**République Algérienne Démocratique et Populaire**

**Ministère de L'Enseignement Supérieur et de la Recherche Scientifique**

**Université Mouloud Mammeri de Tizi-Ouzou**

**Faculté des sciences**

MÉMOIRE

Présenté pour l'obtention du **diplôme** de **Master**

**En :** Mathématiques

**Option : Probabilités et Statistique**

**Par : M**ellah **A**dlane

**Sujet:**

# Extrêmes sous données censurées

Soutenue le 08/10/2020, devant le jury composé de :

| | | | |
|---|---|---|---|
| M. BERKOUN Youcef | Professeur | à l'UMMTO | Président |
| Mme. MERABET Dalila | MCB | à l'UMMTO | Examinatrice |
| Mme. BOUALAM Karima | MCB | à l'UMMTO | Raporteur |

# *Dédicace*

Je dédie ce travail à ma famille, à mes amis et à tous ceux qui me sont chers.

# *Remerciements*

Je tiens à exprimer toute ma reconnaissance à ma directrice de mémoire, Mme Boualam Karima, pour le temps qu'elle m'a consacré durant la rédaction de ce mémoire. Je la remercie de m'avoir encadré, aidé, et conseillé.

Mes remerciements vont également à l'ensemble des membres du jury pour m'avoir fait l'honneur d'évaluer ce travail.

Je tiens aussi à remercier tout le corps professoral de l'université Mouloud Mammeri de Tizi-Ouzou, en particulier l'équipe de formation du master Probabilités et Statistique, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation de bonne qualité.

Enfin, j'adresse mes plus sincères remerciements à ma famille: ma mère, mon père, mes soeurs et mes frères, tous mes proches et amis, qui m'ont accompagné, aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire.

Merci.

Résumé:

Dans ce document, nous nous intéressons à l'estimation de l'indice des valeurs extrêmes (IVE), ainsi que le quantile extrême en précense de censure, autrementdit, dans le cas où certaines données sont incomplètes, en particulier dans le cas de censure aléatoire droite.
Nous commençons par rappeler, dans le premier chapitre, les notions de base de la théorie des valeurs extrêmes (TVE), avec un accent particulier sur les différentes approches utilisées dans l'estimation de l'IVE.

Le deuxième chapitre se divise en trois grandes parties. Dans la première, nous présentons l'analyse des données (durées) de survie, notamment la notion de troncature et de censure. Puis, dans la deuxième, nous définissons le modèle des extrêmes sous censure, en essayant d'appliquer la méthode proposée par Beirland et al (2007,[3]) pour l'estimation de l'IVE et quantile extrême sous censure aléatoire droite. La dernière partie, à vocation pratique est consacrée aux applications et au traitement des données réelles (données de SIDA). Nous illustrons avec des exemples de simulation l'influence du pourcentage de censure sur les estimateurs, ainsi le comportement empirique de ces estimateurs.

`Mots clés:` Théorie des valeurs extêmes, IVE, GEV, GPD, quantile extrême, estimateur de Hill, estimateur des moments, estimateur de Pickands, analyse de survie, censure aléatoire droite, extrêmes sous censure, SIDA.

# *Table des Matières*

# *List of Tables*

# *List of Figures*

# *Notations et Abréviations*

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

| **Abreviations & Notations** | **Explication** |
|---|---|
| TVE | : Théorie des Valeurs Extrêmes. |
| IVE | : Indice des Valeurs Extrêmes. |
| v.a. | : variable aléatoire. |
| iid | : indépendantes et identiquement distribuées. |
| := | : Égalité par définition. |
| al | : Autres. |
| $X \wedge Y$ | : Le minimum de X et Y. |
| $X_{i:n}$ | : La i-ème statistique d'ordre. |
| $X_1, X_2, ..., X_n$ | : Échantillon de taille n de X. |
| $X_{n:n}$ | : Maximum de l'échantillon $(X_1, X_2, ..., X_n)$. |
| $X_{1:n}$ | : Minimum de l'échantillon $(X_1, X_2, ..., X_n)$. |
| TCL | : Théorème Centrale Limite. |
| GEV | : Distribution Généralisée des Valeurs Extrêmes. |
| GPD | : Distribution de Pareto généralisée. |
| $x_{\mathrm{F}}$ | : Le point terminal de F. |
| $F \in D(H_\gamma)$ | : F appartient au domaine d'attraction de loi $H_\gamma$. |
| $\mathscr{N}(0,1)$ | : Loi normale centrée réduite. |
| $S = \overline{F}$ | : Fonction de survie. |
| $\mathbb{R}^+$ | : Ensemble des nombres réels positifs . |
| $\mathbb{P}(A)$ | : La probabilité de réalisation de l'événement A . |
| $\mathbb{B}(T)$ | : Le biais de l'estimateur T. |
| MSE(T) | : Écart quadratique moyen de T . |
| RMSE($T$) | : Racine carrée de l'erreur quadratique moyenne. |
| MAE($T$) | : Erreur absolue moyenne de T. |
| abs($a$) | : La valeur absolue de a. |
| $\mathbb{1}_{\{A\}}$ | : Fonction indicatrice de l'ensemble A. |
| $\lfloor x \rfloor$ | : Partie entière de x. |
| $\xrightarrow{\mathscr{L}}$ | : Convegence en loi. |

| | |
|---|---|
| $\xrightarrow{\text{p.s}}$ | : Convegence prèsque sûre. |
| $\xrightarrow{\mathbb{P}}$ | : Convegence en probabilité. |
| $L \in Rv_0$ | : L est une fonction à variation lente à l'infini. |
| $F \in Rv_\alpha$ | : F est une fonction à variation régulière d'indice $\alpha$ à l'infini. |
| Dl | : Développement limité. |
| Dln | : Développement limité d'orde n. |
| v(a) | : Voisinage de point a. |
| $\approx$ | : Approximativement. |
| $\mathbb{V}(X)$ | : Variance de la v.a X. |
| $\mathbb{E}(X)$ | : Espérance de la v.a Y. |
| resp. | : Respectivement. |
| † C.Q.F.D † | : Ce Qu'il Fallait Démontrer |
| IC. | : Intervalle de Confiance. |
| X $\hookrightarrow$ Exp($\lambda$) | : X est de loi Exponentielle de paramètre $\lambda > 0$. |

# *Introduction générale*

La théorie des valeurs extrêmes est une branche de la statistique apparue entre 1920 et 1943 grâce aux résultats de Fréchet (1927, [18]), Fisher-Tippet( 1928, [17]) et Gnedenko (1943, [19]). Elle a pour but de modéliser et décrire les événements dits extrêmes. ces derniers sont des valeurs beaucoup plus grandes ou plus petites que celles observées habituellement, d'une faible probabilité d'apparition. Cependant les conséquences peuvent s'avérer néfastes aussi bien sur le plan humain qu'économique. C'est pour cette raison, qu'il est important de ce prémunir contre ces risques. La TVE occupe ces dernières années, une place fondamentale dans la prévention et la gestion de ces aléas, elle représente un champs de recherche très actif en théorie et en pratique.

L'étude des données de survie ou l'analyse de survie, est le domaine qui s'intéresse au délai d'apparition d'un événement, au cours d'une période de temps bien déterminée. L'étude de ces données est l'objet de divers domaines comme la médecine, la fiabilité, l'économie, l'assurance, la psychologie...
Dans le domaine biomédical, les données de survie se rencontrent principalement en recherche clinique dans le cadre des essais thérapeutiques et dans les études de cohorte en épidémiologie (cohorte épidémiologique). En fiabilité industrielle, on s'intéresse à la durée de vie des composants d'un système. Les économistes s'intéressent à des durées d'épisodes de chômage. Tandis que les psychologues mesurent le temps nécessaire avec un sujet pour accomplir une tâche donnée. La spécificité des durées de survie est de correspondre à des variables aléatoires positives et que les données receuillées sont souvent incomplètes à cause de deux phénomènes distincts: la censure et la troncature. En effet, dans la plupart des études prospectives, les individus sont suivis pendant une durée d'observation fixée à l'avance. Pour les sujets pour lesquels l'événement d'intérêt a lieu pendant la période d'observation, on dispose de délai exact d'apparition de l'événement d'intérêt. Cependant, à la fin d'observation, certains individus n'auront pas eu l'événement d'intérêt, on aura alors pour ces individus qu'une information partielle, à savoir pour le délai d'apparition de l'événement est plus grand que la durée d'observation. De telles données sont dites censurées à droite et la durée d'observation constitue le délai de censure.

L'analyse des valeurs extrêmes sous censure est un problème de recherche relativement nouveau dans la littérature de la TVE, qui a reçu une grande attention au cours de ces dernières années. Il est mentionné pour la première fois en 1997 avec la sortie du livre Reiss et Thomas ([33]). Beirland et al (2007,[3]) ont proposé une méthode pour l'estimation de l'indice des valeurs extrême (IVE) sous censure aléatoire droite, qui consiste à adapter les estimateurs standards de l'IVE en les divisant par l'estimateur de la proportion de données non censurées

et ils ont illustré le comportement de cet estimateur proposé sur des données du SIDA. Par la suite, Einmahl et al (2008,[15]) ont repris la même méthode pour proposer un estimateur adapté de l'IVE, et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique. La recherche sur la théorie des valeurs extrêmes sous censure est devenue depuis une actualité.

Ce travail est composé de deux chapitres.

Le premier chapitre est consacré à des rappels sur les notions fondamentales de la théorie des valeurs extrêmes, et une présentation des différents estimateurs de l'indice des extrême et du quantile extrême.

Dans le deuxième chapitre, on présente brièvement l'analyse de survie, puis on s'intéresse à l'application de la TVE sur des données incomplètes censurées aléatoirement à droite. On termine ce chapitre par une simulation et une illustration sur des données réelles.
A la fin de ce mémoire, une conclusion qui résume l'intérêt de ce travail et donne quelques perspectives de recherche.

# *Chapter1*

# *Introduction à la théorie des valeurs extrêmes*

**Résumé:**

Le but de ce chapitre est de décrire et présenter les principaux résultats classiques sur la théorie des valeurs extrêmes qui permettent de faciliter la lecture de ce mémoire et à qui on fera appel dans les autres chapitres. Après l'introduction, nous allons donner quelques rappels sur les statistiques d'ordre. La partie (1.3) s'intéresse dans un premier temps au comportement asymptotique du maximum d'un échantillon de variables aléatoires, aux différentes lois limites possibles et domaines d'attraction, puis nous allons traiter dans un second temps (partie (1.4)) le comportement asymptotique des excès au-delà d'un seuil. La partie (1.5) propose une caractérisation des lois associées aux différents domaines d'attraction. Enfin, la partie (1.6) s'attache à décrire les différentes approches utilisées pour l'estimation d'un quantile extrême ainsi l'indice des valeurs extrêmes.

## 1.1 Introduction

*"La loi des grands nombres et la distribution gaussienne, fondements de l'étude statistique des grandeurs moyennes, échouent à rendre compte des événements rares ou extrêmes. Pour ce faire, des outils statistiques plus adaptés existent... mais ne sont pas toujours utilisés !"*

Rama Cont[1]

La plupart des approches statistiques classiques reposent sur l'étude du comportement moyen des phénomènes observés, par le bias d'outils statistiques comme par exemple la loi des grands nombres ou le théorème central limite. La théorie des valeurs extrêmes (TVE) quant à elle s'intéresse à l'étude des plus grandes valeurs des échantillons de données avec pour but d'en comprendre le comportement. La TVE s'attache à répondre aux deux questions suivantes:

(1) Quelle est la probabilité d'observer un évènement d'amplitude supérieure à une valeur x donnée?

(2) Quelle est l'amplitude de l'évènement qui est dépassée avec une faible probabilité p ?

La première question s'intéresse au calcul d'une faible probabilité (probabilité proche de zéro) associée à un évènement extrême. La deuxième question est la question duale de la première, Elle cherche à quantifier la valeur, aussi appelée quantile extrême dans le vocabulaire statistique, d'un évènement extrême. Pour répondre aux questions précédentes, on a l'intuition que ce sont les valeurs les plus extrêmes du jeu de données qui contiennent l'information pertinente, et non pas les valeurs centrales, moyennes, comme dans l'approche statistique classique. C'est avec cette idée que la théorie des valeurs extrêmes s'est développée, en s'appuyant notamment sur les résultats de Fisher et Tippett [1928] [17] et Gnedenko [1943] [19], puis sur le résultat de Pickands [1975] [31] qui donne la convergence en loi des excès au-delà d'un seuil.

## 1.2 Généralités sur les statistiques d'ordre

Les statistiques d'ordre sont utilisées dans divers domaines tels que le génie civil avec Longuet-Higgins [1952] [27] qui utilise la distribution de Rayleigh[2] et les statistiques d'ordre pour modéliser la hauteur des vagues et estimer la rigidité des digues de protection; Bernard et Ahmed [1958] [5] pour la modélisation des données de santé. Castillo et al.[2005] [7] qu'étudient la durée de vie des ampoules et les tests de défaillance par les statistiques d'ordre...etc. Par ailleurs, les statistiques d'ordre décrivant les variables aléatoires dans l'ordre de la magnitude sont très utilisées dans les méthodes statistiques et les inférences. Dans la pratique, les deux statistiques les plus importantes des statistiques d'ordre sont les distributions du minimum et du maximum car elles sont les valeurs critiques utilisées en ingénierie, physique, médecine,...etc.

[1] La statistique face aux événements rares. Voir le lien suivant: https://www.pourlascience.fr/sd/mathematiques/la-statistique-face-aux-evenements-rares-1095.php

[2] la distribution de Rayleigh ou la loi de Rayleigh est une loi de probabilité de densité : $f(x,\sigma^2) = \frac{x}{\sigma^2}\ exp\left(-\frac{x^2}{2\sigma^2}\right)$ pour $x \in ]0,\infty[$. Elle est nommée par le physicien anglais Lord Rayleigh. Voir le lien suivant: www.statisticshowto.com/rayleigh-distribution/

### 1.2.1 Statistiques d'ordre

Soit X une variable aléatoire (v.a.) de distribution commune F avec $F = P(X \leqslant X)$. Considérons $X_1, X_2, ..., X_n$ un échantillon iid de X. Rangeons $X_1, X_2, ..., X_n$ dans l'ordre croissant et soit :

$$X_{1:n} \leqslant X_{2:n} \leqslant ... \leqslant X_{n:n} \tag{1.1}$$

les statistiques d'ordre, telle que:

$$\begin{aligned} X_{1:n} &= \min \{X_1, X_2, ..., X_n\} \\ X_{2:n} &= \text{deuxième plus petite } X_i \\ &\vdots \\ X_{n:n} &= \max\{X_1, X_2, ..., X_n\} \end{aligned}$$

**Définition 1.2.1.** ***(Statistiques d'ordre extrêmes)***
*Les statistiques d'ordre extrêmes sont définies comme terme de maximum et du minimum des n variables aléatoires iid $X_1,\ X_2, ...,\ X_n$. La v.a. $X_{1:n}$ est la plus petite statistique d'ordre et $X_{n:n}$ est la plus grande statistique d'ordre.*

On écrit aisément l'égalité suivante:

$$\min\{X_1, X_2, ..., X_n\} = -\max\{-X_1,\ -X_2, ...,\ -X_n\}$$

### 1.2.2 Loi de la i-ème statistique d'ordre

Le résultat suivant caractérise les fonctions de probabilité d'une statistique d'ordre quelconque. Pour la démonstration voir David et Nagaraja ([2003], [9]).

**Proposition 1.** *Supposons que $X_1, X_2, ..., X_n$ sont des variables aléatoires iid. de fonction de répartition commune F continue. Alors la fonction de répartion de la i-ème statistique d'ordre $X_{i:n}$ est donnée par :*

$$F_{X_{i:n}}(x) = \sum_{j=i}^{n} \binom{n}{j} F^j(x)[1 - F(x)]^{n-j} \tag{1.2}$$

*Sa fonction de densité associée est définie par :*

$$f_{X_{i:n}}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1 - F(x)]^{n-i} f(x) \tag{1.3}$$

**Corollaire 1.** *En particulier, la fonction de répartition de la statistique d'ordre du maximum $X_{n:n}$ est :*

$$F_{X_{n:n}}(x) = [F(x)]^n \tag{1.4}$$

**Remarque 1.1.** On peut trouver le résultat de l'équation (1.4) en se basant sur le fait que les variables aléatoires $X_i$ sont iid, en effet:

$$\begin{aligned} F_{X_{n:n}}(x) &:= \mathbb{P}(X_{n:n} \leqslant x) \\ &= \mathbb{P}(X_1 \leqslant x, X_2 \leqslant x, ..., X_n \leqslant x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x) \\ &= F^n(x) \end{aligned}$$

*Ainsi, on obtient la densité correspondante en dérivant la relation (1.4) et on a :*

$$f_{X_{n:n}}(x) = nf(x)[F(x)]^{n-1} \tag{1.5}$$

### 1.2.3 Distribution conjointe d'un couple de statistiques d'ordre

**Proposition 2.**

1. *La densité jointe de couple $(X_{r:n}, X_{s:n})$ pour $1 \leqslant r < s \leqslant n$ où $x < y$ est donnée par :*

$$f_{(X_{r:n},X_{s:n})}(x,y) = \frac{n!}{(n-s)!(s-r-1)!(n-s)!} F^{r-1}(x) f(x) [F(y) - F(x)]^{s-r-1} [1 - F(y)]^{n-s} f(y) \tag{1.6}$$

2. *La fonction de répartition de couple $(X_{r:n}, X_{s:n})$ pour $1 \leqslant r < s \leqslant n$ où $x < y$ est donnée par :*

$$F_{(X_{r:n},X_{s:n})}(x,y) = \sum_{j=s}^{n} \sum_{i=r}^{j} \frac{n!}{i!(j-i)!(n-j)!} F^i(x) [F(y) - F(x)]^{j-i} [1 - F(y)]^{n-j} \tag{1.7}$$

**Corollaire 2.** *La densité jointe de n statistiques d'ordre $X_{1:n} \leqslant X_{2:n} \leqslant ... \leqslant X_{n:n}$ est donnée par :*

$$f_X(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) ... f(x_n) & Si \;\; x_1 < x_2 < .. < x_n \\ 0 & Sinon. \end{cases} \tag{1.8}$$

*avec* $X = (X_{1:n}, X_{2:n}, ..., X_{n:n})$

Pour la preuve des résultats cité dans la propositions (2) et le corollaire (2), voir David et Nagaraja (2003, [9]).

## 1.3 Comportement asymptotique du maximum d'un échantillon

Soit $X_1, X_2, ..., X_n$ un échantillon iid. de même loi que X, avec X une v.a. continue de fonction de répartition F et de fonction de survie associée $\overline{F}$ définie par :

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

D'après l'équation (1.4), la loi de maximum est liée à celle de F. La difficulté provient du fait que le comportement de cette dernière est en partie connu. En effet, si l'on définit $x_{\mathrm{F}}$ comme étant le point terminal de la fonction F, c'est à dire l'extrémité droite du support:

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

alors, on sait que:

$$\begin{cases} 0 \leqslant F(x) < 1 & \text{Si } x < x_F \\ F(x) = 1 & \text{Si } x \geqslant x_F \end{cases}$$

Par conséquence

$$\begin{aligned} \lim_{n\to+\infty} F_{X_{n:n}}(x) &= \lim_{n\to+\infty} F^n(x) \\ &= \begin{cases} 0 & \text{Si } x < x_F \\ 1 & \text{Si } x \geqslant x_F \end{cases} \end{aligned} \tag{1.9}$$

L'équation (1.9) nous indique donc que la loi limite du maximum est dégénérée. Ce résultat est très peu informatif sur le comportement du maximum et il préférable d'obtenir une loi non-dégénérée comme limite de l'équation (1.9).
Pour que la distribution du maximum ne soit pas dégénérée, il faut trouver une transformation linéaire, à l'image de celle utlisée dans le théorème central limite (TCL). La question est de savoir s'il existe un équivalent du TCL non pas pour la somme, mais pour le maximum de n variables aléatoires. Le théorème suivant donne une réponse positive à cette question.

**Théorème 1.1.** *(Théorème de Fisher-Tippett[1943]-Gnedenko[1975])*
*Soit $X_1, X_2, ..., X_n$ une suite de v.a iid. de fonction de répartition $F$. S'il existe deux suites normalisantes réelles $(a_n > 0,\ b_n \in \mathbb{R},\ n \geqslant 1)$ et une loi non-dégénérée $H$ telle que :*

$$\begin{aligned} \lim_{n\to+\infty} \mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \leqslant x\right) &= \lim_{n\to\infty} F^n(a_n x + b_n) \\ &= H(x),\ \forall x \in \mathbb{R} \end{aligned}$$

*Alors $H$ ne peut être que soit du type :*

- **Gumbel**

$$\Lambda(x) = exp\{-exp(-x)\} \quad \forall x \in \mathbb{R}$$

- **Fréchet**

$$\Phi_\alpha(x) = \begin{cases} 0 & Si\ x \leqslant 0 \\ exp\{-x^{-\alpha}\} & Si\ x > 0,\ \alpha > 0. \end{cases}$$

- **Weibull**

$$\Psi_\alpha(x) = \begin{cases} exp\{-(-x^\alpha)\} & Si\ x \leqslant 0 \\ 1 & Si\ x > 0,\ \alpha > 0. \end{cases}$$

**Remarque 1.2.**

Les trois lois de probabilité ci-dessus sont appelées lois des valeurs extrêmes. Pour la preuve, nous renvoyons le lecteur à Resnick [1987][34] ou encore Embrechis et Colla [2013][16].

Le théorème suivant dû à Von Mises [36] et Jenkinson [25] établit l'unification des trois types de loi en une loi unique dite distribution généralisée des valeurs extrêmes pour le maximum, en anglais GEV ( Generalized Extreme Value distribution).

**Théorème 1.2.** *Représentation de Jenkinson-von Mises*
*Soit $X_1, X_2, ..., X_n$ une suite de v.a iid. de fonction de répartition F. S'il existe deux suites normalisantes réelles ($a_n > 0$, $b_n \in \mathbb{R}$, $n \geqslant 1$) et une loi non-dégénérée $H_\gamma$ telle que :*

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

*Alors :*

$$H_\gamma(x) = \begin{cases} exp\{-(1+\gamma x)^{\frac{-1}{\gamma}}\} & Si\ \gamma \neq 0 \quad et\ 1+\gamma x > 0 \\ exp\{-exp(-x)\} & Si\ \gamma = 0. \end{cases} \tag{1.10}$$

Ce théorème joue un rôle aussi important dans la TVE que le TCL dans la théorie classique. Il indique que le comportement asymptotique du maximum renormalisé est régi par une seule loi $H_\gamma$. Le paramètre $\gamma$ caractérise le comportement de la queue de distribution de F. il est appelé l'indice des valeurs extrêmes, et on distingue trois formes différentes pour la loi $H_\gamma$, selon le signe de $\gamma$ . On parle du domaine d'attraction :

- Si $\gamma = 0$, on dit que F appartient au domaine d'attraction de Gumbel [20]. Ce domaine d'attraction regroupe les lois à queue legère, telle que les lois Normale, Exponentielle, Gamma,...
- Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de Fréchet. [18], Ce domaine d'attraction regroupe les lois à queue lourde, telle que les lois de Paréto, Student,...
- Si $\gamma < 0$, on dit que F appartient au domaine d'attraction de Weibull [37]. Ce domaine d'attraction comporte uniquement des lois dont le point terminal $x_F$ est fini telle que la loi Uniforme, Beta...

**Remarque 1.3.**

- La forme la plus générale de la GEV est :

$$H_{\gamma,\mu,\sigma}(x) = \begin{cases} exp\{-\left(1+\gamma\frac{x-\mu}{\sigma}\right)_+^{\frac{-1}{\gamma}}\} & \text{Si } \gamma \neq 0 \\ exp\{-e^{-\frac{x-\mu}{\sigma}}\} & \text{Si } \gamma = 0. \end{cases} \tag{1.11}$$

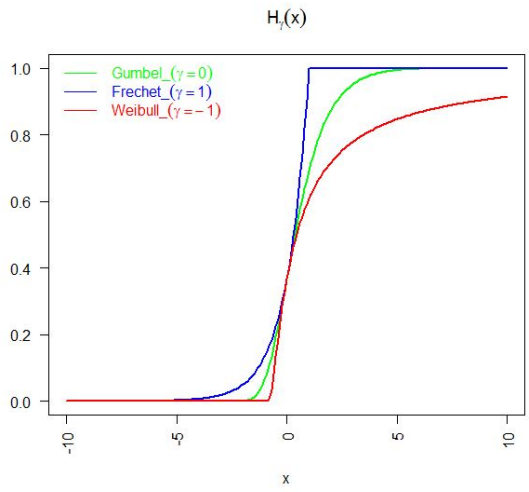
- La densité correspandante à une v.a de loi GEV est donnée dans ce cas par :

$$\begin{aligned} h_{\gamma,\mu,\sigma}(x) &:= (H_{\gamma,\mu,\sigma}(x))' \\ &= \begin{cases} \frac{1}{\sigma}\left(1+\gamma\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\gamma}-1} H_{\gamma,\mu,\sigma}(x) & \text{Si } \gamma \neq 0,\ 1+\gamma\frac{x-\mu}{\sigma} > 0 \\ \frac{1}{\sigma} exp\{-(\frac{x-\mu}{\sigma})\} exp\{-e^{-\frac{x-\mu}{\sigma}}\} & \text{Si } \gamma = 0. \end{cases} \end{aligned} \tag{1.12}$$

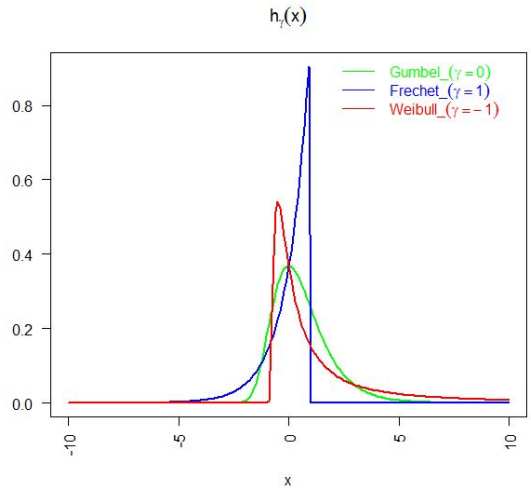
**Exemple 1.**

Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires iid. de loi exponentielle de paramètre $\lambda > 0$. Alors

$$F(x) = \begin{cases} 1 - exp\{-\lambda x\} & \text{Si } x \geqslant 0 \\ 0 & \text{Sinon .} \end{cases}$$




(a) Fonctions de répartitions




(b) fonctions de densités

Figure 1.1: Représentation graphique de la GEV avec $\mu = 0,\ \sigma = 1$ et $\gamma \in \{-1, 0, 1\}$

Si on normalise $X_{n:n}$ en utilisant des suites adéquantes, il vient :

$$\begin{aligned}
\mathbb{P}(X_{n:n} - \frac{logn}{\lambda} \leqslant x) &= \mathbb{P}(X_{n:n} \leqslant x + \frac{logn}{\lambda}) \\
&= F_{X_{n:n}}(x + \frac{logn}{\lambda}) \\
&= F^n(x + \frac{logn}{\lambda}) \\
&= \left(1 - e(-\lambda(x + \frac{logn}{\lambda}))\right)^n \\
&= e^{nlog(1-\frac{1}{n}e^{-\lambda x})} \\
&\rightarrow e^{-e^{-\lambda x}} \text{quand } n \rightarrow \infty \\
&= H_{\gamma=0}(x)
\end{aligned}$$

Par conséquent, la loi exponentielle de paramètre $\lambda > 0$ est dans le domaine d'attraction de Gumbel.

