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 **:**

MASTER en Mathématiques

Option : **Statistique**

Par

Nesserine BENSMAIL

Titre :

Estimation et test de l'indice des valeurs extrêmes en présence de censure

Membres du Comité d'Examen :

Pr.	**BRAHIMI Brahim**	UMKB	Président
Dr.	**BENAMEUR Sana**	UMKB	Encadreur
Dr.	**SOLTANE Louiza**	UMKB	Examinateur

26 Juin 2019

Dédicace

Je dédie ce modeste travail

À Ma chère mère et mon cher père

À mes frères.

Bensmail Nesserine

REMERCIEMENTS

Avant tout, je remercie mon dieu le tout puissant de m'avoir accordé volonté et patience pour accomplir ce travail.

Mes remerciements les plus sincères vont en particulier, à ***Dr. Benameur Sana*** *pour avoir m'encadrer et diriger, et pour l'effort fournit, ces conseils prodigués et sa patience dans le suivie de ce travail.*

Mes remerciements vont également au membres de jury : ***Pr.Brahimi Brahim*** *et* ***Dr. Soltane Louiza*** *pour avoir accepté de juger ce mémoire.*

Je souhaite particulièrement remercier mes parents qui m'ont stimule et encouragé pendant mes études. qui étaient toujours prêts à fournir tous les moyens physiques et morales pour la réussite de ce projet.

Un grand merci aux enseignants(es) et personnel du département de mathématique.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse de survie est la partie de la statistique qui s'intéresse à l'inférence dans le cas d'observations censurées. Elle trouve sa place dans tous les champs d'application où l'on étudie le délai de survenue d'un évènement dans un ou plusieurs groupes d'individus. 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 le décès. Ce trouve aussi beaucoup d'applications dans les sciences sociales, économiques et actuarielles, on l'appelle "analyse de durée". Une des caractéristiques des données de survie est l'existence d'observations incomplètes.

La théorie des valeurs extrêmes (TVE) est une branche de la statistique qui essaie d'amener une solution face à des évènements rares ; c'est à dire des événements dont la probabilité de réalisation est très faible. Cette théorie a été développée dans le contexte d'observations indépendantes ; les auteurs Fisher et Tippet (1928) [14], montrent que sous certaines conditions, les seules distributions limites des extrêmes sont les lois de Fréchet, Gumbel et Weibull. Ceci nous permet de classer la plupart des lois en trois domaines d'attraction où chaque domaine est déterminé par des caractérisations sur les fonctions de répartition. Von Mises (1954) [30] puis Jenkinson (1955) [21] ont rassemblé les distributions de ces trois domaines en une seule écriture (voir Embrechts et al (1997) [12], de Haan et Ferreira (2006) [9]). Cependant, on y retrouve deux modèles : loi généralisée des extrêmes (GEV : Generalized Extreme Value) et loi de Pareto généralisée (GPD : Generalized Pareto Distribution). Divers travaux ont été consacrés à l'estimation de l'indice des extrêmes dont l'objectif revient à construire des estimateurs et étudier leurs propriétés ; citons Hill (1975) [19] et Pickands (1975) [26]. La plupart des estimateurs reposent sur l'utilisation de la statistique d'ordre.

La modélisation des valeurs extrêmes censurées, est un problème très récent dans la littérature, les premiers qui ont mentionné le sujet sont (Reiss et Thomas, 2007 [27]). Les auteurs Beirlant et al en (2007) [3] commencé a proposé une adaptation d'estimation classique de

l'indice des valeurs extrêmes (EVI : Extreme value index) dans le cas où les données sont censurées. Leur estimateur est basé sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dépassant un certain seuil déterministe. En 2008, Einmahl et al [13]ont utilisé le même concept pour proposer un estimateur de l'indice de queue sur les k-plus grandes valeurs et ils ont proposés une méthode unifiée pour établir leur normalité asymptotique.

L'application de la théorie des tests d'hypothèse dans le domaine des valeurs extrêmes censurées a bénéficiée d'une certaine attention dans la littérature récente. L'objectif de ce travail est l'estimation de l'indice des valeurs extrêmes et de construire un test paramétrique pour cet indice, en se basant sur des données censurées ou incomplète.

Ce mémoire est subdivisé en deux chapitres comme suit :

Au premier chapitre, on présente une introduction sur les statistiques d'ordre, nous donnons un aperçu des définitions et des résultats théoriques essentiels de la TVE, nous présentons les résultats sur le comportement du maximum, un rappel sur les deux principaux outils servant à modéliser le comportement des valeurs extrêmes d'un échantillon : la loi des valeurs extrêmes (GEV) et la loi des excès (GPD). Nous nous intéressons ensuite à la caractérisation des domaines d'attraction et les fonctions à variation régulière.

Nous avons consacré le deuxième chapitre pour l'estimation et l'application d'un test paramétrique de l'indice de queue sous données incomplètes, nous nous limitons à un bref rappel sur quelques éléments fondamentaux de l'analyse de survie et sur des notions et définitions de base de la théorie des tests d'hypothèse. Dans la dernière section nous nous intéressons essentiellement a un test basé sur la normalité asymptotiquen de l'indice des valeurs extrêmes, dans le cas de données censurées à droite. En fin, nous présentons des résultats numériques du test à l'aide des simulations sous logiciel **R**.

Chapitre 1

Théorie des valeurs extrêmes

Dans ce chapitre, nous rappelons quelques notions essentielles dans la TVE, nous définissons rapidement les caractérisations des domaines d'attractions et nous étudions le comportement des valeurs extrêmes. On commence dans un premier temps par donner la notion et les propriétés des statistiques d'ordres qu'elles sont un outil essentiel dans cette théorie. Pour plus de détails sur ce chapitre, il se trouve des bonnes références sur la théorie et les applications des valeurs extrêmes suivantes : Embrechts et al (1997) [12], David et Nagaraja (2003) [7].

1.1 Statistiques d'ordre

Pour commencer notre étude et les explications de la théorie des valeurs extrêmes, il faut avoir un grand bagage, alors notre point de départ sera les statistiques d'ordre.

1.1.1 Définitions

Soit $(X_1, X_{2,} \cdots, X_n)$ une suite de variables aléatoires (v.a's) réelles, indépendantes et identiquement distribuées (i.i.d), de fonction de répartition F, telle que :

$$F(x) := P(X \leq x), \text{ pour tout } x \in \mathbb{R}.$$

On appelle statistique d'ordre le réarrangement croissant de $(X_1, X_{2,} \cdots, X_n)$ notée :

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

Pour $1 \leq k \leq n$, la v.a $X_{k,n}$ est appelée la $k^{ème}$ statistique d'ordre.

En particulier, la statistique du minimum et du maximum sont respectivement données par :

$$X_{1,n} := min(X_1, X_{2,} \cdots, X_n) \; et \;\; X_{n,n} := max(X_1, X_{2,} \cdots, X_n).$$

En générale, les résultats des minimaux peuvent être tirés du maximum par la relation suivante :

$$min(X_1, X_{2,} \cdots, X_n) = -max(-X_1, -X_2 \cdots, -X_n).$$

1.1.2 Distribution d'une statistique d'ordre

Distribution du maximum et du minimum

La fonction de répartition du maximum $X_{n,n}$ et du minimum $X_{1,n}$ de l'échantillon $(X_1, \cdots, X_n)$ sont données respectivement par :

$$\begin{aligned} F_{X_{n,n}}(x) &= P(X_{n,n} \leq x) = P(X_1 \leq x, \cdots, X_n \leq x) \\ &= P\left[\bigcap_{i=1}^{n}(X_i \leq x)\right] \\ &= \prod_{i=1}^{n} P(X_i \leq x) \\ &= [F(x)]^n, \end{aligned} \tag{1.1}$$

et

$$\begin{aligned} F_{X_{1,n}}(x) &= P(X_{1,n} \leq x) = 1 - P(X_{1,n} > x) \\ &= 1 - P\left[\bigcap_{i=1}^{n}(X_i > x)\right] \\ &= 1 - \prod_{i=1}^{n}(P(X_i > x) \\ &= 1 - [1 - F(x)]^n. \end{aligned} \tag{1.2}$$

Remarque 1.1 *Pour obtenir la densité du maximum et du minimum, on dérive respective-*

ment (1.1) et (1.2), alors :

$$f_{X_{n,n}}(x) = nf(x)\left\{F(x)\right\}^{n-1},$$
$$f_{X_{1,n}}(x) = nf(x)\left\{1-F(x)\right\}^{n-1}.$$

Proposition 1.1 *La fonction de répartition de la $k^{ème}$ statistique d'ordre est donnée par :*

$$F_{X_{k,n}}(x) := \sum_{i=k}^{n} \binom{n}{i} \left\{F(x)\right\}^{i} \left\{1-F(x)\right\}^{n-i}, \ -\infty < x < \infty.$$

De plus, si F admet une densité f, alors :

$$f_{X_{k,n}}(x) := \frac{n!}{(k-1)!(n-k)!} f(x)F(x)^{k}(1-F(x))^{n-k}, \ -\infty < x < \infty.$$

Preuve. Voir David et Nagaraja (2003) [7]. ■

Théorème 1.1 *Si on suppose l'échantillon $X_1, X_2, \cdots, X_n$ de v.a's i.i.d selon une loi de densité f, alors la densité jointe des n statistiques d'ordre $X_{1,n} \leq \cdots \leq X_{n,n}$ est donnée par :*

$$f_{X_{1,n},\cdots,X_{n,n}}(x_1, x_2, \cdots, x_n) = n! \prod_{i=1}^{n} f\left(x_i\right), \ -\infty < x_1 < \cdots < x_n < \infty.$$

1.2 Distribution asymptotique du maximum

La distribution du maximum devrait nous fournir des informations sur des événements extrêmes, et comme la fonction de répartition de X n'étant pas souvent connue, il n'est généralement pas possible de déterminer cette distribution à partir (1.1). On s'intéresse alors à la distribution asymptotique du maximum. Alors, on a :

$$\lim_{n\to\infty} F_{X_{n,n}}(x) = \lim_{n\to\infty} \left\{F\left(x\right)\right\}^{n} = \begin{cases} 1 & \text{si} \quad F(x) = 1 \\ 0 & \text{si} \quad F(x) < 1 \end{cases}$$

On constate que la distribution asymptotique du maximum donne une loi dégénérée (Elle prend les valeurs 0 et 1 seulement). On cherche alors une loi non dégénérée pour $X_{n,n}$ de façon analogue au théorème centrale limite.

Fisher et Tippet (1928) [14], Gnedenko (1943) [15] et de Hann (1970) ont proposées le théorème suivant qui donne une condition nécessaire et suffisante pour l'existence d'une loi limite non dégénérée pour le maximum.

