Kaplan-Meier

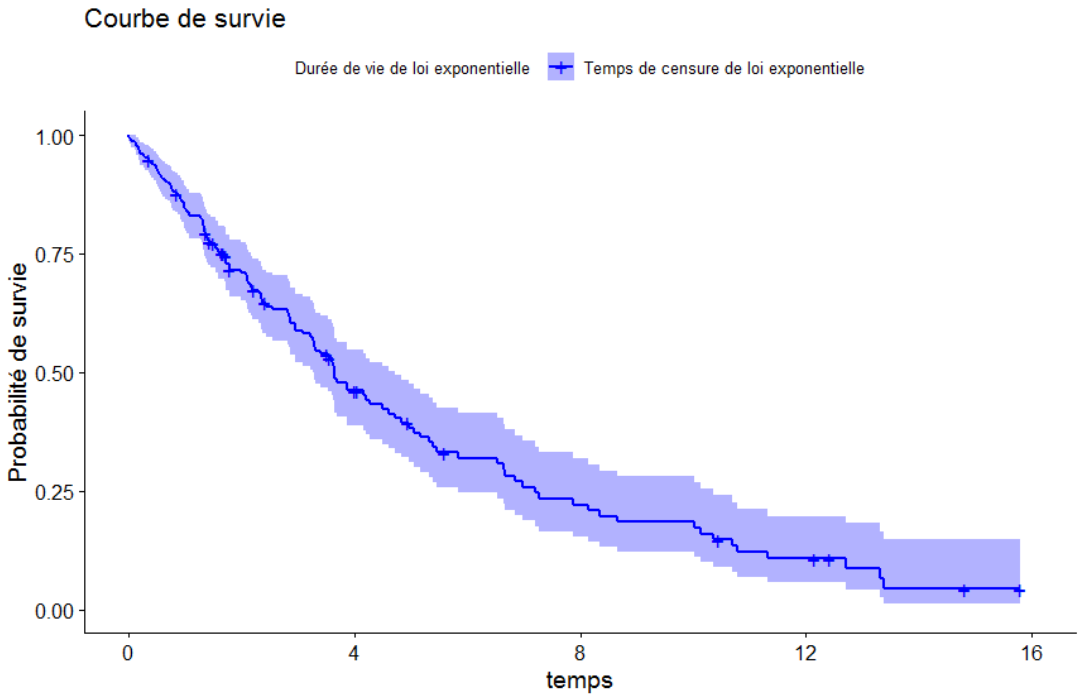

FIGURE 4.1 – Courbe de survie par l'estimateur de KAPLAN-MEIER d'une loi exponenetielle de paramètre $\lambda = 0.2$ censurée par une loi exponenetielle de paramètre $\lambda = 0.2$.

