الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'Enseignément Supérieur et la Recherche Scientifique

Université de Blida 1

Mémoire

Pour l'obtention du diplôme de

Master en Mathématiques

Spécialité : **Modélisation Stochastique et Statistique**

Modélisation des distributions extrêmes et ses application dans l'analyse de survie

Présentée par

BOULARES Adil & MEROUANI Taha

Soutenu le 11/07/ 2021

Devant le jury composé de :

Abdelaziz RASSOUL	Promoteur de Mémoire	Prof.	ENSH, Blida
Omar TAMI	Président	MCB	Univ. Blida 1
Redouane FRIHI	Examinateur	MAA	Univ. Blida 1

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

13 juillet 2021

DÉDICACE

Salutations

à nos pères

mères

et

tous
nos amis.

REMERCIEMENTS

Tout d'abord nous remercions **Allah** qui nous a donné la volonté et le courage pour pouvoir réaliser ce travail. Nos sincéres remerciements à notre encadreur,Monsieur RASSOUL Abdelaziz , Professeur à L'Ecole Nationale Supérieure d'Hydraulique, pour avoir accepté d'encadrer cette mémoire . Un grand merci pour son extrême patience au cours de ces Trois dernières mois. nous somme profondément reconnaissant pour ses nombreux conseils, pour ses corrections et son soutien indéfectible tout au long de ce travail, il me faudrait des pages pour le remercier. nous a donné l'occasion de travailler sur un thème fascinant et intéressant, me permettant de mieux comprendre les concepts théoriques de la statistique par leur application à des cas concrets.

Nous tenons à exprimer notre gratitude à tous les enseignants du département de mathématiques, en particulier les enseignants de Master .

Un merci infini à nos parents pour leur patience avec nous toutes ces années.

ملخص

في هذا العمل ، قمنا بعرض علاقة بين مفهومين للاحصاء و هما نظرية القيم المتطرفة و تحليل البقاء .

في تحليل الموثوقية او تحليل البقاء، تكون معظم المعطيات معالجة بنومذج ويبول الذي يعتبر حالة من حلات قانون الجذب العام لنظرية القيم المتطرفة حسب معلميت الشكل .

في هذا العمل نهتم بعائلة خاصة من القوانين الاحتملاية آلا وهيا التوزيعات ذات الذيل من نوع ويبول ، هذه القوانين لديها دالة بقاء تتانقص بسرعة أسية (نتحدث ايضا على الذيل الخفيف)، امثلة على بعض هذه القوانين التوزيع الأسي و التوزيع الطبيعي و التوزيع غاما،.....الخ

سرعة التقارب لذيل توزيع الاحتمال يتم التحكم به عن طريق معلمية الشكل الذي يسمى مؤشر ذيل ويبول، و في تحليل البقاء نواجه كثيرا من العارقيل من نوع المعطيات الناقصة.

المعطيات الخاصة بالحياة لاتخضع لملاحظة كاملة، هي ليست نادرة و لكنها غير كاملة .

الرقابة و الانقطاع عن الرقابة كلاهما سبب للمعطيات الغير كاملة او الناقصة ،

الرقابة هي آلية تمنع المراقبة الدقيقة للحد الزمني لحدوث الفائدة نحن نستخدم نماذج الموثوقية (نموذج كوكس للمخاطر النسبية ... إلخ) و التقديرات المعروفة (Hill، Pickands، Kaplan-Meier، ... إلخ) وكل هذا في حالة الرقابة العشوائية على جهة اليمين .

ABSTRACT

In this work, we present a relationship between two concepts in statistics, the extreme value theory and survival analysis. In reliability analysis or survival analysis, the data are often treated with a weibull model which is considered as a case in the domains of attraction and the theoretical concepts associated with them in the three cases corresponding to the sign of the shape parameter . we are interested in a particular family of laws : the tail laws of Weibull type. These laws have a survival function which decreases at an exponential rate (we also speak of a light tail). Examples of such laws are the exponential, normal, gamma, etc... The speed of convergence of the tail distribution is controlled by a shape parameter called the Weibull tail index, and in survival analysis it is very common to face the problem of missing data.

The survival data are not fully observed. It is not uncommon, but they are rather incomplete. Censoring and truncation are the two most common causes of incomplete data. Censoring is a mechanism that prevents accurate observation of the time of occurrence of interest. We use reliability models (Cox model proportional hazard ... ect) and known estimates (Hill, Pickands, Kaplan-Meier, ...ect) is all that in the case random censoring to the right .

RÉSUMÉ

Dans ce travail, nous présentons une relation entre deux concepts en statistique, La théorie des valeurs extrême et L'analyse de survie.

En analyse de fiabilité ou L'analyse de survie, les données sont souvent traitées avec un modèle weibull qui est considéré comme un cas dans les domaines d'attraction et les concepts théoriques qui leur sont associés dans les trois cas correspondant au signe du paramètre de forme .

nous nous intéressons à une famille particulèire de lois : les lois à queue de type Weibull . Ces lois possèdent une fonction de survie qui décroit à une vitesse exponentielle (on parle aussi de *queue légère*).Des exemples de telles lois sont les lois exponentielle, normale, gamma, ect ... La vitesse de convergence de la queue de distribution est controlée par un paramètre de forme appelé indice de queue de Weibull , et dans l'analyse de survie, il est très commun de se trouver en face du problème de données manquantes.

Les données de survie ne sont pas totalement observées. Il n'est pas rare, mais elles sont plutôt incomplètes. La censure et la troncature sont les deux causes de données incomplètes les plus répandues. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt.

Nous utilisons des modèles de fiabilité (Cox model proportional hazard ... ect) et des estimations connues (Hill, Pickands, Kaplan-Meier, ...ect) est tous ça dans le cas censure aléatoire à droite .

INTRODUCTION GÉNÉRALE

Pour un profane, la statistique est associée à la notion de moyenne ou d'écart-type. En effet, dans de nombreuses applications, notamment dans les sciences sociales ou sciences physiques , les statistiques se résument parfois au calcul de moyennes ou à l'évaluation de la dispersion d'une série de valeurs autour de leur moyenne.

Par définition, les évènements rares sont des évènements ayant une faible probabilité d'apparition. Lorsque le comportement de ces évènements est dû au hasard on peut étudier leur loi. Ils sont dits extrêmes quand il s'agit de valeurs beaucoup plus grandes ou plus petites que celles observées habituellement.

Les évènements extrêmes et catastrophiques (tremblements de terre, inondations, accidents nucléaires, crises monétaires ou financiéres, krachs boursiers, émergence d'un nouveau phénomène endémique, etc...) dominent l'actualité quotidienne par leur caractère imprévisible.

L'analyse des valeurs extrêmes avec l'analyse de survie est un nouveau sujet de recherche. L'objet de cette mémoire est d'étendre les résultats de la théorie des valeurs extrêmes dans le cas où l'échantillon consiste en un ensemble de données censurées tout en apportant les modifications nécessaires.

Des estimateurs sont proposés dans le bouquin de Reiss et Thomas [87, 2007], mais sans résultats asymptotiques. Un premier pas, dans l'analyse du comportement asymptotique, des estimateurs de l'indice des valeurs extrêmes et des quantiles extrêmes sous censure, est fait par Beirlant et al. [90, 2007]. Leurs estimateurs sont basés sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dépassant un certain seuil donné. L'année suivante Einmahl et al. [89, 2008] ont utilisé le même concept pour proposer un estimateur adapté de l'indice de queue dans le cas où les données sont censurées par un seuil aléatoire et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique.

Dans le cas de non censure, il y a toute une théorie (théorie des valeurs extrêmes (TVE). Pour l'analyse des extrêmes qui se fait selon deux approches. La première, qu'on

appellera approche GEV ; permet de modéliser les block maxima par une distribution GEV (generalized extreme value distribution) et la seconde, appelée approche GPD consiste à ajuster les observations dépassant un certain seuil (peaks over threshold : POT) par une GPD (generalized Pareto distribution). Pour une description détaillée de la TVE, en particulier sur l'estimation de l'indice des valeurs et quantiles extrêmes, consulter les excellents bouquins comme Embrechts et al. [1, 1997], Coles [6, 2001], Beirlant et al. [39, 2006], Reiss et Thomas [87, 2007] . On essaie à travers cette mémoire d'adapter les outils de la TVE avec censure au cas de données à queues lourdes et distribution bornée (distribution de Fréchet et distribution de Weibull) censurées.

Dans l'analyse de survie, il est très commun de se trouver en face du problème de données manquantes. Les données de survie ne sont pas totalement observées. Il n'est pas rare, mais elles sont plutôt incomplètes. La censure et la troncature sont les deux causes de données incomplètes les plus répandues. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt. On sait bien que ce délai appartient à un certain intervalle de temps. La troncature survient qu'on ne peut pas observer les individus de l'échantillon dont le délai de survenue appartient à un certain intervalle de temps, on observe donc un sous-échantillon. Dans ce cas les techniques classiques ne s'adaptent pas correctement aux données incomplètes.

La littérature est beaucoup plus riche en censure que la troncature qui est plus récente. Dans cette mémoire , on va s'intéresser particuliérement à la censure droite dans le cadre d'apporter de nouveaux résultats. Pour des détails complets sur la censure et l'analyse de survie, on réfère aux livres de Cox et Oakes [63, 1984], Kalbfleisch et Prentice [91, 2011], Lee et Wang [70, 2003], Klein et Moeschberger [92, 2003].

En 1958, Kaplan et Meier [82, 1958] ont introduit un estimateur (portant leurs noms) de la fonction de survie des données censurées. Cet estimateur possède des propriétés asymptotiques très populaire (convergence uniforme, presque sure, normalité asymptotique) similaires à celles de la fonction de répartition empirique. Le com- portement asymptotique de l'estimateur de Kaplan-Meier a suscité l'intérêt d'un grand nombre d'auteurs, Breslow et Crowley [93, 1974] sont les premiers à traiter la convergence et la normalité asymptotique de l'estimateur de Kaplan-Meier. Pour plus de détails, on renvoie au livre de Shorack et Wellner [94, 2009].

Dans le chapitre 2 nous nous intéressons à une famille particulèire de lois : les lois à queue de type Weibull . Ces lois possèdent une fonction de survie qui décroit à une vitesse exponentielle (on parle aussi de *queue légère*).Des exemples de telles lois sont les lois exponentielle, normale, gamma, ect ... La vitesse de convergence de la queue de distribution est controlée par un paramètre de forme appelé indice de queue de Weibull. Nous introduisons et étudions le comportement asymptotique d'estimateurs de cet indice et des quantiles extrêmes. Nous présentons aussi une méthode de réduction du biais basée sur un modèle de régression exponentielle.

Dans le chapitre 3 nous nous intéressons à L'analyse de survie est une branche de statistique souvent liée à l'étude des durées de survie dans les applications médicales. En plus de la mort d'organismes biologiques, l'analyse de survie peut s'étendre et s'intéresser à l'échec de systèmes mécaniques et électroniques, dans ce cas on l'appelle "analyse de fiabilité". Ce thème trouve aussi beaucoup d'applications dans les sciences sociales, économiques et actuarielles, où on l'appelle "analyse de durée". Pour des raisons de commodité, les termes propres à la survie biologique sont souvent les plus utilisés.

CHAPITRE 1

ELÉMENTS DE LA THÉORIE DES VALEURS EXTRÊMES

1.1 Introduction

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

La Théorie des Valeurs Extrêmes (TVE) a pour objectif l'étude du comportement asymptotique des grandes ou petites observations d'un échantillon de variables aléatoires indépendantes et identiquement distribuées (iid). L'approche classique en théorie de probabilités s'intéresse au comportement moyen et à la variabilité des phénomènes autour de la moyenne par le biais d'outils probabilistes comme par exemple la loi des grands nombres ou le théorème central limite. Le théorème fondamental de la Théorie des Valeurs Extrêmes, connu sous le nom de théorème de Fisher-Tippett [2, 1928], donne quant à lui les lois limites possibles du maximum de l'échantillon et permet ainsi d'avoir une certaine connaissance sur le comportement stochastique de la queue de distribution.

L'utilisation des lois des valeurs extrêmes repose sur des propriétés des statistiques d'ordre et sur des méthodes d'extrapolation. Plus précisément, elle repose principalement sur les distributions limites des extrêmes et leurs domaines d'attraction. En pratique, on souhaite estimer des petites probabilités ou des quantités dont la probabilité d'observation est très faible, c'est-à-dire proche de zéro. De plus, elle ne vise pas à modéliser ou à estimer la fonction de répartition inconnue F dans son ensemble, mais seulement ses queues de distribution qui sont utiles à la représentation des extrêmes. Les deux approches utilisées sont : la méthode des blocs de maxima et l'ap-

proche de dépassement de seuil ou Peaks Over Threshold (POT) qui modélise la loi des variables excédant un certain seuil fixé. Dans la section (1.2), on s'intéressera à l'approche des blocs de maxima et aux différents résultats théoriques de la théorie des valeurs extrêmes univariées. Avant d'entamer la section(1.2), nous présentons quelques exemples pratiques tirés de [6, 2001] qui illustrent différentes situations dans les modélisations utilisant la TVE.

Exemple 1.1 *La figure (1.1) illustre les maxima annuels du niveau de la mer au Port Pirie dans le sud de l'Australie durant la période 1923-1987. A partir de ces données, nous pouvons faire des prédictions sur le niveau maximal de la mer susceptible de se reproduire dans cette région au cours des 100 ou 1000 ans à venir. Cependant, l'estimation du niveau de retour pour de longues périodes peut être confrontée à des incertitudes dues aux changements climatiques.*

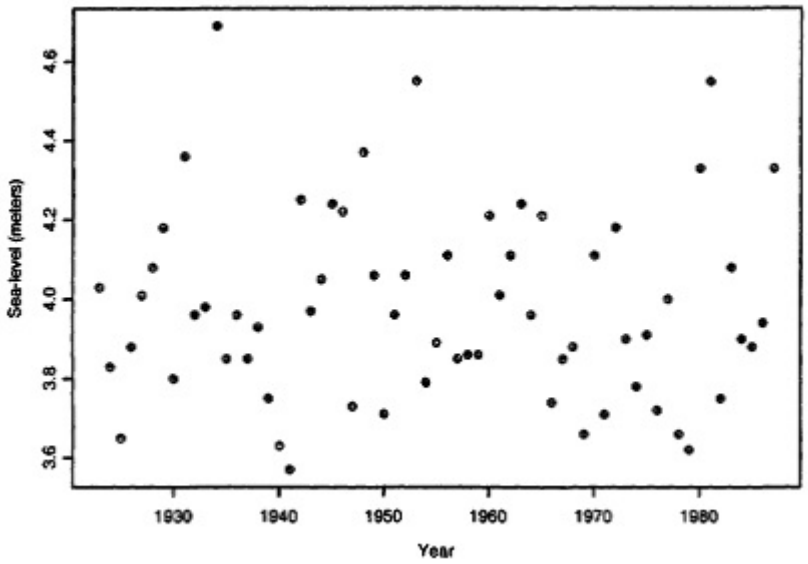

FIGURE 1.1 – Maxima annuels du niveau de la mer au Port Pirie, sud de l'Australie Coles (2001

1.2 Approche des maxima ou des minima par blocs

Dans cette approche, on regroupe les observations par blocs de taille fixée, et on s'intéresse à la loi des maxima (minima) sur chaque bloc pour une loi des valeurs extrêmes généralisées (GEV). Il faut alors trouver un compromis entre la taille des blocs qui doivent être assez grands pour que l'approximation par la loi GEV soit réaliste et le nombre de blocs qui doit être assez grand pour obtenir une estimation précise des trois paramètres de la GEV. En pratique, le choix retenu est souvent de considérer les maxima (minima) annuels pour éviter les effets saisonniers, ce qui nécessite d'observer

le phénomène sur de nombreuses années. Cependant, dans les domaines de l'industrie par exemple, les données ne sont souvent pas à l'échelle annuelle. Cette méthode présente l'inconvénient bien connu de ne pas prendre en compte toutes les informations présentes dans les données sur les évènements extrêmes puisqu'elle ne conserve qu'une valeur par bloc pour l'inférence statistique.La figure (1.2) illustre l'approche par une division en n blocs ; les batons en rouge indiquent les maxima des blocs tandis que ceux en jaune indiquent les minima.

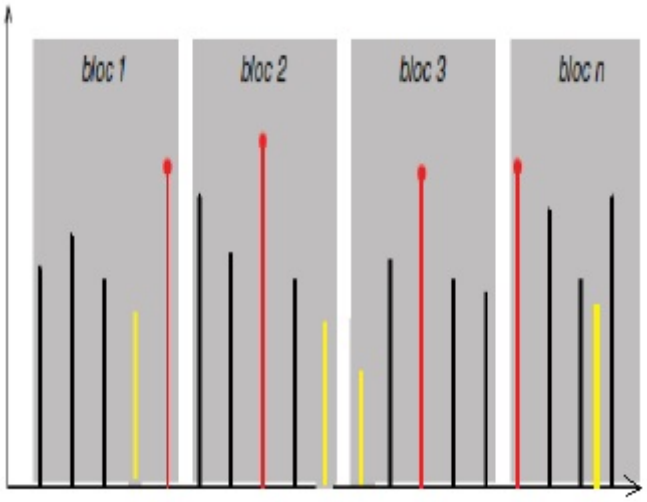

FIGURE 1.2 – Approche des blocs de maxima (minima)

1.2.1 Comportement asymptotique des extrêmes

Nous présentons ici une synthèse de la théorie des valeurs extrêmes univariées. Pour plus de détails et d'éventuelles démonstrations, on pourra se référer à Embrechts et al.[1, 1997] . Soit $(X_n)_{n\geq 1}$ une suite de copies indépendantes d'une variable aléatoire $(v.a)$ X et de fonction de distribution $F(x) = P(X \leq x)$.Rangeons ces variables aléatoires par ordre croissant, on notera dans la suite l'échantillon ordonné :

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

Une manière simple d'étudier le comportement des évènements extrêmes est de considérer les variables aléatoires :

$$M_n = \max\{X_1, ..., X_n\} \text{ et } W_n = \min\{X_1, ..., X_n\}.$$

Il faut noter que M_n et W_n représentent, respectivement, la plus grande et la plus petite valeur observées sur les n valeurs observées $X_1, ..., X_n$. Comme les variables aléatoires sont (iid), on obtient pour toute réalisation x .

$$\begin{aligned} P(M_n \leq x) &= P\left[\cap_{i=1}^{n}(X_i \leq x)\right] \\ &= \prod_{i=1}^{n} P(X_i \leq x) \\ &= \prod_{i=1}^{n} F(x) \\ &= [F(x)]^n . \end{aligned}$$

La difficulté provient du fait que l'on ne connaît pas, en général, la fonction de répartition F . C'est la raison pour laquelle on s'intéresse au comportement asymptotique de la $v.a\ M_n$, Ainsi, en exhibant la famille de loi vers laquelle M_n convenablement normalisée va converger, on pourra remplacer F par cette dernière pour les grandes valeurs de n. De plus, les valeurs extrêmes se trouvent à droite et à la fin du support de la distribution et intuitivement le comportement asymptotique du maximum M_n caractérise la fin de la distribution Deme (2013). On notera par $x_F = sup\{x \in \mathbb{R}, F(x) < 1\}$, le point terminal de F, c'est-à-dire la borne supérieure du support de F, Ce point terminal peut être fini ou infini, pour plus de détails voir Embrechts et al. [1, 1997].

Définition 1.1 *Soient F_1 et F_2 deux fonctions de répartition. On dit que F_1 et F_2 sont du même type si et seulement si il existe $a \in \mathbb{R}_+^*$ et $b \in \mathbb{R}$ tels que $F_1(ax+b) = F_2(x)$.*

Remarque 1.1 *Deux fonctions de répartition de même type sont donc égales modulo un paramètre d'échelle et de position.*

Théorème 1.1 *Fisher et Tippet [2, 1928]; Gnedenko [3, 1943] . Soit $(X_n)_{n\geq 1}$ une suite de n variables aléatoires iid et de même loi de probabilité F telle que$F(x) = P(X \leq x)$. S'il existe deux suites normalisantes réelles $(a_n > 0, b_n \in \mathbb{R}, n \geq 1)$ et une loi non dégénérée G telle que :*

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

alors G est du même type qu'une des trois types de loi suivantes : Loi de Gumbel $(\xi = 0)$:

$$\Lambda_{\mu,\sigma}(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), x \in \mathbb{R} \tag{1.1}$$

Loi de Fréchet$(\xi > 0)$:

$$\Phi_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{\frac{-1}{\xi}}\right) & si\ x > \mu \\ 0 & si\ x < \mu \end{cases}$$

Loi de Weibull($\xi < 0$) :

$$\Psi_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left(-\frac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right) & si\ x < \mu \\ 1 & si\ x \geq \mu \end{cases}$$

Remarque 1.2 *La loi de Weibull définie précédemment n'est pas la loi de Weibull standard qui est plutôt définie sur* $\mathbb{R}_+$.

Les trois lois de probabilité ci-dessus sont appelées lois des valeurs extrêmes. Pour la preuve, nous renvoyons le lecteur à Resnick[4, 1987] et avec des développements dans Embrechts et al.[1, 1997,p.152] . Il faut aussi signaler que chacune des trois lois des valeurs extrêmes peut s'obtenir par une transformation fonctionnelle de l'autre. D'une manière analogue, on définit les lois des valeurs extrêmes assocées au minimum. Le théorème suivant établit l'unification des trois types de loi en une loi unique dite distribution généralisée des valeurs extrêmes pour le maximum

Théorème 1.2 *Resnick[4, 1987] Soit* $(X_n)_{n\geq 1}$ *une suite de variables aléatoires iid de fonction de répartition F. S'il existe deux suites normalisantes réelles* $(a_n)_{n\geq 1} > 0$ *et* $(b_n)_{n\geq 1} \in \mathbb{R}$ *et une loi non dégénérée G telle que :*

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

alors G est donnée par :

$$G_{\mu,\sigma,\xi}(x) = \exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}\right), \tag{1.2}$$

où x est tel que :

$$1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\ ,\ -\infty < \mu < +\infty, \sigma > 0\ \ ,\ \ et\quad -\infty < \mu < +\infty.$$

Les paramètres ξ *et* μ *et* σ *; et sont respectivement les paramètres de position, d'échelle et de forme.*

Une preuve détaillée de ce théorème est donnée dans l'ouvrage de Resnick[4, 1987] et pour plus de détails voir Embrechts et al.[1, 1997] et Galambos [5, 1985]. Le comportement limite du maximum normalisé est ainsi décrit par la fonction de répartition G_ξ pour la plus grande partie des lois usuelles. Ainsi, G_ξ est appelée fonction de répartition de la loi des valeurs extrêmes, en anglais (Generalized Extreme Value distribution) notée GEV. Pour plus de détails voir Resnick [4, 1987], Embrechts et al. [1, 1997]. L'unification de la loi standard de la distribution généralisée des valeurs extrêmes en

une seule fonction de répartition facilite l'étude du comportement du maximum ou du minimum. De plus, pour $\mu = 0$ et $\sigma = 1$; on obtient la forme standard des trois types de loi des valeurs extrêmes. Cette loi dépend du seul paramètre de forme appelé ξ indice des valeurs extrêmes, ainsi la relation devient 1.2 :

$$G_\xi = \exp\left(-(1+\xi x)^{\frac{-1}{\xi}}\right), \xi \neq 0, \text{avec } (1+\xi x) > 0. \tag{1.3}$$

Le cas $\xi = 0$, dans la relation peut être vu comme le cas limite lorsque $\xi \to 0$.On retrouve alors la loi de Gumbel ayant pour fonction de répartition :

$$G_0 = \exp(-\exp(-x)).$$

La figure (1.3,1.4)représente les fonctions de répartition et de densité pour la forme standard unifiée des lois des valeurs extrêmes .

1. Pour $\xi = 1$,loi de Fréchet, en bleu, où le support de la loi est $[-1, +\infty[$.

2. Pour $\xi = -1$,loi de Weibull en rouge où le support de la loi est $]-\infty, 1]$.

3. Pour $\xi = 0$, loi de Gumbel en noir où le support de la loi est $\mathbb{R}$.
   Souvent on indique les limites des supports des lois sur les graphiques.

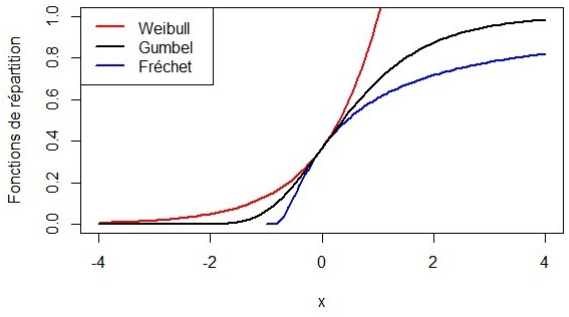

FIGURE 1.3 – fonctions de répartition

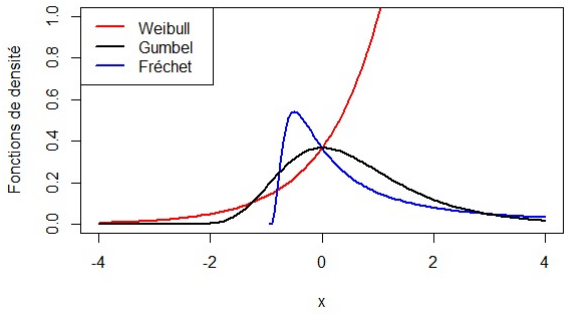

FIGURE 1.4 – fonctions de densité des lois des valeurs extrêmes

En outre, dans la TVE les informations les plus importantes se trouvent dans la-queue de la distribution et sont caractérisées par le paramètre de formeξ . En effet selon son signe, on distingue trois domaines d'attraction que l'on présentera dans la suite.

1.2.2 Domaines d'attraction des lois GEV

Il faut noter que le paramètre de forme ξ conditionne le type de la loi des valeurs extrêmes. Nous présentons dans ce qui suit les domaines d'attraction et les concepts théoriques qui leur sont associés dans les trois cas correspondant au signe du paramètre de forme ξ .

Définition 1.2 *On dit qu'une distribution F appartient au domaine d'attraction de H_ξ , et on note $F \in D\left(H_\xi\right)$ s'il existe des suites réelles $(a_n) > 0$ et $b_n \in \mathbb{R}$ telles que :*

$$\lim_{n\to+\infty} F^n\left(a_n x + b_n\right) = H_\xi\left(x\right)$$

1.2.2.1 Notion de fonction à variation régulière

La notion de fonction à variation régulière est très utilisée dans le contexte de lacaractérisation des domaines d'attraction dans la théorie des valeurs extrêmes. Nous présentons ici quelques résultats principaux, pour plus de détails voir Bingham et al.[7, 1987].

Définition 1.3 *Une fonction $f : \mathbb{R}_+ \to \mathbb{R}_+$mesurable au sens de Lebesgue, est à variations régulières à l'infini si et seulement s'il existe un réel α tel que pour tout $x > 0$*

$$\lim_{t\to\infty} \frac{f\left(tx\right)}{f\left(x\right)} = x^\alpha$$

et on écrit $f \in \Re v_\alpha$, est appelé indice (ou exposant) de la fonction à variation régulière f.

Remarque 1.3 *Pour $\alpha = 0$, on retrouve le cas de la fonction à variation lente dans la définition suivante.*

Définition 1.4 *Une fonction L est dite à variation lente si $L(t) > 0$ pour t assez grand et si pour tout $x > 0$, on a :*

$$\lim_{t\to\infty} \frac{L(tx)}{L(x)} = 1$$

Théorème 1.3 *Resnick [4, 1987] Représentation de Karamata Toute fonction à variation lente L à l'infini s' écrit sous la forme*

$$L(x) = c(x)\exp\left(\int_1^x \Delta(t)\, t^{-1} dt\right), \tag{1.4}$$

où $c(.) > 0$ et $\Delta(.)$ sont deux fonctions mesurables telles que

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

Si la fonction $c(.)$ est une constante, alors on dit que L est normalisée. La relation 1.4 implique que si L est normalisée alors L est dérivable, de dérivée L' avec pour

$$\text{tout } x > 0, L'(x) = \frac{\Delta(x)L(x)}{x}. \text{ en particulier , on a } \lim_{x\to\infty} \frac{xL'(x)}{L(x)} = 0$$

1.2.2.2 Domaine d'attraction de la loi de Gumbel

La loi présente dans la queue une décroissance de type exponentiel, ce qui permet de caractériser dans ce cas, le domaine d'attraction de Gumbel noté $D(\Lambda)$. La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe car il n'y a pas de lien direct entre la queue de la loi et les fonctions à variation lente, pour plus de détails voir Delmas et Jourdain [8, 2006]. VonMises [9, 1936] a donné une caractérisation simple pour le domaine d'attraction de Gumbel, formulée par le biais du théorème suivant :

Théorème 1.4 *VonMises [9, 1936] S'il existe une fonction mesurable R, appelée fonction auxiliaire telle que :*

$$\lim_{x\to x_F} \frac{1 - F(t + xR(t))}{1 - F(t)} = \exp(-x)$$

où $x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$, et le point terminal de F, alors $F \in D(\Lambda)$.

1.2.2.3 Domaine d'attraction de la loi de Weibull

Les lois de ce domaine sont bornées à droite et, par conséquent, le point terminal x_F est fini , le domaine d'attraction de Weibull est noté $D\left(\Psi_\xi\right)$: Une caractérisation

d'appartenance à ce domaine d'attraction est donnée par le théorème suivant, pour la preuve voir Gnedenko [3, 1943] :

Théorème 1.5 *Gnedenko [3, 1943] Une fonction de répartition F appartient au* $D\left(\Psi_\xi\right)$ *si et seulement si* $x_F < \infty$ *et*

$$\overline{F}\left(x_F - \frac{1}{x}\right) = x^{\frac{-1}{\xi}} L(x)$$

avec : $\overline{F}$ *est la fonction de survie donnée par* $\overline{F}(x) = 1 - F(x)$, *L est une fonction à variation lente.*

1.2.2.4 Domaine d'attraction de la loi de Fréchet

Les lois appartenant à ce domaine d'attraction sont caractérisées par une queue à décroissance lente (polynomiale) à l'infini, et un point terminal $x_F = +\infty$: Elles sont dites aussi lois à queues lourdes. Une caractérisation de ce domaine d'attraction noté $D\left(\Phi_\xi\right)$ est donnée par le théorème suivant, pour la preuve voir Gnedenko [3, 1943] :

Théorème 1.6 *VonMises [9, 1936] Une fonction de répartition F appartient au* $D\left(\Phi_\xi\right)$ *si et seulement si sa fonction de survie est donnée par :*

$$\overline{F}(x) = x^{\frac{-1}{\xi}} L(x)$$

où L est une fonction à variation lente.

Remarque 1.4 *Les variables aléatoires dans les trois domaines d'attraction de Gumbel, de Fréchet et de Weibull sont liées par la relation suivante, pour plus de détails voir Embrechts et al. [1, 1997] :*

$$X \in D\left(\Phi_\xi\right) \Longleftrightarrow \log\left(X^{\frac{1}{\xi}}\right) \in D(\Lambda) \Longleftrightarrow -X^{-1} \in D\left(\Psi_\xi\right)$$

Le tableau suivant présente quelques lois qui appartiennent aux domaines d'attraction de Gumbel, Fréchet et Weibull.

