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

Hani Zerroukhi

Thème

Estimation de l'Indice des Valeurs Extrêmes sous Censure

Membres du Comité d'Examen

Dr.	**Sana Benamer**	UMKB	Président
Dr.	**Louiza Soltane**	UMKB	Encadreur
Dr.	**Ghozlene Benbraika**	UMKB	Examinateur

Juin 2019

À *mes chers Parents,*
À mes soeurs et mes frères,
À ma famille et mes amis.

REMERCIEMENTS

$\mathcal{J}$e tient à remercier sincèrement et tout puissant $\mathcal{ALLAH}$ qui m'aide et ma donne la santé , la patience et le courage durant ces langues années d'étude et la force pour finir ce travail.

$\mathcal{E}$n second lieu, je tiens à remercier mon encadreur **$\mathcal{L}$ouiza $\mathcal{S}$oltane**, *pour ses précieux conseils et son aide durant toute la période du travail. Mes vifs remerciements vont également aux membres du jury :* **$\mathcal{S}$ana $\mathcal{B}$enamer** *et* **$\mathcal{G}$hozlene $\mathcal{B}$enbraika** *pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.*

$\mathcal{M}$es remerciements s'étendent également à tous mes enseignants durant les années d'études.

$\mathcal{M}$a famille et mes amis qui par leurs prières et leurs encouragements, j'ai pu surmonter tous les obstacles. Enfin, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

$\mathcal{M}$erci ...

$\mathcal{Z}$. $\mathcal{H}$ani

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

L'analyse de survie est un domaine des statistiques qui trouve sa place dans tous les champs d'applications où l'on étudie la survenue d'un évènement. L'objectif de cette analyse réside dans l'analyse du délai de survie d'un évènement dans un ou plusieurs groupes d'individués. Dans le domaine biomédical, par exemple, plusieurs évènements sont intéressants à étudier :

- Le développement d'une maladie.
- La réponse à un traitement donné.
- La rechute d'une maladie ou décès.

Une des caractéristiques des données de survie est l'existence d'observation incomplète. La censure fait partie du processus générant ce type de donnée.

La théorie des valeurs extrêmes (TVE) représante un outil approprie permettant d'extrapoler le comportement des queues de distrubutions à partir des plus grands (ou plus petites) valeurs observées. Sur le plan statistique, l'étude permet de fournir des outils probabilistes et statistiques qui permettent de modéliser pour prévoir l'occurrence des évènements rares. Ces événements rares sont des phénomènes aléatoires qui ont une faible probabilité d'apparition et sont rencontrés dans plusieurs domaines devenus secteurs d'application de la théorie des valeurs extrêmes. On peut citer entre autres :

- La météorologie : pour l'étude de vitesse de vent, extrême pluviométrique, des températures, ...
- En hydrologie : pour l'étude la probabilité que la hauteur d'eau d'un fleuve dépasse un certain seuil (Voir **Guillou et al** [4]).
- La finance (Marchés financiers) : Pour l'étude quantitative des boums et Krachs boursiers, qui se traduisent par de fortes variations des cours financiers (voir **Login** [22]).
- L'assurance et la Réassurance : Pour l'étude des risques graves ou rares afin d'avoir une certaine stabilité des indicateurs qui traduit une bonne adéquation entre la sinistralité et la tarification (voir **Embrechts et al** [12]).

La théorie des valeurs extrêmes, fondée sur des résultats de la théorie des probabilités, offre un cadre mathématique rigoureux pour l'estimation des probabilités d'événements rares. Il s'agit fondamentalement de modéliser un phénomène aléatoire, en s'intéressant principalement aux quantiles extrêmes et à la queue de distribution souvent modélisée par un indice appelé indice des valeurs extrêmes.

L'utilisation des lois des valeurs extrêmes repose sur des propriétés des statistiques d'ordre et sur des méthodes d'extrapolation. Plus précisément, elle repose sur la convergence des maxima ou des minima des variables aléatoires indépendantes et identiquement distribuées, convenablement renormalisées. C'est dire donc qu'on étudie le comportement asymptotique des lois des extrêmes. Ces lois, appelées lois des valeurs extrêmes sont bien connues et elles sont de trois types : Fréchet, Gubmel, et Weibull.

Les lois des valeurs extrêmes, lorsqu'elle existe, sont indexées par l'indice des valeurs extrêmes (IVE). La connaissance de cet indice est un élément important car il contrôle la "lourdeur" de la queue de distribution. Ainsi de nombreux estimateurs de l'indice des valeurs extrêmes ont été proposé dans la littérature (Estimateur d'Hill, Pikands, ...). L'estimation de cet indice dépend largement de nombre de statistiques d'ordre extrêmes observées. Ce nombre détermine les valeurs qui sont réellement extrêmes.En particulier la théorie des valeurs extrêmes dans le cas censurée est un sujet de recherche et étude de nombreux scientifiques et dévelloppement des nombreuses estimations tel que Reiss et Thomas [28], Einmahl et all [11], Brahimi et all [5]...etc.

Ce mémoire est composé de deux chapitres :

Le premier Chapitre se regroupe en trois sections. Dans la section 1.1, on commençe par quelques rappels et définitions sur la fonction de survie, la fdr, les fonctions empiriques de répartition et de survie et les trois fonction de densité, risque et risque cumulé. Plus dans la section 1.2, on présente un seul cas de données incomplètes : données de censure et ses types (droite, gauche et par intervale). Dans la section 1.3, on présente une synthèse des principaux estimateurs, dont les plus célèbres estimateurs non-paramétriques sont l'estimateur de Kaplan-Meier de fonction de survie et l'estimateur de Nelson-Aalen pour la fonction de hasard cumulée.

Le deuxième Chapitre se compose de quatre Sections. La section 2.1 se compose de deux partie, la première partie, on parle sur les lois des grands nombres et les propriétés asymptotiques de la somme des va's iid (TCL), et la deuxième partie on parle sur les staristiques d'ordres, qui est trés utile en théorie des valeurs extrêmes. Ensuite, la section 2.2 donne des résultats limites sur la distrubution de maximum de l'échantillon. La section 2.3, on discute également la notion des domaines d'attraction d'une distribution selon le paramètre de l'indice de queue. Puis la section 2.4 , se compose deux parties, on donne dans la première partie les estimateurs classiques de l'indice de valeurs extrêmes tels l'estimateur de Hill, de Pickands et des Moments dans le cadre de sans censure.le dernier partie parle sur l'estimation de l'indice des valeurs extrêmes en présence de données censurées aléatoirement à droite.

CHAPITRE 1

GÉNÉRALITÉS SUR L'ANALYSE DE SURVIE

Analyse de données de survie : étude l'apparition d'un évènement au cours du temps, qui n'est pas forcément la mort. Il peut s'agir par exemple :

- Temps de survie aprés le diagnostic d'un cancer du sein.
- Durée de séropositivité sans symptôme de patients infectés par le VIH.
- Durée d'un épisode de chômage, durée de vie d'une entreprise, durée séparant deux sinistres, instant d'un défaut de paiement, durée avant la ruine, ...
- Durée de vie d'une ampoule, d'un pièce mécanique, . . .

Dans tous les domaines où l'on cherche à mesurer l'instant d'arrivée d'un évènement aléatoire (panne, mort, maladie, chômage,....). Dans les trois premiers exemples, la notion de durée sera appelée "durée de survie" ou "durée de vie" (voir **Vivian** [**35**], page 14) et dans le dernier exemple on l'appelle "analyse de fiabilité" (voir **Soltane** [**32**]**,** page 1) celles-ci sont les plus communément utilisées dans la littérature statistique, puisque c'est précisément dans le domaine médical que les avancées méthodologiques liées à ces variables ont été développées en premier.

On commence dans ce chapitre par quelques rappels et définitions couramment utilisées dans les études.

1.1 Rappels et Définitions

On désigne par X une variable aléatoire (va) positive définie sur un espace probalisé $(\Omega, \mathcal{F}, \mathbb{P})$, et représentant une durée j'usqu'à un événment d'intérêt (voir **Jean-François**[**19**], page 13). Soit P_X la loi de X,. alors sa loi de probabilité peut être définie par l'une des fonctions suivantes :

1.1.1 Fonction de Survie

Définition 1.1.1 (Fonction de survie)
La fonction de survie qu'on note par $S(t)$ ou $\overline{F}(t)$ est définie sur $\mathbb{R}_+$ par

$$S(t) = \overline{F}(t) := \mathbb{P}(X > t). \tag{1.1}$$

Pour t fixé c'est la probabilités de survivre jusqu'a l'instant t.

