République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathématiques Appliquées**"

Option : Statistique

Par Mme. CHOURAB RANDA

Titre :

Estimation non paramétrique de la fonction du survie

Devant le Jury :

Mr.	BRAHIMI Brahim	Prof.	U. Biskra	Président
Mr.	YAHIA Djabrane	Prof.	U. Biskra	Rapporteur
Mme.	BENBRIKA Ghozlane	MAA.	U. Biskra	Examinatrice

Soutenu Publiquement le 26/06/2022

Dédicace

Par le soin de Allah et tout le courage et la patience qu'il m'a apporté de ses années d'étude que j'arrive aujourd'hui à voir le mon travail ce modeste memoir.Je dédie ce travail

A mon père qui a beaucoup souffert pour nous que la vie nous soit de mieux

Merci ma Père !

A ma mère qui a donné cher et qui a tant sacrifié pour que j'arrive au bon part

Merci ma Mère !

A'amourde ma vie mon mari Karim et mon cher fils Anas

A les fleures de ma vie a mes freres et mes soeurs ; Rachid ; Nabil ; Omar ; Oussama ; Hatim ; Saif ; Nacira ; Sabah ; Djaouda ; et petite roses Nour Achamse

A tous mes amis surtout Ilyas Bouzina ; Kravey Khawla

Je dédie aussi tous mes respectes et toutes mes appréciations a ceux qui mont .Merci infiniment.

$\mathcal{R}$emerciements

J'adresse mes sincères remerciements, mon appréciation et mon respect à mon encadreur Prof. Yahia Djabrane.

Je tiens à remercier également les membres du comité d'examen, Prof. B. Brahimi (en qualité de président) et Mme. G. Benbrika (examinatrice).

A tous ceux qui ont contribué à la réalisation de ce mémoire :

Merci

Résumé du mémoire

Les fonctions de survie sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines.

L'objectif de ce mémoire est d'étudier l'estimateur non paramétrique de la fonction de survie. Nous étudions les différentes propriétés de cette fonction et nous donnons quelques résultats importants et utiles, des concepts de base sur le contexte des distribution et fonction de la survie.Nous donnons aussi des exemples et nous établissons sa relation avec la théorie des valeurs extrêmes.

Notations et symbols

n	Effectif ou taille de l'échantillon
$\wedge$	Minimum.
$\vee$	Maximum.
va	Variable aléatoire
$1_{(A)}$	Fonction indicatrice.
$S(.)$	Fonction de survie.
$h(.)$	Fonction de risque (hazard)
$H(.)$	Fonction de risque cumulée.
δ_i	Iindicatrice de l'évènement i
∞	Proportionelle à.
$i.i.d$	Indépendants et indentiquement distribution.
$\overset{P}{\rightarrow}$	convergence en probabilités.
$\overset{P.s}{\longrightarrow}$	convergence presque sur.
$E[X]$	Espérance mathématique ou moyenne de la va X.
$V(X)$	Variance de la va X.

EVD	Distribution des valeurs extrêmes.
GPD	Distribution de Pareto généralisée.
IVE	Indice des valeurs extrêmes.
$D(A)$	Domaine d'attraction du maximum
$F_n(.)$	Fonction de répartition empirique.
$\widehat{F_n}$	Estimateur de Kaplan-Meier.
μ	Moyenne d'une va.
Q	Fonction de quantile.
Q_n	Quantile empirique.
$\mathbb{R}$	Ensemble des valeurs réelles.
$N(0.1)$	Loi Normale de moyenne 0 et Variance 1

Table des matières

Table des figures

Introduction

L'estimation est un élément fondamental de la statistique qui permet la généralisation des résultats observés sur un échantillon à la population étudié. On y distingue deux approches : paramétrique qui considère que les modèles sont connus, avec des paramètres inconnus. Dont la loi de la variable étudiée est supposée appartenir à une famille de lois pouvant être caractérisée par une forme fonctionnelle connue (fonction de répartition F, densité f,...) qui dépend d'un ou plusieurs paramètres inconnus à estimer.

L'approche non paramétrique, qui ne fait aucune hypothèse sur la loi, ni sur ses paramètres (on ne fait, extérieurement à l'échantillon, aucune hypothèse restreignant le choix de la distribution à une famille particulière). Nos connaissances ainsi, sur le modèle ne sont pas précises, ce qui est souvent le cas dans la pratique. Dans ce deuxième cas, il est naturel d'estimer une des fonctions décrivant le modèle, comme la fonction de répartition ou la densité.

L'objectif de ce mémoire est d'étudier les méthodes d'estimation non paramètrique de la fonction de survie. Nous étudions les différentes propriétés de cette fonction, nous donnons aussi des exemples et nous établissons sa relation avec la théories des valeurs extrêmes.

La thèse est organisée en deux chapitres comme suit :

Nous donnons dans le premier chapitre quelques résultats importants et utiles, des concepts de base et des exemples sur le contexte de la distribution et fonction du survie.

Le deuxième chapitre est consacré à la présentation du concept de distribution à queue lourde et des différentes classes de ce type de distributions. Les distributions à queue lourde sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines : finance, hydrologie, télécommunications, géologie,...

