République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

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

Département de Mathématiques

Mémoire présenté en vue de l'obtention du

DIPLÔME DE MASTER en MATHÉMATIQES

Option : **Statistique**

Par

Chaima Heddar

Thème

Sur l'estimation de la fonction de survie

Membres du Comité d'Examen

Pr.	**Cherfaoui Mouloud**	UMKB	Président
Dr.	**Louiza Soltane**	UMKB	Encadreur
Dr.	**Dhiabi Samra**	UMKB	Examinateur

Juin 2022

À *l'amour de mon cœur mon père et ma chère mère pour leur soutien permanent,*

À mes soeurs et mes frères,

À ma famille et mes amis,

À tous ceux qui le méritent.

REMERCIEMENTS

Tout d'abord, je remerciais ALLAH, le tout-puissant et le miséricordieux, de m'avoir donné la santé, la volonté et la patience durant ces longues années d'études pour compléter mes études de licence et master.

Je remercie beaucoup mon encadreur Dr. Louiza Soltane qui mon encadreuse pendant la période de la réalisation de ce travail pour sa disponibilité, ses orientations et remarques précieuses qui m'ont aidé à bien présenter ce travail.

Mes remerciements vont également aux membres du jury "le Dr. et le Dr. " pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon mémoire et de l'enrichir par leurs propositions.

Je remercie tous les enseignants qui ont contribué à m'éducatif, ainsi que tous les employés du département de mathématiques.

Je tiens aussi à remercier ma famille, mes amis et collègues en particulier Ramissa, Ikram, grâce à qui ma vie universitaire a été très agréable.

Enfin, je remercie chaleureusement toutes personnes qui m'ont aidé, et qui ont contribué de proche ou de loin à la réalisation de ce travail.

Heddar chaima.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant un évènement d'intérêt, et aide à identifier les conditions et les caractéristiques qui augmentent ou diminuent la probabilité de survie. Elle trouve sa place dans tous les champs d'applications où l'on étudie le délai de survenue d'un évènement dans un ou plusieurs groupes d'individus, par exemple :

- En médecine : durée de guérison d'un patient, durée de progression de la maladie, ...
- En fiabilité : durée de vie d'un matériel, durée de vie d'une lampe , ...
- En biologie : en culture de cellules les durées d'apparition de parasites, ...
- En économie : durée de chômage, durée de la hausse des actions, ...
- En éducation : durée d'apprentissage et l'enseignement d'une langue, ...

La base de toute analyse statistique est l'échantillon à qui il arrive parfois d'être censuré. Il existe plusieurs mécanismes de censure dont la plus couramment rencontrée est la censure aléatoire à droite.

La théorie des valeurs extrêmes est une branche de la statistique apparue entre 1920 et 1943, qui a pour but de modéliser et de décrire la survenue et l'intensité d'évènements dits rares c'est-à-dire qui présentent des variations de très grandes amplitudes (ayant une faible pro-

babilité d'apparition). Il s'agit fondamentalement de modéliser un phénomène aléatoire, en s'intéressant principalement aux quantiles extrêmes et à la queue de distribution souvent modélisée par un indice appelé indice des valeurs extrêmes. Depuis quelques dernières années cette théorie a reçu beaucoup d'attention aussi bien sur le plan théorique que pratique et ne cesse d'élargir son champ d'application à cause de présence de censures, comme récemment dans beaucoup de domaines par exemple : la finance, l'assurance, la météorologie, l'hydrologie,... etc.

Dans ce travail, on s'intéresse à l'estimation non-paramétrique et semi-paramétrique de la fonction de survie en présence de données censurées et pour les distributions à queue lourdes. Dans ce cadre, on a fait une synthèse sur les différentes théories et résultats sur l'analyse de survie et sur la TVE. Ce mémoire est composé de trois chapitres.

- Le premier et le deuxième chapitre contient le cadre théorique du mémoire et comprends quelques rappels et définition de base : va's, la fdr, la fonction de survie,.... Ensuite, on donne généralités sur l'analyse de survie : la définition de données censurées et ses types.... La dernière partie on introduit l'estimation non-paramétriques de Kaplan et Meier [17] de fonction de survie.

  Dans le cas de non censure, il y a toute une théorie des valeurs extrêmes qui donne dans le deuxième chapitre, où on parle sur les statistiques d'ordre, les résultats limites sur la distribution de maximum de l'échantillon. Ensuite on discute également la notion des domaines d'attraction d'une distribution selon le paramètre de l'indice de queue. Puis on parle sur l'estimateur de l'indice des valeurs extrêmes : estimation semi-paramétrique dans les deux cas des données complètes et incomplètes. Enfin on parle sur l'estimation semi-paramétrique de la fonction de survie pour les données extrêmes et sans les données censurées et l'autre cas en présence les deux. Ces définitions et résultats proviennent, entre autres, des références [1], [6], [7], [9], [11], [24], [27], [32], [33],...
- Les concepts théoriques du 1er (vus au Chapter 1) sont appliqués, dans le dernier

chapitre, sur un ensemble de données réelles : les patientes atteintes d'un cancer du sein.

Enfin, il y a lieu de noter que les calculs numériques et les représentations graphiques sont réalisés à l'aide des packages "KMsurv", "survminer", "tidyverse" et "ggplot2" du logiciel d'analyse statistique R (The R Project for Statistical Computing).

CHAPITRE 1

Analyse de survie

1.1 Modélisation non-paramétrique

On commence dans ce chapitre par quelques rappels et définitions couramment utilisées dans les études. Ces définitions peut être trouvés dans tout manuel standard de la théorie des probabilités comme Saporta [27], 2006.

1.1.1 Définitions de base

Définition 1.1.1 (Variable aléatoire)

Une variable aléatoire v.a réelle X est une fonction définie sur un espace probabilisable $(\Omega, \mathcal{F})$ à valeur dans $\mathbb{R}$, et mesurable par rapport à la tribu $B_{\mathbb{R}}$ (tribu borélienne de $\mathbb{R}$)

$$\begin{array}{rccc} X: & \Omega & \rightarrow & \mathbb{R} \\ & w & \rightarrow & X(w). \end{array}$$

Définition 1.1.2 (Loi de probabilité)

On appelle loi de probabilité de X la mesure d'image de P par X et on la note P_X, définie sur l'espace $(\mathbb{R}, B_{\mathbb{R}})$ vers $[0,1]$ par :

$$\begin{aligned} \forall B \in B_{\mathbb{R}} : P_X(B) &= P\left\{X^{-1}(B)\right\}. \\ &= P\left\{X \in B\right\} \end{aligned}$$

Définition 1.1.3 (Fonction de répartition)

Soit X une v.a, on appelle fonction de répartition fdr de X la fonction de $\mathbb{R}$ dans $[0,1]$, définie pour tout $x \in \mathbb{R}$ par :

$$F(x) := P\left(X \leq x\right).$$

Propriété 1.1.1 (Fonction de répartition)

1. *$F(x)$ fonction croissante sur $\mathbb{R}$.*
2. *$F(x)$ fonction continue à droite en tout point de $\mathbb{R}$.*
3. *$\lim\limits_{x \to -\infty} F(x) = 0$ et $\lim\limits_{x \to +\infty} F(x) = 1$.*

Définition 1.1.4 (Fonction de survie)

La fonction de survie qu'on note par $S(x)$ ou $\overline{F}(x)$ est définie sur $\mathbb{R}_+$ par

$$S(x) = \overline{F}(x) = 1 - F(x) := P(X > x). \tag{1.1}$$

Pour t fixé c'est la probabilités de survivre jusqu'a l'instant t.

Propriété 1.1.2 (Fonction de survie)

1. *$\overline{F}$ est aussi appellé fonction de queue.*
2. *$\overline{F}(x)$ fonction décroissante monotone.*
3. *$\overline{F}(x)$ fonction continue à gauche.*

4. $\lim_{x\to 0}\overline{F}(x) = 1$ *et* $\lim_{x\to\infty}\overline{F}(x) = 0$.

Définition 1.1.5 (Fonctions empiriques de répartition et de survie)

Soit $X_1, ..., X_n$ *un échantillon de taille* $n \geq 1$ *d'une v.a positive* X *de fdr* F *et de fonction de survie* $\overline{F}$. *Les fonctions empiriques de répartition et de survie,* F_n *et* $\overline{F}_n$ *sont respectivement définies par :*

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}, \qquad \forall t \geq 0, \tag{1.2}$$

et

$$S_n(t) = \overline{F}_n(t) = 1 - F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > t\}, \qquad \forall t \geq 0, \tag{1.3}$$

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

Pour $1 \leq i \leq n$, les fonctions (1.2) et (1.3) s'écrivent en termes des valeurs des statistiques d'ordre[1] comme suit :

$$F_n(t) = \begin{cases} 0 & \text{si } t \leq X_{1:n}, \\ \dfrac{i}{n} & \text{si } X_{i:n} \leq t \leq X_{i+1:n}, \\ 1 & \text{si } t \geq X_{n:n}, \end{cases} \qquad \text{et} \qquad \overline{F}_n(t) = \begin{cases} 1 & \text{si } t \leq X_{1:n}, \\ 1 - \dfrac{i}{n} & \text{si } X_{i:n} \leq t \leq X_{i+1:n}, \\ 0 & \text{si } t \geq X_{n:n}. \end{cases}$$

Définition 1.1.6 (Fonction de quantile)

Soit X *une v.a et* F *sa fdr, la fonction de quantile est définie pour toute* $0 < p < 1$ *par :*

$$Q(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\} = F^{-1}(p),$$

où F^{-1} *désignant l'inverse généralisé de* F. *On l'exprime en termes de la fonction de queue par :*

$$Q(p) := \inf\left\{x \in \mathbb{R} : \overline{F}(x) \leq 1 - p\right\}, \qquad 0 < p < 1.$$

[1] Les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$, sont obtenues en classant ces va's par ordre croissant $X_{1:n} \leq ... \leq X_{n:n}$. Une brève étude sur ces dernières est donnée dans la Subsection 2.1.1.