TABLE 1.1 – d'attraction et quelques lois associées

Domaines d'attraction	Gumbel $\xi = 0$	Fréchet $\xi > 0$	Weibull $\xi < 0$
Lois	Normale	Cauchy	Uniforme
	Exponentielle	Pareto Généralisée	reverse Burr
	Log normale	Student	Beta
	Weibull	Log-gamma	
	Gumbel		

Exemple 1.2 *Considérons une variable aléatoire X suivant une loi exponentielle de paramètre* 1 *,Sa fonction de répartition est donnée par :*

$$F(x) = \begin{cases} 1 - \exp(-x) si\ x \geq 0 \\ 0 \qquad\qquad si\ x \leq 0 \end{cases}$$

Le support de la loi étant $\mathbb{R}_+$ *on a* $x_F = +\infty$ *d'où* $X_{n:n} \to +\infty$ *Effectuons la normalisation suivante :*

$$\begin{aligned} P\left(\frac{X_{n:n} - \log(n)}{1} \leq x\right) &= P(X_{n:n} \leq x + \log(n)) \\ &= [F(x + \log(n))]^n \\ &= [1 - \exp(-x - \log(n))] \\ &= \left(1 - \frac{\exp(-x)}{n}\right) \\ &\to \exp(-\exp(-x)), n \to +\infty. \end{aligned}$$

Il en résulte donc, d'après la relation 1.1, que $X_{n,n}$ *converge vers une loi de Gumbel après normalisation.*

1.2.3 Méthodes d'estimation des paramètres des lois GEV

Les lois GEV se distinguent fondamentalement selon le paramètre ξ . De ce fait, on distinguera trois cas dans l'estimation des paramètres. Le premier cas correspond à la loi de Gumbel pour $\xi = 0$; le deuxième correspond à celle de Weibull pour $\xi < 0$ et le troisième correspond à celle de Fréchet pour $\xi > 0$. Dans la littérature, il existe plusieurs méthodes d'estimation des paramètres des lois GEV. On peut citer par exemple, les méthodes d'estimation empirique Gumbel et Mustafi [11, 1967], la méthode du maximum de vraisemblance Smith [12, 1987], Prescott et Walden [13, 1980], la méthode des moments Christopeit [14, 1994], la méthode des moments de probabilité pondérés Greenwood et al. [15, 1979]. La comparaison de ces méthodes d'estimation s'effectue généralement dans un cadre empirique, pour plus de détails voir Benkhaled [16, 2007]. Par ailleurs, d'autres approches principalement non paramètriques, ont été proposées pour l'estimation de l'indice de queue. A titre d'exemple, on peut citer l'estimateur de Pickands, Pickands [17, 1975], l'estimateur de Hill, Hill [18, 1975] pour le modèle de Fréchet et l'estimateur de Dekkers, Dekkers et al. [19, 1989].

1.2.3.1 Méthodes d'estimation pour les lois du domaine d'attraction de Gumbel

a) Méthode du maximum de vraisemblance

Cette méthode consiste à chercher les paramètres $\omega = (\mu, \sigma)$ qui maximisent la

fonction de vraisemblance d'un échantillon de taille n est

$$\ell(x_1,...,x_n,\mu,\sigma) = \prod_{i=1}^{n} f(x_1,...,x_n\mu,\sigma)$$

Par suite, sa log-vraisemblance est :

$$L_{MV}(\varpi) = \log l(\varpi)$$

en résolvant le système suivant représentant les conditions nécessaires d'optimalité :

$$\begin{cases} \frac{\partial L_{MV}(\varpi)}{\partial \mu} = 0 \\ \frac{\partial L_{MV}(\varpi)}{\partial \sigma} = 0 \end{cases}$$

on obtient une estimation des paramètres $\varpi = (\mu,\sigma)$. Cependant, la solution n'est pas explicite et il faut passer par une résolution numérique en utilisant par exemple la méthode de Newton-Raphson.

b) Méthode des moments pour la loi Gumbel

La méthode consiste à associer respectivement les deux premiers moments m_1 et m_2, respectivement à la moyenne $\left(\overline{X_n}\right)$ et à la variance empirique S_X^2.Dans le cas d'une variable aléatoire issue d'une loi de Gumbel, la méthode des moments nous conduit au système suivant :

$$\begin{cases} E(X) = \mu + \gamma\sigma \\ V(X) = \frac{1}{6}\pi^2\sigma \end{cases}$$

où $\gamma \approx 0.57721$ représente la constante d'Euler. La résolution de ce système d'équations nous permet d'obtenir les valeurs estimées $\widehat{\mu}$ et $\widehat{\sigma}$:

$$\begin{cases} \widehat{\sigma} = \frac{\sqrt{6}S_X}{\pi} \\ \widehat{\mu} = \overline{X_n} - \gamma\widehat{\sigma} \end{cases}$$

c) Méthode des moments pour la loi Weibull à deux paramètres

La méthode consiste à associer respectivement les deux premiers moments théoriques m_1 et m_2 aux deux premiers moments observés à savoir la moyenne $\left(\overline{X_n}\right)$ et la variance empirique σ^2. La méthode des moments est une autre technique utilisée dans l'estimation des paramètres. Nous présentons ici la procédure pour la loi de Weibull à deux paramètres (k,λ), définie par :

$$\begin{cases} f_{k,\lambda}(y) = \frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1}\exp^{-\left(\frac{y}{\lambda}\right)^k} \\ F_{k,\lambda(y)} = 1 - \exp^{-\left(\frac{y}{\lambda}\right)^k} \end{cases} \tag{1.5}$$

où f et F sont respectivement les fonctions de densité et de répartition. Soit $y_1,...,y_n$ un

ensemble de données pour lesquelles nous allons chercher les deux premiers moments empiriques à partir de la relation 1.6 définie comme suit :

$$\widehat{M_m} = \frac{1}{m}\sum_{i=m}^{m} y_i^m, \tag{1.6}$$

où $\widehat{M_m}$ est une estimation de M_m. Dans la distribution de Weibull, le moment d'ordre m peut être obtenu à travers la fonction Gamma $\Gamma(y)$. De plus, la moyenne des observations peut aussi être exprimée en fonction des paramètres k et λ. Pour la distribution de Weibull donnée dans la relation 1.5, le moment d'ordre m, est défini par :

$$\mu_m = \left(\frac{1}{\lambda}\right)^{\frac{m}{k}} \Gamma\left(1+\frac{m}{k}\right), \tag{1.7}$$

où la fonction Γ est définie par $\Gamma(s) = \int_0^{+\infty} q^{s-1}e^{-q}dq, (s>0)$.A partir de la relation 1.7, nous pouvons trouver les moments d'ordre 1 et 2 comme suit :

$$\mu_1 = \left(\frac{1}{\lambda}\right)^{\frac{1}{k}} \Gamma\left(1+\frac{1}{k}\right) = M_1 \tag{1.8}$$

$$\mu_2^2 = \mu_1^2 + \sigma^2 = \left(\frac{1}{\lambda}\right)^{\frac{2}{k}} \left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right] = M_2 \tag{1.9}$$

où σ^2 est la variance, M_1et M_2 sont respectivement la moyenne arithmétique et le moment centré d'ordre 2. Lorsqu'on divise M_2 par le carré de M_1 ; on obtient l'expression suivante :

$$\frac{M_2}{M_1^2} = \frac{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)}{\Gamma^2\left(1+\frac{1}{k}\right)} \tag{1.10}$$

En prenant la racine carrée de la relation 1.10, on obtient le coefficient de variation noté C_v, défini comme suit :

$$C_v = \frac{\sqrt{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)}}{\Gamma\left(1+\frac{1}{k}\right)}$$

Ainsi, les valeurs de k et λ peuvent être obtenues par le système d'équations suivant :

$$\begin{cases} \sigma = \left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right]^{\frac{1}{2}} \\ C_v = \sqrt{\frac{\Gamma\left(1+\frac{2}{k}\right)}{\Gamma^2\left(1+\frac{1}{k}\right)} - 1} \end{cases} \tag{1.11}$$

Après certaines transformations, voir Justus et al. [20, 1977], Abul et al. [21, 2014],

nous avons :

$$k = \left(\frac{0,9874}{C_v}\right), \lambda = \frac{\overline{y}}{\Gamma\left(1+\frac{1}{k}\right)} \tag{1.12}$$

1.2.3.2 Méthodes d'estimation pour les lois des domaines d'attraction de Fréchet et Weibull

Il existe plusieurs méthodes d'estimation des paramètres pour les lois de Fréchet et de Weibull, telles la méthode du maximum de vraisemblance, la méthode des moments, la méthode des moments de probabilité pondérés, etc. Mais dans cette partie, nous présentons celle du maximum de vraisemblance commune aux deux domaines d'attraction.

1.2.3.2.1 Méthode du maximum de vraisemblance La méthode consiste à chercher les paramètres $\theta = (\mu, \sigma, \xi)$ qui maximisent la fonction de log-vraisemblance. La maximisation conduit à résoudre le système suivant :

$$\begin{cases} \frac{\partial L_{MV}(\theta)}{\partial \mu} = 0 \\ \frac{\partial L_{MV}(\theta)}{\partial \sigma} = 0 \\ \frac{\partial L_{MV}(\theta)}{\partial \xi} = 0 \end{cases}$$

Les problèmes relatifs à l'estimation par la méthode du maximum de vraisemblance ont été étudiés par Smith [12, 1985]. Les résultats obtenus sont :

1. Si $\xi > -0.5$, alors les estimations du maximum de vraisemblance ne possèdent pas de propriété asymptotique telles que la convergence vers la vraie valeur du paramètre inconnu, l'invariance par rapport à une transformation paramétrique et l'efficacité asymptotique.
2. Si $-1 > \xi > -0.5$, alors les estimateurs du maximum de vraisemblance ne possèdent pas de propriétés asymptotiques standards.
3. Si $\xi < -1$ alors l'obtention des estimateurs du maximum de vraisemblance n'est pas garantie.

1.2.4 Approche des dépassements de seuil : loi GPD

L'approche par dépassements de seuil, en anglais (Peaks-Over Threshold) notée POT, repose sur l'utilisation des statistiques d'ordre supérieur de l'échantillon. Elle consiste à ne conserver que les observations dépassant un certain seuil. L'excès au-delà du seuil est défini comme l'écart entre l'observation et le seuil. Considérons un échantillon de variables aléatoires i.i.d $Y_1, ..., Y_n$: Soit u un seuil fixé tel que $u < y_F$ et

les N_u observations $Y_{i_1},....,Y_{i_{N_u}}$ dépassant le seuil u, On appelle excès au-delà du seuil u les Z_j définis par $Z_j = Y_{ij} - u$, pour $j = 1,...,N_n$ voir la figure (1.5).

La figure (1.5) est composée de deux graphiques sur celui de gauche on peut remarquer le seuil u en trait horizontal et les observations au-dessus du seuil, par exemple Y_2 et Y_{10}, Quant à celui de droite, il indique le seuil u et les excès au-dessus de ce seuil.

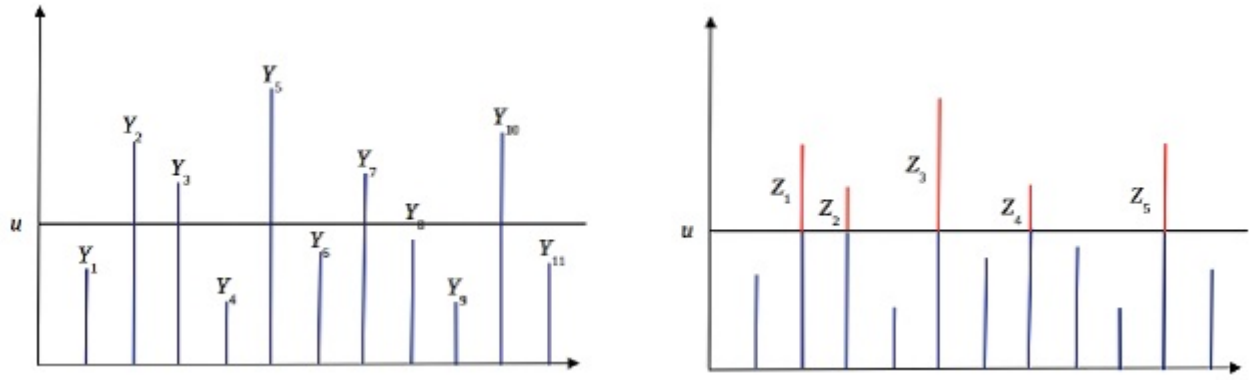

FIGURE 1.5 – Excès au-dessus du seuil

1.2.4.1 Modélisation des excès

L'approche des blocs de maxima conduit souvent à des pertes d'information lors de la sélection des maxima ou des minima par blocs. La prise en compte de cette insuffisance a conduit à une nouvelle approche dite des dépassements de seuil Peaks Over Threshold (POT). De plus, certains blocs peuvent contenir plusieurs valeurs extrêmes provenant de la distribution initiale, alors que d'autres peuvent ne pas en contenir, pour plus de details voir Katz [22, 2002]. Contrairement à l'approche des maxima par blocs, la méthode POT consiste à utiliser toutes les observations, appelées excès, qui dépassent un certain seuil suffisamment élevé. L'objectif est d'analyser leur comportement asymptotique. La méthode POT a été proposée par Pickands [17, 1975] et reprise par NECIR et RASSOUL [23, 2010]. Les premières applications étaient dans le domaine de l'hydrologie. Concernant les aspects théoriques de la méthode, ils ont été abondamment développés par plusieurs auteurs. On peut citer Todorovic et Zelenhasic [25, 1970], Todorovic et Rousselle [24, 1971], Smith [12, 1987], Davison et Smith [26, 1990], Reiss et Thomas [27, 2001]. L'idée d'utiliser un nombre croissant de statistiques d'ordre de l'échantillon a ensuite été plus largement développée dans le cadre de l'approche POT, via l'approximation de la loi des excès au-delà d'un seuil par des Generalized Pareto Distribution (GPD). Plus précisément, soit $u < y_F$ et F_u la fonction de répartition des excès définie par

$$F_u(y) = \Pr(Z \leq y \mid Y > u) = \Pr(Y - u \leq y \mid Y > u) = \frac{F(u+y) - F(u)}{\overline{F}(u)},$$

où $\overline{F} = 1 - F$ est la fonction de survie. On a alors le résultat suivant dû à Balkema et de Haan [28, 1974] et Pickands [17, 1975].

Théorème 1.7 *Balkema et de Haan [28, 1974], Pickands [17, 1975] Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes, (Fréchet, Gumbel ou Weibull), alors il existe une fonction σ_u strictement positive et un $\xi \in \mathbb{R}$ tel que*

$$\lim_{u \to y_F} \sup_{y \in]0, y_F - u[} \left| F_u(y) - G_{\xi,\sigma_u}(y) \right| = 0;$$

où $y_F = \sup\{y \in \mathbb{R}, F(y) < 1\}$ est le point terminal de F et G_{ξ,σ_u} est la fonction de répartition de la loi de Pareto Généralisée définie par :

$$G_{\xi,\sigma_u}(y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma_u} y\right)^{\frac{-1}{\xi}}, si\ \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma_u}\right), \quad si\ \xi = 0 \end{cases} \tag{1.13}$$

Remarque 1.5 *Selon le signe de ξ ; nous avons les cas suivants :*
$\xi > 0$: distribution de type Pareto à queue lourde,
$\xi < 0$: distribution de type Beta bornée au dessus de $u - \frac{\sigma_u}{\xi}$,
$\xi = 0$: distribution de type exponentielle à queue légère.

La figure (1.6 et 1.7) représente les fonctions de répartition et de densité des lois de Pareto pour un paramètre d'échelle fixé à 1. Ainsi, en faisant varier le paramètre de forme entre -1 et 1, on obtient différentes lois GPD comme suit,$\xi = -1$ pour la loi de Pareto II en rouge,$\xi = 0$ pour la loi exponentielle en noir et $\xi = 1$ pour la loi de Pareto en bleu. En remplaçant par exemple dans la relation 1.13 $\xi = 0$ et $\sigma_u = 1$, on obtient respectivement les fonctions de répartition et de densité suivantes :

$$G_{0,1}(y) = 1 - \exp(-y) \text{ et } g_{0,1} = \exp(-y)$$

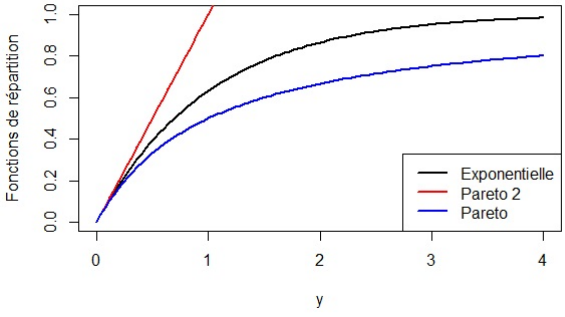

FIGURE 1.6 – fonctions de répartition

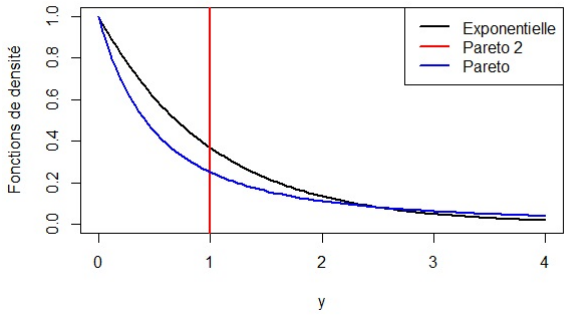

FIGURE 1.7 – fonctions de densité des lois de Pareto Généralisée

1.2.4.2 Méthodes pour la détermination du seuil

La détermination du seuil est l'étape la plus délicate dans l'implémentation de l'approche POT, étant donné que la qualité du modèle en dépend. La convergence des excès vers une GPD passe par la détermination d'un seuil adéquat (pas très bas et pas très haut). Par la suite, il faut estimer les paramètres et les niveaux de retour. Le choix du seuil doit être un compromis de sorte que le seuil déterminé soit suffisamment grand pour pouvoir utiliser les résultats asymptotiques, mais pas trop élevé afin d'obtenir des estimations précises. Cependant, le choix d'un seuil faible peut conduire à des incertitudes sur le nombre d'observations extrêmes et par conséquent produire des estimations biaisées et une mauvaise approximation de la loi asymptotique. Dans cette optique, plusieurs méthodes de détermination de seuils sont proposées dans la littérature. On distingue principalement deux approches, l'approche graphique et celle dite numérique. La plupart de ces méthodes sont subjectives et il est nécessaire de quantifier les incertitudes dues à ces méthodes.

1.2.4.2.1 Méthodes graphiques

a- Fonction moyenne des excès

La fonction moyenne des excès, obtenue par une méthode graphique, peut être illustrée de deux façons grâce au logiciel **R** :

1. Par la fonction Mrlplot « mean excess function »
2. Par la fonction Meplot « mean residual life plot ».

Elle est définie comme suit :

Définition 1.5 *On appelle fonction moyenne des excès, mean excess function ou Mean residual life plot au-dessus d'un seuil u, la fonction $e(u)$ définie par :*

$$e(u) = E[Y - u \mid Y > u]$$

avec $0 \leq u \leq y_F$.

Dans le logiciel R, elle est donnée par la fonction « Meplot ou Mrlplot » ; ces deux fonctions jouent le même rôle mais présentent la linéarité sous deux formes différentes. La moyenne des excès peut être utilisée pour guider le choix du seuil adéquat u^* :

Proposition 1.1 *Si* $[Z = Y - u^* \mid Y > u^*] \backsim G_{\xi,\sigma_{u^*}}, e(u)$ *est linéaire en* u *pour* $u > u^*$,*Coles [6, 2001].*

Si :

$$Z \backsim G_{\xi,\sigma_{u^*}} alors E(Z) = \frac{\sigma_{u^*}}{1-\xi} = C$$

.

Donc pour $u > u^*, E[Y - u \mid Y > u] = \frac{\sigma_{u^*} + \xi u}{1-\xi} = C + \frac{\xi}{1-\xi} u$,où $C = E(Z)$.
Ainsi, si au-dessus d'un seuil u^*, la GPD une approximation de la loi des excès $Z = Y - u^*$ avec $\xi < 1$, alors on a :

$$E[Y - u^* \mid Y > u^*] \approx \frac{\sigma_{u^*}}{1-\xi}.$$

Comme l'approximation par une GPD reste valable pour tout seuil $u > u^*$, alors on a :

$$E[Y - u \mid Y > u] \approx \frac{\sigma_{u^*}}{1-\xi}$$

avec $\sigma_u = \sigma_{u^*} + \xi(u - u^*)$.Ainsi, pour tout $u > u^*$:

$$E[Y - u \mid Y > u] \approx \frac{\sigma_{u^*} + \xi(u - u^*)}{1-\xi}.$$

La fonction $e(.)$ est une fonction linéaire de u. Par conséquent, l'identification du seuil u consiste en la recherche de la linéarité sur le graphe défini par $(u, e(u))$ et appelé « Mean Residual Life Plot ou Mean excess plot ». En pratique, la fonction $e(.)$ est approchée sur la base de l'approximation empirique $\widehat{e_n(.)}$:

$$\widehat{e_n(u)} = \frac{\sum_{i=1}^{n} (Y_i - u)^+}{\sum_{i=1}^{n} \left(I_{Y_i > u}\right)} = \frac{1}{N_u} \sum_{i=1}^{n} (Y_i - u)^+ \tag{1.14}$$

c'est-à-dire la somme des excès au-dessus du seuil u divisée par le nombre N_u de données qui excèdent le seuil u :

b- Méthodes des graphes des paramètres d'échelle et de forme (Parameter stability plot), fonction tcplot

Encore appelée « stable scale and shape parameters », cette méthode permet dedéterminer un seuil requis en ajustant les données à une distribution de GPD enutilisant

un seuil différent. La stabilité des paramètres (forme et échelle) peut alors être contrôlée et localisée. Cette technique est implémentée dans le logiciel R avec des packages spécifiques. Ces paquets disposent d'outils objectifs pour guider le choix du seuil adéquat enexaminant tout simplement la stabilité des paramètres de forme et d'échelle et ξ et σ . Ce graphe établit un lien direct entre les valeurs des paramètres estimés (ξ et σ) et les seuils potentiels u^* . Les paramètres estimés au-dessus des seuils sont ceux pour lesquels le modèle GPD devient valable.

Exemple 1.3 *Les figures (1.8) et (1.9) illustrent le (MRLplot) et le (parameter stability plot) pour les données de pluies anglaises, Coles [6] a montré que le meilleur choix du seuil est approximativement* $u = 30mm$ *et qu'un seuil de* $20mm$ *peut être trop bas pour la validité de l'hypothèse des valeurs extrêmes .*

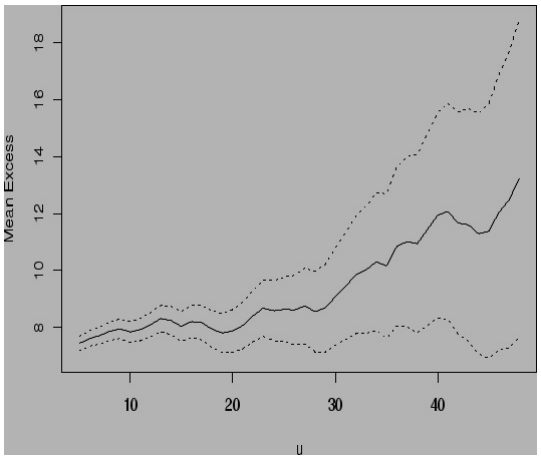

FIGURE 1.8 – Graphe de la fonction moyenne des excès (MRlplot) pour des données de pluie anglaises, (Coles, 2001)

Exemple 1.4 *Sur la figure (1.8), les pointillés encadrant la courbe de la fonction moyenne des excès indiquent les intervalles de confiance associés au seuil.*

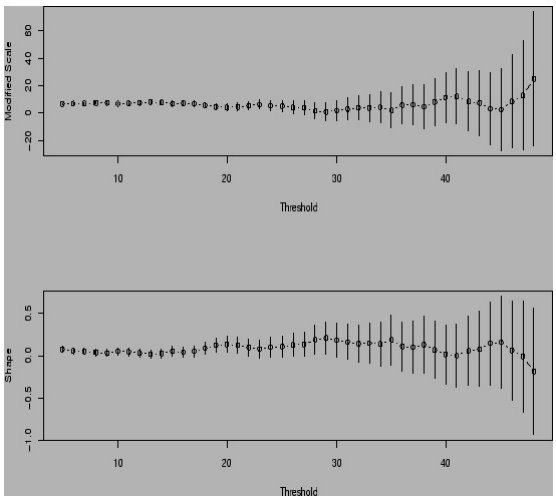

FIGURE 1.9 – Graphe de la stabilité des paramètres (tcplot)

Sur la figure (1.8), on constate une linéarité et une stabilité entre 23 et 32mm. De ce fait, on peut affirmer que le seuil est compris entre 23 et 32mm. En outre, sur la figure (1.9), on remarque une stabilité sur l'intervalle $[25, 32]$; ainsi, le seuil se situer autour de 30mm.

1.2.4.2.2 Une approche de compromis Dans cette approche de compromis, nous combinons les méthodes graphiques et une estimation intensive des paramètres de forme et échelle par différentes méthodes, ainsi que les erreurs standards sur les paramètres estimés pour les seuils potentiels obtenus. Les erreurs standards sur les paramètres estimés pour différents seuils permettent de guider le choix du seuil et de la méthode d'estimation qui sont les principales sources d'incertitudes dans l'approche POT.

1.2.4.3 Estimation des paramètres du modèle GPD

Dans cette sous-section, nous nous intéressons aux différentes méthodes d'estimation du paramètre ξ ou σ qui interviennent dans la distribution asymptotique des valeurs extrêmes. On distingue principalement deux méthodes, à savoir les méthodes paramétriques et non-paramétriques. Parmi les méthodes paramétriques, on peut citer la méthode du maximum de vraisemblance, la méthode des moments, la méthode des moments de probabilités pondérées Hosking et Wallis [29, 1987], les méthodes de régression Beirlant et Goegebeur [30, 2003]. Concernant les méthodes non paramètriques, on peut noter l'estimateur de Pickands [17, 1975], l'estimateur de Hill [18, 1975], l'estimateur des moments Dekkers et al. [19, 1989]. Dans le cadre de ce travail, on s'intéressera aux méthodes du maximum de vraisemblance, des moments et des moments de probabilités pondérées.

1.2.4.3.1 Méthode du maximum de vraisemblance Considérons un échantillon $Y_1, ..., Y_k$ iid de loi GPD G_{ξ,σ_u}, La fonction de log-vraisemblance est obtenue à partir de la loi GPD, ce qui nous donne :

$$\log(L(\xi,\sigma_u)) = -k_n \log(\sigma_u) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k_n}\log\left(1+\frac{\xi}{\sigma_u}Y_i\right), \tag{1.15}$$

lorsque $1+\frac{\xi}{\sigma_u}Y_i > 0$ pour $i = 1,...,k_n$ sinon $(L_{MV}(\xi,\sigma_u)) = -\infty$. La log-vraisemblance définie par la relation 1.15 est maximisée par des méthodes numériques telles que l'algorithme de Newton-Raphson. Les estimateurs du maximum de vraisemblance sont asymptotiquement gaussiens et efficaces pour $\xi > \frac{1}{2}$, pour plus de détails voir Smith [12, 1987]. De plus, Davison et Smith [26, 1990] ont montré que l'estimateur du maximum de vraisemblance a souvent des problémes de convergence et d'efficacité pour des échantillons de petite taille ($n < 50$).

1.2.4.3.2 Méthode des moments

La méthode des moments fut introduite par Hosking et Wallis [29, 1987] pour estimer les paramètres de la loi GPD. La condition d'existence de l'espérance et de la variance d'une variable aléatoire Y de loi GPD G_{ξ,σ_u} est $\xi < \frac{1}{2}$, Pour une GPD, G_{ξ,σ_u} (voir équation 1.13), l'estimation par la méthode des moments (voir Hosking et Wallis [29, 1987]) est basée sur l'hypothèse que

$$E\left[\left(1+\frac{\xi}{\sigma_u}Y\right)^r\right] = \frac{\sigma_u}{1-r\xi}, \text{ si } 1-r\xi > 0$$

Dans ce cas, on a :

$$\begin{cases} E(Y) = \frac{\sigma_u}{1-r\xi} \\ V(Y) = \frac{\sigma_u^2}{(1-\xi)^2(1-2\xi)} \end{cases}$$

On peut donc exprimer les paramètres de la GPD en fonction en fonction de ξ et σen fonction de $E(Y)$ et $V(Y)$, on a donc :

$$\begin{cases} \xi = \frac{1}{2}\left(1-\frac{E(Y)^2}{V(Y)}\right) \\ \sigma = \frac{E(Y)}{2}\left(1+\frac{E(Y)^2}{V(Y)}\right) \end{cases};$$

où $\overline{Y}$ et S_Y^2 sont respectivement les estimateurs empiriques des moments d'ordre 1 et 2 de l'échantillon. Ainsi, en remplaçant $E(Y)$ et $V(Y)$ par leurs estimateurs empiriques

respectifs, il vient que :

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

et on obtient les estimateurs de $\widehat{\xi}$ et $\widehat{\sigma_u}$ définis respectivement pars

$$\begin{cases} \widehat{\xi} = \frac{1}{2}\left(1 - \frac{\overline{Y}^2}{S_y^{*2}}\right) \\ \widehat{\sigma_u} = \frac{1}{2}\overline{Y}\left(\frac{\overline{Y}^2}{S_y^{*2}} + 1\right) \end{cases} \tag{1.16}$$

1.2.4.3.3 Méthode des moments de probabilité pondérés A partir de la théorie des moments pondérés, Hosking et al. (2003) ont proposé une estimation. Cette approche consiste à utiliser les deux moments pondérés M_0et M_1 avec $M_s = E\left[X\left(1 - G_{\xi,\sigma}(Y)\right)^s\right]$. Nous avons

$$M_s = \frac{\sigma_u}{(s+1)(s-\xi+1)}, \text{ si } s > \xi - 1$$

Un estimateur de ce moment est donné par

$$\widehat{M_s} = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1 - \frac{i}{k_n + 1}\right)^s Y_{i,k_n},$$

où $Y_{1,k_n},...,Y_{k_n,k_n}$ sont les statistiques ordonnées associées $Y_1,...,Y_{k_n}$. Cette dernière équation conduit pour $s = 0$ et $s = 1$ aux estimateurs suivants :

$$\begin{cases} \widehat{\xi} = 2 - \dfrac{\widehat{M_0}}{\widehat{M_0} - 2\widehat{M_1}} \\ \widehat{\sigma_u} = 2\dfrac{\widehat{M_0}\widehat{M_1}}{\widehat{M_0} - 2\widehat{M_1}} \end{cases} \tag{1.17}$$

Les estimateurs des moments pondérés sont asymptotiquement gaussiens pour $-1 < \xi < \frac{1}{2}$, voir Hosking et al. [31, 2003]. Ils ont aussi étudié leurs performances avec des simulations. Une extension du domaine de validité à $-1 < \xi < \frac{3}{2}$ a été introduite par Diebolt et al. [32, 2004].

1.2.5 Concepts de quantiles extrêmes

1.2.5.1 Quelques éléments théoriques sur les quantiles extrêmes

Soit F la fonction de répartition associée à une loi X.