Théorème 1.2 (Fisher et Tippet(1928), Gnedenko (1943))
S'il existe deux suites de constantes de normalisation avec $a_n > 0$ *et* $b_n \in \mathbb{R}$ *et une loi non-dégénérée de fonction de répartition* $\mathcal{H}$*, telle que :*

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F_{X_{n,n}}\left(a_n x + b_n\right) = \mathcal{H}_\gamma\left(x\right),$$

alors $\mathcal{H}$ est du même type que l'une des trois distributions suivantes :

$$\textbf{Loi de Gumbel} : \Lambda(x) := \exp\left(-\exp\left(-x\right)\right),\ x \in \mathbb{R}\ et\ \alpha = 0.$$

$$\textbf{Loi de Fréchet} : \Phi_\alpha(x) := \begin{cases} \exp\left(-x^{-\alpha}\right) & , x \geq 0 \\ 0 & , x < 0 \end{cases} \quad et\ \alpha > 0.$$

$$\textbf{Loi de Weibull} : \Psi_\alpha(x) := \begin{cases} \exp(-(-x)^{\alpha}) & , x \leq 0 \\ 1 & , x > 0 \end{cases} \quad et\ \alpha < 0.$$

Preuve. Une preuve détaillée de ce théorème peut être trouvée dans Embrechts et al (1997) [12] ■

Remarque 1.2 (Relation entre Λ, Φ_γ et Ψ_γ)
On peut passer de l'une des trois lois à l'autre comme suit :

$$Y \backsim \Phi_\gamma \iff -Y^{-1} \backsim \Psi_\gamma \iff \ln Y^\gamma \backsim \Lambda.$$

1.2.1 Approximation par GEV

Etant donnée qu'il est difficile de travailler avec les distributions limites, Jenkinson (1955) [21] et Von-mises (1954) [30] ont proposés une famille paramétrique appelée distribution des valeurs extrêmes généralisées standard GEV (Generalized Extreme Value), qui permet d'unifier les trois types des lois extrêmes si dessus, la fonction $\mathcal{H}_\gamma$ devient :

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

où γ est un paramètre qui contrôle la lourdeur de la queue de distribution appelé indice des valeurs extrêmes (EVI). Pour les variables non centrées et non réduites, on peut écrire $\mathcal{H}_\gamma(x)$ sous une forme plus générale, notée $\mathcal{H}_{\mu,\gamma,\sigma}(x)$, en remplaçant x par $(x-\mu)/\sigma$

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

où $\mu \in \mathbb{R}$ est un paramètre de localisation et $\sigma > 0$ est un paramètre d'échelle.
Cette approche est apportée aux données qui consistent en un ensemble des maximums par blocs, dite "Block Maxima" : annuels journaliers, semestriels journaliers, trimestriels journaliers, etc.

Proposition 1.2 (Λ, Φ_α et Ψ_α en terme de $\mathcal{H}_{\mu,\gamma,\sigma}$)
Nous avans, les correspondances suivantes :

$$\begin{aligned} &\mathcal{H}_{\frac{1}{\alpha}}(\alpha(x-1)) = \Phi_\alpha(x) \quad si \quad \alpha > 0, \\ &\mathcal{H}_{-\frac{1}{\alpha}}(\alpha(x+1)) = \Psi_\alpha(x) \quad si \quad \alpha < 0, \\ &\mathcal{H}_0(x) = \Lambda(x) \quad si \quad \alpha = 0. \end{aligned}$$

La densité de la GEV s'écrit pour $\gamma \neq 0$ comme suit :

$$h_{\mu,\gamma,\sigma}(x) := \frac{1}{\sigma}\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\gamma}{\gamma}\right)}_{\mu,\gamma,\sigma} \mathcal{H}(x),$$

la fonction de densité standard correspondante h_γ est

$$h_\gamma(x) := \begin{cases} \mathcal{H}_\gamma(x)(1+\gamma x)^{-\frac{1}{\gamma-1}} & \text{si} \quad \gamma \neq 0,\ 1+\gamma x > 0, \\ \exp(-x-\exp(-x)) & \text{si} \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases}$$

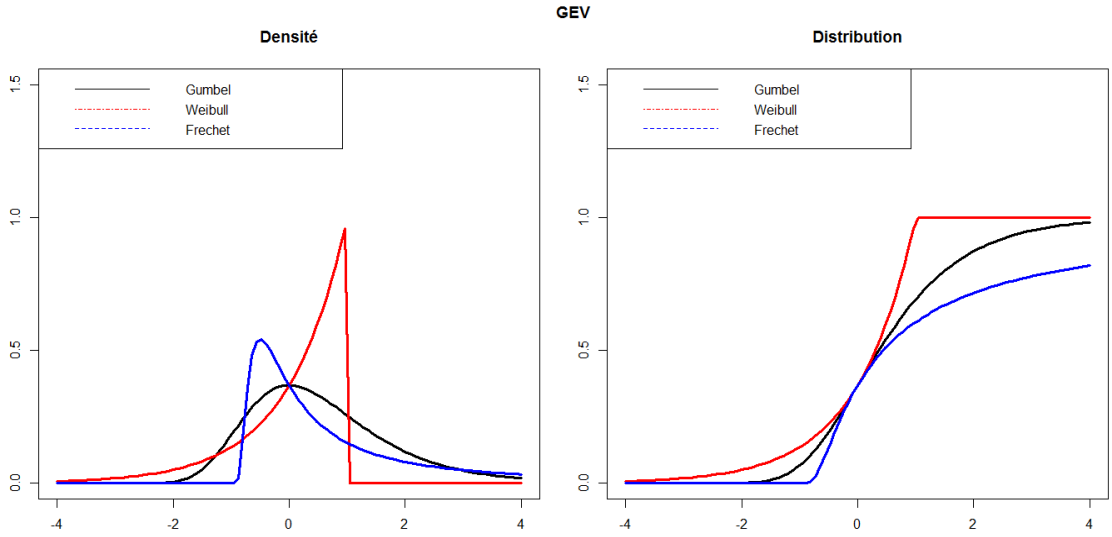

FIG. 1.1 – Représentation de la densité de probabilité et fonction de répartition : Gumbel ($\gamma = 0$), Fréchet ($\gamma = 1$) et Weibull ($\gamma = -1$)

1.2.2 Domaine d'attraction

Dans cette section, nous établirons les conditions nécessaires et suffisantes qui doivent être vérifier par une fonction de distribution F, pour que la loi du maximum converge vers $\mathcal{H}_\gamma$.

Définition 1.1 (Domaine d'attraction)

On dit qu'une distribution appartient au domaine d'attraction de $\mathcal{H}_\gamma$ *(notée* $F \in \mathcal{D}(\mathcal{H}_\gamma)$*), si la distribution du maximum normalisé converge vers* $\mathcal{H}_\gamma$*. Autrement dit, s'il existe des constantes réelles* $a_n > 0$ *et* $b_n \in \mathbb{R}$ *tels que* $\lim_{n\to\infty} F(a_n x + b_n) = \mathcal{H}_\gamma(x)$.

Définition 1.2 (Fonction à variation régulière)

Une fonction h *positive à l'infini (c-à-d : s'il existe* A *tel que* $x \geq A$, $h(x) > 0$*) est dite à variation régulière à l'infini d'indice* $\rho \in \mathbb{R}$ *et on note* $h \in \mathcal{RV}_\rho$ *si et seulement si*

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

Dans le cas particulier où $\rho = 0$*, on dit que* h *est une fonction à variation lente.(notée par* $h \in \mathcal{RV}_0$*). Nous réserverons le symbole* L *pour de telles fonctions.*

Remarque 1.3

On remarque que si $h \in \mathcal{RV}_\rho$ *alors* $\dfrac{h(x)}{x^\rho}$ *est à variation lente, il est facile de montrer qu'une fonction à variation régulière d'indice* ρ *peut toujours s'écrire sous la forme*

$$h(x) = x^\rho L(x),$$

où $L \in \mathcal{RV}_0$.

Proposition 1.3 (Représentation de Karamata)

Soit $h \in \mathcal{RV}_\rho$, $\rho \in \mathbb{R}$. Il existe deux fonctions mesurable $c > 0$ et ε telles que

$$\lim_{x\to\infty} c(x) = c_0 \in]0, +\infty[\text{ et } \lim_{x\to\infty} \varepsilon(x) = \rho$$

et $x_0 > 0$ tels que pour tout $x \geq x_0$,

$$L(x) = c(x) \exp\left\{ \int_{x_0}^{x} \frac{\varepsilon(u)}{u} du \right\}.$$

Remarque 1.4 *Dans le cas où la fonction* $c(.)$ *est constante, la fonction* $L(.)$ *correspondante est dite normalisée.*

1.2.3 Caractérisations des domaines d'attraction

Proposition 1.4 (Caractérisation de $\mathcal{D}(\mathcal{H})$)
$F \in \mathcal{D}(\mathcal{H}_\gamma)$ *si et seulement si, pour une certaine suite* $a_n > 0$ *et* $b_n \in \mathbb{R}$, *on a*

$$\lim_{n\to\infty} n\bar{F}(a_n x + b_n) = -\log \mathcal{H}_\gamma(x), \ x \in \mathbb{R},$$

avec $\bar{F}$ *est la fonction de survie (encore appelée queue de distribution), définie par :*

$$\bar{F}(x) := 1 - F(x).$$

En pratique, il est souvent plus commode, non pas de travailler sur la fonction F elle-même, mais sur la fonction quantile de queue.

Définition 1.3 (Inverse généralisée)

On appelle inverse généralisée (fonction des quantiles) de la fonction F, l'application notée $F^{\leftarrow}$ définie par :

$$Q(p) = F^{\leftarrow}(p) := \inf\{x : F(x) \geq p\}, \ p \in [0,1].$$

Définition 1.4 (Fonction quantile de queue)

On appelle fonction quantile de queue de la distribution F, la fonction $U :]1, +\infty[\to \mathbb{R}$, définit par :

$$U(t) := Q\left(1 - \frac{1}{t}\right) = F^{\leftarrow}\left(1 - \frac{1}{t}\right), \ \forall t > 1.$$

où Q est la fonction des quantiles associée à F.

Dans la suite on note x_F le point terminal de F défini comme suit :

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

Théorème 1.3 (Caractérisation du $\mathcal{D}(\mathcal{H}_\gamma)$)

Pour $\gamma \in \mathbb{R}$, les affirmations suivantes sont équivalentes :

1. $F \in \mathcal{D}(\mathcal{H}_\gamma)$
2. Il existe une fonction mesurable a telle que pour $1 + \gamma x > 0$, on a :

$$\lim_{u \to x_F} \frac{\bar{F}((a(u)x + u))}{\bar{F}(u)} = \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}} & \text{si} \quad \gamma \neq 0, \\ \exp(-x) & \text{si} \quad \gamma = 0. \end{cases}$$

3. Pour $x, y > 0$ et $y \neq 1$, on a

$$\lim_{s \to \infty} \frac{U(sx) - U(s)}{U(sy) - U(s)} = \begin{cases} \dfrac{x^\gamma - 1}{y^\gamma - 1} & \text{si} \quad \gamma \neq 0, \\ \dfrac{\log x}{\log y} & \text{si} \quad \gamma = 0. \end{cases} \tag{1.3}$$

Preuve. Voir Embrechts et al. [12] ■

Théorème 1.4 (Caractérisation du $\mathcal{D}(\Phi_\gamma)$)

Une fonction de répartition F appartient au domaine d'attraction de la loi de Fréchet Φ_γ $(\gamma > 0)$ si et seulement si $x_F = +\infty$ et

$$\bar{F}(x) = x^{-\gamma} L(x),$$

où L est une fonction à variation lente. Dans ce cas en choisis $a_n = F^{\leftarrow}(1-\frac{1}{n}) = U(n)$ et $b_n = 0, \forall n > 0$, alors

$$a_n^{-1} X_{n,n} \xrightarrow{d} \Phi_\gamma, \text{ quand } n \to \infty.$$

Exemple 1.1 *Soit $X_1, \cdots, X_n$ une suite de v.a's i.i.d de loi de Paréto de paramètre $\gamma > 0$, de fonction de répartition $F(x) := 1 - cx^{-\gamma}$, pour $a_n = (cn)^{\frac{1}{\gamma}}$ et $b_n = 0$, on'a :*

$$\begin{aligned}\lim_{n\to\infty} F^n(a_n x + b_n) &= \lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) \\ &= \lim_{n\to\infty} \left(1 - \frac{x^{-\gamma}}{n}\right)^n \\ &= \exp(-x^{-\gamma}).\end{aligned}$$

Donc la loi limite est une loi de Fréchet $F \in \mathcal{D}(\Phi_\gamma)$.

