République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A. MIRA-BEJAIA

Faculté des Sciences Exactes
Département de Mathématiques

Mémoire de fin de cycle

En vue de l'obtention d'un Master en mathématiques
Option
Probabilités Statistique et Applications

Thème

Estimation non paramétrique de la densité de probabilité par la méthode du noyau : cas des données complètes et incomplètes

Présenté par :
Kessaci Sabrina

Soutenu le 12/07/2022 **devant le jury composé de :**

***Mr*. SACI walid**	**Président**	M.C.B	**Université de Béjaïa**
M^{me}. LAGHA Karima	**Examinatrice**	Professeur	**Université de Béjaïa**
M^{me}. TIMERIDJINE Karima	**Encadreur**	Professeur	**Université de Béjaïa**

Année universitaire : 2021/2022

Remerciements

Tout d'abord, je tiens à remercier Dieu le tout puissant de m'avoir donné, la volonté, le courage et la santé afin d'accomplir ce travail.

Je remercie particulièrement ma promotrice $M^{\mathbf{me}}$ **K. TIMERIDJINE** pour sa disponibilité, son soutien et ses remarques précieuses qui m'ont aidé à bien présenter ce travail.

J'adresse aussi mes remerciements aux membres de jury $M^{\mathbf{me}}$ **K. LAGHA** et **Mr. W. SACI** pour avoir accepté d'évaluer mon travail.

Je tiens à remercier toute ma famille, mes amis (es) et mes condisciples de la promotion M2 PSA(2021/2022).

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

Dédicace

Je dedie ce trovail

Á la mémoire de mon père

Á la mémoire de mon très cher grand père

Á ma très chère mère que j'adore.

Á ma très chère grand mère que j'adore.

Á mon très cher frère Omar.

Á mon cher mari Khelifa qui m'a tellement aidé et encouragé.

Á mes très chers enfant : Achraf et Aymen.

Á Toute ma famille.

Á tous mes amis(es).

Sabrina 

Table des matières

Table des figures

Liste des tableaux

Notations

P	Mesure de probabilité.
$\mathbb{R}$	Ensemble de tous les nombres réels.
$v.a$	Variable aléatoire.
$i.i.d$	Indépendant et identiquement distribué.
$f.d.r$	Fonction de répartition.
$\xrightarrow{p.s}$	Convergence presque sûre.
$\xrightarrow{\mathcal{L}}$	Convergence en loi.
$\xrightarrow{P}$	Convergence en probabilité.
$X \wedge C$	Minimum (X, C).
$X \vee C$	Maximum (X, C).
L^2	L'espace des fonctions de carré intégrable.
MSE	Erreur quadratique moyenne.
$MISE$	Erreur quadratique moyenne intégrée.
$AMISE$	Erreur quadratique moyenne intégrée asymptotique..
EKM	Estimateur de Kaplain et Meier
$O(.)$	petit o.
X	Une variable aléatoire
$\mathbb{E}[X]$	Espérance de la variable aléatoire X
$Var[X]$	Variance de la variable aléatoire X
F	Fonction de répartition associée à la densité f
f	Densité de probabilité de X
$\hat{f}_n$	Estimateur de la densité f
h_{opt}	paramètre de lissage (h) optimale
$\mathcal{N}(0,1)$	loi normale standard (centrée réduite)
$\mathcal{N}(\mu, \sigma)$	loi normale (ou de Gauss) à deux paramètre $\mu \in \mathbb{R}$ et $\sigma^2 > 0$
K	noyau
$\mathbf{1}_A$	Fonction indicatrice de l'ensemble A

Introduction

La théorie de l'estimation est l'une des préoccupations majeures des statisticiens. Cette théorie est habituellement divisée en deux composantes principales, à savoir, l'estimation paramétrique et l'estimation non paramétrique. L'estimation paramétrique consiste à estimer un nombre fini de paramètres réels à partir d'un échantillon. En opposition l'estimation non paramétrique estime à partir des observations une fonction inconnue, élément d'une certaine classe fonctionnelle, telle que la fonction densité. Dans ce travail on s'intéresse à l'estimation non paramétrique de la fonction densité de probabilité.

L'estimation de la densité de probabilité est l'un des plus vieux problèmes de l'estimation non paramétrique. En effet, les premiers travaux consacrés à ce sujet remontent à ceux de Karl Pearson [46] en 1902. Ce problème fondamental a connu un développement considérable, différentes méthodes ont été dédiées à l'estimation non paramétrique de la densité de probabilité inconnue f, on peut citer : la méthode de l'histogramme, méthode d'estimation par les série orthogonales et la méthode du noyau. Cette dernière méthode qui fait l'objet de notre travail sera présentée. Rosenblatt [48] en 1956, suivi de Parzen [44] en 1962, sont les premiers à proposer une classe d'estimateurs à noyau d'une densité univariée.
Cet estimateur est une fonction de deux paramètres une fonction K, appelé noyau, et un réel positif h dit paramètre de lissage (appellé aussi largeur de la fenêtre). Rosenblatt reprenait l'idée de Fix et Hodges [20] en 1951, qui consistait à estimer la densité en un point, en comptant le nombre d'observations situées dans l'intervalle de longueur $2h$ et centré en ce point. L'estimateur à noyau d'une densité f est de la forme :

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

généralement les noyaux les plus utilisés sont des noyaux symétriques. Dans le cas où l'hypothèse de symétrie du noyau n'est pas remplie on trouve Chen S, X [12] en 1999 qui a introduit l'estimateur à noyau bêta de la densité et Bouezmarni T. et Rolin J-M. [7] en 2003 ont étudié la consistance de cet estimateur. Cependant, en raison de contrainte de temps nous nous contentons dans ce travail aux noyaux symétriques.

L'estimateur à noyau a connu un très grand succès parmi les estimateurs non paramétriques, ceci est dû à sa simplicité et ses propriétés de convergence vers la densité inconnue f, il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K, mais aussi dans le choix du paramètre de lissage h.

Le choix du paramètre de lissage h est crucial. Plusieurs auteurs ont montré que l'estimateur peut changer dramatiquement pour de petites variations du paramètre de lissage.

Dans le contexte des données complètes, plusieurs travaux ont été réalisés pour l'estimation de la fonction densité par la méthode du noyau. Sous certaines conditions sur le noyau K, Parzen [44], Silverman [51] et Nadaraya [41] ont établis les propriétés de convergence de cet estimateur. Devroye [15] en 1985 a fait une étude complète sur la convergence L^1. Les théorèmes relatifs à l'erreur quadratique asymptotique et l'erreur quadratique intégrée asymptotique ont été obtenus sous forme élémentaire par Parzen [44]. Enfin, c'est Epanechnikov [19] en 1969 qui s'est rendu compte de l'existence d'un noyau asymptotiquement optimal K_e .

Mais il n'est pas rare que les données à traiter ne soient pas complètes. Cette difficulté est souvent rencontrée dans le domaine de l'analyse de survie, où la variable d'intérêt, appelée durée de vie, représente le temps écoulé jusqu'à l'apparition d'un évènement précis. Par exemple en médecine, cette durée peut être le temps jusqu'à la guérison, la rechute, ou encore le décès, d'un patient. Cette variable de durée de vie est en général, et pour diverses raisons, non complètement observée, essentiellement à cause de deux phénomènes distincts : la censure et la troncature, on parle alors de données incomplètes. Il est bien connu que ce type de données nécessite des techniques statistiques plus élaborées afin de les modéliser. Il n'est pas rare dans la réalité, que ces deux types de données incomplètes se croisent simultanément dans un même échantillon (Tronqué à gauche et Censuré à droite).

Dans ce contexte, de nombreux travaux ont été effectués. Par exemple en (1958), Kaplan et Meier [29] proposent d'utiliser dans le domaine médical un estimateur non paramétrique permettant d'intégrer les données censurées en estimant la fonction de survie en présence de censure à droite. Blum et Susarla en 1980 ont introduit l'estimateur à noyau de f, les propriétés asymptotiques ont été étudiées par Diehl.S et Stute.W [16] en 1988. Dans le cadre de la troncature gauche Lynden-Bell [39] en 1971 introduit les estimateurs produits-limites des fonctions de répartition (f.d.r.) de la variable d'intérêt X et de la troncature T. Woodroofe [56] en 1985 établit les conditions d'identifiabilité du modèle ainsi que la convergence presque sûre des estimateurs de Lynden-Bell.

Avoir des données incomplètes constitue une perte d'information, un problème d'identifiabilité se pose alors. Autrement dit, nous nous demandons si la loi des observations nous permet d'identifier la loi de la variable d'intérêt X (exemple : la densité ou la survie) ?

Tout au long de ce mémoire nous considérons la suite $(X_1, X_2, ..., X_n)$ de n variables aléatoires indépendantes et identiquement distribuées (i.i.d) de fonction de répartition F et admettant une densité inconnue f. L'objectif de ce travail est d'estimer d'une part à partir de données complètes et d'autre part à partir de données incomplètes la densité inconnue f en utilisant la méthode du noyau.

Pour répondre à cet objectif, nous avons organisé notre travail comme suit :

Le premier chapitre est consacré à l'estimation de la densité de probabilité par la méthode du noyau dans le cas des données complètes. Nous présentant la méthode du noyau, l'estimateur à noyau et ses différentes propriétés, les noyaux usuels et quelques méthodes de sélection du paramètre de lissage h.

Le deuxième chapitre porte des généralités sur les données incomplètes, les différents types de censure et troncature.

Le troisième chapitre est consacré à l'estimation de la densité de probabilité par la méthode du noyau dans le cas des données incomplètes. En particulier, l'estimateur de Kaplan-Meyer de la fonction de survie, et les estimateurs à noyau de la fonction de densité dans le modèle de censure droite, et troncature gauche.

Nous présentons dans **le quatrième chapitre** une application et une étude comparative réalisée sur les résultats obtenus.

Nous terminerons ce mémoire par une conclusion générale.

Chapitre 1

Estimation non paramétrique de la fonction densité par la méthode du noyau : cas des données complètes

1.1 Introduction

L'estimation non paramétrique de la densité de probabilité a fait l'objet de multiples travaux par des méthodes diverses citons :

— L'estimateur par histogramme

— L'estimateur par les séries orthogonales.

— Les méthodes à base de splines.

— L'estimateur par la méthode du noyau.

L'estimateur le plus ancien est l'histogramme des fréquences. L'origine des histogrammes est attribué à John Graunt au XVII ème siècle répondant à l'objectif d'une représentation de la distribution de données.

Cette méthode consiste à estimer une fonction densité de probabilité en un point x tel que x appartient à un intervalle de longueur h dit paramètre de lissage. En se basant sur un n-échantillon $X_1, X_2 \ldots, X_n$ de la variable X dont la densité de probabilité est f inconnue. Elle est basée sur le choix d'un point d'origine a_0 et d'une partition $B_k = ([a_k, a_{k+1}[)_{k=1,\ldots,p}$ en p intervalles du support de X. Si nous notons n_k le nombre de variables dans la classe $[a_k, a_{k+1}[$ et $h = a_{k+1} - a_k$, on trace alors une boîte en x, dont la largeur est gouvernée par le paramètre h.

L'estimateur de f sur $[a_k, a_{k+1}[$ du type histogramme est :

$$\hat{f}_n(x) = \frac{1}{n} \frac{\text{Card } \{x_i \text{ dans le même intervalle que } x\}}{\text{largeur de l'intervalle contenant } x} \tag{1.1}$$

Si on choisit des classes de même longueur h,

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{nh}\,\text{Card}\,\{i : x_i, \quad x \in [a_k, a_{k+1}[\} \\ &= \frac{n_k}{nh} \qquad \text{pour } x \in [a_k, a_{k+1}[.\end{aligned} \tag{1.2}$$

Les propriétés asymptotiques de cet estimateur ont été détaillées dans le livre de Bosq et Lecoutre [5] en 1987 et le livre de Simonoff [52] en 1996. L'étude asymptotique de l'erreur quadratique moyenne et de l'erreur quadratique moyenne intégrée de $\hat{f}_n$ ont été établies par Lecoutre [35] en 1982. En 1974 Geoffrey [23] a étudié la convergence uniforme et presque complète de cet estimateur.
L'histogramme a de bonnes propriétés statistiques. Néanmoins, ses discontinuités n'apparaissent pas très naturelles et ce qui est plus grave, les points tombant près du bord d'une classe et ceux tombant près du milieu ne sont pas différenciés, ceci explique la variabilité des interprétations statistiques que l'on peut faire d'un histogramme suivant le choix de l'origine et des classes. Pour des densités raisonnablement lisses, l'histogramme apparait donc comme un estimateur sévèrement limité. Afin de résoudre ce problème, une méthode plus robuste a été introduite, c'est la méthode d'estimation par noyau, et elle est très utilisée en estimation non paramétrique.

Dans ce chapitre, nous allons présenter une étude détaillée de l'estimateur par la méthode du noyau ainsi que ses propriétés statistiques et asymptotiques.

1.2 Définitions et critères d'erreur

Avant de présenter la méthode du noyau et définir l'estimateur à noyau de $f(\cdot)$ ainsi que ses propriétés, il est intéressant de citer les différents critères d'erreur et donner quelques définitions et notions sur les noyaux.

Pour mesurer les performances théoriques des estimateurs et identifier le meilleur, il est nécessaire de spécifier un critère d'erreur. Nous considérons la densité de probabilité f inconnue et son estimateur $\hat{f}_n$.

Définition 1.2.1. *L'erreur quadratique intégrée ISE (Integrated Square Error) est définit par :*

$$\text{ISE}(f, \hat{f}_n) = \int |f(x) - \hat{f}_n(x)|^2 dx \tag{1.3}$$

Définition 1.2.2. *L'erreur quadratique moyenne MSE (Mean Square Error) :*

$$\begin{aligned}\text{MSE}(f(x), \hat{f}_n(x)) &= \mathbb{E}(f(x) - \hat{f}_n(x))^2 \\ &= \Big(\textit{biais } (\hat{f}_n(x)) \Big)^2 + Var(\hat{f}_n(x))\end{aligned} \tag{1.4}$$

Définition 1.2.3. *L'erreur quadratique moyenne intégrée MISE (Mean Integrated Square Error) :*

$$\begin{aligned} \mathrm{MISE}(f, \hat{f}_n) &= \int \mathrm{MSE}(f(x), \hat{f}_n(x))dx = \int \mathbb{E}(f(x) - \hat{f}_n(x))^2 dx \\ &= \int [\Big(\mathrm{Biais}(\hat{f}_n(x))\Big)^2 + Var(\hat{f}_n(x))]dx. \end{aligned} \tag{1.5}$$

Définition 1.2.4. *[13] On dit qu'un estimateur $\hat{f}_n$ de f est sans biais si :*

$$\mathbb{E}(\hat{f}_n(x)) = f(x)$$

Il est dit asymptotiquement sans biais si : $\lim_{n\to\infty} \mathbb{E}(\hat{f}_n(x)) = f(x)$, en tout point x de continuité de la densité f. Un estimateur $\hat{f}_n$ de f est dit asymptotiquement uniformément sans Biais si :

$$\lim_{n\to\infty} \sup_x |\mathbb{E}[\hat{f}_n(x) - f(x)]| = 0.$$

Définition 1.2.5. *On dit qu'un estimateur $\hat{f}_n$ de f est ponctuellement consistant en moyenne quadratique si : $\lim_{n\to\infty} \mathrm{MSE}(f(x), \hat{f}_n(x)) = 0$, en tout point x de continuité de la densité f.*

Définition 1.2.6. *On dit qu'un estimateur $\hat{f}_n$ de f est uniformément consistant en moyenne quadratique intégrée si :*

$$\lim_{n\to\infty} \sup_x \mathrm{MSE}(f(x), \hat{f}_n(x)) = 0.$$

Définition 1.2.7. *On dit qu'un estimateur $\hat{f}_n$ de f est asymptotiquement normal si :*

$$\hat{f}_n(x) \xrightarrow[n\to+\infty]{\text{en loi}} \mathcal{N}(\mathbb{E}(\hat{f}_n(x)), Var(\hat{f}_n(x))), \quad \forall x.$$

Notion de noyau

Un noyau est une fonction de pondération utilisée dans les techniques d'estimation non paramétrique. Le noyau intervient dans l'estimation de la densité de probabilité d'une variable aléatoire, il se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support.

Définition 1.2.8. *Un noyau est une fonction, $K : \mathbb{R} \to \mathbb{R}^+$, permettant d'estimer une densité de probabilité à partir d'un n-échantillon.*

Définition 1.2.9. *Un noyau est dit symétrique si, pour tout u dans son ensemble de définition, $K(u) = K(-u)$.*

Définition 1.2.10. *[38] Un noyau sommatif est une fonction $K : S \to \mathbb{R}^+$, qui vérifie la propriété de sommativité $\int_S K(u)du = 1$.*

Définition 1.2.11. *[37]*

Soit $r \geq 1$ un entier. On dit qu'un noyau K est d'ordre r si :

$$\forall j = 1, \ldots, r, \quad \int u^j K(u)du = 0 \quad et \quad \int u^{r+1} K(u)du \neq 0.$$

Dans la suite de ce document, on ne considèrera que des noyaux sur $\mathbb{R}$.

