UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DONNÉES CENSURÉES PAR INTERVALLE : APPLICATION À L'ÉTUDE DE LA PRÉVALENCE DU SURPOIDS APRÈS TRAITEMENT DE LA LEUCÉMIE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MICHAEL EVRARD MAKOUANGOU NGOUMA

FÉVRIER 2019

REMERCIEMENTS

Premièrement, je suis redevable à mes directrices Geneviève Lefebvre et Juli Atherton, de m'avoir donné l'occasion d'élargir mes compétences en faisant un projet en analyse de survie. Je les remercie pour leur attitude accueillante, ainsi que leur patience qui ont développé en moi confiance, espoir et courage pour atteindre mes objectifs. Votre soutien académique, votre relecture attentive du manuscrit ont contribué à la réalisation et à l'amélioration de ces travaux de recherche.

Je remercie aussi le Département de mathématiques de l'UQAM pour son accueil et son assistance. Enfin, je suis redevable à mes parents pour avoir assuré ma bonne éducation, à ma famille pour avoir toujours été la source de ma motivation, de mon inspiration et pour leurs encouragements et leur assistance sans faille.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Le projet s'intéresse à l'estimation non paramétrique d'une fonction de survie lorsque les données sont censurées par intervalle et à une application à l'étude d'une cohorte de survivants de leucémie lymphoblastique aiguë infantile. Nous mettons en forme de manière détaillée la présentation de l'algorithme de Turnbull comme un algorithme EM. Nous comparons cet algorithme à une approche plus naïve basée sur l'estimation de Kaplan-Meier pour données imputées. Cette comparaison repose sur une étude de simulation adaptée à l'application réelle visée. Nous introduisons à cette occasion un modèle de mélange pour la construction d'un taux de survie de forme non usuelle adapté au problème étudié. Puis dans l'application nous étudions la dépendance entre le surpoids et certaines covariables de traitement de la leucémie.

Mots-clés : censure par intervalles, survie, algorithme EM, mélange, estimation non-paramétrique.

INTRODUCTION

Ce mémoire est consacré à l'estimation non paramétrique d'une fonction de survie lorsque les données sont censurées par intervalle. Ce travail est aussi motivé par une application à l'étude d'une cohorte de survivants de leucémie lymphoblastique aiguë infantile à plusieurs phases basée au Centre de santé de l'Université Sainte-Justine (CHU Sainte-Justine, Montréal, Canada) Marcoux *et al.* [2017]. Cette étude est appelée étude PÉTALE pour « Prévenir les effets tardifs des traitements de la leucémie aiguë lymphoblastique chez l'enfant ». Il s'agit d'un projet de recherche multidisciplinaire visant à caractériser de manière exhaustive les effets indésirables tardifs et à identifier les biomarqueurs prédictifs associés chez les survivants de la leucémie lymphoblastique aiguë infantile. L'effet indésirable qui nous intéressera dans ce projet est la survenue éventuelle d'un surpoids lié au traitement reçu.

Ce mémoire est organisé en trois chapitres. Dans le chapitre 1, nous décrivons plus précisement le schéma d'étude utilisé dans PÉTALE jusqu'à la formalisation mathématique de la censure par intervalle. Puis dans le chapitre 2 nous proposons deux approches pour l'estimation non paramétrique de la fonction de survie basée sur les données censurées par intervalle : l'une est basée sur l'estimation de Kaplan-Meier sur données imputées et l'autre sur l'algorithme de Turnbull. Ces deux approches sont comparées à travers une étude de simulations construite spécifiquement à partir d'informations a priori sur les vraies données. Enfin, le chapitre 3 s'intéresse à l'application aux données de la leucémie pédiatrique. Nous présentons une description et une analyse préliminaire des données, puis nous étudions la dépendance entre la survenue du surpoids et certaines variables de traitement en utilisant les outils du chapitre 2.

Ce mémoire se termine par un récapitulatif de nos principales contributions ainsi que par des perspectives de recherche envisagées.

CHAPITRE I

DESCRIPTION ET MODÉLISATION DES DONNÉES RÉELLES

Les données de l'étude PÉTALE considérées dans ce mémoire portent sur 246 individus et 22 variables. Les individus sont des survivants de la leucémie pédiatrique qui ont suivi un traitement basé sur les corticostéroïdes. Il est connu que certains patients ont un gain de poids rapide durant la phase intensive de ce traitement. D'autres patients auront aussi un gain de poids lié naturellement à une augmentation de leur âge. Nous nous intéresserons dans la suite au premier temps d'apparition du surpoids.

Dans ce qui suit, nous présentons schématiquement le déroulement de l'étude PÉTALE. On fait un suivi d'une cohorte de patients depuis leur diagnostic de leucémie. Dans la première étape, certaines caractéristiques des survivants sont extraites des dossiers médicaux, ce qui nous permet de faire un diagnostic du surpoids en mesurant l'indice de masse corporelle (IMC) au diagnostic du cancer. À cette étape, presque aucun patient n'est en surpoids puisque la prévalence du surpoids est relativement faible chez les enfants en bas âge et les prépubères. Le poids du patient est également disponible après la fin du traitement, soit environ deux ans après le diagnostic (deuxième étape). Dans la troisième étape, qui correspond à la date d'entrevue de l'étude PÉTALE, des mesures complètes de la santé de ces survivants sont obtenues. Cette visite médicale intervient au moins 5 ans et en moyenne 15.5 ans après le diagnostic de la leucémie. Pour être considéré dans l'étude, il faut être en rémission et vivre assez longtemps pour participer aux 3 étapes. L'étude clinique est résumée dans la figure 1.1.

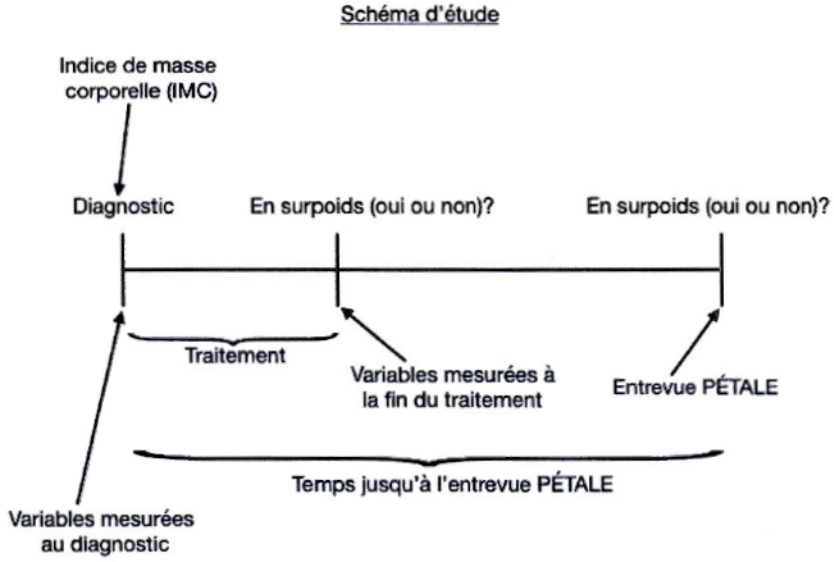


Figure 1.1 Conception de l'étude : Des variables sont mesurées à trois moments. Au moment du diagnostic de la leucémie, de nombreuses variables ont été mesurées, y compris l'IMC. Le traitement dure approximativement 2 ans, à quel point d'autres variables ont été mesurées. Le délai minimum est de 5 ans entre le diagnostic et l'entrevue PÉTALE.

Comme on peut le voir dans la figure 1.1, les médecins ou intervenants en santé peuvent déterminer si le patient est en surpoids à chacune des trois rencontres (diagnostic initial, fin du traitement et début de l'entrevue PÉTALE). Notre intérêt réside dans le temps écoulé entre le diagnostic de leucémie et la date à laquelle un patient devient en surpoids pour la première fois.

Nous notons par T cette variable aléatoire et utilisons des méthodes d'analyse de survie pour modéliser le temps jusqu'au moment où l'on devient en surpoids en fonction des autres variables collectées. Notons que pour simplifier la présentation, nous parlerons souvent de T comme le temps du surpoids.

Les patients qui sont initialement en surpoids au moment du diagnostic n'entrent pas dans notre analyse. Notre analyse porte donc sur les patients qui ne sont pas en surpoids au diagnostic, dont une certaine proportion d'entre eux le deviendra. Le surpoids est censé être lié au traitement ou à l'augmentation de l'âge. Pour les patients de notre base de données qui deviendront en surpoids, le temps écoulé jusqu'au moment où l'on devient en surpoids (à partir du diagnostic initial) est soit censuré par intervalle, soit censuré à droite. Par exemple, un patient qui devient en surpoids après le diagnostic et avant la fin du traitement a son intervalle de temps de surpoids censuré par le diagnostic et la date du premier rendez-vous juste après la fin du traitement. Un patient qui devient en surpoids entre la date de fin du traitement et l'entrevue PÉTALE a également son intervalle de temps de surpoids censuré par la fin du traitement et la date du deuxième rendez-vous (entrevue PÉTALE). D'un autre côté, un patient qui devient en surpoids après son entrevue PÉTALE a un temps de surpoids qui est censuré à droite. Notez que dans notre base de données, nous ne pouvons pas faire la distinction entre les patients qui ne deviennent jamais en surpoids et ceux qui le deviennent et ont un temps de surpoids censuré à droite.

1.1 Notations et exemple

Avant d'introduire formellement la censure par intervalle dans la section 1.2, nous présentons un exemple hypothétique avec quatre patients. Soit G_{ij} la j^{e} date du rendez-vous du i^{e} sujet, avec $j = 1, 2, 3$ et $i = 1, 2, 3, 4$. La figure 1.2 illustre les dates de rendez-vous et les réalisations des dates hypothétiques du surpoids d_i pour chaque sujet i. Le début et la fin de l'étude PÉTALE sont indiqués par des lignes verticales en pointillés tel que décrit dans Marcoux *et al.* [2017] ; l'étude PÉTALE a débuté en 2013 et s'est terminée en 2015.

Pour chaque sujet i le premier intervalle d'observation entre le diagnostic G_{i1} et la fin du traitement G_{i2} est indiqué en bleu, le deuxième intervalle d'observation entre la fin du traitement G_{i2} et le début de l'entrevue PÉTALE G_{i3} est indiqué en noir (ligne épaisse)

tandis que le temps entre le début de l'entrevue PÉTALE G_{i3} et la fin de l'étude (2015) est indiqué en rouge (ligne fine). Lors de cette étude, le statut en surpoids ou non en surpoids de chaque patient i après la fin de l'entrevue PÉTALE en 2015 (lorsqu'on a cessé d'interroger les patients) n'est pas connu. Pour les quatre patients, le délai entre le diagnostic G_{i1} et la fin du traitement G_{i2} est d'environ 2 ans. Par contre, le temps entre la fin du traitement G_{i2} et l'entrevue PÉTALE G_{i3} varie d'un patient à l'autre. Dans notre exemple hypothétique, ce deuxième intervalle d'observation s'étend sur 9 ans pour le patient 1, 10 ans pour le patient 2, 12 ans pour le patient 3 et 13 ans pour le patient 4.

Les figures 1.2 et 1.3 illustrent le suivi de 4 patients.

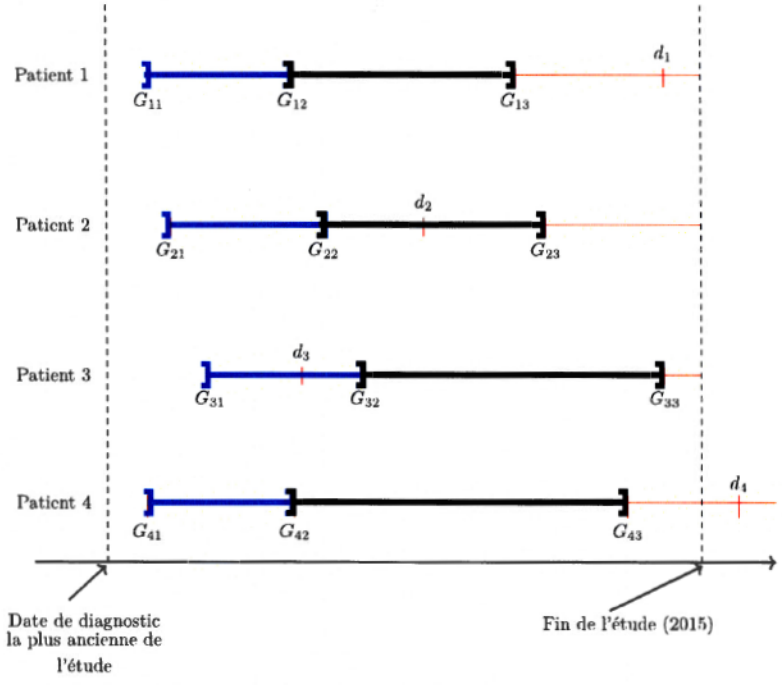


Figure 1.2 Schéma pour quatre patients hypothétiques recrutés pour l'entrevue PÉTALE. Pour chaque patient i, G_{i1} est la date du diagnostic, G_{i2} est la date de fin du traitement et G_{i3} est l'entrevue PÉTALE. La date du surpoids pour chaque patient est indiquée par d_i.

La notation et l'analyse présentées dans la section suivante ne se réfèrent pas aux dates du surpoids, mais au temps écoulé entre le diagnostic et le début du surpoids. Par conséquent, dans la figure 1.3, nous présentons les quatre patients hypothétiques de la figure 1.2 sur une échelle de temps commune où $t = 0$ est le moment du diagnostic pour chaque patient. Maintenant le temps jusqu'au surpoids pour chaque sujet i est noté t_i. Le premier intervalle d'observation (en bleu) pour le sujet i va de 0 à $G_{i2} - G_{i1}$ et le second intervalle d'observation (en noir épais) pour le sujet i va de $G_{i2} - G_{i1}$ à $G_{i3} - G_{i1}$.

La notation $(L_i, R_i]$ dont la valeur de réalisation $(l_i, r_i]$ est l'intervalle contenant t_i (le temps du surpoids) pour chaque sujet i est également introduite dans la figure 1.3. Pour les patients 2 et 3 qui sont censurés par intervalle, nous avons $(l_2, r_2] = (G_{22} - G_{21}, G_{23} - G_{21}]$ et $(l_3, r_3] = (0, G_{32} - G_{31}]$. D'autre part, les patients 1 et 4 deviennent en surpoids après leur entrevue PÉTALE et ont donc des temps de surpoids qui sont censurés à droite. La censure à droite est incluse dans la notation $(L_i, R_i]$ en permettant à R_i de prendre la valeur ∞. Donc pour les patients 1 et 4 on a $(l_1, r_1) = (G_{13} - G_{11}, \infty)$ et $(l_4, r_4) = (G_{43} - G_{41}, \infty)$ respectivement.

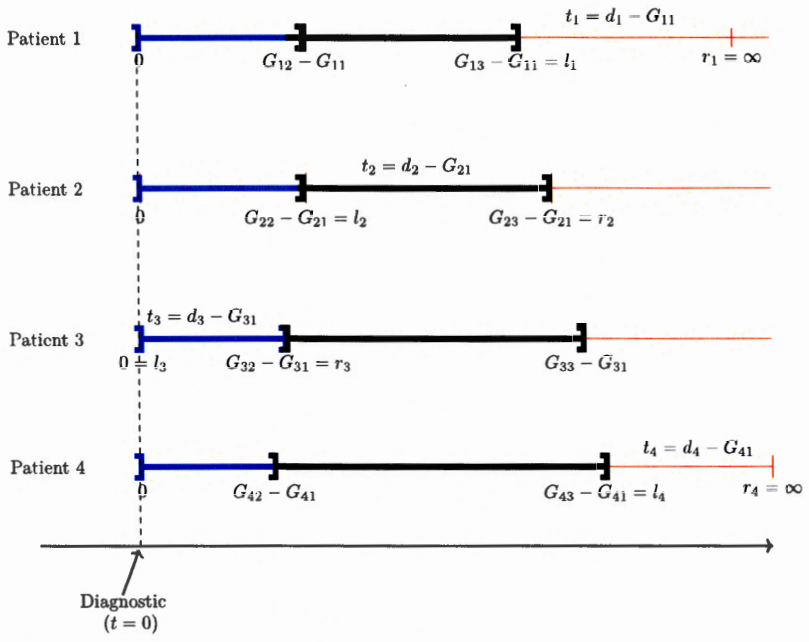


Figure 1.3 Les quatre patients sont représentés sur une échelle de temps commune où $t = 0$ correspond au diagnostic. Les réalisations de valeurs $(l_i, r_i]$ pour chaque sujet i sont également indiquées.

Dans ce mémoire, il est important de rappeler que nous travaillons dans l'échelle de la figure 1.3, autrement dit que le diagnostic a lieu au temps $t = 0$ pour tous les individus.

1.2 Cadre mathématique de la censure par intervalle

Dans cette section, nous allons introduire la censure par intervalle et formaliser les notations introduites dans la section 1.1.

Le temps de défaillance T représente le temps du suivi d'un individu, depuis son diagnostic jusqu'à l'apparition du surpoids. Ce temps a pour fonction de répartition $F(t) = P(T \leq t)$, pour fonction de survie $S(t) = P(T > t) = 1 - F(t)$, et pour densité, $f(t)$, si elle existe.

Dans cette étude, T n'est pas observé directement. En effet, le surpoids peut se produire entre le début du suivi G_1, qui coïncide avec le diagnostic de la leucémie, et la fin du traitement G_2 ou entre celle-ci et l'entrevue PÉTALE G_3. Il est également possible que le surpoids survienne après l'entrevue (fin du suivi) ou ne survienne jamais. On a donc que T est censuré par intervalle, autrement dit, que la seule information disponible pour T constitue un intervalle noté $I = (L, R]$, où les bornes L, R sont aléatoires, tel que $T \in I$. On note $f_{(L,R)}$ la densité conjointe de L et R telle que $P(L \leq R) = 1$. Notons que la censure par intervalle contient aussi la censure à droite comme cas spécial, cette dernière se réalisant lorsque $R = \infty$. Cela correspond au cas où on n'observe pas de surpoids avant la fin du suivi.

Définition 1.2.1. *Un modèle de données censurées par intervalle est décrit par la fonction de répartition conjointe $F_{(L,R,T)}$ entre la variable aléatoire T et l'observation de censure aléatoire $(L, R]$, avec support $\{(l, r, t) : 0 < l < t \leq r \leq \infty\}$,*
c'est-à-dire sous la contrainte que $P(T \in (L, R]) = 1$.

Dans notre étude portant sur n individus, on note $T_1, \ldots, T_n$ les temps de défaillance représentant le moment d'apparition du surpoids depuis le diagnostic de leucémie. Dans ce chapitre, on suppose $T_1, \ldots, T_n$ i.i.d. de même loi que T. Les valeurs de réalisations de $T_1, \ldots, T_n$ sont inconnues et nous observons des intervalles qui contiennent les valeurs non observées de $T_1, \ldots, T_n$:

(i) Soit $\{(L_i, R_i]; i = 1, \ldots, n\}$, les intervalles de censure où L_i est le dernier temps d'observation pour le i^{e} individu avant l'apparition du surpoids et R_i indique la première fois que le surpoids a été observé. De façon formelle, nous observons des vecteurs de censure aléatoire $(L_i, R_i]$, $i = 1, \ldots, n$ i.i.d. de même loi que $(L, R]$.

(ii) Pour un jeu de données fixé, les réalisations $t_1, \ldots, t_n$ des temps de défaillance ne sont pas observées et l'ensemble des données observables est alors

$$O = \{(l_1, r_1], \ldots, (l_n, r_n]\}, \tag{1.1}$$

où l_i, r_i sont des réalisations de L_i, R_i, respectivement.

On suppose ici que la censure est non informative au sens de Peace *et al.* [2012], p. $7-8$, Gómez *et al.* [2004], p. $143-144$ et Klein *et al.* [2013] voir équation (18.3) chapitre 18. En d'autres termes, on a mathématiquement la relation

$$P(T \leq t \mid L = l, R = r, L < T \leq R) = P(T \leq t \mid l < T \leq r). \tag{1.2}$$

L'équation (1.2) signifie que la seule information fournie par l'intervalle de censure $(l, r]$ à propos du temps de survie t est que l'intervalle contient t.

Remarque 1.2.2. *Certains auteurs comme Sun [2007], p.14, parlent de censure indépendante dans le cas (1.2).*

Exemple 1.2.3. ***(Censure non informative)***
On a des visites pré-programmées qui ne sont pas reliées au statut de surpoids du patient.

Vu la manière dont les données ont été collectées (on a trois rendez-vous pour tout le monde), on se limite dans ce chapitre et dans la suite à la censure non informative.

1.3 Vraisemblance

On définit la vraisemblance comme la densité conjointe des observations (L_i, R_i), $i = 1, \ldots, n$, évaluée en les réalisations (l_i, r_i), $i = 1, \ldots, n$. En statistique, il est commun d'utiliser la vraisemblance des données observées pour estimer un paramètre d'intérêt. Pour nous, il s'agira de la fonction de survie S de T. Afin d'exprimer facilement cette vraisemblance en fonction de S, on peut par exemple faire l'hypothèse de somme constante.

Définition 1.3.1. *Un modèle de censure est à somme constante si et seulement si pour tout $t \geq 0$ tel que $S(t) \neq 0$, l'équation suivante est vérifiée*

$$\int\int_{\{(l,r):t\in(l,r]\}} \frac{f_{(L,R)}(l,r)}{P(T \in (l,r])} dl dr = 1. \tag{1.3}$$

La condition (1.3) est toutefois difficile à interpréter concrètement. Aussi, nous avons privilégié dans ce mémoire l'hypothèse de censure non informative, qui est en fait un cas particulier de (1.3).

Théorème 1.3.2. *Si un modèle de censure est à somme constante alors la vraisemblance est proportionnelle à la vraisemblance réduite suivante*

$$\mathcal{L}(S \mid O) = \prod_{i=1}^{n} [S(l_i) - S(r_i)], \tag{1.4}$$

où O est défini en (1.1). Le facteur de proportionnalité est positif et ne dépend pas de S.

Démonstration. Voir Oller Piqué [2006] □

Proposition 1.3.3. *Supposons que nous observons le temps de défaillance T qui tombe dans l'intervalle aléatoire $(L, R]$. Soit $f_{(T,L,R)}(t, l, r)$ la densité conjointe du vecteur partiellement observé (T, L, R) tel que $P(L < T \leq R) = 1$.*

(i) L'équation (1.2) est équivalente à

$$f_{(T|L,R)}(t|l, r) = \frac{f_T(t)}{P(T \in (l, r])} \mathbb{I}_{\{t \in (l,r]\}}. \tag{1.5}$$

(ii) Si un modèle de censure est non informatif, alors le modèle est à somme constante.