Chapitre 1

Distribution et fonction du survie

1.1 Statistique paramétrique et non paramétrique

On dispose d'un échantillon $X_1, \ldots, X_n$ de n observations issu d'une population (i.e, variable aléatoire (va)) X. On cherche souvent à estimer une fonction ou une quantité relative à cette population comme la moyenne, la variance, la densité ou la distribution,... à partir de l'échantillon $X_1, \ldots, X_n$.

1.1.1 Statistique paramétrique

En statistique paramétrique, nous supposons que la fonction à estimer est connue à un vecteur de paramètres $\theta \in \Theta$ près ($\Theta \subset R^d$).

Exemple 1.1.1 *Soit $X_1, \ldots, X_n$ échantillon (suite de va indépendantes et identiquement distribuées (i.i.d)) de distribution $N(m, \sigma^2)$: estimer m et σ^2, sela revient à estimer la densité*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{x-m}{\sigma}\right)^2\right).$$

Il s'agit d'une estimation paramétrique.

Remarque 1.1.1 *Souvent, on ne suppose pas de forme paramétrique pour la fonction à estimer. Par exemple si on veut étudier la surface moyenne d'un logement Y en fonction du salaire X, la fonction à estimer dans ce cas est la fonction de regression :*

$$m(x) = E[Y|X = x].$$

1.1.2 Statistique non paramétrique

En statistique non paramétrique, on s'autorise toutes les formes *a priori* ou on ne fait aucune hypothèse sur la forme, nature ou type de la distribution des variables aléatoires étudiées.

Exemple 1.1.2 *X est va de densité inconue f. Estimer $f \in \mathcal{F}$:*

$$\mathcal{F} = \{ f : [0,1] \to R, \quad f \text{ croissante} \} .$$

On utilise la statistique non paramétrique dans les cas suivants :

- Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
- Quand on n'a aucune idée sur le modèle, ou qu'on ne veut pas avoir une information a priori sur le modèle.
- Quand on ne sait pas combien de composantes on veut mettre dans un modèle à estimer.
- Quand le nombre de variables est trop grand (problème de grande dimension) et qu'un modèle paramétrique est non utilisable car il aurait de toutes façons trop de paramètres.

Remarque 1.1.2 *Les avantages de la statistique non paramétrique est qu'elle a moins d'a priori sur les observations et les modèles sont plus généraux, donc plus robuste. Par contre, l'inconvénient de la statistique non paramétrique est que les vitesses de convergence sont plus lentes par rapport a la statistique paramétrique. Dans ce cas, il faut plus de données pour obtenir une précision équivalente.*

1.2 Distribution empirique

On observe $X_1, \ldots, X_n$ ichantillon issu d'une v.a. réelle, de fonction de répartition (distribution) F :

$$F(x) = P(X \leq x).$$

C'est une fonction croissante, continue à droite et vérifie

$$\lim_{x \longrightarrow 0} F(x) = 0 \quad \text{et} \quad \lim_{x \longrightarrow \infty} F(x) = 1.$$

L'estimateur naturel de F est la distribution empirique notée F_n définie par

$$\begin{aligned} F_n(x) &= \frac{1}{n} \# \{ \textit{nombre de } X_i \leq x \} \\ &= \frac{1}{n} \sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)} \end{aligned} \tag{1.1}$$

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

Remarque 1.2.1 *F_n est la proportion des n variables qui sont inférieurs ou égales à x. Pour $1 \leq i \leq n$, la fonction F_n s'écrit en termes* *des statistiques d'ordre comme suit :*

$$F_n(x) = \begin{cases} 0, & X_{(1)} > x \\ i/n, & X_{(i)} \leq x < X_{(i+1)} \\ 0, & X_{(n)} < x \end{cases}$$

avec

$$\min(X_i) := X_{(1)} \leq \ldots \leq X_{(n)} =: \max(X_i)$$

est la statistique d'ordre associée à l'échantillon $X_1, \ldots, X_n$.

Remarque 1.2.2 *Il est clair que* F_n *est un estimateur non paramétrique de la distribution* F.

Propriété 1.2.1 F_n *est estimateur sans biais de* F.

Preuve. On a

$$E(F_n) = E\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)}\right) = P(X \leq x) = F(x).$$

Donc, $Biais(F_n) = E(F_n) - F = 0$. ■

Propriété 1.2.2 *La variance de* F_n *est* $\frac{1}{n}F(x)(1 - F(x))$.

Preuve. On a

$$\begin{aligned} E\left(F_n^2\right) &= E\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)}\right)^2 \\ &= \frac{1}{n^2} E\left(\sum 1^2_{(X_i \leq x)} + \sum_{i \neq j} 1_{(X_i \leq x)} 1_{(X_j \leq x)}\right) \\ &= \frac{1}{n} F(x) + \frac{n-1}{n} F^2(x). \end{aligned}$$

D'ou,

$$Var(F_n) = E\left(F_n^2\right) - E^2(F_n) = \frac{1}{n} F(x)(1 - F(x)).$$

■

Propriété 1.2.3 *Comme* $Biais(F_n) = 0$ *et* $Var(F_n) \to 0$, *quand* $n \to \infty$, *alors* F_n *est un estimateur convergent. En effet*

$$F_n \to F \quad \textit{en moyenne quadratique.}$$

et en utilisant l'inégalité de Markov,

$$F_n \to F \qquad en \;\; Probabilités. \tag{1.2}$$

D'après le Thèoreme de la limite centrale :

$$\sqrt{n}\left(\frac{F_n \to F}{\sqrt{F(x)(1 - F(x))}}\right) \to N(0, 1), \quad \textit{en loi.} \tag{1.3}$$