1.3 Méthode du noyau

En statistique, la méthode du noyau proposée par Rosenblatt [48] en 1956 et améliorée par Parzen [44] en 1962 est une méthode non paramétrique d'estimation d'une densité de probabilité inconnue f, en se basant sur un n-échantillon d'une variable aléatoire X. En ce sens, cette méthode généralise astucieusement la méthode d'estimation par histogramme, elle est aussi appelée méthode de Parzen-Rosenblatt.

L'idée de l'estimation par la méthode du noyau consiste à évaluer la densité $f(x)$ au point x en comptant le nombre d'observations figurant dans un certain voisinage de x sur $\mathbb{R}$. Cette estimation, présente de bonnes propriétés statistiques car elle permet de retrouver la continuité, contrairement à l'histogramme qui est une fonction étagée et discontinue : pour cela on remplace la loi uniforme de x par une fonction de forme générale $K(x)$ continue (exemple, une gaussienne centrée réduite en x).

L'estimateur ainsi obtenu est appelé estimateur à noyau dit aussi estimateur de Parzen-Rosenblatt, il est une fonction de deux paramètres : le noyau K et le paramètre de lissage h largeur de la fenêtre. Le succès rencontré par cet estimateur s'explique par sa simplicité, sa flexibilité et aussi ses propriétés de convergence, il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K, mais aussi dans le choix cruciale du paramètre de lissage h.

1.3.1 Estimateur de Parzen-Rosenblatt : Construction et définition

Avant la construction de l'estimateur à noyau de f, nous donnons brièvement un aperçu sur l'estimateur empirique de la fonction de répartition.

Définition 1.3.1. *Soit $X_1, \ldots, X_n$ un n-échantillon d'une variable aléatoire X de densité $f(\cdot)$ de probabilité sur $\mathbb{R}$, de fonction de répartition F tel que, $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associé à $X_1, \ldots, X_n$, la fonction aléatoire $F_n : \mathbb{R} \longrightarrow [0,1]$ définie pour tout $x \in \mathbb{R}$ par $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$.*

On peut également écrire de manière équivalente

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\,(X_i)_{]-\infty,x]} \tag{1.6}$$

où $1(.)_{]-\infty,x]}$ *est la fonction indicatrice sur* $]-\infty, x]$.

La loi forte des grands nombres [32] montre que c'est un estimateur fortement consistant c'est à dire :

$$\forall x \in \mathbb{R}, \quad \hat{F}_n(x) \to F(x) \qquad \text{P.S}$$

Le théorème de Glivenko- Cantelli [45] permet d'améliorer ce résultat puisqu'il donne la convergence uniforme.

Théorème 1.3.1. *[45]*
$\forall x \in \mathbb{R}, \quad \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{P.S} 0,$ *quand* $n \to \infty$

Donc $\hat{F}_n$ est un très bon estimateur de F

$$n\hat{F}_n(x) = \sum_{i=1}^{n} 1_{\{x_i<x\}} \xrightarrow{loi} \mathcal{B}(n, F(x)).$$

où $\mathcal{B}$ est la loi binomiale, dont l'espérance et la variance de $\hat{F}_n(x)$ sont donnés respectivement par :

$$\mathbb{E}\left(\hat{F}_n(x)\right) = F(x) \quad \text{et} \quad Var\left(\hat{F}_n(x)\right) = \frac{1}{n}[1 - F(x)]F(x)$$

On peut alors justifier la construction de l'estimateur à noyau de deux façons :
- Une première ideé (developpée par Rosenblatt [48] (1956) en reprenant l'idée de Fix et Hodges [20] en 1951) était de le construire à partir de l'estimateur $\hat{F}_n$ de la fonction de répartition. Pour $h > 0$ assez petit, Rosenblatt (1956) a proposé d'estimer f par :

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}, \tag{1.7}$$

En remplaçant F par l'estimateur $\hat{F}_n$, on obtient

$$\begin{aligned} \hat{f}_n(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} \mathbf{1}_{(x-h<X_i\leq x+h)} \\ &= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} \mathbf{1}_{\left(-1<\frac{x-X_i}{h}\leq 1\right)} \end{aligned} \tag{1.8}$$

En posant :

$$\omega(u) = \begin{cases} 1/2, & -1 < u \leq 1 \\ 0, & \text{sinon.} \end{cases}$$

On peut réécrire (1.8) sous la forme

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}\omega\left(\frac{x - X_i}{h}\right) \tag{1.9}$$

Nous venons de définir l'estimateur à noyau dit de Rosenblatt(uniforme). Parzen [44] en 1962 a étudié une classe générale d'estimateurs. En remplaçant la fonction ω par une fonction densité de forme générale K, d'où l'estimateur à noyau de la densité f (estimateur de Parzen Rosenblatt) définit par :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - X_i}{h}\right) \tag{1.10}$$

Où

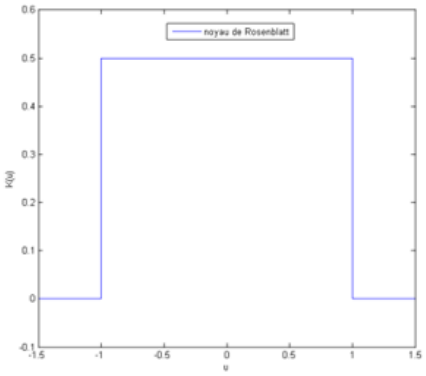
$h = h(n)$ est un reél positif qui est fonction de n, appelé paramètre de lissage vérifiant $nh \longrightarrow \infty, n \longrightarrow \infty$ et K est la fonction noyau satisfaisant certaines hypothèses basiques parmi celles énoncées ci-dessous :

$$\begin{array}{ll}
(i) & \int_S K(u)du = 1. \quad S \text{ est le support de } K \quad (K \text{ densité de probabilité}) \\
(ii) & \int_S u^2K(u)du = \sigma_K^2 < \infty. \\
(iii) & \int_S uK(u)du = 0. \\
(iiii) & K(u) = K(-u),
\end{array} \tag{1.11}$$

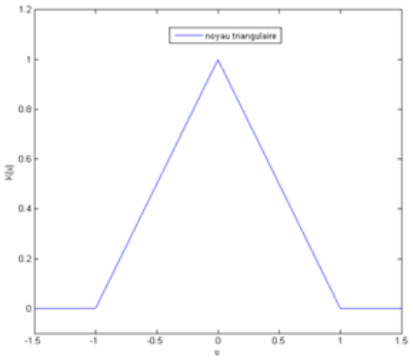
Les noyaux les plus utilisés sont des noyaux symétriques, certains sont donnés dans le tableau suivant :

Noyau	$K(u)$	Domaine de définition
Rectangulaire	$K(u) = 1/2$	$[-1, 1]$
triangulaire	$K(u) = 1 - \|u\|$	$[-1, 1]$
d'Epanechnikov	$K(u) = (3/4)\left(1 - u^2\right)$	$[-1, 1]$
Tukey ou Biweight	$K(u) = (15/16)\left(1 - u^2\right)^2$	$[-1, 1]$
Gaussien	$K(u) = (1/\sqrt{2\pi}) \exp\left(-u^2/2\right)$	$\mathbb{R}$

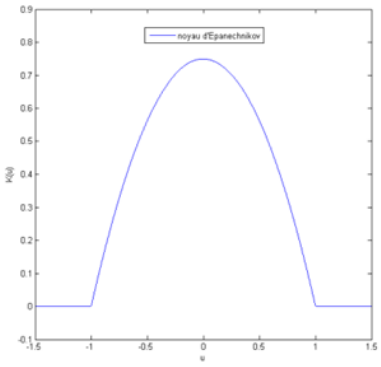
TABLE 1.1 – Exemple de noyaux symétriques



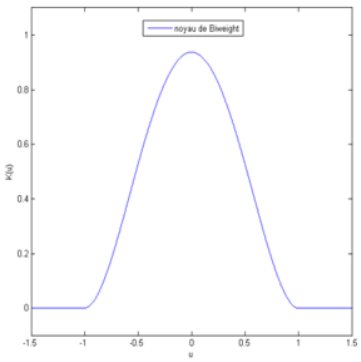
(a) Noyau Rectangulaire.



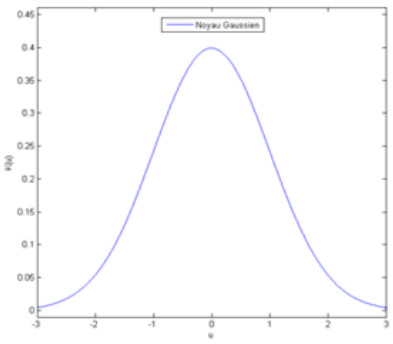
(b) Noyau triangulaire.



(c) Noyau de d'Epanechnikov.



(d) Noyau de Biweight.



(e) Noyau Gaussien.

1.4 Propriétés de l'estimateur de Parzen-Rosenbaltt

Il est facile de voir que l'estimateur à noyau (1.10) possède les propriétés suivantes :

1. Si K est une densité de probabilité, alors $\widehat{f_n}(\cdot)$ est aussi une densité de probabilité. En effet
$$\begin{aligned}\int_{-\infty}^{+\infty} \widehat{f_n}(x)dx &= \int_{-\infty}^{+\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h}\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(u)du = 1 \quad \text{avec } u = \frac{x-X_i}{h}\end{aligned}$$
2. $\widehat{f_n}(\cdot)$ a les même propriétés de continuité et de différentiabilité que K :
   — Si K est continu, $\widehat{f_n}(\cdot)$ sera une fonction continue.
   — Si K est différentiable, $\widehat{f_n}(\cdot)$ sera une fonction différentiable.

   Dans ce qui suit, nous supposons que les dérivées première et seconde de f existent et admettent une intégrale finie sur $\mathbb{R}$.

1.4.1 Espérance, Biais et variance de l'estimateur

Espérance mathématique

Proposition 1.4.1. *Soit x fixe dans $\mathbb{R}$. l'espérance de l'estimateur $\hat{f}_n(x)$ est :*

$$\mathrm{E}\left[\hat{f}_n(x)\right] = f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right) \tag{1.12}$$

Démonstration : Comme les variables aléatoires $X_1, X_2, \ldots, X_n$ sont i.i.d., nous avons :

$$\begin{aligned}\mathbb{E}\left[\widehat{f_n}(x)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] \\ &= \frac{1}{n}\sum_{=1}^{n}\mathbb{E}\left[\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] \\ &= \mathbb{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] \\ &= \int_{-\infty}^{+\infty}\frac{1}{h}K\left(\frac{x-t}{h}\right)f(t)dt\end{aligned}$$

Nous effectuons le changement de variable suivant : $-u = \dfrac{x-t}{h}$ d'où $t = x + uh$
de là, en utilisant l'hypothèse $K(-u) = K(u)$, l'espérance peut s'écrire sous une forme plus

simple, qui ne dépend que du paramètre h. En étulisant le dévelopement limité d'ordre 2

$$f(x+uh) = f(x) + uhf'(x) + \frac{(uh)^2}{2!}f''(x) + o\left(h^2\right).$$

Ainsi, nous obtenons

$$\mathrm{E}\left[\widehat{f_n}(x)\right] = f(x) + hf'(x)\int_{-\infty}^{+\infty} uK(u)du + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right),$$

d'après les conditions : $\int_{-\infty}^{+\infty} K(u)du = 1$ et $\int_{-\infty}^{+\infty} uK(u)du = 0$, l'expression finale de l'espérance est alors :

$$\mathbb{E}\left[\widehat{f_n}(x)\right] = f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right).$$

Le Biais

Proposition 1.4.2. *Soit x dans $\mathbb{R}$. Le Biais de l'estimateur $\widehat{f_n}(x)$ est donné par :*

$$\mathrm{Biais}\left(\widehat{f_n}(x)\right) = \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right). \tag{1.13}$$

Démonstration : Le Biais s'ecrit :

$$\mathrm{Biais}\left(\widehat{f_n}(x)\right) = \mathbb{E}\left[\widehat{f_n}(x)\right] - f(x)$$

En remplaçant l'espérance par l'expression (1.12) on trouve :

$$\mathrm{Biais}\left(\widehat{f_n}(x)\right) = f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right) - f(x),$$

$$\text{d'où} \quad \mathrm{Biais}\left(\widehat{f_n}(x)\right) = \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o\left(h^2\right).$$

Ainsi

$$\mathrm{Biais}\left(\hat{f}_n(x)\right) = \frac{h^2}{2}f''(x)\mu_2(K) + o\left(h^2\right) \quad \text{où } \mu_2(K) = \int_{-\infty}^{+\infty} u^2K(u)du.$$

La variance

Proposition 1.4.3. *Soit x fixe dans $\mathbb{R}$. La variance de l'estimateur $\widehat{f_n}(x)$ est :*

$$\mathrm{Var}\left(\widehat{f_n}(x)\right) = \frac{1}{nh}f(x)\int_{-\infty}^{+\infty} K^2(u)du + o\left(\frac{1}{nh}\right). \tag{1.14}$$

Démonstration : Partant de l'hypothèse d'indépendance entre les X_i, nous avons

$$\begin{aligned}\mathrm{Var}\left(\hat{f}_n(x)\right) &= \mathrm{Var}\left\{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right\} \\ &= \frac{1}{n}\left\{\frac{1}{h^2}E\left(\left[K\left(\frac{x-X}{h}\right)\right]^2\right) - \frac{1}{h^2}E^2\left(K\left(\frac{x-X}{h}\right)\right)\right\}.\end{aligned}$$

Avec le changement de variable, $-u = \dfrac{x-t}{h}$, on obtient :

$$\mathrm{Var}\left(\widehat{f_n}(x)\right) = \frac{1}{nh}\int_{-\infty}^{+\infty}(K(u))^2 f(x+uh)du - \frac{1}{n}\left[\int_{-\infty}^{+\infty}K(u)f(x+uh)du\right]^2$$

Le terme

$$\frac{1}{n}\left[\int_{-\infty}^{+\infty}K(u)f(x+uh)du\right]^2 \to 0 \quad \text{lorsque} \quad n\to\infty$$

En faisant le développent de Taylor à l'ordre 0 au voisinage de x, $f(x+uh) = f(x)+o(1)$. d'où

$$\begin{aligned}\mathrm{Var}\left(\widehat{f_n}(x)\right) &= \frac{1}{nh}\int_{-\infty}^{+\infty}(K(u))^2[f(x)+o(1)]du \\ &= \frac{1}{nh}f(x)\int_{-\infty}^{+\infty}K^2(u)du + o\left(\frac{1}{nh}\right)\end{aligned}$$

Ainsi

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh}f(x)R(K) + o\left(\frac{1}{nh}\right) \text{ Où } \quad R(K) = \int_{-\infty}^{+\infty}K^2(u)du.$$

1.4.2 Critère d'erreur

Erreur quadratique moyenne (MSE) de l'estimateur à noyau

Proposition 1.4.4. *L'erreur quadratique moyenne (MSE) s'écrit sous la forme, pour tout x :*

$$\begin{aligned}\mathrm{MSE}\left(f(x),\hat{f}_n(x)\right) = \frac{1}{nh}f(x)\int_{-\infty}^{+\infty}K^2(u)du + \frac{1}{4}h^4\left(f''(x)\right)^2\left(\int_{-\infty}^{+\infty}u^2K(u)du\right)^2 \\ + o\left(h^4+\frac{1}{nh}\right).\end{aligned} \tag{1.15}$$

Démonstration :

$$\begin{aligned}
\mathrm{MSE}\left(f(x),\hat{f}_n(x)\right) &= E\left(f(x)-\hat{f}_n(x)\right)^2 \\
&= \left(\mathrm{Biais}\left(\hat{f}_n(x)\right)\right)^2 + \mathrm{Var}\left(\widehat{f}_n(x)\right) \\
&= \left(\frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty}u^2K(u)du + o\left(h^2\right)\right)^2 + \frac{1}{nh}f(x)\int_{-\infty}^{+\infty}K^2(u)du + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{nh}f(x)\int_{-\infty}^{+\infty}K^2(u)du + \frac{h^4}{4}\left(f''(x)\right)^2\left(\int_{-\infty}^{+\infty}u^2K(u)du\right)^2 + o\left(h^4+\frac{1}{nh}\right) \\
&= \frac{1}{nh}f(x)R(K) + \frac{h^4}{4}\left(f''(x)\right)^2\left(\mu_2(K)\right)^2 + o\left(h^4+\frac{1}{nh}\right).
\end{aligned}$$

On note par AMSE l'approximation asymptotique de la MSE

$$\mathrm{AMSE}\left(f,\hat{f}_n\right) = \frac{1}{nh}f(x)R(K) + \frac{h^4}{4}\left(f''(x)\right)^2\left(\mu_2(K)\right)^2. \tag{1.16}$$