Le résultat de la proposition 1.3.3 est déjà connu (voir Proposition 1, de Oller *et al.* [2004]). Dans ce qui suit, nous en donnons une autre démonstration.

Démonstration. On note $f_{(T|L,R)}$ la densité conditionnelle de T sachant (L, R) ; $f_{(L,R|T)}$ la densité conditionnelle de L, R sachant T ; $f_{(T,L,R)}$ la densité conjointe de T, L, R.

(i) D'après l'équation (1.2), on a

$$P(T \leq t \mid L = l, R = r, L < T \leq R) = \frac{P(T \leq t \quad \text{et} \quad l < T \leq r)}{P(l < T \leq r)} \tag{1.6}$$

d'où

$$P(T \leq t \mid L = l, R = r, L < T \leq R)$$
$$= \frac{1}{P(T \in (l, r])} \begin{cases} 0, & \text{si } t \leq l \\ F_T(t) - F_T(l) & \text{si } l < t \leq r \\ P(l < T \leq r) & \text{si } t > r. \end{cases} \tag{1.7}$$

En dérivant (1.7) par rapport à t on obtient

$$f_{(T|L,R)}(t|l,r) = \begin{cases} 0, & \text{si } t \leq l \\ \frac{f_T(t)}{P(T \in (l,r])} & \text{si } l < t \leq r \\ 0 & \text{si } t > r, \end{cases}$$

d'où

$$f_{(T|L,R)}(t|l,r) = \frac{f_T(t)}{P(T \in (l,r])} \mathbb{I}_{\{t \in (l,r]\}}. \tag{1.8}$$

La réciproque est immédiate en remontant les calculs de (1.8) à (1.6).

(ii) Pour tous t, l, r tels que $l < t \leq r$, d'après les règles de densité conditionnelle et la condition non informative (1.8), on a

$$\begin{aligned} f_{(L,R|T)}(l,r|t) &= \frac{f_{(T,L,R)}(t,l,r)}{f_T(t)} \\ &= \frac{f_{(T|L,R)}(t|l,r) f_{(L,R)}(l,r)}{f_T(t)} \\ &= \frac{f_T(t) f_{(L,R)}(l,r)}{P(T \in (l,r]) f_T(t)} \\ &= \frac{f_{(L,R)}(l,r)}{P(T \in (l,r])}. \end{aligned} \tag{1.9}$$

Il s'ensuit de l'équation (1.9)

$$\int\int_{\{(l,r):t\in(l,r]\}} \frac{f_{(L,R)}(l,r)}{P(T\in(l,\bar{r}])} dldr = \int\int_{\{(l,r):t\in(l,r]\}} f_{(L,R|T)}(l,r|t)dldr = 1.$$

Par conséquent, la condition de la somme constante est vérifiée.

□

CHAPITRE II

ESTIMATION NON PARAMÉTRIQUE DE LA FONCTION DE SURVIE

Dans ce chapitre, la source principale est Sun [2007]. Nous présentons ici l'estimation non paramétrique de la fonction de survie S pour données censurées par intervalle, où le qualificatif non paramétrique fait référence au fait que nous ne supposons pas un modèle paramétrique quelconque (par exemple, exponentiel, gamma, Weibull, etc.) pour la loi du temps de défaillance. Pour commencer, nous rappelons dans la section 2.1, l'estimateur non paramétrique usuel dans le cas particulier de données censurées à droite : l'estimateur de Kaplan-Meier. Puis dans les sections 2.2 et 2.3 nous présentons deux approches non paramétrique dans le cas général de données censurées par intervalle. La première approche consiste en un algorithme de Kaplan-Meier basé sur des données imputées (imputation par point milieu), puis la deuxième approche est basée sur l'algorithme de Turnbull. À la fin nous comparons les deux approches par une étude de simulation.

2.1 Estimateur de Kaplan-Meier

Dans le cas particulier où les données sont censurées à droite, un bon estimateur non paramétrique de S est l'estimateur de Kaplan-Meier, aussi appelé estimateur produit limite Kaplan et Meier [1958], Klein et Moschberger [2003].

Soit c_i le temps de censure aléatoire pour l'individu i, au lieu d'observer t_i nous observons

le couple (z_i, δ_i), pour le i^e individu où $z_i = \min(t_i, c_i)$ et

$$\delta_i = \begin{cases} 1, \text{si } t_i \leq c_i \\ 0, \text{si } c_i < t_i. \end{cases} \tag{2.1}$$

On appelle l'échantillon des $(z_i, \delta_i)_{i=1,\ldots,n}$ des données censurées à droite.

Considérons une discrétisation du temps formée par les temps d'événements (aussi bien temps de défaillance que censures) et notons $0 = z_{(0)} < z_{(1)} < z_{(2)} < \ldots < z_{(k)}$ les k valeurs distinctes prises dans l'échantillon des (z_i), $i = 1, \ldots, n$ et rangées dans l'ordre croissant. Si par exemple la défaillance correspond à un décès, alors l'estimateur de Kaplan-Meier découle de l'idée que, survivre après un temps $z_{(j)}$, c'est être en vie juste avant $z_{(j)}$ et ne pas mourir au temps $z_{(j)}$, i.e., si $z_{(j-2)} < z_{(j-1)} < z_{(j)}$ on a

$$S(z_{(j)}) = P(\{T > z_{(j)}\} \cap \{T > z_{(j-1)}\}) = P(T > z_{(j)} \mid T > z_{(j-1)}) P(T > z_{(j-1)}).$$

Par récurrence, on obtient

$$\begin{aligned} S(z_{(j)}) &= P(T > z_{(j)} \mid T > z_{(j-1)}) \ldots P(T > z_{(2)} \mid T > z_{(1)}) P(T > z_{(1)}) \\ &= \Pi_{i:z_{(i)} \leq z_{(j)}} P(T > z_{(i)} \mid T > z_{(i-1)}) \\ &= \Pi_{i:z_{(i)} \leq z_{(j)}} (1 - P(T \leq z_{(i)} \mid T > z_{(i-1)})) \\ &= \Pi_{i:z_{(i)} \leq z_{(j)}} \left(1 - \frac{P(z_{(i-1)} < T \leq z_{(i)})}{P(T > z_{(i-1)})}\right) \\ &= \Pi_{i:z_{(i)} \leq z_{(j)}} (1 - q_i), \end{aligned}$$

où $q_i = \frac{P(z_{(i-1)} < T \leq z_{(i)})}{P(T > z_{(i-1)})}$.

Pour estimer S, on estime les q_i à partir des données. Considérons

— $n_j = \sum_{i=1}^{n} \mathbb{I}(z_i \geq z_{(j)})$ qui représente le nombre de sujets à risque à $z_{(j)}$ (ni défaillances, ni censures juste avant $z_{(j)}$),

— $d_j = \sum_{i=1}^{n} \mathbb{I}\left(z_i = z_{(j)}, \delta_i = 1\right)$ qui représente le nombre de temps de défaillance à $z_{(j)}$.

Une estimation de q_j est donnée par

$$q_j = \frac{d_j}{n_j}.$$

Ainsi l'estimateur de la fonction de survie S proposé par Kaplan-Meier est donné par

$$\hat{S}_{KM}(z) = \begin{cases} 1, & \text{si } z < z_{(1)} \\ \prod_{i:z_{(i)} \leq z} \left(1 - \frac{d_i}{n_i}\right), & \text{si } z \geq z_{(1)}. \end{cases} \tag{2.2}$$

L'estimateur de Kaplan-Meier est constant par morceaux, càdlàg (continu à droite avec limite à gauche) et les points de sauts sont les $z_{(j)}$. On déduit de la relation (2.2) que

$$\hat{S}_{KM}(z) = \begin{cases} 1, & \text{si } z < z_{(1)} \\ \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right), & \text{où } j \text{ est défini par } z_{(j)} \leq z < z_{(j+1)}. \end{cases} \tag{2.3}$$

On voit bien que cet estimateur n'est défini que sur $[0, z_{(k)})$ et pas sur $[0, +\infty)$. Ce n'est donc pas une fonction de survie sur $[0, +\infty)$.

2.2 Estimateur de Kaplan-Meier sur données imputées

En général lorsqu'une donnée est manquante, la façon la plus naturelle de procéder consiste à remplacer cette donnée manquante par une valeur supposée l'estimer : c'est ce qu'on appelle de l'imputation (De Waal *et al.* [2011], Carpenter et Kenward [2012]). Dans le contexte de la censure par intervalle, la donnée manquante est l'instant précis de défaillance.

Une pratique courante dans les études médicales et de fiabilité est de simplifier la structure des données censurées par intervalle en une situation de censure à droite en rem-

plaçant les intervalles observés par leur point milieu. Cela s'appelle de l'imputation par point milieu, voir Sun [2007]. En effet, il est plus simple de se ramener à la censure à droite parce que l'on dispose de plus d'outils et techniques dans ce contexte.

Plus précisément, la méthode d'imputation par point milieu consiste à remplacer l'intervalle $(l_i, r_i]$ par

— $z_i^* = \frac{l_i+r_i}{2}$, si r_i est fini ;
— $z_i^* = l_i$, si $r_i = +\infty$.

En posant

$$\delta_i = \begin{cases} 1, \text{si } r_i \text{ est fini} \\ 0, \text{sinon} \end{cases} \tag{2.4}$$

on peut appliquer aux données (z_i^*, δ_i) les techniques pour la censure à droite. Par exemple, on peut estimer la fonction de survie S des T_i par l'estimateur de Kaplan-Meier basé sur ces données imputées. On note cet estimateur $\hat{S}_{KMI}$ et $[0, t_{KMI})$ l'intervalle sur lequel l'estimateur $\hat{S}_{KMI}$ est défini, où t_{KMI} est le max des z_i^*.

Algorithme 1 : Algorithme de Kaplan-Meier basé sur l'imputation par point milieu

1 **pour** $i = 1, \ldots, n$ **faire**
  (i) **Si** $r_i < +\infty$, calculer $z_i^* = \frac{l_i+r_i}{2}$, poser $\delta_i = 1$;
  (ii) **Si** $r_i = +\infty$, calculer $z_i^* = l_i$, poser $\delta_i = 0$
2 **fin**
3 En utilisant (2.3), calculer l'estimateur de Kaplan-Meier $\hat{S}_{KMI}$ de S basé sur le jeu de données censurées à droite (z_i^*, δ_i), pour $i = 1, \ldots, n$.

2.3 Maximisation de la vraisemblance non paramétrique

L'objectif est de trouver un estimateur $\hat{S}_n(t)$ de la fonction de survie S qui maximise la fonction de vraisemblance réduite introduite en (1.4). La maximisation non paramétrique

de la vraisemblance est beaucoup plus difficile avec censure par intervalle qu'avec censure à droite. Deux fonctions de répartition qui prennent les mêmes valeurs en les l_i et r_i, $i = 1, \ldots, n$ ont la même vraisemblance. En particulier, même si l'estimateur du maximum de vraisemblance existe, il peut y avoir des intervalles de temps où la forme de la courbe de survie est ambiguë, du fait que la maximisation de la vraisemblance ne fixe les valeurs de la solution qu'aux l_i, r_i. Par ailleurs, il n'y a pas de forme explicite pour l'estimateur du maximum de vraisemblance. Il existe un algorithme pour obtenir le maximum de vraisemblance non paramétrique de la fonction de survie sous la censure par intervalle. Cet algorithme est appelé algorithme d'auto convergence (« self-consistent algorithm ») et a été suggéré par Turnbull [1976]. Il est basé sur une technique d'optimisation connue sous le nom d'algorithme Espérance - Maximisation (EM) introduite par Dempster *et al.* [1977], méthode qui permet de trouver une approximation de la solution numériquement.

2.3.1 Discrétisation du problème

Considérons une étude de temps de défaillance qui consiste en n sujets indépendants d'une population homogène avec une fonction de survie $S(t)$. Soit t_i, le temps de survie d'intérêt pour le sujet i, $i = 1, \ldots, n$. On note O comme défini en (1.1) l'ensemble des données observables, où $I_i = (l_i, r_i]$ désigne l'intervalle auquel t_i appartient. Pour décrire l'algorithme de Turnbull, nous faisons une partition de $\mathbb{R}_+$, telle que chaque intervalle censuré au cours duquel un événement pourrait se produire est une union d'intervalles de la partition. Désignons par

$$0 = \tau_0 < \tau_1 < \ldots < \tau_{m-1} < \tau_m = \infty, \tag{2.5}$$

les éléments distincts de $\{0, \{l_i\}_{i=1}^n, \{r_i\}_{i=1}^n, \infty\}$ rangés dans l'ordre croissant. En fait, la partition est formée par les intervalles $J_1 = (\tau_0, \tau_1]$, $J_2 = (\tau_1, \tau_2]$, $\ldots$, $J_m = (\tau_{m-1}, \tau_m]$

(voir la figure 2.1). Soit

$$\alpha_{ij} = \mathbb{I}(I_i \ni \tau_j) = \mathbb{I}(J_j = (\tau_{j-1}, \tau_j] \subseteq I_i) = \begin{cases} 1, & \text{si } J_j \subseteq I_i, \\ 0, & \text{sinon,} \end{cases} \tag{2.6}$$

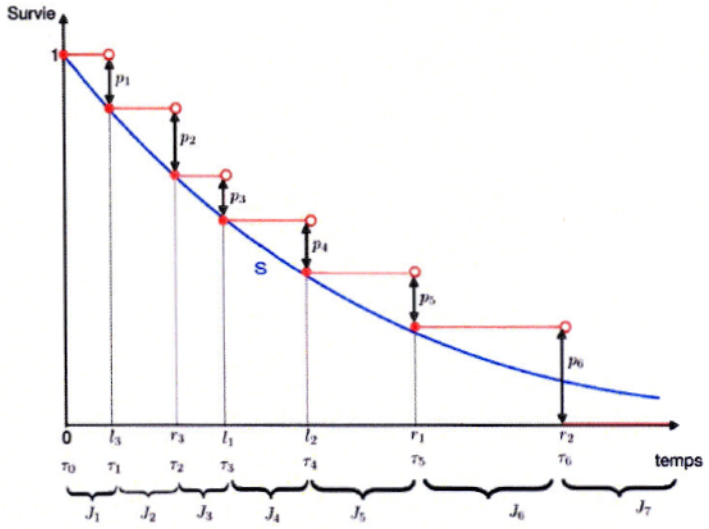
pour $i = 1, \ldots, n$, $j = 1, \ldots, m$. Le poids α_{ij} indique si l'événement qui se produit dans l'intervalle $I_i = (l_i, r_i]$ aurait pu se produire en τ_j. Soient p_j les poids assignés aux classes J_j, $j = 1, \ldots, m$, dans la maximisation de la vraisemblance non paramétrique de S par l'algorithme de Turnbull. On a

$$p_j = p(\tau_{j-1} < T \leq \tau_j) = S(\tau_{j-1}) - S(\tau_j) \geq 0, \tag{2.7}$$

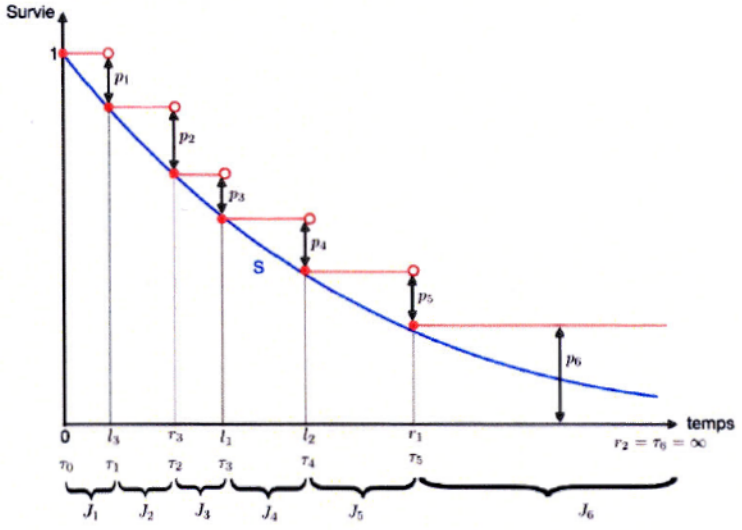
tel que

$$\sum_{j=1}^{m} p_j = S(\tau_0) = 1, \tag{2.8}$$

où $p_j \geq 0$ $(j = 1, \ldots, m)$ (voir la figure 2.1).



(a) Cas où les r_j sont finis (pas de censure à droite). Dans cet exemple, la fonction de survie en escalier est constante égale à zéro à partir de τ_6.



(b) Cas où certains r_j ne sont pas finis (censure à droite). Dans cet exemple, la fonction de survie en escalier est constante (non nulle) entre τ_5 et $\tau_6 = +\infty$.

Figure 2.1 Discrétisation du problème : Représentation des p_j pour une fonction de survie S quelconque (en bleu). En rouge, fonction de survie en escalier qui coïncide avec S en les τ_j. Dans cet exemple $n = 3$.

On peut écrire les intervalles de censure sous la forme d'une union finie d'au plus m intervalles disjoints

$$(l_i, r_i] = \cup_{\{j|\alpha_{ij}=1\}}(\tau_{j-1}, \tau_j]. \tag{2.9}$$

L'intervalle $(l_i, r_i]$ est donc une union finie des $(\tau_{j-1}, \tau_j]$ pour lesquels les $\alpha_{ij} = 1$ (en d'autres termes, les $(\tau_{j-1}, \tau_j]$ sont dans $(l_i, r_i]$). Par conséquent, on a

$$S(l_i) - S(r_i) = \sum_{j=1}^{m} \alpha_{ij}(S(\tau_{j-1}) - S(\tau_j)) = \sum_{j=1}^{m} \alpha_{ij} p_j, \tag{2.10}$$

d'après (2.7). Ceci nous conduit au théorème suivant.

Théorème 2.3.1. *Le problème de maximisation de la fonction de vraisemblance réduite est équivalent au problème de maximisation de*

$$L_\tau(p_1, \ldots, p_m) = \prod_{i=1}^{n} [S(l_i) - S(r_i)] = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} p_j, \tag{2.11}$$

sous les contraintes définies par (2.7) et (2.8), et où α_{ij} est défini en (2.6).

Ce résultat a été mentionné par Peto [1973], puis un peu plus tard formalisé par Turnbull [1976] qui énonce le lemme 2.3.2.

Lemme 2.3.2. *Pour des valeurs fixées de $S(\tau_{j-1})$ et $S(\tau_j)$, $j = 1, \ldots, m$, la fonction de vraisemblance est indépendante du comportement de S à l'intérieur de chaque intervalle $(\tau_{j-1}, \tau_j]$.*

Il découle du lemme 2.3.2 que la fonction de vraisemblance L_τ donnée en (2.11) dépend de S seulement en des valeurs $\{S(\tau_j)\}_{j=1}^{m}$ et non du comportement de S entre les τ_j. En d'autres termes, le maximum de vraisemblance non paramétrique de S peut être

déterminé uniquement en ses valeurs τ_j, et sa détermination est équivalente à maximiser $L_\tau(p)$ par rapport à p sous les contraintes données par (2.7) et (2.8). À partir de l'estimation de vraisemblance maximale $\hat{p}$ de p on a

$$\hat{S}(t) = \begin{cases} 1, & \text{si } t \in [0, \tau_1[, \\ 1 - \sum_{l=1}^{j} \hat{p}_l, & \text{si } \tau_j \leq t < \tau_{j+1} \quad (1 \leq j \leq m-1). \end{cases} \tag{2.12}$$

Pour déterminer une estimation du maximum de vraisemblance non paramétrique, il faudrait maximiser

$$l_\tau(p) = \log L_\tau(p) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \alpha_{ij} p_j \right),$$

sur le sous ensemble de $\mathbb{R}^m$

$$\mathcal{P} = \left\{ p \in [0,1]^m; \sum_{j=1}^{m} p_j = 1, p_j \geq 0 \right\}. \tag{2.13}$$

Or, en notant

$$\begin{aligned} d_j(p) &= \frac{\partial l_\tau(p)}{\partial p_j} \\ &= \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{l=1}^{m} \alpha_{il} p_l}, \end{aligned} \tag{2.14}$$

où $j = 1, \ldots, m$, on se rend compte que trouver une forme explicite pour les points critiques de la log-vraisemblance sous la contrainte (2.13) semble impossible.

2.3.2 Estimateur de Turnbull (EM ou auto convergent)

Une estimation auto convergente se réfère généralement à une estimation qui peut être caractérisée par une équation d'auto convergence et est la limite des itérations obtenue

à partir de cette équation, voir Efron [1967]. Turnbull [1976] a développé un algorithme d'auto convergence pour estimer $S(t)$ avec des données censurées par intervalle. Pour dériver l'équation d'auto convergence pour les données censurées par intervalle, on peut traiter les données censurées par intervalle comme des données incomplètes puis appliquer l'algorithme EM.

2.3.2.1 Les étapes E et M

Dans la dérivation de l'estimateur de Turnbull (EM ou auto convergent) nous introduisons les notations $T_i^{'}$ de T_i pour signifier que l'on fait une approximation et que $T_i^{'}$ suit une loi discrète à valeurs dans $\{\tau_1, \ldots, \tau_m\}$ avec probabilités $p_1, \ldots, p_m$. On définit $T_i^{'}$ à partir de T_i en posant pour $j = 1, \ldots, m$, $\left\{T_i^{'} = \tau_j\right\} = \{T_i \in (\tau_{j-1}, \tau_j]\}$. Voir aussi la figure 2.1.

Pour l'instant, supposons que les données de temps de défaillance exactes $\{t_i\}_{i=1}^n$ sont disponibles. Si on observe $T_i = t$, on peut déterminer dans quel intervalle $(\tau_{j-1}, \tau_j]$ se trouve t, et puis on pose $T_i^{'} = \tau_j$. La fonction de log-vraisemblance pour ces données complètes est

$$\begin{aligned}
\ell_\tau^C(p; t_1^{'}, \ldots, t_n^{'}) &= \sum_{i=1}^{n} \log P(T_i^{'} = t_i^{'}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}(t_i^{'} = \tau_j) \log P(\tau_{j-1} < T_i \leq \tau_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}(t_i^{'} = \tau_j) \log p_j \\
&= \sum_{j=1}^{m} d_j^{'} \log p_j, \qquad (2.15)
\end{aligned}$$

où $d_j^{'} = \sum_{i=1}^{n} \mathbb{I}(t_i^{'} = \tau_j)$, $j = 1, \ldots, m$, est le nombre de personnes avec un temps de défaillance $t_i \in (\tau_{j-1}, \tau_j]$.

