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

GHARBI Djouhaina

Titre :

Estimation des quantiles extrêmes

Membres du Comité d'Examen :

Pr.	**BRAHIMI Brahim**	UMKB	Président
Dr.	**BENAMEUR Sana**	UMKB	Encadreur
Dr.	**DHIABI Samra**	UMKB	Examinateur

21 Septembre 2020

Dédicace

Je dédie ce travail à :

Mes parents

- A mes frères

et mes sœurs, et toute ma famille

- A mes chers amis

- Je tiens à remercier

Mon encadreur : Dr. Benameur Sana

- Tous les membres de ma promotion

et tous mes professeurs

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

Djouhaina Gharbi

REMERCIEMENTS

Je tiens tout d'abord à remercier **Allah** le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, je tiens à remercier mon encadreur : **Dr. Benameur Sana**, pour ses précieux conseils et son aide durant toute la période du travail.

Mes vifs remerciements vont également aux membres du jury : **Pr. Brahimi Brahim et Dr.Dhiabi Samra** pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Mes remerciements s'étendent également à tous mes enseignants durant les années d'études.

Ma famille et mes amis qui par leurs prières et leurs encouragements, on a pu surmonter tous les obstacles.

Enfin, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Merci

Table des matières

Table des figures

Liste des tableaux

Introduction

Ces quatre dernières décennies, il y a eu une recrudescence d'évènements extrêmes tels que les catastrophes naturelles : inondations, séismes de forte intensité, vents violents, crues de rivières inhabituelles, ou bien les crises financières, etc. Ce sont des évènements rares dont les probabilités d'occurrence sont faibles, ainsi qu'ils s'écartent fortement de la moyenne ou de la tendance habituelle et qui ont des conséquences désastreuses pour l'être humain et l'environnement.

La théorie des valeurs extrêmes vient en complément de la statistique classique où il est usuel d'étudier les variables aléatoires autour de leurs moyennes. Nous étudions dans ce travail le comportement de la queue de distribution, en utilisant les deux principaux formalismes issue de cette théorie tels que : la distribution des valeurs extrêmes généralisées (GEV) et la distribution de Pareto généralisé (GPD).

Cette étude va nous permettre d'estimer des quantités appelées quantiles extrêmes dont la probabilité d'observation est très faible, c-à-d l'ordre des quantiles converge vers un quand la taille de l'échantillon tend vers l'infini.

L'objectif principal de ce mémoire est l'estimation des quantiles extrêmes sous données complètes ainsi sous des données censurées. Ce travail est composé de deux chapitres.

Dans le premier chapitre, nous présentons un état de l'art en théorie des valeurs extrêmes et en théorie de censure. Après avoir introduit la notion des statistiques d'ordres,

nous présentons la distribution du maximum d'un échantillon, ensuite les deux principaux outils servant à modéliser le comportement des valeurs extrêmes : la distribution GEV et la GPD. On s'intéresse ensuite à la caractérisation des domaines d'attraction. Enfin, nous énonçons les différentes notions de base sur la censure.

Dans le deuxième chapitre, nous intéressons aux différentes méthodes d'estimation des quantiles extrêmes. On commence par l'estimation sous données complètes : les estimateurs basés sur l'approche semi-paramétrique en utilisant l'estimateur de Hill, de Pickands et de moments pour l'indice de queue, l'estimation basée sur la loi des valeurs extrêmes et celle basée sur la méthode des excès. Ensuite, on introduit l'estimateur de Kaplan-Meier et l'estimateur de Hill adapté afin d'estimer les quantiles extrêmes sous données censurées. Le comportement de ce dernier est illustré sur la base des simulations sous logiciel R.

Chapitre 1

Théorie des valeurs extrêmes et censure

Dans ce chapitre, nous énonçons des fondamentales sur la théorie des valeurs extrêmes (TVE), qui est utilisée pour la modélisation des évènements extrêmes. Cette théorie se repose principalement sur deux approches, la première approche appelée GEV ; elle permet de modéliser les blocks des maximas par une distribution GEV (Generalized Extreme Value) et la seconde, appelée GPD consiste à ajuster les observations dépassant un certain seuil (POT : Peaks Over Threshold) par une GPD (Generalized Pareto Distribution). Pour des descriptions détailler de la TVE, consulter les excellents bouquins comme Embrechtes et al [9], Beirlant et al [2] et Reiss et Thomas [21]. Nous discutons également dans ce chapitre des notions de base sur la censure.

1.1 Définitions et caractéristiques de bases

On va commencer par quelques rappels sur la fonction de répartition, la fonction de survie, le théorème central limite, les lois des grands nombre et les statistiques d'ordre.

Définition 1.1.1 (Fonction de répartition)

La fonction de répartition F d'une variable aléatoire (v.a) X est définie par l'application suivante :

$$\begin{aligned} F: \ \mathbb{R} &\to [0,1], \\ x &\longmapsto F(x) := P(X \leq x). \end{aligned} \tag{1.1}$$

Définition 1.1.2 (Fonction de survie)

La fonction de survie d'une v.a X, aussi appelée queue de distribution, que l'on note par $S(t)$ ou $\bar{F}(t)$ est définie sur $\mathbb{R}_+$ par la probabilité qu'un individu vive au-delà d'une date t :

$$S(t) := \bar{F}(t) = 1 - F(t) = P(X > t). \tag{1.2}$$

Définition 1.1.3 (Fonctions empiriques de répartition et de survie)

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de v.a's de taille $n \geq 1$ indépendantes identiquement distribuées (iid) de fonction de répartition commune F. Les fonctions empiriques de répartition F_n et de survie $\bar{F}_n$ sont définies respectivement par :

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}, \tag{1.3}$$

et

$$\bar{F}_n(x) := 1 - F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}}. \tag{1.4}$$

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

Définition 1.1.4 (Fonction de quantile)

La fonction de quantile est définie par :

$$Q(p) := F^{\leftarrow}(p) = \inf\{t : F(t) \geq p\}, \ 0 < p < 1. \tag{1.5}$$

où $F^{\leftarrow}$ est appelé l'inverse généralisée d'une fonction de répartition F avec convention que $\inf\{\varnothing\} = \infty$ *et* $P(X \leq Q(p)) = p$. *On l'exprime en terme de la fonction de survie par :*

$$Q(p) := \inf\{t : \bar{F}(t) \leq 1 - p\},\ 0 < p < 1. \tag{1.6}$$

Définition 1.1.5 (Fonction de quantile empirique)

La fonction de quantile empirique d'un échantillon $(X_1, X_2, \cdots, X_n)$ *est donnée par :*

$$Q_n(p) := F_n^{\leftarrow}(p) = \inf\{t : F_n(t) \geq p\},\ 0 < p < 1. \tag{1.7}$$

Définition 1.1.6 (Fonction quantile de queue)

La fonction quantile de queue est définie par :

$$U(t) := Q\left(1 - \frac{1}{t}\right) = F^{\leftarrow}\left(1 - \frac{1}{t}\right) = \left(1/\bar{F}\right)^{-1}(t),\ 1 < t < +\infty. \tag{1.8}$$

Définition 1.1.7 (Fonction quantile de queue empirique)

La fonction quantile de queue empirique correspondante est

$$U_n(t) := Q_n\left(1 - \frac{1}{t}\right),\ 1 < t < +\infty. \tag{1.9}$$

Définition 1.1.8 (Point terminal)

Le point terminal de la fonction de répartition F est défini par :

$$x_F := \sup_{x \in \mathbb{R}} \{x : F(x) < 1\} \leq +\infty. \tag{1.10}$$

1.1.1 Théorème central limite

Lorsque l'on s'intéresse à la partie centrale d'un échantillon, le résultat clé est le Théorème Central Limite (TCL) qui joue un rôle capital en statistique. Il établit la converge en loi vers la loi normale d'une somme de v.a's iid sous des hypothèses très peu contraignantes.

Théorème 1.1.1 (TCL)

Soit $X_1, X_2, \cdots, X_n$ une suite de v.a's iid définie sur un même espace de probabilité (Ω, A, P), d'une fonction de répartition F, de moyenne $\mu = E[X_1]$ et de variance finie σ^2 $(E[X_1^2] < \infty)$. Considérons la somme et la moyenne arithmétique correspondantes respectivement :

$$S_n := \sum_{i=1}^{n} X_i, \ \overline{X}_n := \frac{S_n}{n}. \tag{1.11}$$

La variable Y_n converge en loi vers la loi normale centrée réduite, c-à-d :

$$Y_n := \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \ \text{ quand } n \to \infty.$$

1.1.2 Lois des grands nombres

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

Théorème 1.1.2 (Lois des grands nombres)

Si $(X_1, X_2, \cdots, X_n)$ un échantillon provenant d'une v.a X, tel que $\mu := E[X] < \infty$,

alors :

$$\begin{aligned} \textit{Loi faible :} \quad & \overline{X}_n \xrightarrow{p} \mu \quad \textit{quand } n \to \infty. \\ \textit{Loi fort :} \quad & \overline{X}_n \xrightarrow{p.s} \mu \quad \textit{quand } n \to \infty. \end{aligned} \tag{1.12}$$

Théorème 1.1.3 (Glivenko-Cantelli)

La converge presque sûrement uniforme de F_n vers F est définie par :

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p.s} 0 \ \textit{quand } n \to \infty. \tag{1.13}$$

1.2 Statistiques d'ordre

Définition 1.2.1 (Statistique d'ordre)

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de v.a's iid, de fonction de répartition F. En rangeant ces v.a's par ordre croissant, les statistiques d'ordre sont alors définies par :

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

Pour $1 \leq k \leq n$, la variable $X_{k,n}$ s'appelle la $k^{ème}$ statistique d'ordre.

Définition 1.2.2 (Statistiques d'ordre extrêmes)

Les statistiques d'ordre extrêmes $X_{1,n}$ et $X_{n,n}$ sont définies (respectivement) par le minimum et le maximum de l'échantillon $(X_1, X_2, \cdots, X_n)$, comme suit :

$$\begin{aligned} X_{1,n} &:= \min\{X_1, X_2, \cdots, X_n\}, \\ X_{n,n} &:= \max\{X_1, X_2, \cdots, X_n\}. \end{aligned} \tag{1.14}$$

où la variable $X_{1,n}$ est la plus petite statistique d'ordre et la variable $X_{n,n}$ est la plus grande statistique d'ordre.

1.2.1 Distribution des statistiques d'ordre

Soit $X_1, X_2, \cdots, X_n$ une suite de v.a 's iid de fonction de répartition commune F.

