République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Mathématiques

Mémoire présenté en vue de l'obtention du

DIPLÔME DE MASTER en MATHÉMATIQES

Option : **Statistique**

Par

Kalfali Ramissa

Thème

Estimation de la moyenne d'une distribution à queue lourde sous censure aléatoire

Membres du Comité d'Examen

Dr.	Benelmir Imane	UMKB	Président
Dr.	Soltane Louiza	UMKB	Encadreur
Dr.	Touba Sonia	UMKB	Examinateur

Juin 2021

À mes chers Parents,

À ma sœur Ikrame bien-aimé,

À mes chers frères Adel, Fouzi et Mazigh,

À tous les membres de ma grande famille,

Enfin, je ne peux passer outre à ma reconnaissance envers mes très chères amies pour leurs affection et complaisance et envers tous ceux qui m'ont inculqué les bonnes valeurs et la bonne conscience.

REMERCIEMENTS

Je glorifie "Allah" le tout-puissant de m'avoir donnée courage et patience qui m'a permis d'accomplir ce travail.

Je tiens tout d'abord à exprimer toute ma reconnaissance à mon encadreur Dr. Louiza Soltane, pour sa disponibilité, s'écoute, ses conseils éclairés et ses encouragements à moi. Elle a toujours été présente pour m'inculquer sa grande rigueuse. Inestimable a été pour moi le privilégié de l'avoir pour guide. J'espère que nos relations et nos collaborations continuer longtemps après ma soutenue.

Mes remerciements s'étendent également à tous mes enseignants et employés *du département de Mathématiques durant les années de mes études, avec une mention spéciale à Messieurs les membres du jury "Dr. Benelmir Imane et Dr. Touba Sonia" qui nous ont faites l'honneur de participer à notre soutenance.*

Je remercie ma famille à qui je n'ai jamais su dire toute l'affection que j'ai pour eux, mes parents, mes frères, ma sœur et mes amies El Mahdi, Ahlam, Ikram qui ont été et seront toujours présents à mon côté, merci pour votre soutien et vos encouragements.

Kalfali Ramissa.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

La recherche autour des lois de valeurs extrêmes est particulièrement active depuis les années 1970. Nous renvoyons aux ouvrages de Beirlant et al. [2], Reiss et Thomas [29], Embrecht et al. [14] et Cláudia et al. [6] pour une approche détaillée de la théorie des lois de valeurs extrêmes, ainsi que pour des références concernant les applications de cette théorie : en hydrologie, en assurance et en finance pour les calculs de risques, en météorologie pour les événements extrêmes, ... etc.

Théorie des valeurs extrêmes (TVE) est un sujet classique de la théorie des probabilités et des statistiques mathématiques qui s'intéresse aux valeurs extrêmes des distributions de probabilité. Les extrêmes sont des évènements rares qui conduisent à des pertes importantes.

Dans beaucoup de domaines, tels la médecine, la fiabilité, l'économie et la sociologie, la notion de l'analyse de survie est extrêmement importante. En mathématiques, l'analyse de survie est assimilée à une variable aléatoire non négative dont l'étude a reçu une attention particulière de la part des statisticiens. La base de toute analyse statistique est l'échantillon à qui il arrive parfois d'être censuré. Il existe plusieurs mécanismes de censure dont la plus couramment rencontrée est la censure aléatoire à droite.

Depuis quelques années, la statistique des extrêmes sous censure aléatoire a reçu beaucoup

d'attention aussi bien sur le plan théorique que sur le plan pratique (voir par exemple Beirlant et al. [4], Einmahl et al. [13], Soltane et al. [32]).

Dans ce travail, on s'intéresse à l'estimation de la moyenne en présence de données censurées et pour les distributions à queue lourdes. Dans ce cadre, on a fait une synthèse sur les différentes théories et résultats sur la TVE et sur l'analyse de survie. Alors ce mémoire se divise en deux chapitres :

- Le premier chapitre est consacré à l'estimation sans censure, premièrement on donne des définitions de base : fonctions de répartition, du quentile, de queue, variables aléatoires et les différentes initiations de convergence, telle la convergence en probabilité et la convergence en distribution. Ensuite on donne des généralités sur la TVE : statistiques d'ordre, comportement asymptotique des extrêmes et l'estimation de l'IVE. La dernière section, on parle sur l'estimation de la moyenne : Le premier cas le moment centré d'ordre 2 "la variance $E[X^2]$" fini et le deuxième cas $E[X^2]$ infini.
- Dans le deuxième chapitre on présente l'estimation avec censure. On commence par des généralités sur l'analyse de survie : les notation de base (date d'origine, date des dernières nouvelles, ...ect). Puis, on donne les différents types de censure (censure à droite, censure à gauche, censure par intervalle, ...ect). Ensuite, on introduit l'estimation non paramétriques qui sont l'estimateur de Kaplan-Meier [22] de fonction de survie et l'estimateur de Stute [34] pour la moyenne. On présente dans la dernière section le problème de l'estimation de l'IVE et le problème de l'estimation de la moyenne dans le cadre de statistique des extremes sous censure aléatoire.

Enfin, il y a lieu de noter que les représentations graphiques sont réalisés à l'aide du logiciel d'analyse statistique R (The R Project for Statistical Computing).

CHAPITRE 1

Estimation sans censure

L'objectif principal de ce chapitre est de parler sur la méthode d'estimation de la moyenne sans censure, qui on va utiliser la théorie des valeurs extrêmes (TVE) classique, cette dernière, qui est utilisée pour la modélisation des évènements extrêmes. Pour cela on énonce des fondamentales sur la TVE. Pour des présentations plus détaillées sur la TVE, consulter Embrechtes et al [14], Beirlant et al [2] et Reiss et Thomas [29].

1.1 Définitions de base

1.1.1 Fonctions de répartition, quentile et queue

Définition 1.1.1 (Variable aléatoire)

Une variable aléatoire (v.a) X est une fonction de l'ensemble fondamental Ω à valeurs dans $\mathbb{R}$

$$X : \Omega \to \mathbb{R}. \tag{1.1}$$

Définition 1.1.2 (Loi de probabilité)

On appelle loi de probabilité de X notée P_X l'application qui a toute partie A de $\mathbb{R}$ associe :

$$P_X(A) = P(X^{-1}(A)) = P(X \in A). \tag{1.2}$$

Définition 1.1.3 (Fonctions de répartition et empirique)

- *La fonction de répartition (fdr) de la v.a X est définie par :*

$$F_X(x) = F(x) := P(X \leq x), \qquad \forall x \in \mathbb{R}. \tag{1.3}$$

  F est appelée aussi fonction de distribution (fd) ou fonction de distribution cumulée (fdc).

- *La fonction de répartition empirique, notée par F_n et définie par :*

$$F_n(x) := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j \leq x\}, \qquad x \in \mathbb{R}. \tag{1.4}$$

  Où $\mathbb{1}_{\{X_j \leq x\}}$ est la fonction d'indicatrice de l'ensemble A. On appelle la statistique (1.4) l'estimateur de (1.3).

Propriété 1.1.1 (Caractéristiques de $F(x)$)

F est fonction de répartition si :

- $0 \leq F(x) \leq 1$.
- *$F(x)$ fonction croissante et continue à droite.*
- $\lim_{x \to -\infty} F(x) = 0$ *et* $\lim_{x \to +\infty} F(x) = 1$.

Définition 1.1.4 (Fonctions de quantile et quantile empirique)

La fonction du quantile d'une distribution de probabilités est l'inverse généralisé de la fonction

de distribution F

$$Q(p) = F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad 0 < p < 1. \tag{1.5}$$

La fonction du quantile empirique d'un échantillon $X_1, ..., X_n$ est définie par :

$$Q_n(p) = F_n^{-1}(p) := \inf\{x \in \mathbb{R} : F_n(x) \geq p\}, \quad 0 < p < 1. \tag{1.6}$$

Définition 1.1.5 (Fonctions de queue et de queue empirique)

– *La fonction de queue ou de survie, qu'on note par $\overline{F}(x)$ est définie sur $[0,1]$ par :*

$$\overline{F}(x) = 1 - F(x) := 1 - P(X \leq x) = P(X > x).$$

– *La fonction empirique de queue est :*

$$\overline{F}_n(x) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j > x\}, \quad \forall x \in \mathbb{R}. \tag{1.7}$$

Propriété 1.1.2 (Caractéristiques de $S(x)$)

$\overline{F}(x)$ la fonction de survie si :

- *$\overline{F}(x)$ fonction continue à droite, i.e $\overline{F}(x^+) = \overline{F}(x)$, pour $x \geq 0$.*
- *$\overline{F}(x)$ fonction décroissante telle que $\overline{F}(0) = 1$ et $\lim_{x\to\infty} \overline{F}(x) = 0$.*

Définition 1.1.6 (fonctions du quantile de queue et queue empirique)

– *La fonction du quantile de queue est définie par :*

$$U(x) := Q\left(1 - \frac{1}{x}\right), \quad x \in]1, +\infty[. \tag{1.8}$$

– *La fonction empirique de queue correspondante est :*

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

1.1.2 Convergences des suites de variables aléatoires

Soient $X_1, ..., X_n$ une suites de v.a's étant une suite de fonctions de Ω dans $\mathbb{R}$, il existe diverses façons de définir la convergence de $X_1, ..., X_n$ dont certaines jouent un grand rôle en calcul des probabilités. Avant de passer à la théorie asymptotique (ou limite) des extrêmes, on doit nous familiariser avec les concepts suivants de convergence :

– On dit qu'une suite de va's $X_1, ..., X_n$ converge en distribution vers une va X, on la note $X_n \xrightarrow{d} X$, si pour tout $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Ceci est également connu sous le nom de convergence faible.