De plus, d'après le théorème de Glivenko-Contelli :

$$\sup_x |F_n(x) - F(x)| \to 0 \quad P.s. \tag{1.4}$$

1.3 Exemples de distributions empiriques

Nous donnons dans la suite quelques exemples sous logiciel **R**, des distributions empiriques :

- La figure Fig. 1.1 est la courbe de la fonction de distribution d'une loi normale centrée et réduite $N(0,1)$ et sa distribution empirique.
- La figure Fig. 1.2 est la courbe de la fonction de distribution d'une loi exponentielle de paramètre $\lambda = 1$, notée $Exp(1)$ et sa distribution empirique.
- La dernière figure Fig. 1.3 est celle des courbes de la fonction de distribution d'une loi uniforme, notée $Unif(0,1)$ et sa distribution empirique.

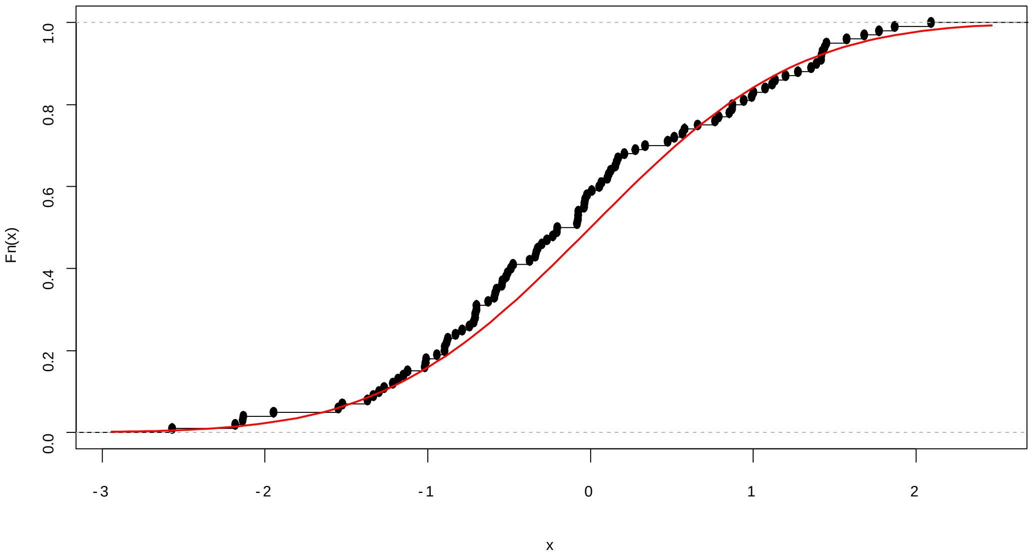

FIG. 1.1 – Distribution empirique, d'une loi $N(0,1)$, $n = 100$.

1.4 Fonction de survie

Une fonction qui joue un rôle important dans les études d'analyse des données de survie est la fonction décrivant la distribution du temps de survie (en en médecine par exemple, la durée de guérison d'un patient où la durée de rémission d'un malade). Elle est définie comme suit :

Définition 1.4.1 *La fonction de survie, aussi appelée queue de distribution, notée souvent $S(t)$ ou $\overline{F}(t)$ est défnie sur R_+ par*

$$S(t) = \overline{F}(t) := 1 - F(t) = \mathbb{P}(X > t). \tag{1.5}$$

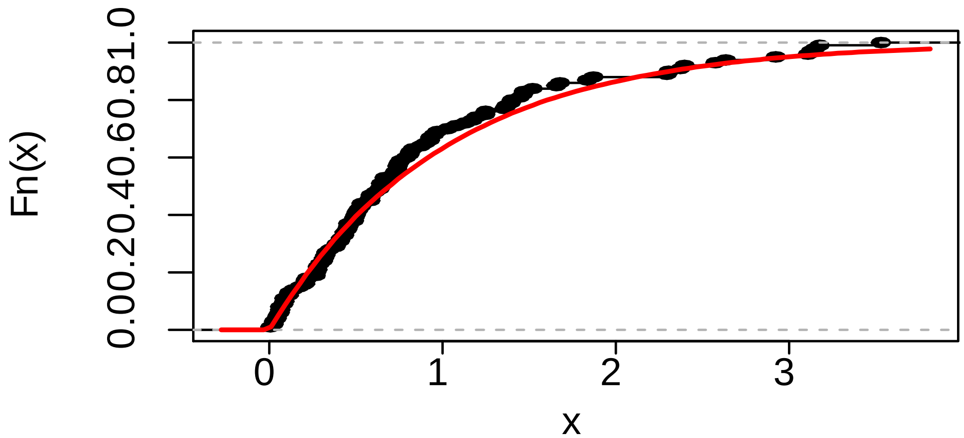

FIG. 1.2 – Distribution empirique, d'une loi $Exp\,(1)$, $n = 100$.

C'est la probabilité qu'un individu vive au-delà d'une date t.

