UNIVERSITE KASDI MERBAH OUARGLA
Faculté des Mathématiques et des Sciences de la Matière

Mémoire
MASTER ACADEMIQUE
Domaine : **Mathématiques**
Spécialité : **Probabilité et Statistique**
Présenté par :
Ramdani Hayat
Thème:

Construction d'un test paramétrique De l'estimateur de l'indice des valeurs extrêmes censurées.

Soutenu publiquement
Le :..29/.05./…2017.
Devant le jury :

Said ZIBAR	M.C.A	Président	UKM Ouargla
Fatima Meddi	Pr	Encadreur/rapporteur	UKM Ouargla
Agoune Rachid	M.C.B	Examinateur	UKM Ouargla

Année universitaire 2016/2017

Dédicaces

Je dédie ce modeste travail

A ma très chère Mère et mon très cher Père

A mes chéres sœurs et frères

A toute la famille RAMDANI

A ceux qui ont veillé pour mon bien être

A ceux qui m'ont toujours encouragé pour que je réussisse dans mes études

A tout ce qui m'ont encouragé lors de la réalisation de mon travail

Sans oublier de dédier ce mémoire à mes très chères amies intimes

et surtout mon amie Ndjet Belatrach

Finalement à tous ceux qui m'ont aidé de proche ou de loin

REMERCIEMENTS

Tout d'abord je remercie ALLAH qui m'a donné la volonté et le courage pour pouvoir réaliser ce travail.

Je voudrais d'abord et avant tout remercier ma encadreur vertueuse Meddi Fatima pour tous ses efforts en vue d'établir ce mémoire, elle a eu le rôle fondamental et essentiel et la grand mérite dans tout ce qui a été réalisé, comme cela avait toujours été, près de moi et me guider et corriger mes erreurs me donner de précieux conseils et les conseils appropriés et les alertes considéres, je la répète mes remerciements pour tout ce qu'elle a fait pour moi à travers la mise en plan à toutes les étapes de la préparation de ce mémoire depuis le début jusqu'à la fin, étape par étape, était que le premier et dernier facteur pour le succès de ce travail, et je lui dis encore une fois merci beaucoup pour votre appréciation profonde.

J'exprime ma gratitude à ma famille qui m'ont toujours soutenue et encouragée dans la voie que je m'étais fixée. Je remercie particulièrement mes parents qui m'ont stimulée et encouragé pendant mes études. qui étaient toujours prêts à fournir tous les moyens physiques et morales pour la réussite de ce projet.

Table des matières

0.1 Notations et conventions

EVD	Distribution des valeurs extrêmes
EVI, γ	Indice des valeurs extrêmes
F	Fonction de répartition
F_n	Fonction de répartition empirique
$F^{\leftarrow}$	Inverse généralisé de F
GEV	Distribution des valeurs extrêmes généralisée
GPD	Distribution de pareto généralisée
G_γ	Famille de la loi de valeurs extrêmes généralisée
$i.i.d$	Indépendantes et identiquement distribuées.
$\mathbb{I}_{\{A\}}$	Fonction indicatrice de l'ensemble A
Λ	Loi de Gumbel
$\ell(x)$	Fonction à variation lente
DA	Domaine d'attraction de maxcimum
$M_n = X_{n,n}$	Maxcimum de $X_1, ..., X_n$
N_u	Nombres des excés qui dépassent du seuil u
POT	Pics au-delà d'un seuil
$p.s$	Prèsque sûre
Φ	Loi de frêchet
Ψ	Loi de weibull
$resp$	Respectivement
$S = \overline{F}$	$1 - F$ fonction de survie
TEV	Théorèm des valeurs extrêmes
$X_{1:n}, ..., X_{n,n}$	Statistique d'ordre associées à $X_1, ..., X_n$
$X \wedge Y$	$\min(X, Y)$
x_F	Point terminal
$\underset{=}{\mathcal{L}}$	Égalité en loi
$:=$	Égalité en définition
$\underset{\rightarrow}{D}$	Converge en distribution

$s.o$	Statistique d'ordre
VR_α	Variation régulière d'indice α
$\hat{\gamma}^{(c,H)}$	Estimateur de Hill avec les données censurées
$al.$	Autres
τ_H	Point terminal
IC	Intervalle de confiance
$v.a$	variable aléatoire
$(\Omega, \mathcal{A}, \mathbb{P})$	Espace probabilisé
TCL	Théorème Centrale Limite
$N(0,1)$	Loi normale standard
inf A	Supremum de l'ensemble A
$\overset{p.s}{\rightarrow}$	Converge prèsque sûre
$\overset{l}{\rightarrow}$	Converge en loi
$\overset{d}{\rightarrow}$	Converge en distribution
$\overset{p}{\rightarrow}$	Converge en probabilité

0.2 Introduction

La théorie des tests est l'une des deux branches de la statistique mathématique. Elle se subdivise en deux volets principaux, les tests non-paramétriques et les tests paramétriques.

Notre choix s'est porté sur des tests paramétriques et plus particulièrement sur les tests d'hypothèses dans le domaine des valeurs extrêmes vu leur importance dans les événements rares. Ces dernières sont des événements dont la probabilité d'apparition est trop faible c'est-à-dire se trouve dans les queues des distributions. Ils apparaissent en général dans des contextes physiques nombreux et variés en particulier les catastrophes naturelles: en hydrologie (crues décennales ou centennales , hauteur des barrages et digues susceptibles des contenir) tempêtes occasionnant d'importants dommages matériels et environnementaux, dans les grands incendies, dans le tremblements de terre, dans les risques financiers (les krachs boursiers, les crises financières), etc.

La théorie des valeurs extrêmes est une branche de la statistique qui essaie d'amener une solution face à ces phénomènes. Elle se repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Cependant, on y retrouve deux modèles: loi généralisée des extrêmes (GEV: « Generalized Extreme Value ») et loi de Paréto généralisée (GPD: « Generalized Pareto Distribution »). Ainsi, tout a commencé avec les auteurs *Fisher et Tippet (1928)* quand ils étudiaient la résistance des fils de coton puis plus tard *Gnedenko (1943)* s'est intéressé à ces distributions. Ils ont énoncé un théorème fondamental avec la création de trois domaines d'attraction: domaine d'attraction de Fréchet, Gumbel et Weibull. Ce théorème intéressant fait référence à un paramètre appelé l'indice de queue qui donne la forme de la queue de distribution. En effet, C'est en ce moment que plusieurs auteurs se sont focalisés aux estimations de l'indice des valeurs extrêmes. Nous pouvons citer Hill (1975) dans le cas où l'indice est positif $\gamma > 0$.

La modélisation des valeurs extrêmes censurées voit le jour la première fois en 1997 dans la littérature des extrêmes avec la sortie du livre Reiss et Thomas. Il a fallu qu'en 2007 Beirlant et al. abordent réellement la statistique non paramétrique des valeurs extrêmes avec des données censurées. Leur traveaux sont basés sur l'estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dépassant un certain seuil donné. *Einmahl et al. (*2008) ont proposé un estimateur de l'indice de queue

sur les k-plus grandes valeurs ensuite ils ont déterminé ses propriétés asymptotiques et enfin illustrer son comportement . Depuis, la recherche sur la théorie des valeurs extrêmes censurées est devenue une actualité.

Ce mémoire est réparti en trois chapitres:

Le premier Chapitre. Dans la *section 1.1*, nous rappelons quelques éléments théoriques essentiels de la théorie des valeurs extrêmes (TVE), nous présontons les résultats sur les distributions asymptotiques en théorie des valeurs extrêmes. On s'intéressera ensuite à la caractérisation des domaines d'attraction. Cette étude faisant appel à la notion de fonctions à variations régulières ensuite l'estimateur classique de l'indice de queue de *Hill* (1975) $\hat{\gamma}^H$, ainsi les propriétés asymptotiques. Dans la *section 1.2*, on présente un seul cas de données incomplètes : censurées (droite, gauche et par intervale). Dans la *section 1.3*, on présente des notions et definitions de base sur la théorie des tests d'hypothèse et quelques éléments fondamentaux.

Le deuxième Chapitre. Dans la *section 2.1,* on va s'intéresser au problème de l'estimation de l'indice de queue en présence de données censurées aléatoirement à droite, *Einmahl et al.* (2008) ont utilisé le concept pour proposer un estimateur de l'indice de queue sur les k-plus grandes valeurs de l'échantillon ensuite déterminer ses propriétés asymptotiques. Dans la *section 2.2,* on appliquer un test d'hypothèse paramétrique sur cet estimateur de l'indice (l'indice des valeurs extrême dans le cas censurée γ_1 inconnu), basé sur sa normalité asymptotique, présentées dans le théorème 1.1.1 *Einmahl et al.* (2008)

Le troisième Chapitre. Nous avons présenté un algorithme détaillé de notre test paramétrique. Ainsi le choix du nombre de statistiques d'ordre extrêmes k optimal utilisés, par trois méthodes *cheng et peng* , *l'erreur quadratique moyenne* et celle de *Reiss et thomes.* Nous avons effictué des simulations par le logiciel R pour rendre notre test plus pratique et utilisable. Nos résultats, nous ont permet de donner une estimation ponctuelle par inervalle pour l'indice γ_1, ce qui semble être trés satisfaisant pour nos attentes.

Chapitre 1

Théorie des valeurs extrêmes et notions de tests paramétriques

Dans cette partie, nous rappelons quelques notions essentielles sur la théorie des valeurs extrêmes et la notion de censure qui permettent de faciliter la lecture du mémoire. Nous définissons rapidement les domaines d'attractions, les fonctions à variations régulières puis nous les caractérisons dans le cas unidimensionnel. Quant à la censure nous présenterons quelques définitions liées à la statistique des durées de survie, oú le censure est fondée avec quelques fonctions telles que la fonction de répartition, la fonction de survie. Comme notre travail porte sur un test paramétrique de l'indice des valeurs extrêmes, nous présontons des definitions de base de la théorie des tests d'hypothèse.

1.1 Présentation de la théorie des valeurs extrêmes

La théorie des valeurs extrêmes (TVE) communément appelée "Extreme Value Theory" (EVT) en anglais, est une vaste théorie dont le but est d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'apparition est faible. Autrement dit, elle essaie d'amener des éléments de réponse aux intempéries, aux inondations, aux catastrophes naturelles, aux problèmes financiers,...etc. les ouvrages de *Reiss et Thomas (1997), Embrechts et al.(2007), Beirlant et al.(2007),* qui font le point sur les différentes techniques existantes.

1.1.1 Statistiques d'ordre

Les statistiques d'ordre jouent un rôle de plus en plus important dans la théorie des valeurs extrêmes, parce qu'ils fournissent des informations sur la distribution de queue (à droite). On les rencontre, en effet, de façon naturelle et depuis long-temps, dans les problèmes de données censurées ou tronquées.

Définition 1.1.1 . *Soit* $X_1, X_2, ..., X_n$ *n variables aléatoires iid de distribution commune* F. *les va's* $X_{1,n}, X_{2,n}, ..., X_{n,n}$ *sont rangés par ordre croissant, soit:*

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

sont appelées les statistiques d'ordre de l'échantillon $X_1, X_2, ..., X_n$.

Deux statistiques d'ordre sont particulièrement intéressantes pour l'étude des événements extrêmes. Ce sont les statistiques d'ordre extrêmes qui sont données par la définition suivante:

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

On note qu'il est très facile de passer de l'un à l'autre à l'aide de la relation :

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

Dans la suite de ce memoire, on se concentrera sur l'étude du maximum.