– La suite $X_1, ..., X_n$ converge en probabilité vers X, on la note $X_n \xrightarrow{P} X$, si pour tout $\epsilon > 0$,

$$\lim_{n\to\infty} (\mid X_n - X \mid > \epsilon) = 0.$$

– La suite $X_1, ..., X_n$ converge presque sûrement vers X, on la note $X_n \xrightarrow{p.s.} X$, si :

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

La convergence presque sûre est également appelée convergence forte.

La convergence de F_n vers F est presque sûrement uniforme c'est-à-dire que :

$$\sup_x \{\mid F_n(x) - F(x) \mid\} \xrightarrow{\boldsymbol{p.s}} 0, \qquad \text{quand } n \to \infty. \tag{1.9}$$

La convergence (1.9) est connue sous le nom de théorème de Glivenko-cantell. Il est l'un des résultats fondamentaux en statistiques non paramétriques. La preuve du résultat (1.9) peut-être trouvés dans tout manuel standard de la théorie des probabilités comme (Pierre [27], page 229).

1.2 Généralités sur la TVE

La TVE (Extreme Value Theory (EVT) en anglais) est une vaste théorie dont le but d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'apparition est faible, cette théorie étude du comportement asymptotique des grandes ou petites observations d'un échantillon de v.a's indépendantes et identiquement distribuées (iid).

1.2.1 Statistiques d'ordre

Soit $X_1, ..., X_n$, n v.a's réelles iid et de même fdr F, on définit les statistiques d'ordre dans les suivantes

Définition 1.2.1 (Statistique d'ordre)

la statistique d'ordre de l'échantillon $X_1, ...X_n$ est le réarrangement croissant de $X_1, ...X_n$, elle notait par $X_{1:n}, ..., X_{n:n}$, on écrit

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

En particulier, on note que :

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

$X_{1:n}$ elle représente la plus petite statistique d'ordre (où statistique du minimum) et $X_{n:n}$ est

la plus grande statistique d'ordre (où statistique du maximum). $X_{j:n}$ elle réprésente le $j^{ème}$ statistique d'ordre (statistique d'ordre j) dans un échantillon de taille n.

Proposition 1.2.1 (Distributions du maximum et du minimum)

La distribution du maximum $X_{n:n}$ et du minimum $X_{1:n}$ sont donnés par les deux formules suivantes :

$$F_{X_{1:n}}(x) = 1 - [1 - F(x)]^n \quad et \quad F_{X_{n:n}}(x) = [F(x)]^n, \quad \forall x \in \mathbb{R}. \tag{1.10}$$

On démontre que les deux expressions de (1.10)

$$\begin{aligned} F_{X_{n:n}}(x) &= P(X_{n:n} \leq x) &&= P(\bigcap_{j=1}^{n} \{X_j \leq x\} \\ &= \prod_{j=1}^{n} P(X_j \leq x) &&= [F(x)]^n, \end{aligned}$$

et

$$\begin{aligned} F_{X_{1:n}}(x) &= P(X_{1:n} \leq x) &&= 1 - P(\min(X_j) \geq x) \\ &= 1 - P(\bigcap_{j=1}^{n} \{X_j \geq x\}) &&= 1 - \prod_{j=1}^{n} (1 - P\{X_j \geq x\}) \\ &= 1 - \left[\overline{F}(x)\right]^n. \end{aligned}$$

David [9] et Balakrishnan et Cohn [1] montre que l'expression de la distribution $X_{j:n}$ est

$$F_{X_{j:n}}(x) = \sum_{r=j}^{n} \binom{n}{r} (F(x))^r (1 - F(x))^{n-r}. \tag{1.11}$$

Remarque 1.2.1

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

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

Ainsi, tous les résultats que nous allons présenter pour les maxima peut-être transposés pour les minima.

1.2.2 Comportement asymptotique des extrêmes

La TVE de suites de variables iid dans $\mathbb{R}$ commence par la connaissance de la loi asymptotique de la suite des maxima $X_{n:n}$ quand n tend vers l'infini sous l'hypothèse que les observations sont iid. Il faut donc une information sur la forme de la distribution en étudiant la loi du maximum. Les définitions et les remarques énoncées dans cette partie sont issues de Cláudia et al. [6]. On remarque d'abord

$$\lim_{n\to\infty} F_{X_{n:n}}(x) = \lim_{n\to\infty} [F(x)]^n = \begin{cases} 1 & x \geq x^F \\ 0 & x < x^F. \end{cases} \tag{1.12}$$

On constate que la distribution asymptotique du maximum donne une loi dégénérée, une masse de Dirac en x^F, puisque pour certaines valeurs de x, la probabilité peut-être égale à 1 dans le cas où x^F est fini et donc $X_{n:n}$ tend vers x^F presque sûrement, avec $x^F \leq \infty$. On donne dans la suivante la définition du point terminal :

Définition 1.2.2 (Point terminal)

Le point terminal supérieur (ou droit) de F est

$$x^F := \sup \{x \in \mathbb{R},\ F(x) \leq 1\}.$$

De façon équivalente, le point terminal inférieur (ou le point terminal gauche) d'une distribution F est désigné par x_F est

$$x_F := \inf \{x \in \mathbb{R},\ F(x) > 0\}.$$

Voir Embrechts et al. (1997) , Exemple 3.3.22, p.139.

Ce fait (1.12) ne fournit pas assez d'informations, d'où l'idée d'utiliser une transformation afin d'obtenir des résultats plus exploitables pour les lois limites des maxima.On s'intéresse par conséquent à une loi non dégénérée pour le maximum, la TVE permet de donner une réponse à cette problématique.Les premiers résultats sur la caractérisation du comportement asymptotique des maxima $X_{n:n}$ convenablement normalisés ont été obtenus par Fisher et Tippet [15], Gnedenko [16]. Ce résultat est analogue au Théorème Central Limite (TCL) (on va donner ce dernier dans le Théoreme 1.3.2) mais pour les phénomènes extrêmes. On a ici le théorème qui donne la loi du maximum.

Théoreme 1.2.1 (Fisher-Tippet-Gnedenko-Mises-Jenkinson)

Sous certaines conditions de régularités sur la fonction F, s'il existe deux constantes de normalisation $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1}$ réelle telle que :

$$\lim_{n\to\infty} P\left(\frac{X_{n:n}-b_n}{a_n}\leq x\right) = \lim_{n\to\infty} F^n\left(a_n x + b_n\right) = \mathcal{H}(x), \quad \forall x\in\mathbb{R}, \tag{1.13}$$

la loi limite converge en distribution vers une loi non dégénérée $\mathcal{H}(x)$. Cette loi de la limite que l'un des trois types suivantes :

$$\begin{array}{lll}
\textit{Gumbel :} \quad \Lambda(x) & = \exp(-e^{-x}), & x\in\mathbb{R} \text{ et } \alpha = 0. \\
\textit{Fréchet :} \quad \Phi_\alpha(x) & = \begin{cases} 0, & x\leq 0 \\ \exp\left(-x^{-\alpha}\right), & x>0 \end{cases} & \alpha > 0. \\
\textit{Weibull :} \quad \Psi_\alpha(x) & = \begin{cases} \exp\left(-\left(-x^{-\alpha}\right)\right), & x\leq 0 \\ 0, & x>0 \end{cases} & \alpha > 0.
\end{array} \tag{1.14}$$

Ces trois distributions limites sont appelées les distributions de valeurs extrêmes standard.

Une démonstration détaillée de ce théorème est donnée dans Resnick [30] et avec des développements dans Embrechts et al. [14], page 152.

Il est possible de rassembler les trois familles de lois en une seule famille paramétrique ($\mathcal{H}_\gamma(x)$, $\gamma \in \mathbb{R}$) de Von Mises-Jenkinson.[[35], [21]], elle est dite famille des lois des valeurs extrêmes généralisées (GEV, Generalized Extreme Value distribution). Notamment, la distribution des valeurs extrêmes généralisées donnée par :

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\left\{-(1+\gamma)^{-1/\gamma}\right\} & \text{si } \gamma \neq 0, \quad 1+\gamma x > 0 \\ \exp\left\{-e^{(-x)}\right\} & \text{si } \gamma = 0. \end{cases} \tag{1.15}$$

Le paramètre $\alpha := 1/\gamma$, il représente l'indice des valeurs extrêmes (IVE) si $\gamma > 0$ ou indice de queue ou paramètre de forme.