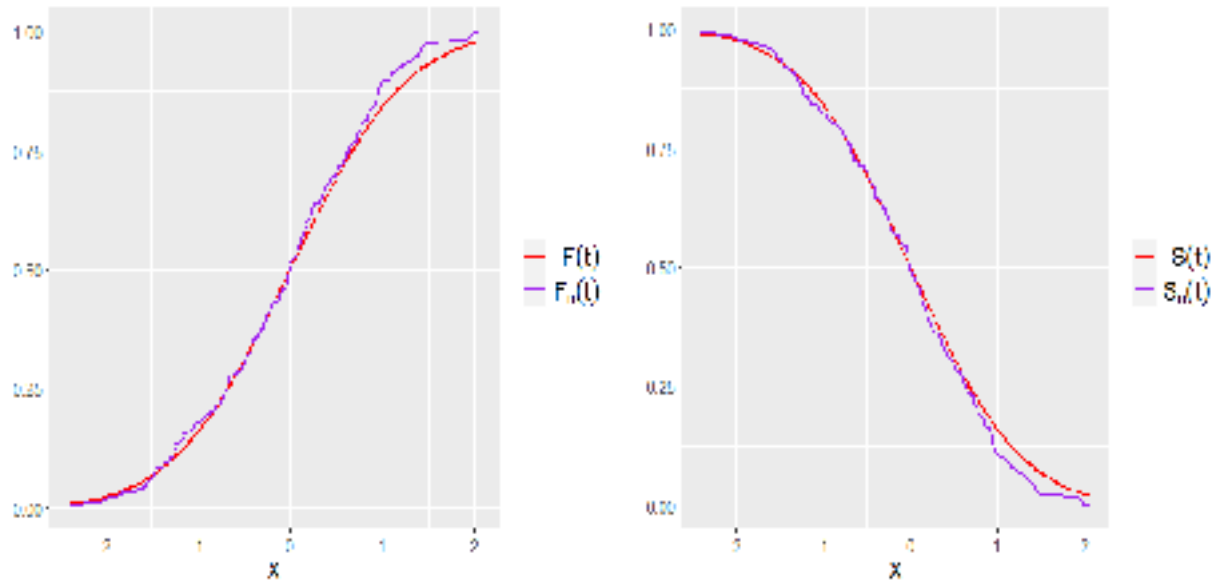

FIG. 1.1 – Fonctions theorique et empirique de repartition (Gauche) et de survie (Droite) d'un echantillon Gaussien standard de taille 200.

Définition 1.1.7 (Fonction de quantile empirique)

La fonction de quantile empirique de l'échantillon $X_1, ..., X_n$ est définie par :

$$Q_n(p) := \inf \{x \in \mathbb{R} : F_n(x) \geq p\} = F_n^{-1}(p), \qquad 0 < p < 1.$$

où

$$Q_n(p) := \inf \left\{x \in \mathbb{R} : \overline{F}_n(x) \leq 1 - p\right\} = \overline{F}_n^{-1}(1 - p), \qquad 0 < p < 1.$$

Remarque 1.1.1

Une fonction, notée par U et (parfois) appelée fonction de quantile de queue, elle est définie par :

$$U(x) := Q\left(1 - \frac{1}{x}\right) = F^{-1}\left(1 - \frac{1}{x}\right) = \left(1/\overline{F}\right)^{-1}(x), \quad 1 < x < \infty,$$

et la fonction empirique correspondante est :

$$U_n(x) := Q_n\left(1 - \frac{1}{x}\right), \qquad 1 < x < \infty.$$

1.1.2 Analyse de survie

La définition suivante vient de mémoire de Zakaria Raiti [33], 2017.

Définition 1.1.8 (Analyse de survie)

L'analyse de survie prend en compte simultanément le nombre d'événements survenus pendant une période donnée, le moment où ces événements se produisent et les sujets pour lesquels l'événement ne s'est pas encore réalisé (données censurées), i.e. d'un passage irréversible entre deux états (décès, guérison, rechute, etc), dans le temps

$$Début \overset{temps}{\Longrightarrow} événement.$$

Exemple 1.1.1

Dans la recherche médicale, l'origine du temps est souvent la date d'enregistrement de l'élément (l'individu) dans une étude, comme les essais médicaux pour comparer deux types de médicaments ou plus (tester l'éficacité d'un médicament) si le point final est la mort du patient, les données résultantes sont les temps de survie (Survival Time), mais si le point final n'est pas la mort, alors les données résultantes sont appelées données de temps sur événement (Time to Event Data).

Remarque 1.1.2

On remarque qu'il existe des nombreux autres noms comme :

- *Dans les statistiques médicales et épidémiologiques, il est connu sous le nom d'analyse de survie ou analyse de risque (Hazard Analysis).*

- *Dans les études d'ingénierie, elle est connue sous le nom d'analyse du temps d'échec (Failure Time Analysis).*
- *Dans les études de psychologie, il est connu comme analyse d'historique d'événement (Event History Analysis) (voir par exemple le livre de Göran Broström [14]).*
- *En économie connue sous le nom d'analyse de transition (Transition Analysis).*

L'exemple suivant vient de Michaël Genin [20], 2015.

Exemple 1.1.2

- *Temps de survie aprés le diagnostic d'un cancer du sein.*
- *Durée de séropositivité sans symptome de patients infectés par le VIH.*
- *Durée de vie d'une ampoule, d'un pièce mécanique, ...*

Vocabulaire associé au contexte de l'analyse de survie

Dans l'étude de survie, nous disposons d'un échantillon de n individus ou bien sujets qui possède des informations principales sur leurs dates et les différentes durées de survie, donc dans cette partie on va parler sur les principaux dates et délais qui caractéristiques des données de survie.

Date d'origine DO C'est le point de départ du suivi du patient. Il doit être le même pour tous les patients, c'est à dire défini de façon précise. Généralement, la date du calendrier varie d'un sujet à l'autre.

Exemple 1.1.3

- *Date de tirage au sort (essai thérapeutique).*
- *Date de diagnostic (étude prospective).*

Date des dernières nouvelles DDN C'est la date à laquelle on a eu pour la dernière fois des nouvelles du sujet : cela peut être la date du décès ou la date de survenue de l'événement étudié (guérison, première rechute, première apparition d'un événement indésirable, ...), mais aussi la date de la dernière consultation si le sujet est perdu de vue ou n'a pas présenté l'événement étudié (décès, guérison, ...) **(**source : Semmari [29], mémoire de master en Statistiques**).**

Date de point DP On ne prouve pas attendre la survenue de l'événement pour tous les sujets pour faire l'analyse des observations, donc il y a ce qu'on appelle la date de point qui est une date au-delà de laquelle on arrête l'observation et on ne tiendra plus compte de l'état du sujet après cet instant.

Recul Noté par L_i la variable de la censure d'un individu i avec $L_i \in [DO, DP]$. c'est le délai écoulé (différence) entre la date d'origine et la date de point c'est-à-dire,le délai maximal d'observation du sujet. Il est dit aussi le délai de la censure, il peut être connu ou inconnu, déterministe ou aléatoire, dépendant de l'individu ou non.

Temps de participation Noté par t_i, c'est le temps écoulé entre :

- la date d'origine et la date de dernières nouvelles, si cette dernière est antérieure à la date du point (décé d'avant la date de point).

$$DDN < DP \Leftrightarrow t_i \in [DO, DDN]$$

- la date d'origine et la date de point, si celle-ci est antérieure à la date de dernières nouvelles (vivant aux dernières nouvelles). **(**source : Semmari [29] et Harrouche [15]).

$$DDN > DP \Leftrightarrow t_i \in [DO, DP]$$

Etat aux dernières nouvelles variable binaire, l'événement s'est produit ou ne s'est pas «encore » produi.

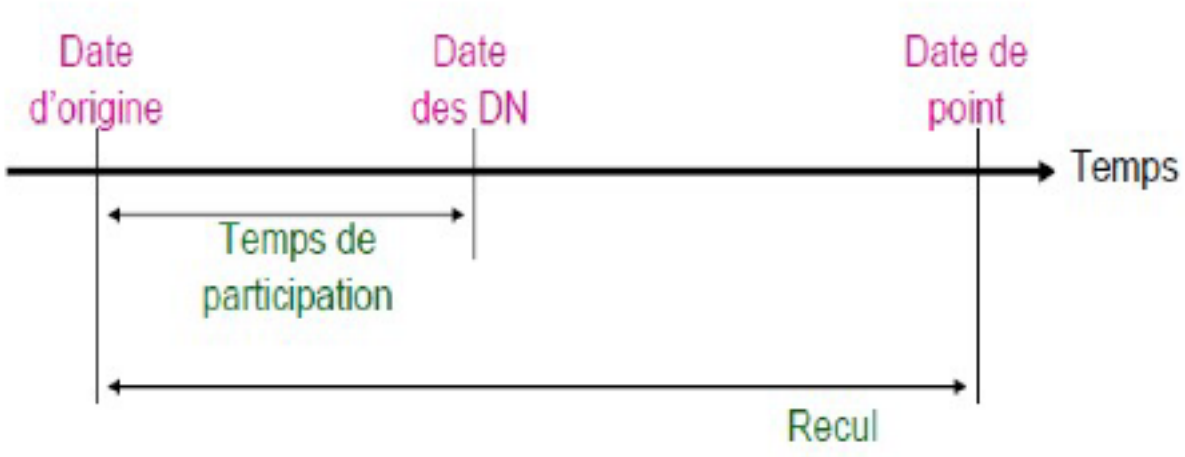

FIG. 1.2 – Schéma représentant les principales définitions relatives à l'analyse de la durée de survie

1.1.3 Notions de censures

Dans l'analyse de survie les données ne sont pas toujours complètement observées, parce que, pour certains individus de l'événement du début et /ou de fin n'est pas observé, c'est-à-dire privées d'une partie de l'information. Dans ce cas les données sont censurées, Il n'est pas rare, mais elles sont plutôt incomplètes. la censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour un individu donné j, on va considérer :

- Son temps de survie X_j.
- Son temps de censure Y_j.
- La durée réellement observée Z_j.