En utilisant l'approche de Lagrange à l'équation (2.15) sous la contrainte $p \in \mathcal{P}$, on maximise

$$\varphi(p, \lambda) = \sum_{j=1}^{m} d_j' \log p_j - \lambda \left(\sum_{j=1}^{m} p_j - 1 \right),$$

tel que

$$\frac{\partial \varphi}{\partial p_j} = 0 \Rightarrow \frac{d_j'}{p_j} = \lambda \tag{2.16}$$

$$\frac{\partial \varphi}{\partial \lambda} = 0 \Rightarrow \sum_{j=1}^{m} p_j = 1. \tag{2.17}$$

De (2.16), on a $\sum_{j=1}^{m} p_j = \sum_{j=1}^{m} d_j'/\lambda$, et de (2.17) on a

$$\lambda = \sum_{j=1}^{m} d_j'. \tag{2.18}$$

Ceci donne $\lambda = \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{I}(t_i \in (\tau_{j-1}, \tau_j]) = n$, car pour chaque i on a soit $t_i \in (\tau_0, \tau_1]$, soit $t_i \in (\tau_1, \tau_2]$, ..., soit $t_i \in (\tau_{m-1}, \infty)$.

De (2.16) et (2.18), on obtient

$$p_j = \frac{d_j'}{\sum_{j=1}^{m} d_j'} = \frac{d_j'}{n}, \tag{2.19}$$

de sorte que l'estimation de p_j correspond à la proportion de temps de défaillance qui tombent dans l'intervalle $(\tau_{j-1}, \tau_j]$.

Revenons au cas où seules les $(l_i, r_i]$ sont observées. Soit $\hat{p}^{init}$ une estimation préliminaire de p, où le symbole *init* veut dire initial. Nous présentons d'abord un résultat qui sera utilisé dans l'algorithme EM.

Sachant que $\alpha_{ij} = \mathbb{I}((l_i, r_i] \ni \tau_j)$ et $P(T_i^{'} = \tau_j|\hat{p}^{init}) = \hat{p}_j^{init}$, on a

$$\begin{aligned}
E(d_j^{'}|\hat{p}^{init}, O) &= \sum_{i=1}^{n} E(\mathbb{I}(T_i^{'} = \tau_j)|\hat{p}^{init}, O) \\
&= \sum_{i=1}^{n} 1.P(T_i^{'} = \tau_j|\hat{p}^{init}, O) \\
&= \sum_{i=1}^{n} P(T_i^{'} = \tau_j|\hat{p}^{init}, l_i < T_i^{'} \leq r_i) \\
&= \frac{\sum_{i=1}^{n} P(T_i^{'} = \tau_j, l_i < T_i^{'} \leq r_i|\hat{p}^{init})}{P(l_i < T_i^{'} \leq r_i|\hat{p}^{init})} \\
&= \sum_{i=1}^{n} \frac{\alpha_{ij}\hat{p}_j^{init}}{\hat{S}^{init}(l_i) - \hat{S}^{init}(r_i)} \\
&= \sum_{i=1}^{n} \frac{\alpha_{ij}\hat{p}_j^{init}}{\sum_{l=1}^{m} \alpha_{il}\hat{p}_l^{init}} \\
&= d_j(\hat{p}^{init})\hat{p}_j^{init}, \qquad (2.20)
\end{aligned}$$

où $d_j(p)$ est défini en (2.14). Voici les étapes de l'algorithme EM :

- À l'étape E : On calcule l'espérance conditionnelle de $\ell_\tau^C(p; T_1^{'}, \ldots, T_n^{'})$ étant donné $\hat{p}^{init}$ et les données observées O, qui prend la forme (voir (2.15))

$$\begin{aligned}
E\left[\ell_\tau^C(p; T_1^{'}, \ldots, T_n^{'})|\hat{p}^{init}, O\right] &= \sum_{j=1}^{m} E(d_j^{'}|\hat{p}^{init}, O) \log p_j \\
&= \sum_{j=1}^{m} d_j(\hat{p}^{init})\hat{p}_j^{init} \log p_j, \qquad (2.21)
\end{aligned}$$

d'après l'équation (2.20).

- À l'étape M : Soit $\hat{p}^{maj}$ l'estimation mise à jour de p, où le symbole maj signifie mise à jour. Il faut maximiser l'espérance conditionnelle donnée en (2.21) sur la région $\mathcal{P}$ (voir (2.13)). Pour maximiser l'espérance conditionnelle donnée en (2.21) sur la région

$\mathcal{P}$, on a par analogie avec (2.19)

$$\hat{p}_j^{maj} = \frac{d_j(\hat{p}^{init})\hat{p}_j^{init}}{\sum_{j=1}^{m} d_j(\hat{p}^{init})\hat{p}_j^{init}}. \tag{2.22}$$

Or d'après (2.14), on a

$$\sum_{j=1}^{m} d_j(\hat{p}^{init})\hat{p}_j^{init} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\alpha_{ij}\hat{p}_j^{init}}{\sum_{l=1}^{m}\alpha_{il}\hat{p}_l^{init}} = n.$$

Par conséquent, (2.22) devient

$$\hat{p}_j^{maj} = \frac{d_j(\hat{p}^{init})\hat{p}_j^{init}}{n}. \tag{2.23}$$

D'après (2.14), l'équation (2.23) devient

$$\hat{p}_j^{maj} = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_{ij}\hat{p}_j^{init}}{\sum_{l=1}^{m}\alpha_{il}\hat{p}_l^{init}}, \tag{2.24}$$

ce qui conduit à l'algorithme de Turnbull pour le maximum de vraisemblance non paramétrique.

Remarque 2.3.3. *Notez que $\hat{p}^{maj}$ est une mise à jour de $\hat{p}^{init}$. En résumé, les étapes E et M de l'algorithme sont telles que si nous avons $\hat{p}^{init}$ une estimation préliminaire de p, elle peut être mise à jour en calculant $\hat{p}^{maj}$. Les étapes E et M de l'algorithme représentent une étape de mise à jour.*

2.3.2.2 Convergence de l'algorithme de Turnbull

Turnbull [1976] définit la notion d'auto convergence de la façon suivante.

Définition 2.3.4. *Un vecteur $p = (p_1, \ldots, p_m)$ est dit auto convergent si*

$$p_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k}, \; j = 1, \ldots, m. \tag{2.25}$$

Remarque 2.3.5. *On remarque qu'un point fixe de l'algorithme EM est auto convergent (comparer (2.24) et (2.25)).*

Supposons maintenant que nous augmentons un p_j donné de $\epsilon > 0$; alors, on doit diviser tout les $p_k (k = 1, \ldots, m$, pour $k \neq j)$ et $p_j + \epsilon$, par $1 + \epsilon$ afin de maintenir la somme des $p_k (k = 1, \ldots, m)$ égale à 1. On définit $e_j(p)$ la dérivée de l_τ (voir (2.14)) par rapport à ϵ, évaluée à $\epsilon = 0$, i.e.,

$$e_j(p) = \frac{\partial l_\tau}{\partial \epsilon} \left(\frac{p_1}{1+\epsilon}, \ldots, \frac{p_{j-1}}{1+\epsilon}, \frac{p_j + \epsilon}{1+\epsilon}, \frac{p_{j+1}}{1+\epsilon}, \ldots, \frac{p_m}{1+\epsilon} \right) |_{\epsilon=0}. \tag{2.26}$$

Théorème 2.3.1. *Turnbull [1976]*

1 *Si $\hat{p}$ est un estimateur du maximum de vraisemblance de p, alors $\hat{p}$ satisfait l'équation d'auto convergence (2.25).*

2 *Inversement, la solution $\hat{p}$ de l'équation d'auto convergence (2.25) est un estimateur non paramétrique du maximum de vraisemblance de p à condition que $e_j(p) \leq 0$ lorsque $p_j = 0$.*

La clé du théorème 2.3.1 est la proposition suivante

Proposition 2.3.6. *Turnbull [1976]*

De l'équation (2.26) , pour $j = 1, \ldots, m$ on a

$$e_j(p) = d_j(p) - \sum_{k=1}^{m} p_k d_k(p), \tag{2.27}$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k} = \left(1 + \frac{e_j(p)}{n} \right) p_j. \tag{2.28}$$

Démonstration. En prenant la dérivée de la fonction composée de l_τ par rapport à p_k (voir (2.26)), on a

$$\begin{aligned} e_j(p) &= \sum_{k=1}^{m} \frac{\partial l_\tau}{\partial p_k} \frac{\partial p_k}{\partial \epsilon}|_{\epsilon=0} \\ &= \frac{\partial l_\tau}{\partial p_1}\left[\frac{\partial}{\partial \epsilon}\left(\frac{p_1}{1+\epsilon}\right)\right] + \ldots + \frac{\partial l_\tau}{\partial p_j}\left[\frac{\partial}{\partial \epsilon}\left(\frac{p_j+\epsilon}{1+\epsilon}\right)\right] + \frac{\partial l_\tau}{\partial p_{j+1}}\left[\frac{\partial}{\partial \epsilon}\left(\frac{p_{j+1}}{1+\epsilon}\right)\right] \\ &+ \ldots + \frac{\partial l_\tau}{\partial p_m}\left[\frac{\partial}{\partial \epsilon}\left(\frac{p_m}{1+\epsilon}\right)\right]|_{\epsilon=0} \\ &= \frac{\partial l_\tau}{\partial p_1}\left(-\frac{p_1}{(1+\epsilon)^2}\right) + \ldots + \frac{\partial l_\tau}{\partial p_j}\left[\frac{\partial}{\partial \epsilon}\left(\frac{p_j}{1+\epsilon}\right) + \frac{\partial}{\partial \epsilon}\left(\frac{\epsilon}{1+\epsilon}\right)\right] \\ &+ \frac{\partial l_\tau}{\partial p_{j+1}}\left(-\frac{p_{j+1}}{(1+\epsilon)^2}\right) + \ldots + \frac{\partial l_\tau}{\partial p_m}\left(-\frac{p_m}{(1+\epsilon)^2}\right)|_{\epsilon=0} \end{aligned} \tag{2.29}$$

et de l'équation (2.29) on a

$$\begin{aligned} e_j(p) &= \frac{\partial l_\tau}{\partial p_1}\left(-\frac{p_1}{(1+\epsilon)^2}\right) + \ldots + \frac{\partial l_\tau}{\partial p_j}\left[-\frac{p_j}{(1+\epsilon)^2} + \frac{(1+\epsilon)-\epsilon}{(1+\epsilon)^2}\right] \\ &+ \frac{\partial l_\tau}{\partial p_{j+1}}\left(-\frac{p_{j+1}}{(1+\epsilon)^2}\right) + \ldots + \frac{\partial l_\tau}{\partial p_m}\left(-\frac{p_m}{(1+\epsilon)^2}\right)|_{\epsilon=0} \\ &= \sum_{k=1}^{m} \frac{\partial l_\tau}{\partial p_k}\left(-\frac{p_k}{(1+\epsilon)^2}\right) + \frac{\partial l_\tau}{\partial p_j}\left[\frac{(1+\epsilon)-\epsilon}{(1+\epsilon)^2}\right]|_{\epsilon=0} \\ &= -\sum_{k=1}^{m} p_k \frac{\partial l_\tau}{\partial p_k} + \frac{\partial l_\tau}{\partial p_j} \\ &= d_j(p) - \sum_{k=1}^{m} p_k d_k(p), \end{aligned} \tag{2.30}$$

par conséquent de (2.30) on a

$$e_j(p) = -\underbrace{\sum_{k=1}^{m}\sum_{i=1}^{n} \frac{\alpha_{ik}p_k}{\sum_{l=1}^{m}\alpha_{il}p_l}}_{=n} + \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{k=1}^{m}\alpha_{ik}p_k}, \quad \text{voir (2.14)}$$

$$\Rightarrow e_j(p)p_j = -np_j + \sum_{i=1}^{n} \frac{\alpha_{ij}p_j}{\sum_{k=1}^{m}\alpha_{ik}p_k}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_{ij}p_j}{\sum_{k=1}^{m}\alpha_{ik}p_k} = \left(1+\frac{e_j(p)}{n}\right)p_j. \tag{2.31}$$

Donc, de l'équation (2.31), on en déduit que si $\hat{p}$ est un maximum de vraisemblance alors

$$e_j(\hat{p}) = 0 \text{ ou } (e_j(\hat{p}) \leq 0 \text{ avec } \hat{p}_j = 0), \text{ pour } j = 1, \ldots, m. \tag{2.32}$$

Ainsi de (2.31)-(2.32), on en déduit que le maximum de vraisemblance $\hat{p}$ satisfait

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_{ij}\hat{p}_j}{\sum_{k=1}^{m}\alpha_{ik}\hat{p}_k} = \hat{p}_j, \text{ pour tout } j,$$

et par conséquent est auto convergent. Inversement, si l'algorithme converge avec une valeur limite $\hat{p}$, alors $\hat{p}$ doit satisfaire (2.32). Un argument de continuité montre que nous ne pouvons pas avoir $e_j(\hat{p}) > 0$ avec $\hat{p}_j = 0$. Voir aussi Gómez *et al.* [2004] (p.149). □

Concernant la convergence de l'algorithme de Turnbull, soit p et p' des approximations successives. De l'équation (2.28) on a

$$p_j' = \left(1+\frac{e_j(p)}{n}\right)p_j, \quad \text{pour } 1 \leq j \leq m. \tag{2.33}$$

Cette équation permet de faire l'analogie avec un estimateur de type produit limite comme l'estimateur de Kaplan-Meier. En utilisant le développement de Taylor, on a

$$
\begin{aligned}
l_\tau(p') - l_\tau(p) &= \sum_{j=1}^{m}(p_j' - p_j)\frac{\partial l_\tau}{\partial p_j} + O(\|p' - p\|^2) \\
&\simeq \frac{1}{n}\sum_{j=1}^{m} p_j e_j(p)\frac{\partial l_\tau}{\partial p_j}, \quad \text{voir (2.33)} \\
&= \frac{1}{n}\sum_{j=1}^{m} p_j d_j(p)\left(d_j(p) - \sum_{k=1}^{m} p_k d_k(p)\right), \quad \text{voir (2.14) et (2.27)} \\
&= \frac{1}{n}\sum_{j=1}^{m} p_j d_j(p)^2 - \frac{1}{n}\left(\sum_{k=1}^{m} p_k d_k(p)\right)^2, \\
&= \frac{1}{n}\sum_{j=1}^{m} p_j \left(d_j(p) - \sum_{k=1}^{m} p_k d_k(p)\right)^2, \qquad (2.34) \\
&= \frac{1}{n}\sum_{j=1}^{m} p_j e_j(p)^2 \geq 0, \quad \text{voir (2.27)} \qquad (2.35)
\end{aligned}
$$

où nous avons négligé les termes de second degré et d'ordre supérieur. Notez que les quatrième et cinquième lignes du calcul précédent ne sont autres que les formules usuelles de la variance. Une autre manière de le voir serait de développer l'équation (2.34) et d'utiliser le fait que $\sum_{j=1}^{m} p_j = 1$.

De l'équation (2.35) on a $l_\tau(p') \geq l_\tau(p)$, avec égalité seulement si, pour chaque j, soit $p_j = 0$ ou $d_j(p) = 0$. Ainsi la vraisemblance croît à chaque itération, au moins pour $p^{(0)}$ assez proche de $\hat{p}$ pour que les termes d'ordre supérieur puissent en effet être négligés. Il est clair que nous devons choisir $\hat{p}_j^{(0)} > 0$ (sinon $\hat{p}_j^{(k)} = 0$ pour tout k). En effet à l'étape d'initialisation de l'algorithme de Turnbull, on choisit les valeurs initiales $\hat{p}^{(0)}$ de p. Celles-ci peuvent être n'importe quel ensemble de nombres positifs sommant à l'unité, par exemple $\hat{p}_j^{(0)} = 1/m$, pour $j = 1, \ldots, m$.

2.3.2.3 Bilan de l'algorithme de Turnbull

Dans cette section, nous faisons un bilan des étapes de la dérivation de l'algorithme de Turnbull (voir algorithme 2) données à la section 2.3.2.1. On construit une suite $\hat{p}^{(k)}$

pour approcher l'estimation de vraisemblance maximale de p, en itérant plusieurs fois les étapes E et M de la partie 2.3.2.1 jusqu'à convergence, c'est-à-dire jusqu'à ce qu'un critère d'arrêt soit vérifié. Nous proposons deux types de critères : l'un est basé sur la stabilisation de la vraisemblance (voir algorithme 2, ligne 7), l'autre est basé sur la stabilisation de la suite $\hat{p}^{(k)}$ (voir algorithme 2, ligne 8). Dans nos simulations, nous utiliserons le premier critère. On en déduit l'estimateur de Turnbull $\hat{S}_T$ de S défini à l'équation (2.12). On définit t_T le plus grand des τ_j finis.

2.4 Construction de données simulées

Dans ce qui suit nous allons faire une étude de simulation afin de comparer les deux méthodes, Turnbull et Kaplan-Meier sur données imputées. Nous allons calculer la distance entre $\hat{S}$ et S pour $n = 250$. Nous expliquons ici comment générer des données censurées par intervalle de telle sorte que les données simulées soient non informatives par rapport à la variable principale de temps d'intérêt au sens décrit dans (1.2). Nous décrivons également la simulation des temps de défaillance selon un taux de survie non conventionnel, pour ainsi mieux correspondre à celui sous-jacent aux données réelles.

2.4.1 Construction du taux de survie

On appelle taux de survie du temps de défaillance T, la fonction h définie par $h(x) = f(x)/S(x)$, avec la convention $h(x) = 0$ lorsque $S(x) = 0$.

Nous allons définir le taux de survie à partir d'un modèle de mélange (voir Marshall et Olkin [2007], $p.120 - p.126$). On a choisi un mélange à deux classes parce qu'on a deux types de patients. Le premier type de patient est susceptible de développer le surpoids dû au traitement et le deuxième non.

La construction du taux de survie se fait ainsi. On a une densité f de la forme

$$f(x) = pf_1(x) + (1-p)f_2(x), \tag{2.36}$$

Algorithme 2 : L'estimateur de Turnbull « auto convergent »

Entrées : À partir d'un ensemble de données, nous construisons la partition de $\mathbb{R}_+$ et α une matrice $(n \times m)$ avec des coefficients α_{ij} tels que décrits en (2.6).

Sortie : On obtient un vecteur $\hat{p}$ de dimension m qui représente les sauts de $\hat{S}_T$ en τ_j, pour $j = 1, \ldots, m$. On a finalement un estimateur non paramétrique de S.

1ère étape de l'algorithme : Calculer les τ_j, pour $j = 1, \ldots, m$;

Initialisation : $\hat{p}^{(0)} = (1/m, \ldots, 1/m)$, $k = 0$;

tant que *les conditions d'arrêt ne sont pas atteintes* **faire**

  pour $j = 1, \ldots, m$ **faire**

$$\hat{p}_j^{(k+1)} = \frac{d_j(\hat{p}^{(k)})\hat{p}_j^{(k)}}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_{ij}\hat{p}_j^{(k)}}{\sum_{l=1}^{m}\alpha_{il}\hat{p}_l^{(k)}};$$

  $k := k + 1$;

  Condition d'arrêt

$$l_\tau(\hat{p}^{(k+1)}) - l_\tau(\hat{p}^{(k)}) = \sum_{i=1}^{n} \log\left(\frac{\sum_{j=1}^{m}\alpha_{ij}\hat{p}_j^{(k+1)}}{\sum_{j=1}^{m}\alpha_{ij}\hat{p}_j^{(k)}}\right) < \epsilon, \quad \text{pour } \epsilon > 0.$$

  Une autre condition d'arrêt pourrait être

$$\sum_{j=1}^{m}(\hat{p}_j^{(k+1)} - \hat{p}_j^{(k)})^2 < \epsilon, \text{ ou } \max_{j=1,\ldots,m}|\hat{p}_j^{(k+1)} - \hat{p}_j^{(k)}| < \epsilon \quad \text{pour } \epsilon > 0,$$

  fin

fin

pour $0 \leq t < \tau_1$ **faire**

  $\hat{S}_T(t) = 1$

fin

pour $j = 1 \ldots m - 1$ **faire**

  pour $\tau_j \leq t < \tau_{j+1}$ **faire**

    $\hat{S}_T(t) = 1 - \sum_{l=1}^{j}\hat{p}_l^{(k+1)}$

  fin

fin

où f_1 et f_2 sont des densités connues et $p \in [0,1]$ (proportion du mélange). Puis on calcule la fonction de survie

$$S(x) = pS_1(x) + (1-p)S_2(x), \tag{2.37}$$

où S_1 et S_2 sont les fonctions de survies de f_1 et f_2, respectivement. Le taux de survie est donné par

$$h(x) = \frac{f(x)}{S(x)} = \frac{pf_1(x) + (1-p)f_2(x)}{pS_1(x) + (1-p)S_2(x)}. \tag{2.38}$$

Ici, on choisit S_1 de la forme

$$S_1(x) = S_0(x)S_2(x). \tag{2.39}$$

Il s'agit de la fonction de survie d'un temps minimum, voir aussi l'algorithme 4.

Remarque 2.4.1. *Certains patients ont un gain de poids rapide durant la phase intensive du traitement basé sur les corticostéroïdes, d'autres patients auront plutôt un gain de poids uniquement lié à une augmentation de leur âge. D'où le choix de la forme de la fonction de survie S_1 (pour le premier type de patient) donnée à l'équation (2.39). Ici S_0 représente la loi du temps d'apparition du surpoids due au traitement, puis S_2 la loi du temps d'apparition du surpoids due à une augmentation de l'âge.*

En substituant l'équation (2.39) dans (2.37), on obtient

$$\begin{aligned} S(x) &= pS_0(x)S_2(x) + (1-p)S_2(x) \\ &= S_2(x)(pS_0(x) + (1-p)). \end{aligned} \tag{2.40}$$

D'autre part, par définition et en vertu de (2.39) on a

$$f_1(x) = -S_1'(x) = -S_0'(x)S_2(x) - S_0(x)S_2'(x)$$
$$= f_0(x)S_2(x) + S_0(x)f_2(x). \tag{2.41}$$

En substituant (2.41) dans (2.36) on a

$$f(x) = f_2(x)(1 - p + pS_0(x)) + pf_0(x)S_2(x). \tag{2.42}$$

Puis en substituant (2.42) et (2.40) dans (2.38) on obtient

$$\begin{aligned} h(x) &= \frac{f_2(x)}{S_2(x)} + \frac{pf_0(x)S_2(x)}{S_2(x)(pS_0(x) + 1 - p)} \\ &= h_2(x) + \frac{ph_0(x)S_0(x)}{(pS_0(x) + 1 - p)}, \end{aligned} \tag{2.43}$$

où $h_2 = \frac{f_2}{S_2}$, $h_0 = \frac{f_0}{S_0}$, et S_0 est la fonction de survie de f_0. Ce qui conduit à la dérivation de l'algorithme suivant.

