République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abderahmane Mira de Béjaia
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

MÉMOIRE de fin d'étude
En vue
de l'obtention du diplôme Master en Recherche Opérationnelle
Option : Modélisation Mathématique et techniques de décision

Thème

Estimation de l'indice des valeurs extrêmes dans le cas des données censurées

Présenté par :

M$_{\text{r}}$ Halkoum Said

Devant le jury composé de :

Présidente :	Mme S. AMROUN	M.A.A	Université de Bejaia
Promotrice :	Mme Y. ZIANE	Docteur	Université de Bejaia
Examinatrice :	Mme L. DJERROUD	Docteur	Université de Bejaia
Examinatrice :	Mme L.HARFOUCHE	Docteur	Université de Bejaia

Année Universitaire 2018-2019

⁕ *Remerciements* ⁕

Au terme de ce travail qui marque la fin du cycle de master de notre formation au sein de l'université Abderahmane Mira de Béjaia, il nous est opportun d'exprimer notre gratitude à tous ceux qui, de loin ou de près, ont matériellement ou moralement contribué â la réalisation de notre modeste travail. Qu'ils trouvent ici l'expression de notre considération.

Au ***ALLAH****, qui nous a donné la vie, qui nous a denné l'intelligence et le courage de réaliser ce travail.*

A nos chers parents par leur affection et amour de nous avoir donné la vie et l'éducation. Voila aujourd'hui nous sommes comptés parmi les hommes intellectuels du monde. Qu'ils se réjouissent du fruit de leur progéniture.

Nous exprimons nos vifs remerciements, notre profonde gratitude et notre reconnaissance à notre encadreur M_{me} ***Y. ZIANE****, qui a dirigé ce travail.*

Nous tenons également à remercier les membres de jury d'avoir accepté de juger notre travail.

Enfin, Nous remercions, de tout coeur, tous ceux qui ont contribué de prés ou de loin à la réalisation de ce travail.

⁕ *Dédicaces* ⁕

***J**e dedie ce modeste travail tout d'abord à nos trés chers parents qui nous ont soutenu tout le long de notre parcour d'études ;*

***À** mes frères et mes sœurs ;*

***À** toute la famille HALKOUM ;*

***À** mes amies ;*

***À** toute la famille RO ;*

***À** toute la promotion 2018-2019 ;*

M. Halkoum Said

Table des matières

Table des figures

Liste des tableaux

Notations et abréviations

$v.a$	variable aléatoire.
fdr	Fonction de répartition.
F_n	Fonction de répartition empirique.
$F^{\leftarrow}$	Inverse généralisée de F.
EVD	Distribution des valeurs extrêmes.
EVI, γ	Indice des valeurs extrêmes.
GEV	Distribution des valeurs extrêmes généralisée
GPD	Distribution de pareto généralisée
POT	Pics au-delà d'un seuil
$\mathcal{H}_\gamma$	Famille de la loi de valeurs extrêmes généralisée .
$\mathbb{P}(A)$	La probabilité de l'événement A.
$\mathbb{E}(X)$	L'espérance mathématique de la $v.a.$ X.
$Var(X)$	La variance de la variable aléatoire X.
$Cov(X, Y)$	La covariance entre X et Y .
i, i, d	Indépendantes et identiquement distribuées.
$\mathbb{1}_{\{A\}}$	Fonction indicatrice de l'ensemble A
$L(x)$	Fonction à variation lente
DA	Domaine d'attraction de maxcimum.
$M_n = X_{n,n}$	Maxcimum de $X_1, ..., X_n$
N_u	Nombres des excés qui dépassent du seuil u
$p.s$	Présque sûre
$\Lambda^{(G)}$	Loi de Gumbel
$\Phi^{(F)}$	Loi de frêchet
$\Psi^{(W)}$	Loi de weibull

$resp$	Respectivement
$S = \overline{F}$	$1 - F$ fonction de survie
TEV	Théorêm des valeurs extrêmes
$X_{1:n}, ..., X_{n,n}$	Statistique d'ordre associéé à $X_1, ..., X_n$
$X \wedge Y$	$\min(X, Y)$
$X \vee Y$	$\max(X, Y)$
x_F	Point terminal
$\overset{\text{loi}}{=}$	Egalité en loi
$:=$	Egalité en défénition
$\underset{\rightarrow}{l}$	Converge en loi
$\underset{\rightarrow}{p}$	Converge en probabilit
$\underset{\rightarrow}{p.s}$	Converge prèsque sûre
$0p(.)$	Converge vers 0 en probabilité
VR_α	Variation régulière d'indice α
$\sup\{A\}$	Supremum de l'ensemble A
$T^* = (T_1^*, ..., T_n^*)$	Echantillon Bootstrap
$\hat{\gamma}_X^c$	Estimateur de Hill avec les données censurées
$\hat{\gamma}_{X,Hill}^{KM}$	Estimateur de Hill par l'intégration de Kaplan-Meier
SD	Erreur standard
IC	Intervalle de confiance
MSE	L'erreur quadratique moyenne

Introduction générale

Au cours des dernières années, nous avons pu observer dans la recherche scientifique, une modélisation des événements rares. Ces événements rares sont des événements dont la probabilité d'apparition est trop faible c'est-à-dire se trouve dans les queues des distributions. Ils apparaissent en général dans de nombreux contextes physiques en particulier les catastrophes naturelles : en hydrologie (crues décennales ou centennales et hauteur des barrages et digues susceptibles de les contenir [2], tempêtes occasionnant d'importants dommages matériels et environnementaux [7], dans les grands incendies [3], dans les tremblements de terre [46], [1], dans les risques financiers (les kraks boursiers, les crises financières, [10]), dans les records sportifs ([24], [25]), mais aussi dans l'étude de la résistance d'un matériau fibreux [15], etc.
La théorie des valeurs extrêmes est une branche des statistique qui essaie d'amener une solution face à ces phénomènes. Elle repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Cependant, on y retrouve deux modèles : loi généralisée des valeures extrêmes (**GEV :** « **Generalized Extreme Value** ») et loi de Paréto généralisée (**GPD :** « **Generalized Pareto Distribution** »). Ainsi, tout a commencé avec les auteurs Fisher et Tippet (1928, [29]) quand ils étudiaient la résistance des fils de coton puis plus tard Gnedenko (1943, [35]) s'est intéressé à ces distributions. Ils ont énoncé un théorème fondamental avec la création de trois domaines d'attraction : domaine d'attraction de Fréchet, Gumbel et Weibull . Ce théorème intéressant fait référence à un paramètre appelé l'indice de queue qui donne la forme de la queue de distribution. 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 (1954, [56]) puis Jenkinson (1955, [43]) ont rassemblé les distributions de ces trois domaines en une seule écriture. C'est en ce moment que plusieurs auteurs se sont focalisés aux estimations de l'indice des valeurs extrêmes. Nous pouvons citer Hill (1975, [41]) dans le cas où l'indice est positif. Puis

Pickands ([42]) dans la même année a proposé un estimateur de l'indice des valeurs extrêmes dans le cas général. Par contre Dekkers et al. ont généralisé l'estimateur de Hill, dénommé l'estimateur des Moments (1989, [17][20]). Beirlant et al. (1996, [6]) ont présenté à leur tour, l'estimateur de l'indice des valeurs extrêmes généré à partir de l'estimateur de Hill

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, 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. En 1958, Kaplan et Meier ([45]) proposent un estimateur jusqu'a l'inconnue de la fonction de répartition dans le cas où les données sont censurées. Ils en profitent pour déterminer ses propriétés asymptotiques qui appelé « l'estimateur de Kaplan-Meir ». La modélisation des valeurs extrêmes censurées voit le jour en 1997 dans la littérature des extrêmes avec la sortie du livre Reiss et Thomas [48]. Il a fallu qu'en 2007 Beirlant et al. [5] abordent réellement la statistique non paramétrique des valeurs extrêmes avec des données censurées. Leur estimateur est basé sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion $\widehat{p}$ représentant des données non censurées dépassant un certain seuil donné. Puis, Einmahl et al en (2008, [25]),ont utilisé le même concept pour proposer un estimateur de l'indice de queue sur les k-plus grandes valeurs, ensuite déterminer ses propriétés asymptotiques . Puis Warms et Warms en (2013,[25]) a pue créer un nouvel indice basé sur l'intégral des Kaplen -Meier. Le domaine d'application des données censurées est trés vaste,
Dans ce mémoire nous allons présenter quelques applications de ces derniers.Pour cela nous avons divsé ce mémoire en troix chapitres :
Ce mémoire est composé d'une introduction générale, de troix chapitres, d'une conclusion générale, et d'une bibliographie, 'Le premier chapitre nous présentons la théorie des valeurs

extrêmes , et nous regroupons les définitions et des résultats sur cette derniére. Aprés avoir introduit le comportement du maximum, on présente les deux principaux outils servant à modéliser le comportement des valeurs extrêmes d'un échantillon : la loi généralisée des valeurs extrêmes (**GEV**) et loi de Paréto généralisée (**GPD**) . On s'intéressera ensuite à la caractérisation des domaines d'attraction. Enfin , on rappele l'estimation de l'indice des valeurs extrêmes γ, on exposera uniquement trois estimateurs de γ l'estimateur de Pickands [42] , l'estimateur de Hill.B [41] et l'estimateur des moments (Dekkers et al.[17]). On donne également certaines de leurs propriétés statistiques. Ces estimateurs sont basées fortement sur les plus grandes statistiques d'ordre $X_{n-k:n}, ..., X_{n:n}$, ou la statistique $X_{n-k:n}$ est alors dite statistique d'ordre intermédiaire. Dans le deuxieme chapitre, on a commencer par presenter les données incomplètes (Censure et troncature), ensuite on s'est intéresser au problème de l'estimation de l'indice de queue en présence de données censurées aléatoirement à droite, nous passons ainsi a la presentation de quelques estimateurs de l'indice des valeurs extrêmes en présence de censure aléatoire à droite. Nous nous focalisons principalement sur les estimateurs suivants : l'estimateur de Hill de l'indice des valeurs extrêmes Einmahl et al (2008,[25]), et l'estimateur de l'indice des valeurs extrêmes d'intégration de Kaplan-Meier Worms ,J., Worms, R., (2013,[58]) .

Dans le chapitre 3,nous avons présenter une application numeriques , dont on a estimé l'indice des valeurs extrêmes par l'estimateur de Hill avec l'application de la méthode ré-échantillonage ainsi l'integration de Kaplan Meier.

Chapitre 1

Notions de base sur le théorie des valeurs extrêmes

1.1 Introduction

Dans ce chapitre, on s'intéresse à la théorie des valeurs extrêmes (**TVE**) dont le probléme de base est de modéliser et prévoir l'occurrence d'événements extrêmes en s'intéressant principalement non pas au "corps" de la distribution mais à sa queue. Cette théorie est à la fois ancienne et moderne , elle a donné lieu à de fructueux développements mathématiques , l'ouvrage synthèse de Gumbel [37], résumé de ses travaux de 1954 et 1958, est souvent identifié à la théorie statistique des extrêmes. Il faut aussi citer les travaux de Fréchet [30] , Fisher et Tippett [29] et De Finetti [28].

1.2 Événements rares

Les évènements rares sont des évènements ayant une faible probabilité d'apparition. Lorsque le comportement de ces évènements est dû au hasard on peut étudier leur loi. Ils sont dits extrêmes quand il s'agit de valeurs beaucoup plus grandes ou plus petites que celles observées habituellement.

Les événements rares (tremblements de terre, inondations, accidents nucléaires, crises monétaires ou financières, crachs boursiers, émergence d'un nouveau phénomène endémique, etc.) dominent l'actualité quotidienne par leur caractère imprévisible. Compte tenu de l'importance des enjeux sociaux et scientifiques, aucun débat sérieux sur le hasard ne saurait être mené sans une réflexion sur les évènements rares et extrêmes.

1.3 Statistique d'ordre

Définition 1.3.1. (les statistiques d'ordres) Soit une suite finie d'observations iid X_i, $1 \leq i \leq n$, classées par ordre croissant . On écrit cette suite d'observations sous la notation $X_{i:n}$ avec,

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

$X_{i:n}$ est donc la $i - ieme$ statistique d'ordre (ou statistique d'ordre i) dans un échantillon de taille n

Définition 1.3.2. (Statistiques d'ordre extrêmes). Les statistiques d'ordre extrêmes sont définies comme termes du maximum et du minimum des n $va's$ $X_1, ..., X_n$. La variable $X_{1:n}$ est la plus petite statistique d'ordre (ou statistique du minimum) et $X_{n:n}$ est la plus grande statistique d'ordre (ou statistique du maximum)

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

nous introduisons l'écart des statistiques d'ordre :

$$W_{i,j:n} = X_{j:n} - X_{i:n} (i < j).$$

L'écart $W = W_{1,n:n} = X_{n:n} - X_{1:n}$ est appelée la déviation extrême.

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

$$F_{X_{i:n}}(x) = \mathbb{P}(X_{i:n} \leq x) = \sum_{i=k}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k} \tag{1.1}$$

$F_{X_{i:n}}$: est la fonction de répartition de la $i^{\text{ème}}$ statistique d'ordre $1 \leq \ i \leq n$: Autrement dit, le nombre d'éléments de l'échantillon inférieurs à x suit une loi binomiale de paramètres n et $F(x)$, puisqu'il s'agit là de n expériences indépendantes, possédant deux issues : « être inférieur à x » et « être supérieur à x », la première des deux issues ayant pour probabilité $F(x)$, et la deuxième issue ayant pour probabilité $1 - F(x)$.
on déduit que la fonction de densité est de la forme suivante :

$f_{X_{i:n}}(x) = \frac{d}{dx} F_{X_{i:n}}(x)$

$$= \sum_{k=i}^{n} \binom{n}{k} \Big(kf(x)F(x)^{k-1}(1-F(x))^{n-k} + F(x)^k(n-k)(-f(x))(1-F(x))^{n-k-1} \Big)$$

$$= \binom{n}{i} if(x)F(x)^{i-1}(1-F(x))^{n-i} + \sum_{k=i+1}^{n} \binom{n}{k} kf(x)F(x)^{k-1}(1-F(x))^{n-k}$$
$$- \sum_{k=i}^{n-1} \binom{n}{k} F(x)^k(n-k)(f(x))(1-F(x))^{n-k-1}$$

$$= \binom{n}{i} if(x)F(x)^{i-1}(1-F(x))^{n-i} + \sum_{k=i+1}^{n} \binom{n}{k} kf(x)F(x)^{k-1}(1-F(x))^{n-k}$$
$$- \sum_{k=i+1}^{n} \binom{n}{k-1} F(x)^{k-1}(n-k+1)(f(x))(1-F(x))^{n-k}$$

$$= \binom{n}{i} if(x)F(x)^{i-1}(1-F(x))^{n-i}$$

$\text{car}(\binom{n}{i}i = \frac{n!i}{i!(n-i)!} = \frac{n!(n-i+1)}{(i-1)!(n-i+1)!} = (n-i+1)\binom{n}{i-1}$.

Finalement :

$$f_{X_{i:n}}(x) = \frac{n!}{(i-1)(n-i)} f(x)F(x)^{i-1}(1-F(x))^{n-i} \tag{1.2}$$

Proposition 1.3.1. *(Distributions du maximum et du minimum). Pour les statistiques d'ordre extrême, nous obtenons des expressions très simples . En utilisant la propriété d'indépendance des variables aléatoires $X_1, ..., X_n$, nous en déduisons que pour la statistique du minimum :*

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

et :

$$f_{X_{1:n}}(x) = f(x)(1-F(x))^{n-1} \tag{1.3}$$

pour la statistique du maximum :

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

et :

$$f_{X_{n:n}}(x) = nf(x)(F(x))^{n-1} \tag{1.4}$$

Définition 1.3.3. Le point terminal d'une distribution F est défini par :

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\} \tag{1.5}$$

Pour plus de détails sur les statistiques d'ordre voir [26]

1.4 Théorie des valeurs extrêmes

Le principal résultat de la théorie des valeurs extrêmes repose sur le Théorème de Fisher et Tippet (1928) (ou ils ont déduit de manière heuristique les lois limites possibles pour le maximum d'une suite des va's indépendantes identiquement distribuées et de même loi) dont la première preuve rigoureuse est due à Gnedenko (1943,[35]).

Lorsque on étudie un phénomène aléatoire, on s'intéresse principalement à la partie dite centrale de la loi modélisant au mieux le phénomène considéré (calcul de l'espérance, la médiane, la variance, utilisation du théorème central limite(**TCL**), etc.).

Cependant, l'étude des "grande" valeurs (ou de manière équivalente des "petites" valeurs) du phénomène est parfois essentielle lorsqu'il s'agit par exemple de quantifier le risque pour une compagnie d'assurance (par exemple connaître la fréquence des crues d'une rivière, etc...).

La théorie des valeurs extrêmes a pour but d'étudier la loi du maximum d'une suite des variables aléatoires réelles même si, et spécialement si, la loi du phénomène n'est pas connue.

Considérons $(X_1, X_2, ..., X_n)$ une suite de n variables aléatoires $(v.a.)$ indépendantes et identiquement distribuées $(i.i.d)$ de fonction de répartition F définie par :

$$F(x) = \mathbb{P}(X_i \leq x) \; pour \;\; i = 1, \ldots, n \tag{1.6}$$

Pour étudier le comportement extrême des événements, on considère la variable aléatoire $\mathbf{M_n} = X_{n:n} = \max(x_1, x_2, ..., x_n)$ le maximum d'un échantillon de taille n .

