RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ FRÈRES MENTOURI CONSTANTINE
FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT DE MATHÉMATIQUES

POLYCOPIÉ DE COURS
MASTER 2 DE PROBABILITÉS-STATISTIQUE

Statistique non paramétrique
Pour des données censurées

Fatiha MESSACI BELALOUI

Année 2016/2017

Contents

Les étudiants de Master 2 de l'option Probabilité-Statistique de l'université frères Mentouri de Constantine, à qui s'adresse ce cours, ont déjà étudié l'inférence statistique dans un cadre paramétrique, c'est à dire lorsque la forme générale de la loi régissant le phénomène étudié est connue et que l'estimation ou le test porte alors sur les paramètres, en nombre fini, de la loi. Mais, ceci est loin d'etre possible dans tous les cas pratiques, nous devons alors inférer dans un cadre non paramétrique. De plus, toute étude statistique se base sur l'observation d'échantillons, qui sont dans le meilleur des cas, de véritables réalisations des phénomènes d'intérêt. Cependant, bien souvent des phénomènes de censure interviennent d'une manière fortuite (accident, migration...) et (ou) planifiée (fixation du temps de l'étude...).
Ce cours se veut donc une initiation à la statistique non paramétrique pour des donnés réelles indépendantes et censurées. Afin de mettre en évidence, d'une manière intuitive et naturelle, le passage de l'étude du cas des données complètes au données censurées, un premier chapitre est consacré au premier cas mais sans trop s'y attarder. Nous nous intéressons à l'estimation des fonctions de répartition, de densité et de régression. La première caractérise complètement la loi de toute variable aléatoire et permet, entre autres, de calculer la probabilité d'appartenance à un intervalle. Néanmoins, c'est la seconde qui permet de mieux visualiser la forme de la loi en donnant des éléments concernant la symétrie, l'aplatissement ou la multimodalité, de plus elle intervient dans l'estimation du taux de hasard qui est une caractéristique fort recherchée en fiabilité. Quant à la fonction de régression, elle permet d'étudier le lien entre deux variables à des fins de prévision. Si c'est sur la méthode à noyaux, qui reste largement utilisée, que nous nous sommes appuyés, l'attention du lecteur a été attirée sur le fait que bien d'autres méthodes existent et des références ont été données. Par ailleurs, c'est pour des raisons historiques (c'est la première censure étudiée) et d'ordre pratique (c'est la censure la plus courante), que la censure à droite a été la plus détaillée. La seule censure à gauche peut se traiter d'une manière symétrique (cf Boukeloua (2013) et les exercices 3.5, 3.6 et 5.3). Par contre, les choses se compliquent quand les deux censures s'invitent dans un même échantillon. Le temps alloué à ce cours permet juste d'initier l'étudiant à ces cas là et de lui citer des travaux de base en éspérant avoir mis à sa disposition des éléments permettant d'éveiller l'envie et l'intérêt chez certains de mener une recherche dans ce domaine.

Fatiha MESSACI BELALOUI

Mes vifs remerciements vont à mes collègues les professeurs N. NEMOUCHI et O. SADKI d'avoir accepté d'expertiser ce document et à mon doctorant M. BOUKELOUA de l'avoir lu attentivement. Ils ont permis l'amélioration substantielle de la version qui leur a été soumise.

Chapter 1

Estimation non paramétrique pour des données complètes

La théorie de l'estimation est une préoccupation majeure de la statistique. Cette théorie est divisée en deux volets principaux, à savoir l'estimation paramétrique et l'estimation non-paramétrique. Cette dernière consiste, généralement, à estimer à partir des observations, une fonction inconnue, élément d'une certaine classe fonctionnelle. Une procédure non-paramétrique est définie indépendamment de la loi de l'échantillon d'observation. En fait, on parle de méthode d'estimation non-paramétrique lorsque celle-ci ne se ramène pas à l'estimation d'un paramètre appartenant à un espace de dimension finie. Plus généralement un modèle de statistique semi paramétrique comporte à la fois une composante paramétrique et une autre non paramétrique (exemple : le modèle de Cox). Un des problèmes centraux en statistique est celui de l'estimation de caractéristiques fonctionnelles associées à la loi des observations, comme par exemple, la fonction de répartition, la densité ou la fonction de régression. Dans le modéle de régression non-paramétrique, on suppose l'existence d'une fonction $r(x)$ qui exprime la valeur moyenne de la variable réponse Y en fonction de la variable d'entrée X.

Par ailleurs des données sont dites complètes si elles correspondent à de véritables réalisations de la variable d'intérêt. Le but de ce chapitre est d'introduire des estimateurs non paramétriques des fonctions de répartition, de densité et de régression, basés sur des données complètes, et de donner quelques unes de leurs propriétés. Ces dernières s'énoncent par un ou plusieurs des modes de convergence, que nous rappelons ci dessous.

Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, (X_n) une suite de variables aléatoires

réelles définies sur Ω et X une variable aléatoires réelle définie sur Ω.

1.1 Modes de convergence

1.1.1 Convergence en loi

(X_n) converge en loi vers X si la suite des fonctions de répartition des variables X_n converge vers la fonction de répartition de X, en tout point de continuité de cette dernière. Ceci est équivalent, grâce au théorème de Paul Lévy, à la convergence des suites des fonctions caractéristiques des variables X_n vers la fonction caractéristique de X, en tout point.
On note $X_n \xrightarrow{\mathcal{L}} X$.

1.1.2 Convergence en probabilité

(X_n) converge en probabilité vers X si

$$\forall \delta > 0, \lim_{n\to\infty} P(|X_n - X| > \delta) = 0.$$

On note $X_n \xrightarrow{\mathcal{P}} X$. Elle correspond à la convergence en mesure.

1.1.3 Convergence en moyenne quadratique

Si $X_n \in L^2$ et $X \in L^2$, on dit que X_n converge en moyenne quadratique vers X si

$$E\,|X_n - X|^2 = \int |X_n - X|^2\, dP \underset{n\to\infty}{\longrightarrow} 0.$$

On note $X_n \xrightarrow{m.q.} X$. Elle correspond à la convergence dans L^2.

1.1.4 Convergence presque sûre

(X_n) converge vers X presque sûrement s'il existe un ensemble P négligeable N tel que

$$\forall \omega \in N^c, |X_n(\omega) - X(\omega)| \underset{n\to\infty}{\longrightarrow} 0.$$

On note $X_n \xrightarrow{p.s.} X$. Elle correspond à la convergence P presque partout.

1.1.5 Convergence presque complète

La suite de v.a.r. $(X_n)_{n\in\mathbb{N}}$ *converge presque complètement vers la v.a.r.* X *si*

$$\forall \varepsilon > 0, \quad \sum_{n\in\mathbb{N}} P(|X_n - X| > \varepsilon) < \infty.$$

On note $X_n \xrightarrow{p.co.} X$.
Il est connu que ces différents modes de convergence sont liés par les relations suivantes.

$X_n \xrightarrow{p.co.} X \Rightarrow X_n \xrightarrow{p.s.} X \Rightarrow X_n \xrightarrow{\mathcal{P}} X \Rightarrow X_n \xrightarrow{\mathcal{L}} X$ et
$X_n \xrightarrow{m.q.} X \Rightarrow X_n \xrightarrow{\mathcal{P}} X$.
Soulignons le fait que toutes les autres implications ne sont pas vraies et qu'en particulier la convergence presque complète n'implique pas la convergence en moyenne quadratique qui, à son tour, n'entraîne pas la convergence presque complète (voir exercice 1.3).

1.2 Estimation de la fonction de répartition

La fonction de répartition caractérise complètement la loi d'une variable aléatoire X et permet, par exemple, de calculer la probabilité que X appartienne à un intervalle. Son estimation est fondamentale en statistique. Soient donc $X_1, ..., X_n$ des variables aléatoires réelles indépendantes et identiquement distribuées, de même loi que X, de fonction de répartition inconnue F. Une estimation naturelle de F se fait par la fonction de répartition empirique donnée par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}},$$

où $1_{\{.\}}$ désigne la fonction indicatrice.
F_n est elle même une fonction de répartition d'une loi discrète.
D'aprés la loi forte des grands nombres, on a pour tout $x \in \mathbb{R}$

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x). \tag{1.1}$$

Donc, F_n est un estimateur fortement consistant de F. De plus, la convergence presque sûre est uniforme en $x \in \mathbb{R}$, d'après le résultat suivant.

Théorème 1 *(Glivenko-Cantelli)*

$$P\left[\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| = 0\right] = 1.$$

Preuve. En tenant compte de la relation (1.1), elle découle immédiatement de l'exercice 1.2. ■

Le taux de convergence est précisé par la loi du logarithme itéré (LIL) qui est donné dans Kiefer (1961) comme suit

$$P\left(\limsup_{n\to\infty}\frac{\sqrt{n}\sup_x|F_n(x) - F(x)|}{\sqrt{\frac{1}{2}\log\log n}} \leq 1\right) = 1.$$

Il est facile de voir que, pour tout x, $F_n(x)$ converge en moyenne quadratique vers $F(x)$ et que $F_n(x)$ est asymptotiquement normal par application du théorème central limite.

1.3 Estimation de la fonction de densité

Supposons que la loi de X est absolument continue, par rapport à la mesure de Lebesgue, de densité de probabilité f_X, que nous cherchons à estimer. Ce problème tire son intérêt du fait qu'on peut mieux visualiser la loi (que par utilisation de la fonction de répartition) en déduisant de l'estimateur des élements concernant la symétrie, la multimodalité ou l'aplatissement de la loi étudiée. De plus cela constitue une aide précieuse au choix d'un modèle de probabilité et permet d'introduire un estimateur non paramétrique de la régression, sur lequel nous revenons à la section suivante.

Une solution intuitive a été proposée par Rosenblatt (1956) comme suit. Pour f_X continue et $h > 0$ assez petit, nous avons

$$f_X(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

En remplaçant F par l'estimateur de la fonction de répartition F_n, nous obtenons l'estimateur suivant de f_X:

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h},$$

qui peut aussi s'écrire sous la forme :

$$f_n(x) = \frac{1}{2nh}\sum_{i=1}^{n} 1_{\{x-h\leq X_i\leq x+h\}} = \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

où $K_0(u) = \frac{1}{2}1_{\{|u|\leq 1\}}$.
Rosenblatt (1956) puis Parzen (1962) ont suggéré une généralisation de cet estimateur, en posant :

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

où $K : \mathbb{R} \to \mathbb{R}$ est une fonction intégrable, telle que $\int K(u)\,du = 1$, c'est **l'estimateur à noyau** de la densité ou **l'estimateur de Parzen-Rosenblatt.**
La fonction K est le noyau (kernel en anglais) et le paramètre h est la fenêtre (bandwidth en anglais).
Voici quelques exemples de noyaux usuels.

- ♠ $K(u) = \frac{1}{2}1_{\{|u|\leq 1\}}$: noyau rectangulaire,
- ♠ $K(u) = (1-|u|)\,1_{\{|u|\leq 1\}}$: noyau triangulaire,
- ♠ $K(u) = \frac{3}{4}(1-u^2)\,1_{\{|u|\leq 1\}}$: noyau d'Epachnikov ou parabolique,
- ♠ $K(u) = \frac{15}{16}(1-u^2)^2\,1_{\{|u|\leq 1\}}$: noyau quadratique,
- ♠ $K(u) = \frac{1}{\sqrt{2\Pi}}\exp(-u^2/2)$: noyau gaussien,
- ♠ $K(u) = \frac{1}{\sqrt{2\Pi}}\exp\left(-|u|/\sqrt{2}\right)\sin\left(|u|/\sqrt{2} + \Pi/4\right)$: noyau de Silverman.

Quant à la fenêtre, nous la faisons dépendre de n et nous la notons h_n, la suite $(h_n)_{n\geq 1}$ tendant vers zéro lorsque n tend vers l'infini.
Intéressons nous maintenant à la convergence en moyenne quadratique de cet estimateur. Pour cela, nous avons besoin du résultat suivant**.**

Théorème 2 *(Bochner) Soit* $K : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *(*$\mathcal{B}(\mathbb{R})$ *étant la tribu borélienne de* $\mathbb{R}$*) une application bornée et intégrable telle que :*

$$|z|K(z) \underset{|z|\to\infty}{\longrightarrow} 0.$$

Par ailleurs, soit $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *une application intégrable. Posons* $g_n(x) = \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{z}{h_n}\right) g(x-z)\, dz$, *où* $0 < h_n \underset{n\to\infty}{\longrightarrow} 0$.
Si g *est continue au point* x *alors* $\lim_{n\to\infty} g_n(x) = g(x) \int_{-\infty}^{+\infty} K(z)\, dz$. *Si* g *est uniformément continue alors la convergence de* g_n *est uniforme.*

Preuve. Preuve : Soit $\varepsilon > 0$, on a : $|g_n(x) - g(x) \int_{-\infty}^{+\infty} K(z)\, dz| =$
$|\frac{1}{h_n} \int_{-\infty}^{+\infty} K(\frac{z}{h_n}) g(x-z)\, dz - \int_{-\infty}^{+\infty} g(x)K(z)\, dz| = \left| \int_{-\infty}^{+\infty} K(y) g(x - h_n y)\, dy - \int_{-\infty}^{+\infty} g(x)K(y)\, dy \right|$.
g étant continue au point x, il vient

$$\forall \varepsilon > 0,\ \exists \delta > 0 \ /\ |x - z| < \delta \Rightarrow |g(x) - g(z)| \leq \varepsilon.$$

Donc :

$$\begin{aligned} & \int |(g(x - h_n y) - g(x))K(y)|\ dy \\ \leq\ & \int_{\left\{y\,:\,|y|<\frac{\delta}{h_n}\right\}} |g(x - h_n y) - g(x)||K(y)|\, dy + \int_{\left\{y\,:\,|y|\geq\frac{\delta}{h_n}\right\}} |g(x - h_n y) - g(x)||K(y)|\, dy \end{aligned}$$

$$\leq \varepsilon \int_{\left\{|y|<\frac{\delta}{h_n}\right\}} |K(y)|\, dy + \int_{\left\{|y|\geq\frac{\delta}{h_n}\right\}} |g(x - h_n y)||K(y)|\, dy + |g(x)| \int_{\left\{|y|\geq\frac{\delta}{h_n}\right\}} |K(y)|, dy$$

Par l' absolue continuité de l'intégrale et le fait que $h_n \underset{n\to\infty}{\longrightarrow} 0$:

$\exists n_0 \in \mathbb{N} \,/\, \forall n \geq n_0 \ :\ \int_{\left\{|y|\geq\frac{\delta}{h_n}\right\}} |K(y)|\, dy < \varepsilon.$
De plus, sur $\left\{|y| \geq \frac{\delta}{h_n}\right\}$, on a $|y|\frac{h_n}{\delta} \geq 1$, d'où
$\int_{\left\{|y|\geq\frac{\delta}{h_n}\right\}} |g(x - h_n y)||K(y)|\, dy \leq \int_{\left\{|y|\geq\frac{\delta}{h_n}\right\}} |y|\frac{h_n}{\delta}|g(x - h_n y)||K(y)|\, dy$
$\leq \frac{1}{\delta} \sup_{|y|\geq\frac{\delta}{h_n}} |y||K(y)| \int |g(z)|dz \to 0$ par hypothèse puisque $h_n \longrightarrow 0$.
Nous pouvons montrer que si g est uniformément continue (et du fait qu'elle soit intégrable) alors $\lim_{|x|\to\infty} g(x) = 0$, g est donc bornée et nous obtenons la seconde assertion du théorème. ■
Nous imposons au noyau K d'être une densité de probabilité (ce qui fait de f_n une densité) et de satisfaire les conditions suivantes.

