République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

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

Département de Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathématiques Appliquées**"

Option : **Statistique**

Par

MOUSSI Selma Hibat Ellah

Titre :

Sur les données doublement tronquées

Devant le Jury :

Mr.	BENATIA Fatah	Prof.	U. Biskra	Président
Mr.	YAHIA Djabrane	Prof.	U. Biskra	Rapporteur
Melle.	SOLTANE Louiza	MCB	U. Biskra	Examinatrice

Soutenu Publiquement le 28/06/2022

Dédicace

Ce humble travail est dédié au maître de la création, le prophète Muhammad (paix et bénédictions d'Allah soient sur lui)

À la bougie brillante qui s'est flétrie pour éclairer mon chemin, ma chère mère Moussi Fatiha

À mon cher Père Moussi Slimane qui m'a donné sa confiance

À tous mes enseignants, sœurs, frères et amis qui m'ont soutenu sur les chemins de la vie

À tous ceux qui sont tombés accidentellement de ma mémoire.

Remerciements

Tout d'abord je tiens à remercier Dieu de m'avoir donné le courage, le morale et la santé pour mener à bien ce travail.

Mes plus vifs remerciements vont à mon cher père et à ma chère mère pour leur patience et leurs sacrifices ainsi d'atteindre cette étape avancée de l'étude.

Je tiens ensuite à exprimer toute ma gratitude envers à mon encadreur Prof. Yahia D. pour son disponibilité, ses encouragements, ses précieux conseils, son sens de l'écoute et pour toute l'aide. Mes remerciements s'adressent également aux membres du jury, qui m'ont fait l'honneur de bien vouloir étudier avec attention mon travail : Prof. Benatia F. (Président) et Melle Soltane L. (Examinatrice).

Je tiens également à exprimer mes remerciements à tous les amis et les collègues du département de mathématiques (Sellami. Ch, Baadache. R,...) et à l'étranger et à tous les membres de Département de Mathématiques.

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

Merci à tous

Notations et symboles

$i.i.d$	Indépendantes et indentiquement distribué
$v.a$	variable aléatoire
$\mathbf{1}_A$	Fonction indicatrice de l'ensemble A
δ_j	Indicateur de censure
$X_{n,n}$	Maximum de $X_1, ..., X_n$
$X_{1,n}$	Minimum de $X_1, ..., X_n$
$X_{j,n}$	$j^{ème}$statistique d'ordre
$X_1, ..., X_n$	Une suite de n v.a
F	Fonction de répartition
F_n	Fonction de répartition empirique
$F_n^{(km)}$	L'estimateur de Kaplan-Meier
$\mathbb{R}$	Ensemble de valeur réelles
$E(x)$	Espérance mathématique
$Var(x)$	Variance mathématique

Θ	espace des paramètre
$ENPMV$	Estimation non paramétrique par maximum de vraisemblance
EMV	Estimation du maximum de vraisemblance
$\overset{p.s}{\rightarrow}$	Converge presque sûre
$\overset{p}{\rightarrow}$	Converge en probabilités
$\overset{Loi}{\rightarrow}$	Converge en loi
$MISE$	L'eureur quadratique moyenne intégrée
MSE	L'eureur quadratique moyenne
$N(\mu, \sigma^2)$	Loi Normale de moyenne μ et de variance σ^2
$Exp(a)$	Loi Exponentielle de paramètre a
Φ	Distribution de la loi normale $N(0,1)$
Y^*	Variable aléatoire d'intérêt
U^*	Variable de troncature gauche
V^*	Variable de troncature droite ($U^* \leq V^*$).
$\frac{d}{\partial u}$	Dirivée partielle
a_W	Point limite gauche de la distribution W
b_W	Point limite droite de la distribution W
$K(.)$	Fonction du Noyau
$h = h_n$	Fenêtre (bandwidth en anglais).
h_{AMISE}	Fenêtre Optimale.

Table des matières

Table des figures

Liste des tableaux

Introduction

Le problème des données manquantes, incomplètes ou erronées est très vaste et a suscité beaucoup d'intérêt des statisticiens ces dernières années. L'attitude de ce type de données a longtemps été soit de les éliminer, soit de minimiser le mauvais impact qu'elles pourraient avoir sur des procédures statistiques adaptées à des données complètes. Dans le domaine des durées de survie, les données sont souvent incomplètes à cause de deux phénomènes distincts : la censure et la troncature, voir Huber-Carol (1994), pour plus d'information et détailles.

Souvent, les chercheurs peuvent rencontrer des difficultés pour acquérir des échantillons dont les mesures sont en dehors des plages spécifiques. Efron et Petrosian (1999) ont fournis un excellent exemple de double troncature, où il est impossible pour les astronomes d'évaluer la luminosité des quasars s'ils sont trop faibles (troncature à gauche) ou trop lumineux (troncature à droite). Le problème de la double-troncature se pose notamment lorsque les chercheurs tentent d'utiliser des données de terrain ou des données d'observation pour faire des inférences sur une population.

La double troncature peut entraîner un biais systématique dans le contenu des données en raison de la perte d'informations. Les données doublement tronquées sont souvent observées dans de nombreux domaines, comme l'économie, la méde-

cine, l'ingénierie,... Dans ce mémoire, nous donnons un aperçu sur les données doublement tronquées. Nous mettons l'accent sur l'estimation de la densité et de la distribution, tout en précisant le cadre statistique des données et leur modèle sous troncature double. Des travaux de simulation à l'aide du logiciel de traitement statistique R sont réalisés pour confirmer le bon comportement et pour évaluer la performance des différents estimateurs étudiés.

Ce mémoire est composé de trois chapitres comme suit :

Le premier chapitre est adopté aux généralités sur les données incomplètes (censure à droite, à gauche et par intervalle) et (troncature à droite, à gauche et par intervalle). Après cela on parle de l'estimation de la fonction de répartition sous le modèle tronqué. Le deuxième chapitre se regroupe en deux sections. Dans la première section nous présentons quelques exemples réels de données doublement tronquées. Puis, dans la deuxième section nous rappelons le modèle de probabilités sous la double troncature. Le dernier chapitre est consacré aux estimateurs de la distribution et de la densité et leurs propriétés asymptotiques sous troncature double. Nous illustrons aussi leurs performances sur quelques exemples de simulation à l'aide du logiciel de traitement statistique R. Enfin, nous décrivons quelques remarques de conclusion et des perspectives de recherche sur le sujet traité.

Chapitre 1

Généralités sur les données incomplètes

Une des caractéristiques des données de survie est l'existence d'observations incomplètes, dont les données sont souvent recueillies partiellement, notamment, à cause des processus de censure ou de troncature. Les données censurées ou tronquées proviennent du fait de non accès à toute l'information. En effet, au lieu d'observer des réalisations indépendantes et identiquement distribuées d'une durée X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié. Pour plus d'information sur les données incomplètes, on recommande de voir Hughes (*1962*), Saint-Pierre (*2015*) et Touraine (*2013*).