Définition 1.1.9 *(Variable de censure)*

La variable de censure Y est définie par la non-observation de l'événement étudié. Si l'on observe Y, et non X, 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, ...). Si l'événement se produit, X est "réalisée". S'il ne se produit pas (l'individu étant perdu de vue, ou bien du vivant), c'est Y qui est "réalisée".

Censure à droite

La variable d'intérêts est dite censurée à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées.

Censure à gauche

Il y a une censure à gauche lorsque l'individu a déjà subi l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue.

Censure double et mixte

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

Censure par intervalle

Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt.

1.1.4 Types de censures

Censure de type I : (fixée)

Soit Y une valeur fixée, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq Y$, sinon on sait uniquement que $X_i > Y$. On observe

donc une variable aléatoire Z_i telle que :

$$Z_i := X_i \wedge Y = \min(X_i, Y), \qquad i = 1, ..., n.$$

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

Exemple 1.1.4

En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps μ sont sacrifiées).

Censure de type II : (attente)

L'expérimentateur fixe a priori le nombre d'événements à observer. La date définie d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie. Soient $X_{i:n}$ et $Z_{i:n}$ sont les statistiques d'ordre associées à la v.a X_i et Z_i respectivement. La date de censure est donc $X_{k:n}$ et on observe les variables suivantes :

$$Z_{1:n} = X_{1:n}, Z_{2:n} = X_{2:n}, ..., Z_{k:n} = X_{k:n}, Z_{k+1:n} = X_{k:n}, ..., Z_{n:n} = X_{k:n}.$$

Exemple 1.1.5

En épidémiologie : on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment-là.

Censure de type III (ou censure aléatoire)

C'est type de ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type, la date d'inclusion du patient dans l'étude est fixé, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'événement observés et la durée totale de l'expérience sont aléatoires. Soit $X_1, ..., X_n$ un

échantillon indépendantes et identiquements distribuées (*iid*) d'une v.a 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 $Y_1, ..., Y_n$ iid, dans ce cas au lieu d'observer les $X_i's$, on observe un couple de va's (Z_i, δ_i) avec

$$Z_i := X_i \wedge Y_i \quad \text{et} \quad \delta_i := \mathbb{1}\{X_i \leq Y_i\}.$$

L'information disponible peut être résumée par :

1. Z_i : la durée réellement observée.
2. δ_i : l'indicateur de censure, où

$$\delta_i = \begin{cases} 1 & \text{si l'évènement est observé (d'où } Z_i = X_i\text{). On observe les durées complètes.} \\ 0 & \text{si l'individu est censuré (d'où } Z_i = Y_i\text{). On observe les durées incomplètes.} \end{cases}$$

L'exemple suivant vient de Zakaria Raiti [33], 2017.

Exemple 1.1.6

Afin d'illustrer ce phénomène, nous étudions l'exemple [i] *simple suivant :*
Imaginez que l'on transplante 4 *nanopuces dans les coeurs de* 4 *bébés tortues marines numérotés de* 1 *à* 4 *dont les oeufs ont éclos tous en même temps puis on les lâche dans l'océan. Ces nanopuces mesurent les battements de coeur de ces bébés tortues et sont connectées de façon continue à un outil de mesure dans lequel on reçoit :*

– *Le nombre de battements de coeur du bébé tortue tant que celui ci est vivant*

– *Le message «Décès» au moment où le coeur du bébé tortue s'arrête de battre*

– *Le message «Erreur» quand le signal avec la nanopuce est perdu*

Soit $X_1, ..., X_4$ les durées de vie de ces bébés tortues. On a suivi cette cohorte et au bout d'un mois on a constaté ce qui suit :

– Le bébé tortue $n^\circ 1$ est décédé au bout de 9 jours

– On a perdu tout contact avec la nanopuce du bébé tortue $n^\circ 2$ au bout de 5 jours

– Le bébé tortue $n^\circ 3$ est décédé au bout de 29 jours

– Le bébé tortue $n^\circ 4$ est toujours en vie

Ainsi, ce que l'on sait sur la durée de vie des ces tortues est comme suit :

– Pour le bébé tortue $n^\circ 1$: on observe $X_1 = 9$ jours qui est sa durée de vie

– Pour le bébé tortue $n^\circ 2$: on observe $C_2 = 5$ jours qui est la durée pendant laquelle il a survécu jusqu'à perte de tout signal avec la nanopuce. même si l'on ne la connait pas avec exactitude, on sait que sa durée de vie X_2 est forcément plus grande que C_2.

– Pour le bébé tortue $n^\circ 3$: on observe $X_3 = 29$

– Pour le bébé tortue $n^\circ 4$: on reçoit toujours des signaux de la part de la nanopuce donc on observe $C_4 = 30$ jours et on sait que forcément $X_4 > C_4$.

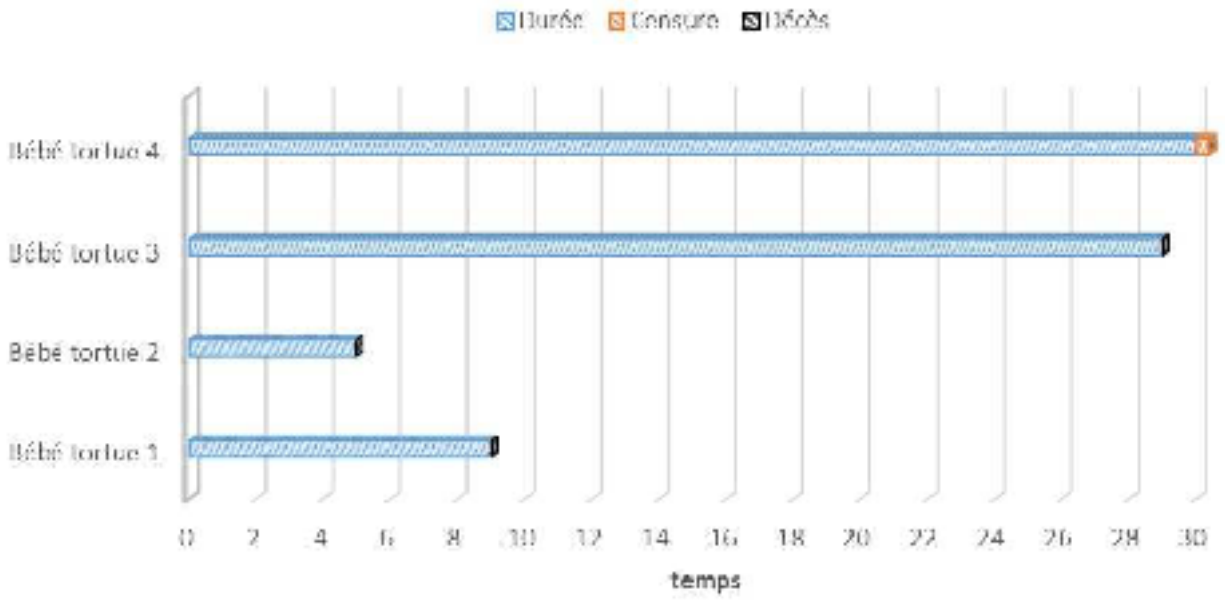

FIG. 1.3 – Illustration de censure aléatoire à droite

Remarque 1.1.3

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.

1.1.5 Estimation non-paramétrique

Les méthodes non-paramétriques sont souvent très faciles et simples a comprendre par rapport aux méthodes paramétriques, est basées sur un tableau des données contient à la fois des données censurés. De plus, l'analyse non-paramétrique est largement utilisée dans des situations, où il y a un doute sur la forme exacte de la distribution.

Estimateur de Kaplan-Meier (KM)

Kaplan et Meier [17] en 1958 ont obtenu un estimateur de la fonction de survie pour des données aléatoirement censurées à droite, nommé aussi estimateur produit-limite. Le principe de la méthode repose sur l'idée qu'être encore en vie aprés un instant t, c'est être en vie juste avant cet instant t et ne pas mourir a cet instant [**26**]. Ainsi, la survie à un instant quelconque est le produit de probabilités conditionnelles de survie de chacun des instants précédents.
On utilise le théorème de probabilité conditionnelle, Soit : $t_i \leq t_{i+1}$

$$\begin{aligned} S(t_i) &= P(X > t_i) \\ &= P(X > t_i, X > t_{i-1}) \\ &= P(X > t_i | X > t_{i-1}) P(X > t_{i-1}) \\ &= P(X > t_i | X > t_{i-1}) P(X > t_{i-1} | X > t_{i-2}) ... P(X > t_0 = 0). \end{aligned}$$

En considérant les temps d'événements (décès et censure) distincts $Z_{i:n} (i = 1, ..., n)$ rangés

par ordre croissante, on obtient :

$$P(X > Z_{i:n}) := \prod_{k=1}^{i} P(X > Z_{k:n} \,|X > Z_{k-1:n}), \qquad i = 1, ..., n$$

avec $Z_{0:n} = 0$, Considérons les notations suivantes :

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

- d_i le nombre de décès en $Z_{i:n}$.

Alors la probabilité p_i de mourir dans l'intervalle $]Z_{i-1:n}, Z_{i:n}[$ sachant que l'on était vivant en $Z_{i-1:n}$, (c'est à dire $p_i = P(X \leq Z_{i:n}|X > Z_{i-1:n})$), peut être estimé par :

$$\widehat{p_i} := \frac{d_i}{n_i}.$$

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

- $d_i = 0$ en cas de censure en $Z_{i:n}$, c'est-à-dire quand $\delta_i = 0$.

- $d_i = 1$ en cas de décès en $Z_{i:n}$, c'est-à-dire quand $\delta_i = 1$.

Si on désigne par $\delta_{[i:n]}$ le concomitant de $Z_{i:n}$ ou les indicateurs de censure (c'est-à-dire : $\delta_{[i:n]} = \delta_j$ si $Z_{i:n} = Z_j$ avec $j = 1, ..., n$), alors :

- $\delta_{[i:n]} = 0$ en cas de censure en $Z_{i:n}$.