## 1.4 Comportement asymptotique des excès au-delà d'un seuil

L'approche basée sur la distribution GEV peut être réductrice du fait que l'utilisation d'un seul maxima conduit à une perte d'information contenue dans les autres grandes valeurs de l'échantillon. La solution est de considérer plusieurs grandes valeurs au lieu de la plus grande. La méthode des dépassement de seuil, en englais "Peaks-Over Threshold" notée POT proposée en 1975 par Pickands [31], est l'approche dual de comportement asymptotique de maximum d'un échantillon. Nous supposons que $X_1, X_2, ..., X_n$ un échantillon iid de variables aléatoires et fixons un seuil $\mu < x_F$. On s'intérèsse qu'aux $N_\mu$ obsérvations $X_{i1}, X_{i2}, ..., X_{i_{N_\mu}}$ qui dépassent le seuil $\mu$. Les excès au-delà du seuil $\mu$ sont alors définis par :

$$Y_j := X_{i_j} - \mu \tag{1.13}$$

avec $j = 1..N_\mu$ (voir la figure 1.2 )

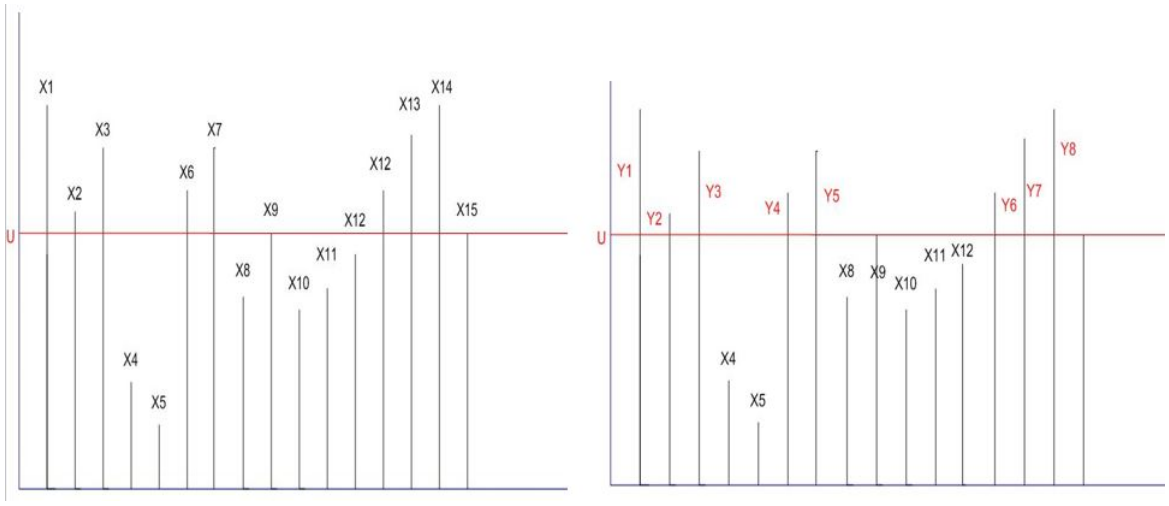



Figure 1.2: Illustration de la définition des excès.

**Définition 1.4.1.** ***Fonction des excès***
*Soit $X$ une v.a. de répartition $F$ et $\mu$ un seuil relativement grand ($\mu < x_{\mathrm{F}}$). On définit alors l'excès $Y$ de la v.a $X$ au-delà du seuil $\mu$ par $X - u$ sachant $X > u$. On appelle fonction des excès au-delà du seuil $\mu$, la fonction notée $F_\mu$ et définie par:*

$$\begin{array}{rccl} F_\mu & : & [0,\infty[ & \longrightarrow [0,1] \\ & & y & \longmapsto F_\mu(y) := \mathbb{P}(Y \leqslant y | X > \mu) = \mathbb{P}(X - \mu \leqslant y | X > \mu) \end{array} \tag{1.14}$$

**Remarque 1.4.** Par définition des probabilités conditionnelles, $F_\mu$ peut être également définie par :

$$\begin{aligned} F_\mu(y) &:= \mathbb{P}(X - \mu \leqslant y | X > \mu) \\ &= \frac{F(\mu + y) - F(\mu)}{1 - F(\mu)} \end{aligned}$$

Ou de manière équivalente:

$$\begin{aligned} \overline{F}_\mu(y) &= 1 - F_\mu(y) \\ &= 1 - \frac{F(\mu + y) - F(\mu)}{\overline{F}(\mu)} \\ &= \frac{\overline{F}(\mu + y)}{\overline{F}(\mu} \end{aligned} \tag{1.15}$$

Ce qui donne la relation:

$$\overline{F}_\mu(y) = \frac{\overline{F}(\mu + y)}{\overline{F}(\mu)} \tag{1.16}$$

Le théorème qui suit établit l'existence d'une loi limite pour les excès en se basant sur la fonction de répartition des excés.

**Théorème 1.3.** *(Théorème de Balkema et Haan [1974]; Pickands [1975])*
*$F$ appartient au domaine d'attraction de $H_\gamma$ si et seulement s'il existe $\sigma > 0$ et $\gamma \in \mathbb{R}$, tels que la loi des excès $F_\mu$ peut être uniformément approché par une loi de Paréto généralisée notée $G\gamma, \sigma$,i.e:*

$$\lim_{\mu \to y_F} \sup_{y \in ]0, x_F - \mu[} |F_\mu(y) - G_{\gamma,\sigma}(y)| = 0 \tag{1.17}$$

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

$$G_{\gamma,\sigma}(y) = \begin{cases} 1-\left(1+\gamma\frac{y}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{Si } \gamma \neq 0 \\ 1-e^{-\frac{y}{\sigma}} & \text{Si } \gamma = 0. \end{cases} \tag{1.18}$$

définie sur $\{y : y > 0 \quad et \quad (1+\gamma\frac{y}{\sigma}) > 0\}$
Elle dépend de deux paramètres :

**(a)** $\sigma > 0$ est un paramètre d'échelle.

**(b)** $\gamma \in \mathbb{R}$ est un paramètre de forme.

**Preuve 1. Preuve du théorème ( 1.3):**
D'après le théorème ( 1.2) on a :

$$\lim_{n\to\infty} F^n(a_n x + b_n) \simeq H_\gamma(x)$$

Donc

$$F^n(x) \simeq H_\gamma(\frac{x-b_n}{a_n})$$

ce qui donne que

$$F^n(\mu) \simeq exp\{-(1+\gamma(\frac{\mu-b_n}{a_n}))^{\frac{-1}{\gamma}}\}$$

En appliquant le logarithme des deux côtés de l'équation, il vient:

$$nlog(F(\mu)) \simeq -(1+\gamma(\frac{\mu-b_n}{a_n}))^{\frac{-1}{\gamma}}$$

un développement de logarithme donne alors, pour $\mu$ assez grand

$$1-F(\mu) \simeq \frac{1}{n}\left(1+\gamma(\frac{\mu-b_n}{a_n})\right)^{\frac{-1}{\gamma}}$$

De même, pour $y > 0$ on a :

$$1-F(\mu+y) \simeq \frac{1}{n}\left(1+\gamma\left(\frac{\mu+y-b_n}{a_n}\right)^{\frac{-1}{\gamma}}\right)$$

En remplaçons dans l'équation ( 1.15) on obtient:

$$\begin{aligned} \overline{F}_\mu(y) &= \frac{\overline{F}(\mu+y)}{\overline{F}(\mu)} \\ &= \frac{\frac{1}{n}\left(1+\gamma\left(\frac{\mu+y-b_n}{a_n}\right)^{\frac{-1}{\gamma}}\right)}{\frac{1}{n}\left(1+\gamma\left(\frac{\mu-b_n}{a_n}\right)^{\frac{-1}{\gamma}}\right)} \\ &= \left(\frac{1+\gamma\left(\frac{\mu+y-b_n}{a_n}\right)}{1+\gamma\left(\frac{\mu-b_n}{a_n}\right)}\right)^{\frac{-1}{\gamma}} \\ &= \left(1+\gamma\frac{y}{\sigma}\right)^{\frac{-1}{\gamma}} \end{aligned}$$

Ce qui démontre que $F_\mu(y) = G_{\gamma,\sigma}(y)$ avec :

$$\sigma = a_n + \gamma(\mu - b_n) \tag{1.19}$$

† C.Q.F.D †

**Remarque 1.5.**

1. Dans le cas où $\gamma = 0$, cette loi correspond à une loi exponentielle de paramètre $\frac{1}{\sigma}$

$$G_{0,\sigma}(y) = 1 - e^{\frac{-y}{\sigma}};\ y \geq 0.$$

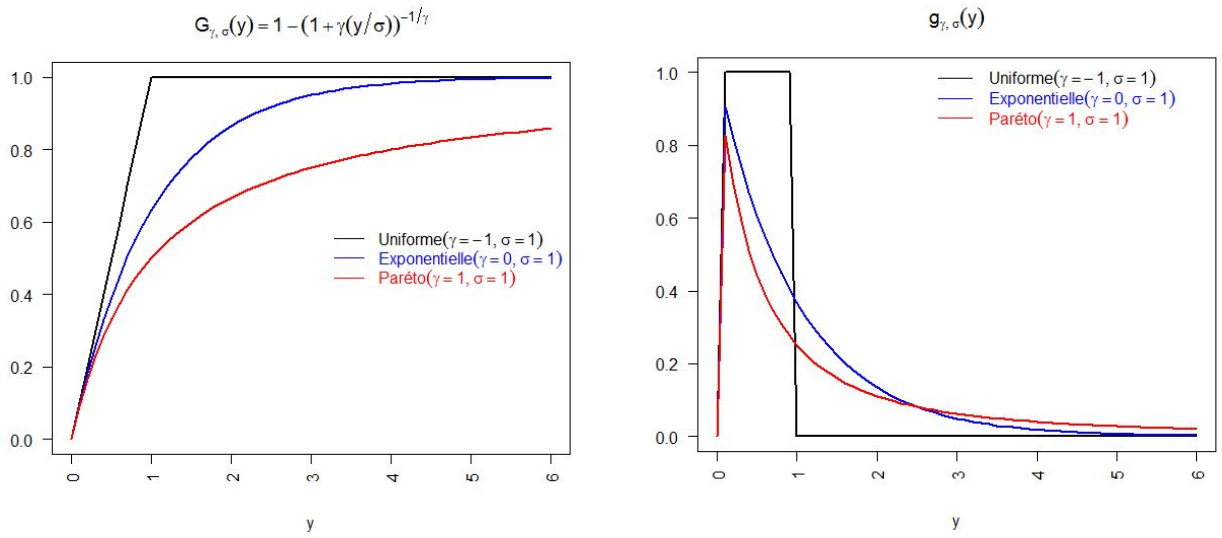
2. Dans le cas où $\gamma = -1$, cette loi correspond à une loi uniforme sur $[0, \sigma]$.

3. La densité de la GPD s'écrit comme suit :

$$g_{\gamma,\sigma}(y) \quad = \quad \begin{cases} \frac{1}{\sigma}\left(1 + \gamma\frac{y}{\sigma}\right)^{-\frac{1}{\gamma}-1} & \text{Si } \gamma \neq 0 \ \text{ et } 1 + \gamma\frac{y}{\sigma} > 0 \\ \frac{1}{\sigma}e^{-\frac{y}{\sigma}} & \text{Si } \gamma = 0. \end{cases} \tag{1.20}$$

4. $G_\gamma(y) = 1 + logH_\gamma(y)$ si $\ logH_\gamma(y) > -1$

5. Le théorème (1.3) nous indique que la loi de Paréto généralisée $G_{\gamma,\sigma}(.)$ régit le comportement des excès au-delà d'un seuil suffisamment grand.




(a) Fonctions de répartitions (b) fonctions de densités

Figure 1.3: Représentation graphique de la loi Pareto généralisée.

**Exemple 2.** Reprenos l'exemple de la loi exponentielle de paramètre $\lambda > 0$.
$\overline{F}(x) = e^{-\lambda x}$, et d'après l'équation 1.15 la fonction de répartition des excès s'écrit :

$$\overline{F}_\mu(y) = \frac{e^{-\lambda(\mu+y)}}{e^{-\lambda\mu}} = e^{-\lambda y} = 1 - G_{0,\frac{1}{\lambda}}(y) \quad \forall y > 0$$

Ce qui veut dire, les excès issus d'une loi exponentielle suivent également une loi exponentielle (Pareto avec $\gamma = 0$)

### 1.4.1 Détermination du seuil $\mu$

La difficulté avec la nouvelle approche basée sur les excès réside dans le choix d'un seuil $\mu$ approprié. En effet, le seuil $\mu$ doit être ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations. Une des méthodes pour la détermination d'un seuil performant est le Mean Excess Plot (MEP) appelée aussi Mean Residual Life Plot (MRLP).

**Définition 1.4.2.** (**Le Mean Excess Plot**)
*Le MEP est le graphe des points* $(\mu,\ e(\mu))$ *où* $e(\mu)$ *représente la moyenne des excès au-delà du seuil* $\mu$.

**Définition 1.4.3.** (Embrecht et al [16])
*Soit* $X$ *une v.a de répartion* $F$ *et d'espérance finie* ($\mathbb{E}(X) < \infty$). *On appelle fonction moyenne des excès (FME), la fonction notée* $e(.)$ *définie par :*

$$\forall \mu < x_F, \quad e(\mu) = \mathbb{E}(X - \mu | X > \mu) = \frac{1}{\overline{F}(\mu)} \int_{\mu}^{x_{\text{F}}} \overline{F}(s) ds \tag{1.21}$$

En pratique, la FME e(.) est estimée par la FME empirique $\widehat{e}_n(.)$ (Embrecht et al, [1997]) [16], telle que:

$$\widehat{e}_n(\mu) = \frac{\sum_{i=1}^{n} x_i \mathbb{1}\{x_i > \mu\}}{\sum_{i=1}^{n} \mathbb{1}\{x_i > \mu\}} - \mu \quad ; \quad \mu < x_{\text{F}}$$

avec $x_i$, $i = 1..n$ est la rèalisation de le i-ème v.a $X_i$ de loi F et $\mathbb{1}_{\{x_i > \mu\}}$ est la fonction indicatrice, définie par:

$$\mathbb{1}_{\{x_i > \mu\}} = \begin{cases} 1 & \text{si } \ x_i > \mu \\ 0 & \text{Sinon.} \end{cases}$$

**Proposition 3.** (Embrecht et al, (1997,[16]))
*Si* $(Y_1, Y_2, ..., Y_{N_\mu})$ *suivent une loi de Pareto généralisée* $G_{\gamma,\sigma}$ *pour* ($\gamma < 1, \quad \sigma > 0$) *alors:*

$$e(\mu) = \mathbb{E}(Y | X > \mu) = \frac{\gamma}{1-\gamma}\ \mu\ +\ \frac{\sigma}{1-\gamma} \quad avec \quad \gamma\mu + \sigma > 0.$$

Dans le cadre de la proposition ci-dessus, on remarque bien la linéarité en $\mu$ de la fonction moyenne des excès $e(\mu)$. Pour déterminer le seuil $\mu$, on trace le graphe suivant :

$$\mathcal{G}_\mu = \left\{ (\mu,\ \widehat{e}_n(\mu), \quad \min_{1 \le i \le n} x_i \le\ \mu\ \le \max_{1 \le i \le n} x_i \right\}$$

Une fois le graphe tracé, nous choisissons comme estimateur du seuil la valeur positive du $\mu$ à partir de laquelle le graphe de la FME empirique $\widehat{e}_n(\mu)$ est approximativement une droite.

**Exemple 3.** La figure suivante (Figure 1.4) présente le FME empirique d'un jeu de données composé de 17531 obsérvations. Ces données représentent l'accumulation quotidienne de pluie dans le sud-ouest de l'Angleterre durant les années $1914 - 1962$ (Coles (2001)[8]). Dans le R [3], ces données sont stockées dans le package ismev sous forme d'un vecteur numérique de nom rain (voir le programme 1.4.1).

```r
library(ismev)       ##Pour charger les donnees (rain)

library(fExtremes)   ## pour utiliser la fonction mrlPlot
                     ## qui renvoie la moyenne des exces

data(rain)

mrlPlot(rain)
```

**Program 1.4.1:** code R qui illustre la moyenne des excès pour les données rain

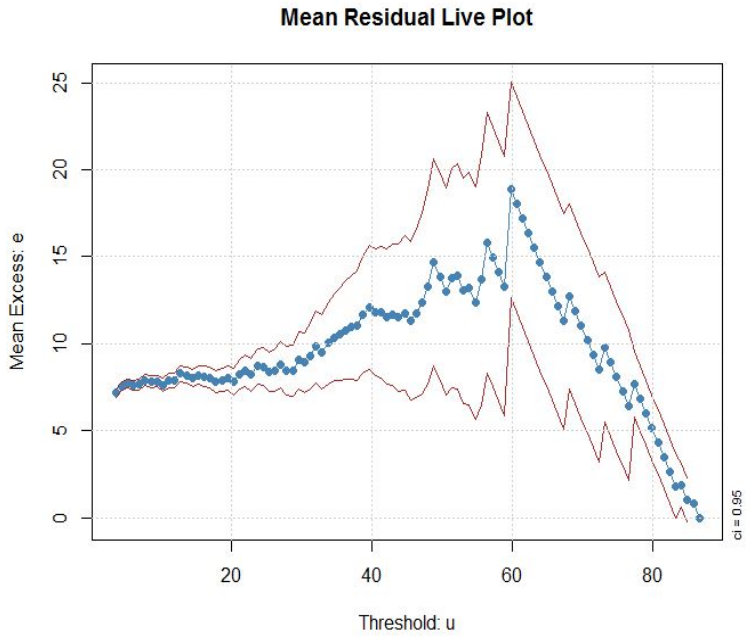



Figure 1.4: La fonction moyenne des excès .

D'après le graphe (Figure 1.4), la valeur de $\mu$ à partir duquel le graphe de la FME empirique est approximativement une droite est égale à 30 (estimation par intervalle de confiance du niveau 95%). Donc la MELP suggère de choisir comme estimation du seuil des excès cette valeur ($\widehat{\mu} = 30$).

## 1.5 Caractérisation des domaines d'attraction

Nous commencons par donner la définition de la notion de fonction à variation régulière et l'inverse généralisée d'une fonction.

### 1.5.1 Fonction à variation régulière

**Définition 1.5.1.** (Bingham et al. (1987, [6]))
*Une fonction positive mesurable $f : [a, \infty[ \longrightarrow \mathbb{R}_+$ $(a > 0)$ est dite à variation régulière à l'infini d'indice $\alpha$ et on note $f \in \mathrm{Rv}_\alpha$ si, pour tout $x > 0$ il existe un réel $\alpha$ tel que :*

$$\lim_{t \to +\infty} \frac{f(tx)}{f(t)} = x^\alpha \tag{1.22}$$

[3] R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données.

**Définition 1.5.2.** (Bingham et al.(1987, [6]))
*Une fonction $L$ est dite à variation lente si $L(t) > 0$ pour $t$ assez grand et si pour tout $x > 0$, on a:*

$$\lim_{t\to+\infty} \frac{L(tx)}{L(t)} = 1 \tag{1.23}$$

**Théorème 1.4.** (Représentation de Karamata, [6])
*La fonction $L$ est dite à variation lente à l'infini si et seulement si elle peut être représentée sous la forme suivante :*

$$\forall x \geq a > 0 \ , \ L(x) = c(x) exp \int_a^x \frac{\sigma(\mu)}{\mu} d\mu$$

*avec $c(.)$ et $\sigma(.)$ sont deux fonctions mesurables telles que :*

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

La proposition suivante nous donne quelques propriétés fondamentales des fonctions à variations lentes.

**Proposition 4.** (Bingham et al. (1987, [6]))

1. *Si $L$ est une fonction à variation lente à l'infini, alors :*

$$\lim_{x\to+\infty} \frac{logL(x)}{logx} = 0$$

2. *Si $L$ est une fonction à variation lente à l'infini et $\rho > 0$, alors :*

$$\lim_{x\to+\infty} x^{\rho} L(x) = +\infty \ et \ \lim_{x\to+\infty} x^{-\rho} L(x) = 0$$

3. *Si $L$ est une fonction à variation lente à l'infini, alors pour tout $\rho \in \mathbb{R}$ :*

$$L^{\rho} : x \longmapsto [L(x)]^{\rho}$$

   *est une fonction à variation lente à l'infini*

4. *Si $L_1$ et $L_2$ sont des fonctions à variation lente à l'infini, alors :*

$$L_1 + L_2 : x \longmapsto L_1(x) + L_2(x)$$

$$L_1 \times L_2 : x \longmapsto L_1(x) \times L_2(x)$$

   *sont à variation lente à l'infini. Si, de plus, $\lim_{x\to+\infty} L_2(x) = +\infty$, alors :*

$$L_1 \circ L_2 : x \longmapsto L_1[L_2(x)]$$

   *est aussi à variation lente à l'infini.*

Le théorème suivant qui suit, établit le lien entre une fonction à variation régulière d'indice $\alpha \in \mathbb{R}$ et les fonctions à variations lentes.

**Théorème 1.5.** (Bingham et al.(1987), [6])
*Soit $f$ : $[a,\infty[\longrightarrow\mathbb{R}_+$ $(a>0)$ une fonction positive mesurable et $\alpha\in\mathbb{R}$. Alors les assertions suivantes sont équivalentes:*

1. $f\in \mathrm{Rv}_\alpha$
2. $\exists L\in \mathrm{Rv}_0$ *telle que :*
$$f(x)=x^\alpha L(x) \tag{1.24}$$

La proposition suivante nous donne quelques propriétés fondamentales des fonctions à variations régulières.

**Proposition 5.** (Bingham et al.(1987), [6])
*Soient $\alpha_1$, $\alpha_2$ et $\alpha_3$ des réels*

1. *Si $f\in \mathrm{Rv}_\alpha$ alors*
$$\lim_{t\to+\infty} f(x)=\begin{cases}0 & Si\ \alpha<0\\ +\infty & Si\ \alpha>0\end{cases}$$
2. *Si $f\in \mathrm{Rv}_\alpha$ alors*
$$\lim_{t\to+\infty}\frac{logf(x)}{logx}=\alpha$$
3. *Si $f\in \mathrm{Rv}_\alpha$ et $\rho\in\ \mathbb{R}$, alors*
$$f^\rho : x\longmapsto[f(x)]^\rho\in\ \mathrm{Rv}_{\alpha\rho}$$
4. *Si $f_1\in \mathrm{Rv}_{\alpha_1}$ et $f_2\in \mathrm{Rv}_{\alpha_2}$ alors :*
$$f_1+f_2 : x\longmapsto f_1(x)+f_2(x)\in\ \ \mathrm{Rv}_{max\{\alpha_1,\alpha_2\}}$$
*et si de plus, $\lim_{x\to+\infty} f_2(x)=+\infty$, alors :*
$$f_1\circ f_2 : x\longmapsto f_1[f_2(x)]\in \mathrm{Rv}_{\alpha_1\alpha_2}$$

### 1.5.2 Inverse généralisée

**Définition 1.5.3.** *Soit $\varphi$ une fonction non-décroissante et continue à droite sur $\mathbb{R}$. L'inverse généralisée de $\varphi$ est définie par:*
$$\varphi^{\leftarrow}(q):=inf\{x|\ \ \varphi(x)\geq q\} \tag{1.25}$$
*avec la convention $inf\{\emptyset\}=+\infty$.*

Notons que l'inverse généralisé correspond à l'inverse classique $\varphi^{-1}$, lorsque la fonction considérée est continue et strictement croissante.

**Proposition 6.** *Soit $\varphi$ une fonction croissante et continue à droite. L'inverse généralisée $\varphi^{\leftarrow}$ est une fonction croissante et continue à gauche. On a les propriétés suivantes :*

1. $\varphi(\varphi^{\leftarrow}(q))\geq q$.
2. $\varphi^{\leftarrow}(q)\leq x\Longleftrightarrow q\leq\varphi(x)$.
3. $x<\varphi^{\leftarrow}(q)\Longleftrightarrow q>\varphi(x)$

### 1.5.3 Domaine d'attraction de Fréchet

▷ **Conditions nécessaires et suffisantes**

**Théorème 1.6.** (De Haan et Ferreira, (2007, [10]))
*Une fonction de répartition F appartient au domaine d'attraction de Fréchet ($F \in DA(Fréchet)$) si et seulement si:*

$$\begin{aligned} & 1) \quad x_F = \infty \\ et \quad & 2) \quad \lim_{t\to+\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma} \ , \ \gamma > 0 \end{aligned}$$

Autrement dit, au vu de la définition (1.5.1), $F \in \text{DA(Fréchet)}$ si et seulement si $\overline{F} \in Rv_{(-1/\gamma)}$. En utilisant la caractérisation donnée par le Théorème (1.5), il vient:

$$F \in \text{DA(Fréchet)} \Longleftrightarrow \overline{F}(x) = x^{-1/\gamma} L(x) \quad \text{avec} \quad L(x) \in Rv_0. \tag{1.26}$$

**Exemple 4.**

-La loi de Pareto, dont la fonction de survie est donnée par $\overline{F}(x) = x^{-\alpha}$ $(x > 0,\ \alpha > 0)$ appartient au domaine d'attraction de Fréchet avec un indice $\gamma = 1/\alpha$.

Le corollaire suivant reformule le théorème précédent en se basant sur la fonction quantile de queue définit par la relation suivante:

$$U(.) := F^{\leftarrow}(1 - 1/.) = \overline{F}^{\leftarrow}(1/.) = q(1/.) \tag{1.27}$$

**Corollaire 3.** (De Haan et Ferreira, (2007, [10]))
*Une fonction de répartition F appartient au domaine d'attraction de Fréchet ($F \in DA(Fréchet)$) si et seulement si:*

$$\begin{aligned} & 1) \quad x_F = \infty \\ et \quad & 2) \quad \forall x > 0, \lim_{t\to+\infty} \frac{U(tx)}{U(t)} = x^{\gamma} \ , \ \gamma > 0 \ (autrement\ U \in \text{Rv}_{\gamma}) \end{aligned}$$

▷ **Constantes de normalisation**

Les constantes de normalisations sont données en fonction de l'inverse généralisé de la répartition F.