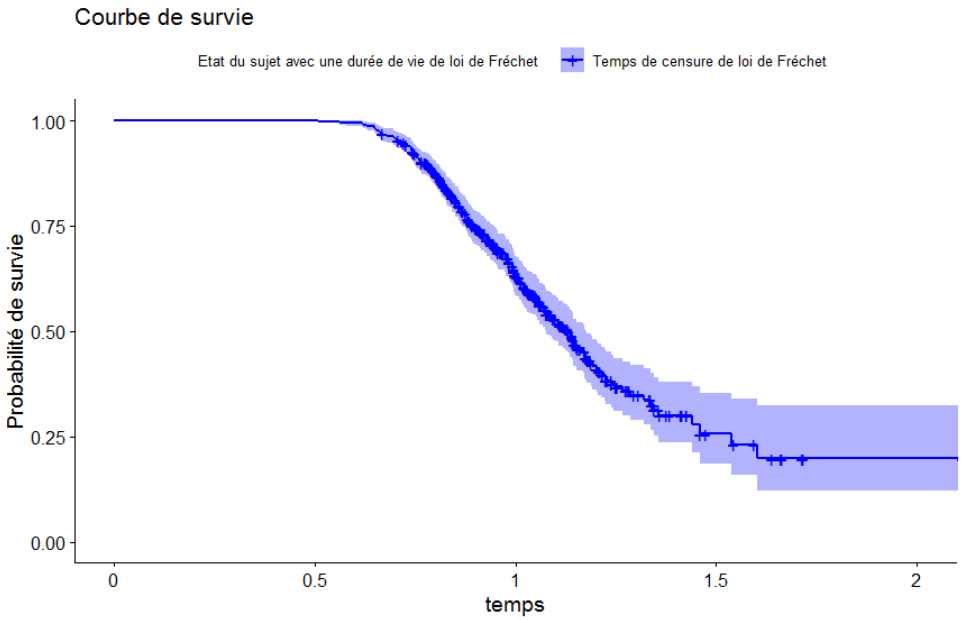

FIGURE 4.2 – Courbe de survie par l'estimateur de KAPLAN-MEIER d'une loi de FRÉCHET de paramètre de forme $\lambda = 3$ censurée par une loi de FRÉCHET de paramètre de forme $\lambda = 5$.

4.2.2 Comportement de l'estimateur de Hill en présence de censure

❶ Comportement de l'estimateur de Hill en fonction de la taille de l'échantillon.

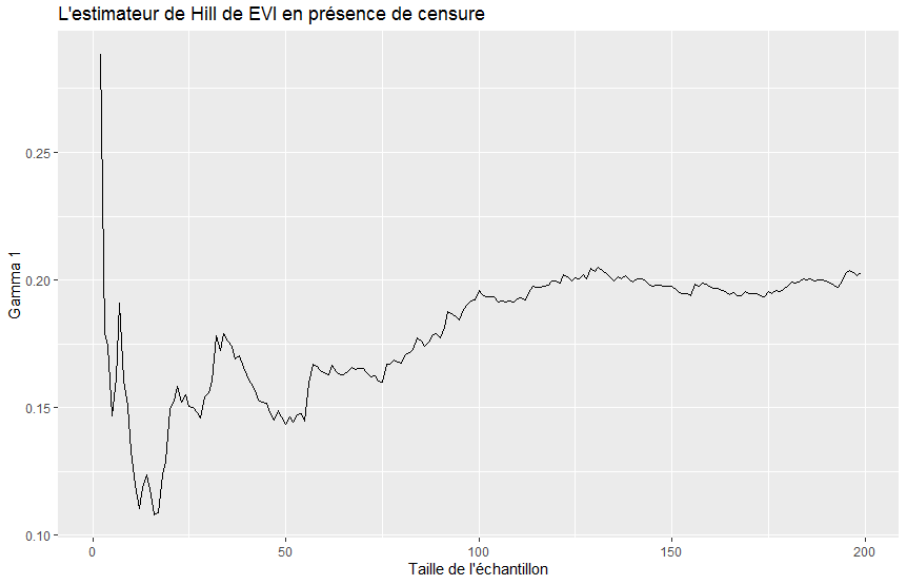

FIGURE 4.3 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(200,5) censurée par une loi de PARÉTO(200,1), avec un taux de censure de 15%.