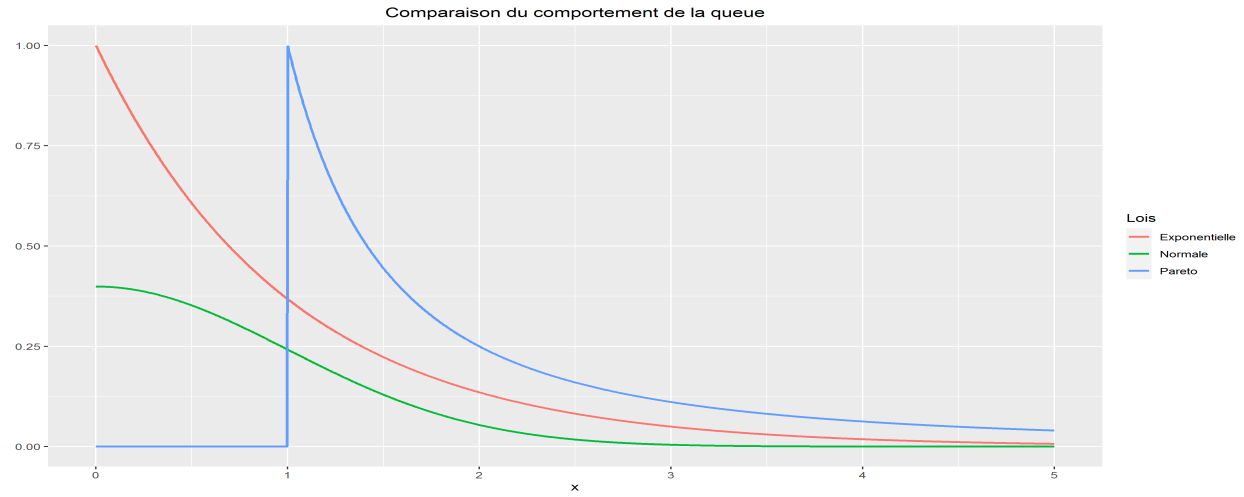

FIG. 1.1 – Comparaison du comportement de queue

La figure ci-dessus, elle représente une comparaison du comportement de queue où les distributions à queue lourdes, elles sont représentées par la courbe verte elle est décroît lentement vers zéro par rapport la distribution exponentielle et la distribution exponentielle que représentée par la courbe rouge et la distribution à queue légère représentée par la courbe bleue.

Distributions à queue lourde

Les distributions à queues lourdes sont liées à la TVE et permettent de modéliser beaucoup de phénomènes que l'on trouve dans différentes disciplines telles que la finance, l'hydrologie, la climatologie épidémiologie, ...etc. Ce type des distributions est défini ainsi :

Définition 1.2.3 (Distribution à queue lourde)

Soit X une v.a de fdr F, donc cette dernière elle est dite distribution à queue lourde, s'il existe un constant positif γ qui représente l'indice de queue et prend la formule suivante :

$$\overline{F}(x) \sim x^{-1/\gamma} l(x), \ \textit{pour } x \to \infty, \tag{1.16}$$

où $l(x)$ la fonction à variation lente au voisinage de l'infini.

Remarque 1.2.2

- *On dit qu'une fonction V est à variations régulière d'indice $\rho \in \mathbb{R}$ (à l'∞) si V est positive à l'infini et si pour tout $t > 0$*
$$\lim_{t\to\infty} \frac{V(tx)}{V(x)} = t^{\rho}.$$
- *On dit qu'une fonction l est à variations lentes à l'infini si $l(x) > 0$ pour x assez grand et si pour tout $t > 0$, on a*
$$\lim_{t\to\infty} \frac{l(tx)}{l(x)} = 1, \qquad t > 0.$$

Le type de distribution (1.16) satisfaite pour tout $x > 0$, les conditions suivantes :

Définition 1.2.4 (Conditions du premier et du second ordre)

(i) *$1 - F(x)$ est une fonction à variation régulière d'indice $-1/\gamma < 0$, on a*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}, \qquad x > 0. \tag{1.17}$$

(ii) $\exists\ \beta \leq 0$ *et une fonction* $A \to 0$ *et ne change pas le signe au voisinage de l'infini, tel que*

$$\lim_{t\to\infty} \frac{1 - F(tx)/1 - F(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma} \frac{x^{\beta} - 1}{\gamma\beta}. \tag{1.18}$$

Remarque 1.2.3

La condition du premier ordre en général n'est pas suffisante pour étudier les propriétés des estimateurs des paramètres de queue, en particulier la normalité asymptotique. Dans ce cas, une condition du second ordre des fonctions à variations régulières est nécessaire en spécifiant le taux de convergence dans (1.17)*. Cette condition vient de de Haan et Ferreira [18] page* 48*.*

Domaines d'attraction

Définition 1.2.5 (Domaines d'attraction)

Si F *vérifie le Théoreme 1.2.1 alors on dit que* F *appartient au Domaine D'attraction* $(\mathcal{DA})$ *du maximum de* $\mathcal{H}_\gamma$ *et on note* $F \in \mathcal{DA}(\mathcal{H}_\gamma)$*, s'il existe deux suites normalisantes* $a_n > 0$ *et* b_n *fournis aussi dans le Théoreme 1.2.1 de telle manière que*

$$\lim_{n\longrightarrow\infty} F^n\left(a_n x + b_n\right) = \mathcal{H}_\gamma(x). \tag{1.19}$$

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

- Si $\gamma > 0$, on dit que $F \in \mathcal{DA}(\Phi_{1/\gamma})$, et F a un point terminal à droite infinie $(x_F = +\infty)$. Ce domaine d'attraction est celui des distributions à queues lourdes, c'est-à-dire qui ont une fonction de survie à décroissance polynomiale.
- Si $\gamma = 0$, on dit que $F \in \mathcal{DA}(\Lambda)$, le point terminal x_F peut alors être fini ou non. Ce domaine d'attraction est celui des distributions à queues légères, c'est-à-dire qui ont une fonction de survie à décroissance exponentielle.
- Si $\gamma < 0$, on dit que $F \in \mathcal{DA}(\Psi_{1/\gamma})$, et F a un point terminal à droite finie $(x_F < +\infty)$.

Ce domaine d'attraction est celui des fonctions de survie dont le support est borné supérieurement.

Exemple 1.2.1 (Loi Pareto)

La fdr de la loi de Pareto est $F(x) = 1 - cx^{-\alpha}$, *avec* $c > 0$ *et* $\alpha > 0$. *On pose* $a_n = \left(\frac{1}{nc}\right)^{-1/\alpha}$ *et* $b_n = 0$ *alors :*

$$\begin{aligned} P\left(\frac{X_{n:n}}{(1/nc)^{-1/\alpha}}\right) &= \lim_{n\to\infty} F^n\left(\left(\frac{1}{nc}\right)^{-1/\alpha} x\right) = \left(1 - c\left(\left(\frac{1}{nc}\right)^{-1/\alpha} x\right)^{-\alpha}\right)^n \\ &= \left(1 - \frac{x^{-\alpha}}{n}\right)^n \to \exp(-x^{-\alpha}) = \Phi_\alpha(x). \end{aligned}$$

Donc le maximum de normalisation de la loi de Pareto converge vers la loi de Fréchet.

Le tableau suivant résume le classement de quelques lois usuelles selon leur appartenance à l'un des domaines d'attraction. (Embrechts et al [14])

Domaines d'attraction	Gumbel ($\gamma = 0$)	Fréchet ($\gamma > 0$)	Weibull ($\gamma < 0$)
Lois	Normal Exponentielle Log-normale Gamma Weibull	Pareto Burr Student Log-gamma Log-logistique	Uniforme Beta ReverseBurr

TAB. 1.1 – Quelques loi usuelles classées en fonction de leur domaine d'attraction

La représentation des trois fonctions de densité de Λ, Φ, Ψ elle est illustrée dans la Figure 1.2. Cette dernière, elle montre que les distributions prototypiques à chaque domaine d'attraction, pour les valeurs sélectionnées de γ. On souligne les queues longues, les queues en décomposition polynomiales présentées par les distributions choisies de Fréchet, contrastant avec les queues courtes délimitées en haut, attribuées au domaine de Weibull (voir Cláudia et al. [6], 2019).

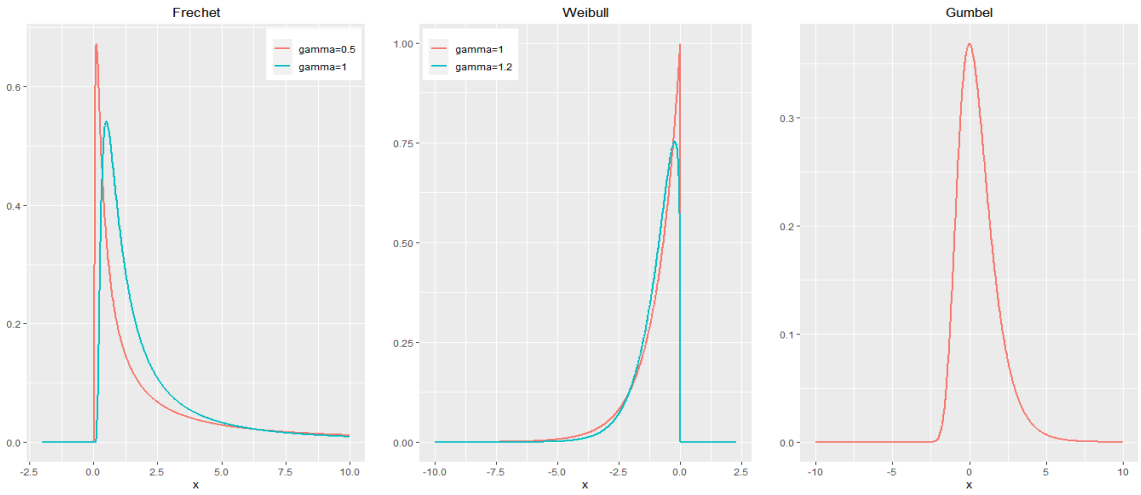

FIG. 1.2 – Densités de lois des valeur extremes avec $1/\mid \gamma \mid, \alpha > 0$.