Propriété 1.4.1 *$\overline{F}$ est une fonction décroissante, continue à gauche telle que*

$$\lim_{t \longrightarrow 0} \overline{F}\,(t) = 1 \quad et \quad \lim_{t \longrightarrow \infty} \overline{F}\,(t) = 0.$$

Exemple 1.4.1 *La figure suivante Fig. 1.4 montre la courbe de la fonction de survie d'une variable $N\,(0, 1)$.*

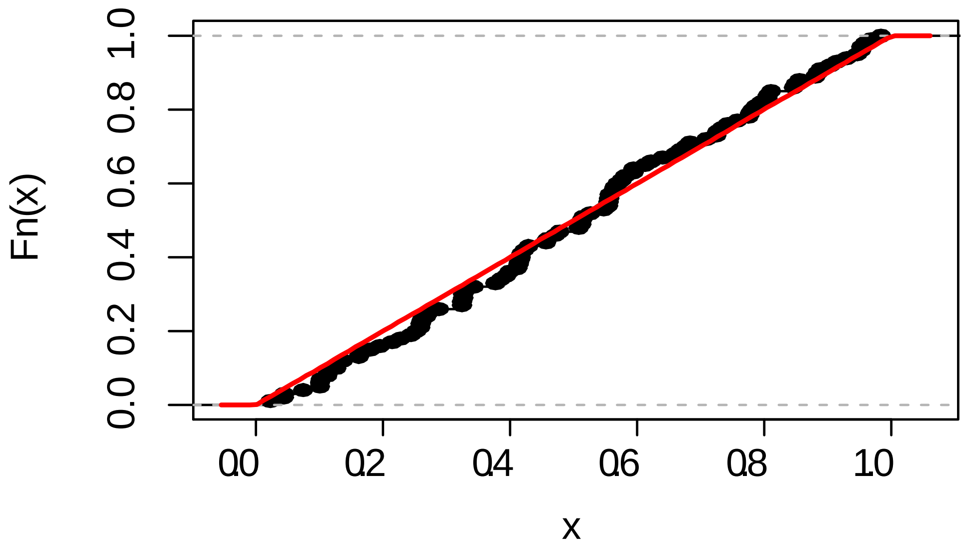

FIG. 1.3 – Distribution empirique, d'une loi $Unif_{(0,1)}$, $n = 100$.

1.5 Fonction de survie empirique

Soit $X_1, ..., X_n$ un échantillon de taille $n \geq 1$ d'une va positive X de distribution F et de fonction de survie $\overline{F}$. En utilisant la définition de la fonction de distribution empirique, en peut définir la fonction de survie empirique notée $\overline{F}_n$ par

$$\begin{aligned}\overline{F}_n(t) &= 1 - F_n(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}} \\ &= \frac{1}{n}\{nombre\ de\ X_i > x\} \\ &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i > t\}}, \qquad \forall t \geq 0.\end{aligned}$$

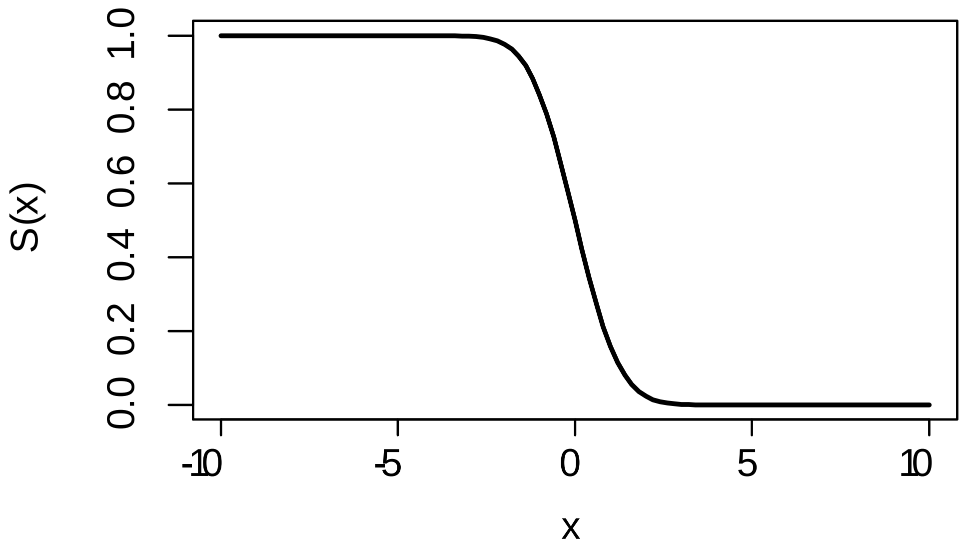

FIG. 1.4 – Fonction de survie d'une variable $N(0,1)$.

$\overline{F}_n(t)$ est la proportion d'observations qui dépasse t.