**Proposition 7.** (De Haan et Ferreira, (2007, [10]))
*Si $F \in DA(Fréchet)$ alors un choix possible de suites normalisantes est:*

$$\begin{cases} a_n & := F^{\leftarrow}(1 - \frac{1}{n}) = U(n) \\ b_n & := 0 \end{cases}$$

Pour la démonstration et plus de détailles sur les résultats de théorème (1.6) , corollaire (3) et la proposition (7) voir De Haan et Ferreira (2007) [10].

### 1.5.4 Domaine d'attraction de Weibull

▷ **Conditions nécessaires et suffisantes**

**Théorème 1.7.** (De Haan et Ferreira, (2007, [10]))
*Une fonction de répartition $F$ appartient au domaine d'attraction de Weibull ($F \in DA(Weibull)$) si et seulement si:*

$$\begin{aligned} &1) \quad x_{\text{F}} < \infty \\ \textit{et} \\ &2) \quad \forall x > 0 \ , \exists \gamma < 0 \quad \textit{tel que} \quad \lim_{t \to 0^+} \frac{1 - F(x_F - tx)}{1 - F(x_F - t)} = x^{-1/\gamma} \end{aligned}$$

▷ **Constantes de normalisation** Un choix possible des constantes de normalisation est donnée dans la proposition suivante.

**Proposition 8.** (De Haan et Ferreira, (2007, [10]))
*Si $F \in DA(Weibull)$ alors un choix possible de suites normalisantes est:*

$$\begin{cases} a_n & := F^{\leftarrow}(1) - F^{\leftarrow}(1 - \frac{1}{n}) = x_{\text{F}} - U(n) \\ b_n & := F^{\leftarrow}(1) \end{cases}$$

### 1.5.5 Domaine d'attraction de Gumbel

Le domaine d'attraction de Gumbel correspond au cas $\gamma = 0$ dans le théorème (1.2).

▷ **Conditions nécessaires et suffisantes**

**Théorème 1.8.** (De Haan et Ferreira, (2007, [10]))
*Une fonction de répartition $F$ appartient au domaine d'attraction de Gumbel et on note ($F \in DA(Gumbel)$) si et seulement si:*

$$\text{a) } \mathbb{E}(X | x > c) < \infty \quad \textit{pour } c < F^{\leftarrow}(1)$$

$$\textit{et} \quad \text{b) } \lim_{t \to x_F} \frac{1 - F(t + xg(t))}{1 - F(t)} = e^{-x}$$

*avec :*

$$g(t) := \mathbb{E}(X - t | x > t) = \frac{\int_t^{x_F} 1 - F(s) ds}{1 - F(t)} \tag{1.28}$$

▷ **Constantes de normalisation**

**Proposition 9.** (De Haan et Ferreira, (2007, [10]))
*Si $F \in DA(Gumbel)$ alors un choix possible de suites normalisantes est:*

$$\begin{cases} a_n & := g(b_n) \\ b_n & := F^{\leftarrow}(1 - \frac{1}{n}) \end{cases}$$

*avec $g$ définie dans l'équation (1.28)*

### 1.5.6 Conditions suffisantes

Le théorème suivant dû à Von Mises(1936) tiré de livre de David et Nagaraja (2003, [9]), nous donne des conditions suffisantes facile à vérifier en pratique.

**Théorème 1.9.** (Conditions de Van-Mises, [9])
*Soit $X$ une v.a de répartition $F$ (continue) et de densité $f$. On pose:*

$$h(x) = \frac{f(x)}{1 - F(x)}$$

1. *$F \in DA(Fréchet)$ si:*

$$1) \quad x_{\text{F}} = \infty$$

$$et \quad 2) \quad h(x) > 0 \quad \forall x > 0 \ et \ \lim_{x \to \infty} xh(x) = 1/\gamma \ , \ (\gamma > 0)$$

2. *$F \in DA(Weibull)$ si:*

$$1) \quad x_{\text{F}} < \infty$$

$$et \quad 2) \quad \lim_{x \to x_{\text{F}}} (x_{\text{F}} - x)h(x) = -1/\gamma \ , \ (\gamma < 0)$$

3. *$F \in DA(Gumbel)$ si: $h(x) > 0$ et dérivable au voisinage de $x_{\text{F}}$, telle que:*

$$\lim_{x \to x_F} \frac{d}{dx} \left( \frac{1}{h(x)} \right) = 0$$

*Dans ce cas, on peut choisir comme constantes de normalisations :*

$$\begin{cases} a_n & = [nf(b_n)]^{-1} \\ b_n & = F^{\leftarrow}(1 - \frac{1}{n}). \end{cases}$$

Le tableau suivant (Tableau 1.1) propose une liste des lois usuelles et leur domaine d'attraction. (Pour plus d'exemples voir Embrechts et al (1997, [16])).

| **Domaines d'attraction** | **Weibull** $\gamma < 0$ | **Gumbel** $\gamma = 0$ | **Fréchet** $\gamma > 0$ |
|---|---|---|---|
| Distribution | Uniforme<br>Beta | Normale<br>Exponentielle<br>Log-normale<br>Gamma<br>Weibull<br>Logistique<br>Gumbel | Fréchet<br>Cauchy<br>Pareto<br>Student<br>Burr<br>khi-deux |

Table 1.1: Lois usuelles et leur domaine d'attraction.

## 1.6 Estimation de quantiles extrêmes

**Définition 1.6.1.** *Soit $X$ une v.a de répartition $F$. Le quantile d'ordre $1-p$ de la fonction de répartion $F$ est défini par :*

$$q(p) := \overline{F}^{\leftarrow}(p) = F^{\leftarrow}(1-p) = inf\{x|\ \ \overline{F}(x) \leq p\} \ \ avec\ p \in ]0,1[ \tag{1.29}$$

*Où $\overline{F}^{\leftarrow}$ représente l'inverse généralisé de $\overline{F}$ (*cf. *équation (1.25)).*

Il est à noter que d'après la propriété 1 de la proposition 6 (cf.(6)) :

$$\mathbb{P}(x > q(p)) = 1 - F(F^{\leftarrow}(1-p)) \leq p \tag{1.30}$$

On a l'égalité si $q(p)$ est un point de continuité de F.

Considérons $X_1, X_2, ..., X_n$ une suite de variables aléatoires iid de fonction de répartition F (continue) et $X_{1:n}, X_{2:n}, ..., X_{n:n}$ leurs statistiques d'ordre associées. On définit la fonction de répartition empirique $F_n$:

$$\begin{aligned} F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_{i:n} \leq x\}} \\ &= \begin{cases} 0 & \text{si } \ x < X_{1:n} \\ \frac{i-1}{n} & \text{si } \ X_{i-1:n} \leq x < X_{i:n} \\ 1 & \text{si } \ x \geq X_{n:n} \end{cases} \end{aligned} \tag{1.31}$$

Une manière d'estimer q(p) est d'inverser la fonction de répartition empirique.

$$\widehat{q}(p) = F_n^{\leftarrow}(1-p) = inf\{x|\ \ F_n(x) \geq 1-p\} \tag{1.32}$$

L'estimateur ci-dessus correspond aussi à la $\lfloor np \rfloor$ ième plus grande obsérvation de l'échantillon. Autrement, l'estimateur de quantile est aussi défini par :

$$\widehat{q}(p) = X_{n-\lfloor np \rfloor + 1:n} \tag{1.33}$$

Où $\lfloor . \rfloor$ représente la fonction partie entière.
Cependant, dans la deuxième question de l'introduction (1.1), nous sommes intéressés par l'estimation d'un quantile extrême dont l'ordre est $1-p_n$ avec $p_n$ très petit ( tend vers zéro). Ce qui implique que le nombre d'observation au-dessus de $q(p_n)$ est égal à un nombre très petit. Autrement, nous cherchons une valeur $q(p_n)$ qui est à droite de toutes ou presque toutes les observations.

**Définition 1.6.2.** ***Quantile extrême***
*Le quantile extrême d'ordre $1-p_n$ de la fonction de répartition $F$ est défini par :*

$$q(p_n) := \overline{F}^{\leftarrow}(p_n) \quad avec \quad \mathrm{p_n} \longrightarrow 0 \quad quand \quad \mathrm{n} \longrightarrow \infty$$

La difficulté de l'estimation des quantiles extrêmes réside dans la vitesse de convergence de $p_n$ vers zéro. En effet, si on cherche la probabilité que le quantile extrême soit plus grand que le

maximum de l'échantillon, sous l'hypothèse que les $X_i$ sont iid et que $p_n \longrightarrow 0$ quand $n \longrightarrow \infty$ alors cette probabilité s'écrit:

$$\begin{aligned}
\mathbb{P}(X_{n:n} < q(p_n)) &= \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i < q(p_n)\}\right) \\
&= \prod_{i=1}^{n}\mathbb{P}(X_i < q(p_n)) \\
&= F^n(q(p_n)) \\
&= (1-p_n)^n \qquad \text{Propriété de l'équation(1.30)} \\
&= e^{nlog(1-p_n)} \\
&\approx e^{n(-p_n+o(1))} \qquad \text{Dl de la fonction } log(1-u) \text{ au v(0)} \\
&\approx e^{-np_n(1+o(1))}.
\end{aligned} \tag{1.34}$$

D'après le résultat de l'équation ci-dessus, on remarque que la probabilité qu'un quantile extrême soit plus grand que le maximum de l'échantillon dépend du comportement asymptotique de la quantité $np_n$. Ainsi, on distingue deux cas: soit $np_n \longrightarrow \infty$ ou $np_n \longrightarrow 0$.

1. **Cas classique :** Si $np_n \longrightarrow \infty$ alors $\mathbb{P}(X_{n:n} < q(p_n)) \longrightarrow 0$
   Dans ce cas, le quantile se trouve presque surement à l'intérieur de l'échantillon. Un estimateur naturel de ce quantile est la $\lfloor np_n \rfloor$ ième plus grande obsérvation de l'échantillon

   $$\widehat{q}(p_n) = X_{n-\lfloor np_n \rfloor + 1:n} \tag{1.35}$$

   **Remarque 1.6.**

   ▷ Sous certains conditions, cet estimateur est asymptotiquement gaussien. (voir De Haan et Ferreira [10]).

   ▷ Dans la suite, nous notons l'ordre de ce quantile par $\alpha_n$ et $q(\alpha_n)$ son quantile associé.

2. **Cas extrême :** Si $np_n \longrightarrow 0$ alors $\mathbb{P}(X_{n:n} < q(p_n)) \longrightarrow 1$
   On cherche un quantile qui se trouve presque sûrement en dehors de l'échantillon. Dans ce cas, l'estimateur $q(p_n)$ ne peut être obtenu en inversant simplement la fonction de répartition empirique . En effet, $F_n(x) = 1$ pour $x > X_{n:n}$. Cela nécessite d'extrapoler les résultats de l'échantillon là où il n'y a pas de données observées. Pour ce faire, on va se servir des résultats des théorèmes (cf.(1.2) et (1.3)).

La figure suivante illustre les deux cas précédents pour un échantillon d'une loi de Weibull. La courbe noire correspond à la fonction de répartition de cette loi et les points noirs en abcisse représentent des réalisations de cette échantillon.

Il existe essentiellement deux méthodes permettant d'approximer la valeur d'un quantile extrême: l'approche par la loi des valeurs extrêmes et celle utilisant la loi de Pareto Généralisée. On considèrera également l'approche non-paramétrique.

### 1.6.1 L'approche par la loi GEV: méthodes des maxima par blocs

D'après le théorème (1.2) on a que

$$\mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \sim H_\gamma(x) \tag{1.36}$$

En appliquant le logarithme des deux côtés de l'équation (1.36), il vient que :

$$\lim_{n\to\infty} nlog(F(a_n x + b_n)) = \lim_{n\to\infty} nlog(1 - \overline{F}(a_n x + b_n)) = log(H_\gamma(x))$$

Sous l'hypothèse que $a_n x + b_n \longrightarrow x_F$ quand $n \longrightarrow \infty$ on aura $\overline{F}(a_n x + b_n) \longrightarrow 0$. Avec un Dl1 de la fonction $log(1-u)$ au voisinage de zéro on obtient:

$$\overline{F}(a_n x + b_n) \approx -\frac{1}{n} log(H_\gamma(x)).$$

En effectuant le changement de variable suivant :

$$z = a_n x + b_n \Longleftrightarrow x = \frac{z - b_n}{a_n}$$

Il vient donc:

$$\begin{aligned} \mathbb{P}(X_{n:n} \leq z) = \overline{F}(z) &\approx -\frac{1}{n} log\left(H_\gamma(\frac{z-b_n}{a_n})\right) \\ &\approx -\frac{1}{n} log\left(H_{\gamma,b_n,a_n}(z)\right) \end{aligned}$$

En remplaçant $H_{\gamma,b_n,a_n}(z)$ par son expression (cf.(1.11)) on obtient ainsi une approximation de la fonction de survie en queue:

$$\overline{F}(z) = \begin{cases} \frac{1}{n}\left(1 + \gamma(\frac{z-b_n}{a_n})\right)_+^{-\frac{1}{\gamma}} & \text{si } \gamma \neq 0. \\ \frac{1}{n} exp(-\frac{z-b_n}{a_n}) & \text{si } \gamma = 0. \end{cases} \tag{1.37}$$

Pour l'extrapolation d'un quantile extrême dont la définition est $q(p_n) = \overline{F}^{\leftarrow}(p_n)$, il nous faut inverser la fonction de survie de l'equation(1.37). Ce qui nous permet d'approcher le quantile extrême $q(p_n)$ par :

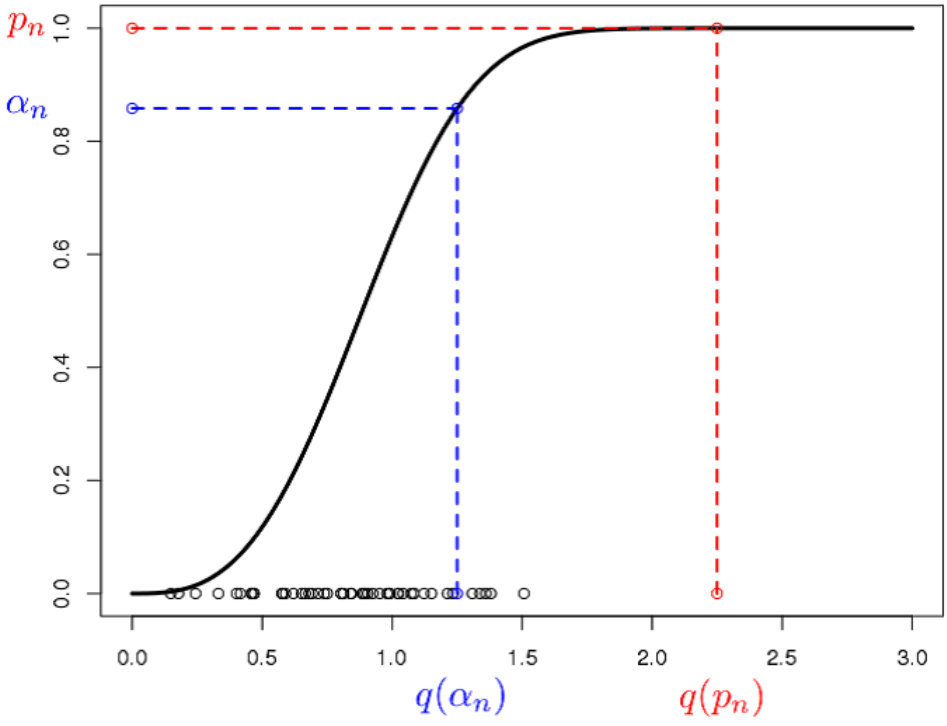



Figure 1.5: Illustration de la notion de quantile pour un échantillon d'une loi de Weibull. Le cas classique en bleu et cas extrême en rouge.

$$q(p_n) = \begin{cases} b_n + \frac{a_n}{\gamma}\left((\frac{1}{np_n})^{\gamma} - 1\right) & \text{si } \gamma \neq 0. \\ b_n - a_n log(np_n) & \text{si } \gamma = 0. \end{cases} \tag{1.38}$$

**Définition 1.6.3.**

*L'estimateur d'un quantile extrême basé sur la loi GEV des valeurs extrême est défini par :*

$$\widehat{q}_{GEV}(p_n) = \begin{cases} \widehat{b}_n + \frac{\widehat{a}_n}{\widehat{\gamma}_n}\left((\frac{1}{np_n})^{\widehat{\gamma}_n} - 1\right) & si\ \gamma \neq 0. \\ \widehat{b}_n - \widehat{a}_n log(np_n) & si\ \gamma = 0. \end{cases} \tag{1.39}$$

*où $(\widehat{a}_n, \widehat{b}_n, \widehat{\gamma}_n)$ sont respectivement des estimateurs des paramètres $(a_n, b_n, \gamma)$.*

Pour mettre cette méthode en pratique, on aura besoin d'un échantillon iid dont la loi est une loi GEV. Pour obtenir un tel échantillon, l'idée est de diviser l'échantillon $X_1, X_2, ..., X_n$ en m sous-échantillons (blocs) disjoints avec m suffisamment grand. Après on extraira le maximum de chaque bloc pour obtenir un nouveau échantillon des maxima noté $Z_1, Z_2, ..., Z_m$. D'après le théorème (1.2), les réalisations des variables aléatoires $Z_1, ..., Z_m$ se comportent comme des réalisations d'une loi GEV. Une fois cet échantillon de maxima obtenu, on peut alors s'en servir pour l'estimation des paramètres $a_n, b_n$ et $\gamma$.

### 1.6.1.1 Estimation des paramètres de la GEV

#### 1.6.1.1.1 Méthode de maximum de vraisemblance (MMV)

Considérons un échantillon $Z_1, Z_2, ..., Z_m$ des maxima et $z_1, z_2, ..., z_m$ des réalisations de cet échantillon. La MMV consiste à maximiser la fonction log-vraisemblance suivante obtenu à partir de la densité de la loi GEV (cf.(1.12)):

$$\begin{aligned} log\left(\mathscr{L}(\gamma, a_n, b_n; z_1, z_2, ..., z_m)\right) = \\ - mlog(a_n) - (1 + \frac{1}{\gamma}) \sum_{i=1}^{m} log\left(1 + \gamma\left(\frac{z_i - b_n}{a_n}\right)\right) \\ - \sum_{i=1}^{m}\left(1 + \gamma\left(\frac{z_i - b_n}{a_n}\right)\right)^{-\frac{1}{\gamma}} \end{aligned} \tag{1.40}$$

avec $1 + \gamma\left(\frac{z_i - b_n}{a_n}\right) > 0 \quad \forall i \in \{1, .., m\}$

Dans le cas $\gamma = 0$, l'expression de la fonction log-vraisemblance s'écrit:

$$log\left(\mathscr{L}(\gamma, a_n, b_n; z_1, z_2, ..., z_m)\right) = -mlog(a_n) - \sum_{i=1}^{m} exp\left(-\frac{z_i - b_n}{a_n}\right) - \sum_{i=1}^{m}\left(-\frac{z_i - b_n}{a_n}\right)$$

L'estimateur par la MMV des paramètres de la GEV n'a pas de forme explicite. La solution de l'equation (1.40) s'obtient à l'aide des méthodes numériques (algorithme). L'une de ces méthode est l'algorithme de Newton-Raphson (voir Hosking et al(1985), [22]).

#### 1.6.1.1.2 Méthode des moments pondérés(MMP)

**Définition 1.6.4.**

*Soit X une v.a de répartition F. On appelle moment pondéré généralisé d'ordre $p, r, s \in \mathbb{N}$, le nombre noté M(p,r,s) défini par :*

$$M(p,r,s) = \mathbb{E}[X^p (F(x))^r (1-\overline{F}(x))^s]$$

En partuclier, M(p,0,0) est le moment d'ordre p classique .

Considérons une v.a Z de loi GEV de répartition $H_{\gamma,b_n,a_n}(z)$. Pour $p = 1$ et $s = 0$ on a:

$$\beta_r = M(1,r,0) = \mathbb{E}[Z(H_{\gamma,b_n,a_n}(z))^r] \tag{1.41}$$

L'expression (1.41) est particulièrement utile dans l'estimation des paramètres de la GEV. Hosking et al [22] ont démontré que pour $\gamma < 1$ :

$$\beta_r = \frac{1}{r+1}\left(b_n - \frac{a_n}{\gamma}\left(1-(r+1)^{\gamma}\Gamma(1-\gamma)\right)\right) \tag{1.42}$$

avec $\Gamma(.)$ est la fonction gamma d'Euler définie par :

$$\Gamma(t) = \int_0^{\infty} u^{t-1}e^{-u}du \ ; \ \forall t > 0$$

En utilisant la formule de l'équation (1.42), les estimateur des paramètres $a_n, b_n$ et $\gamma$ sont obtenus en résolvant le système d'équation suivant :

$$\begin{cases} \beta_0 & = b_n - \frac{a_n}{\gamma}\left(1-\Gamma(1-\gamma)\right) \\ 2\beta_1 - \beta_0 & = -\frac{a_n}{\gamma}(1-2^{\gamma})\Gamma(1-\gamma) \\ \frac{3\beta_2-\beta_0}{2\beta_1-\beta_0} & = \frac{1-3^{\gamma}}{1-2^{\gamma}} \end{cases} \tag{1.43}$$

Où les $\beta_r$, $r \in 0,1,2$ sont respectivement remplacés par leurs estimateurs empiriques donnés par :

$$\widehat{\beta}_r[p_{j,m}] = \frac{1}{m}\sum_{j=1}^{m} p_{j,m}^r Z_{j:m}$$

Avec $Z_{j:m}$, $j = 1...m$ sont les statistiques d'ordre associées à l'échantillon $Z_1$, $Z_2$, ..., $Z_m$ et $p_{j,m}$ est choisi tel que: $p_{j,m} = (j-a)/m$ avec $0 < a < 1$ ou $p_{j,m} = (j-a)/(m+1-2a)$ avec $1/2 < a < 1/2$.
La solution du système linéaire (1.43) est déterminée par des méthodes numériques. Dans le cas $-1/2 < \gamma < 1/2$, on peut approcher la fonction $(1-3^{\gamma})/(1-2^{\gamma})$ par un polynôme linéaire de degré 2 et calculer les estimateurs des moments pondérés des paramètres $a_n, b_n$ et $\gamma$ dont l'expression est analytique (voir Hosting et al (1985))[22].

### 1.6.2 Approche par dépassements de seuil

L'approche par dépassement du seuil se base sur le théorème (1.3) et sur le lien qui unit la fonction de survie $\overline{F}$ et celle des excès $\overline{F}_{\mu}$ (cf.(1.16)). D'après l'équation(1.16), on a pour tout $y \geq 0$:

$$\overline{F}(\mu+y) = \overline{F}(\mu)\overline{F}_{\mu}(y) \tag{1.44}$$

Si l'on pose alors $x = \mu + y$, l'approximation de la queue de distribution donne:

$$\begin{aligned} \overline{F}(x) &= \overline{F}(\mu)\overline{F}_{\mu}(x-\mu) \\ &\approx \overline{F}(\mu)\overline{G}_{\gamma,\sigma}(x-\mu) \end{aligned} \tag{1.45}$$

où $\overline{G}_{\gamma,\sigma}$ est la fonction de survie de la loi Paréto généralisée (cf. Théorème (1.3)).
On introduit alors la probabilité $\alpha_n$ que Y dépasse le seuil $\mu$:

$$\alpha_n = \mathbb{P}(Y > \mu) = \overline{F}(\mu) \Longleftrightarrow \mu = \overline{F}^{\leftarrow}(\alpha_n) \tag{1.46}$$

D'où :

$$\overline{F}(x) = \alpha_n \overline{G}_{\gamma,\sigma}(x-\mu)$$

En remplaçant $\overline{G}_{\gamma,\sigma}(x-\mu)$ par son expression, on obtient une approximation de la queue de distribution :

$$\overline{F}(x) = \begin{cases} \alpha_n \left(1 + \gamma(\frac{x-u}{\sigma})\right)^{\frac{1}{\gamma}} & \text{si } \gamma \neq 0 \\ \alpha_n exp\left(-\frac{x-u}{\sigma}\right) & \text{si } \gamma = 0. \end{cases} \tag{1.47}$$

Afin d'obtenir des quantiles sous cette approche, on inverse l'expression (1.47) ce qui donne:

$$q(p_n) = \begin{cases} q(\alpha_n) + \frac{\sigma}{\gamma}\left((\frac{\alpha_n}{p_n})^{\gamma} - 1\right) & \text{si } \gamma \neq 0 \\ q(\alpha_n) - \sigma log(\frac{\alpha_n}{p_n}) & \text{si } \gamma = 0 \end{cases} \tag{1.48}$$

**Remarque 1.7.**

1. Il y a une similitude entre l'expression du quantile extrême basé sur la loi de la GEV (*cf.*(1.38)) et celle du quantile basé sur la GPD (cf.(1.48)). Il y a trois paramètres inconnus dans chacune d'entre eux:
   - L'IVE $\gamma$ qui est le même dans les deux approches.
   - Le paramètre d'échelle $\sigma$ joue le rôle de $a_n$ dans l'approche basée sur la loi GEV.
   - Le seuil $\mu = q(\alpha_n) = \overline{F}^{\leftarrow}(\alpha_n)$ joue le rôle de $b_n$ dans l'approche de la GEV.
2. Le seuil $\mu = q(\alpha_n)$ représente un quantile se trouvant dans l'échantillon, on peut l'estimer par inversion de la fonction de survie empirique (cf. (1.35)). Ainsi si $\alpha_n = k_n/n$ où $k_n$ est le nombre d'excès (auparavant noté $N_u$) , on peut estimer $\mu$ par la statistique d'ordre $X_{n-k_n+1:n}$. Une fois le seuil $\mu$ déterminé, il nous reste qu'à estimer $\gamma$ et $\sigma$.

**Définition 1.6.5.** *L'estimateur du quantile extrême basé sur la loi GPD est défini par :*

$$\widehat{q}_{GPD}(p_n) = \begin{cases} X_{n-k_n+1:n} + \frac{\widehat{\sigma}_n}{\widehat{\gamma}_n}\left((\frac{k_n}{np_n})^{\widehat{\gamma}_n} - 1\right) & si\ \gamma \neq 0 \\ X_{n-k_n+1:n} - \widehat{\sigma}_n log(\frac{k_n}{np_n}) & si\ \gamma = 0. \end{cases} \tag{1.49}$$

où $(\widehat{\gamma}_n, \widehat{\sigma}_n)$ sont respectivement les estimateurs des paramètres $(\gamma, \sigma)$.