Distributions du maximum et minimum

La distribution du maximum $X_{n,n}$ est définie par :

$$F_{X_{n,n}}(x) := [F(x)]^n. \tag{1.15}$$

En effet,

$$\begin{aligned} F_{X_{n,n}}(x) &:= P(X_{n,n} \leq x) = P(X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_1 \leq x) = [P(X_1 \leq x)]^n = [F(x)]^n. \end{aligned}$$

La distribution du minimum $X_{1,n}$ est définie par :

$$F_{X_{1,n}}(x) := 1 - [1 - F(x)]^n. \tag{1.16}$$

En effet,

$$\begin{aligned} F_{X_{1,n}}(x) &:= P(X_{1,n} \leq x) = 1 - P(X_{1,n} > x) = 1 - P(X_1 > x, X_2 > x, \cdots, X_n > x) \\ &= 1 - \prod_{i=1}^{n} P(X_1 > x) = 1 - \prod_{i=1}^{n} (1 - P(X_1 \leq x)) \\ &= 1 - [1 - P(X_1 \leq x)]^n = 1 - [1 - F(x)]^n. \end{aligned}$$

Si F est continue, alors les fonctions de densité du minimum $f_{X_{1,n}}$ et du maximum $f_{X_{n,n}}$

sont données respectivement par :

$$f_{X_{1,n}}(x) := nf(x)\left[F(x)\right]^{n-1}, \tag{1.17}$$

et

$$f_{X_{n,n}}(x) := nf(x)\left[1-F(x)\right]^{n-1}. \tag{1.18}$$

Distribution de la $k^{ème}$ statistique d'ordre

La distribution de la $k^{ème}$ statistique d'ordre est définie par :

$$F_{X_{k,n}}(x) := \sum_{i=1}^{n}\binom{n}{i}\left[F(x)\right]^{i}\left[1-F(x)\right]^{n-i},\ x\in\mathbb{R}, \tag{1.19}$$

et la densité de la $k^{ème}$ statistique d'ordre est :

$$f_{X_{k,n}}(x) := \frac{n!}{(k-1)!\,(n-k)!}\left[F(x)\right]^{k-1}\left[1-F(x)\right]^{n-k}f(x),\ 1\le k\le n. \tag{1.20}$$

La démonstration de ces formules est détaillée dans l'ouvrage de Reiss et Thomas [21].

1.2.2 Fonctions empiriques et statistique d'ordre

Les fonctions $F_n(t)$ et $\bar{F}_n(t)$ s'écrivent en terme des statistiques d'ordre comme suit :

$$F_n(t) = \begin{cases} 0 & si \quad t < X_{1,n}, \\ \dfrac{i}{n} & si \quad X_{i,n}\le t\le X_{i+1,n}, \\ 1 & si \quad t\ge X_{n,n}. \end{cases} \tag{1.21}$$

et

$$\bar{F}_n(x) = \begin{cases} 1 & si \quad t < X_{1,n}, \\ 1 - \dfrac{i}{n} & si \quad X_{i,n} \leq t \leq X_{i+1,n}, \\ 0 & si \quad t \geq X_{n,n}. \end{cases} \tag{1.22}$$

La fonction de quantile empirique peut être exprimé comme une fonction simple des statistiques d'ordre concernant l'échantillon $(X_1, X_2, \cdots, X_n)$

$$Q_n(p) = X_{i,n} \text{ si } \frac{i-1}{n} \leq p \leq \frac{i}{n}, \ i = 1, \cdots, n.$$

On note que $X_{[np],n}$ est le quantile empirique d'ordre p où $[np]$ est la partie entière de np.

1.3 Distributions des valeurs extrêmes

Si $(X_1, X_2, \cdots, X_n)$ est un échantillon de n v.a's, la limite de la distribution du maximum $X_{n,n}$ est donnée comme suite :

$$\lim_{n\to\infty} F_{X_{n,n}}(x) = \lim_{n\to\infty} F^n(x) = \begin{cases} 1 & si \quad F(x) = 1, \\ 0 & si \quad F(x) < 1. \end{cases} \tag{1.23}$$

Le résultat nous indique que la limite de la distribution de $X_{n,n}$ est une loi dégénérée. Ce résultat fournit très peu d'informations sur le comportement de $X_{n,n}$. On aimerait obtenir une loi non dégénérée pour $X_{n,n}$.

De façons analogue au TCL, les travaux de Fisher et Tippett [11] en (1928), Gnedenko [12] en (1943) et de Haan [8] en (1976) sont établit la loi asymptotique du maximum $X_{n,n}$ convenablement normalisé d'un échantillon.

Théorème 1.3.1 (Fisher & Tippett (1928))

Si $X_1, X_2, \cdots, X_n$ *est une suite de v.a's iid de fonction de répartition commune* F *et de statistique d'ordre associe* $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ *et il existe deux suites normalisantes réelles* $(a_n)_{n\geq 0} > 0$, *et* $(b_n)_{n\geq 0} \in \mathbb{R}$, *telle que :*

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) := \lim_{n\to\infty} F^n\left(a_n x + b_n\right) := H_\gamma(x), \forall x \in \mathbb{R}. \tag{1.24}$$

où H_γ *est une fonction de distribution non dégénérée, appelée distribution des valeurs extrêmes (EVD : Extremes Values Distribution). La distribution* H *est du même type que l'une des trois distributions des valeurs extrêmes standard suivantes :*

Distribution de Gumbel : $\Lambda(x) := \exp\left[-\exp(-x)\right];\ x \in \mathbb{R}.$

Distribution de Fréchet : $\Phi_\alpha(x) := \begin{cases} 0 & si \quad x \leq 0, \\ \exp\left(-x^{-\alpha}\right) & si \quad x > 0. \end{cases},\ \alpha > 0.$

Distribution de Weibull : $\Psi_\alpha(x) := \begin{cases} \exp\left[-(-x)^{\alpha}\right] & si \quad x \leq 0, \\ 1 & si \quad x > 0. \end{cases},\ \alpha > 0.$

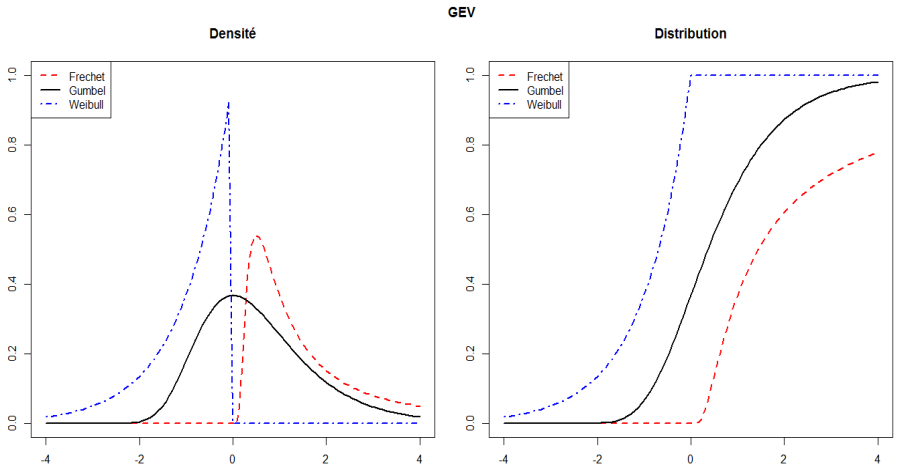

FIG. 1.1 – Densités et Distribution de Lois des Valeurs Extrêmes.

Exemple 1.3.1 (Loi uniforme)

Soit $X_1, X_2, \cdots, X_n$ une suite de v.a's iid selon une loi uniforme standard de fonction de répartition

$$F(x) := \begin{cases} 0 & si \quad x < 0, \\ x & si \quad x \in [0,1], \\ 1 & si \quad x > 1. \end{cases}$$

Prenons $a_n = \dfrac{1}{n}$ et $b_n = 1$, alors $\dfrac{X_{n,n} - b_n}{a_n}$ tend asymptotiquement vers la loi de ***Weibull*** *avec $\alpha = 1$, en effet :*

$$\lim_{n\to\infty} P\left(\frac{X_{n,n}-1}{\frac{1}{n}} \leq x\right) = \lim_{n\to\infty} F^n\left(\frac{1}{n}x+1\right) = \lim_{n\to\infty}\left(\frac{1}{n}x+1\right)^n = \exp(x)$$
$$= \exp\left[-(-x)^1\right] = \Psi_1(x).$$

1.4 Distribution GEV

Définition 1.4.1 (Jenkinson (1955))

La fonction de distribution H_γ de la famille des valeur extrêmes généralisées (GEV : Generalized Extreme Value), pour $\gamma \in \mathbb{R}$, est définie par :

$$H_\gamma(x) := \begin{cases} \exp[-(1+\gamma x)^{-1/\gamma}] & si \quad \gamma \neq 0,\ 1+\gamma x > 0, \\ \exp[-\exp(-x)] & si \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases} \tag{1.25}$$

où γ s'appelle paramètre de forme ou indice des valeur extrême (EVI : Extreme Value Index).

La fonction de densité de probabilité h_γ associée est définie par :

$$h_\gamma(x) := \begin{cases} H_\gamma(x)\,(1+\gamma x)^{-1/\gamma-1} & si \quad \gamma \neq 0,\ 1+\gamma x > 0, \\ \exp[-x-\exp(-x)] & si \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases} \tag{1.26}$$

Pour $\gamma = 0$, il faut lire $H_0(x) = \exp[-\exp(-x)],\ x \in \mathbb{R}$, qui s'obtient dans H_γ en faisant tendre γ vers 0. Les lois des valeurs extrêmes généralisées correspondent, à une translation et un changement d'échelle prés, aux lois des valeurs extrêmes standard. Nous avons alors les correspondances suivantes :

$$\begin{aligned} \Lambda(x) &:= H_0(x). \\ \Phi_\alpha(x) &:= H_{\frac{1}{\alpha}}(\alpha(x-1)),\ x \in \mathbb{R}. \\ \Psi_\alpha(x) &:= H_{\frac{-1}{\alpha}}(\alpha(x+1)),\ x \in \mathbb{R}. \end{aligned}$$

La forme générale de $H_\gamma(x)$ dite forme paramétrée de von Mises dans laquelle on fait apparaître un paramètre de localisation $\mu \in \mathbb{R}$ et un paramètre d'échelle $\sigma > 0$, pour les variables non centrées($\mu \neq 0$), et non réduites ($\sigma \neq 1$), est définie par :

$$H_{\mu,\sigma,\gamma}(x) := \begin{cases} \exp\left\{-\left[1+\gamma\dfrac{(x-\mu)}{\sigma}\right]^{-1/\gamma}\right\} & si \quad \gamma \neq 0,\ 1+\gamma\dfrac{(x-\mu)}{\sigma} > 0, \\ \exp\left\{-\exp\left[-\left(\dfrac{x-\mu}{\sigma}\right)\right]\right\} & si \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases} \tag{1.27}$$

1.5 Domaines d'attraction

Définition 1.5.1 (Domaine d'attraction)

On dit qu'une distribution F appartient au domaine d'attraction de H_γ, et on note $F \in \mathcal{D}(H_\gamma)$ si la distribution du maximum renormalisée converge vers H_γ. Autrement

dit, s'il existe des constantes réelles $a_n > 0$, *et* $b_n \in \mathbb{R}$ *tels que :*

$$\lim_{n\to\infty} F^n\left(a_n x + b_n\right) = H_\gamma(x) = \exp[-(1+\gamma x)^{-1/\gamma}],\ 1+\gamma x > 0,\ \forall x \in \mathbb{R}. \tag{1.28}$$

La caractérisation des domaines d'attraction fait appel à la notion de fonctions à variations régulières.