$K_1 : K$ est bornée et paire;
$K_2 : |x|K(x) \underset{|x|\to\infty}{\longrightarrow} 0.$
Remarquons que les noyaux gaussien, parabolique et rectangulaire (par exemple) vérifient ces conditions.
Nous sommes, maintenant, en mesure d'énoncer et de montrer le résultat visé

Théorème 3 *Si f_X est continue au point x, $h_n \underset{n\to\infty}{\longrightarrow} 0$, $nh_n \longrightarrow \infty$ et si la densité K vérifie les conditions K_1 et K_2, alors $f_n(x)$ est un estimateur convergeant en moyenne quadratique vers $f_X(x)$.*

Preuve. D'abord par equidistribution des X_i, l'application du théoréme de transfert et l'invariance par translation de la mesure de Lebesgue, il vient

$$\begin{aligned} Ef_n(x) &= E\left\{\tfrac{1}{nh_n}\textstyle\sum_{i=1}^n K\left(\tfrac{x-X_i}{h_n}\right)\right\} \\ &= \tfrac{1}{h_n} EK\left(\tfrac{x-X_1}{h_n}\right) = \tfrac{1}{h_n}\textstyle\int K\left(\tfrac{x-y}{h_n}\right) f_X(y)\,dy \\ &= \tfrac{1}{h_n}\textstyle\int K\left(\tfrac{z}{h_n}\right) f_X(x-z)\,dz \longrightarrow f_X(x)\textstyle\int K(z)\,dz = f_X(x), \end{aligned}$$

en vertu du théorème de Bochner (toutes les conditions sont remplies) et du fait que K est une densité. Ceci montre que $f_n(x)$ est un estimateur asymptotiquement sans biais de $f_X(x)$.
Ensuite, par indépendance et équidistribution des X_i, on peut écrire

$$\begin{aligned} Varf_n(x) &= \frac{1}{nh_n^2} var\left\{K\left(\frac{x-X_1}{h_n}\right)\right\} \leq \frac{1}{nh_n^2} E\left\{K^2\left(\frac{x-X_1}{h_n}\right)\right\} \\ &= \frac{1}{nh_n}\left\{\frac{1}{h_n}\int K^2\left(\frac{z}{h_n}\right) f_X(x-z)\,dz\right\} \\ &\underset{n\to\infty}{\longrightarrow} \lim \frac{1}{nh_n}\left\{f_X(x)\int K^2(z)\,dz\right\} = 0, \end{aligned}$$

par application, encore une fois, du théorème de Bochner et du fait que $nh_n \longrightarrow \infty$.

■

Remarque 1 *1) On peut montrer que si K est à variation bornée ($K = K_1 - K_2$ où $K_1 \nearrow$ et $K_2 \nearrow$) et si pour tout $\theta > 0$, $\sum_{i=1}^{\infty} \exp(-n\theta h_n^2) < \infty$*

alors $\sup |f_n(x) - f(x)| \xrightarrow[n\to\infty]{p.s.} 0$*, ssi* f *est uniformément continue (se référer à Prakasa (1983) pour la preuve).*

2) Cette méthode d'estimation (à noyau) se généralise au cas de $\mathbb{R}^p$*. Ainsi, si* $X_1, \dots X_n$*, sont* n *vecteurs aléatoires i.i.d. de* $\mathbb{R}^p$*, de même densité* f_X *inconnue, alors on peut l'estimer par* $f_n(x) = \frac{1}{nh_n^p} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)$*,* $x \in \mathbb{R}^p$*. Le théorème de Bochner reste valable si* K *et* g *sont définies sur* $\mathbb{R}^p$ *et si on pose* $g_n(x) = \frac{1}{h_n^p} \int K\left(\frac{z}{h_n}\right) g(x - z)\, dz$*, et permet de déduire la consistance de* $f_n(x)$*, sous des conditions adéquates sur* f_X*,* h_n *et* K*.*

3) Il existe bien d'autres méthodes d'estimation de la densité, à savoir : la méthode de projection et celle des delta suites, une des plus récentes et des plus générales (elle englobe les précédentes) étant la méthode des ondelettes (cf Doukhan et Léon (1990), Kerkyacharian et Picard (1992) et Walter (1992), mais l'estimation par la méthode des noyaux reste très utilisée.

4) Il est connu que le choix du noyau n'influe pas beaucoup sur l'estimateur à noyau, mais ce dernier est très sensible au choix de la fenêtre. Il existe donc des méthodes qui aident à un choix plus ou moins "optimum", parmi lesquelles la méthode du plug-in ou la méthode du pouce ; une petite initiation à ce problème est donnée à l'exercice 1.5.

1.4 Estimation de la fonction de régression

La méthode connue de la régression linéaire sert à prédire des valeurs pour une variable Y, dite variable expliquée ou dépendante, à partir des valeurs d'une variable X, dite variable explicative ou indépendante. Elle se base sur l'hypothèse d'existence d'une relation linéaire entre X et Y, qui peut être vérifiée graphiquement ou plus rigoureusement par un test. Mais une telle hypothèse n'est pas toujours raisonnable. En statistique non paramétrique, nous introduisons le modèle plus général $Y = r(X) + \varepsilon$ où ε et X sont indépendantes et $E\varepsilon = 0$, ce qui montre que $r(X) = E(Y/X)$. Notre problème se ramène à l'estimation de l'espérance conditionnelle de Y sachant X que nous étudions ci dessous.

1.4.1 Estimateur de Nadaraya-Watson

Soient $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ des couples aléatoires indépendants, à valeurs dans $\mathbb{R}^2$, et de même loi que (X, Y). Le couple de variables aléatoires

(X, Y) est supposé admettre une densité jointe sur $\mathbb{R}^2$ notée $f_{X,Y}$ et nous désignons par $f_X(.)$ la densité marginale de X, donnée par

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)dy.$$

La densité conditionnelle de Y en $X = x$ est

$$f_{Y/X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)} \qquad \text{si} \quad f_X(x) \neq 0.$$

L'éspérance conditionnelle ou la fonction de régression de Y en $X = x$ s'écrit

$$\begin{aligned} r(x) &= E(Y/X = x) = \int_{-\infty}^{+\infty} y f_{Y/X=x}(y)dy \\ &= \int_{-\infty}^{+\infty} \frac{y f_{X,Y}(x, y)}{f_X(x)}dy. \end{aligned}$$

Soit $J(x, y)$ une fonction de densité sur $\mathbb{R}^2$, on pose

$$K(x) = \int_{-\infty}^{+\infty} J(x, y)dy$$

Soit $h_n \to 0$ quand $n \to \infty$, on a alors

$$\frac{1}{nh_n^2} \sum_{j=1}^{n} J\left(\frac{x - X_j}{h_n}, \frac{y - Y_j}{h_n}\right)$$

est un estimateur de $f_{X,Y}(x, y)$. On a aussi

$$f_n(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right).$$

est un estimateur de $f_X(x)$.
Un estimateur naturel de $r(x)$ est alors donné par

$$r_n(x) = \int_{-\infty}^{+\infty} \frac{y \frac{1}{nh_n^2} \sum_{j=1}^{n} J\left(\frac{x-X_j}{h_n}, \frac{y-Y_j}{h_n}\right)}{f_n(x)} dy.$$

En posant $z = \frac{y-Y_j}{h_n}$, il vient

$$r_n(x) = \frac{h_n \sum_{j=1}^n p\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} + \frac{\sum_{j=1}^n Y_j K\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)},$$

où

$$p(x) = \int_{-\infty}^{+\infty} yJ(x,y)dy.$$

En choisissant

$$J(x,y) = K(x)L(y),$$

avec $\int yL(y)dy = 0$, nous obtenons $p(x) = 0$. D'où l'expression de l'estimateur de la fonction de régression :

$$r_n(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} & \text{si } \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \neq 0. \\ 0 & \text{sinon.} \end{cases}$$

C'est l'estimateur de Nadaraya-Watson, introduit simultanément par ces deux auteurs en 1964. Notons que la covariable X peut être à valeurs dans $\mathbb{R}^d$ ($d \in \mathbb{N}^*$), il faut alors aussi prendre un noyau K défini sur $\mathbb{R}^d$.
Plus généralement, des estimateurs à poids de $r(x)$ peuvent s'écrire

$$\hat{r}_n(x) = \sum_{i=1}^n W_{n,i}(x) Y_i,$$

où $W_{n,i}(x)$ sont des poids dépendant de x et de $(X_1, ..., X_n)$. La valeur de $W_{n,i}$ dépend du type d'estimateur considéré, nous citons ci-dessous des exemples connus.

Estimateur des plus proches voisins

Soit k_n un paramètre de l'estimation, on pose

$$W_{n,i}(x) = \begin{cases} \frac{1}{k_n}, & \text{si } X_i \text{ est parmi les } k_n \text{ plus proches voisins de } x \text{ dans } \{X_1, ..., X_n\}, \\ 0, & \text{sinon.} \end{cases}$$

(Nous renvoyons le lecteur à Devroye et al. (1994)).

Estimateur à partitions

On utilise une partition $\pi_n = \{A_{n,j} : j\}$ de $\mathbb{R}$ et on pose

$$W_{n,i}(x) = \sum_j \frac{1_{A_{n,j}}(X_i)}{\sum_{k=1}^n 1_{A_{n,j}}(X_k)} 1_{A_{n,j}}(x).$$

Pour plus de détails, voir Devroye et Györfi (1983).
Signalons que d'autres types d'estimateurs non paramétriques existent, parmi les quels nous citons l'estimateur des moindres carrés, ou plus généralement les estimateurs spline de lissage (cf, par exemple Kohler et Krzyżak (2001), Kohler (1997, 1999) et Györfi et al. (1997)), ainsi que des estimateurs à ondelettes (cf Donoho et al. (1995)).

1.4.2 Consistance de l'estimateur de Nadaraya-Watson

Théorème 4 *Soit K une densité vérifiant les hypothèses K_1 et K_2, si $E(Y^2) < \infty$, $f_X(x)$ est strictement positive, r, f_X et $m(x) = \int y^2 f_{X,Y}(x,y)dy$ sont continues au point x et si $h_n \to 0$, et $nh_n \to +\infty$ (quand $n \to \infty$), alors $r_n(x)$ est un estimateur consistant de $r(x)$, ce qui veut dire qu'il converge en probabilité vers $r(x)$.*

Preuve. Puisque $f_n(x)$ est un estimateur consistant de $f_X(x)$, il suffit donc de montrer que $\Phi_n(x) = \frac{1}{nh_n}\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)$ est un estimateur consistant de

$$\Phi(x) = \int_{-\infty}^{+\infty} y f_{X,Y}(x,y)dy.$$

Nous avons

$$\begin{aligned}
E(\Phi_n(x)) &= E\left[\frac{1}{nh_n}\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)\right] \\
&= \frac{1}{h_n} E\left[Y K\left(\frac{x-X}{h_n}\right)\right] \\
&= \frac{1}{h_n}\int_{\mathbb{R}} E(Y/X=t)K(\frac{x-t}{h_n})f_X(t)dt \\
&= \frac{1}{h_n}\int_{\mathbb{R}} K(\frac{x-t}{h_n})r(t)f_X(t)dt. \\
&= \frac{1}{h_n}\int_{\mathbb{R}} K(\frac{x-t}{h_n})\Phi(t)dt \to \Phi(x)
\end{aligned}$$

par le théorème de Bochner. De plus,

$$\begin{aligned} var(\Phi_n(x)) &= \frac{1}{nh_n^2} var(YK[(x-X)/h_n]) \\ &\leq \frac{1}{nh_n^2} E(Y^2K^2[(x-X)/h_n]) \\ &= \frac{1}{nh_n^2} \int y^2K^2[(x-u)/h_n] f_{X,Y}(u,y)dudy \\ &= \frac{1}{nh_n} \frac{1}{h_n} \int m(u)K^2[(x-u)/h_n]du \to 0 \text{ quand } n \to \infty, \end{aligned}$$

par le théorème de Bochner et le fait que $nh_n \to \infty$.
Ceci implique que

$$\Phi_n(x) \xrightarrow{\mathcal{P}} \Phi(x).$$

D'où

$$r_n(x) = \frac{\Phi_n(x)}{f_n(x)} \xrightarrow{\mathcal{P}} \frac{\Phi(x)}{f_X(x)} = r(x).$$

■ Remarquons que l'exercice 1.6 donne des conditions suffisantes pour que cet estimateur soit asymptotiquement sans biais. Bien d'autres résultats concernant cet estimateur existent (comme la normalité asymptotique, la convergence en moyenne quadratique et la convergence presque complète). Il est impossible de tout citer, nous nous contentons de rapporter le résultat suivant puisque nous l'utiliserons dans la suite.

1.4.3 Résultat de Devroye et Krzyżak (1989)

Motivation

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ et Y une variable aléatoire réelle de carré intégrable ($EY^2 < \infty$). L'objet de l'analyse de régression est d'estimer Y, après l'observation de X. On cherche une fonction f telle que $f(X)$ soit la plus proche possible de Y dans L^2, dans le sens qu'on veut trouver une fonction f^* qui minimise $E|f(X)-Y|^2$, i.e., f^* doit vérifier

$$E\left|f^*(X)-Y\right|^2 = \min_f E\left|f(X)-Y\right|^2$$

En notant $r(x) = E(Y/X = x)$ et μ la loi de X, on a

$$E\left|f(X) - Y\right|^2 = E\left|r(X) - Y\right|^2 + \int_{\mathbb{R}^d} \left|f(x) - r(x)\right|^2 d(\mu(x)),$$

car en conditionnant par rapport à X, il vient

$$E\left((r(X) - Y)(f(X) - r(X))\right) = 0.$$

On en déduit que $f^* = r$. On cherche donc à minimiser $\int_{\mathbb{R}^d} \left|f(x) - r(x)\right|^2 d(\mu(x))$ pour que $E\left|f(X) - Y\right|^2$ soit la plus proche de sa valeur optimale $E\left|r(X) - Y\right|^2$.
Dans le cadre de l'estimation non paramétrique, la loi du couple (X, Y) est inconnue.
Sur la base d'un échantillon $(X_1, Y_1), ..., (X_n, Y_n)$ de la loi de (X, Y), on veut construire un estimateur $r_n(x)$ de r telle que $\int_{\mathbb{R}^d} \left|r_n(x) - r(x)\right|^2 d\mu(x)$ soit petite. Pour cela nous avons besoin que le noyau soit régulier, notion rappelée ci dessous.

Définition 1 *On dit que le noyau positif K est régulier s'il existe une boule S_r, de rayon r et de centre l'origine, et une constante positive b, telles que $K(x) \geq b1_{S_r}(x)$ et $\int \sup_{y \in x + S_r} K(y)dx < \infty$.*

Enoncé du résultat

Théorème 5 *Soit K un noyau régulier et soit $r_n(x)$ l'estimateur à noyau de la fonction de régression $r(x)$. Alors les propositions suivantes sont équivalentes.*

(A) Pour chaque loi de (X, Y) avec $|Y| \leq M < \infty$, et pour tout $\varepsilon > 0$, il existe deux constantes c et n_0 tels que pour tout $n \geq n_0$,

$$P\left[\int_{\mathbb{R}^d} \left[r_n(x) - r(x)\right] \mu(dx) > \varepsilon\right] \leq e^{-cn}.$$

(B) Pour toute loi de (X, Y) avec $|Y| \leq M < \infty$,

$$\int_{\mathbb{R}^d} \left[r_n(x) - r(x)\right] \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

(C) Pour toute loi de (X, Y) avec $|Y| \leq M < \infty$,

$$\int_{\mathbb{R}^d} \left[r_n(x) - r(x)\right] \mu(dx) \to 0 \quad (n \to \infty) \ \textit{en probabilité.}$$

(D)

$$\lim_{n \to \infty} h_n = 0, \qquad \lim_{n \to \infty} nh_n^d = \infty.$$

EXERCICE 1.1
Soit F la fonction de répartition d'une v.a.r. X et soit F^{inv} son inverse généralisée, définie pour $t \in]0,1[$ par

$$F^{inv}(t) = \inf\{x \in \mathbb{R} \ / \ F(x) \geq t\}.$$

1) Montrer qu'on a les propriétés suivantes.
   a) F^{inv} est une fonction croissante.
   b) $\forall t \in]0,1[, \ F(F^{inv}(t)) \geq t$.
   c) $\forall x \in \mathbb{R}, \ F^{inv}(F(x)) \leq x$.
   d) $F(x) \geq t \Leftrightarrow x \geq F^{inv}(t)$.
   e) Si F est bijective alors $F^{inv} = F^{-1}$.
2) Donner la loi de $F^{inv}(U)$ où U suit la loi uniforme sur $[0,1]$.
3) Si F est continue, quelle est la loi de $F(X)$?