Algorithme 3 : taux de survie à partir d'un modèle de mélange

Entrées : $p \in [0,1]$ (proportion du mélange) et $x \in \mathbb{R}$ (point où on évalue la fonction), (f_0, S_0), (f_2, S_2)

Sortie : Taux de survie $h(x)$

1. **Densité :** Calculer $f(x) = pf_1(x) + (1-p)f_2(x)$, où f_1 est donnée par (2.41);
2. **Fonction de survie :** Calculer $S(x) = pS_1(x) + (1-p)S_2(x)$, où S_1 est donnée par (2.39);
3. **Taux de survie :** Calculer $h(x) = h_2(x) + \frac{ph_0(x)S_0(x)}{(pS_0(x)+1-p)}$, où $h_2 = \frac{f_2}{S_2}$, $h_0 = \frac{f_0}{S_0}$, et S_0 est la fonction de survie de f_0 (voir (2.43)).

Exemple 2.4.2. *Dans cet exemple on propose un choix de S_0 et S_2 qui vérifie la remarque 2.4.1. On a pris pour S_0 la fonction de survie d'une loi Log-normale recentrée et renormalisée telle que $S_0(x) = 1 - G_0(\frac{x-m}{\sigma})$, où $m = 1/2$, $\sigma = \{2,3,4,5,6,7\}$ et G_0 est*

la fonction de répartition de la loi Log-normale de paramètre de position 0 *et de paramètre d'échelle* $1/2$. *Ensuite on a posé* $S_2(x) = 1 - G_2(x/15)$ *la fonction de survie d'une loi de Weibull renormalisée, où* G_2 *est la fonction de répartition de la loi de Weibull de paramètre de forme* 2.572 *et de paramètre d'échelle* 4.537. *Ainsi, en prenant* $p = 0.266$, *on obtient les figures suivantes de la fonction de survie* S_0 *et la fonction de survie* S_2 *(voir figure 2.2), la fonction de survie* S *(2.40) (voir figure 2.3), la fonction du taux de survie* h *(voir figure 2.4) et la densité* f *(2.42) (voir figure 2.5).*

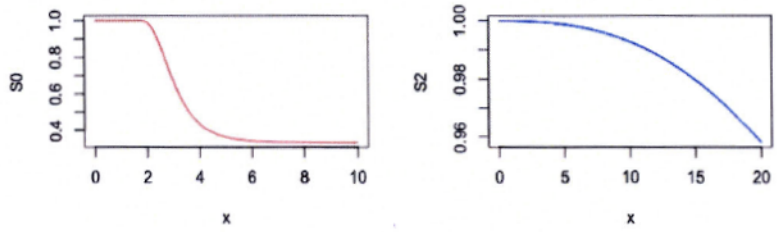


Figure 2.2 Fonction de survie S_0 et S_2

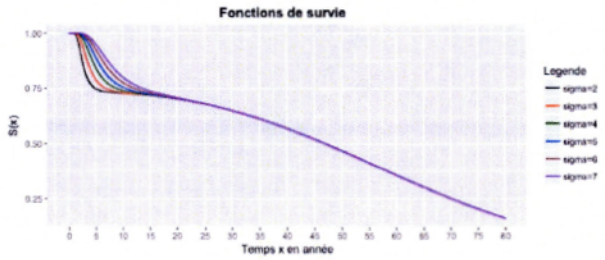


Figure 2.3 Fonction de survie obtenue à partir d'un modèle de mélange

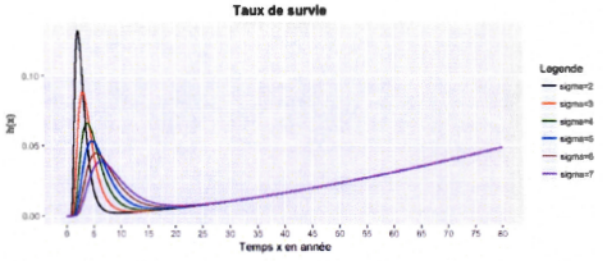


Figure 2.4 Taux de survie obtenu à partir d'un modèle de mélange

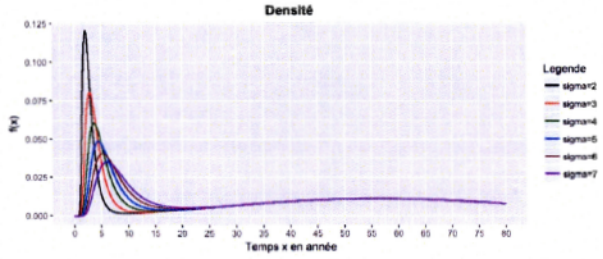


Figure 2.5 Densité f obtenue à partir d'un modèle de mélange

2.4.2 Simulation du temps de défaillance

On rappelle que si $X_0 \sim S_0$ et $X_2 \sim S_2$ sont deux variables aléatoires indépendantes, alors $X_1 = \min(X_0, X_2) \sim S_1$, où S_1 est défini par (2.39). Dans ce qui suit, nous présentons l'algorithme de génération de temps de défaillance pour le taux de survie donné en (2.43).

2.4.3 Simulation de données censurées par intervalle

On suppose que l'on possède un échantillon $t_1, \ldots, t_n$ généré selon l'algorithme 4. Pour générer les intervalles $(l_1, r_1], \ldots, (l_n, r_n]$, le mécanisme de censure de T imite une étude longitudinale avec un suivi périodique et des visites programmées, et ceci selon un modèle inspiré de Schick et Yu [2000]. Dans ce cas, et pour chaque individu i, on définit :

— les temps d'inter-suivi indépendants $\xi_{i1} = G_{i2} - G_{i1}$ et $\xi_{i2} = G_{i3} - G_{i2}$

Algorithme 4 : temps de défaillance pour le taux de survie (2.43)

Entrées : $p \in [0,1]$ (proportion du mélange), S_0 (loi du temps d'apparition du surpoids due au traitement) et S_2 (loi du temps d'apparition du surpoids due à une augmentation de l'âge)

Sortie : temps de défaillance T

1 **On tire :** $X_0 \sim S_0$ et $X_2 \sim S_2$, où X_0 et X_2 sont indépendantes ;

2 **On simule** $Z \sim \mathcal{B}(p)$;

3 **Si** $Z = 1$, on pose $T = \min(X_0, X_2) \sim S_1$;

4 **Si** $Z = 0$, on pose $T = X_2 \sim S_2$.

— les temps écoulés entre chaque rendez-vous et le premier rendez-vous $v_{i0} = 0$, $v_{i1} = G_{i2} - G_{i1} = \xi_{i1}$ et $v_{i2} = G_{i3} - G_{i1} = \xi_{i1} + \xi_{i2}$

— les bornes des intervalles observés

$$l_i = \max\{v_{ia}, a \in \{0,1,2\} : v_{ia} < t_i\}, \text{ et } r_i = \min\{v_{ia}, a \in \{1,2,3\} : v_{ia} \geq t_i\},$$

où $v_{i3} = \infty$. Le paramètre $E(\xi_{ij}) = \mu$, $j = 1,2$ garantit un contrôle du pourcentage d'observations censurées à droite.

2.5 Étude de simulation

Nous présentons ici les résultats de l'étude de simulation conçue pour évaluer la performance de l'algorithme 2, tout en comparant les deux méthodes i.e., l'algorithme 2 et l'algorithme 1 pour $n = 250$. Cette étude de simulation a été conçue en nous basant sur des connaissances a priori sur les données réelles étudiées dans le chapitre suivant (les simulations ont été réalisées avant d'avoir accès à ces données).

2.5.1 Description de l'étude

Le temps de défaillance T est simulé tel que décrit à la section 2.4.2 en utilisant les paramètres p, S_0 et S_2 de l'exemple 2.4.2 avec $\sigma = 7$.

La loi du 1er inter-suivi (entre le diagnostic v_{i0} et le 1er rendez-vous v_{i1}) sera fixe pour la suite : il s'agit de la loi uniforme sur l'intervalle $(2-5/12, 2+5/12)$. Cette loi est centrée en 2 et l'amplitude de son support est de 10 mois. Le 1er rendez-vous intervient donc en moyenne 2 ans après le diagnostic. On fera varier la loi du deuxième inter-suivi (entre le 1er rendez-vous G_2 et le 2e rendez-vous G_3) afin d'étudier l'impact de la longueur du deuxième intervalle sur le comportement des méthodes. Une expérience correspond au choix de la loi du deuxième inter-suivi et de ses paramètres. On répète 500 fois chaque expérience avec un échantillon de taille 250.

2.5.1.1 Loi du deuxième inter-suivi

Les densités des lois utilisées dans les simulations sont données dans la figure 2.6.

Premier cas : loi uniforme. On donne dans le tableau 2.1, les paramètres des lois uniformes utilisées dans l'étude de simulation, et notées U1, U2 et U3.

Deuxième cas : loi exponentielle tronquée. La loi du 2e inter-suivi est la loi exponentielle tronquée d'intensité λ et de support $[a, b]$, avec $a = 0$, $b = 40$. Elle a pour densité

$$f(x) = \begin{cases} \frac{g(x)}{G(b)-G(a)}, & \text{si } a \leq x \leq b \\ 0, & \text{sinon} \end{cases} \tag{2.44}$$

où g et G sont la densité et la fonction de répartition de la loi exponentielle de paramètre λ, respectivement Nadarajah *et al.* [2006]. L'espérance de la loi du 2e inter-suivi est donnée par

$$E = \frac{1}{\lambda} - \frac{b\exp(-\lambda b) - a\exp(-\lambda a)}{\exp(-\lambda a) - \exp(-\lambda b)}. \tag{2.45}$$

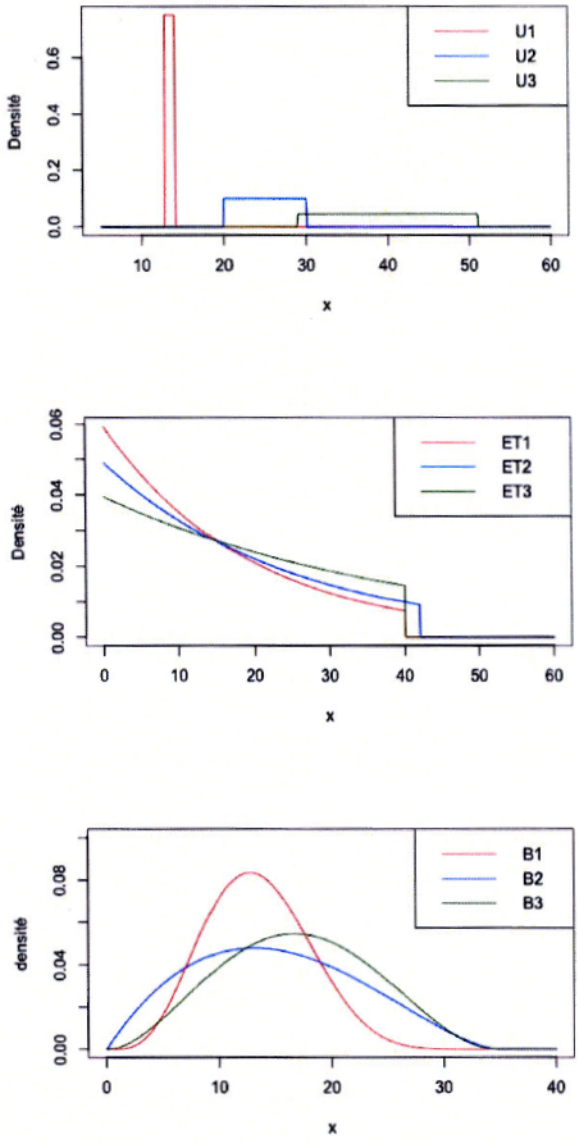


Figure 2.6 Densités des lois des 2^e inter rendez-vous utilisées dans les simulations. En haut loi uniforme, au milieu loi exponentielle tronquée et en bas loi bêta renormalisée.

On peut simuler la loi exponentielle tronquée à partir d'une loi exponentielle usuelle en utilisant l'algorithme 5.

Algorithme 5 : Simulation du deuxième inter rendez-vous selon la loi exponentielle tronquée

Entrées : n (taille de l'échantillon), $a = 0$, $b = 40$, λ (paramètre à spécifier)

1 **Tirer :** $u_1, \ldots, u_n$ i.i.d. de loi uniforme sur $(0, 1)$;

2 **Calculer :** $\xi_{i2} = G^{-1}(G(a) + u_i \times (G(b) - G(a)))$, où G est la fonction de répartition de la loi exponentielle de paramètre λ et G^{-1} son inverse.

On donne dans le tableau 2.2, les paramètres des lois exponentielles tronquées utilisées dans l'étude de simulation, et notées ET1, ET2 et ET3.

Troisième cas : loi bêta. La loi du 2e inter-suivi est la loi bêta de paramètre de forme a et d'échelle b, renormalisée pour être à support dans $[0, 35]$ au lieu de $[0, 1]$. Cette loi a pour espérance $E = 35\frac{a}{a+b}$. On donne dans le tableau 2.3, les paramètres des lois Beta renormaliées utilisées dans l'étude de simulation, et notées B1, B2 et B3.

2.5.2 Mesures de distance

On aimerait savoir lequel de deux estimateurs est le meilleur autrement dit le plus proche de la fonction de survie S. Pour cela on a besoin de définir des distances entre fonction de survies. La première distance utilisée ici est inspirée du test de Kolmogorov-Smirnov (voir Thas [2010], p. 124-125). Étant donné $0 < q^\star \leq +\infty$ et $\hat{S}$ un estimateur de S, on définit

$$d_{KS}(S, \hat{S}, q^\star) = \sup_{0 \leq t \leq q^\star} |\hat{S}(t) - S(t)|.$$

Si de plus, il existe une partition $0 = q_{(1)} < \ldots < q_{(p)} = q^\star$ telle que $\hat{S}$ est constante sur chaque intervalle $[q_{(i)}, q_{(i+1)}[$, alors

Tableau 2.1 Paramètres des lois uniformes du deuxième inter-suivi

Exemple	Support	Espérance (en année)	Amplitude du support
U1	$13.5 \pm 8/12$	13.5	16 mois
U2	25 ± 5	25	10 ans
U3	40 ± 11	40	22 ans

Tableau 2.2 Paramètres des lois exponentielles tronquées du deuxième inter-suivi

Exemple	a	b	λ	Espérance (en année)
ET1	0	40	0.052	13.5
ET2	0	42	0.04	15
ET3	0	40	0.025	17

Tableau 2.3 Paramètres des lois Beta renormalisées du deuxième inter-suivi

Exemple	Support	a	b	Espérance (en année)
B1	$[0, 35]$	5	8	13.5
B2	$[0, 35]$	2	8/3	15
B3	$[0, 35]$	3	3.2	17

$$d_{KS}(S,\hat{S},q^\star) = \max_{0\le i\le p-1}\left\{\sup_{q_{(i)}\le t<q_{(i+1)}}(\hat{S}(t)-S(t)); \sup_{q_{(i)}\le t<q_{(i+1)}}(S(t)-\hat{S}(t))\right\}.$$

Comme $\hat{S}$ est constante sur chaque intervalle $[q_{(i)}, q_{(i+1)}[$ et que $-S$ est croissante, alors

$$\begin{aligned}\sup_{q_{(i)}\le t<q_{(i+1)}}(\hat{S}(t)-S(t)) &= \sup_{q_{(i)}\le t<q_{(i+1)}}(\hat{S}(q_{(i)})-S(t))\\ &= \hat{S}(q_{(i)})-S(q_{(i+1)}),\end{aligned}$$

qui correspond à l'écart entre les points $(q_{(i+1)}, S(q_{(i+1)}))$ et $(q_{(i+1)}, \hat{S}(q_{(i)}))$. De même, comme $\hat{S}$ est constante sur chaque intervalle $[q_{(i)}, q_{(i+1)}[$ et que S est décroissante, alors

$$\sup_{q_{(i)}\le t<q_{(i+1)}}(S(t)-\hat{S}(t)) = S(q_{(i)})-\hat{S}(q_{(i)}),$$

qui correspond à l'écart entre les points $(q_{(i)}, S(q_{(i)}))$ et $(q_{(i)}, \hat{S}(q_{(i)}))$. En conclusion, on a

$$d_{KS}(S,\hat{S},q^\star) = \max\left\{\max_{0\le i\le p-1}\left(\hat{S}(q_{(i)})-S(q_{(i+1)})\right); \max_{0\le i\le p-1}\left(S(q_{(i)})-\hat{S}(q_{(i)})\right)\right\}.$$

Notons que $\hat{S}_{KMI}$ et $\hat{S}_T$ sont constants par morceaux, sur des partitions différentes de $(0,+\infty)$.

La deuxième distance utilisée est la distance renormalisée associée à la norme L_2 sur l'intervalle $(0,q^\star)$, où $q^\star$ est un réel strictement positif. On note

$$d_2(S,\hat{S},q^\star) = \sqrt{\int_0^{q^\star}\left((S(t)-\hat{S}(t))^2/q^\star\right)dt}. \tag{2.46}$$

Dans la pratique on choisit $q^\star = t_T$ (voir section 2.3.2.3) ou $q^\star = t_{KMI}$ (voir section 2.2), et on approche l'équation (2.46) par une somme de Riemann.

2.6 Résultats de simulation

Dans cette section, on présente les résultats des simulations pour l'étude qu'on a décrite à la section 3.1. On donne dans la suite pour chaque cas le tableau de résumé statistique des distances entre l'estimateur $\hat{S}_T$ et la vraie fonction de survie S, puis entre l'estimateur $\hat{S}_{KMI}$ et la vraie fonction de survie S.

Nous présentons des notations qui seront utilisées dans la légende des figures de la partie résultats de simulation. On notera les distances d_{KS} entre $\hat{S}_T$ et S, puis entre $\hat{S}_{KMI}$ et S respectivement par

- $D\_Turnbull = d_{KS}(S, \hat{S}_T, \infty)$,
- $D\_KaplanMeier = d_{KS}(S, \hat{S}_{KMI}, t_{KMI})$,
- $D\_Turnbull\_Mekm = d_{KS}(S, \hat{S}_T, t_{KMI})$.

Les distances associées à la norme L_2 entre $\hat{S}_T$ et S, puis entre $\hat{S}_{KMI}$ et S, sont notées par

- $L2norm\_Turnb = d_2(S, \hat{S}_T, t_T)$ où t_T est le plus grand des τ_j fini,
- $L2norm\_KM = d_2(S, \hat{S}_{KMI}, t_{KMI})$,
- $L2norm\_Turnb\_Mekm = d_2(S, \hat{S}_T, t_{KMI})$.

On notera que par définition t_{KMI} est inférieur à t_T. D'où l'intérêt de calculer les distances entre S et $\hat{S}_T$ ou S et $\hat{S}_{KMI}$ sur l'intervalle $(0, t_{KMI})$.

Une première illustration de cette expérience pour $n = 250$ est donnée dans les figures 2.7, 2.9 et 2.11 pour différentes lois du 2^e inter-suivi, les lois uniformes, exponentielles tronquées, bêta renormalisées introduites dans la section 3.1. Les résultats des 500 simulations pour chaque loi sont présentés sous forme de boîte à moustaches dans les figures 2.8, 2.10, 2.12 et sous forme de tableaux numériques 2.6, 2.5 et 2.6.

Dans le cas de la loi uniforme lorsque l'espérance de la loi du 2^e inter-suivi croît (13.5 ans à 40 ans), les performances de chaque estimateur se dégradent légèrement, voir le

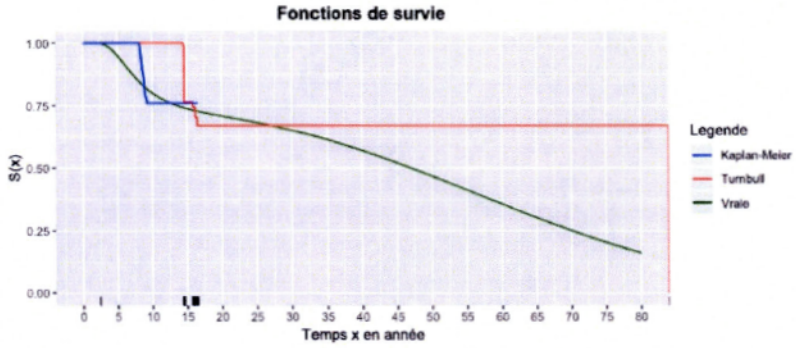
tableau avec valeurs numériques 2.6. Dans le cas de la loi exponentielle tronquée, lorsque l'espérance varie (13.5 ans à 17 ans) cela n'a pas d'effet marquant sur les performances de chaque estimateur. Il en est de même dans le cas de la loi bêta renormalisée. Il semble donc que le fait de faire varier l'espérance de la loi (tout en restant dans une plage de valeurs réalistes pour notre application) n'a qu'un léger impact sur les performances des estimateurs.

Pour toutes les lois considérées dans cette étude, l'estimateur de KMI est meilleur que l'estimateur de Turnbull pour les distances d_{KS} et L_2 sur $[0, t_{KMI}]$; voir les figures 2.8, 2.10, 2.12 et les tableaux. Dans le cas des lois uniformes, la médiane des distances $D\_KaplanMeier$ (respectivement $L2norm\_Turnb\_Mekm$) est 1.3 à 2.5 fois inférieure à la médiane des distances $D\_Turnbull\_Mekm$ (respectivement $L2norm\_KM$), voir tableaux numériques 2.6. Par contre pour les autres lois on ne voit pas de différence à 10^{-2} près.

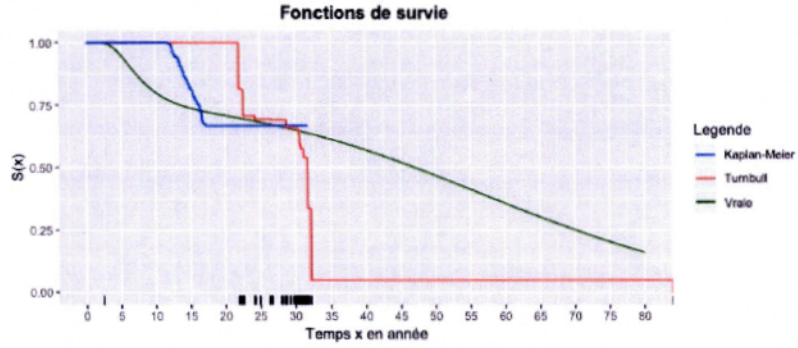
Visuellement, il semble que quelle que soit la loi, l'estimation de Turnbull ait moins de sauts que le KMI. Mais, il semble difficile de dégager une conclusion globale sur la comparaison de l'allure des estimations. Par exemple :

- dans le cas des lois uniforme et Beta renormalisée, on a l'impression que l'estimation de KMI est mieux au début, alors que quand le KMI s'arrête l'estimation de Turnbull continue à donner des détails pertinents, voir les figures 2.11, 2.7 ;
- dans le cas de la loi exponentielle tronquée, visuellement il est difficile de départager les estimations.