1.1 Données censurées

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. on désigne par C le temps de censure et par Z la durée réellement

observée. On dispose donc de trois variables aléatoires (v.a's) X, C et Z à partir desquelles on extrait des échantillons de taille $n > 1$: $(X_1, ..., X_n)$, $(C_1, ..., C_n)$ et $(T_1, ..., T_n)$ respectivement. Pour tout individu i, on a donc :

– son temps de survie X_i,

– son temps de censure C_i,

– la durée réellement observée Z_i,

1.1.1 Censure à droite ou gauche

Définition 1.1.1 *Une durée de vie aléatoire X est dite censurée par une va de censure C si on observe parfois C au lieu de X. L'information donnée par C sur X est :*

$$X > C \text{ s'il y a censure droite} \quad \text{et} \quad X < C \text{ s'il y a censure gauche.}$$

1.1.2 Censure aléatoire a droite

Définition 1.1.2 *La durée Y est dite censurée aléatoirement à droite si au lieu d'observer $Y_1, ..., Y_n$, on observe*

$$Z_i = min(Y_i, C_i) \quad \text{et } \delta_i = \mathbf{1}_{\{Y_i \leq C_i\}} \text{ pour } i = 1, ..., n$$

où C est une censure aléatoire et le δ sert en fait à connaitre la nature de l'observation, il indique si l'on est face à une observation réelle ($\delta = 1$) ou à une censure ($\delta = 0$).

Exemple 1.1.1 *Un exemple typique est celui ou l'événement considéré est le décès d'un patient et la durée d'observation est la durée totale d'hospitalisation.*

Remarque 1.1.1 *On peut aussi observer ce genre de phénomène dans les études de fiabilité quand la panne d'un appareil ne permet pas de continuer l'observation pour un autre appareil. Pour ce type de censure tout ce que l'on sait est que la vraie durée de survie est supérieure à la durée observée.*

Exemple 1.1.2 *Un exemple artificiel de données censurées aléatoire à droite impliquant* $n = 10$ *points de données est donné dans le tableau 1.1. On voit que l'échantillon observé est donné par*

$$(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9, Z_{10}) = (X_1, C_2, X_3, X_4, C_5, X_6, X_7, X_8, X_9, X_{10})$$

pour laquelle $Z_j = \min(X_j, C_j)$. *La variable indicatrice est*

$$\delta_j = \mathbf{1}(X_j < C_j) = (1, 0, 1, 1, 0, 1, 1, 1, 1, 1) = \begin{cases} 1, & \textit{Si } X_j \textit{ a été observé} \\ 0, & \textit{Si } C_j \textit{ est censuré} \end{cases}$$

j	X_j	C_j	$Z_j = \min(X_j, C_j)$	δ_j
1	10.85	11.45	10.85	1
2	7.65	7.54	7.54	0
3	10.02	10.25	10.02	1
4	9.08	9.88	9.08	1
5	11.12	9.15	9.15	0
6	9.68	10.25	9.68	1
7	6.63	7.65	6.63	1
8	4.02	5.41	4.02	1
9	5.03	8.03	5.03	1
10	5.21	5.77	5.21	1

TAB. 1.1 – Un example artificiel de données censurées à droite.

1.1.3 Censure aléatoire à gauche

Définition 1.1.3 *La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu est observé.*

La durée Y est dite censurée aléatoirement à gauche si au lieu d'observer $Y_1, ..., Y_n$, on observe (Z_i, δ_i) où

$$Z_i = max(Y_i, C_i) \text{ et } \delta_i = \mathbf{1}_{\{Y_i \geq C_i\}} \text{ pour } i = 1, ..., n$$

et C_i est une censure aléatoire.

Exemple 1.1.3 *Supposons par exemple, on s'intéresse à l'âge à partir duquel une personne commence à accomplir une certaine tâche. Certaines personnes peuvent ne pas se rappeler et donner juste une valeur supérieure. Cette donnée est donc censurée à gauche.*

1.2 Données tronquées

Une autre situation dans laquelle les données incomplètes apparaissent est celle des données tronquées. La troncature est différente de la censure au sens où elle concerne l'échantillon lui-même. Une observation est dite tronquée si elle est conditionnelle à un autre évènement. On dit que la variable Y de durée de vie est tronquée si Y n'est observable que sous une certaine condition dépendante de la valeur de Y.

1.2.1 Troncature gauche

Définition 1.2.1 *On dit qu'il y a troncature gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à T. T est alors la va de troncature gauche :*

$$X \text{ n'est observée que si } X > T.$$

Exemple 1.2.1 *On étudie la durée de vie après la retraite de sujets qui entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite. Un sujet n'est donc observé que si sa durée de vie après la retraite excède le délai entre sa prise de retraite et l'instant de l'enquête. La durée de vie après la retraite est donc tronquée à gauche par ce délai. Elle peut aussi être censurée à droite si la fin de l'enquête a lieu alors que le sujet est toujours vivant (voir, Chaib,* 2013*).*

1.2.2 Troncature droite

Définition 1.2.2 *On dit qu'il y a troncature droite lorsque X n'est observable que si elle est inférieure à T. T est alors la v.a de troncature droite :*

$$X \text{ n'est observée que si } \ X < T.$$

Exemple 1.2.2 *Klein et Moeschberger (*2003*), présentent des données sur les temps d'infection et l'induction pour* 258 *adultes et* 37 *enfants qui ont été infectés par le virus de SIDA. Ici, le nombre de personnes infectées est inconnu et l'information est disponible seulement pour ceux qui ont été infectés et développés le SIDA dans un certain laps de temps. Ainsi, les personnes qui n'ont pas encore développé le SIDA ne sont pas connues à l'enquêteur et ne sont pas inclues dans l'échantillon. C'est le cas de troncature à droite.*

Exemple 1.2.3 *Un exemple artificiel de données tronquées aléatoire à droite impliquant* $N = 10$ *points de données est donné dans le tableau 1.2. Nous voyons que seulement* $(X_3, X_5, X_6, X_7, X_8, X_9)$ *sont observés, mais pas* (X_1, X_2, X_4, X_{10}). *Dans ce cas, l'échantillon observé est donné par*

$$(X_3^*, X_5^*, X_6^*, X_7^*, X_8^*, X_9^*) = (0.09, 2, 3.55, 6.09, 4.01, 8.18)$$

j	X_j	T_j	X_j^*
1	8.05	0.56	
2	0.11	0.01	
3	0.09	1.2	0.09
4	1.7	1.3	
5	2	3.35	2
6	3.55	4.58	3.55
7	6.09	8.07	6.09
8	4.01	5.03	4.01
9	8.18	9.81	8.18
10	7.15	6.14	

TAB. 1.2 – Un exemple artificiel de données tronquées à droite.

La probabilité de troncature est d'environ 40%, donc :

$$p = P(X \leq T) \simeq (\frac{n}{N}) = (\frac{6}{10}) = 0.6.$$

1.2.3 Troncature par intervalle

Définition 1.2.3 *Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.*

Exemple 1.2.4 *On rencontre ce type de troncature lors de l'étude des patients d'un registre : les patients diagnostiqués avant la mise en place du registre ou répertoriés après la consultation du registre ne seront pas inclus dans l'étude.*

Remarque 1.2.1 *Plus généralement, il y a troncature si l'observation de la variable d'intérêt X n'a lieu que conditionnellement à un évènement B.*

1.3 Estimation de la fonction de répartition sous troncature

Soit $Y_1, ..., Y_N$ une suite de variable aléatoires réelles d'intérêt i.i.d. de fonction de répartition (f.d.r) commune F et de densité continue f. Soit $T_1, ..., T_N$ une suite de variable aléatoires de troncature i.i.d. de f.d.r. continue G. Les variables aléatoires T_i sont supposées être indépendantes des Y_i. Soit $(Y_1, T_1), ..., (Y_n, T_n)$ l'échantillon réellement observé et définissons la probabilité de troncature par $\alpha = P(Y_1 \geq T_1) > 0$. Il est clair que si $\alpha = 0$, aucune donnée peut être observée.