EXERCICE 1.2
Soit F une fonction de répartition et soit $(F_n(x))$ une suite de fonctions croissantes convergeant vers $F(x)$ (en tout point x). Montrer que cette convergence est en fait uniforme, c'est à dire que

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \underset{n \to \infty}{\longrightarrow} 0.$$

EXERCICE 1.3
1) Montrer que la convergence presque complète implique la convergence presque sûre et la convergence en probabilité.
2) Montrer que la convergence presque complète n'implique pas la convergence en moyenne quadratique qui, à son tour, n'entraîne pas la convergence presque complète. Indication : soit l'espace de probabilité $([0,1], \mathcal{B}([0,1]), \lambda)$, où $\mathcal{B}([0,1])$ désigne la tribu borélienne sur $[0,1]$ et λ est la restriction de la mesure de Lebesgue à cet ensemble, considérer les suites $f_n = n1_{[0,\frac{1}{n^2}]}$ et $g_n = 1_{[0,\frac{1}{n}]}$.

EXERCICE 1.4
Soit $\mathcal{C} = \{]a, b], -\infty \leq a \leq b \leq \infty\}$ (pour $b = \infty$, on ouvre l'intervalle) et soit F une application croissante de $\mathbb{R}$ dans $\mathbb{R}$ et continue à droite.
1) Montrer que l'application m_F, définie sur $\mathcal{C}$, par

$$m_F(]a, b]) = F(b) - F(a) \quad (F(\pm\infty) = \lim_{x \to \pm\infty} F(x)) \tag{1.2}$$

est positive et σ additive.
2) En déduire qu'il existe une mesure unique sur $\mathcal{B}(\mathbb{R})$ (boréliens de $\mathbb{R}$), notée aussi m_F , qui vérifie la relation (1.2).
3) Montrer que $\forall\ b \in \mathbb{R},\ m_F(\{b\}) = F(b) - F(b_-)$,
où $F(b_-)$ est la limite à gauche de b. En déduire que F est continue en b ssi $m_F(\{b\}) = 0$.
4) Montrer que $\forall\ a, b \in \mathbb{R}\ a \leq b, \quad m_F(]a, b[) = F(b_-) - F(a), \quad m_F([a, b]) = F(b) - F(a_-)$ et $m_F([a, b[) = F(b_-) - F(a_-)$.
5) Quelle mesure obtient-on quand F est l'application identité ?
Dans toute la suite, soit g une application de $[a, b]$ dans $\mathbb{R}$ intégrable par rapport à la mesure de Lebesgue. On définit G de $[a, b]$ dans $\mathbb{R}$ par
$G(x) = \int_{[a,x]} g(t)d\lambda(t)$, ($\lambda$ est la mesure de Lebesgue)
6) Montrer que G est continue.
7) Montrer que les intégrales $\int_{[a,b]} g(x)F(x)d\lambda(x)$ et $\int_{[a,b]} G(x)dm_F(x)$ ont un sens.
8) Soit $\hat{g}$ l'application définie de $[a, b]^2$ dans $\mathbb{R}$ par

$$\hat{g}(x, y) = \begin{cases} g(y) & \text{si } y \leq x \\ 0 & \text{si } y > x \end{cases},$$

Montrer que $\hat{g}$ est intégrable par rapport à la mesure $m_F \otimes \lambda$ sur $[a, b]^2$.
9) En déduire que
$F(b)G(b) = \int_{]a,b]} g(x)F(x_-)d\lambda(x) + \int_{]a,b]} G(x)dm_F(x)$
(Formule d'intégration par parties).

EXERCICE 1.5
Soit X une v. a. r. de fonction de répartition F et de fonction de répartition empirique F_n.
1) Calculer $varF_n(x)$ et $cov(F_n(x), F_n(y))$ en fonction de F.
2) Soit

$$g_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n} \quad (h_n > 0).$$

Calculer la variance de $g_n(x)$ et $cov(g_n(x), g_n(y))$ en fonction de F.
3) On suppose que X admet une densité f, soit $f_n(x)$ l'estimateur à noyau de f, où le noyau positif K vérifie les conditions suivantes

$K1 : \int K(x)dx = 1$

$K2 : \sup K(x) \leq M < \infty$

$K3 : \int^{x} |x^3| K(x)dx < \infty$

$K4 : \int xK(x)dx = 0.$

On suppose que f a des dérivées d'ordre inférieur ou égal à 3 sur $\mathbb{R}$ (avec $f(x) \neq 0$ et $f^{[3]}$ bornée) et que $h_n \underset{n\to\infty}{\longrightarrow} 0$. On se propose de trouver la "meilleure" fenêtre, au sens MSE, autrement dit on cherche la fenêtre qui minimise l'erreur quadratique moyenne de $f_n(x)$.
Montrer qu'on a asymptotiquement

i)
$$Ef_n(x) = f(x) + \frac{h_n^2}{2}\left\{ f^{[2]}(x) \int x^2 K(x)dx + o(1) \right\}.$$

ii)
$$Var f_n(x) = \frac{1}{nh_n} f(x) \int K^2(x)dx(1 + o(1)).$$

En déduire la valeur de la fenêtre optimale asymptotiquement.

EXERCICE 1.6
Soit $r_n(x)$ l'estimateur de Nadaraya-Watson de la fonction de régression $r(x) = E(Y/X = x)$. Soit

$$\Phi_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K(\frac{x - X_i}{hn})$$

et

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - X_i}{hn}).$$

Supposons que toutes les hypothèses du théorème 4 sont vérifiées.
1) Montrer que si Y est bornée presque sûrement, alors

$$E(r_n(x)) = \frac{E(\Phi_n(x))}{E(f_n(x))} + O(\frac{1}{nh_n}).$$

(Indication : utiliser l'identité $\frac{1}{u} = \frac{1}{c} - \frac{u-c}{c^2} + \frac{(u-c)^2}{uc^2} (u \neq 0,\ c \neq 0)$).

2) En déduire des conditions suffisantes pour que $r_n(x)$ soit un estimateur asymptotiquement sans biais de $r(x)$.

Chapter 2

Introduction à l'analyse de survie

Toute inférence statistique nécessite l'observation d'un échantillon, lequel est dans les meilleurs cas, constitué par des vraies observations de la variable d'intérêt (données complètes). Or en analyse de survie, on ne se trouve justement pas dans ce cas de figure. Un phénomène de censure et (ou) de troncature peut empêcher l'observation de la vraie valeur d'intérêt et ne nous fournit alors qu'une information partielle sur elle, dont la nature nous conduit à différents types de censure et de troncature rapportés ci dessous.

2.1 Différents types de censure et de troncature

Définition 2 *(Censure à droite)*
Il y a censure à droite lorsque nous observons la variable de censure C (et non pas la durée de vie d'intérêt T) et que nous savons que $T > C$. Ce modèle est le plus fréquent en pratique, il est par exemple adapté au cas où l'événement d'intérêt est le temps de survie à une maladie et où la date de fin de l'étude est préalablement fixée ; les patients vivants à la fin de l'étude fournissent des données censurées à droite. Les observations sont des répliques du couple $(T \wedge C, \delta = 1_{\{T \leq C\}})$ où δ vaut 1 quant l'observation est complète, ce qui veut dire qu'elle correspond à une vraie valeur de la variable d'intérêt et vaut 0 sinon (donnée censurée).

Définition 3 *(Censure à gauche)*
Il y a censure à gauche lorsque nous observons la censure C (et non pas la durée de vie T) et que nous savons que $T < C$. Les observations sont des répliques du couple $(T \vee C, \delta = 1_{\{T \geq C\}})$. Un phénomène symétrique au précédent (censure à droite) se produit. Lorsque nous nous intéressons au temps de panne d'un composant placé en parallèle avec un autre composant, une inspection du matériel peut révèler la défaillance du composant d'intérêt à une date antérieure mais inconnue. La censure à gauche est généralement accompagnée de la censure à droite, c'est le cas dans les modèles suivants.

Définition 4 *(Censure par intervalle)*
Il y a censure par intervalle lorsque nous savons seulement que la variable d'intérêt appartient à un intervalle sans connaître sa valeur exacte (une même donnée peut être censurée à gauche et à droite en même temps). Ce modèle est adapté au cas de suivis périodiques de patients (ou de composants en fiabilité) et généralise aussi bien le modèle de censure à droite que celui de censure à gauche.

Définition 5 *(Censure double)*
Il y a censure double lorsque certaines observations sont censurées à droite et d'autres le sont à gauche. Si on s'intéresse à l'âge d'apprentissage d'une certaine tâche, dans certains cas on sait seulement qu'il est inférieur (ou supérieur) à une valeur donnée (l'âge au début ou à la fin de l'étude par exemple).

Définition 6 *(Censure mixte)*
Il y a censure mixte (qualificatif introduit dans ce cours) lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Dans le modèle II décrit dans l'article de Patilea et Rolin (2006), au lieu d'observer un échantillon de la variable d'intérêt T, on observe un échantillon du couple (Z, A) avec $Z = \min(\max(T, G), D)$ et

$$A = \begin{cases} 0 & si\ G < T < D, \\ 1 & si\ D < \max(T, G) \\ 2 & si\ T \leq G \leq D, \end{cases}$$

où G et D sont des variables de censure et A est l'indicateur de censure. Un exemple de ce modèle est donné par un système formé par trois composants,

dont deux sont placés en parallèle (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en série avec ce système en parallèle.

Les modèles de censure à droite peuvent prendre l'une ou l'autre des formes suivantes.

La censure non-aléatoire (ou fixe) de type I Etant donné un nombre positif fixé c et un n-échantillon $T_1, \ldots, T_n$, les observations consistent en (X_i, δ_i), où $X_i = T_i \wedge c$ et $\delta_i = 1_{\{T_i<c\}}$. Ce modèle, souvent utilisé dans les études épidémiologiques, peut correspondre à l'observation de la durée de survie de n patients au cours d'une expérience de durée prédéterminée c.

La censure aléatoire de type I Etant donné un n-échantillon $T_1, \ldots, T_n$, il existe une variable aléatoire n-dimensionnelle $C_1, \ldots, C_n$ de $\mathbb{R}$ telle que les observations consistent en (X_i, δ_i), où $X_i = T_i \wedge C_i$ et $\delta_i = 1_{\{T_i<C_i\}}$. Ce modèle est typiquement utilisé pour les essais thérapeutiques. Nous nous interessons à une cause de décès ayant lieu au bout d'un temps T et nous désirons connaître la loi de T ; cependant, une autre cause aléatoire (décès dû à une autre cause aléatoire, abondon, ...) peut survenir auparavant et empêcher par conséquent l'observation de T.

La censure aléatoire de type II Etant donné un nombre positif fixé r et un n-échantillon $T_1, \ldots, T_n$, les observations consistent en (X_i, δ_i) où $X_i = T_i \wedge T_{(r)}$ et $\delta_i = 1_{\{T_i<T_{(r)}\}}$ avec $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ est la statistiques d'ordre associée à $T_1, \ldots, T_n$. Ce modèle, souvent utilisé dans les études de fiabilité, correspond à l'observation de la durée de fonctionnement de n machines tant que r d'entre elles ne sont pas tombées en panne

Troncature

Nous parlons de troncature à droite (respectivement à gauche) lorsque la variable d'intérêt n'est pas observable quand elle est supérieure (respectivement inférieure) à un seuil C fixé. Dans le cadre de la censure, la variable C est observée alors que dans le cas de la troncature à droite (respectivement à gauche) l'analyse porte uniquement sur la loi de T conditionnellement à

l'événement $\{T < C\}$ (respectivement $\{T > C\}$) et une donnée tronquée ne peut faire partie de l'échantillon. Si une maison de retraite n'accepte que des personnes âgées d'au moins soixante ans, aucun individu décédé avant cet âge n'a la possibilité d'y avoir été admis et est de ce fait tronqué à gauche.

2.2 Fonctions de base

Certaines fonctions jouent un rôle important en analyse de survie, nous allons les introduire et donner les liens existant entre elles.

Soit T une variable aléatoire (v.a.) positive (appelé temps de survie), de fonction de répartition F, sa fonction de survie au point t, notée $S(t)$, est donnée par la probabilité que l'individu survive au delà de t, c'est à dire $S(t) = P(T > t) = 1 - F(t)$. Il est clair que S est une fonction décroissante, continue à droite avec $\lim_{t\to 0} S(t) = 1$ et $\lim_{t\to+\infty} S(t) = 0$.

Avant d'introduire la notion de mesure de hasard, nous allons d'abord étudier deux cas particulièrement intéressants dans la pratique.

2.2.1 Cas d'existence de la densité

Supposons que T a une densité f (par rapport à la mesure de Lebesgue), on définit alors

Définition 7 *On appelle taux de hasard du temps de survie T, la fonction*

$$\lambda(t) = \begin{cases} \frac{f(t)}{S(t)} \text{ si } S(t) \neq 0 \\ 0 \text{ sinon} \end{cases}.$$

La terminologie taux se justifie par le résultat suivant

Théorème 6 *Si f est continue sur $\mathbb{R}_+^*$ alors pour tout $t > 0$ tel que $S(t) \neq 0$, on a*

$\lambda(t) = \lim_{\Delta\to 0} \frac{1}{\Delta} P(t < T \leq t + \Delta / T > t)$.

Preuve. On a

$\lim \frac{1}{\Delta} P(t < T \leq t + \Delta / T > t) = \lim \frac{1}{\Delta} \frac{P(t<T\leq t+\Delta)}{S(t)} = \lim \frac{1}{\Delta} \frac{F(t+\Delta)-F(t)}{S(t)} = \frac{F'(t)}{S(t)} = \lambda(t)$ par continuité de f. ■

Si f est continue, λ l'est aussi et il est possible de définir la fonction de hasard cumulé de T en posant

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

Le résultat suivant donne les relations entre S, λ et Λ.

Théorème 7 *Si f est continue sur $\mathbb{R}_+^*$ et si $A = \{t > 0 \ / \ S(t) \neq 0\}$ alors les propositions suivantes sont équivalentes.*

a) $\forall t \in A,\ \lambda(t) = \frac{f(t)}{S(t)}$.

b) $\forall t \in A,\ \lambda(t) = (-\log\ S(t))'$.

c) $\forall t \in A,\ S(t) = \exp\{-\Lambda(t)\}$.

d) $\forall t \in A,\ f(t) = \lambda(t)\exp\{-\Lambda(t)\}$

Preuve. a)$\Leftrightarrow$ b) car $\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = (-\log\ S(t))'$.

b)$\Leftrightarrow$ c) car $\lambda(t) = (-\log\ S(t))' \Rightarrow \int_0^t \lambda(s)ds = -\log\ S(t)]_0^t = -\log\ S(t)$ puisque $S(0) = 1$ (cas continu).

c)$\Rightarrow$ d) car $S(t) = \exp\{-\Lambda(t)\} \Rightarrow S'(t) = -f(t) = -\lambda(t)\exp\{-\Lambda(t)\}$.

d)$\Rightarrow$ c) car $f(s) = \lambda(s)\exp\{-\Lambda(s)\} \Rightarrow 1 - S(t) = F(t) = \int_0^t f(s)ds = \int_0^t \lambda(s)\exp\{-\Lambda(s)\} = -\exp\{-\Lambda(s)]_0^t\} = -\exp\{-\Lambda(t)\} + 1$. ■

En fiabilité, si T représente le temps de fonctionnement d'un matérial, λ est appelé le taux de défaillance (ou panne) et en général c'est une fonction décroissante pendant la première période de fonctionnemnt (période de rodage), puis elle est approximativement constante (période de vie utile), enfin elle croit (période de vieillissement).