- $\delta_{[i:n]} = 1$ en cas de décès en $Z_{i:n.}$

On obtient alors l'estimateur de Kaplan-Meier, pour $t < Z_{n:n}$, est défini par :

$$S_n^{KM}(t) = \overline{F}_n^{KM}(t) := \prod_{Z_{i:n} \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{Z_{i:n} \leq t} \left(1 - \frac{\delta_{[i:n]}}{n_i}\right) = \prod_{Z_{i:n} \leq t} \left(1 - \frac{\delta_{[i:n]}}{n - i + 1}\right). \quad (1.4)$$

où $Z_{1:n} \leq ... \leq Z_{n:n}$ les statistiques d'ordre associées à $Z_1, ..., Z_n$.

Remarque 1.1.4

Il existe d'autres formes pour l'estimateur de Kaplan-Meier :

$$S_n^{KM}(t) = \prod_{Z_{i:n}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} = \prod_{i=1}^{n}\left(1-\frac{\delta_{[i:n]}}{n-i+1}\right)^{\mathbb{1}_{\{Z_{i:n}\leq t\}}}, \qquad i=1,...,n$$

Une écriture sous la forme d'une somme peut être trouvée dans le livre de Reiss et Thomas [24], page 122

$$S_n^{KM}(t) = \sum_{i=1}^{n}\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}\mathbb{1}_{\{Z_{i:n}\leq t\}}.$$

Remarque 1.1.5

Dans le cas où il y a des ex-aequo :

- *Si ces ex-aequo sont tous des morts, la seule diférence tient à ce que d_i n'est plus égal à* 1 *mais au nombre des morts et l'estimateur de Kaplan Meier devient :*

$$S_n^{KM}(t) = \prod_{i=1}^{n}\left(1-\frac{d_i}{n_i}\right) = \prod_{i=1}^{n}\widehat{p}_i$$

- *Si ces ex-aequo sont des deux sortes, on considère que les observations non censurées ont lieu juste avant les censurées.*

Remarque 1.1.6

- *L'estimateur de Kaplan-Meier est une fonction étagé avec des sauts seulement aux observations non-censurées.*
- *La hauteur des sauts de cet estimateur est aléatoire.*
- *Quand toutes les observations sont non-censurées alors on obtient la fonction de répartition empirique.*

Erreur standard de l'estimateur de Kaplan-Meier Une aide essentielle à l'interprétation d'une estimation de toute quantité est la précision de l'estimation, qui se reflète dans

l'erreur standard de l'estimation. Ceci est défini comme étant de la racine carrée de la variance estimée de l'estimateur et il est utilisé dans la construction d'une estimation d'intervalle de confiance pour une quantité d'intérêt.

Dans cette partie, on va donner l'erreur standard de l'estimateur de Kaplan-Meier, parce que ce dernier est le plus important et le plus utilisé pour la fonction de survie. L'erreur standard de $S_n^{KM}(t)$ il est présenté en détail dans la Section 2.2, page 25. dans le livre de collett [6].

L'estimateur de Kaplan-Meier pour la fonction de survie, pour toute valeur de t dans l'intervalle de t_k à t_{k+1}, on peut s'écrire :

$$S_n^{KM}(t) = \prod_{j=1}^{k} \widehat{p}_j, \qquad k = 1, ..., n.$$

où $\widehat{p}_j = (n_j - d_j)/n_j$ est la probabilité estimée qu'un l'individu survit à travers l'intervalle du temps qui commence à t_j, $j = 1, ..., n$. Prendre des logarithmes,

$$\log S_n^{KM}(t) = \sum_{j=1}^{k} \log \widehat{p}_j,$$

et donc la variance de $\log S_n^{KM}(t)$ est donnée par :

$$var\left\{\log S_n^{KM}(t)\right\} = \sum_{j=1}^{k} var\{\log \widehat{p}_j\}. \tag{1.5}$$

Maintenant, le nombre d'individus qui survivent au début de l'intervalle à t_j peut être supposé avoir une distribution binomiale avec les paramètres n_j et p_j, où p_j est la vraie probabilité de survie à travers cet intervalle. le nombre observé qui survit est $n_j - d_j$, et en utilisant le résultat que la variance d'une v.a binomiale avec les paramètres n et p est $np(1-p)$, la

variance de $n_j - d_j$ est donné par :

$$var(n_j - d_j) = n_j p_j (1 - p_j).$$

Puisque $\widehat{p}_j = (n_j - d_j)/n_j$, la variance de $\widehat{p}_j = (n_j - d_j)/n_j^2$, c'est-à-dire $p_j(1-p_j)/n_j$. La variance de $\widehat{p}_j$ peut alors être estimée par

$$\widehat{p}_j(1 - \widehat{p}_j)/n_j \tag{1.6}$$

Pour obtenir la variance de $\log \widehat{p}_j$, on utilise un résultat général pour la variance approximative d'une fonction d'une v.a. Selon ce résultat, la variance d'une fonction $g(X)$ de la v.a X est donnée par :

$$var\{g(X)\} \approx \left\{ \frac{dg(X)}{dX} \right\}^2 var(X). \tag{1.7}$$

C'est ce qu'on appelle l'approximation de la série de Taylor à la variance d'une fonction d'une v.a. En utilisant l'équation (1.7), la variance approximative de $\log \widehat{p}_j$ est $var(\widehat{p}_j)/\widehat{p}_j^2$, et en utilisant l'expression (1.6), la variance estimée approximative de $\log \widehat{p}_j$ est $(1 - \widehat{p}_j)/(n_j \widehat{p}_j)$, qui, par substitution à $\widehat{p}_j$, se réduit à

$$\frac{d_j}{n_j(n_j - d_j)}. \tag{1.8}$$

En suite, à partir de l'équation (1.5)

$$var\left\{\log S_n^{KM}(t)\right\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}, \tag{1.9}$$

et une nouvelle application du résultat de l'équation (1.7) donne

$$var\left\{\log S_n^{KM}(t)\right\} \approx \frac{1}{\left[S_n^{KM}(t)\right]^2} var\left\{S_n^{KM}(t)\right\}.$$

Donc

$$var\left\{S_n^{KM}(t)\right\} \approx \left[S_n^{KM}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}. \tag{1.10}$$

Enfin, l'erreur standard de l'estimateur de Kaplan-Meier est donné par :

$$S_n^{KM}(t) \approx S_n^{KM}(t) \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}}, \qquad \text{pour } t_k \leq t < t_{k+1}. \tag{1.11}$$

Ce résultat est connu sous le nom de formule de Greenwood. S'il n'y a pas du temps de survie censuré, $n_j - d_j = n_{j+1}$ et l'expression (1.8) devient $(n_j - n_{j+1})/n_j n_{j+1}$. Maintenant

$$\sum_{j=1}^{k} \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^{k} \left(\frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \sum_{j=1}^{k} \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

qui peut s'écrire

$$\frac{1 - S_n^{KM}(t)}{n_1 S_n^{KM}(t)}.$$

Puisque $S_n^{KM}(t) = n_{k+1}/n_1$ pour $t_k \leq t < t_{k+1}$, $k = 1, 2, ..., n-1$. En l'absence de censure. Par conséquent, d'après l'équation (1.11), la variance estimée de $S_n^{KM}(t)$ est $S_n^{KM}(t)\left[1 - S_n^{KM}(t)\right]/n_1$. Il s'agit d'une estimation de la variance de la fonction de survie empirique, donnée dans l'équation (1.11), en supposant que le nombre des individus à risque au temps t a une distribution binomiale avec des paramètres n_1, $S(t)$.

Proposition 1.1.1

1. ***Normalité asymptotique*** *: En tout point de continuité de* S, $t_0 \in [0, \tau]$ *et* $S(\tau^-) > 0$,

$$\sqrt{n}\left(\widehat{S}(t_0) - S(t_0)\right) \longrightarrow^{\mathcal{D}}_{n \longrightarrow \infty} \mathcal{N}\left(0, var^2(t_0)\right),$$

avec

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

où $G(t)$ la fonction de survie de la variable Y.

2. ***Intervalle de confiance** : Dans chaque intervalle de temps, l'estimation de la survie est une proportion. On peut donc, sous les conditions du (1), On peut construire, pour un niveau de signification $\alpha \in]0,1[$ fixé, un intervalle d'estimation $IC(\alpha)$ pour $S(t)$ de la maniére suivante*

$$IC(\alpha) := \left[S_n^{KM}(t) \pm Z_{\frac{\alpha}{2}} \sqrt{\widehat{var}(S_n^{KM}(t))} \right],$$

où $Z_{\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ de loi normale standard $N(0,1)$, c'est-à-dire $Z_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

3. *d'aprés [25], cet intervalle ne peut être utilisé quand $S_n^{KM}(t)$, est proche de 0 ou de 1. En effet l'intervalle étant symétrique autour de $S_n^{KM}(t)$, les bornes peuvent dépasser les valeurs 0 ou de 1. On préfère utiliser l'intervalle de confiance de Rothman qui contournent cette difficulté :*

$$IC(\alpha) := \frac{K}{K + \left(Z_{\frac{\alpha}{2}}\right)^2} \left[S_n^{KM}(t) + \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{2K} \pm Z_{\frac{\alpha}{2}} \sqrt{\widehat{var}(S_n^{KM}(t)) + \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{4K^2}} \right],$$

avec

$$K = \frac{S_n^{KM}(t)\left(1 - S_n^{KM}(t)\right)}{\widehat{var}(S_n^{KM}(t))}.$$

CHAPITRE 2

Théorie des valeurs extrême (TVE)

2.1 Modélisation semi-paramétrique

La TVE (Extreme Value Theory (EVT) en anglais) est une vaste théorie dont le but d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'apparition est faible, cette théorie basée sur l'approximation asymptotique des lois des maxima convenablement normalisés des vecteurs aléatoires dont les composantes sont des variables supposées. Plus de détails sont disponibles sur les livres de [1], **[7] et [11].**

2.1.1 Statistique d'ordre

Pour commencer notre étude et les explications de la théorie des valeurs extrêmes, il faut avoir un grand bagage, alors notre point de départ sera les statistiques d'ordre.