Distributions du maximum et du minimum

Loi de $X_{i,n}$.

$$F_{i,n} = \mathbb{P}\{X_{i,n} \leq x\} = \sum_{r=i}^{n} C_r^n (F(x))^r (1 - F(x))^{n-r}.$$

Nous en déduisons que la fonction de densité est:

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

où $f(x)$ est la densité de probabilité de X_i et F sa fonction de répartition associée.

Loi de $X_{1,n}$.

$$F_{1,n}(x) = \mathbb{P}\{X_{1,n} \leq x\} = 1 - (1 - F(x))^n,$$

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

En utilisant la propriété d'indépendance des variables aléatoires $X_1, ..., X_n$ nous en déduisons que :

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

d'où,

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

Loi de $X_{n,n}$.

$$F_{n,n}(x) = \mathbb{P}\{X_{n,n} \leq x\} = (F(x))^n,$$

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

En utilisant la propriété d'indépendance des variables aléatoires $X_1, ..., X_n$ nous en déduisons que :

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

d'où,

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

Définition 1.1.2 *(La fonction de répartition empirique). La fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ notée F_n est donnée par:*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{]-\infty,x[}(X_i), \quad x \in \mathbb{R}.$$

Il existe une autre version de la définition de F_n en utilisant les $(s.o)$ comme suit:

$$F_n(x) = \begin{cases} 0 & si, \quad x \leq X_{1,n} \\ \frac{i-1}{n} & si, \quad X_{i-1,n} < x \leq X_{i,n}, \quad 2 \leq i < n \\ 1 & si, \quad x > X_{n,n} \end{cases}$$

Définition 1.1.3 *(Les fonctions de quantile et de quantile de queue). On définit la fonction des quantiles Q par :*

$$Q(s) = F^{\leftarrow}(s) := \inf\{x \in \mathbb{R} : F(x) \geq s\}, \quad 0 < s < 1,$$

où $F^{\rightarrow}$ est l'inverse généralisée de F. Dans la théorie des extrêmes une fonction, notée par U et (parfois) appelée la fonction quantile de queue, est utilisée assez souvent et elle est définie comme:

$$U(t) := Q(1 - 1/t) = (1/\overline{F})^{\leftarrow}(t), \quad 1 < t < \infty,$$

où,

$$\overline{F}(t) := 1 - F(t)$$

Définition 1.1.4 *(La fonction de survie). La fonction de survie est pour t fixé, la probabitité de survivre jusqu'à l'instant t, c'est-à-dire : pour* $t \geq 0$

$$\begin{aligned} S(t) &= 1 - F(x) \\ &= P(X > t) \end{aligned}$$

Définition 1.1.5 *(Théorème Central Limite). Soit* $X_1, ..., X_n$ *une suite de variables aléatoires iid de moyenne* μ *et de variance* σ^2 *finie, alors:*

$$\sqrt{n}\frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0;1) \quad quand \quad n \to \infty$$

Comportement asymptotique des extrêmes

Nous posones

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

Nous tirons la conclusion que le maximum M_n est une variable aléatoire dont la fonction de répartition est égale à $(F_X)^n$. La fonction de répartition de X étant souvent inconnue et généralement pas possible d'être déterminée. Notons $x_F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$ le point terminal à droite de la fonction de répartition F_X. Ce point terminal peut être infini ou fini *(Embrechts et al.* 1997*)*. On s'intéresse ici au distribution asymptotique du maximum, en faisant tendre n vers l'infini,

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty} [F_X(x)]^n = \begin{cases} 0 & \text{si } F(x) < 1 \\ 1 & \text{si } F(x) = 1 \end{cases}$$

On constate que la distribution asymptotique du maximum, donne une loi dégénérée, une masse de Dirac en x_F, puisque pour certaines valeurs de x, la probabilité peut être égale à 1 dans le cas où x_F est fini. Donc M_n tend vers x_F presque sûrement, Ce fait ne fournit pas assez d'informations, d'où l'idée d'utiliser une transformation afin d'obtenir des résultats plus exploitables pour les loi limites des maxima M_n. On s'intéresse par conséquent à une loi non dégénérée pour le maximum, la théorie des valeurs extrêmes permet de donner une réponse à cette problématique. Les premièrs résultats sur la caractérisation du comportement asymptotique des maxima M_n convenablement normalisés et donnés par la suite.

Distributions des Valeurs Extrêmes

Théorème 1.1.1 *(Fisher et Tippett, 1928, Gnedenko, 1943). Soit $X_1, ..., X_n$ une suite de n variables aléatoires réelles iid de loi continue P et $M_n = \max(X_i)_{1\leq i\leq n}$. S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$, et une fonction de répartition non-dégénérée G_γ telle que,*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n\to\infty} [F(a_n x + b_n)]^n = G_\gamma(x), \quad \forall x \in \mathbb{R}$$

Alors G_γ est du même type qu'une des trois lois suivantes:

$$\textit{loi de Gumbel : } \Lambda_\gamma(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty,$$

$$\textit{loi de Fréchet : } \Phi_\gamma(x) = \begin{cases} 0 & x < 0 \\ \exp\left(-x^{-1/\gamma}\right) & x \geq 0, \gamma > 0, \end{cases}$$

$$\textit{loi de Weibull : } \Psi_\gamma(x) = \begin{cases} \exp\left(-(-x)^{-1/\gamma}\right) & x < 0, \gamma < 0, \\ 1 & x \geq 0, \end{cases}$$

avec, G_γ est la loi des valeurs extrêmes, γ est l'indice des valeurs extrêmes et a_n et b_n sont des paramètres de normalisation. Ce théorème donne la forme des lois limites de G_γ.

1.1.2 Distributions GEV et GPD

Distributions GEV

Jenkinson (1955) donne l'expression générale notée GEV (Generalized Extreme Value Distribution) des trois distributions par :

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

Remarque 1.1.1 . *Si F vérifie le théorème* 1.1.1. *On dit alors que F appartient au domaine d'attraction de G_γ et on note $F \in D(G_\gamma)$ selon le signe de γ.*

Distribution GPD

Pickands a introduit la méthode POT (Peaks-over-Threshold) encore appelée méthode des excès au-delà d'un certain seuil réel u suffisamment grand, inférieur au point terminal ($u < x_F$). Cette méthode consiste à utiliser les observations qui dépassent un certain seuil, plus particulièrement les différences entre ces observations et le seuil, appelées "excès". Il est clair que cette méthode nécessite la détermination d'un seuil ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations.

Plus précisément, soit un échantillon de va's iid $X_1, ..., X_n$. Soit u un seuil fixé (non aléatoire) tel que $u < x_F$. On note par N_u le nombre d'exceedances $X_1, ..., X_{N_u}$ qui dépassent le seuil u. On appelle excès au-delà du seuil u les $Y_i = X_i - u$, pour $j = 1, ..., Nu$.

Définition 1.1.6 *(La fonction de distribution des excès). Nous définissons la fonction de distribution des excès au dessus du seuil u par:*

$$\begin{aligned} F_u(y) &= \mathbb{P}(Y \leq y \mid X > u) = \mathbb{P}(X - u < y \mid X > u) \\ &= \frac{F(u+y) - F(u)}{1 - F(u)} \end{aligned}$$

Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction $\sigma(u)$ strictement positive et un $\gamma \in \mathbb{R}$ tels que:

$$\lim_{u \uparrow x_F} \sup_{0 \leq y \leq x_F - u} \left| F_u(y) - G_{\gamma, \sigma(u)}(y) \right| = 0$$

où $G_{\gamma,\sigma}$ est la fonction de répartition de la loi de Paréto Généralisée définie par:

$$G_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 - \gamma \frac{y}{\sigma}\right)^{-1/\gamma} & si \ \ \gamma \neq 0 \ , \ \sigma > 0, \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & si \ \ \gamma = 0 \ , \ \sigma > 0, \end{cases}$$

1.1.3 Domaines d'attraction

L'unification du comportement du maximum en une seule fonction de répartition facilite grandement l'étude du comportement du maximum. Cette loi dépend du seul paramètre de forme γ appelé indice des valeurs extrêmes ou indice de queue. Comme on le verra tout au long de ce mémoire, γ est le paramètre clé de toute la théorie des valeurs extrêmes.

L'estimation de γ nous fournira le comportement de la queue de distribution. En effet, selon son signe γ .

• Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de **Fréchet**, que l'on notera $\mathcal{D}$(Fréchet). Il contient les lois dont la fonction de survie est à décroissance polynomiale, i.e. les lois à **queues lourdes** ou lois de type Pareto. Les lois de ce domaine ont un point terminal x_F infini.

• Si $\gamma < 0$, on dit que F appartient au domaine d'attraction de **Weibull**, que l'on notera $\mathcal{D}$(Weibull). Toutes les lois de ce domaine d'attraction ont un point terminal x_F fini.

• Si $\gamma = 0$, on dit que F est dans le domaine d'attraction de **Gumbel**, que l'on notera $\mathcal{D}$(Gumbel). Il contient les lois dont la fonction de survie est à décroissance exponentielle, i.e. les lois à queues légères.

Fonctions à variation régulière

Définition 1.1.7 . *Une fonction mesurable $g : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation régulière à l'infini si et seulement si, il existe un réel α tel que, pour tout $x > 0$,*

$$\lim_{t\to\infty} \frac{g\,(tx)}{g\,(t)} = x^{\alpha}.$$

Et on note $g \in VR_{\alpha}$, α est appelé indice de la fonction à variation régulière.

Proposition 1.1.1 . *Soient $\alpha \in \mathbb{R}$ et $g \in VR_{\alpha}$. Alors il existe une fonction à variation lente ℓ à l'infini telle que:*

$$\forall x > 0, \quad g(x) = x^{\alpha}\ell(x).$$

Définition 1.1.8 . *Une fonction de répartition F sur $\mathbb{R}$ appartient à une classe à variation régulière VR s'il existe $\alpha \geq 0$ tel que $1 - F \in VR_{-\alpha}$ sur $\mathbb{R}$, ou d'une manière équivalente:*

$$1 - F\,(x) \sim x^{-\alpha}\ell\,(x)\,, \text{ quand } x \to \infty,$$

pour certaines $\ell \in RV_0$.

Caractérisation des domaines d'attraction

Domaine d'attraction de Fréchet

Théorème 1.1.2 . *F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrême $\gamma > 0$ si et seulement si $x_F = +\infty$ et $1 - F$ est une fonction à variation régulière d'indice $-1/\gamma$.*

$$1 - F(x) = x^{-1/\gamma} \ell(x).$$

Dans ce cas, un choix possible pour les suites a_n et b_n est:

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) \quad et \quad b_n = 0.$$

Quelques distributions appartenant au domaine d'attraction de Fréchet : Burr, Fréchet, loggamma, loglogistic, pareto.

Domaine d'attraction de Weibull

Théorème 1.1.3 . *F appartient au domaine d'attraction de Weibul avec un indice de valeur extrême $\gamma < 0$ si et seulement si $x_F < +\infty$ et $1 - F^*$ est une fonction à variation régulière d'indice $1/\gamma$ c'est-à-dire,*

$$1 - F = (x_F - x)^{-1/\gamma} \ell\left[(x_F - x)^{-1}\right].$$

avec,

$$F^*(x) = \begin{cases} 0 & si \quad x \leq 0, \\ F(x_F - x^{-1}) & si \quad x > 0. \end{cases}$$

Dans ce domaine d'attraction les suites de normalisation a_n et b_n sont déterminées comme suit :

$$a_n = x_F - F^{-1}\left(1 - \frac{1}{n}\right) \quad et \quad b_n = x_F.$$

Domaine d'attraction de Gumbel

Définition 1.1.9 . *Soit F une fonction de répartition de point terminal x_F fini ou infini. S'il existe $z < x$ tel que*

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

où $c > 0$ et a une fonction positive absolument continue de densité a' vérifiant $\lim_{x \uparrow x_F} a'(x) = 0$. Alors F est une fonction de Von-Mises et a est sa fonction auxiliaire.