Exemple 1 *a) Loi exponentielle*

Soit T une v.a.de loi exponentielle de paramètre α, alors sa densité est $f(t) = \alpha\exp\{-\alpha t\}1_{]0,+\infty[}(t)$. Calculons sa fonction de survie et son taux de hasard. On a

$S(t) = \int_t^\infty \alpha\exp\{-\alpha s\}1_{]0,+\infty[}(s)ds = \exp\{-\alpha t\}$ *et donc* $\lambda(t) = \frac{f(t)}{S(t)} = \alpha$.

Nous remarquons que ce taux de hasard est constant, la loi exponentielle est alors dite sans mémoire. En fait nous pouvons montrer que la loi exponentielle est la seule loi, à densité, qui soit sans mémoire.

b) Loi de Weibull

Soit T une v.a. de loi de Weibull de paramètres (α, β), $\alpha(> 0)$ est le paramètre d'échelle et $\beta(> 0)$ est le paramètre de forme, sa densité est $f(t) = \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}\exp\{-(\frac{t}{\alpha})^\beta\}1_{]0,+\infty[}(t)$. Calculons sa fonction de survie et son taux de hasard. On a

$S(t) = \int_t^\infty \frac{\beta}{\alpha}(\frac{s}{\alpha})^{\beta-1}\exp\{-(\frac{s}{\alpha})^\beta\}1_{]0,+\infty[}(s)ds = -\exp\{-u^\beta\}]_{t/\alpha}^\infty = \exp\{-(\frac{t}{\alpha})^\beta\}$ *et* $\lambda(t) = \frac{f(t)}{S(t)} = \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}$. *Nous en déduisons que* λ *croît si* $\beta > 1$, *décroît si* $\beta < 1$ *et reste constante si* $\beta = 1$ *(on retrouve la loi exponentielle).*
La loi de Weibull est rès utilisée en fiabilité, elle est apparue expérimentalement lors d'études sur le taux de défaillance de matériels. Le théorème des valeurs extrêmes fournit une explication mathématique au fait que cette loi se rencontre dans la nature (pour plus de détails voir Cocozza (1997) page 13).

2.2.2 Cas discret

Soit X une variable aléatoire discrète prenant les valeurs ordonnées, par ordre croissant, $x_1, \ldots, x_n$...Sa fonction de survie est donnée par
$S(t) = P(X > t) = \sum_{j:x_j>t} P(X = x_j)$. Son taux de hasard est $\lambda(x_j) = P(X = x_j/X \geq x_j) = \frac{P(X=x_j)}{S(x_{j-1})} = 1-\frac{S(x_j)}{S(x_{j-1})}$. Mais puisque $S(x) = \prod_{j:x_j\leq x}\frac{S(x_j)}{S(x_{j-1})}$, il vient

$$S(x) = \prod_{j:x_j\leq x}(1-\lambda(x_j)).$$

Exemple
Soit X une v.a. prenant les valeurs 0, 5 et 8 avec probabilité 1/3 pour chacune. Alors $S(x) = \begin{cases} 1 \text{ si } x < 0 \\ 2/3 \text{ si } 0 \leq x < 5 \\ 1/3 \text{ si } 5 \leq x < 8 \\ 0 \text{ si } x \geq 8 \end{cases}$ et $\lambda(x) = \begin{cases} 1/3 \text{ si } x = 0 \\ 1/2 \text{ si } x = 5. \\ 1 \text{ si } x = 8 \\ 0 \text{ ailleurs} \end{cases}$

2.2.3 Cas général

Mesure de hasard

Si T est une v.a.r. de fonction de répartition quelconque F, la notion de taux de hasard se généralise en la notion de mesure de hasard $d\Lambda$ définie par

$$d\Lambda(s) = \frac{dF(s)}{1-F(s_-)},$$

où $F(s_-)$ est la limite de F à gauche de s et dF est la loi de probabilité de T. Nous pouvons remarquer que
- Si $dF(s) = f(s)ds$ alors $d\Lambda(s) = \lambda(s)ds = \frac{f(s)ds}{S(s)}$.
- Si dF est une loi discrète nous retrouvons $\lambda(s) = \frac{P(T=s)}{S(s_-)}$.

- Nous avons donné dans chacun des cas particuliers précédents la formule liant la fonction de survie au taux de hasard, dans le cas général cette relation est donnée à l'exercice 2.3. $S(t)$ est le produit intégral de $1 - d\Lambda$ où ce produit intégral généralise le produit fini au cas infini tout comme l'intégrale généralise la somme finie au cas infini.

2.3 Intégrale de Lebesgue-Stieljes

Partant d'une fonction croissante et continue à droite, nous avons défini l'intégrale de Lebesgue-Stieltjes à l'exercice 1.4, notion que nous allons étendre au cas d'une fonction à variation bornée, cette dernière notion se définissant comme suit.

Définition 8 *Soit f une fonction définie sur un intervalle $[a, b]$, on dit que f est à variation bornée s'il existe une constante C telle que*

$$V_a^b(f) := \sup_{\{t_i\}} \left\{ \sum_{i=1}^{n-1} |f(t_{i+1}) - f(t_i)| \right\} < C,$$

le sup étant pris sur toutes les subdivisions $\{t_i, a = t_1 < t_2 \ldots < t_n = b\}$ de l'intervalle $[a, b]$. $V_a^b(f)$ est alors la variation totale de f sur $[a, b]$. Si f est définie sur $\mathbb{R}$, elle est dite à variation bornée si les nombres $V_a^b(f)$ forment un ensemble borné, dans ce cas

$$V_{-\infty}^{\infty}(f) := lim_{a \to -\infty, b \to \infty} V_a^b(f),$$

est la variation totale de f sur $\mathbb{R}$.

Il est clair que si f est croissante (resp. décroissante) alors elle est à variation bornée et $V_a^b(f) = |f(b) - f(a)|$. Par contre la fonction $f(x)$ définie sur un intervalle $[a, b]$ et prenant la valeur 1 si x est rationnel et la valeur 0 sinon n'est pas à variation bornée.
La variation totale d'une fonction possède les propriétés suivantes.
i) Pour toute constante c, $V_a^b(cf) = |c|V_a^b(f)$.
ii) Si f et g sont à variation bornée alors $f + g$ est aussi à variation bornée et $V_a^b(f + g) \leq V_a^b(f) + V_a^b(g)$.
iii) Si $a < b < c$ alors $V_a^b(f) + V_b^c(f) = V_a^c(f)$.
iv) $x \to V(x) := V_a^x(f)$ est croissante.

Il est donc clair que l'ensemble des fonctions à variation bornée est un espace vectoriel.
Le point 2) du résultat suivant permet de caractériser les fonctions numériques à variation bornée.

Théorème 8 *Si f est une fonction à variation bornée sur $[a,b]$ alors pour tout couple $(x,y) \in \mathbb{R}^2$, on a*
1) $x < y \Rightarrow |f(x) - f(y)| \leq V(y) - V(x)$.
2) $f = f_1 - f_2$ où f_1 et f_2 sont croissantes.

Preuve. 1) Soit $\varepsilon > 0$. par définition de la borne sup, il existe $t_1 < t_2 \ldots < t_n = x$ tels que

$$\sum_{i=1}^{n-1} |f(t_{i+1}) - f(t_i)| > V(x) - \varepsilon.$$

De plus, on a

$$V(y) \geq |f(x) - f(y)| + \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|.$$

Nous obtenons donc pour tout $\varepsilon > 0$,

$$V(y) \geq |f(x) - f(y)| + V(x) - \varepsilon.$$

Le résultat en découle en faisant tendre ε vers 0.
2) Il suffit de poser $f_1 = \frac{V+f}{2}$ et $f_2 = \frac{V-f}{2}$ et 1) permet de déduire que ces deux fonctions sont croissantes. ■
On montre (cf exercice 2.8) que si une fonction à variation bornée est continue à droite alors sa variation $V_a^{\cdot}(f)u$ est aussi continue à droite.
Soit donc F une fonction à variation bornée et continue à droite alors il existe un couple unique (F_1, F_2) de fonctions de $\mathbb{R}$ dans $\mathbb{R}$ croissantes, continues à droite et nulles en 0 telles que

$$F - F(0) = F_1 - F_2 \text{ et } V_F = F_1 + F_2.$$

Par conséquent, nous définissons l'intégrale par rapport à dF en posant pour toute fonction g mesurable et positive ou dF_1 et dF_2 intégrable

$$\int g dF = \int g dF_1 - \int g dF_2.$$

Une propriété fondamentale de cette intégrale est donnée par le résultat suivant.

Théorème 9 *(Formule d'intégration par parties)*
Soient F et G des applications de $\mathbb{R}$ dans $\mathbb{R}$ continues à droite et à variation bornée alors
$F(t)G(t) - F(0)G(0) = \int_{]0,t]} F(x_-)dG(x) + \int_{]0,t]} G(x)dF(x) =$
$\int_{]0,t]} F(x_-)dG(x) + \int_{]0,t]} G(x_-)dF(x) + \sum_{x\in]0,t]} \Delta F(x)\Delta G(x)$,
où $\Delta H(x) = H(x) - H(x_-)$ est l'accroissement de H au point x.

Preuve. Par application du théorème de Fubini, nous pouvons écrire

$$(F(t) - F(0))(G(t) - G(0)) = \int_{]0,t]} dF(x) \int_{]0,t]} dG(y) = \int\int_{]0,t]\times]0,t]} (dF \otimes dG)(x, y)$$

$$= \int_{]0,t]} dF(x) \left[\int_{]0,t]} 1_{\{0<y\leq x\}} dG(y) + \int_{]0,t]} 1_{\{0<x<y\}} dG(y) \right]$$

$$= \int_{]0,t]} dF(x)(G(x) - G(0)) + \int_{]0,t]} dG(y) \int_{]0,y[} dF(x)$$

$$= \int_{]0,t]} (G(x) - G(0))dF(x) + \int_{]0,t]} (F(y-) - F(0))dG(y)$$

$$= \int_{]0,t]} G(x)dF(x) + \int_{]0,t]} F(y-)dG(y) - G(0)[F(t) - F(0)] - F(0)[G(t) - G(0].$$

Par simplification des termes identiques, nous obtenons
$F(t)G(t) - F(0)G(0) = \int_{]0,t]} F(x_-)dG(x) + \int_{]0,t]} G(x)dF(x)$
Il reste à montrer que $\int_{]0,t]} G(x)dF(x) = \int_{]0,t]} G(x_-)dF(x) + \sum_{x\in]0,t]} \Delta F(x)\Delta G(x)$.
Puisque $\Delta G(x) = G(x) - G(x_-)$, nous avons
$\int_{]0,t]} G(x)dF(x) = \int_{]0,t]} G(x_-)dF(x) + \int_{]0,t]} \Delta G(x)dF(x)$.
En introduisant la partie continue de F : $F^c(x) = F(x) - \sum_{s\leq x} \Delta F(s)$,
il vient $\int_{]0,t]} \Delta G(x)dF(x) = \int_{]0,t]} \Delta G(x)dF^c(x) + \sum_{0<x\leq t} \Delta G(x)\Delta F(x)$.
Or $\int_{]0,t]} \Delta G(x)dF^c(x) = 0$ puisque ΔG est une fonction en escalier et F^c étant continue, la mesure (dF^c) de tout singleton est nulle. ■

2.4 Estimation non paramétrique de la fonction de survie

2.4.1 Introduction de l'estimateur de Kaplan-Meier

Nous nous intéressons à l'estimation de la fonction de survie S de la v.a.r. T censurée à droite. C'est à dire que nous observons n variables indépendantes

$(X_i = T_i \wedge C_i, \delta_i = 1_{\{T_i \leq C_i\}})$ de même loi que le couple $(X = T \wedge C, \delta = 1_{\{T \leq C\}})$, où C est une variable de censure et δ est l'indicateur de censure. La méthode d'estimation se base sur la relation suivante.

$$S(t_n) = \prod_{j=1}^{n} P(T > t_j / T > t_{j-1}) S(t_0),$$

où $t_0 < t_1 \ldots < t_n$.
Soit donc $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ la statistique d'ordre associée à $X_1, X_2, \ldots, X_n$ et notons $\delta_{(i)}$ l'indicateur de censure associé à $X_{(i)}$. Nous partons de $P(T > X_{(i)}) = \prod_{j=1}^{i} P(T > X_{(j)} / T > X_{(j-1)}) = \prod_{j=1}^{i} p_j = \prod_{j=1}^{i} (1 - q_j)$, où X_0 est choisi tel que $S(X_{(0)}) = 1$ et
p_j est la probabilité qu'un individu survive pendant l'intervalle $I_j = [X_{(j)}, X_{(j+1)}[$ sachant qu'il était vivant juste avant cet intervalle et q_j est la probabilité qu'un individu meure pendant l'intervalle I_j sachant qu'il était vivant juste avant cet intervalle. Une estimation naturelle de q_j est donnée par

$$\hat{q}_j = \frac{M(X_{(j)})}{R(X_{(j)})},$$

où $M(X_{(j)}) = \sum_{i=1}^{n} \delta_i 1_{\{X_i = X_{(j)}\}}$ est le nombre de morts réelles observées à l'instant $X_{(j)}$ et $R(X_{(j)}) = \sum_{i=1}^{n} 1_{\{X_i \geq X_{(j)}\}}$ est le nombres d'individus à risque (de mourir) juste avant l'instant $X_{(j)}$. L'estimateur $S_n(t)$ de $S(t)$ est alors défini par

$$S_n(t) = \prod_{X_{(j)} \leq t} \left(1 - \frac{M(X_{(j)})}{R(X_{(j)})}\right).$$

C'est l'estimateur introduit dans Kaplan et Meier (1958) et que nous notons dans la suite par L'EKM.

Remarque 2 *Si $X_{(j)}$ est un instant où ne se produisent que des censures alors $M(X_{(j)}) = 0$, donc S_n ne varie qu'aux points où il y a effectivement une mort (donnée complète).*
S'il n ' y a pas d'ex æquo alors $M(X_{(j)}) = 0$ ou $M(X_{(j)}) = 1$ et dans ce dernier cas $R(X_{(j)}) = n - j + 1$, on obtient alors

$$S_n(t) = \prod_{X_{(j)} \leq t} \left(1 - \frac{1}{n - j + 1}\right)^{\delta_{(j)}}.$$

En fait, cette formule est toujours valable en affectant aux ex æquo des rangs différents et successifs. La valeur de S_n à conserver est alors celle trouvée au dernier rang.

Exemple 2 *Sur dix patients atteints de cancer des bronches, on a observé les durées de survie suivantes, exprimées en mois.*
5, 4+, 11, 9, 10+, 13+, 3, 1, 7+, 8,
les données suivies de + sont censurées à droite. Calculons l'estimateur de la loi du temps de survie à cette maladie.