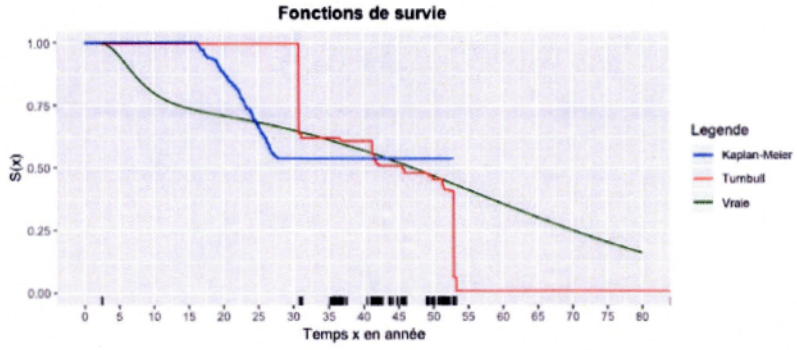
Ces différences semblent liées à la position des données imputées (instants de sauts du KMI) par rapport aux τ_j, $j = 1, \ldots, m-1$ (instants de sauts possibles du Turnbull) représentées dans ces figures par des barres noires (tic).



(a) Loi uniforme de support $(13.5-8/12, 13.5+8/12)$ pour le 2e inter-suivi.

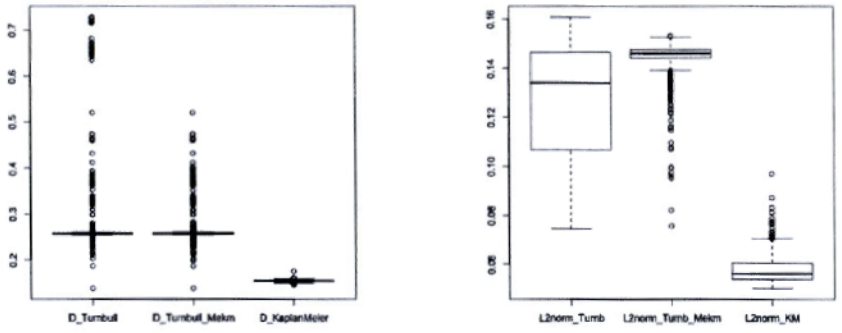


(b) Loi uniforme de support $(25-5, 25+5)$ pour le 2e inter-suivi.

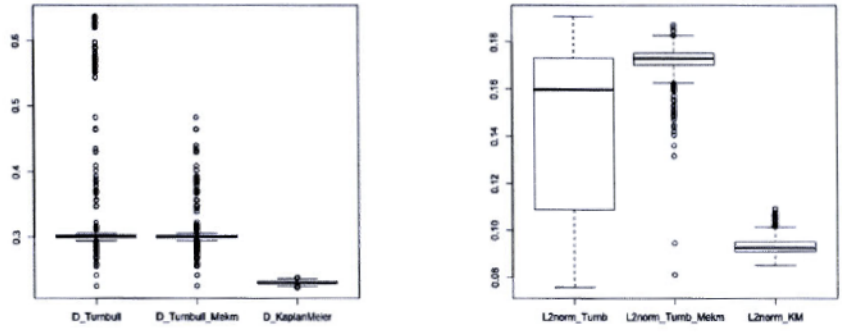


(c) Loi uniforme de support $(40-11, 40+11)$ pour le 2e inter-suivi.

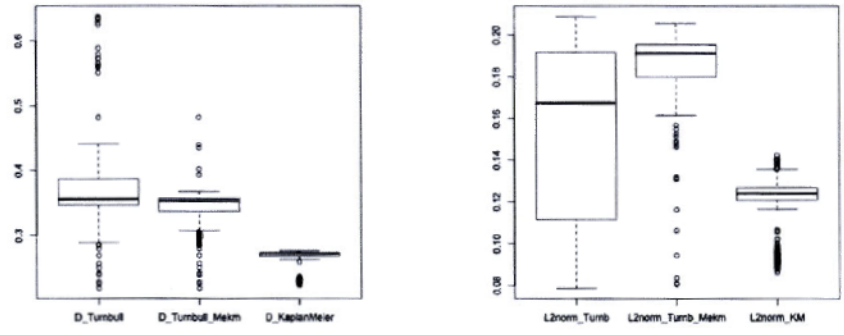
Figure 2.7 Exemple de comparaison entre la vraie fonction de survie S, l'estimateur de survie de Turnbull $\hat{S}_T$ et de Kaplan-Meier $\hat{S}_{KMI}$, pour $n = 250$, et pour différentes lois uniformes du 2e inter-suivi. Les barres noires (tic) représentent les données.



(a) Loi uniforme de support $(13.5 - 8/12, 13.5 + 8/12)$ du 2^e inter-suivi.



(b) Loi uniforme de support $(25-5, 25+5)$ du 2^e inter-suivi.



(c) Loi uniforme de support $(40 - 11, 40 + 11)$ du 2^e inter-suivi.

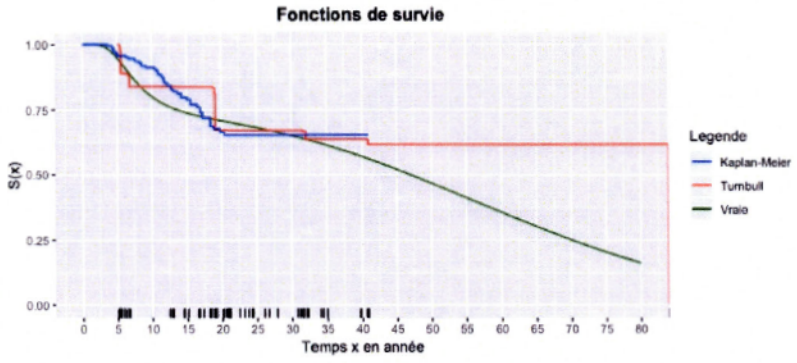
Figure 2.8 Loi uniforme : boîtes à moustaches des distances entre $\hat{S}_T$ et S, puis entre $\hat{S}_{KMI}$ et S pour $n = 250$.

Tableau 2.4 Résumé statistique des distances pour $n = 250$. En haut, au milieu et en bas les lois uniforme respectivement de support $(13.5 - 8/12, 13.5 + 8/12)$, $(25 - 5, 25 + 5)$ et $(40 - 11, 40 + 11)$ du 2e inter-suivi.

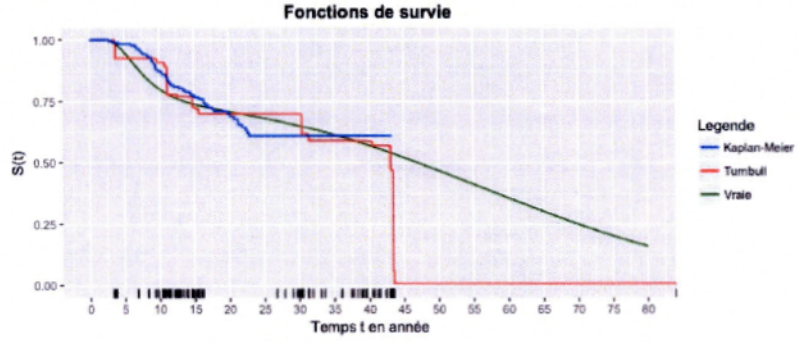
	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.14	0.26	0.26	0.30	0.26	0.73
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.14	0.26	0.26	0.27	0.26	0.52
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.15	0.15	0.15	0.15	0.16	0.18
$d_2(S, \hat{S}_T, \infty)$	0.08	0.11	0.13	0.13	0.15	0.16
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	0.08	0.14	0.15	0.14	0.15	0.15
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	**0.05**	**0.05**	**0.06**	**0.06**	**0.06**	**0.10**

	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.23	0.30	0.30	0.34	0.30	0.64
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.23	0.30	0.30	0.30	0.30	0.48
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.22	0.23	0.23	0.23	0.23	0.24
$d_2(S, \hat{S}_T, \infty)$	**0.08**	0.11	0.16	0.14	0.17	0.19
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.08**	0.17	0.17	0.17	0.18	0.19
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.09	**0.09**	**0.09**	**0.09**	**0.10**	**0.11**

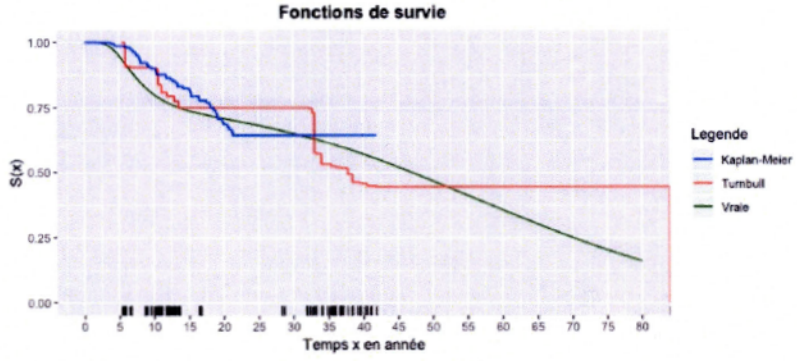
	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.22	0.35	0.36	0.37	0.39	0.64
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.22	0.34	0.35	0.34	0.36	0.48
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.22	0.27	0.27	0.26	0.27	0.28
$d_2(S, \hat{S}_T, \infty)$	**0.08**	**0.11**	0.17	0.15	0.19	0.21
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.08**	0.18	0.19	0.19	0.20	0.21
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.09	0.12	**0.12**	**0.12**	**0.13**	**0.14**



(a) Avec $b = 40$ et $\lambda = 0.052$

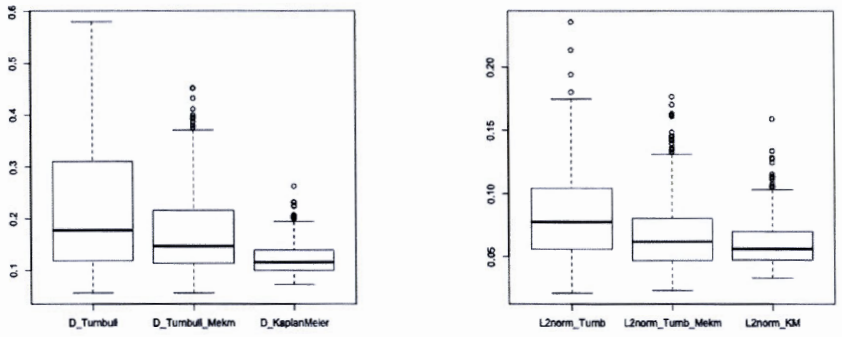


(b) Avec $b = 42$ et $\lambda = 0.04$

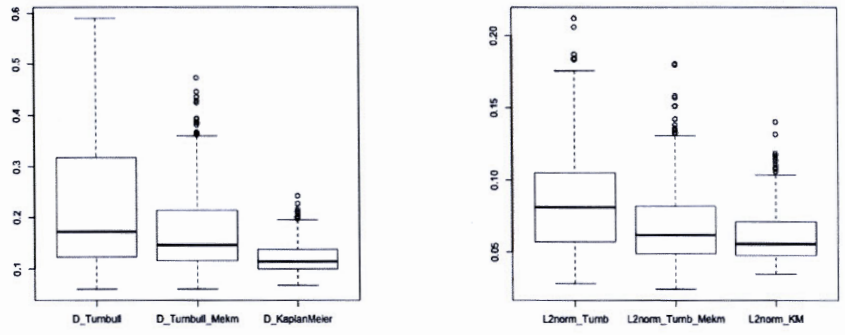


(c) Avec $b = 40$ et $\lambda = 0.025$

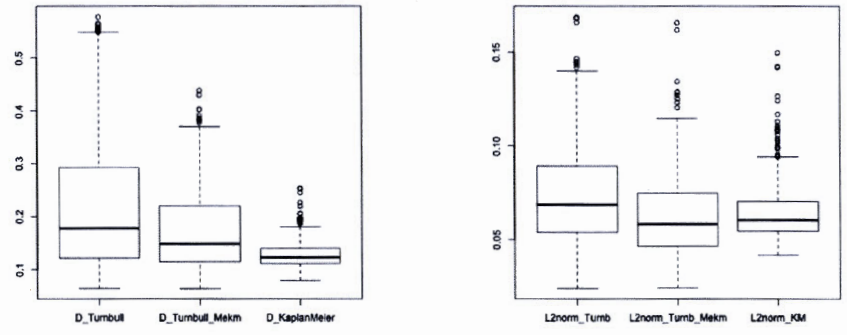
Figure 2.9 Exemple de comparaison entre la vraie fonction de survie S, l'estimateur de survie de Turnbull $\hat{S}_T$ et de Kaplan-Meier $\hat{S}_{KMI}$, pour $n = 250$, et pour différentes lois exponentielles tronquées du 2^e inter-suivi. Les barres noires (tic) représentent les données.



(a) Loi exponentielle tronquée de paramètres ($a = 0$, $b = 40$ et $\lambda = 0.052$) du 2^e inter-suivi.



(b) Loi exponentielle tronquée de paramètres ($a = 0$, $b = 42$ et $\lambda = 0.04$) du 2^e inter-suivi.



(c) Loi exponentielle tronquée de paramètres ($a = 0$, $b = 40$ et $\lambda = 0.025$) du 2^e inter-suivi.

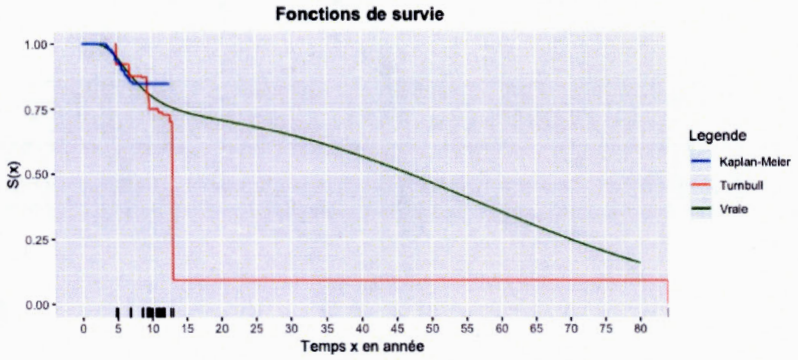
Figure 2.10 Loi exponentielle tronquée : boîtes à moustaches des distances entre $\hat{S}_T$ et S, puis entre $\hat{S}_{KMI}$ et S pour $n = 250$.

Tableau 2.5 Résumé statistique des distances pour $n = 250$. En haut, au milieu et en bas les lois exponentielles tronquées respectivement (de support $[0, 40]$ et d'intensité $\lambda = 0.052$), (de support $[0, 42]$ et d'intensité $\lambda = 0.04$) et (de support $(0, 40)$ et d'intensité $\lambda = 0.025$) du 2e inter-suivi.

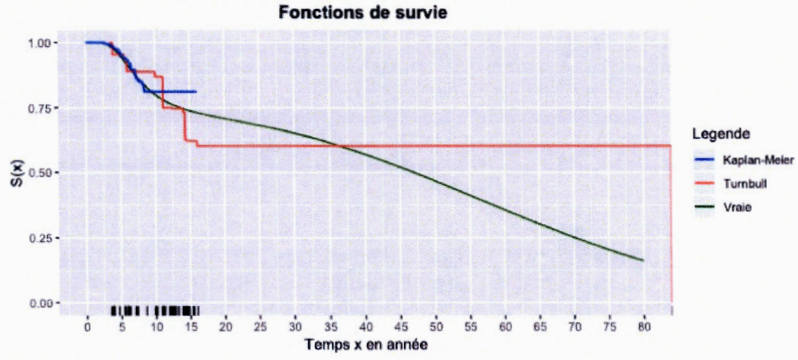
	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.06	0.12	0.18	0.24	0.31	0.58
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.06	0.11	0.15	0.17	0.22	0.45
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.07	0.10	0.12	0.12	0.14	0.26
$d_2(S, \hat{S}_T, \infty)$	**0.02**	0.06	0.08	0.08	0.10	0.24
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.02**	**0.05**	**0.06**	0.07	0.08	0.18
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.03	**0.05**	**0.06**	**0.06**	**0.07**	**0.16**

	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.06	0.12	0.17	0.24	0.32	0.59
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.06	0.12	0.15	0.17	0.21	0.47
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.07	0.10	0.11	0.12	0.14	0.24
$d_2(S, \hat{S}_T, \infty)$	0.03	0.06	0.08	0.08	0.10	0.21
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.02**	**0.05**	**0.06**	0.07	0.08	0.18
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.03	**0.05**	**0.06**	**0.06**	**0.07**	**0.14**

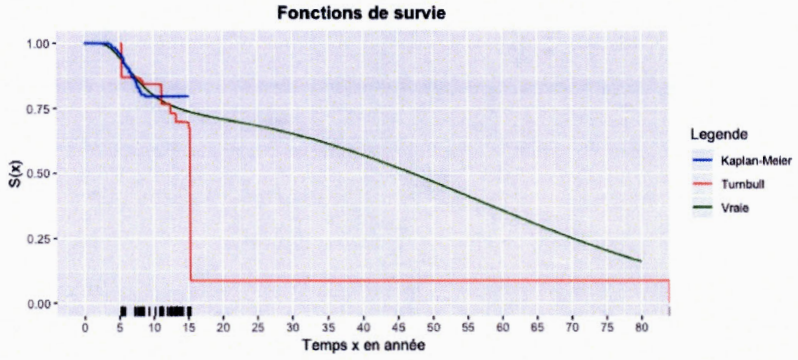
	Min.	1er Qu.	Médiane	Moy.	3e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.06	0.12	0.18	0.23	0.29	0.58
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.06	0.11	0.15	0.17	0.22	0.44
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.08	0.11	0.12	0.13	0.14	0.25
$d_2(S, \hat{S}_T, \infty)$	**0.02**	**0.05**	0.07	0.07	0.09	0.17
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.02**	**0.05**	**0.06**	**0.06**	**0.07**	0.17
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.04	**0.05**	**0.06**	**0.06**	**0.07**	**0.15**



(a) Avec $a = 5$ et $b = 8$

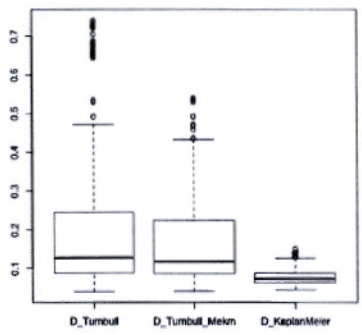


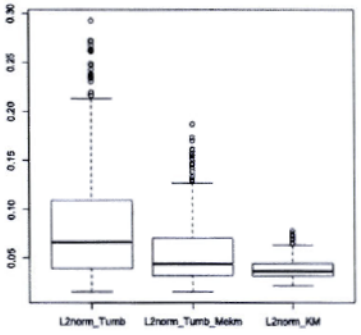
(b) Avec $a = 2$ et $b = 8/3$



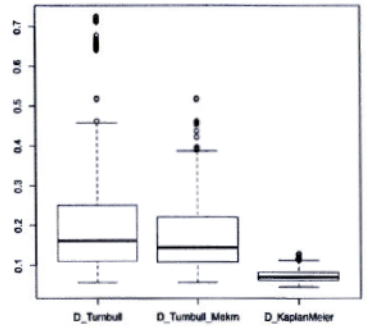
(c) Avec $a = 3$ et $b = 3.2$

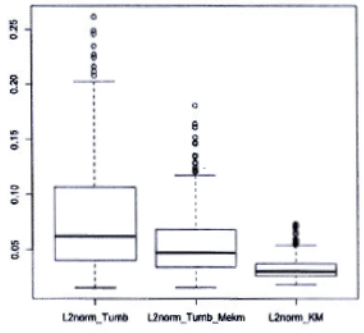
Figure 2.11 Exemple de comparaison entre la vraie fonction de survie S, l'estimateur de survie de Turnbull $\hat{S}_T$ et de Kaplan-Meier $\hat{S}_{KMI}$, pour $n = 250$, et pour différentes lois bêta renormalisées du 2e inter-suivi. Les barres noires (tic) représentent les données.



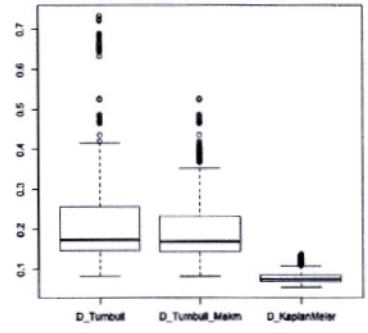


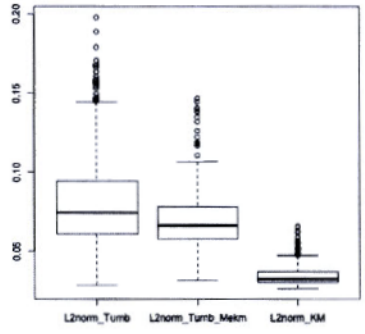
(a) Loi bêta renormalisée de paramètres ($a = 5$, $b = 8$) du 2^e inter-suivi.





(b) Loi bêta renormalisée de paramètres ($a = 2$, $b = 8/3$) du 2^e inter-suivi.





(c) Loi bêta renormalisée de paramètres ($a = 3$, $b = 3.2$) du 2^e inter-suivi.

Figure 2.12 Loi bêta renormalisée : boîtes à moustaches des distances entre $\hat{S}_T$ et S, puis entre $\hat{S}_{KMI}$ et S pour $n = 250$.

Tableau 2.6 Résumé statistique des distances pour $n = 250$. En haut, au milieu et en bas les lois bêta renormalisées respectivement de paramètres ($a = 5$ et $b = 8$), ($a = 2$ et $b = 8/3$) et ($a = 3$ et $b = 3.2$) du 2^e inter-suivi.