Théorème 1.1.4 . *F appartient au domaine d'attraction de Gumbel si et seulement si il existe une fonction de Von-Mises F^* telle que pour $z < x < x_F$ on ait:*

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

où, $c(x) \to c > 0$ lorsque $x \to x_F$.

1.1.4 Estimateur de Hill (1975) $\hat{\gamma}_{k_n}^H$

L'estimateur de Hill de l'indice de queue (également connu sous le nom indice des valeurs extrêmes), uniquement défini pour les indices positifs $\gamma > 0$. La construction de l'estimateur de Hill est basée sur la méthode du Maximum de Vraisemblance où on se sert des statistiques d'ordre supérieur à un certain seuil u, pour ne garder que les observations les plus grandes, de façon à ce quelles suivent approximativement une distribution Pareto. Il est défini par la statistique suivante :

$$\begin{aligned}\hat{\gamma}_{k_n}^H &= \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \qquad (1.1.1)\\ &= \frac{1}{k}\sum_{i=1}^{k} i\left(\log X_{n-i+1,n} - \log X_{n-i,n}\right)\end{aligned}$$

Si on choisit $k, n \to +\infty$ de sorte que $\dfrac{k}{n} \to 0$ alors on peut montrer que

$$\lim_{k\to\infty} \hat{\gamma}_{k_n}^H = \gamma$$

et l'estimateur de Hill est de plus asymptotiquement normal :

$$\sqrt{k}\frac{\hat{\gamma}_{k_n}^H - \gamma}{\gamma} \xrightarrow{d} N(0,1)$$

la convergence étant en loi. Dans le cas général du domaine de Fréchet, la fonction de survie est de la forme $1 - F(x) = x^{-1/\gamma}\ell(x)$ avec ℓ une fonction à variation lente, Cela induit un biais important sur l'estimateur de Hill, qui est donc en pratique d'un maniement délicat. Dans le cas général, la fonction ℓ apparaît comme un paramètre de nuisance de dimension infinie, qui complique l'estimation *(cf. BERTAIL, (2002))*. Pour plus détails sur la consistance de $\hat{\gamma}^H$ *(Beirlant et al.(2007))* Pour cela, nous allons commencer par les conditions du premières et du seconds ordre :

Proposition 1.1.2 *(Condition du première ordre, de Haan et Ferreira (2006)). Les assertions suivantes sont équivalentes :*

1. F est à queue lourde
$$F \in \mathcal{D}(\text{fréchet}),\ \gamma > 0$$
2. $1 - F$ est une fonction à variation régulière à l'infini d'indice $-1/\gamma$
$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma},\ x > 0$$
3. $Q(1 - s)$ est une fonction à variations régulières à zéro d'indice $-\gamma$
$$\lim_{s\to 0} \frac{Q(1 - sx)}{Q(1 - s)} = x^{-\gamma},\ x > 0$$
4. U est une fonction à variation régulière à l'infini d'indice γ
$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}\ ,\ x > 0$$

Proposition 1.1.3 *(Condition du seconde ordre de Haan et Ferreira (2006)). Une fonction de répartition $F(.) \in \mathcal{D}$(fréchet), $\gamma > 0$, admet une condition du seconde ordre à l'infini si elle satisfait à l'une des assertions suivantes :*

1. Il existe un paramètre $\rho \leq 0$, et une fonction $A_1(.)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall x > 0$
$$\lim_{t\to\infty} \frac{(1 - F(tx))/(1 - F(t)) - x^{-1/\gamma}}{A_1(t)} = x^{-1/\gamma} \frac{x^{\rho} - 1}{\rho}.$$
2. S'il existe un paramètre $\rho \leq 0$ et une fonction $A_2(.)$ qui tend vers 0 (ne change pas de signe à zéro) définie par, $\forall x > 0$
$$\lim_{s\to 0} \frac{Q(1 - sx)/Q(1 - s) - x^{-\gamma}}{A_2(s)} = x^{-\gamma} \frac{x^{\rho} - 1}{\rho},$$
3. S'il existe un paramètre $\rho \leq 0$, et une fonction $A(.)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall x > 0$
$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho}$$

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

Les fonctions $A(.)$, $A_1(.)$, $A_2(.)$ sont à variations régulières à l'infini d'indices respectifs ρ, ρ/γ, et $-\rho$, avec $A_1(t) = A(1/(1 - F(t)))$ et $A_2(s) = A(1/s)$.

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

Théorème 1.1.5 *(Propriétés asymptotiques de l'estimateur de Hill). Soit k_n, $n \geq 1$ une suite d'entiers telle que $1 \leq k_n \leq n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.*

1. **Consistance faible**: $\hat{\gamma}^H_{k_n}$ converge en probabilité vers γ.

2. **Consistance forte**: Si de plus $k_n/\log n \log n \to \infty$ quant $n \to \infty$, alors $\hat{\gamma}^H_{k_n}$ converge presque sûrement vers γ.

3. **Normalité asymptotique**: Si la condition

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

est satisfaite avec $\sqrt{k_n}A(n/k_n) \to \lambda \in \mathbb{R}$ quand $n \to \infty$, alors$\sqrt{k_n}(\hat{\gamma}^H_{k_n} - \gamma) \xrightarrow{\mathcal{L}} \mathcal{N}(\lambda/(1-\rho), \gamma^2)$.

1.2 Données de survie

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

Définition 1.2.1 . *La « durée de vie » d'un individu est une variable aléatoire (v.a.) X positive et continue. Sa fonction de répartition*

$$F(t) = \mathbb{P}(X \leq t),$$

est la probabilite que l'événement se produise entre 0 et t.

1.2.1 Données censurées

Définition 1.2.2 . *La variable de censure Y est définie par la non-observation de l'événement étudié. Si au lieu d'observer X , on observe Y, ona :*

1. $X > Y$ est censure à droite.
2. $X < Y$ est censure à gauche.
3. $Y_1 < X < Y_2$ est censure par intervalle.

Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique. Pour un individu donné i, nous allons considérer,

. son temps de survie X_i de fonction de répartition F,

. sa variable de censure Y_i de fonction de répartition G,

. sa variable réellement observée Z_i de fonction de répartition H.

Dans la littérature on distingue trois types de censure :

Censure à droite: La variable d'intérêt est dite censurée à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées. Un exemple typique est celui où l'événement considéré est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation. On trouve aussi ce genre de phénomène dans les études de fiabilité quand la panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant..., L'expérimentateur peut fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

Censure à gauche: Il y a censure à gauche lorsque l'individu a déjà subi l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue. Par exemple si nous voulons étudier en fiabilité un certain composant électronique qui est branché en parallèle avec un ou plusieurs autres composants: le système peut continuer à fonctionner, quoique de façon aberrante, jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). Ainsi donc, la durée observée pour ce composant est censurée à gauche.

Censure par intervalle: Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[y, +\infty[$ et $[0, y]$ respectivement.

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

La censure de type I

La censure est dite non-aléatoire du type I si, étant donné un nombre positif fixé Y et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où

$$\begin{cases} Z_i = X_i \wedge Y \\ \delta_i = \mathbb{I}_{\{X_i \leq Y\}} \end{cases}$$

La censure de type II

La censure est dite non-aléatoire du type II si, étant donné un nombre positif fixé r et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où

$$\begin{cases} Z_i = X_i \wedge X_{(r)} \\ \delta_i = \mathbb{I}_{\left\{X_i \leq X_{(r)}\right\}} \end{cases}$$

La censure de type III:

C'est la version aléatoire du type I

Définition 1.2.3 . *La censure est dite aléatoire du type I si, étant donné un n-échantillons* $X_1, ..., X_n$, *il existe un v.a. n-dimensionnelle* $(Y_1, ..., Y_n)$ *de* $(\mathbb{R}^+)^n$ *telle que les observations consistent en* (Z_i, δ_i) *où*

$$\begin{cases} Z_i = X_i \wedge Y_i \\ \delta_i = \mathbb{I}_{\{X_i \leq Y_i\}} \end{cases}$$

1.3 Introduction générale de test

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Définition 1.3.1 *(test d'hypothése).* *Un test est un mécanisme qui permet de trancher entre deux hypothéses à la vue des résultats d'un échantillon , en quantifiant le risque associé à la décision prise.*

hypothéses de test:

Soient H_0 et H_1 deux hypothéses, une seule est vraie. H_0 joue le plus souvent un rôle prédominant par rapport à H_1 en effet H_0 est l'hypothèse de référence alors que H_1 est l'hypothèse alternative.

- Il y a donc quatre cas possibles qui sont détaillés dans le tableau ci-dessous:

l'hypothese	$H_0 vraie$	$H_1 vraie$
$H_0 acceptée$	$1-\alpha$	β
$H_1 acceptée$	α	$1-\beta$

où α et β sont les risques d'erreur de première et de deuxième espèce.

L'erreur de première espèce :

Est le fait de décider l'hypothése alternative H_1 est accepté, alors qu'en fait, en réalité, c'est l'hypothése nulle H_0 qui est vraie.

- Le **risque d'erreur** associé à cette décision est noté généralement α.
- Il s'agit donc de la probabilité de décider à tort que l'hypothése alternative H_1 est accepté.

L'erreur de deuxiéme espèce :

Est le fait de décider l'hypothése nulle H_0 est accepté alors qu'en fait, en réalité, c'est l'hypothése alternative H_1 qui est vraie.

- Le **risque d'erreur** associé à cette décision est noté généralement β.
- Il s'agit donc de la probabilité de décider à tort que l'hypothése nulle H_0 est accepté.

Puissance de test :

Définition 1.3.2 . *La puissance d'un test est la probabilité d'accepté H_1, alors que H_1 est vraie. C'est-à-dire lorsqu'on est en vérité dans le cadre de l'hypothèse alternative H_1. La puissance du test est donc le complément de l'erreur de deuxième espèce β. On la note* $1-\beta$

Test bilatéral :

S'applique quand vous cherchez une différence entre deux paramétres,ou entre un paramètre et une valeur donnée sans se préoccuper du signe ou du sens de la difference. Dans ce cas la region de l'hypothèse principale se fait de part et d'autre de la distribution de référence.

Test unilatéral :

S'applique quand vous cherchez à savoir si un paramètre est supérieur (ou inférieur) à un autre ou à une valeur donnée. La region de rejet de l'hypothése principale est située d'un seul côté de la distribution de probabilité de référence.

1.3.1 Construction d'un test

Statistique de test

Pour un risque d'erreur de premiére 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 probléme posé, à savoir le choix entre les deux hypothéses. Sa loi doit être parfaitement déterminée dans au moins une des deux hypothéses (le plus souvent sous H_0) afin de ne pas introduire de nouvelles inconnues dans le problème.