Définition 2.1.1 (statistique d'ordre)

Soit $(X_i)_{i\in\mathbb{N}}$ *une suite de v.a's réelles iid définie sur l'espace* $(\Omega, \mathfrak{B})$*, d'une densité commune*

f et d'une fonction de répartition F. La statistique d'ordre de l'échantillon $X_1, ..., X_n$ est le réarrangement croissant de $X_1, ..., X_n$, et que l'on note par :

$$X_{1:n}, ..., X_{n:n}.$$

Pour $1 \leq i \leq n$, la v.a $X_{i:n}$ est appelée la $i^{-iéme}$ statistique d'ordre (ou statistique d'ordre i), et les extrêmes sont comme termes du minimum et du maximum de l'échantillon $X_1, ..., X_n$, tel que :

- *$X_{1:n}$ est la plus petite valeur observée (où statistique du minimum) :*

$$X_{1:n} = \min(X_1, ..., X_n) = \min_{1 \leq i \leq n} X_i.$$

- *$X_{n:n}$ est la plus grand statistique d'ordre (où statistique du maximum) :*

$$X_{n:n} = \max(X_1, ..., X_n) = \max_{1 \leq i \leq n} X_i.$$

Distributions des statistiques d'ordres

Distribution du minimum La distribution du minimum $X_{1:n}$ est donnée par :

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

En effet :

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

La fonction de densité du minimum est :

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

Distribution du maximum La distribution du maximum $X_{n:n}$ est donnée par :

$$F_{X_{n:n}}(x) = \left[F(x)\right]^n.$$

En effet :

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

La fonction de densité du maximum est :

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

Distribution de la $i^{ème}$ statistique d'ordre La distribution de la $i^{ème}$ statistique d'ordre est donnée par :

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

La densité de la $i^{ème}$ statistique d'ordre est :

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

La démonstration de ces formules est détaillée dans l'ouvrage de Reiss et Thomas [24].

Remarque 2.1.1

Les résultats correspondant pour les minima sont facilement accessibles en utilisant l'égalité suivante :

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

2.1.2 Lois des valeurs extrêmes

Les distributions à queues lourdes sont liées à la TVE et permettent de modéliser beaucoup de phénomènes que l'on trouve dans différentes disciplines telles que la finance, l'hydrologie, la climatologie épidémiologie, ...etc. Ce type des distributions est défini ainsi :

Définition 2.1.2 (Distribution à queue lourde)

Soit X une v.a de fdr F, donc cette dernière elle est dite distribution à queue lourde, s'il existe un constant positif γ qui représente l'indice de queue et prend la formule suivante :

$$\overline{F}(x) \sim x^{-1/\gamma} l(x), \qquad \text{pour } x \to \infty,$$

où $l(x)$ la fonction à variation lente au voisinage de l'infini, Ce type de distribution satisfaite pour tout $x > 0$, la condition suivante :

Définition 2.1.3 (Condition du 1^{er} ordre)

$\overline{F}$ est dite variation régulière à l'infini d'indice $-1/\gamma < 0$, on a :

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma}, \qquad x > 0. \tag{2.1}$$

mais, en général cette condition n'est pas suffisante pour étudier les propriétés des estimateurs, en particulier la normaliser asymptotique. Dans ce cas, une condition du second ordre des fonctions à variations régulières est nécessaire en spécifiant le taux de convergence dans

l'Equation Equation 2.1. La définition suivante de cette condition vient de de Haan et Ferreira [7], page 48.

Définition 2.1.4 (Condition du $2^{ème}$ ordre)

$\exists \rho \leq 0$ *et une fonction* $A \to 0$ *et ne change pas le signe au voisinage de l'infini, tel que :*

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma}}{A(t)} = x^{-1/\gamma}\frac{x^{\rho}-1}{\gamma\rho} \tag{2.2}$$

La famille de distribution de queue lourde, elle se caractérise par une décroissance lente vers zéro par rapport à la distribution exponentielle, comme le montre la figure suivante :

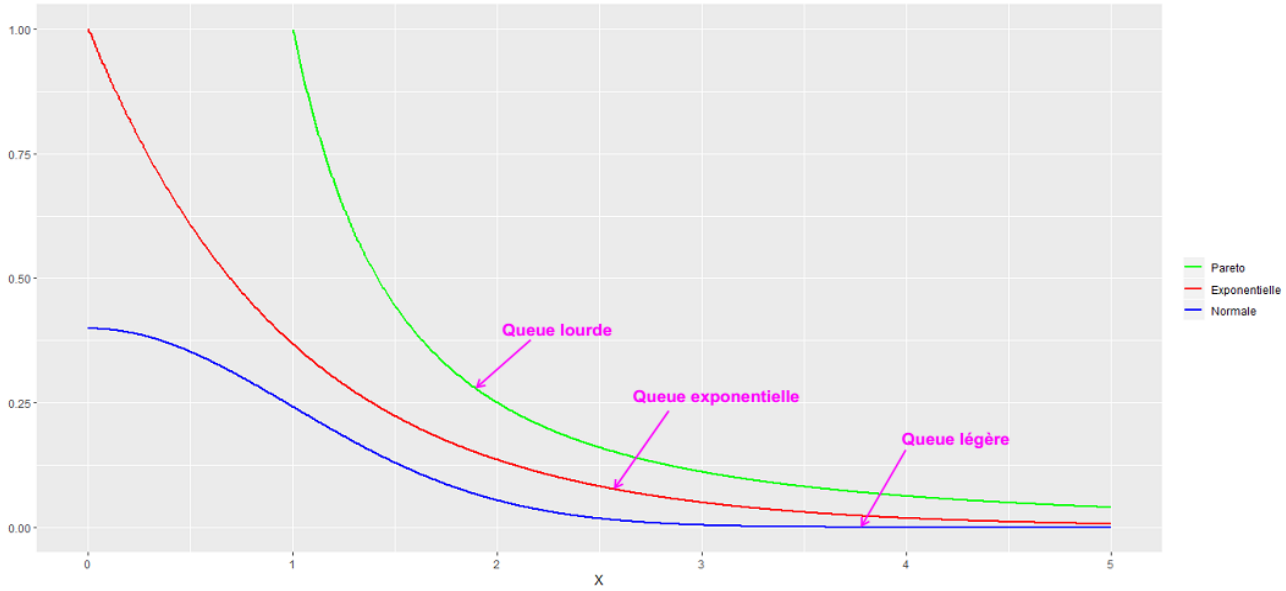

FIG. 2.1 – Comparaison du comportement de la queue.

On a ici les distributions à queue lourdes, elles sont représentées par la courbe verte elle est décroît lentement vers zéro par rapport la distribution exponentielle et la distribution exponentielle que représentée par la courbe rouge et la distribution à queue légère représentée par la courbe bleue.

Fisher et Tippett [12] en 1928, Gnedenko [13] en 1943, ils ont proposées le théorème suivant qui donne une condition nécessaire et suffisante pour l'existence d'une loi limite non-dégénérée pour le maximum.

Théoreme 2.1.1 (Fisher et Tippet(1928), Gnedenko (1943))

Soit $X_1, ..., X_n$, n v.a's iid de fonction de répartition F, S'il existe deux suites de constantes de normalisation avec $a_n > 0$ et $b_n \in \mathbb{R}$ et une loi non-dégénérée de fonction de répartition $\mathcal{H}$ (c'est-à-dire une fonction de répartion qui n'est pas associée à une variable constante presque sûrement), telle que :

$$\lim_{n\to\infty} P\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F_{X_{n:n}}\left(a_n x + b_n\right) = \mathcal{H}(x), \qquad x \in \mathbb{R},$$

alors $\mathcal{H}$ appartient à une des trois types des distributions standard des valeurs extrêmes suivants :

1. *Type I : Gumbel :* $\mathcal{H}_0(x) = \Lambda_0(x) = \exp\left[-\exp\left(-x\right)\right], \qquad x \in \mathbb{R}.$
2. *Type II : Fréchet :* $\mathcal{H}_\alpha(x) = \Phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & x > 0 \end{cases}, \alpha > 0$
3. *Type III : Weibull :* $\mathcal{H}_\alpha(x) = \Psi_\alpha(x) = \begin{cases} \exp\left(-\left(-x\right)^{\alpha}\right) & x \leq 0 \\ 1 & x > 0 \end{cases}, \alpha < 0$

Avec $\Lambda_0(x)$, $\Phi_\alpha(x)$ et $\Psi_\alpha(x)$, où α paramètre de forme, sont les lois limites possibles pour le maximum s'appellent les distributions standard ou traditionnelle des valeurs extrêmes.

Preuve. Une preuve détaillée de ce théorème peut être trouvée dans Embrechts et al. (1997) [11]. ■

Les trois formules précédentes peuvent être combinées en une seule paramétrisation :

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\left(-\left(1+\gamma x\right)^{-1/\gamma}\right), & \text{pour} \quad \gamma \neq 0,\ 1+\gamma x > 0 \\ \exp\left(-\exp\left(-x\right)\right), & \text{pour} \qquad \gamma = 0,\ x \in \mathbb{R}. \end{cases}$$

Où $\mathcal{H}_\gamma$ est une fonction de répartition non-dégénérée de paramètre $\gamma := 1/\alpha$ que l'on appelle indice des valeurs extrêmes (IVE) ou indice de queue.

Cette loi est appelée loi de valeurs extrêmes généralisée (Generalized Extrême Value) que l'on note GEV. La forme la plus générale de la GEV est :

$$\mathcal{H}_{\gamma,\mu,\sigma}(x) := \begin{cases} \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right) & \gamma \neq 0,\ 1+\gamma\frac{x-\mu}{\sigma} > 0 \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) & \gamma = 0,\ x \in \mathbb{R}. \end{cases}$$

Où $\mu \in \mathbb{R}$ et $\sigma > 0$ sont respectivement les paramètres de localisation et de dispersion.

2.1.3 Domaines d'attraction

Définition 2.1.5 (Domaines d'attraction)

On dit qu'une distribution F appartient au domaine d'attraction ($\mathcal{DA}$) de $\mathcal{H}_\gamma$, et on note $F \in \mathcal{DA}(\mathcal{H}_\gamma)$, si la distribution du maximum renormalisée converge vers $\mathcal{H}$. Autrement dit, s'il existe des constantes réelles $a_n > 0$ et $b_n \in \mathbb{R}$, tels que :

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}_\gamma(x).$$

Selon le signe de γ, on distingue trois domaines d'attraction :

- Si $\gamma < 0$, on dit que $F \in \mathcal{DA}(\Psi_\alpha)$, et F a un point terminal [1] à droite finie ($x_F < +\infty$). Ce domaine d'attraction est celui des fonctions de survie dont le support est borné supérieurement.
- Si $\gamma = 0$ on dit que $F \in \mathcal{DA}(\Lambda)$, le point terminal x_F peut alors être fini ou non. Ce domaine d'attraction est celui des distributions à queues légères, c'est-à-dire qui ont une fonction de survie à décroissance exponentielle.