Définition 1.5.2 (Fonction à variation régulière)

On dit qu'une fonction h *mesurable sur* $\mathbb{R}_+$ *est à variation régulière d'indice* $\rho \in \mathbb{R}$ *et on note* $h \in RV_\rho$, *si*

$$\lim_{t\to\infty} \frac{h\left(tx\right)}{h\left(t\right)} = x^\rho,\ x > 0. \tag{1.29}$$

Définition 1.5.3 (Fonction à variation lente)

Si une fonction $L : \mathbb{R} \mapsto \mathbb{R}_+$ *est à variation régulière d'indice* $\rho = 0$ $(L \in RV_0)$, *on dit que* L *est à variation lente, telle que :*

$$\lim_{t\to\infty} \frac{L\left(tx\right)}{L\left(t\right)} = 1,\ x > 0. \tag{1.30}$$

Remarque 1.5.1 *Toute fonction* $h \in RV_\rho$ *peut s'écrire sous la forme :*

$$h\left(x\right) := x^\rho L\left(x\right),\ \text{où } L \in RV_0. \tag{1.31}$$

Définition 1.5.4 (Condition du second ordre)

On dit que la fonction de queue de quantile U *est à variation régulière du second ordre avec le paramètre du premier ordre* $\gamma > 0$ *et le paramètre du second ordre* $\rho \leq 0$, *on écrit* $U \in 2RV_{\gamma,p}$, *s'il existe une fonction* $A^*\left(t\right) \to 0$ *et ne change pas le signe au voisinage*

de ∞*, telles que*

$$\lim_{n\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A^{*}(t)} = x^{\gamma}\frac{x^{\rho}-1}{\rho},\ x > 0. \tag{1.32}$$

où $|A^{*}| \in RV_{\rho}$ *est appelée la fonction auxiliaire de* U*.*

1.5.1 Caractérisation des domaines d'attraction

Théorème 1.5.1 (Caractérisation du $D(\Phi_{\gamma})$**)**

Une fonction de répartition F *appartient au domaine d'attraction de* ***Fréchet*** *de paramètre* $\gamma > 0$ *ssi* $x_F = \infty$ *et*

$$\bar{F}(x) := x^{-\gamma} L(x), \tag{1.33}$$

où la fonction $L \in RV_0$*. De plus si* $F \in \mathcal{D}(\Phi_{\gamma})$*, alors avec* $a_n = U(n) = F^{\leftarrow}(1-1/n)$ *et* $b_n = 0$*, la suite* $(a_n^{-1} X_{n,n})_{n\geq 1}$ *converge en loi vers une v.a de fonction de répartition* Φ_{γ}*, quand* $n \to \infty$*.*

Théorème 1.5.2 (Caractérisation du $D(\Psi_{\gamma})$**)**

Une fonction de répartition F *appartient au domaine d'attraction de* ***Weibull*** *de paramètre* $\gamma > 0$ *ssi* $x_F < \infty$

$$\bar{F}(x_F - 1/x) := x^{-\gamma} L(x), \tag{1.34}$$

où la fonction $L \in RV_0$*. De plus si* $F \in \mathcal{D}(\Psi_{\gamma})$*, alors avec* $a_n = x_F - U(n) = x_F - F^{\leftarrow}(1-1/n)$ *et* $b_n = x_F$*, la suite* $(a_n^{-1}(X_{n,n} - x_F))_{n\geq 1}$ *converge en loi vers une v.a de fonction de répartition* Ψ_{γ}*, quand* $n \to \infty$*.*

Théorème 1.5.3 (Caractérisation du $D(\Lambda)$**)**

Une fonction de répartition F *appartient au domaine d'attraction de* ***Gumbel*** *de paramètre* $\gamma > 0$ *ssi*

$$\bar{F}(x) := c(x) \exp\left[-\int_z^x \frac{g(t)}{a(t)} dt\right],\ z < x < x_F, \tag{1.35}$$

où c et g sont deux fonctions mesurables satisfaites $c(x) \to c > 0$, $g(x) \to 1$ quand $x \to x_F$ et a est une fonction positive, absolument continue (par rapport la mesure de Lebesgue) avec la densité a' ayant $\lim_{n\to\infty} a'(x) = 0$. Dans ce cas, un choix possible pour les suites de normalisation est :

$$a_n := x_F - F^{\leftarrow}(1 - 1/n) \text{ et } b_n = \frac{1}{\bar{F}(x)} \int_{a_n}^{x_F} \bar{F}(y)\, dy. \tag{1.36}$$

Un classement de quelques lois usuelles par domaine d'attraction est présenté dans le tableau 1.1.

Fréchet($\gamma > 0$)	**Gumbel**($\gamma = 0$)	**Weibull**($\gamma < 0$)
Pareto Student Burr Chi-deux Cauchy Fréchet Log-gamma	Gamma Weibull Normale Exponentielle Gumbel Log-normale	Uniforme Beta

TAB. 1.1 – Domaines d'attraction de quelques lois usuelles

1.6 Distribution GPD

1.6.1 Distribution des excès

Nous supposons $X_1, X_2, \cdots, X_n$ est une suite de v.a's iid de loi F et de point terminal x_F. Plutôt que se focaliser sur le maximum de l'échantillon, on étudie les valeurs dépassant un certain seuil à fixer. Cette approche est appelée POT (Peaks Over Threshold ou bien piques au-delà d'un seuil).

Pour un seuil fixé u défini pour $u < x_F$, on définit les excès de la variable X au-dessus

du seuil u par :

$$Y_i := X_i - u, \text{ quand } X > u.$$

Définition 1.6.1 (Distribution des excès)

La fonction de répartition des excès de X au-dessus du seuil u est définie par :

$$F_u(y) := P(X - u \leq y \,|X > u). \tag{1.37}$$

De plus

$$\begin{aligned} F_u(y) &= \frac{P(X - u \leq y, X > u)}{P(X > u)} = \frac{P(u < X \leq y + u)}{1 - P(X \leq u)} = \frac{P(X \leq y + u) - P(X \leq u)}{1 - P(X \leq u)} \\ &= \frac{F(y + u) - F(u)}{1 - F(u)} = \frac{F(y + u) - F(u)}{\bar{F}(u)} = 1 - \frac{\bar{F}(u + y)}{\bar{F}(u)}, \ 0 < y < x_F - u, \end{aligned}$$

1.6.2 Distribution de Paréto généralisée

Lorsque la valeur du seuil est élevé (proche du point terminal), on peut approcher la loi des excès par une distribution de Pareto généralisée (GPD : Generelezed Pareto Distribution) de variance inconnue (dépendant de u).

Définition 1.6.2 (Distribution de Paréto généralisée)

La GPD est donnée par :

$$G_{\gamma,\sigma(u)}(x) := \begin{cases} 1 - \left(1 + \gamma \dfrac{x - u}{\sigma}\right)^{-1/\gamma} & si \quad \gamma \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right) & si \quad \gamma = 0. \end{cases} \tag{1.38}$$

avec

$$\begin{aligned} x \geq 0 \qquad\quad & si \quad \gamma \geq 0, \\ 0 \leq x \leq -\frac{\sigma}{\gamma} \quad & si \quad \gamma < 0. \end{aligned}$$

La fonction de densité de la loi de Paréto généralisée $g_{\gamma,\sigma(u)}$ *est définie par :*

$$g_{\gamma,\sigma(u)}(x) := \begin{cases} \frac{1}{\sigma}\left(1+\gamma\frac{x}{\sigma}\right)^{-1/\gamma-1} & si \quad \gamma \neq 0, \\ \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right) & si \quad \gamma = 0. \end{cases} \tag{1.39}$$

La GPD standard correspond au cas où $u = 0$ *et* $\sigma = 1$*, est définie par :*

$$G_\gamma(x) := \begin{cases} 1-(1+\gamma x)^{-1/\gamma} & si \quad \gamma \neq 0, \\ 1-\exp(-x) & si \quad \gamma = 0. \end{cases} \tag{1.40}$$

avec

$$x \geq 0 \quad si \quad \gamma \geq 0,$$
$$0 \leq x \leq -\tfrac{1}{\gamma} \quad si \quad \gamma < 0.$$

Lorsque le paramètre de localisation est nul $(u = 0)$ *et le paramètre d'échelle est arbitraire* $(\sigma > 0)$*, cette distribution joue un rôle important.*

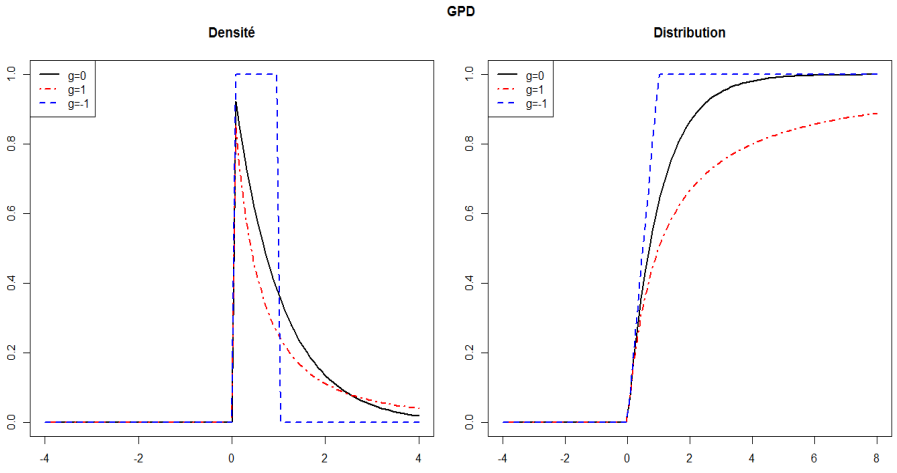

FIG. 1.2 – Densités et Distributions de loi Pareto Généralisées avec différentes valeurs de γ.

Théorème 1.6.1 (Balkema & de Haan(1974), Pickands(1975))

Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes, alors il existe une fonction $\sigma(u)$ positive et un réel γ tel que :

$$\lim_{u \to x_F} \sup_{0<y\leq x_F-u} \left| F_u(y) - G_{\gamma,\sigma(u)}(y) \right| = 0, \tag{1.41}$$

où $G_{\gamma,\sigma(u)}$ est la fonction de répartition de la loi de Paréto généralisée et F_u est la fonction de répartition des excès au-delà du seuil u.

1.7 Nations de base sur la censure

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Définition 1.7.1 (Variable de censure)

La variable de censure C est définie par la non-observation de l'événement étudié. Si au lieu d'observer S, on observe C, et que l'on sait que $S > C$ ($S < C$, $C_1 < S < C_2$ respectivement), on dite qu'il y a censure à droite (censure à gauche, censure par intervalle respectivement).

Pour un individu donné i, on va considérer

– Son temps de survie S_i.

– Son temps de censure C_i.

– La durée réellement observée O_i.

1.7.1 Types de censure

– **Censure à droite**

La variable d'intérêts 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.

– **Censure à gauche**

Il y a une 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.

– **Censure double ou mixte**

On dit qu'on a une censure double ou mixte si on a des données censurées à droite et des données censurées à gauche dans le même échantillon. Plusieurs modèles non-paramétriques ont été présentés pour l'étude de la double censure.

– **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.

Dans la littérature on retrouve les types suivants :

1. **Censure de type I (fixe)**

   Soit C une valeur fixée, au lieu d'observer les variables $S_1, S_2, \cdots, S_n$ qui nous intéressent, on observe S_i que lorsque $S_i \leq C$, sinon on sait uniquement que

$S_i > C$. On observe donc une variable O_i telle que :

$$O_i := \min(S_i, C),\ i = 1, \cdots, n. \tag{1.42}$$

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles.