Par la loi forte des grands nombres on a, lorsque N tend vers ∞ :

$$\tilde{\alpha}_n = \frac{n}{N} \to \alpha, \; P \text{ - } p.s$$

Remarque 1.3.1 *Lemdani et Ould Saïd (*2007*) ont prouvé que la propriété i.i.d de l'échantillon observé de taille n est déduite de celle de l'échantillon de taille N.*

Sous le modèle de troncature à gauche, la distribution conjointe conditionnelle

(voir Stute (*1993*), et Zhou (*1996*)) d'un (Y,T) observé devient

$$\begin{aligned} J^*(y,t) &= P(Y \leq y, T \leq t) = P(Y \leq y, T \leq t | Y \geq T) \\ &= \frac{P(Y \leq y, T \leq t, Y \geq T)}{P(Y \geq T)} \\ &= \frac{P(Y \leq y, T \leq t, T \leq y)}{P(Y \geq T)} \\ &= \alpha^{-1} \int_{-\infty}^{t} G(t \wedge u) dF(u) \\ &= \alpha^{-1} \int_{-\infty}^{t} G(\min(t,u)) dF(u). \end{aligned}$$

Les distributions marginales sont donc définies par :

$$\begin{aligned} F^*(t) &= H^*(y, \infty) \\ &= \alpha^{-1} \int_{-\infty}^{t} G(u) dF(u) \end{aligned}$$

et

$$\begin{aligned} G^*(t) &= J^*(\infty, t) \\ &= \alpha^{-1} \int_{-\infty}^{+\infty} G(t \wedge u) dF(u) \\ &= \alpha^{-1} \int_{-\infty}^{+\infty} \int_{-\infty}^{t \wedge u} dG(v) dF(u) \\ &= \alpha^{-1} \int_{-\infty}^{t} dG(v) \int_{v}^{+\infty} dF(u) \\ &= \alpha^{-1} \int_{-\infty}^{t} (1 - F(v)) dG(v), \end{aligned}$$

qui peuvent être estimés respectivement par

$$F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq y\}} \text{ et } G_n^*(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{T_i \leq t\}}.$$

Soit la fonction $C(.)$ définie par

$$\begin{aligned} C(y) &= P(T \leq y \leq Y | Y \geq T) = G^*(y) - F^*(y) \\ &= \frac{P(T \leq y \leq Y, Y \geq T)}{P(Y \geq T)} = \frac{P(T \leq y \leq Y)}{P(Y \geq T)} \\ &= \alpha^{-1} P(T \leq y) P(Y \geq y) \\ &= \alpha^{-1} G(y)(1 - F(y)), \end{aligned}$$

qui peut être estimée par

$$C_n(y) = G_n^*(y) - F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Ti \leq y \leq Yi\}}$$

Finalement, Lynden-Bell (*1971*) introduit les estimateurs de maximum de vraisemblance non paramétriques de F et G appelés estimateurs produit-limite comme suit :

$$F_n(y) = 1 - \prod_{i;Y_i \leq y} \left[\frac{nj_n(T_i) - 1}{nj_n(T_i)} \right] \text{ et } G_n(t) = \prod_{i:T_i > y} \left[\frac{nj_n(Y_i) - 1}{nj_n(Y_i)} \right].$$

Chapitre 2

Données doublements tronquées

La troncature fait référence à un phénomène selon lequel certains individus de la population ont moins de chances d'être sélectionnés. En particulier, on ne peut pas identifier un individu dont la durée de vie est inférieure ou supérieure à un seuil, appelé limite de troncature.

La double troncature est un type de troncature où la troncature gauche et la troncature droite se produisent simultanément. Par exemple, en astronomie, les objets stellaires dans les galaxies ne sont pas détectés s'ils sont trop brillants ou trop faibles. Cet exemple de quasar est le phénomène étudier initialement par Efron et Petrosian (*1999*), ils ont motivé le développement ultérieur de l'analyse de survie avec la double troncature, bien que la luminosité ne soit pas la durée de vie au sens littéral.

2.1 Exemples de troncature double

2.1.1 Données du cancer infantile

Le terme cancer infantile désigne les cancers qui surviennent entre la naissance et l'âge de 15 ans des enfants. Moreira et de Uña Álvarez (*2010*) ont fourni un ensemble de données sur 406 enfants dans le nord du Portugal, ce cancer a été diagnostiqué au cours d'une période de recrutement de 5 ans (entre le 1er janvier 1999 et le 31 décembre 2003).

▷ Enfants diagnostiqués avant le 1er janvier 1999 ou après le 31 décembre 2003 n'existent pas dans l'ensemble de données, et par conséquent, les données sont doublement tronquées.

Soit y^* l'âge au moment du diagnostic de cancer pour un enfant choisi au hasard parmi une population. Le critère d'inclusion de l'échantillon s'écrit $u^* \leq y^* \leq v^*$, où u^*est l'âge au 1er janvier 1999 et $v^* = u^* + 5$(ans) est l'âge au 31 décembre 2003 (voir Fig. 2.1). Ainsi, la limite de troncature gauche est u^*et la limite de troncature droite est v^* :

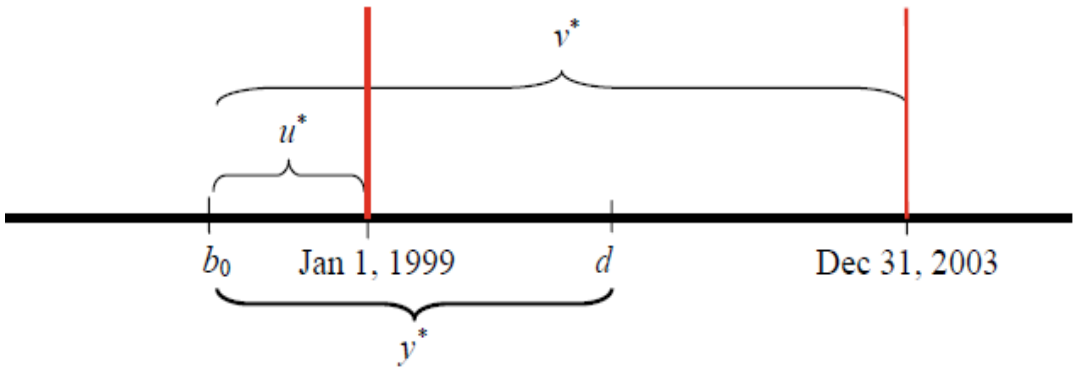

FIG. 2.1 – La journée du cancer infantile au nord du Portugal (Moreira et de Uña-Álvarez, 2010)

b_0 : date de naissance d'un enfant

d : date du diagnostic (date de l'apparition du cancer)

y^* : âge au moment du diagnostic (âge du cancer)

u^* : âge au 1er janvier 1999 (u^* est négatif pour les personnes nées après le 1er janvier 1999)

$v^* = u^* + 5$ (ans) : âge au 31 décembre 2003.

2.1.2 Données d'équipements

Les données d'équipements de Ye et Tang (*2016*) forment un troisième exeùple de données doublement tronquées. Ces données comprennent les temps de défaillance des unités, appelés **équipements-S**, et leurs dates d'installation qui varient de 1977 à 2012. Le suivi a commencé en 1996 lorsque le service de maintenance à détecter l'importance de collecter les données sur l'équipements-S. Par conséquent, une unité est observée en cas d'échec entre 1996 et 2012. La population d'intérêt est l'ensemble de toutes les unités installées. Les unités ayant échoué avant 1996 ou après 2012 n'existent pas dans l'ensemble de données et par conséquent, les données sont doublement tronquées.