	Min.	1er Qu.	Médiane	Moy.	3^e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.04	0.08	0.12	0.20	0.26	0.76
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.04	0.08	0.11	0.16	0.23	0.59
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.05	0.07	0.09	0.09	0.10	0.16
$d_2(S, \hat{S}_T, \infty)$	**0.01**	0.04	0.06	0.08	0.11	0.32
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.01**	**0.03**	0.05	0.06	0.07	0.23
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.02	**0.03**	**0.04**	**0.04**	**0.04**	**0.07**

	Min.	1er Qu.	Médiane	Moy.	3^e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.04	0.09	0.13	0.20	0.24	0.74
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.04	0.09	0.12	0.16	0.22	0.54
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.04	0.06	0.07	0.08	0.09	0.15
$d_2(S, \hat{S}_T, \infty)$	**0.02**	0.04	0.06	0.08	0.11	0.26
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.02**	**0.03**	0.05	0.05	0.07	0.18
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	**0.02**	**0.03**	**0.03**	**0.03**	**0.04**	**0.07**

	Min.	1er Qu.	Médiane	Moy.	3^e Qu.	Max.
$d_{KS}(S, \hat{S}_T, \infty)$	0.05	0.09	0.14	0.20	0.24	0.74
$d_{KS}(S, \hat{S}_T, \max(t_{KMI}))$	0.05	0.09	0.13	0.17	0.23	0.59
$d_{KS}(S, \hat{S}_{KMI}, \max(t_{KMI}))$	0.04	0.06	0.07	0.08	0.09	0.15
$d_2(S, \hat{S}_T, \infty)$	**0.02**	0.04	0.06	0.08	0.10	0.24
$d_2(S, \hat{S}_T, \max(t_{KMI}))$	**0.02**	0.04	0.05	0.06	0.07	0.19
$d_2(S, \hat{S}_{KMI}, \max(t_{KMI}))$	**0.02**	**0.03**	**0.03**	**0.03**	**0.04**	**0.08**

CHAPITRE III

APPLICATION AUX DONNÉES DE LA LEUCÉMIE PÉDIATRIQUE

Dans ce chapitre, nous nous intéressons à un jeu de données réelles portant sur une cohorte d'individus atteints de la leucémie pédiatrique. Pour ces données, nous disposons à la fois du statut de surpoids et des covariables liées au traitement ce qui nous permet d'étudier la dépendance entre les deux. Dans la section 3.1 nous donnons une description plus complète des données, dans la section 3.2 nous faisons une analyse descriptive des données qui nous donne les premiers élements pour mettre en évidence la dépendance entre le surpoids et certaines de ces covariables. Enfin, dans les sections 3.3 et 3.4 nous illustrons l'apport du chapitre précédent pour préciser la dépendance entre le surpoids et deux covariables en particulier.

3.1 Description des données

Les données de la leucémie pédiatrique proviennent du Centre de santé de l'Université Sainte-Justine (CHU Sainte-Justine, Montréal, Canada) et portent sur un diagnostic de surpoids (notamment d'obésité) en mesurant l'indice de masse corporelle (IMC) au diagnostic du cancer, à la fin du traitement et à l'entrevue PÉTALE. L'IMC a été calculé comme le rapport poids / (taille)^2, où le poids est en kilos et la taille en mètres. Pour les sujets de 18 ans et plus, un IMC $\geq$ 25 (kg / m^2) indique le surpoids, tandis que l'IMC $\geq$ 30 (kg / m^2) correspond à l'obésité. Pour les enfants et les adolescents, l'IMC $\geq 85^{e}$ et $< 97^{e}$ percentile correspond à un surpoids et l'IMC $\geq 97^{e}$ percentile correspond à l'obésité selon les graphiques de l'IMC de l'Organisation Mondiale de la Santé, voir

Levy *et al.* [2017].

Au départ, les données se composaient de 246 patients et 22 variables, puis nous avons fait une sélection de patients qui n'étaient pas en surpoids au diagnostic. Les patients dont on ignore le statut de surpoids au diagnostic ont été écartés de notre analyse au regard de la conception de notre modèle décrit dans le chapitre 1. Au final, on a utilisé 179 patients et 22 variables. Nous appellerons « cohorte entière » cet ensemble de 179 patients. D'une part, on a 6 variables permettant de construire les intervalles de censure (information sur le temps de défaillance T) tels que décrits dans le chapitre 2 :

- 3 variables portant sur l'âge au diagnostic, à la fin du traitement et à l'entrevue PÉTALE,
- 3 variables donnant le statut de surpoids au diagnostic, à la fin du traitement et à l'entrevue PÉTALE.

D'autre part, on a une covariable portant sur le sexe (féminin, masculin) et 15 covariables liées au traitement représentant des agents chimiothérapeutiques hormis la radiothérapie :

- dose cumulative de corticostéroïdes (mg / m2),
- radiothérapie crânienne reçue (non, oui),
- dose totale de radiothérapie reçue (Gy),
- dose cumulative d'améthoptérine (mg / m2),
- dose cumulative de cytarabine (mg / m2),
- dose cumulative de dexamethasone (mg / m2),
- dose cumulative de dexrazoxane (mg / m2),
- dose cumulative de doxorubicine (mg / m2),
- dose cumulative de L-asparaginase (UI / m2),
- dose cumulative de leucovorine (mg / m2),
- dose cumulative de mercaptopurine (mg / m2),
- dose cumulative de vincristine (mg / m2),

— dose cumulative d'améthoptérine intrathécal ou IT (mg / m2),

— dose cumulative de cytarabine intrathécal ou IT (mg / m2),

— dose cumulative d'hydrocortisone intrathécal ou IT (mg / m2).

3.2 Analyse préliminaire des données

Le premier rendez-vous intervient en moyenne 2 ans et 1 mois après le traitement avec un écart-type d'environ 3 mois et demi. Le deuxième rendez-vous (entrevue PÉTALE) intervient en moyenne 13 ans et 3 mois après le premier rendez-vous avec un écart-type d'environ 5 ans et 2 mois. L'analyse de la distribution du premier inter rendez-vous indique qu'elle est unimodale (voir figure 3.1), celle du deuxième inter rendez-vous indique qu'elle est bimodale (voir figure 3.2) avec pour modes environ 12 et 17.

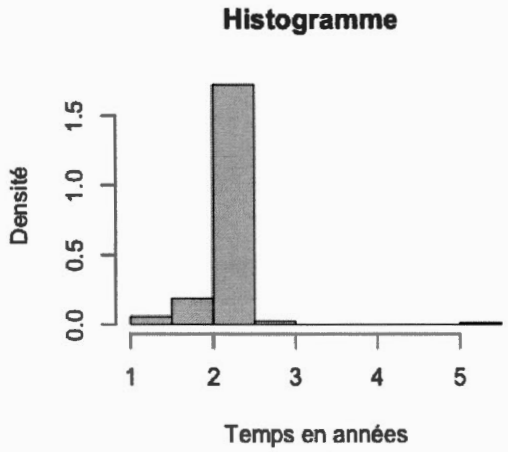


Figure 3.1 Distribution du premier inter rendez-vous.

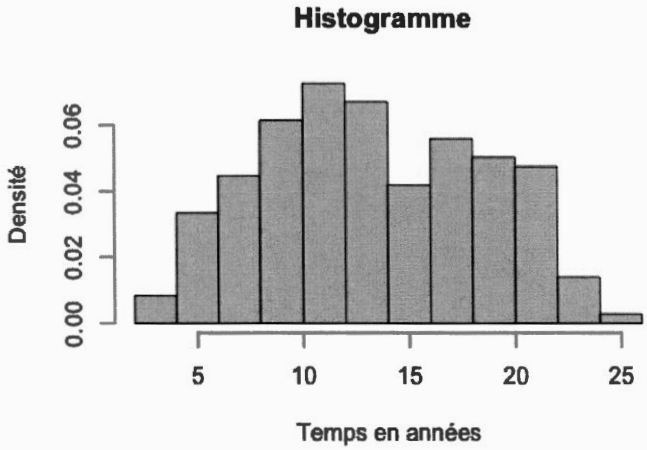


Figure 3.2 Distribution du deuxième inter rendez-vous.

La proportion des patients qui deviendront en surpoids avant la fin de l'étude est d'environ 49%. La proportion de surpoids intervenant dans les premier et deuxième inter rendez-vous sont respectivement d'environ 35% et 14%. Donnons quelques éléments de comparaison entre le sous-groupe des patients qui deviendront en surpoids avant la fin de l'étude et l'autre sous-groupe. Parmi ceux qui deviendront en surpoids, la proportion des filles est d'environ 59% alors que parmi ceux qui ne deviendront pas en surpoids elle est d'environ 47%, voir figure 3.3. Parmi ceux qui deviendront en surpoids, la proportion de ceux qui ont reçu une radiothérapie est d'environ 64%, alors que parmi les autres elle est d'environ 56%, voir figure 3.4. Le surpoids semble donc visuellement lié aussi bien au sexe qu'au fait d'avoir reçu de la radiothérapie crânienne.

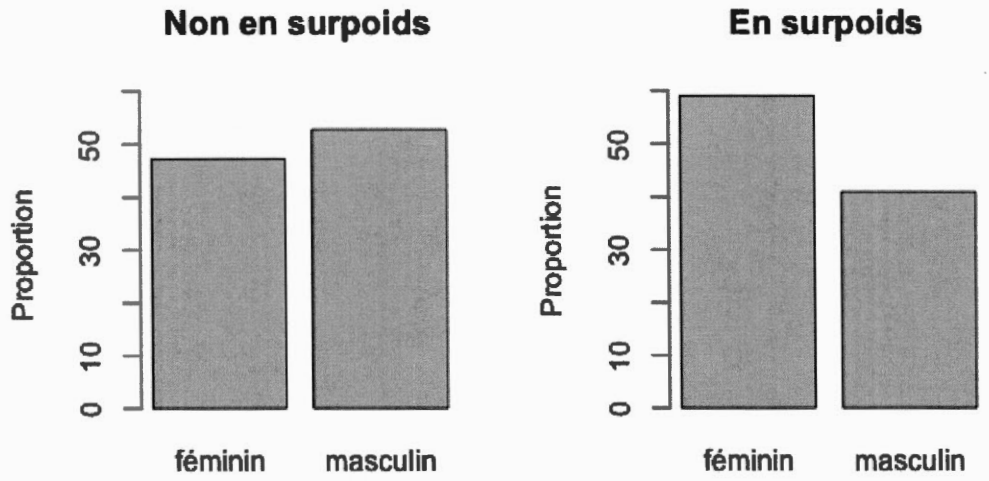


Figure 3.3 Proportion du sexe selon le surpoids avant la fin de l'étude.

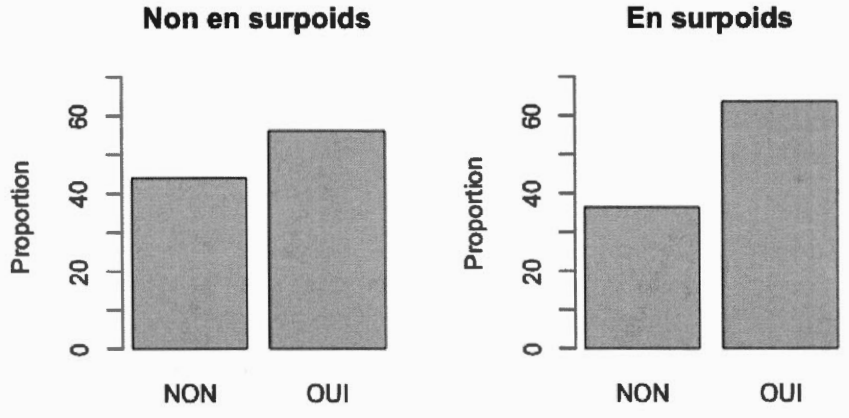


Figure 3.4 Proportion de radiothérapie crânienne selon le surpoids.

Analysons maintenant la dépendance entre le surpoids et les covariables quantitatives. Les covariables de la figure 3.5 sont CRT_dose (dose totale de radiothérapie reçue), Amethopterine (dose cumulative d'améthoptérine), Cytarabine (dose cumulative de cytarabine), Dexamethasone (dose cumulative de dexamethasone), celles de la figure 3.6 sont Dexrazoxane (dose cumulative de dexrazoxane), Leucovorin (dose cumulative de leucovorine), Mercaptopurine (dose cumulative de mercaptopurine), Vincristine (dose cumulative de vincristine), puis celles de la figure 3.7 sont Amethopterine_IT (dose cumulative d'améthoptérine intrathécal), Cytarabine_IT (dose cumulative de cytarabine intrathécal), Hydrocortisone_IT (dose cumulative d'hydrocortisone intrathécal) et Asparaginase (dose cumulative de L-asparaginase).

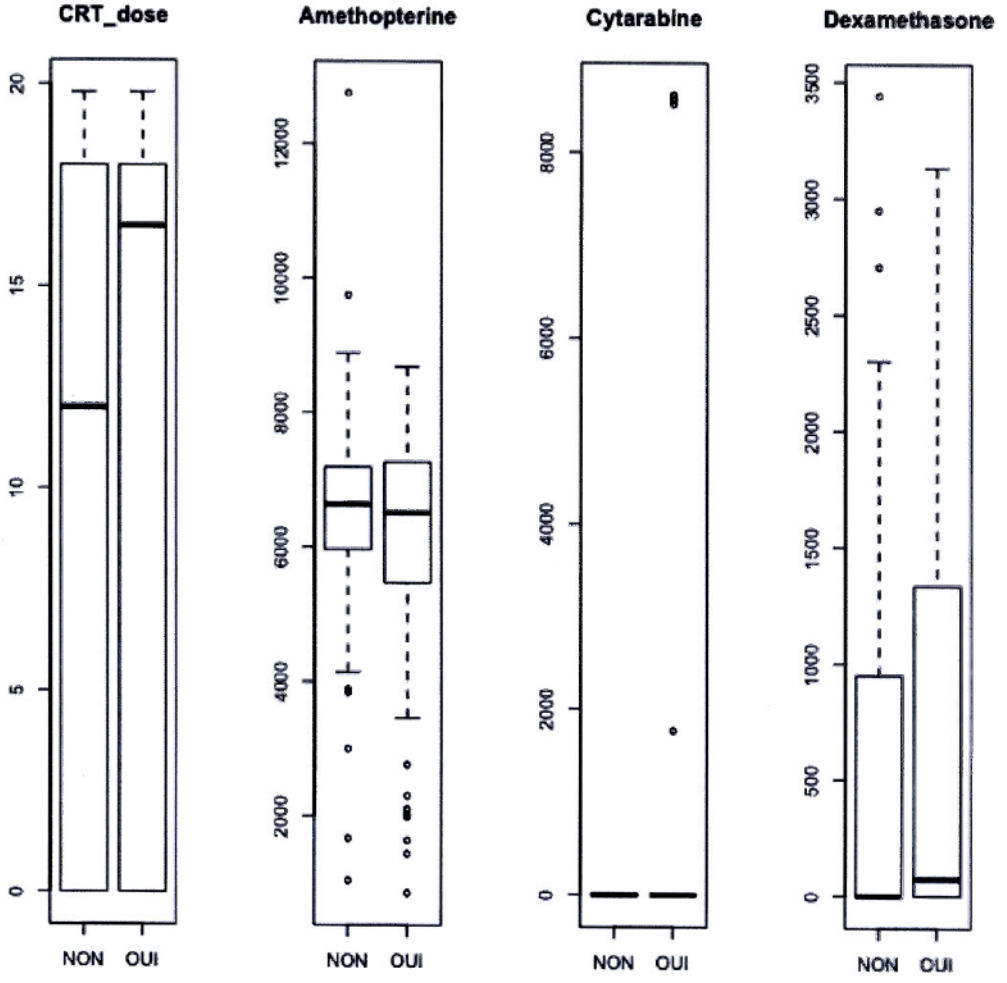


Figure 3.5 Covariables (CRT_dose, Amethopterine, Cytarabine, Dexamethasone) liées au traitement selon le statut de surpoids (NON, OUI)

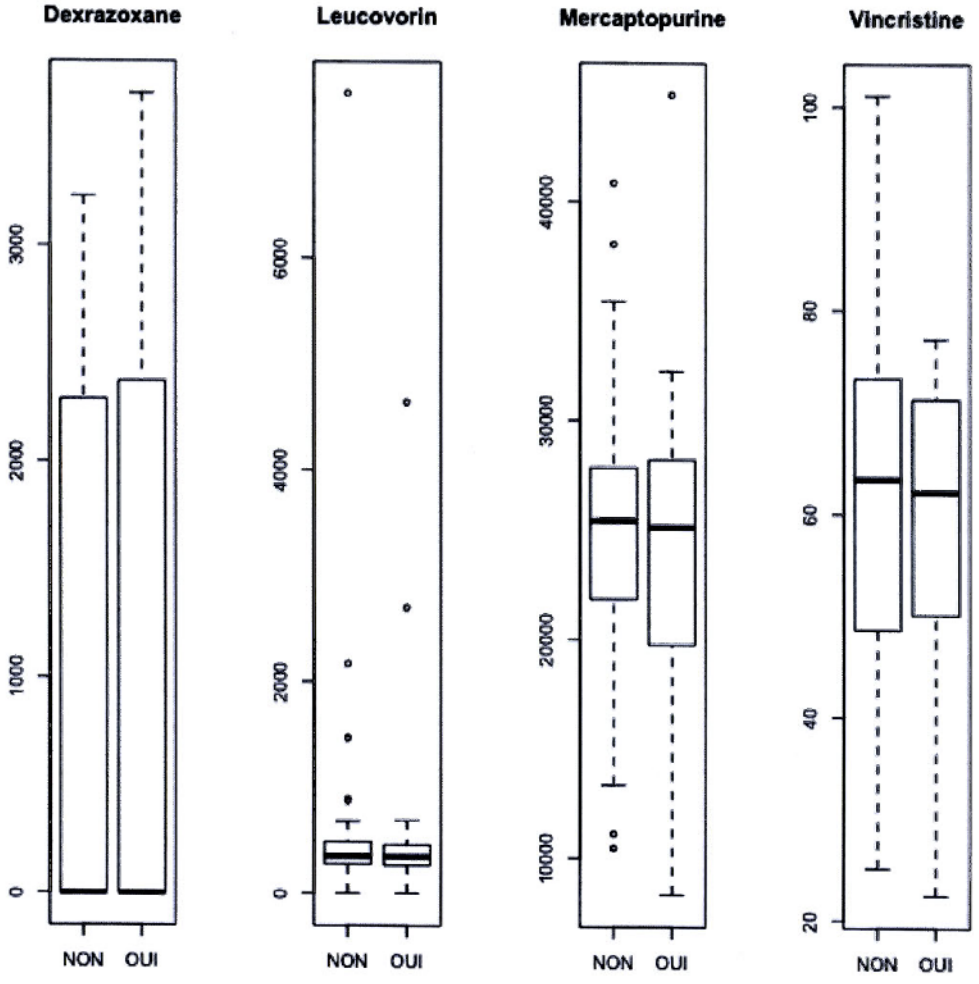


Figure 3.6 Covariables (Dexrazoxane, Leucovorin, Mercaptopurine, Vincristine) liées au traitement selon le statut de surpoids (NON, OUI)

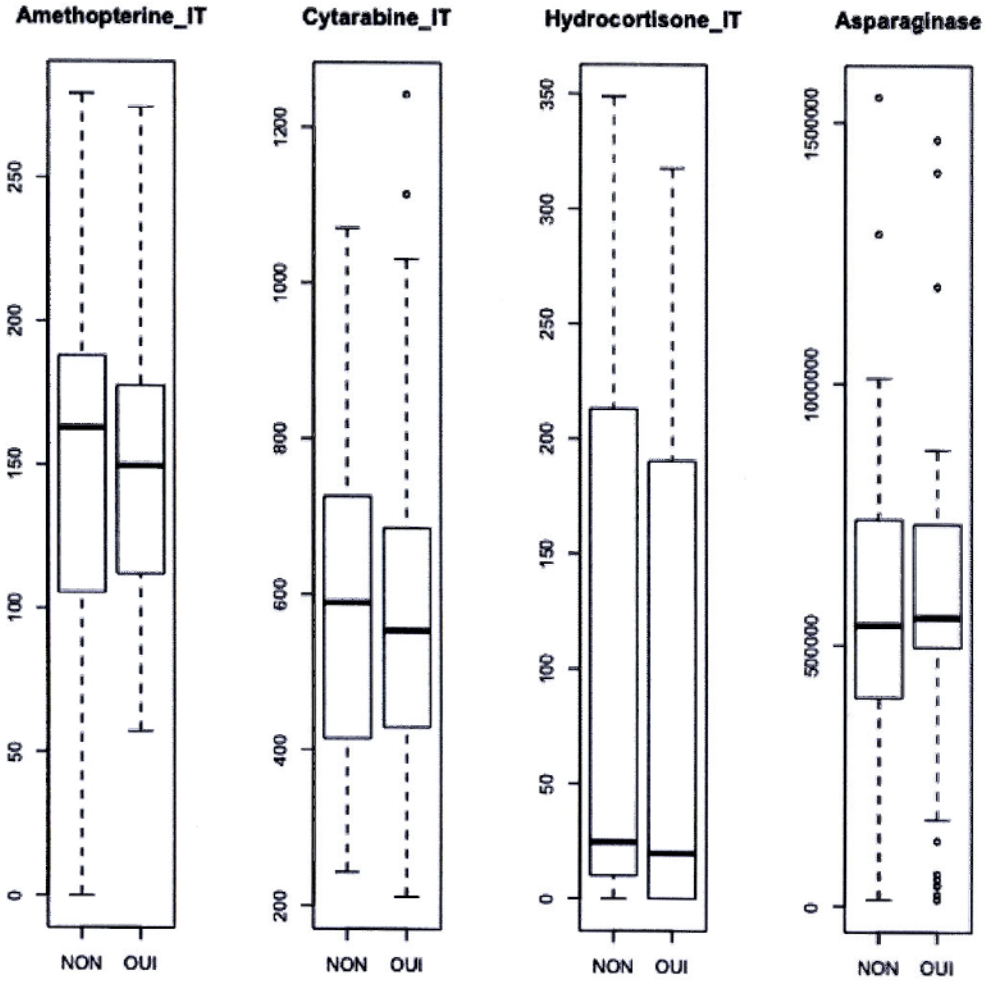


Figure 3.7 Covariables (Amethopterine_IT, Cytarabine_IT, Hydrocortisone_IT, Asparaginase) liées au traitement selon le statut de surpoids (NON, OUI)

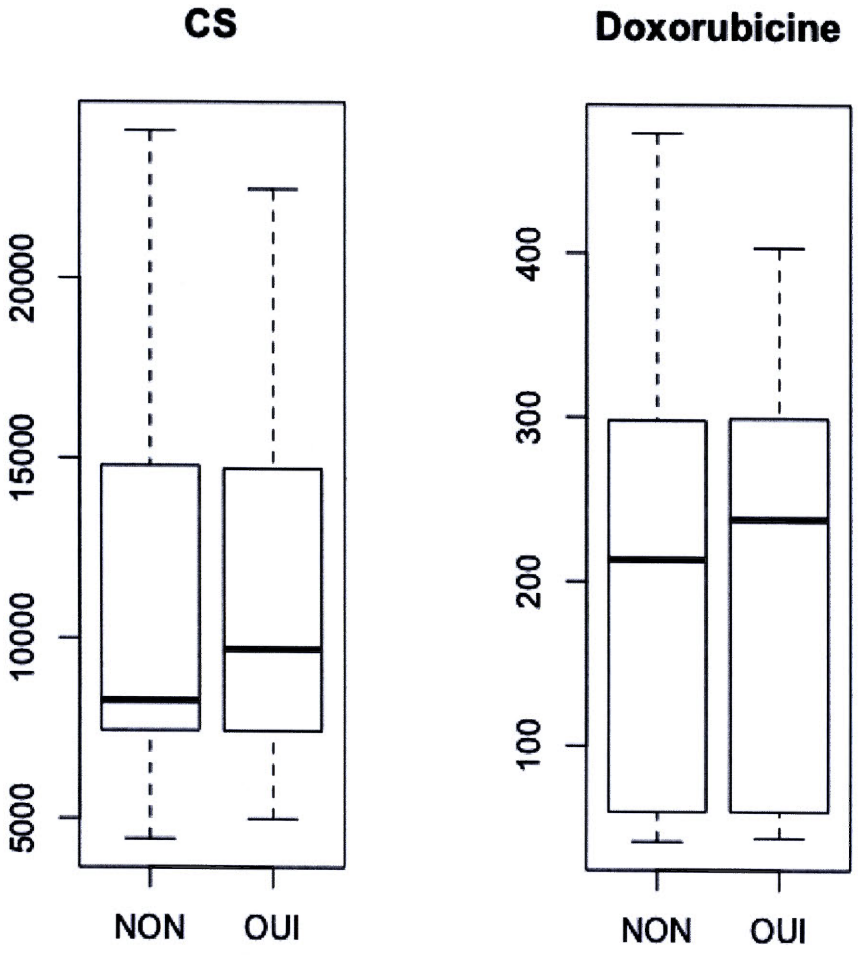


Figure 3.8 Covariables CS (dose cumulative de corticostéroïdes) et Doxorubicine (dose cumulative de doxorubicine) liées au traitement selon le statut de surpoids (NON, OUI)

En comparant les 14 covariables quantitatives liées au traitement selon le statut de surpoids à la fin de l'étude, on se rend compte qu'il y a trois catégories de covariables qui

se distinguent. La première catégorie concerne celles pour lesquelles la quantité de produit a tendance à être plus élevée dans le groupe en surpoids. Il s'agit des covariables CRT_dose, Dexamethasone, Asparaginase, CS, et Doxorubicine. En particulier, concernant la covariable CRT_dose, dans le groupe en surpoids on constate que la médiane est beaucoup plus élevée et la distribution des données est très asymétrique. Pour la covariable Asparaginase, pour les individus en surpoids on constate que le premier quartile est beaucoup plus élevé et les données sont moins dispersées.