#### 1.6.2.1 Estimation des paramètres de la GPD

Considérons un échantillon $X_1, X_2, ..., X_n$ de répartition F telle que $F \in D(H_\gamma)$. Soit alors $Y_1, Y_2, ..., Y_{k_n}$ l'échantillon des excés de loi GPD tel que :

$$\begin{aligned} Y_1 &:= X_{n-k_n+1:n} - X_{n-k_n:n} \\ Y_2 &:= X_{n-k_n+2:n} - X_{n-k_n:n} \\ &\vdots \\ Y_{k_n} &:= X_{n:n} - X_{n-k_n:n} \end{aligned}$$

et $y_1, y_2, ..., y_{k_n}$ des réalisations de cet échantillon.

##### 1.6.2.1.1 Méthode de maximum de vraisemblance

On cherche à maximiser la fonction de vraisemblance suivante:

$$\mathscr{L}(\gamma, \sigma; y_1, y_2, ..., y_{k_n}) = \prod_{i=1}^{k_n} g_{\gamma,\sigma}(y_i) \tag{1.50}$$

avec $g_{\gamma,\sigma}$ est la densité de la loi GPD (cf.(1.18))
maximiser la fonction $\mathscr{L}(.)$ revient à trouver les valeurs de $\gamma$ et $\sigma$ qui maximisent la fonction log-vraisemblance suivante :

$$\begin{aligned} log\mathscr{L}(\gamma, \sigma; y_1, y_2, ..., y_{k_n}) &= \sum_{i=1}^{k_n} log\left(g_{\gamma,\sigma}(y_i)\right) \\ &= \sum_{i=1}^{k_n} log\left(\frac{1}{\sigma}(1+\frac{\gamma}{\sigma}y_i)^{-\frac{1}{\gamma}-1}\right) \\ &= -k_n log\sigma - (1+\frac{1}{\gamma})\sum_{i=1}^{k_n} log(1+\frac{\gamma}{\sigma}y_i) \end{aligned} \tag{1.51}$$

avec $1+\frac{\gamma}{\sigma}y_i > 0$ pour tout $i = 1...k_n$.
La solution de l'équation (1.51) n'est pas explicite. Elle est déterminée par des méthodes numériques telle que l'algorithme de Newton-Raphson (voir [23]).

##### 1.6.2.1.2 Méthode des moments (MOM)

La MOM pour l'estimation des paramètres de la GPD a été proposée par Hosking et Wallis [23]. Dans le cas où $\gamma < 1/2$, l'espérance et la variance d'une v.a Y issue d'une loi GPD $G_{\gamma,\sigma}$ existent, et elles sont données par :

$$\mathbb{E}(Y) = \frac{\sigma}{1-\gamma} \quad \text{et} \quad \mathbb{V}(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}$$

Pour estimer les deux paramètres de la GPD, il faut exprimer $\gamma$ et $\sigma$ comme fonction de ces derniers:

$$\begin{cases} \gamma = \frac{1}{2}\left(1 - \frac{[\mathbb{E}(Y)]^2}{\mathbb{V}(Y)}\right) \\ \sigma = \frac{\mathbb{E}(Y)}{2}\left(1 + \frac{[\mathbb{E}(Y)]^2}{\mathbb{V}(Y)}\right) \end{cases}$$

En remplaçant $\mathbb{E}(Y)$ et $\mathbb{V}(Y)$ par leurs estimateurs empiriques donnés respectivement par ::

$$\overline{Y} := \frac{1}{k_n}\sum_{i=1}^{k_n} Y_i \quad \text{et} \quad S_Y^2 := \frac{1}{k_n - 1}\sum_{i=1}^{k_n}(Y_i - \overline{Y})^2$$

on obtient les estimateurs des moments de $\gamma$ et $\sigma$ suivant:

$$\begin{cases} \widehat{\gamma}_n = \frac{1}{2}\left(1 - \left(\frac{\overline{Y}}{S_Y}\right)^2\right) \\ \widehat{\sigma}_n = \frac{\overline{Y}}{2}\left(1 + \left(\frac{\overline{Y}}{S_Y}\right)^2\right) \end{cases}$$

#### 1.6.2.1.3 Méthode des moments pondérés

La MMP pour l'estimation des paramètres de la GPD a été proposée par Hosking et Wallis [23].

Soit $\alpha_s$ le moment pondéré d'ordre s de la v.a Y de loi $G_{\gamma,\sigma}$ tel que :

$$\alpha_s := M(1,0,s) = \mathbb{E}[Y(1 - G_{\gamma,\sigma}(Y))^s] = \frac{\sigma}{(1+s)(1+s-\gamma)} \quad \text{avec} \quad \gamma < 1$$

Les estimateurs des moments pondérés sont les solution des équations suivantes :

$$\begin{cases} \alpha_0 &= \dfrac{\sigma}{1-\gamma} \\ \alpha_1 &= \dfrac{\sigma}{2(2-\gamma)} \end{cases}$$

Ce qui nous permet d'exprimer $\gamma$ et $\sigma$ en fonction de $\alpha_0$ et $\alpha_1$, soit :

$$\gamma = \frac{4\alpha_1 - \alpha_0}{2\alpha_1 - \alpha_0} \quad \text{et} \quad \sigma = \frac{2\alpha_1\alpha_0}{\alpha_0 - 2\alpha_1} \tag{1.52}$$

Les estimateurs $\widehat{\gamma}_n$ et $\widehat{\sigma}_n$ sont obtenus en remplaçant $\alpha_s,\ s \in \{0,1\}$ par son estimateur empirique:

$$\widehat{\alpha}_s = \frac{1}{k_n}\sum_{j=1}^{k_n}\left(1 - \frac{j}{k_n+1}\right)^s Y_{j:k_n}$$

avec $Y_{1:k_n}, Y_{2:k_n}, ..., Y_{k_n:k_n}$ sont les statistiques d'ordre associées à l'échantillon des excès $Y_1, Y_2, ..., Y_{k_n}$.

### 1.6.3 Estimateurs non-paramétriques

Dans ce paragraphe nous nous intéressons aux estimateurs non-paramétriques les plus courants de l'IVE, ainsi l'estimateur du quantile extrême associé.

#### 1.6.3.1 Estimateur de Hill

Cet estimateur a été introduit par Hill en 1975 [21] pour estimer d'une manière non-paramétrique l'IVE pour les lois appartenant au domaine d'attraction de Fréchet ($\gamma > 0$). Rappelons qu'une fonction de répartition $F \in D(\text{Fréchet})$ vérifie $\overline{F} \in Rv_{-1/\gamma}$:

$$\overline{F}(x) = x^{-1/\gamma}L(x) \quad \text{avec} L \in Rv_0 \tag{1.53}$$

**Définition 1.6.6.**

*Considérons $X_1, X_2, ..., X_n$ une suite de variables aléatoires iid de fonction de répartition $F \in D(H\gamma)$ avec $\gamma > 0$. Soit $k_n$ une suite d'entiers tels que $1 < k_n < n$. L'estimateur de Hill est défini par :*

$$\widehat{\gamma}_n^H := \frac{1}{k_n} \sum_{i=1}^{k_n} log(X_{n-i+1:n}) - log(X_{n-k_n:n}) \tag{1.54}$$

**Propriétés asymptotiques de l'estimateur de Hill**

**Proposition 10.** *Soit $F \in D(H\gamma)$ avec $\gamma > 0$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$ alors on a :*

1. **Convergence en probabilité:** (Mason (1982, [28])*:*

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{H}} \xrightarrow{\mathbb{P}} \gamma \quad quand \quad n \longrightarrow \infty$$

2. **Convergence p.s:** Deheuvels et al (1988, [11])
   *Si $k_n/loglogn \longrightarrow \infty$ quand $n \longrightarrow \infty$ alors :*

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{H}} \xrightarrow{\mathrm{p.s}} \gamma \quad quand \quad n \longrightarrow \infty$$

Pour établir la normalité asymptotique de l'estimateur de Hill, on aura besoin d'une hypothèse sur la fonction à variation régilière.

**Définition 1.6.7.** ***(Condition du second-ordre )***
*On dit qu'une fonction quantile $U$ est à variation régulière du second ordre d'indice $(\gamma, \rho)$, $\rho \leq 0$, s'il existe une fonction $A(.)$ à signe constant avec $A \in \mathrm{Rv}_\rho$ et $A(t) \longrightarrow 0$ quand $t \longrightarrow \infty$, telle que :*

$$\lim_{t\to\infty} \left( \frac{U(tx)/U(t) - x^\gamma}{A(t)} \right) = cx^\gamma \frac{\lambda^\rho - 1}{\rho}, \qquad \forall x > 0, \ c \in \mathbb{R}^* \tag{1.55}$$

**Remarque 1.8.** Le paramètre $\rho < 0$ contrôle la vitesse de convergence dans (1.55). Si $\rho$ est petit (très négatif), la convergence est rapide et l'estimateur de Hill aura un bon comportement. À l'inverse, si $\rho$ est proche de 0, la convergence est lente et l'estimateur de Hill présentera un biais important.

**Théorème 1.10.** *(Normalité asymptotique del'estimateur de Hill* ( Beirlant et al [4])*)*
*Soit $F \in D(H\gamma)$ avec $\gamma > 0$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$. Supposant que $F$ vérifie la condition du second-ordre ci-dessus. Si $\sqrt{k_n} A(\frac{n}{k_n}) \longrightarrow \lambda$ quand $n \longrightarrow \infty$ alors :*

$$\sqrt{\mathrm{k_n}}(\widehat{\gamma}_{\mathrm{n}}^{\mathrm{H}} - \gamma) \xrightarrow{\mathscr{L}} \mathscr{N}(\frac{\lambda}{1-\rho}, \gamma^2)$$

**Remarque 1.9.**

1. L'estimateur de Hill est biaisé et le résultat sur la normalité asymptotique permet de donner un intervalle de confiance (IC) pour l'estimation. Par exemple pour un niveau de confiance de $(1-\alpha)\%$ on a :

$$\gamma \in \left] \widehat{\gamma}_n^H - t_{1-\alpha/2} \frac{\widehat{\gamma}_n^H}{\sqrt{k_n}}, \ \widehat{\gamma}_n^H + t_{1-\alpha/2} \frac{\widehat{\gamma}_n^H}{\sqrt{k_n}} \right[$$

   avec $t_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite.

2. Cet estimateur est influencé par le choix du paramètre $k_n$:

   - Si $k_n$ est trop grand: l'approximation par une loi de Pareto sera mauvaise ($\mathbb{B}[\widehat{\gamma}_n^H]$) est important.
   - Si $k_n$ est très petit: on aura peu d'observations pour l'estimation de $\gamma \Rightarrow \mathbb{V}(\widehat{\gamma}_n^H)$ est importante.

   Le bon choix de $k_n$ est donc celui de meilleur compromis biais/variance de telle sorte: $k \longrightarrow \infty$ et $k_n/n \longrightarrow 0$ (assez grand mais pas trop grand).

Pour illustrer cette situation, on trace le graphe de l'estimateur de Hill en fonction de $k_n$, c'est-à-dire $k_n \longmapsto f(k_n) = \widehat{\gamma}_n^H$ (voir Figure (1.6)). Sur le graphe on observe une grande volatilité pour les petites valeurs de $k_n$ (entre 0 et 60), puis vient une zone de stabilité et enfin pour $k_n > 260$ l'estimateur de Hill devient de plus en plus biaisé. Le meilleur choix de $k_n$ se situe dans la zone de stabilité .

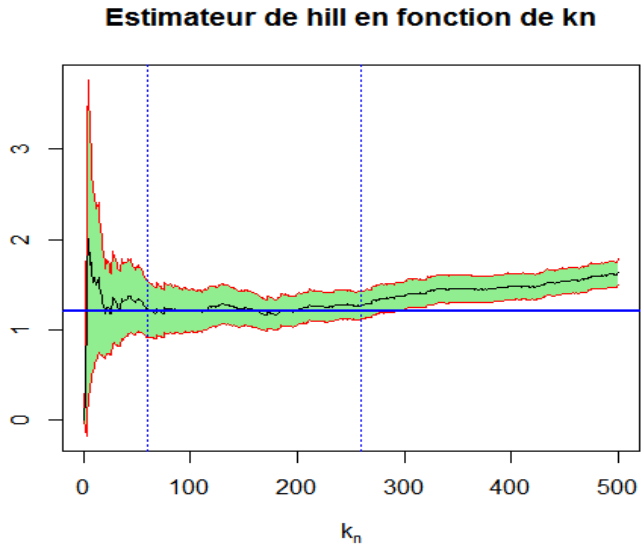



Figure 1.6: Graphique de l'estimateur de Hill $\widehat{\gamma}_n^H$ en fonction de $k_n$ (trait noir) avec son intervalle de confiance à 95% (lignes rouges), $n = 500$ réalisations d'une loi Pareto généralisée pour laquelle $\gamma = 1.2, \mu = 1$ et $\sigma = 2$. La ligne horizontale en blue représente la vraie valeur de gamma, les deux lignes verticales en blue représentent les frontières de la zone de stabilité.

#### 1.6.3.1.1 Pareto quantile plot

Considérons $X_1, X_2, ..., X_n$ n variables aléatoires iid de de répartitions $F \in$ DA(Fréchet). F s'écrit donc comme suit: $\overline{F}(x) = x^{-1/\gamma} l(x)$ avec $l \in Rv_0$. Soit $k_n$ une suite d'entiers telle que: $1 < k_n < n$. D'après le corollaire (3) la fonction quantile vérifie dans ce cas:

$$U(x) = x^{\gamma} l_u(x) \quad \text{avec} \quad l_u \in Rv_0 \tag{1.56}$$

En appliquons le logarithme aux deux côtés de l'équation(1.56), il vient:

$$\begin{aligned} log(U(x)) &= \gamma log x + log(l_u(x)) \\ &= \gamma log x \left(1 + \frac{log l_u(x)}{\gamma log x}\right) \\ &\sim \gamma log x \quad \text{quand} \quad x \longrightarrow \infty \end{aligned} \tag{1.57}$$

Posons $x = \dfrac{n+1}{j}$ avec $j = 1...k_n$, on a donc:

$$logU\left(\frac{n+1}{j}\right) \sim \gamma logx$$

Or $U\left(\dfrac{n+1}{j}\right) := F^{\leftarrow}\left(1 - \dfrac{j}{n+1}\right) = q(\dfrac{j}{n+1})$ (cf.(1.27))
Donc $U\left(\dfrac{n+1}{j}\right)$ représente un quantile puisque $0 < \dfrac{j}{n+1} < 1$, il est estimé par la statistique d'ordre $X_{n-j+1:n}$ (cf. (1.35)). Ce qui nous conduit à la relation suivante :

$$logX_{n-j+1:n} \sim \gamma log\left(\frac{n+1}{j}\right) \tag{1.58}$$

**Définition 1.6.8.** *On appelle graphe "Pareto quantile plot" (PQL) le graphe des points :*

$$\left(log\left(\frac{n+1}{j}\right), log(X_{n-j+1:n})\right) \quad j = 1...k_n \tag{1.59}$$

Ce graphe nous permet de visualiser facilement si les observations sont distribuées selon une loi appartenant au domaine d'attraction de Fréchet. En effet, si un échantillon provient d'une loi appartenant à ce domaine, le PQL doit être approximativement linéaire avec une pente $\gamma$ pour les petites valeurs de j, autrement dit pour les valeurs extrêmes. Il devient linéaire à partir du point $\left(log\left(\dfrac{n+1}{k_n}\right), log(X_{n-k_n+1:n})\right)$. Ce qui nous permet de déterminé graphiquement une estimation de Hill de l'IVE $\gamma$.

La figure (1.7) illustre le PQL pour un échantillon de loi Pareto généralisée de paramètres $\gamma = 1$ et $\sigma = 1$

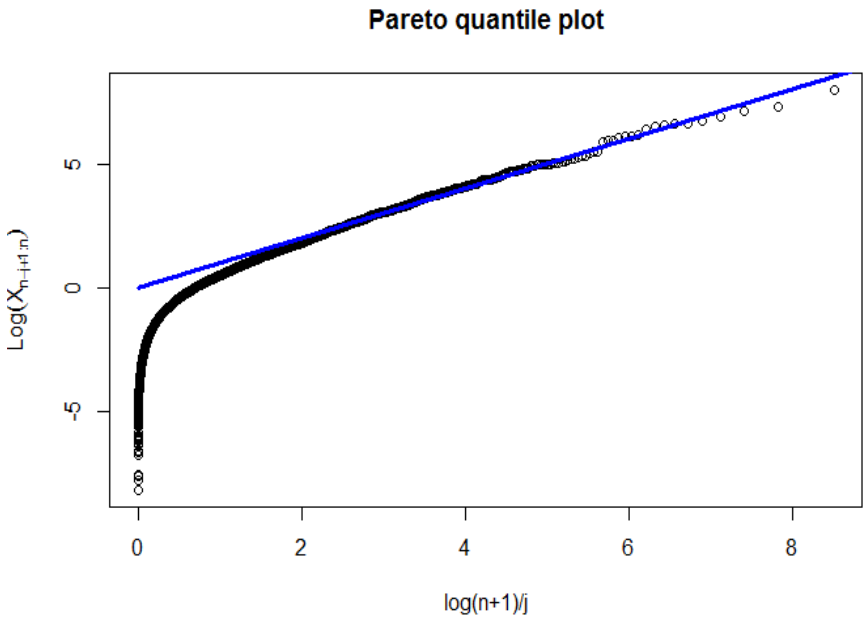



Figure 1.7: Le graphe PQL de n=5000 réalisations d'une loi PGD avec $(\gamma = 1, \sigma = 1)$ à lequel on a ajuster la droite d'équation $logX_{n-j+1:n} = log(n + 1/j)$ ( en blue).

**Définition 1.6.9.** *(Estimateur de Weissman [38])*

*Considérons $X_1, X_2, ..., X_n$ une suite de variables aléatoires iid de fonction de répartition ($F \in D(H\gamma)$) avec $\gamma > 0$. Soit $k_n$ une suite d'entiers tels que $1 < k_n < n$. L'estimateur de Weissman du quantile extrême $q(p_n)$ est défini par :*

$$\widehat{q}_n^W(p_n) := X_{n-k_n+1,n}\left(\frac{k_n}{np_n}\right)^{\widehat{\gamma}_n^H} \tag{1.60}$$

#### 1.6.3.2 Estimateur de Dekkers et al (estimateur des moments)

L'estimateur de l'IVE proposé par Dekkers et al (1989) [13] est une généralisation du célèbre estimateur de Hill ( cf.(1.54)) à tous les domaines d'attraction, permettant ainsi d'estimer $\gamma$ qualque soit sa valeur.

**Définition 1.6.10.** *: Estimateur des moments*
*Soit $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$ et $k_n$ une suite d'entier telle que $1 < k_n < n$. L'estimateur des moments est donné par:*

$$\widehat{\gamma}_n^D = M_{X,n}^{(1)} + S_{X,n} = M_{X,n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_{X,n}^{(1)}\right)^2}{M_{X,n}^{(2)}}\right)^{-1} \tag{1.61}$$

*avec :*

$$\begin{cases} M_{X,n}^{(a)} & = \frac{1}{k_n}\sum_{i=1}^{k_n}(logX_{n-i+1:n} - logX_{n-k_n:n})^a \quad ; \quad a \in \{1,2\} \\ S_{X,n} & = 1 - \frac{1}{2}\left(1 - \frac{\left(M_{X,n}^{(1)}\right)^2}{M_{X,n}^{(2)}}\right)^{-1} \end{cases} \tag{1.62}$$

On note que $M_{X,n}^{(1)}$ correspond exactement à l'estimateur de Hill.(1.54)
Dekkers et al proposent comme estimateur de $\sigma > 0$:

$$\widehat{\sigma}_n = X_{n-k_n+1:n} M_{X,n}^{(1)}(1 - \widehat{\gamma}_n^D + M_{X,n}^{(1)})$$

Ainsi, en remplaçant $\gamma$ et $\sigma$ par leurs estimateurs dans l'équation (1.49), on obtient l'estimateur du quantile extrême .

**Définition 1.6.11.** *L'estimateur de Dekkers et al du quantile extrême est donné par :*

$$\widehat{q}_n^D = X_{n-k_n+1:n} + \widehat{a}_n^M \frac{(k_n/(np_n))^{\widehat{\gamma}_n^D} - 1}{\widehat{\gamma}_n^D} \tag{1.63}$$

*avec :*

$$\widehat{a}_n^M = \frac{M_{X,n}^{(1)}}{\rho(\widehat{\gamma}_n^D)} X_{n-k_n+1:n} \quad , \quad \rho(\gamma) = \begin{cases} 1 & si \ \ \gamma \geq 0 \\ \frac{1}{1-\gamma} & si \ \ \gamma < 0 \end{cases}$$

**Propriétés asymptotiques de l'estimateur $\widehat{\gamma}_n^M$** (Dekkers et al [13]).

**Proposition 11.** *Considérons $F \in D(H\gamma)$, $\gamma \in \mathbb{R}$, $k_n \longrightarrow \infty$ et $k_n/n \longrightarrow 0$ quand $n \longrightarrow \infty$ alors :*

1. **Convergence en probabilité:**

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{D}} \xrightarrow{\mathbb{P}} \gamma \quad quand \quad n \longrightarrow \infty$$

2. **Convergence p.s:** *Si $k_n/(logn)^\delta \longrightarrow \infty$ quand $n \longrightarrow \infty$, pour $\delta > 0$ alors :*

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{D}} \xrightarrow{\mathrm{p.s}} \gamma \quad quand \quad n \longrightarrow \infty$$

3. **Normalité asymptotique, (voir théorème 3.1 et corollaire 3.2 de [13])**

$$\sqrt{\mathrm{k_n}}(\widehat{\gamma}_{\mathrm{n}}^{\mathrm{D}} - \gamma) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \eta^2)$$

*où*

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

#### 1.6.3.3 Estimateur de Pickands

L'estimateur de Pickands de l'IVE a été proposé en 1975 par James Pickands[31]. Cet estimateur a l'avantage d'être valable quel que soit le domaine d'attraction de la distribution F ($\gamma \in \mathbb{R}$)

**Définition 1.6.12.** *(Estimateur de Pickands)*

*Considérons $X_1, X_2, ..., X_n$ une suite de variables aléatoires iid de fonction de répartition ($F \in D(H\gamma)$), où $\gamma \in \mathbb{R}$. Soit $k_n$ une suite d'entiers tels que $1 < k_n < \lfloor \frac{n}{4} \rfloor$. L'estimateur de Pickands de l'IVE est défini par :*

$$\widehat{\gamma}_n^P := \frac{1}{log2} log\left(\frac{X_{n-k_n+1:n} - X_{n-2k_n+1:n}}{X_{n-2k_n+1:n} - X_{n-4k_n+1:n}}\right) \tag{1.64}$$

**Propriétés asymptotiques de l'estimateur de Pickands:**

**Proposition 12.** *Si les hypothèses de la définition ci-dessus sont satisfaites et que $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} k_n/n = 0$ alors on a:*

1. **Convergence en probabilité:** Pickands (1975, [31])*:*

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{P}} \xrightarrow{\mathbb{P}} \gamma \quad quand \quad n \longrightarrow \infty$$

2. **Convergence presque sûre:** Dekkers et de Haan (1989, [12])

*Si de plus $\lim_{n\to\infty} \dfrac{k_n}{loglogn} = \infty$ alors :*

$$\widehat{\gamma}_{\mathrm{n}}^{\mathrm{P}} \xrightarrow{\mathrm{p.s}} \gamma \quad quand \quad n \longrightarrow \infty$$

3. **Normalité asymptotique:** *Sous certaines hypothèses supplémentaires sur la suite $k_n$ et sur $F$ que l'on pourra consulter dans Dekkers et de Haan [12], on a:*

$$\sqrt{\mathrm{k_n}}(\widehat{\gamma}_{\mathrm{n}}^{\mathrm{P}} - \gamma) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \nu^2)$$

*avec* $\nu^2 = \dfrac{\gamma^2(2^{2\gamma+1}+1)}{(2(2^\gamma-1)log2)^2}$

Comme l'estimateur de Hill, cet estimateur est biaisé et le résultat sur la normalité asymptotique nous permet de donner un IC pour l'estimation. Par exemple pour un niveau de confiance de $(1-\alpha)\%$ on a :

$$\gamma \in \left]\widehat{\gamma}_n^P - t_{1-\alpha/2}\frac{\widetilde{\nu}}{\sqrt{k_n}},\ \widehat{\gamma}_n^P + t_{1-\alpha/2}\frac{\widetilde{\nu}}{\sqrt{k_n}}\right[$$

avec $t_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite. et $\widetilde{\nu}$ est donné par la relation suivante:

$$\widetilde{\nu} = \frac{\widehat{\gamma}_n^P}{2(2^{\widehat{\gamma}_n^P}-1)log2}\sqrt{2^{(2\widehat{\gamma}_n^P+1)}+1}$$

**Remarque 1.10.** Dans le langage R, la fonction pickandsplot() de package evmix [4] nous permet de visualer le graphe de $\widehat{\gamma}_n^P$ en fonction de $k_n$. La figure (1.8) correspond à un exemple d'application de cette fonction sur un échantillon de taille $n = 5000$ d'une loi de Pareto généralisée avec $\gamma = -2$.