Remarque 1.1.1
La fonction de survie d'une va X est décroissante monotone continue à gauche et vérifie

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

1.1.2 Fonction de Répartition

Définition 1.1.2 (Fonction de répartition)
La fonction de répartition (fdr ou fd) de X ou de sa loi P_X est la fonction sur $\mathbb{R}_+$ définie par

$$\begin{array}{rccl} F: & \mathbb{R}_+ & \longrightarrow & [0,1] \\ & t & \longrightarrow & \mathbb{P}\left(X \leq t\right), \end{array}$$

Pour t fixé, c'est la probabilité de mourir avant l'instant t.

Remarque 1.1.2

1. *F est aussi appellé fonction de distribution et $S(t)$ est la queue de distribution.*
2. *On peut définir $F(t) := 1 - S(t)$, pour tout $t \geq 0$,*

alors que $\overline{F}$ est une fonction croissante monotone continue à droite telle que

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

1.1.3 Fonctions Empiriques de Répartition et de Survie

Définition 1.1.3 (*Fonctions empiriques de répartition et de survie*)
Soit $X_1, ..., X_n$ un échantillon de taille $n \geq 1$ d'une va positive X de fdr F et de fonction de survie S. Les fonctions empiriques de répartition et de survie, F_n et S_n sont respectivement définies par

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \leq t\right\}, \quad \forall\ t \geq 0, \tag{1.2}$$

et

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

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

On peut écrire (1.2) et (1.3) en termes des valeurs des statistiques d'ordre[1] comme suite$F$

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

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

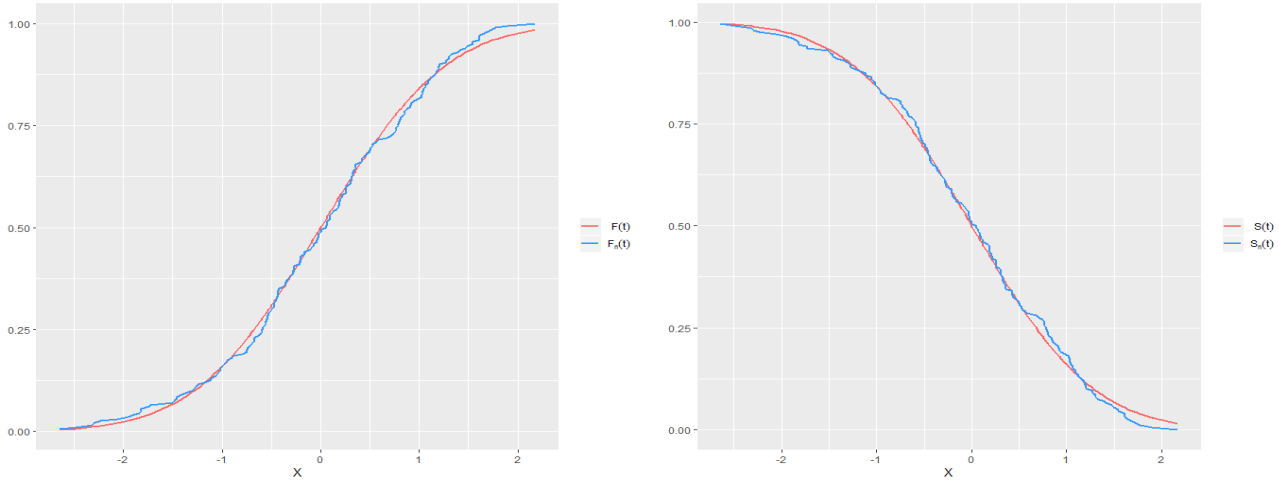

FIG. 1.1 – Fonction Empirique de Répartition (Droite) et de Survie (Gauche) d'un Echantillon Gaussien Standard de Taille 200.

Proposition 1.1.1 (Propriétés asymptotiques de F_n)

1. *$F_n(t)$ c'est un éstimateur sans bians de $F(t)$, c-à-d que*

$$E\left[F_n(t)\right] = F(t), \quad \text{quand } n \to \infty.$$

2. *La convergence de F_n vers F est presque sûrement uniforme, c-à-d que :*

$$\sup_t \left|F_n(x) - F(x)\right| \xrightarrow{p.s} 0 \text{ quand } n \longrightarrow \infty. \tag{1.4}$$

La convergence (1.4) est connue sous le nom de théorème de Glivenko-Cantelli. Il est l'un des résultats fondamentaux en statistiques non-paramétriques. La preuve du résultat (1.4) peut être trouvés dans tout manuel standard de la théorie des probabilités comme (Pierre [36], page 229).

[1] Une brève étude sur les statistiques d'ordre est donnée dans la sous-section 2.1.3.

Définition 1.1.4 (Fonction de quantile)
Pour tout $0 < s < 1$, *la fonction de quantile associe à* F *est définie par*

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

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

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

Définition 1.1.5 (*Quantile empirique*)
La fonction de quantile empirique de l'échantillon $X_1, ..., X_n$ *est définie par*

$$Q_n(s) = F_n^{-1}(s) := \inf\{t : F_n(t) \geq s\}, \quad 0 < s < 1, \tag{1.5}$$

où

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

Remarque 1.1.3

1. *Le quantile d'ordre* p $(p \in]0.1[)$ *est définie par* $x_p = F^{-1}(p)$.
2. *Une fonction parfois appelée fonction de quantile de queue, notée par* U, *est définie par* $U(t) := F^{-1}(1 - 1/t) = (1/\overline{F})^{-1}(t), \quad t \geq 1,$
   et la fonction empirique correspondante est

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

1.1.4 Fonction de Densité

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

Définition 1.1.6 (*Fonction de densité*)
Si F *admet une dérivée par rapport à la mesure de Lebesgue sur* $\mathbb{R}_+$, *la fonction de densité de probabilité existe, elle est définie pour tout* $t \geq 0$, *par*

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} = \lim_{dx \longrightarrow \infty} \frac{\mathbb{P}(t \leq X \leq t + dt)}{dt}. \tag{1.6}$$

Pour t fixée, la densité de probabilité caractérise la probabilité de mourir dans un petit intervalle de temps aprés l'instant t.

1.1.5 Fonctions de Risque et de Risque Cumulé

Appelée selon les domaines d'application : *"taux instantané de défaillance"*, *"taux de risque"*, *"le taux de hasad"* ou encor, *quotient de mortalité"*. Pour plus de détails, on renvoie au livre de Saporta [31], page 19.

Définition 1.1.7 (*Fonction de risque*)
Si X est une variable continue positive représentant une durée. La fonction de risque, notée par $h(t)$, est définie par

$$h(t) = \lim_{dx \longrightarrow 0} \frac{\mathbb{P}(t \leq X < t + dx / X \geq t)}{dx} = \frac{f(t)}{S(t)}, \tag{1.7}$$

et la fonction de risque cumulé, qu'on note par $H(t)$, c'est l'intégral de fonction du risque

$$H(t) := \int_0^t h(u)du = \int_0^t \frac{dF(u)}{S(u)}. \tag{1.8}$$

$h(t)$ caractérise la loi de X car on peut retrouver $F(t)$ à partir de $h(t)$

$$h(t) = -\frac{d}{dt} \ln S(t). \tag{1.9}$$

Il est facile de deduire de (1.9) que

$$F(t) = 1 - \exp\left\{ -\int_0^t \frac{dF(u)}{S(u)} \right\} = 1 - \exp(-H(t)). \tag{1.10}$$

Cette égalité présente une caractérisation de distribution et une fonction de survie par l'intermédiaire de fonction de risque. Toutes ces fonctions (1.1) , (1.6), (1.7) et (1.8) sont donc liées entre elles. En d'autres termes, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également défiies. Pour plus d'illustrer sur les relations d'équivalence entre ces fonctions précédentes on réfère à, par exemple, (Lee et Wang [21], Exemple 2.2, page 17), (Wienke [37], Exemple 2.1, page 17).et (Gilbert [15] page 21).

1.2 Censures

Dans l'analyse de survie les données ne sont pas toujours complètement observées, parce que, pour certains individus l'évènement du début et /ou de fin n'est pas observé, c'est-à-dire privées d'une partie de l'information. Dans ce cas les données sont censurées, Il n'est pas rare, mais elles sont plutôt incomplètes.