Rang	*Temps de survie*	$S_n(t)$
1	1	$1-(1/10)=0.9$ *sur* $[1,3[$
2	3	$(9/10)\times(8/9)=0.8$ *sur* $[3,5[$
3	4+	
4	5	$(9/10)\times(8/9)\times(6/7)=0.69$ *sur* $[5,8[$
5	7+	
6	8	$(9/10)\times(8/9)\times(6/7)\times(4/5)=0.56$ *sur* $[8,9[$
7	9	$(9/10)\times(8/9)\times(6/7)\times(4/5)\times(3/4)=0.42$ *sur* $[9,11[$
8	10+	
9	11	$(9/10)\times(8/9)\times(6/7)\times(4/5)\times(3/4)\times(1/2)=0.21$ *sur* $[11,\infty[$
10	13+	

2.4.2 Autre approche : self consistance

Supposons que S est continue et que T et C sont des variables indépendantes. Soient $N_t(s)$ le nombre de T_i dépassant strictement s et $N_x(s)$ le nombre de X_i dépassant strictement s, il est naturel "d'estimer" S par $\frac{N_t(s)}{n}$ mais la censure empêche l'observation de $N_t(s)$ pour les observations censurées. Puisque $t_i \geq x_i$, nous avons $N_t(s) \geq N_x(s)$. L'ambiguité s'introduit lorsque $x_i < s$ et il y a censure, on ne sait pas si $t_i > s$ mais on a au vu des hypothèses

$$P(T_i > s/X_i < s, \delta_i = 0) = \frac{P(T_i > s, C_i < s)}{P(T_i > x_i, C_i < s)} = \frac{P(T_i > s)}{P(T_i > x_i)} = \frac{S(s)}{S(x_i)}.$$

S étant inconnue, on se donne n'importe quel estimateur de S, noté S_1, on estime alors $\frac{S(s)}{S(x_i)}$ par $\frac{S_1(s)}{S_1(x_i)}$. On en déduit un second estimateur S_2 de S en imposant

$$nS_2(s) = N_x(s) + \sum_{x_i<s,\delta_i=0} \frac{S_1(s)}{S_1(x_i)}. \tag{2.1}$$

Par itération, nous obtenons une suite $S_1, S_2, S_3, \ldots$ qui converge vers une fonction $\hat{S}$ qui ne varie plus par application de l'equation (2.1). Une telle fonction vérifie

$$n\hat{S}(s) = N_x(s) + \sum_{x_i<s, \delta_i=0} \frac{\hat{S}(s)}{\hat{S}(x_i)}.$$

Un estimateur satisfaisant une telle equation s'appelle estimateur self-consistant. C'est Efron (1967) qui a introduit cette approche et il a montré que l'estimateur $\hat{S}$ coincide avec l'estimateur de Kaplan-Meier et qu'il consiste à commencer par attribuer à chaque observation la masse $1/n$, puis on parcourt les observations et à chaque fois que l'on rencontre une donnée censurée, on lui enlève sa masse et on la distribue équitablement sur toutes les observations qui lui sont supérieures (voir aussi Peterson (1977)).

EXERCICE 2.1
Soit U une fonction continue sur $\mathbb{R}_+$ et à variation bornée sur tout intervalle de longueur finie. Montrer que pour tout $n \in \mathbb{N}^*$, on a

$$U^n(t) = U^n(0) + \int_{]0,t]} nU^{n-1}(s)dU(s),$$

et que pour tout réel a

$$\exp\{aU(t)\} = \exp\{aU(0)\} + \int_{]0,t]} a\exp\{aU(s)\}dU(s)$$

EXERCICE 2.2
Soit U une fonction de $\mathbb{R}_+$ dans $\mathbb{R}$, continue à droite, à variation bornée sur tout intervalle de longueur finie. Montrer que pour tout réel a l'équation :

$$Z(t) = Z(0) + a \int_{]0,t]} Z(s_-)dU(s) \tag{2.2}$$

a une et une seule solution bornée sur tout compact et cette solution est :

$$Z(t) = Z(0)\{\prod_{0<s\leq t} (1 + a\Delta U(s))\} \exp\{aU^c(t)\}, \tag{2.3}$$

où $U^c(t) = U(t) - \sum_{s\leq t} \Delta U(s)$ est la partie continue de U.

En particulier si U est en plus continue et nulle en zéro, l'équation :

$$Z(t) = Z(0) + a \int_{]0,t]} Z(s_-) dU(s), \tag{2.4}$$

admet une unique solution bornée sur tout compact, donnée par :

$$Z(t) = Z(0) \exp \left\{ a \int_{]0,t]} dU(s) \right\}. \tag{2.5}$$

EXERCICE 2.3
1) Montrer que si T est une v.a.r. quelconque, de fonction de survie S et de mesure de hasard $d\Lambda$ alors :

$$S(t) = (1 - P(T = 0)) \prod_{0<s\leq t} (1 - \Delta\Lambda(s)) \exp \left\{ - \int_0^t d\Lambda^c(s) \right\}.$$

2) Si en plus, T est une v.a.r. de loi diffuse (i.e. $\forall x \in \mathbb{R},\ P(T = x) = 0$), montrer que pout tout $t > 0$, on a:

$$S(t) = \exp \left\{ - \int_0^t d\Lambda(s) \right\}.$$

EXERCICE 2.4
Dix patients atteints d'un cancer sont suivis jusqu'à la mort. Leurs temps de survie, en mois, sont : 4 5 6 8 8 8 10 10 11 12.
Estimer la fonction de survie du temps de survie à ce cancer et tracer la courbe de cet estimateur.

EXERCICE 2.5
Les temps de rémission en mois, après l'atteinte par un cancer, ont été observés chez dix patients. Six ont rechuté après 3.0, 6.5, 6.5, 10, 12 et 15 mois. Un patient a été perdu de vue après 8.4 mois et trois patients sont encore en période de rémission à la fin de l'étude après 4.0, 5.7 et 10 mois.
Estimer la fonction de survie du temps de rémission de ce cancer et tracer la courbe de cet estimateur.

EXERCICE 2.6
Donner une condition nécessaire et suffisante pour que l'estimateur de Kaplan-Meier s'annule.

EXERCICE 2.7
Voir quelles sont les propriétés de l'intégrale de Lebesgue (par rapport à une mesure positive) qui restent valables pour l'intégrale de Lebesgue Stieltjes associée à une fonction à variation bornée.

EXERCICE 2.8
Montrer que si une fonction f à variation bornée sur $[a, b]$ est continue à droite (resp. à gauche) alors sa variation totale $V_a^{\cdot}(f)$ est aussi continue à droite (resp. à gauche).

EXERCICE 2.9
1) Soit f l'application définie sur $\mathbb{R}$ par $f(x) = \begin{cases} \frac{1}{x-1} & \text{si } x \neq 1 \\ 0 & \text{si } x = 1 \end{cases}$.
Montrer que f est à variation bornée sur tout intervalle $[a, b]$ satisfaisant $[a, b] \subset]0, 1[$. f est elle à variation bornée sur $[0, 1]$?
2) Soit g l'application définie sur $\mathbb{R}$ par $g(x) = \begin{cases} \sin(\pi/x)\sqrt[3]{x} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$.
g est elle continue sur $[0, 1]$? Montrer que g n'est pas à variation bornée sur $[0, 1]$.

Chapter 3

Propriétés de L'estimateur de Kaplan-Meier

Dans toute la suite de ce document, pour toute variable aléatoire réelle V, de fonction de répartition F_V, notons

$$T_V := T_{F_V} = \sup\{t : F_V(t) < 1\} \quad \text{et} \quad I_V := I_{F_V} = \inf\{t : F_V(t) \neq 0\},$$

les points terminaux du support de V.
Il est clair que si $t > T_V$ alors $S_V(t) = 0$ et si $t < I_V$ alors $F_V(t) = 0$.
Soit T une variable aléatoire positive, censurée à droite par une variable aléatoire C positive et indépendante de T. Nous observons l'échantillon $(X_i = T_i \wedge C_i, \delta_i = 1_{\{T_i \leq C_i\}})_{1 \leq i \leq n}$ de n couples de variables aléatoires i.i.d. et de même loi que $(X = T \wedge C, \delta = 1_{\{T \leq C\}})$, et nous notons F, G et H les fonctions de répartition respectives de T, C et X; S, $\overline{G}$ et $\overline{H}$ leurs fonctions de survie respectives.
Soit $(Z_j)_{1 \leq j \leq m}$ les valeurs distinctes des $(X_j)_{1 \leq j \leq n}$ rangées dans l'ordre croissant. L'estimateur de Kaplan-Meier de S est donné pour tout $t \in \mathbb{R}$ par :

$$S_n(t) = \prod_{j/Z_j \leq t} \left(1 - \frac{M(Z_j)}{R(Z_j)}\right),$$

où $M(Z_j) = \sum_{i=1}^{n} \delta_i 1_{\{X_i = Z_j\}}$ est le nombre de morts exactes au jème instant et $R(Z_j) = \sum_{i=1}^{n} 1_{\{X_i \geq Z_j\}}$ est le nombre d'individus à risque juste avant le jème instant.

Grâce à la théorie des martingales appliquée aux processus ponctuels, la consistance de cet estimateur a été prouvée d'une manière assez laborieuse (le lecteur peut se reporter au livre de Fleming et Harrington (1991)pour des détails sur ce point). En s'appuyant sur une preuve donnée dans Shorack et Wellner (1986), nous allons montrer sa consistance forte d'une manière élégante et rapide. Pour cela, nous avons d'abord besoin du lemme suivant, dont la preuve peut se faire comme celle de l'exercice 2.2.

Lemme 1 *(cf Shorack et Wellner (1986) page 302)*
Si A et B sont deux fonctions croissantes et continues à droite sur $[0,+\infty[$ avec $A(t) = B(t) = 0$ pour $t < 0$ et $\Delta A \leq 1$ et $\Delta B < 1$ sur $[0,+\infty[$ et si $\theta_B = \inf\{t \in \mathbb{R}/B(t) = +\infty\}$, alors la seule solution locale bornée Z de l'équation

$$Z(t) = Z(0) - \int_{[0,t]} \frac{Z(x_-)}{1-\Delta B(x)} d(A(x) - B(x))$$

sur $[0,\theta_B[$ est donnée par

$$Z(t) = Z(0)\exp(B^c(t) - A^c(t)) \frac{\prod_{0\leq x\leq t}(1-\Delta A(x))}{\prod_{0\leq x\leq t}(1-\Delta B(x))}.$$

Introduisons les statistiques suivantes

$$N_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t, \delta_i = 1\}} \text{ et } Y_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \geq t\}}.$$

$N_n(t)$ et $Y_n(t)$ sont respectivement les versions empiriques associées à
$H_1(t) = P(X \leq t, \delta = 1)$ (sous loi de X correspondant à une donnée complète) et à $\overline{H}(t_-)$.
Nous pouvons aisément déduire que
$H_1(t) = P(X \leq t, \delta = 1) = E(N_n(t)) = \int_{[0,t]} \overline{G}(x_-)dF(x)$ et $E(Y_n(t) = P(X \geq t) = \overline{H}(t_-) = \overline{G}(t_-)S(t_-)$.
Par ailleurs, la fonction de hasard cumulé de T est donnée pour tout $t < t_H$ par

$$\Lambda(t) = \int_{[0,t]} \frac{dF(x)}{S(x_-)} = \int_{[0,t]} \frac{dEN_n(x)}{EY_n(x)}.$$

Ceci suggère de l'estimer par

$$\Lambda_n(t) = \int_{[0,t]} \frac{dN_n(x)}{Y_n(x)},$$

c'est l'estimateur de Nelson Aalen, qui s'écrit

$$\Lambda_n(t) = \sum_{j/Z_j \leq t} \frac{\Delta N_n(Z_j)}{Y_n(Z_j)} = \sum_{j/Z_j \leq t} \frac{M(Z_j)}{R(Z_j)}.$$

Donc, $\Delta\Lambda_n(Z_j) = \frac{M(Z_j)}{R(Z_j)}$.
Nous pouvons donc voir que l'estimateur de Kaplan-Meir s'écrit

$$S_n(t) = \prod_{s \leq t} (1 - \Delta\Lambda_n(s)). \tag{3.1}$$

Nous sommes maintenant en mesure d'énoncer et de montrer le résultat suivant.

Théorème 10 *Pour tout $\theta \in]0, T_H[$, on a*

$$\sup_{0 \leq t \leq \theta} |\Lambda_n(t) - \Lambda(t)| \xrightarrow{p.s.} 0.$$

et

$$\sup_{0 \leq t \leq \theta} |S_n(t) - S(t)| \xrightarrow{p.s.} 0.$$

Preuve. Soit $t \in [0, \theta]$.
• Commençons par traiter Λ_n.
Nous avons

$$\begin{aligned}
|\Lambda_n(t) - \Lambda(t)| &= \left| \int_{[0,t]} \frac{dN_n(x)}{Y_n(x)} - \int_{[0,t]} \frac{dH_1(x)}{\overline{H}(x_-)} \right| \\
&= \left| \int_{[0,t]} \frac{dN_n(x)}{Y_n(x)} - \int_{[0,t]} \frac{dH_1(x)}{\overline{H}(x_-)} + \int_{[0,t]} \frac{dN_n(x)}{\overline{H}(x_-)} - \int_{[0,t]} \frac{dN_n(x)}{\overline{H}(x_-)} \right| \\
&\leq \left| \int_{[0,t]} \left(\frac{1}{Y_n(x)} - \frac{1}{\overline{H}(x_-)} \right) dN_n(x) \right| + \left| \int_{[0,t]} \frac{1}{\overline{H}(x_-)} d(N_n(x) - H_1(x)) \right|.
\end{aligned}$$

Par la formule d'intégration par parties, nous obtenons

$$\left|\int_{[0,t]} \frac{\overline{H}(x_-) - Y_n(x)}{Y_n(x) \times \overline{H}(x_-)} dN_n(x)\right| + \left|\int_{[0,t]} \frac{1}{\overline{H}(x_-)} d(N_n(x) - H_1(x))\right|$$

$$\leq \sup_{0\leq u\leq\theta} \left|\frac{\overline{H}(u_-) - Y_n(u)}{Y_n(u) \times \overline{H}(u_-)}\right| N_n(t) + \left|\frac{N_n(t) - H_1(t)}{\overline{H}(t)}\right| + \left|\int_{[0,t]} (N_n(x) - H_1(x)) d\left(\frac{1}{\overline{H}(x)}\right)\right|$$

$$\leq \frac{1}{Y_n(\theta) \times \overline{H}(\theta_-)} \sup_{0\leq u\leq\theta} |Y_n(u) - \overline{H}(u_-)| + \frac{1}{\overline{H}(\theta)} \sup_{0\leq u\leq\theta} |N_n(u) - H_1(u)| +$$

$$\sup_{0\leq u\leq\theta} |N_n(u) - H_1(u)| \left(\frac{1}{\overline{H}(t)} - 1\right)$$

$$\leq \frac{1}{Y_n(\theta) \times \overline{H}(\theta_-)} \sup_{0\leq u\leq\theta} |Y_n(u) - \overline{H}(u_-)| + \left(\frac{2}{\overline{H}(\theta)} - 1\right) \sup_{0\leq u\leq\theta} |N_n(u) - H_1(u)|$$

et comme $Y_n(\theta) \xrightarrow{p.s.} \overline{H}(\theta_-) \neq 0$, on a $\frac{1}{Y_n(\theta)} \xrightarrow{p.s.} \frac{1}{\overline{H}(\theta_-)}$
Donc, $\exists\, C(\theta) > 0 / \frac{1}{Y_n(\theta)} \leq C(\theta)\, p.s.$ et par conséquent :
$|\Lambda_n(t) - \Lambda(t)| \leq \frac{C(\theta)}{\overline{H}(\theta_-)} \sup_{0\leq u\leq\theta} |Y_n(u) - \overline{H}(u_-)| + \left(\frac{2}{\overline{H}(\theta)} - 1\right) \sup_{0\leq u\leq\theta} |N_n(u) - H_1(u)|\, p.s.$
et ceci pour tout $t \in [0, \theta]$, autrement dit

$$\sup_{0\leq t\leq\theta} |\Lambda_n(t) - \Lambda(t)| \leq \frac{C(\theta)}{\overline{H}(\theta_-)} \sup_{0\leq t\leq\theta} |Y_n(t) - \overline{H}(t_-)|$$
$$+ \left(\frac{2}{\overline{H}(\theta)} - 1\right) \sup_{0\leq t\leq\theta} |N_n(t) - H_1(t)|\, p.s.$$