Remarque 1.5 *Toutes les distributions appartenant au domaine d'attraction de Fréchet sont dites de type Paréto et leur queue de distribution $\bar{F}$ s'écrit, pour x très grand comme suit :*

$$\bar{F}(x) = Cx^{-\gamma}, \ C, \gamma > 0.$$

Théorème 1.5 (Caractérisation du $\mathcal{D}(\Psi_\gamma)$)
On dit que F appartient au domaine d'attraction de la loi de Weibull Ψ_γ ($\gamma < 0$) si et seulement si $x_F < +\infty$ et

$$\bar{F}(x_F - x^{-1}) = x^{-\gamma} L(x),$$

pour certains fonction à variation lente L. Dans ce cas en peut choisir les constantes de normalisation $a_n = x_F - U(n)$ et $b_n = x_F, \forall n > 0$, alors :

$$a_n^{-1}(X_{n,n} - x_F) \xrightarrow{d} \Psi_\gamma, \text{ quand } n \to \infty.$$

Exemple 1.2 *Soit $X_1, \cdots, X_n$ une suite de v.a's i.i.d de loi uniforme sur $[0,1]$, de fonction de répartition $F(x) = x$, pour $a_n = \dfrac{1}{n}$ et $b_n = 0$, on'a :*

$$\begin{aligned}\lim_{n\to\infty} F^n(a_n x + b_n) &= \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \\ &= \exp(x) \\ &= \exp(-(-x))\end{aligned}$$

Donc la loi limite est une loi de Weibull $F \in \mathcal{D}(\Psi_1)$.

Définition 1.5 (Fonction de von-Mises)

Soit F une fonction de répartition de point terminale x_F , on suppose qu'il existe $z < x < x_F$, tel que :

$$\bar{F}(x) = c \exp\left\{ -\int_z^x \frac{1}{a(t)} dt \right\},$$

où $c > 0$ et a est une fonction positive absolument continue avec la densité á vérifiant $\lim_{x \longrightarrow x_F} a'(x) = 0$. Alors F est appelé fonction de von-Mises et a est une fonction auxiliaire de F.

Théorème 1.6 (Caractérisation du $\mathcal{D}(\Lambda)$)

La fonction de répartition F appartient au domaine d'attraction de Gumbel $\mathcal{D}(\Lambda)$ avec le point terminale $x_F \leq \infty$ si et seulement s'il existe $z < x_F$ tel que :

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

où c et g sont des fonctions mesurables satisfaisant $c(x) \to c > 0$, $g(x) \to 1$ quand $x \to x_F$, et a est une fonction auxiliaire. Dans ce cas on peut choisir les constantes de normalisation $b_n = F^{\leftarrow}(1 - \frac{1}{n})$ et $a_n = a(b_n)$. Un choix possible pour la fonction a est :

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

Dans ce tableau (1.1) , nous avons regroupé quelques lois et leur domaine d'attraction

Domaine d'attraction	Gumbel($\gamma = 0$)	Fréchet ($\gamma > 0$)	Weibull ($\gamma < 0$)
Lois	Gumbel	Burr	Beta
	Weibull	Pareto	Uniforme
	Gamma	Cauchy	inverse de Burr
	Normale	Student	inverse de Pareto
	Exponentielle	Log-gamma	
		Log-logistique	

TAB. 1.1 – Exemple sur les lois classées selon leurs domaines d'attraction

1.2.4 Approximation par GPD

Cependant, l'approche par GEV a été critiquée dans la mesure où l'utilisation d'un seul maxima conduit à une perte d'information contenue dans les autres grandes valeurs de l'échantillon. Pour pallier ce problème, Pickands en (1975) [26] a introduit une nouvelle approche dans la TVE connue par la méthode POT : Peaks Over Threshold (ou des excès au-delà d'un seuil). Cette méthode consiste à observer non pas le maximum ou les plus grandes valeurs, mais toutes les valeurs des réalisations qui excèdent un certain seuil élevé bien déterminé. Elle a été développée par divers auteurs tels que Smith en (1987) [29], et Reiss et Thomas en (2007) [27].

La GPD peut être utilisée pour modéliser des queues de distributions, c'est-à-dire pour des données dépassant un certain seuil Pour être plus précis, on définit un nombre réel u suffisamment élevé appelé seuil avec $N_u = card\{i : i = 1, \cdots, n,\ X_i > u\}$ est le nombre de dépassements du seuil et $Y_i = X_i - u > 0$ pour $1 \leq i \leq N_u$ pour les $(X_i)_{1 \leq i \leq n}$ et $Y_1, \cdots, Y_{N_u}$ les excès correspondants.

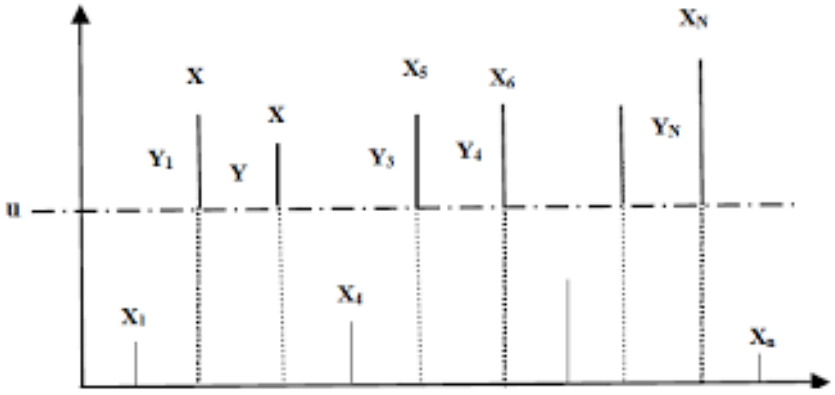

FIG. 1.2 – Schéma représentatif de la méthode POT : u est le seuil, Y : excès de X au-delà de u.

Définition 1.6 (Fonction de répartition et moyenne des excès)

Soit X une v.a de fonction de répartition F et de point terminal x_F.pour tout $u < x_F$, la fonction

$$F_u(y) := P\left(X - u \leq y \,|X > u\right) = \frac{F(y+u) - F(u)}{\bar{F}(u)},\ 0 < y < x_F$$

est appelée fonction de répartition des excès au-dessus du seuil u (ou encore appelée fonction de distribution excédentaire).

De même, si X est intégrable, la fonction de moyenne des excès de X est donnée par :

$$e_u(x) := E(X - u \mid X > u) = \int x dF_u(y), u < x_F,$$

qui peut s'écrire aussi comme suit :

$$e_u(x) = \frac{1}{\bar{F}(u)} \int_u^{\infty} \bar{F}(x)dx, u < x_F$$

Définition 1.7 (GPD)

La loi de Pareto Généralisée (GPD), de paramètres $\gamma \in \mathbb{R}$ et $\sigma > 0$, est définie par sa fonction de répartition, donnée par :

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \frac{\gamma}{\sigma}x\right)^{-\frac{1}{\gamma}} & si \quad \gamma \neq 0 \\ 1 - \exp(-x/\sigma) & si \quad \gamma = 0 \end{cases}, \ pour\ x \in \left\{x \in \mathbb{R}; 1 + \frac{\gamma}{\sigma}x > 0\right\} \cap [0; \infty].$$

La densité de la distribution GPD s'écrit comme suit :

$$g_{\gamma,\alpha}(x) := \begin{cases} \frac{1}{\alpha}\left(1 + \frac{\gamma}{\sigma}x\right)^{-\frac{1}{\gamma}-1} & si \quad \gamma \neq 0, \\ \frac{1}{\sigma}\exp(- - x/\sigma) & si \quad \gamma = 0. \end{cases}$$

Remarque 1.6 *Selon les valeurs du paramètre de forme γ, la GPD regroupe les trois distributions suivantes*

$$si\ \gamma > 0\ alors\ G_{\gamma,\alpha} \longmapsto loi\ Pareto$$
$$si\ \gamma < 0\ alors\ G_{\gamma,\alpha} \longmapsto loi\ Béta$$
$$si\ \gamma = 0\ alors\ G_{\gamma,\alpha} \longmapsto loi\ exponentiel$$

Théorème 1.7 (Pickands-Balkema et de Haan (1975))

Si la fonction de répartition F de point terminal x_F appartient au domaine d'attraction de G_γ, alors il existe une fonction mesurable et positive $\sigma(u)$, tel que :

$$\lim_{u \to x_F} \sup_{0<y<x_F-u} |F_u(y) - G_{\gamma,\sigma(u)}(y)| = 0. \tag{1.4}$$

L'interprétation de ce théorème est la suivante : On peut modéliser asymptotiquement la distribution conditionnelle de dépassements au-dessus d'un seuil par la GDP.

Preuve. La preuve de ce théorème doit être trouvé dans Embrecht et al.(1997) [12] ■

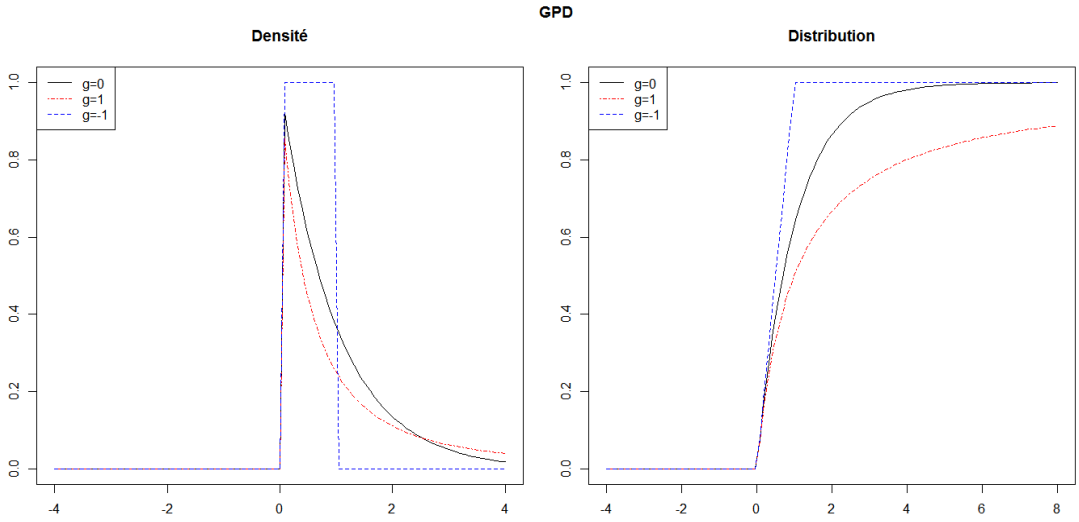

FIG. 1.3 – Représentation graphique de la densité et distribution de lois Paréto généralisée

Exemple 1.3 *Pour la loi exponentielle standard, la fonction de répartition est* $F(x) = 1 - \exp(-x)$, $x > 0$, *alors :*

$$F_u(y) = \frac{\exp(-u) - \exp(-u-y)}{\exp(-u)} = 1 - \exp(-y).$$

Ceci corresponds à $\gamma = 0$ et $\sigma = 1$ dans (1.4).