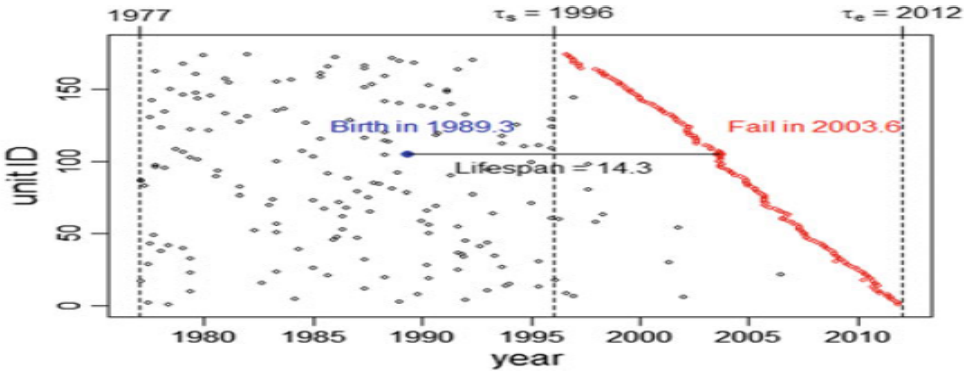

FIG. 2.2 – Données d'équipements de Ye et Tang (2016).

La figure *2.2* représente les données équipements de Ye et Tang (*2016*). Par exemple, une unité a l'année d'installation en 1989.3 et l'année de défaillance en 2003.6. Par conséquent, la durée de vie est de $2003,6 - 1989,3 = 14,3$ (années). On peut définir la limite de troncature gauche notée u_i, le temps de défaillance y_i et la limite de troncature droite notée v_i de manière similaire aux exemples *2.1.1* et *2.1.3*. Ensuite, les données sont constituées de (u_i, y_i, v_i).à $u_i \leq y_i \leq v_i$ pour $i = 1, 2, ...n$, où $n = 174$.

Remarque 2.1.1 *Les exemples 2.1.1, 2.1.3 et 2.1.2 ont la même structure mathématique pour les limites de troncature, à savoir, $v^* = u^* + d$ pour une constante fixe $d > 0$. Cependant, dans l'exemple 2.1.1, le support de y^* est connu pour être $[0, 15]$ depuis l'enfance, les cancers surviennent entre la naissance et l'âge de 15 ans. Les données observées couvre principalement cette plage en $min(y_i^*) = 0,0164$ et $max(y_i^*) = 14.997$. Cependant, dans les exemples 2.1.3 et 2.1.2, la borne supérieure du support de y^* serait l'infini étant donné que certaines entreprises ou machines peuvent avoir une durée de vie assez longue.*

2.1.3 Données de l'entreprise allemande

Dörre (*2020*) à étudier les données de l'entreprise allemande comme exemple de double troncature. Ces données incluent les entreprises allemandes qui ont été déclarées insolvables lors de la collecte des données sur la période entre le 1er septembre 2013 et le 31 mars 2014 (voir Fig. 2.3). La variable y^* est la durée de vie d'une entreprise sélectionnée au hasard (en années). Cependant, les données ne contiennent aucune information sur les entreprises déclarées insolvables :

b_0 : date de fondation d'une entreprise (naissance)

d : date de l'insolvabilité (décès)

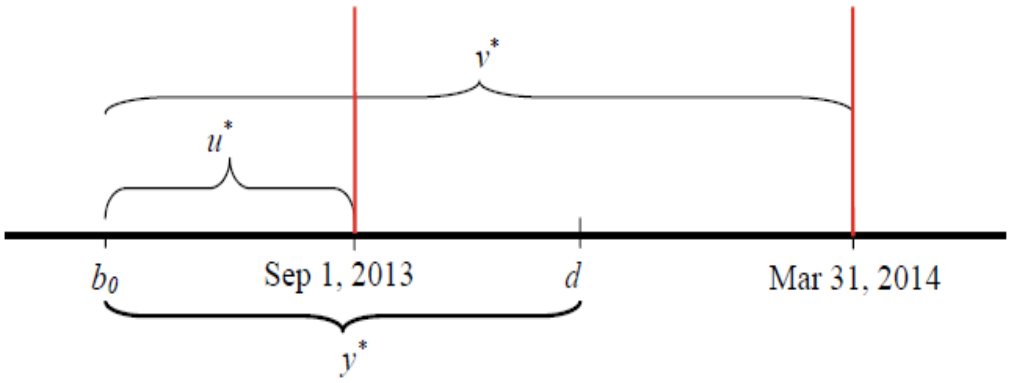

FIG. 2.3 – Les données sur les entreprises allemandes Dörre (*2020*)

y^* : durée de vie d'une entreprise (en années)

u^* : âge d'une entreprise au 1er septembre 2013 (en années)

$v^* = u^* + 7/12$: âge d'une entreprise au 31 mars 2014 (en années).

Remarque 2.1.2 *En dehors de cette période, on ne peut pas déterminer les entreprises qui ont été créées puis insolvables avant la période (par exemple, les petites entreprises). En outre, on ne peut pas déterminer les entreprises qui poursuivent leurs activités au-delà de la période (par exemple, grandes entreprises stables). Formellement, le critère d'inclusion de l'échantillon s'écrit* $u^* \leq y^* \leq v^*$*, où* u^* *est l'âge d'une entreprise en septembre* 2013 *et* $v^* = u^* + 7/12$ *pour la période de* 7 *mois. Ainsi, la limite de troncature gauche est* u^* *et la limite de troncature à droite est* v^**. L'ensemble de données contient* $n = 4139$ *(entreprises formées après* 2000*).*

2.2 Modèles de probabilités pour la double troncature

Pour fournir une structure probabiliste de double troncature, nous considérons trois variables aléatoires :

- y^* : une durée de vie continue ayant une fonction de densité f, à savoir $f(y) = \frac{d}{dy}P(y^* \leq y)$.
- u^* : une limite de troncature à gauche.
- v^* : une limite de troncature à droite.

Étant donné que les données doublement tronquées incluent les individus satisfaisant $u^* \leq y^* \leq v^*$, nous définissons la probabilité d'inclusion ou la probabilité de sélection définie comme $P(u^* \leq y^* \leq v^*)$.

On suppose typiquement l'indépendance entre y^* et (u^*, v^*), à savoir

$$P(y^* \in A, (u^*, v^*) \in B) = P(y^* \in A) \times P((u^*, v^*) \in B)$$

pour les événements A et B. Une condition équivalente est

$$P(y^* \leq y, u^* \leq u, v^* \leq v) = P(y^* \leq y) \times P(u^* \leq u, v^* \leq v), \qquad \forall (y, u, v).$$

Cette hypothèse signifie que les durées de vie des populations ne sont pas influencées par les limites de troncature. En effet, une durée de vie plus courte s'accompagne d'une plage de limites de troncature plus courte.