[1] On appelle le point terminal ou le point le plus à droite de la fonction de distribution F, noté x_F la borne supérieure du support de F définit par : $x_F = \sup\{x \in \mathbb{R},\ F(x) < 1\} \leq \infty$

– Si $\gamma > 0$ on dit que $F \in \mathcal{DA}(\Phi_\alpha)$, et F a un point terminal à droite infinie ($x_F = +\infty$). Ce domaine d'attraction est celui des distributions à queues lourdes, c'est-à-dire qui ont une fonction de survie à décroissance polynomiale.

Voici quelques lois et leurs domaines d'attraction :

Domaines d'attraction	Fréchet($\gamma > 0$)	Gumbel($\gamma = 0$)	Weibull($\gamma < 0$)
Lois	Burr	Gamma	Uniforme
	Student	Normale	Beta
	Pareto	Exponentielle	Reverse Burr
	Log gamma	Lognormale	
	Chi-deux	Weibull	
	Cauchy	Logistique	

TAB. 2.1 – Domaines d'attraction des lois usuelles

Exemple 2.1.1 (loi exponentielle)

La fdr de la loi exponentielle de paramètre $\lambda > 0$ est :

$$F(x) = 1 - \exp(-\lambda x), \qquad 0 \leq x \leq \infty,$$

on pose $a_n = \frac{1}{\lambda}$ et $b_n = \frac{1}{\lambda}\ln(n)$ alors :

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{X_{n:n} - \frac{1}{\lambda}\ln(n)}{\frac{1}{\lambda}} \leq x\right) &= \lim_{n\to\infty} F^n\left(\tfrac{1}{\lambda}x + \tfrac{1}{\lambda}\ln(n)\right) \\ &= \left(1 - \exp\left(-\tfrac{1}{\lambda}\lambda x - \lambda\tfrac{1}{\lambda}\ln(n)\right)\right)^n \\ &= \left(1 - \tfrac{\exp(-x)}{n}\right)^n \to \exp\left(-\exp\left(-x\right)\right) = \Lambda(x).\end{aligned}$$

D'où le maximum normalisé de la loi exponentielle converge vers la loi de Gumbel, et $F(x)$ appartient au domaine d'attraction de Gumbel.

2.1.4 Estimation de l'IVE

Estimation de l'IVE sans censure

Dans la littérature de la TVE, il existe plusieurs méthodes et téchniques pour l'estimation de l'IVE, dans cette partie on reste limiter à trois méthodes, l'estimateur de Pickands **[23]**, l'estimateur de Hill [16] et l'estimateur des moments (**Dekkers et al.,** [9]). Ces estimateurs sont basées fortement sur les plus grandes statistiques d'ordre $X_{n-k:n} \leq ... \leq X_{n:n}$, où la statistique $X_{n-k:n}$ est alors dite statistique d'ordre intermédiaire, où k est donné par la définition suivante :

Définition 2.1.6 (Nombre k de statistiques d'ordre)

Soit $X_1, ..., X_n$ de v.a's iid et $X_{1:n} \leq ... \leq X_{n:n}$ les statistiques d'ordre associées. $k = k_n$ une suite d'entier satisfaisant :

$$1 < k < n, \quad k \longrightarrow \infty \quad et \quad \frac{k}{n} \longrightarrow 0 \ quand \ n \longrightarrow \infty. \tag{2.3}$$

Estimateur de Pickands Cet estimateur a été introduit en 1975 par James Pickands [23], pour toute $\gamma \in \mathbb{R}$.

Définition 2.1.7 (Estimateur de Pickands)

Soit $X_1, ..., X_n$ de v.a's iid de fdr $F \in \mathcal{DA}(\mathcal{H}_\gamma)$, l'estimateur de Pickands est défini par :

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

L'auteur a démontré la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par **Dekkers et al.,** [9] et de Haan et Ferreira [7].

Estimateur de Hill Cet estimateur a été défini par Hill [16] en 1975. C'est un estimateur très populaire et beaucoup utilisé qui est défini seulement pour les valeurs positives de γ, qui correspond aux distributions appartenant au domaine d'attraction de Fréchet.

Définition 2.1.8 (Estimateur de Hill)

Soit $X_1, ..., X_n$ de v.a's iid de fdr $F \in \mathcal{DA}\left(\oplus_{1/\gamma}\right)$, l'estimateur de Hill est défini par :

$$\widehat{\gamma}^H = \widehat{\gamma}^H(k) := \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n}. \tag{2.4}$$

Un grand nombre de travaux théoriques ont été consacrés à l'étude des propriétés de l'estimateur de Hill. Mason [19] en 1982 *a démontré La consistance faible, La consistance forte fut établie en* 1988 *par Deheuvels et al* ***[8].***

Estimateur des moments Cet estimateur a été défini par **Dekkers et al.,** [9] en 1989. C'est une généralisation directe de l'estimateur de Hill. il est défini pour toutes $\gamma \in \mathbb{R}$.

Définition 2.1.9 (Estimateur des moments)

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

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

avec :

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

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

Les propriétés asymptotiques de cet théorème ont été étudiées par **Dekkers et al. [9].**

Estimation de l'IVE avec censure

Dans cette partie on va parler sur l'estimation de l'IVE en présence de censure de type III. Ce problème est très récent dans la littérature, les premiers qui ont mentionné ce sujet sont Beirlant et al**.** [**4**], 1996 et Reiss et Thomas [24], 1997, mais sans résultats asymptotiques. Certains estimateurs des paramètres de queue ont été proposés par Beirlant et Guillou [2], 2001 mais pour les données tronquées et étendues à la censure aléatoire par Beirlant et al. [3], 2007 et l'année suivante par Einmahl et al. [10], 2008.

En réalité, l'estimation des valeurs extrêmes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $X_1, ..., X_n$, (les durées réelles de vie) n'est pas observé, mais qu'il est censuré par un deuxième échantillon $Y_1, ..., Y_n$, qui est supposé être indépendant du premier, où les X_i et Y_i sont des v.a's *iid* de lois F et G respectivement. Toutefois, il convient de signaler que, les différents estimateurs proposés de l'indice des valeurs extrêmes en prenant en considération la présence des censures ont été tous construits de la même manière. Alors l'estimateur d'Hill adapté est basé sur un estimateur standard de l'indice de queue divisé par la proportion de données non censurées dans les plus grands k v.a's $Z_1, ..., Z_n$.

$$\widehat{\gamma}_1^{(c,.)}(k) := \frac{\widehat{\gamma}^{(.)}(k)}{\widehat{p}(k)},$$

où $\widehat{p} = \widehat{p}(k) := \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]}$, $\delta_{[j:n]}$ est le concomitant de la $i^{ème}$ statistique d'ordre, c'est-à-dire, $\delta_{[j:n]} = \delta_i$ si $Z_{[i:n]} = Z_i$, $1 \leq i \leq n$. $\widehat{\gamma}^{(.)}$ peut être n'importe quel estimateur pas adapté à la censure, en particulier l'estimateur de Hill $\widehat{\gamma}^H$, moment $\widehat{\gamma}^M$,...

Dans les années récentes, le problème de l'étude des phénomènes extrêmes et de l'estimation de l'IVE pour des données censurées a attiré l'attention d'un nombre croissant des chercheurs, en raison des nombreuses applications qui appellent des solutions concrètes. Pour plus de detail sur ce sujet on peux réfère aux **[32],** [[21],[22]], **[30],** [5] et [**?**].

2.1.5 Estimation semi-paramétrique

Dans cette partie on va parler sur l'estimation de la fonction de survie dans les deux cas, le premier cas pour les données complètes et en présence les extrêmes et le deuxième cas pour les données censurées et en présence les extrêmes.

Soit $X_1, ..., X_n$ une suite de v.a's iid de fdr commune F, on suppose que $F \in \mathcal{DA}\left(\Phi_{1/\gamma}\right)$ (F est à queue lourde). L'estimateur de $\overline{F}$ est donné par :

$$\widehat{\overline{F}}(x) = 1 - \widehat{F}(x) = 1 - F(\widehat{x}_p), \qquad x \to \infty,$$

où $\widehat{x}_p$ est un estimateur du quantile extrême $x_p := F^{-1}(1-p)$, $p \to 0$. Cet estimateur est donnée pour l'estimateur de Hill ($\gamma > 0$) par la formule suivante :

$$\widehat{x}_p^H := X_{n-k:n} \left(\frac{k}{np}\right)^{\widehat{\gamma}^H}, \tag{2.5}$$

Par conséquent

$$\widehat{\overline{F}}(x) = p, \qquad x \to \infty.$$

Et comme

$$\widehat{x}_p^H := X_{n-k:n} \left(\frac{k}{np}\right)^{\widehat{\gamma}^H}. \tag{2.6}$$

En exposant (2.6) par $1/\widehat{\gamma}^H$, on trouve

$$\widehat{x}_p^{1/\widehat{\gamma}^H} = \left(X_{n-k:n}\right)^{1/\widehat{\gamma}^H} \left(k/np\right), \qquad p \to 0,$$

Ce qui nous donne :

$$p = \frac{k}{n} \left(X_{n-k:n}\right)^{1/\widehat{\gamma}^H} \widehat{x}_p^{-1/\widehat{\gamma}^H}, \qquad p \to 0,$$

alors

$$p = \frac{k}{n} \left(X_{n-k:n}\right)^{1/\widehat{\gamma}^H} x^{-1/\widehat{\gamma}^H}, \qquad x \to \infty.$$