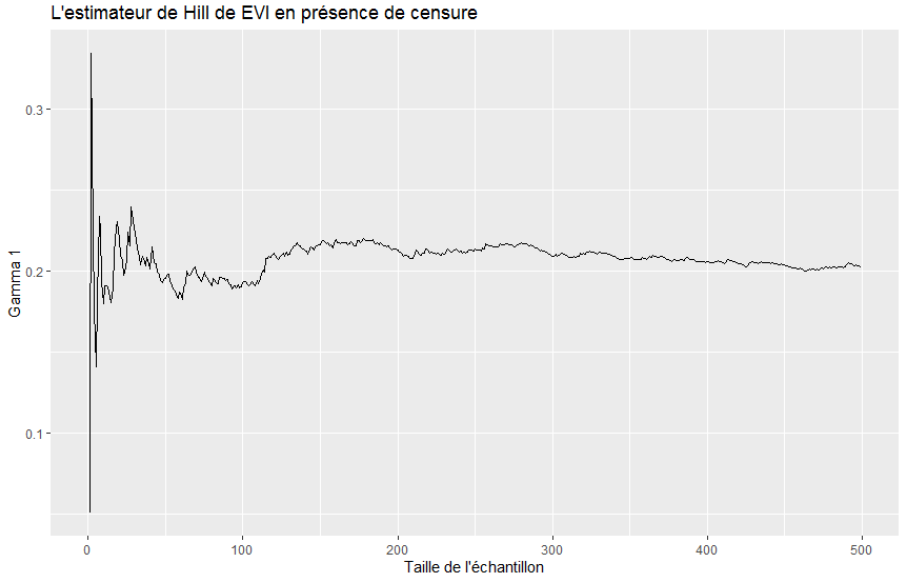

FIGURE 4.4 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(500,5) censurée par une loi de PARÉTO(500,1), avec un taux de censure de 15%.

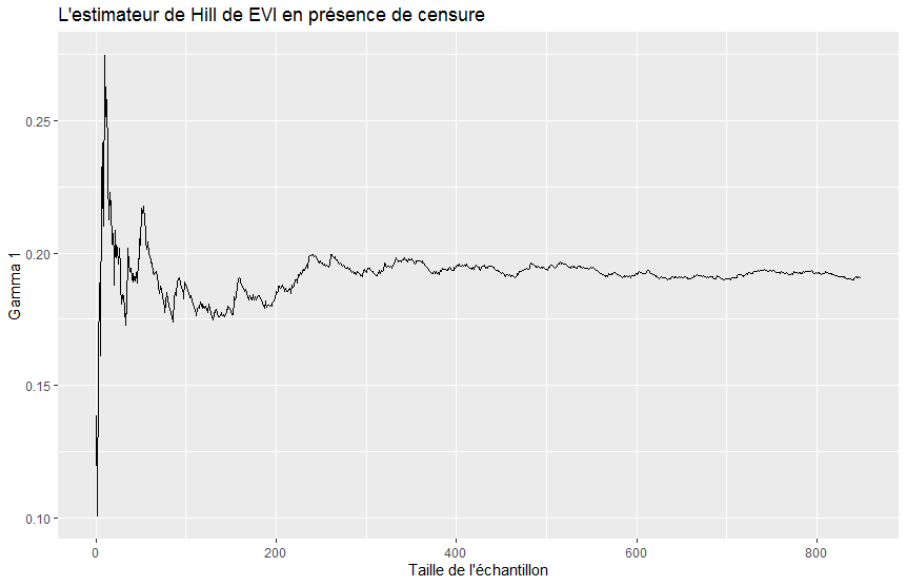

FIGURE 4.5 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(850,5) censurée par une loi de PARÉTO(850,1), avec un taux de censure de 15%.

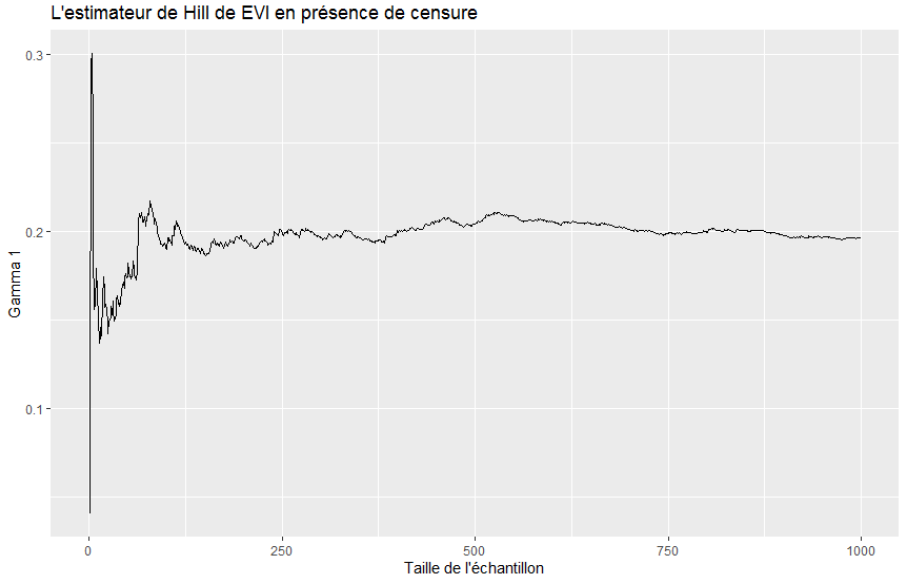

FIGURE 4.6 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(1000,5) censurée par une loi de PARÉTO(1000,1), avec un taux de censure de 15%.