D'après les travaux d'Efron et Petrosian (*1999*), Martin et Betensky (*2005*), Emura et Wang (*2010*) et Shen (*2011*) la double troncature aléatoire signifie

que u^* et v^* sont des variables aléatoires continues ayant une densité jointe

$$k(u,v) = \frac{d^2}{\partial u \partial v} P(u^* \leq u, v^* \leq v).$$

Alors, la probabilité d'inclusion est

$$\begin{aligned} P(u^* \leq u, v^* \leq v) &= \int\int_{u \leq v} [\int_u^v f(y)dy] k(u,v)dudv \\ &= \int [\int\int_{u \leq y \leq v} k(u,v)dudv] f(y)dy. \end{aligned}$$

Exemple 2.2.1 *Soient y^*, u^* et v^* trois v.a's de lois normale, de même variance et de moyennes différentes :*

$$y^* \sim N(\mu, 1), \quad u^* \sim N(\mu_u, 1) \quad et \quad v^* \sim N(\mu_v, 1).$$

La probabilité d'inclusion est

$$P(u^* \leq u, v^* \leq v) = \int_{-\infty}^{+\infty} \Phi(y - \mu_u)\{1 - \Phi(y - \mu_v)\Phi(y - \mu)dy$$

Si $\mu_u = \mu - 0,91$ et $\mu_v = \mu + 0,91$, alors

$$P(u^* \leq y^* \leq v^*) \approx 0,5.$$

La double troncature de longueur fixe est donnée par la relation $v^* = u^* + d$, où d est une valeur déterministe (non aléatoire) et u^* est une variable aléatoire continue ayant une fonction de densité

$$g(u) = \frac{d}{du} P(u^* \leq u).$$

Les exemples *2.1.1*, *2.1.3* et *2.1.2* sont des cas de double troncature de longueur fixe. Dans ces exemples, la densité de u^* décrit le processus de naissance et peut avoir une distribution connue (par exemple la distribution uniforme). La probabilité d'inclusion est

$$P(u^* \leq y^* \leq v^*) = \int [\int_u^{u+d} f(y)dy]g(u)du = \int [\int_{y-d}^{y} g(u)du]f(y)dy.$$

Chapitre 3

Estimation de la densité sous troncature double

Dans ce chapitre nous essayons d'étudier le problème d'estimation de la fonction de densité pour les données doublement tronquées. Soit respectivement. Ce qui signifie que le triplet (U^*, Y^*, V^*) est observé si et seulement si $U^* \leq Y^* \leq V^*$, bien qu'aucune information ne soit disponible lorsque $Y^* < U^*$ou $Y^* > V^*$.

On suppose que (U^*, V^*) est indépendant de Y^*. En l'absence d'une telle hypothèse d'indépendance, la récupération de la distribution de Y^* n'est pas possible en général. Martin et Betensky (*2005*) a discuté d'une procédure de test pour la quasi-indépendance, une plus faible hypothèse sous laquelle les méthodes discutées ici sont toujours cohérentes. Laisser $(U_i, Y_i, V_i), i = 1, ..., n$, désignent les informations d'échantillonnage, ce sont des données i.i.d avec la même distribution de (U^*, Y^*, V^*) étant donné $U^* \leq Y^* \leq V^*$.

Introduire $\alpha = P(U^* \leq Y^* \leq V^*)$, la probabilité de non-troncature. Pour tout distribution W, les extrémités (point limite) gauche et droite de son support notées

respectivement par

$$a_W = inf\{t : W(t) > 0\} \quad et \quad b_W = inf\{t : W(t) = 1\}.$$

Soit $T_1(u) = T(u, \infty)$ et $T_2(v) = T(\infty, v)$ les lois marginales de U^* et V^* respectivement. Lorsque $T_1 \leq a_F \leq a_{T2}$ et $b_{T1} \leq b_F \leq b_{T2}$, F et T sont tous deux identifiables (voir, Woodroofe, *1985*).

Dans les deux sections suivantes, nous introduisons respectivement les estimateurs non-paramétrique et semi-paramétrique de la distribution de Y^*. Ensuite, nous considérons le problème de l'estimation de la fonction de densité sur la base de ces deux estimateurs.

3.1 Estimateur non paramétrique de la distribution

Premièrement, rappelons l'estimateur non paramétrique par maximum de vraisemblance noté $(ENPMV)$ de Efron et Petrosian(*1999*). Soit $\varphi = (\varphi_1, ..., \varphi_n)$ une distribution mettant la probabilité φ_i sur Y_i, $i = 1, ..., n$. De même, soit $\psi = (\psi_1, ..., \psi_n)$ une distribution mettant la probabilité conjointe ψ_i sur $(U_i, V_i), i = 1, ..., n$. De l'hypothèse d'indépendance entre Y^* et (U^*, V^*), la vraisemblance $\mathcal{L}(\varphi, \psi)$, peut être décomposée comme un produit de la vraisemblance conditionnelle du Y_i notée $\mathcal{L}_1(\varphi)$, et la probabilité marginale de (U_i, V_i) notée $\mathcal{L}_2(\varphi, \psi)$:

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\varphi_j}{\Phi_j} \times \prod_{j=1}^{n} \frac{\Phi_j \psi_j}{\sum_{i=1}^{n} \Phi_i \psi i} = \mathcal{L}_1(\varphi) \times \mathcal{L}_2(\varphi, \psi) \tag{3.1}$$

où Φ_i est défini par $\Phi_i = \sum_{m=1}^{n} \varphi_m J_{im}$, $i = 1, ..., n$ avec $J_{im} = \mathbf{1}_{[U_i \leq Y_m \leq V_i]} = 1$ si $U_i \leq Y_m \leq V_i$ et égal à zéro sinon.

L'ENPMV conditionnel de F (Efron et Petrosian,*1999*) est défini comme le maximum de $\mathcal{L}_1(\varphi)$ dans l'équation (3.1) :

$$\hat{\varphi} = Arg \max_{\varphi} \mathcal{L}_1(\varphi).$$

Shen (*2010*) a prouvé que le NPEMV $F_n(x) = \sum_{i=1}^{n} \hat{\varphi}_i \mathbf{1}_{[Y_i \leq x]}$ maximise en effet la vraisemblance, qui peut être aussi écrite comme le produit

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\psi_j}{\Psi_j} \times \prod_{j=1}^{n} \frac{\Psi_j \varphi_j}{\sum_{i=1}^{n} \Psi_i \varphi i} = \mathcal{L}_1^*(\psi) \times \mathcal{L}_2^*(\psi, \varphi)$$

où $\Psi_i = \sum_{m=1}^{n} \Psi_m \mathbf{1}_{[U_m \leq X_i \leq V_m]} = \sum_{m=1}^{n} \Psi_m J_{mi}$, pour $i = 1..., n$. Ici,$\mathcal{L}_1^*(\psi)$ désigne la probabilité conditionnelle des (U_i, V_i) étant donné les X_i et $\mathcal{L}_2^*(\psi, \varphi)$ fait référence à la probabilité marginale des X_i. De même, notons $\hat{\psi} = (\hat{\psi}_1, ..., \hat{\psi}_n)$ en tant que maximum de $\mathcal{L}_1^*(\psi)$, alors, $T_n(u, v) = \sum_{i=1}^{n} \hat{\psi}_i \mathbf{1}_{[U_i \leq u, V_i \leq v]}$ est le NPEMV de la distribution T (voir, Shen, *2010*).

L'ENPMV de F admet également la représentation

$$F_n(y) = \alpha_n \int_{\alpha F}^{x} \frac{F_n^*(dt)}{G_n(t)}$$

où F_n^* est la distribution empirique ordinaire des Y_i,

$$G_n(t) = \int_{\{u \leq t \leq v\}} T_n(du, dv)$$

est un estimateur non paramétrique de la probabilité conditionnelle d'échantillonnage d'une durée de vie $Y^* = t$, c'est-à-dire $G(t) = P(U^* \leq t \leq V^*)$, et

$$\alpha_n = (\int_{a_F}^{\infty} G_n^{-1}(t) F_n^*(dt))^{-1}$$

est un estimateur pour α.

Remarque 3.1.1 *Shen (*2010*) a établi la convergence forte uniforme et la convergence faible de* F_n.