Définition 1.2.1 (variable de censure)
La variable de censure Y est définie par la non-observation de l'événement étudié. Si au lieu d'observer X, on observe Y, et que l'on sait que $X > Y$ (respectivement $X < Y$, $Y_1 < X < Y_2$), on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).

Pour un individu donné i, on va considérer
- son temps de survie X_i, de fonction de répartition F.
- son temps de censure Y_i, de fonction de répartition G
- la durée réellement observée Z_i de fonction de répartition N.

1.2.1 Types de Censures

Les observations peuvent présenter différents types de censure
- **Censure à droite :** Une durée de vie est dite censure à droite si l'individu n'a pas connu l'évènement d'intérêt à sa dernière visite.
- **Censure à gauche :** Une durée de vie est dite censure à gauche si l'individu a déjà subi l'évènement avant qu'il ne soit observé.
- **Censure double :** La censure double (ou mixte) ce type de censure c'est un mélange entre les deux censures, la censure à droite et la censure à gauche, dans le même échantillon.
- **Censure par intervalle :** Comme son nom indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt.

L'expérience elle-même peut engendrer cette censure
- Censure de type I : fixée
- Censure de type II : attente
- Censure de type III : aléatoire

Pour des détails complets sur les types de censure on réfère aux livres de (Pierre [30], page 7) et (**Vivian [35],** page 14). Dans ce travail, on s'intéresse uniquement au modèle de censure à droite du type aléatoire. Celui-ci correspond à un modèle fréquemment utilisé en pratique (voir aussi **Soltane [32]**, page 13).

1.2.2 Modèle de Censure Aléatoire à Droite

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

$$Z_i := \min(X_i, Y_i) \quad \text{et} \quad \delta_i := \mathbb{1}\{X_i \leq Y_i\} \quad \text{pour } i = 1, ...n, \tag{1.11}$$

où δ_i l'indicateur de censure, qui détermine si X a été censuré ou non :
- si $\delta_i = 1$, la durée d'intérêt est observée $(Z_i = X_i)$.
- si $\delta_i = 0$, elle est censurée $(Z_i = Y_i)$. On observe des durées incomplètes.

Exemple 1.2.1
On considére une étude relative à la durée de survie de patients soumis à un traitement particulier. L'évènement d'intérêt est la mort de la patient. Tous les individus sont suivis pendant les 10 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. On peut citer certaines causes entraînant la censure aléatoire :

1. *"Perdu de vue" : le patient peut décider de se faire soigner ailleurs et on ne le revoit plus.*
2. *Arrêt du traitement : suite à des effets secondaires le traitement est arrêté.*
3. *Fin de l'étude : l'étude se termine et certains patients sont toujours vivants.*

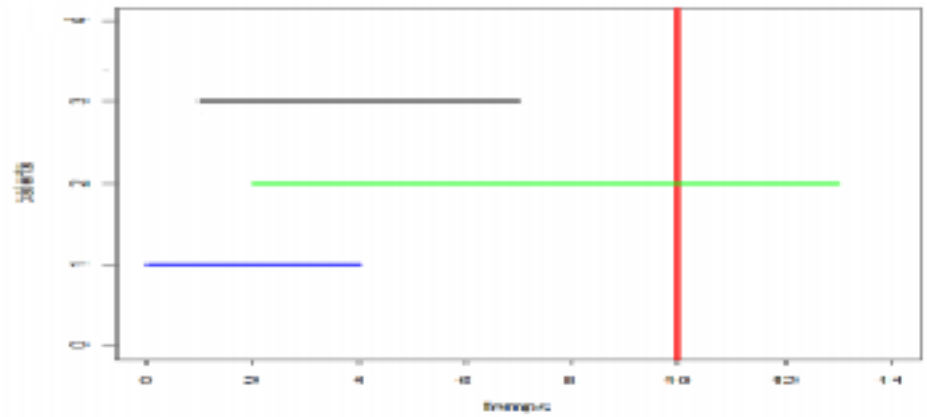

FIG. 1.2 – Schema Represantant les Cas de Censure Aléatoire (Source [15]) .

Dans la Figure 1.2 on représente le suivi de trois patients. Le premier patient est décédé 4 semaines après le début du traitement. Il s'agit d'une observation non censurée. Le deuxième patient est vivant au terme des 14 semaines d'observation. L'information de ce patient n'est pas connue lorsque la constitution de la base de données est arrêtée (en $t = 10$). il est donc censuré. Quand au troisième patient, il a été perdu de vue à $t = 7$, donc il est censuré à $t = 7$.

Remarque 1.2.1
Les données censurées ne sont pas le type unique de données incomplètes. L'autre cas classique de données incomplètes est celui des données dites tronquées. Le phénomène de troncature est très différent de la censure.

1.3 Estimation non Paramétrique

1.3.1 Estimateur de Kaplan-Meier

Soit $\{(Z_i, \delta_i)\,,\ 1 \leq i \leq n\}$ l'échantillon réellement observé défini par (1.11) . Lorsque les données sont censurées, il est impossible d'utiliser l'estimateur (1.3) puisqu'il fait intervenir des

quantités non observées. Dans ce cas, Kaplan et Merie [20] en 1958 ils ont proposé un estimateur, appelé estimateur de Kaplan-Meier ou estimateur du produit limite car il s'obtient comme la limite d'un produit (voir **Gassom Z [14]** L'idée de la construction de cet estimateur est donnée par la probabilité conditionnelle. Soient $0 < t^{"} < t^{'} < t$, on a

$$\begin{aligned}\mathbb{P}(X > t) &= \mathbb{P}(X > t^{'}, X > t) \\ &= \mathbb{P}\left(X > t \backslash X > t^{'}\right) \times \mathbb{P}\left(X > t^{'}\right) \\ &= \mathbb{P}\left(X > t \backslash X > t^{'}\right) \times \mathbb{P}\left(X > t^{'} \backslash X > t^{"}\right) \times \mathbb{P}\left(X > t^{"}\right).\end{aligned}$$

En considérant les temps dévènements (décès et censure) distincts $Z_{i,n} \quad (i = 1, ..., n)$ rangés par ordre croissante, on obtient

$$\mathbb{P}(X > Z_i, n) := \prod_{k=1}^{i} \mathbb{P}\left(X > Z_k, n \backslash X > Z_{k-1,n}\right),$$

avec $Z_{0,n} = 0$. Si on note par n_i le nombre d'individus à risque de subir l'évènement (mourir) juste avant le temps $Z_{i,n}$ et par d_i le nombre de décès en $Z_{i,n}$, alors la probabilité p_i de mourir dans l'intervalle $]Z_{i-1,n}, Z_{i,n}]$ sachant que l'on était vivant en $Z_{i-1,n}$, c'est à dire $p_i = \mathbb{P}(X \leq Z_{i,n} \backslash X > Z_{i-1,n})$, peut être éstimée par

$$\hat{p}_i := \frac{d_i}{n_i}.$$

Comme les temps des évènements sont supposés distincts, on a $d_i = 0$ en cas de censure en $Z_{i,n}$ et $d_i = 1$ en cas de décès en $Z_{i,n}$. Si on désigne par $\delta_{[i,n]}$ le concomitant de la $i^{ème}$ satatistique d'ordre $Z_{i,n}$ défini par $\delta_{[i,n]} = \delta_j$ si $Z_{i,n} = Z_j$, alors on a $\delta_{[i,n]} = 0$ en cas de censure en $Z_{i,n}$ et $\delta_{[i,n]} = 1$ en cas de décès en $Z_{i,n}$, Ainsi, l'estimateur de Kaplan-Meier, pour $t < Z_{n:n}$, est défini par

$$\hat{S}^{KM}(t) = \widehat{\overline{F}}_n(t) := \prod_{i:Z_{i,n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n_i}\right) = \prod_{i:Z_{i,n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n - i + 1}\right),$$

où $Z_{1:n} \leq ... \leq Z_{n:n}$ les statistiques d'ordre associées à $Z_1, ..., Z_n$.