❷ Comportement de l'estimateur de Hill en fonction de taux censure

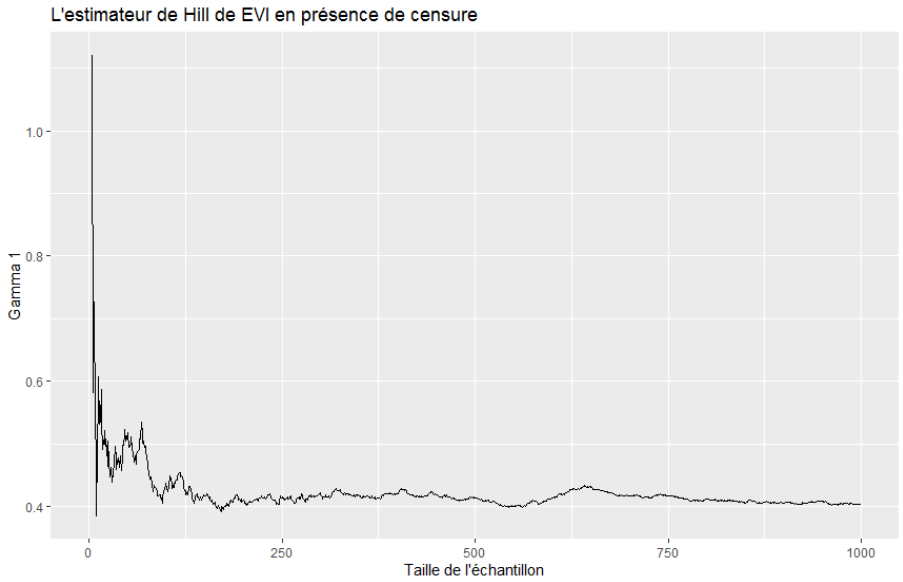

FIGURE 4.7 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(1000,5) avec un taux de censure de $T_c = 50\%$.

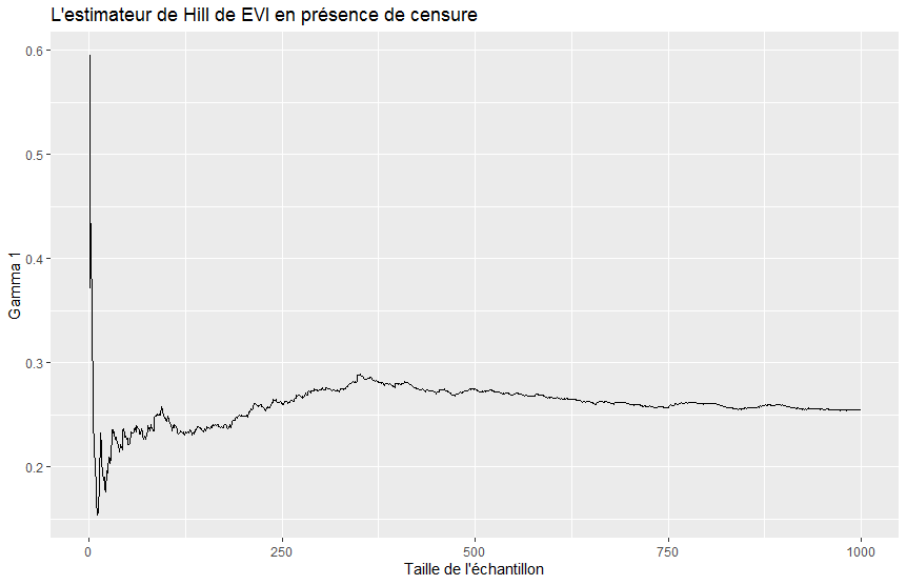

FIGURE 4.8 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(1000,5) avec un taux de censure de $T_c = 25\%$.