Exemple 1.4 *Pour la loi de Fréchet standard st, la fonction de répartition est* $F(x) = \exp(-1/x)$ *pour* $x > 0$, *alors :*

$$F_u(y) = \frac{1 - \exp\left(\frac{-1}{u+y}\right)}{1 - \exp(-\frac{1}{u})} = 1 - \left(1 + \frac{x}{u}\right)^{-1}.$$

Ceci corresponds à $\gamma = 1$ et $\sigma\,(u) = u$ dans (1.4).

Chapitre 2

Test paramétrique de l'indice des valeurs extrêmes sous censure

Dans ce chapitre, nous nous intéressons essentiellement à l'estimation et à un test paramétrique de l'indice des valeurs extrêmes sous données censurées. Dans la première section, nous énonçons les estimateurs les plus célèbres de l'indice de queue sous données complètes tels que l'estimateur Hill qui a été appliqué seulement dans le domaine d'attraction de Fréchet et celui de Pickands qui est valable pour les différentes lois limites des extrêmes, avec leurs propriétés asymptotiques. Nous rappelons dans la deuxième section les différentes fonctions utilisées en analyse de survie et les différents types de censures, et comme notre travail porte sur la construction d'un test paramétrique pour l'EVI à la fin de ce chapitre, nous présentons premièrement un certain nombre de généralités autour des tests d'hypothèses. En fin, nous présentons des résultats numériques du test sur des données simulées sous logiciel R. Pour plus détails sur ce chapitre, nous referons aux ouvrages de référence sur l'estimation de l'EVI sous donnée complète et censuré Einmahl et al (2008) [13], Beirlant et al (2007) [3] et de Haan et Ferreira (2006) [9].

2.1 Estimateur de l'indice des valeurs extrêmes en l'absence de censure

2.1.1 Estimateur de Pickands $\gamma \in \mathbb{R}$

L'estimateur de Pickands est le plus simple et le plus ancien a été proposé par Pickands en (1975) [26], cet estimateur a l'avantage d'être valable quel que soit le domaine d'attraction de $\mathcal{H}_\gamma$, $\gamma \in \mathbb{R}$.

Définition 2.1 (Estimateur de Pickands)
Soit $X_1, X_2, \cdots, X_n$ une suite de v.a's i.i.d de fonction de répartition $F \in \mathcal{D}(\mathcal{H}_\gamma)$, où $\gamma \in \mathbb{R}$. Soit une suite d'entier $k = k(n) \to \infty$ quand $n \to \infty$. L'estimateur de Pickands est donné par la statistique suivante :

$$\hat{\gamma}^p_{k(n)} := \frac{1}{\log 2} \log \frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}$$

Il a prouvé la convergence faible de son estimateur, la convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan (1989) [11].

Construction de l'estimateur de Pickands

D'après la relation (1.3), pour $\gamma \in \mathbb{R}$ avec le choix $t = 2s$, $x = 2$ et $y = 1/2$, on a

$$\lim_{t\to\infty} \frac{U(t) - U(t/2)}{U(t/2) - U(t/4)} = 2^\gamma. \tag{2.1}$$

En remplaçant dans (2.1) U par $U_n = (1/1 - F_n)^\leftarrow$ (F_n étant la fonction de répartition empirique) et on choisit $t = \dfrac{n}{k}$, où k est une suite intermédiaire, on a :

$$\lim_{n\to\infty} \frac{U_n\left(\dfrac{n}{k}\right) - U_n\left(\dfrac{n}{2k}\right)}{U_n\left(\dfrac{n}{2k}\right) - U_n\left(\dfrac{n}{4k}\right)} = 2^\gamma,$$

pour n assez grand du fait que $U_n(x) = X_{\left(\frac{n}{x}\right)}$, alors :

$$\frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}} = 2^\gamma,$$

l'estimateur de Pickands est alors la solution de cette dernière équation.

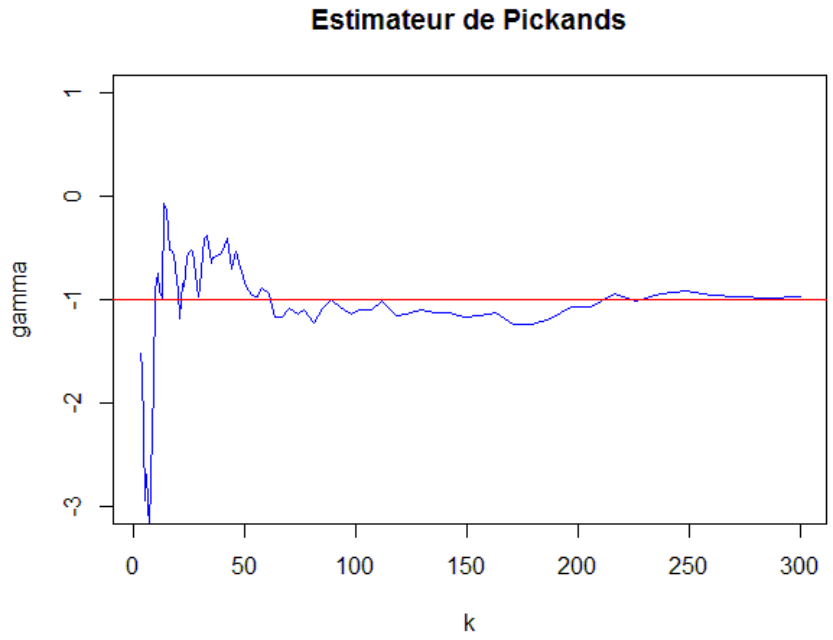

FIG. 2.1 – Représentation graphique de l'estimateur de Pickands

Propriétés asymptotiques de l'estimateur de Pickands

Théorème 2.1

Soit $X_1, X_2, \cdots, X_n$ *une suite de va's i.i.d de fonction de répartition* $F \in \mathcal{D}(\mathcal{H}_\gamma)$*, où* $\gamma \in \mathbb{R}$*. Si* $k = k(n) \to \infty$ *et* $\dfrac{k(n)}{n} \to 0$ *quand* $n \to \infty$*, alors*

– Consistance faible : $\hat{\gamma}_n^P$ converge en probabilité vers γ

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

– Consistance forte : Si $k/\log\log n \to \infty$ quand $n \to \infty$ alors $\hat{\gamma}_n^P$ converge presque sûrement vers γ

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

– Normalité asymptotique : La normalité asymptotique nécessite des conditions addition-

nelles sur la suite intermédiaire $k = k(n)$ et la fonction de répartition F

$$\sqrt{k}\left(\hat{\gamma}_n^p - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \eta^2\right) \ , \text{ quand } n \to \infty,$$

où

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

2.1.2 Estimateur de Hill $\gamma > 0$

L'estimateur de l'EVI le plus célèbre est l'estimateur de Hill qui a été introduit en 1975 et qu'il a été largement étudié dans le cas des v.a's i.i.d. Pour construire cet estimateur on utilise deux approche, la première approche est basée sur la GEV et la seconde approche POT basée sur la GPD a été démontré par de Haan et Ferreira [9].

Définition 2.2 (Estimateur de Hill)
Soit $X_1, X_2, \cdots, X_n$ une suite de v.a's i.i.d de fonction de répartition F appartenant au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $\gamma > 0$, soit une suite d'entier $k = k(n) \to \infty$ et $k/n \to 0$ quand $n \to \infty$. L'estimateur de Hill est défini par la statistique :

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

Mason (1982) [22] et Deheuvels et al.(1988) [10] ont montrés respectivement la consistance faible et forte qui ne dépend que du comportement de la suite $k(n)$. Pour établir la normalité asymptotique, on a besoin de supposer que la fonction de répartition F est à variation régulière du second ordre. Plusieurs auteurs ont obtenus cette normalité, notamment Davis et Resnick (1984) [8], Csorgö et Mason (1985) [6], Haeusler et Teugels (1985) [17].

Proposition 2.1 (Condition du premier ordre de Haan et Ferreira (2006)**)**
Les assertions suivantes sont équivalentes :

a) F à queue lourde

$$F \in \mathcal{D}\left(\Phi_{1/\gamma}\right),\ \gamma > 0.$$

b) $\bar{F}$ est à variation régulière à ∞ d'indice $-1/\gamma$

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

c) $Q(1-s)$ est à variation régulière à 0 d'indice $-\gamma$

$$\lim_{s\to 0}\frac{Q(1-sx)}{Q(1-s)} = x^{-\gamma},\ x>0.$$

d) U est à variation régulière à ∞ d'indice γ

$$\lim_{t\to\infty}\frac{U(tx)}{U(t)} = x^{\gamma},\ x>0.$$

Définition 2.3 (Fonction à variation régulière du second ordre)
On dit que la queue de $F \in \mathcal{D}(\Phi_\alpha)$, avec $\alpha = 1/\gamma$, est à variation régulière du second ordre, d'indice (γ,ρ) avec le paramètre du second ordre $\rho \leq 0$ (on note $\bar{F} \in 2\mathcal{RV}_{(\gamma,\rho)}$), à l'infinie si l'une des conditions équivalentes est satisfaite :

a) Il existe un paramètre $\rho \leq 0$ et une fonction A^* satisfaisant $\lim_{t\to\infty} A^*(t) = 0$ ne change pas son signe près de ∞, telle que pour tout $x>0$

$$\lim_{t\to\infty}\frac{(1-F(tx)/1-F(t)) - x^{-\alpha}}{A^*(t)} = x^{-\alpha}\frac{x^\rho - 1}{\rho}.$$

b) Il existe un paramètre $\rho \leq 0$ et une fonction A^{**} satisfaisant $\lim_{t\to\infty} A^{**}(t) = 0$ et ne change pas son signe près de 0 , telle que pour tout $x>0$

$$\lim_{s\to\infty}\frac{\left(Q\left(1-sx\right)/Q(1-s)\right) - x^{-\gamma}}{A^{**}(t)} = x^{-\gamma}\frac{x^\rho - 1}{\rho},\ \forall\ x>0$$

c) Il existe un paramètre $\rho \leq 0$ et une fonction A^* satisfaisant $\lim_{t\to\infty} A^*(t) = 0$ ne change pas son signe près de ∞, telle que pour tout $x>0$

$$\lim_{t\to\infty}\frac{\dfrac{U(tx)}{U(t)} - x^{\gamma}}{A(t)} = x^{\gamma}\frac{x^\rho - 1}{\rho}. \tag{2.2}$$

Si $\rho = 0$, $x^\rho - 1/\rho$ s'interprète comme $\log x$.
A, A^* et A^{**} sont des fonction à variation régulière avec $A^*(t) = A(1/\bar{F}(t))$ et $A^{**}(s) = A(1/s)$. Leurs rôle est de contrôler la vitesse de convergence.