1.2.3 Estimation de l'IVE

On donne trois familles d'estimateurs du paramètre de la loi de valeurs extrêmes généralisées. Il en existe de nombreux autres, voir les monographies [2] et [14].

Estimateur de Pickands

Cet estimateur a été introduit en 1975 par James Pickands dans [28], pour toute $\gamma \in \mathbb{R}$.

Définition 1.2.6 (Estimateur de Pickands)

Soient $X_1, ..., X_n$, n v.a's iid de fdr $F \in \mathcal{DA}(\mathcal{H}_\gamma)$, où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suites d'entiers avec $1 < k < n$, $k/n \to 0$, quand $n \to \infty$. L'estimateur de Pickands est défini par :

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

L'auteur a démontré la consistance faible d'estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers [12] et de Haan [18].

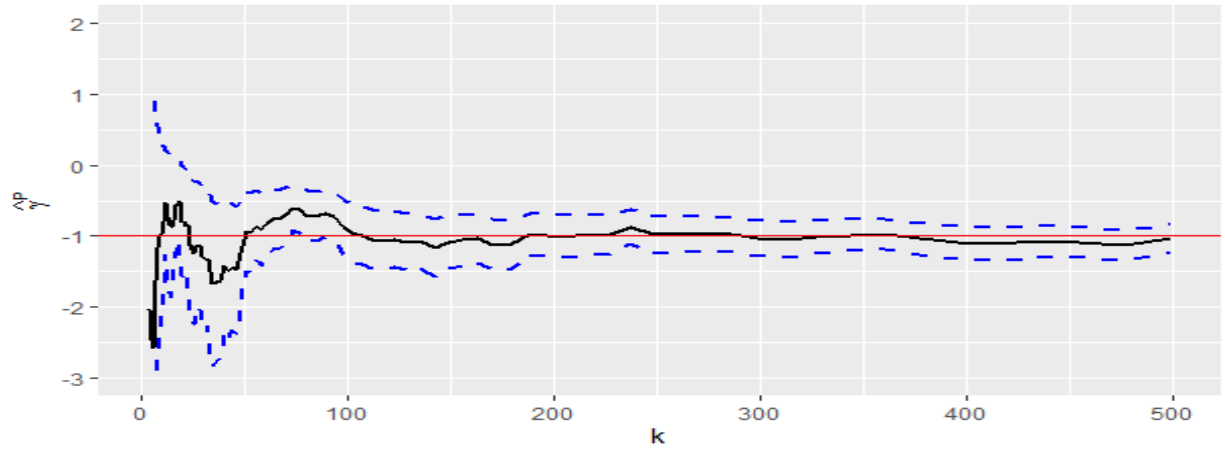

FIG. 1.3 – Estimateur de Pickands, avec un intervall de confiance de niveau 95%, pour l'EVI de distribution de Pareto standard ($\gamma = -1$) basé sur 1000 échantillons de 5000 observations.

Estimateur de Hill

Cet estimateur a été introduit par Hill [20] en 1975. C'est un estimateur simple et largement utilisé. Cet estimateur n'est applicable que dans le cas où l'EVI, est connu pour être positif, ce qui répond à des distributions appartenant au domaine d'attraction de type Fréchet ($\gamma > 0$).

Définition 1.2.7 (Estimateur de Hill)

Soient $X_1, ..., X_n$, n v.a's iid de fdr $F \in \mathcal{DA}(\Phi_{1/\gamma})$, où $\gamma > 0$. soit $k = k_n$ une suites d'entiers avec $1 < k < n$, $k/n \to 0$, quand $n \to \infty$. L'estimateur de Hill est défini par :

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

Un grand nombre des travaux théoriques ont été consacrés à l'étude des propriétés de l'estimateur de Hill [20]. La consistance faible a été établie par (Mason [23],1982), la consistance forte fut établie par (Deheuvels et al.[11], 1988) et plus récemment par (Necir [25], 2006) .La normalité asymptotique est due entre autres à (Davis, Resnick [10], 1984), (Csörgő et Mason [8], 1985) et (Häusler et Teugels [17], 1985).

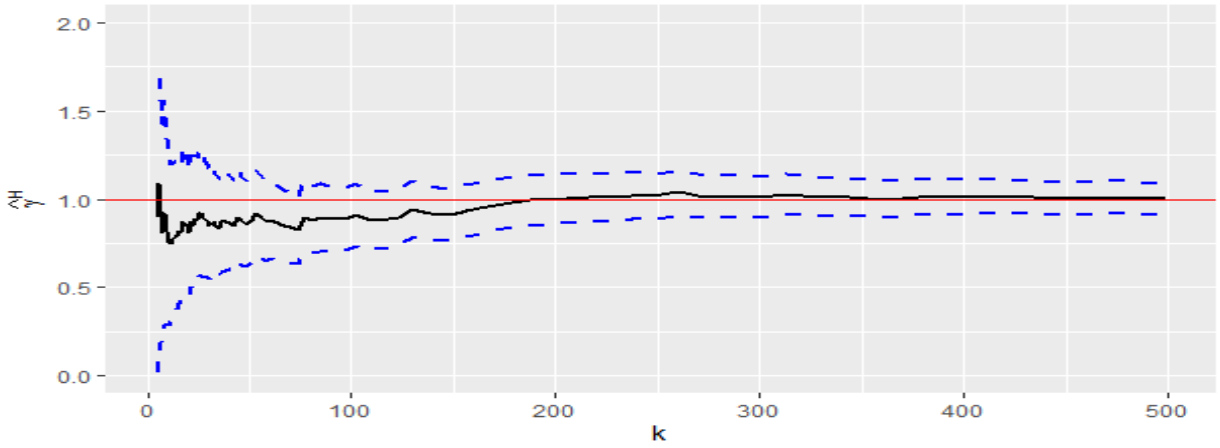

FIG. 1.4 – Estimateur de Hill, avec un intervalle de confiance de niveau 95%, pour l'EVI de distribution de Pareto standard ($\gamma = 1$) basé sur 1000 échantillons de 5000 observations.

Estimateur des moments

Pour $\gamma \in \mathbb{R}$, en 1989, Dekkers et al. [12] ont proposé une extension de tous types de distribution, appelé estimateur des moments.

Définition 1.2.8 (Estimateur des moments)

Pour $\gamma \in \mathbb{R}$, soit $k = k_n$ une suites d'entiers avec $1 < k < n$, $k/n \to 0$, quand $n \to \infty$. L'estimateur des moments est

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

avec

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

Où $M_n^{(1)}$ est l'estimateur de Hill $\hat{\gamma}^M$.

Les propriétés asymptotiques de cet estimateur ont été étudiées par Dekkers et al. [12], 1989.

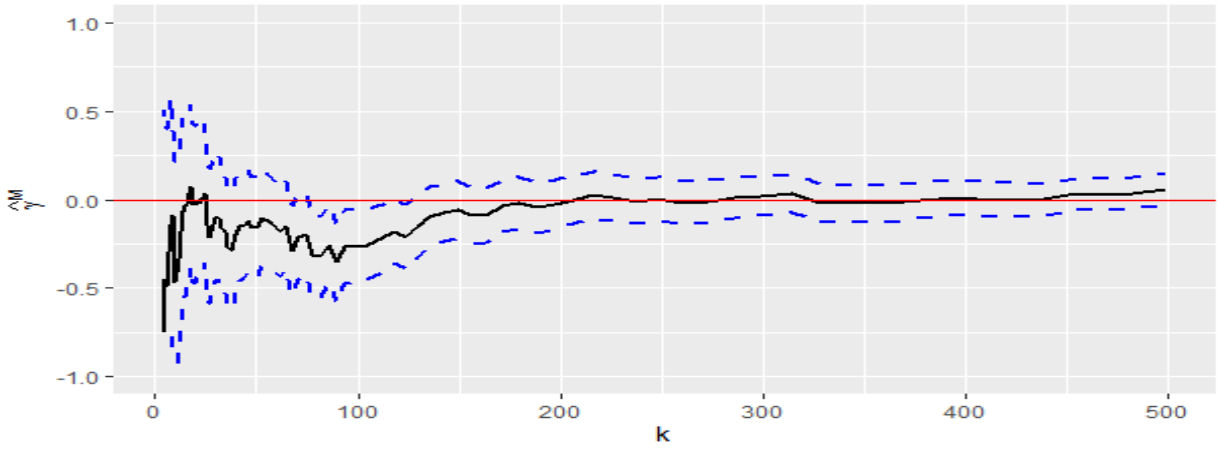

FIG. 1.5 – Estimateur de Moments, avec un intervalle de confiance de niveau 95%, pour l'EVI de distribution de Gumbel standard ($\gamma = 0$) basé sur 1000 échantillons de 5000 observations.

1.3 Estimation de la moyenne

Une loi de probabilité peut être caractérisée par certaines valeurs typiques associées aux notions de valeur centrale, de dispersion et de forme de la distribution. La définition suivante devient de livre de Saporta [31], page 22.