✠ Erreur quadratique moyenne intégrée (MISE) [3, 28]

Proposition 1.4.5. *L'erreur quadratique moyenne intégrée (MISE) s'écrit sous la forme :*

$$\mathrm{MISE}\left(f,\hat{f}_n\right) = \frac{1}{nh}R(K) + \frac{h^4}{4}\left(\mu_2(K)\right)^2\int_{-\infty}^{+\infty}\left(f''(x)\right)^2dx + o\left(h^4+\frac{1}{nh}\right). \tag{1.17}$$

Démonstration : En utilisant l'expression du critère MSE (1.15), nous avons :

$$\begin{aligned}
\mathrm{MISE}\left(f,\hat{f}_n\right) &= \int_{-\infty}^{+\infty}\mathrm{MSE}\left(f(x),\hat{f}_n(x)\right)dx \\
&= \int_{-\infty}^{+\infty}\left[\frac{1}{nh}f(x)R(K) + \frac{h^4}{4}\left(f''(x)\right)^2\left(\mu_2(K)\right)^2 + o\left(h^4+\frac{1}{nh}\right)\right]dx \\
&= \frac{1}{nh}R(K)\int_{-\infty}^{+\infty}f(x)dx + \frac{h^4}{4}\left(\mu_2(K)\right)^2\int_{-\infty}^{+\infty}\left(f''(x)\right)^2dx + o\left(h^4+\frac{1}{nh}\right) \\
&= \frac{1}{nh}R(K) + \frac{1}{4}h^4\left(\mu_2(K)\right)^2R\left(f''\right) + o\left(h^4+\frac{1}{nh}\right).
\end{aligned}$$

Avec :

$$\mu_2(K) = \int_{-\infty}^{+\infty}u^2K(u)du, \quad R(K) = \int_{-\infty}^{+\infty}K^2(u)du, \quad \boldsymbol{R}\left(f''\right) = \int_{-\infty}^{+\infty}\left(f''(x)\right)^2dx.$$

On note par AMISE l'approximation asymptotique de l'MISE

$$\mathrm{AMISE}\left(f,\hat{f}_n\right) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\left(\mu_2(K)\right)^2\int_{-\infty}^{+\infty}\left(f''(x)\right)^2dx. \tag{1.18}$$

1.4.3 Propriétés asymptotique du Biais et de la variance de l'estimateur

Dans ce qui suit, on suppose que les conditions suivantes sont vérifiées :

$$\lim_{n\to\infty} h(n) = 0, \lim_{u\to\infty} |uK(u)| = 0, \quad \sup_u |K(u)| < \infty, \quad \int_{-\infty}^{+\infty} K(u)du = 1 \tag{1.19}$$

✠ Comportement asymptotique du Biais

Théorème 1.4.1. *(Parzen [44])*
Si la fonction K satisfait les conditions (1.19) : Alors, l'estimateur $\hat{f}_n(x)$ est asymptotiquement sans Biais c'est-à-dire : $\lim_{n\to\infty} \mathbb{E}\left(\hat{f}_n(x)\right) = f(x)$, en tout point x pour lequel la densité f est continue.

Démonstration : voir : (Parzen [44]) **✠ Comportement asymptotique de la variance**

Théorème 1.4.2. *(Parzen [44])*
Si la fonction K satisfait les conditions (1.19) alors

$$\lim_{n\to\infty} nhV\left(\hat{f}_n(x)\right) = f(x)\int_{-\infty}^{+\infty} K^2(u)du,$$

en tout point x pour lequel la densité f est continue.

Démonstration : voir (Parzen [44])

Les convergences au sens de l'erreur quadratique moyenne et de l'erreur quadratique moyenne intégrée ont été établies respectivement par Parzen [44] et Tiago de Oliveira [55] pour l'estimateur $\hat{f}_n(x)$ de Parzen-Rosenblatt défini dans (1.10).

1.4.4 Critères de convergence

✠ Convergence en moyenne quadratique

Théorème 1.4.3. *(Parzen [44])*
Si $\lim_{n\to\infty} nh(n) = \infty$ et K satisfait aux conditions (1.19), alors l'estimateur $\hat{f}_n(x)$ est consistant en moyenne quadratique, c'est-à-dire :

$$\lim_{n\to\infty} \mathrm{MSE}\left(f(x), \hat{f}_n(x)\right) = 0,$$

en tout point x de continuité de la densité f.

✠ Convergence en moyenne quadratique intégrée

Théorème 1.4.4. *(Parzen [44])*
Si K est un noyau de Parzen - Rosenblatt, on a :
Si

$$\lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} nh(n) = \infty$$

Alors,

$$(\forall f \in \mathbb{L}^p)\,, \lim_{n\to\infty} \text{MISE}\left(f(x), \hat{f}_n(x)\right) = 0$$

Où $\mathbb{L}^p$ est l'ensemble des fonctions réelles de puissance p ^{ième} integrable, c'est-à-dire l'ensemble des fonctions f définies sur $\mathbb{R}$, telles que $\int |f(x)|^p dx < \infty$.

✠ Convergence uniforme en probabilité

La convergence uniforme en probabilité a été obtenue par Parzen [44] en 1962.

Théorème 1.4.5. *(Parzen [44])*
Si $\lim_{n\to\infty} nh(n)^2 = \infty$, et si la fonction K satisfait les conditions (1.19). Et si la transformé de Fourier

$$\tilde{K}(z) = \int_{-\infty}^{+\infty} \exp(-izu)K(u)du$$

est absolument intégrable, alors, $\hat{f}_n(x)$ est un estimateur uniformement consistant en probabilité, c'est-à-dire :

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\hat{f}_n(x) - f(x)\right| < \epsilon\right) = 1.$$

démonstration 1. *voir (Parzen [44])*

✠ Convergence uniforme prèsque complète

La convergence presque complète a eté obtenue par Nadaraya(1965).

Théorème 1.4.6. *(Nadaraya [41])*
Si K est un noyau positif à variation bornée et f est uriformément continue,
si

$$\lim_{n\to\infty} h(n) = 0 \text{ et } \sum_{n=1}^{\infty} \exp\left(-\gamma nh(n)^2\right) < \infty, \quad \forall \gamma > 0$$

Alors,

$$\sup_{x\in\mathbb{R}} \left|\hat{f}_n(x) - f(x)\right| \to 0 \text{ avec une probabilite 1.}$$

Silverman [51] a donné le même théorème sur la convergence presque complète en remplaçant la condition, $\sum_{n-1}^{\infty} \exp\left(-\gamma nh(n)^2\right) < \infty$ par les deux conditions suivantes :

$$\lim_{n\to\infty} h(n) = 0 \text{ et } \lim_{n\to\infty} \frac{\log n}{nh(n)} = 0.$$

Théorème 1.4.7. *(Silverman [51])*

Si on a :

$$\lim_{n\to\infty} h(n) = 0 \text{ et } \lim_{n\to\infty} \frac{\log n}{nh(n)} = 0.$$

et K satisfait aux conditions suivantes :

- *K est uniformement continu et à variation bornée sur $\mathbb{R}$.*
- *supposons aussi que f est uniformément continu,*
- $\int_{-\infty}^{+\infty} |K(u)|du < \infty, \quad \int_{-\infty}^{+\infty} \sqrt{|u\log(u)|}|dK(u)| < \infty,$
- $\int_{-\infty}^{+\infty} K(u)du = 1.$

Alors,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left[\hat{f}_n(x) - f(x)\right] = 0 \quad \textit{Presque sûrement}$$

Démonstartion : voir Silverman [51].

✠ Convergence en loi

La convergence en loi a été établie par Parzen 1962.

Théorème 1.4.8. *(Parzen [44])*

Si $\lim_{n\to\infty} h(n) = 0$, $\lim_{n\to\infty} nh(n) = \infty$ et K satisfait les conditions du théorème 1.4.1, alors $\hat{f}_n(x)$ est un estimateur asymptotiquement normal en tout point x pour lequel la densité f est continue, c'est à dire :

$$\frac{\hat{f}_n(x) - E\left\{\hat{f}_n(x)\right\}}{\sqrt{Var\left\{\hat{f}_n(x)\right\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0.1).$$

où $\xrightarrow{\mathcal{L}}$ désigne la convergence en loi et $\mathcal{N}(0.1)$ est la loi normal standard.

✠ Convergence $\mathbb{L}^1$ presque complète

Théorème 1.4.9. *(Deuroye [15])*

Si

$$\lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} nh(n) = \infty$$

alors

$$(\forall f \in \mathcal{F}), \lim_{n\to\infty} \int \left|\hat{f}_n(x) - f(x)\right| dx = 0, \ \textit{Presque Complètement,}$$

où $\mathcal{F}$: Ensemble des densités de probabilité.

1.5 Choix du noyau K et du paramètre de lissage h

L'estimateur de Parzen- Rosenblatt (1.10) de la fonction densité de probabilité étant une fonction du noyau K et du paramètre de lissage h, il est nécessaire de faire un bon choix pour ces deux paramètres pour avoir un estimataur avec de bonnes propriétés statistiques et asymptotiques.

1.5.1 Choix du paramètre de lissage h

1.5.1.1 Choix théorique

On suppose que la densité à estimer f et le noyau K sont des fonctions de carré intégrable. de sorte que l'MISE soit finie. Rappelons que

$$\text{MISE}\left(f, \hat{f}_n\right) = \frac{h^4}{4} R\left(f''\right) \left(\mu_2(K)\right)^2 + \frac{1}{nh} R(K) + o\left(\frac{1}{nh} + h^4\right)$$

L'approximation asymptotique du l'MISE est donnée par

$$\text{AMISE}\left(f, \hat{f}_n\right) = \frac{h^4}{4} R\left(f''\right) \left(\mu_2(K)\right)^2 + \frac{1}{nh} R(K)$$

On constate que l'AMISE s'écrit en fronction du biais et de la variance, le biais est une fonction croissante en h alors que le terme en variance est une fonction décroissante en h, si h est grand la variance sera petite (faible) et le biais sera fort, donc la valeur optimale de h qui minimise l'erreur quadratique moyenne intégrée MISE réalise un compromis entre le Biais et la variance.

On calcule le h optimal qui minimize l'AMISE comme suit :

$$\begin{aligned}
&\frac{\partial}{\partial h} \text{AMISE}\left(f, \hat{f}_n\right) = 0 \\
\Rightarrow & h^3 \mu_2^2(K) \int_{-\infty}^{+\infty} \left(f''(x)\right)^2 dx - \frac{1}{nh^2} \int_{-\infty}^{+\infty} K^2(u) du = 0 \\
\Rightarrow & h^5 = \frac{\int_{-\infty}^{+\infty} K^2(u) du}{n\mu_2^2(K) \int_{-\infty}^{+\infty} \left(f''(x)\right)^2 dx}
\end{aligned}$$

D'où

$$h_{opt} = \left[\frac{\int_{-\infty}^{+\infty} K^2(u) du}{\mu_2^2(K) \int_{-\infty}^{+\infty} \left(f''(x)\right)^2 dx}\right]^{1/5} n^{-\frac{1}{5}}$$

donc :

$$h_{opt} = \left[\frac{R(K)}{\mu_2^2(K) R\left(f''\right)}\right]^{1/5} n^{-\frac{1}{5}} \tag{1.20}$$

sous condition que $f''(x) \neq 0$.

$$\frac{\partial^2}{\partial^2 h}\ \text{AMISE}\ \left(f(x), \hat{f}_n(x)\right) = 3h^2\mu_2^2(K)R\left(f''\right) + \frac{2}{nh^3}R(K) > 0 \Rightarrow h_{opt}$$

minimise la valeur de AMISE.

En substituant h_{opt} dans la formule AMISE on obtient :

$$\text{AMISE}_{h_{\text{opt}}} = \frac{5}{4}C(K)R\left(f''\right)^{\frac{1}{5}} n^{-\frac{4}{5}} \tag{1.21}$$

Avec

$$C(K) = \left(\mu_2^2(K)R^4(K)\right)^{1/5}$$

Exemple. *[4] Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f suite une loi normale, donc iI reste le paramètre h à estimer.*
Si $f \sim \mathcal{N}\left(\mu; \sigma^2\right)$ suite loi normale alors $f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$, *si on pose* $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, *et*

$$f''(x) = \frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right), \quad \varphi''(x) = \frac{1}{\sqrt{2\pi}}\left(x^2-1\right)e^{-x^2/2}$$

La quantité inconnue $R\left(f''\right)$ sécrit alors

$$\begin{aligned} R\left(f''\right) &= \int_{-\infty}^{+\infty}\left[f''(x)\right]^2 dx \\ &= \frac{1}{\sigma^6}\int_{-\infty}^{+\infty}\left\{\varphi''\left(\frac{x-\mu}{\sigma}\right)\right\}^2 dx \\ &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\left\{\varphi''(v)\right\}^2 dv \end{aligned}$$

Nous avons :

$$\begin{aligned} \varphi(v) &= \frac{1}{\sqrt{2\pi}}e^{-v^2/2} \\ &\Rightarrow \varphi'(v) = -\frac{v}{\sqrt{2\pi}}e^{-v^2/2} \\ &\Rightarrow \varphi''(v) = \frac{1}{\sqrt{2\pi}}\left(v^2-1\right)e^{-v^2/2}. \end{aligned}$$

$$\begin{aligned} R\left(f''\right) &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\left\{\frac{1}{\sqrt{2\pi}}\left(v^2-1\right)e^{-v^2/2}\right\}^2 dv \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{\int_{-\infty}^{+\infty}v^4e^{-v^2}dv - 2\int_{-\infty}^{+\infty}v^2e^{-v^2}dv + \int_{-\infty}^{+\infty}e^{-v^2}dv\right\} \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}v^2e^{-v^2}dv + \int_{-\infty}^{+\infty}e^{-v^2}dv\right\} \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}\frac{u^2}{2}e^{-u^2/2}\frac{1}{\sqrt{2}}du + \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2}}e^{-u^2/2}du\right\} \quad \text{avec } u = \sqrt{2}v \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{4}\sqrt{\pi} + \sqrt{\pi}\right\} \\ &= \frac{1}{\sigma^5}\frac{3}{8\sqrt{\pi}} \end{aligned}$$

Donc, l'expression du paramètre de lissage optimal devient

$$h_{opt} = \left[\frac{8\sqrt{\pi}R(K)}{3\left[\mu_2(K)\right]^2}\right]^{\frac{1}{5}} \hat{\sigma} n^{-1/5} \tag{1.22}$$

où $\hat{\sigma}$ est l'estimateur de σ, donné par :

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \text{ et } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On a $K \to N(0,1)$ alors

$$\begin{aligned} R(K) &= \int_{-\infty}^{+\infty} [K(u)]^2 du \\ &= \int_{-\infty}^{+\infty} \left[\frac{1}{\sqrt{2\pi}} e^{-u^2/2}\right]^2 du \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\pi} e^{-u^2} du \\ &= \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}} \end{aligned}$$

Nous avons $\mu_2(K) = \int_{-\infty}^{+\infty} u^2 K(u) du = 1$. Nous remplaçons dans l'équation (1.22) nous obtenons dans le cas du noyau gaussien :

$$h_{opt} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \hat{\sigma} n^{-1/5} = 1.06 \hat{\sigma} n^{-1/5} \tag{1.23}$$

Exemple. *[4] Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales $\mathcal{N}\left(\mu; \sigma^2\right)$, soit K un noyau d'Epanechnikov.*

$$\text{Si} \begin{cases} X_i & \sim \mathcal{N}\left(\mu; \sigma^2\right) \\ K(u) & = \frac{3}{4}\left(1 - u^2\right), 1_{|u| \leq 1} \end{cases}$$

Alors :

$$h_{opt} = 2.34 \hat{\sigma} n^{-1/5} \tag{1.24}$$

1.5.1.2 Choix pratique

•Méthode de validation croisée par moindre carrés [24]

Cette méthode appellée aussi méthode de validation croisée non biaisée (Unbiaised CrossValidation "UCV") proposée par Rudemo [50] en 1982 et Bowman [8] en 1984. Le principe de cette méthode est la minimisation de l'erreur quadratique moyenne intégrée MISE de l'estimateur à

noyau, on a :

$$\begin{aligned}\text{MISE}\left(f,\widehat{f_n}\right) &= \mathbb{E}\int_{-\infty}^{+\infty}\left\{\widehat{f_n}(x)-f(x)\right\}^2 dx \\ &= \mathbb{E}\int_{-\infty}^{+\infty}\left[\widehat{f_n}(x)\right]^2 dx - 2\mathbb{E}\int_{-\infty}^{+\infty}\widehat{f_n}(x)f(x)dx + \mathbb{E}\int_{-\infty}^{+\infty}[f(x)]^2 dx\end{aligned}$$

Le dernier terme ne dépend pas de h, pour minimiser $\text{MISE}\left(f(x),\widehat{f_n}(x)\right)$ il suffit de minimiser l'expression

$$J(h) = \mathbb{E}\left(\int_{-\infty}^{+\infty}\left[\widehat{f_n}(x)\right]^2 dx\right) - 2\mathbb{E}\int_{-\infty}^{+\infty}\widehat{f_n}(x)f(x)dx$$

Puisque J dépend de la densité inconnue f donc on propose de l'estimer et de, choisir h qui minimise son estimateur.