Niveau de signification

L'erreur de première espèce est limitée à un niveau dit **niveau de signification .**

• Le risque cousenti à l'avance et que nous notous α de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie, s'appelle le **seuil de signification** de test et s'énonce en probabilité ainsi :

$$\alpha = \mathbb{P}\left(rejeter H_0 \setminus H_0 vraie\right)$$

Région critique

Notée **W,** est égale à l'ensemble des valeurs de la variable de décision que conduisent à écarter H_0 au profit de H_1. Dans la plupart des situatious que nous rencontrerons dans la suite, la région critique $\boldsymbol{W}$ peut être reliée au risque d'erreur de premiére espéce α par :

$$\mathbb{P}_{H0}\left(\mathbf{W}\right) = \alpha$$

• Sur la distribution correspondra aussi une région complémentaire, dite région d'acceptation de H_0 de probabilité $1 - \alpha$.

$$\mathbb{P}_{H0}\left(\bar{\mathbf{W}}\right) = 1 - \alpha$$

• $1 - \alpha$ est parfois appelée niveau de confiance du test.

• les seuils de signification les plus utilisés sont $\alpha = 0.05, 0.01, 0.025$ dépendant des conséquences de rejetes à tort l'hypothése H_0.

Chapitre 2

Test paramétrique de l'estimateur de l'indice de queue dans le cas censure

On a vu dans le Chapitre 1 que pour la majorité des fonctions de repartitions F la loi asymptotique du maximum $X_{n,n}$ (convenablement normalisée) est une loi des valeurs extrêmes qui étant indexée par le paramètre de queue γ, ce paramètre apporte une information sur la forme de la queue de distribution de F. Notamment, selon que $\gamma > 0$, $\gamma < 0$ ou $\gamma = 0$. On distingue trois domaines d'attraction : Frechet, Weibull et Gumbel. Dans la littérature de la TVE de nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes. On va s'intéresser dans première Section au problème de l'estimation de l'IVE et cela en présence de données censurées aléatoirement à droite. Ce problème est très récent dans la littérature, au meilleur de notre connaissance, les premiers qui ont mentionné le sujet sont *Beirlant et al.(2007)*, mais sans résultats asymptotiques. Puis certains estimateurs des paramètres de la queue ont été proposées par *Beirlant et Guillou(2005),* pour les données tronquées et étendu à la censure aléatoire à droite par *Beirlant et al.(2007)*, et l'année suivante par *Einmahl et al. (2008).*

2.1 Estimation de l'indice des valeurs extrêmes censurés

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

$$1 - F \in RV(-1/\gamma_1).$$

Soit le deuxiéme échantillon $Y_1, ..., Y_n$ des variables aléatoires iid, de fonction de répartition G et de queue de distribution:

$$1 - G \in RV(-1/\gamma_2).$$

Alors les variables Z_i définies par :

$$Z_i = X_i \wedge Y_i \quad , \quad i = 1, ..., n,$$

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

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

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

$$1 - F(x) = x^{-\frac{1}{\gamma_1}} \ell_1(x) \quad \text{et} \quad 1 - G(x) = x^{-\frac{1}{\gamma_2}} \ell_2(x).$$

avec $\ell_1(x)$ et $\ell_2(x)$ des fonction à variation lente. Alors,

$$\begin{aligned} 1 - H(x) &= (1 - F(x))(1 - G(x)) \\ &= x^{-\frac{1}{\gamma_1}} \ell_1(x) \, x^{-\frac{1}{\gamma_2}} \ell_2(x) \\ &= x^{-(\frac{1}{\gamma_1} + \frac{1}{\gamma_2})} \ell_1(x) \, \ell_2(x) \\ &= x^{-(\frac{\gamma_1 + \gamma_2}{\gamma_1 \gamma_2})} \ell(x) \\ &= x^{-\frac{1}{\gamma}} \ell(x) \quad \text{avec } \ell(x) = \ell_1(x) \, \ell_2(x) \end{aligned}$$

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

$$1 - H(x) \in RV(-1/\gamma),$$

avec,

$$\gamma = \frac{\gamma_1 \gamma_2}{(\gamma_1 + \gamma_2)}$$

Si F et G sont supposées être dans le dommaine d'attraction maximale $D(G_{\gamma_1})$ et $D(G_{\gamma_2})$ respectivement où $\gamma_1, \gamma_2 \in \mathbb{R}$ avec points terminales τ_F et τ_G, où $\tau_F = \sup\{x, F(x) < 1\}$, alors cela signifie que $H \in D(G_\gamma)$. *Einmahl et al.*(2008) ont examiné les trois cas les plus intéressants suivants:

$$\begin{cases} cas\ 1 \ : \ \gamma_1 > 0,\ \gamma_2 > 0, & \gamma = \dfrac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2} \\ cas\ 2 \ : \ \gamma_1 < 0,\ \gamma_2 < 0, \quad \tau_F = \tau_G\ , & \gamma = \dfrac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2} \\ cas\ 3 \ : \ \gamma_1 = \gamma_2 = 0 \quad \tau_F = \tau_G = \infty & \gamma = 0 \end{cases} \tag{2.1.1}$$

Soit $\{(Z_i, \delta_i), 1 \leq i \leq n\}$ un échantillon de couple de variables aléatoires. $Z_{1,n} \leq ... \leq Z_{n,n}$ représentent les statistiques d'ordre associées à l'échantillon $(Z_1, ..., Z_n)$. Dans le cas général on obtient :

$$\hat{\gamma}^{(c,.)}_{Z,k,n} = \frac{\hat{\gamma}^{(.)}_{Z,k,n}}{\hat{p}}, \tag{2.1.2}$$

où,

$$\hat{p} = \frac{1}{k} \sum_{i=1}^{k} \delta_{[n-i+1:n]}, \tag{2.1.3}$$

avec k est le nombre des valeurs extrêmes. $\hat{p}$ estime $p = \dfrac{\gamma_2}{\gamma_1 + \gamma_2}$ (où p représente la proportion de données observées dans la queue à droite de la distribution). $\hat{\gamma}^{(.)}_{Z,k,n}$ peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill $\hat{\gamma}^{(H)}_{Z,k,n}$. Pour adapter l'estimateur de Hill dans le cas censuré nous allons diviser cet estimateur par la proportion de données non censurées des k plus grandes valeurs de Z, Alors l'estimateur de Hill adapté de l'indice de queue γ_1 est difinit par:

$$\hat{\gamma}^{(c,H)}_{Z,k,n} = \frac{\hat{\gamma}^{H}}{\hat{p}},$$

où,

$$\hat{\gamma}^{H} = \frac{1}{k} \sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}$$

alors,

$$\hat{\gamma}_1^{(c,H)} = \frac{k^{-1}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}}{k^{-1}\sum_{i=1}^{k} \delta_{[n-i+1:n]}} \tag{2.1.4}$$

$\hat{\gamma}_1^{(c,H)}$ estime $\gamma_1 = \gamma/p$ avec $\delta_{[1:n]}, ..., \delta_{[n:n]}$ les indicateurs de censure retenues respectivement avec l'échantillon $Z_{1:n}, ..., Z_{n:n}$. *Einmahl et al* (2008) ont établi de façon unifiée, la normalité asymptotique de tout estimateur de l'indice des valeurs extrêmes écrit sous la forme $(2, 1, 2)$. Cet estimateur est basé sur les obsevations Z_i.

Dans notre travail, nous nous intersessons à construire un test d'hypothèse paramétrique de l'estimateur de l'indice des valeurs extrêmes sous des données censure noté préalablement $\hat{\gamma}_1$. Pour cela nous déterminous les propriétés asymptotiques de l'estimateur pricétée nous aurons besoin de quelques conditions de régularité. Comme suit:

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

$$\lim_{t\to\infty} \frac{H^{\leftarrow}(1 - \frac{1}{tu})/H^{\leftarrow}(1 - \frac{1}{t}) - u^{\gamma}}{b(t)} = u^{\gamma}\frac{u^{\rho} - 1}{\rho} \tag{2.1.5}$$

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

$$1 < k < n;\ k \to \infty \ \text{ et } k/n \to 0, n \to \infty \tag{2.1.6}$$

$\mathbb{C}2$: $\sqrt{k}b(\frac{n}{k}) \to \alpha_1 \in \mathbb{R}$

$\mathbb{C}3$: $\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1 - \frac{i}{n})) - p\right] \to \alpha_2 \in \mathbb{R}$

$\mathbb{C}4$: Soit $c > 0$ et $\mathcal{A}(s,t) := \left\{1 - k/n \le t < 1,\ |t - s| \le C\sqrt{k}/n,\ s < 1\right\}$ si $n \to \infty$,

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

Sous ces conditions, nous avons les résultats asymptotiques suivants.

Théorème 2.1.1 . *Sous les condition* $\mathbb{C}1 - \mathbb{C}4$ *et s'il existe* b_0 *et* σ *telles que*

$$\sqrt{k}(\hat{\gamma}^{(.)}_{Z,k,n} - \gamma) \xrightarrow{d} \mathcal{N}(\alpha_1 b_0, \sigma^2). \tag{2.1.7}$$

Alors, nous avons

$$\sqrt{k}\left(\hat{\gamma}^{(c,.)}_{Z,k,n}-\gamma_1\right)\xrightarrow{d}\mathcal{N}\left(\frac{1}{p}(\alpha_1 b_0-\gamma_1\alpha_2),\frac{\sigma^2+\gamma_1^2 p(1-p)}{p^2}\right).\ quand\ \ n\to\infty.$$

Les résultats asymptotiques de l'estimateurs de Hill, est donnés par

$$\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)\xrightarrow{d}\mathcal{N}\left(\mu^{(c,H)};\frac{\gamma_1^3}{\gamma}\right) \tag{2.1.8}$$

où

$$\mu^{(c,H)}:=-\frac{\gamma_1\alpha_2}{p}+\frac{\alpha_1}{p}\frac{\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}$$

Ce corollaire se déduit directement du théorème précédent en notant,

$$b_0=1/(1-\rho)\ et\ \ \sigma^2=\gamma^2,$$

Pour déterminer les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes nous avons besoin de la fonction suivante comme difinie dans (*Einmahl et al.*(2008)),

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

Nous pouvons l'écrire d'une autre manière,

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

où f et g désignent respectivement les densités de F et G et on a:

$$\lim_{z\to\tau_H}p(z)=\frac{\gamma_2}{\gamma_1+\gamma_2}:=p$$

Supposons X et Y sont respectivement de Pareto (γ_1) et Pareto (γ_2), C'est-à-dire pour tout $x\geq 1$.

$$F_X(x)=1-x^{-1/\gamma_1},\qquad \gamma_1>0$$
$$F_Y(x)=1-x^{-1/\gamma_2},\qquad \gamma_2>0$$

on obtient:

$$\begin{aligned}F_Z(z)&=\mathbb{P}(\min(X,Y\leq z))\\&=1-\mathbb{P}(X>z)\mathbb{P}(Y>z)\\&=1-z^{-1/\gamma_1}z^{-1/\gamma_2}\\&=1-z^{-\frac{\gamma_1+\gamma_2}{\gamma_1\gamma_2}},\end{aligned}$$

ce qui implique $Z \frown$ Pareto $(\gamma_1\gamma_2/(\gamma_1+\gamma_2))$. Nous pouvons à présent calculer la fonction $p(z)$

$$\begin{aligned} p \equiv p(z) &:= \frac{(1-F_Y(z))f_X(z)}{(1-F_Y(z))f_X(z)+(1-F_X(z))f_Y(z)} \\ &= \frac{z^{-1/\gamma_2}\frac{1}{\gamma_1}z^{-1/\gamma_1}}{z^{-1/\gamma_2}\frac{1}{\gamma_1}z^{-1/\gamma_1}+z^{-1/\gamma_1}\frac{1}{\gamma_2}z^{-1/\gamma_2}} \\ &= \frac{\frac{1}{\gamma_1}z^{-1/\gamma_1-1/\gamma_2}}{(\frac{1}{\gamma_1}+\frac{1}{\gamma_2})z^{-1/\gamma_1-1/\gamma_2}} \\ &= \frac{\frac{1}{\gamma_1}}{\frac{1}{\gamma_1}+\frac{1}{\gamma_2}} \\ &= \frac{\gamma_2}{\gamma_1+\gamma_2}. \end{aligned}$$

2.2 Test de l'estimateur de queue $\hat{\gamma}^{(c,H)}$

2.2.1 Constraction du test

Notre test est basé sur les observations Z_i. On s'intéresse à tester les valeurs prises par γ_1, l'indice des valeurs extrême de l'échantillon $X_1, ..., X_n$ dans le cas des données censurées, selon son estimateur $\hat{\gamma}^{(c,H)}$ en se basant sur sa normalité asymptotique (théoréme 2.1.1)

$$H_0 : \gamma_1 = \gamma_0$$

contre l'hypothèse

$$H_1 : \gamma_1 \neq \gamma_0$$

avec γ_0 est une valeur spécifique.