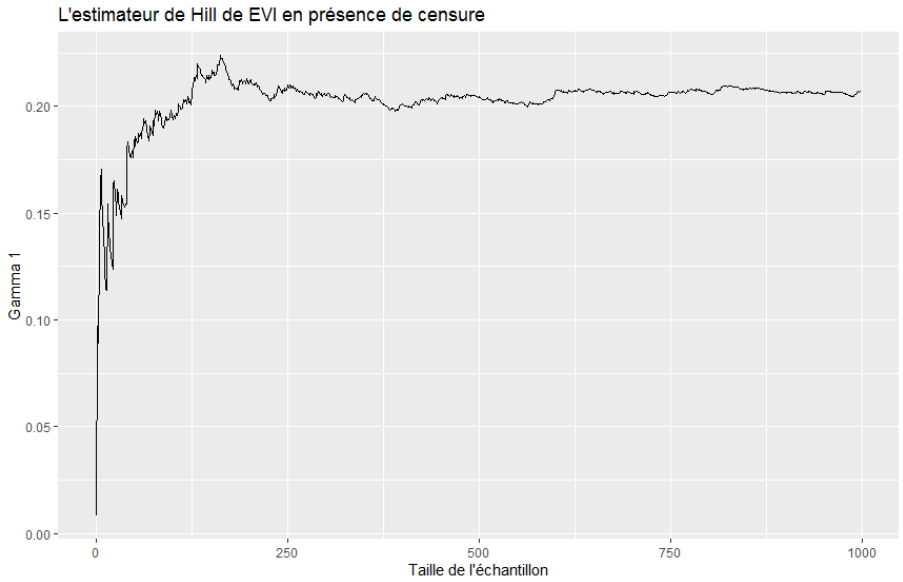

FIGURE 4.9 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(1000,5) avec un taux de censure de $T_c = 10\%$.

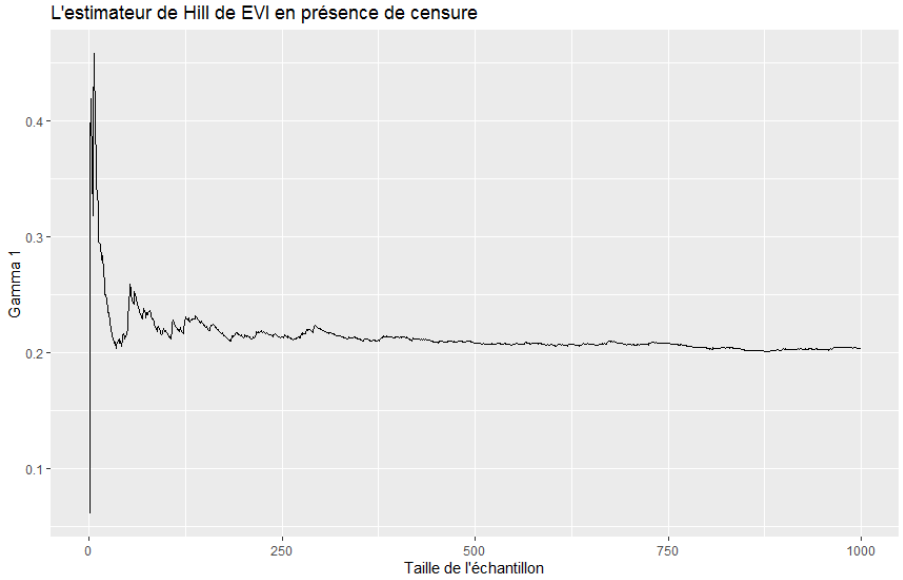

FIGURE 4.10 – L'estimateur de HILL de l'indice de queue d'une loi de PARÉTO(1000,5) avec un taux de censure de $T_c = 5\%$.

On peut en tirer de la figure (4.7), (4.8), (4.9), et (4.10) que l'estimateur de HILL s'approche mieux en déminant le taux de censure T_c vers la vraie valeur de l'indice de queue qui est 0.2 l'inverse du paramètre de forme 5 de la loi de PARÉTO utilisée.

