République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : statistique

Par

BRIKEL Nasreddine

Titre :

Estimation non paramétrique pour des données doublement censurées

Membres du jury de soutenance :

Pr.	YAHIA Djabrane	UMKB	Président
Pr.	BENATIA Fateh	UMKB	Encadreur
Dr.	CHINE Amel	UMKB	Examinatrice

19 Juin 2023

Dédicace

Je dédie ce travail à :

- Maman et papa

- A mes frères

et mes sæurs, et toute ma famille

- A mes chers amis

- Je tiens à remercier

Mon encadreur : Dr. Benatia Fatah

- Tous les membres de ma promotion

et tous mes professeurs

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

Brikel Nasreddine

REMERCIEMENTS

Gloire soit rendue à **Allah** tout puissant, qui m'a donné force et patience d'accomplir mon mémoire de fin d'étude.Mes sincères remerciements à mon encadreur **M.BENATIA Fateh** pour son soutien et son attention exceptionnels durant ces derniers mois. Je tiens à lui témoigner ma gratitude pour le choix du thème. Et pour son suivi attentif et pertinent qui a mené à l'acheminement de ce travail.Je remercie également tous les membres de jury, **M. YAHIA Djabrane** et **M.CHINE Amel** pour avoir accepté d'évaluer mon travail. Mes vifs remerciements à tout le corps professoral qui a contribué à l'acheminement de cette formation pluridisciplinaire, avec dévouement et grande patience Enfin, à toute personne ayant contribué, de prés ou de loin, à l'élaboration de ce mémoire de fin d'études. Veuillez bien trouver ici l'expression de mes sincères remerciements.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les statistiques non paramétriques sont un ensemble de méthodes alternatives qui sont utilisées dans les cas où les hypothéses sur la population statistique à partir de laquelle l'échantillon est tiré ne sont pas remplies (par exemple,modèle exponentiel) qui reléve des statistiques paramétriques.

Pratiquement, il arrive qu'un phénomène de censure empêche l'observation complète de la variable d'intérêt. Par exemple quand on s'intéresse au temps de survie à une maladie grave, la fixation du temps de l'étude va introduire une censure à droite. En effet, à la fin de l'étude, il est possible que certains malades soient encore vivants (heureusement pour eux.). Mais le statisticien ne disposera que de l'information partielle que leurs temps de survie dépassent les valeurs observées.

On rencontre les données censurées dans différent domaines de recherche, en médecine, en biologie, en économie, en fiabilité,. . . Des observations sont dites censurées lorsque la variables étudiée représente la durée à un évènement terminal ; et que l'étude est limitée dans le temps. Cette variable est dite de survie. elle est dite censurée si elle n'est pas intégralement observée. On s'intéresse dans ce mémoire à l'estimation non paramétrique pour des données doublement censurées. En fixant l'étude sur les estimateurs des fonctions de survie, de densité et du taux de hasard. et cela dans les chapitres suivants :

chapitre 1 : Ce premier chapitre est consacré aux rappels sur quelques notions de base sur la censure et ses différents types, ainsi que quelques définitions que nous

utiliserons dans la suite.

chapitre 2 : nous présentons dans ce deuxième chapitre le modèle de censure à droite ainsi que les estimateurs de la fonction de servie, de la densité et du taux de hasard pour ce modèle. Nous rappelons également quelques résultats de convergence de ces estimateurs.

Ensuite, nous étudions de simulation avec des données numériques sur l'estimateur de Kaplan-Meier de la fonction de survie.

chapitre 3 : ce dernier chapitre est consacré en premier lieu aux modèles de censure double. Nous introduisons les estimateurs d'intérêt (des fonctions de répartition, de densité et de hasard) pour tous ces modèles et nous rappelons quelques résultats concernant leurs propriétés asymptotiques est présenté pour clôturer ce mémoire.

Chapitre 1

Généralités sur la censure

Comme préliminaire du chapitres suivants nous présentons dans ce chapitre un rappel des notions essentielles sur la censure. et ses propriétés.

1.1 La notion de censure

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données incomplètes proviennent du fait qu'on n'a pas accès à toute l'information. Au lieu d'observer des réalisations $i.i.d$ de durée X, on observe la réalisation de la variable X soumise à diverses perturbations indépendantes ou non de l'événement étudié.

Définition 1.1.1 [12]*La variable de censure D est définie par la non-observation de l'événement étudié. Si au lieu d'obsever X, on observe D, et que l'on sait que si $X > D$ alors c'est une censure à droite, et si $X < D$ alors c'est une censure à gauche, le cas $D_1 < X < D_2$est une censure par intervalle.*

Pour un individu donné j, on va considérer :

►Temps de survie X_j.

►Son temps de censure D_j.

►La durée réellement observée Z_j.

Définition 1.1.2 (fonction de survie S) [14]

La fonction de survie S représente la probabilité de survivre au moins jusqu'au temps t. Autrementdit, la probabilité de ne pas avoir observé l'événement d'intérêt jusqu'à l'instant t. Elle est définie comme suit :

$$S(t) = \bar{F}(t) = 1 - F(t) = 1 - P(X \leq t) = P(X > t),\ t \geq 0.$$

Remarque 1.1.1 *$S\ (t)$ est une fonction monotone décroissante et continue telle que* $S(0) = 1 et \lim_{t\to\infty} S\ (t) = 0.$

1.2 Types de censures

1.2.1 Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées, pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue. Soit D une variable aléatoire de censure, au lieu d'observer la variable X qui nous intéresse, on observe le couple de variables (Z, δ) avec $Z = \min(X, D)$ et $\delta = I_{\{X \leq D\}}$. δ est appelé indicateur de censure puisque ses valeurs nous informent sur le fait que l'observation est complète (si $\delta = 1$) ou censurée à droite (si $\delta = 0$).

Un exemple illustratif est lorsqu'on s'intéresse à la durée de vie d'un genre de machines précis mais que ces dernières tombent en panne s'il se produit une surtension

d'électricité. Ici la durée de vie de la machine est censurée à droite par l'instant auquel se produit la surtension.

1.2.2 Censure à gauche

Une durée de survie est dite censurée à gauche si l'individu a déjà subi l'événement d'intérêt avant l'entrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple $(Z;\delta)$:

$$Z = X \vee G = \max(X, G) \text{ et } \delta = I_{\{G \geq X\}}.$$

Notons G l'âge au quel une certaine maladie apparait pour la première fois chez un individu. Aprés un examen médical on a reçu deux types de réponses :

1. l'individu a déjà été malade mais l'âge exact de la première apparition n'a pas été retenu : dans ce cas on n'a pas observé G mais on sait que G est inférieur à l'âge de l'individu lors de l'examen X. Il s'agit d'une observation censurée à gauche.
2. l'individu n'a jamais eu de maladie : dans ce cas on sait seulement que G est supérieur l'âge de l'individu, donc on a une observation censurée à droite.