[4] https://www.rdocumentation.org/packages/evmix/versions/2.12/topics/pickandsplot

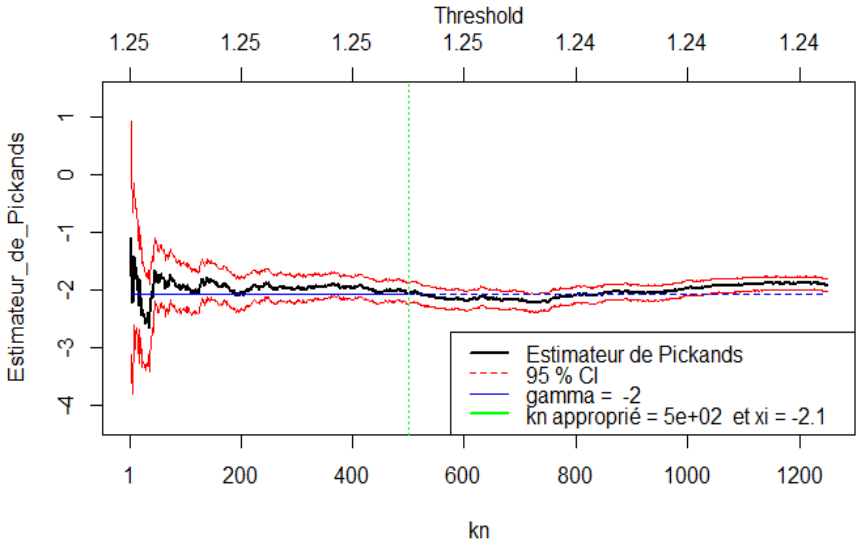



Figure 1.8: Graphique de l'estimateur de Pickands $\widehat{\gamma}_n^P$ en fonction de $k_n$ (trait noir) avec son intervalle de confiance à 95%, $n = 5000$ réalisations d'une loi GPD pour laquelle $\gamma = -2$. La ligne horizontale en blue représente la vraie valeur de $\gamma$. L'intersection de la ligne verticale (en vert) avec l'axe des $k_n$ représente la valeur appropriée de ce dernier et la valeur xi= $-2.1$ représente l'estimateur de Pickands de l'IVE $\gamma$ avec ce choix de $k_n$

**Définition 1.6.13.** *Soit $X$ une v.a de répartition $F_\theta = F(.,\theta)$ avec $\theta$ un paramètre inconnu. Considérons $T$ un estimateur de ce paramètre. On appelle écart quadratique moyen (Mean square error) de $T$ la quantité notée* MSE(T) *défnit par :*

$$\text{MSE(T)} = \mathbb{E}[(T-\theta)^2] = \mathbb{V}(T) + [\mathbb{B}(T)]^2 \tag{1.65}$$

*avec $\mathbb{B}(T)$ représente le biais de l'estimateur $T$ .*

**Remarque 1.11.**

Si $T_1$ et $T_2$ sont deux estimateurs de paramètre $\theta$. L'estimateur $T_1$ domine l'estimateur $T_2$ si:

$$\text{MSE}(\text{T}_1) < \text{MSE}(\text{T}_2)$$

En vertu de la définition (1.6.13) et la remarque (1.11), l'estimateur de Hill est plus efficace que l'estimateur de Pickands dans le cas d'une fonction F appartenant au domaine d'attraction de Fréchet, puisque asymptotiquement les deux estimateurs ont le même biais mais la variance de $\widehat{\gamma}_n^H$ est inférieure à la variance de $\widehat{\gamma}_n^P$ (voire la Figure (1.9))

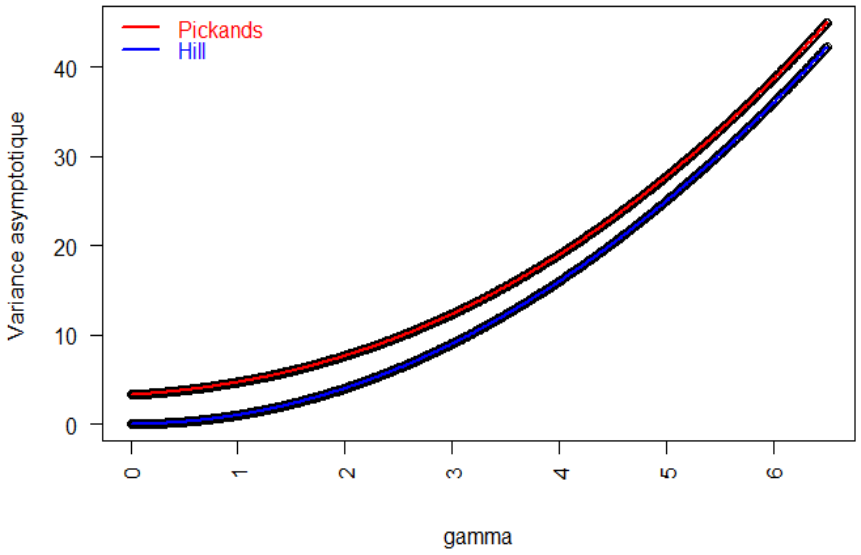



Figure 1.9: Variance asymptotique de l'estimateur de Hill $\widehat{\gamma}_n^H$ (en blue) et celle de l'estimateur de Pickands $\widehat{\gamma}_n^P$ (en rouge) en fonction de $\gamma$.

**Définition 1.6.14.** *Sous les conditions de la définition* (1.6.12)*, L'estimateur de Pickands pour un quantile extrême est donné par :*

$$\widehat{q}_n^P(p_n) := X_{n-k_n+1:n} + \frac{(\frac{k_n}{np_n})^{\widehat{\gamma}_n^P} - 1}{1 - 2^{-\widehat{\gamma}_n^P}} (X_{n-k_n+1:n} - X_{n-2k_n+1:n}) \tag{1.66}$$

*avec* $\widehat{\gamma}_n^P$ *est l'estimateur de Pickands de l'IVE.*

# *Chapter2*

# *Valeurs extrêmes sous données censurées*

**Résumé:**

Ce chapitre introduit le contexte des données censurées (tronquées). Après l'introduction 2.1, nous allons aborder les données incomplètes qu'on rencontre assez souvent dans l'analyse de survie. Puis, dans la partie (2.3) nous allons parler de la censure et de la troncature en présentant les différentes catégories et types de ces phénomènes avec illustrations par des exemples. Dans la partie (2.4), nous allons parler des extrêmes sous censure, en essayant d'adapter les notions de la TVE au cas de données censurées, en particulier: le quantile extrême et l'IVE. Dans la partie (2.5), nous illustrons le comportement des estimateurs sous censure par des exemples de simulation et application sur des données réelles.

## 2.1 Introduction

Le problème des données manquantes ou incomplètes est très vaste et a suscité beaucoup l'intérêt des statisticiens. L'attitude vis-à-vis de ce type de données a longtemps été soit de les éliminer, soit de minimiser l'effet qu'elles pourraient avoir sur des inférences statistiques adaptées à des données complètes. Ce type de problème est surtout rencontré dans le domaine de l'analyse de survie. En effet, dans la plupart des études prospectives, les données sont recueillies partiellement, notamment, à cause de deux phénomènes distincts: la censure et la troncature. Avant d'aborder ces deux phénomènes, nous allons rappeler quelques concepts de base de l'analyse des durées de survie.

## 2.2 Analyse des durées de survie

Dans la suite de se chapitre, pour simplifier l'exposé des notions relatives à l'étude des durées de survie, nous nous plaçons dans le cadre médicale, et donc nous notons par X le temps écoulé entre une date d'origine (par exemple l'initiation à un traitement) et la survenue de l'évènement observé ( évènement d'intérêt) qui est généralement le décès.

### 2.2.1 Données indispensables pour l'analyse de la survie

Les informations suivantes sont essentielles dans les études de survie :

1. **Date d'origine :** elle correspond à la date à laquelle a débuté l'observation. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude.
2. **Date des dernières nouvelles:** Elle correspond à la date la plus récente où des informations sur un sujet ont été recueillies. Autrement, c'est la date de décès pour les patients décédés ou la date à laquelle on dispose des dernières données relatives à l'état du patient sachant qu'il n'est pas décédé.
3. **un événement « en tout ou rien »(binaire):** Correspondant à la survenue ou non de l'événement à la date des dernières nouvelles.
4. **la date de point:** C'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

### 2.2.2 Fonctions d'intérêt

**Définition 2.2.1.** ***Fonction de répartition***
*La fonction de répartition de la v.a X (continue) est définie de* $\mathbb{R}$ *dans* $[0,1]$ *par :*

$$F(x) := \mathbb{P}(X \leq x) \tag{2.1}$$

$F(x)$ représente la probabilité de décèder avant le temps $x \geq 0$.

**Remarque 2.1.**

Si F est dérivable au temps x alors on note par f la fonction de densité de la v.a X telle que :

$$f(x) = \frac{dF(x)}{dx} := \lim_{dx \to 0} \frac{\mathbb{P}(x < X \leq x + dx)}{dx}$$

**Définition 2.2.2.** ***Fonction de survie***
*On appelle fonction de survie, la fonction notée S(.) ou $\overline{F}(.)$ définie par :*

$$S \ : \begin{array}{rl} \mathbb{R}^+ & \longrightarrow [0,1] \\ x & \longmapsto S(x) = \overline{F}(x) := 1 - F(x) = \mathbb{P}(X > x) \end{array} \tag{2.2}$$

c'est-à-dire pour x fixé, S représente la probabilté de vivre au-delà d'une date x.
Cette fonction définit une loi de probabilité sur $\mathbb{R}^+$ et :

$$\mathbb{P}(x_1 \leq X < x_2) = S(x_1) - S(x_2)$$

**Remarque 2.2.** S(x) est une fonction décroissante qui vaut 1 au temps d'origine et tend vers 0 quand t tend vers l'infini.

**Définition 2.2.3.** ***Taux instantané de défaillance h (ou taux de hasard)***
*Si X est une variable continue positive représentant une durée, on définit la fonction suivante:*

$$h(x) := \lim_{dx \to 0} \frac{\mathbb{P}(x < X < x + dx | X > x)}{dx} \tag{2.3}$$

appelée selon les domaines d'application : « taux instantané de défaillance », « fonction de hasard ») ou encore « quotient de mortalités »). Dans notre contexte, h(x) représente la probabilité pour qu'un sujet décède dans un laps de temps très petit ($dx \to 0$), conditionnement au fait que ce sujet était encore vivant jusqu'à l'instant x.
La fonction h vérifie les relations suivantes :

$$h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{d}{dx}\ log(S(x)). \tag{2.4}$$

L'équation (2.4) nous montre que la fonction h(x) caractérise la loi de X car on peut retrouver F(x) à partir de h(x):

$$h(x) = -\frac{d}{dx}\ log(1 - F(x))$$
$$F(x) = 1 - exp(-\int_0^x h(t)dt\ )$$

**Remarque 2.3.**

- Une fonction h(x) croissante est caractéristique d'un phénomène de vieillissement.
- Si $h(x) = c$ (c constante), il y a absence de vieillissement, le décès est dû à des causes aléatoires externes, car :
  X suit dans ce cas une loi exponentielle de répartition $F(x) = 1 - exp(-cx)$. Par conséquent, la fonction de survie est donnée par :

  $$S(x) = \mathbb{P}(X > x) = exp(-cx)$$

  Donc La probabilité conditionnelle de défaillance (ou de décès) entre $x_1$ et $x_2$ sachant que l'individu a déjà fonctionné (ou vécu ) jusqu'à l'instant $x_1$ est:

  $$\mathbb{P}(x_1 \leq X < x_2 | X > x_1) = \frac{S(x_1) - S(x_2)}{S(x_1)}$$

  Ce qui revient à dire que la probabilité conditionnelle de défaillance vaut:

  $$1 - exp(-c(x_2 - x_1)) = \mathbb{P}(X < x_2 - x_1)$$

  il n'y a pas de vieillissement: la probabilité de fonctionner pendant $x_2 - x_1$ à partir de la durée $x_1$ est la même qu'au démarrage. Ce modèle est couramment utilisé en électronique.

**Définition 2.2.4.** ***Fonction de risque cumulé*** *La fonction de risque cumulé* $\Phi$ *est définie par :*

$$\Phi(x) = \int_0^x h(u)du \tag{2.5}$$

$\Phi$ *est reliée à la fonction de survie par la relation:* $\Phi(x) = -log(S(x))$.

## 2.3 Données incomplètes

Cette partie est essentiellement inspirée de la thèse de Ndao [29] et l'article de Huber-Carol, (1994) [24]
Comme nous l'avons déjà mentionné dans l'introduction 2.1, la particularité des données de survie est qu'elles présentent des données incomplètes dûe aux phénomènes de la censure ou de troncature. Ces deux notions ont la signification suivante:

- **La troncature:** On n'observe X que si elle appartient à un sous-ensemble B de ses valeurs possibles. On dit que X est tronquée par B.
- **La censure:** Même dans le cas où X appartient à B, on n'observe pas X complètement ; on sait seulement de cette variable qu'elle appartient à un sous-ensemble A de B. Dans ce cas on dit que X est censurée par A.

Dans le cas où nous considérons un échantillon $X_1,\ X_2, ...,\ X_n$ de la variable durée de survie X, alors à chacune des $X_i$ sont associées deux ensembles $B_i$ et $A_i$. Le premier qui tronque $X_i$ et le second qui le censure. Généralement $B_i$ et $A_i$ sont des demi-droites du type $]-\infty,\ c_i]$ ou $[c_i,\ +\infty[$, ce qui correspond à des censures ou troncatures dites gauches ou droites.

### 2.3.1 Données censurées

Pour un individu (sujet) i, considérons :

- ▷ Son temps de survie $X_i$
- ▷ Son temps de censure $C_i$
- ▷ La durée réellement observée $T_i$

Il existe plusieurs catégories de modèle de censure, parmi eux, mentionnons les suivants:

- **Censure droite ou gauche:** Une durée de vie X est dite censurée par une variable aléatoire de censure C si on observe parfois C au lieu de X. L'information donnée par C sur X est :

$$\begin{array}{ll} X > C & \text{s'il y a censure à droite} \\ X < C & \text{s'il y a censure à gauche} \end{array}$$

**Exemple 5.** : Censure à droite

Un exemple classique de censure droite est celui où l'étude porte sur la durée de survie X de patients atteints d'une certaine maladie. Pour les patients perdus de vue (voir § 2.3.1 on page 40) au bout du temps C alors qu'ils étaient encore vivants, C censure X à droite puisque X est inconnue mais supérieure à C : $X > C$.

**Exemple 6.** : Censure à gauche

Un ethnologue étudie la durée d'apprentissage d'une tâche. Cette durée est une variable aléatoire X et C est l'âge de l'enfant. Pour les enfants qui savent déjà accomplir la tâche, C censure X à gauche car pour eux X est inconnu mais inférieur à C: $X < C$.

- **Censure par intervalle :** Si, au lieu de X, on observe $C_1 < C_2$ tels que $C_1 < X < C_2$ (X non observé), on dit qu'il y a censure par intervalle. En particulier, la censure à gauche peut être considérée comme une censure par un intervalle tel que $C_1 = -\infty$ $(C_1 = 0)$, et la censure à droite par un intervalle tel que $C_2 = +\infty$. La censure par intervalle est due au fait que les individus (sujets) sont observés de manière non continue. Par exemple, lors de suivi de patients (voir le cas $X_7$, Figure2.1). Le patient $X_7$ n'est observé que pendant les temps de visites $V_0, V_1, ..., V_N$. Si lors de la $i^{\text{ème}} + 1$ visite $(V_{i+1})$ on remarque l'apparition de l'évènement d'intérêt, alors la seule information qu'on a est que cet évènement s'est produit entre les deux temps de visites $V_i$ et $V_{i+1}$.

Les 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 :

- **Censure de type I**

  Soit C une valeur fixée, au lieu d'observer les variables aléatoires $X_1, X_2, ..., X_n$ qui nous intéressent, on n'observe $X_i$ uniquement lorsque $X_i < C$, sinon la seule information que nous avons est que $X_i > C$. Donc pour chaque individu (sujet) i, la variable aléatoire a observée est donnée par la relation suivante :

  $$T_i = X_i \wedge C = min(X_i, C)$$

  Ce type de censure est fréquent dans les aplications industrielles. Par exemple, on peut tester la durée de vie de n objets identiques (ampoules par exemple) sur un intervalle d'observations fixé $[0, C]$.

- **Censure de type II**

  L'expérimentateur fixe a priori le nombre d'événements à observer. Par conséquent, la date de fin d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce mécanisme de censure est souvent présent en fiabilité, épidémiologie... Par exemple, en épidémiologie, on décide d'observer les durées de survie de n patients jusqu'à ce que k $(k < n)$ d'entre eux soient décédés et on arrête l'étude à ce moment là.

Considérons alors $X_{i:n}$ et $T_{i:n}$ les statistiques d'ordre des variables aléatoires $X_i$ et $T_i$ (resp). La date de censure est donc $X_{k:n}$ et on observe les variables suivantes:

$$\begin{cases} T_{i:n} = X_{i:n} & \text{Si } \ i \leq k \\ T_{i:n} = X_{k:n} & \text{Si } \ i > k \end{cases}$$

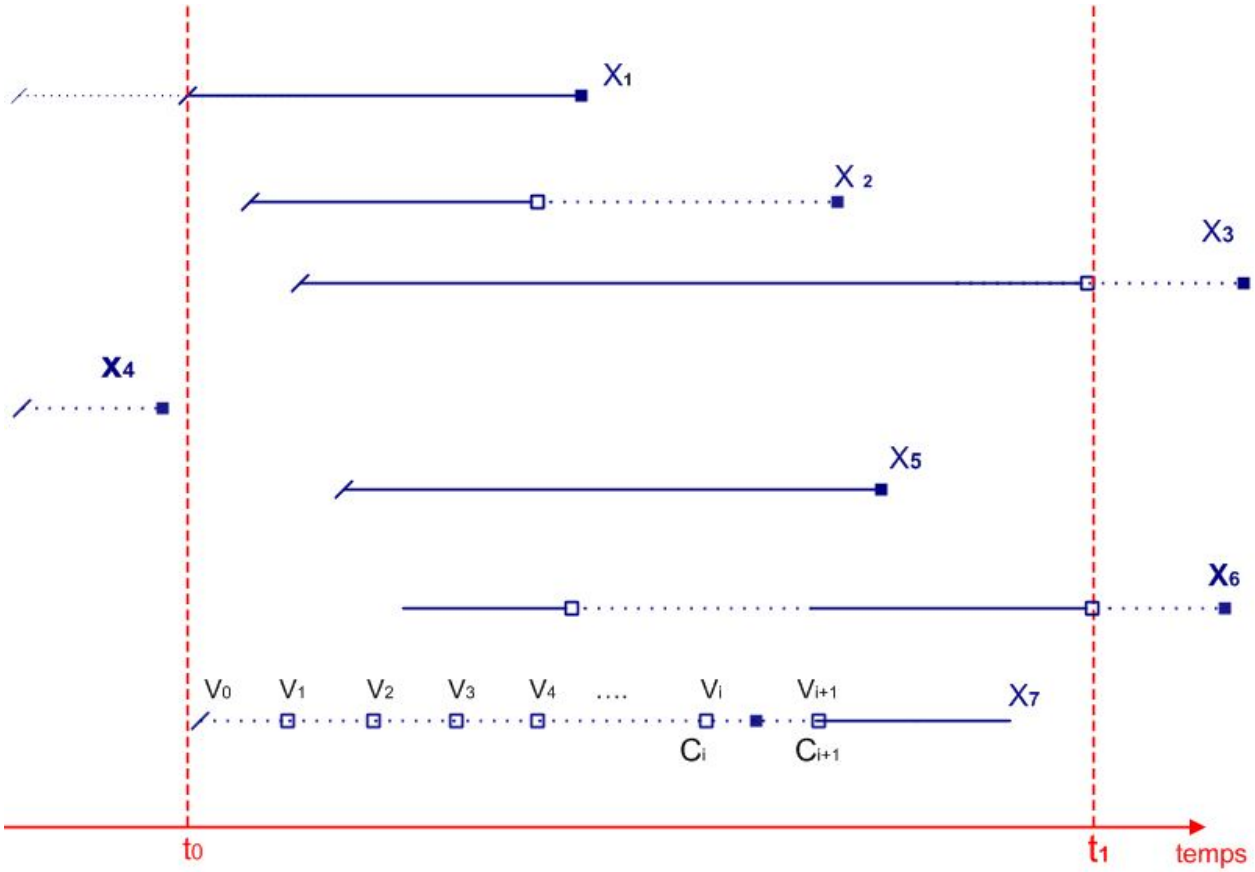



Figure 2.1: Illustration de la censure:
Les lignes pleines (resp. discontinues) représentent les durées observées (resp. non observées); les carrés pleins représentent l'évènement d'intérêt (décès) et les carrés vides représentent la valeur de la censure. $X_1 ... X_7$ sont un échantillon iid de la v.a durée de survie. L'étude a lieu pendant la durée $[t_0, t_1]$. $X_1$ et $X_5$ sont observées (non censurées); $X_2, X_3$ et $X_6$ sont censurées à droite; $X_4$ est censurée à gauche; et $X_7$ est censurée par l'intevalle $[C_i, C_{i+1}]$

- **Censure de type III** (Ou censure aléatoire de type I)
  Considérons $X_1, X_2, ..., X_n$ un échantillon de la v.a X (durée de survie) et soient $C_1, C_2, ..., C_n$ des variables aléatoires iid. Dans ce type de censure aléatoire, au lieu d'observer les $X_i$, on observe un couple de v.a $(T_i, \delta_i)$ avec :

$$\begin{cases} T_i &= X_i \wedge C_i \\ \\ \delta_i &= \mathbb{1}_{\{X_i<C_i\}} \end{cases} \qquad i = 1...n \tag{2.6}$$

  Donc l'information disponible peut être résumée par :

  - ▷ La durée réellement observée $T_i$.
  - ▷ Une v.a booléenne (indicateur) $\delta_i = \mathbb{1}_{\{X_i<C_i\}}$ avec:

    $\delta_i = 1$ si l'événement est observé (d'où $T_i = X_i$), autrement on observe des durées complètes (non censurées).

    $\delta_i = 0$ si l'individu i est censuré par $C_i$ (d'où $T_i = C_i$), dans ce cas on observe des durées non complètes (censurées).

  La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par:

  **(a)** La perte de vue: le patient quitte l'étude en cours (déménagement, soins dans un autre hôpital,...). Ce genre de patients est dit : "perdu de vue".

  **(b)** Arrêt ou changement de traitement: à cause des effets secondaires ou l'inefficacité du traitement. Ces patients sont des "exclus de l'étude".

  **(c)** La fin de l'étude: l'étude se termine alors que certains patients sont toujours vivants (non subit de l'événement d'intérêt). Ces patients sont dits " exclus vivants".

  Les "perdus de vue" (et les exclusions) et les "exclus vivants" correspondent à des observations censurées mais les deux mécanismes sont de nature différente.

**Remarque 2.4.**

1. L'analyse des données censurées nécessite une méthodologie adaptée permettant de prendre en compte l'information contenue dans le délai de censure. Les procédures usuelles qui tiennent compte de la censure supposent souvent l'indépendance des variables durées de survie et délais de censure. Cette hypothèse est très utile de point de vue statistique, et elle est indispensable aux modèles classiques d'analyse de survie. Dans la suite de ce mémoire, l'hypothèse d'indépendance entre $X_i$ et $C_i$ est supposée vérifiée. Cette hypothèse est raisonnable (naturelle) si la censure est causée par un déménagement ou par la fin de l'étude, elle est moins évidemment vérifiée quand les données censurées correspondent à des sujets qui sont "perdus de vue".

2. La censure à droite de type aléatoire (censure aléatoire à droite) est le cas le plus courant dans les études prospectives. Très peu de travaux s'intéressent à la censure à gauche ou par intervalle car elles sont moins fréquentes. C'est pour cette raison que nous considérons que le cas de censure aléatoire à droite dans toute la suite.

### 2.3.2 Données tonquées

Le second phénomène qui cause le problème des données incomplètes est la troncature. Généralement on dit qu'une variable aléatoire X est tronquée lorsque X n'est observée que s'elle est dans un sous-ensemble B de ces valeurs possible, B est généralement un intervalle de la forme $[Z_g, Z_d]$. Si X n'est pas dans cet intervalle, alors elle est non observée, et dans ce cas on perd complètement l'information X. La troncature est souvent confondue avec la censure, alors que ce phénomène diffère de celui de la censure. En effet, dans le cas de la censure, on a connaissance du fait qu'il existe une information, mais on ne connaît pas sa valeur précise, simplement le fait qu'elle excède un seuil ; dans le cas de la troncature on ne dispose pas de cette information. On peut rencontrer aussi d'autres types de troncature, dans le cas où $Z_d = +\infty$, on parle de troncature gauche et dans le cas où $Z_g = 0$ on dit qu'il y troncature droite.

- **Troncature gauche:** On dit qu'il y a troncature gauche, lorsque la variable d'intérêt X n'est observable que si elle est supérieure à Z. Z est alors la variable aléatoire de troncature gauche :

$$\text{X n'est observée que si:} \quad X > Z$$

- **Troncature droite:** On dit qu'il y a troncature droite, lorsque X n'est observable que si elle est inférieure à Z. Z est alors la variable aléatoire de troncature droite :

$$\text{X n'est observée que si:} \quad X < Z$$

- **Troncature par intervalle:** Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle. C'est-à-dire la v.a d'intérêt X n'est observée que si elle est dans l'intervalle $[Z_g, Z_d]$

$$\text{X n'est observée que si:} \quad Z_g \leq X \leq Z_d$$

**Exemple 7. Durée de vie après la retraite.**
On étudie la durée de vie après la retraite de sujets qui entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite. Un sujet n'est donc observé que si sa durée de vie après la retraite excède le délai entre sa prise de retraite et l'instant de l'enquête. La durée de vie après la retraite est donc tronquée à gauche par ce délai. Elle peut aussi être censurée à droite si la fin de l'enquête a lieu alors que le sujet est toujours vivant.