Remarque 1.3.1
Dans le cas où il y a des ex-aequo, si ces ex-aequo sont tous des morts, la seule différence tient à ce que d_i n'est plus égal à 1 mais au nombre des morts et l'estimateur de Kaplan-Meier devient

$$\hat{S}^{KM}(t) = \prod_{i:Z_{i,n} \leq t} (1 - \frac{d_i}{n_i}) = \prod_{i:Z_{i,n} \leq t} (1 - \hat{p}_i).$$

Remarque 1.3.2

1. *L'estimateur de kaplein-Meier peut être construit comme estimateur du maximum de vrisemblance.*
2. *$\hat{S}^{KM}(t)$ est une fonction en escalier décroissante, continue à droite.*

1.3.2 Estimateur de Nelson-Aalen

L'estimateur de Kaplan-Meier est un estimateur non-paramétrique de la fonction de survie..L'estimateur de Nelso-Alen aussi est un estimateur non-parametrique pour la fonction de hasard cumulée,.Il est intoduit par Nelson [24] en 1972 et Alen [1] en 1978. D'aprés la proprieté de l'independance entre X et Y, on peut écrit $N(t)$ comme suit :

$$N(t) := 1 - (1 - F(t))(1 - G(t)) = N^{(0)}(t) + N^{(1)}(t), \tag{1.12}$$

où

$$N^{(0)}(t) = \mathbb{P}(Z \leq t; \delta = 0) = \int_0^t \overline{F}(x) dG(x) \quad \text{et} \quad N^{(1)}(t) = \mathbb{P}(Z \leq t; \delta = 1) = \int_0^t \overline{G}(x) dF(x).$$

La fonction de hasard cumulée (1.8) peut s'exprimer de la forme suivante :

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

Définition 1.3.1 (Estimateur de Nelson-Alen)
L'estimateur de Nelson-Aalen $H_n(t)$ de H basé sur l'échantillion $\{(Z_i, \delta_i),\ 1 \leq i \leq n\}$ donné par :

$$H_n(t) = \int_0^t \frac{dN_n^{(1)}(x)}{\overline{N}(x)} := \begin{cases} \sum\limits_{Z_{i,n} \leq t}^{n} \dfrac{\delta_{[i:n]}}{n - i + 1} & si\ t < Z_{i,n}, \\ 1 & si\ t \geq Z_{i,n}, \end{cases}$$

où

$$N_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\} \quad et \quad N_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \mathbb{1}\{Z_i \leq t\},$$

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

Remarque 1.3.3

En remplaçant $H(t)$ part $H_n(t)$ dans (1.10) on obtient un nouvel estimateur de la fonction de survie S :

$$\hat{S}_n^{NA}(t) := \begin{cases} \prod\limits_{Z_{i:n} < t} \exp\left\{ -\dfrac{\delta_{[i:n]}}{n - i + 1} \right\} & \text{si } t < Z_{n:n}, \\ 0 & \text{sinon.} \end{cases}$$

CHAPITRE 2

THÉORIE DES VALEURS EXTRÊMES

Depuis quelques années, la théorie des valeurs extrêmes a reçu beaucoup d'attention de nombreux statisticiens, ingénieurs et scientifiques tant le champ d'application qu'elle touche est vaste : Hydrologie, biologie, ingénierie, météorologie, gestion de l'environnement, finance, assurance, ...etc.

L'objectif essentiel de ce chapitre est de présenter les définitions de base et les résultats principaux sur la théorie des valeurs extrêmes (TVE) dans le cas unidimensionnel. Pour plus de détails sur ce thème voir (Beirlant et al. [3], 2006)) et.(Reiss et Thomas [28], 2007).

2.1 Définitions de Base

Définition 2.1.1 (Somme et moyenne arithmétique)
Soit $X_1, ..., X_n$ une suite de va's iid. Pour tout entier $n \geq 1$, on définit la somme partielle et la moyenne arithmétique correspondantes respectivement par

$$S_n := \sum_{i=1}^{n} X_i \quad et \quad \overline{X}_n := \frac{S_n}{n}.$$

$\overline{X}_n$ s'appelle la moyenne empirique.

2.1.1 Lois des Grands Nombres

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

Théoreme 2.1.1 (Lois des Grands Nombres)
Si $X_1, ..., X_n$ un échantillon provenant d'une va X tel que $\mu := E[X] < \infty$, alors

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

2.1.2 Théorème Central Limite

L'étude de sommes de variables indépendantes et de même loi joue un rôle capital en statistique. Le théorème suivant connu sous le nom de Théorème Centrale Limite (TCL) établit la convergence en loi vers la loi.de Gauss.

Théoreme 2.1.2 (Théorème Central Limite)
Soit $X_1, ..., X_n$ est une suite de va's iid de moyenne μ et de variance σ^2 définie, alors

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

2.1.3 Statistique d'Ordre

Définition 2.1.2 (Statistique d'ordre)
Soit $X_1, ..., X_n$ une suite de va' s iid, classée par ordre croissant. On écrit cette suite d'observation sous la notation $X_{i:n}$ tel que

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

où

- $X_{i:n}$ *: la $i^{ème}$ statistique d'orde (statistique d'ordre i) dans un échantillion de taille n.*
- $X_{1:n}$ *: la plus petite valeur observée (où statistique de minimum) avec*

$$X_{1:n} = min(X_{1:n}, ..., X_{n:n})$$

- $X_{n:n}$ *: la plus grand statistique d'ordre (où statistique de maximum) avec*

$$X_{n:n} = max(X_1, X_2, ..., X_n).$$

David [8] et Balakrishnan et Cohn [2] montre que l'expression de la distribution de $X_{i:n}$ est

$$F_{i:n} = \mathbb{P}\left\{X_{i:n} \leq x\right\} = \sum_{r=i}^{n} \binom{n}{r} \left(F\left(x\right)\right)^r \left(1 - F\left(x\right)\right)^{n-r},$$

alors, on déduit que la fonction de densité est de la forme suivante :

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

Pour les statistique d'ordre extrême, on obtient les expressions suivantes :

$$F_{1:n}(x) = \mathbb{P}\{X_{1:n} \leq x\} = 1 - (1 - F(x))^n \quad \text{et} \quad F_{n:n} = \mathbb{P}\{X_{n:n} \leq x\} = (F(x))^n .$$

Les expression de $F_{1:n}$ et $F_{n:n}$ peuvent s'obtenir trés facillement en considéront les relation

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

et

$$\begin{aligned}\{X_{n:n} \leq x\} &\Leftrightarrow \{\max(X_1, ..., X_n) \geq x\} \\ &\Leftrightarrow \overset{n}{\underset{i=1}{\cap}} \{X_i \leq x\} .\end{aligned}$$

En utilisant la proprieté d'indépendance de va's $X_1, ..., X_n$ nous en déduisons que,

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

et

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

Dans la suite de ce travail, on ne présente que les résultats concernant le maximum, puisque les résultats relatifs au minimum se déduisent de l'égalité suivante :

$$\min(X_1, ..., X_n) = -\max(-X_1, ..., -X_n).$$

2.2 Lois des Valeurs Extrêmes

Les distributions à queues lourdes jouent un rôle important dans la théorie des valeurs extremes. Elles ont été acceptées comme des modèles appropriés de divers phénomènes on peut citer dans le cas de montage de grand sinister en assurance les fluctuations des prix en finance, ... etc. Mathématiquement, ce type des distributions est défini ainsi :
Soit X une va de fdr F, donc cette dernière elle est dite distribution à queue lourde, s'il existe un un constant positif γ qui représente l'indice de queue (indice des valeurs extrêmes, IVE) et prend la formule suivante :

$$\overline{F}(x) \sim x^{-1/\gamma} l(x), \text{ pour } x \to \infty, \tag{2.1}$$

où $l(x)$ la fonction à variation lente au voisinage de l'infini, et satisfaite pour tout $x > 0$, la condition suivante :

Définition 2.2.1 (Condition du 1^{er} ordre)
$\overline{F}$ est dite variation régulière à l'infini d'indice $-1/\gamma < 0$, si pour tout $x > 0$, on a

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma}. \tag{2.2}$$

mais, en général cette condition n'est pas suffisante pour étudier les propriétés des estimateurs, en particulier la normaliser asymptotique. Dans ce cas, une condition du second ordre des fonctions à variations régulières est nécessaire en spécifiant le taux de convergence dans l'Équation 2.2. La définition suivante de cette condition vient de de Haan et Ferreira [17], page 48.