Pour $1 \leq i \leq n$, la fonction $\overline{F}_n$ s'écrit en termes des valeurs des statistiques d'ordre comme suit

$$\overline{F}_n(t) = \begin{cases} 0 & \text{si} \quad t < X_{(n)}, \\ 1 - \frac{i}{n} & \text{si} \quad X_{(i)} \leq t < X_{(i+1)}, \\ 1 & \text{si} \quad t > X_{(n)}. \end{cases}$$

1.5.1 Propriétés de la fonction de survie empirique

En utilisant les propriétés de la distribution empirique F_n, on déduit les propriétés suivantes :

1) $\overline{F}_n(x)$ est un estimateur sans biais de $\overline{F}(x)$: On a

$$E\left(\overline{F}_n(x)\right) = E\left(1 - F_n(x)\right) = 1 - E\left(F_n(x)\right) = 1 - F(x) = \overline{F}(x)$$

car, $F_n(x)$ est un estimateur sans biais de $F(x)$. Alors,

$$Biais\left(\overline{F}_n\right) = E\left(\overline{F}_n(x)\right) - \overline{F}(x) = 0.$$

2) La variance de $\overline{F}_n(x)$ est donnée par

$$\begin{aligned} Var\left(\overline{F}_n(x)\right) &= Var\left(1 - F_n(x)\right) = Var\left(F_n(x)\right) \\ &= \frac{1}{n} F(x)\, \overline{F}(x). \end{aligned}$$

3) $\overline{F}_n$ est un estimateur convergent de F :

$$F_n(x) \xrightarrow{P} F(x), \quad \text{donc} \;\; 1 - F_n(x) = \overline{F}_n(x) \xrightarrow{P} \overline{F}(x) = 1 - F(x).$$

4) Théorème central limite :

$$\sqrt{n}\left(\overline{F}_n(x) - \overline{F}(x)\right) \xrightarrow{Loi} N\left(0, F(x)\, \overline{F}_n\right), \qquad \text{quand } n \to \infty.$$

5) De plus, d'après le théorème de Glivenko-Contelli :

$$\sup_x |F_n(x) - F(x)| \to 0 \;\; \text{P.s.}$$

On a la convergence en norme infinie de $\overline{F}_n(x)$ vers $\overline{F}(x)$:

$$\begin{aligned}\sup_x |\overline{F}_n(x) - \overline{F}(x)| &= \sup_x |(1 - F_n(x)) - (1 - F(x))| \\ &= \sup_x |F_n(x) - F(x)| \to 0 \;\; \text{P.s.}\end{aligned}$$

1.6 Applications

Dans cette section, nous donnons des applications dans lesqueles nous utilisons la fonction de la survie.

1.6.1 Temps moyen et variance du survie

Soit X une durée de survie (i.e ; une va de distribution F). On définie les paramètres de position et de dispersion liés à la durée de survie comme suit :

Définition 1.6.1 *Le temps moyen du survie ou bien l'espérance de la durée de survie X, est donnée par :*

$$E(X) = \int_0^\infty t f(t)\, dt = \int_0^\infty \overline{F}(t)\, dt.$$

Définition 1.6.2 *La variance du survie est donnée par :*

$$\begin{aligned} Var\left(X\right) &= E\left(X^2\right) - E\left(X\right)^2 \\ &= 2\int_0^{+\infty} t\overline{F}\left(t\right)dt - E\left(X\right)^2. \end{aligned}$$

1.6.2 Fonction des quantiles

Définition 1.6.3 *Les quantiles de la durée de survie pour* $0 \leq p \leq 1$, *notés* q_p, *sont donnés par la formule :*

$$q_p = \inf_x \left\{x|F\left(x\right) \geq p\right\} = \inf_x \left\{x|\overline{F}\left(x\right) \leq 1-p\right\}.$$

Remarque 1.6.1 (Quantile empirique) *La fonction des quantiles empiriques notée* Q_n *de l'échantillon* X_1 *, ...,* X_n *est définie par*

$$\begin{aligned} Q_n\left(t\right) &= F_n^{-1}\left(t\right) = \inf_x \left\{x : F_n\left(x\right) \geq t\right\}, \quad 0 < t < 1 \\ &= \overline{F}_n^{-1}\left(1-s\right) := \inf_t \left\{t : \overline{F}_n\left(t\right) \geq 1-s\right\}, \quad 0 < s < 1 \end{aligned}$$

Q_n *peut être exprimée comme une fonction simple des statistiques d'ordre. On a*

$$Q_n\left(s\right) =: X_{i:n} \quad si \quad \tfrac{i-1}{n} < s \leq \tfrac{i}{n} \quad i = 1, ..., n$$

Remarque 1.6.2 *On note que pour* $0 < s < 1$, $X_{\lceil ns \rceil + 1:n}$ *est le quantile empirique d'ordre* s, *ou* $\lceil a \rceil$ *désigne la partie entiére de* a.

Définition 1.6.4 *La fonction des quantiles du queue, notée* U *est définie par*

$$U(t) := F^{-1}\left(1 - \frac{1}{t}\right) = \left(\frac{1}{\overline{F}}\right)^{-1}(t), \qquad t \geq 1.$$

La fonction empirique correspondante est :

$$U_n(t) := Q_n(1 - 1/t), \; t \geq 1.$$

Ces différentes fonctions ont des applications importantes dans la théorie des valeurs extrêmes. C'est le sujet du prochain chapitre.