Propriétés asymptotiques de l'estimateur de Hill

Théorème 2.2

Soit $X_1, X_2, \cdots, X_n$ *une suite de v.a's i.i.d de fonction de répartition* $F \in \mathcal{D}(\Phi_{\frac{1}{\gamma}})$. *Supposons que pour une suite intermédiaire* $k = k(n) \to \infty, \dfrac{k(n)}{n} \to 0$, *et* $k(n+1)/k(n) \to 1$, *quand* $n \to \infty$ *alors :*

– Consistance faible : $\hat{\gamma}_n^H$ converge en probabilité vers γ

$$\hat{\gamma}_n^H \xrightarrow{P} \gamma \ , \text{ quand } n \to \infty.$$

– Consistance forte : Si $k/\log\log n \to \infty$ alors $\hat{\gamma}_n^H$ converge presque sûrement vers γ

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

– Normalité asymptotique : Si la condition (2.2) est satisfaite avec $\lim_{n\to\infty} \sqrt{k} A\left(\dfrac{n}{k}\right) = \lambda$, alors

$$\sqrt{k}\left(\hat{\gamma}_n^H - \gamma\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right) \ , \text{ quand } n \to \infty.$$

Preuve. Voir le livre de de Haan, L. et Ferreira, A. (2006) [9]. ■

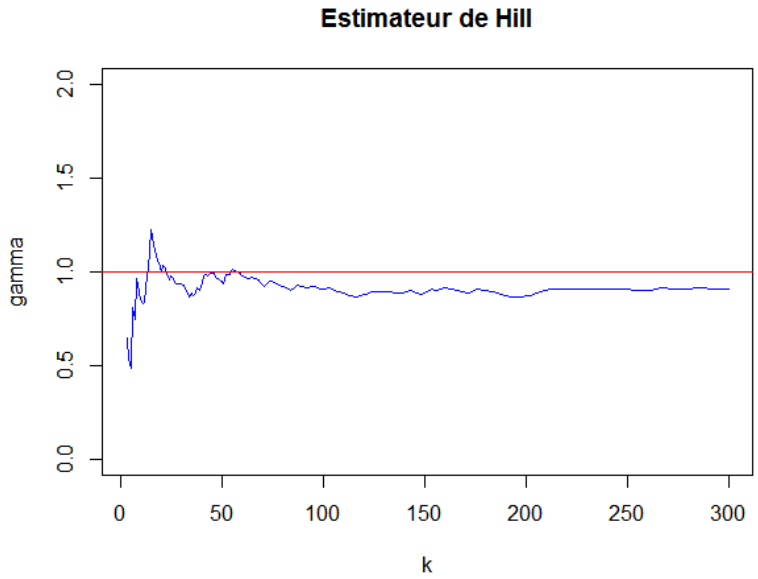

FIG. 2.2 – Représentation graphique de l'estimation de Hill

2.2 Notions de censure

2.2.1 Données de survie

L'analyse de survie est utilisée pour analyser des données dans lesquelles le temps jusqu'à la survenue d'un évènement est d'intérêt. La réponse est souvent appelée comme un temps d'échec ou de survie d'un évènement.

Exemple 2.1

- Temps jusqu'à la récurrence de la tumeur
- Temps jusqu'à la mort cardiovasculaire après un traitement
- Temps avant le sida pour les patients VIH

Définition 2.4 (Fonction de survie)
Soit X une v.a. positive et continue dite "durée de vie". Pour t fixé, la fonction de survie est la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = \bar{F}(t) = 1 - F(t) = 1 - P(X \leq t) = P(X > t)$$

2.2.2 Données de censure

La censure est présente lorsque nous avons des informations sur l'évènement d'intérêt, mais nous ne connaissons pas l'heure exacte de cet évènement.

Pour l'individu i, considérons

- son temps de survie X_i
- son temps de censure Y_i
- la durée réellement observée Z_i

Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'évènement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

La censure de type I

Soit Y une valeur fixée, au lieu d'observer les variables $X_1, \cdots, X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq Y$, sinon on sait uniquement que $X_i > Y$. on utilise la notation suivante :

$$Z_i := X_i \wedge Y = \min(X_i, Y).$$

Exemple 2.2 *Dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note X la durée d'apprentissage de cette langue. Pour certains étudiants nous allons observer leurs durées X_i d'apprentissage de la langue par contre pour d'autres leurs X_i ne seront pas observées car le stage est limité dans le temps.*

La censure de type II

Elle présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment-là. Soient $X_{i,n}$ et $Z_{i,n}$ les statistiques d'ordre des v.a's X_i et Z_i. la date de censure est donc $X_{k,n}$ et on observe les variables suivantes :

$$\begin{aligned} Z_{1,n} &= X_{1,n} \\ &\vdots \\ Z_{k,n} &= X_{k,n} \\ Z_{k+1,n} &= X_{k,n} \\ &\vdots \\ Z_{n,n} &= X_{k,n}. \end{aligned}$$

La censure de type III (ou censure aléatoire de type I)

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expérience, la date d'inclusion du patient dans l'étude est fixé, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'évènement observés et la durée totale de l'expérience sont aléatoires.

Soient $Y_1, \cdots, Y_n$ des v.a's i.i.d. On observe les variables

$$Z_i := X_i \wedge Y_i.$$

L'information disponible peut être résumée par :

1. la durée réellement observée Z_i.

2. un indicateur $\delta_i := \mathbb{1}_{\{X_i \leq Y_i\}}$, où $\mathbb{1}_{\{.\}}$ est une fonction indicatrice.
   - $\delta_i = \begin{cases} 1 & \text{si l'évènement est observé (d'où } Z_i = X_i\text{). On observe les durées complètes,} \\ 0 & \text{si l'individu est censure (d'où } Z_i = Y_i\text{). On observe les durées incomplètes.} \end{cases}$

Exemple 2.3 *Lors d'un essai thérapeutique, on peut citer certaines causes entraînant la censure aléatoire :*

1) *Perdu de vue : le patient peut décider de se faire soigner ailleurs à cause d'un déménagement et on le revoit plus*

2) *Arrêt du traitement : suite à des effets secondaire le traitement est arrêté.*

3) *Fin de l'étude : l'étude se termine et certains patients soit toujours vivants (exclus-vivants).*

Censure à gauche

Une durée de survie est dite censurée à gauche si l'individu a déjà subi l'évènement d'intérêt avant l'entrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple (Z, δ) :

$$Z := X \vee Y = \max(X, Y),$$
$$\delta := \mathbb{1}_{\{X \geq Y\}}.$$

Exemple 2.4

Notons X l'âge à laquelle une certaine maladie apparaît pour la première fois chez un individu. Après un examen médical on a reçu deux types de réponses :

1) *l'individu a déjà été malade mais l'âge exact de la première apparition n'a pas été retenu : Dans ce cas on n'a pas observé X mais on sait que X est inférieur à l'âge de l'individu lors de l'examen Y. Il s'agit d'une observation censurée à gauche.*

2) *l'individu n'a jamais eu de maladie : Dans ce cas on sait seulement que X est supérieur à l'âge de l'individu, donc on a une observation censurée à droite.*

Remarque 2.1 *Si les variables de censures sont dégénérée (c'est-à-dire constantes), alors on dit que la censure est fixée.*

Censure par intervalle

Une situation plus générale de la censure se produit lorsque la durée de survie n'est pas connue mais on sait seulement qu'il appartient à un certain intervalle. Ceci est le cas lorsque les patient dans les essais clinique ont des suivis périodiques, par exemple chaque six mois, si une maladie surgis, on sait seulement qu'elle est produite dans l'intervalle de temps séparant deux visites. Ce type de censure peut aussi apparaitre dans les expériences industrielles où il y a des inspections périodiques des machines.

- Dans ce mémoire, on s'intéresse uniquement au cas des censures à droite de type 3 (censure aléatoire à droite).

2.3 Estimateur de l'indice des valeurs extrêmes à la présence de censure

Nous travaillons dans l'espace de probabilité $(\Omega, \mathcal{A}, P)$. Soit $(X_1, \ldots, X_n)$ un échantillon des durées réelles de vie, n'est pas observé, mais qu'il est censuré par un deuxième échantillon $(Y_1, \ldots, Y_n)$, qui est supposé être indépendant du premier, où X_i et Y_i sont des v.a's i.i.d de fonction de répartition respectives F et G absolument continues, de points terminaux x_F et x_G (où $x_F = \sup\{x,\ F(x) < 1\}$) . Nous observons, de ce fait, uniquement $Z_i = X_i \wedge Y_i$, $i = 1, 2, \cdots, n$ et d'autre part des indicateurs de censure $\delta_i = \mathbb{1}_{\{X_i \leq Y_i\}}$. On note alors H la fonction de répartition commune des Z_i, de point terminale $x_H = \sup\{x,\ H(x) < 1\}$, satisfaisant :

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

On suppose que F et G appartiennent au domaine d'attraction de Fréchet ; $F \in \mathcal{D}\left(\Phi_{1/\gamma_1}\right)$ et $G \in \mathcal{D}\left(\Phi_{1/\gamma_2}\right)$, telles que :

$$\bar{F}(x) = x^{-1/\gamma_1} L_1(x) \quad \text{et} \quad \bar{G}(x) = x^{-1/\gamma_2} L_2(x),$$

avec $L_1(.)$ et $L_2(.)$ sont des fonctions à variations lentes. Alors :

$$\begin{aligned}\bar{H}(x) &= \bar{F}(x)\,\bar{G}(x)\\ &= x^{-1/\gamma_1}L_1(x)x^{-1/\gamma_2}L_2(x)\\ &= x^{-(\gamma_1+\gamma_2)/\gamma_1\gamma_2}\tilde{L}(x)\\ &= x^{\frac{-1}{\gamma}}\tilde{L}(x),\end{aligned}$$

où $\gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$, avec $\tilde{L}(x) = L_1(x)L_2(x)$. Par conséquent, H appartenant au domaine d'attraction de Fréchet d'indice $1/\gamma$ ($H \in \mathcal{D}\left(\Phi_{1/\gamma}\right)$).

Si F et G appartiennent au domaine d'attraction du maximum $F \in \mathcal{D}(H_{\gamma_1})$ et $G \in \mathcal{D}(H_{\gamma_2})$ respectivement, pour certain $\gamma_1, \gamma_2 \in \mathbb{R}$. Einmahll et al en 2008 [13], ont proposés les trois cas les plus intéressant suivants :

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

La méthode générale existante pour l'estimation de l'indice de queue en présence de censure à droite aléatoire, apparue d'abord dans Beirlant et al. (2007) [3] et développée dans Einmahl et al. (2008) [13], est à considérer tout estimateur consistant $\hat{\gamma}$ de l'EVI γ appliqué à l'échantillon $(Z_1, \ldots, Z_n)$ et deviser par la proportion $\hat{p}$ d'observation non censurées dans les plus grandes k valeurs de Z :

$$\hat{\gamma}^{(c,.)}_{z,k,n} := \frac{\hat{\gamma}^{(.)}_{z,k,n}}{\hat{p}}, \quad \text{où } \hat{p} := \frac{1}{k}\sum_{j=1}^{k}\delta_{[n-j+1,n]},$$

avec $\delta_{[1,n]}, \ldots, \delta_{[n,n]}$ les indicateurs de censure retenues correspondant à la statistique d'ordre $(Z_{1,n}, \ldots, Z_{n,n})$, respectivement. Il sera suivre que $\hat{p}$ estime $p = \dfrac{\gamma_2}{\gamma_1+\gamma_2}$.