Définition 1.6 *Le quantile d'ordre* $1-\alpha$ *de la fonction de répartition* F *est défini par :*

$$q(\alpha) = \overline{F}^{\leftarrow}(\alpha) = \inf\left\{y : \overline{F}(y) \leq \alpha\right\},$$

où $\overline{F}^{\leftarrow}$ *est l'inverse généralisée de* $\overline{F}$.
Rappelons que l'inverse généralisée d'une fonction coïncide avec l'inverse classique lorsque celle-ci existe.

Définition 1.7 *Dans le cas d'une variable aléatoire continue* Y , $P(Y = y) = 0$. *Cependant il y a une probabilité* $\overline{F}(y) = P(Y \geq y)$ *que la variable aléatoire soit supérieure ou égale à* y. *Ainsi, on a alors la fonction appelée période de retour donnée par :*

$$T(y) = \frac{1}{\overline{F}(y)}.$$

Exemple 1.5 *Considérons un dé à six faces parfaitement équilibré. Cet exemple est un cas classique en théorie des probabilités d'une variable aléatoire discrète suivant une loi uniforme sur l'ensemble* $\{1,2,3,4,5,6\}$, *La probabilité d'obtenir l'une des six faces est* $\frac{1}{6}$, *Prenons par exemple la face relative au chiffre* 4, *on peut s'attendre à obtenir cette face au bout d'un certain nombre de lancers ou plus précisément on peut s'attendre à l'obtenir en moyenne tous les* $T = 6$ *lancers.* T *est appelée période de retour.*

Définition 1.8 *Le niveau de retour* y_N *est le niveau que l'on peut s'attendre à atteindre ou dépasser, en moyenne une fois toutes les* N *années.*

Cette fonction représente le nombre d'observations tel que, en moyenne, il y ait une observation égale ou supérieure à y. Il en résulte que la période de retour augmente lorsque y augmente. On peut donc définir la fonction niveau de retour comme l'inverse de la période de retour :

$$y(T) = \overline{F}^{\leftarrow}\left(\frac{1}{T}\right) = q\left(\frac{1}{T}\right).$$

Exemple 1.6 *Supposons que la figure (1.10) représente la distribution de survie d'une variable aléatoire continue* Y *définie sur* $[0,10]$ *: Sur la figure (1.10), on peut lire par exemple pour* $\alpha = 0.1$, $q(0,1) = 8$, *Dans le contexte du niveau d'un cours d'eau, si le nombre d'observations est annuel et le niveau du cours d'eau en mètres, alors on aura :* $T(8) = \frac{1}{\overline{F}(8)} = 10ans$ *et* $q\left(\frac{1}{10}\right) = 8m$. *On peut donc s'attendre à ce que le niveau du cours d'eau atteigne ou dépasse* $8m$ *dans dix ans.*

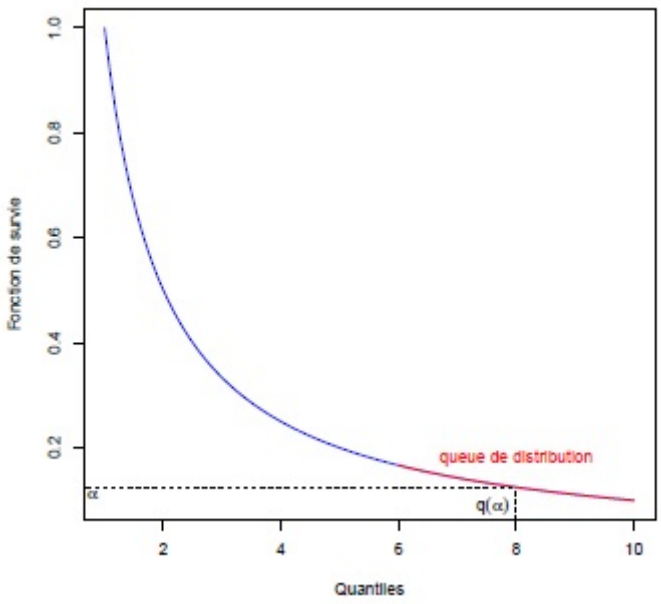

FIGURE 1.10 – Quantiles extrêmes et queue de distribution.

La figure (1.10) montre la relation entre fonction de survie et quantile extrême. La queue de la distribution est donnée sur le graphique en rouge. On remarque donc que pour une valeur de la fonction de survie inférieure ou égale à 0.1, le quantile (niveau de retour) d'ordre 0.1 est supérieur ou égal à 8.

1.2.5.2 Estimation de quantiles extrêmes pour la loi GPD

Nous rappelons que la loi des excès au delà d'un seuil u suffisamment élevé est donnée par la relation suivante :

$$F_u(y) = \Pr(X - u \leq y \mid X > u) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(u+y) - F(u)}{\overline{F}(u)},$$

où $\overline{F} = 1 - F$ est la fonction de survie .
De ce précède , on peut écrire :

$$F_u(y)\overline{F}(u) = F(u+y) - F(u) = -1 + F(u+y) + 1 - F(u),$$

par suite, on a

$$F_u(y)\overline{F}(u) = \overline{F}(u) - \overline{F}(u+y),$$

donc , on aura :

$$\overline{F}(u+y) = \overline{F}(u)\overline{F}_u(y) \tag{1.18}$$

Un résultat limite pour $F_u(y)$ fut proposé par Balkema et de Haan ([28, 1974]) et Pickands ([17, 1975]). De ce fait, on a l'approximation suivante, pour u assez grand

$F_u(y) \approx G_{\xi,\sigma}(y), u \to \infty, y \geq 0$. Ce dernier résultat implique aussi l'expression suivante

$$\overline{F_u}(y) \approx \overline{G_{\xi,\sigma}}(y), u \to \infty, y \geq 0 \tag{1.19}$$

Un estimateur naturel pour $\overline{F_u}$ est la fonction de survie empirique

$$\widehat{\overline{F}}(u) = \widehat{\overline{F}}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Y_i>u\}} = \frac{N_u}{n} \tag{1.20}$$

où $I_{\{Y_i>u\}}$ est une fonction indicatrice ayant l'expression suivante

$$I_{\{Y_i>u\}} \begin{cases} 1 \text{ si } Y_i > u, \forall i = 1,...,n \\ 0 \qquad\qquad\qquad \text{sinon} \end{cases}$$

Par ailleurs, à partir de l'approximation (1.19), on peut proposer un estimateur pour $\overline{F_u}(y)$, soit

$$\widehat{\overline{F}}_u(y) = \widehat{\overline{G_{\xi,\sigma}}}(y) = \left(1 + \frac{\widehat{\xi}}{\widehat{\sigma_u}} y\right)^{\frac{-1}{\widehat{\xi}}} \tag{1.21}$$

En utilisant les relations (1.18), (1.20) et (1.21), on obtient

$$\widehat{\overline{F}}(u+y) = \widehat{\overline{F}}(u)\widehat{\overline{F}}_u(y) = \frac{N_u}{n}\left(1 + \frac{\widehat{\xi}}{\widehat{\sigma_u}} y\right)^{\frac{-1}{\widehat{\xi}}} = \frac{1}{m}, \text{ avec } y > u$$

Le niveau de retour y_N dépassé en moyenne toutes les N années est alors obtenu comme la solution de l'équation (1.22) :

$$\frac{N_u}{n}\left(1 + \frac{\widehat{\xi}}{\widehat{\sigma_u}} y\right)^{\frac{-1}{\widehat{\xi}}} = \frac{1}{m}, \text{ avec } y > u \tag{1.22}$$

Selon les valeurs de en résolvant l'équation (1.22), on a :

$$\widehat{y_N} = \begin{cases} u + \frac{\widehat{\xi}}{\widehat{\sigma_u}}\left[\left(m\frac{N_u}{n}\right)^{\widehat{\xi}} - 1\right], \text{ si } \xi \neq 0 \\ u + \widehat{\sigma_u}\log\left(m\frac{N_u}{n}\right), \text{ si } \xi = 0 \end{cases} \tag{1.23}$$

Remarque 1.6 *Il est souvent adéquat d'estimer le niveau de retour en termes d'échelle annuelle. S'il existe n_y observations par an sur une période de N années avec $m = N * n_y$*

mesures, la relation (1.23) devient

$$\widehat{y_N} = \begin{cases} u + \frac{\widehat{\xi}}{\widehat{\sigma_u}}\left[\left(N n_y \frac{N_u}{n}\right)^{\widehat{\xi}} - 1\right], \; si \; \xi \neq 0 \\ u + \widehat{\sigma_u} \log\left(N n_y \frac{N_u}{n}\right), \; si \; \xi = 0 \end{cases} . \tag{1.24}$$

1.3 Adéquation des modèles des valeurs extrêmes

1.3.1 Introduction

Dans la pratique, il est souvent nécessaire d'analyser la validité d'un modèle retenu pour des prédictions. Il s'agit donc de s'assurer que les modèles sélectionnés décrivent bien les données étudiées et produisent de bonnes estimations afin de donner plus de précision aux prédictions et estimations. De ce fait, nous disposons de certains outils de diagnostic tels que les graphiques Probabilité-Probabilité (P-P plot) et Quantile-Quantile (Q-Q plot) ; il y a aussi les tests usuels d'adéquation comme ceux de Kolmogorov Smirnov (KS), Anderson Darling (AD), Wald, rapport de vraisemblance et autres pour juger de la pertinence des modèles.

1.3.2 Probabilité et Quantile Plot

Les (P-P plot) et (Q-Q plot) sont très utilisés pour tester l'adéquation des modèles des valeurs extrêmes. Ces outils graphiques permettent de comparer la fonction de répartition empirique pour un (P-P plot) ou les quantiles extrêmes (Q-Q plot) d'un échantillon de données à ceux d'un échantillon provenant d'une loi théorique (GPD ou GEV) par exemple. Lorsque la loi observée est la même que la loi théorique, les points sur ces graphiques sont confondus avec la première bissectrice dans le plan. Supposons donc que les valeurs observées $y_{(1)},...,y_{(n)}$ sont ordonnées dans l'ordre croissant, une (P-P plot) est l'ensemble des points

$$\left\{\left(\hat{F}\left(y_{(i)}\right), \frac{i}{n+1}\right) : i = 1,...,n\right\},$$

où $\hat{F}$ est la distribution empirique estimée et $i/(n+1)$ est la valeur prise par la fonction de distribution empirique dans l'intervalle$\left[y_{(i)}, y_{(i+1)}\right[$. Concernant le (Q-Q plot), comme son nom l'indique, il compare les quantiles extrêmes. Il est constitué de l'ensemble des couples

$$\left\{\left(\hat{F}^{-1}\left(\frac{i}{n+1}\right), x_i\right) : i = 1,...,n\right\},$$

Pour plus de détails, voir Coles [6, 2001].

1.3.3 Tests d'adéquation de Kolmogorov Smirnov

Soient $x_1,...,x_n$, n réalisations d'une variable aléatoire X et de fonction de répartition F, On se demande s'il est raisonnable de supposer que X suit la loi caractérisée par F et on pose les deux hypothèses de test :

1. H_0 : X suit la loi de F,
2. H_1 : X suit une autre loi.

Ainsi, on utilise la fonction de répartition empirique F_n de F définie par $F_n(x) = \frac{Card(\{i \mid x_i \leq x\})}{n}$: La statistique de test utilisée est définie par :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

On compare la valeur obtenue D_n à une valeur critique $D_\alpha(n)$ fournie par les tables de Kolmogorov-Smirnov. Si $D_n > D_\alpha(n)$, on rejette l'hypothèse H_0 avec un risque α de se tromper.

1.3.4 Quantification de l'incertitude par la méthode du Bootstrap dans la théorie des valeurs extrêmes

On distingue principalement deux approches : l'approche paramétrique qui est une méthode dans laquelle des hypothèses de base sont émises sur la distribution de l'échantillon de base et l'approche non paramétrique. Dans cette dernière approche, il peut y avoir des hypothèses sur la distribution de l'échantillon de base ; cependant, elle ne nécessite pas que la loi parent appartienne à une famille paramétrique. La technique du bootstrap a été introduite par Efron [33, 1979]. C'est la méthode de réplication des échantillons la mieux fondée théoriquement. Elle consiste à créer, à partir d'un échantillon de base, un grand nombre d'échantillons par tirage aléatoire avec remise. Sur chaque échantillon, les statistiques auxquelles on s'intéresse sont calculées, ce qui permet d'approcher leur dispersion. De ce fait, on peut estimer la variance ou la loi des paramètres caractéristiques de la distribution de l'échantillon et construire des intervalles de confiance, lorsque la distribution des paramètres est analytiquement complexe. Cette méthode semble donc adéquate pour la quantification des incertitudes dans la théorie des valeurs extrêmes. Dans ce travail, on s'intéressera uniquement à l'approche non paramétrique pour quantifier les incertitudes sur les estimations.

Définition 1.9 *Le bootstrap est une technique de reéchantillonnage permettant de simuler la distribution d'un estimateur quelconque pour en apprécier le biais, la variance, l'erreur quadratique moyenne ou encore pour en estimer un intervalle de confiance, même si la loi théorique est inconnue..*

1.3.4.1 Approche non paramétrique du Bootstrap

Cette approche est généralement utilisée dans les situations où l'on ne peut pas faire l'hypothèse que la distribution des observations de certains paramètres appartient à une famille connue. Ici, notre approche consiste à analyser les incertitudes sur les estimations des quantiles extrêmes (niveaux de retour), des paramètres de forme et d'échelle à travers l'approche non paramétrique. Cette analyse se base sur la comparaison des résultats observés et simulés**.**

1.3.4.2 Bootstrap pour l'estimation d'une erreur standard

Dans l'estimation d'une erreur standard par bootstrap, les statistiques d'intérêt peuvent être une moyenne, une médiane, un coefficient de corrélation, un coefficient de variation, une erreur moyenne quadratique, etc.

Remarque 1.7 *L'estimation du bootstrap de l'erreur standard correspond à l'écarttype des réplications bootstrap, ceci est obtenu par l'équation suivante :*

$$\sqrt{\frac{\sum_{b=1}^{B}\left[S\left(X^{*b}\right)-S^{*}\right]}{B-1}},$$

avec $S^{*}=\dfrac{\sum_{b=1}^{B}S\left(X^{*b}\right)}{B}$ *et S peut être une moyenne, un coefficient de variation, une erreur quadratique moyenne, etc.*

1.3.4.3 Quantification des incertitudes sur les estimations dans l'approche POT

La quantification des incertitudes porte sur les niveaux de retour x_T estimés et les intervalles de confiance respectifs associés , ainsi que les paramètres de forme ξ et d'échelle σ de la GPD.

Définition 1.10 *L'erreur quadratique moyenne en anglais (root mean square error, rmse), est une mesure de la différence entre les valeurs prédites et observées. Elle est définie par :*

$$rmse=\sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{y_i}-y_i\right)}{n}}$$

où $\widehat{y_i}$ *est la i-ème valeur estimée,* y_i *celle observée et n la taille de l'échantillon.*

L'algorithme du bootstrap peut être résumé pour l'approche POT comme suit :

1. Les données sont considérées comme un vecteur z_{obs} de n observations indépendantes rangées dans l'ordre croissant.

2. Un échantillon de n observations pris au hasard avec répétition d'éléments du vecteur z_{obs} pour obtenir les données de bootstrap est noté Z^* : Mais dans notre approche, les données sont constituées d'un certain nombre de valeurs au-dessus d'un seuil sélectionné. Ainsi, un échantillon aléatoire de taille le nombre d'excès est selectionné avec remise à partir des données initiales pour obtenir la base de données du bootstrap.
3. Une méthode d'estimation des paramètres de forme ξ et d'échelle σ de la GPD est utilisée afin de déterminer la statistique d'intérêt pour différentes périodes de retour.
4. Déterminer la statistique d'intérêt, $\theta^* = \theta(Z)$.
5. Répéter les étapes 3 et 4 un grand nombre de fois (B fois) pour obtenir une estimation de la distribution du bootstrap.
6. Calculer le($rmse$) pour les paramètres de la GPD et la statistique d'intérêt.
7. Déterminer l'intervalle de confiance pour la statistique d'intérêt.

1.4 Conclusion

Dans ce chapitre, nous avons présenté les deux approches de la théorie des valeurs extrêmes et les éléments théoriques y afférant. Il ressort de cette présentation et de la littérature concernant ces deux approches que l'approche POT est la plus utilisée mais très sensible au choix du seuil pour obtenir une bonne approximation des excès au-dessus du seuil considéré. En outre, dans les domaines de l'industrie ou de l'aéronautique où les données ne sont pas toujours saisonnières ou annuelles, les divisions en blocs s'avèrent problématiques. Ainsi, l'approche POT semble être la plus adaptée pour de nombreuses situations telles que l'analyse des mesures de surface de rugosité, la prédiction des pièces de rechange et la gestion des risques. Dans l'optique de la quantification des incertitudes dues à l'approche POT, nous avons abordé l'approche du bootstrap non-paramétrique dans la TVE ; plus précisément avec l'approche POT. Le chapitre 3 porte sur l'analyse de survie dans le cadre des modèles à risques proportionnels. Il établit le lien entre la théorie des valeurs extrêmes et les modèles de fiabilité.

CHAPITRE 2

ESTIMATION DE L'INDICE ET QUANTILES EXTRÊMES POUR LES LOIS À QUEUE DE TYPE WEIBULL

2.1 Introduction

Dans cet chapitre, nous étudions le comportement des valeurs extrêmes d'un échantillon de variables aléatoires unidimensionnelles. Nous nous concentrons essentiellement sur une famille particulière de lois : les lois à queue de type Weibull. Ces lois ont une fonction de survie qui décroit vers zéro à la vitesse exponentielle. Nous donnerons une définition plus précise de cette famille dans le section 2.2. Notre principal objectif est de proposer des estimateurs de quantiles extrêmes. Plus précisément, disposant d'un échantillon $X_1, ..., X_n$ de n variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition commune $F(.)$, nous souhaitons estimer le réel $q(\alpha_n)$ défini par

$$q(\alpha_n) = \overline{F}^{\leftarrow}(\alpha_n), \text{ avec } \alpha_n \to 0 \text{ lorsque } n \to \infty,$$

où (α_n) est une suite connue et $\overline{F}^{\leftarrow}(u) = \inf\left\{x, \overline{F} \leq u\right\}$ est l'inverse généralisée de la fonction de survie $\overline{F}(.) = 1 - F(.)$. Remarquons que $q(\alpha_n)$ est le quantile d'ordre $1 - \alpha_n$ de la fonction de répartition F. Un problème similaire à l'estimation de $q(\alpha_n)$ est l'estimation de "petites probabilités" p_n. Autrement dit, pour une suite de réels (x_n) fixée, nous souhaitons estimer la probabilité p_n définie par

$$p_n = \overline{F}(x_n), \text{avec } x_n \to \infty \text{ lorsque } n \to \infty$$

Ce sont les hydrologues qui ont été parmi les premiers à s'intéresser à ces deux problèmes. Disposant d'un échantillon de hauteurs d'un cours d'eau, ils se sont posés les deux questions suivantes :

1) quelle est la hauteur d'eau qui est atteinte ou dépassée pour une faible probabilité donnée ?

2) pour une "grande" hauteur d'eau fixée, qu'elle est la probabilité d'observer une hauteur d'eau qui lui sera supérieure ?

Les questions (1) et (2) se rapportent donc respectivement à l'estimation d'un quantile extrême (ou niveau de retour en hydrologie) et d'une "petite probabilité" (ou de façon équivalente en hydrologie, période de retour). La difficulté principale réside dans le fait que l'on considère un ordre de quantile $\alpha_n \to 0$ (ou de manière équivalente un seuil $x_n \to \infty$). En effet, si par exemple $n\alpha_n \to 0$ lorsque $n \to \infty$, il est clair que

$$P\left(X_{n,n} < q\left(\alpha_n\right)\right) = F^n\left(q\left(\alpha_n\right)\right) = \left(1-\alpha_n\right)^n \to 1$$

où $X_{1,n} \leq \ldots \leq X_{n,n}$ est l'échantillon ordonné associé à $X_1, \ldots, X_n$. La quantité $q(\alpha_n)$ n'appartient donc pas à l'intervalle de variation de nos observations. En conséquence, l'estimateur de $q(\alpha_n)$ ne peut être obtenu en inversant simplement la fonction de répartition empirique

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{X_i \leq x\right\},$$

car $\widehat{F}_n(x) = 1$ pour $x \geq X_{n,n}$. L'estimation de quantiles extrêmes et/ou de "petites probabilités" est requise dans de nombreux domaines d'application parmi lesquels citons la fiabilité [34], la finance [1], les assurances [35, 1992], [36] et la climatologie [37]. Pour répondre à ces deux questions, nous devons donc étudier de près le comportement de la queue de distribution de $F(.)$ en utilisant la théorie des valeurs extrêmes. Nous présentons les éléments essentiels de cette théorie chapitre 1 . Dans le paragraphe 2, nous proposons des estimateurs de quantiles extrêmes pour la famille des lois à queue de type Weibull.

2.1.1 Estimateur de Pickands

Il est défini par la statistique :

$$\widehat{\xi}^{P}_{k_n,n} = \frac{1}{\ln(2)} \ln\left(\frac{X_{k_n,n} - X_{2k_n,n}}{X_{2k_n,n} - X_{4k_n,n}}\right)$$

Il présente l'intérêt d'être valable quelle que soit la distribution des extrêmes (Gumbel,Weibull ou Fréchet).La représentation graphique de cet estimateur en fonction du

nombre k d'observations considérées montre un comportement en général très volatil au départ, ce qui nuit à la lisibilité du graphique. De plus, cet estimateur est très sensible à la taille de l'échantillon sélectionné, ce qui le rend peu robuste. Il est donc d'un maniement délicat. On peut noter qu'il est asymptotiquement normal, avec :

$$\sqrt{k_n}\frac{\widehat{\xi}^P_{k_n,n}-\xi}{\sigma(\xi)} \to \mathcal{N}(0,1)$$

lorsque $k_n \to +\infty$ la variance asymptotique étant donnée par :

$$\sigma(\xi) = \frac{\xi\sqrt{2^{2\xi+1}+1}}{2\left(2^{\xi}-1\right)\ln(2)}$$

2.1.2 Estimateur de Hill

L'estimateur de Hill n'est utilisable que pour les distributions de Fréchet (donc telles que $\xi > 0$) pour lesquelles il fournit un estimateur de l'indice de queue plus efficace que l'estimateur de Pickands.Il est défini par la statistique suivante :

$$\widehat{\xi}^H_{k_n,n} = \frac{1}{k_n-1}\sum_{j=1}^{k_n-1}\ln\left(\frac{X_{j,n}}{X_{k_n,n}}\right)$$

Si on choisit $k_n, n \to +\infty$ de sorte que $\frac{n}{k_n} \to +\infty$ alors on peut montrer que $\lim_{k_n\to\infty}\widehat{\xi}^H_{k_n,n} = \xi$ et l'estimateur de Hill est le plus asymptotiquement normal :

$$\sqrt{k_n}\frac{\widehat{\xi}^H_{k_n,n}-\xi}{\xi} \to \mathcal{N}(0,1)$$

la convergence étant en loi. Cet estimateur est l'estimateur du maximum de vraisemblance dans le cas particulier du modèle $S(x) = 1-F(x) = Cx^{\frac{-1}{\xi}}$, on reconnaît ici une distribution de Pareto d'indice $\alpha = \frac{1}{\xi}$.Dans le cas général du domaine de Fréchet, la fonction de survie est de la forme $S(x) = 1-F(x) = x^{-\frac{1}{\xi}}L(x)$ avec L une fonction à variation lente. Cela induit un biais important sur l'estimateur de Hill, qui est donc en pratique d'un maniement délicat.Dans le cas général, la fonction L apparaît comme un paramètre de nuisance de dimension infinie, qui complique l'estimation.

2.1.3 Le passege de domaine d'attraction Fréchet vers Weibull

Le résultat suivant (voir Gnedenko [3], Resnick [4, Proposition 1.13]) montre que l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.voir le subsection 1.2.2.4.

où $\alpha \in [0,1]$. De nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes et des quantiles extrêmes $q(\alpha_n)$ pour des lois à queue lourde. L'estimateur le plus connu de $\xi > 0$ est l'estimateur proposé par Hill [18] et défini par

$$\widehat{\xi}_n^H = \frac{1}{k_n}\sum_{i=1}^{k_n} \log\left(X_{n-i+1,n}\right) - \log\left(X_{n-k_n,n}\right),$$

où (k_n) est une suite d'entiers telle que $1 < k_n < n$. D'autres estimateurs de cet indice ont été proposés notamment par Beirlant et al. [51],[50] qui utilisent un modèle de régression exponentiel pour débiaiser l'estimateur de Hill et par Feuerverger et al. [52] qui introduisent un estimateur des moindres carrés. L'utilisation d'un noyau dans l'estimateur de Hill a été étudiée par Csörgõ et al. [71]. Un estimateur efficace de l'indice des valeurs extrêmes a été proposé par Falk et al. [72]. Une liste plus détaillée des différents travaux sur l'estimation de l'indice des valeurs extrêmes est effectuée par Csörgo et al. [71]. Concernant l'étude du quantile extrême $q(\alpha_n)$, l'estimateur le plus fréquemment utilisé a été proposé par Weissman [57]. Il est défini par :

$$\widehat{q}_n^W(\alpha_n) = X_{n-k_n+1,n}\left(\frac{k_n}{n\alpha_n}\right)^{\widehat{\xi}_n^H}. \tag{2.1}$$

On peut trouver de nombreux autres estimateurs du quantile extrême dans le livre écrit par de Haan et Ferreira [15].

2.2 Inférence pour les lois à queue de type Weibull

Les lois à queue de typeWeibull correspondent au cas particulier, est appelé l'indice de queue de Weibull :

$$\overline{F}(x) = \exp\left\{-x^{1/\theta}L(x)\right\}, L(.) \in \mathcal{RV}_0 \text{ où } \theta > 0. \tag{2.2}$$

Une fonction de répartition s'écrivant selon le modèle (2.2) est dite à queue de type Weibull d'indice $\theta > 0$. L'équation (2.2) équivalente à

$$q(\alpha) = (-\log\alpha)^{\theta}\,\ell(-\log\alpha), \ell(.) \in \mathcal{RV}_0, \tag{2.3}$$

où $\alpha \in [0,1]$. Cette famille de lois contient par exemple les lois normale, Gamma, exponentielle, etc ... Par contre, la loi log-normale qui appartient au domaine d'attraction de Gumbel n'est pas une loi à queue de type Weibull. Dans la suite, on considére un échantillon $X_1,...,X_n$ de variables aléatoires indépendantes et distribuées selon le modèle (2.2), le paragraphe 2.2.1 est consacré à l'estimation de l'indice de queue de

Weibull. L'estimation des quantiles extrêmes est discutée dans le paragraphe 2.2.2.

2.2.1 Estimation de l'indice de queue de Weibull

Les lois à queue de type Weibull appartiennent évidemment au domaine d'attraction de Gumbel (i.e. avec un indice des valeurs extrêmes $\xi = 0$). L'indice des valeurs extrêmes ne fournit donc aucune information sur la vitesse de décroissance de la fonction de survie à l'intérieur de cette famille de loi. C'est l'indice de queue de Weibull qui nous donne cette information : une valeur de proche de zéro (resp. l'infini) correspond à une décroissance rapide (resp. lente) de la queue de distribution. La connaissance de ce paramètre est donc essentielle si l'on souhaite par exemple estimer un quantile extrême. Il existe dans la littérature de nombreux estimateurs de l'indice .Berred [38, 1991] propose un estimateur basé sur des valeurs records mais la majorité des estimateurs utilise les kn plus grandes observations de l'échantillon. Parmi ceux-ci citons Beirlant et al. [39], [40], [41], Broniatowski [42], Diebolt et *al*. [43], Dierckx et *al*. [44], Gardes et *al*. [45, 2006], [46, 2008], Girard [47, 2004] et Goegebeur et *al*. [48, 2010], [49, 2010]. Le plus simple d'entre eux a été proposé dans [41]. Il est défini par :

$$\widehat{\theta}_n^B = \frac{\sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log\left(X_{n-K_n+1,n}\right)}{\sum_{i=1}^{k_n-1} \left(\log\log\left(\frac{n}{i}\right) - \log\log\left(\frac{n}{k_n}\right)\right)}, \tag{2.4}$$

et rappelons que (k_n) est une suite d'entiers tels que $1 < k_n < n$. Son expression est proche de celle de l'estimateur de l'indice des valeurs extrêmes proposé par Hill [18]. Elle est basée sur la remarque suivante (2.3) :

$$\frac{\log q(\alpha)}{\log\log\left(\frac{1}{\alpha}\right)} = \theta + \frac{\log \ell\left(\log\left(\frac{1}{\alpha}\right)\right)}{\log\log\left(\frac{1}{\alpha}\right)}.$$

Ainsi, comme $\frac{\log(\ell(x))}{\log(x)} \to 0$ lorsque $x \to \infty$, on en déduit que pour

$$\log q(\alpha) \sim \theta \log\log\left(\frac{1}{\alpha}\right). \tag{2.5}$$

Ainsi, les points $\left(\log\log\left(\frac{n}{i}\right), \log(X_{n-i+1,n})\right)$, $i = 1, ..., k_n - 1$ sont approximativement répartis sur une droite de pente θ. Les résultats asymptotiques sont obtenus (entre autres) sous les hypothèses suivantes.

(H.1) La suite (k_n) vérifie $k_n \to \infty$ et $\frac{n}{k_n} \to \infty$ lorsque $n \to \infty$.

(H.2) Il existe un paramètre $\rho < 0$ et une fonction $b(.)$ vérifiant $b(x) \to 0$ lorsque

$x \to \infty$ tels que pour tout $1 < A < \infty$

$$\lim_{x\to\infty} \sup_{\lambda\in[1,A]} \left| \frac{\log\left(\ell(\lambda x)/\ell(x)\right)}{b(x)\,K_\rho(\lambda)} - 1 \right| = 0,$$

où $K_\rho(\lambda) = \int_1^\lambda t^{\rho-1}dt$ et $\ell(.)$ est la fonction à variations lentes introduite dans (2.3).

L'hypothèse **(H.1)** assure que le nombre de statistiques d'ordre conservées k_n est assez grand ($k_n \to \infty$) pour obtenir des estimateurs stables, mais pas trop $\left(\frac{n}{k_n} \to \infty\right)$ pour que les observations utilisées restent dans la queue de distribution. Le choix de la suite (k_n) est donc un compromis entre le biais et la variance de l'estimateur.
L'hypothèse **(H.2)** est très souvent utilisée pour étudier le comportement asymptotique d'estimateurs d'indice ou de quantiles extrêmes. Elle est notamment nécessaire pour démontrer la normalité asymptotique de l'estimateur de Hill . On peut montrer que la fonction $b(.)$ (appelée aussi fonction de biais) est à variations régulières d'indice $\rho < 0$. Le paramètre (appelé paramètre du second ordre) contrôle donc la vitesse de convergence de $\frac{\ell(\lambda x)}{\ell(x)}$ vers 1. Une valeur ρ de proche de 0 implique une faible vitesse de convergence. Quelques exemples en sont donnés dans la Table (2.1).

TABLE 2.1 – Paramètres θ,ρ et fonction $b(.)$ associés aux lois usuelles. Les paramètres α et λ sont respectivement des paramètres de forme et d'échelle.