Chapitre 2

Fonction du survie et théorie des valeurs extremes

Considèrons n va's réelles $X_1, X_2, ..., X_n$ iid de fonction de répartition F. La théorie des valeurs extremes (TVE) s'intérsse à des valeurs qui sont supérieures à un certain quantile x_p, à savoir :

$$P(X_i \geq x_p) < \delta, \quad \text{pour } \delta \text{ très petit}$$

Si x_p est le maximum de l'échantillon, alors pour estimer la probabilité précédente c'est d'observer le phénoméne extreme ayant une valeur plus grande que le maximum x_p de l'échantillon. Cette probabilité peut etre exprimée sous forme d'un quantile dont l'estimation dépend du paramètre de la queue de distribution.

2.1 Distributions à queue lourde

Les distributions à queue lourde sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines : finance, hydrologie, télécommunications, géologie ... etc.

En théorie statistique, les distributions à queue lourde sont des fonctions dont la queue est plus lourde que la distribution exponentielle (voir, Fig.2.1).

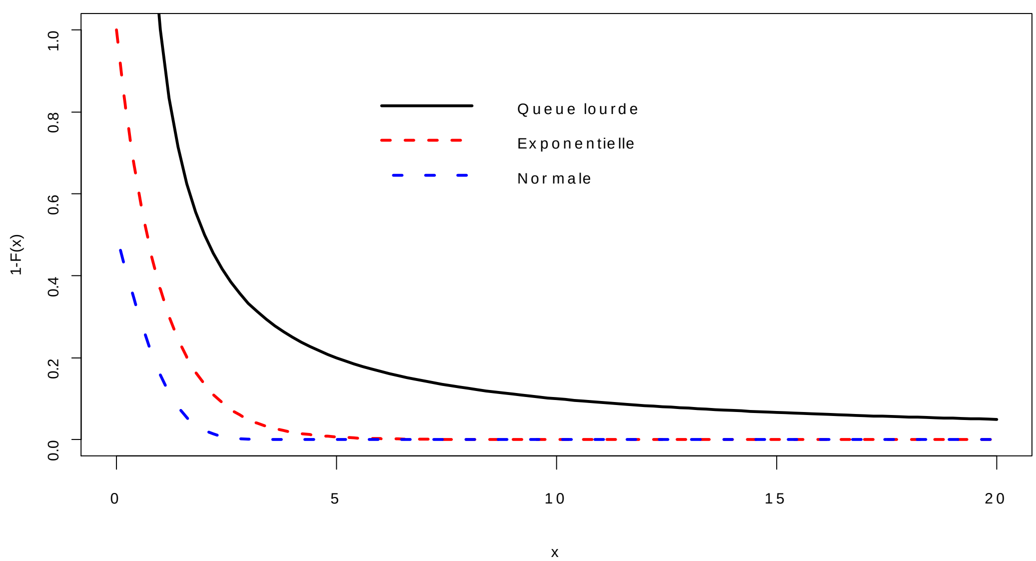

FIG. 2.1 – Décroissance des fonctions de queue : normale, exponentielle et lourde.

Il existe plusieurs définitions différentes de distribution à queue lourde, toutes sont liées à la décroissance de la fonction de queue ou survie $\overline{F}$.

La queue d'une distribution représente la probabilité pour les grandes valeurs (extrêmes) de la variable X. Lorsque ces valeurs extrêmes apparaissent dans un jeu de données, leurs probabilités d'occurrence ne sont pas nulles :

$$\overline{F}(x) = P(X > x) \neq 0.$$

Définition 2.1.1 *Soit X une va de fonction de distribution F. On dit que cette distribution a une queue lourde si*

$$\bar{F}(x) = P(X > x) \sim x^{-\alpha}, \text{ quand } x \to \infty,$$

où le paramètre $\alpha > 0$ est appelé l'indice du queue.

Remarque 2.1.1 *Ce type de distribution est appelée aussi les distributions de type-Pareto. Puisque, par définition, la fonction du survie $\bar{F}$ est équivalante à celle de Pareto. En effet, si X une va de distribution Pareto de paramètre α :*

$$F(x) = 1 - x^{-\alpha}, \quad \alpha > 0 \quad \text{et } x > 1.$$

donc,

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

Remarque 2.1.2 *Le paramètre α mesure la lourdeur de la queue de la distribution (voir, Figure 2.2).*