2.2.2 Statistique du test

sous l'hypothèse H_0, la variable aléatoire $\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)$ suit une loi normale $\mathcal{N}\left(\mu^{(c,H)};\frac{\gamma_1^3}{\gamma}\right)$ avec

$$\begin{aligned}\mu^{(c,H)} &= \mathbb{E}\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)\right)\\ &= -\frac{\gamma_1\alpha_2}{p}+\frac{\alpha_1}{p}\frac{\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\end{aligned}$$

et,

$$\mathbb{V}ar\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)\right)=\frac{\gamma_1^3}{\gamma}$$

et par la conséquent la statistique de test est :

$$T_{k,n}=\frac{\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)-\mathbb{E}\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)\right)}{\sqrt{\mathbb{V}ar\left(\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_1\right)\right)}}$$

qui suit une loi normale centrée réduite $\mathcal{N}(0;1)$.

Sous l'hypothése H_0 ; $\gamma_1=\gamma_0$ et donc la statistique de test noté $T_{k,n}$ et par le théorème $(2,1,1)$, la statistique de test vérifiée :

$$T_{k,n}=\frac{\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_0\right)-\mu^{(c,H)}}{\sqrt{\frac{\gamma_0^3}{\gamma}}}\xrightarrow{d}\mathcal{N}(0;1) \tag{2.2.1}$$

En simplifiant la valeur $T_{k,n}$, on a :

$$\begin{aligned}T_{k,n} &= \frac{\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_0\right)-\mu^{(c,H)}}{\sqrt{\frac{\gamma_0^3}{\gamma}}}\\ &= \frac{\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_0\right)-(-\frac{\gamma_0\alpha_2}{p}+\frac{\alpha_1}{p}\frac{\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})})}{\sqrt{\frac{\gamma_0^3}{\gamma}}}\end{aligned}$$

On pose: $\theta_{k,n}:=\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n}-\gamma_0\right)$ donc on a :

$$T_{k,n}:=\theta_{k,n}\sqrt{\frac{\gamma}{\gamma_0^3}}-\frac{\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha_1\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\right)+\frac{\sqrt{\gamma}(\gamma_0\alpha_2)}{\sqrt{\gamma_0^3}}$$

On pose $T_{k,n} = T_{1,n} + T_{2,n} + T_{3,n}$ avec,

$$\begin{aligned} T_{1,n} &= \theta_{k,n}\sqrt{\frac{\gamma}{\gamma_0^3}}, \\ T_{2,n} &= \frac{-\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha_1\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\right), \\ T_{3,n} &= \frac{\sqrt{\gamma}(\gamma_0\alpha_2)}{\sqrt{\gamma_0^3}}. \end{aligned}$$

par calcul simple, on obtient:

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

$$\begin{aligned} T_{2,n} &= \frac{-\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\alpha_1\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\right) \\ &= \frac{-\sqrt{\gamma}}{p\sqrt{\gamma_0^3}}\left(\frac{\sqrt{k}b(\frac{n}{k})\gamma}{\tilde{\rho}+\gamma(1-\tilde{\rho})}\right) \\ &= -\frac{\sqrt{p}}{\gamma_0}\frac{\sqrt{k}b(\frac{n}{k})\gamma}{p\left(\tilde{\rho}+\gamma(1-\tilde{\rho})\right)} \\ &= -\frac{\gamma_0\sqrt{k}b(\frac{n}{k})}{\left(\tilde{\rho}+\gamma(1-\tilde{\rho})\right)}\frac{\sqrt{p}}{\gamma_0} \\ &= -\frac{\sqrt{k}b(\frac{n}{k})\sqrt{p}}{\tilde{\rho}+\gamma(1-\tilde{\rho})}. \end{aligned}$$

$$\begin{aligned} T_{3,n} &= \frac{\sqrt{\gamma}(\gamma_0\alpha_2)}{p\sqrt{\gamma_0^3}} \\ &= \frac{\gamma_0\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\right](\gamma_0+\gamma_2)}{\gamma_2}\frac{\sqrt{p}}{\gamma_0} \\ &= \frac{\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\right]\sqrt{p}}{p} \\ &= \frac{\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\right]}{\sqrt{p}} \end{aligned}$$

Finalment, on obtient

$$T_{k,n} := \frac{\theta_{k,n}\sqrt{p}}{\gamma_0} - \frac{\sqrt{k}b(\frac{n}{k})\sqrt{p}}{\tilde{\rho}+\gamma(1-\tilde{\rho})} + \frac{\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\right]}{\sqrt{p}} \tag{2.2.2}$$

La statistique $T_{k,n}$ sous l'hypothése H_0, est une variable aléatoire distribuée asymptotiquement comme une loi normale centrée réduite. Quand $n \to \infty$.

2.2.3 Région critique du test

Un test bilatéral se base donc sur la région critique

$$W = \{|T_{k,n}| > q_{1-\frac{\alpha}{2}}\}$$

Pour un risque d'erreur α fixé, on a :

$$\begin{aligned} \mathbb{P}_{H_0}(W) &= \alpha \\ \mathbb{P}_{H_0}(|T_{k,n}| &> q_{1-\frac{\alpha}{2}}) = \alpha \end{aligned}$$

avec, $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\dfrac{\alpha}{2}$ de la loi normal,(voir le tableau de loi normal $N(0,1)$). la région critique est donnée par :

$$\left]-\infty, -q_{1-\frac{\alpha}{2}}\right[\cup \left]q_{1-\frac{\alpha}{2}}, +\infty\right[$$

2.2.4 Région d'acceptation du test

La région d'acceptation notée $\bar{W}$, (ou aussi région de confiance) est :

$$\begin{aligned} \mathbb{P}_{H_0}(\bar{W}) &= 1-\alpha \\ \mathbb{P}_{H_0}(|T_{k,n}| &\leq q_{1-\frac{\alpha}{2}}) = 1-\alpha \end{aligned}$$

avec $1-\alpha$ est niveau de confiance du test. On obtient, donc,

$$RC = \left[-q_{1-\frac{\alpha}{2}} \ ; \ q_{1-\frac{\alpha}{2}}\right]$$

2.2.5 Prise de décision

Si $T_{k,n} \leq q_{1-\frac{\alpha}{2}}$, on accepte l'hypothèse H_0, que l'échantillon réelle proviennent de la distribution de pareto de paramètre γ. Sinon on rejette.

2.2.6 Puissance du test

$$\begin{aligned} 1-\beta &= \mathbb{P}_{H_1}(accepte\ H_1/H_1 vraie) \\ 1-\beta &= \mathbb{P}_{H_1}(W) \\ 1-\beta &= \mathbb{P}_{H_1}\left(|T_{k,n}| \leq z_\alpha - \left(\frac{\sqrt{k}\left(\hat{\gamma}^{(c,H)}_{Z,k,n} - \gamma_0\right) - \gamma_0}{\sqrt{\frac{\gamma_1^3}{\gamma}}}\right)\sqrt{n}\right) \end{aligned}$$

Chapitre 3

Simulations du test paramétrique pour γ_1

On procéde à notre test paramétrique sous l'hypothèse $H_0 : \gamma_1 = \gamma_0$

3.1 Algorithme du test

Échantillons et paramétres de simulations

Pour générer un échantillon de variables aléatoires de taille n d'une loi connue F il faut tout d'abord calculer la fonction inverse généralisée F^{-1}, ensuite déterminer ce dernier à partir d'un autre échantillon issu d'une loi uniforme $U([0,1])$ par la relation:

$$F^{-1}(u) = x$$

La loi de simulation utilisée d'une maniére génerale est une loi de Pareto de paramètre γ de fonction de répartition,

$$F(x) = 1 - x^{-1/\gamma}, \qquad \gamma > 0.$$

Distributions de simulation

Nous avons généré un échantillon $(X_i)_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_1)$ de taille $n = 1000$, à partir d'une variable u de loi uniforme $U([0,1])$, le modèle ajusté sera:

$$F^{-1}(u) = (1-u)^{-\gamma_1}, \qquad \gamma_1 > 0.$$

L'échantillon $(X_i)_{1\leq i\leq n}$est censuré par un deuxième échantillon $(Y_i)_{1\leq i\leq n}$ $\backsim$Pareto(γ_2) à partir d'une variable v de loi uniforme $U([0,1])$:

$$G^{-1}(v) = (1-v)^{-\gamma_2}, \qquad \gamma_2 > 0.$$

Les variables que nous observons sont d'une part les $Z_i \backsim$Pareto(γ) définies par:

$$Z_i = X_i \wedge Y_i.$$

Les indicateurs de censure sont,

$$\delta_i = \mathbb{I}_{\{X_i \leq Y_i\}}$$

Générer le nombre des valeurs extrêmes k

Commentaires Pour le choix du meilleur k optimal qui sera adaptatif à la normalité asymptotique et qui donne une meilleur valeur de l'estimateur $\hat{\gamma}_1^{(c,H)}$. Pour un bon déroulement de notre test nous avons choisit le nombre de valeurs extrêmes selon un moyen assez utilisé en simulation :

$$kopt = [n]^{\theta}, \qquad \text{pour } 0 < \theta < 1.$$

On prend $\theta = 0.45$.

Générer la variable aléatoire de test $\theta_{k,n}$

On ordonne l'échantillon Z_i pour générer les variables

$$\begin{aligned}
\theta_{k,n} &= \sqrt{k}\left(\hat{\gamma}_{Z,k,n}^{(c,H)} - \gamma_0\right) \\
&= \sqrt{k}\left(\frac{k^{-1}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}}{k^{-1}\sum_{i=1}^{k}\delta_{[n-i+1:n]}} - \gamma_0\right)
\end{aligned}$$

Calcul de la statistique de test $T_{k,n}$

Aprés avoir simplifier la statistique de test $T_{k,n}$ qui suit une loi $\mathcal{N}(0,1)$ dans le chapitre 2, on doit la calculer pour la prise de decision,

$$T_{k,n} := \frac{\theta_{k,n}\sqrt{p}}{\gamma_0} - \frac{\sqrt{k}b(\frac{n}{k})\sqrt{p}}{\tilde{\rho}+\gamma(1-\tilde{\rho})} + \frac{\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left[p(H^{\leftarrow}(1-\frac{i}{n}))-p\right]\right]}{\sqrt{p}}$$

avec, $p = \dfrac{\gamma_2}{\gamma_0+\gamma_2}$, $\gamma = \dfrac{\gamma_2\gamma_0}{\gamma_0+\gamma_2}$. Ainsi $b(n/k)$ est une fonction à variation lente, $\tilde{\rho}$ paramétre de seconde ordre et $H^{\leftarrow}(.)$ est l'inverse de $H(.)$ (pour plus de détails voir la section suivante).