### 2.3.3 Estimateurs non-paramétrique des fonctions d'intérêts en présence de censure

#### 2.3.3.1 Estimateur de Kaplan-Meier (EKM)

Soit X la v.a d'intérêt (durée de survie) de répartition F. Considérons $X_1, X_2, ..., X_n$ un échantillon iid de la v.a X. Soit C la v.a de censure de répartition G, et considérons $C_1, C_2, ..., C_n$ un échantillon iid de la v.a C. Supposons que l'hypothèse d'indépendance entre X et C est vérifiée. Soit $\{(T_i, \delta_i) |\ \ 1 \leq i \leq n\}$ l'échantillon réellement observé (cf. équation (2.6)) et $\{(T_{i:n}, \delta_{[i:n]}) |\ \ 1 \leq i \leq n\}$ sa statistique d'ordre croissant.
Dans le cas de données non censurées, un estimateur naturel de la fonction de survie de la v.a d'intérêt X est la fonction de survie empirique :

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}} \tag{2.7}$$

Dans le cas de présence de censure aléatoire à droite, on n'observe plus X mais le couple $\{T_i, \delta_i\}$. L'estimateur de Kaplan-Meier (1958,[26]) de la fonction de survie S est donné par :

$$\widehat{S}_n^{KM}(x) = \begin{cases} 1 & \text{Si } \; x < T_{1:n} \\ \prod_{i=1}^{n} \left( \frac{n-i}{n-i+1} \right)^{\delta_{[i:n]} \mathbb{1}_{\{T_{i:n} \leq x\}}} = \prod_{i=1}^{n} \left( 1 - \frac{\delta_{[i:n]} \mathbb{1}_{\{T_{i:n} \leq x\}}}{n-i+1} \right) & \text{Sinon.} \end{cases} \tag{2.8}$$

**Remarque 2.5.**

1. Cet estimateur est une fonction en escalier décroissante.
2. Il n'atteint 0 que si la plus grande observation $X_{n:n}$ correspond à une observation non censurées.
3. En absence de censure, on retrouve l'estimateur de survie empirique (cf.(2.7))

#### 2.3.3.1.1 Propriétés asymptotiques de l'EKM

La normalité asymptotique de l'EKM est donnée dans le théorème suivant:

**Théorème 2.1.** *Droesbeke et Saporta* (2011,[14])

*Si les fonctions de répartitions de la survie et de la censure ( respectivement F et G) n'ont aucune discontinuité commune, alors on a :*

1. *Convergence uniforme:* $\underset{x \geq 1}{Sup} |\widehat{S}_n^{KM}(x) - S(x)| \xrightarrow{\text{p.s}} 0.$
2. *Normalité asymptotique: Pour tout* $x \geq 0$ *on a:*

$$\sqrt{\mathrm{n}}(\widehat{\mathrm{S}}_{\mathrm{n}}^{\mathrm{KM}}(\mathrm{x}) - \mathrm{S}(\mathrm{x})) \xrightarrow{\mathscr{L}} \mathrm{Z_x}$$

*où* $\{Z_x\}_{x \geq 0}$ *est un processus gaussien centré qui vérifie pour tous t et s strictement positifs*

$$Cov(Z_t, Z_s) = S(t)S(x) \int_0^{t \wedge s} \frac{dF(u)}{(1-F(u))^2(1-G(u))}$$

**Remarque 2.6.**

1. De la relation $\Phi(x) = -log(S(x))$ on peut dériver de l'EKM un estimateur du risque cumulé $\Phi$ : cet estimateur $\widehat{\Phi}_{Br} = -log(\widehat{S}_n^{KM}(x))$ est connu sous le nom d'estimateur de Breslow du risque cumulé.
2. Un estimateur plus connu du risque cumulé est celui de Nelson-Aalen (voir Nelson [30],Aalen [1]) défini par:

$$\widehat{\Phi}_{HN}(x) = \begin{cases} 0 & \text{Si } \; x < T_{1:n} \\ \sum_{i=1}^{n} \frac{\delta_{[i:n]} \mathbb{1}_{\{T_{i:n} \leq x\}}}{n-i+1} & \text{Sinon.} \end{cases} \tag{2.9}$$

3. On peut à partir de $\widehat{\Phi}_{HN}$ obtenir un autre estimateur de la fonction de survie S d'après la relation $S(x) = exp(-\Phi(x))$, connu sous le nom d'estimateur de Harrington et Fleming.

## 2.4 Les extrêmes sous censure

Le traitement des événements extrêmes sous censure est un problème du recherche très récent qui a attiré l'attention de nombreux chercheurs, en raison de ces applications dans divers domaines: assurance, l'analyse de survie, la fiabilité... Les premiers qui l'ont mentionné sont Reiss et Thomas (voir [33], Section 6.1) où ils ont proposé un estimateur de l'IVE dans le cas où celui-là est positif, mais sans donné les propriétés asymptotiques de cet estimateur. Beirland et al (2007,[3]); Einmahl et al (2008,[15]) ont proposé par la suite une méthode générale d'adaptation des estimateurs classiques de l'IVE ainsi des quantiles extrêmes au cas de données censurées aléatoirement à droite, ainsi que les propriétés asymtotiques.

### 2.4.1 Modèle pour les extrêmes sous censure

Soient X et C deux variables aléatoires de répartitions $F \in D(H_{\gamma_1})$ et $G \in D(H_{\gamma_2})$ où F, G sont absolument continues et $\gamma_1, \gamma_2 \in \mathbb{R}$.
Considérons $(X_i)_{1\leq i\leq n}$ et $(C_i)_{1\leq i\leq n}$ deux échantillons de variables iid de même loi que X et C respectivement. Soit $(T_i)_{1\leq i\leq n} = (X_i \wedge C_i)_{1\leq i\leq n}$ et $\delta_i = \mathbb{1}_{\{X_i\leq C_i\}}$. Soit H la répartition de la v.a $T = X \wedge C$. Ainsi, par hypothèse d'indépendance des v.a X et C, on a:

$$\begin{aligned} 1 - H(t) &= \mathbb{P}(T > t) \\ &= \mathbb{P}((X \wedge C) > t) \\ &= \mathbb{P}(X > t) \cap (C > t)) \\ &= \mathbb{P}(X > t)\ \mathbb{P}(C > t) \qquad \text{(Indépendance entre X et C)} \\ &= (1 - F(t))(1 - G(t)) \end{aligned} \tag{2.10}$$

Dans le cas de la censure aléatoire droite, au lieu d'observer l'échantillon $(X_i)_{1\leq i\leq n}$, nous observons l'échantillon $\{(T_i, \delta_i) \ \ |1 \leq i \leq n\}$. Soit $\{(T_{i:n}, \delta_{[i:n]}) \ \ |1 \leq i \leq n\}$ l'échantillon ordonné, où $T_{1:n}, T_{2:n}, ..., T_{n:n}$ sont les statistiques d'ordre associées à l'échantillon $T_1, T_2, ..., T_n$ et $\delta_{[1:n]}, \delta_{[2:n]}, ..., \delta_{[n:n]}$ sont les indicateurs correspondant aux statistiques d'ordre $T_{1:n}, T_{2:n}, ..., T_{n:n}$ respectivement.

### 2.4.2 Définitions des estimateurs

Considérons $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$ l'échantillon observé. L'objectif de cette partie est de déterminer un estimateur de l'IVE $\gamma_1$ de la fonction de répartition F et son quantile extrême associé. Soit $x_{\text{F}}$ (resp. $x_{\text{G}}$ et $x_{\text{H}}$) les points terminaux des fonctions F (reps. G et H). L'IVE de H existe et il est noté $\gamma$ et selon le signe de $\gamma_1$ et $\gamma_2$ Einmahl et al (2008,[15]) ont distingué trois cas intéressants :

$$\begin{cases} 1^{\text{èr}} \text{ cas} & : \ \gamma_1 > 0,\ \gamma_2 > 0 & & \text{F et G} \in \text{DA(Fréchet)} & \gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2} \\ 2^{\text{ème}} \text{ cas} & : \ \gamma_1 < 0,\ \gamma_2 < 0 & x_{\text{F}} = x_{\text{G}} & \text{F et G} \in \text{DA(Weibull)} & \gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2} \\ 3^{\text{ème}} \text{ cas} & : \ \gamma_1 = \gamma_2 = 0 & x_{\text{F}} = x_{\text{G}} = +\infty & \text{F et G} \in \text{DA( Gumbel)} & \gamma = 0 \end{cases} \tag{2.11}$$

**Remarque 2.7.**

- ▷ Dans le troisième cas, nous difinissons également pour une présentation commode:

$$\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} = 0$$

- ▷ Les autres cas, $\{\gamma_1 > 0,\ \gamma_2 < 0\}$ et $\{\gamma_1 < 0,\ \gamma_2 > 0\}$, ne sont pas intéressants. Le premier correspond au cas où les données sont complètement non censurées (absence de ccensure), qui est déja traité dans le chapitre 01; le dernier correspond au cas où les données sont complètement censurées, ce qui rend l'estimation de l'IVE impossible.

La méthode la plus générale utilisée pour l'estimation de $\gamma_1$ est apparue en premier avec Beirland et al (2007,[3]), et développée par la suite par Einmahl et al (2008,[15]). Cette méthode consiste à adapter les estimateurs standards (Hill, estimateur des moments, Pickands... (cf.(1.6))) au cas des données censurées. Elle se base sur le fait que $\gamma_1$ s'écrit :

$$\gamma_1 = \frac{\gamma}{p} \quad \text{avec} \quad p = \frac{\gamma_2}{\gamma_1+\gamma_2} \tag{2.12}$$

**Remarque 2.8.** (Voir Einmahl et al (2008,[15])

p représente la proportion de non-censure, elle est définie par :

$$\begin{aligned} p \equiv p(t) &:= \mathbb{P}(X \le C | T = t) \\ &= \frac{\mathbb{P}(X \le C, T = t)}{\mathbb{P}(T = t)} \quad \text{avec} \quad T = min(X, C) \\ &= \frac{\overline{G}(t) f(t)}{\mathtt{h_T}(t)} \quad \text{car X et C sont indépendantes} \\ &= \frac{(1-G(t))f(t)}{(1-G(t))f(t) + (1-F(t))g(t)} \end{aligned} \tag{2.13}$$

avec f(.), g(.) et $\mathtt{h_T}(.)$ sont les densités asociées à F, G et H, respectivement.
Notons que dans le premier et le deuxième cas de l'équation (2.11), $\lim\limits_{t\to\infty} p(t)$ existe et elle est donnée par :

$$\lim_{t\to\infty} p(t) = \frac{\gamma_2}{\gamma_1+\gamma_2} =: p \in [0,1] \tag{2.14}$$

En effet; d'après le résultat du théorème (1.3), si nous posons: $Y = T - \mu$ avec $\mu$ un seuil suffisamment grand, alors la loi des excès de F (resp. G) notée $F_\mu$ (resp. $G_\mu$ ), peut être approchée par une loi de Paréto généralisée.
En remplaçant $F_\mu$ et $G_\mu$ par leurs expressions, nous obtenons :

$$1 - H_\mu(y) = (1 - F_\mu(y))(1 - G_\mu(y)) \tag{2.15}$$

$$= (1+\gamma_1 y)^{\frac{-1}{\gamma_1}} (1+\gamma_2 y)^{\frac{-1}{\gamma_2}} \tag{2.16}$$

En dérivant $F_\mu$ (resp. $G_\mu$ et $H_\mu$) pour obtenir $f_\mu(.)$ (resp. $g_\mu(.)$ et $\mathtt{h_u}(.)$), et en remplaçant ces fonctions par leurs expressions dans l'équation (2.13), il vient:

$$\begin{aligned}
\lim_{t\to\infty} p(t) &:= \lim_{t-\mu\to\infty} \frac{\overline{G}_\mu(t-\mu) f_\mu(t-\mu)}{\mathtt{h_u}(t-\mu)} \\
&= \lim_{y\to\infty} \frac{\overline{G}_\mu(y) f_\mu(y)}{\mathtt{h_u}(y)} \\
&= \lim_{y\to\infty} \frac{\overline{G}_\mu(y) f_\mu(y)}{\overline{G}_\mu(y) f_\mu(y) + \overline{F}_\mu(y) g_\mu(y)} \\
&= \lim_{y\to\infty} \frac{1}{1 + \dfrac{\overline{F}_\mu(y) g_\mu(y)}{\overline{G}_\mu(y) f_\mu(y)}} \\
&= \lim_{y\to\infty} \frac{1}{1+\beta} \quad \text{avec } \beta = \frac{\overline{F}_\mu(y) g_\mu(y)}{\overline{G}_\mu(y) f_\mu(y)} \longrightarrow \frac{\gamma_1}{\gamma_2} \qquad (2.17)
\end{aligned}$$

(2.18)

Ce qui nous donne : $\lim_{t\to\infty} p(t) = \frac{\gamma_2}{\gamma_2+\gamma_1} =: p$

Pour avoir l'estimateur adapté de $\gamma_1$ nous allons :

1. Estimer $\gamma$ avec l'un des estimateurs standards donné dans le chapitre 1 (cf.(1.6)) ,(en particulier les estimateurs non-paramétriques: l'estimateur des moments, Hill, Pickands (cf.(1.6.3))), obtenu à partir de l'échantillon $T_1, T_2, ..., T_n$. Nous notons par $\widehat{\gamma}_n^\bullet$ cet estimateur.

2. Estimer p par la proportion des données non censurées des $k_n$ plus grandes valeurs de X. Notons par $\widehat{p}$ cet estimateur qu'est donné par la relation suivante :

$$\widehat{p} = \frac{1}{k_n} \sum_{i=1}^{k_n} \delta_{[n-i+1:n]} \qquad (2.19)$$

**Définition 2.4.1.** *L'estimateur de $\gamma_1$ adapté à la censure est noté par $\widehat{\gamma}_1^{(c,\bullet)}$ et définit par :*

$$\widehat{\gamma}_1^{(c,\bullet)} = \frac{\widehat{\gamma}_n^\bullet}{\widehat{p}} \qquad (2.20)$$

**Exemple 8.**

L'estimateur de Hill adapté à la censure aléatoire droite est donné par :

$$\widehat{\gamma}_1^{(c,H)} = \frac{\widehat{\gamma}_n^H}{\widehat{p}} = \frac{\frac{1}{k_n}\sum_{i=1}^{k_n} log(T_{n-i+1:n}) - log(T_{n-k_n:n})}{\frac{1}{k_n}\sum_{i=1}^{k_n} \delta_{[n-i+1:n]}}$$

Pour voir la différence entre l'estimateur de Hill $\widehat{\gamma}_n^H$ et l'estimateur de Hill adapté à la censure aléatoire droite $\widehat{\gamma}_1^{(c,H)}$, nous allons simuler deux échantillons de même taille $n = 1000$, un échantillon de la variable d'intérêt X de loi GPD avec $\gamma_1 = 2$ et $\sigma_1 = 1$ qui sera censuré par un autre échantillon C (de la variable de censure) de loi GPD du paramètres $\gamma_2 = 1$ et $\sigma_2 = 1$. Sur la figure (2.2), nous allons tracer le graphe de ces deux estimateurs en fonction de $k_n$.

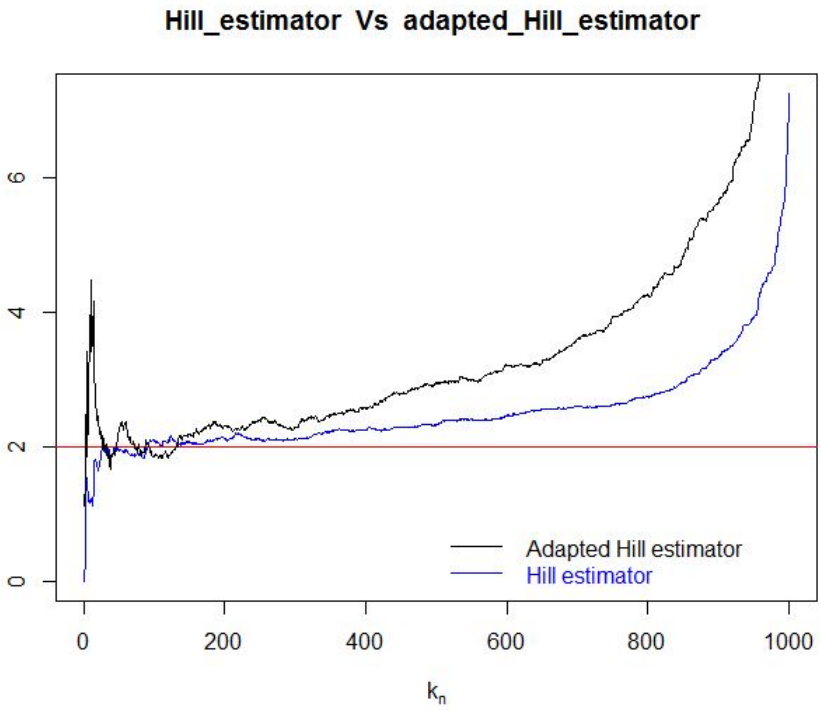



Figure 2.2: Graphe de l'estimateur de Hill adapté (et non) à la censure en fonction de nombre de statistiques d'ordre $k_n$. La ligne horizontale en rouge représente la vraie valeur de $\gamma_1$.

#### 2.4.2.1 Normalité asymptotique des estimateurs adaptés $\widehat{\gamma}_1^{(c,\bullet)}$

Le théorème suivant énoncé par Einmahl et al (2008, [15]) décrit la normalité asymptotique de l'IVE $\widehat{\gamma}_1^{(c,\bullet)}$.

**Théorème 2.2.** *Pour $n \longrightarrow \infty$ et sous les hypothèses:*

**(H1):** $\sqrt{k_n}A(n/k_n) \longrightarrow \lambda \in \mathbb{R}$

**(H2):** $\dfrac{1}{k_n}\displaystyle\sum_{i=1}^{k_n}[p(H^{\leftarrow}(1-\frac{i}{n})) - p] \longrightarrow \alpha \in \mathbb{R}$

**(H3):** *pour* $1 - \dfrac{k_n}{n} < t < 1$ *et* $|t-s| \leq c\,\dfrac{k_n}{n}$, *telle que:* $s < 1$, $c > 0$ *et:*

$$\sqrt{k_n}\underset{t,s}{\mathrm{Sup}}|p(H^{\leftarrow}(t)) - p(H^{\leftarrow}(s))| \longrightarrow 0$$

*On a:*

$$\sqrt{\mathrm{k_n}}(\widehat{\gamma}_1^{(\mathrm{c},\bullet)} - \gamma_1) \xrightarrow{\mathscr{L}} \mathscr{N}\left(\frac{1}{p}(\lambda b_0 - \gamma_1\alpha),\ \frac{\sigma^2 + \gamma_1^2 p(1-p)}{p^2}\right)$$

*où $\lambda b_0$ (resp. $\sigma^2$) représente la moyenne (resp. la variance) de $\sqrt{k_n}(\widehat{\gamma}_1^{(c,\bullet)} - \gamma_1)$.*

Pour la démonstration de ce théorème voir Einmahl et al (2008, [15]).

### 2.4.3 Quantile extrême sous données censurées

Nous souhaitons estimer le quantile extrême d'ordre $(1-p_n)$ dans le cas de données censurées. Pour ce faire, nous allons adapter les estimateurs standards qui existent dans la littérature de la TVE.

**Définition 2.4.2.** *(Einmahl et al [15])*
*L'estimateur de quantile extrême d'ordre* $(1-p_n)$ *sous censure aléatoire droite est donnée par :*

$$\widehat{q}_{(c,\bullet)}(p_n) := T_{n-k_n:n} + \widehat{a}^{(c,\bullet)} \frac{\left(\frac{1-\widehat{F}_n^{KM}(T_{n-k_n:n})}{p_n}\right)^{\widehat{\gamma}_1^{(c,\bullet)}} - 1}{\widehat{\gamma}_1^{(c,\bullet)}} \tag{2.21}$$

*avec:*

$$\widehat{F}_n^{KM} = 1 - \widehat{S}_n^{KM} \quad \text{où } \widehat{S}_n^{KM} \text{ est l'estimateur de Kaplan-Meier (cf.(2.8)).}$$

$$\widehat{a}^{(c,\bullet)} = \frac{T_{n-k_n:n} M_{Z,n}^{(1)} (1-S_{Z,n})}{\widehat{p}} \quad \text{où } M_{Z,n}^{(1)} \text{ et } S_{Z,n} \text{ sont définis dans l'équation (1.61)}$$

Pour voir la différence entre l'estimateur d'un quantile extrême avec et sans censure, nous traçons le graphe de l'estimateur de Weissman (cf.(1.6.9)) à partir des données 'Aids2' (voir 2.5.2), pour un $p_n = 10^{-3}$. D'après les résultats graphique (figure 2.3), on constate que pour un $k_n$ entre 300 et 400, la différence entre les deux estimateurs est environ 10 ans.

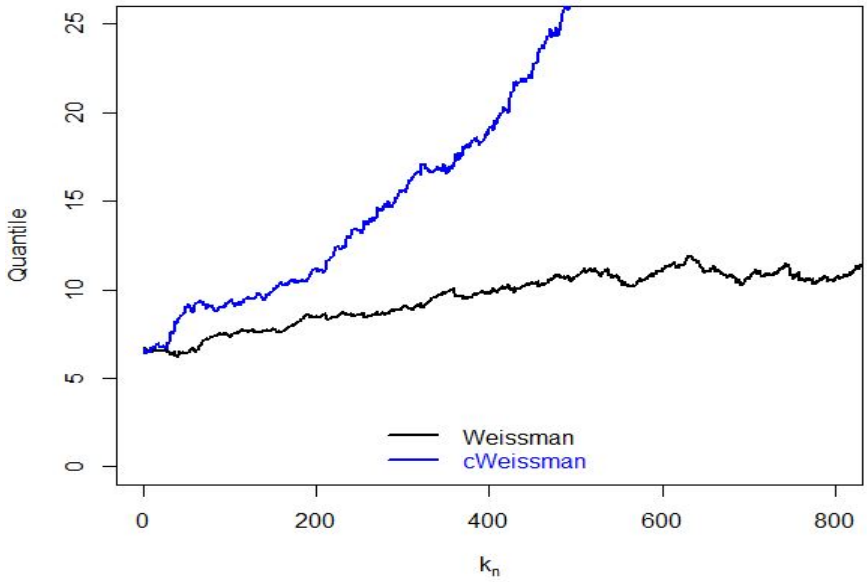



Figure 2.3: Estimation de Weissman d'un quantile extrême à partir des données 'Aids2'. Le graphe en blue c'est dans le cas de la censure, le graphe en noir (absence de censure).

### 2.4.4 Estimateur du maximum de vraisemblance de l'IVE en présence de données censurées

Dans ce paragraphe nous allons adapter l'estimateur du maximum de vraisemblance de l'IVE, du quantile extrême, au cas de données censurées aléatoirement à droite en se plaçant dans

l'approche par dépassement du seuil. Considérons $X_1, X_2, ..., X_n$ un échantillon iid de la v.a d'intérêt X de loi $F \in D(H_{\gamma_1})$. Fixons un seuil $\mu < x_{\mathbb{F}}$. Soit $Y_1, Y_2, ..., Y_{k_n}$ l'échantillon des excès, avec $\{Y_j;\ j = 1...k_n\}$ est défini dans l'équation (cf.(1.13)) et $k_n$ représente le nombre d'excès au-delà du seuil $\mu$.
D'après le théorème de Balkema−De Haan−Pickands (cf.(1.3)) la loi des excès notée $F_\mu(y)$ peut être approchée par une loi de Pareto généralisée $G_{\gamma_1,\sigma_1}(y)$ et dans ce cas (absence de censure), la vraisemblance s'écrit sous la forme suivante :

$$\mathscr{L}(\gamma_1, \sigma_1; y) = \prod_{j=1}^{k_n} g_{\gamma_1,\sigma_1}(y_j) \quad \text{avec } y = y_1, y_2, ..., y_{k_n}. \quad \text{(cf.(1.50))}$$

Dans le cas de données censurées aléatoirement à droite, au lieu d'observer l'échantillon $(X_i)_{1\leq i\leq n}$ nous observons des copies iid $(T_1, \delta_1), ..., (T_n, \delta_n)$ de couple aléatoire $(T, \delta)$. Nous définissons alors les excès au-delà du seuil $\mu$ ,adaptées à la censure, comme suit:

$$Y_j = T_j - \mu \quad \text{avec } j = 1...k_n$$

L'écriture de la vraisemblance adapté à la censure aléatoire droite est donnée par l'expression suivante, ( voir [3]):

$$\widetilde{\mathscr{L}}(\gamma_1, \sigma_1; y) = \prod_{j=1}^{k_n} [g_{\gamma_1,\sigma_1}(y_j)]^{\delta_j} [1 - G_{\gamma_1,\sigma_1}(y_j)]^{1-\delta_j} \tag{2.22}$$

avec $y = y_1, y_2, ..., y_{k_n}$; $G_{\gamma_1,\sigma_1}(.)$ représente la répartition de la GPD (cf.(1.18)) et $g_{\gamma_1,\sigma_1}(.)$ représente la densité associée à cette loi (cf.(1.20)).
Ainsi, en remplaçant $G_{\gamma_1,\sigma_1}$ et $g_{\gamma_1,\sigma_1}$ par leurs expressions dans l'équation (2.22) il vient:

$$\widetilde{\mathscr{L}}(\gamma_1, \sigma_1; y) = \prod_{j=1}^{k_n} \left[\frac{1}{\sigma_1}\left(1 + \gamma_1 \frac{y_j}{\sigma_1}\right)^{-\frac{1}{\gamma_1}-1}\right]^{\delta_j} \left[\left(1 + \gamma_1 \frac{y_j}{\sigma_1}\right)^{-\frac{1}{\gamma_1}}\right]^{1-\delta_j} \tag{2.23}$$

Maximiser $\widetilde{\mathscr{L}}$ revient à trouver les valeurs de $\gamma_1$ et $\sigma_1$ qui maximisent la fonction log-vraisemblance suivante:

$$log\left(\widetilde{\mathscr{L}}(\gamma_1, \sigma_1; y)\right) = \sum_{j=1}^{k_n} \delta_j \left[-log\sigma_1 - \left(1 + \frac{1}{\gamma_1}\right) log\left(1 + \frac{\gamma_1}{\sigma_1} y_j\right)\right] - \sum_{j=1}^{k_n} \frac{1}{\gamma_1}(1 + \delta_j)\, log\left(1 + \frac{\gamma_1}{\sigma_1} y_j\right) \tag{2.24}$$

En dérivant l'équation précédente par rapport aux deux paramètres $\gamma_1$ et $\sigma_1$, nous obtenons le système à deux équations suivant:

$$\begin{cases} \widetilde{\mathscr{L}}'_1(\gamma_1, \sigma_1) = \frac{\partial log\left(\widetilde{\mathscr{L}}(\gamma_1,\sigma_1;y)\right)}{\partial \gamma_1} = \frac{1}{\gamma_1^2} \sum_{j=1}^{k_n} log\left(1 + \frac{\gamma_1}{\sigma_1} y_j\right) - \frac{1}{\gamma_1} \sum_{j=1}^{k_n} \left(\frac{1}{\gamma_1} + \delta_j\right) \frac{\gamma_1 \frac{y_j}{\sigma_1}}{1 + \gamma_1 \frac{y_j}{\sigma_1}} \\ \widetilde{\mathscr{L}}'_2(\gamma_1, \sigma_1) = \frac{\partial log\left(\widetilde{\mathscr{L}}(\gamma_1,\sigma_1;y)\right)}{\partial \sigma_1} = -\frac{1}{\sigma_1} \sum_{j=1}^{k_n} \delta_j + \frac{1}{\sigma_1} \sum_{j=1}^{k_n} \left(\frac{1}{\gamma_1} + \delta_j\right) \frac{\gamma_1 \frac{y_j}{\sigma_1}}{1 + \gamma_1 \frac{y_j}{\sigma_1}} \end{cases} \tag{2.25}$$

Dans le cas où $\gamma_1 = 0$, le système précédent s'écrit comme suit:

$$\begin{cases} \widetilde{\mathscr{L}}'_1(0, \sigma_1) = \dfrac{\partial log\left(\widetilde{\mathscr{L}}(0, \sigma_1; y)\right)}{\partial \gamma_1} = -\dfrac{1}{2} \displaystyle\sum_{j=1}^{k_n} \left(\frac{y_j}{\sigma_1}\right)^2 - \sum_{j=1}^{k_n} \delta_j \frac{y_j}{\sigma_1} \\ \widetilde{\mathscr{L}}'_2(0, \sigma_1) = \dfrac{\partial log\left(\widetilde{\mathscr{L}}(0, \sigma_1; y)\right)}{\partial \sigma_1} = -\dfrac{1}{\sigma_1} \displaystyle\sum_{j=1}^{k_n} \delta_j + \frac{1}{\sigma_1^2} \sum_{j=1}^{k_n} y_j \end{cases} \tag{2.26}$$

La solution du système (2.25) n'est pas explicite, elle est déterminée par des méthodes numériques. Sous R, la fonction cGPDmle (ou cPOT) du package ReIns [1] nous permet de déterminer un estimateur du maximum de vraisemblance adapté à la censure aléatoire droit pour l'IVE $\gamma_1$ ainsi $\sigma_1$, dans la suite nous notons ces estimateurs par $\widehat{\gamma_1}^{(c,\mathsf{GPDmle})}$ (resp. $\widehat{\sigma_1}^{(c,\mathsf{GPDmle})}$).