2. **Censure de type II (attente)**

   L'expérimentateur fixe a priori le nombre d'événements à observer. La date définie d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité d'épidémiologie. Par exemple en épidémiologie on décide d'observer les durées de survie des n patients jusqu'à ce que k ($k = \overline{1, n}$) d'entre eux soient décédés et d'arrêter l'étude à ce moment-là. Soient $S_{i,n}$ et $O_{i,n}$ les statistiques d'ordre des variables S_i et O_i. La date de censure est donc $S_{i,n}$ et on observe

$$O_{i,n} := \begin{cases} S_{i,n} & si \quad i \leq k, \\ S_{k,n} & si \quad i \geq k. \end{cases} \tag{1.43}$$

3. **Censure de type III (aléatoire)**

   Il existe C une v.a's iid positive d'échantillon $(C_1, C_2, \cdots, C_n)$, on observe un couple de v.a's (O_i, δ_i) avec

$$O_i := \min(S_i, C_i) \text{ et } \delta_i := \mathbb{1}_{\{S_i \leq C_i\}},\ i = 1, \cdots, n. \tag{1.44}$$

où δ d'indicateur de censure, qui détermine si S a été censuré ou non :

$$\delta_i := \begin{cases} 1 & \text{d'où} \quad O_i = S_i \text{ (la durée d'intérêt est observée),} \\ 0 & \text{d'où} \quad O_i = C_i \text{ (elle est censurée).} \end{cases} \tag{1.45}$$

La censure aléatoire est la plus courante, et la plus considérée en analyse de survie.

Chapitre 2

Estimation des quantiles extrêmes

Dans ce chapitre on s'intéresse à l'estimation des quantiles extrêmes dont la probabilité d'observation est très faible (proche de zéro) quand la taille de l'échantillon tend vers l'infini. En théorie des valeurs extrêmes, il existe différentes approches pour l'estimation de ces quantités, nous présentons dans la suite trois approches sous données complètes et une approche basée sur l'estimateur de Hill adapté sous données censurées.

2.1 Estimation des quantiles extrêmes sous données complètes

Nous commençons par donner la définition du quantile d'ordre $(1-p)$ et du quantile extrême.

Définition 2.1.1 (Quantile d'ordre $1-p$)

Soit $X_1, X_2, \cdots, X_n$, n v.a's iid de fonction de répartition commune F. On appelle quantile ou fractile d'ordre $1-p$ de la fonction de répartition F, le nombre x_p défini

par :

$$x_p := \bar{F}^{\leftarrow}(p) = inf\{x \in \mathbb{R} : \bar{F}(x) \leq p\}, \ \textit{avec } p \in]0,1[\,.$$

Remarque 2.1.1

Si F est strictement croissante et continue, alors x_p est l'unique nombre réel tel que :

$$\bar{F}(x_p) = 1 - p.$$

Définition 2.1.2 (Quantile extrême)

Le quantile extrême d'ordre $1 - p_n$ de la fonction de répartition F est défini par :

$$x_{p_n} := \bar{F}^{\leftarrow}(p_n) \ \textit{ avec } p_n \to 0 \ \textit{quand } n \to \infty.$$

Dans tout ce qui suit, on suppose que $F \in \mathcal{D}(H_\gamma)$ pour certain $\gamma \in \mathbb{R}$. Afin d'estimer le quantile extrême, on introduit le résultat suivant.

Lemme 2.1.1

Si $x_{p_n} \to 0$ et $np_n \to c$ (non nécessairement fini) quand $n \to \infty$, alors

$$P(X_{n,n} < x_{p_n}) \to e^{-c}.$$

D'après le Lemme 2.1.1, on doit faire la distinction entre deux situations en fonction de c.

Si $c = \infty$ alors $P(X_{n,n} < x_{p_n}) = 0$. Dans un tel contexte, un estimateur classique de x_{p_n} est le quantile empirique $X_{(n-[pn],n)}$.

Si $c = 0$ alors $P\left(X_{n,n} < x_{p_n}\right) = 1$. La quantité x_{p_n} est en dehors de l'échantillon. Par conséquent, on ne peut pas estimer le quantile de manière empirique. Pour résoudre ce problème, nous présentons les méthodes suivantes.

2.1.1 Estimation basée sur l'approche semi-paramétrique

Les méthodes d'estimation des quantiles extrêmes sont basées sur des estimateurs des paramètres des lois GEV et GPD. Dans cette section, nous faisons en particulier un rappel sur les estimateurs de l'indice des valeurs extrêmes les plus célèbres dans la littérature.

Estimateur de Hill

L'estimateur de Hill [15] est un estimateur simple et largement utilisé. Cet estimateur n'est applicable que dans le cas où l'EVI, γ est connu pour être positif, ce qui répond à des distributions appartenant au domaine d'attraction de type **Fréchet** ($\gamma > 0$).

Définition 2.1.3 (Estimateur de Hill $(\gamma > 0)$**)**

Soit $X_1, X_2, \cdots, X_n$, n v.a's iid de fonction de répartition $F \in \mathcal{D}(H_{1/\gamma})$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Hill est défini par :

$$\hat{\gamma}_n^{(H)} = \hat{\gamma}_n^{(H)}(k) := \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}. \tag{2.1}$$

Théorème 2.1.1 (Propriétés asymptotiques de $\widehat{\gamma}_n^{(H)}$)

Supposons que $F \in \mathcal{D}\left(\Phi_{1/\gamma}\right), \gamma > 0,\ k \to \infty$ et $k/n \to 0$ quand $n \to \infty$.

i) Consistance faible

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

ii) Consistance forte : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors :

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

iii) Normalité asymptotique : Supposons que F satisfaisant 1.32 si $\sqrt{k}A(n/k) \to \lambda$ quand $n \to \infty$, alors :

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

Ce dernier résultat permet de calculer des intervalles de confiance pour γ. Par exemple, à un niveau de confiance de $(1-\alpha)\,\%$, on a pour $\lambda = 0$

$$\gamma \in \left[\hat{\gamma}_n^{(H)} - q_{1-\alpha/2}\frac{\hat{\gamma}_n^{(H)}}{\sqrt{k}}; \hat{\gamma}_{k(n)}^{H} + q_{1-\alpha/2}\frac{\hat{\gamma}_n^{(H)}}{\sqrt{k}}\right],$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi normale centrée réduite.

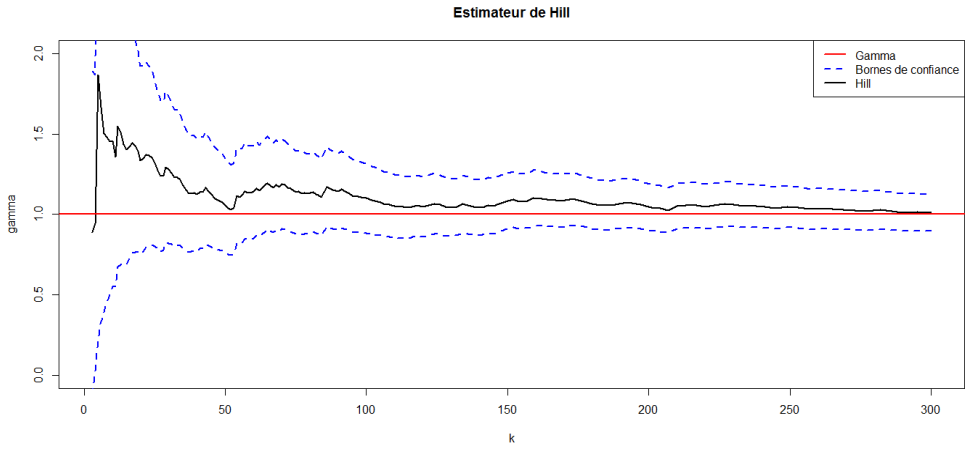

FIG. 2.1 – Estimateur de Hill pour l'EVI de la distribution de Paréto standard ($\gamma = 1$) basé sur 300 échantillons de 3000 observations.

Concernant l'étude du quantile extrêmes x_p, l'estimateur semi-paramétrique le plus fréquemment utilisé, associé à l'estimateur de Hill 2.1, a été proposé par Weissman (1978) [26].

Définition 2.1.4 (Estimateur de Weissman)

L'estimateur de Weissman est défini par :

$$\hat{x}_p^W := X_{n-k,n} \left(\frac{k}{np} \right)^{\hat{\gamma}_n^{(H)}} .$$

Estimateur de Pickands

L'estimateur de Pickands a été introduit en 1975 par Pickands [20], pour toute $\gamma \in \mathbb{R}$.

Définition 2.1.5 (Estimateur de Pickands)

Soit $X_1, X_2, \cdots, X_n$, n v.a's iid de fonction de répartition $F \in \mathcal{D}(H_\gamma)$, où $\gamma \in \mathbb{R}$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Pickands est défini par la statistique :

$$\hat{\gamma}^P = \hat{\gamma}^P(k) := \frac{1}{\log 2} \log \left(\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}} \right). \tag{2.2}$$

La consistance faible et la consistance forte ce cet estimateur a été obtenue par Pickands(1975) [20] et la normalité asymptotique a été démontré par Dekkers et de Haan (1989) [6].

Théorème 2.1.2 (Propriétés asymptotiques de $\hat{\gamma}^P$)

Soit $F \in \mathcal{D}(H_\gamma)$, où $\gamma \in \mathbb{R}$. Si $k \to \infty$ et $k/n \to 0$ quand $n \to \infty$.

i) Consistance faible

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

ii) Consistance forte : Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors

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

iii) Normalité asymptotique : Sous des conditions additionnelles sur la suite k_n et la fonction de répartition F

$$\sqrt{k}\left(\hat{\gamma}^P - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\gamma^2\left(2^{2\gamma+1}+1\right)}{4\left(2^\gamma - 1\right)^2 (\log 2)^2}\right) \text{ quand } n \to \infty.$$

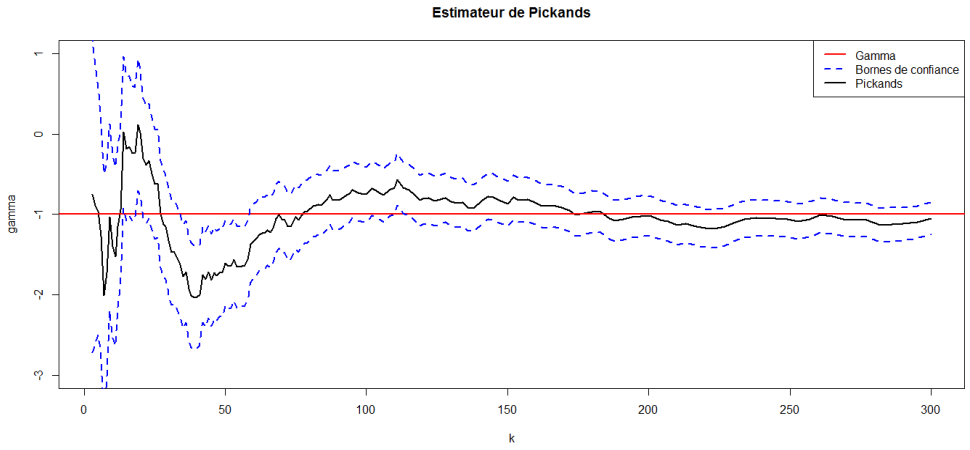

FIG. 2.2 – Estimateur de Pickands pour l'EVI de la distribution uniforme standard ($\gamma = -1$) basé sur 300 échantillons de 3000 observations.