Comme les variables aléatoires sont $(i.i.d)$, alors la fonction de répartition de $\mathbf{M}_n$ est donnée par :

$$\mathbf{F}_{\mathbf{M}_n}(x) = \mathbb{P}(\mathbf{M}_n \leq x) = (F(x))^n \tag{1.7}$$

Remarque 1.4.1. On obtient la correspondance entre minimum et maximum par la relation suivante :

$\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n)$. Ainsi, tous les résultats que nous allons présenter pour les maxima pourront être transposés pour les minima.

Remarque 1.4.2. La distribution asymptotique du maximum, déterminée en faisant tendre n vers l'infini, donne une loi **dégénérée**. En effet

$$\lim_{n\to+\infty} \mathbf{F}_{\mathbf{M}_n}(x) = \begin{cases} 0 & si\ F(x) < 1, (x < x_F) \\ 1 & si\ F(x) = 1, (x \geq x_F) \end{cases} \tag{1.8}$$

où $x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$, est le point terminal de la loi F avec la convention $\sup\{\emptyset\} = \infty$. Le point terminal x_F représente la borne supérieure du support de la loi. Le résultat (1.8) nous indique que la distribution du maximum $X_{n:n}$ est une loi dégénérée. Ce résultat fournit très peu d'informations sur le comportement de $X_{n:n}$. On aimerait obtenir une loi non dégénérée pour le maximum.

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

Définition 1.4.1. (lois de même type).

On dit que deux variables aléatoires réelles X et Y sont de même type s'il existe des constantes réelles $a > 0$ et $b \in \mathbb{R}$ tels que $Y = aX + b$
i.e. Si F et H sont des lois respectives des variables Y et X alors on a $F(ax + b) \overset{\text{loi}}{=} H(x)$
Autrement dit, les variables « de même type » ont la même loi, à un facteur de localisation et d'échelle près.

De façon analogue au théorème central limite (TCL), peut-on trouver des constantes de normalisation : a_n et b_n avec $a_n > 0$ et $b \in \mathbb{R}$ et une loi non-dégénérée de loi H telle que :

$$\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq x\Big] = \Big[F(a_n x + b_n)\Big]^n \to H(x) \quad , \quad x \in \mathbb{R}$$

Fisher et Tippett trouvent en 1928 une solution à ce problème au moyen d'un théorème qui porte leur nom et qui est l'un des fondements de la théorie des valeurs extrêmes.

1.4.1 Distribution généralisée des valeurs extrêmes (GEV)

L'approche classique de la théorie des valeurs extrêmes, consiste à étudier le comportement asymptotique du maximum de l'échantillon et ensuite déduire celui de la queue de distribution par extrapolation. Nous exposons maintenant le résultat principale de la théorie des valeurs extrêmes [59].

Théorème 1.4.1. *(Fisher and Tippet (1928,[29]) et Gnedenko (1943,[35])) Soit $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes et identiquements distribuées (i.i.d) de fonction de répartition commune F et*

$$M_n = max(X_1, X_2, ...X_n)$$

si la loi limite de M_n existe alors, on peut trouver deux constantes de normalisation $a_n > 0$ et $b_n \in R$ telle que :

$$\lim_{n\to+\infty} \mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq x\Big] = \lim_{n\to+\infty} F^n(a_n x + b_n) = \mathcal{H}_\gamma(x), x \in \mathbb{R} \tag{1.9}$$

alors $\mathcal{H}$ est l'une des trois lois limites :

$$\mathcal{H}_\gamma(x) = \begin{cases} \Lambda^{(G)}(x) = \exp\{-\exp\{-x\}\}, & si\, x \in \mathbb{R}, \gamma = 0, \textit{(Gumbel) ;} \\ \Phi_\gamma^{(F)}(x) = \exp\{(-x)^{-\frac{1}{\gamma}}\}, & si\ \ x > 0, \gamma > 0, \textit{ (Frechet) ;} \\ \Psi_\gamma^{(W)}(x) = \exp\{-(-x)^{-\frac{1}{\gamma}}\}, & si\, x \leq 0, \gamma < 0, \textit{ (Weibull).} \end{cases} \tag{1.10}$$

La distribution $H_\gamma(x)$ peut être écrite aussi de la façon suivante :

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\{-(1+\gamma x)_+^{-\frac{1}{\gamma}}\} & si\gamma \in \mathbb{R}^*, w_+ = \max(0.w) \\ \exp\{-\exp(-x)\} & si\gamma = 0, x \in \mathbb{R} \end{cases} \tag{1.11}$$

Où la distribution $\mathcal{H}_\gamma$ est dite distribution généralisée des valeurs extrêmes **(GEV)**, c'est une famille de distribution dont les caractéristiques sont assez différentes. Le paramétre réel γ est applé indice des valeurs extrémes ou indice de queue (Tail index) Cette écriture de la distribution $\mathcal{H}_\gamma$ est due à Von Mises (1936,[56]) et Jenkinson (1955,[43]). Plus l'indice γ est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On parle alors de distributions à "queues épaisses".

Remarque 1.4.3. Ce théorème est aussi valable pour les variables aléatoires faiblement dépendantes, ce résultat est du à Leadbetter (1983)

Remarque 1.4.4.

1. Le cas $\gamma = 0 \rightarrow \mathcal{H}_0(x) = \Lambda^{(G)}(x)$ correspond à la loi Gumbel
2. Le cas $\gamma > 0 \rightarrow \Phi_\alpha^{(F)}(x) = \mathcal{H}_{\frac{1}{\alpha}}(\alpha(X-1))$ correspond à la loi de Fréchet , Si X suit une loi de Fréchet de paramètre $\alpha > 0$ alors $\alpha(X-1)$ suit la loi des **GEV** de paramétre $\gamma = \frac{1}{\alpha}$
3. Le cas $\gamma < 0 \rightarrow \Psi_\alpha^{(W)}(x) = \mathcal{H}_\alpha(-\alpha(X+1))$ correspond à la loi de Weibull , Si X suit une loi weidull de paramètre $\alpha < 0$ alors $-\alpha(X+1)$ suit la loi des **GEV** de paramétre $\gamma = \frac{1}{\alpha}$

1.4.2 Distribution des excès (P.O.T)

L'approche par dépassements de seuil, en anglais "Peaks-Over-Threshold approach" notée **POT**. Cette méthode initialement développée par Pickands (1975,[8]) et abondamment étudiée par divers auteurs tels que de Smith (1987), Davison et Smith (1990), ou Reiss et Thomas (2001,) pour de plus références, repose sur l'utilisation des statistiques d'ordre supérieur de l'échantillon [7]. Elle consiste à ne conserver que les observations dépassant un certain seuil. L'excès au-delà du seuil est défini comme l'écart entre l'observation et le seuil. Plus précisément, soit un échantillon de variables aléatoires indépendantes, identiquement distribuées *iid* $X_1, X_2, ..., X_n$ de fonction de répartition F et x_F un point terminal. Alors,pour un seuil $u < x_F$ fixé, considérons les n observations $x_{i_1}, ..., x_{i_n}$ dépassant le seuil u. On définit la variable $Y_j = X_{ij} - u$, $j = 1, ..., n$, l'excès au-dessus du seuil u. voir Figure 1.1 ci-dessous :

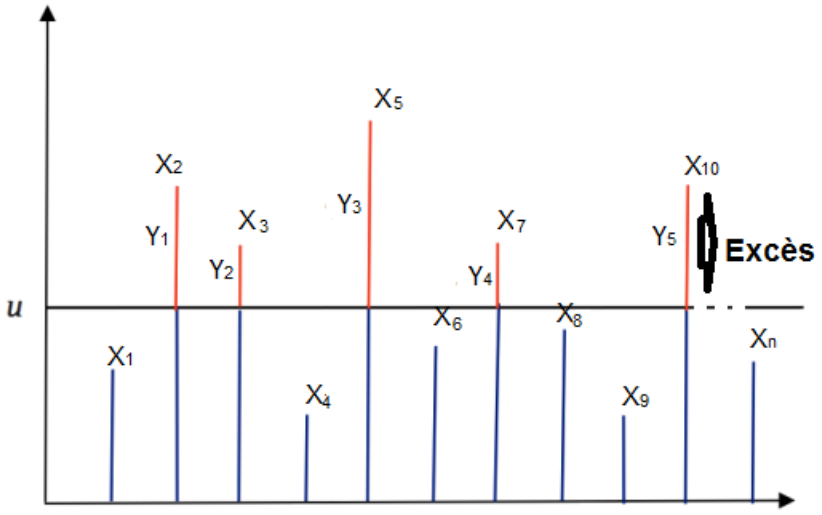

FIGURE 1.1 – Dépassement de seuil (POT)

On note F_u la fonction de répartition de l'excès Y au delà du seuil u. La loi des excès est celle de variables aléatoires *iid.* admettant pour fonction de répartition :

$$F_u(y) = \mathbb{P}(Y \leq x/X > u)$$

représentant la probabilité que la variable aléatoire X ne dépasse pas le seuil u d'au moins une quantité x sachant qu'elle dépasse u. F_u décrit ainsi la loi de Y sachant que $X > u$. On peut la réécrire en fonction de F à l'aide du résultat suivant [49] .

On a pour $x \geq 0$:

$$F_u(x) = \mathbb{P}(Y \leq x/X > u) = \mathbb{P}(X - u \leq x/X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}$$

ou de manière équivalente pour la fonction de survie :

$$\overline{F}_u(x) := \mathbb{P}(Y > x/X > u) = 1 - F_u(x) = \frac{\overline{F}(x+u)}{\overline{F}(u)} \tag{1.12}$$

Le théorème de Pickands-Balkema-de Haan ci-après donne la forme de la loi limite pour les valeurs extrêmes : sous certaines conditions de convergence, la loi limite est une loi de Pareto généralisée que l'on notera `GPD` (Generalized Pareto Distribtution)

Théorème 1.4.2. *(Théorème de Pickands [42])*

Soit $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires iid. suivant la loi F. F appartient au domaine d'attraction DA si et seulement si il existe une fonction $\sigma(u)$ positive telle que

$$\lim_{x \to x*} \sup_{0<x<x*} \{\mathbb{P}(X - u \leq x/X > u) - \mathcal{H}_{\gamma,\sigma(u)}(x)\} = 0$$

, où $x* = x_F = \{\sup x \ , \ F_u < 1\}$ et $\mathcal{H}_{\gamma,\sigma}$ est la distribution de Pareto généralisée (GPD) définie par

$$\mathcal{H}_{\gamma,\sigma}(x) = \begin{cases} 1 - (1 + \frac{\gamma x}{\sigma})^{-\frac{1}{\gamma}}, & si\ \gamma \neq 0\ ,x \geq 0\ et\ x < -\frac{\sigma}{\gamma}\ \ si\gamma < 0, \\ 1 - \exp(-\frac{x}{\sigma}), & si\gamma = 0,x \geq 0. \end{cases}$$

*Ce théorème signifie que si F appartient au domaine d'attraction DA alors il existe une fonction $\sigma(u)$ positive et un réel γ tels que la loi des excès F_u peut être uniformément approchée par une distribution de Pareto généralisée **(GPD)** notée $\mathcal{H}_{\gamma,\sigma}$*

1.5 Domaine d'attraction

Définition 1.5.1. On appelle domaine d'attraction de $F_i(i = 1, 2, 3)$ l'ensemble $D(F_i)$ des lois de probabilité F pour lesquelles $(X_1, X_2, ..., X_n)$ étant un échantillon extrait de la loi F , $M_n = \max_{1 \leq i \leq n}(X_i)$ va converger en loi vers la loi des extrêmes de type F_i.

Nous rappelons les conditions nécessaires et suffisantes sur la fonction de répartition F pour qu'elle appartienne à l'un des domaines d'attraction de l'une des trois lois limites des valeurs extrêmes.

1.5.1 Fonctions à variation régulière

Dans cette partie, nous allons définir quelques notions essentielles de la théorie des valeurs extrêmes permettant de caractériser les domaines d'attraction. Cependant, les fonctions à variations régulières permettent d'avoir une écriture unique pour chaque domaine d'attraction.

Condition du première ordre, de Haan et Ferreira (2006,[38])

Définition 1.5.2. Une fonction mesurable positive U, est dite à variations régulières d'indice $\alpha \in R$ à l'infini si pour tout $t > 0$

$$\lim_{x \to +\infty} \frac{U(tx)}{U(x)} = t^{\alpha} \tag{1.13}$$

On notera dans la suite $\mathcal{RV}_\alpha$ l'ensemble des fonctions à variations régulières d'indice α. Une fonction à variations régulières d'indice $\alpha \in \mathbb{R}$ se comporte asymptotiquement comme la fonction $x \longrightarrow x^{\alpha}$.

Définition 1.5.3. Si une fonction $L : \mathbb{R} \longrightarrow [0, 1[$ est à variations régulières d'indice 0 ($\mathcal{RV}_0$), on dit que L est à variations lentes. Une fonction à variations lentes est "presque constante" asymptotiquement.

Un exemple d'une fonction à variations lentes est la fonction logarithme. En effet, soit $t > 0$,

$$\lim_{x \to +\infty} \frac{\log(tx)}{\log(x)} = 1 + \lim_{x \to +\infty} \frac{\log(t)}{\log(x)} = 1$$

Il est facile de remarquer que toute fonction à variations régulières U d'indice α peut s'écrire sous la forme

$$U(x) = x^{\alpha} L(x), ou\ L \in \mathcal{RV}_0$$

Représentation de Karamata

Définition 1.5.4. (Resnick [49]) Soit H une fonction à variation régulière d'indice α.

En utilisant le fait que $U(x) = x^{\alpha}L(x)$, on déduit facilement que pour tout $x > 0$

$$U(x) = c(x)\exp\left\{\int_{1}^{x} t^{-1}\Delta(t)d(t)\right\} \tag{1.14}$$

où c et Δ sont des fonctions positives telles que $\lim_{x\longrightarrow\infty} c(x) = c \in]0, +\infty[$ *et* $\lim_{x\longrightarrow\infty} \Delta(t) = \alpha$ si $\Delta(t) \longrightarrow \pm\infty$ alors U est une fonction à variations rapides d'indice $\pm\infty$.

On notera (abusivement) $\mathcal{RV}_{\pm\infty}$ l'ensemble des fonctions à variations rapides d'indice $\pm\infty$ admettant la représentation (1.14).

Si la fonction c est constante, on dit que U est une fonction à variations régulières (ou rapides) normalisée

Définition 1.5.5. Une fonction mesurable U est dite à variations rapides d'indice $-\infty(resp. + \infty)$ si elle est positive et si

$$\lim_{x\to+\infty}\frac{U(tx)}{U(x)} = \begin{cases} +\infty(resp.0) & \text{si } 0 < t < 1 \\ 0(resp. + \infty) & \text{si } t > 1 \end{cases} \tag{1.15}$$

Comme exemple de fonction à variations rapides d'indice $+\infty$, on a la fonction exponentielle. La fonction $x \longrightarrow \exp(-x)$ est une fonction à variations rapides d'indice $-\infty$.

Proposition 1.5.1. *(Inverse d'une fonction à variations régulières)*

- *Si $U \in \mathcal{RV}_{\alpha}$ avec $\alpha = 0$ est une fonction croissante telle que $U(x) \longrightarrow \infty$ lorsque $x \longrightarrow \infty$ alors l'inverse généralisée de U est à variations régulières d'indice $\frac{1}{\alpha}$ ($U^{\leftarrow} \in \mathcal{RV}_{\frac{1}{\alpha}}$). Dans le cas $\alpha = 0$, cette condition est essentielle. Par exemple la fonction $L(x) = 1 - x^{-1} \in \mathcal{RV}_0$ admet pour inverse la fonction $L^{-1}(x) = (1-x)^{-1}$ qui n'est pas à variations rapides. Par contre la fonction $log(x) \in \mathcal{RV}_0$ admet bien pour inverse la fonction $\exp(x)$ qui est une fonction à variations rapides.*
- *Si U est à variations régulières d'indice $\alpha > 0$, alors $U^{\rightarrow}(x)$ est à variations régulières d'indice $1/\alpha$.*
- *Si U est à variations régulières d'indice $\alpha < 0$, alors $U^{\rightarrow}(1/x)$ est à variations régulières d'indice $-1/\alpha$.*

Condition du seconde ordre de Haan et Ferreira (2006,[38])

Une fonction de répartition $F(.) \in$ **DA(fréchet)**, $\gamma > 0$, admet une condition du seconde ordre à l'infini si elle satisfait à l'une des assertions suivantes :

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

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

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

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

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

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

si $\rho = 0$, on remplace $(x^\rho - 1)/\rho$ par $\log(x)$

Les fonctions $A(.)$,$A_1(.)$, $A_2(.)$ sont à variations régulières à l'infini d'indices respectifs ρ,ρ/γ,et $-\rho$, avec

$$A_1(t) = A(1/(1 - F(t))) \ et \ A_2(s) = A(1/s)$$

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

1.5.2 Caractérisation des domaines d'attraction

Domaine d'attraction de Fréchet

Théorème 1.5.1. *F appartient au domaine d'attraction de Fréchet $\Phi_\gamma^{(F)}$ avec un indice de valeur extrême $\gamma > 0$ si et seulement si $x_F = +\infty$ et $\overline{F} = 1 - F$ est une fonction à variation régulière d'indice $(-\frac{1}{\gamma})$, $F \in \mathcal{RV}_{-\frac{1}{\gamma}}$. Dans ce cas, un choix possible pour les suites $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ est :*

$$a_n = \overline{F}^{\leftarrow}(\frac{1}{n}) = F^{\leftarrow}(1 - \frac{1}{n}) \ et \ b_n = 0 \ , \ \forall n > 0$$

Autrement dit, une fonction de répartition F appartenant au domaine d'attraction de Fréchet $\Phi_\gamma^{(F)}$ s'écrit sous la forme :

$$F(x) = 1 - x^{-\frac{1}{\gamma}} L(x) \ , L \in \mathcal{RV}_0 \tag{1.16}$$

$F^{\leftarrow}(x) := \sup\{y \in \mathbb{R}, F(y) \leq x\}$ où $F^{\leftarrow}$ est l'inverse généralisée de F. Pour plus de détails le lecteur pourra consulter les ouvrages d'Embrechts et al. (1997, [26]), de Bingham et al. (1987, [8]), de Coles (2001, [13]).