Afin de traiter ci dessous le terme $S_n(t) - S(t)$, nous pouvons montrer de la même façon que

$$\sup_{0\leq t\leq\theta} |\Lambda_n(t_-) - \Lambda(t_-)| \leq \frac{C(\theta)}{\overline{H}(\theta_-)} \sup_{0\leq t\leq\theta} |Y_n(t) - \overline{H}(t_-)|$$
$$+ \frac{1}{\overline{H}(\theta_-)} \sup_{0\leq t\leq\theta} |N_n(t_-) - H_1(t_-)| + \left(\frac{1}{\overline{H}(\theta_-)} - 1\right) \sup_{0\leq t\leq\theta} |N_n(t) - H_1(t)|\, p.s.$$

D'après le théorème de Glivenko-Cantelli tous les sup, figurant aux membres de droite des deux inégalités précédentes, tendent vers zéro, d'où le résultat visé pour $\Lambda_n(t)$ et $\Lambda_n(t_-)$.
• Passons maintenant à la convergence de S_n.
Puisque $S(t) = 1 - P(T \leq t) = 1 - \int_{[0,t]} dF(x) = 1 - \int_{[0,t]} S(x_-)\, d\Lambda(x)$, le

lemme 1 donne:

$$S(t) = \prod_{x \leq t} (1 - \Delta\Lambda(x)) \exp(-\Lambda^c(t)). \tag{3.2}$$

Les relations (3.2) et (3.1) montrent, d'après le lemme 1, que $\frac{S_n(t)}{S(t)}$ vérifie :

$$\frac{S_n(t)}{S(t)} = 1 - \int_{[0,t]} \frac{S_n(x_-) d(\Lambda_n(x) - \Lambda(x))}{S(x_-)(1 - \Delta\Lambda(x))}$$

$$\Rightarrow S_n(t) - S(t) = -S(t) \int_{[0,t]} \frac{S_n(x_-) d(\Lambda_n(x) - \Lambda(x))}{S(x_-)(1 - \Delta\Lambda(x))}$$

$$\Rightarrow |S_n(t) - S(t)| \leq \left| \int_{[0,t]} \frac{S_n(x_-)}{S(x_-)} dK_n(x) \right| \text{ où } K_n(t) = \int_{[0,t]} \frac{d(\Lambda_n(x) - \Lambda(x))}{1 - \Delta\Lambda(x)}.$$

Nous en déduisons, en appliquant la formule d'intégration par parties et le fait que les fonctions S et S_n sont bornées par 1, que :

$$\begin{aligned}
|S_n(t) - S(t)| &\leq \frac{S_n(t)}{S(t)} |K_n(t)| + \left| \int_{[0,t]} K_n(x) d\left(\frac{S_n(x)}{S(x)} \right) \right| \\
&\leq \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |K_n(u)| + \left| \int_{[0,t]} K_n(x) S_n(x_-) d\left(\frac{1}{S(x)} \right) \right| + \left| \int_{[0,t]} \frac{K_n(x)}{S(x)} dS_n(x) \right| \\
&\leq \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |K_n(u)| + \sup_{0 \leq u \leq \theta} |K_n(u)| \left(\frac{1}{S(t)} - 1 \right) + \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |K_n(u)| |S_n(t) - 1| \\
&\leq \left(\frac{3}{S(\theta)} - 1 \right) \sup_{0 \leq u \leq \theta} |K_n(u)|.
\end{aligned}$$

Or $S(x) = S(x_-)(1 - \Delta\Lambda(x))$, d'où

$$\frac{1}{1 - \Delta\Lambda(x)} = \frac{S(x_-)}{S(x)} = 1 - \frac{\Delta S(x)}{S(x)}.$$

Donc $K_n(u) = \int\limits_{[0,u]} (1 - \frac{\Delta S(x)}{S(x)}) d(\Lambda_n(x) - \Lambda(x))$.
Par conséquent et en utilisant le fait que $\sum |\Delta S(x)| \leq 1$, nous obtenons que

$$\begin{aligned}
|K_n(u)| &\leq \left| \int\limits_{[0,u]} d(\Lambda_n(x) - \Lambda(x)) \right| + \left| \int\limits_{[0,u]} \frac{\Delta S(x)}{S(x)} d(\Lambda_n(x) - \Lambda(x)) \right| \\
&\leq \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \sum_{\substack{x \in [0,u] \\ \Delta S(x) > 0}} \left| \frac{\Delta S(x)}{S(x)} \right| |\Delta \Lambda_n(x) - \Delta \Lambda(x)| \\
&\leq \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |\Delta \Lambda_n(u) - \Delta \Lambda(u)| \sum_{\substack{x \in [0,u] \\ \Delta S(x) > 0}} |\Delta S(x)| \\
&\leq \left(1 + \frac{1}{S(\theta)} \right) \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |\Lambda_n(u_-) - \Lambda(u_-)|
\end{aligned}$$

et ceci pour tout $u \in [0, \theta]$, d'où

$$\sup_{0 \leq u \leq \theta} |K_n(u)| \leq \left(1 + \frac{1}{S(\theta)} \right) \sup_{0 \leq u \leq \theta} |\Lambda_n(u) - \Lambda(u)| + \frac{1}{S(\theta)} \sup_{0 \leq u \leq \theta} |\Lambda_n(u_-) - \Lambda(u_-)|.$$

Ce qui nous permet de conclure que

$$\sup_{0 \leq t \leq \theta} |S_n(t) - S(t)| \leq \frac{(3 - S(\theta))(1 + S(\theta))}{S(\theta)^2} \sup_{0 \leq t \leq \theta} |\Lambda_n(t) - \Lambda(t)| + \frac{3 - S(\theta)}{S(\theta)^2} \sup_{0 \leq t \leq \theta} |\Lambda_n(t_-) - \Lambda(t_-)|. \tag{3.3}$$

Ce qui prouve, en appliquant le résultat montré pour $\Lambda_n(t)$ (qui marche aussi pour $\Lambda_n(t_-)$), la seconde partie du théorème. ■

Remarque 3 *Le point T_H peut être atteint sous certaines conditions, par exemple il est énoncé au corollaire 1.3 de l'article de Stute et Wang (1993) que si F et G n'ont pas de saut en commun alors*

$$\sup_{0 \leq t \leq T_H} |S_n(t) - S(t)| \xrightarrow{p.s.} 0,$$

ssi $\Delta F(T_H) = 0$ ou ($\Delta F(T_H) > 0$ mais $G(T_H-) < 1$).

La LIL permet de préciser le taux de convergence presque sûre, il est de l'ordre de $\sqrt{(\log \log n)/n}$. Elle est donnée, par exemple, dans Földes et Rejtő (1981).

L'EKM possède bien d'autre propriétés, dont sa normalité asymptotique qui a été montrée dans Breslow et Crowly (1974), dont voici l'énoncé.

Théorème 11 *Si $T^* < \infty$ avec $H(T^*) < 1$ et si F et G sont continues alors pour tout $t \leq T^*$, nous avons*

$$\sqrt{n}(S_n(t) - S(t)) \xrightarrow{\mathcal{L}} Z,$$

où Z suit la loi normale centrée et de variance $S^2(t) \int_0^t \frac{dH_1(s)}{\overline{H}^2(s)}$.

Nous citons aussi le résultat suivant, qu'on peut trouver à la page 238 de Földes et al. (1980), pour son utilité dans la suite.

Lemme 2 *Si $T_F < \infty$, $G(T_{F-}) < 1$ et si (ε_n) est une suite de nombres positifs tendant vers zéro avec pour tout $B > 0$, $\sum_{n=1}^{\infty} \varepsilon_n^{-4} \exp(-Bn\varepsilon_n^8) < \infty$ alors*

$$\sum_{n=1}^{\infty} P[\sup_{t \in \mathbb{R}} |S_n(t) - S(t)| > \varepsilon_n] < \infty.$$

EXERCICE 3.1
Soit $X_1, X_{2,\cdots}, X_n$ n v.a.r i.i.d. Posons $T_n = \sup_{1 \leq i \leq n} X_i$. Montrer que
1) Si $T_{X_1} < \infty$, alors $T_n \xrightarrow[n \to \infty]{p.s.} T_{X_1}$.
2) Si X et Y sont deux v.a.r. vérifiant ($X \leq Y$ p.s.), alors $T_X \leq T_Y$.
3) Soient X et Y deux v.a.r. indépendantes, montrer que
$T_{X \wedge Y} = T_X \wedge T_Y$ et $T_{X \vee Y} = T_X \vee T_Y$.

EXERCICE 3.2
Pour tout réel x et tout réel strictement positif t, posons

$$T_{[0,t]}(x) = \begin{cases} t & \text{si} \quad x > t \\ x & \text{si} \quad 0 \leq x \leq t \\ 0 & \text{si} \quad x < 0 \end{cases} .$$

Montrer que
1) $\forall t_1 > 0, t_2 > 0$ et $\forall x \in \mathbb{R}, \left|T_{[0,t_1]}(x) - T_{[0,t_2]}(x)\right| \leq |t_1 - t_2|$.
2) $\forall t > 0$ et $\forall x, y \in \mathbb{R}, \left|T_{[0,t]}(x) - T_{[0,t]}(y)\right| \leq \min(|x - y|, t)$.

EXERCICE 3.3
En utilisant les notations de ce chapitre, a t'on toujours
$S_n(T_H) \longrightarrow S(T_H)$? (indication : considérer le cas $T_H < \infty$, $G(T_{H-}) = 1$ et $\Delta F(T_H) > 0$).

EXERCICE 3.4
Cet exercice utilise aussi les notations de ce chapitre.
Harington et Fleming ont proposé d'estimer $S(t)$ par $S_n^{HF}(t) = \exp(-\Lambda_n(t))$.
Justifier l'introduction de cet estimateur.
En supposant que T admet une densité, montrer que pour tout $t < T_H$, $S_n^{HF}(t) \xrightarrow{p.s.} S(t)$, où $T_H = sup\{t/H(t) < 1\}$.
Une étude portant sur le temps de rémission T après l'atteinte par une maladie a conduit à la collecte des données suivantes (en années) (où + indique une donnée censurée à droite) : 3, 10, 12, 4+, 10+, 0.3, 15, 0.3+, 0.2+, 0.3.
i) Estimer la fonction de hasard cumulé de T.
ii) Estimer la fonction de survie de T par deux méthodes différentes.

EXERCICE 3.5
Soit (T_i, C_i) n paires de v.a. i.i.d. telles que pour tout i, T_i (resp. C_i) positive, de fonction de répartition F (resp. G) avec T_i indépendante de C_i. On suppose que T_i admet une densité continue. Nous observons $X_i = T_i \vee C_i$ et $\delta_i = 1_{(T_i \geq C_i)}$. Introduisons les fonctions de répartitions empiriques suivantes.

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}},\ N_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t, \delta_i = 1\}}.$$

1) Calculer $EZ_n(t)$, $EN_n(t)$, $VAR[\sqrt{n}(Z_n(t) - EZ_n(t))]$, $VAR[\sqrt{n}(N_n(t) - EN_n(t))]$ et $COV(\sqrt{n}N_n(t), \sqrt{n}Z_n(t))$ en fonction des lois de T_i et C_i.
2) En déduire que, pour tout t tel que $F(t)G(t) \neq 0$, $\Gamma(t) = \int_{]t,+\infty[} \frac{dEN_n(s)}{EZ_n(s)}$, où $\Gamma(t) = \int_{]t,+\infty[} \frac{dF(s)}{F(s)}$.
3) Quel estimateur $\hat{F}_n(t)$ vous semble naturel pour estimer $F(t)$?
4) Une étude s'est intéressée au temps de survie à une maladie, dont voici les résultats
$10, 3, 4+, 12, 15, 10+, 0.3+, 0.3, 0.2+, 0.3.$
i) En supposant que les données suivies de + sont censurées à droite, estimer la loi du temps de survie à cette maladie et tracer la courbe de cet estimateur
ii) En supposant que les données suivies de + sont censurées à gauche, cal-

culer $\hat{F}_n(t)$ et tracer sa courbe.

EXERCICE 3.6
Soit T une v.a.r. positive, de fonction de répartition F, nous allons définir le taux de hasard inverse de T, noté h.
1) Si T a pour densité f, posons
$h(t) = \begin{cases} \frac{f(t)}{F(t)} \text{ si } F(t) \neq 0 \\ 0 \text{ sinon} \end{cases}$.
Supposons que f est continue sur $\mathbb{R}_+^*$.
a) Montrer que pout tout t tel que $F(t) \neq 0$, on a
$h(t) = \lim\limits_{\Delta \overset{>}{\to} 0} \frac{P(t-\Delta<T\leq t \ / \ T\leq t)}{\Delta}$.
b) Ecrire F en fonction de h.
2) Si T est une variable discrète prenant les valeurs ordonnées dans l'ordre croissant $x_1, x_2, \ldots, x_n, \ldots$posons
$h(x_j) = P(T = x_j \ / \ T \leq x_j)$.
Montrer que pour tout $x \geq 0$, $F(x) = \prod\limits_{j/x_j>x} (1 - h(x_j))$.
3) En vous inspirant de la méthode d'introduction de l'estimateur de Kaplan-Meier au chapitre 2 du cours, proposer un estimateur de F lorsque T est censurée à gauche. Réecrire cet estimateur lorsqu'il n' y a pas d'ex æquo, déduire alors une condition nécessaire et suffisante pour qu'il s'annule.

Chapter 4

Estimation des fonctions de densité et de régression pour des données censurées à droite

4.1 Estimation de la fonction de densité

Soit T une variable aléatoire positive, censurée à droite par une variable aléatoire C positive et indépendante de T. Rappelons que cela veut dire que nous observons l'échantillon $(X_i = T_i \wedge C_i, \delta_i = 1_{\{T_i \leq C_i\}})_{1 \leq i \leq n}$ de n couples de variables aléatoires i.i.d. et de même loi que $(X = T \wedge C, \delta = 1_{\{T \leq C\}})$. Notons F, G et H les fonctions de répartition respectives de T, C et X. Supposons que $f = F'$ est la densité de probabilité de T. Etendant le cas des données complètes, Földes et al. (1981) ont proposé d' estimer f comme suit:

$$f_n(x) = \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) dF_n(y),$$

où $F_n = 1 - S_n$ (S_n étant l'EKM), K est une densité de probabilité (noyau) et (h_n) est une suite de nombres strictement positifs (fenêtre).
Il es clair que nous retrouvons l'estimateur de Parzen-Rosenblatt si les données sont complètes.

Introduisons les quantités

$$\bar{f}_n(x) = \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) dF(y) = \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) f(y)dy$$

et

$$\tilde{K}(y) = K\left(\frac{x-y}{h_n}\right).$$

Remarquons que $\bar{f}_n(x) = Ef_n(x)$ dans le cas des données complètes mais ceci n'est pas le cas pour des données censurées. Cependant, $\bar{f}_n(x)$ tend vers $f(x)$ (presque sûrement) sous certaines conditions comme nous allons le voir. Dans toute la suite, sauf indication contraire, nous utilisons les mêmes notations que dans les chapitres précédents. En particulier, l'estimation se base sur l'échantillon $(X_i = T_i \wedge C_i, \delta_i = 1_{\{T_i \leq C_i\}})_{1 \leq i \leq n}$, G est la fonction de répartition de la variable de censure et T_F est le point terminal de la variable d'intérêt.
Les résultats suivants donne la consistance forte de l'estimateur f_n (aussi bien la convergence ponctuelle qu'uniforme).