Définition 2.2.2 (Condition du $2^{ème}$ ordre)
$\exists\ \rho \leq 0$ et une fonction $A \to 0$ et ne change pas le signe au voisinage de l'infini, tel que

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma}\frac{x^{\rho}-1}{\gamma\rho}. \tag{2.3}$$

La famille de distribution de queue lourde, elle se caractérise par une décroissance lente vers zéro par rapport à la distribution exponentielle, comme le montre la figure suivante :

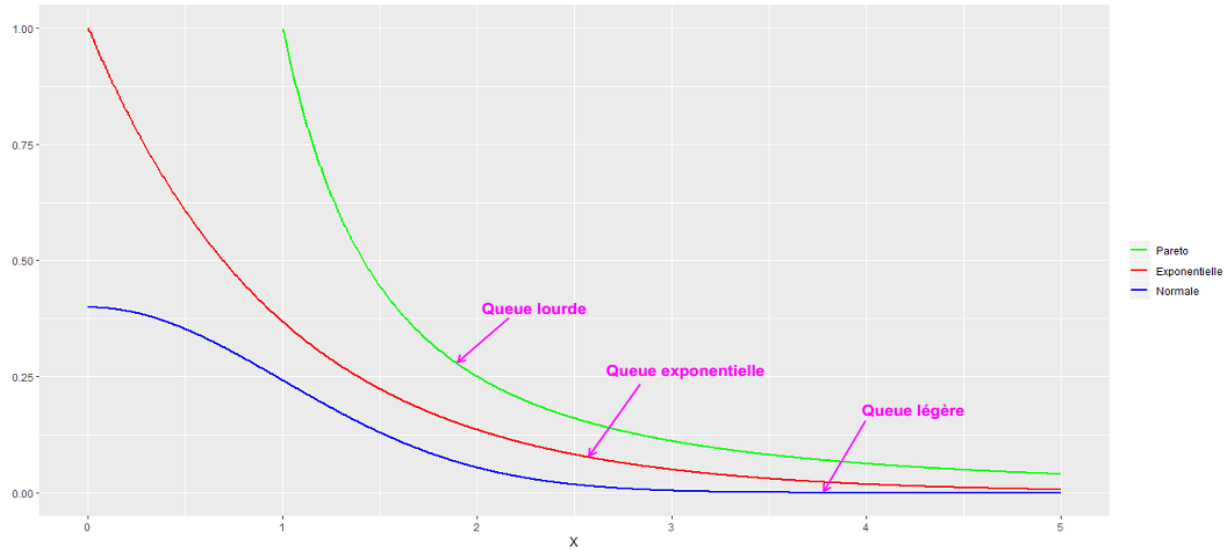

FIG. 2.1 – Comparaison du Comportement de la Queue.

On ici 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. Une autre définition de distribution à queue lourde peut être dérivée dans TVE, cette théorie étudie le comportement asymptotique du maximum et déduire celui la queue de distribution par extrapolation. Le résultat principal de cette théorie est le théorème de Ficher et Tippett.

Théoreme 2.2.1 (Fisher et Tippett (1928), Gnedenko (1943))
Supposons que n va' s X_i, $i = 1, ..., n$ indépendant et de même loi de distribution F et $X_{n:n}$ le maximum de l'échantillon $X_1, ..., X_n$ S'il existe deux constantes a_n et b_n et une distribution ***limite non-dégénéré*** *$\mathcal{G}_\gamma$ telle que :*

$$\lim_{n\longrightarrow\infty} \mathbb{P}\left[\frac{X_{n:n} - b_n}{a_n}\right] = \lim_{n\longrightarrow\infty} F^n(a_n x + b_n) = \mathcal{G}_\gamma(x), \quad \forall x \in \mathbb{R}, \tag{2.4}$$

et la distribution $\mathcal{G}_\gamma(x)$ elle est distribution généralisée des valeurs extrêmes (GEV), le paramètre γ il déja appelle indice des valeurs extrêmes (IVE) et a_n et b_n sont des paramètres de normalisation. Cette distribution elle prend différentes formes suivantes :

$$\mathcal{G}_\gamma(x) = \begin{cases} \exp(-(1-\gamma x)^{-1/\gamma}) & \forall x \in \mathbb{R}, \quad 1+\gamma x \quad si\ \gamma \neq 0 \\ \exp(\exp(-x)) & \forall x \in \mathbb{R}, \quad si\ \gamma = 0. \end{cases}$$

Pour plus de détail sur la démonstration du Théorème 2.2.1 le lecteur pourra se référer aux ouvrages de Resnick [29] et Charpentier et Denuit [10].

2.3 Domaines d'attractions

Selon le singe de paramètre γ on peut distinguer trois domaines d'attractions ($\mathcal{DA}$) :

$$\mathcal{G}_\gamma(x) = \begin{cases} \exp(\exp(-x)) & -\infty < x < +\infty, \quad \text{Gumbel} \\ \begin{cases} 0 & x < 0 \\ \exp(-x^{-1/\gamma}) & x \geq 0, \gamma > 0 \end{cases} & , \quad \text{Fréchet} \\ \begin{cases} \exp(-(-x^{-1/\gamma})) & x < 0, \gamma < 0 \\ 1 & x \geq 0. \end{cases} & , \quad \text{Weibull} \end{cases}$$

Remarque 2.3.1

1. *Pour distinguer les trois distributions en utilisent généralement les notations suivantes : Λ pour la distribution de Gumbel, Φ_γ pour la distribution de Fréchet et Ψ_γ pour la distribution de Weibull.*
2. *Les distributions Λ, Φ et Ψ sont appellé les distributions des valeurs extrêmes et va' s correspondant sont les va's extrémales.*

La représentation des trois fonctions de densité de Λ ,Φ_1 et Ψ_{-1} elle est illustrée dans la Figure 2.2.

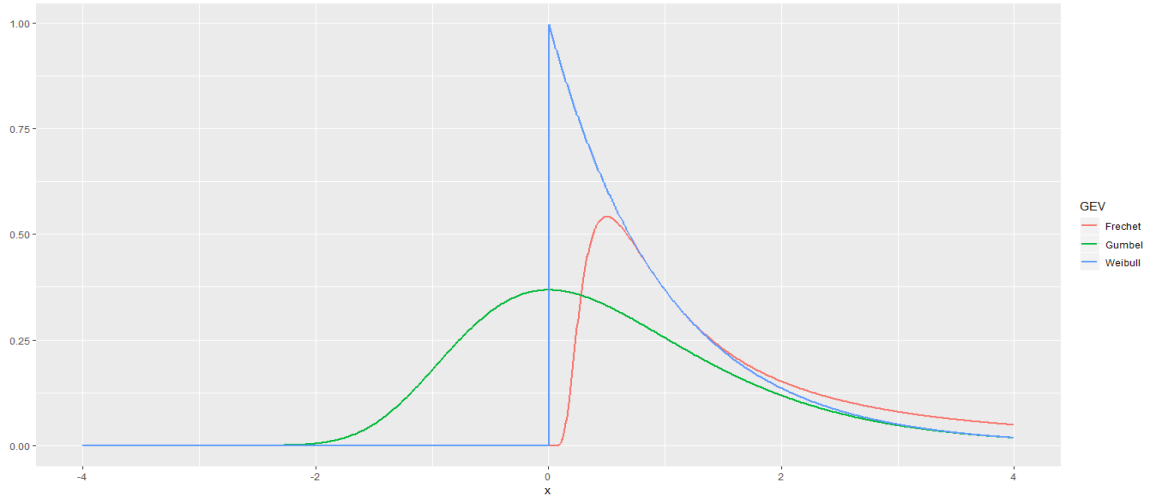

FIG. 2.2 – Densités des Lois des Valeurs Extrêmes.