**Définition 2.4.3.** *(Quantile extrême basé sur l'approche POT en présence de censure [3]) L'estimateur du quantile extrême d'ordre $1-p_n$ ($p_n \to 0$ quand $n \to \infty$) basé sur l'approche POT dans le cas de données censurées aléatoirement à droite est donnée par :*

$$\widehat{q}_{\mathsf{cPOT}}(p_n) = \mu + \widehat{\sigma_1}^{(c,\mathsf{GPDmle})} \frac{\left(\frac{\widehat{S}_n^{KM}(\mu)}{p_n}\right)^{\widehat{\gamma_1}^{(c,\mathsf{GPDmle})}} - 1}{\widehat{\gamma_1}^{(c,\mathsf{GPDmle})}} \tag{2.27}$$

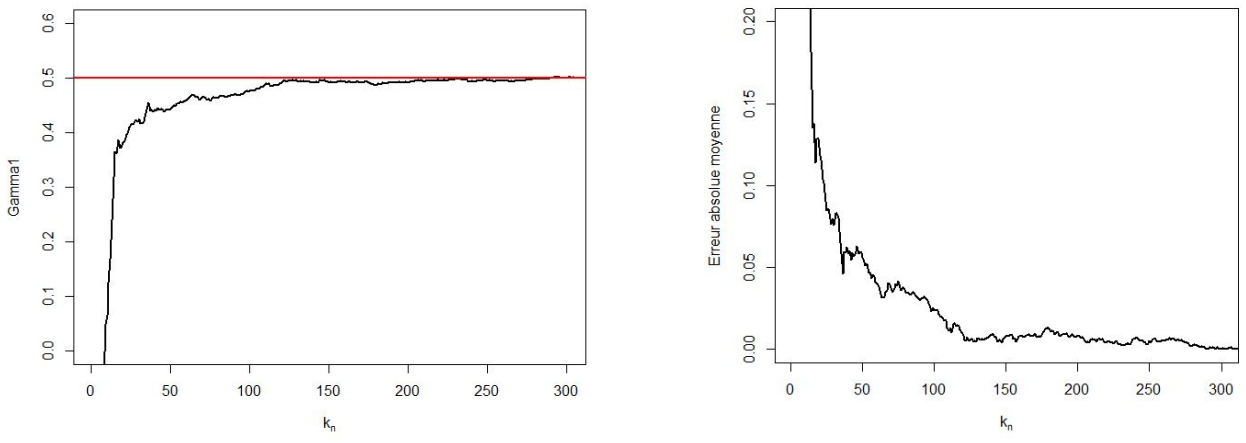
*avec:*

- ▷ *Le seuil $\mu$ peut être estimé par la statistique d'odre $T_{n-k_n+1:n}$.*
- ▷ *$\widehat{S}_n^{KM}$ est l'estimateur de Kaplan-Meier de la fonction de survie $S = 1 - F$* (cf.(2.8))

**Exemple 9.** Dans cet exemple, nous allons appliquer la méthode de maximum de vraisemblance pour estimer l'IVE de la loi de Pareto standard dans le cas de données censurées aléatoirement à droite. Pour cela, on génère $n = 100$ échantillons de taille $m = 500$ de la v.a X de loi Pareto standard de paramètre ($\alpha = 2$ donc $\gamma_1 = 0.5$) censurée par une autre v.a de même loi avec $\gamma_2 = 2$. La figure (2.4) montre le graphe de la moyenne des ces estimateurs $\widehat{\gamma_1}^{(c,\mathsf{GPDmle})}$ en fonction $k_n$ (2.4(a)) et l'erreur absolue moyenne (2.4(b)) noté MAE, défini par:

$$\mathrm{MAE}(\theta) = \mathrm{abs}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_i - \theta\right)\right) \tag{2.28}$$

avec n est le nombre d'échantillons et $\widehat{\theta}_i$ est le ième estimateur de $\theta$ obtenu à partir de ième échantillon.




(a) (b)

Figure 2.4: Estimation de l'IVE pour une loi de Pareto standard avec la méthode de maximum de vraisemblance. La ligne horizontale en rouge représente la vraie valeur de $\gamma$.

[1] https://cran.r-project.org/web/packages/ReIns/ReIns.pdf

## 2.5 Simulation et illustration sur des données réelles

### 2.5.1 Simulation

Dans cette partie, nous allons essayer d'illustrer:

- L'effet du pourcentage de censure sur les estimations.
- La performance des estimateurs les plus fréquents de l'IVE, (resp. quantile extrême), adaptés à la censure aléatoire droite, en traçant les graphes de ces estimateurs en fonction du nombre de statistiques d'ordre $k_n$, i.e.:

$$k_n \mapsto f(k_n) = \widehat{\gamma}_1^{(c,\bullet)} \quad \text{et} \quad k_n \mapsto g(k_n) = \widehat{q}_{(c,\bullet)} \tag{2.29}$$

$$\text{avec } \widehat{\gamma}_1^{(c,\bullet)} \in \Omega_1 = \{\widehat{\gamma}_1^{(c,H)}, \widehat{\gamma}_1^{(c,D)}, \widehat{\gamma}_1^{(c,\mathsf{GPDmle})}\} \qquad \text{et} \qquad \widehat{q}_{(c,\bullet)} \in \Omega_2 = \{\widehat{q}_{(c,W)}, \widehat{q}_{(c,D)}, \widehat{q}_{\mathsf{cPOT}}\} \tag{2.30}$$

Poue cela, on génère $s = 100$ échantillons de couple aléatoire $\{(X_i, C_i), \quad i = 1...t\}$ de taille $t = 500$, à partir de distributions données dans le tableau (2.1). Avec X représente la v.a d'intérêt d'IVE $\gamma_1$ et C est la v.a de censure d'IVE $\gamma_2$. $\gamma_2$ est choisi selon le pourcentage de censure:

$$\gamma_2 = \begin{cases} 2 & \text{Si le \% de censure est faible, } (1-p)\% = 11\% \\ 7/12 & \text{Si le \% de censure est moyen, } (1-p)\% = 30\% \\ 1/8 & \text{Si le \% de censure est fort, } (1-p)\% = 67\% \end{cases} \tag{2.31}$$

| **Distribution** | $1 - F(z)$ | $\gamma$ |
|---|---|---|
| Burr$(\beta, \tau, \lambda)$ | $\left(\frac{\beta}{\beta + z^\tau}\right)^\lambda, \quad z > 0; \ \beta, \tau, \lambda > 0$ | $\frac{1}{\tau\lambda}$ |
| Exp$(\lambda)$ | $exp^{-\lambda z}, \quad z > 0; \ \lambda > 0$ | $\gamma = 0$ |

Table 2.1: Distributions

▷ **Premier cas :** $X \hookrightarrow$ Burr avec $\gamma_1 > 0$, censurée par $C \hookrightarrow$ Burr avec $\gamma_2 > 0$.

1. **Indice extrême:** La figure (2.5) illustre l'influence de pourcentage de la censure sur le comportement empirique des estimateurs $\widehat{\gamma}_1^{(c,\bullet)} \in \Omega_1$, en se basant sur $s = 100$ échantillons de taille 500. Le panneau gauche (figure (2.5)), représente la moyenne empirique de ces estimateurs pour différent niveau de censure (resp. 11%, 30% et 67%). Tandis que, le panneau droit, représente la racine carrée de l'erreur quadratique moyenne de chaque estimateur, ( qu'on note RMSE(.)), défini par :

$$\text{RMSE}(\widehat{\gamma}_1^{(c,\bullet)}) := \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(\widehat{\gamma}_1^{(c,\bullet)} - \gamma_1\right)^2} \tag{2.32}$$

avec $\widehat{\gamma}_1^{(c,\bullet)} \in \Omega_1$, s: le nombre d'échantillons simulés ( ici $s = 100$) et $\gamma_1$ représente la vraie valeur de $\gamma_1$.

D'après les résultats graphiques de la simulation (figure (2.5)), nous constatons que:

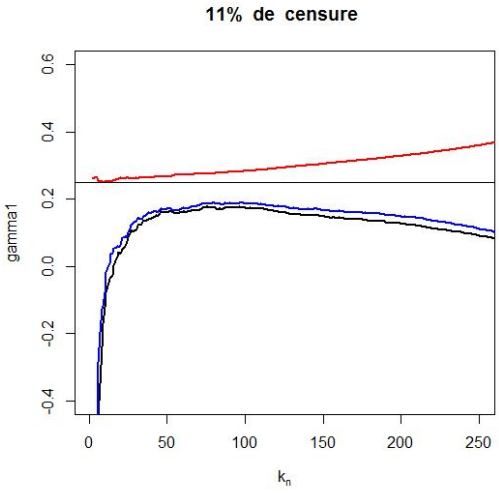
- L'estimateur de Hill adapté $\widehat{\gamma}_1^{(c,H)}$ semble être de meilleure qualité en terme de RMSE et d'erreur absolue moyenne (MAE) que les deux autres estimateurs $\widehat{\gamma}_1^{(c,D)}$ et $\widehat{\gamma}_1^{(c,\mathsf{GPDmle})}$.
- Au fure et à mesure que l'on augumente le pourcentage de censure (de 11% à 67%), l'adéquation des estimateurs se dégradent en terme de MAE et RMSE, ce qui est traduit par l'éloignement des courbes des estimateurs de l'IVE de la vraie valeur de $\gamma_1 = 0.25$ et le RMSE des estimateurs qui devient de plus en plus important.

2. **Quantile extrême:** La figure (2.6) illustre l'influence du pourcentage de la censure sur le comportement empirique des quantiles extrêmes $\widehat{q}_{(c,\bullet)}$. La ligne horizontale en rouge sur les graphes de panneau gauche, représente la vraie valeur du quantile d'ordre $p_n = 1/50$ issu de la loi de Burr$(10, 4, 1)$, ( $q(1/50) = Burr^{\leftarrow}(1 - 1/50) = 4.704885$, (cf. (1.29)).
D'après les résultats graphiques de la simulation (2.6), nous constatons que :

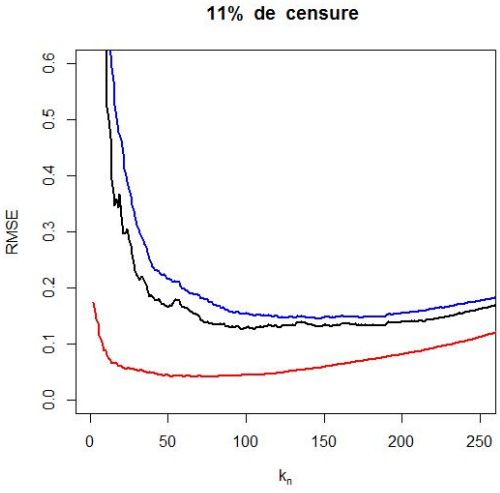
   - Les deux estimateurs $\widehat{q}_{(c,D)}$ et $\widehat{q}_{\mathsf{cPOT}}$ sont plus performants que l'estimateur de Weissman adapté à la censure $(\widehat{q}_{(c,W)})$.
   - Au fure et à mesure que le pourcentage de censure augmente, l'adéquation des estimateurs se dégradent en terme de MAE et RMSE: les courbes des estimateurs s'éloignent de la vraie valeur de $q(1/50)$ et le RMSE devient de plus en plus important.

▷ **Deuxième cas :** $X \hookrightarrow \text{Exp}(1)$ avec $\gamma_1 = 0$ censurée par $X \hookrightarrow \text{Exp}(2)$ avec $\gamma_2 = 0$.

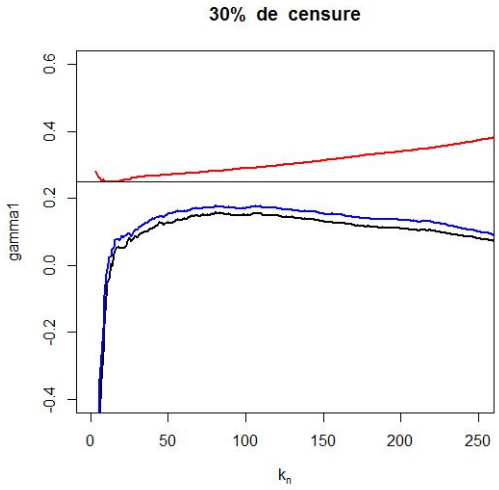
Dans ce cas, le pourcentage de censure est égale à 50% ($p = 1/2$). La figure (2.7) illustre le comportement empirique des estimateurs en fonction de nombre de statistiques d'ordre $k_n$ en se basant sur les 100 échantillons. Sur le panneau gauche, nous présentons les graphes de l'IVE $\widehat{\gamma}_1^{(c,\bullet)}$ (resp. MAE$(\widehat{\gamma}_1^{(c,\bullet)})$ et le RMSE$(\widehat{\gamma}_1^{(c,\bullet)})$) en fonction de $k_n$. Sur le panneau droit, nous présentons les graphes de $\widehat{q}_{(c,\bullet)}$ ( resp. MAE$(\widehat{q}_{(c,\bullet)})$ et RMSE$(\widehat{q}_{(c,\bullet)})$) en fonction de $k_n$. L'ordre de quantile extrême est $p_n = 1/50$, la ligne horizontale en rouge représente la vraie valeur de ce quantile.
D'après les résultats graphiques de la simulations, nous constatons que :
Pour l'IVE (resp. quantile extrême), l'estimateur $\widehat{\gamma}_1^{(c,\mathsf{GPDmle})}$ (resp. $\widehat{q}_{\mathsf{cPOT}}$) semble être de meilleur qualité que l'estimateur $\widehat{\gamma}_1^{(c,D)}$ (resp. $\widehat{q}_{(c,D)}$) en terme de l'erreur absolue moyenne (MAE) et le RMSE.



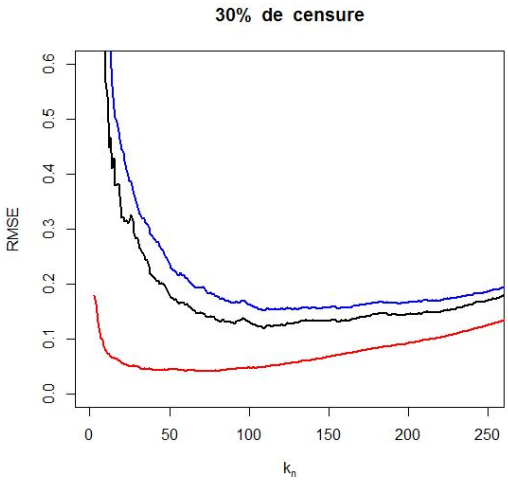

(a) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 0.5)$



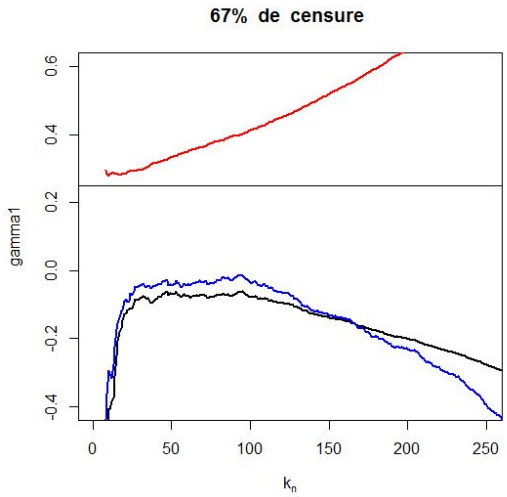

(b) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 0.5)$



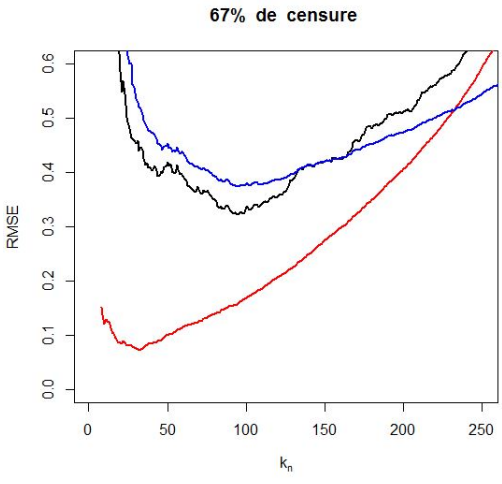

(c) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 5/7)$




(d) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 5/7)$

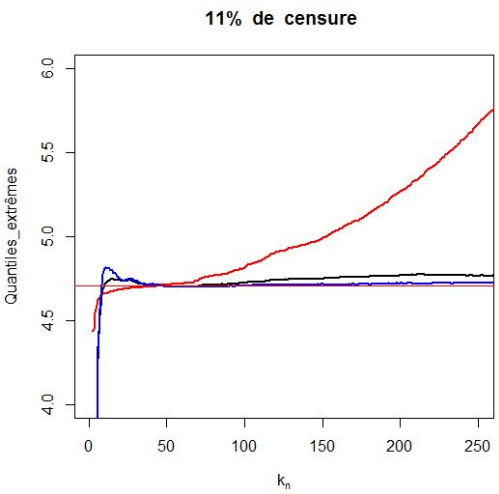



(e) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 2, 8)$

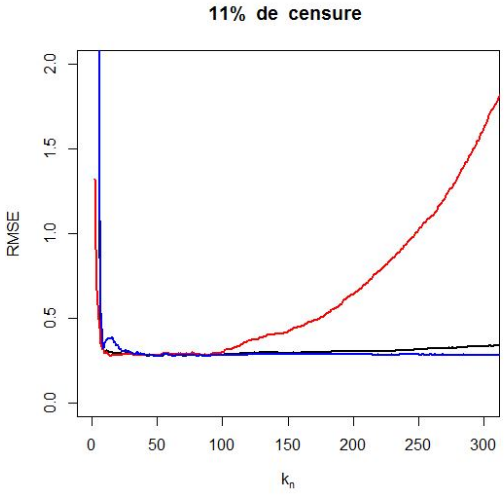



(f) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 2, 8)$

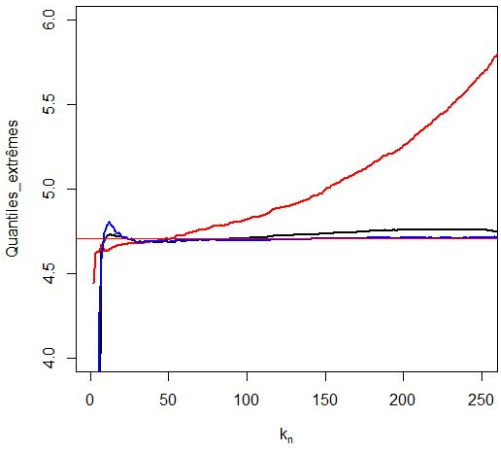
Figure 2.5: Influence de pourcentage de censure sur le comportement empirique des estimateurs de l'IVE de $\Omega_1$. En noir: $\widehat{\gamma}_1^{(c,D)}$, en rouge: $\widehat{\gamma}_1^{(c,H)}$ et en blue: $\widehat{\gamma}_1^{(c,\mathsf{GPDmle})}$. La ligne horizontal en noir représente la vrai valeur de $\gamma_1 (= 0.25)$



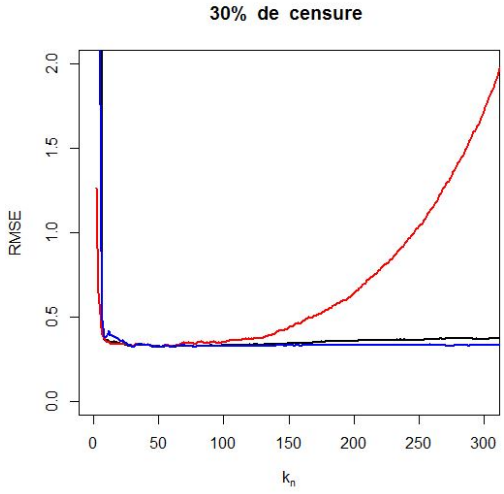

(a) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 0.5)$



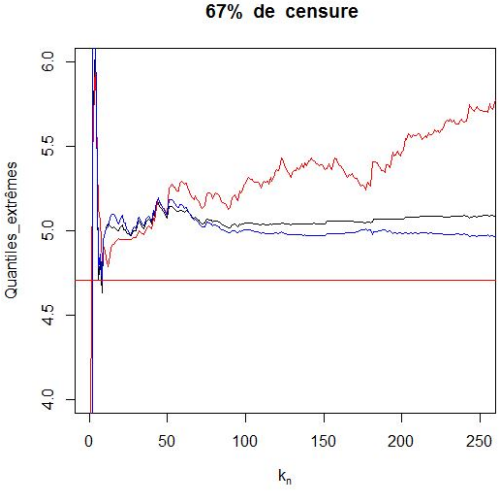

(b) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 0.5)$



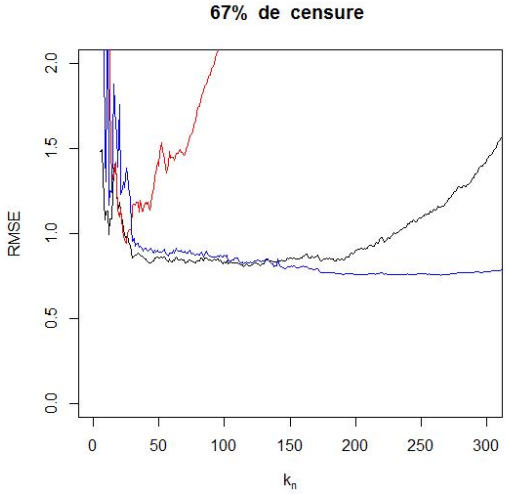

(c) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 5/7)$




(d) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 1, 5/7)$

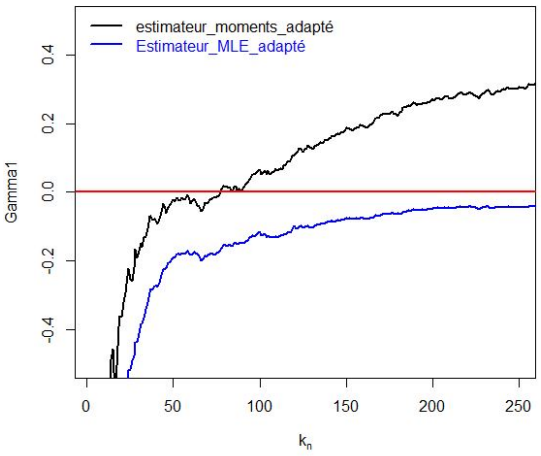



(e) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 2, 8)$

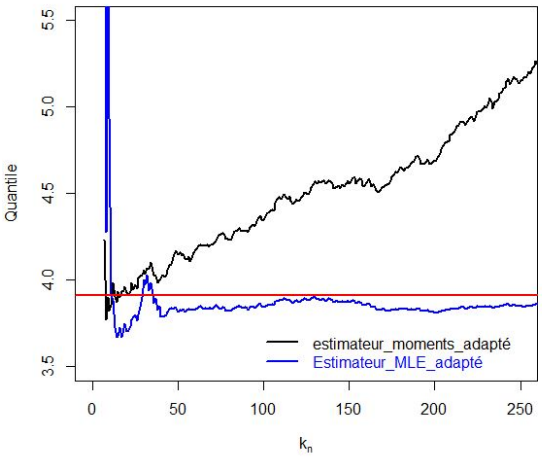



(f) X$\hookrightarrow$ Burr$(10, 4, 1)$ et C $\hookrightarrow$ Burr$(10, 2, 8)$

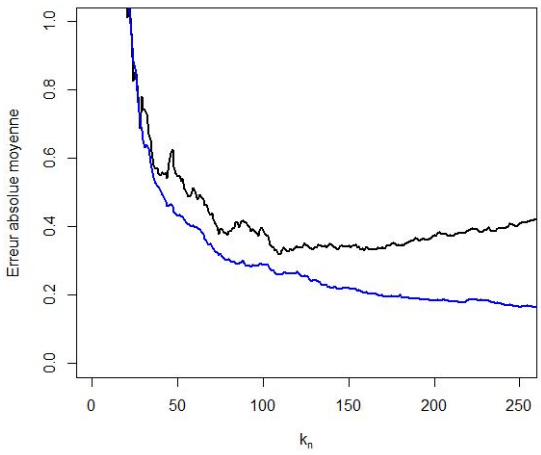
Figure 2.6: Influence de pourcentage de censure sur le comportement empirique des quantiles extrêmes $\widehat{q}_{(c,\bullet)}$. En noir: $\widehat{q}_{(c,D)}$, en rouge $\widehat{q}_{(c,W)}$ et en blue : $\widehat{q}_{\mathsf{cPOT}}$.



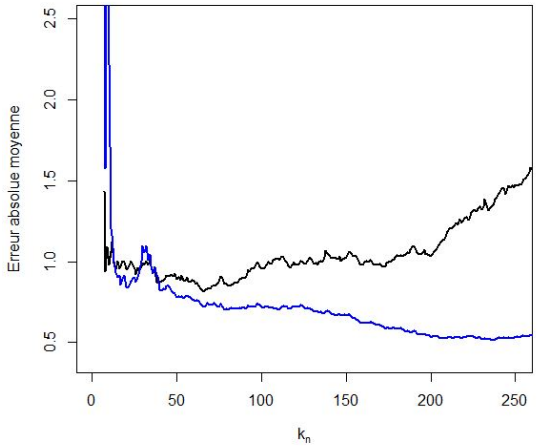

(a) X ↪ Exp(1) et C ↪ Exp(2)



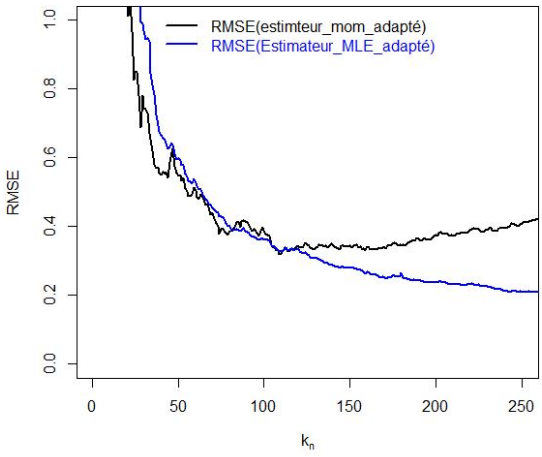

(b) X ↪ Exp(1) et C ↪ Exp(2)



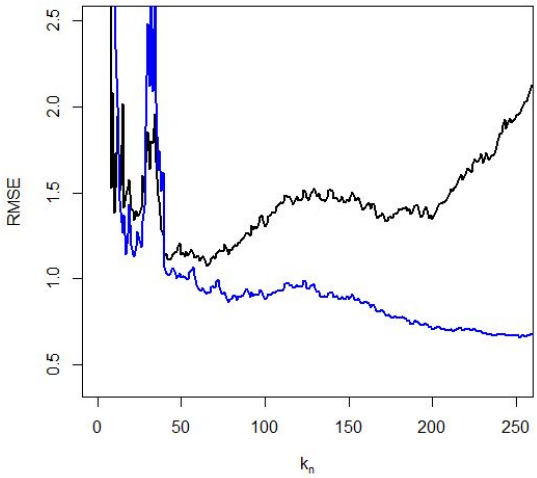

(c)




(d)




(e)




(f)

Figure 2.7: Estimation de $\widehat{\gamma}_1^{(c,\bullet)}$ et $\widehat{q}_{(c,\bullet)}$ dans le cas de $\gamma_1 = \gamma_2 = 0$.