3.2 Estimateur semi-paramétrique de la distribution

Dans l'approche semi-paramétrique, on suppose que la distribution T appartient à une famille de distribution paramétrique $\{T_\theta\}_{\theta \in \Theta}$, où θ est un vecteur de paramètres et Θ représente l'espace paramétrique. En conséquence, $G(t)$ est paramétré comme suit :

$$G_\theta(t) = \int_{\{u \leq t \leq v\}} T_\theta(du, dv)$$

Le paramètre θ est estimé par la fonction de maximisation $\hat{\theta}$ de la vraisemblance conditionnelle des (U_i, V_i) étant donné les Y_i,

$$\mathcal{L}_1^*(\psi) \equiv \mathcal{L}_1^*(\theta) = \prod_{i=1}^{n} \frac{g_\theta(U_i, V_i)}{G_\theta(X_i)}$$

où

$$g(u, v) = \frac{\partial^2}{\partial u \partial v} P(U^* \leq u, V^* \leq v) = T_\theta(du, dv)$$

représente la densité jointe de (U^*, V^*). Une fois θ estimé, un estimateur semi-paramétrique pour F est introduit par

$$F_{\hat{\theta}}(y) = \alpha_{\hat{\theta}} \int_{a_F}^{x} \frac{F_n^*(dt)}{G_{\hat{\theta}}(t)}, \quad \text{où} \quad \alpha_{\hat{\theta}} = \left(\int_{a_F}^{\infty} G_n^{-1}(t)\, F_n^*(dt) \right)^{-1}.$$

Remarque 3.2.1 *Moreira et de Uña-Álvarez (*2010*) ont montré la normalité asymptotique des $\hat{\theta}$ et de $F_{\hat{\theta}}$. Ils ont également montré par des simulations que $F_{\hat{\theta}}$ peut fonctionner beaucoup plus efficacement que l'ENPMV. Comme un inconvénient, l'estimateur semi-paramétrique nécessite une spécification préliminaire d'une famille paramétrique $\{T_\theta\}_{\theta \in \Theta}$, qui peut éventuellement introduire un terme du biais lorsqu'elle est éloignée de la réalité (voir Moreira et de Uña-Álvarez,* 2010*) pour plus de détails sur le sujet.*

3.3 Estimation non paramétrique de la densité

Soit Y^* une variable d'intérêt de densité de probabilité inconnue f par rapport à la mesure de Lebesgue sur R. La fonction de distribution correspondante est

$$F(x) = \int_{-\infty}^{x} f(t)\, dt,$$

Pour h suffisamment petit et en remplaçant F par l'estimation F_n nous définissons un estimateur non paramétrique (à noyau) de f :

$$f_h(y) = \int K_h(x - t)\, F_n(dt) = \alpha_n \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)\, G_n(X_i)^{-1} \tag{3.2}$$

où $K_h(t) = K(t/h)/h$, et $K(.)$ est la fonction dite noyau et $h = h_n$ est la fenêtre (bandwidth en anglais), avec $h_n \to 0$.

D'autre part, en utilisant l'estimateur semi-paramétrique $F_{\hat{\theta}}$ de F nous définissons l'estimateur semi-paramétrique de la densité f :

$$f_{\hat{\theta},h}(y) = \int K_h(y-t)\, F_{\hat{\theta}}(dt) = \alpha_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^{n} K_h(y - Y_i)\, G_{\hat{\theta}}(Y_i)^{-1} \tag{3.3}$$

Remarque 3.3.1 *1) Notons que les deux estimateurs (3.2) et (3.3) corrigent la double troncature par une sous-pondération des Y_i en fonction d'une estimation de la probabilité d'échantillonnage $G(Y_i)$.*

2) G_n et G_b sont deux estimateurs convergent de G. Pour $G_{\hat{\theta}}$, cela découle de la convergence de $\hat{\theta}$, à condition que G_θ soit une fonction lisse de θ (Moreira et de Uña-Álvarez, 2010).

3) Selon Shen (2010), puisque G_n et $G_{\hat{\theta}}$ approchent G à un taux $\sqrt{n}$, ce qui est plus rapide que la vitesse non paramétrique $\sqrt{nh}$, les propriétés asymptotiques de f_h et $f_{\hat{\theta},h}$ seront les mêmes, et coïncident avec ceux de l'estimateur basé sur le vrai G.

Considérons la version asymptotiquement équivalente de f_h et $f_{\hat{\theta},h}$à travers :

$$\bar{f}_{\hat{\theta},h}(y) = \int K_h(y-t)\, \bar{F}_n(dt) = \alpha \frac{1}{n} \sum_{i=1}^{n} K_h(y - Y_i)\, G(Y_i)^{-1} \tag{3.4}$$

où

$$\bar{F}_n(y) = \alpha \frac{1}{n} \sum_{i=1}^{n} G(Y_i)^{-1}\, \mathbf{1}_{[Y_i \leq x]}$$

Dans le résultat suivant, nous donnons la convergence forte et la normalité asymptotique de $\bar{f}_h(y)$. Nous supposons implicitement que $G(y) > 0$, tout au long de cette section.

Théoréme 3.3.1 (Moreira et de Uña-Álvarez, 2012.) *(i) Si K est borné sur un support compact, h est tel que*

$$\sum_{n=1}^{\infty} exp^{-\eta hn} < \infty, \quad pour\ chaque\ \ \eta > 0,$$

G est continue en y. Alors,

$$\bar{f}_h(y) \to f(y) \quad avec\ \ probabilités\ 1.$$

(ii) Si, en plus de la conditions (i), K est une fonction paire, $h = o(n^{-1/5})$, $G^{-1}f$ a une dérivée seconde et bornée au voisinage de y et $f(y) > 0$. Alors,

$$(nh)^{1/2}\left(\bar{f}_h(y) - f(y) \overset{Loi}{\to} N(0, \alpha G(y)^{-1} f(y) R(K)\right),$$

avec $R(K) = \int K(t)^2 dt$.

La moyenne asymptotique et la variance de (3.4) sont données dans le résultat suivant. Nous nous référons aux hypothèses de régularité standard suivantes :

(A1) La fonction noyau K est une fonction de densité avec

$$\int tK(t)dt = 0, \quad \mu_2(K) = t^2 K(t)dt < \infty \quad \text{et} \quad R(K) = \int K(t)^2 dt < \infty.$$

(A2) La suite (fenêtre) $h = h_n$ satisfait $h \to 0$ et $nh \to \infty$ quand $n \to \infty$.

(A3) Les fonctions f et $G^{-1}f$ sont deux fois continuement différentiables autour de y.

Théoréme 3.3.2 (Moreira et de Uña-Álvarez, 2012.) *Sous (A1)-(A3) nous*

avons, quand $n \to \infty$,

$$E\left[\bar{f}_h(y)\right] = f(y) + \frac{1}{2}h^2 f''(y)\mu_2(k) + o(h^2),$$
$$Var\left[\bar{f}_h(y)\right] = (nh)^{-1}\alpha G(y)^{-1} f(y) R(k) + o((nh)^{-1})$$

Preuve. La preuve suit les étapes standard. Un développement de Taylor d'ordre deux de f au voisinage de y est utilisé, et les hypothèses sur le noyau et la fenêtre h permet de conclure. Voir par exemple Wand et Jones (*1985*), pour les détails. ■

Habituellement, on sera intéressé par l'erreur globale de $\bar{f}_h$ en tant qu'estimateur de la courbe entière f. Cela peut être mesuré par la MSE intégrée, à savoir :

$$MISE(\bar{f}_h) = \int MSE(\bar{f}_h(y))dy = \int [E\bar{f}_h(y) - f(y)]^2 dy + \int Var\left(\bar{f}_h(y)\right) dy.$$

Sous régularité, nous avons à partir des résultats précédents les expression asymptotiques suivantes pour la $MISE(\bar{f}_h)$:

$$MISE(\bar{f}_h) = \frac{1}{4}h^4 R(f'')\mu_2(K)^2 + (nh)^{-1}\alpha R(K)\int G^{-1} f$$

où $R(f") = \int f"(t)^2 dt$.