1.2.3 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. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite aprés que l'évènement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou plus rarement à gauche. Un avantage de ce type est qu'il

permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[a_j, b_j]$.

Dans la censure il n'ya pas qu'un seul type de censure mais plusieures qui nous allons présentés comme suite :

censure de type 1 : (fixée)

Soit C un nombre positif fixé. Au lieu d'observer les variables $G_1, G_2, ..., G_n$ qui nous intéressent, on observe G_j que lorsque $G_j \leq C$, si non on sait seulement que $G_j > C$. L'observation est alors $X_i = \min(G_j, C) = G_j \wedge C$. C'est le cas lorsqu'on décide à l'avance que le nombre C est la durée de l'étude.

Dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note G la durée d'apprentissage de cette langue. Pour certains étudiants nous allons observer leurs durées G_j d'apprentissage de la langue par contre pour d'autres leurs G_j ne seront pas observées car le stage est limité dans le temps.

censure de type 2 :

L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant, quant à lui, non aléatoire. Ce modéle est souvent utilisé dans les études de fiabilité.

censure de type 3 : (aléatoire)

A chaque individu i, est associé un couple de v.a (G_j, C_j) positives où G_j est son temps de survie et C_j son temps de censure, tel que seule la plus petite est observée,

c'est-à-dire $X_j = G_j \wedge C_j$

$$\delta_j = I_{\{Xj \leq G_j\}} = I_{\{G_j \leq C_j\}} = \begin{cases} 1 & \text{si non censure.} \\ 0 & \text{si censure.} \end{cases}$$

En pratique la censure aléatoire peut avoir plusieurs causes : par exemple perte de vue, arrêt du traitement ou bien fin de l'étude, Alors ce qu'on observe c'est le couple (X_i, δ_i) et $\delta_j = I_{\{X_j \leq G_j\}}$(l'indicatrice de non censure).

1.2.4 Censure mixte

Nous disons qu'il y a censure mixte lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénoméne d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Comme dans le modèle décrit dans l'article de [Patilea et Rolin (2006)] , au lieu d'observer un échantillon de la variable d'intérêt Y, on observe un échantillon du couple $(Z; A)$ avec $Z = \max(\min(X; D); G)$ et

$$A := \begin{cases} 0 & \text{si } G < X \leq D. \\ 1 & \text{si } G < D < X. \\ 2 & \text{si } \min(X; D) \leq G. \end{cases}$$

A prend la valeur 0 lorsque$Z = X$ (donnée complète) et la valeur 1 lorsque $Z = D$ (donnée censurée à droite) et la valeur 2 lorsque $Z = G$ (donnée censurée à gauche). Un exemple de ce modèle est donné par un système formé par trois composants, dont deux sont placés en série (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en parallèle avec ce système en série. Ici, il est clair qu'il n'est pas raisonnable de supposer que le temps de fonctionnement d'un composant soit inférieur à un autre.

Chapitre 2

Estimation non paramétrique dans un modèle de censure à droite

Le but de ce chapitre est de rappeler les principales méthodes et résultats de l'estimation non paramétrique pour des données censurées à droite. Mais comme ces méthodes sont généralement inspirées de celles connues dans le cas des données complétes, nous commençons d'abord par regarder ce cas.

2.1 Le cas des données complêtes

2.1.1 Estimation de la fonction de répartition.

Définition 2.1.1 (fonction de répartition F) [9]

La fonction de répartition de X, notée F, est définie comme suit :

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

elle désigne la probabilité que l'événement d'intérêt ait lieu avant t.

Soit $X_1, ...X_n$un échantillon issu d'une $v.a.X$ de loi absolument continue et de fonction de répartition $F_X(t)$. L'estimation de cette dernière tient une place importante dans l'étude de nombreux phénomènes de nature aléatoire. Le point de départ de l'estimation non-paramétrique de la fonction de répartition fut l'introduction de la fonction de répartition empirique qui se calcule sur la base de véritables observations de la variable d'intérêt X. C'est une fonction en escalier, limitée à gauche et continue à droite qui met un poids $1/n$ sur chaque point X_j telle que :

$$F_n^c(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_j \leq t\}}.$$

Cette estimation est d'excellente qualité. On se place pour commencer dans le cas où les données sont indépendantes. La loi forte des grands nombres indiqué que cet estimateur est fortement consistant sur tout $\mathbb{R}$. Le théorème de Glivenko-Contelli améliore ce résultat en donnant la convergence uniforme. Chang [1949]introduit la loi du logarithme itéré pour cet estimateur dans $\mathbb{R}$. Puis Kiefer [1961] a précisé le taux de cette convergence dans $\mathbb{R}^m$.

2.1.2 Estimateur à noyau de la densité

Définition 2.1.2 (Fonction de Noyau) [14]*Soit* $K : \mathbb{R} \to \mathbb{R}$.*on dit que* K *est un noyau si et seulement si :*

$$\int K(t)dt = 1$$

▶ K est dit positif si $K(t) \geq 0\ \forall t$.

▶ K est dit symétrique si $K(t) = K(-t)\ \forall t$

Définition 2.1.3 (Estimateur à noyau) [14]*Un estimateur à noyau noté* f_n^c*de la*

fonction f est défini par :

$$f_n^c(t) = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{X_j - t}{h_n}),$$

où$\{h_n\}_{n\geq 1}$est une suite de réels positifs appelés "Paramètres de lissage "ou "largeur de la fenêtre" , qui tend vers 0 quand n tend vers l'infini.

Comme nous allons le voir par la suite, si le noyau K est une fonction de densité alors l'estimateur à noyau f_n^c est lui aussi une fonction de densité. De plus , ce dernier posséde les propriétés de continuité et de différentiabilité. De sorte que si, par exemple, K est la densité normale alors f_n^c possède des dérivées de tout ordre.