### 2.5.2 Application sur des données réelles

DANS cette partie, nous allons appliquer les résultats de la TVE sous censure pour étudier le temps de survie des patients atteints de SIDA. Les données sont recueilliées par le Dr.P.J. Solomon et le NCHECR[2] pendant la période de 24–09–1982 (date d'origine), jusqu'à la date de 30–06–1991 qui correspond à la date de point, (voir Venables et Ripley (2002,[35])). Dans le R, ces données sont stockées dans le package MASS, sous forme d'une liste de classe "data.frame", nommée 'Aids2'. Le jeu de données contient 2843 lignes, qui correspondent aux 2843 patients traités pendant la durée de l'étude qui est égale à peu près sept ans. Les colonnes de ce jeu de données contiennent les informations relatives à chaque patient: date du diagnostic, la date du décès ( ou la fin d'observation), un indicateur qui vaut 1 si le patient est décédé ou 0 si le patient est encore vivant à la date de point (fin d'observation),...etc
Parmi les 2843 patients, 1708 sont décédés et les autres temps de survie sont donc censurées à droite. Sur le nombre total des patients, 2754 étaient des hommes, dont 1708 sont morts et les restants (1046) sont censurés. Dans notre étude, nous nous intéressons qu'aux patients du sexe masculin. Ces données sont déjà étudiées dans la litérature des valeurs extrêmes par plusieur auteurs: Einmahl et al [15], Nado [29], Richard et al (2017,[32])...

#### 2.5.2.0.1 Modélisation des données et estimation des paramètres

Notons par X la v.a qui représente le temps de survie d'un patient atteint de SIDA, F (inconnue) la fonction de répartition de X et $\gamma_1$ son IVE. Nous allons appliquer les estimateurs donnés dans $\Omega_1$ (cf.(2.30)) pour estimer l'IVE de F, ainsi les estimateur de $\Omega_2$ (cf.(2.30)) pour l'extrapolation un quantile extrême $F^{\leftarrow}(1-p_n)$ avec $p_n = 1/100$, ce qui va nous donner une indication sur la durée qu'un sujet diagnostiqué positif à ce virus pourra survivre.

**Choix de proportion de non-censure $\widehat{p}$**

Comme l'ont déjà montré les résultats de simulation, les estimateurs de l'IVE ,(ainsi quantile extrême), sont sensibles à la proportion de censure dans la queue de distribution. Par conséquence, il est nécessaire de choisir un $\widehat{p}$ approprié dans l'estimation. La figure (2.8) montre le graphe de la proportion de non-censure en fonction de nombre de statistique d'ordre $k_n$. Einmahl et al ont pris comme valeur de $\widehat{p}$, la valeur pour laquelle ce graphe est plus stable:

$$\widehat{p} = 0.28 \quad \text{qui correspnd à} \qquad 60 \leqslant k_n < 200 \tag{2.33}$$

**Estimation de l'IVE et quantile extrême**

- **Résultats graphiques:**

  La figure 2.9(a) (resp. 2.9(b)) illustre les différents estimateurs de l'IVE (resp. quantile extrême d'ordre $p_n = 1/100$) en fonction de $k_n$.

[2]National Centre in HIV Epidemiology and Clinical Research (Australia).

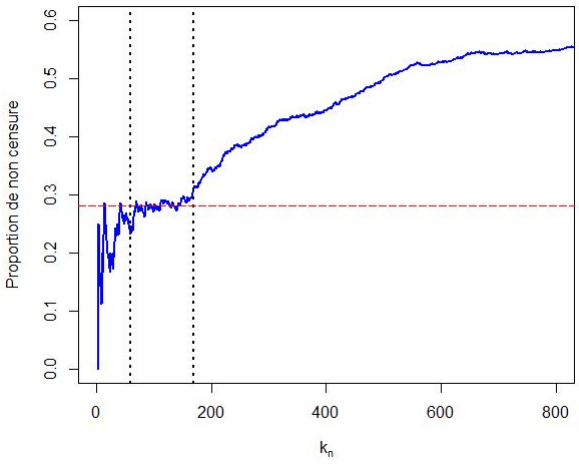
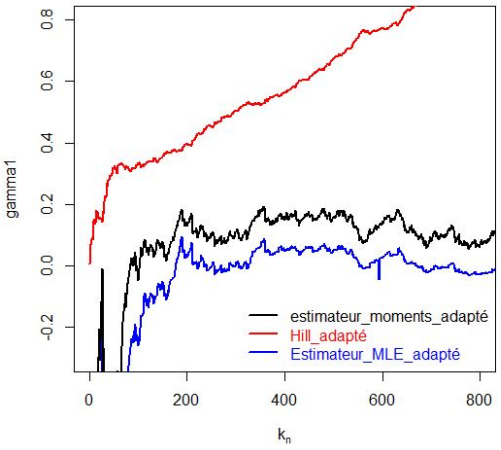
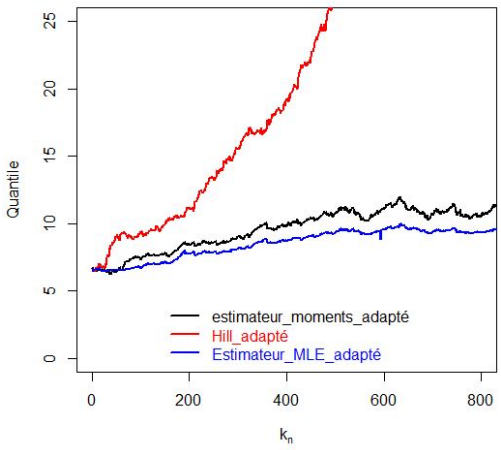



Figure 2.8: Proportion de censure en fonction de $k_n$ pour les patients de sexe masculin atteint de SIDA (données: 'Aids2'). Les deux lignes verticales représentent la zone de stabilité de $\widehat{p}$ ($\widehat{p} = 0.28$).




(a) $\widehat{\gamma}_1^{(c,\bullet)}$ en fonction de $k_n$




(b) $\widehat{q}_{(c,\bullet)}$ en fonction de $k_n$

Figure 2.9: Estimation de $\gamma_1$ (a) et le quantile extrême d'ordre $p_n = 1/100$ (b) à partir de données 'Aids2'.

- **Résultats numériques:**

  Pour les résultats numériques, nous commonçons d'abord par le choix du nombre optimal des excès $k_n$. Pour cela, nous utilisons l'approche heuristique modifiée donnée par Richard et al (2017,[32]).
  Considérons $\{\widehat{\gamma}_1^{(i)} = \widehat{\gamma}_1^{(c,\bullet)}, \quad i \in \omega_1 = \{1,2,3\}\}$ avec $\widehat{\gamma}_1^{(c,\bullet)} \in \Omega_1$(cf.(2.30)). La valeur optimale de $k_n$ est choisie comme:

$$k_{\text{opt}} = \underset{k_n}{\text{argmin}} \sqrt{\sum_{\substack{(i,j)\in\omega_1 \\ i\neq j}} \left(\widehat{\gamma}_1^{(i)} - \widehat{\gamma}_1^{(j)}\right)^2} \tag{2.34}$$

  La figure (2.10), illustre l'approche heuristique utilisé pour le choix de $k_{\text{opt}}$. Sur le graphe, on peut constater que le minimum de $\sqrt{\sum(\widehat{\gamma}_1^{(i)} - \widehat{\gamma}_1^{(j)})^2}$ est atteint pour $k_n$ entre 150 et 250.

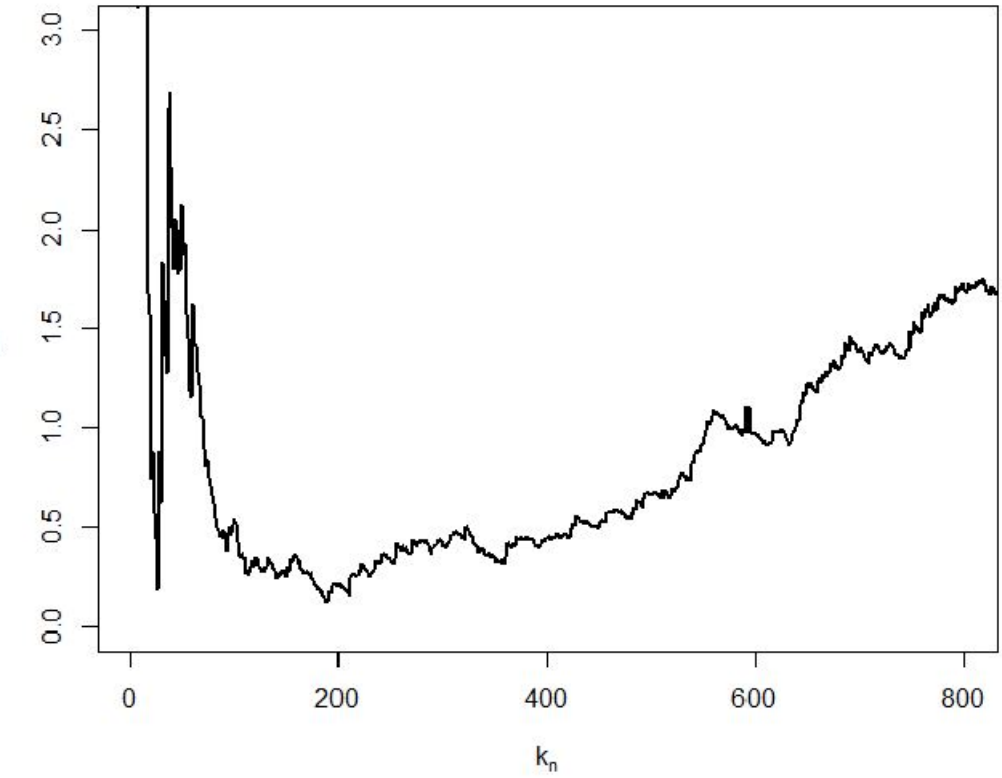



Figure 2.10: Graphe de $\sqrt{\sum(\widehat{\gamma}_1^{(i)} - \widehat{\gamma}_1^{(j)})^2}$ en fonction de $k_n$

Le résultat numérique obtenu avec le R, (voir le programme (2.5.1)), en appliquant l'approche donnée par l'expression (2.34) est :

$$\boxed{k_{\text{opt}} = 190}$$

Ainsi, avec cette valeur de $k_n$, nous obtenons les résultats numériques de l'estimation de l'IVE (tableau 2.2) et le quantile extrême avec $p_n = 1/100$, (voir tableau 2.3).

| $\widehat{\gamma}_1^{(c,H)}$ | $\widehat{\gamma}_1^{(c,D)}$ | $\widehat{\gamma}_1^{(c,\textbf{GPDmle})}$ |
|---|---|---|
| 0.374 | 0.183 | 0.095 |

Table 2.2: Estimation de l'IVE

| $\widehat{q}_{(c,W)}$ | $\widehat{q}_{(c,D)}$ | $\widehat{q}_{\textbf{cPOT}}$ |
|---|---|---|
| 10.500 | 8.605 | 7.981 |

Table 2.3: Quantile extrême

D'après les résultats de tableau (2.3), on constate que la durée de survie d'un patient atteint du SIDA peut atteindre 11 ans.

```
library(MASS)

library(ReIns)  #To use the estimators of EVT adapted to censore

data=Aids2

#############################################################################
###                          Extraire les donnees de sexe  masculin       ##
#############################################################################

m=2754        ## nbr de patients du sexe masculin
nn=length(data[,1])
X=array(1:m,dim=m)
Y=array(1:m,dim=m)
censure=array(1:m,dim=m)
sexe=data[,2]
a=1
for(j in 1:nn)
   {
    if (sexe[j]=="M")
      {
        X[a]=c(data[j,3])
        Y[a]=c(data[j,4])
        censure[a]=c(data[j,5])
         a=a+1
      }
   }

#############################################################################

X=as.Date(X , origin=as.Date("1960-01-01")) ##~Calenderier Julien-->Gregorien
Y=as.Date(Y , origin=as.Date("1960-01-01"))
T=Y-X
n=length(T)

tab =array(1:n,dim=c(1,n))
tab2 = array(1:n,dim=c(1,n))
tab3 = array(1:n,dim=c(1,n))

for( i in 1:m)
    {
      if (censure[i]==2)    ##2: !censored
            {
                delta=1
            }
      if(censure[i]==1)     ##1: censored
            {
               delta=0
            }
tab2[i]=c(delta)
     if (T[i] !=0)
            {
              tab[i]=c(T[i])
             }
     else {tab[i]=c(0.0001)}
    }
```

```
##############################################################################
###                  Proportion de censure en fonction de kn                ###
##############################################################################

####      fonction qui retourne la propo de censure en fonction de kn    ####

proportion=function(data,censure0)
    {
       s =sort(data,index.return =TRUE)  ## ordre croissant des observation

       X =s$x

       censored = censure0
       deltaa = !(censored[s$ix])

       n =length(X)
       prop= numeric(n)
       K = 1:(n-1)

       prop[K]= cumsum(deltaa[n - K + 1])/K

 return(prop)

    }


tab=as.numeric(tab)
tab2=as.numeric(tab2)
p_censure=proportion(tab,tab2) ##proportion de censure


##############################################################################
########    Graphe de proportion de non-censure en fct de kn           ######
##############################################################################


plot(1-(p_censure),type="l",lty=1,col='blue',lwd=2,
  "xlab"=expression(list(k[n])), ylab="Proportion_de_non_censure",
   xlim=c(1,800),ylim=c(0,0.6))

abline(h=0.28,col="red",lty=5,lwd=0.5)
abline(v = 60, lty = 3,lwd= 2.5, col = "black")
abline(v = 170, lty = 3,lwd = 2.5, col="black")

##############################################################################
```

```
#############################################################################
######                          IVE et Quantile sous censure    ###############
#############################################################################

tab=as.numeric(tab/365)  ## convertir les durees de survie en annee

tab2=as.numeric(tab2)
hill_adapted=cHill(abs(tab), censored=tab2, plot=FALSE)
c_Moment=cMoment(abs(tab), censored=tab2, plot=FALSE)
cGPD_mle=cGPDmle(abs(tab), censored=tab2, plot=FALSE)

plot(cGPD_mle$gamma,type="l",lty=1,col='blue',lwd=2,
"xlab"=expression(list(k[n])),ylab="gamma1",xlim=c(1,800),ylim=c(-0.3,0.8))

lines(hill_adapted$gamma,type="l",lty=1,col='red',lwd=2)

lines(c_Moment$gamma,type="l",lty=1,col='black',lwd=2)

legend('bottomright',legend=c("estimateur_moments_éadapt","Hill_éadapt",
"Estimateur_MLE_éadapt"),lty=c(1,1,1),lwd=c(2,2,2),bty = "n",
col=c('black','red','blue'), text.col =c('black','red','blue'))


quantile_GPD_mle=cQuantGPD(abs(tab), gamma1=cGPD_mle$gamma1,
   sigma1=cGPD_mle$sigma1,censored=tab2, p=p, plot=FALSE)

quantile_hill=cQuant(abs(tab), gamma1=hill_adapted$gamma,
                         censored=tab2, p=p, plot=FALSE)

quantile_moments=cQuantMOM(abs(tab), censored=tab2,
                 gamma1=c_Moment$gamma1, p=p, plot=FALSE)

X11()
plot(quantile_moments$k,quantile_moments$Q, type="l",lty=1,col='black',lwd=2,
    "xlab"=expression(list(k[n])),ylab="Quantile",ylim=c(0,25),xlim=c(1,800))

lines(quantile_hill$k,quantile_hill$Q,type="l",lty=1,col='red',lwd=2)

lines(quantile_GPD_mle$k,quantile_GPD_mle$Q,type="l",lty=1,col='blue',lwd=2)

legend('bottom',
legend=c("estimateur_moments_éadapt","Hill_éadapt",
"Estimateur_MLE_éadapt"), lty=c(1,1,1),lwd=c(2,2,2),bty = "n",
col=c('black','red','blue'), text.col =c('black','red','blue'))
```

```
#############################################################################
#######                  Choix optimal du nombre d'exces kn            ######
#############################################################################

alpha = array(4:n-1,dim=c(4,n-1))
alpha[1,]=hill_adapted$k

alpha[2,]=(hill_adapted$gamma -cGPD_mle$gamma1)^2
alpha[3,]=(hill_adapted$gamma -c_Moment$gamma1)^2
alpha[4,]=(cGPD_mle$gamma1 -c_Moment$gamma1)^2

argmin = array(2:n-1,dim=c(2,n-1)) argmin[1,]=alpha[1,]
argmin[2,]=alpha[2,]+alpha[3,]+alpha[4,]
k=1 beta=argmin[2,1]

for (i in 2:(n-1))
    {
          if (beta>argmin[2,i])
              { beta=argmin[2,i]
                k=i
              }
    }
k     ## afficher le kn optimal

plot(argmin[1,],argmin[2,],type="l",lty=1,col='black',lwd=2,
     "xlab"=expression(list(k[n])),"ylab"=expression(k[list(opt)](kn)),
                                          xlim=c(1,800),ylim=c(0,3))


#############################################################################
###    Affichage des resultats numeriques d'estimation pour le k_opt      ###
#############################################################################

ive_hill = array(1:n-1,dim=c(1,n-1))
ive_mom = array(1:n-1,dim=c(1,n-1))
ive_MLE = array(1:n-1,dim=c(1,n-1))

quantil_hill = array(1:n-1,dim=c(1,n-1))
quantil_mom = array(1:n-1,dim=c(1,n-1))
quantil_mle = array(1:n-1,dim=c(1,n-1))

ive_hill=hill_adapted$gamma
ive_mom=c_Moment$gamma1
ive_MLE=cGPD_mle$gamma1

quantil_hill =quantile_hill$Q
quantil_mom =quantile_moments$Q
quantil_mle =quantile_GPD_mle$Q

#############################################################################
ive_hill[k]
ive_mom[k]
ive_MLE[k]

quantil_hill[k]
quantil_mom[k]
quantil_mle[k]
```

**Program 2.5.1:** Le code source utilisé dans l'application sur les données Aids2.

# *Conclusion*

L'objectif de ce travail est d'étudier les estimateurs de l'IVE et quantile extrême dans le cas de données censurées aléatoirement à droite.

Pour faciliter la lecture de ce document, nous avons rappelé, dans le premier chapitre, les principaux résultats et fondements de la théorie des valeurs extrêmes, avec un accent particulier sur les estimateurs les plus fréquents de l'IVE et quantile extrême, où nous avons illustré le comportement de ces estimateurs avec des exemples de simulation.

Le deuxième chapitre est consacré pour les extrêmes sous censure. Après avoir abordé le sujet de l'analyse de survie et les données censurées, nous avons présenté la méthode de Beirland et al (2007,[3]) pour l'estimation de l'IVE et du quantile extrême sous données censurées aléatoirement à droite.
À la fin de notre travail, nous avons illustrer le comportement et l'influence de pourcentage de la censure sur le comportement de ces estimateurs.

Il serait intéressant d'étendre ce travail à plusieurs perspectives comme:

- Étudier la consistance des estimateurs de IVE et quantile extrême sous des données incomplètes.
- Étudier le comportement des estimateurs sous l'hypothèse de dépendance (le mélange, association , dépendance faible au sens de Doukhan...) pour des données censurées.

# *Bibliographie*


[1] Aalen, O. (1978).*Nonparametric Inference for a Family of Counting Processes, The Annals of Statistics.* vol.6, n°4, p.701–726. 42

[2] Alberti.C, Timsit.J.-F, et S. Chevret. [2005].*Analyse de survie : comment gérer les données censurées ?, Revue des Maladies Respiratoires* ,vol. 22, N° 2-C1,p. 333–337. URL https://www.em-consulte.com/rmr/article/157019

[3] Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A. (2007).*Estimation of the extreme value index and extreme quantiles under random censoring. Extremes,* 11, p.151–174.. xi,43,44

[4] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., de Waal, D. et Ferro, C. (2004b).*Statistics of Extremes : Theory and Applications.* John Wiley & Sons. 26

[5] Bernard, G. et Ahmed, S. (1958). *Applications of order statistics to health data.* A.J.P.H, 48(10):1388–1394.2

[6] Bingham, N., C. Goldie et J. Teugels. (1987). *Regular Variation, Encyclopedia of Mathematics and its application, vol.27, Cambridge University Press* 13,13.

[7] Castillo, E., Hadi, A., Balakrishnan, N. et Sarabia, M. (2005). *Extreme Value and Related Models with Applications in Engineering and the Sciences.* Wiley, WILEY-INTERSCIENCE.2

[8] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values,* Springer, London.11.

[9] David, H.A., et Nagaraja, H.A. (2003). *Order Statistics Third Edition.* JOHN WILEY & SONS, INC., PUBLICATION.3,4,17

[10] De Haan, L. et A. Ferreira. (2007). *Extreme value theory : an introduction,* Springer Science & Business Media.15,15,19.

[11] Deheuvels, P., Häeusler, E., and Mason, D. M. (1988).*Almost sure convergence of the Hill estimator, dans Mathematical Proceedings of the Cambridge Philosophical Society, vol.104,Cambridge University Press,* p.371–381. 26

[12] Dekkers, A. L., et De Haan, L. (1989).*On the estimation of the extreme value index and large quantile estimation, Ann. Statist.,*vol.17, n°364, p.1795–1832. 30

[13] Dekkers, A. L.,Einmahl, J. H., et De Haan, L. (1989). *A moment estimator for the index of an extreme-value distribution, The Annals of Statistics,*vol. 17, n°4, p.1833–1855. 29

[14] Droesbeke, J.J., Saporta, G. (2011).*Approches non parametriques en régression. Editions Technip* 42

[15] Einmahl, J. H., Fils-Villetard, A., and Guillou, A. (2008). *Statistics of extremes under random censoring. Bernoulli,* vol.14,n°1, p.207–227. xii,43,43,44,55.

[16] Embrechts, P., C.Klüppelberg et T.Mikosch. (1997). *Modelling extremal events: for insurance and finance, Springer Science BusinessMedia.* vol.33. 5,17,11.

[17] Fisher, R. A. et L. H. C. Tippet. (1928). *Limiting forms of the frequency distribution of the largest or smallest member of a sample, dans Mathematical Proceedings of the Cambridge Philosophical Society,* vol.24, Cambridge University Press,180–190. xi,2

[18] Fréchet, M. (1927). *Sur la loi de probabilité de l'ecart maximum, dans Annales de la société Polonaise de Mathématique,* vol.6, p.93–116. xi,6

[19] Gnedenko, B. (1943). *Sur la distribution limite du terme maximum d'une serie aléatoire*, Annals of Mathematics, vol.44, 3:423–453. 2

[20] Gumbel, E. J. (1958). *Statistics of extremes, Columbia Univ. press, New York,* vol.247. 6

[21] Hill, B. (1975).*A simple general approach to inference about the tail of a distribution, The Annals of Statistics,*vol. 3,n°5, p.1633–1174. 25

[22] Hosking, J. R., J. R. Wallis et E. F. Wood. (1985).*Estimation of the generalized extreme-value distribution by the method of probability-weighted moments,*Technometrics, vol. 27,n°3,p.251–261. 21,22

[23] Hosking, J. et Wallis, J. (1987).*Parameter and quantile estimation for the generalized Pareto distribution. Technometrics,*vol. 29,n°3,p.339–349. 24,25

[24] Huber-Carol.C (1994).*durées de survie tronquées et censurées, Journal de la société statistique de Paris* tome. 135, n°4,p.3–23. URL http://www.numdam.org/item?id=JSFS_1994__135_4_3_0. 37

[25] Jenkinson, A. F. (1955).*The frequency distribution of the annual maximum (or minimum) values of meteorological elements. The Quarterly Journal of the RoyalMeteorological Society,* vol.81, n°384, p.141–160. 5

[26] Kaplan, E. L. et P. Meier. (1958).*Nonparametric Estimation from Incomplete Observations,Journal of the American Statistical Association,* vol.53, n°282, p.457–481. 42

[27] Longuet-Higgins, M. (1952). *On the statistical distribution of the heights of sea waves*, Journal of marine research, 9:245–266. 2

[28] Mason, D. M. (1975).*Laws of large numbers for sums of extreme values, The Annals of Probability,*vol.10, n°3, p.754–764. 26

[29] Ndao, P. (2015).*Modélisation de valeurs extrêmes conditionnelles en présence de censure.* université Gaston Berger de Saint-Louis,Senegal,Thèse de doctorat. 37

[30] Nelson, W. (1972).*Theory and Applications of Hazard Plotting for Censored Failure Data, Technometrics.* vol.4, n°14, p.945–966. 42

[31] Pickands, J. (1975). *Statistical inference using extreme order statistics*, The Annals of Statistics, vol.3, p.119–131. 2, 7, 30

[32] Richard.M,Tertius.W et Kwabena.D-W. (2017). *On Extreme Value Index Estimation under Random Censoring*,DOI: 10.16929/ajas/419.223. 55.

[33] Reiss, R.-D. and Thomas, M. (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Basel: Birkhäuser.* xi, 43.

[34] Resnick,Sidney I. (1987). *Extreme Values, Regular Variation and Point Processes. Springer Verlag, New-York.* 5

[35] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S. Springer, New York, NY, 4 edition.* 55

[36] Von-MISES, R. (1936).*La distribution de la plus grande de n valeurs. Revue deMathématique Union Interbalcanique,* vol.1, p.141–160. 5

[37] Weibull, W. (1951). *"Wide applicability", Journal of Applied Mechanics,* vol.103, p.293–297. 12

[38] Weissman, I. (1978).*Estimation of parameters and large quantiles based on the k largest observations, Journal of the American Statistical Association,*vol.54, n°364, p.812–815. 29