Si F est une fonction de répartition continue et strictement croissante, alors la fonction inverse généralisé $F^{\leftarrow}$ est équivalente à l'application réciproque F^{-1} .

Domaine d'attraction de Weibull

Le résultat suivant montre que l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.

Théorème 1.5.2. *(Embrechts et al. [26])*

Une fonction de répartition F appartient au domaine d'attraction de $\Phi_\gamma^{(W)}$ (Weibull, avec un indice des valeurs extrêmes $\gamma < 0$) si et seulement si son point terminal x_F est fini et si la fonction de répartition F^* définie par

$$F^* = \begin{cases} 0 & si\ x < 0 \\ F(x_F - \frac{1}{x}) & si\ x \geq 0 \end{cases} \tag{1.17}$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $-\gamma > 0$. Des suites possibles de normalisation (a_n) et (b_n) sont données par $a_n = x_F - \overline{F}^{\leftarrow}(\frac{1}{n})$ et $b_n = x_F$.

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

$$F(x) = 1 - (x_F - x)^{-\frac{1}{\gamma}} L((x_F - x)^{-1}), L \in \mathcal{RV}_0 \tag{1.18}$$

Domaine d'attraction de Gumbel

La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe.

Théorème 1.5.3. *(Fonction de Von Mises)[55] Une fonction de répartition F appartient au domaine d'attraction de $\Lambda^{(G)}$ (Gumbel) si et seulement si il existe $z < x_F \leq \infty$ que*

$$\overline{F}(x) = c(x) \exp\Big\{ - \int_z^x \frac{1}{a(t)} dt \Big\} \; si \; z < x \leq x_F. \tag{1.19}$$

où $c(x) \to c > 0$ lorsque $x \to x_F$ et a est une fonction positive et dérivable de dérivée $a^{'}$ telle que $a^{'}(x) \to 0$ lorsque $x \to x_F$ .Un choix possible pour la fonction a est :

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

Des suites possibles de normalisation (a_n) et (b_n) sont données par $b_n = \overline{F}^{\leftarrow}(\frac{1}{n})$ et $a_n = a(b_n)$

Définition 1.5.6. Le paramètre γ du Théorème 1.4.1 ou de la Théorème 1.4.2 est un paramètre de forme que l'on appelle « indice des valeurs extrêmes » ou « indice de queue ».

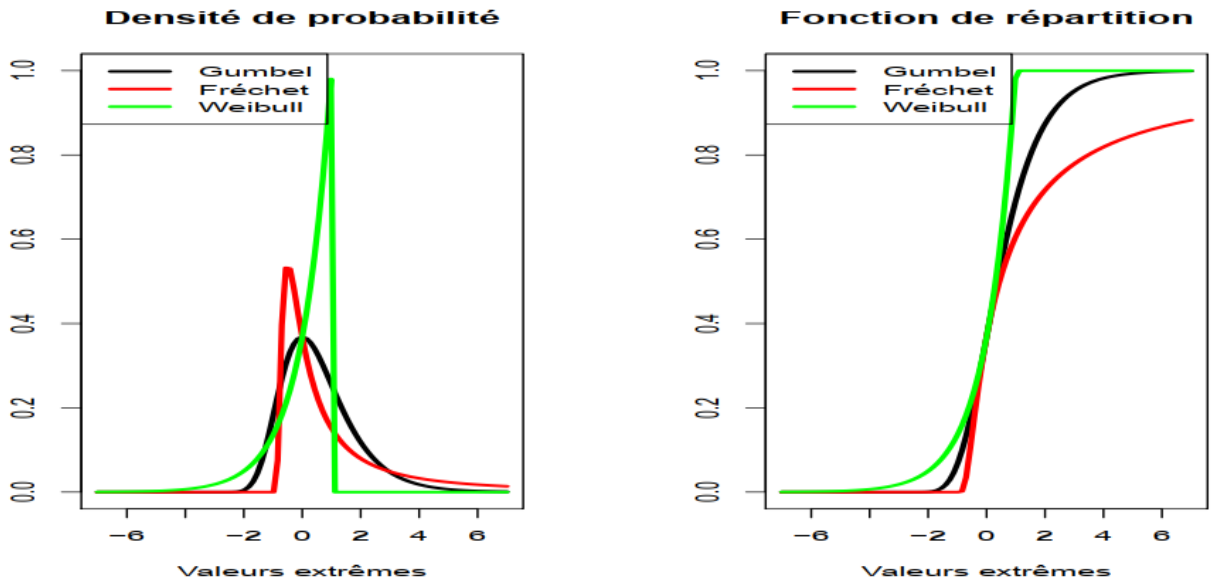

FIGURE 1.2 – Représentation de la densité et fonction de répartition , avec $\gamma = 1$ pour la loi de `Fréchet`, $\gamma = 0$ pour la loi de `Gumbel` et $\gamma = -1$ pour la loi de `Weibull`.

Les tableaux 1.1 , 1.2 , 1.3 regroupe quelques lois usuelles classées en fonction de leur domaine d'attraction

Distribution	1-F(x)	γ
Burr $(\beta, \tau, \lambda), \lambda > 0, \tau > 0, \beta > 0$	$(\frac{\beta}{\beta+x^{\tau}})^{\lambda}$	$1/\lambda\tau$
Fréchet $(1/\alpha), \alpha > 0$	$1 - \exp(-x^{-\alpha})$	$1/\alpha$
Loggamma $(m, \lambda), m > 0, \lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^{\infty} (\log u)^{m-1} u^{-\lambda-1} du$	$1/\lambda$
Loglogistique $(\beta, \alpha), \beta > 0, \alpha > 1$	$\frac{1}{1+\beta x^{\alpha}}$	$1/\alpha$
Pareto $(\alpha), \alpha > 0$	$x^{-\alpha}$	$1/\alpha$

TABLE 1.1 – Quelques distributions de type Pareto associés á un indice positif.

Distribution	1-F(x)	γ
ReverseBurr$(\beta, \tau, \lambda, \tau_F)\beta > 0, \tau > 0, \lambda > 0$	$\left(\frac{\beta}{\beta+(\tau_F-x)^{-\tau}}\right)^{\lambda}$	$-1/\lambda\tau$
Uniforme $(0, 1)$	$1 - x$	-1

TABLE 1.2 – Quelques distributions associés á un indice négatif.

Distribution	1-F(x)	γ
Gamma $(m, \lambda), m \in N, \lambda > 0$	$\frac{\lambda^{m-1}}{\Gamma(m)} \int_x^{\infty} u^{m-1} exp(-\lambda u) du$	0
Gumbel $(\mu, \beta), \mu \in R, \beta > 0$	$\exp(-\exp(-\frac{x-\mu}{\beta}))$	0
Logistique	$\frac{2}{1+\exp(x)}$	0
Lognormale $(\mu, \delta), \mu \in R, \delta > 0$	$\frac{1}{\sqrt{2\pi}} \int_x^{\infty} \frac{1}{u} \exp(-\frac{1}{2\delta^2}(log(u) - \mu)^2) du$	0
Weibull $(\lambda, \tau), \lambda > 0, \tau > 0$	$\exp(-\lambda x^{\tau})$	0

TABLE 1.3 – Quelques distributions associés á un indice nul.

Exemples de comportements limites

a.Loi exponentielle Considérons la loi exponentielle de paramétre $\lambda = 1$, La fonction de répartition de la loi exponentielle est donnée par

$$F(x) = \begin{cases} 1 - e^{-x}, & \text{si } x \geq 0\,; \\ 0, & \text{si } x < 0. \end{cases}$$

Le support de la loi étant $\mathbb{R}^+$ on a $x_F = +\infty$ d'où $X_{n:n} \xrightarrow{\text{p}} +\infty$. Effectuons la normalisation suivante :

$$\begin{aligned} \mathbb{P}\left(\tfrac{X_{n:n}-b_n}{a_n} \leq x\right) &= \mathbb{P}\left(X_{n:n} \leq a_n x + b_n\right) \\ &= F^n(a_n x + b_n) \\ &= [1 - \exp(-a_n x - b_n)]^n \end{aligned}$$

Si nous ont posons $a_n = 1$ et $b_n = \log n$, nous avons

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left(\tfrac{X_{n:n}-b_n}{a_n} \leq x\right) &= \lim_{n\to\infty} \left(1 - \tfrac{\exp(-x)}{n}\right)^n \\ &= \exp(-\exp(-x)) \\ &= \Lambda(x) \end{aligned}$$

$$\textit{puisque} \quad \exp(x) = 1 + x + \frac{x^2}{2!} + ... = \lim_{n\to+\infty} \left(1 + \frac{x}{n}\right)^n$$

c'est-à-dire que le maximum convenablement normalisé de la loi exponentielle converge vers la loi de Gumbel. Cette loi appartient au domaine d'attraction maximal de Gumbel.

b.loi de Pareto : Considérons la loi de Pareto de fonction de répartition :

$$F(x) = 1 - cx^{-\alpha}, \quad \textit{avec} \quad c > 0 \quad \textit{et} \quad \alpha > 0$$

En posant, $b_n = 0$ et $a_n = (nc)^{\frac{1}{\alpha}}$ alors on a pour $x > 0$

$$\begin{aligned} F^n(a_n x + b_n) &= \left(1 - c(a_n x)^{-\alpha}\right)^n = \left(1 - c(a_n^{-\alpha} x^{-\alpha})\right)^n \\ &= \left(1 - \tfrac{x^{-\alpha}}{n}\right)^n \\ &\to \exp(-x^{-\alpha}) \\ &= \Phi_\alpha(x) \end{aligned}$$

qui est la loi de Fréchet. La loi de Pareto appartient au domaine d'attraction maximal de Fréchet.

Aussi, les lois dans le domaine d'attraction de la loi de Fréchet sont parfois appelées lois de type Pareto.

c. Loi uniforme (loi à support borné à droite) La distribution de la loi uniforme sur [0,1] est : $F(x) = x$ si $0 \leq x \leq 1$: Pour $x < 0$ et $n > -x$, posons $a_n = \frac{1}{n}$ et $b_n = 1$, alor

$$\begin{aligned} F^n(a_n x + b_n) &= F^n(\tfrac{x}{n} + 1) = (\tfrac{x}{n} + 1)^n \\ &\to \exp(x) \\ &= \Psi_{-1}(x) \end{aligned}$$

La derniére est la loi de Weibull avec $\gamma = -1$.

d. Loi normale La fonction de répartition de la loi normale centrée et réduite $N(0,1)$est : $F(x) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy, \quad x \in \mathbb{R}$

$$F(-x) = \frac{1}{\sqrt{2\Pi}} \int_{x}^{\infty} e^{-\frac{y^2}{2}} dy, \quad x > 0$$

.

$$\leq \frac{1}{\sqrt{2\Pi}} \int_{x}^{\infty} \frac{y}{x} e^{-\frac{y^2}{2}} dy = \frac{1}{x\sqrt{2\Pi}} e^{-\frac{x^2}{2}}$$

D'aprés l'inégalité de Gordon [36] on aura

$$\frac{x}{1+x^2} \frac{1}{\sqrt{2\Pi}} e^{-\frac{x^2}{2}} \leq F(-x) \leq \frac{1}{x\sqrt{2\Pi}} e^{-\frac{x^2}{2}}$$

$$1 - F(x) \sim \frac{1}{x\sqrt{2\Pi}} e^{-\frac{x^2}{2}} \quad \textit{pour} \quad x \to +\infty$$

On en déduit que

$$\frac{1 - F(u + \frac{z}{u})}{1 - F(u)} \sim \Big[\Big(1 + \frac{z}{u^2}\Big)^{-1} \exp\Big(- \frac{1}{2(u + \frac{z}{u})^2} + \frac{1}{2} u^2 \Big) \Big] \sim \exp(-z), \quad \textit{pour} \quad u \to +\infty$$

Soit b_n la solution de $F(b_n) = 1 - \frac{1}{n}$ et posons $a_n = \frac{1}{b_n}$, alors

$$n(1 - F(a_n x + b_n)) = \frac{1 - F(a_n x + b_n)}{1 - F(b_n)} \to \exp(-x)$$

d'où

$$\lim_{n \to \infty} F^n(a_n x + b_n) = \lim_{n \to \infty} \Big(1 - \frac{\exp(-x)}{n}\Big)^n = \exp(-\exp(-x)) = \Lambda(x)$$

La limite est la loi de Gumbel qui correspond à celle obtenue dans le cas de la loi exponentielle.

1.6 Estimateurs de l'indice des valeurs extrêmes γ

On a vu dans ce chapitre que pour la majorité des fdr's F la loi asymptotique du maximum $X_{n:n}$ est une loi des valeurs extrêmes qui étant indexée par le paramètre de queue γ , ce paramètre apporte une information sur la forme de la queue de distribution de F : Notamment, selon que $\gamma > 0$, $\gamma < 0$ ou $\gamma = 0$, on distingue trois domaines d'attraction : Frechet, Weibull et Gumbel. Il existe dans la littérature de la TVE de nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes. Dans cette Section, on exposera uniquement trois estimateurs de γ , l'estimateur de `Pickands` [42], l'estimateur de Hill.B [17] et l'estimateur des moments (Dekkers et al.[20]). On donne également certaines de leurs propriétés statistiques. Ces estimateurs sont basés fortement sur les plus grandes statistiques d'ordre $X_{n-k:n} \leq ... \leq X_{n:n}$, où la statistique $X_{n-k:n}$ est alors dite statistique d'ordre intermédiaire.

1.6.1 Estimateur de Hill

Cet estimateur est le plus utilisé en théorie des valeurs extrêmes et a été largement étudié dans le cas de variables aléatoires i.i.d. Mason [47] et Deheuvels et al.[19] ont montré respectivement la consistance faible et forte qui ne dépend que du comportement de la suite k. Pour établir la normalité asymptotique, on a besoin de supposer que la fonction de répartition F est à variation régulière du second ordre. Plusieurs auteurs ont obtenu cette normalité, notamment Dekkers et al. [20], Csörgó et Mason [14], Davis et Resnick [49], Geluk et De Haan[31], Haeusler et Teugels [40].

Définition 1.6.1. Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et de même fonction de répartition F cette derniére appartenant au **DA(Fréchet)** avec un indice des valeurs extrêmes $\gamma > 0$. L'estimateur de Hill de γ est

$$\widehat{\gamma}_n^{(Hill)}(k) := \frac{1}{k}\sum_{i=1}^{k} \log(X_{n-i+1:n}) - \log(X_{n-k:n})$$

où $1 \leq k \leq n$ est une valeur à choisir par l'utilisateur.

Propriétés asymptotiques $\gamma_{k,n}^{Hill}$

Supposons $F \in$ **DA(Fréchet)** de paramètre de forme $\gamma > 0$, $k \longrightarrow \infty$ et $\frac{k}{n} \longrightarrow 0$ si $n \longrightarrow \infty$

(a) Consistance faible :

$$\widehat{\gamma}_{k,n}^{Hill} \xrightarrow{\mathbb{P}} \gamma, \quad n \longrightarrow \infty$$

(b) Consistance forte : Supposons que $k/\log\log(n) \longrightarrow \infty$ si $n \longrightarrow \infty$

$$\widehat{\gamma}_{k,n}^{Hill} \xrightarrow{\text{p.s}} \gamma, \quad n \longrightarrow \infty$$

(c) Normalité asymptotique : Pour établir la normalité asymptotique de l'estimateur $\gamma_{n,k}^{Hill}$ on a besoin d'une hypothèse sur la fonction à variation lente L. Il est en effet nécessaire d'imposer une condition qui spécifie la vitesse de convergence du rapport des fonctions à variations lentes vers 1 .