Exemples de noyaux

les noyaux les plus utilisées dans l'estimateur à noyaux sont :

1. noyau quadratique $\rightarrow$ $K(t) = \begin{cases} \frac{15}{16}(1-t^2)^2 \text{ si } \{|t|\leq 1\} \\ 0 \text{ sinon} \end{cases}$

2. noyau gaussien $\rightarrow$ $K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2).$

3. noyau triangulaire $\rightarrow$ $K(t) = \begin{cases} (1-|t|) \text{ si } \{|t|\leq 1\} \\ 0 \text{ sinon} \end{cases}$

4. noyau uniforme(rectangulaire) $\rightarrow$ $K(t) = \begin{cases} \frac{1}{2} \text{ si } \{|t|\leq 1\} \\ 0 \text{ sinon} \end{cases}$

5. noyau d'Epanechnicov $\rightarrow$ $K(t) = \begin{cases} \frac{3}{4}(1-t^2) \text{ si } \{|t|\leq 1\} \\ 0 \text{ sinon} \end{cases}$

L'estimateur f_n^c, de la fonction densité a été largement étudié dans la littérature. Rosenblatt(1956) donna dans son article l'erreur quadratique moyenne relative à l'estimation de la densité dans le cas d'observation univariées $(i.i.d)$ et où le noyau uniforme $K = \frac{1}{2}I_{\{|t|\leq 1\}}$. Parzen(1962) généralisa ce résultat en considérant une classe

trés vaste de noyaux et établit aussi la normalité asymptotique, puis le cas multivarié fut traité par Cacoullos(1966).

Rosenblatt(1971) a donné par la suite les conditions de convergence en moyenne quadratique de l'estimateur de la densité. Deheuvels(1974) a, en outre, étudié les convergences ponctuelle et uniforme presque sûre.

Sa convergence uniforme faible et forte a été considérée aussi par plusieurs auteurs comme Schuster(1969), Van Ryzin(1969), Rosenblatt(1971) et Silverman(1978). la loi du logarithme itéré a été établie par Deheuvels(1991).

Rappelons que cette méthode d'estimation (à noyau) se généralise au cas de $\mathbb{R}^p$. Ainsi,si $X_1, ..., X_n$ sont n vecteurs aléatoires $i.i.d.$de $\mathbb{R}^p$, de même densité inconnue, alors on peut l'estimer par $f_n^c(t) = \frac{1}{nh_n^p} \sum_{i=1}^n K(\frac{X_j - t}{h_n}), t \in \mathbb{R}^p$.

2.2 Le cas des données censurées à droite

Pratiquement, il n'est pas toujours possible de disposer d'un échantillon de données complètes. Une variable de censure D peut empêcher l'observation de la vraie variable d'intérêt et ne nous fournit alors qu'une information partielle sur elle comme nous l'avons indiqués dans le chapitre précédent. Il existe plusieurs types de censure dont la censure à droite, qui est l'objet d'étude dans cette section.

2.2.1 L'estimateur de Kaplan-Meier

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

$$\begin{aligned} P(X > t) &= P(X > t', X > t) \\ &= P(X > t \ / \ X > t') \times P(X > t') \\ &= P(X > t \ / \ X > t') \times P(X > t' \ / \ X > t") \times P(X > t"). \end{aligned}$$

En considérant les temps d'évènements (décès et censure) distincts $Z_{(j)}(j = 1, ..., n)$ rangés par ordre croissant, $Z_{(0)} = 0$ on obtient

$$P(X > Z_{(j)}) = \prod_{k=1}^{j} P(X > Z_{(k)} \ / \ X > Z_{(k-1)}).$$

Considérons les notations suivantes :

D_j le nombre d'individus à risque de subir l'événement juste avant le temps $Z_{(j)}$, d_j le nombre de décès en $Z_{(j)}$. Alors la probabilité p_j de mourir dans l'intervalle $\left]Z_{(j-1)}, Z_{(j)}\right]$ sachant que l'on était vivant en $Z_{(j-1)}$, i.e.

$$p_j = P(X \leq Z_{(j)} \setminus X > Z_{(j-1)}),$$

peut être estimée par $p_j = \frac{d_j}{D_j}$. Comme les temps d'événements sont supposés distincts, on a

► $d_j = 0$ en cas de décès en $Z_{(j)}$,i.e.quand $\delta_j = 0$,

► $d_j = 1$ en cas de non décès en $Z_{(j)}$,i.e.quand $\delta_j = 1$,

On obtient alors l'estimateur de Kaplan-Meier :

$$\begin{aligned}\hat{S}_{d_n}(t) = 1 - \hat{F}_n^d(t) &= \prod_{j=1,n\ ,\ Z_{(j)}\leq t}^{n}\left(1 - \frac{\delta_j}{D_j}\right) \\ &= \prod_{j:Z_{(j)}\leq t}^{n}\left(1 - \frac{\delta_j}{n-(j-1)}\right) \\ &= \prod_{j=1}^{n}\left[\frac{n-1}{n-j+1}\right]^{\delta_j}.\end{aligned}$$

L'estimateur $\hat{S}_d(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance. $\hat{S}_d(t)$ est une fonction en escalier décroissante, continue à droite. La figure suivante représente un exemple de l'estimateur Kaplan-Meier de la fonction de survie du temps de la récurrence du cancer du poumon.

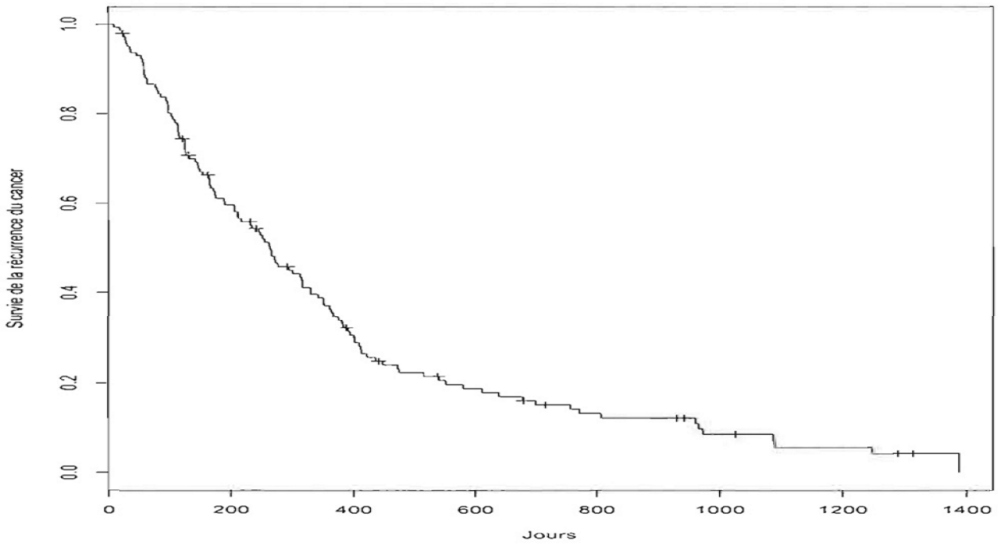

FIG. 2.1 – Courbe de Kaplan-Meier de la fonction de survie $S(t)$

Théorème 2.2.1 *Étant donnée une variable d'intérêt X de fonction de répartition F et une censure D de fonction de répartition R indépendantes, alors :*

$$\sup_{0\leq t\leq X_F}\left|F_n^d(t)-F(t)\right|\overset{p.s}{\to}0.\ quand\ n\to\infty.$$

On note par F (respectivement R) la fonction de répartition de la variable d'intérêt X(respectivement de la variable de censure D), nous pouvons énoncer le résultat suivant :

Théorème 2.2.2 [15]*on suppose que F et R sont continues ,et si le réel x est tel que $R(x)<1$,alors*

$$P(\sup_{-\infty<t\leq x^*}\left|\overline{F^d}_n(t)-\bar{F}(t)\right|=o(\frac{\log\log n}{n}))=1.$$

où $x^*=\min(x,x_{_F})$.