- Si $\gamma > 0$, F appartient au $\mathcal{DA}$ de **Fréchet**, et l'on note $F \in \mathcal{DA}(\Phi_{1/\gamma})$ et F a un point terminal[1] à droite infinie ($x_F = +\infty$). Il contient toutes les lois dont la fonction de survie décroît comme une fonction puissance. Ce type de distributions elle est dite distribution à queue lourde, ces distributions de ce domaine elles sont beaucoup utilisées en fiabilité mécanique, dans les phénomènes climatiques tels que la météorologie, l'hydrologie, la vitesse du vent enregistrée en continu dans les aéroports et en finance dans les études de risque.Comme exemple de lois appartenant à ce domaine d'attraction on a les lois de Cauchy, de Pareto, du Chi-deux, de Student, de Fréchet, ...etc.
- Si $\gamma = 0$, F appartient au $\mathcal{DA}$ de **Gumbel**, et l'on note $F \in \mathcal{DA}(\Lambda)$,et F a un point terminal à droite x_F peut alors être fini ou non.Ce sont les lois dont la fonction de survie décroît vers zéro à une vitesse exponentielle. Ces distributions sont souvent utilisées pour faire des prévisions dans les événements environnementaux comme le tremblement de terre, l'hydrologie...etc. Ce domaine regroupe les lois Normale, Exponentielle, Log-normale, Gamma, ...etc.
- $\gamma < 0$, F appartient au $\mathcal{DA}$.de **Weibull**, et l'on note $F \in \mathcal{DA}\left(\Psi_{1/\gamma}\right)$ et F a un point terminal à droit fini ($x_F < +\infty$). Ce domaine regroupe toutes les lois dont le point terminal, ce type de distribution sont souvent utilisées pour décrire la résistance mécanique d'un matériau ou encore le temps de fonctionnement d'un appareil électronique ou mécanique. Comme un exemple sur ce domaine on trouve les lois Uniforme, Beta, ...etc.

[1]Le point terminal d'une distribution F définit par :

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

2.4 Estimation de l'Indice des Valeurs Extrême

2.4.1 Estimation de l'IVE en présence de Données Complètes

Dans la littérature de la TVE, il existe plusieurs méthodes et téchniques pour l'estimation de l'IVE, dans cette partie on reste limiter à trois méthodes.

Soit $X_1, ..., X_n$ de va's iid et $X_{1:n} \leq ... \leq X_{n:n}$ les statistique d'ordre associées. $k = k_n$ une suite d'entier satisfaisant :

$$1 < k < n,\ k \longrightarrow \infty \text{ et } \frac{k}{n} \longrightarrow 0 \text{ quand } n \longrightarrow \infty. \tag{2.5}$$

Estimateur de Pickands

L'estimateur de Pickands (1975) [27] est le premier estimateur suggéré pour le paramètre $\gamma \in \mathbb{R}$ et il est donné par la formule suivante :

$$\widehat{\gamma}_n^{(p)} := \frac{1}{\log 2} \frac{X_{n-k:n} - X_{n-2k:n}}{X_{n-2k:n} - X_{n-4k:n}}. \tag{2.6}$$

Théoreme 2.4.1 (Propriétés asymptotiques de $\widehat{\gamma}_n^{(p)}$)
Pour $\gamma \in \mathbb{R}$, $F \in \mathcal{DA}(\mathcal{G}_\gamma)$, et k que vérifie la condition (2.5), on a

***(a) Consistance faible** :*

$$\widehat{\gamma}_n^{(p)} \xrightarrow{\mathbb{P}} \gamma.$$

***(b) Consistance forte** : $k/\log\log(n) \longrightarrow \infty$, pour $n \longrightarrow \infty$, alors*

$$\widehat{\gamma}_n^{(p)} \xrightarrow{p.s} \gamma.$$

***(c) Normalité asymptotique** : sous certaines conditions sur k et F on a :*

$$\sqrt{k}(\widehat{\gamma}_n^{(p)} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \eta^2) \quad \text{quand } n \longrightarrow \infty,$$

où

$$\eta^2 := \frac{\gamma\sqrt{(2^{2\gamma+1}+1)}}{(2(2^\gamma - 1)\log 2)}.$$

La cosistance faible a été démontrée par Pikinds [27] en 1975 et la consistance forte et ainsi la normalité asympototique ont été démontrées par Dekkers et de Haan [9] en 1986.

Estimateur de Hill

Cet estimateur qui a été présenté par Hill [18] en 1975 et l'estimateur le plus célèbre parmi tous les estimateurs de l'indice de queue. Il s'applique seulement dans le cas où l'indice de queue est positif ($\gamma > 0$), qui correspond aux distributions appartenant au domaine d'attraction de Fréchet. Cet estimateur donné par la définition suivante :

Définition 2.4.1 (Estimateur de Hill)
Pour $\gamma > 0$ l'estimateur de Hill est défini par

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

Pour obtenir l'estimateur de $\widehat{\gamma}_n^{(H)}$, en effet la condition (2.2) a une forme équivalente

$$\lim_{t\longrightarrow\infty} \frac{1}{\overline{F}(t)} \int_t^{\infty} \frac{F(x)}{x} dx = \gamma.$$

Par l'integration par partie, on obtient

$$\lim_{t\longrightarrow\infty} \frac{1}{\overline{F}(t)} \int_t^{\infty} \log \frac{x}{t} dF(x) = \gamma.$$

On remplace F par F_n et $t = X_{n-k:n}$, l'estimateur de Hill $\widehat{\gamma}_n^{(H)}$ défini par

$$\widehat{\gamma}_n^{(H)} := \frac{1}{\overline{F}(X_{n-k:n})} \int_{X_{n-k:n}}^{\infty} \log \frac{x}{X_{n-k:n}} dF_n(x),$$

où

$$\widehat{\gamma}_n^{(H)} := \frac{1}{k}\sum_{i=0}^{k} \log X_{n-i+1:n} - \log X_{n-k:n}.$$

Théoreme 2.4.2 (Propriétés asymptotiques de $\widehat{\gamma}_n^{(H)}$)
Pour $\gamma > 0$, $F \in \mathcal{DA}(\Phi_{1/\gamma})$ et k que vérifie 2.5 on a :

(a)consistance faible

$$\widehat{\gamma}_n^{(H)} \xrightarrow{\mathbb{P}} \gamma \quad quand\ n \longrightarrow \infty.$$

(b)consistance forte

$$\widehat{\gamma}_n^{(H)} \xrightarrow{p.s} \gamma \quad quand\ n \longrightarrow \infty.$$

(c)Normalité asymptotique *: Si $lim_{t\longrightarrow\infty}(U(tx)/U(t) - x^{-\gamma}/A(t)) = x^{\gamma}(x^{\rho} - 1/\rho)$ est satisfaite avec $\sqrt{k}A(k/n) \longrightarrow \lambda \in \mathbb{R}$, quand $n \longrightarrow 1$, alors :*

$$\sqrt{k}(\widehat{\gamma}_n^{(H)} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(\frac{\lambda}{1-\rho}, \gamma^2) \quad quand\ n \longrightarrow \infty.$$

La consistance faible a été démontrée par Mason [23] en (1982) et Deheuvels et al [7] en (1989) ont établi la consistance forte.

Estimateur des Moments

Un autre estimateur qui peut être considéré comme une adaptation de l'estimateur de Hill, pour obtenir la consistance pour quelque soit le signe de l'indice γ, a été proposé par Dekkers et all [9] c'est l'estimateur de moment.

Définition 2.4.2 (Estimateur des Moments)
Pour $\gamma \in \mathbb{R}$, l'estimateur de moment est défini par

$$\widehat{\gamma}_n^{(M)} := M_1 + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1},$$

où

$$M_n^{(r)} = \frac{1}{k}\sum_{i=0}^{k}\left(\log X_{n-i+1:n} - \log X_{n-k:n}\right)^r ; \quad r = 1, 2$$

Théoreme 2.4.3 (Propriétés asymptotiques de $\widehat{\gamma}_n^{(M)}$)
Pour $\gamma \in \mathbb{R}$, $F \in \mathcal{DA}(\Phi_{1/\gamma})$ et k que vérifie la condition (2.5) on a :

(a) Consistance faible

$$\widehat{\gamma}_n^{(M)} \xrightarrow{\mathbb{P}} \gamma \quad quand\ n \longrightarrow \infty.$$

(b) Consistance forte : *Si $k/(log(n))^\delta \longrightarrow \infty$ quand $n \longrightarrow \infty$ avec $\delta > 0$ alors*

$$\widehat{\gamma}_n^{(M)} \xrightarrow{p.s} \gamma \quad quand\ n \longrightarrow \infty.$$

(c) Normalité asymptotique

$$\sqrt{k}(\widehat{\gamma}_n^{(M)} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \eta^2) \quad \text{quand } n \longrightarrow \infty,$$

où

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

Les propriétés asymptotiques de cet estimateur ont été étudiées dans Dekkers et al [9].