4.1.1 convergence presque sûre

Théorème 12 *(cf Földes et al. (1981) page 22)*
Si f est bornée, $G(T_F^-) < 1$, K est une densité continue à droite (c.a.d.) et à variation bornée et si $h_n \to 0$, $h_n(\frac{n}{\log n})^{1/8} \to \infty$, on a
1) Si f est continue au point x alors $f_n(x) \to f(x)$ p.s.
2) Si $-\infty \leq a \leq b \leq \infty$ et si f est uniformément continue sur $]a, b[$ alors pour tout $c > 0$ tel que $a + c < b - c$

$$\sup_{x \in]a+c, b-c[} |f_n(x) - f(x)| \to 0 \quad p.s.$$

3) Si $-\infty \leq a \leq b \leq \infty$ et si f admet une dérivée bornée sur $]a, b[$ alors

$$\sup_{x \in]a,b[} |f_n(x) - f(x)| \to 0 \quad p.s.$$

Preuve. La preuve est basée sur l'inégalité suivante.

$$|f_n(x) - f(x)| \leq |\bar{f}_n(x) - f(x)| + |f_n(x) - \bar{f}_n(x)|.$$

a) Etude du terme $|\bar{f}_n(x) - f(x)|$
f étant bornée posons $M = \sup |f(x)|$.
• Sous les hypothèses données en 1)
f étant continue au point x, pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que pour tout y / $|y| \leq \delta$, on ait $|f(x-y) - f(x)| \leq \epsilon/2$

Puisque K est une densité, il vient

$$|\bar{f}_n(x) - f(x)| \leq \frac{1}{h_n} \int |f(x-y) - f(x)| K(\frac{y}{h_n}) dy.$$

Il s'ensuit que

$$|\bar{f}_n(x) - f(x)| \leq \frac{1}{h_n} \int_{|y| \leq \delta} |f(x-y) - f(x)| K(\frac{y}{h_n}) dy + \frac{1}{h_n} \int_{|y| > \delta} |f(x-y) - f(x)| K(\frac{y}{h_n}) dy.$$

Donc

$$|\bar{f}_n(x) - f(x)| \leq \epsilon/2 + 2M \int_{|y| > \delta/h_n} K(y) dy.$$

Mais, puisque $h_n^{-1} \to \infty$ l'ensemble $\{|y| > \delta/h_n\} \to \varnothing$ et l'absolue continuité de l'intégrale de Lebesgue entraîne l'existence d'un entier n_0 tel que pour tout $n \geq n_0$, on ait

$$\int_{|y| > \delta/h_n} K(y) dy \leq \frac{\epsilon}{4M}.$$

• Sous les hypothèses données en 2)
Soit $c > 0$, pour $\epsilon > 0$, $\exists \delta > 0$ tel que $\forall x, x' \in]a, b[$ vérifiant $|x - x'| \leq \delta$, on a $|f(x) - f(x')| \leq \epsilon/2$.
Si $x \in]a + c, b - c[$ et $|y| \leq \delta$ (qu'on peut prendre inférieur à c pour assurer le fait que $x - y \in]a, b[$) alors $|f(x) - f(x-y)| \leq \epsilon/2$.
Nous terminons comme en 1) puisque $\int_{|y| > \delta/h_n} K(y) dy$ ne dépend pas de x.
• Sous les hypothèses données en 3)
Soit $L = \sup_{x \in]a,b[} |f'(x)|$ et soient $x \in]a, b[$ et $|y| \leq \epsilon/2L$. On a
$f(x) - f(x-y) = y f'(\theta)$, où θ est compris entre x et $x - y$, donc

$$|f(x) - f(x-y)| \leq |y f'(\theta)| \leq L|y| \leq \epsilon/2,$$

et nous continuons comme ci dessus.
b) Etude du terme $|\bar{f}_n(x) - f_n(x)|$
K vérifiant les conditions requises, la formule d'intégration par parties donne

$$|f_n(x) - \bar{f}_n(x)| = |\frac{1}{h_n} \int \tilde{K}(y) dF_n(y) - \frac{1}{h_n} \int \tilde{K}(y) dF(y)|$$

$$\leq \frac{1}{h_n} \int |F_n(y) - F(y)| d\tilde{K}(y) \leq \frac{1}{h_n} \sup_{y \in \mathbb{R}} |F_n(y) - F(y)| V_K,$$

où V_K est la variation totale de K sur $\mathbb{R}$.
Pour $\epsilon > 0$, posons $\epsilon_0 = \epsilon V_K^{-1}$ et $\epsilon_n = \epsilon_0 h_n$. Notons $\alpha_n = h_n(\frac{n}{\log n})^{1/8}$ alors $\alpha_n \to \infty$ par hypothèse et pour $B > 0$, $\epsilon_n^{-4} \exp\{-Bn\epsilon_n^8\} = \epsilon_0^{-4}(\frac{n}{\log n})^{1/2}\alpha_n^{-4} n^{-B\epsilon_0^8\alpha_n^8}$. L'hypothèse sur α_n permet d'une part de borner α_n^{-4} et d'autre part de choisir la puissance de n dans la formule précédente aussi petite que nous voulons de sorte à obtenir une série de Bertrand convergente, autrement dit $\sum \epsilon_n^{-4} \exp\{-Bn\epsilon_n^8\} < \infty$, le lemme 2 permet alors d'écrire

$$\sum P[\sup |f_n(x) - \bar{f}_n(x)| > \epsilon] \leq \sum P[h_n^{-1} V_K \sup |F_n(x) - F(x)| > \epsilon]$$

$$= \sum P[\sup |F_n(x) - F(x)| > \epsilon_n] < \infty.$$

Finalement le lemme de Borel-Cantelli permet de conclure que

$$\sup_{x \in \mathbb{R}} |f_n(x) - \bar{f}_n(x)| \to 0 \text{ p.s.}$$

■

4.1.2 Convergence presque complète

La convergence presque complète implique la convergence presque sûre et se prête bien aux calculs faisant intervenir des sommes de variables aléatoires. Elle est surtout utilisée en statistique non-paramétrique .

Rappelons que la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \to \infty$ si

$$\forall \epsilon > 0, \sum_{n \in \mathbb{N}} P\left[|X_n - X| > \epsilon\right] < \infty.$$

On dit que la vitesse de convergence presque complète de la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ vers X est d'ordre (u_n) si

$$\exists \epsilon_0 > 0, \quad \sum_{n \in \mathbb{N}} P[|X_n - X| > \epsilon_0 u_n] < \infty.$$

Cette définition du taux a été introduite par Ferraty et Vieu (2006). Elle a l'avantage théorique d'impliquer les deux vitesses de convergence classiques en probabilité et presque sûre, et l'avantage pratique d'être souvent plus facile à démontrer.

Convergence presque complète

Nous allons montrer la convergence presque complète de f_n en un point $x < T_H$, en précisant éventuellement le taux de convergence, sous différentes conditions de régularité de f. Pour cela considérons les hypothèses suivantes.

H1 f est continue au point x.
H2 f est de classe $\mathcal{C}^2$ au voisinage de x.
H3 $\exists k, p, \varepsilon_0 \in \mathbb{R}_+^*$, $\forall y \in]x-\varepsilon_0, x+\varepsilon_0[$, $|f(x)-f(y)| \leq k|x-y|^p$.
H4 $h_n \to 0$ et $nh_n^2/\log n \to \infty$.
H5 K est une densité continue à droite, à variation bornée sur $\mathbb{R}$ et telle que : $\exists M > 0$, $\forall u \in \mathbb{R}$, $|u| \geq M \Rightarrow K(u) = 0$.
H6 K est bornée.
H7 $\int uK(u)\,du = 0$ et $\int u^2K(u)\,du < \infty$.

Théorème 13 *(Boukeloua (2013))*
Soit $x < T_H$,

i) Sous (**H1**), (**H4**), (**H5**) *et* (**H6**), *nous avons:*

$$f_n(x) \xrightarrow{p.co.} f(x),$$

ii) Sous (**H2**), (**H4**), (**H5**) *et* (**H7**), *nous avons:*

$$f_n(x) - f(x) = O_{p.co.}\left(h_n^2 + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),$$

iii) Sous (**H3**), (**H4**) *et* (**H5**), *nous avons:*

$$f_n(x) - f(x) = O_{p.co.}\left(h_n^p + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Ce théorème découle des deux lemmes suivants:

Lemme 3 *Sous* (**H4**) *et* (**H5**), *on a pour tout $\theta < T_H$:*

$$\sup_{x\leq\theta}|f_n(x) - \bar{f}_n(x)| = O_{p.co.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),$$

où, rappelons le, $\bar{f}_n(x) = \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right)dF(y)$.

Preuve. Soient $\theta < T_H$ et $x \leq \theta$, nous avons,

$$|f_n(x) - \bar{f}_n(x)| = \frac{1}{h_n} \left| \int K\left(\frac{x-y}{h_n}\right) d(F_n(y) - F(y)) \right|$$

En posant $u = \frac{x-y}{h_n}$, nous obtenons,

$$|f_n(x) - \bar{f}_n(x)| = \frac{1}{h_n} \left| \int_{-M}^{M} K(u) d(\widetilde{F}_n(u) - \widetilde{F}(u)) \right|$$

avec

$$\widetilde{F}_n(u) = F_n(x - uh_n) \text{ et } \widetilde{F}(u) = F(x - uh_n).$$

En intègrant par parties, il vient

$$|f_n(x) - \bar{f}_n(x)| = \frac{1}{h_n} \left| \int_{-M}^{M} K(u) d(\widetilde{F}_n(u) - \widetilde{F}(u)) \right| \leq \frac{1}{h_n} \left| \int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u)) dK(u) \right|.$$

De plus, K étant à variation bornée et continue à droite s'écrit $K = K_1 - K_2$ où K_1 et K_2 sont deux fonctions croissantes et continues à droite, d'où

$$\begin{aligned}
|f_n(x) - \bar{f}_n(x)| &\leq \frac{1}{h_n} \left| \int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u)) dK_1(u) \right| + \frac{1}{h_n} \left| \int_{-M}^{M} (\widetilde{F}_n(u) - \widetilde{F}(u)) dK_2(u) \right| \\
&\leq \frac{1}{h_n} \sup_{u>-M} |\widetilde{F}_n(u) - \widetilde{F}(u)| \int_{-M}^{M} dK_1(u) + \frac{1}{h_n} \sup_{u>-M} |\widetilde{F}_n(u) - \widetilde{F}(u)| \int_{-M}^{M} dK_2(u) \\
&\leq \frac{V_K}{h_n} \sup_{u>-M} |\widetilde{F}_n(u) - \widetilde{F}(u)| = \frac{V_K}{h_n} \sup_{u>-M} |F_n(x - uh_n) - F(x - uh_n)|,
\end{aligned}$$

où V_K est la variation totale de K sur $\mathbb{R}$.
Puisque $h_n \downarrow 0$, $\exists n_0 \in \mathbb{N} / \forall n \geq n_0 : Mh_n < \theta^* - \theta$ où $\theta^* \in]\theta, T_H[$, donc $u > -M \Rightarrow x - uh_n < \theta^*$ et par conséquent $\sup_{u>-M} |F_n(x-uh_n) - F(x-uh_n)| \leq \sup_{t<\theta^*} |F_n(t) - F(t)|$.
Donc pour tout $x \leq \theta$ nous avons

$$|f_n(x) - \bar{f}_n(x)| \leq \frac{V_K}{h_n} \sup_{t<\theta^*} |F_n(t) - F(t)|. \tag{4.1}$$

Il reste à appliquer un résultat de type borne DKW pour l'EKM, énoncé comme suit.

Théorème 14 *(Theorem 1 dans Bitouzé et al. (1999))*
Il existe une constante $C > 0$ telle que pour tout $\lambda > 0$, on a

$$P\left(\sqrt{n}\sup_{t\in\mathbb{R}}((1-G(t))|F_n(t)-F(t))| > \lambda\right) \leq 2.5\, e^{-2\lambda^2+C\lambda},$$

nous pouvons déduire que

$$P\left(\sup_{t<\theta^*}|F_n(t)-F(t)| > \frac{1}{1-G(\theta^*)}\sqrt{\frac{\log n}{n}}\right) \leq 2.5\, e^{-2\log n+C\sqrt{\log n}}.$$

En choisissant $\alpha < 1$ et pour n assez grand, nous obtenons $e^{C\sqrt{\log n}} \leq n^{\alpha}$, ce qui implique que

$$\sup_{t<\theta^*}|F_n(t)-F(t)| = O_{p.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{4.2}$$

Ceci combiné avec (4.1) donne

$$\sup_{x\leq\theta}|f_n(x)-\bar{f}_n(x)(x)| \leq \frac{V_K}{h_n}\sup_{t<\theta^*}|F_n(t)-F(t)| = O_{p.co.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right). \tag{4.3}$$

■

Lemme 4 *Soit $x < T_H$,*

i) Sous (**H1**), (**H4**), (**H5**) *et* (**H6**), *nous avons:*

$$\bar{f}_n(x) \underset{n\to\infty}{\longrightarrow} f(x),$$

ii) Sous (**H2**), (**H4**), (**H5**) *et* (**H7**), *nous avons:*

$$\bar{f}_n(x) - f(x) = O(h_n^2),$$

iii) Sous (**H3**), (**H4**) *et* (**H5**), *nous avons:*

$$\bar{f}_n(x) - f(x) = O(h_n^p).$$

Preuve.

i) En utilisant le changement de variable $z = x - y$, nous pouvons écrire

$$\bar{f}_n(x) = \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) f(y)dy = \frac{1}{h_n}\int K\left(\frac{z}{h_n}\right) f(x-z)dz \underset{n\to\infty}{\longrightarrow} f(x),$$

d'après le théorème de Bochner.

ii) En utilisant le changement de variable $u = \frac{x-y}{h_n}$, il s'ensuit que:

$$\begin{aligned}
|\bar{f}_n(x) - f(x)| &= \left|\frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right) f(y)dy - f(x)\right| \\
&= \left|\int K(u)f(x-uh_n)du - f(x)\right| \\
&= \left|\int_{-M}^{M} K(u)f(x-uh_n)du - f(x)\int_{-M}^{M} K(u)du\right| \\
&= \left|\int_{-M}^{M} K(u)(f(x-uh_n) - f(x))du\right|, \qquad (4.4)
\end{aligned}$$

comme f est de classe $\mathcal{C}^2$ au voisinage de x, le développement de Taylor à l'ordre 2 donne

$$\begin{aligned}
|\bar{f}_n(x) - f(x)| &= \left|\int_{-M}^{M} K(u)\left[f(x) - uh_n f'(x) + \frac{u^2h_n^2}{2}f''(\eta_n) - f(x)\right]du\right| \\
&= \frac{h_n^2}{2}\left|\int_{-M}^{M} f''(\eta_n)u^2K(u)du\right|, .
\end{aligned}$$

où η_n est compris entre x et $x - uh_n$.
Par ailleurs, f'' étant continue au point x, pour tout $\varepsilon > 0$
$\exists\delta > 0/\forall y, |y - x| < \delta \Rightarrow |f''(y) - f''(x)| < \varepsilon$,
et comme $h_n \downarrow 0$, $\exists n_0 \in \mathbb{N}/\forall n \geq n_0 : h_n < \frac{\delta}{M}$, donc pour tout $n \geq n_0$
nous avons: $|\eta_n - x| \leq |u|h_n < \delta \Rightarrow |f''(\eta_n) - f''(x)| < \varepsilon$, d'où:

$$\begin{aligned}
|\bar{f}_n(x) - f(x)| &\leq \frac{h_n^2}{2}\int_{-M}^{M} |f''(\eta_n) - f''(x)|u^2K(u)du + \frac{|f''(x)|h_n^2}{2}\int_{-M}^{M} u^2K(u)du \\
&\leq \left[\frac{\varepsilon + |f''(x)|}{2}\int_{-M}^{M} u^2K(u)du\right]h_n^2 = O(h_n^2). \qquad (4.5)
\end{aligned}$$

iii) Selon (4.4), nous avons:

$$|\bar{f}_n(x) - f(x)| \leq \int_{-M}^{M} K(u)|f(x - uh_n) - f(x)|du,$$

et comme $h_n \downarrow 0$, $\exists n_1 \in \mathbb{N}/\forall n \geq n_1 : h_n < \frac{\varepsilon_0}{M}$, donc pour tout $n \geq n_1$ nous avons: $|x - uh_n - x| = |u|h_n < \varepsilon_0 \Rightarrow |f(x - uh_n) - f(x)| \leq k|u|^p h_n^p \leq kM^p h_n^p$, d'où:

$$|\bar{f}_n(x) - f(x)| \leq kM^p h_n^p \int_{-M}^{M} K(u)du = kM^p h_n^p = O(h_n^p).$$

■ Dans la suite, nous allons donner une version uniforme du théorème 13 (les preuves sont identiques en appliquant dans chaque cas les hypothèses correspondantes). Soit C un compact inclus dans $]-\infty, T_H[$ et considérons les hypothèses suivantes de régularité de f sur C:

H8 f est continue sur C.