L'estimateur de Kaplan-Meier est asymptotiquement gaussien, précisément on a le résultat suivant :

Théorème 2.2.3 *(Droesbeke et Saporta (2011,* [5]*))*

Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sup_{t\geq1}\left|\hat{S}_{d_n}(t)-S(t)\right|\overset{p.s}{\to}0,$$

et pour tout $t\geq0$,

$$\sqrt{n}(\hat{S}_{d_n}(t)-S(t))\overset{d.}{\to}W_t,$$

où $(W_t)_{t\geq0}$ est un processus gaussien centré qui vérifie pour tous t et s strictement positifs

$$Cov(W_s,W_t)=S(s)S(t)\int_0^{t\wedge s}\frac{dF(u)}{(1-F(u))^2(1-G(u))}.$$

2.2.2 Estimation de la densité.

Définition 2.2.1 (densité de probabilité f) [14]

Si F admet une dérivée au point t, notons cette dérivée f, alors f est appelée densité de probabilité de X, définie sur$[0, +\infty[$ par

$$\begin{aligned} f(t) &= \lim_{dt \to 0} \frac{P(t \leq X \leq t + dt)}{dt} \\ &= \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt}. \end{aligned}$$

Elle désigne que l'évènement d'intérêt ait lieu aprés, dans un petit intervalle de temps $[t \ ; \ t + dt]$.

Remarque 2.2.1 *Puisque X est une variable aléatoire absolument continue, les notations*

$$F(t) = P(X \leq t) \ \textit{et} \ F(t) = P(X < t).$$

sont identiques.

Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau réel, c'est-à-dire une fonction intégrable, d'intégrale 1, on supposera que K est continue, symétrique, à support compact, et à variations bornées. Soit, de plus une suite de paramétres positifs $(h_n)_{n \geq 1}$, dite "fenêtres", et qui vérifie $h_n \to 0$, soit enfin $\tau < \inf \{t \mid P(X > t) = 0\}$.

S'il n'y a pas de censures, l'estimateur à noyau de f au point t est la convolution du noyau K avec la fonction de survie empirique S_{d_n}, i.e.

$$\begin{aligned} \hat{f}_n^d(t) &= \frac{1}{h_n} \int K(\frac{t-u}{h_n}) \hat{F}_n^d dt = -\frac{1}{h_n} \int K(\frac{t-u}{h_n}) \hat{S}_{d_n} dt \\ &= \frac{1}{n h_n} \sum_{j=1}^{n} K(\frac{t - X_i}{h_n}). \end{aligned}$$

En présence de données censurées, l'estimateur empirique naturelle de la fonction de distribution survie est $\widehat{S}_{KM}$ ce qui nous donne :

$$\begin{aligned}\hat{f}_n(t) &= -\int \frac{1}{h_n} K(\frac{t-u}{h_n}) \hat{S}_{KM} dt \\ &= \frac{1}{nh_n} \sum_{j=1}^{n} K(\frac{t-X_{(i)}}{h_n}) \frac{\delta_{(i)}}{n-i+1} \hat{S}_{KM}(-X_{(i)}).\end{aligned}$$

2.2.3 Estimation du taux de hasard.

Définition 2.2.2 (taux de hasard ou risque instantané λ) [14]

Le taux de risque instantané λ est la probabilité qu'un évènement survienne dans un petit intervalle de temps aprés t, sachant qu'il n'a pas eu lieu avant t il est défini par :

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq X < t + dt \mid X \geq t)}{dt}.$$

L'estimation du taux de hasard, de part la variété de ses possibilités d'application, est une question importante en statistique. Une des techniques les plus courantes pour construire des estimateurs de λ_X est basée sur sa définition donnée par la relation suivant :

$$\lambda_X(t) = \begin{cases} \frac{f_X(t)}{S_X(t)} & \text{si } s_X(t) \neq 0, \\ 0 & \text{sinon,} \end{cases}$$

et qui consiste à étudier un quotient entre un estimateur de f_X et un estimateur de S_X. L'article de Patilea et al. [1994] fait une présentation générale de ces techniques d'estimation. Les méthodes non-paramétriques basées sur les idées de noyau, qui sont connues pour leur bon comportement dans les problèmes d'estimation de densité, sont ainsi abondamment utilisées en estimation non-paramétrique de la fonction de hasard. Un large éventail de la littérature dans ce domaine est fourni par les revues de Singpurwalla et Wong [1983], Hassani et al. [1986], Izenman [1991], Gefeller et

Michels [1992] et Pascu et Vaduva [2003].

Il est donc tout naturel de construire un estimateur de la fonction λ_x en s'inspirant de ces idées de la manière suivante :

$$\hat{\lambda}_n^d(t) = \begin{cases} \frac{\hat{f}_n^d(t)}{\widehat{S}_{d_n}(t)} & \text{si } \hat{S}_{d_n}(t) \neq 0, \\ 0 & \text{sinon.} \end{cases}$$

Les propriétés de l'estimateur de la fonction de hasard s'obtiennent relativement facilement à partir de la littérature connue en matière d'estimation des fonctions de répartition et de densité.

2.3 Aplication

2.3.1 Estimateur de Kaplan-Meier de la fonction de la survie.

Afin de comparer la période de rémission des patients atteints de leucémie pendant des semaines et s'ils ont reçu le médicament "6-MP" selon le traitement.

en 1963, le scientifique Frichet a mené une expérience thérapeutique pour étudier les résultats de cette expérience. Comme le montre le tableau suivant :

$6-MP$	6	6	6	6^+	7	9^+	10	10^+	11^+
13	16	17^+	19^+	20^+	22	23	25^+	32^+	32^+
placebo	1	1	2	2	3	4	4	5	5
8	8	8	8	11	11	12	12	15	17

TAB. 2.1 – données de Frichet

Le signe $^+$ indique les patients qui sont sortis à la date spécifiée.