Loi	θ	$b(x)$	ρ
Normale $\mathcal{N}\left(\mu,\sigma^2\right)$	1/2	$\frac{1}{4}\frac{\log(x)}{x}$	-1
Gamma $\xi\,(\alpha\neq 1,\lambda)$	1	$(1-\alpha)\frac{\log(x)}{x}$	-1
Weibull $\mathcal{W}(\alpha,\lambda)$	$1/\alpha$	0	$-\infty$

2.2.1.1 Un estimateur de θ débiaisé

Considérons à présent un estimateur débiaisé de l'indice de queue de Weibull . Il est basé sur un modèle de régression exponentiel inspiré de ceux proposés par Beirlant et *al.* [50, 1999], [51, 2002] et Feuerverger et *al.* [52, 1999] afin d'estimer l'indice des valeurs extrêmes pour des lois du domaine d'attraction de Fréchet. Plus précisément, on définit les variables aléatoires

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

Le modèle suivant peut être établi (2.6) :

$$Z_j = \left(\theta + \left(\frac{\log(n/k_n)}{\log(n/j)}\right) b\left(\log\left(\frac{n}{k_n}\right)\right)\right) f_j + o_p\left(b\left(\log\left(\frac{n}{k_n}\right)\right)\right), j = 1,...,k_n. \tag{2.6}$$

où f_j, $j = 1,...,k_n$ sont des variables aléatoires indépendantes de loi exponentielle de paramètre 1 et le terme $o_p\left(b\left(\log\left(\frac{n}{k_n}\right)\right)\right)$ ne dépend pas de j. On obtient à partir du modèle (2.6) l'approximation

$$Z_j \approx \theta + b\left(\log\left(\frac{n}{k_n}\right)\right)x_j + \eta_j, j = 1,...,k_n \tag{2.7}$$

où η_j est un terme d'erreur aléatoire centré et $x_j = \log\left(\frac{n}{k_n}\right)/\log\left(\frac{n}{j}\right)$. En estimant les paramètres θ et $b\left(\log\left(\frac{n}{k_n}\right)\right)$ du modèle de régression linéaire (2.7) par la méthode des moindres carrés ordinaires, on obtient un estimateur de débiaisé :

$$\widehat{\theta}_n^D = \frac{1}{k_n}\sum_{j=1}^{k_n} Z_j - \frac{\widehat{b}\left(\log\left(\frac{n}{k_n}\right)\right)}{k_n}\sum_{j=1}^{k_n} x_j, \tag{2.8}$$

où

$$\widehat{b}\left(\log\left(n/k_n\right)\right) = \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=i}^{k_n} x_j\right)Z_j / \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=1}^{k_n} x_j\right)^2. \tag{2.9}$$

La normalité asymptotique de $\widehat{\theta}_n^D$ est donnée par le résultat ci-dessous (voir(2.6))

Théorème 2.1 *On se place sous le modèle (2.2) et on suppose que les conditions **(H.1)** et **(H.2)** sont satisfaites. Supposons de plus que la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$ avec*

$$\frac{k_n^{1/2}}{\log\left(\frac{n}{k_n}\right)} b\left(\log\left(\frac{n}{k_n}\right)\right) \to \widetilde{\Lambda} \in \mathbb{R}.$$

Supposons enfin que,si $\widetilde{\Lambda} = 0$, $\frac{\log(k_n)}{\log\left(\frac{n}{k_n}\right)} \to 0$ et $\frac{k_n^{1/2}}{\log\left(\frac{n}{k_n}\right)} \to \infty$. On a :

$$\frac{k_n^{1/2}}{\log\left(\frac{n}{k_n}\right)}\left(\widehat{\theta}_n^D - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \theta^2\right).$$

L'hypothèse $x|b(x)| \to \infty$ implique que dans la condition **(H.2)** la vitesse de convergence est lente (et plus particuliérement que $\rho \geq -1$). Les estimateurs non débiaisés θ de auront donc tendance à avoir un biais important dans ce cas. On peut montrer (voir 2.1) que les lois normale, Gamma satisfont cette hypothèse mais pas les lois de Weibull.

2.2.1.2 Choix du nombre k_n de statistiques d'ordre

En ne tenant pas compte du terme de biais dans le modèle de régression (2.6), on obtient un estimateur non débiaisé de θ défini par :

$$\frac{1}{k_n}\sum_{j=1}^{k_n} Z_j.$$

L'erreur moyenne quadratique asymptotique (AMSE) de cet estimateur est

$$AMSE(k_n) = \frac{\theta^2}{k_n} + \left(\frac{b\left(\log\left(\frac{n}{k}\right)\right)}{k_n}\sum_{j}^{k_n}\frac{\log\left(\frac{n}{k_n}\right)}{\log\left(\frac{n}{j}\right)}\right)^2.$$

Un choix possible pour k_n est alors de prendre $k_n^{opt} = \arg\min_{k_n} AMSE(k_n)$. Nous pouvons estimer cette erreur par la quantité $\widehat{AMSE}(k_n)$obtenue en remplaçant θ et $b\left(\log\left(\frac{n}{k}\right)\right)$ par les estimateurs $\widehat{\theta}_n^D$ et $\widehat{b}\left(\log\left(\frac{n}{k}\right)\right)$ définis précédemment (voir les équations (2.8) et (2.9)). Le nombre k_n^{opt} est estimé par :

$$\widehat{k_n} = \arg\min_{k_n} \widehat{AMSE}(k_n).$$

Comme l'ont fait remarquer récemment Asimit et al. [53, 2010], $AMSE(k_n) \backsim \frac{\theta^2}{k_n + b^2(\log(n))}$. Ainsi, la sélection du nombre d'observations k_n n'est pas justifiée théoriquement puisque $k_n^{opt} \sim n$.Une méthode alternative a récemment été proposée [54] basée sur les idées de [56, 1998] elles-mêmes inspirées d'une variante de la méthode de Lepski [55].

2.2.2 Estimation de quantiles extrêmes

Toujours pour des lois à queue de type Weibull, nous nous intéressons à présent au probléme d'estimation d'un quantile extrême $q(\alpha_n)$ lorsque l'ordre α_n converge vers zéro. Le principal estimateur de $q(\alpha_n)$ disponible dans la littérature a été proposé par Beirlant et al. [41]. Il est basé sur l'approximation (2.5) qui assure que pour n assez grand, on a sous l'hypothèse **(H.1)** :

$$\log q(\alpha_n) \approx \theta \log\log\left(\frac{1}{\alpha_n}\right) \text{ et } \log q\left(\frac{k_n}{n}\right) \approx \theta \log\log\left(\frac{n}{k_n}\right).$$

En soustrayant membre à membre les deux approximations ci-dessus et en appliquant la fonction exponentielle, on montre facilement que :

$$q(\alpha_n) \approx q\left(\frac{k_n}{n}\right)\left(\frac{\log\left(\frac{1}{\alpha_n}\right)}{\log\left(\frac{n}{k_n}\right)}\right)^{\theta}.$$

En estimant $q\left(\frac{k_n}{n}\right)$ par $X_{n-k_n+1,n}$ qui est le quantile associé à la fonction de répartition empirique et θ par $\widehat{\theta}_n^B$ (voir équation (2.4)), Beirlant et al. [41] proposent l'estimateur suivant :

$$\widehat{q}^B(\alpha_n) = X_{n-k_n+1,n}\left(\frac{\log\left(\frac{1}{\alpha_n}\right)}{\log\left(\frac{n}{k_n}\right)}\right)^{\widehat{\theta}_n^B}.$$

La construction de cet estimateur est similaire à celle de l'estimateur proposé par Weissman [57, 1978] (voir équation (2.1)) pour des lois du domaine d'attraction de Fréchet. Un autre estimateur de $q(\alpha_n)$ a été proposé par Beirlant et al. [40, 1995]. Il est défini par :

$$\widehat{q}^{B_*}(\alpha_n) = X_{n-k_n+1,n}\left(1 + \frac{\widehat{\sigma}_n \log\left(\frac{k_n}{(n\alpha_n)}\right)}{\widehat{\theta}_n^{B_*} X_{n-k_n+1,n}}\right)^{\widehat{\theta}_n^{B_*}},$$

avec

$$\widehat{\sigma}_n = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1}\left(X_{n-i+1,n} - X_{n-k_n+1,n}\right) \text{ et } \widehat{\sigma}_n^{B*} = \frac{\log\left(\frac{n}{k_n}\right)}{X_{n-k_n+1,n}}\widehat{\sigma}_n.$$

Etant donné que

$$1 + \frac{\widehat{\sigma}_n \log\left(\frac{k_n}{(n\alpha_n)}\right)}{\widehat{\theta}_n^{B_*} X_{n-k_n+1,n}} = \frac{\log\left(\frac{1}{\alpha_n}\right)}{\log\left(\frac{n}{k_n}\right)},$$

l'estimateur $\widehat{q}^{B_*}(\alpha_n)$ est en fait l'estimateur $\widehat{q}^B(\alpha_n)$ pour lequel l'indice θ n'est pas estimé par $\widehat{\theta}_n^B$ mais par $\widehat{\theta}_n^{B_*}$. L'étude du comportement asymptotique de ces deux estimateurs peut alors être unifiée [58]. Plus généralement, intéressons-nous à la famille d'estimateurs du quantile extrême $q(\alpha_n)$ définie par

$$\mathcal{Q}_{\alpha_n} = \left\{\widehat{q}\left(\alpha_n, \widehat{\theta}_n\right), \widehat{\theta}_n \text{ estimateur de } \theta\right\},$$

avec

$$\widehat{q}\left(\alpha_n, \widehat{\theta}_n\right) = X_{n-k_n+1,n}\tau^{\widehat{\theta}}, \tau_n = \frac{\log\left(\frac{1}{\alpha_n}\right)}{\log\left(\frac{n}{k_n}\right)}, \tag{2.10}$$

où $\widehat{\theta}_n$ est un estimateur quelconque de θ.

2.2.3 Un estimateur de $q(\alpha_n)$ débiaisé

Pour proposer un estimateur débiaisé du quantile extrême $q(\alpha_n)$, on se base sur le résultat suivant établi dans [59] : sous la condition **(H.2)**, si $\tau_n \to \tau \in]1,\infty[$, on a lorsque $n \to \infty$

$$q(\alpha_n) \sim q\left(\frac{k_n}{n}\right)\tau_n^{\theta}\exp\left\{b\left(\log\left(\frac{n}{k_n}\right)\right)K_\rho(\tau_n)\right\}.$$

En estimant $q\left(\frac{k_n}{n}\right)$ par la statistique d'ordre $X_{n-k_n+1,n}$, θ, par $\widehat{\theta}_n^D$ (voir équation (2.8)), $b\left(\log\left(\frac{n}{k_n}\right)\right)$ par $b\left(\log\left(\frac{n}{k_n}\right)\right)$ (voir équation (2.9)) et ρ par un estimateur $\widehat{\rho}_n$, on obtient l'estimateur

$$X_{n-k_n+1,n}\tau^{\widehat{\theta}}\exp\left\{\widehat{b}\left(\log\left(\frac{n}{k_n}\right)\right)K_{\widehat{\rho}_n}(\tau_n)\right\}.$$

Si on néglige le terme de correction $\exp\left\{\widehat{b}\left(\log\left(\frac{n}{k_n}\right)\right)K_{\widehat{\rho}_n}(\tau_n)\right\}$ dans l'expression ci-dessus, on retrouve l'estimateur non débiaisé $\widehat{q}\left(\alpha_n,\widehat{\theta}_n^D\right)$ appartenant à la famille$\mathcal{Q}_{\alpha_n}$ (voir équation (2.10)). Concernant le paramètre ρ , plusieurs estimateurs ont été proposés pour des modèles différents (citons les travaux de Gomes [60], Gomes et al. [61], Feuerverger et al. [52] Peng et al. [62] et Beirlant et al. [50]). Le résultat suivant (voir 2.7), montre que l'on peut remplacer $\widehat{\rho}_n$ par une valeur arbitraire $\rho' < 0$ et obtenir un estimateur

$$\widehat{q}^D(\alpha_n) = X_{n-k_n+1,n}\tau^{\widehat{\theta}_n^D}\exp\left\{\widehat{b}\left(\log\left(\frac{n}{k_n}\right)\right)K_{\rho'}(\tau_n)\right\}$$

asymtotiquement normal.

On se place sous le modèle (2.2) et on suppose que les conditions **(H.1)** et **(H.2)** sont satisfaites. Supposons que la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$ avec

$$\frac{k_n^{\frac{1}{2}}}{\log\left(\frac{n}{k_n}\right)}b\left(\log\frac{n}{k_n}\right) \to \widetilde{\Lambda} \in \mathbb{R}.$$

Supposons de plus que, si

$$\widetilde{\Lambda} = 0, \frac{\log^2(k_n)}{\log\left(\frac{n}{k_n}\right)} \to 0 \text{ et } \frac{k_n^{\frac{1}{2}}}{\log\left(\frac{n}{k_n}\right)} \to \infty. \text{ Alors, si } \tau_n \to \tau \in]1,\infty[\text{ on a :}$$

$$\frac{k_n^{\frac{1}{2}}}{\log\left(\frac{n}{k_n}\right)}\left(\frac{\widehat{q}^D(\alpha_n)}{q(\alpha_n)} - 1\right) \xrightarrow{d} \mathcal{N}\left(\widetilde{\Lambda}\mu(\tau), \theta^2\sigma^2(\tau)\right),$$

avec

$$\sigma^2(\tau) = \left(K_{p'}(\tau) - \log(\tau)\right)^2 \text{ et } \mu(\tau) = \left(K_{\rho'}(\tau) - K_\rho(\tau)\right)^2$$

Si $\widetilde{\Lambda} \neq 0$ et si$\rho' = \rho$ alors l'estimateur $\widehat{q}^D(\alpha_n)$ est sans biais avec une vitesse de convergence de l'ordre de $\log^{-\rho'}(n)\ell^*(\log(n))$ où $\ell^*(.)$ est une fonction à variations lentes. Evi-

demment un mauvais choix de ρ' conduit à un estimateur du quantile extrême biaisé. Notons cependant que les lois à queue de type Weibull usuelles (loi normale, Gamma) ont un paramètre du second ordre $\rho = -1$. En pratique, on prendra donc une valeur ρ' gale à -1.

CHAPITRE 3

ANALYSE DE SURVIE DANS UN CADRE EXTRÊME

3.1 Introduction

Le terme analyse de fiabilité tire sa source de l'analyse de survie. L'analyse de survie concerne les êtres vivants, par exemple le suivi d'un patient après un traitement spécifique. Quant à l'analyse de fiabilité, elle s'intéresse aux systèmes matériels tels que les outils électriques, mécaniques ou tout autre système industriel. L'analyse de survie est un domaine important des biostatistiques qui consiste à étudier le fonctionnement, les durées de vie ou l'évolution des systèmes. La théorie de la fiabilité sert à étudier l'aptitude de systèmes à fonctionner correctement durant une période donnée. Plus précisément, la fonction de fiabilité est la probabilité pour qu'un matériel ou un système fonctionne sans défaillance pendant une durée de temps. Ce matériel ou un système peut être technique (une machine, un bâtiment, un outil industriel, etc), biologique (plantes, êtres vivants, etc).

3.2 Concepts de base de l'analyse de Fiabilité

La fiabilité a pris son développement depuis la dernière guerre mondiale. Elle est devenue une science à part entière avec des applications dans de nombreux domaines. Elle a pour fondements mathématiques la statistique et le calcul de probabilités qui sont nécessaires à la compréhension et à l'analyse des données de fiabilité. La détermination de la fiabilité d'un système électronique, mécanique ou autre nécessite tout d'abord de connaître la loi de fiabilité (ou la loi de défaillance) de chacun des composants intervenant dans le système. Certaines terminologies sont couramment utilisées pour indiquer la fin d'une durée de vie telles que par exemple :

* une défaillance par l'ingéniérie de la fiabilité,

* décès dans le domaine des sciences actuarielles et biostatistiques,
* une époque pour les processus ponctuels.

De plus, certains concepts sont associés à des objets, comme par exemple :

* un système ou un composant dans l'ingéniérie de la fiabilité,
* un individu pour les sciences actuarielles,
* un organisme en biostatistiques.

3.2.1 Fonctions décrivant l'évolution d'un système

On distingue principalement cinq fonctions représentant l'évolution d'un système, chacune de ces fonctions peut avoir une dénomination particulière selon le domaine d'application :

1. le complémentaire de la fonction de répartition notée S ou R. On l'appelle aussi fonction de survie ou fonction de fiabilité,
2. la fonction de répartition notée F, elle est aussi connue sous le nom de fonction de défaillance en fiabilité des systèmes,
3. la fonction de densité de probabilité notée f, elle est aussi connue sous le nom de densité de défaillance,
4. la fonction de risque notée h, elle est aussi appelée taux de défaillance instantané ou risque instantané,
5. la fonction de risque cumulée notée H, elle est aussi connue sous le nom de fonction de risque intégrée.

3.2.1.1 Fonction de survie S ou de fiabilité R

Soit X une variable aléatoire modélisant l'évolution d'un système. La fonction de survie S est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = P(X > t), t > 0.$$

Remarque 3.1 *Dans un contexte multivarié* $S(t) \neq 1 - F(t)$ *car* $\overline{F}(t) = P(X > t) \neq 1 - F(t)$.

3.2.2 Fonction de répartition F

La fonction de répartition ou « *cumulative distribution en anglais* » représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à -dire

$$F(t) = P(X > t) = 1 - S(t),$$

et donc $S(t) = 1 - F(t) = \overline{F}(t)$.

3.2.2.1 Fonction de densité de probabilité f

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$,

$$F(t) = \int_0^t f(u)\,du.$$

Si la fonction de répartition admet une dérivée au point t, alors

$$f(t) = \lim_{h \to 0} \frac{p(t \leq X \leq t+h)}{h} = F'(t) = -S'(t).$$

Pour t fixé, la fonction de densité représente la probabilité que l'évènement d'intérêt se produise dans un intervalle de temps rapporté à la largeur de cet intervalle.

3.2.2.2 Fonction de risque instantané

Le risque instantané ou encore taux d'incidence, pour t fixé, caractérise la probabilitéde mourir ou d'avoir une défaillance dans un intervalle de temps après t, sous l'hypothèse d'avoir survécu jusqu'au temps t rapporté à la largeur de cet intervalle :

$$h(t) = \lim_{x \to 0} \frac{p(t \leq X \leq t+h \mid X \geq t)}{x} = \frac{f(t)}{S(t)}.$$

3.2.2.3 Fonction de risque cumulé

La fonction ou taux de risque cumulé est l'intégrale du risque instantané h défini par :

$$H(t) = \int_0^t h(u)\,du = -\log[S(t)]. \tag{3.1}$$

On peut déduire de la relation (3.1), une expression de la fonction de survie en fonction du risque cumulé :

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u)\,du\right).$$

On en déduit que

$$f(t) = h(t)\exp\left(-\int_0^t h(u)\,du\right).$$

3.2.3 Fonction de fiabilité d'un système

Le terme fiabilité est un néologisme introduit il y a plus de cinquante ans pour traduire le terme anglo-saxon reliability. La Commission Électronique Internationale donne à la fiabilité la définition suivante :

Définition 3.1 *La fiabilité notée R (Reliability) caractérise l'aptitude d'un système ou d'un*

matériel à accomplir une fonction requise dans des conditions données pendant un intervalle de temps donné.

Considérons un système matériel dont on s'intéresse à la fiabilité. Soit T la variable aléatoire associée au temps de fonctionnement du système. On peut distinguer deux évènements A et B définis comme suit :

* A : le système est en bon état de fonctionnement à l'instant t.
* B : le système est défaillant à l'instant $t+\Delta t$.

On a alors : $P(A) = P(T > t)$ et $P(B) = P(T \leq t+\Delta t)$.Par conséquent,

$$\begin{aligned} P(A \cap B) &= P(t < T \leq t+\Delta t) \\ &= F(t+\Delta t) - F(t) \\ &= [1 - R(t+\Delta t)] - [1 - R(t)] \\ \text{soit } P(A \cap B) &= R(t) - R(t+\Delta t). \end{aligned}$$

On en déduit donc que :

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{R(t) - R(t+\Delta t)}{R(t)}.$$

Remarque 3.2 *On appelle fonction de défaillance la fonction F définie par :*

$$F(t) = P(T \leq t), t \geq 0.$$

où T représente la durée de vie pour une v.a. Le nombre $F(t)$ représente la probabilité que le système ait une défaillance avant l'instant t.La figure (3.1) illustre la fonction de fiabilité et de défaillance.

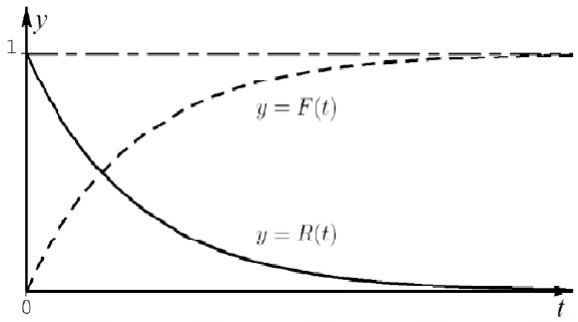

FIGURE 3.1 – Fonction de fiabilité et de défaillance

3.2.4 Fonction de Hasard

Appelée selon les domaines d'application : *"taux instantané de défaillance"*, *"taux de risque"*, ou encore *"quotient de mortalité"*

Définition 3.2 *Soit un système ayant pour fonction de fiabilité R, le taux de défaillance instantané à l'instant t, est noté h(t) , et défini par :*

$$h(t) = \lim_{\Delta t \to 0}\left[\frac{1}{\Delta t} \times \frac{R(t) - R(t+\Delta t)}{R(t)}\right] t \geq 0.$$

De plus, le terme $h(t)\Delta t$, mesure la probabilité qu'une défaillance du système se produise dans l'intervalle de temps $[t, t+\Delta t]$ sachant que le système a bien fonctionné jusqu'à l'instant t. On a aussi la fonction taux de défaillance cumulé, noté $H(t)$ et défini par :

$$H(t) = \int_0^t h(x)\,dx.$$

Ainsi on peut aussi écrire,

$$\begin{aligned} h(t) &= -\frac{dR(t)}{dt} \times \frac{1}{R(t)} \\ &= \frac{dF(t)}{dt} \times \frac{1}{R(t)} \\ h(t) &= \frac{f(t)}{1 - F(t)}. \end{aligned}$$

Remarque 3.3 *On peut exprimer les fonctions de fiabilité et de défaillance en fonction du risque instantané ou du risque cumulé comme suit*

$$R(t) = \exp\left(-\int_0^t h(x)\,dx\right) \text{ et } F(t) = 1 - \exp\left(-\int_0^t h(x)\,dx\right).$$

Proposition 3.1 *Si h est une fonction de risque, alors elle vérifie les propriétés suivantes :*

1. $h(t) \geq 0$, pour tout $t \geq 0$,

2. $\int_0^{+\infty} h(t)\,d(t) = +\infty$.

La fonction de fiabilité ou de survie examine le risque de défaillance d'un organisme, d'un système matériel, etc pouvant se produire sur une période donnée. Pour suivre la durée de vie d'un système à travers la distribution associée à l'évolution du système, la fonction de risque est utilisée. En fait, la fonction de risque est plus informative sur la défaillance des systèmes que les autres distributions caractérisant la durée des systèmes. C'est dans cette optique que Cox et Oakes [63, 1984] donnent les raisons pour lesquelles l'utilisation de la fonction de risque peut être une bonne idée :

1. Elle peut mettre en évidence le fait de considérer le risque immédiat lié à un système d'être en vie à un âge t.
2. la comparaison de groupes d'individus est souvent faite par la fonction de risque,

3. les modèles de risque sont souvent adéquats lorsqu'il y a des données censurées ou plusieurs types de défaillance,
4. la comparaison avec une distribution exponentielle est particulièrement simple en termes de fonction de risque.

La fonction de risque peut prendre plusieurs dénominations selon le domaine d'intérêt, ainsi

* dans les sciences de l'ingénierie, on parle de taux de défaillance,
* en actuariat, elle est connue sous le nom de force de mortalité,
* dans les sciences de la vie et les sciences sociales, on parle d'âge ou taux de décès,
* en économie, elle est connue sous le nom de ratio de Mills, voir Mills [64, 1926],
* concernant les processus ponctuels et la Théorie des Valeurs Extrêmes on parle plutôt de taux de fonction ou intensité de fonction.

Par souci de simplification, l'hypothèse selon laquelle le taux de défaillance est constant et indépendant du temps est très souvent admise. Mais l'expérience a montré que, pour la plupart des composants, bien que cette hypothèse soit acceptable pendant une durée assez longue entre leur jeunesse et leur vieillesse, on constate souvent une évolution de ce taux en fonction du temps en forme de « baignoire » comme le montre la figure (3.2) [65],[66].

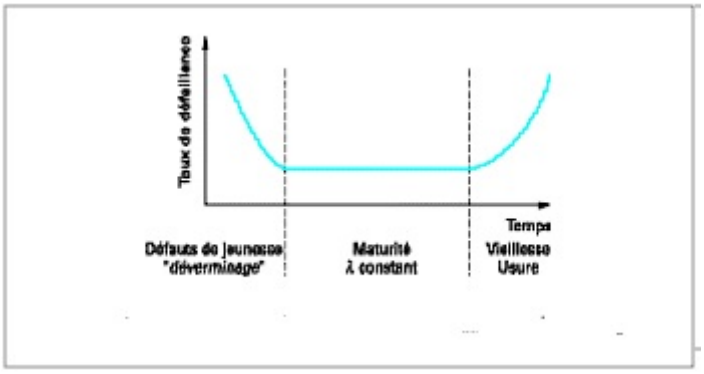

FIGURE 3.2 – Evolution du taux de défaillance en fonction du temps

Selon cette courbe, le taux de défaillance des composants passe par trois périodes :

1. La première période caractérisée par une décroissance du taux de défaillance est dite période de « défaillance précoce ». Elle correspond aux erreurs de conception ou de fabrication, à des composants mal utilisés ou insuffisamment vérifiés. Pour éliminer ces défaillances, le produit subit un déverminage ou rodage [67].
2. La deuxième période correspond à la zone de vie utile des composants où le taux de défaillance est constant. Les mécanismes de défaillance dans cette période sont indépendants du temps.

3. Durant la troisième et dernière période, le taux de défaillance augmente. Les défaillances dans cette période sont majoritairement causées par l'usure et le vieillissement du composant.

3.2.4.1 MTTF (Mean Time To Failure), ou durée moyenne de fonctionnement avant défaillance

est ecrire comme suit :

$$MTTF = E[T] = \int_0^\infty t f((t))\, dt = \int_0^\infty R(t). \tag{3.2}$$

où $E[T]$ représente l'espérence mathématique des durées de vie.

le MTBF (Mean Time Between Failure), ou durée moyenne séparant deux défaillances consécutives dans le cas d'un composant réparable. Dans le cas particulier de dispositifs non réparables, les MTTF et MTBF sont confondus.

3.3 Données Incomplètes

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

3.3.1 Données Censurées

Le phénomène de censure est lié aux évènements perturbateurs qui peuvent se produire dans le laps de temps nécessaire au recueil d'une donnée. Il intervient donc fréquemment lors de mesures qui portent sur les variables modélisant le temps écoulé entre deux évènements : durée de vie d'un individu, durée entre le début d'une maladie et la guérison, durée d'un épisode de chômage, ... etc. Ces perturbations empêchent l'observateur d'accéder à la totalité de l'information concernant le phénomène qu'il étudie et conduit à l'apparition d'observations incomplètes dites censurées. La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Définition 3.3 (variable de censure) *La variable de censure Y est définie par la non-observation de l'événement étudié.Si au lieu d'observation X , on observer Y , et que l'on sait que $X > Y$ (respectivement $X < Y$, $Y_1 < X < Y_2$), on dit qu'il y a censure à droite (*

respectivement censure à gauche ,censure par intervalle). Pour un individu donné j , on va considérer :

— Temps de survie X_j
— Sont temps de censure Y_j
— La durée réellement observée Z_j

3.3.2 Types de Censures

3.3.2.1 Données complètes :

le temps exact de défaillance est observé, cette donnée représente donc la durée de vie du composant étudié dans le cas des composants non réparables, ou la durée entre deux défaillances dans le cas d'un composant réparable. C'est le type de données le plus informatif en fiabilité.

3.3.2.2 Données censurées à droite :

si on décide d'arrêter l'observation à la date t_d et qu'à cette date, le composant n'a pas encore eu de défaillance, la date t_d sera une donnée censurée à droite, et on
n'aura qu'une seule information sur le temps de défaillance $t_d < t$.

3.3.2.3 Données censurées à gauche :

si on décide d'observer l'état du composant à partir de la date t_g et qu'on constate que le composant a déjà été défaillant, t_g sera une donnée censurée à gauche et la seule information sur le temps de défaillance t sera : $t < t_g$.

3.3.2.4 Données censurées par intervalle :

si le composant a eu une défaillance entre deux dates t_d et t_g connues, il s'agit de données censurées par intervalle, et on a seulement l'information : $t_d < t < t_g$.

3.3.2.5 Données censure double :

La censure double (ou mixte) ce type de censure c'est un mélange entre les deux censures, la censure à droite et la censure à gauche, dans le méme échantillon.
L'expérience elle-même peut engendrer cette censure.

3.3.3 Données censure de type 1 : fixée

L'expérimentateur fixe une valeur (une date par exemple non aléatoire de fin d'expérience). Par exemple en épidémiologie on fixe la durée maximale de participation

et vaut, pour chaque observation, la différence entre la date de fin d'expérience et la date d'entrée du patient dans l'étude. Le nombre d'évènements observés est, quant à lui, aléatoire. Soit Y une valeur fixée. Par exemple en censure à droite, au lieu d'observerles variables $X_1,...,X_n$ qui nous intéressent, on observe X_j que lorsqu'elle est inférieure ou égale à une durée fixée Y. On observe donc une variable Z_j telle que $Z_j := \min\left(X_j, Y\right), j = 1,...,n$.Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles.

3.3.4 Données censure de type 2 : attente

L'expérimentateur fixe a priori le nombre d'évènements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'évènements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie. Par exemple en épidémiologie on décide d'observer les durées de survie des n patients jusqu'à ce que r $(1 \leq r \leq n)$ d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{j:n}$ et $Z_{j:n}$ les statistiques d'ordre desvariables X_j et Z_j . La date de censure est donc $X_{r:n}$ et on observe

$$\begin{cases} Z_{j:n} = X_{j:n} \text{ si } j \leq r \\ Z_{j:n} = X_{j:n} \text{ si } j \geq r \end{cases}.$$

3.3.5 Données censure de type 3 : aléatoire

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques.Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient).

soit $X_1,...,X_n$ un échantillon d'une va positive X, on dit qu'il y a censure aléatoire de cet échantillon s'il existe une autre va positive elle aussi Y d'échantillon $Y_1,...,Y_n$ dans ce cas au lieu d'observer les $X_j^{'}s$, on observe un couple de va's $\left(Z_j, \delta_j\right)$ avec

$$Z_j := \min\left(X_j, Y_j\right) \text{ et } \delta_j := \mathbf{1}\left\{X_j \leq Y_j\right\} \text{ pour } j = 1,...,n. \tag{3.3}$$

où δ_j l'indicateur de censure, qui détermine si X a été censuré ou non :

— si $\delta_j = 1$, la durée d'intérêt est observée $\left(Z_j = X_j\right)$.