La deuxième catégorie concerne les variables pour lesquelles la quantité de produit a tendance à être moins élevée dans le groupe en surpoids. Il s'agit des covariables Amethopterine, Mercaptopurine, Vincristine, Amethopterine_IT, Cytarabine_IT et Hydrocortisone_IT. En particulier, concernant les covariables Amethopterine et Mercaptopurine, dans le groupe en surpoids on constate que la médiane est à peu près au même endroit, mais le premier quartile est beaucoup plus petit, les données sont plus dispersées et asymétriques vers la gauche.

Enfin, la troisième catégorie concerne les variables pour lesquelles il n'y a pas vraiment de différence marquée entre le groupe de patients en surpoids et les patients qui ne sont pas en surpoids. Il s'agit des covariables Cytarabine, Dexrazoxane et Leucovorin. En particulier, concernant la covariable Cytarabine, la dose est de zéro pour tous les individus non en surpoids et en surpoids sauf pour environ 5% des individus en surpoids pour lesquels la dose est plus élevée et qui représentent des points aberrants.

3.3 Surpoids et radiothérapie crânienne

D'après l'analyse préliminaire (voir figure 3.4) il semble y avoir une liaison entre le fait de devenir en surpoids et le fait d'avoir reçu une radiothérapie crânienne. Nous proposons ici d'appliquer les techniques d'estimation du chapitre 2 pour étudier cette liaison.

Pour commencer, nous allons estimer la fonction de survie du premier temps d'apparition du surpoids. Nous avons fait trois estimations de cette fonction en utilisant l'algorithme

de Turnbull, l'estimateur de Kaplan-Meier sur données imputées et l'estimation d'un mélange, voir figure 3.9. Notre modèle de mélange est le mélange (2.40) donné par

$$S(x) = S_2(x)(pS_0(x) + (1-p)).$$

En s'inspirant de l'exemple 2.4.2 déjà utilisé dans l'étude de simulation 2.5, nous choisissons

- $S_0(x) = 1 - G_0(\frac{x-m}{\sigma})$, avec G_0 la fonction de répartition de la loi Log-normale de paramètre de position θ et de paramètre d'échelle $1/2$;
- $S_2(x) = 1 - G_2(x/15)$, où G_2 est la fonction de répartition de la loi de Weibull de paramètre de forme a et de paramètre d'échelle b.

Dans ce modèle de mélange les paramètres à estimer sont p, m, σ, θ, a et b. L'estimation des paramètres du mélange a été obtenue par maximisation de la vraisemblance réduite (1.4) en utilisant un algorithme de type Newton (fonction optim de `R`).

Visuellement les trois estimations ont à peu près la même allure. On remarque une décroissance rapide entre 0 et 2 ans environ, puis une décroissance lente. Cependant, à partir de la 13^e année, on constate que l'estimation de Kaplan-Meier sur données imputées reste constante tandis que les deux autres estimations continuent de décroître lentement, avec une différence nette par rapport à l'estimation KMI à partir de la 23^e année. Cette différence est peut-être due au fait qu'il y a moins de données sur la fin (51% de censure à droite).

Nous allons comparer le groupe des patients qui ont reçu de la radiothérapie versus ceux qui n'ont pas reçu de la radiothérapie. Pour cela nous construisons les estimations de la fonction de survie du premier temps d'apparition du surpoids des patients dans chacun des deux sous-groupes.

Comme le suggère la figure 3.10 (estimations de Turnbull) le temps d'apparition du surpoids des patients qui ont reçu une radiothérapie crânienne a tendance à intervenir

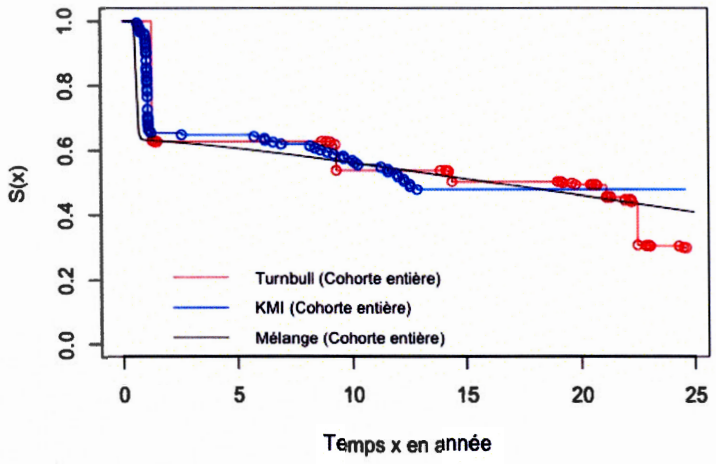


Figure 3.9 Comparaison de trois estimations de la fonction de survie du premier temps d'apparition du surpoids pour la cohorte entière.

un peu plus tôt durant la phase intensive de ce traitement (au bout de 2 ans). Ce temps d'apparition du surpoids des patients semble se stabiliser après 2 ans de traitement. Entre 13 (voire 14) et 21 ans la différence n'est pas visuellement significative. Puis après 21 années on voit une détérioration rapide pour ceux qui ont reçu une radiothérapie crânienne avant une nouvelle phase de stabilisation. Quant au deuxième groupe (pas de radiothérapie crânienne) on constate un gain de poids rapide au bout de 23 années environ. En comparant à l'estimation de Turnbull de la cohorte entière on voit également que l'apparition du surpoids a tendance à intervenir plus tôt chez les patients ayant reçu une radiothérapie crânienne. C'est le contraire pour l'autre sous-groupe.

En comparant les estimations de Kaplan-Meier sur données imputées (voir figure 3.11) on obtient pratiquement les mêmes conclusions, la principale différence étant que les estimations restent constantes à partir de la 14^{e} année.

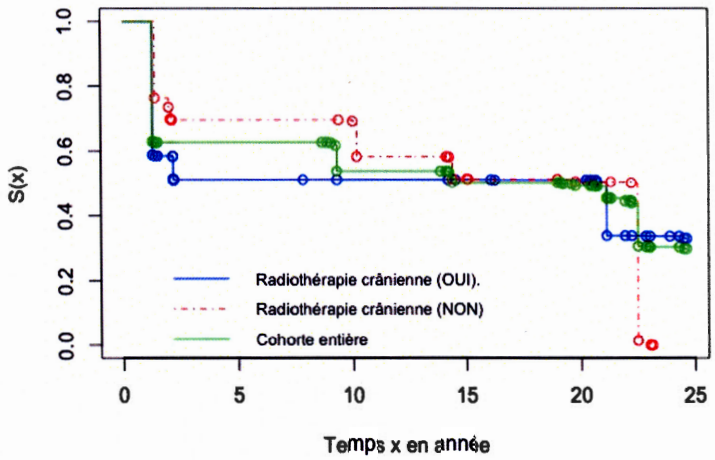


Figure 3.10 Comparaison non paramétrique (Turnbull) du premier temps d'apparition du surpoids de patients qui ont reçu ou non une radiothérapie crânienne.

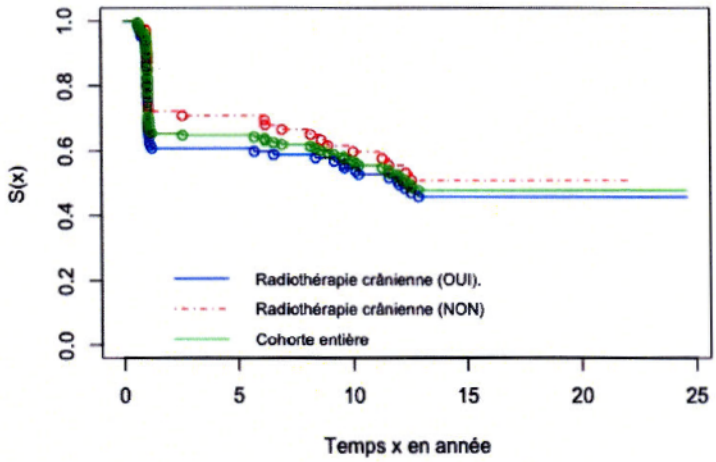


Figure 3.11 Comparaison non paramétrique (Estimation de la fonction de survie de Kaplan-Meier sur données imputées) du premier temps d'apparition du surpoids de patients qui ont reçu ou non une radiothérapie crânienne.

3.4 Surpoids et dose cumulative de corticostéroïdes

Nous avons créé deux sous-groupes à partir de la covariable CS (dose cumulative de corticostéroïdes) en partageant la cohorte entière selon la médiane de cette covariable qui vaut 8867.27 mg / m2.

On obtient une analyse similaire à celle de la section 3.3 (voir figures 3.12 et 3.13), c'est-à-dire le surpoids a tendance à apparaître plus tôt chez les individus ayant reçu une forte dose de corticostéroïdes entre 0 et 23 ans environ après le diagnostic. Puis la tendance s'inverse légèrement après 23 ans.

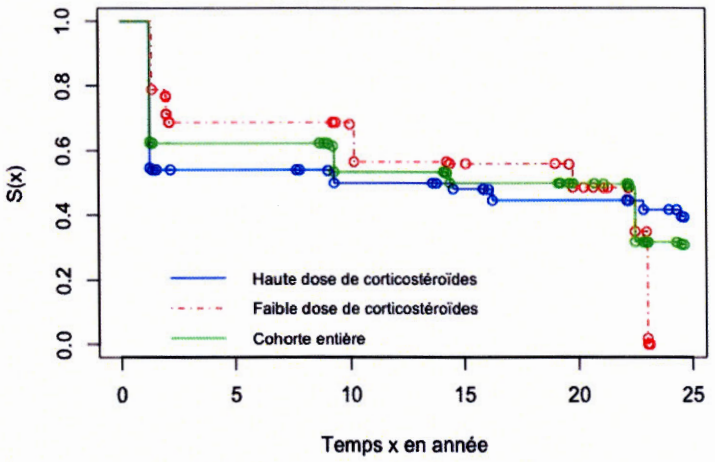


Figure 3.12 Comparaison non paramétrique (Turnbull) du premier temps d'apparition du surpoids selon la dose de corticostéroïdes.

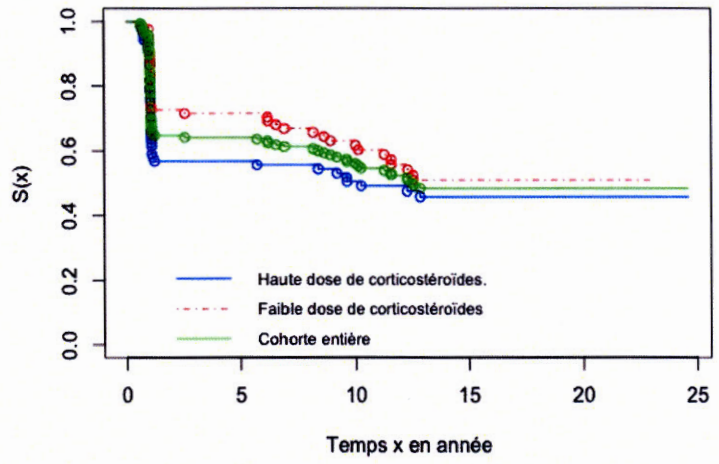


Figure 3.13 Comparaison non paramétrique (Estimation de la fonction de survie de Kaplan-Meier sur données imputées) du premier temps d'apparition du surpoids selon la dose de corticostéroïdes.

CONCLUSION

Ce mémoire porte sur la comparaison de deux estimateurs non paramétriques de la fonction de survie pour des données censurées par intervalle et propose également une application sur les données réelles. Les principales contributions sont décrites plus bas.

En premier, d'un point de vue théorique, nous avons proposé une autre démonstration du fait que la censure non informative est un cas particulier du modèle à somme constante. De plus nous avons mis en forme de manière détaillée la présentation de l'algorithme de Turnbull comme un algorithme EM.

Ensuite, d'un point de vue computationnel, nous avons réalisé une étude de simulation afin de comparer l'estimateur de Kaplan-Meier sur données imputées et l'estimateur de Turnbull. Dehghan et Duchesne [2011] proposent une comparaison par simulation de l'estimateur KMI et d'une généralisation de l'estimateur de Turnbull (pour une fonction de survie avec une covariable), mais avec des instants de rendez-vous simulés selon un processus de Poisson. Dans ce mémoire, nous avons construit une étude de simulation adaptée à l'application visée, avec un nombre fixe de rendez-vous et en tenant compte de contraintes sur les inter-suivis. Nous avons également proposé un modèle de mélange pour la construction d'un taux de survie de forme non usuelle adapté au problème étudié dans le chapitre 3.

Enfin, d'un point de vue pratique, pour l'application de l'algorithme de Turnbull, nous avons utilisé comme critère d'arrêt la stabilisation de la log-vraisemblance, critère d'arrêt différent de celui utilisé dans les références. Notre étude nous a permis de mettre en évidence que la forme de la loi du 2^e inter-suivi a un impact sur la qualité des estimateurs ; et que même si l'estimateur KMI est meilleur du point de vue de deux distances (Kolmogorov-Smirnov et L_2), visuellement les deux estimateurs peuvent apporter des

informations complémentaires. Enfin, concernant l'application aux données réelles nous avons mis en évidence la dépendance entre le surpoids et la radiothérapie crânienne, et entre le surpoids et la dose de corticostéroïdes, en utilisant la comparaison d'estimations des fonctions de survie dans des sous-groupes différents.

Les conclusions de ce mémoire mènent à quelques pistes de recherche. Il pourrait être intéressant de faire de nouvelles simulations pour comparer KMI et Turnbull avec des lois d'inter-suivis plus près des vraies données, qui sont bimodales. En utilisant l'estimateur de Turnbull, on pourrait également construire un test statistique de comparaison basé sur des données censurées par intervalle afin de déterminer si deux groupes d'individus sont significativement différents, par exemple dans l'application du chapitre 3, groupe des patients ayant reçu une radiothérapie crânienne et groupe des patients n'en ayant pas reçu (Peace *et al.* [2012], Self et Grossman [1986], Zhang et Sun [2010]). Il serait aussi intéressant d'exploiter le modèle de mélange paramétrique pour faire de la comparaison de fonctions de survie comme suggéré dans le chapitre 3. Les méthodes utilisées dans ce projet, ne permettent pas de prendre en compte simultanément toute les covariables. Aussi, il serait intéressant d'ajuster des modèles de régression à ces données. Nous pensons par exemple au modèle de Cox avec, possiblement une régularisation de type Ridge pour prendre en compte la corrélation entre les covariables, voir Finkelstein [1986], Wu et Cook [2015]. On pourrait également prévoir la probabilité de devenir en surpoids à la fin de l'étude à partir des covariables. Pour cela, il faudrait adapter le modèle classique de régression logistique aux données censurées par intervalle.

ANNEXE A : CODE R

```
#########################################################################
## Implémentation de L'estimateur de Turnbull (Algorithme 1)
#########################################################################
## arguments :
## data, les données censurées par intervalles
## left, borne inférieure des données data
## right, borne supérieure des données data
## tau, partition de R+ comme ordre unique des données data
## alpha, matrice (nxm) avec des coefficients alpha(ij)
## p_hat :=p, nous notons l'estimateur "p_hat" de p par p
#########################################################################

## I- Fontions utiles à utiliser dans la fonction
## de Turnbull à la section II :

require(spam)
require(survival)

## (i) Definir tau, l'ordre unique des données "data"

Ordre.tau <- function(data){
  l <- data$left
  r <- data$right
  tau <- sort(unique(c(l,r[is.finite(r)])))
  tau<-c(tau,Inf)
```

```
  return(tau)
}


## (ii) Alpha, une matrice (nxm) avec des coefficients alpha(i,j)

mat.Alpha <- function(data,tau){
  tau12 <- cbind(tau[-length(tau)],tau[-1])
  interv <- function(x,inf,sup) ifelse(x[1]>=inf & x[2]<=sup,1,0)
  alpha <- apply(tau12,1,interv,inf=data$left,sup=data$right)
  alpha<-spam(x=as.numeric(alpha), nrow = nrow(alpha), ncol = ncol(alpha))
  return(alpha)
}

## (iii) Initialisation de p_hat0

p_hat.ini <- function(tau){
  m<-length(tau)
  ekm<-survfit(Surv(tau[1:m-1],rep(1,m-1))~1)
  So<-c(1,ekm$surv)
  p <- -diff(So)
  return(p)
}



## II- Fonction de l'algorithme de Turnbull :

Turnbull_Old <- function(data, tau, p, Alpha, tol=1e-3,
```

```
maxit=1e4, quiet=FALSE)
{
  p<-matrix(p,ncol=1)
  n<-nrow(Alpha)
  m<-ncol(Alpha)
  iterating=TRUE
  k=0
  p.new <- p
  tol.adjusted=tol        ## Seuil d'ajustement
  C<-tcrossprod.spam(Alpha, t(p))
  #C<-MatMult(Alpha, p)
  if(!quiet) cat("k=", k, "log.vrai(Alpha,p)=", sum(log(C)), "\n")
  totalchange=0
  l_tau <-c(0)
  dist1<-c()
  dist2<-c()
  l_diff<-c()
  Est_Survie <- c()
  while(iterating)
  {
    k=k+1
    Q<-((crossprod.spam(Alpha,1/C))/n)
    p.new <-p*Q
    # }

    ## Actualiser la valeur de la log-vraisemblance
    C<-tcrossprod.spam(Alpha, t(p.new))

    #C<-MatMult(Alpha, p.new)
```

```
  l_tau.new <- sum(log(C))

  ## Stabilité de la log-vraisemblance
  # log_diff<- (l_tau.new-log.vrai(Alpha,p))
  log_diff<- (l_tau.new-l_tau[k])
  totalchange <- abs(log_diff)

  ## Mise a jour de p
  d1<-max(abs(p.new-p))
  dist1<-c(dist1,d1)
  d2<-sum((p.new-p)^2)
  dist2<-c(dist2,d2)
  l_diff<-c(l_diff,log_diff)
  p <-p.new
  l_tau <- c(l_tau,l_tau.new)
  if( ( totalchange < tol.adjusted) | (k >= maxit))
    iterating=FALSE

  if(!quiet) cat("k=", k, "log.vrai(Alpha,p)=", sum(log(C)), "\n")
  #if(!quiet) cat("k=", k, "p.new=", p.new, "\n")
}
## Faire de p un vecteur au lieu d'une matrice à une colonne
# p=as.numeric(p)

cat("Iterations = ", k,"\n")
cat("diff de vraisemblance = ", totalchange,"\n")
cat("Critère de convergence : diff de vraisemblance < tol.adjusted","\n")
# dimnames(p)<-list(NULL,c("Estimateur de P"))
survtau<-round(c(1,1-cumsum(p)),digits=5)
```

```
  #  right <- data$right
  #   if(any(!(is.finite(right)))){
  #     t <- max(right[is.finite(right)])
  #     return(list(time=tau[tau<t],surv=surv[tau<t],phat=p,
   k=k, l_tau=l_tau))
  #   }
  #
  #   else
  return(list(data=data, tau=tau, survtau=survtau, phat=p,
   k=k, l_tau=l_tau,
              dist1=dist1, dist2=dist2, l_diff=l_diff))
}
```

RÉFÉRENCES

Carpenter, J. et Kenward, M. (2012). *Multiple imputation and its application.* John Wiley & Sons.

De Waal, T., Pannekoek, J. et Scholtus, S. (2011). *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons.

Dehghan, M. H. et Duchesne, T. (2011). On the performance of some non-parametric estimators of the conditional survival function with interval-censored data. *Computational Statistics & Data Analysis*, *55*(12), 3355–3364.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Efron, B. (1967). The two sample problem with censored data. Dans *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, 831–853.

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*(4), 845–854.

Gómez, G., Calle, M. L. et Oller, R. (2004). Frequentist and bayesian approaches for interval-censored data. *Statistical Papers*, *45*(2), 139–173.

Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Klein, J. P. et Moschberger, M. L. (2003). Survival analysis : Techniques for censored and truncated data.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. et Scheike, T. H. (2013). *Handbook of Survival Analysis.* CRC Press.

Levy, E., Samoilenko, M., Morel, S., England, J., Amre, D., Bertout, L., Drouin, S., Laverdière, C., Krajinovic, M., Sinnett, D. *et al.* (2017). Cardiometabolic risk factors in childhood, adolescent and young adult survivors of acute lymphoblastic leukemia–a petale cohort. *Scientific reports*, *7*(1), 17684.

Marcoux, S., Drouin, S., Laverdière, C., Alos, N., Andelfinger, G. U., Bertout, L., Curnier, D., Friedrich, M. G., Kritikou, E. A., Lefebvre, G. *et al.* (2017). The petale study : Late adverse effects and biomarkers in childhood acute lymphoblastic leukemia survivors. *Pediatric blood & cancer*, *64*(6), e26361.

Marshall, A. W. et Olkin, I. (2007). *Life distributions*, volume 13. Springer.