Le premier terme admet : $\int_{-\infty}^{+\infty}\left[\widehat{f_n}(x)\right]^2 dx$ comme estimateur
Pour le second terme on peut monter qu'il admet comme estimateur sans biais la quantité :

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{(n,-i)}(X_i) \tag{1.25}$$

avec

$$\widehat{f}_{(n,-i)}(X_i) = \frac{1}{(n-1)h}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) \tag{1.26}$$

où $\hat{f}_{(n,-i)}$ est l'estimateur de f privé d 'une observation.
On a comme les X_i sont i.i.d,

$$\begin{aligned}\mathbb{E}\{\widehat{G}\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{n,-i}(X_i)\right\} \\ &= \mathbb{E}\left\{\widehat{f}_{(n,-1)}(X_1)\right\} \\ &= \mathbb{E}\left\{\frac{1}{(n-1)h}\sum_{j=1,j\neq 1}^{n} K\left(\frac{X_1 - X_j}{h}\right)\right\} \\ &= \frac{1}{(n-1)}\sum_{j=1,j\neq 1}^{n}\mathbb{E}\left\{\frac{1}{h}K\left(\frac{X_1 - X_j}{h}\right)\right\} \\ &= \mathbb{E}\left\{\frac{1}{h}K\left(\frac{X_1 - X}{h}\right)\right\} \\ &= \frac{1}{h}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K\left(\frac{x-t}{h}\right)f(t)f(x)dtdx \\ &= \frac{1}{h}\int_{-\infty}^{+\infty} f(x)\int_{-\infty}^{+\infty} K\left(\frac{x-t}{h}\right)f(t)dtdx\end{aligned}$$

d'autre part,

$$\begin{aligned}\mathbb{E}\left\{\int_{-\infty}^{+\infty}\widehat{f}_n(x)f(x)dx\right\} &= \mathbb{E}\left\{\int_{-\infty}^{+\infty}\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \quad f(x)dx\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\int_{-\infty}^{+\infty}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)f(x)dx\right\}\\ &= \frac{1}{h}\mathbb{E}\left\{\int_{-\infty}^{+\infty}K\left(\frac{x-X}{h}\right) \quad f(x)dx\right\}\\ &= \frac{1}{h}\int_{-\infty}^{+\infty}f(x)\int_{-\infty}^{+\infty}K\left(\frac{x-t}{h}\right) \quad f(t)dtdx\end{aligned}$$

Ce qui implique que $\mathbb{E}(\widehat{G}) = \mathbb{E}\int_{-\infty}^{+\infty}\widehat{f}_n(x)f(x)dx$. En définitif, l'estimateur sans biais de $J(h)$ est donnée par

$$\text{UCV}(h) = \int_{-\infty}^{+\infty}\left[\widehat{f}_n(x)\right]^2 dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{(n,-i)}(X_i) \tag{1.27}$$

et le paramètre de lissage du type " Unbiaised Cross-Validation " est la valeur de h qui minimise la quantité $UCV(h)$, c'est-à-dire

$$h_{UCV} = \underset{h>0}{\text{argmin}} UCV(h) \tag{1.28}$$

- **Méthode du maximum de vraisemblance par validation croisée** [24]

La méthode du maximum de vraisemblance avec validation croisée (Maximum Likelihood Cross-Validation "MLCV") est proposée par Habbema, Hermans et Van den Broek (1974) [27] et Duin (1976) [18]. Ils choisissent h de sorte que la pseudo-vraisemblance $\prod_{i=1}^{n}\widehat{f}_n(X_i)$ soit maximale. Cependant, le maximum est atteint en $h = 0$. Donc le principe de validation croisée est introduit par le remplacement de $\widehat{f}_n(x)$ par $\widehat{f}_{n,-i}(x)$, où

$$\widehat{f}_{(n,-i)}(X_i) = \frac{1}{h(n-1)}\sum_{j\neq i}K\left(\frac{X_i - X_j}{h}\right) \tag{1.29}$$

Le paramètre de lissage donné par la méthode du maximum de vraisemblance avec validation croisée est le paramètre qui maximise l'expression :

$$\text{MLCV}(h) = \left(\frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{j\leq i}K\left(\frac{X_i - X_j}{h}\right)\right] - \log\left[(n-1)/h\right]\right) \tag{1.30}$$

C'est-à-dire

$$h_{MLCV} = \underset{h>0}{\text{argmax}}\,\text{MLCV}(h) \tag{1.31}$$

1.5.2 Le choix du noyau

On rappelle de l'expression asymptotique de l'erreur quadratique intégrée $\text{AMISE}_{h_{opt}}$

$$\text{AMISE}_{h_{opt}} = \frac{5}{4} C(K) R\left(f''\right)^{\frac{1}{5}} n^{\frac{-4}{5}}$$

Avec

$$C(K) = \left(\mu_2^2(K) R^4(K)\right)^{1/5}$$

Trouver le noyau optimal revient donc à minimiser $C(K)$. Epanechnikov [19] (1969) a trouvé parmi les noyaux symétriques le noyau optimal au sens de l'AMISE sous la forme :

$$K_e(u) = \frac{3}{4}\left(1 - u^2\right) 1_{\{-1 \leq u \leq 1\}}, \qquad \text{[Noyau d'Epanechnikov]}.$$

D'après les travaux de Tsybakov [54] (2004), on peut considérer l'eficacité de chacun des noyaux symétriques présentés dans le tableau (1.2), en comparant avec le noyau d'Epanechnikov. On définit l'efficacité d'un noyau par :

$$\begin{aligned} eff(K) &= \left\{ \frac{C\left(K_e\right)}{C(K)} \right\}^{5/4} \\ &= \frac{3}{5\sqrt{5}} \frac{1}{\sqrt{\int_{-\infty}^{+\infty} u^2 K(u) du} \int_{-\infty}^{+\infty} K(u)^2 du} \leq 1 \end{aligned} \tag{1.32}$$

Noyau	Efficacité
d'Epanechnikov	≈ 1.000
Biweight	≈ 0.9939
triangulaire	≈ 0.9859
Gaussien	≈ 0.9512
Rectangulaire	≈ 0.9295

TABLE 1.2 – Efficacité des noyaux continus symétriques.

Remarque. *On remarque que les valeurs d'efficacité sont très proches de 1 et qu'il ya très peu de différence entre les différents noyaux sur la base de l'erreur quadratique moyenne intégrée. Par conséquent, Le choix du noyau n'est pas très important.*

Chapitre 2

Généralités sur les données incomplètes

2.1 Introduction

Les données manquantes, incomplètes ou erronées sont fréquemment rencontrées dans différents domaines, tels que la médecine, la biologie, la santé publique, l'épidémiologie, l'astronomie, l'économie, la fiabilité, etc...Remarquant que dans chacun de ces domaines, on s'intéresse à des variables aléatoires généralement positives représentant la durée de temps jusqu'à l'apparition d'un certain événement, appelé durée de survie. Les données de survie se caractérisent par l'existence d'observations incomplètes. En effet les données sont souvent recueillies partiellement, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information.

Dans ce chapitre, nous introduisons quelques définitions de base dans l'analyse de survie, nous nous intéressons en particulier aux notions de censure et troncature dans les données, dont nous donnons quelques exemples pour rendre la lecture plus facile.

2.2 Définitions

Quelques définitions sont couramment utilisées dans les études de survie.

Modèle de survie : Un modèle de survie est l'observation d'un n-échantillon d'une variable aléatoire X représentant la durée de vie d'un individu, d'une machine, d'un portefeuille, etc...

L'analyse des données de survie est l'étude des délais de survenue d'un événement.

Durée de survie : La durée de survie noté par la variable aléatoire X (positive), est le temps qui s'écoule depuis une date d'origine (début du traitement,...) jusqu'à la survenue d'un événement d'intérêt (décès, guérison,...)

Date d'origine : Elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente.

Date de point : C'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

Date des dérnières nouvelles : Cest la date la plus récente où des informations sur un sujet ont été recueillies.

2.3 Fonctions de base en analyse de survie

Dans cette partie, nous allons définir des fonctions jouant un rôle très important en analyse de survie et nous allons voir comment elles sont interdépendantes.
Soit X une variable aléatoire non négative et continue qui représente la durée de survie d'un sujet dans une expérience. Plusieurs fonctions caractérisent la distribution de X.

2.3.1 La fonction densité

On note $f(\cdot)$ la fonction densité de X à valeurs dans $\mathbb{R}^+$ définie par

$$f(t) = \lim_{dt \to 0} \frac{\mathrm{P}(t \leqslant X < t + dt)}{dt}, \quad t \in \mathbb{R} \text{ où } t > 0. \tag{2.1}$$

2.3.2 La fonction de répartition

La fonction de répartition $F(\cdot)$ mesure la probabilité de décéder entre 0 et t :

$$F(t) = \mathrm{P}(X \leqslant t) = \int_0^t f(x)dx \tag{2.2}$$

La fonction de répartition est croissante, telle que :

$$\lim_{t \to 0^+} F(t) = 0$$
$$\lim_{t \to \infty} F(t) = 1$$

2.3.3 La fonction de survie

La fonction de survie, notée par $S(t)$ est définie comme :

$$\begin{aligned} S(t) &= \mathrm{P}(\text{un individu survit au-delà du temps } t) \\ &= \mathrm{P}(X > t) \end{aligned} \tag{2.3}$$

A partir de la définition de fonction de répartition F de X

$$\begin{aligned} S(t) &= 1 - \mathrm{P}(\text{un individu décède entre 0 et } t) \\ &= 1 - F(t), \quad t > 0 \end{aligned}$$

Notons que S est une fonction monotone décroissante avec les propriétés :

$$\lim_{t\to 0} S(t) = 1$$
$$\lim_{t\to \infty} S(t) = 0$$

2.3.4 Taux de hasard

Cette fonction est aussi appelée fonction de risque instantané de décès. Le taux de hasard, noté $\Lambda(.)$, est la probabilité pour q'un sujet décéde au temps t sachant qu'il est encore vivant juste avant t, définie par :

$$\Lambda(t) = \begin{cases} 0 & \text{si} \quad S(t) = 0 \\ \dfrac{f(t)}{S(t)} & \text{si} \quad S(t) \neq 0, \quad t > 0 \end{cases} \tag{2.4}$$

Le taux de hasard se justifie par le fait que si f est continue, alors, pour tout $t > 0$ on a

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

En effet $\mathrm{P}(t \leq X < t + dt \mid X \geq t) = \dfrac{P(t \leq X < t + dt)}{S(t)}$

$$\begin{aligned} \lim_{dt\to 0} \frac{\mathrm{P}(t \leq X < t + dt \mid X \geq t)}{dt} &= \lim_{dt\to 0} \frac{1}{dt} \frac{\mathrm{P}(t \leq X < t + dt)}{S(t)} \\ &= \lim_{dt\to 0} \frac{1}{dt} \frac{F(t + dt) - F(t)}{S(t)} \\ &= \frac{F'(t)}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

2.3.5 La fonction de hasard cumulée

Supposons que Λ est intégrable, on définit la fonction de hasard cumulée, notée H , en posant

$$H(t) = \int_0^t \Lambda(x)dx \tag{2.5}$$

Remarque. *Toutes les fonctions* (f, F, S, Λ, H) *permettent de décrire la distribution de la durée de survie* X.

Remarque. *Si on connait une de ces fonctions, les autres peuvent être déterminées.*

On va le voir dans la proposition suivante :

Proposition 2.3.1. *[6]*

Supposons que X admette une densité f qui soit une fonction continue sur $\mathbb{R}^+$, et soit $A = \{t > 0; S(t) \neq 0\}$, alors les propriétés suivantes sont équivalentes :

1. $\forall t \in A, \Lambda(t) = f(t)/S(t)$.
2. $\forall t \in A, \Lambda(t) = (-\log S(t))'$.
3. $\forall t \in A, S(t) = \exp(-H(t))$.
4. $\forall t \in A, f(t) = \Lambda(t)\exp(-H(t))$.

Démonstration :

1) $\Rightarrow$ 2)

$$\Lambda(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = \frac{-S'(t)}{S(t)} = (-\log S(t))'$$

2) $\Rightarrow$ 3)

$$\begin{aligned}\Lambda(t) &= (-\log S(t))' \\ &= \frac{-S'(t)}{S(t)} \\ \Rightarrow S(t) &= \exp\left(-\int_0^t \Lambda(x)dx\right) \\ &= \exp(-H(t))\end{aligned}$$

3) $\Rightarrow$ 4) on a

$$S(t) = \exp(-H(t))$$

par ailleurs

$$\Lambda(t) = \frac{f(t)}{S(t)}$$

d'où

$$\begin{aligned}f(t) &= \Lambda(t)S(t) \\ &= \Lambda(t)\exp(-H(t))\end{aligned}$$

4) $\Rightarrow$ 1) On a

$$\begin{aligned}f(t) &= \Lambda(t)\exp(-H(t)) \\ \Rightarrow \Lambda(t) &= \frac{f(t)}{\exp(-H(t))}\end{aligned}$$

Par ailleurs

$$\begin{aligned}S(t) &= 1 - F(t) \\ &= 1 - \int_0^t f(x)dx \\ &= \int_t^{+\infty} f(x)dx\end{aligned}$$

D'après 4), on a

$$\begin{aligned} S(t) &= \int_t^{+\infty} \Lambda(x)\exp(-H(x))dx \\ &= -\exp(-H(x))|_t^{+\infty} \end{aligned}$$

comme

$$\begin{aligned} H(t) &= -\log S(t) \\ &= -\log[1-F(t)] \end{aligned}$$

par ailleurs

$$F(t) \to 1, \text{ quand } t \to +\infty$$

Ainsi

$$\log[1-F(t)] \to -\infty \text{ et donc } H(t) \to +\infty, \text{ quand } t \to +\infty$$

D'où
$\exp(-H(x)) \to 0$, quand $t \to +\infty$

Donc $S(t) = \exp(-H(t))$.

2.4 Moyenne et variance de la durée de survie

Théorème 2.4.1. [43]

1.

$$\mathbb{E}(X) = \int_0^{+\infty} S(t)dt \tag{2.6}$$

2.

$$\text{Var}(X) = 2\int_0^{+\infty} tS(t)dt - \left(\int_0^{+\infty} S(t)dt\right)^2 \tag{2.7}$$

Démonstration :

1. Remarquons que $f(t)$ peut s'écrire $f(t) = -\frac{d}{dt}(1-F(t)) = -\frac{d}{dt}S(t)$. En faisant une intégration par partie on a

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{+\infty} tf(t)dt = -\int_0^{+\infty} t\frac{d}{dt}S(t)dt = -\underbrace{tS(t)|_0^{+\infty}}_{0} + \int_0^{+\infty} S(t)dt \\ &= \int_0^{+\infty} S(t)dt \end{aligned}$$

   car $S(+\infty) = \mathbb{P}(X > +\infty) = 0$ (événement impossible).

2. En procédant de la même manière on montre que $\mathbb{E}\left(X^2\right) = 2\int_0^{+\infty} tS(t)dt$ d'où le résultat.

2.5 Censure et troncature

2.5.1 Données censurées

Une donnée est dite "censurée" si la valeur exacte n'est pas connue, seules des bornes supérieures et (ou) inférieures pour cette valeur sont disponibles. La censure peut se manifester pour différentes raisons : l'événement d'intérêt n'est pas survenu au moment de l'analyse, un sujet peut être perdu de vue avant d'avoir expérimenté l'événement d'intérêt, etc...
Le phénomène de censure est plus couramment rencontré lors du recueil de données de survie. Le modèle de censure à droite est la forme de censure la plus comune dans les études médicales. Un ouvrage qui fait autorité sur le sujet est le livre de Andersen et al. (1992). On peut citer aussi Survival Analysis écrit par Klein et Goel (1991), les livres de Klein et Moeschberger (1997)[31] et de Com-Nougué et al.(1999)[14].

Définition 2.5.1. *[10] Une durée de vie aléatoire X est dite censurée par une variable aléatoire de censure C si on observe parfois C au lieu de X.*

Définition 2.5.2. *Soit $X_1, \ldots, X_n$ un n-échantillon de X. On appelle statistique d'ordre $X_{(i)}, \quad i = \overline{1,n}$ la suite $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ où $X_{(1)} = \min_{1\leq i\leq n} X_i$ et $X_{(n)} = \max_{1\leq i\leq n} X_i$.*

Dans ce qui suit, on considère pour chaque "individu" i :

- Son temps de survie X_i ;
- Son temps de censure C_i ;
- La durée réellement observée T_i.