— si $\delta_j = 0$, elle est censurée $\left(Z_j = X_j\right)$. On observe des durées incomplètes.

3.3.6 Données Tronquées

Les données censurées ne sont pas le type unique de données incomplètes. L'autre cas classique de données incomplètes est celui des données dites tronquées. Le phénomène de troncature est très différent de la censure. La troncature, quant à elle, élimine

de l'étude une partie des X_j . Lors d'une étude pratique sur les durées de vie, il n'est pas rare que la variable d'intérêt X ne soit pas observable quand elle est inférieure à un seuil aléatoire Y , ce qui aura pour conséquence que l'analyse ne pourra porter que sur la loi conditionnelle de X sachant $X > Y$. Il y a trois types de troncature : troncature à gauche, à droite et par intervalle.

— **Troncature à gauche** : Soit Y est une va indépendante de X, on dit qu'il y a troncature à gauche lorsque X (la durée de survie) n'est observable que si $X > Y$. On observe le couple (X, Y) , avec $X > Y$.
— **troncature à droite :** De même, il y a troncature à droite lorsque X n'est observable que si $X < Y$.
— **troncature par intervalle :** Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.

Pour des détails complets sur les types de censure on réfère aux livres de (Pierre [83, page 7]) et (Vivian [84, page 14]). Dans ce travail, on s'intéresse uniquement au modèle de censure à droite du type aléatoire. Celui-ci correspond à un modèle fréquemment utilisé en pratique (voir aussi Soltane [85, page 13]).

En général, une donnée censurée est une donnée pour laquelle on ne connaît pas la date exacte de défaillance. Le traitement des données censurées est une des préoccupations majeures des fiabilistes. Dans notre étude nous considérons que les composants que nous testons sont non réparables. De plus, les tests sont configurés de façon à pouvoir observer les défaillances des composants testés dans un temps relativement court (les tests sont dits tests de vieillissement accéléré). Par conséquent, le temps exact de défaillance du composant testé (donc sa durée de vie étant donné que le composant est non réparable) est observé et toutes les données sont non censurées.

3.4 Lois de distribution des durées de vie

Après avoir introduit les notions générales utilisées en fiabilité, nous pouvons remarquer que toutes les fonctions évoquées sont naturellement liées entre elles : la connaissance de $R(t)$ implique celle de $f(t)$ et donc celle de $\lambda(t)$. Ainsi, il suffit de spécifier la distribution des durées de vie (variable aléatoire T) pour déterminer toutes les grandeurs d'intérêt relatives à la fiabilité. Parmi les lois de distributions statistiques, certaines sont d'un intérêt particulier dans le cas d'étude des durées de vie (variable aléatoire continue et positive), vue la validité de leur application dans la majorité des cas [65], [68], [69], [70] : la loi exponentielle, la loi de Weibull et la loi log-normale.

3.4.1 Loi exponentielle

La loi exponentielle est la seule loi caractérisée par un taux de défaillance constant λ qui est son unique paramètre. Les fonctions et grandeurs caractéristiques de la loi exponentielle sont exprimées dans le tableau 3.1 La figure 3.3 illustre, pour différentes valeurs du paramètre λ, les fonctions densité de probabilité et les fonctions de répartition de la loi exponentielle.

TABLE 3.1 – Loi exponentielle

$f(t) = \lambda \exp(-\lambda t)$	$R(t) = \exp(-\lambda t)$
$\lambda(t) = \lambda$	$MTTF = \frac{1}{\lambda}$

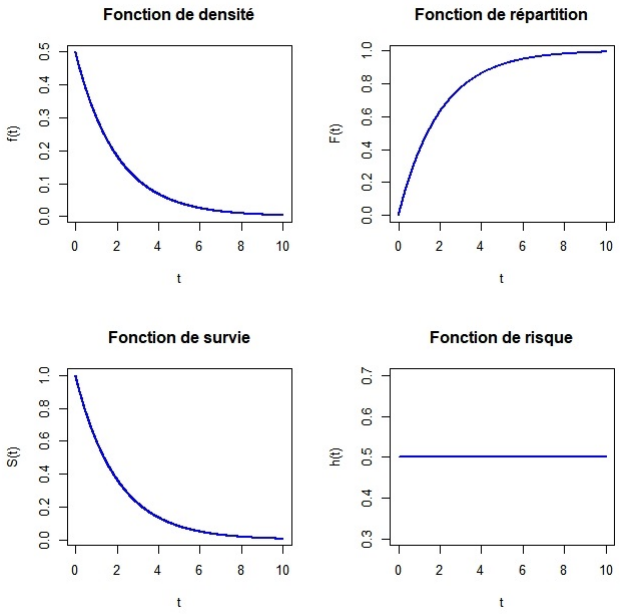

FIGURE 3.3 – Fonction de densité et Fonction de répartition et de Fonction de risque et Fonction de survie de Loi exponentielle

Historiquement, la loi exponentielle était la première loi utilisée en fiabilité, pour sa simplicité d'une part, et parce qu'elle permet de modéliser la fiabilité des composants dans leur période de vie utile la plus longue [69]. Cependant, la loi exponentielle décrit un processus sans mémoire caractérisé par un taux de défaillance constant. Afin d'illustrer cette propriété, nous pouvons vérifier que pour deux durées t_1 et t_0 telles que $t_1 > t_0$:

$$\Pr(T > t_1 \mid T > t_0) = \Pr(T > t_1 - t_0)$$

3.4.2 Loi de Weibull

La loi de Weibull est caractérisée par deux paramètres :

— un paramètre d'échelle η dont l'unité est homogène à celle de la durée de vie.
— un paramètre de forme β qui détermine la forme de variation du taux de défaillance : monotone décroissant ($0 < \beta < 1$), monotone croissant ($\beta > 1$) ou constant ($\beta = 1$, on retrouve alors la loi exponentielle avec un paramètre $\lambda = \frac{1}{\eta}$).

Le tableau 3.2 résume les fonctions et grandeurs caractéristiques de la loi de Weibull. La figure 3.4 illustre, pour un paramètre d'échelle fixe $\eta = 1$ et différentes valeurs du paramètre de forme β, les

fonctions densité de probabilité, les fonctions de répartition et le taux de défaillance associés à la loi de Weibull.

TABLE 3.2 – La loi de Weibull est caractérisée par deux paramètres

$f(t) = \frac{\beta}{\eta}(\frac{t}{\eta})^{\beta-1}\exp(-\frac{t}{\eta})^{\beta}$	$R(t) = \exp(-\frac{t}{\eta})^{\beta}$
$\lambda(t) = \frac{\beta}{\eta}(\frac{t}{\eta})^{\beta-1}$	$MTTF = \eta\Gamma(1+\frac{1}{\beta})$ ou $\Gamma(n) = \int_0^{\infty}\exp(-x)x^{n-1}dx$

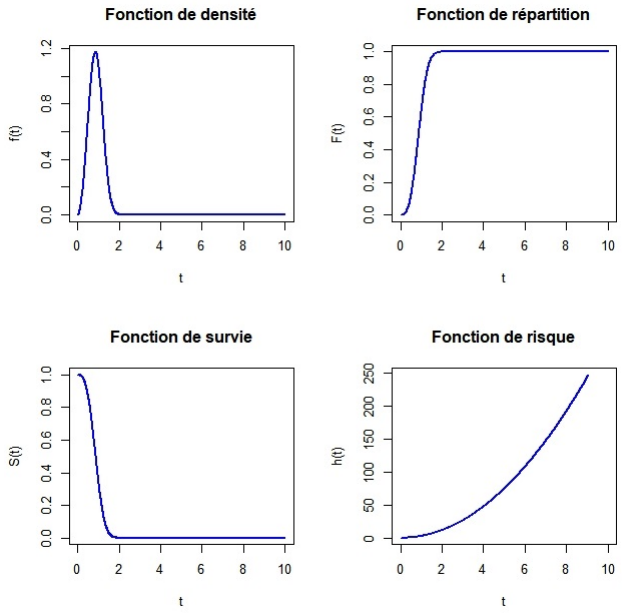

FIGURE 3.4 – Fonction de densité et Fonction de répartition et de Fonction de risque et Fonction de survie de loi de Weibull

En associant les trois possibilités évoquées de la forme du taux de défaillance, la loi de Weibull peut représenter les trois étapes de la vie d'un composant : défaillance précoce ($0 < \beta < 1$), vie utile ($\beta = 1$) ou vieillissement ($\beta > 1$). Ainsi, la loi de Weibull est très souvent utilisée en fiabilité pour modéliser la distribution de la durée de vie des composants et ceci est valable dans plusieurs domaines d'application (électronique, mécanique, etc.) [73],[74].

3.4.3 Loi log-normale

Une variable aléatoire positive T suit une loi log-normale de paramètres μ et σ^2 si son logarithme suit une loi normale de moyenne μ et de variance σ^2. La fonction de fiabilité et, par conséquent, le taux de défaillance de la loi log-normale n'ont pas d'expressions explicites. Ils peuvent être calculés numériquement. En effet, la fonction de fiabilité de la loi log-normale fait intervenir la fonction de répartition de la gaussienne standard :

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-\frac{u^2}{2}) du \tag{3.4}$$

D'autre part, il a été démontré dans [69] que pour $\sigma \leq 1$ le taux de défaillance de la loi log-normale admet une forme générale de cloche : il vaut 0 pour $t = 0$, croit et atteint un maximum, puis décroit pour atteindre 0 quand t tend vers l'infini. Pour $\sigma > 1$, le taux de défaillance est décroissant.

Le tableau 3.3 donne les expressions de la fonction densité de probabilité et du MTTF de la loi lognormale, et celle des fonctions de fiabilité et du taux de défaillance en fonction de ϕ. .La distribution Log-normale est un modèle fréquemment utilisé en fiabilité, car elle concerne des variables aléatoires positives, et le paramètre de forme σ lui permet des représentations variées [75] [76].

TABLE 3.3 – Caractéristiques de loi log-normale

$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\frac{\ln(t)-\mu}{\sigma})^2\right]$	$R(t) = 1 - \phi(\frac{\ln(t)-\mu}{\sigma})$
$\lambda(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\frac{\ln(t)-\mu}{\sigma})^2\right]}{1-\phi(\frac{\ln(t)-\mu}{\sigma})}$	$MTTF = \exp(\mu + \frac{1}{2}\sigma^2)$

3.5 Introduction de co-variables et modèles paramétriques de la durée de vie

3.5.1 Les modèles composites

L'objet de cette section est de décrire les principales caractéristiques des modèles de base couramment utilisés dans un cadre paramétrique ou semi-paramétrique, et faisant appel à un degré de sophistication supérieur à la simple analyse d'un échantillon iid de loi paramétrique fixée a priori. Il s'agit de modèles que l'on rencontre en général lorsque l'on est confronté à une population hétérogène, composées d'individus avec des lois de survie différentes, on a donc choisi de désigner ces modèles sous le nom générique de « *modèles composites* », et ils diffèrent par la manière dont l'hétérogénéité est

prise en compte. Les modèles purement non paramétriques seront étudiés par ailleurs , ils ne sont pas évoqués ici.

3.5.2 Les mélanges de lois

3.5.3 Exemple introductif

On considère un système composé de deux éléments indépendants montés en parallèle, chacun des éléments ayant une durée de vie de loi exponentielle, avec des paramètres λ_1 et λ_2 .La durée de vie de l'équipement est mesurée par

$$T = T_1 \vee T_2,$$

la loi de T s'obtient facilement en observant que

$$1 - S(t) = (1 - \exp(-\lambda_1 t))(1 - \exp(-\lambda_2 t)).$$

On en déduit que dans le cas général la fonction de hasard est d'abord croissante, puis décroissante ; si $\lambda_1 = \lambda_2$, la fonction de hasard est croissante. L'indépendance temporelle est donc une propriété peu stable et elle se perd rapidement.On va voir qu'elle se perd également dans le cas de l'agrégation de lois.

3.5.4 Agrégation de lois

Il arrive souvent en pratique que les durées que l'on observe résultent de l'agrégation de sous-populations ayant chacune un comportement spécifique, souvent inobservable. On parle alors d'hétérogénéité. On suppose ici que la fonction de survie dépend d'un paramètre aléatoire ν , ce paramètre étant distribué selon une loi π . D'un point de vue heuristique, on se trouve en présence de sous-populations à l'intérieur desquelles la loi de survie est homogène et décrite par la loi de survie conditionnelle au fait que la valeur du paramètre soit ν , $S(t, \nu)$, la loi π décrivant le poids respectif de chaque sous-population dans la population totale.

On a donc la forme suivante pour la fonction de survie initiale de la population totale :

$$\begin{aligned} S(t) &= \int S(t, \nu)\pi d\nu : \\ S(t) &= \Pr(T > t) = E_\nu[\Pr(T > t \mid \nu)] = \int S(t, \nu)\pi d\nu . \end{aligned}$$

La distribution d'hétérogénéité dépend a priori de t, puisque les individus des différentes sous-populations ne sortent pas du groupe à la même vitesse. À la date t, et

en supposant la taille de la population infinie, on a ainsi :

$$\pi_t(d\nu) = \frac{S(t,\nu)}{S(t)}\pi d\nu \ .$$

La fonction de hasard à la date t s'écrit alors $h(t) = \int h(t,\nu)\pi_t d\nu$. En effet, il suffit de remarquer que :

$$u^{-1}\Pr(T \leq t+u \mid T > t) = \int u^{-1}\Pr(T \leq t+u \mid T > t\ ,\nu)\pi_t(d\nu)\ ,$$

puis de faire tendre u vers 0. Dans le cas particulier où $S(t,\nu) = \exp(-\lambda(\nu)t)$, c'est à dire où chaque sous-population est décrite par une loi exponentielle de paramètre $h(t,\nu) = \lambda(\nu)$, la fonction de survie agrégée s'écrit :

$$S(t) = \int \exp(-\lambda(\nu)t)\pi d\nu$$

D'après l'expression ci-dessus de la fonction de hasard s'écrit donc $h(t) = \int \lambda(\nu)\pi_t(d\nu)$ et on en déduit que :ν

$$\frac{dh(t)}{dt} = -\int \lambda^2(\nu)\pi_t(d\nu) + (\lambda(\nu)\pi_t(d\nu))^2$$

En effet, de l'expression de $\pi_t(d\nu) = \frac{S(t,\nu)}{S(t)}\pi d\nu$ il découle :

$$\frac{\partial}{\partial t}\pi_t(d\nu) = \frac{\frac{\partial}{\partial t}S(t,\nu)\times S(t) - S(t,\nu)\times\frac{d}{dt}S(t)}{S(t)^2}\pi d\nu$$

avec

$$\frac{\partial}{\partial t}S(t,\nu) = -\lambda(\nu)S(t,\nu)$$

et

$$\frac{s'(t)}{S(t)} = -h(t) = -\int \lambda(\nu)\pi_t(d\nu).$$

On en déduit :

$$\frac{\partial}{\partial t}\pi_t(d\nu) = \frac{-\lambda(\nu)\times S(t,\nu)}{S(t)}\pi(d\nu) + \frac{S(t,\nu)\times h(t)}{S(t)}\pi(d\nu) = -\lambda(\nu)\pi_t(d\nu) + h(t)\pi_t(d\nu)$$

En écrivant

$$\frac{d}{dt}h(t) = \int \lambda(\nu)\frac{\partial}{\partial t}\pi_t(d\nu)$$

on trouve donc finalement :

$$\frac{dh(t)}{dt} = -\int \lambda^2(\nu)\pi_t(d\nu) + h(t)^2$$

Ce qui est le résultat attendu. Cette égalité implique par l'inégalité de Schwarz (ou en remarquant que $\frac{dh(t)}{dt} = -V_{\pi_t}(\lambda(\nu))$que $\frac{dh(t)}{dt} \leq 0$; l'agrégation de fonctions de hasard constantes conduit donc à une fonction de hasard globale décroissante. Ce phénomène s'explique par le fait que les individus ayant une valeur élevée de $\lambda(\nu)$ sortent en premier et il reste donc proportionnellement plus d'individus à $\lambda(\nu)$ faible lorsque le temps s'écoule. Le taux de sortie est donc logiquement décroissant. Ce phénomène porte le nom de « *biais d'hétérogénéité* », ou « *mobile-stable* ».

Exemple 3.1 *mélange de 2 lois exponentielles La durée est ici une variable exponentielle de paramètre λ_1 avec la probabilité p et λ_2 avec la probabilité $1-p$, soit :*

$$S(t) = p\exp(-\lambda_1 t) + (1-p)\exp(-\lambda_2 t)$$

La fonction de hasard a alors l'allure suivante :

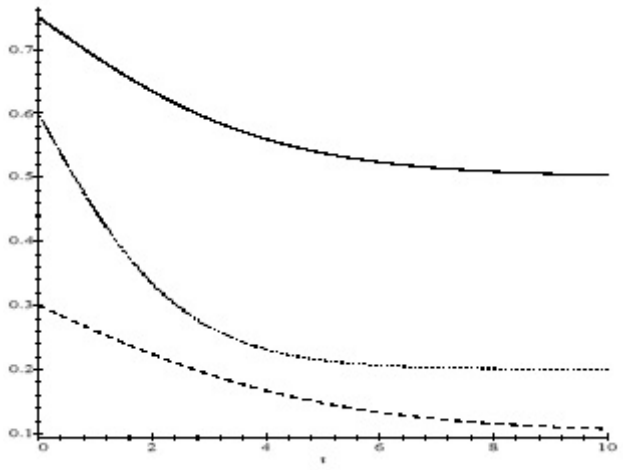

FIGURE 3.5 – Mélange de 2 lois exponentielles

On voit que le risque instantané peut être rapidement décroissant, alors même que les 2 fonctions d'origine sont à risque constant.

3.6 Modèles à durée de vie accélérée AFT

Dans ce qui précède, les données étaient considérées comme des réalisations d'une variable aléatoire d'une seule distribution. Cependant, en réalité, plusieurs facteurs contribuent à la dégradation du composant ou de l'entité en question. Les réalisations

d'une variable aléatoire d'une même distribution sont donc nécessairement soumises aux mêmes conditions, aux mêmes facteurs de dégradation. Par conséquent, les durées de vie doivent être expliquées par les différentes variables (facteurs de dégradation) contribuant à la défaillance. Ainsi, bien qu'en général un seul type de distribution puisse être considéré pour un type de données en fiabilité, les paramètres de cette distribution doivent être exprimés en fonction des variables explicatives.

Dans cette approche paramétrique, le modèle à variables explicatives le plus souvent utilisé en fiabilité est le modèle à durée de vie accéléré ou modèle AFT par référence au terme anglais (Accelerated Failure Time model) [78], [78]. Ce modèle suppose que, pour deux populations différentes de la variable aléatoire T de fonctions de survie R_1 et R_2, si T subit un changement d'échelle par un facteur multiplicatif $c > 0$ indépendant du temps (dit facteur d'accélération), la probabilité de survie reste la même. En introduisant p variables explicatives sous forme d'un vecteur X affecté d'un vecteur de coefficients ρ, le facteur c peut s'écrire sous la forme :

$$c(X) = \exp(\rho X) \tag{3.5}$$

L'égalité des fonctions de survie s'écrit :

$$R_1(T) = R_2(ct) \tag{3.6}$$

Ainsi, les co-variables affectent uniquement l'échelle de temps : si $c(X) < 1$, l'effet des co-variables est de ralentir le temps, sinon les co-variables accélèrent le temps.

En supposant que ε est une variable aléatoire centrée réduite, et que $R_2(T)$est la fonction de survie de la variable aléatoire $\exp(\mu + \sigma\varepsilon)$, μ et σ étant des constantes, nous pouvons écrire :

$$R_1(T) = R_2(ct) = \Pr(\exp(\mu + \sigma\varepsilon) > ct) = \Pr(\exp(\mu + \sigma\varepsilon) > t\exp(\rho X)) \tag{3.7}$$

$$= pr > t = pr(T > t) \tag{3.8}$$

Ainsi, en effectuant le changement de variable $\beta = -\rho$, $R_1(T)$ est la fonction de survie de la variable aléatoire T telle que :

$$\ln(T) = \mu + X\beta + \sigma\varepsilon \tag{3.9}$$

Ainsi, avec le modèle AFT, nous obtenons une relation linéaire entre le logarithme de la durée de vie et les variables explicatives avec un terme d'erreur ε dont les composantes sont indépendantes et identiquement distribuées (i.i.d.) selon une loi connue a priori, à partir de la loi de la durée de vie T. Dans ce modèle, μ représente le logarithme de la durée de vie quand les facteurs sont nuls (constante à l'origine), β est le

vecteur des coefficients du modèle, et σ est le paramètre d'échelle de la loi de $\ln(T)$.

La loi de distribution de T définit celle de $\ln(T)$ et par conséquent celle de ε. La correspondance entre les lois est donnée dans le tableau 3.4 [69]

TABLE 3.4 – Lois de distribution de T et leurs correspondants par transformation logaritmique

Loi de distribution de T	Loi de distribution de ε
Exponentielle	Valeur extrême à 1 paramètre
Weibull	Valeur extrême à 2 paramètre
Log-normale	Normale

Les paramètres des lois de distribution et les coefficients β peuvent être estimés par maximum de vraisemblance, en maximisant la log-vraisemblance de la variable aléatoire :

$$\varepsilon = \frac{\ln(T) - \mu - X\beta}{\sigma} \tag{3.10}$$

Les lois de distributions de $\ln(T)$ dans un modèle AFT sont du type localisation-échelle (location-scale). Ce type de distribution a une fonction de survie de la forme :

$$R(y,u,b) = R_0\left(\frac{y-u}{s}\right) \qquad -\infty < y < +\infty \tag{3.11}$$

Ce type de distribution a deux paramètres en général : un paramètre de localisation u $(-\infty < u < +\infty)$, et un paramètre d'échelle $s(s > 0)$. Par exemple pour la loi normale, u représente la moyenne, et s l'écart-type. Dans le modèle AFT, u correspond à $(\mu + X\beta)$ et s à σ. Ainsi, dans les modèles AFT, seul le paramètre de localisation u de la loi des durées de vie transformées $\ln(T)$ est fonction des variables explicatives, le paramètre d'échelle s étant constant. Par conséquent, pour une entité soumise à un ensemble de contraintes, la loi de distribution des durées de vie est la même et subit uniquement une translation (changement de localisation) quand les contraintes varient.

Ainsi, dans une approche paramétrique, l'évaluation de la fiabilité d'un composant par l'estimation de sa durée de vie nécessite la spécification de la loi de distribution statistique des données et de la relation analytique entre la durée de vie et le(s) stress appliqué(s), appelée modèle durée de vie-stress.

Ces deux modèles sont naturellement reliés entre eux puisque, d'une part, les paramètres de la loi de distribution sont estimés conditionnellement aux facteurs de stress appliqués et que d'autre part, le modèle durée de vie-stress relie un quantile de la distribution considérée ou un de ses paramètres (par exemple la moyenne, la médiane ou le paramètre de localisation) aux contrainte(s) appliquée(s).

Si nous revenons au modèle AFT à distribution log-normale et après transformation logarithmique, nous remarquons que la moyenne conditionnelle aux variables explicatives est le paramètre de la loi normale qui est représenté dans l'expression de la

relation linéaire entre la durée de vie et le(s) stress :

$$E\left[\ln\left(T \setminus X\right)\right] = \mu = X\beta \tag{3.12}$$

Si par ailleurs nous prenons l'exemple d'un modèle AFT à distribution de Weibull et après transformation logarithmique, le paramètre de la loi intervenant dans le modèle durée de vie-stress est le paramètre de localisation de la loi des valeurs extrêmes $u = \ln(\eta)$, tel que, pour des contraintes données :

$$\ln(\eta) = u = X\beta \tag{3.13}$$

L'objectif de notre travail est de développer une méthode générale pour la modélisation de la durée de vie en fonction des facteurs de stress appliqués et qui prend en compte différentes contraintes expérimentales et économiques que nous explicitons dans les chapitres suivants. Dans une première approche paramétrique, les modèles de durée de vie seront développés selon la forme générale (3.9) des modèles AFT. Cependant, le logarithme décimal sera utilisé au lieu du logarithme népérien pour une interprétation plus facile des logarithmes des durées de vie. Il est donc nécessaire de vérifier au préalable l'adéquation des données de durée de vie qui seront utilisées pour l'estimation des paramètres du modèle aux deux lois de distribution (log-normale et Weibull). Ceci revient alors à tester l'adéquation des logarithmes des durées de vie aux lois normale et valeur extrême respectivement. Nous présentons alors les résultats de ces tests dans le chapitre suivant.

3.7 Les modèles à hasard proportionnel

Il s'agit d'un modèle semi-paramétrique dans lequel on se donne une fonction de survie de base, $B(t)$ et on fait l'hypothèse que la fonction de survie du phénomène observé est de la forme $S_\theta(t) = B(t)^\theta$, pour un paramètre $\theta > 0$ inconnu. Il est immédiat que la densité sousjacente s'écrit $f_\theta = \theta B(t)^{\theta-1} f(t)$, et la fonction de hasard est donc de la forme :

$$h_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)} = \theta \frac{f(t)}{B(t)} = \theta h(t)$$

La fonction de hasard est ainsi proportionnelle à la fonction de hasard de base associée à $\theta = 1$, d'où la dénomination de « *modèle à hasard proportionnel* ». Le modèle exponentiel constitue un cas particulier de modèle à hasard proportionnel dans lequel la fonction de hasard de base est constante égale à l'unité.

On peut remarquer que ces modèles satisfont la propriété suivante : si la variable

aléatoire T_θ est associée à la fonction de survie

$$S_\theta(T) = B(t)^\theta,$$

alors

$$E(T_\theta) = \int_0^{+\infty} S_\theta(t)\,dt = \int_0^{+\infty} B(t)^\theta\,dt$$

or on reconnaît dans

$$\rho_\theta(T) = \int_0^{+\infty} B(t)^\theta\,dt$$

la mesure de risque de Wang PLANCHET et al.[77, 2011] associée à la fonction de distorsion $g_\theta(x) = x^\theta$ (appelée PH-tranform de paramètre $\frac{1}{\theta}$) .

En spécifiant différentes formes pour le coefficient de proportionnalité, on est conduit à définir différentes classes de modèles.

3.7.1 Le modèle de Cox

Le modèle de Cox [80, 1972], ou "*modèle continu semi-paramétrique à risques proportionnels*", est un modèle de régression en temps continu. L'objectif est de modéliser le logarithme du rique instantané en fonction d'un ensemble de variables explicatives x dont la valeur peut éventuellement varier au fil du temps :

$$\ln h(t,X) = \beta_0(t) + \sum_k \beta_k X_k(t) = \beta_0(t) + X'\beta \tag{3.14}$$

De façon équivalente :

$$h(t,X) = h_0(t)\exp\left(\sum_k \beta_k X_k(t)\right) = h_0(t)\exp\left(X'\beta\right) \tag{3.15}$$

- Le terme $h_0(t)$ est un risque de base indépendant des facteurs explicatifs du modèle.
- Aucune hypothèse n'est faite sur la distribution des durées, c'est-à-dire sur la forme de $h(t)$ ou $R(t)$.
- Le modèle de Cox est capable d'approximer correctement des modèles paramétriques (Weibull, exponentiel,...).
- Si le vrai modèle paramétrique est inconnu, Cox est une bonne alternative. Sinon, il vaut mieux utiliser le modèle paramétrique.
- Comme pour une courbe de survie, il est nécessaire de disposer de deux variables particulières pour estimer un modèle de Cox :
  - une variable indiquant la durée de temps jusqu'à la survenance de l'évènement ou jusqu'à la fin de la période d'observation dans le cas de données

censurées .
— une variable codée 1 si l'évènement a eu lieu et zéro sinon.
— Lorsque toutes les variables explicatives sont invariantes dans le temps, le modèle peut être calculé directement sur le jeu de données, alors que lorsque certains facteurs évoluent au cours du temps, une procédure particulière est appliquée au préalable sur les données
— S'il n'y a pas de facteurs explicatifs évoluant dans le temps, le modèle s'apparente à une "simple" régression linéaire.

3.7.2 Les modèles de fragilité

Dans le modèle de Cox on cherche à modéliser l'effet de variables explicatives connues sur le niveau de la fonction de risque, dans certaines situations, ces variables sont inobservables, et on souhaite tout de même évaluer les conséquences de ces variables inobservables sur la forme de la fonction de survie .

On repart de la formulation

$$S_\theta(t) = S(t \mid \theta) = B(t)^\theta \tag{3.16}$$

ou, de manière équivalente, $h_\theta(t) = \theta h(t)$, d'un modèle à hasard proportionnel, et on considère que le paramètre θ est une variable aléatoire ; en d'autres termes on se donne la loi de survie conditionnelle au paramètre, et la loi globale s'obtient donc par intégration :

$$S(t) = E\left[B(t)^\theta\right] \tag{3.17}$$

l'espérance étant calculée par rapport à la loi de θ. Cette expression est analogue à l'expression $S(t) = \int S(t,v)\pi(dv)$ obtenue à la section 3.5.4. Le paramètre θ s'appelle la « *fragilité* ». Ces modèles sont également parfois appelés « *modèles à effets aléatoires* ».

3.8 Les modèles à causes de sortie multiples

Dans certaines situations on est amené à distinguer entre différentes causes de sortie, par exemple en décès on s'intéresse à la cause du décès, en arrêt de travail au motif de la sortie d'incapacité (retour au travail ou passage en invalidité), etc. C'est typiquement ce qu'on fait lorsqu'on interprète le modèle de Makeham.

Si on note $T_1, ..., T_n$ les variables de durée associées à chacune des causes étudiées, la survie globale est simplement $T = T_1 \wedge .. \wedge T_n$, sous l'hypothèse d'indépendance des différentes composantes le modèle est simple et la fonction de hasard globale est la somme des fonctions de hasard. Mais l'hypothèse d'indépendance peut être parfois

restrictive, et les modèles de fragilité fournissent un moyen simple de la relâcher. Cette approche a été proposée initialement par OAKES [79, 1989].

On suppose donc que les durées associées à chaque cause, $T_1, ..., T_n$ sont indépendantes conditionnellement à θ et que les marginales (conditionnelles) sont de la forme

$$S_i(t \mid \theta) = B_i(t)^{\theta}.$$

On est alors ramené à des calculs proches de la section 3.7.2 ci-dessus et on trouve :

$$S(t_1, ..., t_n) = E\left[\prod_{i=1}^{n} B_i(t_i)^{\theta}\right]$$

Exemple 3.2 *avec deux causes de sortie distribuées chacune suivant une loi de Weibull et une distribution du paramètre de mélange selon une loi stable de paramètre a, on trouve*

$$S(t) = \exp\{-(\lambda_1 t^{\alpha_1} + \lambda_2 t^{\alpha_2})\},$$

qui est une conséquence immédiate de

$$E(\exp(-x\theta)) = \exp\left(-x^{\theta}\right)$$

et de l'expression de la fonction de survie de la loi de Weibull

$$S(t) = \exp(-\lambda t^{\alpha}).$$

3.9 Les modèles à choc commun

L'idée est ici que la durée de survie dépend de deux facteurs, l'un propre à l'individu et l'autre affectant la population dans son ensemble. Ce second facteur peut être un facteur accidentel ou environnemental. On considère le modèle :

$$T_i = X_i \wedge Z$$

avec S_i la fonction de survie de X_i et S_z la fonction de survie de Z. La loi conjointe du vecteur $(T_1, ..., T_n)$ s'obtient en observant que l'évènement $\{X_i \wedge Z > t\}$ est égal à $\{X_i > t\} \cap \{Z > t\}$, ce qui conduit à :

$$S(t_1, ..., t_n) = \prod_{i=1}^{n} S_i(t_i) \times S_z(\max(t_1, ..., t_n)).$$

MARSHALL et OLKIN [81, 1967] proposent par exemple une distribution exponentielle pour Z.