Condition : Il existe une constante réelle $\rho < 0$ et une fonction $\Delta(x) \to \infty$ quand $x \to \infty$, telle que pour tout $\lambda > 1$

$$log\frac{L(\lambda x)}{L(x)} \sim \Delta(x)\frac{\lambda^{\rho}-1}{\rho} \ \ quand \ \ x \to \infty$$

Cette condition appelée condition du second ordre est satisfaite pour la plupart des lois appartenant au **DA(Fréchet)**

Théorème 1.6.1. *(Beirlant et al.[6]) Soit $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes et identiquements distribuées (i.i.d) de fonction de répartition F avec $F \in$ **DA(Fréchet)**. Soit $(k_n)_{n\geq 1}$ une suite d'entiers telle que $1 < k_n \leq n$. Si $k_n \to \infty$ et $\frac{k_n}{n} \to \infty$ quand $n \to \infty$ et si la condition de second ordre est satisfaite avec $(\frac{k_n}{n}) \to \infty$ quand $n \to \infty$ Alors*

$$\sqrt{k_n}(\gamma_{k_n,n}^{Hill} - \gamma) \xrightarrow{\text{D}} \mathcal{N}\Big(0, \gamma^2\Big).$$

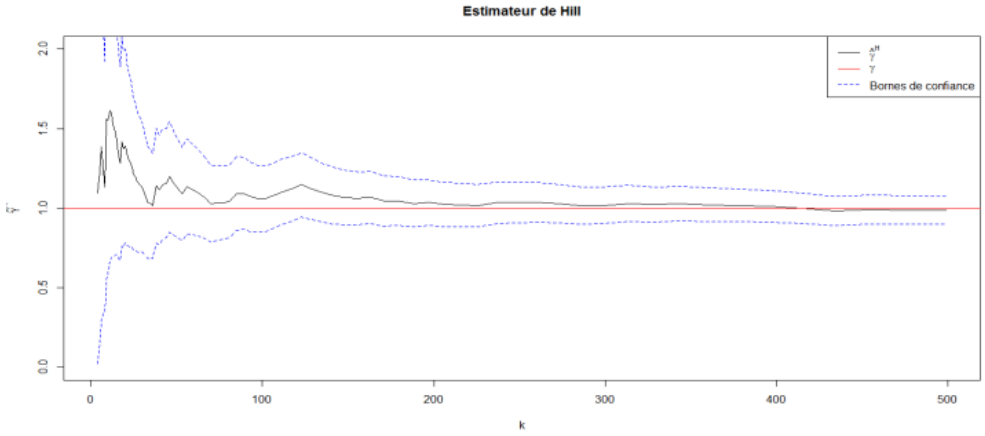

FIGURE 1.3 – Estimateur de Hill et l'intervalle de confiance de niveau 95%, pour l'IVE de la loi de Pareto standard ($\gamma = 1$) basé sur 1000 échantillons de 5000 observations [52].

1.6.2 Estimateur de Pickands

L'estimateur de `Pickands` a été introduit en 1975 par `James Pickands` dans [42] pour toute $\gamma \in \mathbb{R}$.

Définition 1.6.2. (`Estimateur de Pickands`). Soit $X_1, ..., X_n$, n va's *iid* de fdr $F \in \mathcal{DA}(\mathcal{H}_\gamma)$, où $\gamma \in R$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, l'estimateur de Pickand est défini pour tout $\gamma \in \mathbb{R}$ par :

$$\widehat{\gamma}^P = \widehat{\gamma}^P(k) = \frac{1}{\log 2} \log \left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}} \right), \ 1 < k < n \tag{1.20}$$

L'auteur démontre la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan [21]. Des améliorations de cet estimateur ont été introduites notamment par Drees [22] et Segers [50].

Théorème 1.6.2. *(`Pickands (1975)`, [42], `Dekkers et de Haan (1989)`, [17]) Supposons que* $F \in DA(\mathcal{H}_\gamma)$,$\gamma \in \mathbb{R}$ *et* $\frac{k}{n} \longrightarrow 0$ *si* $n \longrightarrow \infty$.

(a) Consistance faible :

$$\widehat{\gamma}^P_{k,n} \xrightarrow{\mathbb{P}} \gamma, \quad n \longrightarrow \infty$$

(b) Consistance forte Supposons que : $k/\log, \quad \log n \to \infty$ *si* $n \longrightarrow \infty$

$$\widehat{\gamma}^P_{k,n} \xrightarrow{\text{p.s}} \gamma, \quad n \longrightarrow \infty$$

(c) Normalite asymptotique :

$$\sqrt{k}(\gamma^P_{k,n} - \gamma) \xrightarrow{\text{D}} \mathcal{N}\left(0, \sigma^2\right) \quad n \longrightarrow \infty$$

où $\sigma^2 = \frac{\gamma^2(2^{2\gamma+1}+1)}{\left(2(2^\gamma-1)log2\right)^2}$

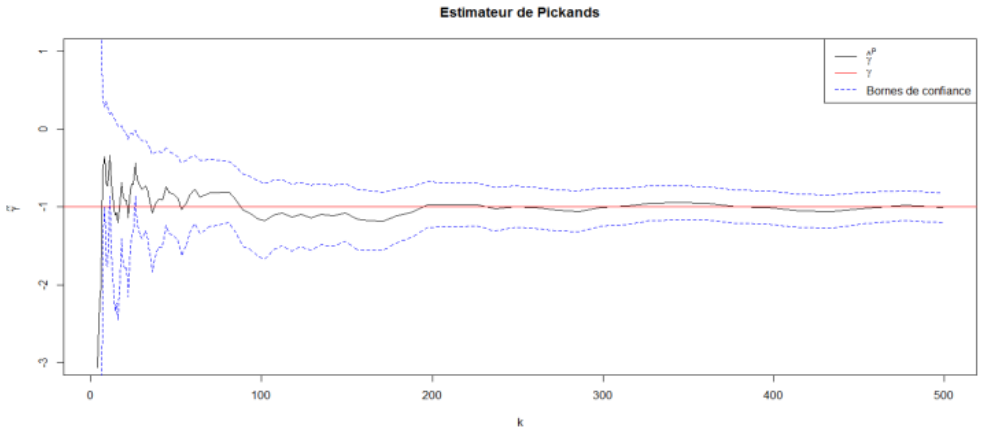

FIGURE 1.4 – Estimateur de Pickands avec l'intervalle de confiance au niveau 95% pour γ basés sur 1000 échantillons de taille 5000 pour la loi uniform standard ($\gamma = 1$)[52].

.

1.6.3 Estimateur des Moments

Un inconvénient de l'estimateur de Hill est qu'il est conçu seulement pour l'indice des valeurs extêmes des distributions à queues lourdes. En 1989, Dekkers et al. ont proposé dans [20] une extension de tout type de distribution, appelée estimateur de moment

Définition 1.6.3. (`Estimateur des Moments`). Pour $\gamma \in \mathbb{R}$, l'estimateur des moments est

$$\begin{aligned}\widehat{\gamma}_{k,n}^{M} &= M_{k,n}^{(1)} + 1 + \tfrac{1}{2}\left(1 - \frac{\left(M_{k,n}^{(1)}\right)^2}{M_{k,n}^{(2)}}\right)^{-1} \\ &= \widehat{\gamma}_{k,n}^{(Hill)} + 1 + \tfrac{1}{2}\left(1 - \frac{\left(\widehat{\gamma}_{k,n}^{(Hill)}\right)^2}{M_{k,n}^{(2)}}\right)^{-1}\end{aligned}$$

où

$$M_{k,n}^{(r)} = \frac{1}{k}\sum_{i=1}^{k}\Big(\log(X_{n-i+1:n}) - \log(X_{n-k:n})\Big)^r, \;\; r = 1, 2$$

Théorème 1.6.3. *(`Dekkers al. (1989)`, [20])) Supposons que $F \in DA(\mathcal{H}_\gamma)$,$\gamma \in \mathbb{R}$ et $\frac{k}{n} \longrightarrow 0$ si $n \longrightarrow \infty$.*

(a) Consistance faible :

$$\widehat{\gamma}_{k,n}^{M} \xrightarrow{\mathbb{P}} \gamma, \quad n \longrightarrow \infty$$

(b) Consistance forte Supposons que : $k/(\log n)^{\delta} \to \infty$ *si* $n \to \infty$, *avec delta* > 0

$$\widehat{\gamma}_{k,n}^{M} \xrightarrow{\text{p.s}} \gamma, \quad n \longrightarrow \infty$$

(c) Normalité asymptotique : *(voir Théorème 3.1 et corollaire 3.2 de [17])*

$$\sqrt{k}(\gamma_{k,n}^{M} - \gamma) \xrightarrow{\text{D}} \gamma\mathcal{N}\Big(0, \sigma^2\Big) \quad n \longrightarrow \infty$$

où

$$\sigma^2 = \begin{cases} 1 + \gamma^2 & si \sigma \geq 0. \\ (1+\gamma^2)(1-2\gamma)\big(4 - 8\frac{1-2\gamma}{1-3\gamma} + \frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\big) & si \gamma < 0. \end{cases} \tag{1.21}$$

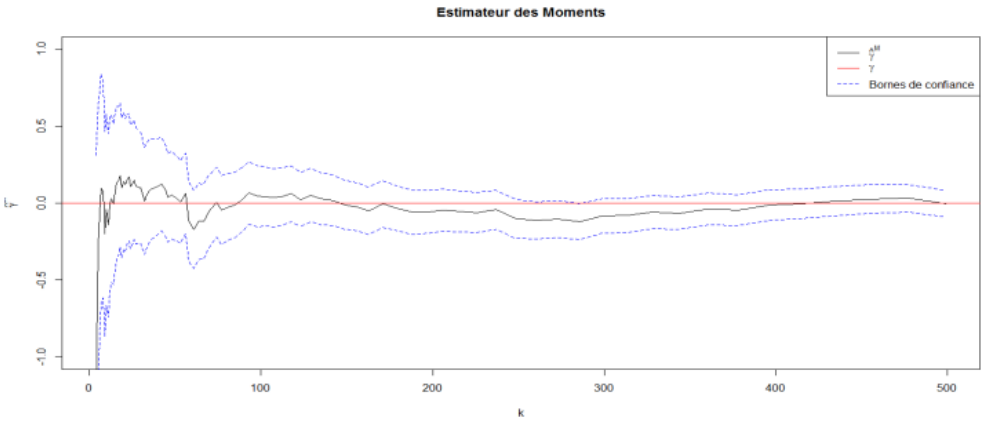

FIGURE 1.5 – Estimateur des Moments et l'intervalle de confiance de niveau 95%, pour l'IVE de la loi de Gumbel ($\gamma = 0$) basés sur 1000 échantillons de taille 5000 [52]

.

1.7 Conclusion

Dans ce chapitre, nous avons fait un rappel sur la théorie des valeurs extrémes, en mentionnant les différentes caractéristiques et les notions de base qui sont trés utiles pour l'estimation de l'indice des valeurs extrêmes γ

Chapitre 2

Estimation de l'indice des valeurs extrémes avec les données censurées

2.1 Introduction

Le problème des données manquantes, incomplètes ou erronées est très vaste et a suscité beaucoup d'intérêt parmi les statisticiens ces dernières années. L'attitude vis-à-vis de ce type de données a longtemps été soit de les éliminer, soit de minimiser le mauvais impact qu'elles pourraient avoir sur des procédures statistiques adaptées à des données complètes. Dans le domaine des durées de survie, les données sont souvent incomplètes à cause de deux phénomènes distincts : la censure et la troncature

2.2 Données de survie

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information. Au lieu d'observer des réalisations indépendantes et identique ment distribuées de durée X, on observe la réalisation de la variable X soumise à diverses perturbations indépendantes ou non de l'événement étudié.

2.3 Données Censurées

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

Définition 2.3.1. {Variable de censure}. La variable de censure C est définie par la non-observation de l'événement étudié. Si au lieu d'observer X, on observe C

1. $X > C$ est censure à droite .
2. $X < C$ est censure à gauche .
3. $C_1 < X < C_2$ est censure par intervalle .

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

- son temps de survie X_i , de fonction de répartition F
- son temps de censure C_i , de fonction de répartition G
- la durée réellement observée T_i , de fonction de répartition H.

2.3.1 Censure à droite

Les données sont censurées à droite en cas d'absence d'information sur le statut de l'individu après sa dernière observation. Deux cas sont possibles : l'individu n'a pas réalisé l'événement à la fin de l'étude (exclu-vivant) ou il a quitté l'étude en cours de route (perdu de vue ou exclu). Ainsi,en présence de censure à droite les variables d'intérêt ne sont pas toutes observées. En présence de censure à droite, l'information disponible pour chaque individu est $\{T_i, \delta_i\}$ avec :

- T_i , la durée réellement observée : $T_i = X_i \wedge C_i = \min(X_i, C_i)$.
- $\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$, un indicateur qui vaut 1 si l'événement est observé et 0 si l'individu est censuré.

Un exemple typique est celui où l'événement considéré est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation. On trouve aussi ce genre de phénomène dans les études de fiabilité quand la panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant. On peut aussi trouver ces genres de phénomènes en hydrologie, en pluviométrie, ... L'expérimentateur peut fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

`Exemple de censure droite :` Un exemple classique de censure droite est celui où l'étude porte sur la durée de survie X de patients atteints d'une certaine maladie. Pour les patients perdus de vue au bout du temps C alors qu'ils étaient encore vivants, C censure X à droite puisque, pour eux, X est inconnue mais supérieure à C , $X > C$.

Exemple 2.3.1. Considérons une étude relative à la durée de survie de patient soumis à un traitement particulier. L'évènement auquélle ou s'intéresse est la mort de patient. Tous les individus sont suivis pendant les 52 semaines suivant la première administration du traitement. On considère plus particulièrement 3 sujets qui vont permettre d'illustrer certaines des caractéristiques les plus fréquentes des données de survie et notamment deux cas possibles de censure à droite

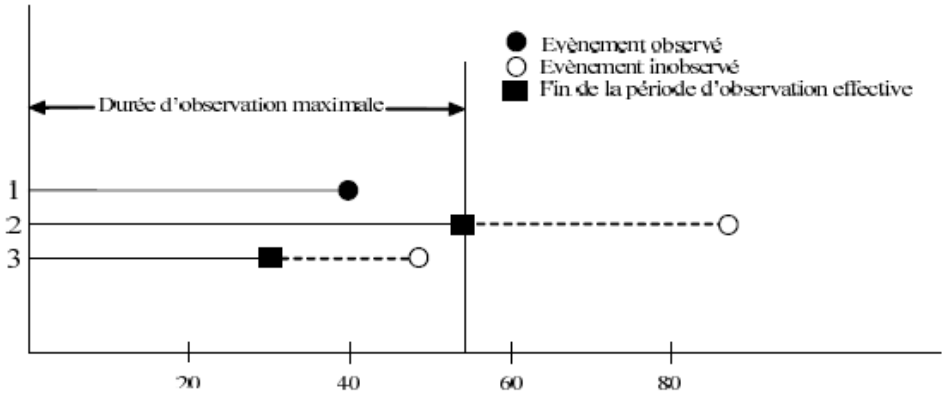

FIGURE 2.1 – Exemple de censure droite

.

- L'individu 1 est décédé 40 semaines après le début du traitement. Il s'agit d'une observation non censurée.
- La deuxième personne est toujours vivante au terme des 52 semaines d'observation. Elle décédera après 90 semaines mais cette information n'est pas connue lorsque la

constitution de la base de données est arrêtée. Même incomplète, l'information est utile puisque l'on sait que le temps de survie réel est supérieur à 52 semaines. Il ne faut donc pas l'éliminer de la base sous peine par exemple de biaiser vers le bas l'estimation de la durée moyenne de survie. Il s'agit d'une censure déterministe car elle ne dépend pas de l'individu considéré mais des conditions de l'expérimentation.

– La troisième personne décède après 50 semaines mais cet évènement n'est pas enregistré dans la base de données car le patient concerné n'a pas pu être effectivement suivi que pendant 30 semaines. C'est un exemple de censure aléatoire car elle échappe au contrôle de l'expérimentateur. Là encore l'information est incomplète mais non nulle. Par exemple savoir que cet individu a survécu au moins 30 semaines est pertinent pour l'estimation du taux de survie à 20 semaines.

2.3.2 Censure à gauche

Il y a censure à gauche lorsque l'individu a déjà subi l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue En présence de censure à gauche, l'information disponible pour chaque individu est $\{T_i, \delta_i\}$ avec :

– T_i , la durée réellement observée : $T_i = X_i \vee C_i = \max(X_i, C_i)$.

– $\delta_i = \mathbb{1}_{\{X_i \geq C_i\}}$, un indicateur qui vaut 1 si l'événement est observé et 0 si l'individu est censuré.

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 exemplelors 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.

2.3.3 Censure double ou mixte

On dit qu'on a une censure double ou mixte si on a des données censurées à droite et des données censurées à gauche dans le même échantillon. Plusieurs modèles non-paramétriques

ont été présentés pour l'étude de la double censure. Par exemple, le modèle de Turnbull [54] est le plus utilisé, et plusieurs travaux sont basés sur ce modèle.

`Exemple de censure double ou mixte` : Un ethnologue étudie la durée d'apprentissage d'une tâche. Cette durée est une variable aléatoire X et C est l'âge de l'enfant. Pour les enfants qui savent déjà accomplir la tâche, C censure X à gauche car pour eux X est inconnu mais inférieur à C , $X < C$. D'ailleurs cet exemple comporte aussi des censures droites. En effet, les enfants qui ne savent pas encore accomplir la tâche en question lors du départ de l'ethnologue sont censurés à droite par la durée d'apprentissage C' observée par l'ethnologue, car $X > C'$.

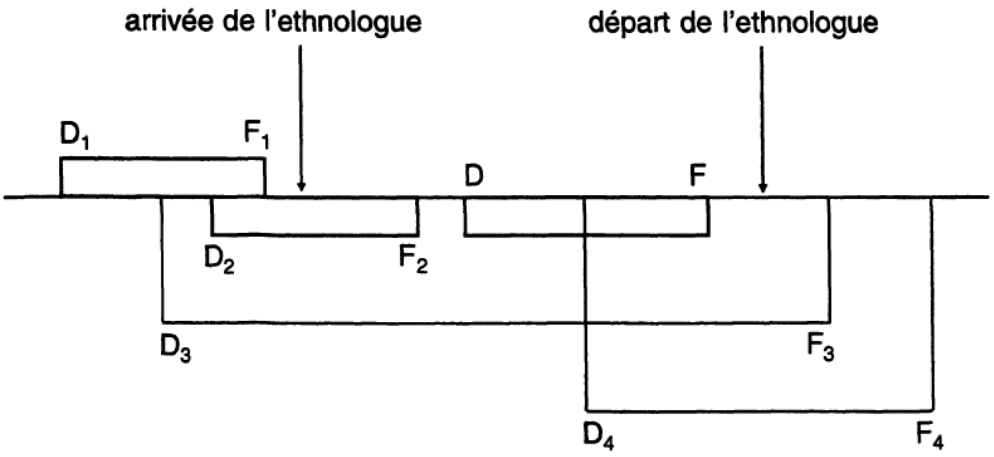

FIGURE 2.2 – Exemple de censures droite et gauche

Dans la figure 2.2, D_i est le début de l'apprentissage, F_i la fin, pour le sujet i.