2.5.1.1 Censure à droite

Définition 2.5.3. *[10] Une durée de vie aléatoire X est dite censurée à droite par une variable aléatoire de censure C si on observe parfois C au lieu de X. L'information donnée par C sur X est tel que $X > C$. Autrement dit, l'individu n'a pas connu l'événement à sa dernière observation, on sait seulement que sa durée de vie est supérieure à une certaine valeur connue.*

Un exemple [43] typique est celui où l'événement d'intérêt est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation (on n'a plus de nouvelles après l'hospitalisation).

Types de censure à droite :

1. **La censure de type I (censure fixe)** [47]
   Soit C une valeur fixée, au lieu d'observer les variables $X_1, \ldots, X_n$ qui nous intéressent, on n'observe X_i que si $X_i \leqslant C$, sinon on sait uniquement que $X_i > C$. On utilise la notation suivante :
   $$T_i = X_i \wedge C = \min(X_i, C). \tag{2.8}$$

La date de fin d'étude constitue un exemple de censure fixe.

2. **La censure de type II (censure d'attente)** [47]
   Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables X_i et T_i. La date de censure est donc $X_{(k)}$ et on observe les variables suivantes

$$T_{(1)} = X_{(1)}, \ldots, T_{(k)} = X_{(k)}, T_{(k+1)} = X_{(k)}, \ldots, T_{(n)} = X_{(k)}$$

   Exemple. *[43] En contrôle de qualité lorsqu'on observe la durée de vie d'un n-échantillon de lampes produites par une machine : dès qu'on atteint une certaine proportion de lampes défectueuse (par exemple on admet 5%), on arrête de vérifier les lampes restantes et on considère que les durées de vie restantes sont toutes égales à la dernière durée observée, donc il y a censure de type II.*

3. **La censure de type III (censure aléatoire de type I)** [43]
   Soient $C_1, \ldots, C_n$ des variables aléatoires i.i.d. qui représentent le temps de censure pour chaque individu. Donc pour chaque individu i, soit on observe sa vraie durée de vie X_i, soit on observe son temps de censure C_i. Donc l'échantillon de taille n qui est en réalité observé est formé par les couples de variables (T_i, δ_i) où

$$T_i = \min(X_i, C_i) \text{ et } \delta_i = \mathbf{1}_{\{X_i \leq C_i\}} \tag{2.9}$$
$$= \begin{cases} 1 & \text{si } X_i \leq C_i : \text{ le } i-ème \text{ individu n'est pas censuré} \\ 0 & \text{si } X_i > C_i : \text{ le } i-ème \text{ individu est censuré} \end{cases}$$

   $\delta_i = \mathbf{1}_{\{X_i \leq C_i\}}$ (l'indicatrice de non censure).
   La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être provoquée par :
   - Arrêt du traitement médical car non supporté,
   - Déménagement,
   - Décès par une cause indépendante de la maladie étudiée,
   - Fin de l'étude.

2.5.1.2 Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà connu l'événement avant qu'il soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu i, on observe

$$T = X \vee C = \max(X, C) \quad \text{et} \quad \delta = \mathbf{1}_{\{X \geq C\}} \tag{2.10}$$

Exemple. *[47] Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).*

Exemple. *[43] Si on veut savoir à quel âge X les enfants d'une maternelle donnée sont capables de lire les lettres de l'alphabet. Au début de l'étude, certains enfants d'âge C (observé) sont déjà capables de les lire, et pour eux $X \leq C$: il s'agit d'une censure gauche.*

2.5.1.3 Censure double

Il y a des situations où dans le même échantillon on peut trouver des données censurées à droite et d'autres censurées à gauche.

Exemple. *Leiderman et al. (1973)* [34, 10] *ont étudié l'âge auquel les enfants d'une communauté africaine apprennent à accomplir certaines tâches. Au début de l'étude, certains enfants savaient déjà effectuer les tâches étudiées, on sait seulement alors que l'âge où ils ont appris est inférieur à leur âge à la date du début de l'étude (censure à gauche). A la fin de l'étude, certains enfants ne savaient pas encore accomplir ces tâches et on sait alors seulement que l'âge auquel ils ont appris est supérieur à leur âge à la fin de l'étude (censure à droite). Dans cet exemple, on trouve dans un même échantillon des données censurées à gauche aussi bien que des données censurées à droite.*

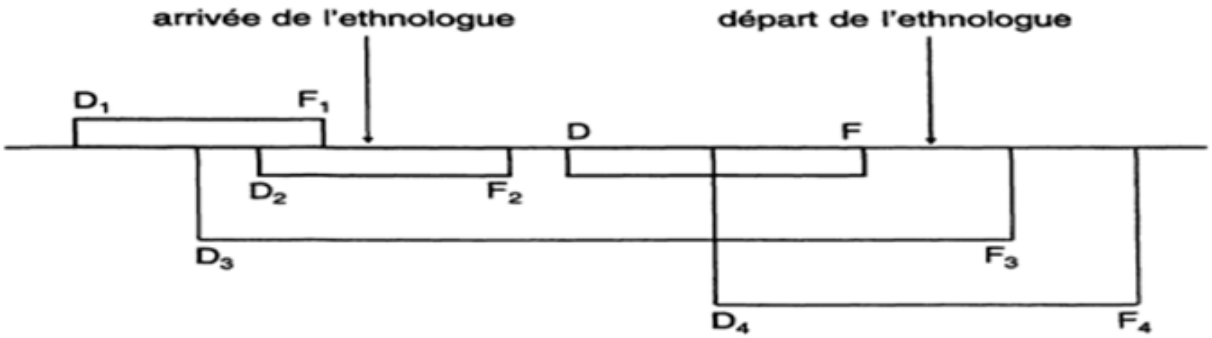


FIGURE 2.1 – Exemple de censures droite et gauche

Dans la figure 2.1 D_i est le début de l'apprentissage, F_i la fin, pour le sujet i.

D_1F_1 : est censuré à gauche par l'âge C de l'enfant.

D_2F_2, D_3F_3, D_4F_4 : bien qu'étant de trois types différents, ils sont tous les trois censurés à droite : le premier par l'âge de l'enfant, le second par la durée de séjour de l'ethnologue et le troisième par la durée d'aprentissage observée par l'ethnologue.

DF : n'est pas censuré.

2.5.1.4 Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues

$$C_1 \leq T \leq C_2. \tag{2.11}$$

Exemple. *[6] Un patient se rend à l'hôpital à des dates régulières : s'il ne se présente pas à un rendez-vous, on sait seulement que son décès s'est produit dans l'intervalle entre la dernière visite et le rendez-vous.*

Exemple. *[34] Pour détecter les composants défectueux d'un processus de production industriel, on effectue des contrôles selon des dates aléatoires. Lorsqu'on constate qu'un composant est à changer, on sait seulement qu'il est tombé en panne entre les dates de deux contrôles successifs.*

Schéma récapitulatif [49] : Une étude de survie a débuté en 1984 et terminé en 2006.

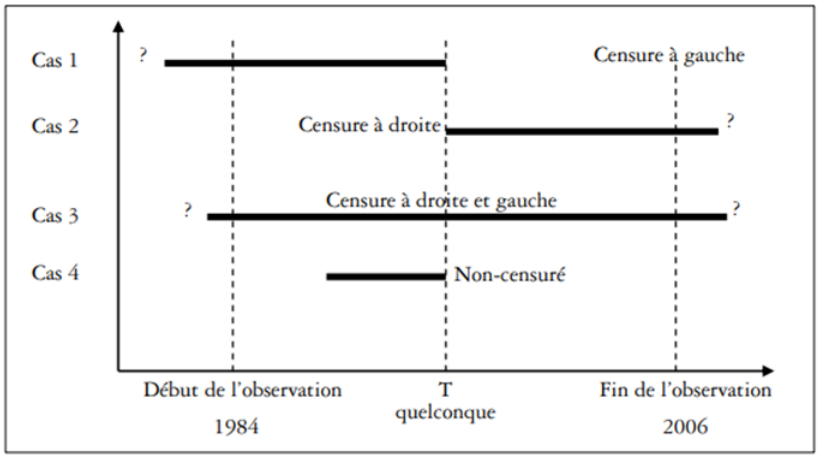


FIGURE 2.2 – Différents types de censure

2.5.2 Données tronquées

Définition 2.5.4. *On dit que la variable d'intérêt X de durée de vie est tronquée si X n'est observable que sous une certaine condition dépendante de la valeur de X [11].*
Plus généralement, il y a troncature si l'observation de la variable d'intérêt X n'a lieu que conditionnellement à un événement B[10].

Un bref historique

Le modèle de troncature est apparu tout d'abord en astronomie, où les échantillons sont composés d'objets astraux d'une certaine zone. Les luminosités absolues et apparentes d'un objet astral sont respectivement définies comme étant sa brillance observée à une distance fixe et depuis la terre et l'on observe que les objets qui sont suffisamment brillants, c'est-à-dire ceux

pour lesquels la luminosité $M \geq m$, m étant la variable de troncature [25]. La troncature est observée aussi dans plusieurs domaines comme la médecine, l'épidémiologie, la biométrie et l'économie. De nombreux travaux ont été effectués sur l'analyse de données tronquées (Klein et Moeschberger, 1997)[31], (Lynden-Bell, 1971)[39].
Il existe plusieurs types de troncature : la troncature à droite, à gauche et par intervalle.

2.5.2.1 Troncature à droite

Définition 2.5.5. *On dit qu'il y a troncature à droite lorsque la variable d'interêt X n'est observable que si elle est inférieure à T. T est alors la variable aléatoire de troncature droite :*

$$X \text{ n'est observée que si } X < T \tag{2.12}$$

Exemple. ***Le problème relatif au SIDA acquis par transfusion (voir Klein et Moeschberger (1997))*** *[6, 10]*
Lagakos et al. (1998) 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 infectés 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 inclus dans l'échantillon. C'est le cas de troncature à droite.

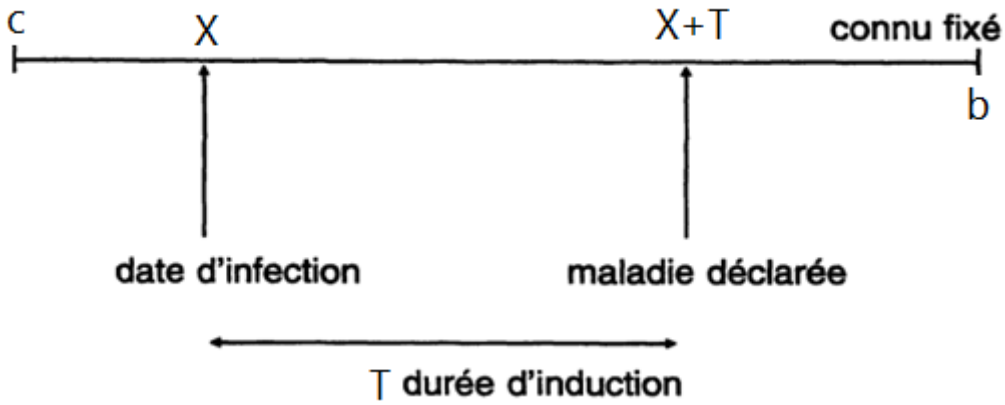


FIGURE 2.3 – Schéma correspondant au SIDA.

La variable d'intérêt est ici la durée d'induction T de la maladie, durée qui s'écoule entre la date d'infection X et la date $(X+T)$ de déclaration de la maladie. On suppose que l'observation a lieu entre deux dates fixes c et b. Il y a donc troncature puisqu'on n'observe T que conditionnellement à l'événement $B = \{c < X + T < b\}$.

2.5.2.2 Troncature à gauche

Définition 2.5.6. *[10] On dit qu'il y a troncature à gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à T. T est alors la variable aléatoire de troncature gauche :*

$$X \text{ n'est observée que si } X > T \tag{2.13}$$

Exemple. *(voir Klein et Moeschberger (1997) [31])*
Une étude de survie des résidents du centre de retraite en Californie est décrite. On enregistre l'âge au décès et l'âge d'entrée au centre. Un individu doit survivre à un âge suffisant pour entrer dans le centre, tous les individus qui sont morts ne seront pas entré dans le centre et sont donc hors de la connaissance de l'enquêteur, ces personnes n'ont aucune chance d'être dans l'étude et sont considérées comme tronquées à gauche.

Exemple. *[42, 17] Des chercheurs travaillant pour la Société de Protection des Animaux veulent étudier la durée de vie des chats. Ils ont à leur disposition des données de la **SPA** concernant les animaux confiés à cette institution. On dispose pour chaque chat de son âge à son arrivée à la SPA et de son âge à la sortie (adoption).* Si *on note T_i l'âge du ième chat (T_i durée de vie) et l'intervalle $B_i = [E_i, F_i]$ où E_i est la date d'arrivée et F_i est la date de sortie de la **SPA**. Pour certains chats qui sont adoptés on a leurs durées de vie $T_i \notin B_i = [E_i, F_i]$ ie T_i n'est pas observé. Donc $T_i \in A_i = [F_i, \infty[$. Donc on a une troncature à gauche par la v.a. F_i.*

2.5.2.3 Troncature par intervalle

Définition 2.5.7. *[47] Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle.*

Exemple. *[47] 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.*

Chapitre 3

Estimation de la densité de probabilité avec la méthode du noyau : cas des données incomplètes

3.1 Introduction

Les durées de survie se caractérisent par l'existence d'observations incomplètes. Cela est dû aux phénomènes de censure et/ou de troncature déjà cités dans le chapitre 2. Les procédures statistiques classiques ne sont alors plus valables. Un problème d'identifiabilité se pose alors. Autrement dit :

— Est-il possible d'estimer la densité de probabilité inconnue f à partir des données incomplètes ?

— Peut-on appliquer la méthode du noyau pour ce type de données ?

Ces problèmes ont suscité l'intérêt de nombreux auteurs vu que la modélisation de ce type de données nécessite des techniques statistiques plus complexes. De nombreux estimateurs ont été développés dans ce cas, afin de considérer les mécanismes de censure et troncature.

Nous présentons dans ce chapitre l'estimation de la densité de probabilité inconnue f par la méthode du noyau dans le cas de données incomplètes. En particulier l'estimateur de Kaplan-Meier (1958) de la fonction de survie, et les estimateurs à noyau de la fonction densité dans les modèles de censure droite, et troncature gauche.

3.2 Estimation de la densité dans le cas des données censurées à droite

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires positives indépendantes et identiquement distribuées (i.i.d) désignant des durés de survie d'un évenement donné de fonction de répartition F. Soit $C_1, C_2, \ldots, C_n$ une suite de variables aléatoires de censures, positives (i.i.d) et de fonction de répartition G. Généralement, les variables aléatoires C_i sont supposées être indépendantes des X_i. Soit $(T_i, \delta_i)_{i=1,\ldots,n}$ l'échantillon réellement observé, où

$$T_i = \min(X_i, C_i) \quad \text{et} \quad \delta_i = \mathbf{1}_{\{X_i \leq C_i\}} \tag{3.1}$$

δ_i l'indicatrice de non censure.

3.2.1 Estimateur de Kaplan-Meier

Cet estimateur (que l'on notera EKM) est aussi appelé estimateur Product Limit (PL) car il s'obtient comme limite d'un produit. L'ideé de la construction de l'EKM est la suivante, pour $t_1 < t$ la probabilité de survivre au-delà de l'instant t est égale à :

$$\begin{aligned} S(t) &= P(X > t) \\ &= \mathbb{P}(X > t, X > t_1) \\ &= \mathbb{P}(X > t / X > t_1)\, S(t_1) \end{aligned} \tag{3.2}$$

Si l'on renouvelle l'opération en choisissant une date t_2 antérieure à t_1 on aura de même

$$S(t_1) = \mathbb{P}(X > t_1 / X > t_2)\, S(t_2)$$

D'où

$$S(t) = \mathbb{P}(X > t / X > t_1)\, \mathbb{P}(X > t_1 / X > t_2)\, S(t_2) \tag{3.3}$$

Donc, à partir de (3.2) et (3.3) si on a $T_0 < T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)} < t$, on obtient

$$S(t) = \mathrm{P}\left(X > t / X > T_{(n)}\right) \mathrm{P}\left(X > T_{(n)} / X > T_{(n-1)}\right) \cdots \mathrm{P}\left(X > T_{(1)} / X > T_{(0)}\right) \mathrm{P}\left(X > T_{(0)}\right) \tag{3.4}$$

Si on choisit pour dates où l'on conditionne celles où il s'est produit un évènement (décès ou censure) i.e $T_{(i)}$, on estime seulement des quantité de la forme

$$p_i = \mathbb{P}\left(X > T_{(i)} / X > T_{(i-1)}\right) \tag{3.5}$$

Or p_i est la probabilité de survivre au-delà de l'intervalle de temps $I_i =]\, T_{(i-1)}, T_{(i)}\,]$ sachant qu'on ai vivant au début de l'intervalle.