L'estimateur du quantile extrême, associé à l'estimateur de Pickands 2.2, est donnée par :

$$\hat{x}_p^P := X_{n-k+1,n} + \frac{(np/k)^{-\hat{\gamma}^P} - 1}{1 - 2^{-\hat{\gamma}^P}} \left(X_{n-k+1,n} - X_{n-2k+1,n}\right).$$

Estimateur des moments

En 1989, Dekkers et al. ont proposés dans [7] une extension de l'estimateur de Hill en l'estimateur des moments, qui lui valable quelque soit le signe de l'indice γ.

Définition 2.1.6 (Estimateur des moments)

Pour $\gamma \in \mathbb{R}$, l'estimateur des moments est donnée par :

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

avec

$$M_n^{(r)} = M_n^{(r)}(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)^r, \ r = 1, 2, \tag{2.3}$$

où $M_n^{(1)}$ est l'estimateur de Hill $\hat{\gamma}_n^{(H)}$.

La consistance faible et forte de cet estimateur a été avérée par ses créateurs Dekkers et al (1989) [7].

Théorème 2.1.3 (Propriétés asymptotiques de $\hat{\gamma}^M$)

Soit $F \in \mathcal{D}(H_\gamma)$, $\gamma \in \mathbb{R}, k \to \infty$ et $k/n \to 0$ quand $n \to \infty$.

i) Consistance faible :

$$\hat{\gamma}^M \xrightarrow{p} \gamma \text{ quand } n \to \infty.$$

ii) Consistance forte : Si $k/(\log n)^\delta \to \infty$ quand $n \to \infty$., pour $\delta > 0$, alors

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

iii) Normalité asymptotique : (Voir Théorème 3 :1 et corollaire 3.2 de Dekkers et al.[7])

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

avec

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

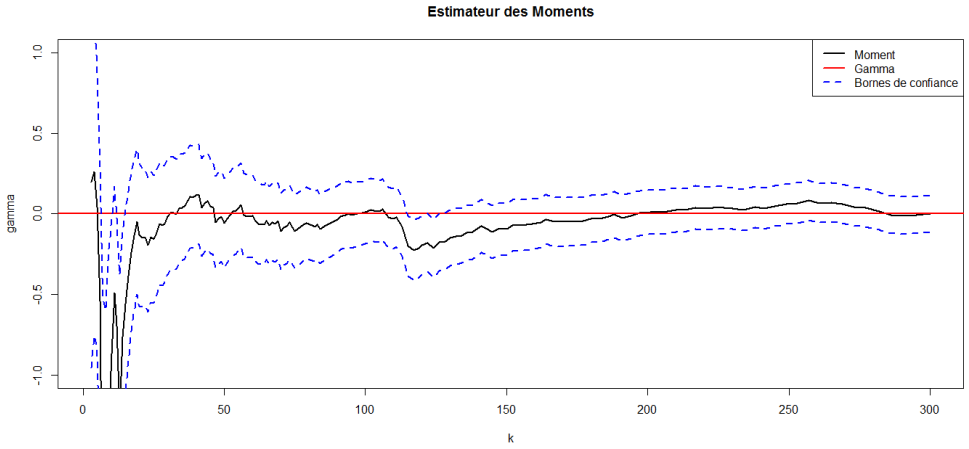

FIG. 2.3 – Estimateur de Moment pour l'EVI de la distribution de Gumbel ($\gamma = 0$) basé sur 300 échantillons de 3000 observations.

Un estimateur pour des quantiles extrêmes sur la base de l'estimateur des moments est donné par :

$$\hat{x}_p^M := X_{n-k,n} + \frac{\left(\dfrac{np}{k}\right)^{-\hat{\gamma}^M} - 1}{\hat{\gamma}^M} \frac{X_{n-k,n} M_n^{(1)}}{\rho\left(\hat{\gamma}^M\right)}, \tag{2.4}$$

où $M_n^{(1)}$ est définie par (2.3) et

$$\rho\left(\hat{\gamma}^M\right) := \begin{cases} 1, & \gamma \geq 0, \\ \dfrac{1}{1-\gamma} & \gamma < 0. \end{cases} \tag{2.5}$$

2.1.2 Estimation basée sur la GEV

Nous supposons que l'échantillon de maximum suit exactement une loi GEV. Pour obtenir les estimateurs des quantiles extrêmes $\widehat{x}_p$, il suffit d'inverser la fonction $H_{\mu,\sigma,\gamma}$ donnée par 1.27. Ils se présentent comme suit :

$$\hat{x}_p = H^{-1}_{\widehat{\mu},\widehat{\sigma},\hat{\gamma}}(1-p) := \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\gamma}}\left(1 - (-\log(1-p))^{-\hat{\gamma}}\right) & si \quad \gamma \neq 0, \\ \hat{\mu} - \hat{\sigma}\log(-\log(1-p)) & si \quad \gamma = 0, \end{cases} \tag{2.6}$$

Supposons que $F \in \mathcal{D}(H_\gamma)$. Dans le cas où le $(1-p)$-quantile est à l'intérieur des données (c-à-d : $p \geq 1/n$), il peut être estimé par :

$$\hat{x}_p := \widehat{b}_n + \frac{\hat{a}_n}{\hat{\gamma}}\left(\left(\frac{1}{np}\right)^{\hat{\gamma}} - 1\right),$$

où $\hat{\gamma}$, $\widehat{a}_n$ et $\widehat{b}_n$ sont des estimations appropriées (basées sur les k plus grande statistiques d'ordre) de l'indice de queue de distribution, et des constantes de normalisation a_n et b_n respectivement.

Dans le cas où le $(1-p)$-quantile est à l'extérieur des données (c-à-d $p < 1/n$), il peut être estimé par :

$$\hat{x}_p := \hat{a}_{n/k}\frac{(np/k)^{-\hat{\gamma}} - 1}{\hat{\gamma}} + \hat{b}_{n/k}.$$

2.1.3 Estimation basée sur la GPD

Une estimation des quantiles extrêmes au-dessus du seuil u ($x_p > u$) est obtenu en inversant l'expression de l'estimateur GPD donnée dans 1.38

$$\hat{x}_p := u + \frac{\hat{\sigma}_u}{\hat{\gamma}_u}\left(\left(\frac{N_u}{np}\right)^{\hat{\gamma}_u} - 1\right), \; p < \frac{N_u}{n}$$

avec $\hat{\sigma}_u$ et $\hat{\gamma}_u$ les estimateurs des paramètres de la loi GPD et N_u le nombre d'excès.

Dans le cas ($\gamma < 0$) le point terminal de la distribution est estimé par

$$\hat{x}_F := u + \frac{\hat{\sigma}_u}{\hat{\gamma}_u}.$$

Le seuil u est souvent choisi égal à une des statistiques d'ordre $X_{1,n} \leq \cdots \leq X_{n-k,n} \leq \cdots \leq X_{n,n}$. Si l'on choisit comme seuil $u = X_{n-k,n}$ à la $(k+1)^{éme}$ plus grande observation alors $N_u = k$ et l'estimateur des quantiles aux ordres élevés se réécrit par la forme suivante

$$\hat{x}_p^{(POT)} := X_{n-k,n} + \frac{\hat{\sigma}_u^{(POT)}}{\hat{\gamma}_u^{(POT)}} \left(\left(\frac{k}{np} \right)^{\hat{\gamma}_u^{(POT)}} - 1 \right), \; p < \frac{k}{n},$$

où $\hat{\gamma}^{(POT)}$, $\hat{\sigma}^{(POT)}$ sont les estimateurs résultants de γ et σ respectivement.

Le point final aux ordres élevés est estimé par

$$\hat{x}_F^{(POT)} := X_{n-k,n} + \frac{\hat{\sigma}_u^{(POT)}}{\hat{\gamma}_u^{(POT)}}.$$

2.2 Estimation des quantiles extrêmes sous données censurées

Dans un premier temps, nous rappelons les notations nécessaires, relatives au cas des données censurées, et que nous avons, en grande partie, introduites dans le premier chapitre. Nous supposons que l'échantillon de base (non observé directement) est constitué de copies indépendantes (S_i, C_i, O_i),$i = 1, 2, ...$, du vecteur aléatoire (S, C, O). Ici, S désigne la variable d'intérêt, C la variable de censure, et O un vecteur de variables concomitantes, dont le lien avec S est à étudier.

Avant de présenter l'estimateur du quantile extrême en présence de censure, il est nécessaire d'introduire l'estimateur de la fonction de survie et l'estimateur de l'EVI

2.2.1 Estimateur de Kaplan-Meier

Soit $(S_1, S_2, \cdots, S_n)$ et $(C_1, C_2, \cdots, C_n)$ deux échantillons de v.a's iid de fonction de répartition commune absolument continues F et G respectivement (avec x_F et x_G sont les point terminaux respectivement). On suppose aussi que ces variables sont indépendantes. Soit $\{(Oi, \delta i); 1 \leq i \leq n\}$ l'échantillon réellement observé définie par (1.44), dans la suite on suppose que la variable O a comme fonction de répartition H (x_H est le point terminal du support) définie par :

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

et $O_{1,n} \leq O_{2,n} \leq \cdots \leq O_{n,n}$ les statistiques d'ordre lui associées. Avec $\delta_{[1,n]}, \ldots, \delta_{[n,n]}$ sont les indicateurs de censure retenues avec ces dernières $\left(\delta_{[i,n]} = \delta_j \text{ si } O_{i,n} = O_j\right)$.

En présence de censure, la fonction de survie $S(t)$ définie par (1.2) peut être estimé grâce à plusieurs méthodes non paramétriques dont la plus intéressante est celle de Kaplan-Meier (1958) [18]. Cet estimateur est aussi appelé un estimateur à limite produit (P-L estimateur), car il s'obtient comme un produit. L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t. Si $t_0 = 0 < t_1 < t_2 < ... < t_k$ où $1 \leq k \leq n$

$$\begin{aligned} P(S > t_k) &= P(S > t_k, S > t_{k-1}) \\ &= P(S > t_k \mid S > t_{k-1}) \times P(S > t_{k-1}) \\ &\vdots \\ &= P(S > t_k \mid S > t_{k-1}) \times \ldots \times P(S > t_2 \mid S > t_1) \times P(S > t_1). \end{aligned}$$

On considère les temps d'événements (décès et censure) distincts $O_{i,n}$ $(1 \leq i \leq n)$ ordonnés par ordre croissant, on obtient :

$$P(S > O_{i,n}) = \prod_{k=1}^{i} P(S > O_{k,n} \mid S > O_{k-1,n}), \ i = 1, ..., n,$$

avec $O_{0,n} = 0$.

Considérons les notations suivantes :

- n_i le nombre d'individus à risque de subir l'événement juste avant le temps $O_{i,n}$.

- d_i le nombre de décès en $O_{i,n}$.