4.3 Étude de données réelles : Transplantation d'un rein

La base de données kidtran du package KMsurv (du logiciel R) regroupe les durées de survie (en jours) de 863 patients ayant subit une transplantation d'un rein. Pour ce base de données, plusieurs covariables sont disponibles tel que le sexe, l'âge ou le groupe ethnique.

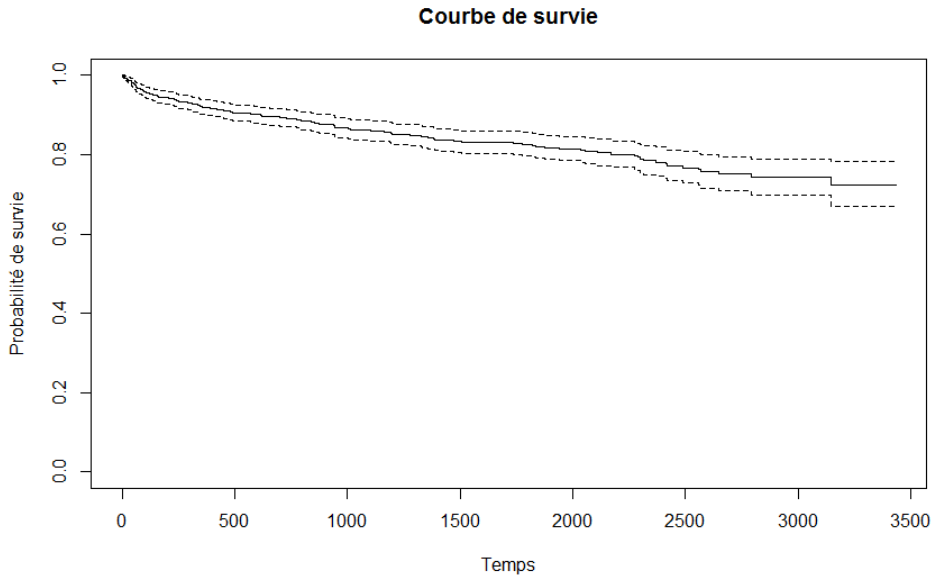

FIGURE 4.11 – Courbe de survie par l'estimateur de KAPLAN-MEIER des durées de vies de 863 patients ayant subit une transplantation d'un rein.

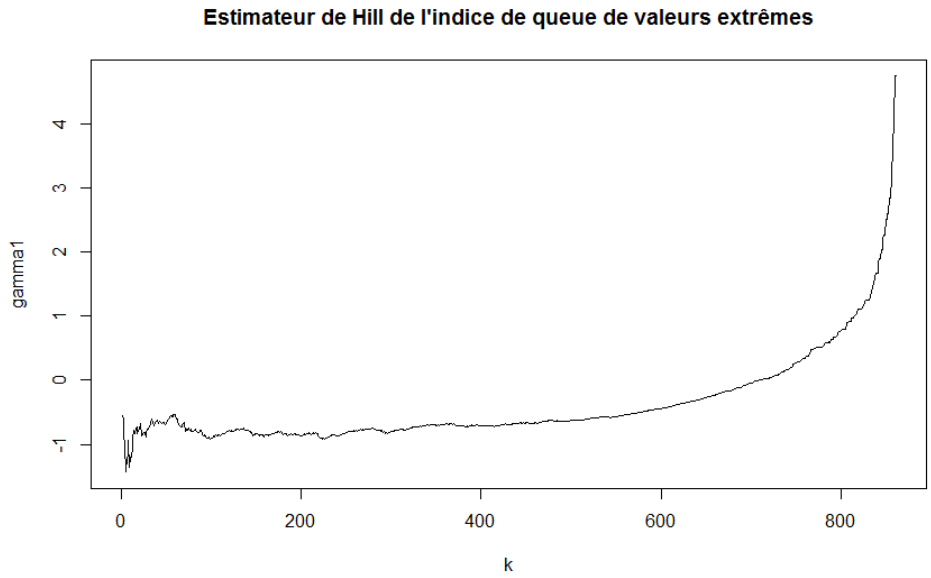

FIGURE 4.12 – Estimateur de HILL de l'indice de queue de valeurs extrêmes des durées de vies de 863 patients ayant subit une transplantation d'un rein.