Notons R_i le nombre des sujets qui sont vivants(donc "à risque" de mourir) juste avant l'instant $T_{(i)}$ et M_i le nombre de morts à l'instant $T_{(i)}$.
On pose $q_i = 1 - p_i$ = *la probabilité de mourir durant l'intervalle* I_i *sachant que l'individu était vivant au début de cet intervalle.* Alors l'estimateur naturel de q_i est

$$\widehat{q_i} = \frac{M_i}{R_i} = \frac{\text{nombre de mort observés à l'instant } T_{(i)}}{\text{nombre de sujets à risque}} \tag{3.6}$$

Supposons qu'il n'y ait pas d'ex-aequo (c-à-d tous les $T_{(i)}$ sont différents). Si $\delta_{(i)} = 1$ dans ce cas il n'y a pas de censure à l'instant $T_{(i)}$, implique $M_i = 1$. Et si $\delta_{(i)} = 0$ il ya censure à l'instant $T_{(i)}$, implique $M_i = 0$. On a alors

$$\widehat{q_i} = \begin{cases} \frac{1}{R_i} & \text{si} \quad \delta_{(i)} = 1 \\ 0 & \text{si} \quad \delta_{(i)} = 0 \end{cases} \tag{3.7}$$

$\Longrightarrow$

$$\widehat{p_i} = 1 - \widehat{q_i} = \begin{cases} 1 - \frac{1}{R_t} & \text{si} \quad \delta_{(i)} = 1 \\ 1 & \text{si} \quad \delta_{(i)} = 0 \end{cases}$$

$\Longrightarrow$

$$\widehat{p_i} = \left(1 - \frac{1}{R_i}\right)^{\delta_{(i)}} \tag{3.8}$$

Comme $R_i = n - i + 1$, (car il y'a eu "$i-1$" dècès ou censures avant $T_{(i)}$ et il y'a n individus dans l'étude). On obtient finalement l'EKM pour la fonction de survie de la variable durée de vie X :

$$\widehat{S}_{KM}(t) = 1 - \widehat{F}_{KM}(t) = \begin{cases} \prod\limits_{\substack{i=1 \\ T_i \le t}}^{n} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} & \text{si} \quad t < T_{(n)} \\ 0 & \text{si} \quad t \ge T_{(n)} \end{cases} \tag{3.9}$$

et donc on a aussi l'EKM pour la fonction de survie de la variable de censure C

$$\bar{G}_n(t) = 1 - \hat{G}_{KM}(t) = \begin{cases} \prod\limits_{\substack{i=1 \\ T_i \le t}}^{n} \left(1 - \frac{1}{n-i+1}\right)^{1-\delta_{(i)}} & \text{si} \quad t < T_{(n)} \\ 0 & \text{si} \quad t \ge T_{(n)} \end{cases} \tag{3.10}$$

où les $\delta_{(i)}$ sont les indicatrices de censure correspondantes.

Remarque. *l'estimateur de Kaplan Meier peut aussi se mettre sous la forme suivante*

$$\widehat{S}_{KM}(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right)^{\mathbf{1}_{\{T_{(i)} \leq t\}}} & si \quad t < T_{(n)} \\ 0 & si \quad t \geq T_{(n)} \end{cases} \tag{3.11}$$

et

$$\bar{G}_n(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1-\delta_{(i)}}{n-i+1}\right)^{\mathbf{1}_{\{T_{(i)} \leq t\}}} & si \quad t < T_{(n)} \\ 0 & si \quad t \geq T_{(n)} \end{cases} \tag{3.12}$$

On en déduit un estimateur non paramétrique de la fonction de répartition de X :

$$\begin{aligned} \hat{F}_n(t) &= \begin{cases} 1 - \prod_{\substack{i=1 \\ T_i \leq t}}^{n} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right) & \text{si} \quad t < T_{(n)} \\ 1 & \text{si} \quad t \geq T_{(n)} \end{cases} \\ &= \begin{cases} 1 - \prod_{\substack{i=1 \\ T_i \leq t}}^{n} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} & \text{si} \quad t < T_{(n)} \\ 1 & \text{si} \quad t \geq T_{(n)} \end{cases} \end{aligned} \tag{3.13}$$

Exemple. [1] *Sur 10 patients atteints de cancer des bronches, on a observé les durées de survie suivantes exprimées en mois :* $1 \quad 3 \quad 4^+ \quad 5 \quad 7^+ \quad 8 \quad 9 \quad 10^+ \quad 11 \quad 13^+$. *Les données suivies du signe + correspondent à des patients qui ont été perdus de vue à la date fournie ainsi que l'existence (statut=0) ou non (statut =1) d'une censure à droite.*

Patient i	1	2	3	4	5	6	7	8	9	10
Durées T_i	1	3	4	5	7	8	9	10	11	13
Statut M_i	1	1	0	1	0	1	1	0	1	0

Temps T_i	R_i	M_i	$\widehat{S}(T_i)$	Intervalle
0	0	0	1	$[0;1[$
1	10	1	$\left(1-\frac{1}{10}\right)\widehat{S}(0)=0.9$	$[1;3[$
3	9	1	$\left(1-\frac{1}{9}\right)\widehat{S}(1)=0.8$	$[3;4[$
4	8	0	$\left(1-\frac{0}{8}\right)\widehat{S}(3)=0.8$	$[4;5[$
5	7	1	$\left(1-\frac{1}{7}\right)\widehat{S}(4)=0.7$	$[5;7[$
7	6	0	$\left(1-\frac{0}{6}\right)\widehat{S}(5)=0.7$	$[7;8[$
8	5	1	$\left(1-\frac{1}{5}\right)\widehat{S}(7)=0.6$	$[8;9[$
9	4	1	$\left(1-\frac{1}{4}\right)\widehat{S}(8)=0.5$	$[9;10[$
10	3	0	$\left(1-\frac{0}{3}\right)\widehat{S}(9)=0.5$	$[10;11[$
11	2	1	$\left(1-\frac{1}{2}\right)\widehat{S}(10)=0.25$	$[11;13[$
13	1	0	$\left(1-\frac{0}{1}\right)\widehat{S}(11)=0.25$	$[13;\infty[$

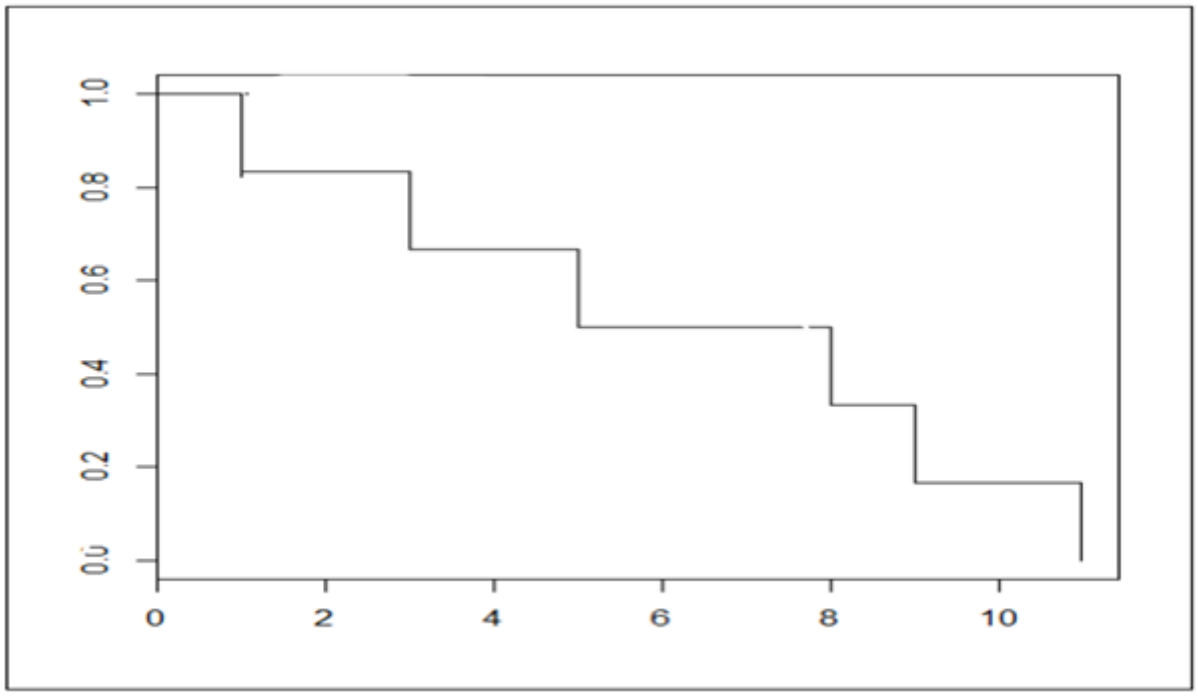


FIGURE 3.1 – Le graphe de la fonction de survie.

Remarque. *L'estimateur de Kaplan-Meier est une fonction en escalier.*

3.2.2 Sauts de l'estimateur de Kaplain Maier (EKM)

Comme $\hat{F}_n(t)$ est une fonction en escalier on devrait donc pouvoir l'écrire sous la forme

$$\hat{F}_n(t) = \sum_{i=1}^{n} W_{i,n} \mathbf{1}_{\{T_i \leq t\}} \tag{3.14}$$

où $W_{(i,n)}$ est un poids qui tient compte de la censure, et correspond à $\left(T_{(i)}, \delta_{(i)}\right)$ tel que

$$W_{(j,n)} = \frac{\delta_{(j)}}{n\bar{G}_n\left(T_{(j-1)}\right)} \Rightarrow \hat{F}_n\left(T_{(j)}\right) = \sum_{i=1}^{n} \frac{\delta_{(i)}}{n\bar{G}_n\left(T_{(i-1)}\right)} \mathbf{1}_{\{T_{(i)} \leq T_{(j)}\}}$$

où

$$\bar{G}_n(T_{(j-1)}) := 1 - G_n(T_{(j-1)})$$

et grâce à (3.14)

$$\hat{F}_n(t) = \sum_{i=1}^{n} \frac{\delta_{(i)}}{n\tilde{G}_n\left(T_{(i)}\right)} \mathbf{1}_{\{T_i \leq t\}} \tag{3.15}$$

Démonstration :voir (Pr. SADKI Ourida [43])

3.2.3 Propriétés de l'estimateur de Kaplain Meier

— L'estimateur de Kaplain Maier $\hat{F}_n(t)$ est asymptotiquement sans biais.

$$\text{Biais}(\hat{F}_n(t)) \xrightarrow[n\to\infty]{} 0 \tag{3.16}$$

Démonstration : voir (Pr. SADKI Ourida [43])

3.2.4 Estimateur de la densité pour des variables censurées

Soit K un noyau réel et h un réel positif. On définit l'estimateur à noyau de la densité dans le cas d'un modèle de survie censuré par :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{(i)}}{\bar{G}_n\left(X_i\right)} K\left(\frac{x - X_i}{h}\right) \tag{3.17}$$

3.2.5 Biais de l'estimateur de la densité

Sous l'hypothèse d'indépendance entre X et C

$$\begin{aligned}\mathbb{E}\left(\hat{f}_n(t)\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\bar{G}\left(X_i\right)}K\left(\frac{x-X_i}{h}\right)\right)\\ &= \mathbb{E}\left(\frac{\delta_1}{\bar{G}\left(X_1\right)}K\left(\frac{x-X_1}{h}\right)\right) = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_1\leq C_1\}}}{\bar{G}\left(X_1\right)}K\left(\frac{x-X_1}{h}\right)\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}_{\{X_1\leq C_1\}}}{\bar{G}\left(X_1\right)}K\left(\frac{x-X_1}{h}\right)|X\right)\right)\end{aligned}$$

$$\begin{aligned}\int_0^{+\infty}\mathbb{E}\left(\frac{\mathbf{1}_{\{u\leq C_1\}}}{\bar{G}(u)}K\left(\frac{x-u}{h}\right)\mid X_1=u\right)f_{X_1}(u)du &= \int_0^{+\infty}\frac{1}{\bar{G}(u)}E\left(I_{\{C_1\geq u\}}K\left(\frac{x-u}{h}\right)\mid X_1=u\right)f_{X_1}(u)du\\ &= \int_0^{+\infty}\frac{1}{\mathrm{P}\left(C_1\geq u\right)}\mathrm{P}\left(C_1\geq u\right)K\left(\frac{x-u}{h}\right)f_{X_1}(u)du\\ &= \int_0^{+\infty}K\left(\frac{x-u}{h}\right)f_{X_1}(u)du\\ &= \mathbb{E}(K(x))\end{aligned}$$

3.3 Estimation de la densité dans le cas de données tronquées à gauche

Soit $X_1,\ldots,X_N$, une suite de variables aléatoires réelles d'intérêt i.i.d., de f.d.r. commune F. Soit $T_1,\ldots,T_N$ une suite de variables aléatoires de troncature i.i.d de f.d.r. continue G. Les T_i sont aussi supposées être indépendantes des X_i. Ainsi la f.d.r. conjointe de X et T est

$$\begin{aligned}H(x,t) &= \mathbb{P}(X\leq x, T\leq t)\\ &= F(x)G(t)\end{aligned} \tag{3.18}$$

Dans le modèle de troncature gauche, la variable aléatoire d'interet X est interferée par une variable de troncature T, tel que les quantitées X et T sont observables seulement si $X \geq T$ tandis que rien n'est observé si $X < T$. On note $\{(X_i,T_i)\,;i=1,\ldots,n\}\,,(n\leq N)$ l'échantillon observé (i.e. $X_i \geq T_i$) parmi le N-échantillon d'origine.

Une conséquence de la troncature est que la taille vraiment observé n est une variable aléatoire distribuée selon la loi Binomiale de paramètre N et μ où μ est la probabilité de troncature définit par

$$\mu = \mathbb{P}(X\geq T)$$

Il est clair que si $\mu = 0$, aucune donnée ne peut être observée. Pour cela, nous supposons, dorénavant, que $\mu \neq 0$. Par la loi forte des grands nombres on a, lorsque N tend vers ∞

$$\hat{\mu}_n = \frac{n}{N} \to \mu, \mathbb{P}-p.s. \tag{3.19}$$

Sous ce modèle, les résultats ne seront pas établis par rapport à la probabilité $\mathbb{P}$ (relative au N-échantillon) mais par rapport à une nouvelle probabilité $\mathbf{P}$ (relative au n-échantillon) définit par

$$\mathbf{P}(.) = \mathbb{P}(. \mid X \geq T)$$

Nous noterons respectivement $\mathbb{E}$ et $\mathbf{E}$ les espérances relatives aux probabilitées P et $\mathbf{P}$. Dans la suite, nous dénotons par un exposant (*), toute caractéristique liée aux données observées (i.e. conditionnellement à la valeur de n).