La région critique du test W

Pour une erreur de premiére espèse α fixée, on a:

$$P_{H_0}\left(T_{k,n} \in W\right) = \alpha,$$

où,

$$W = \left]-\infty, -q_{1-\frac{\alpha}{2}}\right[\cup \left]q_{1-\frac{\alpha}{2}}, +\infty\right[,$$

est la région critique de test, avec $q_{1-\frac{\alpha}{2}}$est le quantile d'ordre $\dfrac{\alpha}{2}$ d'une loi normale centré réduite $\mathcal{N}(0,1)$.

La prise de décision

$$\begin{cases} \text{on accepte } H_0 & \text{si} \qquad -q_{1-\frac{\alpha}{2}} \leq T_{k,n} \leq q_{1-\frac{\alpha}{2}} \\ \text{on rejéte } H_0 & \text{sinon} \end{cases}$$

3.2 Génération de l'échantillons (Z_i, δ_i) sous R

paramètres de simulation

n=2000

gamma0<-

gamma1<-?

gamma2<-?

échantillon de simulation

```
u<-runif(n,0,1)
x<-(1-u)^(-gamma1)
v<-runif(n,0,1)
y<-(1-v)^(-gamma2)
z<-pmin(x,y)
delta=as.numeric(x<=y)
z
delta
r<-rank(z)
#ordonnées les données

Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for (i in 1:length(z))
{
L[i]<-(delta[i]+length(z))*r[i]
}
L
A<-sort(L)
A
for (i in 1:length(z))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(z)
Delta
Z<-log(Y)
```

3.3 Choix du nombre des valeurs extrêmes optimal k

Les résultats asymptotiques concernant les estimateurs de l'indice des valeurs extrêmes sont obtenus lorsque $k \to \infty$ et $k/n \to 0$. La difficulté en pratique consiste à choisir le nombre d'extrêmes k utilisé dans les estimations. L'issue est importante: l'extrême volatilité du graphe $((k,\ \hat{\gamma}_k),\quad k = 1, ..., n-1)$. Où $\hat{\gamma}_k$ représente n'importe quel estimateur introduit précédement, rend difficile l'utilisation de l'estimateur en pratique si aucune indication sur le choix de k n'est donnée. Des travaux ont montré qu'en utilisant trop d'observations, dans la procédure d'estimation de γ, on observe un biais substantiel tandis que l'utilisation de peu d'observations conduit à une variance considérable. Ce probème a été longuement abordé dans la littérature, voir par exemple *Reiss et Thomas (1997), de Haan et Peng (1998), Drees et Kaufmann (1998), Danielsson et al. (2001), Cheng et Peng (2001) Beirlant et al. (2002), Beirlant et al. (2004), etc.* Nous allons utiliser dans nos simulations les méthodes suivants *Cheng et Peng* et *l'erreur quadratique moyenne* et *Reiss et Thomas* pour déterminer la valeur optimale de k correspendante à l'estimateur $\hat{\gamma}_1^{(c,H)}$.

3.3.1 Méthode de Cheng et Peng

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

$$k_n^{opt} := \begin{cases} \left(\frac{1+2z_\alpha^2}{3\hat{\delta}(1+2\hat{\rho})}\right)^{1/(1+\hat{\rho})} & n^{\hat{\rho}/(1+\hat{\rho})},\quad si\ \ \hat{\delta} > 0 \\ \\ \left(\frac{1+2z_\alpha^2}{-3\hat{\delta}}\right)^{1/(1+\hat{\rho})} & n^{\hat{\rho}/(1+\hat{\rho})},\quad si\ \ \hat{\delta} < 0 \end{cases} . \tag{3.3.1}$$

Où z_α est le quantile de la distribution standard .

$$\hat{\rho} \ : = -\log\left(\left|\frac{M_n^{(2)}(n/2\sqrt{\log n}) - 2\{\hat{\gamma}_1^{(c,H)}(n/2\sqrt{\log n})\}^2}{M_n^{(2)}(n/\sqrt{\log n}) - 2\{\hat{\gamma}_1^{(c,H)}(n/\sqrt{\log n})\}}\right|\right)/\log 2,$$

$$\hat{\delta} \ : = (1+\hat{\rho})(\log n)^{\hat{\rho}/2}\frac{M_n^{(2)}(n/\sqrt{\log n}) - 2\{\hat{\gamma}_1^{(c,H)}(n/\sqrt{\log n})\}}{2\hat{\rho}\{\hat{\gamma}_1^{(c,H)}(n/\sqrt{\log n})\}^2},$$

avec,

$$M_n^{(2)} = \frac{\frac{1}{k}\sum_{i=1}^{k}(\log Z_{n-i+1,n} - \log Z_{n,k,n})^2}{\hat{p}}$$

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

Calcul de k^{opt} sous R

Le programme suivant sous R calcul directement la valeur de k^{opt}.

```
## kopt par Cheng et Peng ##
alpha=0.05
q<-qnorm(1-alpha)
W<-n/sqrt(log(n))
V<-W/2
w<-floor(W)
v<-floor(V)
K11<-(((1/w)*sum(Z[n-w+1]))-Z[n-w])/((1/w)*sum(Delta[n-w+1]))
K12<-(((1/v)*sum(Z[n-v+1]))-Z[n-v])/((1/v)*sum(Delta[n-v+1]))
K21<-(1/w)*sum((Z[n-w+1])-(Z[n-w]))^2
K22<-(1/v)*sum((Z[n-v+1])-(Z[n-v]))^2
ro<-abs((1/log(2))*log(abs((K22-2*(K12^2))/(K21-2*(K11^2)))))
d<-(K21-2*(K11^2))*(1+ro)*((sqrt(log(n)))^ro)/(2*(K11^2)*ro)
if(d>0){
K<-(((1+2*(q^2))/(3*d*(1+2*ro))))^(1/(1+ro))*(n^(ro/(1+ro)))
}else{
K<-((1+2*(q^2))/(-3*d))^(1/(1+ro))*(n^(ro/(1+ro)))
}
kopt<-ceiling(K)
kopt
```

3.3.2 Méthode basée sur l'erreur quadratique moyenne

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

Pour toute réplication R nous estimons γ_1 et soit $\hat{\gamma}_k^{(c,H),j}$ l'estimateur de γ_1 obtenu à la $j-$ *iéme* réplication $(j = 1, ..., R)$ avec $(k = 1, ..., n-1)$. Il semble donc naturel de trouver une valeur k^{opt} qui minimise les valeurs de l'erreur quadratique moyenne $\{(k, MSE(k), \quad k = 1, ..., n-1)\}$ par rapport à k. La valeur optimale de k est donnée par:

$$k^{opt} := \underset{1\leq k\leq n-1}{\arg\min} \left\{ \frac{1}{R} \sum_{j=1}^{R} \left(\hat{\gamma}_k^{(c,H),j} - \gamma_1 \right)^2 \right\}. \tag{3.3.2}$$

Il est donc facile de voir que la MSE de $\hat{\gamma}_{k_n}^{(c,H)}$, qui est en fonction de k_n n'est rien d'autre que le carré du biais plus la variance de l'estimateur, ils est nécessaire de trouver un compromis entre le biais et la variance. Il semble raisonnable qu'une minimisation du MSE permet de trouver une valeur intermédiaire entre les composantes du biais et de la variance pour ce compromis.

Calcul de k^{opt} par minimisation du MSE sous R.

Le programme suivant sous R calcul directement la valeur de k^{opt}.

```
R=300
a<-floor(n/5)
b<-floor(n-n/5)
EQM<-seq(1:b)
for (k in 1:b)
{
hills<-seq(1,R)
for(j in 1:R){
indice<-sample(1 :length(z),length(z),replace=TRUE)
zboot<-z[indice];deltaboot<-delta[indice]
zboot
```

```
deltaboot
r<-rank(zboot)
Y<-sort(zboot)
r1<-sort(r)
L<-seq(1,length(zboot))
B<-seq(1,length(zboot))
for (i in 1:length(zboot))
{
L[i]<-(deltaboot[i]+length(zboot))*r[i]
}
L
A<-sort(L)
for (i in 1:length(zboot))
{
A
B[i]<-A[i]/r1[i]
}
B
Deltaboot<-B-length(zboot)
Deltaboot
Zboot<-log10(Y)
l<-seq(1,k)
hills[j]<-(((1/k)*sum(Zboot[n-l+1]))-Zboot[n-k])/((1/k)*(sum(Deltaboot[n-l+1])))
hills
}
hills
EQM[k]<-(1/R)*sum((hills-gamma1)^2)
}
EQM
kopt<-which.min(EQM[a:b])+a-1
```

3.3.3 Méthode de Reiss et Thomas

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

$$\frac{1}{k}\sum_{i\leq k} i^{\beta}\left|\hat{\gamma}_n^{(c,H)}(j) - med\left(\hat{\gamma}_{1,n}^{(c,H)},...,\hat{\gamma}_{k,n}^{(c,H)}\right)\right|, \qquad 0\leq\beta\leq 1/2$$

Calcul de k^{opt} par la méthode sous R

```
rm(list=ls())
    n=1000
    gamma1<-0.35
    gamma2<- 2.5
    p<-gamma2/(gamma1+gamma2)
    gamma<-(gamma1*gamma2)/(gamma1+gamma2)
    gamma
    # géneration d'une loi Pareto de paramétre#
    u<-runif(n,0,1)
    x<-(1-u)^(-gamma1)
    v<-runif(n,0,1)
    y<-(1-v)^(-gamma2)
    z<-pmin(x,y)
    delta=as.numeric(x<=y)
    z
    delta
    r<-rank(z)
    #ordonnées les données
```

```
Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for (i in 1 :length(z))
{
L[i]<-(delta[i]+length(z))*r[i]
}
L
A<-sort(L)
A
for (i in 1 :length(z))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(z)
Delta
Z<-log(Y)
a=n-2
b=n-1
## calcul du vecteur de l'estimateur de Hill dans le cas cencurée de k=1,...n-1, (hill1)
hill1<-seq(1,b)
med<-seq(1,b)
for(j in 1:b)
{
s1<-0
for(i in 1:j)
{
s1<-s1+((1/j)*(Z[n-i+1]-Z[n-j]))/((1/j)*sum(Delta[n-j+1]))
```

```
}
h<-s1
hill1[j]<-h
med[j]<-median(hill1[1:j])
}
hill1
## l'equation de Reiss and Thomas (2003) page 697, equation (10), (M)
Mi<-seq(1,b)
for(j in 1:b)
{
s2<-0
for(i in 1:j)
{
s2<-s2+((1/j)*(i^0.3)*(abs(hill1[i]-med[j])))
}
m<-s2
Mi[j]<-m
}
Mi
## le nombre de valeurs extremes est l'entier K qui minimise M
## ou bien directement K<-which.min(Mi[2:j])
L<-seq(1,a)
for(i in 1:a)
{
L[i]<-Mi[i+1]
}
L
Kopt<-which.min(L)
```