►Dans le groupe traité par le $6-MP$, 19 patients et 10 données censurées. La fonction de survie va être estimée de façon différente dans les 2 groupes. On note S_{6-MP}la fonction de survie des patients traités par le $6-MP$.

►Dans le groupe *placebo*, il y a 19 patients et aucune donnée censurée. On note $S_{placebo}$la fonction de survie des patients traités par le *placebo*.

groupe 6-MP "Estimateur de Kaplan-Meier"

L'idée est d'écrire :

$$\begin{aligned}
&P(\text{être en rémission à la ième semaine})\\
&= P(\text{être en rémission à la ième semaine sachant}\\
&\text{qu'il n'ya pas eu rechute à la}(j-1)\text{ème semaine})\\
&* P(\text{être en rémission à la } (j-1)\text{ème semaine})
\end{aligned}$$

on a $0 = X_{(0)} < X_{(1)} < ... < X_{(l)}$ avec $l \leq n$.

$$P(\tilde{X} > t_{(j)}) = \underbrace{P(\tilde{X} \leq t_{(j)} \mid \tilde{X} > t_{(j-1)})}_{pj} \times P(\tilde{X} > t_{(j-1)})$$

$$S\left(t_{(i)}\right) = p_j \times S\left(t_{(j-1)}\right)$$

$$S\left(t_{(j)}\right) = p_j \times p_{j-1} \times ... \times p_1 \times S\left(t_{(0)}\right)$$

on estime $p_j = P(\tilde{X} \leq t_{(j)} \mid \tilde{X} > t_{(j-1)})$ par

$$\hat{p}_j = (1 - \frac{\delta_j}{D_j}),$$

où

► δ_j est le nombre de rechutes observées au temps $t_{(j)}$.

► D_j est le nombre d'individus à risque de rechute (individus toujours en rémission) juste avant $t_{(j)}$.

L'estimateur de Kaplan-Meier est une fonction en escalier qui s'écrit :

$$\hat{S}_{6-MP}(t) = \prod_{j=1}^{n} \left(1 - \frac{\delta_j}{D_j}\right), \text{où } X_{(j)} \leq t < X_{(j+1)}.$$

et de la fonction de survie S de groupe de 19 malades traite par le traitement $6-MP$ donne le tableau suivant :

Temps t_j	n_j	δ_j	$\hat{S}_{6-MP}(t_j)$
0	19	0	1
6	19	3	$(1 - 3/19) * 1 - 0.857$
7	17	1	$(1 - 1/17) * 0.857 - 0.807$
10	15	1	$(1 - 1/15) * 0.807 - 0.753$
13	12	1	$(1 - 1/12) * 0.753 - 0.690$
16	11	1	$(1 - 1/11) * 0.690 - 0.627$
22	7	1	$(1 - 1/7) * 0.627 - 0.538$
23	6	1	$(1 - 1/6) * 0.538 - 0.448$

TAB. 2.2 – l'estimateur empirique pour le groupe traité par un 6-MP

groupe placebo

Dans le groupe traité par un *placebo*, la fonction de survie $S_{placebo}(t)$ est simplement estimée par

$$\hat{S}_{placebo}(t) = \frac{1}{n}\sum_{j=1}^{n} l\,(X_j > t)$$

$$= \text{proportion d'individus tels que} X_j > t.$$

Idée : on estime $P(X > t) = P(\text{ne pas rechuter avant } t)$ par la proportion de patients n'ayant pas rechutés avant t.

l'estimateur empirique pour le groupe traite par un *placebo* (pas de censurée) donne le tableau saivant :

Semaine j	Nombre de rémissions àla Semaine j	$\hat{S}_{placebo}(t)$
0	19	1
1	18	0.0047
2	17	0.0095
3	16	0.76
4	14	0.66
8	8	0.38
12	4	0.19
22	1	0.05
23	0	0

TAB. 2.3 – l'estimateur empirique pour le groupe traité par placebo

et si on compare ces chiffres aux valeurs critiques afférentes aux seuil de risque usuels de distribution de KHI-DEUX à un degré de liberté on conclut que lavantage de traitement est significatif.

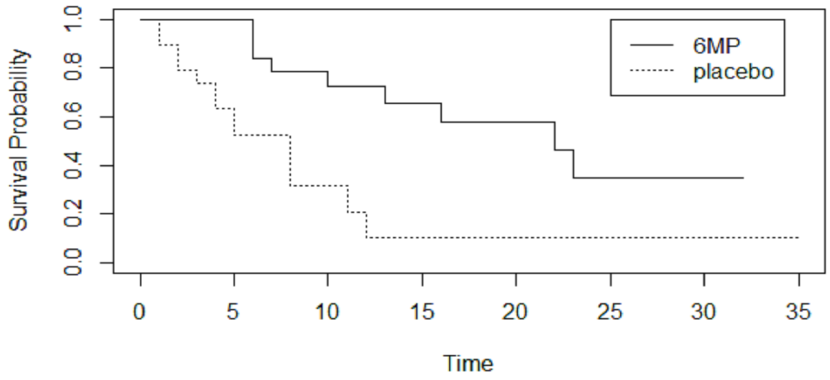

FIG. 2.2 – courbes de survie de deux traitements différents

Code R

```
X = c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32,
32, 34, 35, 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12)
D = c(1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
t=c(rep("6MP", 19), rep("placebo", 19))
f=data.frame(X, D,t)
library(survival)
s=survfit(Surv(X, D)~t , data=f)
plot(s, lty=c(1, 3), xlab="Time", ylab="Survival Probability")
legend(25,1.0, c("6MP", "placebo") , lty=c(1, 3) )
```

Ce code R effectue une analyse de survie sur un ensemble de données composé de durées de survie, d'indicateurs de censure et de types de traitement. Le jeu de données est créé en combinant trois vecteurs : X, D et t.

X contient les temps de survie pour chaque observation, où une observation correspond à un patient ou à un individu. D contient les indicateurs de censure, qui indiquent si une observation a été censurée ou non. Un indicateur de censure de 1 signifie que l'observation a été censurée, tandis qu'un indicateur de censure de 0 signifie que l'observation n'a pas été censurée.

t est une variable factorielle qui indique quel traitement chaque observation a reçu. Dans ce cas, il y a deux traitements : $6-MP$ et $Placebo$.Les 19 premières observations ont reçu le traitement $6 - MP$, tandis que les 19 secondes ont reçu le traitement $Placebo$.

Le code charge ensuite la bibliothèque de survie et utilise la fonction surfit pour ajuster une courbe de survie de Kaplan-Meier aux données. L'argument de la formule spécifie que l'objet de survie doit être modélisé en fonction de la variable de traitement t.