3.10 Estimateur de Kaplan-Meier de la survie

3.10.1 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier [82] découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t, c'est-à-dire, si $t^{''} < t^{'} < t$

$$\begin{aligned}\Pr(X > t) &= \Pr\left(X > t^{'}, X > t\right)\\ &= \Pr\left(X > t \mid X > t^{'}\right) \times \Pr\left(X > t^{'}\right)\\ &= \Pr\left(X > t \mid X > t^{'}\right) \times \Pr\left(X > t^{'} \mid X > t^{''}\right) \times \Pr\left(X > t^{''}\right)\end{aligned}$$

En considérant les temps d'évènements (décès et censure) distincts $T_{(i)}\,(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$\Pr\left(X > T_{(i)}\right) = \prod_{k=1}^{j} \Pr\left(X > T_{(k)} \mid X > T_{(k-1)}\right),$$

avec $T_{(0)} = 0$. Considérons les notations suivantes :

— Y_i le nombre d'individus à risque de subir l'évènement juste avant le temps $T_{(i)}$,

— d_i le nombre de décès en $T_{(i)}$.

Alors la probabilité p_i de mourir dans l'intervalle $\left]T_{(i-1)}, T_{(i)}\right[$ sachant que l'on était vivant en $T_{(i-1)}$, *i.e.*$p_i = \Pr\left(X \leq T_{(i)} \mid X > T_{(i-1)}\right)$, peut être estimée par

$$\widehat{p_i} = \frac{d_i}{Y_i}.$$

Comme les temps d'évènements sont supposés distincts, on a

$d_i = 0$ en cas de censure en $T_{(i)}$, *i.e.* quand $\delta_i = 0$,

$d_i = 1$ en cas de décès en $T_{(i)}$, *i.e.* quand $\delta_i = 1$.

On obtient alors l'estimateur de Kaplan-Meier :

$$\widehat{S}(t) = \prod_{\substack{i=1,...,n \\ T_{(i)} \leq t}} \left(1 - \frac{\delta_i}{Y_i}\right) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{\delta_i}{n-(i-1)}\right) = \prod_{i:T_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}.$$

L'estimateur $\widehat{S}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier [82] est un estimateur du maximum de vraisemblance. $\widehat{S}(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier [82] dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car les

temps de décès ne sont pas connus).

Remarque 3.4 *Dans le cas où il y a des ex-aequo :*

— si ce sont des évènements de nature di érente, on considère que les observations non censurées ont lieu avant les censurées,
— si il y a plusieurs décès au même temps $T_{(i)}$, alors $d_i > 1$ et on a

$$\widehat{S}(t) = \prod_{\substack{i=1,\dots,n \\ T_{(i)} \leq t}} \left(1 - \frac{d_i}{Y_i}\right)$$

Remarque 3.5 *Estimation empirique :*

Pour un échantillon *i.i.d.* de durées non censurées $(X_i)_{i=1,\dots,n}$, un estimateur naturel de la survie de la variable X est la survie empirique

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i > x\}}.$$

Cet estimateur a de bonnes propriétés en terme de convergence : convergence *p.s* (Glivenkocantelli), convergence en loi du processus empirique associé vers un pont brownien.Néanmoins, dans le cas des données censurées, la variable d'intérêt n'est plus la variable observée. Ainsi estimer la survie S par la survie empirique des données observées $(T_i)_{i=1,\dots,n}$ $\left(S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i > x\}}\right)$fournit une estimation biaisée de S (les censures (qui ne sont pas des décès) sont considérées comme des décès : il y a une sous estimation de la survie) : Il en est de même si on estime la fonction de survie par la survie empirique des données observées non censurées (échantillon tronqué). Notons que quand il n'y a pas de censure, l'estimateur de Kaplan-Meier se réduit à la fonction de survie empirique.

3.10.2 Estimation de la variance de $\widehat{S}(t)$

L'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier est

$$\widehat{Var}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{i:T(i)\leq t} \frac{d_i}{Y_i\left(Y_i - d_i\right)}.$$

Il est obtenu en utilisant l'approximation suivante,

$$\widehat{Var}\left(\log\left(\widehat{S}(t)\right)\right) \approx \sum_{i:T(i)\leq t} \frac{d_i}{Y_i\left(Y_i - d_i\right)},$$

et en appliquant la delta-méthode $\left(Var(f(Z)) \approx \left[f^{'}(E(Z))\right]^2 Var(Z)\right)$ pour montrer que

$$\widehat{Var}\left(\log\left(\widehat{S}(t)\right)\right) \approx \frac{1}{\widehat{S}(t)^2}\widehat{Var}\left(\widehat{S}(t)\right).$$

Remarque 3.6 *Ce résultat s'obtient, de manière théorique, de la propriété de normalité asymptotique de l'estimateur de Kaplan-Meier.*

Théorème 3.1 *En tout point de continuité de S, $t_0 \in [0,\tau]$ et $S(\tau^-) > 0$,*

$$\sqrt{n}\left(\widehat{S}(t_0) - S(t_0)\right) \underset{n\to\infty}{\overset{\mathfrak{L}}{\to}} \mathcal{N}\left(0, V^2(t_0)\right),$$

avec

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0} \frac{S(du)}{S^2(u)G(u)},$$

$G(t)$ la fonction de survie de la variable C.

Considérons les quantités $H(t) = \Pr(T > t)$ et $H_1(t) = \Pr(T > t, \delta = 1)$. D'aprèsl'hypothèse d'indépendance, on obtient les égalités suivantes

$$\begin{aligned} H(t) &= \Pr(T > t) = P(X > t, C > t) = S(t)G(t) \\ H_1(t) &= \Pr(T > t, \delta = 1) = P(X > t, C \geq X) = E\left(1\!1_{\{X>t\}}G(X^-)\right) \\ \int_t^\infty G(u^-)f(u)d(u) &= -\int_t^\infty \int_t^\infty G(u^-)S(du). \end{aligned}$$

Par conséquent, $H_1(dt) = G(t^-)S(dt)$ et on peut ainsi écrire

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0} \frac{H_1(du)}{S(u)H(u)G(u^-)}.$$

En remplaçant les fonctions H et H_1 par leurs équivalents empiriques (calculables car les variables T et δ sont observées),

$$\widehat{H}(u) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{T_i>u\}} \text{ et } \widehat{H}_1(u) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{T_i>u,\delta_i=1\}}$$

et S par $\widehat{S}$, on obtient l'estimateur suivant

$$\widehat{V}^2(t_0) = -\widehat{S}^2(t_0)\int_0^{t_0} \frac{\widehat{H}_1(du)}{\widehat{H}(u)\widehat{H}(u)}$$

Un estimateur de la variance de l'estimateur de Kaplan-Meier (qui converge presque

sûrement vers la variance asymptotique de $\widehat{S}$) est

$$\widehat{Var}\left(\widehat{S}(t)\right) = \frac{1}{n}\widehat{V}^2(t).$$

Avec les notations, Y_i le nombre d'individus à risque de subir l'évènement juste avant letemps $T_{(i)}$ et d_i le nombre de décès en $T_{(i)}$, on remarque que

$$\begin{aligned}
\widehat{H}(u) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{T_{(i)}>u\}} = \frac{Y_i - d_i}{n}, \\
\widehat{H}(u^-) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{T_{(i)}\geq u\}} = \frac{Y_i}{n}, \\
\widehat{H}_1(u) &= -\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{T_{(i)}\in[u,u+du],\delta_i=1\}} = -\frac{d_i}{n}.
\end{aligned}$$

Ainsi, on obtient

$$\widehat{Var}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{i:T_{(i)}\leq t}\frac{d_i}{Y_i(Y_i - d_i)}$$

Dans chaque intervalle de temps, l'estimation de la survie est une proportion. On peut donc, sous certaines conditions, faire une approximation par la loi normale et ainsi obtenir un intervalle de confiance,

$$Ic(\alpha) = \left[\widehat{S}(t) \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\widehat{S}(t)\right)}\right].$$

Il ne faut pas utiliser cet intervalle quand $\widehat{S}(t)$ est proche de 0 ou de 1. En e et, l'intervalle étant symétrique autour de $\widehat{S}(t)$, les bornes peuvent dépasser les valeurs 0 ou 1. On préfère utiliser l'intervalle de con ance de Rothman qui contournent cette di culté :

$$Ic(\alpha) = \frac{K}{K + \left(z_{\frac{\alpha}{2}}\right)^2}\left[\widehat{S}(t) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{2K} \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\widehat{S}(t)\right) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{4K^2}}\right],$$

avec

$$K = \frac{\widehat{S}(t)\left(1 - \widehat{S}(t)\right)}{\widehat{Var}\left(\widehat{S}(t)\right)}.$$

Remarque 3.7 *On montre que, sous certaines conditions, l'estimateur de Kaplan-Meier est uniformément consistant, asymptotiquement normal et presque sans biais quand le nombre d'individu à risque est grand.*

3.10.3 Estimation de l'IVE avec censure

Des techniques statistiques pour analyser les ensembles de données censurées sont maintenant très bien étudiées, mais elles concernent principalement des caractéristiques centrales de la distribution sous-jacente. On va s'intéresser dans cette Section au problème de l'estimation de l'IVE et cela en présence de données censurées aléatoirement à droite. Ce problème est très récent dans la littérature, les premiers qui ont mentionné le sujet sont BEIRLANT ET AL [86, 2016]. et REISS ET THOMAS [87, 2007], mais sans résultats asymptotiques. Puis certains estimateurs des paramètres de la queue ont été proposées par BEIRLANT ET GUILLOU [88, 2001], pour les données tronquées et étendu la censure aléatoire à droite par BEIRLANT ET AL [86, 2016] . et l'année suivante par EINMAHL ET AL [89, 2008].

Dans le cas de censure, on suppose disposer de deux échantillons $X_1,...,X_n$ et $Y_1,...,Y_n$ ces deux échantillons sont formés de variable aléatoire i.i.d de loi F et G respectivement et que $F \in \mathcal{D}(\mathcal{H}\xi_1)$ et $G \in \mathcal{D}(\mathcal{H}\xi_2)$ pour certains $\xi_1, \xi_2 \in \mathbb{R}$ Soit $\left\{\left(Z_j, \delta_j\right), 1 \leq j \leq n\right\}$ l'échantillon réellement observé défini par (3.3). Il clair que les Z_j sont des variables indépandents de loi H liée à F et G. L'IVE de H la fdr de Z, existe et il est notée par ξ où $\xi = \frac{\xi_1\xi_2}{\xi_1+\xi_2}$. Soit F, G et r les points terminaux du support de F, G, et h respectivement. EINMAHL ET AL [89, 2008]. ont fourni une adaptation générale des estimateurs existants de L'IVE dans les cas suivants :

$$\begin{cases} cas1 : \xi_1 > 0, \xi_2 > 0, & \text{dans ce cas } \xi = \frac{\xi_1\xi_2}{\xi_1+\xi_2} \\ cas2 : \xi_1 < 0, \xi_2 < 0, x_F = x_G & \text{dans ce cas } \xi = \frac{\xi_1\xi_2}{\xi_1+\xi_2} \\ cas3 : \xi_1 = \xi_2 = 0, x_F = x_G = \infty & \text{dans ce cas } \xi = 0 \end{cases} .$$

Dans le cas 3, nous définissons également, pour une présentation pratique, $\frac{\xi_1\xi_2}{\xi_1+\xi_2} = \xi = 0$.Les autres possibilités ne sont pas très intéressantes. Typiquement, elles sont très proches du " cas non censuré ", qui a été étudiée en détail dans la littérature (elle est notamment valable lorsque $\xi_1 < 0$ et $\xi_2 > 0$) ou de la "situation complétement situation censurée",dans laquelle l'estimation est impossible (ceci est vrai, en particulier, lorsque $\xi_1 > 0$ et $\xi_2 < 0$).

Le premier point important qu'il convient de mentionner est le fait que tous les estimateurs précédents(hill,...)ne sont évidemment pas cohérents s'ils sont basés sur l'échantillon.(Hill,...)ne sont évidemment pas cohérents s'ils sont basés sur l'échantillon $Z_1,...,Z_n$ c'est-à-dire si la censure n'est pas prise en compte. En effet, ils convergent tous vers ξ , l'indice des valeurs extrêmes de l'échantillon Z.l'indice des valeurs extrêmes de l'échantillon Z, et non vers ξ_1, l'indice des valeurs extrêmes de F. Par conséquent, nous devons adapter tous ces estimateurs à la censure. Nous allons diviser tous ces estimateurs par la proportion d'observations non censurées dans les k plus grands

Z :

$$\widehat{\xi}^{(c,\cdot)}_{Z,k,n} = \frac{\widehat{\xi}^{(\cdot)}_{Z,k,n}}{\widehat{p}} \quad où \quad \widehat{p} = \frac{1}{k}\sum_{j=1}^{k} \delta_{[n-j+1,n]},$$

avec $\delta_{[1,n]},..,\delta_{[n,n]}$ étant les δ correspondant à $Z_{1,n},...,Z_{n,n}$ respectivement $\widehat{\xi}^{(\cdot)}_{Z,k,n}$. Pourrait étre tout estimateur non adapté à la censure, en particulier. Il s'ensuit que suit que $\widehat{p}$ estime $\frac{\xi_2}{\xi_1\xi_2}$,d'où $\widehat{\xi}^{(\cdot)}_{Z,k,n}$ estimations ξ divisé par $\frac{\xi_2}{\xi_1\xi_2}$ qui est égal à ξ_1.

3.10.4 Le test statistique du logrank pour comparer les courbes de Kaplan-Meier

Le test statistique quasiment exclusivement utilisé pour comparer des courbes de survie est le test du logrank. Le principe est le suivant : si en vrai, il n'existait aucune différence (réelle) des taux d'incidence entre les groupes comparés, les courbes de survie se chevaucheraient parfaitement. Le test du logrank va donc quantifier l'ensemble des écarts entre les courbes de survie, et va tester si l'ensemble de ces écarts est significativement différent de 0. L'hypothèse nulle H_0 de ce test est donc : " tous les écarts entre toutes les courbes de survie sont nuls " ou bien " toutes les courbes de survie se chevauchent parfaitement ". Le rejet de H_0 est donc : " il existe au moins une courbe de survie significativement différente des autres ".

H_0 : Les deux coubes de survie sont similaires : $S_A(t) = S_B(t)$.

H_1 : Les deux coubes de survie sont différentes : $S_A(t) \neq S_B(t)$.

CHAPITRE 4

SIMULATIONS ET APPLICATIONS À DES DONNÉES RÉELLES

4.1 Introduction

Dans cette partie du mémoire on va étudier le comprotement de l' estimateur de Hill et Pickands , en premier lieu, en fonction de la taille de l'échantillon générée ($n = 1000$) par une loi de Paréto de paramètre de forme égale à 1 et paramètre d'échelle respectivement $(1, 5, 10, 20, 50)$ et une loi censure de Paréto de la taille ($n = 500$) de paramètre de forme égale 1 pour l'échantillon aléatoire de Paréto noté X et de paramètre forme égale 2 pour la variable de censure noté Y.L'estimateur de Hill généralisé adapté aux données censurées à droite est égal à l'estimateur de Hill généralisé ordinaire divisé par la proportion des k plus grandes observations non censurées.

L'analyse de survie est généralement définie comme un ensemble de méthodes d'analyse des données où la variable de résultat est le temps écoulé jusqu'à l'apparition d'un évènement d'intérêt. Cet évènement peut être un décès, l'apparition d'une maladie, un mariage, un divorce, etc. L'analyse de survie est utilisée parce qu'elle est équipée pour traiter les données censurées (condition dans laquelle la valeur d'une mesure ou d'une observation n'est que partiellement connue), ce qui n'est pas le cas d'autres techniques analytiques telles que la régression linéaire. Il existe trois techniques différentes dans l'analyse de survie : les équations non paramétriques(**Kaplan-Meier Plots**), les équations semi-paramétriques (**Cox Proportional Hazard Plots**) et les équations paramétriques (**Kaplan-Meier Plots weibull Distribution**).

4.1.1 Data and Package

Package.evd : Fonction de densité, fonction de distribution, fonction quantile et génération aléatoire pour la distribution des valeurs extrêmes généralisées (GEV) avec

des paramètres de localisation, d'échelle et de forme.

Package.evir : Tracez l'estimation de Hill de l'indice de queue des données à queue lourde, ou d'une estimation quantile associée. **Package.evmix :**Trace le MLE des paramètres GPD en fonction du seuil et Hill plot.

Package.survival : contient les fonctions et arguments nécessaires pour effectuer une analyse de survie en R

lung dataset : mesure la survie des patients atteints d'un cancer du poumon avancé, provenant du North Central Cancer Treatment Group. Les scores de performance évaluent la capacité du patient à effectuer les activités quotidiennes habituelles.

Package.survminer : contient le **Package.ggsurvplot()** pour dessiner facilement de belles courbes de survie

Package.knitr : permet à l'utilisateur de produire un tableau (un tableau très simple).

Package.flexsurv : permet une modélisation paramétrique de la survie.

4.2 Simulation et estimation

Dans la suite on simule le loi Paréto et on applique l'estimateur de Hill avec les données simulée .

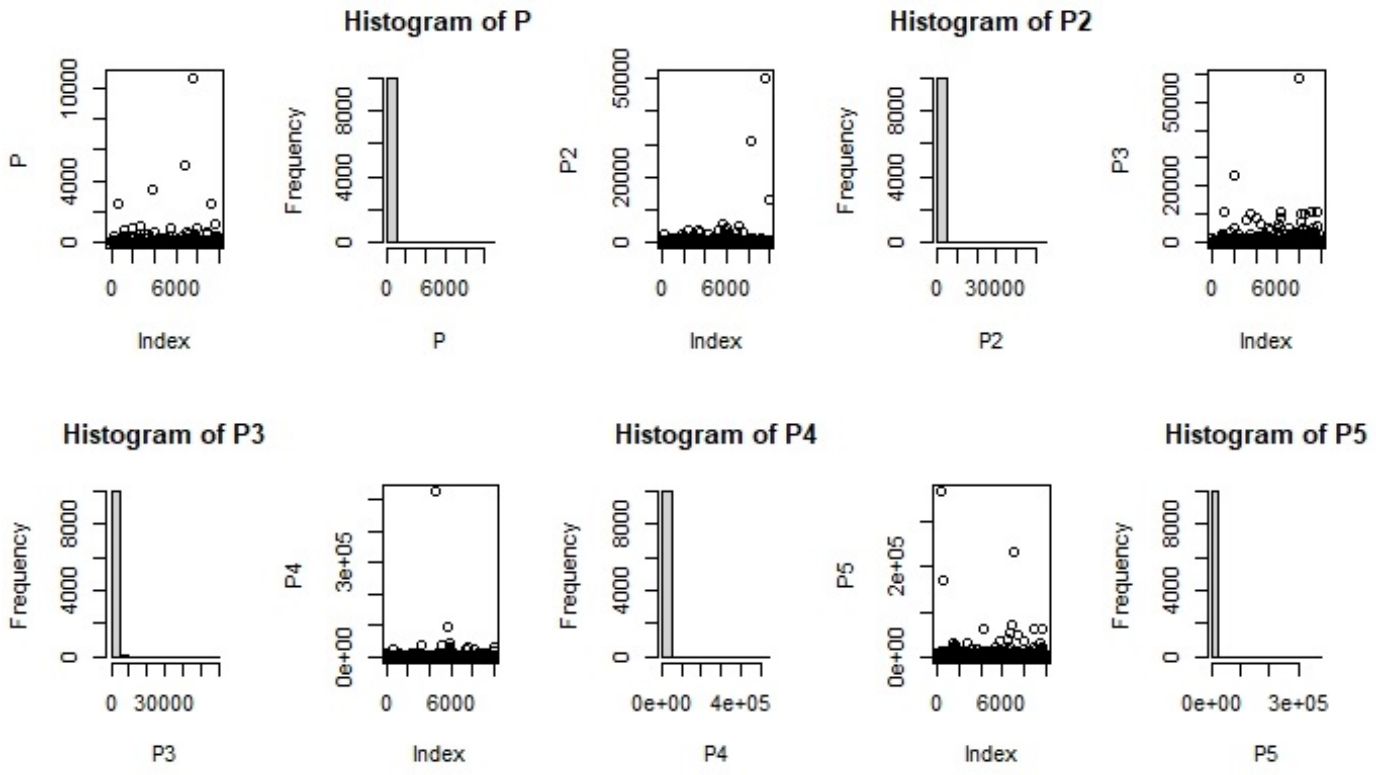

FIGURE 4.1 – Histograme et Ghraphe de loi de Paréto a plusieure paramèter

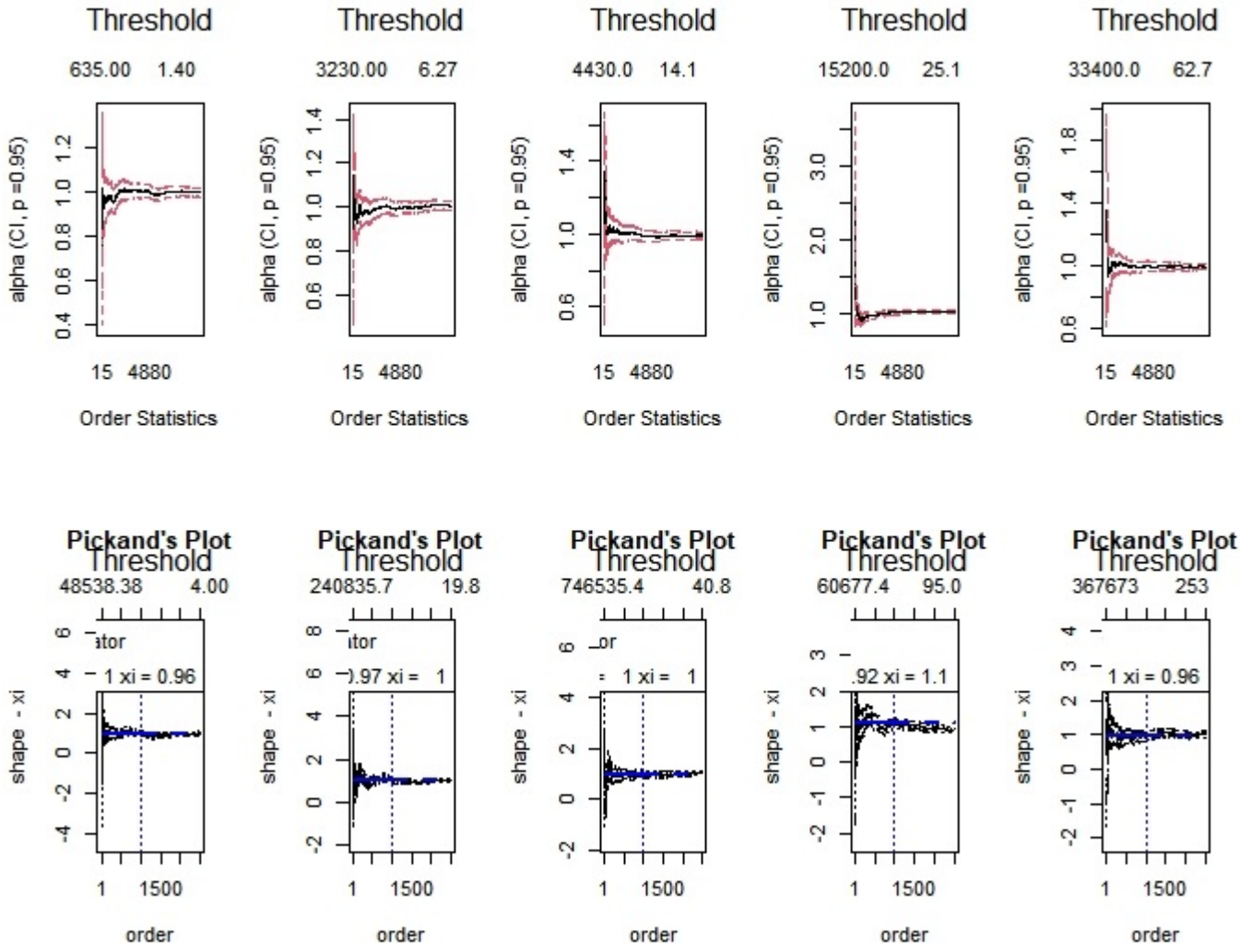

FIGURE 4.2 – L'estimateur de Hill et Pickands avec les données simulées.

Calcule l'estimateur de Hill pour les indices de valeurs extrêmes positives, adapté à la censure de droite, en fonction du paramètre de queue k (Beirlant et al.[]). En option, ces estimations sont tracées en fonction de k.

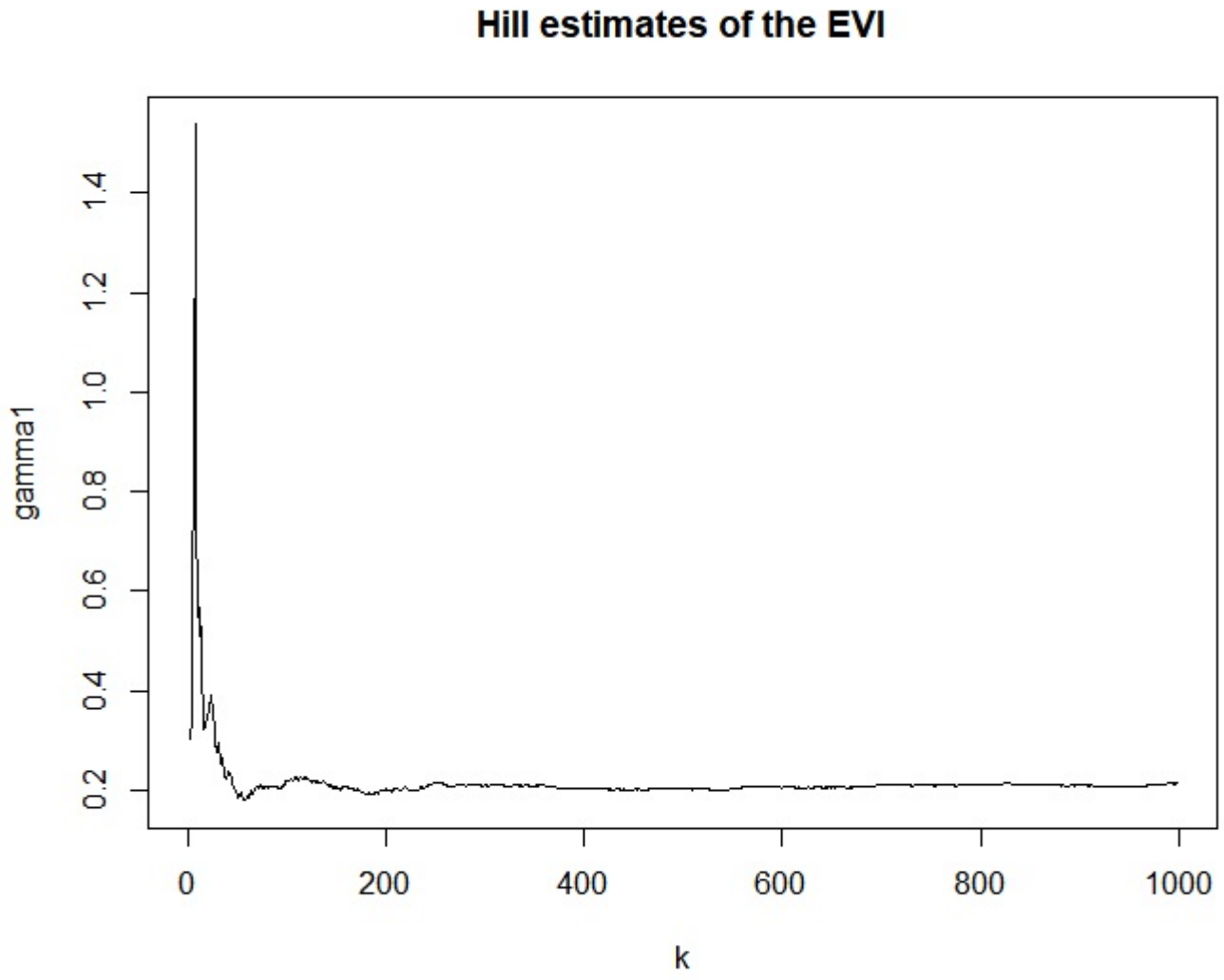

FIGURE 4.3 – L'estimateur de Hill de l'indice queue d' une loi de Paréto $(1000, 5)$ avec taux de censure $T_c = 50\%$ hill censure 0.5 n=1000

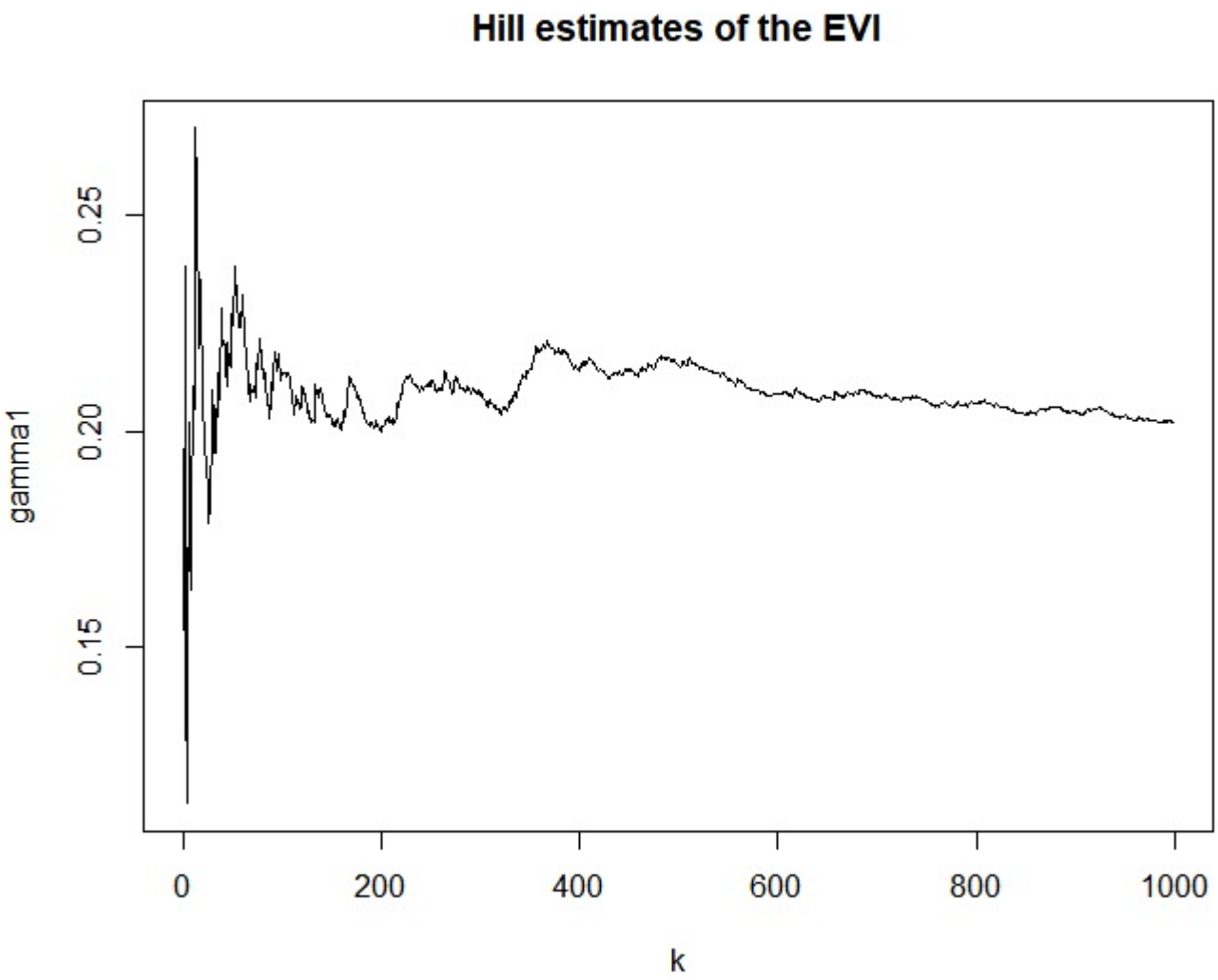

FIGURE 4.4 – L'estimateur de Hill de l'indice queue d' une loi de Paréto $(1000, 5)$ avec taux de censure $T_c = 10\%$