D_1F_1 est censuré à gauche par l'âge C de l'enfant.

D_2F_2 , D_3F_3 , D_4F_4 bien qu'étant de trois types différents, ils sont tous les trois censurés à droite : le premier par l'âge de l'enfant, le second par la durée de séjour de l'ethnologue et le troisième par la durée d'aprentissage observée par l'ethnologue.

DF n'est pas censuré.

2.3.4 Censure par intervalle

[44] Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit.

Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[c, +\infty[$ et $[0, c]$ respectivement.

L'exemple qui sera présenté par la suite est constitué d'études de suivi de patients prenant des biothérapies. L'apparition d'anticorps anti-biothérapies ou Anti Drug Antibody (ADA) chez un patient ne peut être constaté que lors d'une visite. Sur le graphique ci-dessous, l'individu 2 a effectué deux visites : la première en c_1 et la seconde, au temps c_3. Si lors de la seconde visite, il est déclaré ADA-positif, la seule information sur la date de positivité du patient est un intervalle de temps entre les deux visites, soit $X_2 \in [c_1, c_3]$.

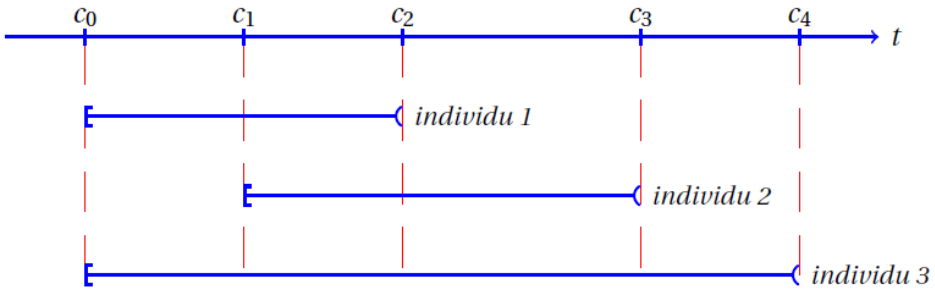

FIGURE 2.3 – Censure par intervalle

Remarque 2.3.1. Si, au lieu de X, on observe $C_1 < C_2$ tels que $C_1 < X < C_2$ (X non observé), il y a censure par intervalle. En particulier, la censure gauche peut être considérée comme une censure par un intervalle tel que $C_1 = -\infty$, et la censure droite par un intervalle tel que $C_2 = +\infty$.

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

- **Censure de type I (fixée) :** L'expérimentateur fixe une valeur (une date par exemple non aléatoire de fin d'expérience). Soit C une valeur fixée. Par exemple en censure à droite, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on observe X_i que lorsqu'elle est inférieure ou égale à une durée fixée C. On observe donc une variable T_i telle que $T_i := min(X_i, C)$, $i = 1, ..., n$. Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles . Par exemple, on peut tester la durée de vie de n objet identiques (ampoules) sur un intervalle d'observation fixé $[0, c]$. En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps c sont sacrifiées).

- **Censure de type II (attente) :** L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie. Par exemple en épidémiologie on décide d'observer les durées de survie des n patients jusqu'à ce que r $(1 \leq r \leq n)$ d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{i:n}$ et $T_{i:n}$ les statistiques d'ordre des variables X_i et T_i . La date de censure est donc $X_{r:n}$ et on observe

$$\begin{cases} T_{i:n} = X_{i:n}, & \text{si } i \leq r\,; \\ T_{i:n} = X_{r:n}, & \text{si } i \geq r. \end{cases}$$

- **La censure de type III (ou censure aléatoire de type I) :** Soient $C_1, ..., C_n$ des variables aléatoires *i.i.d.* On observe les variables $T_i = X_i \wedge C_i$: L'information disponible peut être résumée par :
  - la durée réellement observée T_i,
  - un indicateur $\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$:

$$\begin{cases} \delta_i = 1, & \text{si l'événement est observé (d'où } T_i = X_i\text{). On observe les "vraies" durées ou} \\ & \text{les durées complètes. ;} \\ \delta_i = 0, & \text{si l'individu est censuré (d'où } T_i = C_i\text{). On observe des durées incomplètes} \\ & \text{(censurées). ;} \end{cases}$$

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

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

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

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

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

2.4 Troncature

La troncature est une variante de censure ce qui se produit à la conception de l'étude lorsque la nature de l'observation incomplète est du a un inhérent de processus de sélection systématique. On a trois types de troncature, comme suit Données tronquées

2.4.1 Données tronquées

Une autre situation dans laquelle les données incomplètes apparaissent est celle des données tronquées. Les troncatures différent des censures au sens où elles concernent l'échantillonnage lui même. Une observation est dite tronquée si elle est conditionnelle à un autre événement. On dit que la variable X de durée de vie est tronquée si X n'est observable que sous une certaine condition dépendante de la valeur de X

1. **La troncature à gauche** Soit Z une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X n'est observable que si $X > Z$, On observe le couple (X, Z), avec $X > Z$,

   Par exemple, si la durée de vie d'une population est étudiée à partir d'une cohorte tirée au sort dans cette population, seule la survie des sujets vivants à l'inclusion pourra être étudiée (il y a troncature à gauche car seuls les sujets ayant survécu jusqu'à la date d'inclusion dans la cohorte sont observables).

2. **La troncature à droite** De même, il y a troncature à droite lorsque X n'est observable que si $X < Z$:

3. **La troncature par intervalle** Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle. Par exemple, on rencontre ce type de troncature lors de l'étude des patients d'un registre . les patients diagnostiqués avant la mise en place du registre ou répertoriés après la consultation du registre ne seront pas inclus dans l'étude.

2.5 Distributions de la durée de survie

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

2.5.1 Fonction de survie S

Pour un t fixé la fonction de survie est la probabilité de survivre jusqu'à l'instant t , aussi appelée queue de distribution, qu'on note par $S(t)$ ou $\overline{F}(t)$ est définie sur $\mathbb{R}_+$ par

$$S(t) = \overline{F}(t) = 1 - F(t) = \mathbb{P}(X > t). \tag{2.1}$$

Définition 2.5.1. (*Fonctions de répartition empiriques et de survie*).

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

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X \leq t\}}, \forall t \geq 0, \tag{2.2}$$

et

$$\overline{F}_n(t) = 1 - F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X > t\}}, \forall t \geq 0, \tag{2.3}$$

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

Remarque 2.5.1.

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

Pour $1 \leq i \leq n$, il existe une autre version de la définition des fonctions (2.2) et (2.3) en utilisant les statistiques d'ordres comme suit :

$$F_n(t) = \begin{cases} 0 & \text{si} \quad t \leq X_{1:n} \\ \frac{i}{n} & \text{si} \quad X_{i:n} \leq t \leq X_{i+1:n}, \\ 1 & \text{si} \quad t < X_{n:n} \end{cases} \tag{2.4}$$

et

$$\overline{F}_n(t) = \begin{cases} 1 & \text{si} \quad t \leq X_{1:n} \\ 1 - \frac{i}{n} & \text{si} \quad X_{i:n} \leq t \leq X_{i+1:n}, \\ 0 & \text{si} \quad t < X_{n:n} \end{cases} \tag{2.5}$$

Pour une représentation graphique de ces deux fonctions, voir la Figure (2.4).

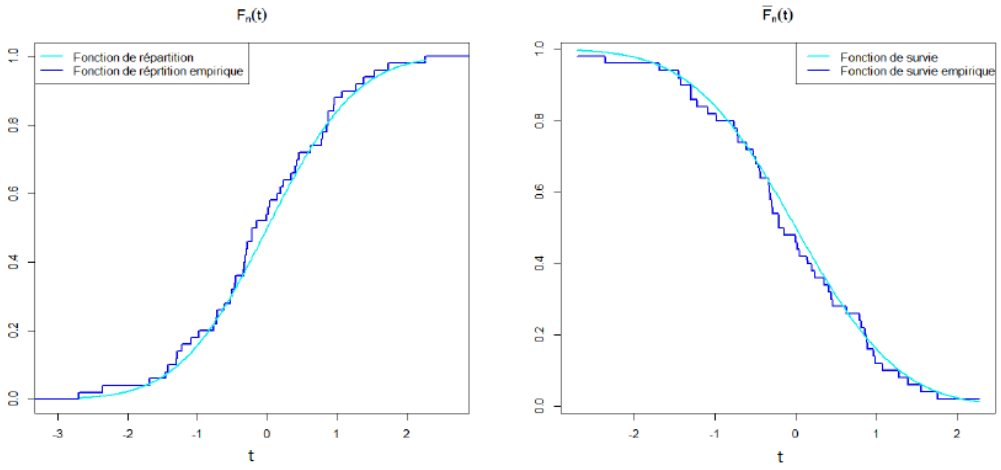

FIGURE 2.4 – Fonctions de répartition empiriques (gauche) et de survie (droite) d'un échantillon Gaussien standard de taille 50.

2.5.2 Fonction de Densité f

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

Définition 2.5.2. (*Fonction de densité*). Si F admet une dérivée par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$, la fonction de densité de probabilité existe, elle est définie pour tout $t \geq 0$, par

$$f(t) = -S'(t) = \frac{dF(t)}{dt} := \lim_{h \longrightarrow 0} \frac{P(t < X < t + h)}{h},$$

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

2.6 Estimateur de Kaplan-Meier

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Soit $X_1, ..., X_n$ une suite de va's d'intérêt *iid* positives, de fdr commune F et $C_1, ..., C_n$ une suite de va's de censure *iid* positives, de fdr continue G.

On suppose aussi que C_i sont indépendantes des X_i : Soit $\{(T_i, \delta_i), 1 \leq i \leq n\}$ l'échantillon réellement observé défini dans le cadre le plus fréquent d'une censure à droite aléatoire de type I , dans la suite on supposera que la variable T a comme fdr H.

Malheureusement, en présence de censure, la fonction de survie empirique (2.3) de la variable X n'est plus valable car elle dépend de va's parmi X, $X_1, ..., X_n$ qui ne sont pas observées. Afin d'estimer la loi de X, il a été donc nécessaire de construire un estimateur de fonction de survie (2.1) en présence de données censurées. En 1958, Kaplan et Meier [45] ont introduit les estimateurs non-paramétiques du maximum de vraisemblance de F et G Deheuvels et Einmahl [18])

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t, c'est-à-dire, si $t'' < t' < t$

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

En considérant les temps d'événements (décès et censure) distincts $T_{i:n}$, $(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$\mathbb{P}(X > T_{i:n}) = \prod_{k=1}^{i} \mathbb{P}(X > T_{k:n} \mid X > T_{k-1:n}), \ \ avec \ T_{0:n} = 0$$

Considérons les notations suivantes :

- Y_i le nombre d'individus à risque de subir l'événement juste avant le temps $T_{i:n}$,
- d_i le nombre de décès en $T_{i:n}$,

Alors la probabilité p_i de mourir dans l'intervalle $]T_{i-1:n}, T_{i:n}]$ sachant que l'on était vivant en $T_{i-1:n}$, i.e.$p_i = P(X \leq T_{i:n} \mid X > T_{i-1:n})$, peut être estimée par

$$\widehat{p_i} = \frac{d_i}{Y_i},$$

Comme les temps d'événements sont supposés distincts, on a

$d_i = 0$ en cas de censure en $T_{i:n}$, *i.e.* quand $\delta_i = 0$,

$d_i = 1$ en cas de décès en $T_{i:n}$, *i.e.* quand $\delta_i = 1$.

On obtient alors l'estimateur de Kaplan-Meier :

$$\widehat{S}_n(t) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(1-\frac{\delta_{[i:n]}}{Y_i}\right) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(1-\frac{\delta_{[i:n]}}{n-(i-1)}\right) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}}, pour\, t < T_{n:n} \tag{2.6}$$

et

$$\widehat{G}_n(t) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}(1-\frac{1-\delta_{[i:n]}}{Y_i}) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(1-\frac{1-\delta_{[i:n]}}{n-(i-1)}\right) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_{[i:n]}}, pour\, t < T_{n:n} \tag{2.7}$$

$\delta_{[i:n]}$ le concomitant de la $i^{\text{éme}}$ statistique d'ordre, c'est-à-dire, $\delta_{[i:n]} = \delta_j$ si $T_{(i:n)} = T_j$, $1 \leq j \leq n$ On présente ici une autre écriture de l'estimateur de Kaplan-Meier sous forme de somme. Cet écriture peut être trouvée dans le livre de Reiss et Thomas [48]

$$\widehat{F}_n(t) = 1 - \widehat{S}_n(t) := \sum_{i=2}^{n} W_{i:n}\mathbb{1}_{\{T_{i:n}\leq t\}}. \tag{2.8}$$

par des raisonnements combinatoires, Stute et Wang [53] obtiennent l'expression suivante des sauts de l'estimateur de Kaplan-Meier

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

où $W_{i:n}$ est le saut à la $i^{\text{ème}}$ observation $T_{i:n}$ dans l'échantillon ordonné

Remarque 2.6.1. Les estimateurs (2.6) et (2.7) sont parfois écris de la manièr suivante :

$$\widehat{S}_n(t) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(1-\frac{\delta_{[i:n]}}{n-(i-1)}\right) = \prod_{i=1}^{n}\left(\frac{\delta_{[i:n]}}{n-i+1}\right)^{\mathbb{1}_{T_{i:n}\leq t}}, pour\, t < T_{n:n}$$

et

$$\widehat{G}_n(t) = \prod_{\substack{i=1 \\ T_{i:n}\leq t}}^{n}\left(1-\frac{1-\delta_{[i:n]}}{n-(i-1)}\right) = \prod_{i=1}^{n}\left(1-\frac{1-\delta_{[i:n]}}{n-i+1}\right)^{\mathbb{1}_{T_{i:n}\leq t}}, pour\, t < T_{n:n}$$

Dans la littérature de l'analyse de survie, un grand nombre des auteurs ont été consacrés à l'étude des propriétés asymptotiques de l'estimateur de Kaplan-Meier. Par exemple. La consistance uniforme a été étudiée par Shorack et Wellner [51], Wang [57], Stute et Wang [53] et plus récemment par Gill [34]. La normalité asymptotique a été étudiée par Breslow et Crowley [11], Gill [33] et Gill [32].

Proposition 2.6.1. *(Propriétés asymptotiques de $\widehat{S}_n$).*

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

$$\boldsymbol{E}[\widehat{S}_n(t)] = S(t), \quad quand \;\; n\to\infty.$$

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

$$\sup_{0\leq t<x_H} \mid \widehat{S}_n(t) - S(t) \mid \xrightarrow{\text{p.s}} 0, \quad quand \; n\to\infty.$$

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

$$\sqrt{n}(\widehat{S}_n(t) - S(t)) \xrightarrow{\text{D}} \mathbb{X}_t.$$

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

$$cov(\mathbb{X}_s, \mathbb{X}_t) = S(s)S(t) \int_0^{\min(s,t)} \frac{dH(t)}{(S(t))^2}$$

L'estimateur $S(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance. $S(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car les temps de décès ne sont pas connus)

2.7 Estimation de l'IVE en présence de censure $\widehat{\gamma}_X^c$

Nous travaillons dans l'espace probabilisé (Ω, A, P) et soit l'échantillon $X_1, ..., X_n$ de variables aléatoires définit sur (Ω, A, P). Sa fonction répartition F et sa queue de distribution :

$$1 - F \in RV_{(-1/\gamma_X)}.$$

Soit le deuxiéme échantillon $C_1, ..., C_n$ des variables aléatoires iid, de fonction de répartition G et de queue de distribution :

$$1 - G \in RV_{(-1/\gamma_C)}.$$

Alors les variables T_i définies par :

$$T_i = X_i \wedge C_i \;\; , \;\; i = 1, ..., n$$

avec, T_i sont des variables indépendantes de loi H liées à F et G par la relation :

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

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

$$1 - F(x) = x^{-\frac{1}{\gamma_X}} L_X(x) \;\; et \;\; 1 - G(x) = x^{-\frac{1}{\gamma_C}} L_C(x)$$

avec $L_X(x)$ et $L_C(x)$ des fonction à variation lente. Alors,

$$\begin{aligned} 1 - H(x) \; &= (1 - F(x))(1 - G(x)) \\ &= x^{-\frac{1}{\gamma_X}} L_X(x) x^{-\frac{1}{\gamma_C}} L_C(x) \\ &= x^{-(\frac{1}{\gamma_X} + \frac{1}{\gamma_C})} L_X(x) L_C(x) \\ &= x^{-\frac{\gamma_X + \gamma_C}{\gamma_X \gamma_C}} L(x) \\ &= x^{-\frac{1}{\gamma}} L(x) \;\; avec \;\; L(x) = L_X(x) L_C(x) \end{aligned}$$

Donc, H est une fonction de répartition appartenant au domaine d'attraction de Fréchet :

$$1 - H(x) \in RV_{(-1/\gamma)}$$

avec

$$\gamma := \frac{\gamma_X \gamma_C}{\gamma_X + \gamma_C}$$

Si F et G sont supposées être dans le dommaine d'attraction maximale $DA(\mathcal{H}_{\gamma_X})$ et $DA(\mathcal{H}_{\gamma_C})$ respectivement où $\gamma_X, \gamma_C \in \mathbb{R}$ avec points terminales x_F et x_G , où $x_F = \sup\{x, F(x) < 1\}$, alors cela signifie que $H \in DA(\mathcal{H}_\gamma)$. Einmahl et al.(2008,[25]) ont examiné les trois cas les plus intéressants suivants :

$$\begin{cases} cas1, \quad \gamma_X > 0 \ , \ \gamma_C > 0, & \gamma = \frac{\gamma_X \gamma_C}{\gamma_X + \gamma_C} \\ cas2, \quad \gamma_X < 0 \ , \ \gamma_C < 0, x_F = x_G \ , & \gamma = \frac{\gamma_X \gamma_C}{\gamma_X + \gamma_C} \\ cas3, \quad \gamma_X = \gamma_C = 0 \ , \ x_F = x_G = \infty. & \gamma = 0 \end{cases} \tag{2.9}$$

Dans le cas 3, nous définissons également, pour une présentation commode, $\gamma := \frac{\gamma_X \gamma_C}{\gamma_X + \gamma_C} = 0$. Les autres possibilités ne sont pas trés intéressantes. Pratiquement, ils sont trés proches du cas non censuré, qui a été étudié en détail dans la littérature (cela arrive, en particulier, quand $\gamma_X > 0$ et $\gamma_C < 0$) ou la situation complétement censurée, où l'estimation est impossible (cela arrive, en particulier, quand $\gamma_X < 0$ et $\gamma_C > 0$).