Définition 1.3.1 (Espérance mathématique)

X étant un v.a réelle définie sur $(\mathbb{R}, \mathcal{A}, P)$, L'espérance mathématique de X est, si elle existe, l'intégrale de X par rapport il la mesure F :

$$\mu = E[X] := \int_{\mathbb{R}} X dF.$$

D'après le théorème de la mesure image, on a :

$$\mu = \int_{\mathbb{R}} x dF_X(x), \tag{1.20}$$

Remarque 1.3.1 (Espérance par quantile)

On peut donne une autre formule d'espérance par rapport au quantile si on pose $F(X) = t$ *avec changement de condition limite d'intégral on trouve qu'il définit par :*

$$\mu := \int_0^1 Q(t)dt. \tag{1.21}$$

1.3.1 Estimation de la moyenne dans le cas $E[X^2]$ fini

Définition 1.3.2 (La moyenne empirique)

On appelle moyenne de l'échantillon ou moyenne empirique ou encore moyenne statistique, notée $\overline{X}$ *définie par*

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \quad et \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2, \tag{1.22}$$

où S^2 *est la variance empirique.*

Proposition 1.3.1

Soit X *une v.a de moyenne* μ *et d'écart-type* σ*.On a* $E\left[\overline{X}\right] = \mu$ *et* $Var\left[\overline{X}\right] = \sigma^2/n$*.*

Il y a deux méthodes pour obtenir l'estimateur de la moyenne, la première méthode on utilise la formule (1.20) de Définition 1.3.1 et la deuxième méthode on utilise la formule (1.21) de la Remarque 1.3.1. En effet, en substation de Q par Q_n dans (1.21) , on obtient

$$\begin{aligned}
\widehat{\mu} &= \int_0^1 X_{j:n}dx, \qquad \text{où } \frac{j-1}{n} \leq x \leq \frac{j}{n} \\
&= \sum_{j=1}^{n}\int_{j-1/n}^{j/n} X_{j:n}dx, \\
&= \frac{1}{n}\sum_{j=1}^{n} X_{j:n}, \\
&= \overline{X}.
\end{aligned}$$

Théoreme 1.3.1 (Lois des grands nombres)

Soit $X_1, ..., X_n$ indépendantes d'espérance $E[X] < \infty$, et var $[X] < \infty$ tel que :

$$\textit{Loi faible :} \quad \overline{X} \xrightarrow{\mathbf{p}} \mu \quad \textit{quand } n \to \infty.$$

$$\textit{Loi forte :} \quad \overline{X} \xrightarrow{\mathbf{p.s}} \mu \quad \textit{quand } n \to \infty.$$

Le théorème suivant connaître sous le nom de théorème central-limite (il vaudrait mieux dire théorème de la limite centrée) établit la convergence vers la loi normale sous des hypothèses peu contraignantes. Comme un exemple la Figure 1.6 montre que $\overline{X}$ convergent vers la vraie valeur de μ pour la loi gaussienne

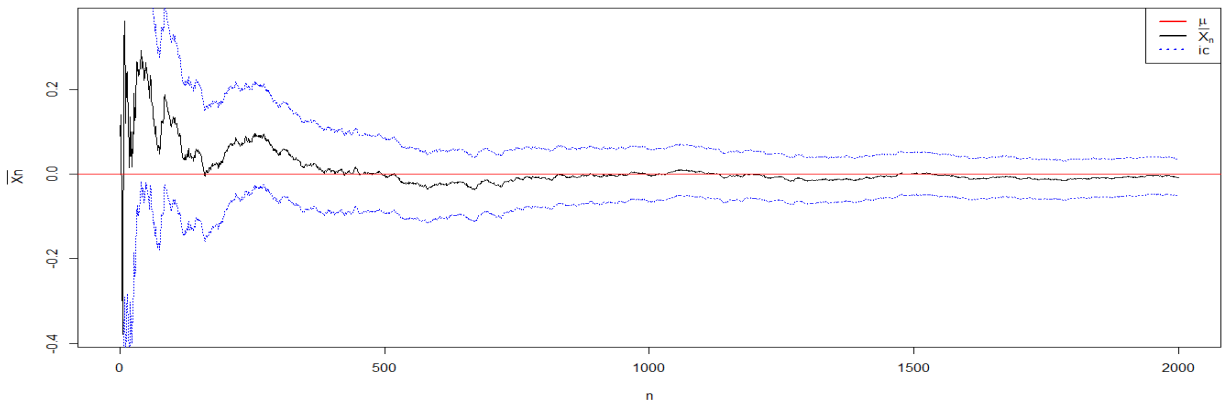

FIG. 1.6 – Illustration de la loi des grands nombres : moyenne empirique d'unéchantillon Gaussien standard de taille 2000.

Théoreme 1.3.2 (Théorème central limite)

Soit $(X_n)_{n\geq 1}$ une suites de v.a's iid d'espérance μ et de variance σ^2 finie, alors

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + ... + X_n - n\mu}{\sigma}\right) = \sum_{j=1}^{n} \frac{X_j - \mu}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1), \quad \textit{quand } n \to \infty.$$

L'intervalle de confiance de seuil λ pour le paramètre μ de la loi $\mathcal{N}(\mu, \sigma^2)$ est :

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_\lambda,\ \overline{X} + \frac{\sigma}{\sqrt{n}} z_\lambda\right].$$

La preuve du Théoreme 1.3.2 peut être trouvée dans n'importe quel livre standard des statistiques (voir par exemple, Saporta [31], page 66).

1.3.2 Estimation de la moyenne dans le cas $E[X^2]$ infini

On a vu dans la Subsection 1.3.1, pour $\alpha = 2$ (F appartient a au domaine d'attraction d'une distribution normale.) la moyenne de X existe et elle est égale au paramètre de localisation μ, la variance de X est finie et elle est égale à $2\sigma^2$ dans ce cas l'estimateur naturel de μ est la moyenne empirique $\overline{X}$, et en vertu au Théoreme 1.3.2 l'estimation de X est asymptotiquement normale. Mais, dans le cas $1 < \alpha < 2$ (on dite que F appartient au domaine d'attraction d'une loi stable), le Théoreme 1.3.2 n'est pas applicable parce que la variance de X est infinie. Par conséquent, la normalité asymptotique de la moyenne de l'échantillon $\overline{X}$ n'est pas établie. Pour pallier ce problème, Peng [26] en 2001 a proposé un estimateur asymptotiquement normal basé sur la théorie des valeurs extrêmes, comme suit :

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

où

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

$$\widehat{\mu}_n^{(1)} = \widehat{\mu}_n^{(1)}(k) := \frac{k}{n} X_{k:n} \frac{\widehat{\alpha}_n^{(1)}}{\widehat{\alpha}_n^{(1)} - 1} \quad \text{et} \quad \widehat{\mu}_n^{(3)} = \widehat{\mu}_n^{(3)}(k) := \frac{k}{n} X_{n-k+1:n} \frac{\widehat{\alpha}_n^{(3)}}{\widehat{\alpha}_n^{(3)} - 1},$$

avec

$$\widehat{\alpha}_n^{(1)} = \widehat{\alpha}_n^{(1)}(k) := \left(\frac{1}{k} \sum_{j=1}^{k} \log\left(-X_{j:n}\right) - \log\left(-X_{k:n}\right)\right)^{-1},$$

et

$$\widehat{\alpha}_n^{(3)} = \widehat{\alpha}_n^{(3)}(k) := \left(\frac{1}{k} \sum_{j=1}^{k} \log\left(-X_{n-j+1:n}\right) - \log\left(-X_{n-k:n}\right) \right)^{-1}.$$

On note que $\widehat{\alpha}_n^{(1)}$ et $\widehat{\alpha}_n^{(3)}$ sont également des estimateurs convergents en probabilité vers α (see Mason [23], 1982) et la convergence presque sûre est établie dans Necir [25] en 2006.

Théoreme 1.3.3 (Normalité asymptotique)

Sous certaines conditions, Peng [26] a montré que, quand $k = o(n^{-2\rho(\alpha-2\rho)})$, *avec le paramètre du second ordre* $\rho < 0$,

$$\frac{\sqrt{n}}{\sigma(k/n)} \left(\widehat{\mu}_n^P - \mu \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^2) \quad \textit{quand } n \to \infty, \tag{1.23}$$

ou de façon équivalente

$$\frac{\sqrt{n}}{\delta\sigma(k/n)} \left(\widehat{\mu}_n^P - \mu \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad \textit{quand } n \to \infty, \tag{1.24}$$

où

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

et

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

Preuve. Voir Peng [26], 2001. ■

CHAPITRE 2

Estimation avec censure

L'analyse des modèles de survie est une branche de la statistique qui a pris son développement depuis la Deuxième Guerre mondiale. Cette analyse s'intéresse à l'étude statistique des données qui proviennent des expériences sur des durées de vie ou durée de fonctionnement. Le terme de cette durée 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 habilité) ou la survenue d'un sinistre (temps entre deux sinistres, en actuariat). Pour plus de detail sur ce thème on peux refère au Collett [7] .

Dans ce chapitre, on va adresser toutes les méthodes statistiques qui va utiliser dans la partie principale de ce chapitre qu'est la Subsection 2.2.2.