Remarque 3.3.2 *En raison de l'équivalence entre* $G_n, G_{\hat{\theta}}$ *et* G*, la même expression asymptotique tiendra pour* f_h *et* $f_{\hat{\theta},h}$ *dans des conditions appropriées.*

La minimisation de la $AMISE(\bar{f}_h)$ par rapport à h conduit à la fenêtre asymptotiquement optimale :

$$h_{AMISE} = Arg\min_h AMISE(\bar{f}_h) = \left[\frac{\alpha R(K)\int G^{-1} f}{R(f'')\mu_2(K)^2}\right]^{1/5} n^{-1/5}. \qquad (3.5)$$

Remarque 3.3.3 *Bien entendu, cette dernière expression (3.5) dépend des quantités inconnues qui doivent être estimées en pratique. Il existe plusieurs critères pour sélectionner la fenêtre optimale. Pratiquement, la méthode la plus utilisée est la validation croisée.*

3.4 Étude de simulation

Dans cette section[1], nous illustrons le comportement asymptotique de l'estimateur à noyau (non paramétrique) noté f_h défini par (3.3) de la densité. Nous examinons la normalité asymptotique dans différentes situations.

Considérons deux modèles simulés (de cas de la densité à estimer f) :

Modèle 1 (cas de décroissance exponentielle) : La variable Y est distribuée comme une normale $\mathcal{N}(\mu = 2, \sigma^2 = 1)$, de densité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \qquad x \in R.$$

Modèle 2(cas de queue lourde) : La variable Y est distribuée sous la forme d'un Weibull à trois paramètres $(\lambda = 1, \beta = 4, \gamma = 1)$ avec densité :

$$f(x) = \left(\frac{\beta}{\lambda}\right)\left(\frac{x-\gamma}{\lambda}\right)^{\beta-1} \exp\left\{-\left(\frac{x-\gamma}{\lambda}\right)^{\beta}\right\} 1_{(x\geq 0)}$$

où $\gamma \in \mathbb{R}$ est le paramètre de localisation (ou durée de vie sans défaillance), $\beta > 0$ est le paramètre de forme et $\lambda > 0$ est le paramètre d'échelle ou durée de vie caractéristique de la distribution.

[1]Les résultats de cette section ont été obtenus avec le package DATA du logiciel R (voir, Moreira et al., *2012*, disponible sur le site : *http ://CRAN.R-project.org/package=DTDA*.

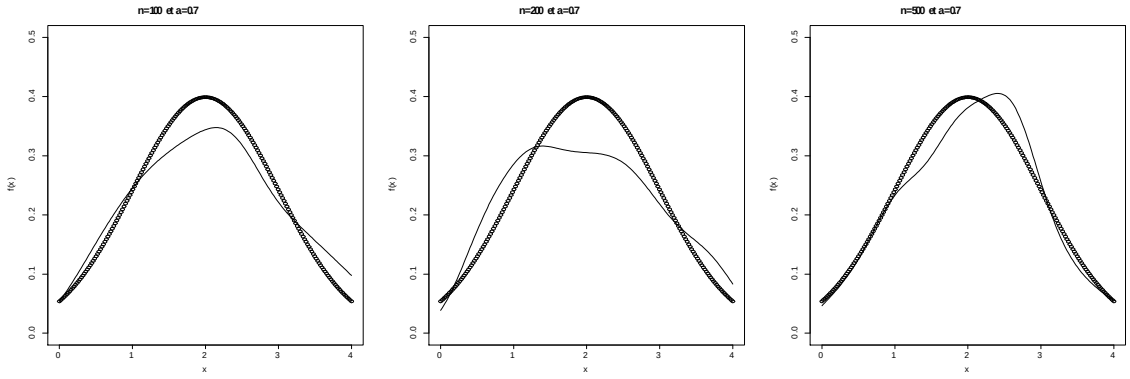

FIG. 3.1 – Estimation de la densité sous double troncature : **Modèle(1)**, $\alpha = 0.7$.

Pour la double troncature, supposons que U^* et V^* sont indépendantes l'une de l'autre. On prend $U^* \sim Exp(b)$ et $V^* \sim Exp(c)$, où b et c sont choisis pour obtenir les pourcentages de troncature suivants : 30% et 10% correspondent à $\alpha = 0,7$ et $\alpha = 0,9$ resp. La troncature se produit lorsque $U^* \leq Y^* \leq V^*$ est violé.

En utilisant ce schéma, $B = 500$ échantillons indépendants de taille $n = 100$, $n = 200$ et $n = 500$ ont été générés pour chaque modèle. Cela signifie que, pour chaque replication, le nombre de données simulées N est beaucoup plus grand que n, en fait $\alpha \approx (\frac{n}{N})$, où α représente la proportion de non troncature.

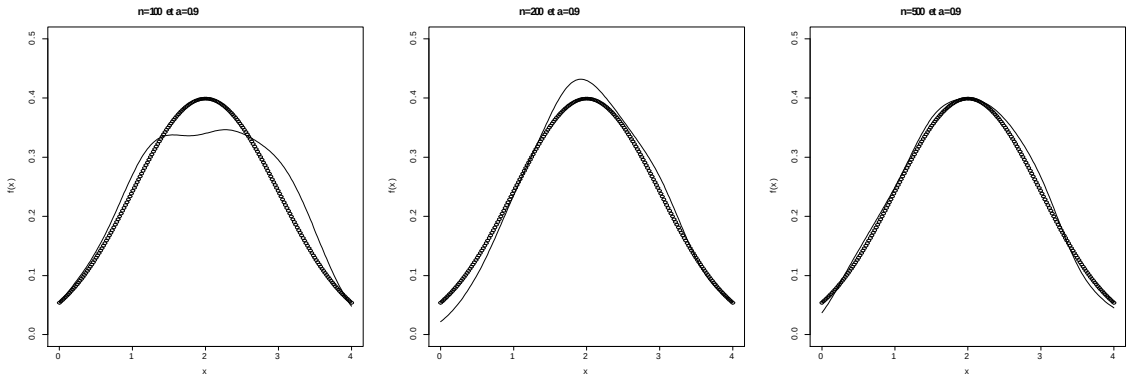

FIG. 3.2 – Estimation de la densité sous double troncature : **Modèle(1)**, $\alpha = 0.9$.

Pour chaque échantillon, en utilisant l'estimation du plug-in, nous avons estimé

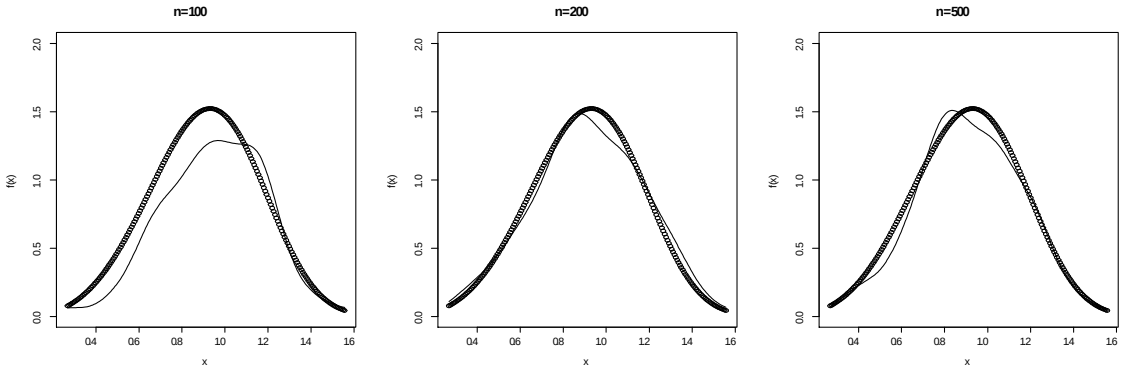

FIG. 3.3 – Estimation de la densité sous double troncature : **Modèle(2)**, $\alpha = 0.7$.

la densité. Les résultats sont affichés en figures 3.1, 3.2, 3.3 et 3.4. Rappelons que dans l'estimation non paramétrique, l'optimalité (au sens MSE) n'est pas sérieusement influencée par le choix du noyau K mais est affectée par le choix de h. Dans cette étude, h est choisie pour satisfaire les hypothèses ci-dessus, et le noyau K est gaussien. La fenêtre h que nous avons utilisée pour estimer la densité est celle utilisée par Moreira et de Uña-Álvarez (*2012*), elle est basée sur la minimisation de l'$AMISE(f)$, ce qui conduit à la fenêtre asymptotiquement optimale qui a été donnée par l'expression (3.5).