Définition 2.7.1. Soit $\{(T_j, \delta_j) \ , \ 1 \leq j \leq n\}$ un échantillon de couple de $v.a's$ (T, δ) : Soient $T_{1:n} \leq ... \leq T_{n:n}$ représente les statistiques d'ordre associées à l'échantillon $(T_1, ..., T_n)$: Avec les $\delta_{[1:n]}, ..., \delta_{[n:n]}$ sont les indicateurs de censure retenues respectivement avec l'échantillon $T_{1:n}, ..., T_{n:n}$,

$$\delta_{[i:n]} = \delta_j \quad si \quad T_{i:n} = T_j$$

Beirlant et al (2007[5]). ont proposé différents estimateurs de γ_X, l'indice des valeurs extrêmes associé à F dans le cas des données censurées ces derniers sont tous construits de façon similaire, à partir d'un estimateur non adapté à la censure, par exemple l'estimateur de Hill. Ces estimateurs basés sur les observations T_j, estiment par conséquent l'indice γ de H. Il s'agit alors de les modifier de façon à estimer γ_X et non γ. Une façon de procéder consiste à diviser ces estimateurs usuels (non adaptés à la censure) par la proportion de données non censurées au-delà d'un seuil u, c'est-à-dire à utiliser

$$\widehat{\gamma}_X^{(\bullet,c)} = \widehat{\gamma}_X^{(\bullet,c)}(k) := \frac{\widehat{\gamma}_T^{\bullet}}{\widehat{p}(k)} \tag{2.10}$$

où

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

avec,

k est le nombre d'excés au-delà de $u = T_{n-k:n}$ (le nombre des valeurs extrêmes). $\widehat{p}$ estime $p = \frac{\gamma_C(x)}{\gamma_X(x)+\gamma_C(x)}$ ($\widehat{p} \to p$, quand $n \to \infty$),(où p représente la proportion de données observées

dans la queue à droite de la distribution), par conséquent $\widehat{\gamma}_T^{(\bullet)}(k)$ estimateur de γ divisé par $\frac{\gamma_C(x)}{\gamma_X(x)+\gamma_C(x)}$ qui est égal à γ_X.

$\widehat{\gamma}_T^{(\bullet)}(k)$ peut être n'importe quel estimateur non adapté à la censure, en particulier l'estimateur de Hill $\widehat{\gamma}_T^{(Hill)}(k)$.

Pour adapter l'estimateur de Hill dans le cas censuré nous allons diviser cet estimateur par la proportion de données non censurées des k plus grandes valeurs de T, Alors l'estimateur de Hill adapté à l'indice de queue $\widehat{\gamma}_X^c$ est difini par

$$\widehat{\gamma}_X^c = \frac{\widehat{\gamma}_T^{Hill}(k)}{\widehat{p}}$$

où

$$\widehat{\gamma}_T^{Hill}(k) = \frac{1}{k}\sum_{i=1}^{k} \log T_{n-i+1:n} - \log T_{n-k:n}$$

alors,

$$\widehat{\gamma}_X^c = \frac{\frac{1}{k}\sum\limits_{i=1}^{k} \log T_{n-i+1:n} - \log T_{n-k:n}}{\frac{1}{k}\sum\limits_{i=1}^{k} \delta_{[n-i+1:n]}}$$

Einmahl et al (2008,[25]) ont établi de façon unifiée, la normalité asymptotique de tout estimateur de l'indice des valeurs extrêmes écrit sous la forme (2.10) dans le cas où le seuil choisi, u est aléatoire et égal à $T_{n-k:n}$ la $n-k$-ième statistique d'ordre de l'échantillon $T_1, ..., T_n$.

Propriété asymptotique de l'estimateur de l'indice des valeurs extrêmes

Pour déterminer les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes nous avons besoin de la fonction suivante comme définie dans [25] :

$$p(z) = \mathbb{P}(\delta = 1 | T = z)$$

Nous pouvons l'écrire d'une autre manière

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

où f et g désignent respectivement les densités conditionnelles de X et C, alors,

$$\lim_{z\to\infty} p(z) = \frac{\gamma_C}{\gamma_X + \gamma_C} = p.$$

Nous proposons quelques exemples avec des lois à « queues lourdes ».

Exemple

Supposons X et C sont respectivement de Pareto(γ_X) et Pareto(γ_C), c'est-à-dire pour tout $x \geq 1$,

$$F_X(x) = 1 - x^{-\frac{1}{\gamma_X}}$$

et

$$G_C(x) = 1 - x^{-\frac{1}{\gamma_C}}$$

où γ_X et γ_C sont des paramètres fonctionnels positifs.
On obtient ,

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

ce qui implique T suit une Pareto $(\frac{\gamma_X\gamma_C}{\gamma_X+\gamma_C})$.

Nous pouvons à présent calculer la fonction $p(.)$

$$\begin{aligned} p(z) &= \frac{(1-G(z))f(z)}{(1-G(z))f(z)+(1-F(z))g(z)} \\ &= \frac{z^{-1/\gamma_C}\frac{1}{\gamma_X}z^{-1/\gamma_X}}{z^{-1/\gamma_C}\frac{1}{\gamma_X}z^{-1/\gamma_X}+z^{-1/\gamma_X}\frac{1}{\gamma_C}z^{-1/\gamma_C}} \\ &= \frac{\frac{1}{\gamma_X}z^{-1/\gamma_X-\frac{1}{\gamma_C}}}{(\frac{1}{\gamma_X}+\frac{1}{\gamma_C})z^{-1/\gamma_X-1/\gamma_C}} \\ &= \frac{\frac{1}{\gamma_X}}{(\frac{1}{\gamma_X}+\frac{1}{\gamma_C})} \\ &= \frac{\gamma_C}{\gamma_X+\gamma_C} \end{aligned}$$

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

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

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

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

$$1 < k < n, \quad k \to \infty et k/n \to 0 \quad , \quad n \to \infty$$

- $\mathbb{C}2 : \sqrt{k}b(\frac{n}{k}) \to \alpha_1 \in \mathbb{R}$
- $\mathbb{C}3 : \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \Big[p(H^{\leftarrow}(1-\frac{i}{n})) - p\Big] \to \alpha_2 \in \mathbb{R}$
- $\mathbb{C}4$: Soit $c > 0$ et $\mathcal{A}(s,t) := \Big\{1 - k/n \leq t < 1; \mid t - s \mid \leq C\sqrt{k}/n, s < 1\Big\}$ si $n \to \infty$,

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

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

Théorème 2.7.1. *Sous les condition* $\mathbb{C}1$ - $\mathbb{C}4$ *et s'il existe* b_0 *et* σ *telles que*

$$\sqrt{k}(\widehat{\gamma}^{(c,.)}_{T,k,n} - \gamma) \longrightarrow \mathcal{N}\Big(\frac{1}{p}\alpha_1 b_0, \sigma^2\Big)$$

Alors, nous avons

$$\sqrt{k}(\widehat{\gamma}^{(c,.)}_{T,k,n} - \gamma_X) \longrightarrow \mathcal{N}\Big(\frac{1}{p}(\alpha_1 b_0 - \gamma_X \alpha_2), \frac{\sigma^2 + \gamma_X^2 p(1-p)}{p^2}\Big)$$

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

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

$$\sqrt{k}\Big(\widehat{\gamma}^{(c.Hill)}_{T} - \gamma_X\Big) \xrightarrow{\mathbb{D}} \mathcal{N}\Big(u^{(c.Hill)}, \frac{\gamma_X^3}{\gamma}\Big)$$

On a :

$$\mathbb{E}\Big(\sqrt{k}\Big(\widehat{\gamma}^{(c,H)}_{T,k,n} - \gamma_X\Big)\Big) = \mu^{(c,H)} := -\frac{\gamma_X \alpha_2}{p} + \frac{\alpha_1}{p}\frac{\gamma}{\bar{\rho}+\gamma(1-\bar{\rho})}$$

$$\mathbb{V}\Big(\sqrt{k}\Big(\widehat{\gamma}^{(c,H)}_{T,k,n} - \gamma_X\Big)\Big) = \frac{\gamma_X^3}{\gamma}$$

2.8 Estimations de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier $\widehat{\gamma}^{KM}_{X,Hill}$

Julien Worms et Rym Worms (2013,[58]) ont proposé deux nouvelles approches pour l'estimation de l'indice des valeurs extrêmes dans le cadre de la censure aléatoire (à droite) des échantillons, sur la base des idées d'intégration de Kaplan-Meier.

Ces idées sont développées dans le cas des distributions à queue lourde, et pour l'adaptation de l'estimateur de Hill, dont la consistance est prouvée sous la conditions du premier ordre. Le premier point de la nouvelle approche de départ est le résultat bien connu suivant, qui est la base des méthodes de régression censurés : Si ϕ est une fonction réel positif,

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

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

$$\begin{aligned}
\mathbb{E}\Big[\tfrac{\delta}{1-G(T)}\phi(T)\Big] &= \int \mathbb{1}_{x<y}\frac{\delta}{1-G(x)}dF(x)dG(y) \\
&= \int \phi(x)(1-G(x))^{-1}\Big(\int \mathbb{1}_{y>x}dG(y)\Big)dF(x) \\
&= \int \phi(x)dF(x) \\
&= E[\phi(X)]
\end{aligned}$$

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

$$F \in DA(\mathcal{H}_{\gamma_x}), G \in DA(\mathcal{H}_{\gamma_C}) \quad \gamma_X > 0 \ et \ \gamma_C > 0, \tag{2.12}$$

qui, comme indiqué plus haut, implique que $H \in DA(\mathcal{H}_\gamma)$ avec

$$\gamma = \frac{\gamma_X\gamma_C}{\gamma_X + \gamma_C}$$

Dans ce cas, il est bien connu que (voir remarque 1.2.3 dans Haan et Ferreira [38])

$$\lim_{u\to\infty} \mathbb{E}[\log\Big(\frac{X}{u}\Big)|X > u] = \gamma_X \tag{2.13}$$

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

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

Alors nous pouvons définir une version aléatoire de ϕ

$$\phi(x) = (\mathbb{P}(X > u))^{-1} \log(x/u) \mathbb{1}_{\{x>u\}}$$

avec un seuil aléatoire $u = T_{n-k:n}$

$$\widehat{\phi}_n(x) := \frac{1}{1 - \widehat{F}(T_{n-k:n})} \log\Big(\frac{x}{T_{n-k:n}}\Big) \mathbb{1}_{\{x>T_{n-k:n}\}} \tag{2.15}$$

Par conséquent, en combinant (2.11) et (2.13) avec cette fonction $\widehat{\phi}_n$,

$$\int \widehat{\phi}_n d\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} w_{i:n} \widehat{\phi}_n(T_i) \ \ ou \ \ w_{i:n} = \frac{\delta_i}{1 - \widehat{G}_n(T_i)}$$

L'adaptation premiére de l'estimateur de Hill est valable dans le cas (2.12),

$$\widehat{\gamma}^{KM}_{X,Hill} := \frac{1}{n(1 - \widehat{F}_n(T_{n-k:n}))} \sum_{i=1}^{k_n} \frac{\delta_{n-i+1:n}}{1 - \widehat{G}_n(T^-_{n-i+1:n})} \log\Big(\frac{T_{n-i+1:n}}{T_{n-k:n}}\Big) \tag{2.16}$$

où $\widehat{F}_n$ et $\widehat{G}_n$ représentent les estimations de Kaplan-Meier de F et G , respectivement . Notez que nous prenons $\widehat{G}_n(T^-_{n-i+1:n})$ au lieu de $\widehat{G}_n(T_{n-i+1:n})$, dans la défnition de $\widehat{\gamma}^{KM}_{X,Hill}$, parce que $1 - \widehat{G}_n(T_{n:n})$ peut étre nul. Le théoréme suivant fournit la consistance de cet estimateur.

A cet effet, il faut deux hypothéses supplémentaires sur le comportement de la fonction $poH^{\leftarrow} = p(H^{\leftarrow}(.))$, qui sont similaires à celles utilisées dans Einmahl et al[25]. :

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

$$\frac{1}{\sqrt{k_n}} \sum_{i=1}^{k} \Big[p(H^{\leftarrow}(i - \frac{i}{n})) - p \Big] \xrightarrow{\mathbb{P}} c \in \mathbb{R} \tag{2.17}$$

$$\sqrt{k_n} \sup |p(H^{\leftarrow}(t)) - p(H^{\leftarrow}(s))| \to 0 \ \ C > 0 \tag{2.18}$$

où $C_n = \{(s,t)$ tel que $s < 1 \, , 1 - k_n/n \leq u < 1 \ \ , \mid u - s \mid \leq C\sqrt{k_n}/n \ \}$

Théorème 2.8.1. *Sous les hypothéses (2.12) , (2.14) , (2.17), (2.18) si l'on suppose en outre que, pour $\delta > 0$,*

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

et que $\gamma_X < \gamma_C$*,puis,loreque* n *tend vers* $+\infty$

$$\widehat{\gamma}_{X,Hill}^{KM} \xrightarrow{\mathbb{P}} \gamma_X$$

Remarque 2.8.1. Ce théoréme n'a été prouvé que pour $\gamma_X < \gamma_C$ (une condition utilisée à plusieurs reprises dans la preuve, voir remarque suivante), qui peut être interprété comme une légére censure dans la queue (finalement, pas plus de 50% des observations dans la queue sont censurées). en fait, si nous utilisons notre estimateur dans le cas de censure forte ($\gamma_X \geq \gamma_C$), les simulations semblent montrer que la performance (en termes de MSE) est, étonnamment, pire que celle de l'affaire $\gamma_X < \gamma_C$ (voir chapitre 3)

Remarque 2.8.2. La condition $\gamma_X < \gamma_C$ provient essentiellement du fait que l'etimateur $\widehat{\gamma}_{X,Hill}^{KM}$ est convergent vers la puissance α d'une lois i.i.d. standard de Pareto, où α est proche de γ/γ_C , cet exposant $\gamma/\gamma_C = \frac{\gamma_X}{\gamma_X+\gamma_C}$ est toujours plus petit que 1, mais (pour les conditions de moment) nous avons été amenés à supposer qu'il soit en fait plus petit que 1/2, c-à-d. $\gamma_X < \gamma_C$.

2.9 Conclusion

Dans ce chapitre on a introduit les données incomplètes et puis on est passer au problème de l'estimation de l'indice de queue en présence de données censurées aléatoirement à droite, et nous nous sommes interesser a quelques estimateurs de l'indice des valeurs extrêmes en présence de censure aléatoire à droite qu'on a juger imporatant dans notre études. Il existe cependant plusieurs autre estimateurs comme (Dekkers(moments), Pickands, Rapport des moments, Peng et W-estimateur, noyau ...) Nous nous focalisons principalement sur les estimateurs suivants : l'estimateur de Hill classique de l'indice des valeurs extrêmes Einmahl et al (2008,[25]), et l'estimateur de l'indice des valeurs extrêmes d'intégration de Kaplan-Meier Worms ,J., Worms, R., (2013,[58]) .

Chapitre 3

Comparaison des estimateurs par simulation

3.1 Principe du Bootstrap

Les techniques de "ré-échantillonnage", appeleés aussi en anglais methode du "bootstrap" a été proposée par Bradley Efron (1979,[27]), cette méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles.

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

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

3.1.1 Choix du B, le nombre de Bootstraps

L'importance des valeurs extrêmes de la statistique γ étudiée est un facteur important dans la détermination du choix de B(le nombre de réplications), plus ces valeurs sont fréquentes,

plus B devrait très grand. On notera cependant que certaines autres applications du bootstrap exigent un B beaucoup plus grand (ce sera en particulier le cas pour l'application à la construction d'intervalles de confiance). Selon B.Efron [23], il est rarement nécessaire d'utiliser plus de $B = 200$ échantillons bootstrap pour estimer une variance, dans bien des cas, $B = 50$ ou 100 sont suffisants .

3.1.2 Application de bootstrap sur des données censurées

Le bootstrap est bien validée par de nombreuses études statistiques et une des premiéres applications du bootstrap a été faite dans le contexte d'analyse de la survie (Efron, 1981) pour répondre à certaines questions notamment pour construire les bandes de confiance ([9],[27]).