2.1 Généralités sur l'analyse de survie

Dans cette section on va faire des préliminaires sur les modèles de survie, on introduit les notions de base en analyse de survie, on donne aussi les différents types de censure et on va présenter les principaux estimateurs non paramétriques et semi-paramétriques.

2.1.1 Notions de base en analyse de survie

Définition 2.1.1 (Analyse de survie)

L'analyse de survie est le terme utilisé pour décrire l'analyse des données qu'ils sont sous forme de temps (times) allant de l'origine du temps (time origin) à la survenance d'un événement ou d'un point final spécifique (end point).

Définition 2.1.2 (Date d'origine)

C'est la date de début de l'observation, c'est-à-dire l'origine du début de l'analyse de survie, elle correspond au temps égal à zéro $t = 0$.

Exemple 2.1.1

- *Date de tirage au sort (essai thérapeutique).*
- *Date de diagnostic (étude prospective).*

Définition 2.1.3 (Date des dernières nouvelles)

Pour faire l'analyse des résultats, chaque individu dispose d'une date des dernières nouvelles, qui est la date la plus récente où les informations ont été recueillies. Par exemple, dans la case où le sujet a subi l'événement, sa date de dernière nouvelle est la date de survenue de cet événement.

Définition 2.1.4 (Date de point)

On ne prouve pas attendre la survenue de l'évènement pour tous les sujets pour faire l'analyse des observations, donc il y a ce qu'on appelle la date de point qui est une date au-delà de

laquelle on arrête l'observation et on ne tiendra plus compte de l'état du sujet après cet instant.

Définition 2.1.5 (Durée de survie)

C'est la durée entre la date d'origine et la survenue de l'évènement d'intérêt, c'est-à-dire du décès. Elle correspond au temps de suivi lorsque le décès est observé avant la date de point.

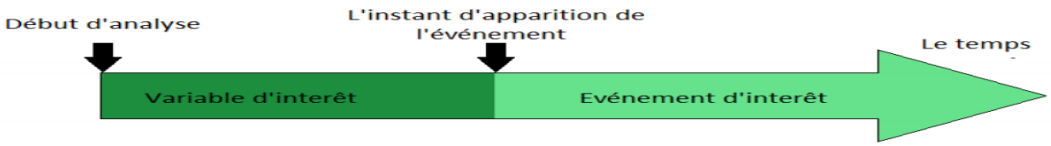

FIG. 2.1 – Schéma représentant les principales définitions relatives à l'analyse de durée de survie. (source : Harrouche [19], mémoire de master en Statistiques et Probabilité.

2.1.2 Censure

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique. Pour un individu donné j, on va considérer

- Son temps de survie X_j de fdr F.
- Son temps de censure Y_j de fdr G.
- La durée réellement observée Z_j de fdr H.

Définition 2.1.6 (Variable de censure)

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

Quelques détails de différents cas de censure suivent :

Censure à droite

La variable d'intérêt est dite censure à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées.

Censure à gauche

Il y a censure à gauche lorsque l'individu déja l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue.

Censure double ou mixte

Il y a censure mixte (double) dans un échantillon de données s'il y a à la fois censure à gauche et à droite dans cet échantillon. Les données sont censurées à la fois à droite et à gauche. Plusieurs modèles non paramétriques ont été présentés pour l'étude de la double censure.

Censure par intervalle

Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôles périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. on a aussi ce genre de donnée 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 intervalles du type $[c, +\infty[$ et $[0, c]$ respectivement.

Il existe différents types de censures : censure du type I, si le temps de censure est fixé par le chercheur comme étant la fin de l'étude. La censure du type II, se caractérise par le fait que l'étude cesse aussitôt que se produisent un nombre d'événements prédéterminés par l'expérimentateur. Une autre possibilité de censure aléatoire est que la censure n'est plus du

tout sous le contrôle du chercheur et (ou) que le temps d'entrée varie aléatoirement (voir Ndao [24] : thèse de doctorat de Probabilité et Statistique). Le type le plus intéressant dans les études est la censure aléatoire à droite.

Censure aléatoire à droite

Soient $Y_1, ...Y_n$ des v.a iid. On observe les variables

$$Z_j = X_j \wedge Y_j := \min(X_j, Y_j) \text{ et } \Delta_j = \mathbb{1}_{\{X_j \leq Y_j\}} \qquad \text{pour } 1 \leq j \leq n. \tag{2.1}$$

L'information disponible peut être résumée par :

- la durée réellement observée Z_j,
- $\Delta_j = 1$ si l'événement est observé (d'où $Z_j = X_j$). On observe les "vraies" durées ou les durées complètes.
- $\Delta_j = 0$ si l'individu est censuré (d'où $Z_j = Y_j$). On observe des durées incomplètes (censurées).

Pour une présentation détaillée des différents types de censure voire comme un exemple le livre de Collett [7].

2.1.3 Estimation non paramétrique

En l'absence de censure, la fdr F s'estime de manière très simple en utilisant la fdr empirique usuelle (1.4). Malheureusement dans le cas où les données sont censurées, il est impossible d'utiliser cette dernière puisqu'elle fait intervenir des quantités non observées, car tous les Z_i censurés ne sont pas observés. On estime alors généralement F en utilisant l'estimateur Kaplan-Meier (1958), [22]. Ce dernier est l'outil de bas en statistiques pour estimer de manière non paramétrique la distribution d'une v.a X censurée à droite.

Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier (KM) est le modèle de durée le plus utilisé en pratique. Il intervient dans toutes les applications qui requièrent la modélisation de durées. Soit $(Z_j, \Delta_j)_{1\leq j\leq n}$, l'échantillon réellement observé défini par (2.1) et soit $Z_{1:n} \leq ... \leq Z_{n:n}$ sont des statistiques d'ordre. L'estimateur de Kaplan-Meier (1958) est défini :

$$\begin{aligned} \overline{F}_n^{KM}(x) &= \prod_{j=1}^{n} \left(\frac{n-j}{n-j+1}\right)^{\Delta_{[j:n]}\mathbf{1}_{\left\{Z_{j:n}\leq x\right\}}} \\ &= \prod_{j=1}^{n} \left[1 - \frac{\Delta_{[j:n]}\mathbb{I}_{\{Z_{j:n}\leq x\}}}{n-j+1}\right]. \end{aligned} \tag{2.2}$$

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

Remarque 2.1.1

- Cet estimateur de Kaplan-Meier est une fonction étagée avec des sauts seulement aux observations non censurées.
- La hauteur des sauts de cet estimateur est aléatoire.
- Quand toutes les observations sont non censurées alors on obtient la fonction de répartition empirique, un extrait de thèse de Ndao [24].

Estimation de la moyenne

Lorsque la variable X est censurée l'estimateur cité ci-dessus (1.22) ne fonctionne pas car il est basé sur la totalité des observations, dans ce cas, Stute [34] a introduit un estimateur qui s'appelle les intégrales de Kaplan-Meier pour des quantités plus générales que la moyenne.

$$\widetilde{\mu} := \int \varphi(x) dF(x), \quad \text{pour } \varphi(x) = x.$$

Définition 2.1.7 (Moyenne empirique sous censure aléatoire)

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

$$\widetilde{\mu}_n := \sum_{j=1}^{n} W_{j,n} Z_{j:n},$$

où $W_{j,n} := \dfrac{\Delta_{[j:n]}}{n-j+1} \prod_{i=1}^{j-1} \left(\dfrac{n-i}{n-i+1} \right)^{\Delta_{[i:n]}}$.

Stute [34] a montré que cet estimateur est asymptotiquement normale sous les deux conditions suivantes :

$$I_1 = \int_0^{\infty} x^2 \Gamma_0^2(x) dH^{(1)}(x) < \infty, \tag{2.3}$$

et

$$I_2 = \int_0^{\infty} x \left(\int_0^{\infty} \frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2} \right)^{1/2} dF(x) < \infty, \tag{2.4}$$

où

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

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

Théoreme 2.1.1 (Normalité asymptotique)

Sous les deux conditions (2.3) *et* (2.4) *, on a*

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

où $\sigma^2 := var Z_1 \Gamma_0\left(Z_1\right) \Delta_1 + \Gamma_1\left(Z_1\right)\left(1 - \Delta_1\right) - \Gamma_2 Z_1.$

Preuve. Voir Stute [34]. ■

2.2 Statistique des extremes sous censure aléatoire

On va s'intéresser dans cette partie au problème de l'estimation de l'IVE et le problème de l'estimation de la moyenne cela en présence de données censurées aléatoirement à droite.