$\hat{\gamma}^{(c,.)}_{z,k,n}$ peut être n'importe quel estimateur non adapté à la censure. En particulier, une adaptation de l'estimateur de Hill $\hat{\gamma}^{(H)}_{z,k,n}$ de l'indice γ_1 dans le cas de censure est défini par :

$$\hat{\gamma}^{(c,H)}_{z,k,n} := \frac{\hat{\gamma}^{H}_{n}}{\frac{1}{k}\sum_{j=1}^{k}\delta_{[n-j+1,n]}}$$

où

$$\hat{\gamma}_n^H := \frac{1}{k}\sum_{i=1}^{k} \log(Z_{n-i+1,n}) - \log(Z_{n-k,n}), 1 \leq k \leq n.$$

Alors

$$\hat{\gamma}_1^{(c,H)} = \frac{\sum_{i=1}^{k} \log(Z_{n-i+1,n}) - \log(Z_{n-k,n})}{\sum_{j=1}^{k} \delta_{[n-j+1,n]}}.$$

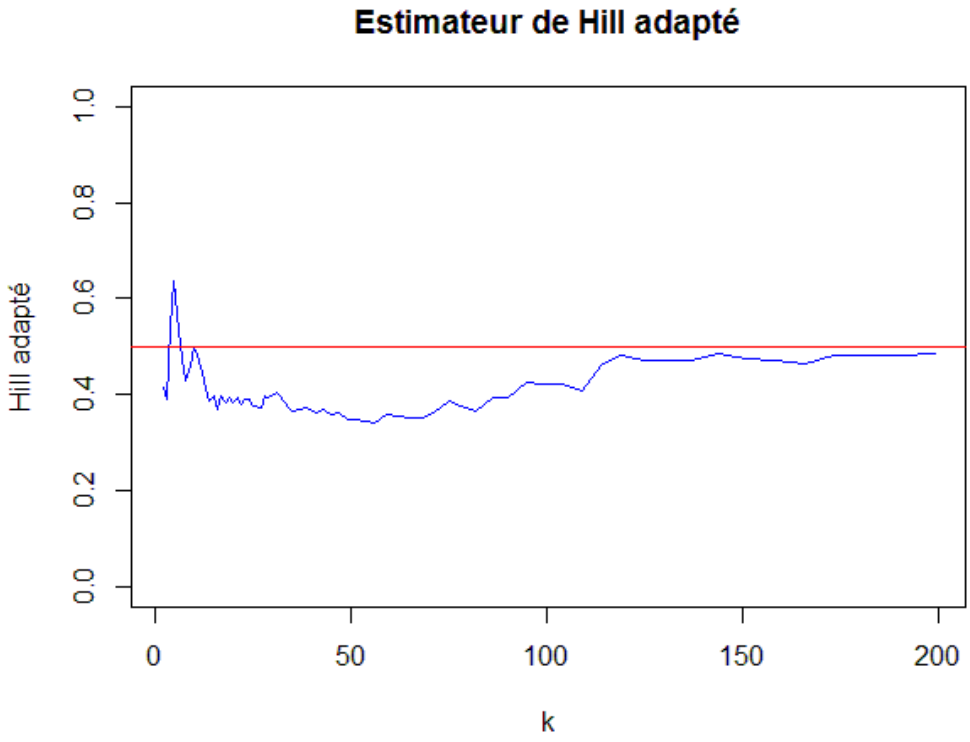

FIG. 2.3 – Représentation grafique de l'esimateur de Hill adapté

2.3.1 Propriétés asymptotique de l'estimateur de Hill adapté

Pour déterminer les propriétés asymptotiques de l'estimateur de l'EVI, adapté au cas de censure basé sur l'estimateur de Hill, nous avons besoin de la sous distribution des observations non-censurées, définie dans Einmahl et al (2008) [13] comme suit :

$$p(z) := P(X \leq Y \,|Z = z\,) = P(\delta = 1 \mid Z = z).$$

Nous pouvons l'écrire d'une autre manière

$$p(z) = \frac{\bar{G}(z)f(z)}{\bar{G}(z)f(z) + \bar{F}(z)g(z)}.$$

où f et g désignent respectivement les densités de X et Y, alors :

$$\lim_{z\to\infty} p(z) = \frac{\gamma_2}{\gamma_1 + \gamma_2} = p. \tag{2.3}$$

Pour expliquer la formule précédente (2.3), en supposant que nous sommes dans les cas (1) X et Y sont de *Pareto* (γ_1) et *Pareto* (γ_2) respectivement, c'est-à-dire pour tout $x \geq 1$, $\bar{F}(x) = x^{-1/\gamma_1}$ et $\bar{G}(y) = y^{-1/\gamma_2}$ avec $\gamma_1, \gamma_2 > 0$

$$\begin{aligned}\bar{H}(z) &= \bar{F}\,(z)\,\bar{G}(z)\\ &= z^{-1/\gamma_1} z^{-1/\gamma_2}\\ &= z^{-(\gamma_1+\gamma_2)/\gamma_1\gamma_2},\end{aligned}$$

ce qui implique que Z suit une loi de *Pareto* $(\gamma_1\gamma_2/\gamma_1 + \gamma_2)$. D'autre part maintenant, on trouve

$$p(z) = \frac{\gamma_2}{\gamma_1 + \gamma_2}.$$

Théorème 2.3

Soit $k = k(n)$ une suite intermédiaire telle que $1 < k \leq n$, $k \to \infty$ et $k/n \to 0$, si les conditions suivantes sont satisfaites :

C1 : Il existe $\rho < 0$ et une fonction à variations régulières $b(\cdot)$ d'indice ρ telles que pour tout

$x > 0$

$$\lim_{t\to\infty} \frac{H^{\leftarrow}\left(1-\frac{1}{tx}\right)/H^{\leftarrow}\left(1-\frac{1}{x}\right) - x^{\gamma}}{b(t)} = x^{\gamma}\frac{x^{\rho}-1}{\rho}.$$

C2 : $\sqrt{k}b\left(\frac{n}{k}\right) \to \alpha \in \mathbb{R}$.

C3 : $\sqrt{k}\sup_{\left\{1-k/n\leq t\leq 1, |t-s|\leq c\sqrt{k}/n, s<1\right\}} |p(H^{\leftarrow}(t)) - p\left(H^{\leftarrow}(s)\right)| \to 0$, pour $c > 0$

C4 : $\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n})) - p\right] \to \beta \in \mathbb{R}$.

Alors

$$\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,.)} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\frac{1}{p}(\alpha b_0 - \gamma_1\beta), \left(\frac{\sigma^2 + \gamma_1^2 p(1-p)}{p^2}\right)\right), \text{ quand } n \to \infty. \tag{2.4}$$

où αb_0, σ^2 représente le biais et la variance respectives de $\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(.)} - \gamma_1\right)$, avec

$$b_0 = \frac{1}{1-p},$$

et

$$\sigma^2 = \gamma^2.$$

Cela conduit au corolaire suivant, dont la preuve est plutôt simple pour l'estimateur de Hill :

Corollaire 2.1 (Normalité asymptotique de l'estimateur de Hill adapté)
Sous les conditions du théorème précédent, nous avons

$$\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\mu^{(c,H)}, \frac{\gamma_1^3}{\gamma}\right), \text{ quand } n \to \infty,$$

où

$$\mu^{(c,H)} := E\left(\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_1\right)\right) = -\frac{\gamma_1\beta}{p} + \frac{\alpha}{p}\frac{\gamma}{\tilde{\rho} + \gamma(1-\tilde{\rho})},$$

et

$$Var\left(\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_1\right)\right) = \frac{\gamma_1^3}{\gamma}$$

Pour spécifier le biais asymptotique de l'estimateur de Hill adapté, nous utilisons une condition de second ordre, s'il existe une fonction a positive et une fonction auxiliaire a_2 positive

telle que $\lim_{t\to\infty} a_2(t) = 0$, alors la fonction quantile de queue vérifiée

$$\lim_{t\to\infty} \frac{1}{a_2(t)} \left\{ \frac{H^{\leftarrow}\left(1-\frac{1}{tx}\right) - H^{\leftarrow}\left(1-\frac{1}{t}\right)}{a(t)} - h_\gamma(x) \right\} = k(x),$$

où $x > 0$ et $h_\gamma(x) = \int_1^x z^{\gamma-1} dz$.

S'il existe une constante $c \in \mathbb{R}$ et un paramètre de second ordre $\rho \leq 0$ alors

$$\lim_{t\to\infty} \frac{1}{a_2(t)} \left\{ \frac{a(tx)}{a(t)} - x^\gamma \right\} = c x^\gamma h_\gamma(x).$$

Si $0 < \gamma < -\rho$, alors nous avons une autre représentation de la fonction quantile de queue

$$U_H(t) = H^{\leftarrow}\left(1-\frac{1}{t}\right) = l_+ t^\gamma \left\{ \frac{1}{\gamma} + D t^{-\gamma} + \frac{A}{\gamma+\rho} a_2(t) (1 + o(1)) \right\},$$

où $A \neq 0$ et $D = \frac{1}{l_+} \lim_{t\to\infty} \left\{ U_H(t) - \frac{a(t)}{\gamma} \right\}$, $D \in \mathbb{R}$ et l_+ est une constante positive.

Dans l'énoncé de nos résultats, nous utilisons les notations suivantes :

$$b(x) := \begin{cases} \dfrac{A\rho\left[\rho + \gamma(1-\rho)\right]}{(\gamma+\rho)(1-\rho)} a_2(x) & \text{si} \quad 0 < -\rho < \gamma \text{ ou } 0 < \gamma < -\rho \text{ avec } D = 0, \\ -\dfrac{\gamma^3}{(1+\gamma)} x^{-\gamma} L_2(x) & \text{si} \quad \gamma = -\rho, \\ -\dfrac{\gamma^3 D}{(1+\gamma)} x^{-\gamma} & \text{si} \quad 0 < \gamma < \text{-}\rho \text{ avec } D \neq 0, \end{cases} \tag{2.5}$$

avec $L_2 = B + \int_1^x (A + o(1)) \dfrac{l_2(t)}{t} dt + o(l_2(x))$ pour un constant B et l_2 est une fonction à variation régulière.

Et

$$\tilde{\rho} := \begin{cases} \rho & \text{si} \quad 0 < -\rho < \gamma \text{ ou } 0 < \gamma < -\rho \text{ avec } D = 0, \\ -\gamma & \text{si} \quad 0 < \gamma < -\rho \text{ avec } D \neq 0. \end{cases} \tag{2.6}$$

Pour plus de détails vous pouvez consulter le papier de Einmahl et al (2008) [13].

2.4 Généralités sur les tests statistiques

Les tests statistiques constituent une approche décisionnelle de la statistique inférentielle. Un tel test a pour objet de décider (accepter ou rejeter) sur la base d'un échantillon aléatoire si une caractéristique de la population ou de la loi répond ou non à une certaine spécification que l'on appelle hypothèse. Cette dernière notée H, est une proposition logique contenant les caractéristque d'une ou plusieurs population, telle que des valeurs pour les paramètres, la forme de la distribution, l'indépendance entre deux variables, etc.
Soit l'hypothèse H_0 (dite hypothèse nulle) et l'hypothèse H_1 (dite hypothèse alternative), un test statistique a pour but de fournir une règle de décision permettant de choisir entre H_0 et H_1.