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

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

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

-$d_i = 0$ en cas de censure en $O_{i,n}$, i.e : quand $\delta_{[i,n]} = 0$,

-$d_i = 1$ en cas de décès en $O_{i,n}$, i.e : quand $\delta_{[i,n]} = 1$.

Pour $1 \leq i \leq n$, $\delta_{[i,n]}$ le concomitant de la $i^{éme}$ statistique d'ordre $O_{i,n}$, c'est-à-dire, $\delta_{[i,n]} = \delta_j$ donnée par (1.45) si $O_{i;n} = O_j$;$1 \leq j \leq n$. On obtient alors l'estimateur de Kaplan-Meier

$$\widehat{\overline{F}}^{KM}(t) := \prod_{\substack{i=1,\ldots,n \\ O_{i,n} \leq t}} \left(1 - \frac{\delta_{[i,n]}}{n_i}\right) = \prod_{i:O_{i,n} \leq t} \left(1 - \frac{\delta_{[i,n]}}{n-(i-1)}\right) = \prod_{O_{i,n} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{[i,n]}}.$$

Cet estimateur peut aussi être s'écrire de la manière suivante :

$$\widehat{\overline{F}}^{KM}(t) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i,n]}}{n-i+1}\right)^{\mathbf{I}\{O_{i,n} \leq t\}}, \text{ pour } t < O_{n,n}.$$

Remarque 2.2.1

1. Les points de discontinuité de cette fonction correspondent aux observations non-censurées.
2. L'hauteur des sauts de $\widehat{\overline{F}}^{KM}(t)$ est aléatoire.
3. En l'absence de censures, on retrouve la fonction de survie empirique (1.4).

Proposition 2.2.1 (Propriétés asymptotiques de $\widehat{\overline{F}}^{K}_{n}$)

i) Absence de biais : Pour tout t, on absence de biais, on a $\widehat{\overline{F}}^{KM}(t) \overset{p.s}{\to} \overline{F}_n(t)$, i.e :

$$E\left[\widehat{\overline{F}}^{KM}(t)\right] = \overline{F}(t) \text{ quand } n \to \infty,$$

ii) Consistance uniforme : Soit $x_H = H^{-1}(1) := \inf\{t : H(t) = 1\}$, Alors

$$\sup_{0 \leq t < x_H} |\widehat{\overline{F}}^{KM}(t) - \overline{F}(t)| \overset{p.s}{\to} 0 \text{ quand } n \to \infty.$$

iii) Normalité asymptotique : Pour tout $t \geq 0$, on a

$$\sqrt{n}\left(\widehat{\overline{F}}^{KM}(t) - \overline{F}(t)\right) \xrightarrow{d} N\left(0, V^2(t)\right). \tag{2.7}$$

avec

$$V^2(t) := -\overline{F}^2(t) \int_0^t \frac{\overline{F}(ds)}{\overline{F^2}(s)\, G(s)}.$$

iv) Estimation de la variance de $\widehat{\overline{F}}_n^{KM}$: L'estimateur de Greenwood de la variance de $\widehat{\overline{F}}_n^{KM}$ est

$$\widehat{Var}(\widehat{\overline{F}}_n^{KM}) := \left(\widehat{\overline{F}}_n^{KM}\right)^2 \sum_{i:O_{i,n}\leq t}^{n} \frac{d_i}{n_i\,(n_i - d_i)}.$$

Exemple 2.2.1 (Données simulées)

On génère un échantillon de v.a S, qui représente la durée de survie, issu d'une loi uniforme standard. Cet échantillon est censuré à droite par un échantillon de v.a C uniformément distribué sur $[0, 0.6]$, qu'il s'agit de la durée de censure. Les deux échantillons ont la même taille $n = 20$. Les résultats numériques de l'estimation de la fonction de survie sont résumés dans le tableau 2.1, où :

- t_i : Les 5 durées réellement observées, parmi les 20 individus qui sont analysés, ce qui est équivalent à 75% de censure.
- n_i : Le nombre d'individu à risque sur l'intervalle de temps écoulé.
- d_i : Indique le nombre d'évènements observé d_i.
- $\widehat{\overline{F}}^{KM}$: L'estimateur de la fonction de survie de Kaplan-Meier.
- $\widehat{\sigma}_\varepsilon$: L'erreur standard associée à l'estimateur de Kaplan-Meier pour chaque t_i.
- CI (95%) : L'intervalle de confiance à 95% associée à l'estimateur de Kaplan-Meier pour chaque t_i.

t_i	n_i	d_i	$\widehat{\overline{F}}_n^K$	$\widehat{\sigma}_\varepsilon$	CI (95%)
0.0489	19	1	0.947	0.0512	$0.852 - 1.000$
0.0513	18	1	0.895	0.0704	$0.767 - 1.000$
0.2698	11	1	0.813	0.1006	$0.638 - 1.000$
0.3398	8	1	0.712	0.1296	$0.498 - 1.000$
0.4657	3	1	0.474	0.2121	$0.198 - 0.975$

TAB. 2.1 – Résultats de l'estimateur de Kaplan-Meier relatifs aux données simulées de 20 observations uniformes standards censurées par une variable uniforme sur [0,0.6]

Le figure 2.4 donne l'estimateur de Kaplan-Meier (ligne continue) et les bornes inférieures et supérieures de l'intervalle de confiance à 95% (lignes en tirets) de la fonction de survie sous données simulées .Nous observons que la courbe de cet estimateur est en escalier décroisant. On peut aussi observer 15 valeurs à travers un croix rouge (+) qui représentent les durées censurées.

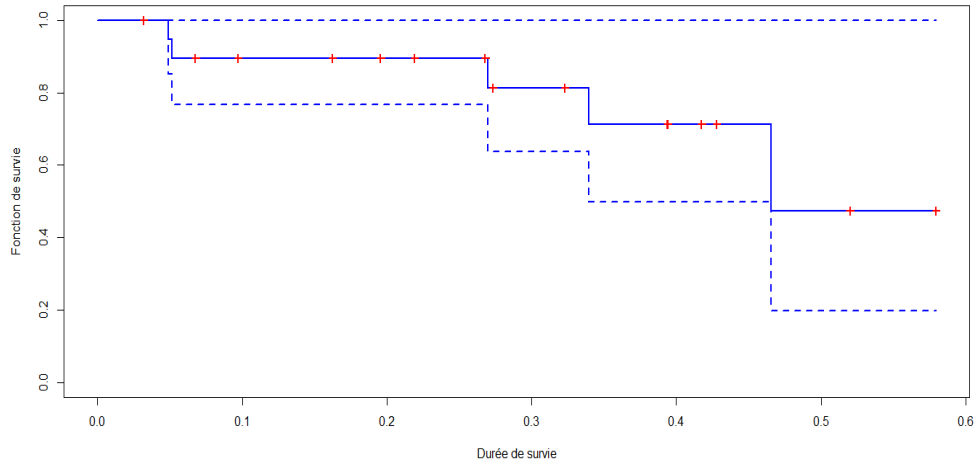

FIG. 2.4 – Estimateur de Kaplan-Meier (ligne continue) et bornes de confiance à 95% (lignes en tirets) de la fonction de survie sous données simulées

Exemple 2.2.2 (Données de ventilateurs)

Le tableau 2.2 contient la durée de vie de 70 ventilateurs, en milliers d'heures de fonctionnement (+ indique une donnée censurée).

4.5	4.6^+	11.5	11.5	15.6^+	16.0	16.6^+	18.5^+	18.5^+	18.5^+
18.5^+	18.5^+	20.3^+	20.3^+	20.3^+	20.7	20.7	20.8	22.0^+	30.0^+
30.0^+	30.0^+	30.0^+	31.0	32.0^+	34.5	37.5^+	37.5^+	41.5^+	41.5^+
41.5^+	41.5^+	43.0^+	43.0^+	43.0^+	43.0^+	46.0	48.5^+	48.5^+	48.5^+
48.5^+	50.0^+	50.0^+	50.0^+	61.0^+	61.0	61.0^+	61.0^+	63.0^+	64.5^+
64.5^+	67.0^+	74.5^+	78.0^+	78.0^+	81.0^+	81.0^+	82.0^+	85.0^+	85.0^+
85.0^+	87.5^+	87.5	87.5^+	94.0^+	99.0^+	101.0^+	101.0^+	101.0^+	115.0^+

TAB. 2.2 – Les durées de fonctionnement de 70 ventilateurs

L'estimateur de Kaplan-Meier de la fonction de survie des données de ventilateurs est illustré dans la figure 2.5.

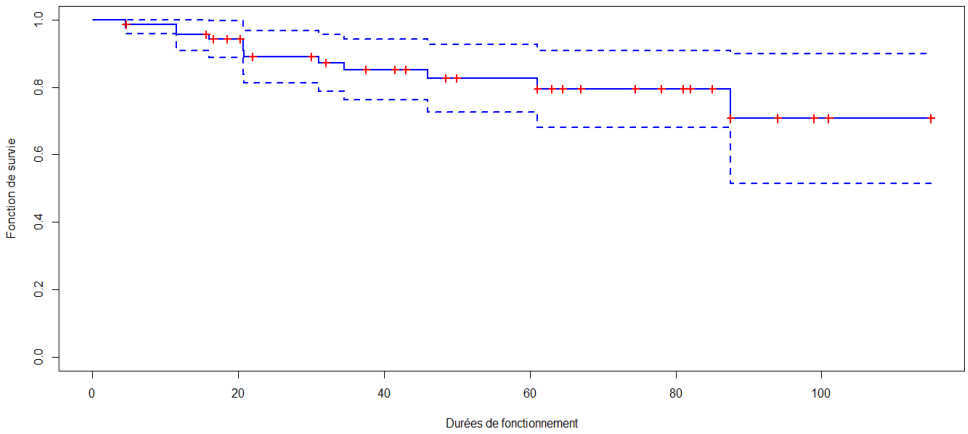

FIG. 2.5 – L'estimateur de Kaplan-Meier des données de ventilateurs.

Par exemple $\widehat{\overline{F}}_n^K(60 \times 10^3) = 0.8$ ce qui donne 80% des ventilateurs vont fonctionner au-delà de 60×10^3.