3.2 Choix du nombre des valeurs extrêmes optimal k_n

Les résultats concernant les estimateurs de l'indice des valeurs extrêmes sont asymptotiques, ils sont obtenus lorsque $k_n \to \infty$ et $k_n/n \to \infty$. La difficulté en pratique consiste à choisir le nombre d'extrêmes k_n utilisé dans les estimations. L'issue est importante : l'extrême volatilité du graphe $\{(k_n, \widehat{\gamma}_{k_n}) \ , \ 1 \leq k_n < n\}$. Où $\widehat{\gamma}_{k_n}$ représente n'importe quel estimateur introduit précédement, rend difficile l'utilisation de l'estimateur en pratique si aucune indication sur le choix de k_n n'est donnée. Des travaux ont montré qu'en utilisant trop d'observations, dans la procédure d'estimation de γ, on observe un biais substantiel tandis que l'utilisation de peu d'observations conduit à une variance considérable. Ce probème a été longuement abordé dans la littérature, voir par exemple Reiss et Thomas (1997,[48]), de Haan et Peng (1998,[39]), Drees et Kaufmann (1998,[22]), Danielsson et al. (2001,[16]), Cheng et Peng (2001,[12]) Beirlant et al. (2002,[16]),Beirlant et al. (2006,[4]), etc. Nous allons utiliser dans nos simulations les méthodes suivantes : Méthode Cheng et Peng et l'erreur quadratique moyenne et Reiss et Thomas pour déterminer la valeur optimale de k_n correspendante à l'estimateur $\widehat{\gamma}_X^{(Hill,c)}$.

3.2.1 Méthode de Cheng et Peng

La valeur optimale de k_n peut être obtenue par la méthode de Cheng et Peng. La valeur optimale de k_n est donnée par :

$$k_n^{opt} := \begin{cases} \left(\frac{1+2z_\alpha^2}{3\widehat{\delta}(1+2\widehat{\rho})}\right)^{1/(1+\widehat{\rho})} n^{\widehat{\rho}/(1+\widehat{\rho})}, & \text{si } \widehat{\rho} > 0, \\ \left(\frac{1+2z_\alpha^2}{3\widehat{\delta}}\right)^{1/(1+\widehat{\rho})} n^{\widehat{\rho}/(1+\widehat{\rho})}, & \text{si } \widehat{\rho} < 0. \end{cases} \tag{3.1}$$

Où z_α est le quantile de la distribution standard

$$\widehat{\rho} := -\log\left(\left|\frac{M_n^{(2)}(n/2\sqrt{\log n}) - 2\{\widehat{\gamma}_X^{(c.Hill)}(n/2\sqrt{\log n})\}^2}{M_n^{(2)}(n/\sqrt{\log n}) - 2\{\widehat{\gamma}_X^{(c.Hill)}(n/\sqrt{\log n})\}}\right|\right)/\log 2$$

$$\widehat{\delta} := (1+\widehat{\rho})(\log n)^{\widehat{\rho}/2}\frac{M_n^{(2)}(n/\sqrt{\log n}) - 2\{\widehat{\gamma}_X^{(c.Hill)}(n/\sqrt{\log n})\}}{2\widehat{\rho}\{\widehat{\gamma}_X^{(c.Hill)}(n/\sqrt{\log n})\}^2}$$

avec

$$M_n^{(2)} = \frac{\frac{1}{k}\sum_{i=1}^{k}\Big(\log(T_{n-i+1:n}) - \log(T_{n-k:n})\Big)^2}{\widehat{p}}$$

$$\widehat{\gamma}_X^{(c.Hill)} = \frac{\widehat{\gamma}^{(Hill)}}{\widehat{p}}$$

3.2.2 Méthode basée sur l'erreur quadratique moyenne

La valeur optimale de k_n peut être obtenue par la minimisation de l'erreur quadratique moyenne de l'estimateur γ_X . On peut se baser sur des méthodes de Bootstrap pour calculer (MSE).

Pour toute réplication R nous estimons γ_X et soit $\widehat{\gamma}_{k_n}^{(Hill,c),j}$ l'estimateur de γ_X obtenu à la j-ieme réplication $(j = 1, ..., R)$ avec $(k = 1, ..., n-1)$. Il semble donc naturel de trouver une valeur k_n^{opt} qui minimise les valeurs de l'erreur quadratique moyenne $\{(k_n, MSE(k_n), k_n = 1, ..., n-1)\}$ par rapport à k_n, La valeur optimale de k_n est donnée par

$$k_n^{opt} := arg\min_{1\le k_n\le n-1}\left\{\frac{1}{R}\sum_{j=1}^{R}(\gamma_{k_n}^{(Hill,c),j} - \gamma)^2\right\} \tag{3.2}$$

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

entre le biais et la variance. Il semble raisonnable qu'une minimisation du MSE permet de trouver une valeur intermédiaire entre les composantes du biais et de la variance pour ce compromis.

3.2.3 Méthode de Reiss et Thomas

Reiss et Thomas (1997) ont proposé une méthode heuristique de choisir le nombre des extrêmes pour utiliser dans l'estimation de l'indice de queue dans le cas censuré . Reiss et Thomas ont basé leur approche de choisir le nombre adéquat de plus grandes observations sur un moyen de minimiser la distance résumant un terme de pénalité. Dans un certain sens, ce coefficient est prévu pour être plus sévère en ce qui concerne des estimations de γ_x avec l'origine dans les observations prises plus loin de la queue réelle. Ils proposent une manière automatique de choisir k_n^{opt} en minimisant :

$$\frac{1}{k_n}\sum_{i=1}^{k_n} i^{\beta}\left|\widehat{\gamma}^{c,Hill}(i) - mod(\widehat{\gamma}^{c,Hill}(1),.....,\widehat{\gamma}^{c,Hill}(k_n))\right| \longrightarrow \min,\ \ 0 \leq \beta \leq \frac{1}{2}$$

3.3 Bootstrap des l'estimateurs

Soit $X_1, ..., X_n$ n variables représentant les durées de vie de n sujets, sont des variables aléatoires positives, indépendantes et de fonction de répartition F , et indépendamment d'elles, et $C_1, ..., C_n$ les instants de censures associés, positives, de fonction de répartition G : On note $\{(T_1, \delta_1), ..., (T_n, \delta_n)\}$ l'échantillon réellement observé et pour $i \leq n$

$$T_i = X_i \wedge C_i \ \ et \ \ \delta_i = \mathbb{I}_{\{X_i \leq C_i\}},$$

avec H la fonction de distribution de échantillons T. Soit $\{(T_{1:n}, \delta_{1:n}), ..., (T_{n:n}, \delta_{n:n})\}$ l'échantillon ordonné suivant les valeurs de T_i. Efron (1981) suggére le plan du ré-échantillonnage suivant : On génére un échantillon bootstrapé

$$(T_1^*, \delta_1^*), ..., (T_n^* \delta_n^*)$$

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

$$(T_1, \delta_1), ..., (T_n, \delta_n)$$

et soit $(T^*_{i:n}, d^*_{i:n})_{i=1,...,n}$,l'échantillon ordonné suivant les valeurs de T^*_i :

Si F et G sont supposées être dans le domaine d'attraction maximales tel que :
$F \in DA(\mathcal{H}_{\gamma_X})$, $G \in DA(\mathcal{H}_{\gamma_C})$, $\gamma_X > 0$,$\gamma_C > 0$, comme indiqué plus haut,implique que

$$H \in DA(\mathcal{H}_\gamma) \ avec \ \gamma = \gamma_X\gamma_C/(\gamma_X + \gamma_C)$$

L'estimateur de Hill bootstrapé de l'indice de valeurs extrêmes $\widehat{\gamma}^{*c}_X$ et l'estimateur de Kaplan-Meier bootstrapé $\widehat{\gamma}^{*KM}_{X,Hill}$ construit avec les données $(T^*_{i:n}, \delta^*_{i:n})_{i=1,...,n}$ s'écrit :

$$\widehat{\gamma}^{*c}_X = \frac{\frac{1}{k}\sum_{i=1}^{k} \log T^*_{n-i+1:n} - \log T^*_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k} \delta^*_{[n-i+1:n]}} \tag{3.3}$$

et

$$\widehat{\gamma}^{*KM}_{X,Hill} := \frac{1}{n(1-\widehat{F}_n(T^*_{n-k:n}))} \sum_{i=1}^{k_n} \frac{\delta^*_{n-i+1:n}}{1-\widehat{G}_n(T^{*-}_{n-i+1:n})} \log\left(\frac{T^*_{n-i+1:n}}{T^*_{n-k:n}}\right) \tag{3.4}$$

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

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

$$T = \{(T_1, \delta_1), ..., (T_n, \delta_n)\}$$

Chaque échantillion

$$T^{*b} = \{(T^{*b}_1, \delta^{*b}_1), ..., (T^{*b}_n, \delta^{*b}_n)\}$$

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

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

On obtient alors un échantillon de B valeurs

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

3.4.1 L'erreur standard de estimation Bootstrap

Définition 3.4.1. On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs $\widehat{\gamma}_X^{*c}(b)$, la moyenne est :

$$\widehat{\gamma}_X^{*c}(.) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_X^{*c}(b) \tag{3.5}$$

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

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

où B est le nombre total d'échantillons bootstrap

Algorithme 3.4.1. estimation bootstrap de l'erreur standard

Variable

B :entier assez grand

Début

pour $b = 1$ à B ***faire***

on calcule l'estimateur de queue :

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

On obtient alors un échantillon de B valeurs

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

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

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

avec $\widehat{\gamma}_X^{*c}(.) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_X^{*c}(b)$

fin pour

Fin.

3.4.2 Biais de estimation bootstrap

Le biais d'un estimateur s'exprime comme

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

Définition 3.4.2. On appelle estimateur bootstrap du biais, l'estimateur de l'indice de queue pour les données observées

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

Algorithme 3.4.2. Estimation bootstrap du biais **Variable**

B : entier assez grand.

Début

Pour $b = 1$ à B **faire**

Générer $T^{*b} = \{(T_1^{*b}, \delta_1^{*b}), ..., (T_n^{*b}, \delta_n^{*b})\}$

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

Calculer $\widehat{\gamma}_X^{*c}(b) = \#\{j, (T_i^{*b}, \delta_i^{*b}) = (T_i, \delta_i)\}/n \quad$ pour tout i.

Fin pour Calculer $\widehat{\gamma}_X^{*c}(.) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_X^{*c}(b) \quad$ pour tout i.

Retourner

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

Fin.

3.4.3 L'erreur quadratique moyenne de estimation Bootstrap

Définition 3.4.3.

Monde classique : l'erreur quadratique moyenne (MSE) de $\widehat{\gamma}_X^c$ est égale à

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

Monde bootstrap : l'estimateur bootstrap de l'erreur quadratique moyenne de $\widehat{\gamma}_X^{(c,Hill)}$ est défini par :

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

Algorithme 3.4.3. estimation bootstrap de la MSE

Variable

B : entier assez grand

Début

Pour b variant de 1 à B

Générer T^{*b} réalisation d'un échantillon bootstrap

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

Fin Pour

Retourner

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

Fin

3.4.4 Estimation des Intervalles de confiance

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

Algorithme 3.4.4. Estimation Bootstrap de l'intervalle de confiance

Variable

B : entier assez grand.

Début

Pour $b = 1$ à B **faire**

Générer $T^{*b} = \{(T_1^{*b}, \delta_1^{*b}), ..., (T_n^{*b}, \delta_n^{*b})\}$ réalisation d'un echantillon bootstrap,

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

Fin pour

Retourner

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

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

Fin.

3.5 Propriétés de l'estimateur de l'indice de queue $\widehat{\gamma}^{*KM}_{X,Hill}$

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

$$T = \{(T_1, \delta_1), ..., (T_n, \delta_n)\}$$

Chaque échantillion

$$T^{*b} = \{(T_1^{*b}, \delta_1^{*b}), ..., (T_n^{*b}, \delta_n^{*b})\}$$

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

$$\widehat{\gamma}^{*KM}_{X,Hill} \quad , \quad b = 1, ..., B,$$

3.5.1 L'erreur standard de estimation Bootstrap

Définition 3.5.1. On définira maintenant la moyenne bootstrap. Pour un ensemble d'estimateurs $\widehat{\gamma}^{*KM}_{X,Hill}(b)$,la moyenne est :

$$\widehat{\gamma}^{*KM}_{X,Hill} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*KM}_{X,Hill}(b) \tag{3.7}$$

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

$$\widehat{se} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\gamma}^{*KM}_{X,Hill}(b) - \widehat{\gamma}^{*c}_{X}(.)\right)^2} \tag{3.8}$$

où B est le nombre total d'échantillons bootstrap

Algorithme 3.5.1. estimation bootstrap de l'erreur standard

Variable

B :entier assez grand

Début

pour $b = 1$ à B ***faire***

on calcule l'estimateur de queue :

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

On obtient alors un échantillon de B valeurs

$$\widehat{\gamma}^{*KM}_{X,Hill}(1), \widehat{\gamma}^{*KM}_{X,Hill}(2), ..., \widehat{\gamma}^{*c}_{X}(B)\}$$

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

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

avec $\widehat{\gamma}^{*KM}_{X,Hill}(.) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}^{*c}_{X}(b)$

fin pour

Fin.

3.5.2 Biais de estimation bootstrap

Le biais d'un estimateur s'exprime comme

$$\widehat{Biais}(\widehat{\gamma}^{*KM}_{X,Hill}) = \mathbb{E}_{\widehat{F}}[\widehat{\gamma}^{*KM}_{X,Hill}] - \widehat{\gamma}^{*KM}_{X,Hill}.$$

Définition 3.5.2. On appelle estimateur bootstrap du biais, l'estimateur de l'indice de queue pour les données observées

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

Algorithme 3.5.2. Estimation bootstrap du biais **Variable**

B : entier assez grand.

Début

Pour $b = 1$ à B **faire**

Générer $T^{*b} = \{(T^{*b}_1, \delta^{*b}_1), ..., (T^{*b}_n, \delta^{*b}_n)\}$

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

Calculer $\widehat{\gamma}^{*KM}_{X,Hill} = \#\{j, (T^{*b}_i, \delta^{*b}_i) = (T_i, \delta_i)\}/n$ pour tout i.

Fin pour Calculer

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

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

Fin.

3.5.3 L'erreur quadratique moyenne de estimation Bootstrap

Définition 3.5.3.

Monde classique : l'erreur quadratique moyenne (MSE) de $\widehat{\gamma}^{*KM}_{X,Hill}$ est égale à

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

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

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

Algorithme 3.5.3. estimation bootstrap de la MSE

Variable

B : entier assez grand

Début

Pour b variant de 1 à B

Générer T^{*b} réalisation d'un échantillon bootstrap

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

Fin Pour

Retourner

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

Fin

3.5.4 Estimation des Intervalles de confiance

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

Algorithme 3.5.4. Estimation Bootstrap de l'intervalle de confiance

Variable

B : entier assez grand.

Début

Pour $b = 1$ à B **faire**

Générer $T^{*b} = \{(T_1^{*b}, \delta_1^{*b}), ..., (T_n^{*b}, \delta_n^{*b})\}$ réalisation d'un echantillon bootstrap,

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

Fin pour

Retourner

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

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

Fin.