2.4.1 Région critique

La région critique W d'un test est l'ensemble de décision qui conduit à écarter H_0 au profit de H_1 et la région d'acceptation $\bar{W}$ de H_0 est le complément de la région critique. Que l'on accepte ou on rejette H_0, on prend le risque de commettre l'une des erreurs suivantes :

1. H_0 est rejetée à tort.

2. H_0 est retenue de façon injustifiée.

Nous sommes donc face à deux erreurs.

Risque d'erreur de 1ère espèce

Il représente la probabilité de rejeter H_0 alors qu'elle est vraie, en d'autre terme, accepter H_1 alors qu'elle est fausse. Il s'écrit

$$P(W\,|H_0\,) = \alpha.$$

Risque d'erreur de 2ème espèce

Il représente la probabilité d'accepter H_0 alors qu'elle est fausse, c'est-à-dire H_1 est vraie. Il s'écrit

$$P(\bar{W}\,|H_1\,) = \beta.$$

Niveau de signification

La valeur α est dite seuil ou niveau de signification du test. Les seuils de signification les plus courantes sont $\alpha = 1\%$ $\alpha = 5\%$ et $\alpha = 10\%$ et dépendant des conséquences de rejeter à tort l'hypothèse H_0.
La région d'acceptation $\bar{W}$ peut être reliée au risque d'erreur de première espèce α par :

$$P(\bar{W}\,|H_0\,) = 1 - \alpha,$$

cette valeur s'appelle niveau de confiance du test.

Puissance du test

L'erreur β est généralement exprimé par son complément à 1, appelée puissance du test (ou encore appelé la p-value), c'est la probabilité de rejeter H_0 tout en ayant raison.

$$P(W\,|H_1\,) = 1 - \beta.$$

Remarque 2.2

- Si la p-value est supérieure à α, il n'est pas exceptionnel sous H_0 d'observer la valeur effectivement observée. Par conséquent, H_0 n'est pas rejeté.
- Si la p-value est inférieure à α, la valeur observée est jugée exceptionnelle sous H_0. On décide alors de rejeter H_0 et de valider H_1.

Statistique du test

Le risque d'erreur de première espèce α étant fixé, il faut choisir une variable de décision encore appelée statistique de test. Cette variable est construite afin d'apporter de l'information sur le problème posé, à savoir le choix entre les deux hypothèses.
Pour résumer, la démarche d'un test est la suivante :

1) Choix de H_0 et H_1.
2) Détermination de la statistique du test.

3) Calcul de la région critique en fonction de α

4) Calcul éventuel de la puissance $1-\beta$

5) Calcul de la valeur expérimentale de la variable de décision.

6) Conclusion : rejet ou acceptation de H_0.

2.5 Test paramétrique de l'estimateur de Hill adapté

2.5.1 Formulation des hypothèses

Soit $X_1, \cdots, X_n$ une suite de v.a's dans le cas des données censurées aléatoirement à droite. Nous allons nous intéresser ici à un test paramétrique bilatéral de l'indice des valeurs extrêmes γ_1, selon son estimateur $\hat{\gamma}_{z,k,n}^{(c,H)}$ en se basant sur sa normalité asymptotique.(2.4), proposé par Meddi et al (2017) [23]. Les hypothèses à tester sont :

$$\begin{cases} H_0 : \gamma_1 = \gamma_0, \\ H_1 : \gamma_1 \neq \gamma_0, \end{cases}$$

avec γ_0 est une valeur spécifique.

2.5.2 Statistique du test

Le point de départ était le résultat du corolaire (2.1) et sous l'hypothèse de la normalité asymptotique de l'estimateur de Hill adapté, c'est-à-dire

$$\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(\mu^{(c,H)}, \frac{\gamma_1^3}{\gamma}\right), \text{ quand } n \to \infty.$$

La statistique qui convient pour ce test de γ_1, notée $S_{k,n}$, est déterminée par l'écart réduit

$$S_{k,n} := \frac{\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_1\right) - \mu^{(c,H)}}{\sqrt{\frac{\gamma_1^3}{\gamma}}},$$

où $S_{k,n}$ est aussi appelée fonction discriminante du test $\left(S_{k,n} \xrightarrow{d} \mathcal{N}(0,1), \text{ quand } n \to \infty\right)$.

Sous l'hypothèse H_0. En simplifiant la valeur $S_{k,n}$, on a :

$$\begin{aligned} S_{k,n} &= \frac{\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_0\right) - \mu^{(c,H)}}{\sqrt{\frac{\gamma_0^3}{\gamma}}} \\ &= \frac{\sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_0\right) - \left(-\frac{\gamma_1\beta}{p} + \frac{\alpha}{p}\frac{\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\right)}{\sqrt{\frac{\gamma_0^3}{\gamma}}}. \end{aligned}$$

On pose

$$\phi_{k,n} := \sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_0\right).$$

On obtient alors :

$$S_{k,n} := \phi_{k,n}\sqrt{\frac{\gamma}{\gamma_0^3}} - \frac{\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha\gamma}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)}\right) + \frac{\sqrt{\gamma}\left(\gamma_0\beta\right)}{p\sqrt{\gamma_0^3}}$$

On remarque que $S_{k,n} = S_{1,n} + S_{2,n} + S_{3,n}$ avec

$$\begin{aligned} S_{1,n} &= \phi_{k,n}\sqrt{\frac{\gamma}{\gamma_0^3}}, \\ S_{2,n} &= -\frac{\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha\gamma}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)}\right), \\ S_{3,n} &= \frac{\sqrt{\gamma}\left(\gamma_0\beta\right)}{p\sqrt{\gamma_0^3}}. \end{aligned}$$

Pour calculer $S_{1,n}$, $S_{2,n}$ et $S_{3,n}$, on remplace γ par $\dfrac{\gamma_0\gamma_2}{\gamma_0+\gamma_2}$, alors :

$$\begin{aligned} S_{1,n} &= \phi_{k,n}\sqrt{\frac{\frac{\gamma_0\gamma_2}{\gamma_0+\gamma_2}}{\gamma_0^3}} \\ &= \phi_{k,n}\sqrt{\frac{\gamma_2}{\gamma_0^2\left(\gamma_0+\gamma_2\right)}} \\ &= \frac{\phi_{k,n}}{\gamma_0}\sqrt{\frac{\gamma_2}{\gamma_0+\gamma_2}} \\ &= \frac{\phi_{k,n}}{\gamma_0}\sqrt{p}. \end{aligned}$$

Sous la condition (C2) du théorème (2.3), on a

$$\begin{aligned} S_{2,n} &= -\frac{\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha\gamma}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)}\right) \\ &= -\frac{1}{p}\sqrt{\frac{\gamma}{\gamma_0^3}}\left(\frac{\sqrt{k}b\left(\frac{n}{k}\right)\gamma}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)}\right) \\ &= -\frac{\sqrt{p}}{\gamma_0}\frac{\sqrt{k}b\left(\frac{n}{k}\right)\frac{\gamma_0\gamma_2}{\gamma_0+\gamma_2}}{p\left(\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)\right)} \\ &= -\frac{\sqrt{kp}b\left(\frac{n}{k}\right)}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)}. \end{aligned}$$

Sous la condition (C4) du théorème (2.3), on a

$$\begin{aligned} S_{3,n} &= \frac{\sqrt{p}}{\gamma_0}\frac{\gamma_0\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]}{p} \\ &= \frac{\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]}{\sqrt{p}}. \end{aligned}$$

Finalement, on obtient

$$S_{k,n} := \frac{\phi_{k,n}}{\gamma_0}\sqrt{p} - \frac{\sqrt{kp}b\left(\frac{n}{k}\right)}{\tilde{\rho}+\gamma\left(1-\tilde{\rho}\right)} + \frac{\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]}{\sqrt{p}}. \tag{2.7}$$

2.5.3 Région critique du test

Sous l'hypothèse H_0, pour une valeur fixé de l'erreur de première espèce α, on a :

$$\begin{aligned} P(W\,|H_0) &= \alpha \\ P(|S_{k,n}| &> z_{\frac{\alpha}{2}}) = \alpha \end{aligned}$$

avec $z_{\frac{\alpha}{2}}$ le quantile d'ordre $1-\dfrac{\alpha}{2}$ de la distribution normale centré réduite, la région critique est donnée par :

$$\left]-\infty\ ;\ -z_{\frac{\alpha}{2}}\right[\cup \left]z_{\frac{\alpha}{2}}\ ;\ \infty\right[$$

La région d'acceptation du test, notée par $\bar{W}$, est

$$P(\bar{W} \mid H_0) = 1 - \alpha$$
$$P(|S_{k,n}| \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

avec $1 - \alpha$ est le niveau de confiance du test

$$RC = \left[-z_{\frac{\alpha}{2}} \; ; \; z_{\frac{\alpha}{2}}\right]$$

2.5.4 Puissance du test

$$1 - \beta = P(\text{rejette } H_0 \mid H_0 \text{ est fausse})$$
$$= P\left(\bar{W} \; / \; H_1\right)$$

2.5.5 Règle de décision

On rejette H_0 au niveau de signification α si et seulement si la variable de décision $S_{k,n}$ est supérieur ou égal à la valeur de quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite. Si non on l'accepte, c'est que l'échantillon provient d'une distribution de paramètre γ_0.

2.6 Exemple et résultats de simulation

Dans cette section, nous illustrons la performance du test paramétrique de L'EVI présenté dans ce mémoire par une application numérique sur des données simulées. Toutes les simulations ont été effectuées sous le logiciel R.