Enfin, la fonction plot est utilisée pour créer un tracé des courbes de survie, et la fonction legend est utilisée pour ajouter une légende au tracé.

Le graphique résultant montre la probabilité de survie estimée pour chaque traitement au fil du temps. La légende indique quelle ligne correspond à quel traitement. Dans l'ensemble, le code est utilisé pour comparer les propriétés de survie de deux traitements différents et visualiser les résultats.

Chapitre 3

Estimation non paramétrique pour des données doublement censurées

Dans ce chapitre, nous fommes intéressés par deux modéles qui sont trés utilisés dans la littérature dans ce domaine. à savoir le modéle de Turnbull(1974) connu par le modéle de censure double et le modéle de Patilea et Rolin(2006) que nous appelons le modéle de censure mixte.

3.1 Le modèle de Turnbull(1974)

Plusieurs modèles non paramétriques ont été proposés pour l'étude de la censure double. Par exemple, le modèle de Turnbull (1974) est le plus utilisé, et plusieurs travaux sont basés sur ce modèle.

3.1.1 Les estimateurs self-Consistantes

Soient $\hat{S}_n, \hat{S}_{d_n}, \hat{S}_{g_n}$.Les estimateurs non paramétriques pour S_X, S_det S_g.(respectivement), ces estimateurs sont proposé par Turnbull(1974). L'inconvénient de ces estimateurs est le fait qu'il n'existe pas de formules explicites connues pour eux, mais ils sont

définis comme les solutions d'équations intégrales appelées les équations de self-Consistantes. Ces équations sont données dans Ren(1997) pour tout $t \geq 0$ par

$$\hat{S}_n(t) = Q^{(n)}(t) - \int_{\{x \leq t\}} \frac{\hat{S}_n(t)}{\hat{S}_n(x)} dQ_1^{(n)}(x) + \int_{\{x>t\}} \frac{1-\hat{S}_n(t)}{1-\hat{S}_n(x)} dQ_2^{(n)}(x),$$

$$\hat{S}_{d_n} = 1 + \int_{\{x \leq t\}} \frac{dQ_1^{(n)}(x)}{\hat{S}_n(x)}, t < B_n,$$

$$\hat{S}_{g_n} = -\int_{\{x>t\}} \frac{dQ_2^{(n)}(x)}{1-\hat{S}_n(x)}, t \geq C_n,$$

où

$Q_j^{(n)}(t) := \frac{1}{n} \sum_{j=1}^{n} I_{\{Z_j > t, A_j = j\}}$, $j \in \{0,1,2\}$,

$Q^{(n)} := \sum_{j=0}^{2} Q_j^{(n)}$,

$C_n := \min \left\{ Z_j \mid \hat{S}_n(Z_j^-) < 1 \right\}$,

$B_n : \max \left\{ Z_j \mid \hat{S}_n(Z_j^-) > 0 \right\}$.

Les properiétés asymptotiques de $\hat{S}_n$(convergence presque sûre, convergence faible, efficacité asymptotique,...) ont fait l'objet de divers travaux, parmi lesquels ceux de Tsai et Crowley(1985), Chang et Yang(1987),Chang(1990),Gu et Zhang(1993)et Bih-Sheue et Cheun-Der(2004).

Nous allons expliciter deux de ces résultats dont nous aurons besoin dans la suite. Pour cela considérons les hypothéses suivant citeés dans...

$H_1 : \forall t \geq 0, S_d(t) - S_g(t) > 0.$

$H_2 : S_{_X}, S_d$et S_g sont des fonction continues de t pour $t \geq 0$, et $0 < S_X(t) < 1$pour $t > 0$.

$H_3 : S_X(0) = S_d(0) = 1$ et $\lim_{u \to \infty} S_X(x) = \lim_{u \to \infty} S_d(x) = \lim_{u \to \infty} S_g(x) = 0$.

H_4 : il existe αet β, $0 < \alpha < \beta < \infty$,tels que $P(G \in]0, \alpha[) = 0$et $P(G \leq \beta) = 1$.

Théorème 3.1.1 *Sous les hypothése H_1 et H_2 nous avons*

$$\sup_{t\geq 0}\left|\hat{S}_n(t)-S_X(t)\right|\overset{p.s.}{\to}0,$$
$$\sup_{t\geq 0}\left|\hat{S}_{d_n}(t)-S_d(t)\right|\overset{p.s.}{\to}0,$$
$$et\ \sup_{t\geq 0}\left|\hat{S}_{g_n}(t)-S_g(t)\right|\overset{p.s.}{\to}0.$$

Preuve. Voir Change et Yang(1987) Théorème 4.2 page 1546. ■

Notons $D[0,b]$ l'espace des fonction réelles définies sur $[0,b]$, continues à droite et ayant des limites à gauche en tout point de $[0,b]$(les fonctions cadlag). Le résultat suivant concerne la convergence faible des estimateurs $\hat{S}_n, \hat{S}_{d_n}$et $\hat{S}_{g_n}$.

Théorème 3.1.2 *Sous les hypothéses H_1 et H_2, le processus*

$$x^{(n)}=\sqrt{n}\left(\hat{S}_n-S_X,\hat{S}_{d_n}-S_d,\hat{S}_{g_n}-S_g\right)$$

converge faiblement vers un processus gaussien sur $D[0,b]\times D[0,b]\times D[0,b]$.

Preuve. Voir Change (1990) Théorème 3.1 page 399. ■

3.1.2 Estimation de la densité et du Taux hasard

Dans le cas de données complétes et censurées à droite. Ren(1997) a proposé d'estimer la densité f de X par

$$\hat{f}_n(t):=\frac{1}{h_n}\int K(\frac{t-x}{h_n})d\hat{F}_n(x),$$

où $\hat{F}_n:=1-\hat{S}_n$.

Ren(1997) a montré la convergence presque sûre uniforme et la normalité asymptotique de $\widehat{f_n}$.