Finalement, on déduit l'estimateur de la queue de distribution pour les distributions à queue lourde est :

$$\widehat{\overline{F}}_n(x) = \frac{k}{n} \left(X_{n-k:n}\right)^{1/\widehat{\gamma}^H} x^{-1/\widehat{\gamma}^H}, \qquad x \to \infty.$$

Dans le contexte d'observations censurées à droite, l'estimateur du type Weissman (voir [31]) pour la queue de distribution $\overline{F}$ comme suit :

$$\overline{F}_n^C(x) = \left(\frac{x}{Z_{n-k:n}}\right)^{-1/\widehat{\gamma}_1^{(H,c)}} \overline{F}_n^{KM}\left(Z_{n-k:n}\right),$$

où $\overline{F}_n^{KM}$ l'estimateurde Kaplan and Meier qui défini par (1.4) et $\overline{F}_n^{KM}$ au point $Z_{n-k:n}$ donne comme suit

$$\overline{F}_n^{KM}(Z_{n-k:n}) = \prod_{i=1}^{n-k} \left(1 - \frac{\delta_{[i:n]}}{n-i+1}\right).$$

Ainsi, l'estimateur de queue de distribution est de la forme :

$$\overline{F}_n^C(x) := \left(\frac{x}{Z_{n-k:n}}\right)^{-1/\widehat{\gamma}_1^{(H,c)}} \prod_{i=1}^{n-k} \left(1 - \frac{\delta_{[i:n]}}{n-i+1}\right).$$

CHAPITRE 3

Exemple d'application

Dans ce chapitre, on applique quelques unes des méthodes vues aux Chapter 1 sur un exemple de données réelles. Les résultats numériques et les représentations graphiques sont obtenus à l'aide du package "KMsurv", "survminer", "tidyverse" et "ggplot2" du logiciel d'analyse statistique R.

3.1 Présentation des données

Un sous-ensemble des données de l'exemple suivant, présentées dans le Table 3.1, sont prises du package "KMsurv" du logiciel R, plus précisément ces données ont été sélectionnés dans le registre du cancer des hôpitaux de l'université d'état de l'Ohio Sedkmak et al. [28]. Ces données, très célèbres dans l'illustration des outils de l'analyse de survie, peuvent être trouvées dans plusieurs documents comme par exemple, le livre de Klein and Moeschberger **[18],** la page 7. Elles concernent les résultats d'un essai de voir les temps d'avant la mort sur les patientes atteintes d'un cancer du sein. 45 femelles atteintes d'un cancer du sein avec

ganglions lymphatiques axillaires négatifs, des parties identiques de ganglions lymphatiques négatifs ont été examinées séquentiellement par microscopie à lumière standard (SLM) et immunohistochimique (IH). Les temps de survie (en mois) pour les deux groupes de patientes sont indiqués au Table 3.1. Le mois au cours de laquelle de chaque patiente est mort a été enregistré et les patientes qui ont survécus jusqu'aux 189 mois (fin de l'étude) ont été enregistrés comme étant censurés (pour eux, Indicateur de décès (0 = vivant, 1 = mort)). Ces patientes contribuent donc à des temps de survie censurés à droite.

ID	**Temps**	**Décès**	**Im**
20	130	0	1
21	133	0	1
22	134	0	1
23	136	0	1
24	141	0	1
25	143	0	1
26	148	0	1
27	151	0	1
28	152	0	1
29	153	0	1
30	154	0	1
31	156	0	1
32	162	0	1
33	164	0	1
34	165	0	1
35	182	0	1
36	189	0	1
37	22	1	2
38	23	1	2
39	38	1	2
40	42	1	2
41	73	1	2
42	77	1	2
43	89	1	2
44	115	1	2
45	144	0	2

TAB. 3.1 – Temps d'avant décès (en mois) chez les patientes atteintes d'un cancer du sein ayant des réponses immunohistochimiques différentes. Seules les données des 20 derniers des patientes sont présentées.

3.2 Statistiques descriptives

Ici, on va utiliser l'approche descriptive pour décrire les données d'étude à l'aide de tableaux ou de représentations graphiques qui décrivent les caractéristiques générales des données. Cette approche, elle représente une étape nécessaire et constitue un premier pas dans l'analyse et l'interprétation des données. Pour ce là, les statistiques descriptives élémentaires des données sont résumées dans le Table 3.2.

	Temps	**im**
Minimum	19.00	1 : 36
1er Q_1	51.00	2 : 9
Médiane	89.00	
Moyenne	98.33	
Écart-type	51.84	
3ème Q_3	144.00	
Maximum	189.00	

TAB. 3.2 – Résumé statistiques descriptives élémentaires des patientes atteintes d'un cancer du sein.

On constate que le plus petit temps d'avant décès $Min = 19$, le plus grand temps est $Max = 189$. On sait qu'il y a 45 patientes et on note à partir le tableau au-dessus il y a 9 patientes étaient immunohistochimique positifs, et les 36 autres patientes étaient négatifs, les méthodes de détection IH peuvent être un complément important dans la stadification des patientes atteintes de cancer du sein. On a $l'ecart - type = 51.84$, il est fort, donc les valeurs de temps ne sont pas dispersées autour de la moyenne cela signifie que les temps des décès hétérogène.

3.3 Statistique inférentielle

Ici, on va utiliser l'approche inférentielle (inductive), elle a pour objectif de mettre en place des règles de décision afin de réaliser de l'inférence statistique, qui réside dans l'étudier de

l'analyse de survie et son utilisation pour trouver l'estimation de la fonction de survie

La relation de l'estimateur (1.4) permet de calculer l'estimation de la probabilité de survie par la méthode de KM au Table 3.1. Les résultats obtenus sont donnés dans le Table 3.3, où la $1^{ère}$ colonne contient les 24 durées réellement observées, parmi les 45 patientes qui sont analysés. La $2^{ème}$ colonne indique le nombre des patientes n_i à risque sur l'intervalle de temps écoulé. Le nombre d'évènements observé d_i est indiqué dans la $3^{ème}$ colonne. Dans la $4^{ème}$ et $5^{ème}$ colonne on donne l'estimation de KM pour la fonction de survie $\mathbf{S}_n^{KM}(t)$ et l'erreur standard (err.std**),** où la dernière colonne contient, pour chaque instant t_i (Temps), l'intervalle d'estimation à 95% de confiance (voir la partie 1.1.5). La représentation graphique associée au Table 3.3 est présentée dans la Figure 3.1, où la courbe de survie estimée pour les 45 données de patientes en escaliers décroissants (avec les bornes de confiance à 95%), sur lequel on constate que la plus grande des valeurs non censurées pour estimation $\mathbf{S}_n^{KM}(115) = 0.467 \neq 0$. Ceci témoigne de l'existence de données censurées au delà de 115. Le Table 3.4 contient la médiane et l'IC estimées associée au Figure 3.1.

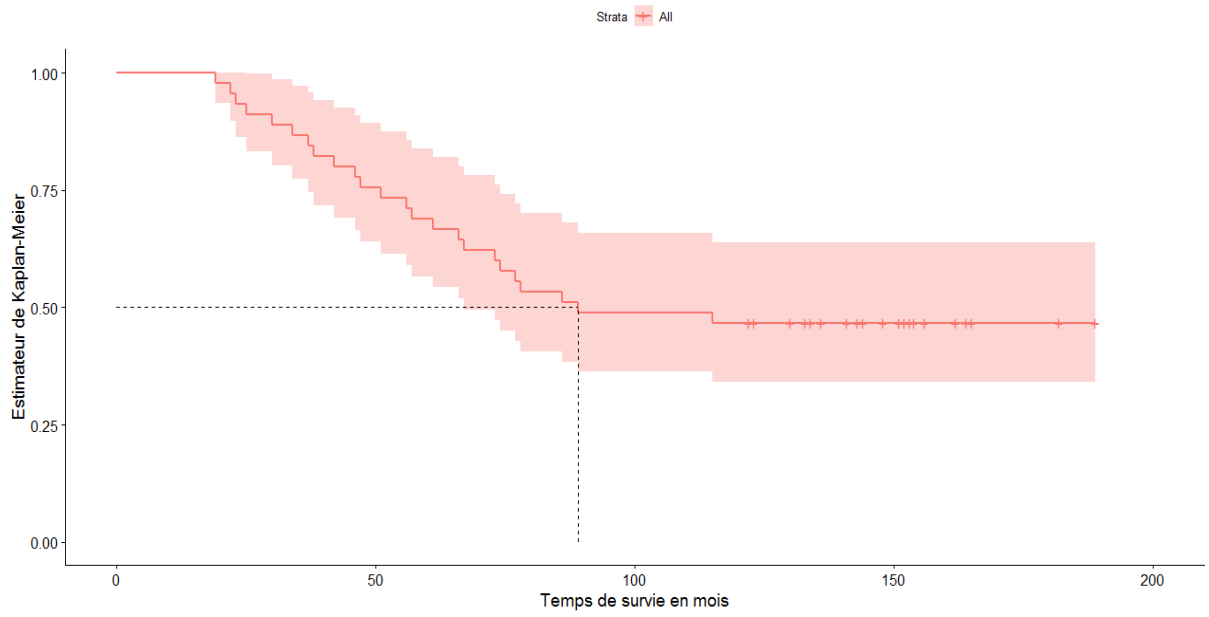

FIG. 3.1 – Courbe de survie estimée pour les 45 données de patientes.