3.4 la fonction de variation régulère $b(.)$ et le valeur $\tilde{\rho}$

$$b(n/k_{opt}) = \begin{cases} \dfrac{A\rho[\rho+\gamma(1-\rho)]}{(\gamma+\rho)(1-\rho)} a_2(n/k_{opt}), & si\ 0<-\rho<\gamma\ ou\ 0<\gamma<-\rho \quad avec\ D=0 \\ \\ \frac{-\gamma^3}{(1+\gamma)}(n/k_{opt})^{-\gamma} L_2(n/k_{opt}), & si \quad \gamma=-\rho \\ \\ \frac{-\gamma^3 D}{(1+\gamma)}(n/k_{opt})^{-\gamma}, & si\ 0<\gamma<-\rho \qquad D\neq 0, \end{cases}$$

avec, $A \neq 0$ et $D \in \mathbb{R}$. Pour la fonction a_2 à variation régulière avec l'indece ρ. Comme d'habitude, nous supposerons que $\rho < 0$ et nous supposerons également que la partie à variation lente de a_2 est asymptotiquement équivalente à une constante positive, qui peut et sera toujours égale à 1, et le paramètre du second ordre peut être fixé à une valeur ρ, les meilleurs résultats sont obtenus si elle est fixée (généralement on prend ρ égale -1). (*Einmahl et al.*(2008)). C'est à dire:

$$\begin{aligned} \lim_{x\to\infty} a_2(x) &= 0,\ \text{et}\ a_2(x) = x^{\rho}\ell(x)\ \ \text{avec}\ \ell(x)=1, \\ a_2(n/kopt) &= (kopt/n) \end{aligned}$$

$$\tilde{\rho} = \begin{cases} -\gamma, & si \quad 0<\gamma<-\rho \qquad avec\ D\neq 0 \\ \\ \rho, & si\ -\rho\leq\gamma\ ou\ \ 0<\gamma<-\rho \quad avec\ D=0 \end{cases}$$

3.4.1 Programme sous R

Simulation la fonction $b(.)$ n=?

kopt=?

rho=-1

A=?

gamma=?

L2=?

```
if (-rho<=gamma){
b<-((-gamma^3)/(1+gamma))*((n/kopt)^(-gamma))*L2
}else if(-rho<gamma){
```

```
b<-(A*rho*(rho+gamma*(1-rho)))*((kopt/n)^(-gamma))/((gamma+rho)*(1-rho))
}else{
D<-?
b<-((D*(-gamma^3)) /(1+gamma))*((n/kopt)^(-gamma))
}
b
```

simulation le valeur $\tilde{\rho}$ gamma= ?

```
rho=-1
if (-rho<gamma){
trho<-rho
}else if(D<-0){
trho<-rho
}else {
trho<-(-gamma)
}
trho
```

3.5 La fonction quantile de la queue $H^{\leftarrow}(.)$

La loi de simulation utilisée dans ce cas est une loi de Pareto de paramètre γ, Un échantillon $(X_i)_{1\leq i\leq n} \frown \text{Pareto}(\gamma_1)$ de fonction répartition :

$$F(x) = 1 - x^{-1/\gamma_1}.$$

L'échantillon $(Y_i)_{1\leq i\leq n} \frown \text{Pareto}(\gamma_2)$ de fonction de répartition:

$$G(x) = 1 - x^{-1/\gamma_2}.$$

Les variables que nous observons sont d'une part les $Z_i \backsim$ Pareto(γ) de fonction de répartition connue, définie par:

$$\begin{aligned} H(x) &= 1 - x^{-1/\gamma}, \text{ on pose } y = H(x) \\ y &= 1 - x^{-1/\gamma} \\ x &= (1-y)^{-\gamma}, \text{ et } x = H^{-1}(y) \\ H^{-1}(y) &= (1-y)^{-\gamma}, \text{ on pose } y = 1 - \frac{1}{x} \\ H^{-1}(1-\frac{1}{x}) &= (1-1+\frac{1}{x})^{-\gamma} \\ &= (\frac{1}{x})^{-\gamma} \end{aligned}$$

Dans la cas générale, la distribution de la fonction de quantile de queue est inconnue. On utilisée la condition de second ordre formulée en fonction de la fonction de quantile de la queue $U_H(n) = H^{\leftarrow}(1-\frac{1}{n})$. De la théorie de la variation régulière généralisée de second ordre décrit dans de *Haan et Stadtmüller* (1996). Pour $U_H(n) \in GRV_2(\gamma, \rho, \ell_+ n^\gamma, a_2(n), A)$ on a :

$$U_H(n) = H^{\leftarrow}(1-\frac{1}{n}) \begin{cases} \ell_+ n^\gamma \left\{ \frac{1}{\gamma} + \frac{A}{\gamma+\rho} a_2(n)(1+\circ(1)) \right\} & , si \;\; 0 < -\rho < \gamma \\ \ell_+ n^\gamma \left\{ \frac{1}{\gamma} + n^{-\gamma} L_2(n) \right\} & \gamma = -\rho \\ \ell_+ n^\gamma \left\{ \frac{1}{\gamma} + D n^{-\gamma} + \frac{A}{\gamma+\rho} a_2(n)(1+\circ(1)) \right\} & , si \;\; 0 < \gamma < -\rho \end{cases}$$

avec , $D = \frac{1}{\ell_+} \lim_{n\to\infty} \{U_H(n) - a(n)/\gamma\}$ et $\ell_+ > 0$, $D \in \mathbb{R}$ et

$$L_2(n) = B + \int_1^n (A + \circ(1)) \frac{\ell_2(t)}{t} dt + \circ(\ell_2(n))$$

B est constant et ℓ_2 est une fonction à variation lente.

programme sous R

```
#calcule la fonction quantile de la queue #
    for(i in 1:kopt){
```

```
l[i]<-(i/n)^(-gamma)
}
l
H<-sum((p*(i/n)^(-gamma))-p)
H
```

3.6 Déroulement du test

3.6.1 Pour γ_0 fixé

Sous l'hypothèse $H_0 : \gamma_1 = \gamma_0$, avec un seuil de signification $\alpha = 0.05$:

```
#paramétre de simulation#
rm(list=ls())
n=1000
alpha=0.05
gamma0<-  ?
gamma1<- ?
gamma2<- 2.5
p<-gamma2/(gamma0+gamma2)
gamma<-(gamma0*gamma2)/(gamma0+gamma2)
gamma
A<-3     # valeur positive
D<-1      #valeur positive voir.J. EINMAHL(2008)#
L2<-1    #valeur voir.J. EINMAHL(2008)#
rho<-(-1) #paramétre de second order negative#
#échantillon de simulation#
u<-runif(n,0,1)
x<-(1-u)^(-gamma1)
v<-runif(n,0,1)
y<-(1-v)^(-gamma2)
```

```
z<-pmin(x,y)
delta=as.numeric(x<=y)
z
delta
r<-rank(z)
#ordonnées les données
Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for (i in 1 :length(z))
{
L[i]<-(delta[i]+length(z))*r[i]
}
L
A<-sort(L)
A
for (i in 1 :length(z))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(z)
Delta
Z<-log(Y)
K<- n^(0.45)
kopt<-ceiling(K)
kopt
#variable de test #
l<-seq(1,kopt)
```

```
hill1<-(((1/kopt)*sum(Z[n-l+1]))-Z[n-kopt])/((1/kopt)*sum(Delta[n-l+1]))
hill1
acen<-(sqrt(kopt))*(hill1-gamma0)
acen
#calcule la fonction à variation lente #
if (-rho==gamma){
b<-((-gamma^3)/(1+gamma))*((n/kopt)^(-gamma))*L2
}else if(-rho<gamma){
b<-(A*rho*(rho+gamma*(1-rho)))*(kopt/n)/((gamma+rho)*(1-rho))
}else{
b<-((D*(-gamma^3))/(1+gamma))*((n/kopt)^(-gamma))
}
b
#calcule la teld rho #
if (-rho<gamma){
trho<-rho
}else{
trho<-(-gamma)
}
trho
#calcule la fonction inverse #
for(i in 1:kopt){
l[i]<-(i/n)
}
l
H<-sum(((p)*((i/n)^(-gamma)))-p)
H
# calcule la statistique de test#
sT<-((acen)*((sqrt(p))/(gamma0)))+(((1/sqrt(kopt))*(H))/(sqrt(p)))
-((sqrt(kopt))*(sqrt(p))*(b))/((trho)+(gamma*(1-(trho))))
```

```
sT
T<-abs(sT)
T
#Qauntile de test #
Q<-qnorm (1-(alpha/2))
Q
#decision de test #
if (T<Q){
"accept H0"
}else{
"rejet H0"
}
```

Résultats pour γ_0 fixé

On test l'hypothèse:

$$\begin{aligned} \gamma_1 &= 1.05 \\ \gamma_1 &\neq 1.05 \end{aligned}$$

aprés à l'aide des logiciels d'analyse statistique R, pour le traitement des données (calculs numérique) on observée les résultat suivants :

- Pour $kopt$ dans ce cas $kopt$ (nombre des valeurs extrêmes).

[1] 100

- $\hat{\gamma}_1^{(c,H)} = hill1$ (l'estimateur de l'indice de queue censuré).

[1] 0.8328943

- la fonction $b(n/kopt) = b$

[1] -0.101467

- la valeur $\tilde{\rho} = trho$

[1] -0.7394366

- $H^{\leftarrow}(1 - \frac{i}{n}) = l[i]$

[1] 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030 0.0035 0.0040 0.0045

[10] 0.0050 0.0055 0.0060 0.0065 0.0070 0.0075 0.0080 0.0085 0.0090

[19] 0.0095 0.0100 0.0105 0.0110 0.0115 0.0120 0.0125 0.0130 0.0135

[28] 0.0140 0.0145 0.0150 0.0155 0.0160 0.0165 0.0170 0.0175 0.0180

[37] 0.0185 0.0190 0.0195 0.0200 0.0205 0.0210 0.0215 0.0220 0.0225

[46] 0.0230 0.0235 0.0240 0.0245 0.0250 0.0255 0.0260 0.0265 0.0270

[55] 0.0275 0.0280 0.0285 0.0290 0.0295 0.0300 0.0305 0.0310 0.0315

[64] 0.0320 0.0325 0.0330 0.0335 0.0340 0.0345 0.0350 0.0355 0.0360

[73] 0.0365 0.0370 0.0375 0.0380 0.0385 0.0390 0.0395 0.0400 0.0405

[82] 0.0410 0.0415 0.0420 0.0425 0.0430 0.0435 0.0440 0.0445 0.0450

[91] 0.0455 0.0460 0.0465 0.0470 0.0475 0.0480 0.0485 0.0490 0.0495

[100] 0.0500

- T statistique de test est :

  [1] 0.5071794

- Q = Qauntile de test :

  [1] 1.959964

- Decision de test:

  [1] "accept H0"