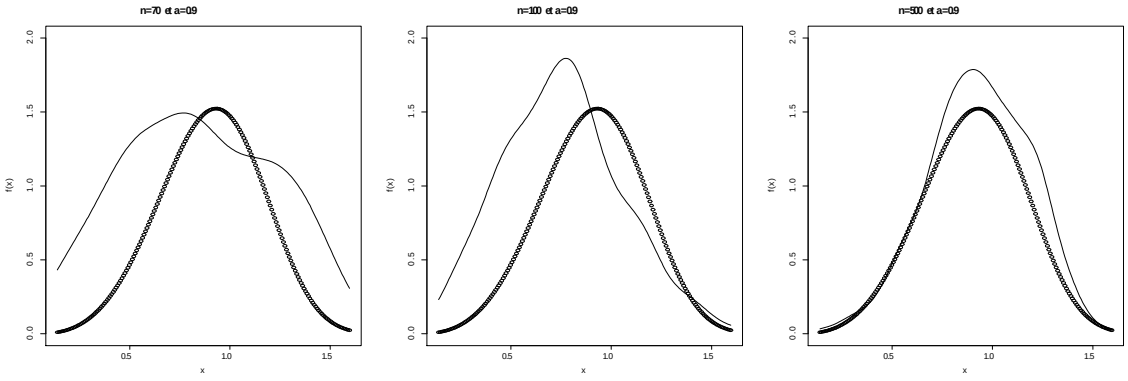

FIG. 3.4 – Estimation de la densité sous double troncature : **Modèle(2)**, $\alpha = 0.9$.

Conclusion

Les données doublement tronquées est un cas particulier de données incomplètes, dont la littérature reste encore d'actualité. Le traitement de ces données occupe une place importante dans les nouveaux recherches (ces dernières années). Plusieurs exemples sont cités dans ce mémoire (Chapitre 2) :

- *Données du cancer infantile dans le nord du Portugal (Moreira et de Uña Álvarez, 2010).*
- *Données d'équipements de Ye et Tang (2016).*
- *Données de l'entreprise allemande étudier par Dörre (2020).*

Ce mémoire est donc un aperçu sur ce type de données. En particulier, nous rappelons (chapitre 3) les estimateurs non paramétriques de la densité et de la distribution et le cadre statistique des données et leur modèle sous troncature double, leur comportement asymptotique et des de simulations pour confirmer et évaluer la performance des différents estimateurs étudiés.

Comme future travail, nous envisagions l'estimation d'autres paramètres fonctionnels tels que la fonction de régression, les quantiles, les lois conditionnelles,... et des applications sur des données réelles.

Bibliographie

[1] Chaib, Y. (2013). Estimation de la fonction mode pour des données tronquées et censurées. *Thèse de Doctorat, Université de Annaba. Algérie.*

[2] Dörre, A. (2020). Bayesian estimation of a lifetime distribution under double truncation caused by time-restricted data collection. *Statistical Papers*, 61(3), 945–965.

[3] Dörre, A., Emura, T. (2019). *Analysis of doubly truncated data : an introduction*. Singapore : Springer Singapore.

[4] Efron, B., Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association,* 94(447), 824–834.

[5] Emura, T., Wang, W. (2010). Testing quasi-independence for truncation data. *Journal of Multivariate Analysis,* 101(1), 223–239.

[6] Huber-Carol, C. (1994). Durée de survie tronquées et censurées. *Journal de la société française de statistique,*135(4),3–23

[7] Hughes, E. J. (1962). Maximum likelihood estimation of distribution parameters from incomplete data. *Lowa State University.*

[8] Klein, J. P., Moeschberger, M. L. (2003). *Survival analysis : techniques for censored and truncated data*. New York : Springer.

[9] Lemdani, M., Ould-Saïd, E. (2007). Asymptotic behavior of the hazard rate kernel estimator under truncated and censored data. *Communications in Statistics-Theory and Methods*, 36(1), 155–173.

[10] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1), 95–118.

[11] Martin, E. C., Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *Journal of the American Statistical Association*, 100(470), 484–492.

[12] Moreira, C., de Uña-Álvarez, J. (2010). A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine*, 29(30), 3147–3159.

[13] Moreira, C., de Una-Alvarez, J. (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics*, 6, 501–521.

[14] *Moreira, C., de Uña-Álvarez, J. Crujeiras Casais, R. M. (2010).* DTDA : an R package to analyze randomly truncated data. *J. Stat. Softw.* 37(7) 1–20.

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

[16] Shen, P. S. (2010). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, 62(5), 835–853.

[17] Shen, P. S. (2011). Testing quasi-independence for doubly truncated data. *Journal of Nonparametric Statistics*, 23(3), 753–761.

[18] Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics*, 146–156.

[19] Touraine, C. (2013). Modèles illness-death pour données censurées par intervalle : application à l'étude de la démence. *Thèse de Doctorat, U. Bordeaux 2, France.*

[20] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1), 163–177.

[21] Ye, Z. S., Tang, L. C. (2016). Augmenting the unreturned for field data with information on returned failures only. *Technometrics*, 58(4), 513–523

[22] Zhou, Y. (1996). A note on the TJW product-limit estimator for truncated and censored data. *Statistics & Probability Letters*, 26(4), 381–387.

[23] Wand, M. P., Jones, M. C. (1994). *Kernel smoothing*. CRC press.

ملـــخـــــص

في هذه المذكرة، نقدم لمحة عامة على البيانات المقتطعة بشكل مزدوج. نؤكد على تقدير دالتي الكثافة والتوزيع، مع تحديد الإطار الإحصائي للبيانات ونموذجها تحت الاقتطاع المزدوج. استنادًا إلى أمثلة بيانات تمت محاكاتها باستخدام البرنامج R، تم تأكيد السلوك الجيد وتقييم أداء مختلف المقدرات المدروسة.

Résumé

*Dans ce mémoire, nous donnons un aperçu sur les données doublement tronquées. Nous mettons l'accent sur l'estimation de la densité et de la distribution, tout en précisant le cadre statistique des données et leur modèle sous troncature double. Des travaux de simulation à l'aide du logiciel de traitement statistique **R** sont réalisés pour confirmer le bon comportement et pour évaluer la performance des différents estimateurs étudiés.*

Abstract

*In this memory, we give an overview on doubly truncated data. We emphasize the estimation of density and distribution, while specifying the statistical framework of the data and their model under doubly truncation. Simulation study using the statistical software **R** is carried out to confirm the good behavior and to evaluate the performance of the different estimators studied.*