Temps mois ($\mathbf{t}_i$)	**n.risque** $\mathbf{n}_i$	**n.évén.** $\mathbf{d}_i$	**pro.survie** $\mathbf{S}_n^{KM}(t)$	**err.std.**	IC (95%)
19	45	1	0.978	0.0220	0.936 – 1.000
22	44	1	0.956	0.0307	0.897 – 1.000
23	43	1	0.933	0.0372	0.863 – 1.000
25	42	1	0.911	0.0424	0.832 – 0.998
30	41	1	0.889	0.0468	0.802 – 0.986
34	40	1	0.867	0.0507	0.773 – 0.972
37	39	1	0.844	0.0540	0.745 – 0.957
38	38	1	0.822	0.0570	0.718 – 0.942
42	37	1	0.800	0.0596	0.691 – 0.926
46	36	1	0.778	0.0620	0.665 – 0.909
47	35	1	0.756	0.0641	0.640 – 0.892
51	34	1	0.733	0.0659	0.615 – 0.875
56	33	1	0.711	0.0676	0.590 – 0.857
57	32	1	0.689	0.0690	0.566 – 0.838
61	31	1	0.667	0.0703	0.542 – 0.820
66	30	1	0.644	0.0714	0.519 – 0.801
67	29	1	0.622	0.0723	0.496 – 0.781
73	28	1	0.600	0.0730	0.473 – 0.762
74	27	1	0.578	0.0736	0.450 – 0.742
77	26	1	0.556	0.0741	0.428 – 0.721
78	25	1	0.533	0.0744	0.406 – 0.701
86	24	1	0.511	0.0745	0.384 – 0.680
89	23	1	0.489	0.0745	0.363 – 0.659
115	22	1	0.467	0.0744	0.341 – 0.638

TAB. 3.3 – Estimateur de Kaplan-Meier de fonction de survie (en mois) pour les 45 patientes atteintes d'un cancer du sein.

records	**n.max**	**n.start**	**events**	**rmean**	**se(rmean)**	**median**	IC (95%)
45.00	45.00	45.00	24.00	117.38	10.33	89.00	$67.00 - NA$

TAB. 3.4 – La médiane et l'intervalle de confiance estimées par KM

Conclusion

L'analyse de survie est un domaine actif de recherche, il utilise des variables de durées qui sont dans la plupart des cas incomplets (ou bien qui ne sont pas totalement observées : censurées ou tronquées), dans ce mémoire on a étudié en cas de censure

Grâce à la fonction de survie on peut être estimée la médiane du temps de survie pour les patients actuels ou futurs, où les résultats estimés que ce soit l'estimation non-paramétrique ou semi-paramétrique sont particulièrement utiles pour concevoir un système de traitement ou conseiller le patient lors du diagnostic.

Il arrive assez souvent qu'une série statistique, ayant des données extrêmes, présente des données tronquées. Il serait intéressant de faire le même travail, par ce que ce cas mérite d'être considéré attentivement vu.

BIBLIOGRAPHIE

[1] Beirlant, J, Goegebeur, Y, Segers, J, & Teugels, J. (2006). Statistics of Extrêmes : Theory and Applications. John Wiley.

[2] Beirlant, J., & Guillou, A. (2001). Pareto index estimation under moderate right censoring. Scand. Actuar. J., 111 – 125.

[3] Beirlant, J., & Guillou, A. Dierckx, G., & Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. Extremes, 10(3), 151 – 174.

[4] Beirlant, J., Teugels, J. L., & Vynckier, P. (1996). Practical analysis of extreme values. Leuven University Press.

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

[6] Collett, D. (2015). Modelling survival data in medical research. Third Edition. CRC press. aylor & Francis Group, A chapman & hall book.

[7] de Haan, L. & Ferreira, A. (2006). Extreme Value Theory : An Introduction. Springer-Verlag, New York.

[8] Deheuvels, P., Häeusler, E.,& Mason, D. M. (1988). Almost sure convergence of the Hill estimator.Mathematical Proceedings of the Cambridge Philosophical Society, 104(02), $371 - 381$.

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

[10] Einmahl, J. H., Fils-Villetard, A., & Guillou, A. (2008). Statistics of extremes under random censoring. Bernoulli, 14(1), $207 - 227$.

[11] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance, Springer-Verlag, Berlin.

[12] Fisher R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Cambridge Philos. Soc., 24(02), $180 - 190$.

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

[14] Göran. B., (2012). Event history analysis with R. (Chapman & Hall/CRC The R Series). Taylor & francis group, LLC.

[15] Harrouche L., (2018). Analyse statistique des modèles de survie. Mémoire de master. Université Mouloud Mammeri de Tizi-Ouzou.

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

[17] Kaplan, E. L. & Meier, P., (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc., 53(282), $457 - 481$.

[18] Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis : techniques for censored and truncated data (Vol. 2, pp. 3-5). New York : Springer.

[19] Mason, D. M. (1982). Laws of large numbers for sums of extreme values. Ann. Probab., $754 - 764$.

[20] Michaël G., (2015), Introduction à l'analyse de survie. Université de Lille 2, EA 2694 - Santé Publique : Epidémiologie et Qualité des soins.

[21] Ndao, P., Diop. A., & Dupuy, J. F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. Comput. Statist. Data Anal., 79, $63 - 79$.

[22] Ndao, P., Diop. A., & Dupuy, J. F. (2016). Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. J. Statist. Plann. Inference, 168, $20 - 37$.

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

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

[25] Saint Pierre, P. (2015). Introduction à l'analyse des durées de survie. Université Pierre et Marie Curie, France.

[26] Samartzis, L., $(2005 - 2006)$. Survival and censored data. École polytechnique fédérale de Lausanne.

[27] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[28] Sedmak, D. D., Meineke, T. A., Knechtges, D. S., & Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metastases. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc, 2(5), $516 - 520$.

[29] Semmari M., (2014). Sur l'analyse de survie et applications. Mémoire de master. Université Mohamed Khider, Biskra.

[30] Stupfler, G. (2016). Estimating the conditional extreme-value index under random right-censoring. J. Multivariate Anal., 144, $1 - 24$.

[31] Weissman, I., (1978). Estimation of parameters and large quantiles based on the k largest observations. J. Amer. Statist. Assoc. 73, $812 - 815$.

[32] Worms, J. & Worms, R., (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes, $17, 337 - 358$.

[33] Zakaria Raiti, (2017). Modélisation de la durée de maintien en arrêt de travail d'une population de travailleurs non-salariés. Mémoire de Statisticien Mention Actuariat et l'admission à l'Institut des Actuaires. Institut des Actuaires, ISUP.

[34] Zouadi, N., & Saidi, G., (2018). Estimation de l'indice des valeurs extrêmes en présence des données censurées Étude de cas : Les durées de chômage en Algérie. Revue des Sciences Economiques, de Gestion et Sciences Commerciales. $520 - 535$.

Annexe B : Abréviations et Notations

Les différentes abréVariations et notation utulisées tout au long de cette thèse sont expliquées ci-dessous.

$v.a$	:	Variable aléatoire.
F	:	Fonction de répartition.
F_n	:	Fonction de répartition empirique.
F^{-1}	:	Inverse généralisé de F.
S ou $\overline{F}$	:	Fonction de queue ou Fonction de survie.
S_n	:	Fonction de survie empirique.
S_n^{KM}	:	Estimateur de Kaplan-Meier.
Q et Q_n	:	Fonctions du quantile et quantile empirique
$X \Lambda Y$	:	$\min(X, Y)$.
iid	:	Indépendantes et identiquement distribué.
$\mathbb{1}_A$	:	Fonction indicatrice de l'ensemble A.
$:=$	:	Egalité par définition.
$\mathcal{DA}(.)$	:	Domaine d'attraction.
$l(t)$	:	Fonction à variation lente.

TEV	:	Theorie des valeurs extremes.
IVE	:	Indice des valeurs extrêmes.
$N(0,1)$	:	Loi normale standard.
x_F	:	Point terminal.
$\widehat{\gamma}^P$	:	Estimateur de Pickands.
$\widehat{\gamma}^H$	:	Estimateur de Hill.
$\widehat{\gamma}^M$	:	Estimateur des Moments.
$\mathbb{R}$	:	Ensemble des valeurs réelles.
SLM	:	microscopie à lumière standard
IH	:	immunohistochimique
Im	:	Immunohistochimique

مُلَخَّص

يعطينا تحليل البقاء على قيد الحياة إحصاءات مهمة من خلال نمذجة الظواهر في العديد من المجالات مثل: البيولوجيا الطبية والموثوقية لتحسين نتائجها. الهدف النهائي من هذه الأطروحة هو تقدير وظيفة البقاء على قيد الحياة تحت رقابة عشوائية للبيانات من اليمين. ندرس في الفصل الأول التقدير اللامعلمي الذي يمثله عادة مقدر كابلان ميار ، في الفصل الثاني سنتحدث عن التقدير شبه المعلمي وسيكون اهتمامنا الرئيسي هو توسيع نتائج نظرية القيم المتطرفة في حالة وجود رقابة على البيانات. في الفصل الأخير، نطبق ما سبق على مجموعة حقيقية من البيانات: مرضى سرطان الثدي.

الكلمات المفتاحية: وظيفة البقاء، تحليل البقاء، الرقابة العشوائية، مقدر كابلان-ميار، نظرية القيم المتطرفة، مؤشر القيم المتطرفة، التقدير اللامعلمي، التقدير شبه المعلمي.

Résumé

L'analyse de survie on donne des statistiques importantes en modélisant des phénomènes dans plusieurs domaines comme : la biologie médicale et la fiabilité pour améliorer leurs résultats. L'objectif final étant de ce mémoire est l'estimation de la fonction de survie sous données censurées aléatoirement à droite. On étudie dans la première partie l'estimateur non-paramétrique qui est généralement représentée par l'estimateur de Kaplan-Meier. Dans la deuxième partie, on parle de l'estimateur semi-paramétrique et notre principale préoccupation sera étendu les résultats de la théorie des valeurs extrêmes en cas de présence de censure des données. Dans la dernière partie, on applique ce qui précède sur un ensemble de données réelles : les patientes atteintes d'un cancer du sein.

Les mots clés : Fonction de survie, Analyse de survie, censure aléatoire, Estimateur de Kaplan-Meier, TVE, Indice des valeurs extrêmes, Estimation non-paramétrique, Estimation semi-paramétrique.

Abstract

Survival analysis gives us important statistics by modeling phenomena in several areas such as: medical biology and reliability to improve their results. The final objective of this thesis is the estimation of the survival function under randomly censored data on the right. We study in the first part of the non-parametric estimator which is usually represented by: Kaplan-Meier estimator. In the second part, we will talk about the semi-parametric estimator and our main concern will be to expand the results of the theory of extreme values in case of data censorship. In the last part, we apply the above one real data set: breast cancer patients.

Key words: Survival function, Survival analysis, Random censuring, Kaplan-Meier estimator, EVT, Extreme value index, Non-parametric estimation, Semi-parametric estimation.