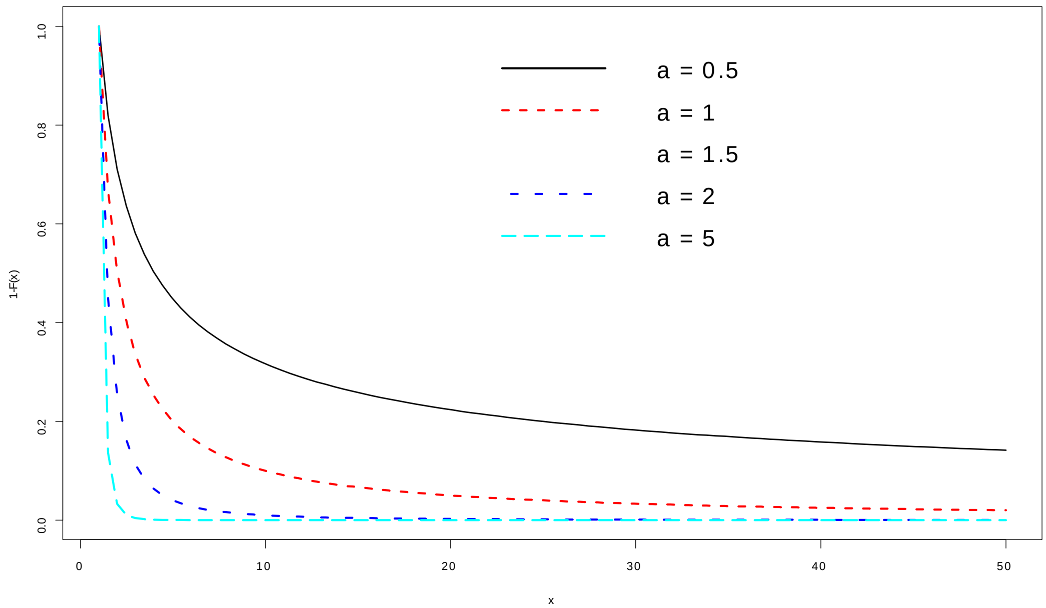

FIG. 2.2 – Fonction du queue de la distribution de Pareto pour $\alpha = 0.5, 1, 1.5, 2$ and 5.

2.2 Exemples de distributions à queue lourde

Exemple 2.2.1 (Distribution de Pareto) *Cette distribution a une fonction de queue donnée par*

$$\bar{F}(x) = \left(\frac{c}{x+c}\right)^{\alpha},$$

de paramètres $c > 0$ *et* $\alpha > 0$.

Remarque 2.2.1 *Clairement nous avons,*

$$\bar{F}(x) \sim \left(\frac{x}{c}\right)^{-\alpha} \qquad \textit{quand } x \to \infty.$$

La distribution de Pareto a tous les moments d'ordre $\lambda < \alpha$ finis, tandis que tous les moments d'ordre $\lambda \geq \alpha$ sont infinis :

$$E\left(X^{\lambda}\right) = \begin{cases} < \infty, & si\ \lambda < \alpha \\ \infty, & si\ \ \lambda \geq \alpha. \end{cases}$$

Exemple 2.2.2 (Distribution de Burr) *Cette distribution est utilisée pour modéliser les pertes en assurance. Elle a une fonction de queue donnée par*

$$\bar{F}(x) = \left(\frac{c}{x^{\tau} + c}\right)^{\alpha},$$

de paramètres $\alpha, c, \tau > 0$.

Remarque 2.2.2 *Nous avons*

$$\bar{F}(x) \sim c^{\alpha} x^{-\tau\alpha}, \quad \textit{quand } x \to \infty,$$

donc la distribution de Burr est similaire dans sa queue à celle de Pareto, dont elle est par ailleurs une généralisation. Tous les moments *d'ordre $\lambda < \alpha\tau$ sont finis, tandis que ceux d'ordre $\lambda \geq \alpha\tau$ sont infinis :*

$$E\left(X^{\lambda}\right) = \begin{cases} < \infty, & si\ \lambda < \alpha\tau \\ \infty, & si\ \ \lambda \geq \alpha\tau. \end{cases}$$

Remarque 2.2.3 *Pour de nombreuses applications, il est important de connaître α ainsi que de déterminer le nombre de moments finis. Par exemple, si*

$$\alpha < 2, \quad alors \ \ E\left(X^{2}\right) = \infty$$

est vrai et donc le théorème de la limite centrale ne tient pas.

2.3 Distribution à variation régulière

Une classe importante de distributions à queue lourde est la classe des distributions à variation régulière.

Définition 2.3.1 *Une distribution F est appelée à variation régulière à l'infini d'indice $(\gamma =: \alpha^{-1})$, si*

$$\bar{F}(x) = P(X > x) = x^{-1/\gamma} L(x), \qquad \forall x > 0,$$

avec $\gamma > 0$ est ce que l'on appelle l'indice des valeurs extrêmes (ive). On note dans ce cas

$$\overline{F} \in RV_{-1/\gamma}.$$

et $L(x)$ est une fonction à variation lente i.e, $L \in \mathcal{RV}_0$.

Remarque 2.3.1 *Une fonction positive et mesurable $L(x)$ sur $(0, \infty)$ est appelée une fonction à variation lente à l'infini si*

$$\lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1, \quad \forall t > 0.$$

$\log(x)$, $\log(\log(x))$ *sont des exemples de $L(x)$. Généralement, $L(x)$ est toute fonction convergente vers une constante positive.*

Remarque 2.3.2 *Pour les fonctions de distribution à variation régulière, le moment d'ordre k n'existe pas dès que $k \geq 1/\gamma$. Cela a quelques implications importantes : Lorsque nous considérons la somme des va iid qui ont des distributions à variation régulière avec un indice de queue $\alpha < 2$, la variance de ces va est infinie, et donc le théorème de la limite centrale ne tient pas.*

Exemple 2.3.1 *Si X suit une loi de Pareto avec la fonction de queue*

$$\bar{F}(x) = \left(\frac{c}{x + c} \right)^{\alpha},$$

alors, $\overline{F} \in RV_{-\alpha}$.

Exemple 2.3.2 *Si X a une distribution de Burr :*

$$\bar{F}(x) = \left(\frac{c}{x^{\tau} + c} \right)^{\alpha},$$

alors, $\overline{F} \in RV_{-\tau\alpha}$.