3.3.1 Estimation des fonctions de répartitions

$$\begin{aligned} H^*(x,t) &= \mathbf{P}(X \leq X, T \leq t) \\ &= \mathbb{P}(X \leq x, T \leq t \mid X \geq T) \\ &= \frac{\mathbb{P}(X \leq x, T \leq t, X \geq T)}{\mathbb{P}(X \geq T)} \\ &= \frac{\mathbb{P}(X \leq x, T \leq t, T \leq x)}{\mathbb{P}(X \geq T)} \\ &= \frac{1}{\mu} \int_{-\infty}^{x} G(t \wedge u) dF(u). \end{aligned} \tag{3.20}$$

où $t \wedge u := \min(t,u)$. Les f.d.r. marginales respectives de X et T sont donc définies par

$$F^*(x) := H^*(x;\infty) \tag{3.21}$$

$$= \frac{1}{\mu} \int_{-\infty}^{x} G(u) dF(u) \tag{3.22}$$

$$\begin{aligned} G^*(t) &= H^*(\infty,t) \\ &= \frac{1}{\mu} \int_{-\infty}^{\infty} G(t \wedge u) dF(u) \\ &= \frac{1}{\mu} \int_{-\infty}^{\infty} \int_{-\infty}^{t \wedge u} dG(v) dF(u) \\ &= \frac{1}{\mu} \int_{v}^{\infty} \int_{-\infty}^{t} dG(v) dF(u) \\ &= \frac{1}{\mu} \int_{-\infty}^{t} dG(v) \int_{v}^{\infty} dF(u) \\ &= \frac{1}{\mu} \int_{-\infty}^{t} (1 - F(v)) dG(v), \end{aligned} \tag{3.23}$$

Qui peuvent être estimées respectivement par les estimateurs empiriques suivants :

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}} \tag{3.24}$$

$$G_n^*(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i \leq t\}} \tag{3.25}$$

Remarque. *F_n^* et G_n^* sont les estimateurs de F^* et G^* mais pas de F et G.*

Pour toute f.d.r. W, nous noterons par a_W et b_W, respectivement, les bornes inférieure et supérieur du support de W définies respectivement par

$$a_W = \inf\{u : W(u) > 0\} \text{ et } \quad b_W = \sup\{u : W(u) < 1\} \tag{3.26}$$

Afin d'assurer l'identifiabilité du modèle, Woodroofe (1985) [56] a proposé,

$$a_G \leq a_F, b_G \leq b_F \text{ et } \int_{a_F}^{\infty} \frac{dF}{G} < \infty \tag{3.27}$$

Cette dernière condition est requise pour le cas $a_G = a_F$. Sous les conditions (3.27),

$$\mu = \mathbb{P}(X \geq T) = \int_0^{\infty} G(u)dF(u) \tag{3.28}$$

Maintenant, pour définir les estimations de F et G, nous avons besoin d'introduire la fonction $C(.)$ définit pour un $x \in [a_F, \infty[$ par

$$\begin{aligned} C(x) &= \mathbf{P}(T \leqslant x \leqslant X) \\ &= G^*(x) - F^*(x) \\ &= \mathbb{P}(T \leqslant x \leqslant X \mid X \geqslant T) \qquad (3.29) \\ &= \frac{\mathbb{P}(T \leqslant x \leqslant X, X \geqslant T)}{\mathbb{P}(X \geqslant T)} \\ &= \frac{\mathbb{P}(T \leqslant x \leqslant X)}{\mathbb{P}(X \geqslant T)} \\ &= \frac{1}{\mu}\mathbb{P}(T \leqslant x)\mathbb{P}(X \geqslant x) \\ &= \frac{1}{\mu}\mathrm{G}(x)(1 - F(x)) \qquad (3.30) \end{aligned}$$

A partir de (3.24) et (3.25) $C(x)$ peut être estimée par

$$\begin{aligned} \hat{C}_n(x) &= G_n^*(x) - F_n^*(x) \qquad (3.31) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i \leqslant x \leqslant X_i\}} \end{aligned}$$

3.3.2 Estimation de fonction de hasard cumulative

En rappelant que la fonction de hasard cumultative d'une fonction de distribution F est définit par :

$$\begin{aligned} H(x) &= \int_0^x \Lambda(t)dt \qquad (3.32)\\ &= \int_0^x \frac{dF(t)}{1-F(t)}dt \quad 0 < x < \infty \\ &= \frac{\mu}{\mu}\int_{aF}^x \frac{G(t)dF(t)}{G(t)\bar{F}(t)}dt \quad 0 < x < \infty \end{aligned}$$

Où a_F la borne inférieure de F, $\bar{F} = 1 - F(t)$. On pose $C(t) = \frac{1}{\mu}G(t)\bar{F}(t)$ et $F^*(x) = \frac{1}{\mu}\int_{a_F}^x G(t)dF(t)$ et donc $dF^*(x) = \frac{1}{\mu}G(t)dF(t)$ alors

$$\begin{aligned} H(x) &= \frac{1}{\mu}\int_{a_F}^x \frac{G(t)dF(t)}{C(t)}dt \quad 0 < x < \infty \qquad (3.33)\\ &= \int_{aF}^x \frac{dF^*(t)}{C(t)} \qquad (3.34) \end{aligned}$$

Ceci suggère d'estimer H par

$$\hat{H}_n(x) = \int_{aF}^x \frac{dF_n^*(u)}{C_n(u)} = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{1}_{\{X_i \le x\}}}{C_n(X_i)} \qquad (3.35)$$

Où F_n^* et C_n sont les estimateurs empiriques de F^* et C déjà définis.

3.3.3 Estimation de la probabilité de troncature

Soit $\mu = P(X \ge T)$ la probabilité que la variable aléatoire X soit observable.
Si $\mu = 0$, alors les données ne sont pas observable, nous supposons que $\mu > 0$ un estimateur de

μ est donné par $\frac{n}{N}$, mais cet estimateur ne peut pas être calculé vu que N est inconnu.

$$\begin{aligned}
\mu &= \mathbb{P}(X \geq T) \\
&= \mathbb{E}\left(\mathbf{1}_{\{X \geq T\}}\right) \\
&= \mathbb{E}\left[\mathbb{E}\left(\mathbf{1}_{\{X \geq T\}} \mid X\right)\right] \\
&= \int_{a_F}^{+\infty} \mathbb{E}\left(\mathbf{1}_{\{X \geq T\}} \mid X = x\right) dF(x) \\
&= \int_{a_F}^{+\infty} \mathbb{P}(X \geq T \mid X = x) dF(x) \\
&= \int_{a_F}^{+\infty} \mathbb{P}(T \leq x) dF(x) \text{ car } X \text{ et } T \text{ sont independantes.} \\
&= \int_{a_F}^{+\infty} G(x) dF(x)
\end{aligned}$$

Keiding et Gill (1990) [31] suggèrent d'estimer μ par

$$\hat{\mu}_n = \int_{a_F}^{+\infty} \hat{G}_n(x) d\hat{F}_n(x) \tag{3.36}$$

Où $\hat{G}_n$ et $\hat{F}_n$ sont les estimateurs du maximum de vraisemblance de Lynden-Bell (1971) [39] de F et G respectivement.

He et Yang (1988) [26] proposent l'estimateur pour μ :

$$\tilde{\mu}_n = G_n(x)[1 - F_n(x)]C_n^{-1}(x). \tag{3.37}$$

3.3.4 Les estimateurs du maximum de vraisemblance de F et G (Lynden-Bell (1971)

Dans le cas absolument continu, l'estimateur de F s'obtient à partir de :

$$H(x) = \int_0^x h(t)dt$$

avec

$$\begin{aligned}
H_n(x) &= \int_0^x \frac{dF_n(t)}{1 - F_n(t)} \\
&= -\log\left(1 - F_n(x)\right)
\end{aligned}$$

d'où

$$\begin{aligned}\hat{F}_n(x) &= 1 - \exp\left[-H_n(x)\right] \\ &= 1 - \exp\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}_{\{X_i\leq x\}}}{C_n(X_i)}\right] \\ &= 1 - \prod_{i=1}^{n}\exp\left[-\frac{1}{n}\frac{\mathbf{1}_{\{X_i\leq x\}}}{C_n(X_i)}\right] \\ &= 1 - \prod_{\substack{i=1\\X_i\leq x}}^{n}\exp\left[-\frac{1}{nC_n(X_i)}\right]\end{aligned}$$

Puisque $\exp\left[-\frac{1}{nC_n(X_i)}\right] \simeq 1 - \frac{1}{nC_n(X_i)} = \frac{nC_n(X_i)-1}{nC_n(X_i)}$

D'où l'estimateur de Lynden-Bell (1971) de F :

$$\hat{F}_n(x) = 1 - \prod_{\substack{i=1\\X_i\leq x}}^{n}\left(1 - \frac{1}{nC_n(X_i)}\right) \tag{3.38}$$

En suivant les mêmes étapes l'estimation produit-limite $\hat{G}_n$ de G est

$$\hat{G}_n(x) = \prod_{\substack{i=1\\T_i>x}}^{n}\left(1 - \frac{1}{nC_n(T_i)}\right) \tag{3.39}$$

3.3.5 Propriétés asymptotiques de l'estimateur de Lynden-Bell

Woodroofe (1985) [56] établie les propriétés asymptotiques de l'estimateur de Lynden-Bell [39] ainsi que celles de l'estimateur similaire pour G. Stute (1993) [53] améliore les hypothèses et établi la normalité asymptotique de l'estimateur.

Corollaire 3.3.1. *Woodroofe (1985) [56]*
Si F et G sont continues telles que $a_G \leq a_F, b_G \leq b_F$ et $\int_{a_F}^{+\infty}\frac{1}{G}dF < \infty$ alors

$$\begin{aligned}&\sup_{x>a_F}\left|\hat{F}_n(x) - F(x)\right| \to 0, \\ &\sup_{t>a_G}\left|\hat{G}_n(t) - G(t)\right| \to 0\end{aligned} \tag{3.40}$$

en $\mathbf{P}$-probabilité quand $n \to \infty$ (où a_G, b_G et a_F, b_F désignent les points finaux de G et F respectivement.)

Démonstration :voir (Woodroofe [56] (1985)).

3.3.6 Estimateur à noyau de la densité pour les données tronquées à gauche

Soit K un noyau réel et h une suite qui tend vers 0. On définit l'estimateur à noyau de la densité dans le cas d'un modèle de troncature gauche par

$$\hat{f}_n(x) = \frac{\hat{\mu}_n}{nh} \sum_{i=1}^{n} \frac{1}{\hat{G}_n(X_i)} K\left(\frac{x - X_i}{h}\right) \tag{3.41}$$

avec $\hat{\mu}_n$ l'estimateur de He et Yang (1998) de μ la probabilité de non-troncature et $\hat{G}_n$ l'estimateur de Lynden Belle de G.

3.3.6.1 Propriétés de l'estimateur de la densité

Théorème 3.3.1. *Si*

i) Le noyau K est une densité de probabilité sous les conditions :

$$\int_{-\infty}^{+\infty} uK(u)du = 0, \mu_2(K) = \int_{-\infty}^{+\infty} u^2K(u) < \infty, \quad R(K) = \int_{-\infty}^{+\infty} K(u)^2du < \infty.$$

ii) Les fonctions f et G^{-1} sont deux fois continument différentiables autour de x.
alors

$$\text{Var}\left[\hat{f}_n(x)\right] = (nh)^{-1}\mu G(x)^{-1} f(x) R(K) + o\left((nh)^{-1}\right).$$

Démonstration : Voir (Carla Moreira et Jacobo de Uña-Álvarez 2012 [9])

L'erreur quadratique asymtotique moyenne est :

$$AMISE\left(\hat{f}_n(x)\right) = \frac{1}{4}h^4 R\left(f''\right)\mu_2(K)^2 + (nh)^{-1}\mu R(K) \int G^{-1}(x) f(x)$$

Démonstration : Voir (Carla Moreira et Jacobo de Uña-Álvarez 2012 [9])

Chapitre 4

Application

4.1 Introduction

Dans ce chapitre, notre objectif est de comparer les estimateurs à noyau de la densité inconnue f sur la base d'un échantillon gaussien, dans le cas des données complètes et le cas des données incomplètes tronquées à gauche, et comparer par la suite ces deux estimateurs à la densité normale centrée réduite. On va aussi calculer l'erreur quadratique moyenne MSE dans les deux cas et pour différentes tailles de l'échantillon. On termine par une étude comparative sur les résultats obtenus.
Pour cela nous avons effectué des simulations à l'aide du langage Matlab **R**2012**a**.

4.2 Plan de simulation

Soit T la variable de troncature, et X une variable aléatoire de loi normale centrée réduite. Avant la troncature, la taille initiale de l'échantillon était N.

— On simule, dans un premier temps, N valeurs $(X_i, T_i), i = 1, .., N$ du couple de variables aléatoires (X, T) avec T indépendante de X.

— On introduit dans le programme la contrainte de troncature gauche

$$X_i \geq T_i \tag{4.1}$$

— On ne tient compte que des couples vérifiant cette contrainte.

— Cette procédure est répétés pour différentes valeurs $N = 100, 200, 300$ jusqu'à obtenir l'échantillon final avec la taille n. $\{(X_i, T_i), 1 \leq i \leq n\}$ avec $n \leq N$, $T_i = \min(X_i, C_i)$. Cet échantillon représente théoriquement l'échantillon observé.

— On introduit l'estimateur à noyau de la densité de la variable X dans le cas des données

complètes définit par :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{4.2}$$

Dans le cas des données tronquées à gauche, il est définit comme suit

$$\tilde{f}_n(x) = \frac{\hat{\mu}_n}{nh}\sum_{i=1}^{n} \frac{1}{\hat{G}_n(X_i)} K\left(\frac{x - X_i}{h}\right) \tag{4.3}$$

avec $\hat{\mu}_n$ l'estimateur de He et Yang (1998) de μ la probabilité de non-troncature.
Où l'estimateur de maximum de vraisemblance de G ((Lynden-Bell) est définit par :

$$\hat{G}_n(x) = \prod_{\substack{i=1 \\ T_i > x}}^{n} \left(1 - \frac{1}{n\hat{C}_n(T_i)}\right) \tag{4.4}$$

et l'estimateur empirique de $C(x)$ est défini par :

$$\hat{C}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i \le x \le X_i\}} \tag{4.5}$$

Basant sur un noyau gaussien, nous calculons pour $\hat{f}_n$ estimateur de f, l'erreur quadratique moyenne MSE. Elle est défini comme suit

$$\text{MSE} = \frac{1}{n}\sum_{k=1}^{n} \left[\hat{f}_n(X_k) - f(X_k)\right]^2 \tag{4.6}$$

On a utilisé le paramètre de lissage optimal pour la loi Normale centrée réduite f dans le cas complet ; $h = 1.06N^{-1/5}$ et le paramètre de lissage $h_1 = \left(\frac{\log n}{n}\right)^{\frac{1}{5}}$ dans le cas tronqué, h_1 satisfait la convergence de l'estimateur à noyau dans le cas tronqué.

4.3 Résultats de simulation

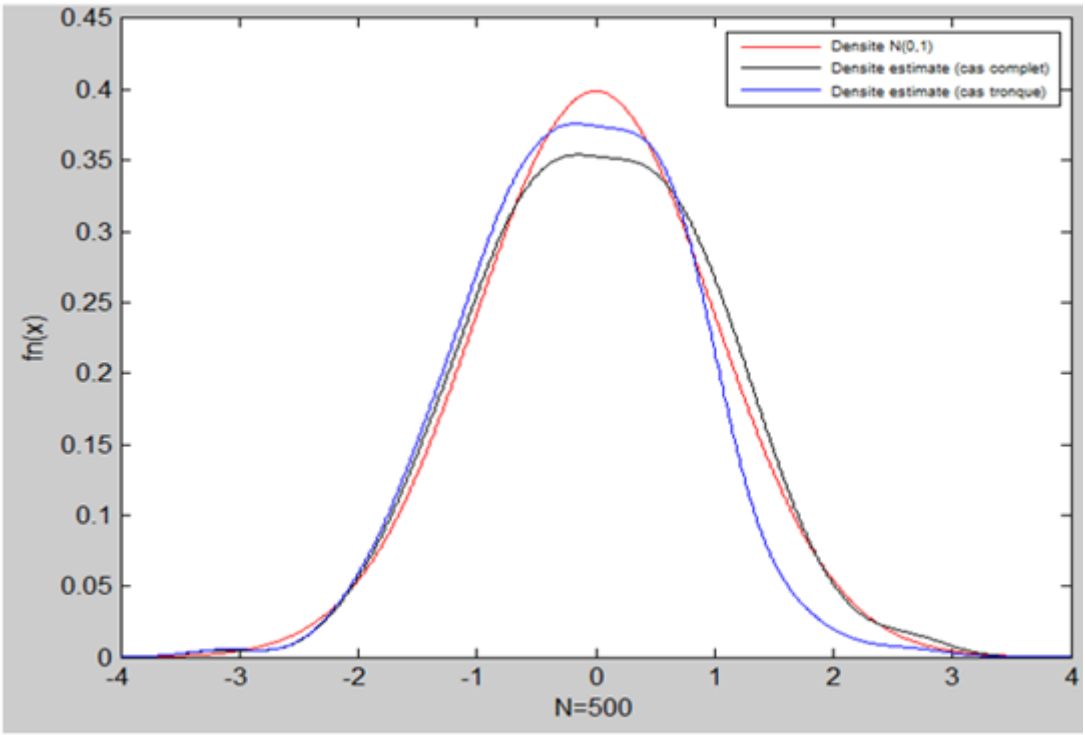
Le tableau suivant représente l'erreur quadratique moyenne MSE de l'estimateur à noyau de la densité dans le cas des données complètes et le cas des données tronquées à gauche.

N	n	μ	h	h_1	MSE(données complètes)	MSE(données tronquées)
100	68	0.68	0.42	0.57	3.56×10^{-3}	2.56×10^{-2}
300	225	0.75	0.34	0.47	2.43×10^{-3}	1.86×10^{-2}
500	300	0.6	0.31	0.45	2.65×10^{-2}	2.86×10^{-2}

TABLE 4.1 – Résultats de la simulation

4.4 Interprétation des résultats

A travers, les résultats obtenus représentés graphiquement par les figures suivantes et le calcul du MSE, on remarque que l'estimateur à noyau de la densité f dans les deux cas des données complètes et données tronquées à gauche est proche de la densité exacte $N(0,1)$. On dit alors qu'on a une bonne approximation pour la loi normale.