2.2.1 Estimateur de Hill adapté

Le problème de l'estimation de l'IVE est très récent dans la littérature, les premiers qui ont mentionné le sujet sont (Beirlant et al. [5], 1996) et (Reiss et Thomas [29], 2007) , mais sans résultats asymptotiques. Certains estimateurs des paramètres de la queue ont été proposés par (Beirlant et Guillou [3], 2001) pour les données tronquées et étendues à la censure aléatoire par (Beirlant et al. [4], 2007) et l'année suivante par (Einmahl et al. [13], 2008) .En réalité, l'estimation des valeurs extrêmes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $X_1, ..., X_n$ n'est pas observé totalement, mais qu'il est censuré par un deuxième échantillon $Y_1, ..., Y_n$, qui est supposé être indépendant du premier, où les X_j et Y_j sont des v.a's iid de lois F et G respectivement. Toutefois, il convient de signaler que, les différents estimateurs proposés de l'IVE en prenant en considération avec censure ont été tous construis de la même manière. Alors l'estimateur d'Hill adapté est basé sur un estimateur standard de l'indice de queue divisé par la proportion de données non censurées dans les plus grands k v.a's $Z_1, ..., Z_n$. Se référer à l'article de Zouadi et al [36]

$$\widehat{\gamma}_1^{(c,.)}(k) = \frac{\widehat{\gamma}^{(.)}(k)}{\widehat{p}}, \tag{2.5}$$

où $\widehat{p} = \widehat{p}(k) = \frac{1}{n}\sum_{j=1}^{k} \Delta_{[n-j+1:n]}$, $\Delta_{[i:n]}$ est le concomitant de la $i-$éme statistique d'ordre, c'est-à-dire, $\Delta_{[i:n]} = \Delta_j$ si $Z_{[i:n]} = Z_j$, $1 \leq j \leq n$ et $\widehat{\gamma}^{(.)}$ peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill, moment,... (basé sur les observations complètes). (voir Zouadi : [36]).

2.2.2 Estimation de la moyenne

Cette étude a été réalisée par l'équipe de Soltane et al. [32] en 2015, l'objectif principal de cette partie est de présenter l'estimateur de la moyenne pour les distributions de queue lourde sous censure.

Soit $X_1, ..., X_n$ et $Y_1, ..., Y_n$, $n \geq 1$ sont des deux v.a positives iid de deux fdr F et G respectivement. On note $\{(Z_1, \Delta_1), ..., (Z_n, \Delta_n)\}$ l'échantillon réellement observé, pour 2.1 avec $1\mathrm{I}_{\{.\}}$ indiquant la fonction de l'indicateur. Ce dernier indique s'il y a censure ou non. Si l'on note H la fdr de Z, par l'indépendance de X et Y, on a $(1-H) = (1-F)(1-G)$. Dans cette morceau, on utilise la notation $\overline{N}(x) = N(\infty) - N(x)$, pour toute fonction N on suppose F et G sont à queue lourde ou en d'autres termes, que F et G variations régulièrement à l'infinité avec des indices négatifs $-1/\gamma_1$ et $-1/\gamma_2$ respectivement.

$$\lim_{x\to\infty} \frac{\overline{F}(xl)}{\overline{F}(l)} = x^{-1/\gamma_1} \text{ et } \lim_{x\to\infty} \frac{\overline{G}(xl)}{\overline{G}(l)} = x^{-1/\gamma_2}, \qquad \text{pour } x > 0.$$

Par conséquent, H est aussi à queue lourde avec un indice de queue $\gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.
Il existe des constantes $\beta_j < 0$ et des fonctions A_j, $j = 1, 2$ tend vers zéro, ne changeant pas de signe près de l'infinité telle que pour tout $x > 0$

$$\begin{aligned} \lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma_1}}{A_1(t)} &= x^{-1/\gamma_1} \frac{x^{\beta_1/\gamma_1 - 1}}{\gamma_1\beta_1}, \\ \lim_{t\to\infty} \frac{\overline{G}(tx)/\overline{G}(t) - x^{-1/\gamma_2}}{A_2(t)} &= x^{-1/\gamma_2} \frac{x^{\beta_2/\gamma_2 - 1}}{\gamma_2\beta_2}. \end{aligned} \tag{2.6}$$

Pour obtenir l'estimateur de la moyenne qui définit dans $\mu = \int_0^\infty \overline{F}(x)dx$, tel que $x \geq 0$. On a μ est un somme de deux termes :

$$\mu = \int_0^h \overline{F}(x)dx + \int_h^\infty \overline{F}(x)dx \quad = \mu_1 + \mu_2,$$

Intégrer la première intégrale par parties et changer les variables dans la deuxième respectivement, on obtient

$$\mu_1 = h\overline{F}(h) + \int_0^h xF(x)\,dx \quad \text{et} \quad \mu_2 = h\overline{F}(h)\int_1^\infty \frac{\overline{F}(hx)}{\overline{F}(h)}dx.$$

En substituant h et $F(x)$ par $Z_{n-k:n}$ et $\overline{F}_n^{KM}(x)$ dans l'équation précédente. On a (voir Soltane et al [32])

$$\widehat{\mu}_1 = \sum_{j=1}^{n-k}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} Z_{n-k} + \sum_{j=1}^{n}\frac{\delta_{[j:n]}}{n-j+1}\prod_{i=1}^{j-1}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} Z_{j:n}, \tag{2.7}$$

comme estimateur de μ_1. Concernant μ_2, on applique le théorème bien connu de Karamata (voir, par exemple, de Hann et Ferreira 2006 [18], page 363)

$$\mu_2 \sim \frac{\gamma_1}{1-\gamma_1}h\overline{F}(h), \text{ quand } n \to \infty, \qquad 0 < \gamma_1 < 1. \tag{2.8}$$

Les quantités h et $\overline{F}(h)$ sont, comme ci-dessus, naturellement estimées par $Z_{n-k:n}$ et

$$\overline{F}_n^{KM}(Z_{n-k}) = \prod_{i=1}^{j-k}\left(\frac{n-i}{n-i+1}\right)^{\Delta_{[i:n]}},$$

pour dériver un estimateur de μ_2, il faut utiliser l'estimateur de γ_1 qui donnait dans (2.5). En remplaçant, dans (2.8), F et γ_1 par sont des estimateurs $\overline{F}_n^{KM}$ et $\widehat{\gamma}_1^{(c.H)}$.

$$\widehat{\mu}_2 := \frac{\widehat{\gamma}_1^{(H,c)}}{1-\hat{\gamma}_1^{(c,.)}}Z_{n-k}\prod_{i=1}^{j-k}\left(\frac{n-i}{n-i+1}\right)^{\Delta_{[j:n]}}, \qquad \widehat{\gamma}_1^{(c,H)} < 1. \tag{2.9}$$

Enfin, avec (2.7) et (2.9), l'estimateur $\widehat{\mu}_n$ de la moyenne μ est :

$$\widehat{\mu}_n = \frac{1}{k}\sum_{i=1}^{n-k}\frac{\Delta_{[i:n]}}{n-i+1}\prod_{i=1}^{j-k}\left(\frac{n-i}{n-i+1}\right)^{\Delta_{[i:n]}} Z_{i:n} + \prod_{i=1}^{n-k}\left(\frac{n-i}{n-i+1}\right)^{\Delta_{[i:n]}}\frac{Z_{n-k}}{1-\widehat{\gamma}_1^{(H,c)}}.$$

Théoreme 2.2.1 (Normalité asymptotique)

On suppose que les conditions du deuxième ordre de variation régulière (2.6) *vérifiées avec* $\gamma_2/(1+2\gamma_2) < \gamma_1 < 1$. *Soit* $k = k_n$ *une suites d'entières avec* $1 < k < n$, $k \to \infty$ *et* $k/n \to \infty$,

$$\lim \sqrt{k} A_1(h) < \infty \text{ et } \sqrt{k} h \overline{F}(h) \to \infty.$$

Alors il existe des constantes finies m *et* $\sigma^2 > 0$ *telles que*

$$\frac{\sqrt{k}(\widehat{\mu}_n - \mu)}{Z_{n-k:n}\overline{F}_n(Z_{n-k:n})}, \qquad \text{quand } n \to \infty,$$

où $m := \dfrac{\omega_1}{(1-p\tau_1)(1-\gamma_1)^2} + \dfrac{\omega_1}{(\gamma_1+\tau_1-1)(1-\gamma_1)}$, *avec* $\omega_1 := \lim \sqrt{k} A_1(h)$.

Conclusion

A travers ce mémoire, on a tenté de faire une synthèse sur les différentes cas de l'estimtion de la moyenne μ (paramètre de localisation d'une distribution) que ce soit dans le cas où les données sont complètes (l'estimation classique non paramétrique c'est le $\overline{X}$ et semi paramétrique basé sur la TVE qui a été proposé par Peng [26] en 2001) ou dans le cas les données sont incomplètes, où on s'est concentré notre étude sur le type de données censurées (l'estimation non paramétrique qui a été proposé par Stute [34] en 1995 en se basant sur les résultats de Kaplan-Meier [22] (1958) et l'estimation semi paramétrique qui a été proposé par Soltane et al. [32] en 2015).

- Le premier point qui mérite d'être considéré est de chercher d'autre méthodes dans le but d'améliorer l'estimation de la moyenne qui a proposé par Soltane et al. [32], avec une illustration à travers une application sur une série de données réelles censurées.
- Le deuxième point à envisager est, il arrive assez souvent qu'une série statistique, ayant des données tronquées (un autre type de données incomplètes) et en présence des valeurs extrêmes. Il serait intéressant de faire le même travail, par ce que ce cas mérite d'être considéré attentivement vu.

BIBLIOGRAPHIE

[1] Balakrishnan, N.& Cohen, A.C. (1991) Order statistics and inference : estimation methods.statist.model.decis.Sci.Academic Press.

[2] Beirlant, J, Goegebeur, Y, Segers, J, & Teugels, J. (2006). Statistics of extrêmes : theory and tpplications. John Wiley.