4.4 Conclusion

On peut conclure de la figure (4.3),(4.4), (4.5) et (4.6) que l'estimateur de HILL converge, en augmantant la taille de l'échantillon, vers la vraie valeur de l'indice de queue qui est 0.2 l'inverse du paramètre de forme 5 de la loi de PARÉTO utilisée.
Selon la figure (4.12), l'indice de que est quasi-négatif, ce qui nous pousse à conclure que la loi est du domaine d'attraction de WEIBULL. Pour aller plus loin dans cette étude, le lecteur pourra à partir de ce dernier résultat faire un test d'adéquation à l'une des distributions qui appartiennent au domaine d'attraction de Weibull mentionnées dans le tableau (2.1).

Conclusion générale

Le premier objectif de ce mémoire est de recueillir et simplifier ce qui a été fait sur l'étude des valeurs extrêmes et d'en faire de ce mémoire une référence à tous ceux qui veulent se lancer dans cette branche récente et dynamique de la statistique qui s'intéresse au événements rares, la cause de toutes les catastrophes qu'on connaît et de toute crise économique...

Le deuxième objectif est de recueillir des connaissances sur les durées de vie et les distributions de survie et de leurs estimateurs et ses applications. La nécessité de ce recueil est qu'en pratique, les statistiques ne sont jamais complètes donc il y a toujours une censure ou troncature.

Le troisième objectif est de jumeler les deux problèmes qui sont les valeurs extrêmes et données censurées ce qui forme un problème à double complexité, la première est que les données sont très rares ce qui nous donnes des statistiques de tailles réduites, la deuxième est de diminuer les tailles des statistiques étudiées à cause de la censure ce qui donne des échantillons de taille nulle, c'est à dire aucune donnée observée. Ce problème reste ouvert en pratique. Une perspective très intéressante de ce travail est d'appliquer ces connaissances sur les mesures de risques.

Bibliographie

[1] **Aalen, O.1978**. Nonparametric estimation of partial transition probabilities in multiple decrement models.*Ann. Statist.,* **534-545 .**

[2] **Arnold.B.C, N. Balakrishnan and H.N. Nagaraja.** *A First Course in Order Statistics.* Classics In Applied Mathemtics, Society for Industrial and Applied Mathematics. Philadelphia. **2008.**

[3] **Beirlant, J., Bardoutsos, A., de Wet, T., and Gijbels, I. 2016** Bias reduced tail estimation for censored Paréto type distributions. *Statist. Probab. Lett.,* **109, 7888.**

[4] **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.**

[5] **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.**

[6] **Bingham, N. H., Goldie, C. M., and Teugels, J. L. 1987** *Cambridge University Press* **.**

[7] **Beirlant, J., and Guillou, A. 2001** . Paréto index estimation under moderate right censoring. *Scand. Actuar. J.,* **111-125.**

[8] **Csorgo, S., Deheuvels, P., and Mason, D. 1985** . Kernel estimates of the tail index of a distribution. *Ann. Statist.,* **1050-1077.**

[9] **Cox, D. R., and Oakes, D. 1984** . Analysis of survival data *CRC Press* **.**

[10] **Charpentier.A et Denuit.M 2005** . Mathématiques de l'assurance non-vie *tome 2, Tarification et provisionnzment* **.**

[11] **Coles, S. 2001** . An Introduction to Statistical Modeling of Extreme Values. *Springer, London* **.**

[12] **Drees, H.1995**. Rened Pickands estimators of the extreme value index. *Ann. Statist,* **2059-2080.**

[13] **David, H. A., and Nagaraja, H. N.2003**. Order Statistics, Third Edition. *John Wiley.*

[14] **Dekkers, A. L., and De Haan, L.1989**. On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.,* **1795-1832.**

[15] **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.**

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

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

[18] **Embrechts, P., Klppelberg, C. and Mikosch, T.1997**. Modelling Extremal Events for Insurance and Finance, *Springer-Verlag, Berlin.*

[19] **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.**

[20] **Fleming, T. R., and Harrington, D. P. 1984**. Nonparametric estimation of the survival distribution in censored data. . *Comm. Statist. Theory Methods.,* **2469-2486.**