2.6.1 Simulation des données

Exemple 1

On génère un échantillon de v.a's X, qui représente la durée de survie, issu d'une loi de $burr(1/\eta, \eta/\gamma_1)$, de taille $n = 1000$ et de fonction de répartition

$$F(x) := 1 - (1 + x^{\frac{1}{\eta}})^{-\frac{\eta}{\gamma_1}},\ x \geq 0 \text{ et } \gamma_1 > 0.$$

Cet échantillon est censuré à droite par un autre échantillon de v.a's Y, qu'il s'agit de la durée de censure, à partir d'une loi de $burr(1/\eta, \eta/\gamma_2)$, de la même taille, de fonction de répartition

$$G(x) := 1 - (1 + x^{\frac{1}{\eta}})^{-\frac{\eta}{\gamma_2}},\ x \geq 0 \text{ et } \gamma_2 > 0.$$

La génération de ces deux échantillons se fait à partir d'un autre échantillon de v.a's U et V respectivement, uniformément distribuées sur $[0, 1]$, à travers le calcul de la fonction inverse généralisée, telle que

$$F^{-1}(u) = \left((1-u)^{-\frac{\gamma_1}{\eta}} - 1\right)^{\eta} \text{ et } G^{-1}(v) = \left((1-v)^{-\frac{\gamma_2}{\eta}} - 1\right)^{\eta}.$$

Les variables que nous observons $Z_i = X_i \wedge Y_i$ sont d'une loi de $burr\,(1/\eta, \eta/\gamma)$ de fonction de répartition connue, définie par :

$$\begin{aligned} \bar{H}_Z(x) &= \bar{F}(x)\bar{G}(x) \\ &= (1 + x^{\frac{1}{\eta}})^{-\frac{\eta}{\gamma}}. \end{aligned} \tag{2.8}$$

Exemple 2

On génère un échantillon de v.a's X, issu d'une loi de $Paréto\,(\gamma_1)$, censuré à droite par un autre échantillon de v.a's Y à partir d'une loi de $Paréto\,(\gamma_2)$. Les deux échantillons ont la même la même taille $n = 1000$, les fonctions de répartition de X et Y sont respectivement :

$$F(x) := 1 - x^{-\frac{1}{\gamma_1}} \text{ et } G(x) := 1 - x^{-\frac{1}{\gamma_2}}.$$

La génération de ces deux échantillons se fait par la méthode d'inverse, donc on génère deux échantillons de v.a's U et V, issus d'une loi uniforme sur $[0, 1]$, on calcul la fonction inverse

généralisée

$$F^{-1}(u) = (1-u)^{-\gamma_1} \text{ et } G^{-1}(v) = (1-v)^{-\gamma_2}.$$

Les variables $Z_i = X_i \wedge Y_i$ sont d'une loi de *paréto* (γ) de fonction de répartition, définie par :

$$\begin{aligned} \bar{H}_Z(x) &= \bar{F}(x)\bar{G}(x) \\ &= x^{-\frac{1}{\gamma}}. \end{aligned} \tag{2.9}$$

2.6.2 Choix du kopt

Pour déterminer le nombre optimal de valeurs extrêmes k, utilisé dans le calcul de l'estimateur de Hill, on applique l'algorithme de Reiss et Thomas. Il s'agit d'une méthode heuristique pour choisir le nombre des extrêmes utilisés dans l'estimation de l'indice du queue. Il suffit de choisir d'une façon automatique pour k_{opt} la valeur de k qui minimise :

$$\frac{1}{k}\sum_{i\leq k} i^{\beta} \mid \hat{\gamma}_n^{(c,H)}(i) - med(\hat{\gamma}_{1,n}^{(c,H)}, ..., \hat{\gamma}_{k,n}^{(c,H)}) \mid, 0 \leq \beta \leq 1/2, \tag{2.10}$$

où med dénote la médiane

2.6.3 Résultats de simulation du test

Exemple 1

Rappelons la forme simplifiée de la statistique du test

$$S_{k,n} := \frac{\phi_{k,n}}{\gamma_0}\sqrt{p} - \frac{\sqrt{kp}b\left(\frac{n}{k}\right)}{\tilde{\rho} + \gamma\left(1-\tilde{\rho}\right)} + \frac{\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n})) - p\right]}{\sqrt{p}}. \tag{2.11}$$

avec $p = \dfrac{\gamma_2}{\gamma_1+\gamma_2}$, on choisie la valeur 0.36 et 4 pour γ_1 et γ_2 respectivement.

$\phi_{k,n} = \sqrt{k}\left(\hat{\gamma}_{z,k,n}^{(c,H)} - \gamma_0\right)$, $\hat{\gamma}_{z,k,n}^{(c,H)}$ est l'estimateur de Hill adapté et $\gamma = \dfrac{\gamma_2\gamma_0}{\gamma_0+\gamma_2}$.

Ainsi que $b\left(\frac{n}{k}\right)$ est une fonction à variation régulière calculer à partir (2.5), avec $\rho = -1$, voir Einmahl et al (2008) [13].

Fixons $\eta = 1/4$. En déduire la fonction quantile de queue à partir (2.8)

$$H^{-1}(x) = \left[\left(1-x\right)^{-\frac{\gamma}{\eta}} - 1\right)\right]^{\eta}.$$

Pour γ_0 fixé, les hypothèses à tester sont :

$$\begin{aligned} \gamma_1 &= 0.36 \\ \gamma_1 &\neq 0.36 \end{aligned}$$

Le test paramétrique de l'estimateur de Hill sous données aléatoirement censurées à droite, issu d'une loi de burr, donne les résultats numériques suivants :

	résultats numérique
kopt	100
γ_1	0.3552032
$b(n/kopt)$	-0.01265935
$\tilde{\rho}$	-0.3302752
$H^{\leftarrow}(1-\frac{1}{i})$	1.021388
$\phi_{k,n}$	-0.04796821
$\lvert S_{k,n}\rvert$	1.090604
Qauntile de test	1.959964

TAB. 2.1 – Résultats numériques du test paramétrique de l'indice de valeurs extrêmes en présence de censure à droite

- Decision de test : Donc on accepté l'hypothèse $H_0 : \gamma_0 = \gamma_1 = 0.36$.

Exemple 2

Pour calculer la statistique du test, on refait la même procédure que pour l'exemple 1, mais cette fois ci par l'utilisation de la fonction quantile de queue déduit à partir (2.9)

$$H^{-1}(x) = (1-x)^{-\gamma}.$$

Pour γ_0 fixé, les hypothèses à tester sont :

$$\begin{aligned} \gamma_1 &= 0.31 \\ \gamma_1 &\neq 0.31 \end{aligned}$$

Le test paramétrique de l'estimateur de Hill sous données aléatoirement censurées à droite, issu d'une loi de paréto, donne les résultats numériques suivantes :

	résultats numérique
kopt	21
γ_1	0.3006385
$b(n/kopt)$	-0.006085762
$\tilde{\rho}$	-0.287703
$H^{\leftarrow}(1-\frac{1}{i})$	0.180913
$\phi_{k,n}$	-0.04289962
$\lvert S_{k,n}\rvert$	0.2322475
Qauntile de test	1.959964

TAB. 2.2 – Résultats numériques du test paramétrique de l'indice de valeurs extrêmes en présence de censure à droite

- Decision de test : Donc on accepté l'hypothèse $H_0 : \gamma_0 = \gamma_1 = 0.31$.

Conclusion

Dans ce mémoire, nous avons présentés une adaptation de l'estimation semi-paramétrique de l'indice des valeurs extrêmes proposée par Einmahl et al (2008) et nous somme intéressées a un test paramétrique de cet indice, sous le domaine d'attraction de Fréchet, en présence des données i.i.d censurées aléatoirement à droite. Notre point de départ est le travail de Meddi et al (2017), avec un autre choix du nombre optimal de valeurs extrêmes basé sur l'algorithme de Reiss et Thomas.

A la lumière des résultats obtenus, on peut tirer comme conclusion que, le test présenté dans ce mémoire et qui assure la propriété de la normalité asymptotique de l'estimateur de Hill adapté, nous fournit un moyen utile d'ajustement optimal d'un échantillon de variables aléatoires de distributions extrêmes (appartient a un domaine d'attraction de Fréchet dans notre travail) lorsque le paramètre γ_1 est inconnu.

Bibliographie

[1] Anis Borchani,(2011) Statistiques des valeurs extrêmes dans le cas de lois discrètes HAL Id : hal-00572559.

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

[3] Beirlant et al (2007) Estimation of the extreme value index and high quantiles under random censoring.Article Université Leuven et Université Paris.

[4] Boualam,K.2017.Etude de l'estimateur de Hill sous dépendance faible,.Thése de doctorat de Université Mouloud Mammeri,Tizi-ouzou, Algeria.

[5] Biost (2004) Introduction to survival analysis, blast 515, Lecture 15.

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

[7] David, H. A., and Nagaraja, H. N. (2003). Order Statistics, Third Edition.john Wiley.

[8] Davis, R., and Resnick, S. (1984). Tail estimates motivated by extreme value theory. The Annalsof Statistics. 12(4). 1467-1487.

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

[10] Deheuvels, P., Haeusler, E., and Mason, D. (1988) Almost sure convergence of the hill estimator. Mathematical Proceedings of the Cambridge Philosophical Society. 104(2). 371-381.

[11] Dekkers.A.L.M, and Haan.L, D. (1989). On the estimation of the extreme-value index and largequantile estimation. The Annals of Statistics. 17(4). 1795-1832.

[12] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). Modelling Extremal Event for Insurance and Finance. Springer, Berlin.

[13] Einmahl et al, (2008). Statistics of extremes under random censoring.Bernoulli 14 :1, 207-227.

[14] Fisher, R., Tippet, L.(1928) Limitting forms the frequency distribution of the largestor smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24 :180–190.

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

[16] Gilbert Saporta (2006) Probabilité analyses des données et statistique . Edition thechnip 27 rue Ginoux, 75737 Paris Cedex 15, France.

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

[18] Halima Boudada (2012) .Quantile conditionnel pour des données incomplétes et dependentes. Mémoire de magistère de Universté Mentouri - Constantine.

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

[20] http ://www.statelem.com/statistique_d_ordre.php.

[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] Mason, D. (1982), Laws of large numbers for sums of extreme values. The Annals of Probability.10(3), 754-764.

[23] Meddi et al (2017). Construction d'un test paramétrique De l'estimateur de l'indice des valeurs extrêmes censurées. Mémoire de master Université Kasdi Merbah Ourgla.

[24] Ndao,P.(2015), Modélisation de valeurs extrêmes conditionnelles en présence de censure,thése de doctorat,université Gaston berger de Saint-Louis.

[25] Philippe.S.P, (2015) Introduction à l'analyse des données de survie, Université Pierre et Marie curie.

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

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

[28] Samah BATEKA (2010). Determination du nombre de statistiques d'ordre extrême. Thése de doctorat de université Mohamed khider, Biskra, Algeria.

[29] Smith, R.L., 1987. Estimating tails of probability lois. The Annals of Statistics. 3, 1174-1207

[30] Von Mises R. (1954), La distribution de la plus grande de n valeurs. Amer. Math. Soc.,2 :271–294

[31] Worms ,J., Worms, R., 2013, New estimators of the extreme value index under randomn right censoring, for heavy-tailed distributions.

[32] Zouadi Nihad et SAIDI Ghania (2018) Estimation of extreme values index in the presence of censored data .Volume : 11 / N° : 02(2018), p 520- 535.

Annexe : Abréviations et Notations

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

$\mathcal{D}$	Domaine d'attraction
EVI, γ	Indice des valeurs extrêmes
F	Fonction de répartition
F_n	Fonction de répartition empirique
$F^{\leftarrow}$	Inverse généralisée de F
GEV	Distribution des valeurs extrêmes généralisés
GPD	Distribution de Pareto généralisée
$\mathcal{H}_\gamma$	Distribution des valeur extrême
$i.i.d$	Indépendente et identiquement distribuées
$\mathbb{1}_A$	Fonction indicatrice
Λ	Loi Gumbel
$X_{n,n}$	Maximum de $X_1, \cdots, X_n$
POT	Pics au_delà d'un seuil
Φ	Loi de Fréchet
Ψ	Loi de Weibull
$S = \bar{F}$	Fonction de survie
$\mathcal{N}(0,1)$	Loi Gauss
$v.a$	variable aléatoire

$(\Omega, \mathcal{A}, P)$	Espace de probablité
x_F	Point terminale
$:=$	Egalité en définition
$\xrightarrow{d}$	Converge en distribution
$\xrightarrow{p.s}$	Converge prèsque sûre
$\xrightarrow{p}$	Converge en probabilité
$\mathcal{RV}_\alpha$	Variation régulière
$\hat{\gamma}_n^{(c,H)}$	Estimateur de Hill sous données censurées
al	Autres