Nadarajah, S., Kotz, S. *et al.* (2006). R programs for truncated distributions. *Journal of Statistical Software*, *16*(c02).

Oller, R., Gómez, G. et Calle, M. L. (2004). Interval censoring : model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics*, *32*(3), 315–326.

Oller Piqué, R. (2006). *Survival analysis issues with interval-censored data.* Universitat Politècnica de Catalunya.

Peace, K. E., Sun, J. et Chen, D.-G. D. (2012). *Interval-censored time-to-event data : methods and applications.* Chapman and Hall/CRC.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, 86–91.

Schick, A. et Yu, Q. (2000). Consistency of the gmle with mixed case interval-censored data. *Scandinavian Journal of Statistics*, *27*(1), 45–55.

Self, S. G. et Grossman, E. A. (1986). Linear rank tests for interval-censored data with application to pcb levels in adipose tissue of transformer repair workers. *Biometrics*, 521–530.

Sun, J. (2007). *The statistical analysis of interval-censored failure time data.* Springer Science & Business Media.

Thas, O. (2010). *Comparing distributions.* Springer.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.

Wu, Y. et Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression : Selection of hla markers in psoriatic arthritis. *Biometrics*, *71*(3), 782–791.

Zhang, Z. et Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, *19*(1), 53–70.

ERRATA : MÉMOIRE

DONNÉES CENSURÉES PAR INTERVALLE : APPLICATION À L'ÉTUDE DE LA PRÉVALENCE DU SURPOIDS APRÈS TRAITEMENT DE LA LEUCÉMIE

PAR
MICHAEL EVRARD MAKOUANGOU NGOUMA

ANCIEN	NOUVEAU
page 1 : La suite de ce mémoire	page 1 : Ce mémoire
page 3 : soit en moyenne deux ans	page 2 : soit environ deux ans
page 5 : Une certaine proportion	page 4 : dont une certaine proportion
page 5 : Ici j=1,2,3 et i=1,2,3,4	page 4 : , avec j=1,2,3 et i=1,2,3,4
page 6 : l'étude PETALE débute en 2013 et se termine en 2015	page 4 : l'étude PETALE a débuté en 2013 et s'est terminée en 2015
page 6 : l'entrevue PETALE G_{i3} et la fin de l'entrevue PETALE (2015)	page 5 : l'entrevue PETALE G_{i3} et la fin de l'étude (2015)
page 9 : figure 2.3	page 7 : figure 1.3
page 9 : pour tout les individus	page 7 : pour tous les individus
page 10 : un intervalle noté $I = (L, R]$ tel que $T \in I$	page 8 : un intervalle noté $I = (L, R]$ où les bornes L, R sont aléatoires et tel que $T \in I$
page 11 : de réalisations de $T_1, \ldots, T_n$ sont inconnus	page 8 : de réalisations de $T_1, \ldots, T_n$ sont inconnues
page 11 : (i) Soient $\{(L_i, R_i]; i = 1, \ldots, n\}$	page 8 : (i) Soit $\{(L_i, R_i]; i = 1, \ldots, n\}$
page 11 : au sens des références Peace et al. [2012}, p. 7-8, Gomez et al. [2004], p. 143 - 144, Klein et al. [2013]	page 9 : au sens de Peace et al. [2012}, p. 7-8, Gomez et al. [2004], p. 143 - 144, et Klein et al. [2013]
page 12 : qui n'ont rien à voir avec le statut	page 9 : qui ne sont pas reliées au statut
page 11 : L'équation (1.2) implique que les paramètres de la distribution de L et R sont indépendants des paramètres de la distribution de T, ce qui veut dire que la seule information fournie par l'intervalle de censure (l, r] à propos du temps de survie t est que l'intervalle contient t.	page 9 : L'équation (1.2) veut dire que la seule information fournie par l'intervalle de censure (l, r] à propos du temps de survie t est que l'intervalle contient t.
page 12 : densité jointe	page 9 : densité conjointe
page 12 : Pour nous, il s'agira de la distribution S de T	page 9 : Pour nous, il s'agira de la fonction de survie S de T

page 12 : Définition 2.3.2	page 9 : Définition 1.3.1
page 12 : Théorème 2.3.1	page 10 : Théorème 1.3.2
page 13 : Supposons que nous avons un individu pour lequel nous avons observé	page 10 : Supposons que nous observons
page 13 : dans ce qui suit nous en donnons	page 10 : Dans ce qui suit, nous en donnons
page 15 : Dans ce chapitre la source principale est	page 13 : Dans ce chapitre, la source principale est
page 15 : gamma ou Weibull) pour la distribution du temps de défaillance	page 13 : gamma, Weibull, etc) pour la loi du temps de défaillance
page 15 : Pour commencer nous rappelons	page 13 : Pour commencer, nous rappelons
page 17 : ni défaillances ni censures	page 15 : ni défaillances, ni censures
page 17 : et ne peut sauter qu'en les $z_{(j)}$	page 15 : et les points de sauts sont les $z_{(j)}$
page 18 : De Waal et al. [2011], Carpenter et Kenward [2012]	page 15 : (De Waal et al. [2011], Carpenter et Kenward [2012])
page 18 : En effet, il est plus simple de se ramener à la censure à droite parce que l'on dispose de plus d'outils, techniques et publications à ce sujet	page 16 : En effet, il est plus simple de se ramener à la censure à droite parce que l'on dispose de plus d'outils et techniques dans ce contexte.
page 18 : Plus précisement la méthode	page 16 : Plus précisément, la méthode
page 18 : $z_i^* = l_i$ et $r_i^* = +\infty$, sinon	page 16 : $z_i^* = l_i$, si $r_i^* = +\infty$
page 19 : en les l_i et r_i ont la même vraisemblance	page 17 : en les l_i et r_i, $i = 1,\ldots,n$ ont la même vraisemblance
page 19 : En particulier même si	page 17 : En particulier, même si
page 19 : Par ailleurs il n'y a pas de forme	page 17 : Par ailleurs, il n'y a pas de forme
page 19 : Espérance - Maximisation (EM) Dempster et al. [1977], qui permet	page 17 : Espérance - Maximisation (EM) introduite par Dempster et al. [1977], méthode qui permet
page 21 : (en bleue)	page 19 : (en bleu)
page 24 : en posant pour $j = 1,\ldots,m,$ on a	page 22 : en posant pour $j = 1,\ldots,m,$
page 28 : donné de $\varepsilon > 0$, alors on doit	page 26 : donné de $\varepsilon > 0$; alors, on doit
page 31 : on a soient p et p'	page 28 : soit p et p'
page 31 : En utilisant le développement de Taylor, on	page 28 : En utilisant le développement de Taylor, on a
page 33 : le surpoids due au traitement	page 30 : le surpoids dû au traitement
page 35 : S1(x)=S0(x)S2(x).	page 32 : S1(x)=S0(x)S2(x). Il s'agit de la fonction de survie d'un temps minimum, voir aussi l'algorithme 4
page 36 : D'autre part, par définition et de	page 32 : D'autre part, par définition et en vertu de

page 36 : Puis on a pris pour	page 36 : Ensuite on a posé
page 37 : Sortie : Taux de survie	page 36 : Sortie : Taux de survie h(x)
page 38 : Fonctions de survies	page 37 : Fonctions de survie
page 38 : Taux de survies	page 37 : Taux de survie
page 38 : Simulation de temps de défaillance	page 37 : Simulation du temps de défaillance
page 38 : sont indépendants	page 37 : sont deux variables aléatoires indépendantes
page 38 : Dans ce qui suit nous présentons	page 37 : Dans ce qui suit, nous présentons
page 39 : Dans ce cas et pour chaque individu i on définit	page 38 : Dans ce cas, et pour chaque individu i, on définit
page 40 : Étude de simulations	page 39 : Étude de simulation
page 40 : de simulations conçues	page 39 : de simulation conçue
page 40 : Cette étude de simulations	page 39 : Cette étude de simulation
page 41 : et en bas loi beta renormalisée	page 41 : et en bas loi bêta renormalisée
page 42 : Troisième cas : loi Beta	page 42 : Troisième cas : loi Bêta
page 42 : est la loi beta de paramètre	page 42 : est la loi bêta de paramètre
page 45 : les résultats de simulations	page 44 : les résultats des simulations
page 45 : décrite au paragraphe 4.1.	page 44 : décrite à la section 4.1.
page 46 : seront notées par	page 45 : sont notées par
page 46 : voir tableaux numériques	page 45 : voir le tableau avec valeurs numériques
page 46 : de la loi Beta	page 46 : de la loi Bêta
page 47 : Pour tous les exemples de lois	page 46 : Pour toutes les lois
page 47 : les figures 3.8, 3.10, 3.12	page 47 : les figures 2.8, 2.10, 2.12 et les tableaux
page 49 : La figure 3.8	page 48 : des (a), (b) et (c) ont été ajoutés à la figure 2.8
page 52 : La figure 3.10	page 51 : des (a), (b) et (c) ont été ajoutés à la figure 2.10
page 55 : La figure 3.12	page 54 : des (a), (b) et (c) ont été ajoutés à la figure 2.12
page 57 : Nous nous intéressons dans ce chapitre	page 56 : Dans ce chapitre, nous nous intéressons
page 57 : Pour ces données nous disposons	page 56 : Pour ces données, nous disposons
page 57 : la taille en metre	page 56 : la taille en mètres
page 58 : Au départ les données	page 57 : Au départ, les données
page 58 : Au final les données se composent maintenant de 179 patients	page 57: Au final, on a utilisé 179 patients et 22 variables

page 58 : Nous appellerons cohorte entière	page 57 : Nous appellerons « cohorte entière »
page 58 : tels que décris dans le chapitre 1 et 2	page 57 : tels que décrits dans le chapitre 2
page 58 : liées au traitement	page 57 : liées au traitement représentant des agents chimiothérapeutiques hormis la radiothérapie
page 40 : décrit dans le paragraphe 3.4.2	page 39 : décrit à la section 2.4.2
page 67 : l'étude de simulations 3.5	page 66 : l'étude de simulation 2.5
page 67 : de position t	page 66 : de position θ
page 67 : sont p, m, σ, θ, a et b	page 66 : sont p, m, σ, θ, a et b
page 69 : sur données imputées (voir figure 4.13)	page 68 : sur données imputées (voir figure 3.11)
page 70 : la section 4.3, voir figure 4.12 et 4.13, le surpoids ayant	page 69 : la section 4.3 (voir figures 4.12 et 4.13), c'est-à-dire le surpoids a
page 4 : Figure 2.1 Conception de l'étude : Au moment du diagnostic de la leucémie, de nombreuses variables ont été mesurées, y compris l'IMC. Le traitement dure approximativement 2 ans, à quel point d'autres variables ont été mesurées. Le délai minimum est de 5 ans entre le diagnostic et l'entrevue PÉTALE	page 4 : Figure 1.1 Conception de l'étude : Des variables sont mesurées à 3 moments. Au moment du diagnostic de la leucémie, de nombreuses variables ont été mesurées, y compris l'IMC. Le traitement dure approximativement 2 ans, à quel point d'autres variables ont été mesurées. Le délai minimum est de 5 ans entre le diagnostic et l'entrevue PÉTALE.
page 5 : et la date du premier rendez-vous après le traitement	page 5 : et la date du premier rendez-vous juste après la fin du traitement
page 5 : la jème date	page 5 : la *j*ème date
page 10 : formaliser la notation introduite	page 9 : formaliser les notations introduites
page 10 : On note $f_{[L,R]}$ la densité conjointe	page 9 : On note $f_{(L,R)}$ la densité conjointe (parce que les parenthèses font bien références au couple (L,R))
page 11 : L'équation (1.2) implique que les paramètres de la distribution de L et R sont indépendants des paramètres de la distribution de T, ce qui veut dire que la seule information fournie par l'intervalle de censure (l, r] à propos du temps de survie t est que l'intervalle contient t	page 10 : L'équation (1.2) veut dire que la seule information fournie par l'intervalle de censure (l, r] à propos du temps de survie t est que l'intervalle contient t
page 11 : comme Sun (voir Sun [2007], p.14)	page 10 : comme Sun [2007], p.14,
page 12 : la densité jointe	page 11 : la densité conjointe
page 12 : il s'agira de la distribution S de T	page 11 : il s'agira de la fonction de survie S de T
page 14 : $f_{[T\|L,R]}$	page 13 : $f_{(T\|L,R)}$
page 14 : $f_{[L,R\|T]}$	page 13 : $f_{(L,R\|T)}$

page 14 : $f_{[T,L,R]}$	page 13 : $f_{(T,L,R)}$
page 13 : Démonstration. (i) D'après l'équation (1.2), on a	page 12 : Démonstration. On note $f_{(T\|L,R)}$ la densité conditionnelle de T sachant (L,R); $f_{(L,R\|T)}$ la densité conditionnelle de L, R sachant T; $f_{(T,L,R)}$ la densité conjointe de T, L, R. (i) D'après l'équation (1.2), on a
page 16 : L'estimateur de Kaplan-Meier découle de l'idée que (page 16) devient : Si par exemple la défaillance correspond à un décès, alors l'estimateur de Kaplan-Meier découle de l'idée que	page 15 : Si par exemple la défaillance correspond à un décès, alors l'estimateur de Kaplan-Meier découle de l'idée que
page 18 : c'est ce qu'on appelle de l'imputation (De Waal et al. [2011], Carpenter et Kenward [2012])	page 17 : c'est ce qu'on appelle de l'imputation (De Waal et al. [2011], Carpenter et Kenward [2012]). Dans le contexte de la censure par intervalle, la donnée manquante est l'instant précis de défaillance
page 18 : $z_i^* = l_i$, si $r_i^* = +\infty$	page 17 : $z_i^* = l_i$, si $r_i = +\infty$
page 33 : 3.4.1 Construction du taux de survie : Nous allons définir le taux de survie à partir d'un modèle de mélange...	page 32 : 2.4.1 Construction du taux de survie On appelle taux de survie du temps de défaillance T, la fonction h définie par h(x) = f(x)/S(x), avec la convention h(x) = 0 lorsque S(x) = 0...
page 36 : On a pris pour S0...	page 35 : Dans cet exemple on propose un choix de S_0 et S_2 qui vérifie la remarque 2.4.1. On a pris pour S_0...
page 40 : Cette étude de simulations a été réalisée en nous basant sur des connaissances a priori sur des données réelles et avant d'avoir accès au jeu de données étudié dans le chapitre suivvant	page 39 : Cette étude de simulation a été conçue en nous basant sur des connaissances a priori sur les données réelles étudiées dans le chapitre suivant (les simulations ont été réalisées avant d'avoir accès à ces données)
page 45 : Dans la pratique on choisit $q^*=t_T$ (voir section 3.3.2.3)	page 44 : Dans la pratique on choisit $q^*=t_T$ (voir section 2.3.2.3), et on approche l'équation (2.46) par une somme de Riemann.

page 72 : Voici nos principales contributions dans ce mémoire	page 72 : Ce mémoire porte sur la comparaison de deux estimateurs non paramétriques de la fonction de survie pour des données censurées par intervalle et propose également une application sur les données réelles. Les principales contributions sont décrites plus bas
page 72 : D'un point de vue théorique	page 72 : En premier, d'un point de vue théorique
page 72 : D'un point de vue computationnel	page 72 : Ensuite, d'un point de vue computationnel
page 74 : nous avons réalisé une étude de simulations	page 72 : nous avons réalisé une étude de simulation
page 72 : afin de comparer Kaplan-Meier sur données imputées et Turnbull	page 72 : afin de comparer l'estimateur de Kaplan-Meier sur données imputées et l'estimateur de Turnbull
page 72 : Certes, il existe déjà au moins une étude de simulations dans ce sens. En effet, Dehghan et Duchesne dans Dehghan et Duchesne [2011], proposent une comparai- son sur simulation du KMI et d'une généralisation de Turnbull (pour une fonction de survie avec une covariable), mais avec des instants de rendez-vous simulés se- lon un processus de Poisson	page 72 : Dehghan et Duchesne [2011] proposent une comparaison par simulation de l'estimateur KMI et d'une généralisation de l'estimateur de Turnbull (pour une fonction de survie avec une covariable), mais avec des instants de rendez-vous simulés selon un processus de Poisson
page 72 : Dans ce projet	page 72 : Dans ce mémoire
page 72 : D'un point de vue pratique	page 72 : Enfin, d'un point de vue pratique
page 33 : Bilan algorithme de Turnbull	page 32 : Bilan de l'algorithme de Turnbull
page 33 : Dans cette section, nous faisons un bilan des étapes de la dérivation de l'algorithme de Turnbull données à la section 3.3.2.1.	page 32 : Dans cette section, nous faisons un bilan des étapes de la dérivation de l'algorithme de Turnbull (voir algorithme 2) données à la section 2.3.2.1.

page 33 : En iterant plusieurs fois les étapes E et M de la partie 3.3.2.1 jusqu'à convergence (voir algorithme 2)	page 32 : On construit une suite $p^{(k)}$ pour approcher l'estimation de vraisemblance maximale de p, en itérant plusieurs fois les étapes E et M de la partie 2.3.2.1 jusqu'à convergence, c'est- à-dire jusqu'à ce qu'un critère d'arrêt soit vérifié. Nous proposons deux types de critères : l'un est basé sur la stabilisation de la vraisemblance (voir algorithme 2, ligne 7), l'autre est basé sur la stabilisation de la suite $p^{(k)}$ (voir algorithme 2, ligne 8). Dans nos simulations, nous utiliserons le premier critère
page 33 : on obtient l'estimateur	page 32 : On en déduit l'estimateur
page 73 : même si l'estimateur KMI est meilleur du point de vue des distances	page 73 : même si l'estimateur KMI est meilleur du point de vue de deux distances (Kolmogorov-Smirnov et L_2)
page 73 : Voici maintenant quelques pistes de recherches inspirées par les conclusions de ce mémoire	page 75 : Les conclusions de ce mémoire mènent à quelques pistes de recherche
page 73 : On pourrait également construire un test de comparaison de deux groupes pour les données censurées par intervalle en utilisant les estimateurs de Turnbull, voir Peace et al. [2012], Self et Grossman [1986], et Zhang et Sun [2010] et évaluer ses performances sur données simulées	page 73 : En utilisant l'estimateur de Turnbull, on pourrait également construire un test statistique de comparaison basé sur des données censurées par intervalle afin de déterminer si deux groupes d'individus sont significativement différents, par exemple dans l'application du chapitre 3, groupe des patients ayant reçu une radiothérapie crânienne et groupe des patients n'en ayant pas reçu (Peace et al. [2012], Self et Grossman [1986], Zhang et Sun [2010])
page 73 : On pourrait également prévoir la probabilité de devenir en surpoids à la fin de l'étude en utilisant un modèle de régression logistique adaptée aux données censurées par intervalle	page 73 : On pourrait également prévoir la probabilité de devenir en surpoids à la fin de l'étude à partir des covariables. Pour cela, il faudrait adapter le modèle classique de régression logistique aux données censurées par intervalle
RESUME	RÉSUMÉ
page 30 : par consequent de (3.30) on a	page 29 : par conséquent de (2.30) on a

page 31 : Donc, de l'équation (3.31) on en	page 29 : Donc, de l'équation (2.31), on en
page 58 : qui ne sont pas	page 57 : qui n'étaient pas
page 68 : Cette différence est peut-être due au fait qu'il y a moins de données sur la fin.	page 64 : Cette différence est peut-être due au fait qu'il y a moins de données sur la fin (51 % de censure à droite).

Remarques :

- Réponse à la question peut-on travailler avec $[L_i, R_i)$? (page 8) : Les patients qui sont initialement en surpoids au moment du diagnostic n'entrent pas dans notre analyse. Par conséquent le premier intervalle de censure doit être ouvert à gauche en zéro. Dans la littérature, les intervalles de censure sont de la forme $(L_i, R_i]$.
- Remarque page 10 : Comme l'application sur le surpoids est vraiment la motivation du mémoire et qu'on a familiarisé le lecteur avec cette application dans la section 2.1, il me semble pas utile de tenir compte de la première remarque de la page 10
- Réponse à la remarque page 15 (bien arranger les références ?) : ça se fait automatiquement, on a bien utilisé la commande \cite{} et ça met automatiquement le nom de l'auteur et la date.
- Réponse au commentaire de la page 15 : L'événement d'intérêt dans ce chapitre 2, ça reste la défaillance. L'idée de mort est utilisée dans une phrase pour expliquer l'idée derrière le Kaplan-Meier
- Réponse au commentaire de la page 17 : càdlàg c'est l'abréviation usuelle pour continu à droite avec limite à gauche
- Réponse au commentaire de la page 36 (je n'utiliserais pas Exemple) : ceci est bel et bien un exemple
- Réponse au commentaire peut-on justifier ces choix de paramètres? référence? (page 37) : il y a une référence pour cette idée de construction du mélange qu'on a déjà donné au début du paragraphe, mais le choix des lois et paramètres n'est pas celui de Marshall Olkin, nous avons fait un choix adapté à nos données réelles.
- Réponse au commentaire ça me semble long (page 69) : c'est ce qu'on constate sur le graphique, le Turnbull on a l'impression qu'il termine puis d'un coup il y a encore un saut. Le problème c'est qu'on ne peut pas savoir s'il s'agit d'un défaut de l'estimation ou bien d'un phénomène qui aurait du sens d'un point de vue médical. Il nous faudrait un retour d'un médecin sur la question, pour l'instant on ne sait pas. On ne peut donc que se limiter à la constatation.
- Réponse au commentaire stabilisation de la log-vraisemblance (page 72) : voir page 34
- Réponse au commentaire le dernier chapitre est assez pauvre en commentaires de base sur les résultats obtenus en analyse de données (Rapport d'arbitre sur le mémoire) : l'objectif principal du mémoire c'était d'étudier l'estimation non paramétrique de la fonction de survie pour des données censurées par intervalle et de l'appliquer aux données de leucémie pour déterminer des covariables liées au surpoids. Cette approche est pertinente lorsque la covariable est catégorielle (permet de définir les sous-groupes), puisque dans ce cas on peut comparer les fonctions de survie selon les sous-groupes. Pour ces données, la majorité des covariables sont continues. Dans le chapitre 3, nous avons surtout voulu illustrer l'intérêt du Turnbull, ce qui justifie d'avoir étudié le surpoids en fonction du dose cumulative de corticostéroïdes (élevée ou faible) et puis radiothérapie crânienne (Oui ou Non). Pour les

covariables continues, nous avons essayé quand même d'étudier la dépendance même de manière sommaire en faisant les boxplots selon surpoids (Oui ou Non). Des nouveaux commentaires à propos des boxplots ont été ajoutés dans le mémoire.

- Réponse aux corrections remplacer estimation par estimé ou estimateur (page 67, 68, 73) : On a gardé le terme usuel d'estimation lorsqu'on a évalué l'estimateur sur le jeu de données réelles.