[21] **Gnedenko, B.1943**. Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.,* **423-453.**

[22] **Gill, R. D.1994**. Glivenko-Cantelli for Kaplan-Meier. *Math. Methods Statist.,* **3(1), 76.**

[23] **Haan, L. and Ferreira, A. 2006**. . Extreme Value Theory : An Introduction. *Springer-Verlag, New York.*

[24] **Huang, X., and Strawderman, R. L. 2006**. A note on the Breslow survival estimator. *J. Nonparametr. Stat.,* **18(1), 45-56 .**

[25] **Heusler, E., and Teugels, J. L. 1985**.On asymptotic normality of Hills estimator for the exponent of regular variation. *Ann. Statist.,* **743-756 .**

[26] **Huang, X., and Strawderman, R. L. 2006**. A note on the Breslow survival estimator. *J. Nonparametr. Stat.,* **18(1), 45-56.**

[27] **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.**

[28] **Kaplan, E. L., and Meier, P.1958**. Non parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.,* **53(282), 457-481.**

[29] **Kalbfleisch, J. D., and Prentice, R. L.2011**. The statistical analysis of failure time data. *Wiley, New York* **.**

[30] **Klein, J. P., and Moeschberger, M. L. 2005**. Survival analysis : techniques for censored and truncated data. *Springer* **.**

[31] **Lee, E. T., and Wang, J. 2003**. Statistical methods for survival data analysis. *John Wiley* **.**

[32] **Mason, D. M. 1982**. Laws of large numbers for sums of extreme values. *Ann. Probab.,* **36, 394-419.**

[33] **Matthys, G., andBeirlant, J. 2003**. Estimating the extreme value index and high quantiles with exponential regression models. *Statist. Sinica .,* **853-880.**

[34] **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.**

[35] **Nelson, W. 1972**. A short life test for comparing a sample with previous accelerated test results. *Technometrics,* **14(1), 175-185.**

[36] **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.**

[37] **Novak, S. Y. 2011**. Extreme value methods with applications to nance. *CRC Press* **.**

[38] **Pickands III, J. 1975**. Statistical inference using extreme order statistics. *Ann. Statist.,* **119-131.**

[39] **Reiss, R.D.1989**. Approximate distributions of order statistics. *Springer, New York* **.**

[40] **Resnick, S.I.1987**. Extreme values, regular variation, and point processes. *Springer, New York* **.**

[41] **Reiss, R.D., and Thomas, M. 2007**. Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. *Birkhuser, Basel* **.**

[42] **Rényi, A. 1966**. Calcul des Probabilités. *Dunod, Paris* **.**

[43] **Stute, W. 1995**. The central limit theorem under random censorship. *Ann. Statist.,* **422-439 .**

[44] **Segers, J. 2001**. Residual estimators. *J. Statist. Plann. Inference.,* **98(1), 15-27.**

[45] **Stupfler, G. 2016**. Estimating the conditional extreme-value index under random right-censoring. *J. Multivariate Anal.,* **144, 1-24.**

[46] **Stute, W., and Wang, J. L. 1993**. A strong law under random censorship. *Ann. Statist.* **1591-1607.**

[47] **Shorack, G.R., and Wellner, J.A. 1986**. Empirical Processes with Applications to Statistics. *John Wiley & Sons* **.**

[48] **Turnbull, B. W. 1974**. Nonparametric estimation of a survivorship function with doubly censored data *J. Amer. Statist. Assoc.,* **69(345), 169-173.**

[49] **Von Mises, R. 1936**. La Distribution de la plus grande des n valeurs. Selected papers, . *Amer. Math. Soc.,* **271-294.**

[50] **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.**

[51] **Wienke, A. 2010**. Frailty models in survival analysis. *CRC Press* **.**

[52] **Wienke and Hanagal 1985**. Estimating a distribution function with truncated data. *Ann.Statist.,* **163-177.**

[53] **Wang, J. G. 1987**. A note on the uniform consistency of the Kaplan-Meier estimator. *Ann. Statist.,* **1313-1316 .**