2.4.2 Estimation de l'IVE sous Données Incomplètes

Dans cette partie on va intéresser au problème de l'estimation de l'IVE mais cette fois c'est en présence de données censurées aléatoirement à droite. Ce problème,.est très récent dans la littérature, les premiers qui ont mentionné ce sujet sont Reiss et Thomas [28] en (2007) mais sans résultat asymptotique. En (2001) , Beirlant et Guillou [4] ils ont proposé un estimateur

mais pour les données tronquées. En 2007, Beirlant et al [3] ils ont introduit une méthode pour les estimateurs de Hill et de moment,... de plus, ils ont proposé les estimateurs des quantiles extrêmes et ont discuté leurs propriétés asymptotiques lorsque les données sont censurées pour un seuil déterministe et l'année suivante Einmahl et al.**[11]** ils ont adapté différents estimateurs de l'IVE au cas où les données sont censurées par un seuil aléatoire et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique.

Soient $\overline{F}$ et $\overline{G}$ sont à queues lourdes avec les indices $-1/\gamma_1$ et $-1/\gamma_2$ où $\gamma_1 > 0$, $\gamma_2 > 0$ de l'échantillon $X_1, ..., X_n$ et $Y_1, ...Y_n$. Soit $\{(Z_i, \delta_i)\,,\ 1 \leq i \leq n\}$ l'échantillon réellement observé défini par (1.11) . Il est clair que les $Z_i's$ sont des variables indépendantes de loi N liée à F et G par la relation (1.12) . L'IVE de la fdr $N(t)$ de Z, existe et il est notée par γ où $\gamma := \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$. Soit x_F, x_G et x_N les points terminaux du support de F, G et N respectivement. Einmahl et al.[11] ont proposé une adaptation générale des estimateurs existants dans les cas suivants :

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

Leurs estimateurs sont basés sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dans le plus grand k de Z

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

où

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

avec k est le nombre des valeurs extrêmes, $\hat{p}$ estime $p = \dfrac{\gamma_2}{\gamma_1+\gamma_2}$, où p représente la proportion des données observée dans la queue à droite de la distribution. $\hat{\gamma}^{(\cdot)}$ peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill, moment, ...

Brahimi et al. [5] ont également établi la consistance de $\hat{p}$ sous la condition du premier ordre sur les fdr's F et G. Ils ont aussi conclu, que l'estimateur de $\hat{\gamma}_1^{(c,)}(k)$ consistant de γ_1 pour l'estimateur de Hill. En autre Brahimi et al [5] ont utilisé la théorie des processus empiriques pour approcher l'estimateur de Hill adaptée en termes de processus Gaussien. Dans les années récentes, le problème de l'étude des phénomènes extrêmes et de l'estimation de l'IVE pour des données censurées a attiré l'attention d'un nombre croissant des chercheurs, en raison des nombreuses applications qui appellent des solutions concrètes. Worms et Worms [[38],] et Beirlant et al. [3] ont proposé une approche davantage axée sur l'analyse de la survie, les deux premières étant limitées au cas de la queue lourde**.** Ndao et al. [[25], [26]] ont adressée l'estimation non paramétrique de l'IVE conditionnel et son quantile pour les distributions à queue lourde qui ont été récemment généralisées par Stupfler [33] pour les trois domaines d'attraction des extrêmes, à savoir les types de distributions de Féchet, Gumbel et Weibull. Dans son document de travail, **Stupfler [34]** examiné le schéma de censure à

droite aléatoirement dépendant et développer un nouveau sujet intéressant. Plus récemment, **Brahimi et al.[6]** ont proposé un nouvel estimateur du type Hill (pondéré) pour l'IVE positif. Ils ont aussi proposé sa consistance et sa normalité asymptotique sont prouvées au moyen du processus mentionné dans leur travail dans le cadre des conditions de second ordre de variation régulière. Ils ont également donné une étude de simulation comparative, leur estimateur nouvellement défini est vu de meilleures performances que celles déjà existant (Einmahl et al [11], Worms et Worms **[38]**) en matière de biais et l'erreur quadratique moyenne (mse).

2.4.3 Application

Pour les données simulées et pour voir la performance de l'estimateur de Hill adapté $\widehat{\gamma}_1^{(H,c)} = \widehat{\gamma}^H/\widehat{p}$, on a réalisé une étude de simulation basée sur $1000, 2000$ et 5000 échantillons de la loi de Burr de paramètre γ_1 censurées par une autre variable de Burr de paramètre $\gamma_2 = p\gamma_1/(1-p)$, où p représente la proportion de données observées dans la queue à droit de distribution.

$$F(x) := 1 - \left(1 - x^{1/\eta}\right)^{-\eta/\gamma_1}, \quad G(x) := 1 - \left(1 - x^{1/\eta}\right)^{1/\gamma_2}, \quad x \geq 0,$$

avec $\eta, \gamma_1, \gamma_2 > 0$. On prend $\eta = 1/4$ et on choisit les valeurs 0.4, 1 pour γ_1. La proportion des valeurs extrêmes réellement observées, on prend $p = 0.3$ et 0.9.Les résultats numériques sont obtenus en faisant les moyennes sur les 1000 réplications. Pour ces résultats, on commence par déterminer le nombre optimal d'observations extrêmes utilisées dans le calcul de $\widehat{\gamma}_1^{(H,c)}$. Pour cela, on applique l'algorithme de Reiss et Thomas (voir [28]). Les résultats obtenus de $\widehat{\gamma}_1^{(H,c)}$, biais et l'erreur moyenne quadratique (mean squared error : mse) respectivement définis par

$$biais\left(\widehat{\gamma}_1^{(H,c)}\right) := \frac{1}{M}\sum_{h=1}^{M}\left(\widehat{\gamma}_1^{(H,c)}(h) - \gamma_1^{(H,c)}\right) \text{ et } mse := \frac{1}{M}\sum_{h=1}^{M}\left(\widehat{\gamma}_1^{(H,c)}(h) - \gamma_1^{(H,c)}\right)^2$$

sont résumés dans la Tableau 2.1 et Tableau 2.1 avec différents choix de l'indice γ_1 et du pourcentage p. Pour faciliter la lecture de ces tableaux, on remarque que l'estimation de $\widehat{\gamma}_1^{(H,c)}$ est meilleure pour un grande valeur de p.

$p = 0.3$								
$\gamma_1 = 0.4$					$\gamma_1 = 1$			
n	Ko	$\widehat{\gamma}_1$	$biais$	mse	Ko	$\widehat{\gamma}_1$	$biais$	mse
1000	30.000	0.555	0.155	0.125	28.000	1.160	0.160	0.351
2000	59.000	0.486	0.086	0.032	72.000	1.085	0.085	0.103
5000	150.00	0.453	0.053	0.007	165.000	1.015	0.015	0.037

TAB. 2.1 – Biais et mse de l'estimation de γ_1, basée sur 1000, 2000 et 5000 échantillons de la loi de Burr de paramètre γ_1 censurée par une variable de Burr de paramètre γ_2 avec p=0.3

$p = 0.9$								
$\gamma_1 = 0.4$					$\gamma_1 = 1$			
n	Ko	$\widehat{\gamma}_1$	$biais$	mse	Ko	$\widehat{\gamma}_1$	$biais$	mse
1000	67.000	0.409	0.009	0.006	70.000	0.983	-0.017	0.023
2000	125.000	0.403	0.003	0.002	127.000	0.978	-0.022	0.020
5000	326.000	0.405	0.005	0.001	330.000	0.989	-0.011	0.005

TAB. 2.2 – Biais et mse de l'estimation de γ_1, basée sur 1000, 2000 et 5000 échantillons de la loi de Burr de paramètre γ_1 censurée par une variable de Burr de paramètre γ_2 avec p=0.9

CONCLUSION

Dans ce mémoire on a abordé différent aspect de la théorie des valeurs extrêmes à la présence de censure aléatoire droite. Après avoir dans le premier chapitre quelques concepts de base et des résultats essentiels de l'analyse de survie, et dans le deuxième chapitre rappelé certaines notations clés en théorie des valeurs extremes .Ensuite, on a passé aux techniques pour l'estimateur de l'indice de queue γ et aussi, on a intéressé à la famille des lois à queue de type de Fréchet, ce type de lois intervient dans les nombreuses applications pour estimer l'indice de queue. On a présenté trois estimateurs (Pikinds, Hill et moment) pour les données complètes et présenté aussi l'estimateur de Einmahl et al. pour les données censurées.

Pour conclure, on signale que ce mémoire n'est qu'un point de départ pour mieux connaitre ce monde immense des valeurs extrêmes.

BIBLIOGRAPHIE

[1] Aalen, O. (1978). Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models. Ann. Statist., 534545.