3.6 Simulations

3.6.1 Echantillon initial et paramétres de simulations

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

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

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

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

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

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

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

$$Ti = X_i \wedge C_i$$

les indicateurs de censure sont

$$\delta_i = \mathbb{I}\{X_i \leq C_i\}$$

3.7 Résultats des simulations

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

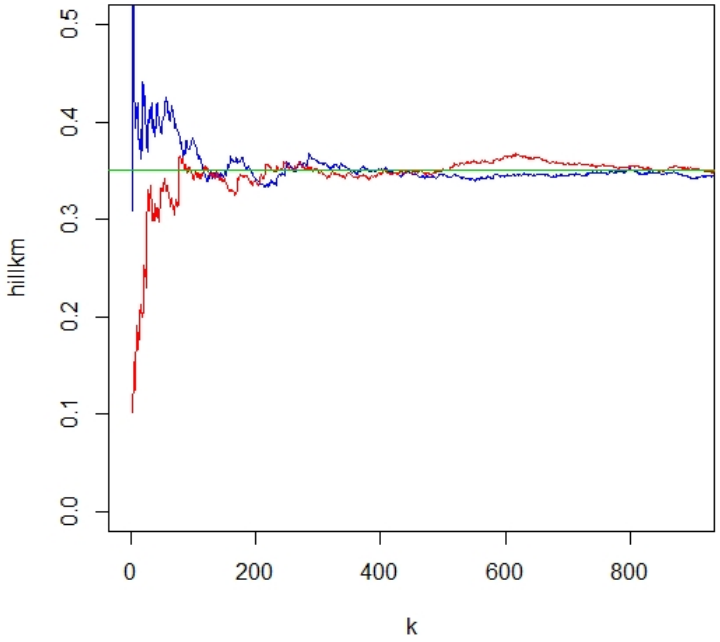

FIGURE 3.1 – Comportement graphique de $\widehat{\gamma}_X^{c.Hill}$ (en rouge) et $\widehat{\gamma}_{X,Hill}^{KM}$ (en bleu) vs k issue de la distribution de Pareto ($\widehat{\gamma}_X = 0.35$) censurées par Pareto ($\gamma_C = 2.5$) , (10% de censure)

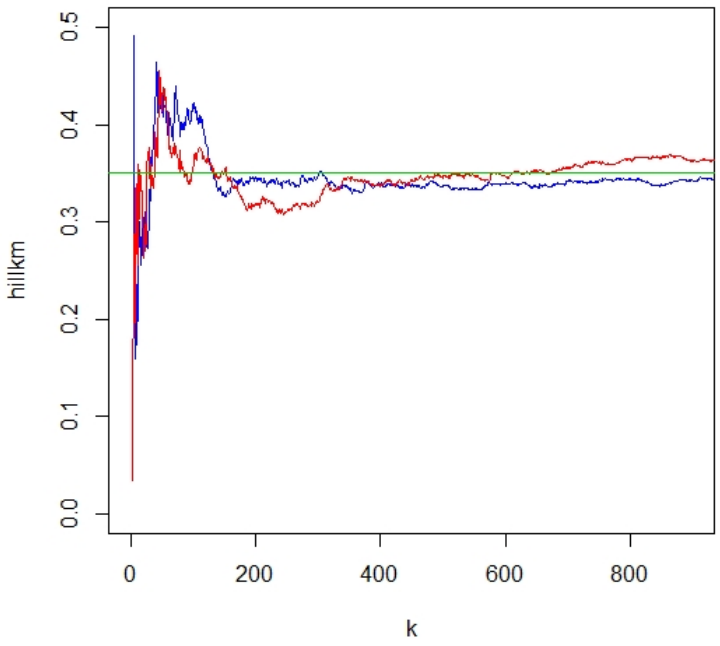

FIGURE 3.2 – Comportement graphique de $\widehat{\gamma}_X^{c,Hill}$ (en rouge) et $\widehat{\gamma}_{X,Hill}^{KM}$ (en bleu) vs k issue de la distribution Pareto ($\widehat{\gamma}_X = 0.35$) censur´ees par Pareto ($\gamma_C = 0.5$) , (40% de censure)

Pour un pourcentage réduit de censure (10%) : on voit que l'estimateur $\widehat{\gamma}_X^{c,Hill}$ et meilleur que $\widehat{\gamma}_{X,Hill}^{KM}$ jusqu'a ($k \leq \frac{n}{2}$) et inversement (Figure 3.1), pour un pourcentage grand de censure (40%) on doit que l'estimateur $\widehat{\gamma}_{X,Hill}^{KM}$ et mailleur que $\widehat{\gamma}_X^{c,Hill}$ sue jus qu'a ($k \leq \frac{n}{2}$) et inversement (Figure 3.2), en général les deux estimateurs sont stables à partir de ($k \geq \frac{n}{2}$) Les résultats de notre simulation bootstrap sont illustrés dans le tableau 3.1 et 3.2 Nous avons simuler deux échantillons de taille $n = 1000$ de pareto ($\gamma_X = 0.35$), censuré par un échantillon de taille $n = 1000$ de pareto. pour différentes valeurs du paramètre ($\gamma_C = 0.5, 2.5$). Le caractére c représente le pourcentage de censure (10%,40%). Les remarques qu'on peut tirées sont :

- A chaque fois le pourcentage de censure augmente, le k optimal décroit.
- L'erreur quadratique moyenne (MSE) de l'indice des valeurs extrèmes avec l'estimateur de l'integralité KM $\widehat{\gamma}_{X_{boot},Hill}^{KM}$ est meilleur que celui de Hill classique $\widehat{\gamma}_{X_{boot}}^{c,Hill}$ pour un pourcentage de censure égale à 10%. Contrairement au pourcentage de censure 40%, l'etimateur Hill classique $\widehat{\gamma}_{X_{boot}}^{c,Hill}$ est meilleur que l'estimateur de l'integralité KM $\widehat{\gamma}_{X_{boot},Hill}^{KM}$.
- L'erreur quadratique moyenne (MSE) de l'indice des valeurs extrèmes augmente avec l'augmention du pourcentage de censure pour les deux estimateurs $\widehat{\gamma}_{X_{boot}}^{c,Hill}$ et $\widehat{\gamma}_{X_{boot},Hill}^{KM}$.

	$c = 10\%$	$c = 40\%$
k^{opt}	675	646
$\widehat{\gamma}_X^{c.Hill}$	0.3578585	0.3527234
$\widehat{\gamma}_{X_{boot}}^{c}$	0.3636489	0.3575239
sd	0.01304067	0.01384914
$biais$	0.005790359	0.004800579
$EMQ(MSE)$	0.0002030204	0.0002142049
IC	$icinf$ $icsup$	$icinf$ $icsup$
	0.338413 0.3889364	0.329958 0.3844768

TABLE 3.1 – Les résultats de l'estimateur de l'indice de queue $\widehat{\gamma}_X^{c.Hill}$ par simulation bootstrap.

	$c = 10\%$	$c = 40\%$
k^{opt}	792	768
$\widehat{\gamma}_{X,Hill}^{KM}$	0.3481659	0.3433215
$\widehat{\gamma}_{X_{boot}}^{KM}$	0.3496989	0.355045
sd	0.01280803	0.01163419
$biais$	0.001533028	0.01172349
$EMQ(MSE)$	0.0001658489	0.0002723434
IC	$icinf$ $icsup$	$icinf$ $icsup$
	0.3250906 0.3733657	0.3326816 0.3770174

TABLE 3.2 – Les résultats de l'estimateur de l'indice des valeurs extrêmes par l'intégration de Kaplan-Meier $\widehat{\gamma}_{X,Hill}^{KM}$ par simulation bootstrap.

Simulation bootstrap de l'estimateur $\widehat{\gamma}_X^{c.Hill}et\widehat{\gamma}_{X,Hill}^{KM}$ vs n

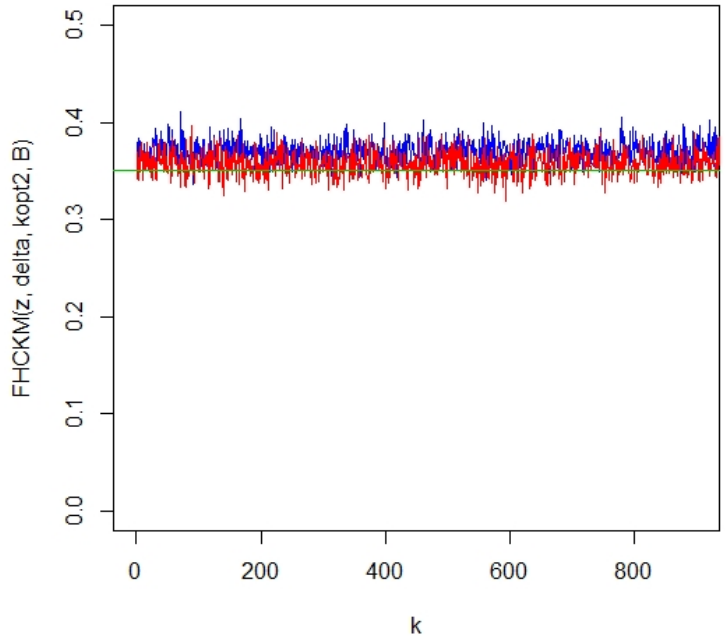

FIGURE 3.3 – $\widehat{\gamma}_X^{c.Hill}$ et $\widehat{\gamma}_{X,Hill}^{KM}$ bootstrap de 1000 répétitions de Pareto ($\widehat{\gamma}_X = 0.35$) censurée par Pareto ($\widehat{\gamma}_C = 1$) 25% de censure

3.8 Conclusion

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

Conclusion générale

Ce mémoire constitue une sorte de mariage entre deux branches de la statistique : la théorie des valeurs extrêmes et l'analyse de survie . L'intérêt principal est d'étendre les résultats de la théorie des valeurs extrêmes au cas où les données sont censurées. Nous nous concentrons sur le domaine d'attraction de Fréchet dans le cas des distributions à queues lourdes lorsque les données sont incomplètes. Nous nous intéressons à effectuer une comparaison de l'indice des valeurs extrêmes dans le cas des données censurées à droite de l'indice de queue $\gamma_X^{c,Hill}$ et celui de l'intégration de Kaplan-Meier $\gamma_{X,Hill}^{KM}$.

Notre travaille se décompose en trois chapitres, dans le premier chapitre est une sorte de rappel des concepts de base et des résultats essentiels de la théorie des valeurs extrêmes. Parmi ces résultats, on a décrit les limites possibles de la loi du maximum d'un échantillon. Ces lois sont indexées par un paramètre appelé indice des valeurs extrêmes ou indice de queue γ. Dans le deuxième Chapitre nous avons faits un rappel sur les domaines où l'on rencontre les données incomplètes (censurées-tronquées) avec une attention particulière sur les données censurées. Pour faciliter la lecture du document, on a rappelé quelques notions fondamentales de l'analyse de survie. Nous nous focalisons essentiellement sur les estimateurs suivants : l'estimateur de Hill de l'indice des valeurs extrêmes , et l'estimateur l'indice des valeurs extrêmes l'intégration de Kaplan-Meier . Dans le dernier Chapitre nous avons utilisé la méthode ré-échantillonnage (Bootstrap) sur les deux estimations $\gamma_X^{c,Hill}$ et $\gamma_{X,Hill}^{KM}$ dons le domaine d'attraction de Fréchet pour les données aléatoires censurées a droite dans l'objectif d'observer le comportement de chacun des deux estimateurs .

Bibliographie

[1] J. N. Al-Abbasi and K. J. Fahmi. Estimating maximum magnitude earthquakes in iraq using extreme value statistics. *Geophysical journal of the Royal Astronomical Society*, (82) :535–548, 1985.

[2] A. A. Aldama and A. I. Ramirez. Dam design flood estimation based on bivariate extreme-value distributions. in the extremes of the extremes : extraordinary floods. *Wallingford, IAHS Press. 1*, page 257–262, 2002.

[3] E. Alvarado, D. V. Sandberg, and S. G. Pickford. Modeling large forest fires as extreme events. *Northwest Science*, (1) :66–75, 1998.

[4] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. Statistics of extremes : Theory and applications. *J. Wiley Sons*, 2006.

[5] J. Beirlant, A. Guillou, and G. Dierckx A. Fils-Villetard. Estimation of the extreme value index and extreme quantiles under random censoring. *Springer Science+Business Media*, 2007.

[6] J. Beirlant, P. Vjnckier, and J. L. Teugels. Tail index estimation, pareto quantile plots, and regression diagnostics. *J. Amer Statist. Assoc.*, 1996.

[7] A. Bengtsson and C. Nilsson. Extreme value modelling of storm damage in swedish forests. *Nat. Hazards Earth Syst. Sci*, pages 515–521, 2007.

[8] N. Bingham, C. Goldie, and J. Teugels. Regular variation. *Cambridge University Press*, page 12, 1987.

[9] A. Borchani. Statistiques des valeurs extrmes dans le cas de lois discretes. *Centre de recherche de l'ESSEC ISSN*, 2010.

[10] J. P. Bouchaud and M. Potters. Theory of financial risk and derivative pricing : from statistical physics to risk management. *Cambridge University Press, Cambridge,and edition*, 2003.

[11] N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist*, pages 437 – 453, 1974.

[12] S. Cheng and L. Peng. Con.dence intervals for the tail index. *Bernoulli*, 2001.

[13] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An Introduction to Statistical Modeling of Extreme Values*. Lecture Notes in Control and Information Sciences. Springer, 2001 (Cité en page 12.).

[14] S. Csörgö and D. Mason. Central limit theorems for sums of extreme values. *Mathematical Proceedings of the Cambridge Philosophical Society*, (1985).

[15] H. E. Daniels. The statistical theory of the strength of bundles and threads. *Proc.Royal Soc*, page 405–435, 1945.

[16] J. Danielsson, L. Haan, and L Peng C. G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 2001.

[17] A. L. M. Dekkers L. de Haan. On the estimation of the extreme value index and large quantile estimation. *The Annals of Statistics*, 1989.

[18] P. Deheuvels and J. H. Einmahl. Functional limit laws for the in- crements of kaplan-meier product-limit processes and applications. *Ann.Probab*, 2000.

[19] P. Deheuvels, E. Haeusler, and D. Mason. Almost sure convergence of the hill estimator. *Mathematical Proceedings of the Cambridge Philosophical Society*, (1988).

[20] A. Dekkers, J. Einmalh, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, (1989).

[21] G. Dekkers and L. De Haan. On the estimation of the extremevalue index and large quantile estimation. *Ann. Statist*, (1989).

[22] H. Drees. Pickands estimators of the extreme value index. *Ann. Statist*, (1995).

[23] B. Efron. Bootstrap methods : Another look at the jackknife. *Ann. Statist*, 1979.

[24] J. F. Eichner, J. W. Kantelhardt, and A. Bunde S. Havlin. Extreme value statistics in records with long-term persistence. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, pages 016–130, 2006.

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

[26] P. Embrechts, C. Kluppelberg, and T. Mikosch. Modelling extremal events for insurance and finance. *Springer-Verlag. Berlin*, (1997).

[27] B. Efron end R. Bradley. An introduction to the bootstrap. *Tibshirani*, 1993.

[28] B. De Finetti. la loi de probabilité des extrêmes. *Metron .Rome*, (1932).

[29] R. Fisher and L. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, (1928).

[30] M. Frechet. La loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, (1927).

[31] J. Geluk and L. De Haan. Regular variation, extensions and tauberian theorems. centrum voor wiskunde en informatica. *Amsterdam*, (1987).

[32] R. Gill. Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, pages 49–58, 1983.

[33] R. D. Gill. Censoring and stochastic integrals,"mathematical centre tracts. *Mathematisch Centrum, Amsterdam*, 1980.

[34] R. D. Gill. Glivenko-cantelli for kaplan-meier. math. *Methods Statist*, 1994.

[35] B. Gnedenko. la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math*, (1943).

[36] R. Gordon. Values of mill's ratio area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 1941.

[37] E. Gumbel. Statistics of extremes. columbia university press. *New York*, 1958.

[38] L. Haan and A. Ferreira. *Extreme Value Theory : An Introduction.* Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[39] L. Haan and L. Peng. Camparison of tail index estimators. *StatisticaNeerlandica*, 1998.

[40] E. Haeusler and J. Teugels. On asymptotic normality of hill's estimator for the exponent of regular variation. *The Annals of Statistics*, (1985).

[41] B. Hill. A simple general approach to inference about the tail of a distribution.the annals of statistics. *The Annals of Statistics*, page 1163–1174, (1975).

[42] J. Pickands III. Statistical inference using extreme order statistics. *Ann. Statist*, (1975).

[43] A. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *The Quarterly Journal of the Royal Meteorological Society*, (1955).

[44] S. F. Jonas. Méthodes de comparaisons de deux ou plusieurs groupes de données censurées par intervalle. avec application en immunologie clinique. *Méthodes et statistiques. Université Paris Saclay*, 2018.

[45] E. L. Kaplan and P. Meier. Non parametric estimation from incomplete observations. *J. Amer. Statist. Assoc*, (1958).

[46] B. H. Lavenda and E. Cipollone. Extreme value statistics and thermodynamics of earthquakes. *Annali di Geofisica*, (43) :469–484, 2001.

[47] D. Mason. Laws of large numbers for sums of extreme values. *The Annals of Probability.*

[48] R. D. Reiss and M.Thomas. Statistical analysis of extreme values with applications to insurance. *Finance, Hydrology and Other Fields. Birkhäuser, Basel*, (2007).

[49] S. Resnick. *Extreme values regular variation and point processes.* Springer Verlag, New-York, (1987).

[50] J. Segers. Residual estimators. *J. Statist. Plann. Inference*, 1999.

[51] G. R. Shorack and J. A. Wellner. Empirical processes with applications to statistics. *John Wiley and Sons*, 1986.

[52] L. Soltane. Analyse des valeurs extrêmes en présence de censure. *These Doctorat,université. M. Khider Biskra, Algeria*, 2016.

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

[54] B. W. Turnbull. Non parametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc*, (1974).

[55] R. Von-Mises. La distribution de la plus grande de n valeurs. *Revue de Mathématique Union Interbalcanique*, (1936).

[56] R. Von-Mises. La distribution de la plus grande de n valeurs. *Amer. Math. Soc*, 1954.

[57] J. G. Wang. A note on the uniform consistency of the kaplan-meier estimator. *Ann. Statist.*, 1987.

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

[59] Y. ZIANE. Sur l'estimation non paramétrique de l'indice de variabilité et la distribution des densités heavy tailed. *These Doctorat,université.A.Mira*, 2016.

Résumé

Dans ce travail, nous nous somme interéssé à la théorie des valeurs extrèmes (TVE) en présence des données censurées. Pour cela, nous avons présenté les notions de base de TVE en présence des données censurées ainsi que les méthodes qui correspendent à l'estimation de l'indice de ces valeurs extrèmes (γ) (Hill, Pickands et la méthode des moments). Une étude de simulation à été réalisé pour l'estimation de (γ) par l'estimateur de Hill avec l'application de la méthode ré-échantillonage (Bootstrap) ainsi que l'integration de Kaplan-Meier

Mots clés : Théorie des valeurs extrèmes(TVE), indice des valeurs extrèmes, données censuées, méthode de Hill, Bootstrap, Intergration Kaplan Meier.

Abstract

In this work, we were interested in the extreme values theory (EVT) in the presence of censored data. To do this, we have presented the basic concepts of EVT in the presence of censored data as well as the methods that correspond to the estimation of the index of these extremes values (γ) (Hill, Pickands et la méthode des moments). A simulation study was realized for the estimation of (γ) by the Hill estimator with the application of the resampling method (Bootstrap) as well as the integration of Kaplan Meier. keywords : Extreme Values Theory(EVT), extreme value index, Censored data , Hill method, Bootstrap, Kaplan-Meier Intergration

Key words : Extreme Values Theory(EVT), extreme value index, Censored data , Hill method, Bootstrap, Kaplan-Meier Intergration