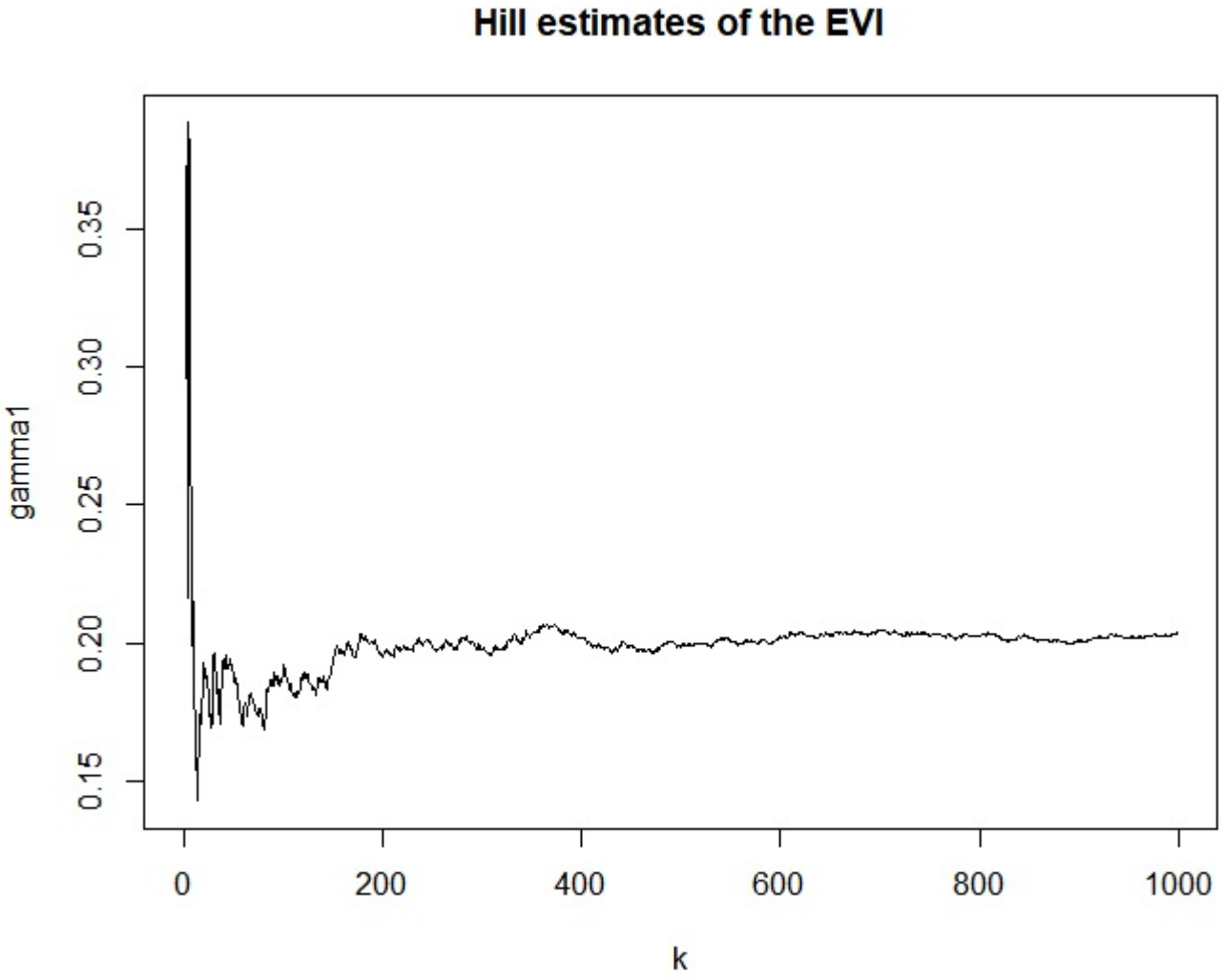

FIGURE 4.5 – L'estimateur de Hill de l'indice queue d' une loi de Paréto $(1000, 5)$ avec taux de censure $T_c = 30\%$

4.3 Applications à des données réelles (cancer du poumon avancé)

On mesure la survie des patients atteints d'un cancer du poumon avancé, selon le groupe de traitement du cancer du Centre-Nord.

4.3.1 Kaplan-Meier courbe de survie

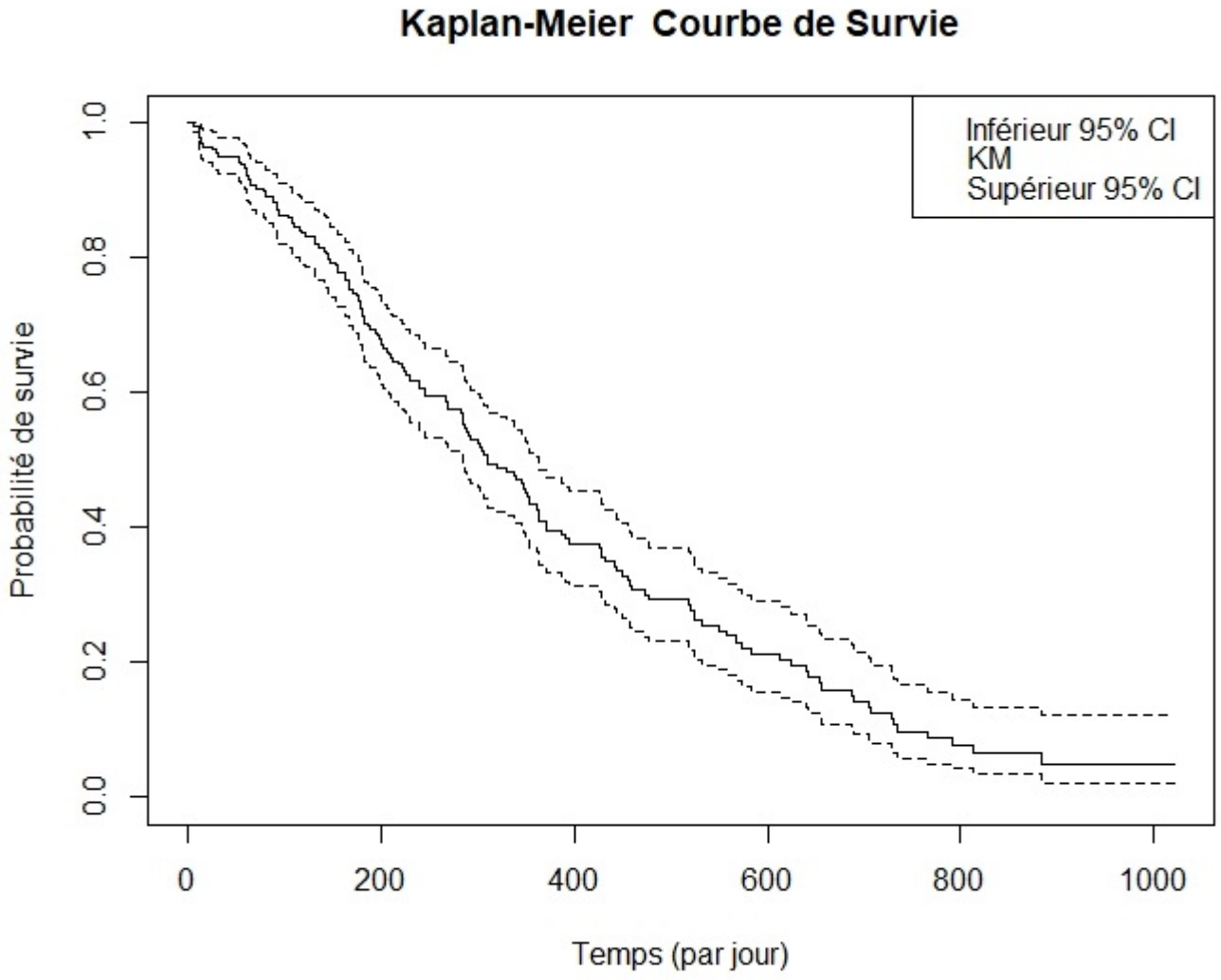

FIGURE 4.6 – Kaplan-Meier courbe de survie

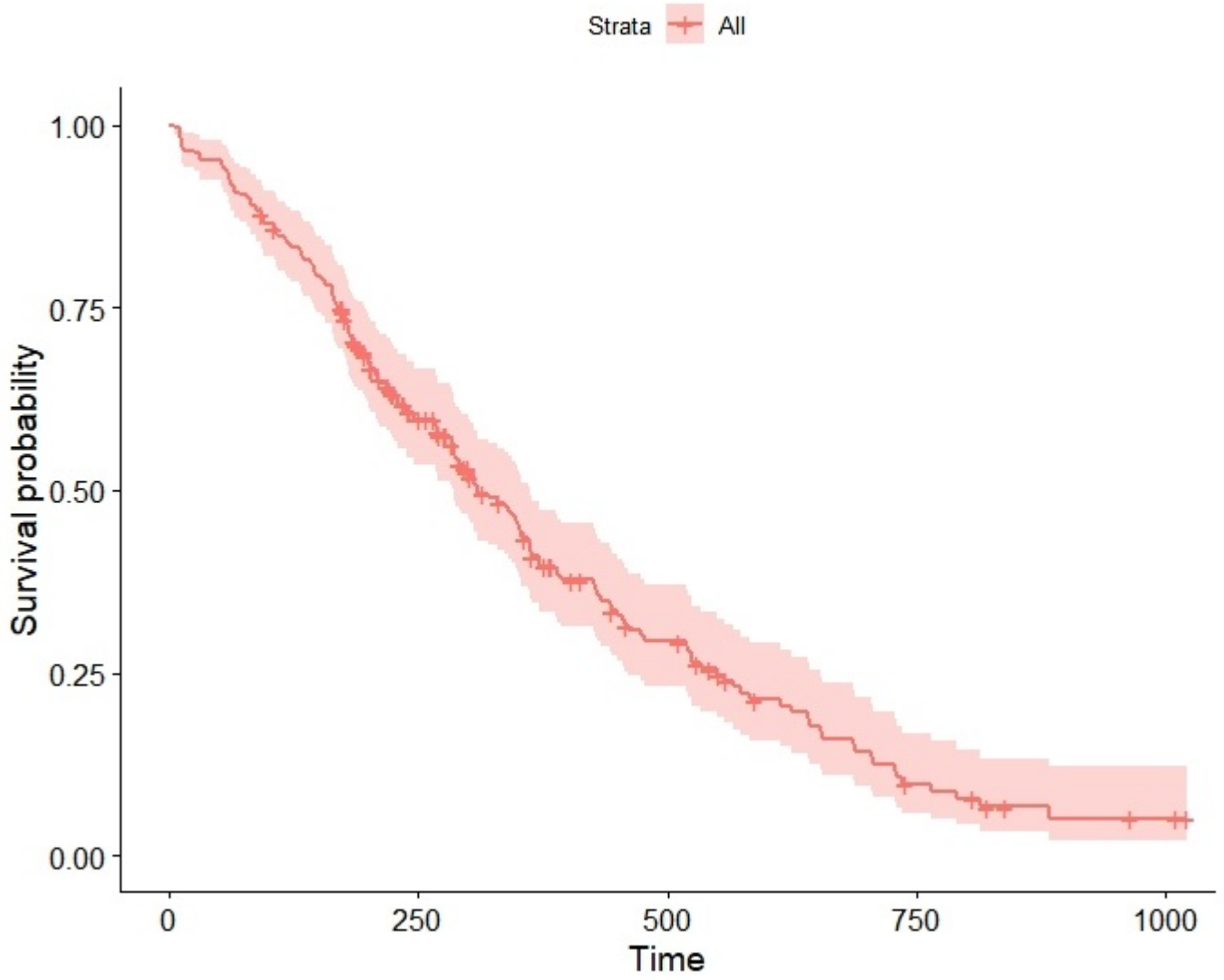

FIGURE 4.7 – Kaplan-Meier courbe de survie avec ggsurvplot.

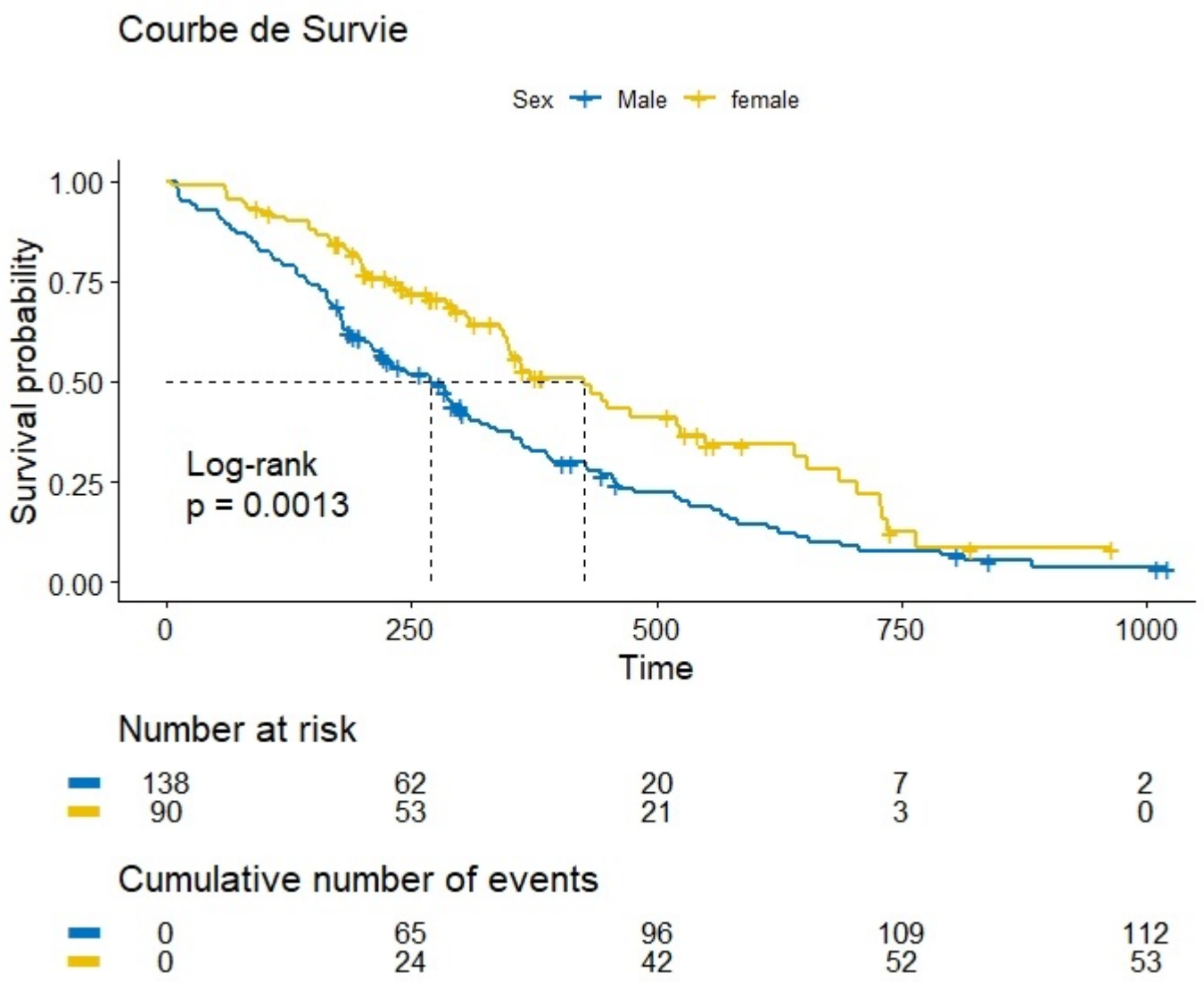

FIGURE 4.8 – Il s'agit du même graphique KM que ci-dessus, mais le graphique ci-dessous montre l'intérêt de l'utilisation du package survminer pour ce type d'analyses.

Comme vous pouvez le constater, il existe une différence radicale entre le tracé X-Y de base et le ggsurvplot. Le tracé de Kaplan-Meier montre qu'aux alentours de 250 , la probabilité de survie est de 55 %, de 25 % à 500 et qu'elle continue de diminuer à partir de là. Le graphique montre qu'au cours de l'essai, environ 10 % des patients ont vécu jusqu'au bout et qu'aucun évènement n'est survenu. En regardant la valeur p, nous voyons qu'elle est inférieure à 0.05 , ce qui signifie que nous rejetons l'hypothèse nulle selon laquelle la différence de sexe n'a pas d'importance. Cela signifie qu'il existe une différence de survie entre les hommes et les femmes, et que les femmes ont un meilleur taux de survie sur l'ensemble de l'essai.

4.4 Ajustement à des données par modèle de Cox

Les modèles semi-paramétriques ne font pas non plus de prédictions sur la forme de la fonction de risque ou de la fonction de survie, mais font seulement une hypothèse forte sur la façon dont les covariables affectent la forme. Lorsqu'on utilise le modèle de Cox pour l'analyse de survie, il existe deux hypothèses clés qui doivent être valides pour que les résultats du modèle puissent être appliqués en toute sécurité aux données de l'analyse de survie. La première hypothèse concerne la question de la censure non informative ; la conception de l'étude sous-jacente doit garantir que les mécanismes donnant lieu à la censure de sujets individuels ne sont pas liés à la probabilité de survenue d'un évènement. Par exemple, la poursuite des suivis doit se poursuivre même si un patient ne présente plus de statut positif. Dans notre cas, nous n'avons pas à nous en préoccuper car les données ont été conduites de cette manière.

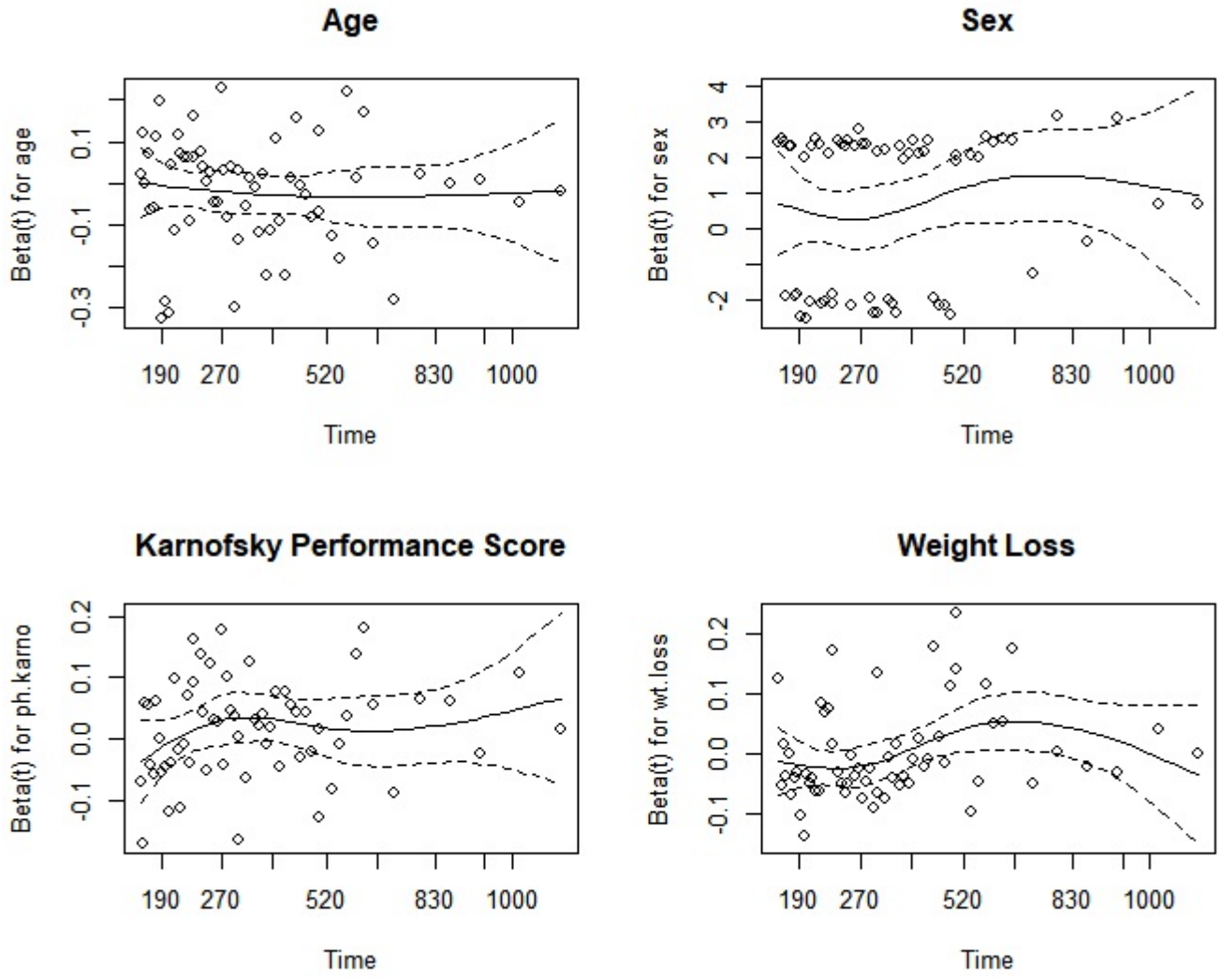

FIGURE 4.9 – Cox graphique

4.4.1 Tracer avec (ggcoxadjustedcurves)

Maintenant que nous savons à quoi ressemble un tracé X, Y normal pour l'analyse de survie, nous allons passer à la méthode la plus facile et la plus agréable visuellement. (ggcoxadjustedcurves) trace des courbes de survie ajustées pour le modèle Coxph.

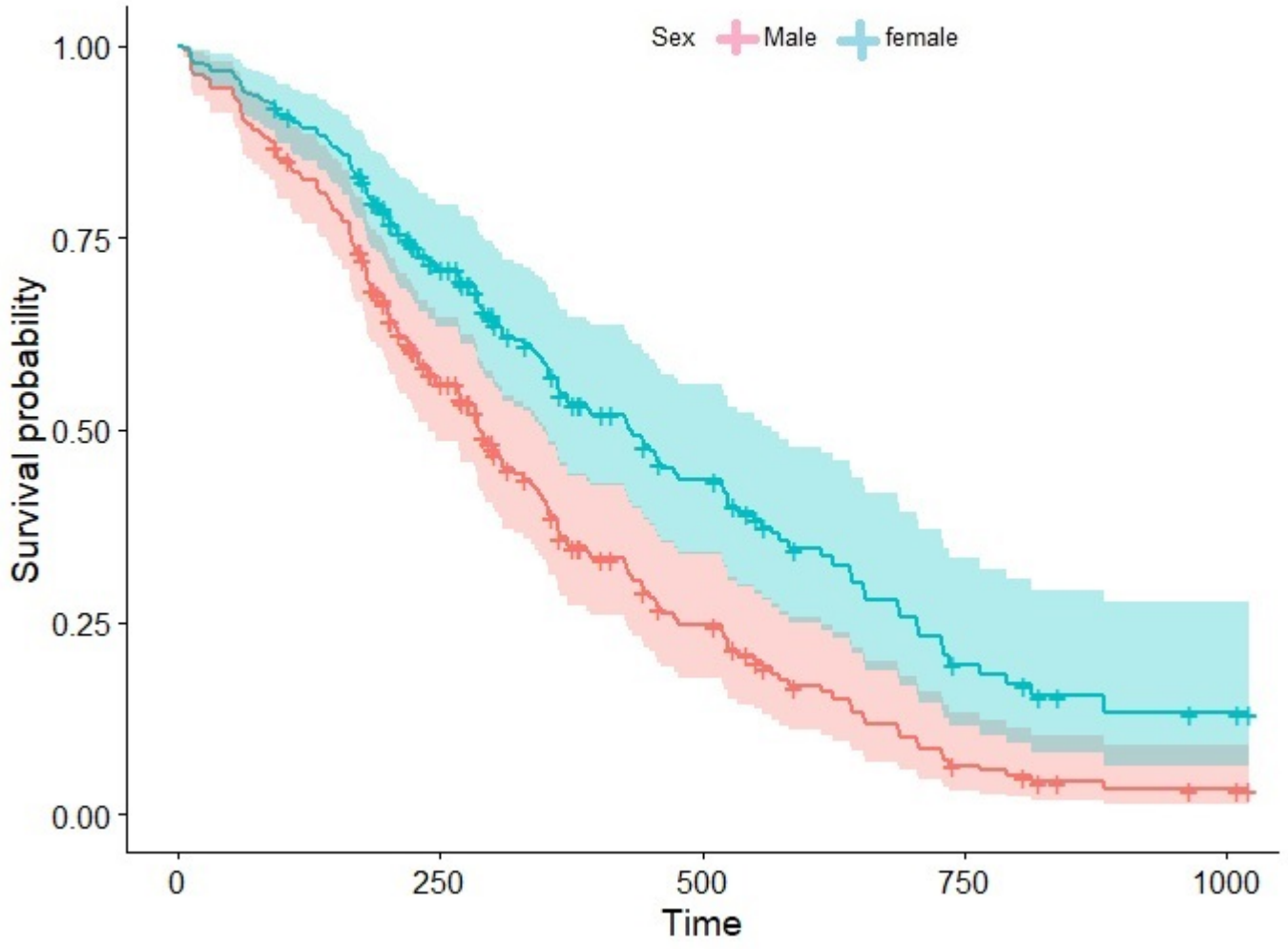

FIGURE 4.10 – Probabilité de survie de modèle de Cox

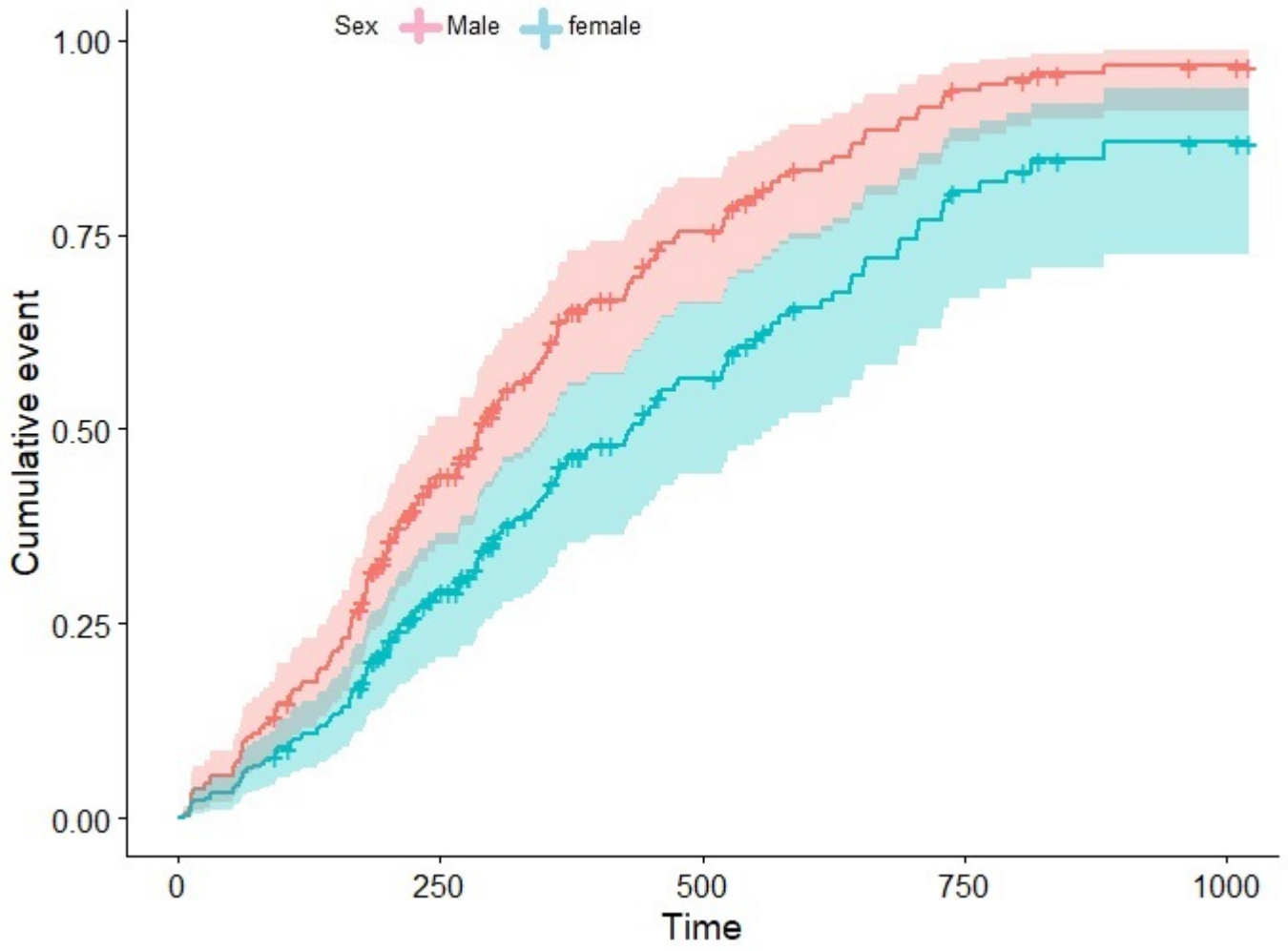

FIGURE 4.11 – Les évènments cumulés où modèle de Cox cumulative

4.5 Ajustement à des données par modèle de Weibull

TABLE 4.1 – Ajustement à des données pour modèle Weibull

Distribution	AIC
weibull	2311.702
gengamma.orig	2313.380
gengamma	2313.380
gamma	2313.469
gompertz	2314.711
genf	2315.153
genf.orig	2315.153
llogis	2325.862
exp	2326.676
lnorm	2342.538

Lorsque l'on compare l'AIC des différentes distributions pouvant être utilisées dans R, la distribution de Weibull présente l'AIC le plus faible de 2312, ce qui signifie qu'il s'agit de la meilleure distribution pour ajuster les données. En observant la distribution de Weibull graphiquement, nous obtenons.