N=500 et μ=0.6

FIGURE 4.1 – Courbe pour $N = 500$ et $\mu = 0.6$

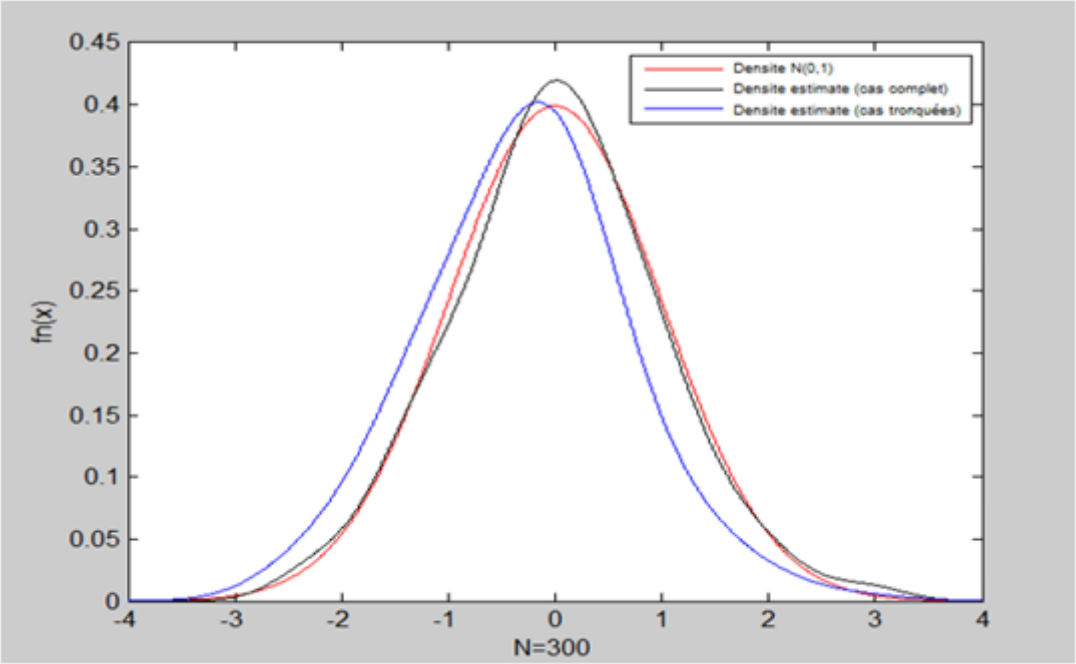


FIGURE 4.2 – Courbe pour $N = 300$ et $\mu = 0.75$

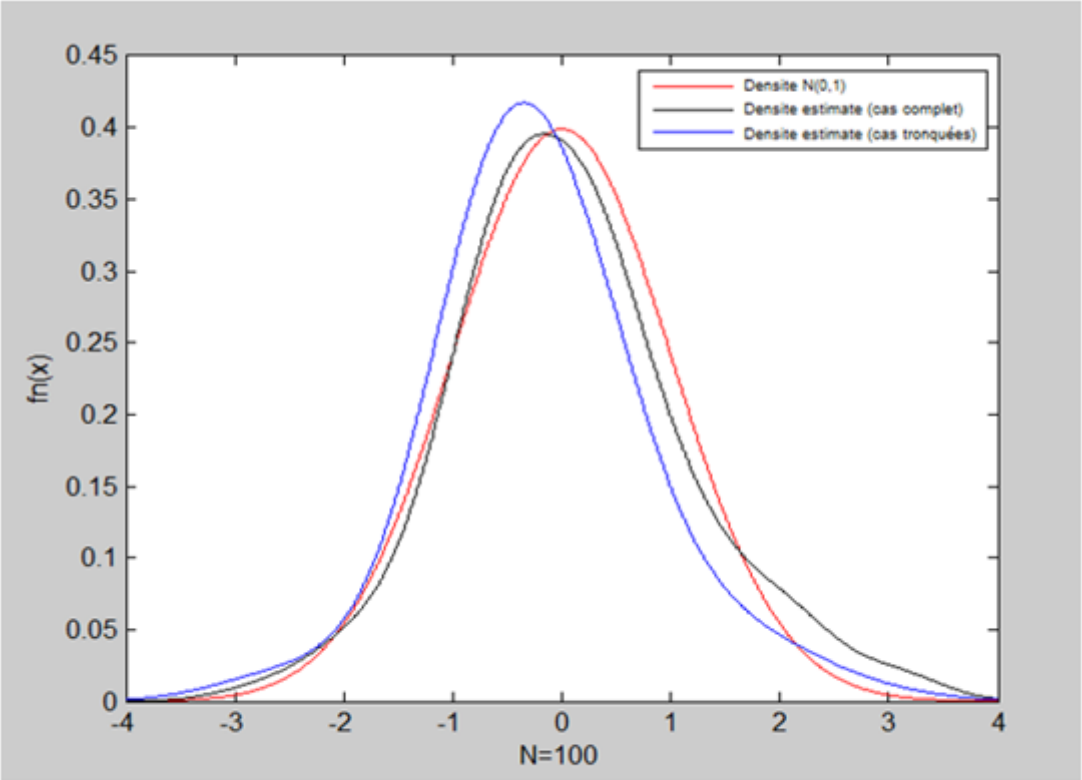


FIGURE 4.3 – Courbe pour $N = 100$ et $\mu = 0.68$

Conclusion

Dans ce mémoire nous avons fait une synthèse des résultats existant concernant l'estimation non paramétrique de la densité inconnue f avec la méthode du noyau sur des données complètes et incomplètes.

Dans un premier temps, nous avons présenté l'estimateur à noyau dans le cas de données complètes, et nous avons donné ses propriétés statistiques et asymptotiques. Nous avons aussi cité quelques méthodes de sélection du paramètre de lissage h. Par la suite nous avons abordé l'estimation non paramétrique de la densité inconnue f avec des données incomplètes en utilisant la méthode du noyau.

Des estimateurs non paramétriques à noyau dans le cas de données incomplètes ont été cités, ainsi que leurs propriétés. Enfin une étude de simulation est conçue pour confirmer les résultats théoriques présentés dans ce mémoire. Pour différentes tailles d'échantillon nous avons représenté graphiquement les courbes des densités (théorique et estimée). Cependant, en raison de contrainte de temps nous nous somme limité uniquement aux cas de données incomplètes tronqués à gauche. Les résultats obtenus montrent que presque pour toutes les tailles de l'échantillon, l'erreur MSE dans les deux cas est petite. Cette constatation est confirmée graphiquement.

Perspectives

Parmi les perspectives de ce travail, nous pouvons dégager des points intéressants :

— Il serait intéressant d'appliquer cette étude sur des données réelles.

— Appliquer la méthode du noyau dans l'autre cas de données incomplètes censurées, dont on n'a pas eu suffisamment de temps pour l'effectuer.

— Il est aussi intéressant d'appliquer cette méthode dans le cas où les données présentent une forme de dépendance.

— Et surtout utiliser la méthode du noyau pour des densités multidimensionnelles.

Bibliographie

[1] Abdelaziz, S., Sissaoui, A. Estimation pour des données censurées. Mémoire de master, Université Mohamed Seddik Ben Yahai Jijel, 2020.

[2] Anderson, P.K, Borgan, Gill R.D et Keiding, N. Statistical model based on counting processesses. New-York : Springer-Velrag, (1993).

[3] Belkacem, A. Etude d'une classe d'estimateurs à noyau de la densité d'une loi de probabilité. Thèse pour l'obtention du grade de (Ph. D).Ecole des Gradués de l'université Laval, 1989.

[4] Belahcene, I. Estimation non paramétrique de la fonction densité de probabilité avec un noyau. Mémoire de Master, Université Kasdi Merbah Ouargla, 2016/2017, pp. 14-16.

[5] Bosq, D. et Lecoutre, J. P., Théorie de l'Estimation Fonctionnelle. Economica, Paris, (1987).

[6] Boudada, H. Quantile conditionnel pour des données incomplètes et dépendantes. Mémoire de magister, Université de Constantine, 2012.

[7] Bouezmarni, T. and Rolin, J-M. Consistency of the beta kernel density function estimator. Canad. J. Statist. 2003, 31, No. 1, 89-98.

[8] Bowman, A. W. An alternative method of crooss-validation for the smoothing density estimates estimator : Biometrika, 1984, Vol. 71, pp. 353-360.

[9] Carla, M et Jacobo, U-A. Kernel density estimation with doubly truncated data. University of Vigo, Spain, 2012.

[10] Catherine, H-C. Durées de survie tronquées et censurées Journal de la société statistique de Paris, 1994, tome 135, no 4, pp. 3-23

[11] Chaib, Y. Estimation de la fonction mode pour des données tronquées et censurées. Thèse de doctorat, Université de Baji Mokhtar, 2013.

[12] Chen, S. X. Beta kernel estimators for density functions. Comput Statist. Data Anal, 1999, 31, pp.131-145.

[13] Cherfaoui, M. Bootstrap dans l'estimation non paramétrique de la densité de probabilité et la courbe de regression de la moyenne. Mémoire de magister, Université de Béjaia, 2009.

[14] Com-nougué, C., Hill, C., Kramar, C. et Moreau, T. Analyse statistique des données de survie. Statistique en Biologie et en Medecine. Médecine Sciences Publications, 1999, ISBN : 2257123107.

[15] Devroye, L. The equivalence of weak, strong and complete convergence in L^1 for kernel density estimates. The Annals of Statistics, 1983, 11, pp. 896 – 904.

[16] Diehl, S. and Stute, W. Kernel Density and Hazard Function Estimation in the Presence of Censoring. Journal of Multivariate Analysis, 1988, Vol. 25, pp. 299-310.

[17] Djeladj, W. Comportement asymptotique d'un estimateur à noyau du quantile pour des données censurées et associées. Thèse de doctorat, USTHB d'Alger, Algérie 2019.

[18] Duin, R. P. W. On the choice of smoothing parameters of Parzen estimators of probability density function IEEE Transactions on Computers, 1976, $C-25$, ISBN-11751179.

[19] Epanechnikov, A. A. Nonparametric estimation of a multidimensional probability density. Theory Probab. Appl, 1969, 14, pp.153-158.

[20] Fix, E. et Hodges, J. R. Discriminatory analysis, nonparametric discrimination : consistency proprieties. Technical report, Report N°4. USAF School of aviation Medicine, Randolph Field, Texas, 1951.

[21] Francial Giscard Baudin Libengué, D. K. Méthode non paramétrique des noyaux associés mixtes et application. THÈSE EN CO-TUTELLE , Université de Franche-Comté, 2013.

[22] Guessoum, Z. Regression non paramétrique dans les modèle censurés. Thèse de doctorat, USTHB d'Alger, 2009.

[23] Geffroy. Sur l'estimation de la densité dans un espace metrique. C.R. Acad. Sci. Paris Sér A-B, 1973, 278, pp. 1449-1452.

[24] Ghettab, S. Estimation non paramétrique par la méthode du noyau : Applications à des données censurées. Mémoire de Magister, USTHB d'Alger, 2014.

[25] Hamrani, F. Estimation non paramétrique pour les données incomplètes, Thèse de doctorat.USTHB d'Alger, 2017.

[26] He, S. and Yang, G. Estimation of the truncation probability in the random truncation model, The Annals of Statistics, 1998, 26, pp. 1011 - 1027 .

[27] Habbema, J. D. F., Hermans, J., and Van den Broek, K. A stepwise discrimination analysis program using density estimation. Compstat : Proceedings in Computational Statistics. Physica Verlag, Vienna. 1974.

[28] Imen, B. K. Estimation non-paramétrique par noyaux associés et données de panel en marketing. Projet de Fin d'Etude, Université du 7 Novembre à Carthage, 2007 – 2008.

[29] Kaplan, E. L., Meier, R. Non parametric estimation from incomplete observations, journal of the american statistical association, 1958, pp. 457-481.

[30] Keiding, N. and Gill, R. D. Random truncation models and Markov processes, Ann. Statist, 1990, 18, pp. 582 – 602.

[31] Klein, JP. et Moeschberger, ML., Survival analysis : techniques for censored and truncated data. Soringer-Verlag. New York, 1997.

[32] Kolmogorov, A. N., & Castelnuovo, G. Sur la loi des grands nombres. G. Bardi, tip. della R. Accad. dei Lincei, 1929.

[33] Leiderman, P. H., Babu, D., Kagia, J., Kraemer, H. C. et Leiderman, G. F. Afriean infant precocity and some social influences during the first year. Nature, 1973, pp. 242, 247 – 249.

[34] L. Aouicha. Estimation non paramétrique du mode conditionnel dans un modèle de censure. Thèse de doctorat, université de Constantine, 2019.

[35] Lecoutre, J. Contribution à l'estimation non paramétrique de la régression. PhD thesis, Université de Pierre et Marie Curie-ParisVI-France, 1982.

[36] Lejeune, M. Statistique : La therorie et ses applications, Springer, France, 2004.

[37] Lejeune, M. Estimation non-paramétrique par noyaux : régression polynomiale mobile Revue de statistique appliquée, , tome 33, n°3, 1985, pp. 43-67

[38] Loquin, K. et Strauss, O. On the granularity of summative kernels. Fuzzy sets and systems, Thèse pour obtenir le grade de Docteur, 2008, 159, pp. 1952-1972.

[39] Lynden-Bell, D. A method of allowing for known observational selection in small samples applied to 3 CR quasars, Monthly Notices Royal Astronomy Society, 1971, 155, pp. 95−118.

[40] Meddour, K. Estimation non paramétrique des données tronquées par les polynômes locaux. Mémoire de magister, USTHB d'Alger, 2014.

[41] Nadaraya, E. A. On nonparametric estimation of density function and regression. Theory. Probab. Appl, 1965, 10, pp 186 – 190, 1965.

[42] Naouel, S. Modèle de survie estimation et Aplications. Mémoire de Master, Université Abou Bekr Belkaid , Tlemcen, 2021.

[43] Ourida, S. Estimation non paramétrique pour des données incomplètes : Workshop sur les interactions entre Equations Différentielles et Probabilités Statistique. Université de Bejaia, 2021.

[44] Parzen, E. On estimation of a probability density function and mode. The Anals of Mathematical Statistics. 33, 1962, pp. 1065 – 1076.

[45] Patrick, B. Convergence of probability measures. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley and Sons, Inc., New York, second edition, A Wiley-Interscience, Publication, 1999.

[46] Pearson, K. On the systematic fitting of curves to observations and measurements. Biometrica, 1, pp.265-303. 2, pp.1-23.

[47] Philippe Saint, P. Introduction à l'analyse des durées de survie, Université Pierre et Marie Curie, Févier 2015.

[48] Rosenblatt, M. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 1956, 27, pp, 832-837.

[49] Rousselière, D., Joly, I. À propos de la capacité à survivre des coopératives : une étude de la relation entre âge et mortalité des organisations coopératives agricoles françaises. Revue d'études en Agriculture et Environnement, 2011, Vol. 92, N°3, pp. 259-289.

[50] Rudemo, M. Empirical choice of histogram and kernel density estimators. Scandinavian Journal of Statistics, 1982, Vol. 9, N°2, pp. 65-78.

[51] Silverman, B. W. Weak and strong uniform consistency of the kernel estimate of density function and its derivates. Ann. Statist, 1978, 6, pp. 177 – 184.

[52] Simonoff, J. S. Smoothing Methods in Statistics. Springer-Verlag, 1996.

[53] Stute, W. Almost sure representations of the product-limit estimator for truncated data. Ann Stat, 1993, 21, pp. 146 – 156.

[54] T. Sybakov, A. B. Introduction à l'Estimation Non Paramétrique. Springer, New York, 2004.

[55] Tiago, de Oliveira, J. Estatistica de densidades. Resultados Assintoticos, Rev. Fac. Ci. Université. Lisboa, 1963, Ser A, 9, pp. 111-206.

[56] Woodroofe, M. Estimating a distribution function with truncated data. Ann Stat, 1985, Vol. 13, N°. 1, pp. 163-177.

[57] Zhou, Y. A note on the TJW product-limit estimator for truncated and censored data, Statis, Probab. Lett., 1996 26, pp. 381 – 387.

Résumé

L'objectif principal de ce travail est l'estimation non paramétrique de la densité de probabilité inconnue f par la méthode du noyau, sur des données complètes et incomplètes.
Dans un premier temps, nous avons présenté la méthode du noyau pour l'estimation de la densité de probabilité d'une variable aléatoire X dans le cas de données complètes. Nous avons donné les propriétés de l'estimateur tel que : le biais, la variance, les critères d'erreur MSE et MISE..., et quelques méthodes pour le choix du paramètre de lissage h. Nous nous sommes ensuite intéressés à l'estimation de la densité avec des données incomplètes en utilisant sur la méthode du noyau.
Enfin une étude de simulation est conçue pour comparer les estimateurs à noyau de la densité f, dans le cas des données complète et le cas des données tronqué à gauche, et comparer par la suite ces deux estimateurs à la densité normale centrée réduite.

Mots clés : Densité, Estimateur à noyau de Parzen-Rosenblatt, noyau, paramètre de lissage, données incomplètes, données tronquées, données censurées.

Abstract

The main objective of this work is the nonparametric estimation of the unknown probability density f by the kernel method, on complete and incomplete data.
First, we presented the kernel method for estimating the probability density of a random variable X in case of complete data. We studied the properties of the estimator such as : the bias, the variance, the error criteria MSE, MISE..., and some methods for the choice of the smoothing parameter h. We then focused on estimating the density with incomplete data, emphasizing the kernel method.
Finally, a simulation study is designed to compare the kernel estimators of the density f, in the case of the full data and the case of the left-truncated data, and then compare these two estimators to the reduced centered normal density.