[3] Beirlant, J., & Guillou, A. (2001). Pareto index estimation under moderate right censoring. Scand. Actuar. J., 111 125.

[4] Beirlant, J., Guillou, A., Dierckx, G., & Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. Extremes, 10(3), 151 174.

[5] Beirlant, J., Teugels, J. L., & Vynckier, P. (1996). Practical analysis of extreme values. Leuven University Press

[6] Cláudia, N , and Maria, J. (2019) . Forecasting and assessing risk of individual electricity peaks : mathematics of planet earth, Springer, International Publishing.

[7] Collett, D. (2015). Modelling survival data in medical research. Third Edition. CRC press. aylor & Francis Group, A chapman & hall book.

[8] Csörgő, S., Mason, D.M. (1985). Central limit theorems for sums of extreme values.Mathematical Proceedings of the Cambridge Philosophical Society, 98(3), 547 − 558.

[9] David, H. A., & Nagaraja, H. N. (2004). Order statistics. Encyclopedia of statistical sciences.

[10] Davis, R., Resnick, S. (1984). Tail estimates motivated by extreme value theory. The Annals of Statistics, 12(4), 1467 − 1487.

[11] Deheuvels, P., Häeusler, E.,& Mason, D.M. (1988). Almost sure convergence of the Hill estimator.Mathematical Proceedings of the Cambridge Philosophical Society, 104(02), 371 − 381

[12] Dekkers, A. L., Einmahl, J. H., & De Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. Ann. Statist., 18331855.

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

[14] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). Modelling extremal events for insurance and finance, springer-verlag, berlin.

[15] Fisher R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Cambridge Philos. Soc., 24(02), 180 − 190

[16] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. Ann. Math., 423 453

[17] Haeusler, E., Teugels, J.L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. The Annals of Statistics, 13(2), 743 − 756.

[18] de Haan, L. & Ferreira, A. (2006). Extreme value theory : an introduction. Springer-Verlag, New York.

[19] Harrouche, L. (2018). Analyse statistique des modèles de survie. Mémoire de Master de probabilités et statistiques.Université Mouloud Mammeri de Tizi-Ouzou. Algériae.

[20] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. Ann. Statist., 3(5), 1163 1174

[21] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quarterly J. R. Methodol.Soc., 81(348), $158 - 171$

[22] Kaplan, E. L., & Meier, P. (1958). Non parametric estimator from incomplete observation, J. Amer. Statist. Assoc. 53.

[23] Mason, D. M. (1982). Laws of large numbers for sums of extreme values. The Annals of Probability, $754 - 764$.

[24] Ndao, P. (2016). Modélisation de Valeurs extrêmes conditionnelles en présence de censure, these de doctorat de probabilités et statistiques Université Gaston Berger

[25] Necir, A. (2006). A functional law of the iterated logarithm for kernel-type estimators of the tail index. Journal of statistical planning and inference, $136(3)$, $780 - 802$.

[26] Peng, L. (2001). Estimating the mean of a heavy tailed distribution. Statistics & Probability Letters, $52(3)$, $255 - 264$.

[27] Pierre,B. (1988). Introduction aux probabilités modélisation des phénoménes aléatoires Springer.

[28] Pickands III, J. (1975). Statistical inference using extreme order statistics. Ann. Statist., 119 131

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

[30] Resnick, S. (1987). Extreme values, regular variation and point processes. Springer Series in Operations Research and Financial Engineering.

[31] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[32] Soltane, L., Meraghni, D., & Necir, A. (2015). Estimating the mean of a heavy-tailed distribution under random censoring. arXiv preprint arXiv :1507.03178.

[33] Soltane, L. (2017). Analyse des valeurs extrêmes en présence de censure, these de doctorat de probabilités et statistiques Université Mohamed Khider Biskra, Algériae

[34] Stute, W. (1995). The central limit theorem under random censorship. The Annals of Statistics, $422 - 439$.

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

[36] Zouadi, N., & Saidi, G. (2018). Estimation De L'indice Des valeurs extrêmes en présence des données censurées étude de Cas : Les Durées De Chômage En Algérie.

Annexe B : Abréviations et Notations

Les différentes abréVariations et notation utulisées tout au long de cette thèse sont expliquées ci-dessous.

fdr ou fd ou fdc	:	La fonction de repartition.
iid	:	Indépendantes et identiquement distribué.
IVE	:	Indice des valeurs extrêmes.
TEV	:	Theorie des valeurs extremes.
TCL	:	Théorème Central Limite.
v.a	:	variable aléatoire.
$(\mathbb{R}, \mathcal{A}, P)$	:	Espace probabilité.
$:=$	:	Egalité par défnition.
$\mathbb{1}_A$	:	Fonction indicatrice de l'ensemble A.
Δ_j	:	Indicateur de censure.
F	:	Fonction de répartition.
F_n	:	Fonction de répartition empirique.
F^{-1}	:	Inverse généralisé de F.
$\mathcal{H}_\gamma$	:	Famille de la loi de valeurs extrêmes généralisée.

$\mathcal{DA}(.)$	:	Domaine d'attraction.
$X_{n:n}$	:	Maximum de $X_1, ..., X_n$.
$X_{1:n}$	:	Minimum de $X_1, ..., X_n$.
$X_{j:n}$	:	$j^{ème}$ statistique d'ordre.
x^F	:	Point terminal supérieur.
x_F	:	Point terminal inférieur.
$\overset{\boldsymbol{p.s}}{\rightarrow}$	:	Converge presque sûrement.
$\overset{\boldsymbol{p}}{\rightarrow}$	:	Converge en probabilité.
$\overset{\boldsymbol{D}}{\rightarrow}$	:	Converge en distribution.
$X_1, ..., X_n$	:	Une suite de n v.a.
Q	:	Fonction du quantile.
Q_n	:	Fonction quantile empirique.
$\overline{F}$	:	Fonction de queue ou Fonction de survie.
$\overline{F}_n^{KM}$	:	L'estimateur de Kaplan-Meier.
Λ	:	Loi de Gumbel.
Φ	:	Loi de Fréchet.
Ψ	:	Loi de Weibull.
$\widehat{\gamma}^p$	:	Estimateur de Pickands.
$\widehat{\gamma}^H$	:	Estimateur de Hill.
$\widehat{\gamma}^M$	:	Estimateur des Moments.

مُلَخَّص

الوسط الحسابي μ هو معلمة لموقع التوزيع، كما أنها مهمة جدا لوصف قيم التوزيع. لسوء الحظ، هي لا تكفي لوصف البيانات، ولهذا نحتاج إلى معلمة إضافية تسمى "التباين". التباين إذن هو اللحظة المركزية من الدرجة الثانية للتوزيع وهو يقيس تشتت X حول μ. الملاحظة الرئيسية لـ μ هي انه لا يمكن أن يوجد دائما و ذلك وفقا لاستقرار α، $\alpha = 1/\gamma$، حيث γ هو مؤشر القيم المتطرفة.

هدفنا الرئيسي من هذه الأطروحة هو تقدير μ، ولهذا نعطي أولا التقدير الكلاسيكي $\bar{X}$ و شبه البراميتري القائم على أساس القيم المتطرفة الذي اقترحه Peng (2001) في إطار التقدير ببيانات مكتملة. ثانيا تم إجراء تجميع و تلخيص لتقدير Stute (1995) و تقدير Soltane et al. (2015) استنادا على نتائج Kaplan-Meier (1958) و القيم المتطرفة في ظل وجود بيانات غير مكتملة.

الكلمات المفتاحية: تحليل البقاء، الرقابة العشوائية، مقدر هيل، مقدر كابلان ماير، تقدير الوسط الحسابي، التقارب الطبيعي.

Résumé

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

Notre objectif principal de ce mémoire est l'estimation du μ, pour cela on donne dans un premier temps, l'estimation classique le $\bar{X}$ et semi paramétrique basé sur la TVE qui a été proposé par Peng (2001) dans le cadre de l'estimation sans censure. Dans un second temps, on effectue une synthèse sur l'estimation de Stute (1995) et l'estimation de Soltane et al. (2015) en se basant sur les résultats de Kaplan-Meier (1958) et la TVE dans le cadre de l'estimation en présence de censure.

Les mots clés : Analyse de survie, Censure aléatoire, Estimateur de Hill, Estimateur de Kaplan-Meier, Estimation de la moyenne, Normalité asymptotique**.**

Abstract

The mean μ is a localization parameter of a distribution, it is also very important in the synthesis of the distribution values. Unfortunately, it is not sufficient to describe the data and for that we need an additional parameter called “the variance”. The variance is therefore the second-order centered moment of the distribution and a measure of the dispersion of X around the μ. The main remark about μ is that it cannot always exist depending on the stability of α, $\alpha = 1/\gamma$ where γ is the EVI.

Our main objective of this thesis is the estimate of μ, for that at first, give the classical estimate of the $\bar{X}$ and semi-parametric based on the EVT which was proposed by Peng (2001) as part of uncensored estimation Secondly, a synthesis is carried out on the estimate of Stute (1995) and the estimate of Soltane and al. (2015) based on the results of Kaplan-Meier (1958) and EVT in the context of the estimation in the presence of censorship.

Key words: Survival analysis, Random censuring, Hill estimator, Kaplan-Meie estimator, Estimating the mean, Asymptotic normality.