Les résultats numériques de l'estimateur de Kaplan-Meier de la fonction de survie relatives aux données de ventilateurs sont résumées dans le tableau 2.3.

t_i	n_i	d_i	$\widehat{\overline{F}}_n^K$	$\widehat{\sigma}_\varepsilon$	CI (95%)
4.5	70	1	0.986	0.0142	$0.958 - 1.000$
11.5	68	2	0.957	0.0244	$0.909 - 1.000$
16.0	65	1	0.942	0.0282	$0.887 - 0.997$
20.7	55	2	0.908	0.0361	$0.837 - 0.978$
20.8	53	1	0.891	0.0392	$0.814 - 0.968$
31.0	47	1	0.872	0.0427	$0.788 - 0.955$
34.5	45	1	0.852	0.0460	$0.762 - 0.942$
46.0	34	1	0.827	0.0510	$0.727 - 0.927$
61.0	26	1	0.795	0.0581	$0.682 - 0.909$
87.5	9	1	0.707	0.0980	$0.515 - 0.899$

TAB. 2.3 – Résultats de l'estimateur de Kaplan-Meier relatifs aux données de ventilateurs

2.2.2 Estimateur de Hill adapté

Si F et G sont absolument continue et que $F \in D(H_{\gamma_1})$, $G \in D(H_{\gamma_2})$ et $H \in D(H_\gamma)$ pour $\gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ pour certains γ_1, $\gamma_2 \in \mathbb{R}$. Pour tout $\gamma \in \mathbb{R}$, Einmahl et al. (2008) [10] ont proposés les trois cas suivants :

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

Beirlant et al (2007) [1] *ont proposés différents estimateurs de l'indice des valeurs ex-*

trêmes γ_1 associé à F en présence de censure, ces derniers sont tous construits de façon similaire, à partir d'un estimateur usuel (non adapté à la censure). Ces estimateurs basés sur les observations O_i, estiment par conséquent l'indice γ de H. Il s'agit alors de les modifier de façon à estimer γ_1 et non γ. Une façon de procéder consiste à diviser ces estimateurs usuels par la proportion de données non censurées au-delà d'un seuil u, c'est-à-dire à utiliser

$$\hat{\gamma}_1^{(\cdot,c)} = \hat{\gamma}_1^{(\cdot,c)}(k) := \frac{\hat{\gamma}^{(.)}}{\hat{p}}, \tag{2.8}$$

où

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

avec k est le nombre des excès au-delà de u et $\hat{p}$ estime $p = \dfrac{\gamma_2}{\gamma_1+\gamma_2}$($\hat{p} \to p$ quand $n \to \infty$), par conséquent $\hat{\gamma}^{(.)}$ l'estimateur de γ divisé par $\dfrac{\gamma_2}{\gamma_1+\gamma_2}$qui est égal à γ_1.

Pour adapter l'estimateur de Hill dans le cas de censure nous allons diviser cet estimateur $\hat{\gamma}_n^{(H)}$ par la proportion de données non censurées des k plus grandes valeurs de O, alors l'estimateur de Hill adapté de l'indice de queue $\hat{\gamma}_1^c$ est défini par :

$$\hat{\gamma}_1^c := \frac{\hat{\gamma}_n^{(H)}}{\hat{p}}, \tag{2.9}$$

où

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

alors

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

Einmahl et al (2008) [10] ont établis de façon unifiée, la normalité asymptotique de tout estimateur de l'indice des valeurs extrêmes écrit sous la forme (2.8) dans le cas où le seuil choisi u est aléatoire et égal à $O_{n-k,n,}$ la $(n-k)^{ème}$ statistique d'ordre de l'échantillon $(O_1, O_2, \cdots, O_n)$.

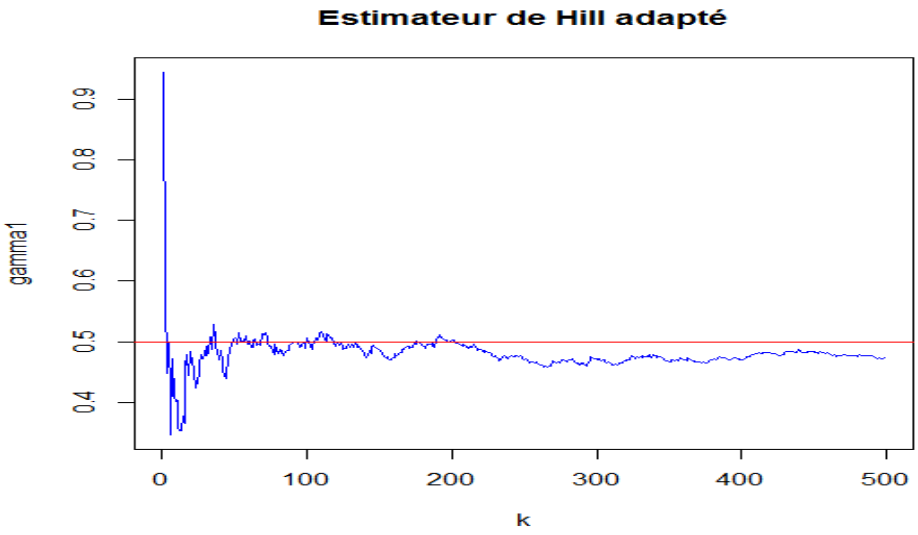

FIG. 2.6 – Estimateur de Hill adapté pour un échantillon de 500 observations de la loi de Paréto ($\gamma_1 = 0.5$) censuré par un échantillon de taille 500 de la loi de Paréto ($\gamma_2 = 1$).

2.2.3 Estimateur des quantiles extrêmes

Le principal estimateur des quantile extrêmes x_s d'ordre $(1-s)$ sous censure aléatoire disponible dans la littérature a été proposé par Beirlant et al en 2007 [1] et par Einmahl et al en 2008.[10]. Il est donné par la définition suivante.

Définition 2.2.1 (Estimation du quantile extrême)

L'estimation du quantile extrême sous censure aléatoire est défini par :

$$\widehat{x}_s^{(\cdot,c)} := O_{n-k,n} + \hat{a}^{(\cdot,c)} \frac{\left(\left(1 - \widehat{\overline{F}}_n^{KM} \left(O_{n-k,n} \right) \right) / s \right)^{\hat{\gamma}_1^{(\cdot,c)}} - 1}{\hat{\gamma}_1^{(\cdot,c)}},$$

où $\hat{a}^{(\cdot,c)} = O_{n-k,n} M_n^{(1)} \left(\frac{1}{2} \left(1 - \frac{\left(M_n^{(1)} \right)^2}{M_n^{(2)}} \right)^{-1} \right) / \hat{p}$. *avec* $M_n^{(r)}, r = 1, 2$ *est défini dans (2.3).*

2.3 Simulations

Les simulations ont été réalisées sur le logiciel R, version 3.6.2, on ce qui concerne l'estimateur de la fonction de survie par Kaplan-Meier, on a utilisé le package "survival", et pour l'estimateur de Hill adapté et l'estimateur des quantiles extrêmes on a utilisé le package "ReIns".

Génération d'un échantillon censuré

La loi de simulation utilisée dans cette section est la loi de *Paréto* de paramètre $\alpha > 0$, et de fonction de survie :

$$\bar{F}(x) = x^{-\alpha}, x > 1. \text{ avec } \gamma = \frac{1}{\alpha}.$$

On génère un échantillon de v.a S_i d'une loi de Paréto de parmètre $\alpha_1 = 2$, censuré par un deuxième échantillon de v.a C_i d'une loi de Paréto standard $\alpha_2 = 1$. Ces deux échantillons sont de taille $n = 500$. La proportion de données non censurées est 67%.

On simule des données censurées aléatoirement à droite. Les variables que nous observons sont les O_i d'une loi de Paréto et les indicateurs de censure δ_i, telles que :

$$O_i = \min(S_i, C_i) \text{ et } \delta_i = \mathbb{1}_{\{S_i \leq C_i\}}, \ i = 1, \cdots, n. \tag{2.10}$$

Estimateur de Hill adapté sous des données simulées

L'estimateur de Hill adapté $\hat{\gamma}_1^c$, définie par 2.9 qui correspond aux données censurés, est donnée sur la figure 2.7 en fonction du nombre de statistique d'ordre k.

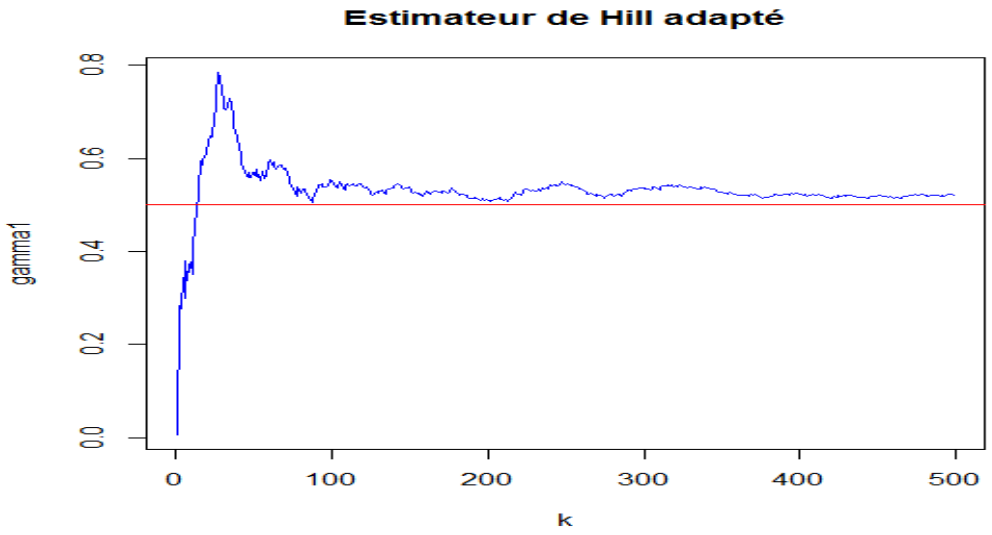

FIG. 2.7 – Estimateur de Hill adapté issue d'un échantillon d'une loi de Paréto $(500, 2)$ censuré par un échantillon d'une loi de Paréto $(500, 1)$. La ligne horizontale représente la vraie valeur de γ_1.

On observe que pour k petit, il y a de grandes oscillations avec un intervalle de confiance large, et pour k grand l'intervalle de confiance devient plus étroit. De plus, cette figure montre que cet estimateur est très stable à partir d'une certaine valeur qui représente le nombre optimal de statistiques d'ordre extrêmes ($k \geq 100$).

On peut également remarquer que l'estimateur de Hill adapté $\hat{\gamma}_1^c$ présente une allure très proche de la ligne horizontale, qui représente la valeur réelle de l'indice $\gamma_1 = 0.5$ l'inverse du paramètre $\alpha_1 = 2$ de la loi de Paréto utilisée.

Estimateur de Kaplan-Meier sous des données simulées

La courbe en escaliers décroissants de l'estimateur de la fonction de survie de Kaplan-Meier, associé aux données simulées, en fonction de durée de survie est illustré dans la figure (2.8). L'intervalle de confiance au niveau 95% de la survie est également représenté sur cette figure (lignes en tirets). Les petits croix rouge (+) indiquent des censures ont été observées.

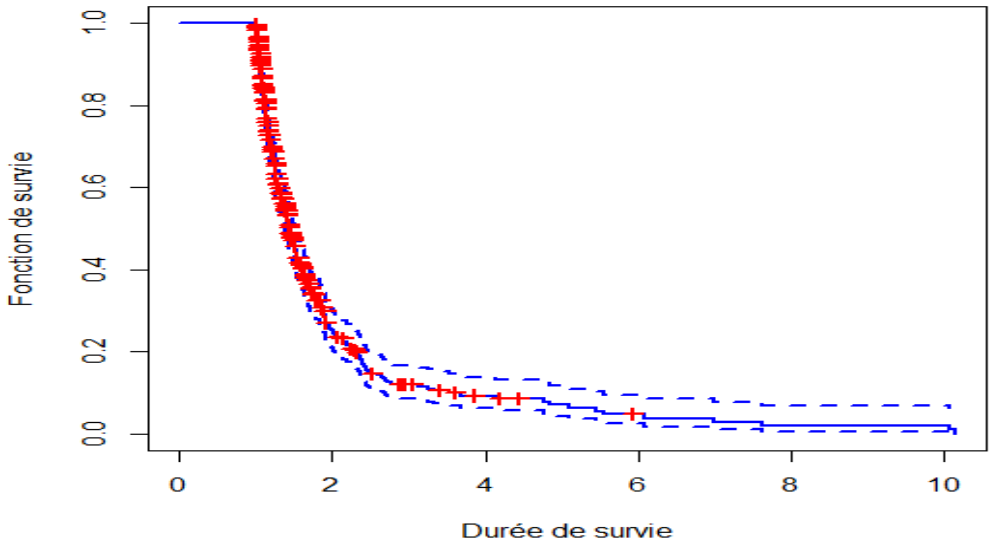

FIG. 2.8 – Estimateur de Kaplan-Meier issue d'un échantillon d'une loi de Paréto $(500, 2)$ censurée par un échantillon d'une loi de Paréto $(500, 1)$.