Donc on accepté l'hypothése $H_0 : \gamma_1 = \gamma_0 = 1.05$

3.6.2 Pour γ_0 varié

```
rm(list=ls())
    n=1000
    alpha=0.05
    gamma1<-1.05
    gamma2<- 2.5
    D<-1#valeur positive, voir J. EINMAHL(2008)#
    L2<-1#valeur, voir J. EINMAHL(2008)#
    rho<-(-1)#paramétre de second order negative#
    #échantillon de simulation#
    u<-runif(n,0,1)
```

```
x<-(1-u)^(-gamma1)
v<-runif(n,0,1)
y<-(1-v)^(-gamma2)
z<-pmin(x,y)
delta=as.numeric(x<=y)
z
delta
r<-rank(z)
#ordonnées les données
Y<-sort(z)
r1<-sort(r)
L<-seq(1,length(z))
B<-seq(1,length(z))
for (i in 1 :length(z))
{
L[i]<-(delta[i]+length(z))*r[i]
}
L
A<-sort(L)
A
for (i in 1 :length(z))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(z)
Delta
Z<-log(Y)
#koptimal#
K<- n^(0.45)
```

```
kopt<-ceiling(K)
kopt
#variable de test #
l<-seq(1,kopt)
hill1<-(((1/kopt)*sum(Z[n-l+1]))-Z[n-kopt])/((1/kopt)*sum(Delta[n-l+1]))
hill1
a<-0.1
d<-3.5
m<-50
gamma0<-seq(a,d,by=((d-a)/(m-1)))
acen<-seq(1,m)
gamma<-seq(1,m)
sT<-seq(1,m)
p<-seq(1,m)
H<-seq(1,m)
b<-seq(1,m)
for(j in 1:m){
p[j]<-gamma2/(gamma0[j]+gamma2)
gamma[j]<-((gamma0[j])*gamma2)/((gamma0[j])+gamma2)
acen[j]<-(sqrt(kopt))*(hill1-(gamma0[j]))
#calcule la fonction à variation lente #
if (-rho==gamma[j]){
b[j]<-(((-gamma[j])^3)/(1+gamma[j]))*((n/kopt)^(-gamma[j]))*L2
}else if(-rho<gamma[j]){
A<-3#valeur positive JOHN H.J. EINMAHL(2008)#
b[j]<-(A*rho*(rho+gamma[j]*(1-rho)))*(kopt/n)/((gamma[j]+rho)*(1-rho))
}
else{
b[j]<-((D*((-gamma[j])^3))/(1+gamma[j]))*((n/kopt)^(-gamma[j]))
}
```

```
b
#calcule la teld rho #
if (-rho<gamma[j]){
trho<-rho
}else{
trho<-(-gamma[j])
}
trho
#calcule la fonction inverse #
for(i in 1:kopt){
l[i]<-(i/n)
}
H[j]<-sum(((p[j])*((i/n)^(-gamma[j])))-p[j])
# calcule la statistique de test#
sT[j]<-((acen[j])*((sqrt(p[j]))/(gamma0[j])))+(((1/sqrt(kopt))*(H[j]))/(sqrt(p[j])))
-((sqrt(kopt))*(sqrt(p[j]))*(0.5))/((trho)+((gamma[j])*(1-(trho))))
T[j]<-abs(sT[j])
}
#Qauntile de test #
Q<-qnorm (1-(alpha/2))
Q
plot(gamma0,sT,ylim=c(-20,40),main='statistique du test',ylab='quantile'
,xlim=c(0.1,3.5),xlab='gamma0',type='l')
abline(h=Q,col=4)
abline(h=(-Q),col=4)
abline(v= gamma1,col=3)
#decision de test #
kk=as.numeric(T<=Q)
kk
```

Résultat pour γ_0 varié

> kk=as.numeric(T<=Q)

> kk

[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

[39] 0 0 0 0 0 0 0 0 0 0 0 0

Alors 1= accepte et 0=rejet

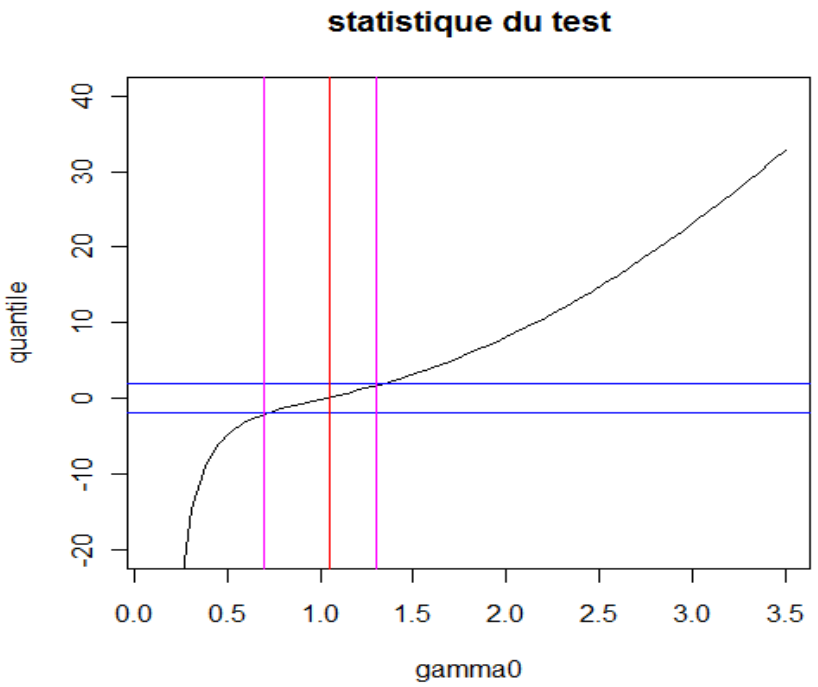

Fig1: $\gamma_1 = 1.05$ avec $[0.7; 1.3]$ l'intervale d'accepté les valeurs γ_0

Commentaires On voit sur le graphique ci-dessus que notre test est efficace est trés satisfisant, car, il accepte les valeurs proche de la vrais valeur de γ_1 et refuse en contre partie celles qui s'eloignent. Ce qui rend notre statistique de test trés signifiante.

3.7 Conclusion

L'objectif de ce mémoire est de proposer un nouveau test d'hypothèse paramétrique sur l'indice de queue (ou l'indice des valeurs extrêmes) dans le cas du domaine d'attraction de fréchet pour des données censurées à droite.

On test l'hypothèse

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

avec, $\hat{\gamma}_1^{(c,H)}$ est l'estimateur en question de γ_1 *Einmahl et al.* (2008)

$$\hat{\gamma}_1^{(c,H)} = \frac{\hat{\gamma}_1^H}{\hat{p}}.$$

Notre but essentiel était de retrouver toute les valeurs γ_0 acceptées par notre test et garantir en même temps les conditions de la normalité asymptotique de $\hat{\gamma}_1^{(c,H)}$. Nous avons mentionné dans ce travail que le nombre des valeurs extrêmes k est très important pour la prise de décision du test. Nous avons considéré trois méthodes pour le choix du meilleur k optimal qui sera adaptatif à la normalité asymptotique et donne une meilleur valeur de l'estimateur $\hat{\gamma}_1^{(c,H)}$ qui converge en loi vers γ_1. Pour un bon déroulement de notre test nous avons choisit le nombre de valeurs extrêmes selon un moyen assez utilisé en simulation, pour $0 < \theta < 1$, $kopt = [n]^\theta$. Ce nombre de valeur extrêmes qui assure un bon fonctionnement de notre test reste un sujet pour travailler dessus en signalant d'après nos simulations que le nombre k doit être relativement petit par rapport à la taille de l'échantillon n autrement dit lorsque k/n tends vers 0 quand n tends vers l'infini.

L'efficacité de ce test consiste en sa capacité de regrouper toute les valeurs de γ_0 sous forme d'intervalle à accepter pour la détermination de la loi de l'échantillon réelle par l'intermédiaire de la loi limite de $\hat{\gamma}_1^{(c,H)}$.

Comme conclusion à tirer de ce test, nous dirons que ce dernier fournit un moyen d'ajustement optimal d'un échantillon de variables aléatoires de loi de Pareto dans notre cas, lorsque le paramètre γ_1 est inconnue, en quelque sorte une estimation ponctuelle de γ_1 par intervalle.

Bibliographie

[1] A.Ameraoui, K. Boukhetala, J.Fran¸c.Dupuy « Bayesian estimation of the tail index of a heavy tailed distribution under random censoring » Article HAL .USTHB Algerie Et Université France 2014

[2] B.Souad « Statistics of incomplete data » thèse De Doctorat, Université Biskra 2016

[3] Cheng, S. and Peng, L. (2001). Confidence Intervals for the Tail Index. Bernoulli **7**, $751 - 760$.

[4] de Haan, L., and Ferreira, A., 2006, Extreme Values Theory : An introduction. New York,Springer.

[5] Deme, E.H., 2013, Quelques contributions à la Théorie univariée des Valeurs Extrêmes et Estimation des mesures de risque actuariel pour des pertes à queues lourdes,Université Gaston Berger.

[6] Einmahl, J.H.J.,Fils-Villetard, A., Guillou, A,2008, Statistics of extremes under random censoring. Bernoulli, 14:207–227.

[7] EL H. DEME « Quelques Contributions À La ThÈorie univariee Des Valeurs Extrêmes Et Estimation Des Mesures De Risque Actuariel Pour Des Pertes À Queues Lourdes » thèse de doctorat Université Gaston Berger De Saint-Louis 2013.

[8] Frédéric B., Tests paramètriques, université de Strasbourg, France.

[9] J. Beirlant, G . Dierckxa ,A. Guillou « Estimation of the extreme-value index and generalized quantile plots » Article , Université Leuven et Université Paris 2005.

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

[11] Ivette Gomes,M.,and Manuela Neves,M.,2010,Estimation of the Extreme Value Index for Randomly Censored Data.

[12] Gaston Berger, Saint Louis, Sénégal et université France 2014.

[13] L. GARDES « Estimation d'une fonction quantile extrême » Formation Doctorale ,UNIVERSITÈ MONTPELLIER II 2013.

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

[15] Reiss,R.-D.,Thomas,M.S.,2007, Statistical analysis of extreme values. From insurance,finance, hydrology and other fields. Birkhäuser Verlag, Basel., Boston, Berlin

[16] Soltane,Louiza,2016, Analyse des Valeurs Extrêmes en présence de censure. Thése de doctorat de université Mohamed khider, Biskra, Algeria.

[17] Touba,S., Sur L'estimation des parametres des lois stables,thése de doctorat de université Mohamed khider, Biskra, Algeria.

[18] Toulemonde,G.,2008, Estimation et tests en théorie des valeurs extrêmes,thése de doctorat de l'université Paris VI.

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

[20] Inférence Statistique Assistée par Ordinateur - 2A, 2006-2007.

Résumé

Résumé

JOAN A.J.E et al (2008) ont proposé un estimateur de l'indice de queue par le biais de l'estimateur de Hill (1975) sous des données censurées, ceci en effectuant sa loi limite. Dans notre travail on appliquer un test d'hypothèse paramétrique sur cet estimateur de l'indice des valeurs extrême précité. Nous allons présenter l'algorithme correspondant basé sur sa normalité asymptotique. L'hypothèse H_0 sera de tester d'accepter ou rejeter n'importe quell valeurs proposé de l'indice de queue, Nous illustrerons notre étude par des simulations afin de monter l'utilité de notre test proposé.

Mots clés: Théorie des valeurs extrems , censure aléatoire à droite, indice des valeurs extremes, estimateur de Hill, test paramétrique, statistique du test

Abstract

JOAN A.J.E et al (2008) proposed an estimator of the tail index through the Hill estimator (1975) under censored data, by carrying out its limiting law. In our work we apply a parametric hypothesis test on this estimator of the extreme value index mentioned above. We will present the corresponding algorithm based on its asymptotic normality. The hypothesis H_0 will be to test to accept or reject any proposed values of the tail index, We will illustrate our study by simulations in order to test the utility of our proposed test.

Key words: Extreme value theory, right random censoring, extreme value index, Hill estimator, parametric test, test statistic