Théorème 3.1.3 *Soit $t \geq 0$,sous les hypothéses H_1 et H_4et si*

► f est bornée, $f(t) > 0$ et au voisinage de t,la dérivée seconde de $f(S_d - S_g)$ existe et elle est bornée,

► K est une densité paire,bornée et à support dans $[-1, 1]$,

► $nh_n^3 \underset{n\to\infty}{\to} 0$ et $nh_n \underset{n\to\infty}{\to} \infty$,

nous avons

$$\sqrt{nh_n}\left(\hat{f}_n(t) - f(t)\right) \xrightarrow{D} N(0, \frac{f(t)}{S_d(t) - S_g(t)} \int K^2(x)dx).$$

Ren(1997) a également étudié l'estimateur suivant du taux de hasard λ

$$\hat{\lambda}_n(t) := \frac{\hat{f}_n(t)}{\hat{S}_n(t)},$$

il a montré que sous les hypothéses H_1 et H_4

$$\sqrt{nh_n}\left(\hat{\lambda}_n(t) - \lambda(t)\right) \xrightarrow{D} N(0, \frac{f(t)}{S_Y^2(t)(S_d(t) - S_g(t))} \int K^2(x)dx).$$

3.2 Le modéle de Patilea et Rolin(2006)

Nous nous interessons ici au modèle de censure double proposé par Patilea et Rolin (2006), et qu'on nomme ici par le modèle de censure mixte.

3.2.1 L'estimateur de Patilea et Rolin

Dans toute la suite, pour toute variable aléatoire X, nous notons F_X sa fonction de répartition et $S_X := 1 - F_X$ sa fonction de survie. De plus, on définit le point initial du support de X, noté I_X, par : $I_X := \inf\{t \in \mathbb{R}/F_X(t) > 0\}$, et le point terminal du support de X, noté T_X par :$T_X := \sup\{t \in \mathbb{R}/F_X(t) < 1\}$.$I_X$ etT_X possédent les

propriétés suivantes :

► $I_X \leq X \leq T_X$ *p.s.*

►si Y est une *v.a.r.*indépendante de X, on a :$I_{X\wedge Y} = I_X \wedge I_Y$, $I_{X\vee Y} = I_X \vee I_Y$, $T_{X\wedge Y} = T_X \wedge T_Y$, $T_{X\vee Y} = T_X \vee T_Y$.

Patilea et Rolin (2006) ont proposé un estimateur produit limite de F en supposant qu'il n'y a pas d'ex-æquo parmi les X_j, ce qui est naturel dans notre situation d'existence de la densité, et qui s'écrit sous la forme suivante :

$$\hat{F}_n^p(t) = 1 - \hat{S}_n^p(t) := 1 - \prod_{k:Z_k\leq t} \left\{ 1 - \frac{I_{\{A_j=0\}}}{n(\hat{F}_{g_n}(Z_k^-) - \hat{F}_{Z_n}(Z_k^-))} \right\},$$

où $\hat{F}_{Z_n}(t) := \frac{1}{n}\sum_{i=1}^n I_{\{Z_j\leq t\}}$est la version empirique de F_Z.
et $\hat{F}_{g_n}(t) := \prod_{j:Z'_j>t} (1 - \frac{\sum_{i=1}^n I_{\left\{Z_j=Z'_j, A_j=2\right\}}}{n\hat{F}_{Z_n}(Z'_j)})$,
$((Z'_j)_{1\leq j\leq m}(m \leq n)$ étant les valeurs distinctes des $(Z_j)_{1\leq j\leq n})$ est l'estimateur produit limite de la fonction de répartition de G qui est censuré à gauche par $\min(X, D)$. Cet estimateur peut être déduit de celui de Kaplan-Meier en inversant le temps.

Patilea et Rolin (2006) ont montré la convergence presque sûre et uniforme de $\hat{F}_n^p(t)$, la vitesse de cette convergence a été précisée par la loi du logarithme itéré de Massaci et Nemouchi(2011,2013). Kitouni et al.(2015)ont donné un taux de la convergence presque complète uniforme, de l'ordre de $\sqrt{\log n/n}$. Notons par $H_k(t) := P(Z \leq t, A = k), k \in \{0, 1, 2\}$les sous loi de Z et $I_{H_0} := \inf\{t \in \mathbb{R}/H_0(t) > 0\}$le point initial du support de H_0. Dans la suite nous aurons besoin des deux résultats suivants de la convergence faible.

Théorème 3.2.1 *Sous les hypothéses*

C_1 : $I_g \leq I_X$ et $T_X \leq T_d$,

C_2 : $\int_{\left\{x>I_{H_0}\right\}} \frac{dH_2(x)}{(F_Z(x))^2} < \infty$,

nous avons

1. $\sqrt{n}(\hat{F}_{g_n} - F_g)$ converge faiblement vers un processus gaussien centré dans $D\left[I_{H_0}, \infty\right]$,
2. $\sqrt{n}(\hat{S}_n - S_X)$ converge faiblement vers un processus gaussien centré dans $D\left[0, \tau\right]$,où τ est tel que $I_{H_0} < \tau$ et $H_0(\tau) + H_1(\tau) < 1$.

Preuve.

1. Voir Patilea et Rolin(2006) Lemme 7.2.page 935.
2. Voir Patilea et Rolin(2006) Théoreme 7.3.page 937.

■

3.2.2 Estimation de la densité et du Taux hasard

Kitouni et al.(2015) et par analogie avec les cas précédents,ont propsé l'estimateur suivant pour f

$$\hat{f}_n^p(t) := \frac{1}{h_n} \int K(\frac{t-x}{h_n}) d\hat{F}_n^p(x),$$

et ils ont montrés sa convergence presque complète énoncée comme suit :

Soit C un compact inclus dans $[0, \min(T_x, T_d)[$,sous les hypothéses

► $\max(I_g, I_d) < I_x$,

► $h_n \to 0$ et $nh_n^2 / \log n \to \infty$,

► K est une fonction continue à droite, à variation bornée, à support compact et telle que $\int K(t)dt = 1$,

nous avons :

1. S'il existe un entier $r \geq 2$ telle que f est r fois continûment différentiable

autour de C et $\int t^j K(t)dt = 0, \forall 1 \leq j \leq r-1$, alors

$$\sup_{t\in C}\left|\hat{f}_n^p(t) - f(t)\right| = o_{p.co.}(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}).$$

2. S'il existe $\alpha, \beta, \varepsilon \in \mathbb{R}_+^*$ tels que $\forall t \in C, \forall x \in]t-\varepsilon, t+\varepsilon[\,, |f(t)-f(x)| \leq \alpha\,|t-x|^\beta$,alors

$$\sup_{t\in C}\left|\hat{f}_n^p(t) - f(t)\right| = o_{p.co.}(h_n^\beta + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}).$$

Kitouni et al.(2015)ont également montré des résultats similaires pour l'estimateur du taux de hasard suivant

$$\hat{\lambda}_n^p(t) := \frac{\hat{f}_n^p(t)}{\hat{S}_n^p(t) + u_n},$$

où $u_n > 0$ est le terme général d'une suite qui converge vers zéro et qui sert à éviter la division par zéro.