2.4 Théorème fondamental de la TVE

Connu sous le nom de théorème de Fisher et al.(1928), le théorème fondamental de la TVE donne les lois limites possibles du maximum de l'échantillon et permet d'avoir une certaine connaissance sur la queue de la distribution.

Notons par $X_{(n)}$ la plus grande statistique d'ordre (ou statistique du maximum) :

$$X_{(n)} = \max\left(X_1, ..., X_n\right) = \max\left(X_{(1)}, ..., X_{(n)}\right).$$

Etant donnée la fonction F et que les variables aléatoires sont i.i.d. Alors, la fonction de répartition du maximum est donnée par

$$\begin{aligned} F^n(x) &= \mathbb{P}\left(X_{(n)} \leq x\right) = \mathbb{P}\left(X_1 \leq x, ..., X_n \leq x\right) \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq x\right) = \left(F(x)\right)^n. \end{aligned}$$

Puispue F étant souvent inconnue, on ne peut généralement déterminer la distribution F^n du maximum à partir du résultat précédent.

Théor■me 2.4.1 (Fischer-Tippett, 1928 et Gnedenko, 1943) *Soient deux suites $a_n \geq 0$ et b_n telles que pour tout $x \in R$,*

$$\frac{X_{(n)} - b_n}{a_n} \overset{Loi}{\to} H(x) \Leftrightarrow F^n\left(a_n x + b_n\right) \longrightarrow H(x),$$

H une fonction de répartition non dégénérée. On dit alors que F appartient au domaine d'attraction de H, noté $DAM(H)$. On a de plus, si $F \in DAM(H)$, alors

$$nF(a_n x + b_n) \longrightarrow -\ln H(x).$$

H appartient à l'un des trois types suivants de domaine d'attraction :

$$\begin{aligned}
Type\ I\ \ (Gumbel): &\ \exp\left(-e^{-x}\right), \quad \forall x \in \mathbb{R}, \\
Type\ II\ \ (Fréchet): &\begin{cases} 0 & si \quad x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & si \quad x \geq 0 \end{cases} \\
Type\ III\ \ (Weibull): &\begin{cases} \exp\left(-\left(-x\right)^{\alpha}\right) & si \quad x \leq 0 \\ 1 & si \quad x \geq 0 \end{cases}
\end{aligned}$$

Remarque 2.4.1 *Ce théorème donne l'ensenble des lois limites H en considérant les variables du type $dX+m$, où X suit une loi de Weibull, de Gumbel ou de Fréchet et d, m des réels.*

Remarque 2.4.2 *Généralement, voir par exemple de Haan et Ferreira* (2006), *on nomme la distribution des valeurs extremes (EVD) :*

$$H_{\gamma}(x) = \begin{cases} \exp\left(-\left(1+\gamma x\right)^{-1/\gamma}\right), & \gamma \neq 0 \quad 1+\gamma x \geq 0 \\ \exp\left(-\exp\left(-x\right)\right), & \gamma = 0 \quad \infty \leq x \leq \infty \end{cases}$$

Remarque 2.4.3 *Le théorème précédent est vrai pour la plupart des lois usuelles.*

2) La fonction de distribution $H_\gamma(x)$ est appelée distribution généralisée des valeurs extrêmes.

3) Si F vérifie le théorème précédent, on dit que F appartient au domaine d'attraction de $H_\gamma(x)$.

Remarque 2.4.4 *Les trois domaines d'attractions associés à γ sont :*

■ *Fréchet $(\gamma \geq 0)$: Ce domaine d'attraction contient les fonctions de distribution à queue lourde (avec décroissance polynomiale) telles que la distribution de Cauchy, les distributions Pareto, Burr et log gamma...*

■ *Weibull $(\gamma \leq 0)$: contient la majorité des fonctions de distribution dont le point final est fini comme la loi uniforme, la loi $Beta(p, q)$ et la distribution de Weibull,...*

■ *Gumbel $(\gamma = 0)$: Ce domaine d'attraction contient les fonctions de distribution (avec décroissance exponentielle) telles que gaussienne, exponentielle, gamma, log normale,...*

Conclusion

La fonction de survie constitue une caracteristique importante des modèles de durée de vie et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines.

Dans ce mémoire nous avons étudié l'estimateur non paramétrique de cette fonction de survie. Nous avons rappeler ces différentes propriétés et les résultats importants, utiles et des concepts de base sur le contexte de la distribution et de la fonction de survie. Des exemples sont donnés et la relation avec la théorie des valeurs extrêmes est établie.

Bibliographie

[1] Castillo, E. (1988). Extreme value theory in engineering. Statistical Modeling and Decision Science. Academic Press, Inc.

[2] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, New York.

[3] Embrechts, P. Klueppelberg ,C. and Mikosch T. (1997). Modelling Extremal Events for Insurance and Finance. Springer.

[4] Fisher, R. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of largest or smallest member of a sample. Proceedings of the Cambridge philosophical society. 24,180–190.

[5] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of mathematics, 423-453.

[6] de Haan, L. and Ferreira, A. (2006). Extreme Value Theory : An Introduction. Springer.

[7] Smith, R. (1987). Estimating tails of probability distributions. The Annals of Statistics, 1174-1207.