Les résultats numériques montrent que 349 valeurs réellement observées, parmi les 500 valeurs qui sont analysés, ce qui est équivalent à 30.2% de censure.

Estimation des quantiles extrêmes sous des données simulées

La figure (2.9) représente le comportement de la courbe de l'estimateur des quantiles extrêmes d'ordre $1-p$ où $p = 10^{-3}$, en fonction du nombre de statistique d'ordre extrêmes k.

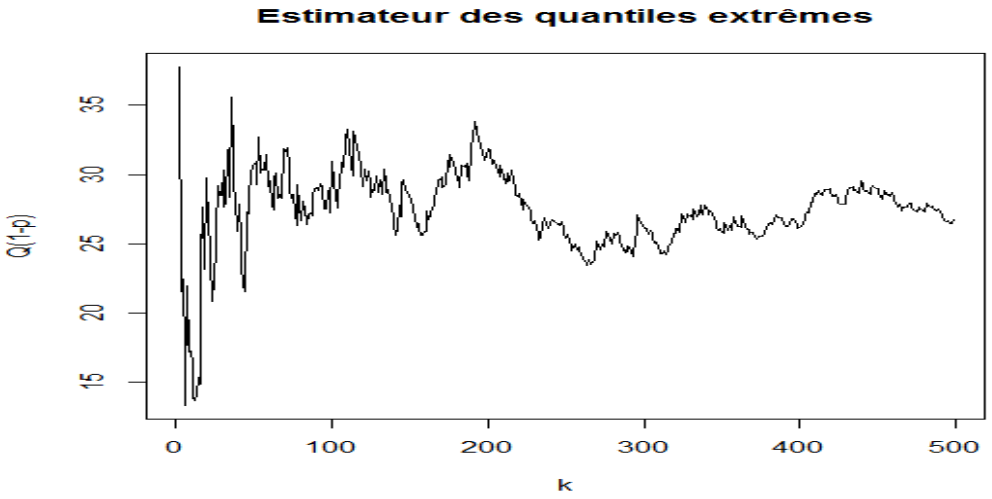

FIG. 2.9 – Comportement graphique de l'estimateur des quantiles extrêmes

Conclusion

Dans ce mémoire nous avons présentés une introduction à la théorie des valeurs extrêmes ainsi qu'à la théorie de censure dont l'objectif principal était l'estimation des quantiles extrêmes. Nous avons constatés une variété de méthodes d'estimation de ces quantités que ce soit pour les données complètes ou pour les données censurées.

Les estimateurs des quantiles extrêmes présentés dans ce travail dépendent de l'indice de queue, par conséquent une telle méthode reste influencée par l'estimateur de cet indice.

Les points abordés dans ce mémoire nous ouvrent la voie à d'autre piste de recherche intéressantes qui méritent d'être considérées.

Le premier point qui mérite d'être considéré est d'étudier le comportement asymptotiques de ces estimateurs tel que : la convergence faible et forte et la normalité asymptotique.

Le deuxième point à envisagé est de chercher d'autre méthodes dans le but d'améliorer l'estimation des quantiles extrêmes, avec une illustration à travers une application sur une série de données réelles complètes ou censurées.

Il serait aussi intéressant de présenter et de proposer un estimateur des quantiles extrêmes sous données tronqués.

Bibliographie

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

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

[3] Brahimi, B., Meraghni, D., and Necir, A. (2015). Gaussian approximation to the extreme value index estimator of a heavy-tailed distribution under random censoring. Math. Methods Statist., 24(4), 266 -279.

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

[5] Davis, R., and Resnick, S. (1984). Tail estimates motivated by extreme value theory. Ann. Statist., 1467-1487.

[6] Dekkers, A. L., and De Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. Ann. Statist., 1795-1832.

[7] Dekkers, A. L., Einmahl, J. H., and De Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. Ann. Statist., 1833-1855.

[8] de Hann, L. (1976). Sample extremes : an elementry intruduction. Stat.Neerl., 30(4),161-172.

[9] Embrechts, P., Klüppelberg, C., Mikosch,T. (1997). Modelling Extremal Events, in : Applications in Mathematics. Springer-Verlag, New York. vol 33.

[10] Einmahl, J.H.J., Fils-Villetard, A., Guillou. (2008) Statistics of extremes under random censoring ; Bernoulli.

[11] Fisher, R.A. and Tippett, L.H.C. (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society **24**, 180-190.

[12] Gnedenko, B.V. (1943). Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. Annales de Mathématiques **44**, 423-453.

[13] Gill, R. D. (1994). Glivenko-Cantelli for Kaplan-Meier. Math. Methods Statist., 3(1), 76.

[14] de Haan, L., Ferreria, A. (2006). Extreme Values Theory : An introduction. Springer.

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

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

[17] Jankinson, A. F.(1955). The Frequency Distribution of the Annual Maximum (or Minimum) of the Meteorological Elements Quarterly Jornal ofthe Royel Meteorological society 81, 185-171.

[18] Kaplan, E.L, Meier, P. (1958). Nonparamertic estimation from incomplete observations. Jornal of American Statistical Association and, 53 :457- 481.

[19] Meraghni, D. (2008). Modelling Distribution Tails. Thèse de Doctorat, Université de Biskra, Algérie.

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

[21] Reiss, R.D. and Thomas, M. (1997). Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. Birkhauser, Basel.

[22] Soltane,L (2016). Analyse des Valeurs Extrêmes en présence de censure.Thése de doctorat de université Mohamed khider, Biskra, Algeria.

[23] Stute, W. and Wang, J-L. (1993). The strong law under censorship. Ann. Statist, 21. 1591-1607.

[24] Stute, W. (1994). The bias of Kaplan-Meier integrals. Scan. J. Statist, 21. 475-484.

[25] Stute, W. (1995). The central limit theorem under random censorship. The Annals of Statistics, 23. 422-439.

[26] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. Journal of the American Statistical Association, 73(364) :812

Annexe : Abréviations et Notations

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

$v.a$	:	variable aléatoire
$\xrightarrow{p}$	:	convergence en probabilité.
$\xrightarrow{p.s}$	:	Convergence presque sur.
$\hat{\gamma}_n^{(H)}$	:	Estimateur de Hill.
$\hat{\gamma}^M$	:	Estimateur de moments
$\hat{\gamma}^P$	:	Estimateur de Pickands.
$\widehat{\overline{F}}_n^{KM}$	:	Estimateur de Kaplan-Meier.
iid	:	Indépendanteset identiquements distribuées.
$\hat{\gamma}_1^c$	:	Estimateur de Hill adapté
GEV	:	Distribution des valeurs extrêmes généralisées.
GPD	:	Distribution de Paréto généralisée.
$\xrightarrow{d}$	:	convergence en loi.
$\mathbb{1}_A$	:	Fonction indicatrice de l'ensemble A.
F_n	:	Fonctionde répartition empirique.
(Ω, A, P)	:	Espace probabilité.
EVI	:	Indice des valeurs extrêmes.
$D(.)$	:	Domaine d'attraction.

TCL	:	Théorème Centrale Limite.
S	:	Fonction de survie $\left(\overline{F}\right)$
F^{-1}	:	Inverse généralisé de F.
Q	:	Fonction de quantile.
Q_n	:	Fonction de quantile empirique.
H_γ	:	Famille de la loi de valeurs extrêmes généralisées.
RV_ρ	:	Variation régulière au ∞ avec l'indice ρ.
$E[X]$	:	Espérence mathématique de X.
x_F	:	Point terminal.
μ	:	Espérence ou moyenne d$'$une v.a.
$\mathbb{R}$	:	Ensemble des valeurs réelles.
$\mathcal{N}(0,1)$	:	Loi normale standard.
F	:	Fonctionde répartition.
f	:	Densité de probabilité d$'$une v.a.
fdr	:	Fonction de répartition.
σ^2	:	Variance d'une variable aléatoire.
$X_{n,n}$	:	Maximum de $X_1, X_2, ..., X_n$.
$X_{1,n}$	:	Minimum de $X_1, X_2, ..., X_n$.
$X_{k,n}$	:	k$^{\text{éme}}$ statistique d'ordre.
$(X_1, X_2, ..., X_n)$	:	échantillon de taille n de$X$.
$\overline{X}$	:	Moyenne arithmétique.
$:=$	:	égalité par définition.

Résume

Le thème du présent travail fait partie de deux branches très importantes en statistique : la théorie des valeurs extrêmes et la théorie de censure, dont notre objectif principal est l'estimation des quantiles extrêmes.

Nous donnons dans un premier temps, quelques rappels et définitions sur la théorie des valeurs extrêmes et de la censure. Dans un second temps, nous effectuons une synthèse sur les différentes méthodes d'estimation des quantiles extrêmes sous données complètes et sous données censurées.

Nous terminons par des simulations pour illustrer le comportement de ces estimateurs en présence de censure, avec un intérêt particulier au cas de censure à droite.

Mots-clés : Statistiques d'ordre, Censure, Quantiles extrêmes, Estimateur de Hill adapté, Estimateur de Kaplan-Meier.

Abstract

The topic of this work belongs to two very important branches in statistics: the extreme value theory and the theory of censorship, where our principal objective is the extreme quantiles estimation.

We first give a few recalls and definitions on the extreme values theory and censorship. In the second time, we make a synthesis on the different methods of estimation of the extreme quantiles under complete data and censured data.

We finish by simulations to illustrate the behavior of these estimators in the presence of censorship, with particular interest in the case of right censoring.

Keywords: Order statistics, Censorship, Extreme quantiles, Adapted Hill estimator, Kaplan-Meier

ملخص

موضوع هذا العمل هو جزء من فرعين مهمين للغاية في الإحصاء: نظرية القيم المتطرفة ونظرية الرقابة، بحيث هدفنا الرئيسي هو تقدير الكميات المتطرفة.

نقدم أولاً بعض التذكيرات والتعاريف حول نظرية القيم المتطرفة والرقابة. في الخطوة الثانية، نقوم بتجميع الطرق المختلفة لتقدير الكميات المتطرفة في ظل البيانات الكاملة والبيانات الخاضعة للرقابة.

ننتهي بالمحاكاة لتوضيح سلوك هاته المقدرات في وجود رقابة ، مع اهتمام خاص في حالة الرقابة من اليمين.

الكلمات المفتاحية: إحصائيات االترتيب، الرقابة ، الكميات المتطرفة ، مقدر هيل المعدل ، مقدر كابلان ماير