Conclusion

Dans notre étude, nous avons considéré différents aspects de l'estimation non paramétrique dans le cas de données censurées. Dans le premier chapitre on a une présentation simple des concepts de base sur la censure et ses types, ainsi que quelques définitions qui sont nécessaire pour le deuxieme chapitre, qui est l'etude des estimations non paramétriques pour des données censurées à droite. Parmi ces estimateurs, il y a l'estimateur de Kaplan-Meier et l'estimation du taux de hasard. Comme nous le verrons au troisiéme chapitre, d'autres estimateurs non paramétriques, dans le cas du double censure sont cités comme l'estimateur de Patilea et Rolin ainsi que l'estimation de la densité et du taux de hasard.

Bibliographie

[1] BENABED, R. Estimation de la moyenne d'une distribution à queue lourde en présence de censure (Doctoral dissertation, UNIVERSITÉ KASDI MERBAH OUARGLA).

[2] Benhaoued, M. Etude théorique de l'estimation de la régression dans le modèle de données censurées (Doctoral dissertation, UNIVERSITÉ KASDI MERBAH OUARGLA).

[3] Chang,M.N(1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. The annals of statistics,18(1),391-404.

[4] Chang,M.N.,et Yong G.L.(1987). Stronge consistency of a non parametric estimtor of the survival function with doubly censored data. The Annals of statistics, 15(4), 1536-1547.

[5] Droesbeke, J.J., Saporta, G. Approches non parametriques en régression. Editions Technip, 2011. (Cité en pages 28, 30, 32 et 33.)

[6] FÖldes, A. and RejtŐ, L. (1981). A lil type result for the product limit estimator. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 56(1), 75–86.*

[7] Kaplan, E. L et Meier, P. (1958). Estimation nano paramétrique à partir d'observations incomplètes. J. Amer. Statist Assoc ; 53(282) ; 457-481.

[8] Kitouni, A. Boukeloua, M. et Messaci, F. (2015). Rate of strong consistency for non parametric estimators based on twice censored data. *Statistics and probability letters, 96, 255-261.*

[9] Boukeloua, M., Messaci, F., & Keziou, A. (2017). Etude de modèles semi et non paramétriques pour des données censurées (Doctoral dissertation, Université Frères Mentouri-Constantine 1).

[10] Rouabah, N. E. H., & Nemouchi, N. (2019). Estimation non paramétrique dans un modèle de censure et de dépendance (Doctoral dissertation, Université Frères Mentouri-Constantine 1).

[11] GEURRICHA, N. L'estimateur de la régression en utilisant la méthode de noyau dans les modèles de censurées, mémoire de Master, UNIVERSITÉ KASDI MERBAH OUARGLA..

[12] Patilea,V.,et Rolin,J,-M.(2006).Product limit estimators of the survival function with twice censored data. the Annals of Statistique, 34(2), 925-938.

[13] Ren, J-J. (1997). On self-consistent estimators and kernel density estimators with doubly censored data. *Journal of Statistical Planning and Inference, 64, 27-43.*

[14] Soltane, L. (2017). Analyse des valeurs extrêmes en présence de censure (Doctoral dissertation, Université Mohamed Khider-Biskra).

[15] Abdelaziz, S., & Sissaoui, A. (2020). Estimation pour des données censurées (Doctoral dissertation, University of Jijel).

[16] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association, 96, 169-173.*

[17] Rabhi, Y. (2006). Modèles de survie avec un point de rupture (Doctoral dissertation, Université du Québec à Montréal).

Annexe A : Logiciel R

3.3 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

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

$i.i.d$	:	independant et identiquement distribué.
$v.a.r$	:	variable(s) aléatoire(s) reelle(s).
$f.d.r$	:	fonction de répartition.
$p.s.$	:	presque sûrement.
$X \xrightarrow{d} Y$	:	convergence en loi.
$\bar{F}$	:	$S(t) = 1 - F,$ fonction de survie.
$u_n = o(v_n)$	:	$\forall \varepsilon > 0, \exists N \in \mathbb{N}^* / \forall n \geq N : u_n \leq \varepsilon v_n.$
$u_n = O(v_n)$	:	$\exists \gamma > 0 / u_n \leq \gamma v_n,$ pour un n assez grand.
$\mathbb{R}$	:	Ensemble des valeurs réelles.
F	:	Fonction de répartition.
F_n	:	Fonction de répartition empirique.
F_V	:	Fonction de répartition de$V$.
S_V	:	Fonction de survie de$V$.
f_V	:	Densité de probabilité de$V$.
λ_V	:	Taux de hasard de$V$.

$P(V \in A)$	:	La probabilité que V appartient à l'ensemble$A$.
I_A	:	fonction indicatrice de l'ensemble$A$.
$p.co.$	:	presque continue

Résumé

Dans ce mémoire nous avons présentés en premier lieu la fonction de survie, particulièrement dans le cas de censure à droite, l'estimateur de la distribution de survie pour les données censurées à droite, plus connue sous le nom de l'estimateur de Kaplan-Meier.

Cet estimateur n'est plus valable dans le cas de censure mixte, d'où l'intérêt de présenté l'estimateur qui convient à cette situation et qu'on appelle estimateur de Patilea et Rollin(2006) et qui comble cette lacune.

Mots clés : Distribution de survie, données avec censure simple, et censure double ou mixte, taux de hasard, estimateurs de Kaplan Meier et Patilea et Rollin.

Abstract

In this thesis we first presented the survival function, particularly in the case of right censoring, the estimator of the survival distribution for right censored data, better known as the Kaplan- Meier.

This estimator is no longer valid in the case of mixed censoring, hence the interest of presenting the estimator which is suitable for this situation and which is called the estimator of Patilea and Rollin (2006) and which fills this gap.

Key words: Survival distribution, data with single censoring, and double or mixed censoring, hazard rate, Kaplan- Meier and Patilea and Rollin estimators.

ملخص

في هذه الأطروحة ، قدمنا أولاً وظيفة البقاء على قيد الحياة ، لا سيما في حالة الرقابة الصحيحة ، مقدر توزيع البقاء على قيد الحياة للبيانات الخاضعة للرقابة الصحيحة, والمعروفة باسم مقدرKaplan-Meier

لم يعد هذا المقدّر صالحًا في حالة الرقابة المختلطة، ومن هنا جاءت الفائدة في تقديم المقدّر المناسب لهذه الحالة والذي يسمى مقدر(2006) Patilea et Rollin, هذا والذي يملأ هذه الفجوة.

الكلمات المفتاحية : توزيع البقاء, البيانات مع رقابة واحدة, والرقابة المزدوجة أو المختلطة, نسبة المجازفة, مقدرات كابلان ماير وباتيليا ورولين.