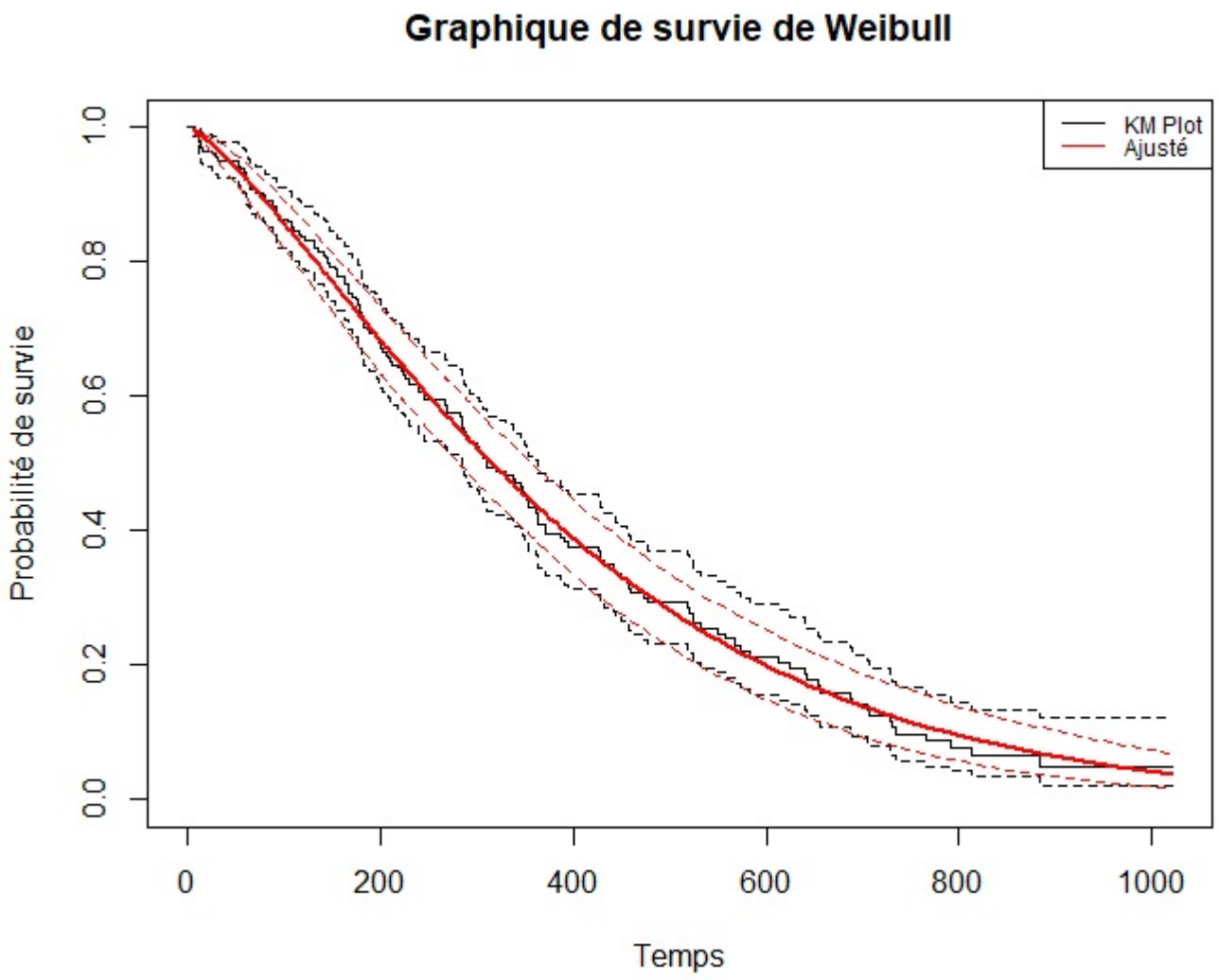

FIGURE 4.12 – Ajustement à des données pour modèle Weibull

Sur la base du graphique, nous pouvons voir la différence de précision entre le tracé de Kaplan-Meier et le tracé ajusté de Gompertz, en particulier avec la différence autour de la marque des 375 jours. Le tracé KM et le tracé ajusté de Weibull sont très similaires en raison de la petite taille de l'ensemble de données, mais le modèle paramétrique permet des estimations plus précises des paramètres et une modélisation plus prédictive. Comme vous pouvez le constater, avec un grand ensemble de données, il serait beaucoup plus facile de tracer une distribution et de la suivre qu'une ligne de Kaplan-Meier.

4.6 Simulation et estimations à des données réelles (Hill et Pickans)

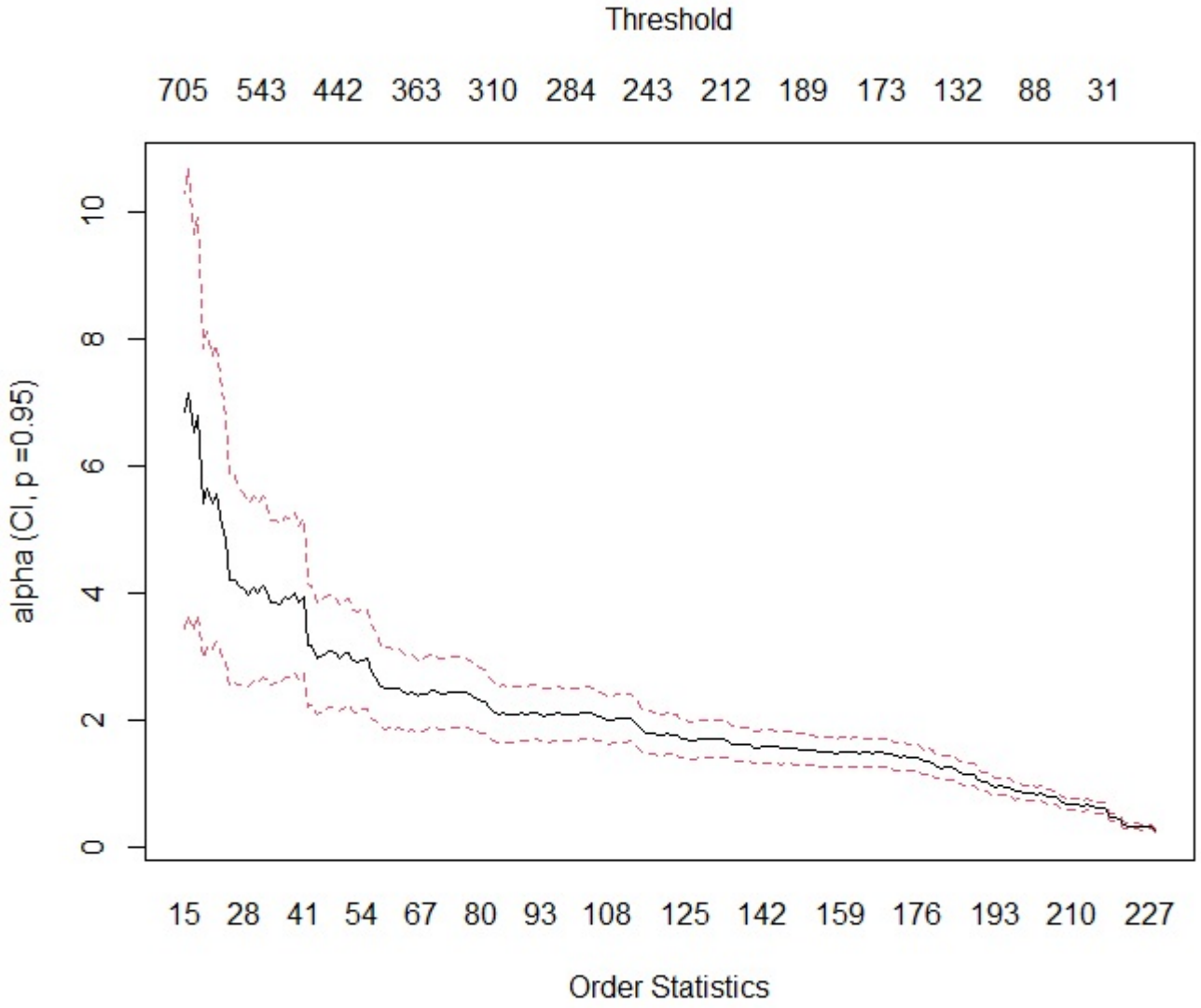

FIGURE 4.13 – L' estimateur de Hill (Threshold) par package.evir .

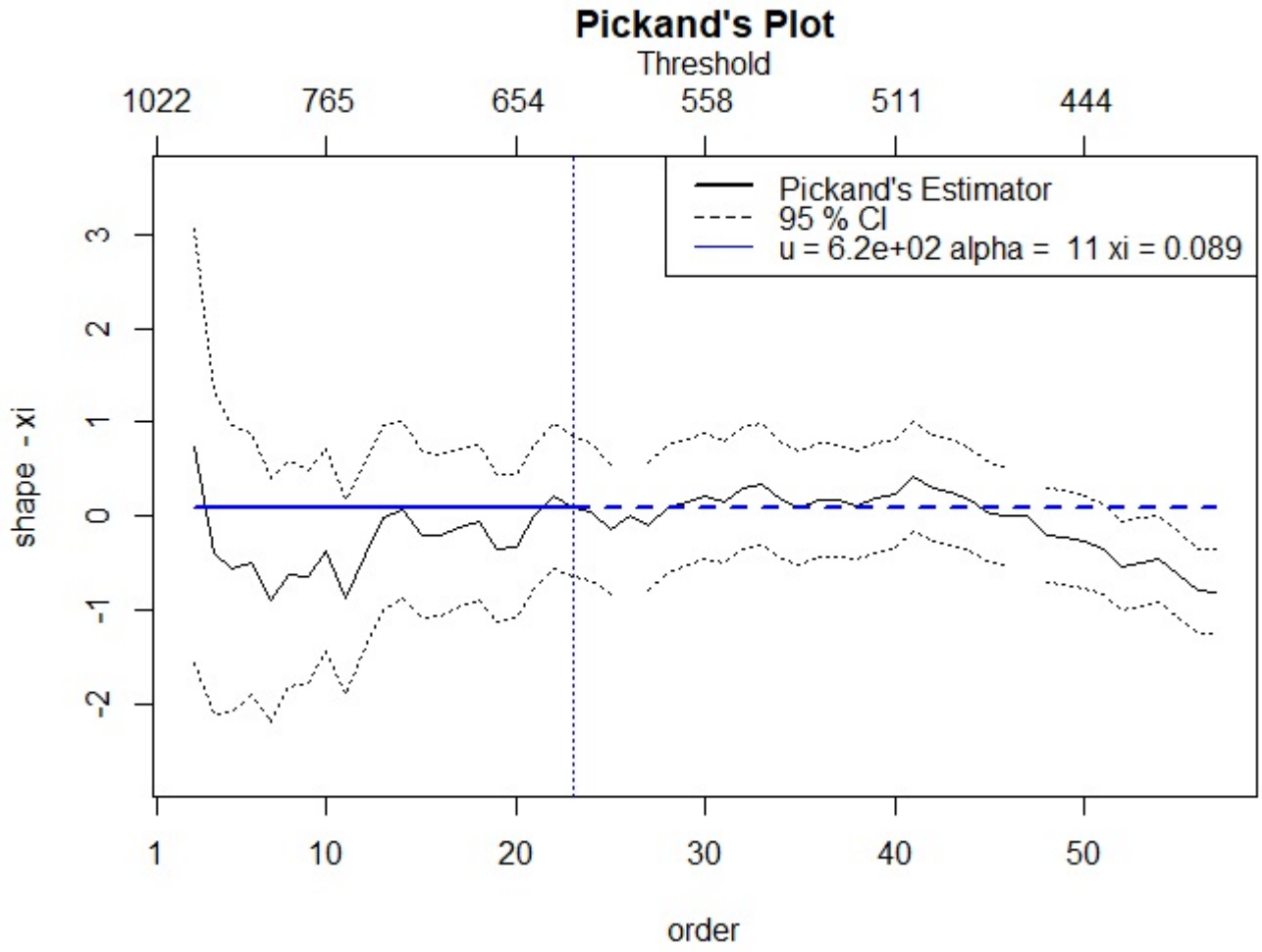

FIGURE 4.14 – L'estimateur de Pickands par package.evir

CONCLUSION GÉNÉRALE

L'objectif principal de ce mémoire est présenté l'analyse de Survie dans le cadre extrême . nous nous intéressons à une famille particulèire de lois : les lois à queue de type Weibull . Ces lois possèdent une fonction de survie qui décroit à une vitesse exponentielle (on parle aussi de *queue légère*).Des exemples de telles lois sont les lois exponentielle, normale, gamma, ect ... La vitesse de convergence de la queue de distribution est controlée par un paramètre de forme appelé indice de queue de Weibull. Nous introduisons et étudions le comportement asymptotique d'estimateurs de cet indice et des quantiles extrêmes. Nous présentons aussi une méthode de réduction du biais basée sur un modèle de régression exponentielle.

Dans l'analyse de survie, il est très commun de se trouver en face du problème de données manquantes. Les données de survie ne sont pas totalement observées. Il ne sont pas rare, mais elles sont plutôt incomplètes. La censure et la troncature sont les deux causes de données incomplètes les plus répandues. La censure est un mécanisme qui empêche l'observation exacte du délai de survenue d'intérêt. On sait bien que ce délai appartient à un certain intervalle de temps. La troncature survient qu'on ne peut pas observer les individus de l'échantillon dont le délai de survenue appartient à un certain intervalle de temps, on observe donc un sous-échantillon. Dans ce cas les techniques classiques ne s'adaptent pas correctement aux données incomplètes.

Et tout ce travail Pour l'analyse de survie lié à deux approches la théorie des valeurs extrême . La première, qu'on appellera approche GEV ; permet de modéliser les block maxima par une distribution GEV (generalized extreme value distribution) et la seconde, appelée approche GPD consiste à ajuster les observations dépassant un certain seuil (peaks over threshold : POT) par une GPD (generalized Pareto distribution). Pour une description détaillée de la TVE, en particulier sur l'estimation de l'indice des valeurs et quantiles extrêmes.

Sur le plan pratique, nous avons utilisé les données de survie des patients atteints d'un cancer du poumon avancé du North Central Cancer Treatment Group. Les scores

de performance évaluent dans quelle mesure un patient effectue des activités quotidiennes normales.En utilisons des packages spécifiques et Nous utilisons des modèles de fiabilité (Cox model proportional hazard ... ect) et des estimations connues (Hill, Pickands, Kaplan-Meier,...ect) est tous ça dans le cas censure à droite. Cancer du poumon donne un exemple de la relation entre une fonction de survie qui décroit à une vitesse exponentielle (on parle aussi de queue légère), c.est-à-dire la relation entre les lois à queue de type Weibull et L'analyse de survie en cas censure.

BIBLIOGRAPHIE

[1] EMBRECHTS, Paul ; KLÜPPELBERG, Claudia ; MIKOSCH, Thomas. Modelling extremal events : for insurance and finance. Springer Science & Business Media, 2013.

[2] FISHER, Ronald Aylmer ; TIPPETT, Leonard Henry Caleb. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In : Mathematical proceedings of the Cambridge philosophical society. Cambridge University Press, 1928. p. 180-190.

[3] GNEDENKO, Boris. Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of mathematics, 1943, 423-453.

[4] RESNICK, Sidney I. Extreme values, regular variation and point processes. Springer, 2013.

[5] GALAMBOS, Janos. The asymptotic theory of extreme order statistics. 1978.

[6] COLES, Stuart, et al. An introduction to statistical modeling of extreme values. London : Springer, 2001.

[7] BINGHAM, Nicholas H., et al. Regular variation. Cambridge university press, 1989.

[8] DELMAS, Jean-Franois ; JOURDAIN, Benjamin. Modèles Aléatoires. Springer-Verlag Berlin Heidelberg, 2006

[9] VON MISES, Richard. La distribution de la plus grande de n valuers. Rev. math. Union interbalcanique, 1936, 1 : 141-160.

[10] GNEDENKO, Boris. Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of mathematics, 1943, 423-453.

[11] GUMBEL, Emil Julius ; MUSTAFI, Chandan K. Some analytical properties of bivariate extremal distributions. Journal of the American Statistical Association, 1967, 62.318 : 569-588.

[12] SMITH, Richard L., et al. Estimating tails of probability distributions. The annals of Statistics, 1987, 15.3 : 1174-1207.

[13] PRESCOTT, P. ; WALDEN, A. T. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. Biometrika, 1980, 67.3 : 723-724.

[14] CHRISTOPEIT, Norbert. Estimating parameters of an extreme value distribution by the method of moments. Journal of Statistical Planning and inference, 1994, vol. 41, no 2, p. 173-186.

[15] GREENWOOD, J. Arthur, LANDWEHR, J. Maciunas, MATALAS, Nicolas C., et al. Probability weighted moments : definition and relation to parameters of several distributions expressable in inverse form. Water resources research, 1979, vol. 15, no 5, p. 1049-1054

[16] BENKHALED, A. Statistical distributions of annual maximum rainfall in the Cheliff region, comparison of techniques and results. Courrier du Savoir, 2007, vol. 8, p. 83-91.

[17] PICKANDS III, James, et al. Statistical inference using extreme order statistics. Annals of statistics, 1975, vol. 3, no 1, p. 119-131.

[18] HILL, Bruce M. A simple general approach to inference about the tail of a distribution. The annals of statistics, 1975, p. 1163-1174.

[19] DEKKERS, Arnold LM, EINMAHL, John HJ, et DE HAAN, Laurens. A moment estimator for the index of an extreme-value distribution. The Annals of Statistics, 1989, p. 1833-1855.

[20] JUSTUS, C. G., HARGRAVES, W. R., MIKHAIL, Amir, et al. Methods for estimating wind speed frequency distributions. Journal of applied meteorology, 1978, vol. 17, no 3, p. 350-353.

[21] AZAD, Abul Kalam, RASUL, Mohammad Golam, et YUSAF, Talal. Statistical diagnosis of the best weibull methods for wind power assessment for agricultural applications. Energies, 2014, vol. 7, no 5, p. 3056-3085.

[22] KATZ, Richard W. Techniques for estimating uncertainty in climate change scenarios and impact studies. Climate research, 2002, vol. 20, no 2, p. 167-185.

[23] NECIR, Abdelhakim, RASSOUL, Abdelaziz, et ZITIKIS, Riďcardas. Estimating the conditional tail expectation in the case of heavy-tailed losses. Journal of Probability and Statistics, 2010, vol. 2010.

[24] TODOROVIC, P. et ROUSSELLE, J. Some problems of flood analysis. Water Resources Research, 1971, vol. 7, no 5, p. 1144-1150.

[25] TODOROVIC, P. et ZELENHASIC, E. A stochastic model for flood analysis. Water Resources Research, 1970, vol. 6, no 6, p. 1641-1648.

[26] DAVISON, Anthony C. et SMITH, Richard L. Models for exceedances over high thresholds. Journal of the Royal Statistical Society : Series B (Methodological), 1990, vol. 52, no 3, p. 393-425.

[27] REISS, R. D. et THOMAS, M. Statistical Analysis of Extreme Values. Birkhäuser-Verlag. Basel, Switzerland, 2001.

[28] BALKEMA, August A. et DE HAAN, Laurens. Residual life time at great age. The Annals of probability, 1974, p. 792-804.

[29] HOSKING, Jonathan RM et WALLIS, James R. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 1987, vol. 29, no 3, p. 339-349.

[30] BEIRLANT, Jan et GOEGEBEUR, Yuri. Regression with response distributions of Pareto-type. Computational statistics & data analysis, 2003, vol. 42, no 4, p. 595-619.

[31] HOSKING, Jonathan Richard Morley, WALLIS, James R., et WOOD, Eric F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics, 1985, vol. 27, no 3, p. 251-261.

[32] DIEBOLT, Jean, GUILLOU, Armelle, et RACHED, Imen. A new look at probability-weighted moments estimators. Comptes Rendus Mathematique, 2004, vol. 338, no 8, p. 629-634.

[33] EFRON, B. Bootstrap methods : another look at the jackknife. e Annals of Statistics, 7 (1) : 1–26. URL http ://www. jstor. org/stable/2958830, 1979.+

[34] DITLEVSEN, Ove. Distribution arbitrariness in structural reliability. Structural Safety and Reliability, 1994, p. 1241-1247.

[35] BEIRLANT, Jan et TEUGELS, Jozef L. Modeling large claims in non-life insurance. Insurance : Mathematics and Economics, 1992, vol. 11, no 1, p. 17-29.

[36] BRODIN, Erik et ROOTZEN, Holger. Univariate and bivariate GPD methods for predicting extreme wind storm losses. Insurance : Mathematics and Economics, 2009, vol. 44, no 3, p. 345-356.

[37] ROOTZEN, Holger et TAJVIDI, Nader. Can losses caused by wind storms be predicted from meteorological observations ?. Scandinavian Actuarial Journal, 2001, vol. 2001, no 2, p. 162-175.

[38] BERRED, M. Record values and the estimation of the Weibull tail-coefficient. Comptes rendus de l'Académie des sciences. Série 1, Mathématique, 1991, vol. 312, no 12, p. 943-946.

[39] BEIRLANT, Jan, BOUQUIAUX, Christel, et WERKER, Bas JM. Semiparametric lower bounds for tail index estimation. Journal of Statistical Planning and Inference, 2006, vol. 136, no 3, p. 705-729.

[40] BEIRLANT, Jan, BRONIATOWSKI, Michel, TEUGELS, Jozef L., et al. The mean residual life function at great age : Applications to tail estimation. Journal of Statistical Planning and Inference, 1995, vol. 45, no 1-2, p. 21-48.

[41] BEIRLANT, Jan, TEUGELS, Jozef L., et VYNCKIER, Petra. Practical analysis of extreme values. Leuven : Leuven University Press, 1996.

[42] BRONIATOWSKI, Michel. On the estimation of the Weibull tail coefficient. Journal of Statistical Planning and Inference, 1993, vol. 35, no 3, p. 349-365.

[43] DIEBOLT, Jean, GARDES, Laurent, GIRARD, Stéphane, et al. Bias-reduced estimators of the Weibull tail-coefficient. Test, 2008, vol. 17, no 2, p. 311-331.

[44] DIERCKX, Goedele, BEIRLANT, Jan, DE WAAL, Dan, et al. A new estimation method for Weibull-type tails based on the mean excess function. Journal of Statistical Planning and Inference, 2009, vol. 139, no 6, p. 1905-1920.

[45] GARDES, Laurent et GIRARD, Stéphane. Comparison of Weibull tail-coefficient estimators. arXiv preprint arXiv :1104.0764, 2011.

[46] GARDES, Laurent et GIRARD, Stephane. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. Journal of Statistical Planning and Inference, 2008, vol. 138, no 5, p. 1416-1427.

[47] GIRARD, Stéphane. A Hill type estimator of the Weibull tail-coefficient. Communications in Statistics-Theory and Methods, 2004, vol. 33, no 2, p. 205-234.

[48] GOEGEBEUR, Yuri, BEIRLANT, Jan, et DE WET, Tertius. Generalized kernel estimators for the Weibull-tail coefficient. Communications in Statistics—Theory and Methods, 2010, vol. 39, no 20, p. 3695-3716.

[49] GOEGEBEUR, Yuri et GUILLOU, Armelle. Goodness-of-fit testing for Weibull-type behavior. Journal of Statistical Planning and Inference, 2010, vol. 140, no 6, p. 1417-1436.

[50] BEIRLANT, Jan, DIERCKX, Goedele, GOEGEBEUR, Yuri, et al. Tail index estimation and an exponential regression model. Extremes, 1999, vol. 2, no 2, p. 177-200.

[51] BEIRLANT, Jan, DIERCKX, Goedele, GUILLOU, A., et al. On exponential representations of log-spacings of extreme order statistics. Extremes, 2002, vol. 5, no 2, p. 157-180.

[52] FEUERVERGER, Andrey, HALL, Peter, et al. Estimating a tail exponent by modelling departure from a Pareto distribution. The Annals of Statistics, 1999, vol. 27, no 2, p. 760-781.

[53] ASIMIT, Alexandru V., LI, Deyuan, et PENG, Liang. Pitfalls in using Weibull tailed distributions. Journal of Statistical Planning and Inference, 2010, vol. 140, no 7, p. 2018-2024.

[54] MERCADIER, Cécile et SOULIER, Philippe. Optimal rates of convergence in the Weibull model based on kernel-type estimators. Statistics & Probability Letters, 2012, vol. 82, no 3, p. 548-556.

[55] LEPSKI, Oleg V., MAMMEN, Enno, et SPOKOINY, Vladimir G. Optimal spatial adaptation to inhomogeneous smoothness : an approach based on kernel estimates with variable bandwidth selectors. The Annals of Statistics, 1997, p. 929-947.

[56] DREES, Holger et KAUFMANN, Edgar. Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Processes and their applications, 1998, vol. 75, no 2, p. 149-172.

[57] WEISSMAN, Ishay. Estimation of parameters and large quantiles based on the k largest observations. Journal of the American Statistical Association, 1978, vol. 73, no 364, p. 812-815.

[58] GARDES, Laurent et GIRARD, Stéphane. Estimating extreme quantiles of Weibull tail distributions. Communications in Statistics—Theory and Methods, 2005, vol. 34, no 5, p. 1065-1080.

[59] DIEBOLT, Jean, GARDES, Laurent, GIRARD, Stephane, et al. Bias-reduced extreme quantile estimators of Weibull tail-distributions. Journal of Statistical Planning and Inference, 2008, vol. 138, no 5, p. 1389-1401.

[60] GOMESA, M. Ivette et MARTINS, M. João. "Asymptotically unbiased" estimators of the tail index based on external estimation of the second order parameter. Extremes, 2002, vol. 5, no 1, p. 5-31.

[61] GOMES, M. Ivette, MARTINS, M. Joao, et NEVES, Manuela. Improving second order reduced bias extreme value index estimation. Revstat, 2007, vol. 5, no 2, p. 177-207.

[62] PENG, Liang et QI, Yongcheng. Estimating the First-and Second-Order Parameters of a Heavy-Tailed Distribution. Australian & New Zealand Journal of Statistics, 2004, vol. 46, no 2, p. 305-312.

[63] COX, David Roxbee et OAKES, David. Analysis of survival data. Chapman and Hall/CRC, 2018.

[64] MILLS, John P. Table of the ratio : area to bounding ordinate, for any portion of normal curve. Biometrika, 1926, p. 395-400.

[65] NELSON, Wayne B. Accelerated testing : statistical models, test plans, and data analysis. John Wiley & Sons, 2009.

[66] MORTUREUX, Yves. La sûreté de fonctionnement : méthodes pour maîtriser les risques. 2001.

[67] GIRAUD, Marc. Sûreté de fonctionnement des sysèmes-Croissance de fiabilité et management. 2007.

[68] SAULL, John W. Practical Reliability Engineering–Fourth edition PDT O'Connor et al John Wiley and Sons, Baffins Lane, Chichester, West Sussex P019 1UD, UK 2002. 513pp. Illustrated.£ 29.95. ISBN 0-470-84463-9. The Aeronautical Journal, 2003, vol. 107, no 1067, p. 63-63.

[69] LAWLESS, Jerald F. Statistical models and methods for lifetime data. John Wiley & Sons, 2011.

[70] LEE, Elisa T. et WANG, John. Statistical methods for survival data analysis. John Wiley & Sons, 2003.

[71] CSORGO, Sandor, DEHEUVELS, Paul, et MASON, David. Kernel estimates of the tail index of a distribution. The Annals of Statistics, 1985, p. 1050-1077.

[72] FALK, Michael et MAROHN, Frank. Efficient estimation of the shape parameter in Pareto models with partially known scale. Statistics & Risk Modeling, 1997, vol. 15, no 3, p. 229-240.

[73] WU, Jingshu. Survival analysis of real-world tire aging data. National Highway Traffic Safety Administration. United States : National centre for Statistics and Analysis, 2007.

[74] ZHOU, Dan, LI, C., et al. Comparison of parameter estimation methods for transformer Weibull lifetime modelling. Gaodianya Jishu/High Voltage Engineering, 2013, vol. 39, no 5, p. 1170-1177.

[75] RATNAPARKHI, Makarand V. et PARK, Won J. Lognormal distribution-model for fatigue life and residual strength of composite materials. IEEE transactions on reliability, 1986, vol. 35, no 3, p. 312-315.

[76] MULLEN, Robert E. The lognormal distribution of software failure rates : origin and evidence. In : Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No. 98TB100257). IEEE, 1998. p. 124-133.

[77] PLANCHET, Frédéric, THEROND, Pierre-E., et al. Optimal strategies for hedging portfolios of unit-linked life insurance contracts with minimum death guarantee. Insurance : Mathematics and Economics, 2011, vol. 48, no 2, p. 161-175.

[78] COLLETT, David. Modelling survival data in medical research. CRC press, 2015.

[79] OAKES, David. Bivariate survival models induced by frailties. Journal of the American Statistical Association, 1989, vol. 84, no 406, p. 487-493.

[80] COX, David R. Regression models and life-tables. Journal of the Royal Statistical Society : Series B (Methodological), 1972, vol. 34, no 2, p. 187-202.

[81] MARSHALL, Albert W. et OLKIN, Ingram. A generalized bivariate exponential distribution. Journal of Applied Probability, 1967, p. 291-302.

[82] KAPLAN, Edward L. et MEIER, Paul. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 1958, vol. 53, no 282, p. 457-481.

[83] SAINT PIERRE, Philippe. Introduction à l'analyse des durées de survie. Université Pierre et Marie Curie, France, 2015.

[84] VIALLON, Vivian. Processus empiriques, estimation non paramétrique et données censurées. 2006. Thèse de doctorat. Université Pierre et Marie Curie-Paris VI.

[85] SOLTANE, Louiza. Analyse des Valeurs Extrêmes en présence de censure. 2017. Thèse de doctorat. Université Mohamed Khider-Biskra.

[86] BEIRLANT, Jan, BARDOUTSOS, Anastasios, DE WET, T., et al. Bias reduced tail estimation for censored Pareto type distributions. Statistics & Probability Letters, 2016, vol. 109, p. 78-88.

[87] REISS, Rolf-Dieter et THOMAS, Michael. Flood Frequency Analysis. Statistical Analysis of Extreme Values : with Applications to Insurance, Finance, Hydrology and Other Fields, 2007, p. 337-351.

[88] BEIRLANT, Jan et GUILLOU, Armelle. Pareto index estimation under moderate right censoring. Scandinavian Actuarial Journal, 2001, vol. 2001, no 2, p. 111-125.

[89] EINMAHL, John HJ, FILS-VILLETARD, Amélie, GUILLOU, Armelle, et al. Statistics of extremes under random censoring. Bernoulli, 2008, vol. 14, no 1, p. 207-227.

[90] BEIRLANT, Jan, GUILLOU, Armelle, DIERCKX, Goedele, et al. Estimation of the extreme value index and extreme quantiles under random censoring. Extremes, 2007, vol. 10, no 3, p. 151-174.

[91] KALBFLEISCH, John D. et PRENTICE, Ross L. The statistical analysis of failure time data. John Wiley & Sons, 2011.

[92] KLEIN, John P. et MOESCHBERGER, Melvin L. Survival analysis : techniques for censored and truncated data. New York : Springer, 2003

[93] BRESLOW, Norman et CROWLEY, John. A large sample study of the life table and product limit estimates under random censorship. The Annals of statistics, 1974, p. 437-453.

[94] SHORACK, Galen R. et WELLNER, Jon A. Empirical processes with applications to statistics. Society for Industrial and Applied Mathematics, 2009.