[2] Balakrishnan, N and Cohen, A. C. (1991). Order Statistics and Inference :Estimation Methods. Statist. Model. Decis. Sci. Academic Press.

[3] Beirlant, J, Goegebeur, Y, Segers, J, and Teugels, J. (2006). Statistics of Extrêmes : Theory and Applications. John Wiley.

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

[5] Brahimi, B., Meraghni, D., and Necir, A. (2015). Gaussian approximation to the extreme value index estimator of a heavy-tailed distribution under random censoring. Math. Methods Statist., 24(4), 266279.

[6] Brahimi, D. Meraghni, A. Necir and L. Soltane. (2018). Tail empirical process and a weighted extreme value index estimator for randomly right-censored data. Unpublished manuscript, available on the ArXiv archive : https ://arxiv.org/abs/1801.00572.

[7] Deheuvels, P., Häeusler, E, and Mason, D. M. (1988). Almost sure convergence of the Hill estimator. Math. Proc. Cambridge Philos. Soc., **104**(02), 371381.

[8] David, H.A. (1970). Order Statistics. John Wiley & Sons, Inc., New York-London Sydney.

[9] Dekkers, A. L. M., Einmahl, J. H. J. & de Haan, L., (1989). A Moment Estimator for the Index of an Extreme-Value Distribution. Ann. Statist. **17**, 1833 − 1855.

[10] Denuit M, Charpentier A. (2005). Mathematiques de l'Assurance non-vie. Tome 2. tarification et provisionnement,Paris : Economica.

[11] Einmahl, J. H., Fils-Villetard, A., and Guillou, A. (2008). Statistics of Extremes Under Random Censoring. Bernoulli, **14**(1), 207227.

[12] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance, Springer-Verlag, Berlin.

[13] Fisher, R. A., and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Cambridge Philos. Soc., **24**(02), 180190.

[14] Gassom Zohra. (2006).Regression nom Parametrique dans les Modèles Censurées. University Houari Boumedianne. Alger.

[15] Gilbert. C.(2015). Modèle de Survie. Note de Cours Master 2. ESA.

[16] Guillou A. and Willems P. (2006). Application de la théorie des valeurs extrêmes en hydrologie. Statistique Appliquée, $5-31$.

[17] de Haan, L. and Ferreira, A. (2006). Extreme Value Theory : An Introduction. Springer-Verlag, New York.

[18] Hill, B. M.(1975). A simple general approach to inference about the tail of a distribution. Ann. Statist., **3**(5), 11631174.

[19] Jean-François D. (2002) Modélisation Conjointe de Données Longitudinales et de Durees de Vie.Mathematics.Université Rene Descartes - Paris French.<tel00002667 >.

[20] Kaplan, E.L, and Meier, P. (1958). Non parametric estimation from incomplete observations. J. Amer. Statist. Assoc, **53**(282), 457481.

[21] Lee, E. T and Wang, J. (2003). Statistical Methods for Survival Data Analysis. John Wiley.

[22] Longin, F. (1995) La théorie des valeurs extrêmes : présentation et premières applications en finance. Journal de la Société de Statistique de Paris, tome 136, N1.

[23] Mason, D. M.(1982). Laws of large numbers for sums of extreme values. Ann. Probab., 754764.

[24] Nelson, W. (1972). A Short Life Test For Comparing a Sample with Previous Accelerated Test Results. Technometrics, **14**(1), 175185.

[25] Ndao, P., Diop, A., and Dupuy, J. F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. Comput. Statist. Data Anal., **79**, 6379.

[26] Ndao., A. Diop, and J-F. (2016) Dupuy Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. In Journal of Statistical Planning and Inference, 168, pages $20-37$.

[27] Pickands, J. (1975). Statistical Inference Esing Extreme Order Statistics. Ann. Statist.$3, 119-13$.

[28] Reiss, R.D, and Thomas, M.(2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields.Birkhäuser, Basel.

[29] Resnick, S.I. (1987). Extreme Values, Regular Variation, and point Processes. Springer, New York.

[30] saint. P. P. (2012). Introduction à l'analyse des durées du survie.

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

[32] Soltan, L.(2017). Analyse des Valeurs Extrême en Presence de Censure.Thèse de doctorat.Université de Biskra, Algerie.

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

[34] Stupfler, G., (2017). On the study of extremes with dependent random right-censoring. Working paper : https ://hal.archives-ouvertes.fr/hal-01450775/document.

[35] Vivian Viallo (2006). Processus Empirique,Estimation non Parametrique et donnée censurées.Paris.

[36] Pierre, B. (1988). Introduction aux Probabilités, Modalisation des Phénomènes Aléatoires. Springer. New York.

[37] Wienke, A. (2010). Frailty models in survival analysis. CRC Press.

[38] Worms, J. and Worms, R. (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes, $17, 337358$

ANNEXE A : LOGICIEL *R*

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques ceux-ci sont visualisés immédiatement dans une fenêtre propre et peuvent être exportés sous divers formats (par exemple jpg, bmp, eps, ou wmf avec Windows, ps, pictex avec Unix). R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données, ... Plusieurs paquets, tels actuar, VGEM, MASS, multivariate, scatterplot3d et rgl entre autres sont destinés à l'analyse des données statistiques multidimensionnelles.

Il a été initialement créé, en 1996, par *Robert Gentleman* et *Ross Ihaka* du département de statistique de l'Université d'Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "*R Core Team*" qui développe *R*. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation *Unix*, *Linux*, *Windows* et *MacOS*.

Un élément clé dans la mission de développement de *R* est le *Comprehensive R Archive Network* (CRAN) qui est un ensemble de sites qui fournit tout ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à Vienne, on peut y accéder par l'URL : *http ://cran.r-project.org/*. Les autres sites du CRAN, appelés sites miroirs, sont répandus partout dans le monde.

R est un logiciel libre distribué sous les termes de la "GNU Public Licence". Il fait partie intégrante du projet GNU et possède un site officiel à l'adresse *http ://www.R-project.org*. Il est souvent présenté comme un clone de *S* qui est un langage de haut niveau développé par les *AT&T Bell Laboratories* et plus particulièrement par *Rick Becker, John Chambers* et *Allan Wilks*. *S* est utilisable à travers le logiciel *S-Plus* qui est commercialisé par la société Insightful *http ://www.splus.com/*.

ANNEXE B : ABRÉVIATIONS ET NOTATIONS

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

IVE	:	Indice des valeurs extrêmes.
TEV	:	Theorie des valeurs extremes.
F	:	Fonction de répartition.
F_n	:	Fonction de répartition empirique.
F^{-1}	:	Inverse généralisé de F.
$\mathcal{G}_\gamma$	:	Famille de la loi de valeurs extrêmes généralisée.
iid	:	Indépendantes et identiquement distribué.
$\mathbb{1}_A$	:	Fonction indicatrice de l'ensemble A.
f	:	Fonction de densitée.
Λ	:	Loi de Gumbel.
Φ	:	Loi de Fréchet.
Ψ	:	Loi de weibull.
$l(t)$	:	Fonction à variation lente.
$\mathcal{DA}()$	:	Domaine d'attraction de maximum.
$\max(X, Y)$	:	Maximum de $X_1, ..., X_n$.
$\xrightarrow{\mathcal{D}}$	:	Converge en distribution.
$\xrightarrow{\mathbb{P}}$	:	Converge en probabilité.
$\xrightarrow{p.s}$	:	Convegence prèsque sûre.
$X_{1:n}, ..., X_{n:n}$	:	Statistique d'ordre associées à $X_1, ..., X_n$.
$X \Lambda Y$	:	$min(X, Y)$.
x_F	:	Point terminal.
$:=$	:	Égalité en définition.
$\overline{X}$	:	Moyenne arithmétique.
al	:	Autres.

va	:	variable aléatoire.
va's	:	variables aléatoires.
$(\Omega, \mathcal{F}, \mathbb{P})$	:	Espace probabilisé.
TCL	:	Théorème Centrale Limite.
$\mathcal{N}(0,1)$	:	Loi normale standard.
$\mathbb{R}$	:	Ensemble des valeurs réelles.
$H_n(t)$	:	Estimateur de Nelson-Aalen
$\widehat{F}_n$	:	Estimateur de Kaplan-Meier.
$S = \overline{F}$	:	Fonction de survie.
S_n	:	Somme arithmétique.
Q	:	Fonction de quantile.
Q_n	:	Fonction de quantile empirique.