H9 f est de classe $\mathcal{C}^2$ sur un ouvert contenant C.

H10 $\exists k, p, \varepsilon_0 \in \mathbb{R}_+^*, \forall x \in C, \forall y \in]x - \varepsilon_0, x + \varepsilon_0[, |f(x) - f(y)| \leq k|x - y|^p$.

Théorème 15 *(Boukeloua (2013))*

i) *Sous* (**H8**), (**H4**), (**H5**) *et* (**H6**), *nous avons:*

$$\sup_{x \in C} |f_n(x) - f(x)| \xrightarrow{p.co.} 0,$$

ii) *Sous* (**H9**), (**H4**), (**H5**) *et* (**H7**), *nous avons:*

$$\sup_{x \in C} |f_n(x) - f(x)| = O_{p.co.}\left(h_n^2 + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),$$

iii) *Sous* (**H10**), (**H4**) *et* (**H5**), *nous avons:*

$$\sup_{x \in C} |f_n(x) - f(x)| = O_{p.co.}\left(h_n^p + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Signalons que bien d'autres travaux concernant cet estimateur existent, puisqu' il n'est pas possible de tous les citer contentons nous de faire remarquer que sa convergence en moyenne quadratique est donnée dans Kagba (2004) et que sa normalité asymptotique est prouvée dans Mielniczuk (1986) et améliorée dans Diehl et Stute (1988).

4.2 Estimation de la fonction de régression

Dans ce paragraphe, la variable réponse à valeurs dans $\mathbb{R}$ est notée Y et la covariable à valeurs dans $\mathbb{R}^d$ ($d \in \mathbb{N}^*$) est notée V. Nous nous proposons d'estimer $r(x) = E(Y/V = x)$, les estimateurs donnés précédemment ne peuvent plus être utilisés puisque Y, étant censurée à droite, n'est pas toujours observable. L'estimation se base sur l'échantillon $(V_i, Z_i, \delta_i)_{(1\leq i\leq n)}$ de v.a. i.i.d., tiré de (V, Z, δ) où $Z = Y \wedge C$ et $\delta = 1_{\{Y\leq C\}}$ est l'indicateur de censure et Y et C sont supposées indépendantes.

4.2.1 Principe de l'estimation

L'idée, introduite par Carbonez et al. (1995), et reprise par Kohler et al. (2002), est de remplacer $W_{n,i}(x)Y$ par un estimateur déduit de l'estimation de sa moyenne.
Soient $S(t) = P(Y > t)$ et $\overline{G}(t) = P(C > t)$ les fonctions de survie respectives de Y et C.
On suppose que

$$(H_1) : \begin{cases} (H_{1.1}) & C \text{ et } (V, Y) \text{ sont indépendants et } \overline{G} \text{ est continue,} \\ (H_{1.2}) & T_Y < \infty \text{ et } \overline{G}(T_Y) > 0. \end{cases}$$

Remarquons que la condition $\overline{G}(T_Y) > 0$ implique que $T_Y < T_C$.
Soit h une fonction mesurable de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. On se propose d'estimer la moyenne $E\{h(V, Y)\}$ sur la base de l'échantillon des données censurées à droite. Un "estimateur" sans biais de $E\{h(V, Y)\}$ est donné par :

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i h(V_i, Z_i)}{\overline{G}(Z_i)}.$$

En effet, l'utilisation de l'indépendance de (V, Y) et de C et des propriétés de l'espérance conditionnelle, nous permet d'écrire

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i h(V_i, Zi)}{\overline{G}(Z_i)}\right\} &= E\left\{\frac{1_{\{Y_1\leq C_1\}} h(V_1, Y_1)}{\overline{G}(Y_1)}\right\} \\ &= E\left(\frac{h(V_1, Y_1)}{\overline{G}(Y_1)} E\left(1_{\{Y_1\leq C_1\}}/(V_1, Y_1)\right)\right) \\ &= E\left(h(V_1, Y_1)\right), \end{aligned}$$

où la dernière étape est explicitée à l'exercice 4.2.
Le problème est que la fonction $\overline{G}$ est, en général, inconnue. On l'estime par l'EKM donné par :

$$\hat{G}_n(t) = \begin{cases} \prod_{i=1}^{n} \left[1 - \frac{1-\delta_{(i)}}{n-i+1}\right]^{1_{[Z_{(i)} \leqslant t]}}, & \text{si } t < M_n, \\ \lim_{s \to M_n, s < M_n} \hat{G}_n(s), & \text{si } t \geqslant M_n, \end{cases},$$

où $M_n = \max\{Z_1, ..., Z_n\}$ et les paires $(Z_{(i)}, \delta_{(i)})$, $i = 1, \ldots, n$ sont les n paires observées (Z_i, δ_i) ordonnées en $Z_{(i)}$, i.e. $Z_{(1)} \leqslant Z_{(2)} \leqslant ... \leqslant Z_{(n)} = M_n$ et $\delta_{(i)}$ est le concomitant de $Z_{(i)}$.
Remarquons que $\hat{G}_n$ a été légèrement modifié afin de ne pas s'annuler.
Cela suggère d'estimer $r(x)$ par

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x) \frac{\delta_i Z_i}{\hat{G}_n(Z_i)}, \tag{4.6}$$

où $W_{n,i}(x)$ est la fonction poids de Nadaraya-Watson, définie comme suit,

$$W_{n,i}(x) = \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})}.$$

Pour $f : \mathbb{R}^d \to \mathbb{R}$ définissons $T_{[0,t]}f : \mathbb{R}^d \to \mathbb{R}$ par $(T_{[0,t]}f)(x) = T_{[0,t]}(f(x))$ où l'application $T_{[0,t]}(x)$ a été définie à l'exercice 3.2.
Du fait que $0 \leqslant Y \leqslant T_Y < \infty$ p.s, on a $0 \leqslant r(x) \leqslant T_Y$, on estime donc $r(x)$ plutôt par

$$r_n(x) := T_{[0,M_n]}(\hat{r}_n(x)) = \begin{cases} M_n & \text{si } \hat{r}_n(x) > M_n, \\ \hat{r}_n(x) & \text{si } 0 \leqslant \hat{r}_n(x) \leqslant M_n, \\ 0 & \text{si } \hat{r}_n(x) < 0. \end{cases} \tag{4.7}$$

Par analogie avec (4.7), posons

$$r_n^*(x) := T_{[0,T_Y]}(\hat{r}_n(x)) = \begin{cases} T_Y & \text{si } \hat{r}_n(x) > T_Y, \\ \hat{r}_n(x) & \text{si } 0 \leqslant \hat{r}_n(x) \leqslant T_Y, \\ 0 & \text{si } \hat{r}_n(x) < 0. \end{cases}$$

4.2.2 Propriétés de l'estimateur

Nous allons faire usage du lemme suivant.

Lemme 5 *Sous l'hypothèse* $(H_{1.2})$, *on a*

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

si et seulement si

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ \ p.s.$$

où μ *représente la loi de* V.

Preuve.
Puisque

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx)$$

et

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |r_n^*(x)(x) - r(x)|^2 \mu(dx),$$

il suffit de voir que :

$$\int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ \text{ p.s.}$$

Or, on a $M_n \leqslant T_Y$ p.s, l'application de l'exercice 3.2 permet alors d'écrire

$$|r_n^*(x) - r_n(x)| = |T_{[0,T_Y]}(\hat{r}_n(x)) - T_{[0,M_n]}(\hat{r}_n(x))| \leq T_Y - M_n.$$

Donc

$$\int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) \leq \int_{\mathbb{R}^d} (T_Y - M_n)^2 \mu(dx) \to 0,$$

car sous $H_{1,2}$, $M_n \to T_Y$ p.s. $(n \to \infty)$ (voir l'exercice 3.1).
La condition suffisante se montre de la même manière en rajoutant et retranchant $r_n^*(x)$ dans $|r_n(x) - r(x)|^2$. ■
Nous sommes maintenant en mesure de montrer le théorème suivant.

Théorème 16 *(Kohler et al. (2002))*
Si K est un noyau régulier, $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d = \infty$ et si l'hypothèse (H_1) est satisfaite, alors l'estimateur $r_n(x)$ défini en (4.6) et (4.7) vérifie:

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

Preuve. D'après le lemme précédent, il suffit de montrer que :

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

Posons

$$\bar{r}_n(x) = T_{[0,T_Y]} \left(\sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} \frac{\delta_i Z_i}{\overline{G}(Z_i)} \right).$$

Nous avons

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \, \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n^*(x) - \bar{r}_n(x) \mu(dx) + 2 \int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx). \tag{4.8}$$

Commençons par majorer le second terme du membre de droite de l'inégalité précédente.
Puisque $0 \leq r(x) \leq T_Y$ p.s. alors $r(x) = T_{[0,T_Y]}(r(x))$ p.s., nous obtenons donc en vertu de l'exercice 3.2

$$\int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx) \leqslant \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} \frac{\delta_i Z_i}{\overline{G}(Z_i)} - r(x)|^2 \mu(dx).$$

L'application de l'hypothèse $H_{1,2}$ permet d'écrire
$0 \leqslant \delta_i Z_i / \overline{G}(Z_i) \leqslant T_Y / \overline{G}(T_Y)$ *p.s.* et grâce à l'hypothèse $H_{1,1}$, nous avons $E(1_{\{Y_1 \leq C_1\}} | (V_1, Y_1) = \overline{G}(Y_1)$ (voir exercice 4.2), donc

$$E\left\{\frac{\delta_1 Z_1}{\overline{G}(Z_1)}/V_1\right\} = E\left\{\frac{1_{\{Y_1\leq C_1\}}Y_1}{\overline{G}(Y_1)}/V_1\right\}$$
$$= E\left(\frac{Y_1}{\overline{G}(Y_1)}E\left(1_{\{Y_1\leq C_1\}}/(V_1,Y_1)\right)/V_1\right)$$
$$= E\left(Y_1/V_1\right) = r(V_1),$$

le théorème de Devroye et Krzyżak (1989) (voir théorème 5) entraîne que

$$\int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} . \frac{\delta_i Z_i}{\overline{G}(Z_i)} - r(x)|^2 \mu(dx) \to 0 \ (n\to\infty) \ p.s \tag{4.9}$$

Reste à majorer le premier terme du second membre de l'inégalité donnée à la formule (4.8). En vertu de la question 2) de l'exercice 3.2, il vient

$$\int_{\mathbb{R}^d} |r_n^*(x) - \bar{r}_n(x)|^2 \mu(dx)$$
$$\leq T_Y \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} \frac{\delta_i Z_i}{\hat{G}_n(Z_i)} - \sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} \frac{\delta_i Z_i}{\overline{G}(Z_i)} | \mu(dx)$$

$$\leq T_Y \int_{\mathbb{R}^d} \sum_{i=1}^{n} \frac{K(\frac{(x-V_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-V_j)}{h_n})} T_Y | \frac{1}{\hat{G}_n(Zi)} - \frac{1}{\overline{G}(Z_i)} | \mu(dx)$$
$$\leq T_Y^2 \frac{1}{\hat{G}_n(T_Y)G(T_Y)} \sup_{t\leq T_Y} |\overline{G}(t) - \hat{G}_n(t)| \to 0 \ \ (n\to\infty) \ p.s.$$

en raison de $(H_{1,2})$ et du fait que $\sup_{t\leq T_Y} |\overline{G}(t) - \hat{G}_n(t)| \to 0$ p.s. (le point T_Y est inclus en application de la remarque 3 du chapitre précédent). ■
Signalons le fait que Guessoum et Ould-Said (2008) ont modifié légèrement $\hat{G}_n$ en lui imposant de s'annuler à partir de l'observation la plus grande. Ils

ont alors, d'une part établi la convergence presque sûre uniforme sur des compacts de l'estimateur ainsi obtenu, donné des vitesses de convergence et prouvé d'autre part sa normalité asymptotique. Remarquons que le résultat donné au théorème précédent a été aussi prouvé dans Kohler et al. (2002) pour des estimateurs à poids (plus proches voisins et à partitions) ainsi que pour des estimateurs des moindres carrés et des estimateurs spline de lissage dans un modèle de censure à droite.

EXERCICE 4.1
Soit l_x et l_y deux nombres réels et (u_n) une suite de nombres réels tels que $\lim_{n\to\infty} u_n = 0$, montrer que si

1. Si $\lim_{n\to\infty} X_n = l_x$ *p.co.* et $\lim_{n\to\infty} Y_n = l_y$ *p.co.*, alors
   - $\lim_{n\to\infty}(X_n + Y_n) = l_x + l_y$ *p.co.*,
   - $\lim_{n\to\infty}(X_n Y_n) = l_x l_y$ *p.co.*,
   - $\lim_{n\to\infty} \frac{1}{X_n} = \frac{1}{l_x}$ *p.co.* lorsque $l_x \neq 0$.
2. Si $X_n - l_x = O_{p.co}(u_n)$ *et* $Y_n - l_y = O_{p.co}(u_n)$, alors
   - $(X_n + Y_n) - l_x - l_y = O_{p.co.}(u_n)$,
   - $(X_n Y_n) - l_x l_y = O_{p.co.}(u_n)$,
   - $\frac{1}{X_n} - \frac{1}{l_x} = O_{p.co.}(u_n)$ lorsque $l_x \neq 0$.
3. Si $X_n = O_{p.co.}(u_n)$ et $\lim_{n\to\infty} Y_n = l_y$ p.co., alors
   - $X_n Y_n = O_{p.co.}(u_n)$,
   - $\frac{X_n}{Y_n} = O_{p.co.}(u_n)$, lorsque $l_y \neq 0$.

EXERCICE 4.2
En utilisant les notations du cours, montrer que si h est une application mesurable de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ alors sous l'hypothèse $H_{1,1}$, nous avons

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(V_i, Z_i)}{\overline{G}(Z_i)}\right) = E(h(V_1, Y_1)).$$

EXERCICE 4.3
Nous proposons l'estimation du taux de hasard λ d'une v.a.r. T censurée à droite par une v.a.r. C qui lui est indépendante. Il s'agit de montrer la consistance des estimateurs introduits.

En notant S_n l'estimateur de Kaplan Meier (L'EKM) et f_n l'estimateur à noyau de la densité f basé sur des données censurées à droite, nous proposons d'estimer λ par

$$\lambda_n(x) = \frac{f_n(x)}{S_n(x) + u_n},$$

où la suite de réels $(u_n)_{n\in\mathbb{N}}$ est telle que $\forall n,\ u_n > 0$ et $u_n \underset{n\to\infty}{\longrightarrow} 0$.
1) Comment justifiez vous l'introduction de l'estimateur λ_n pour estimer λ ?
2) Montrer qu'on a pour $S(x) \neq 0$,

$$\lambda_n(x) - \lambda(x) = \frac{f_n(x) - f(x)}{S_n(x) + u_n} + (S(x) - S_n(x) - u_n)\frac{f(x)}{S(x)(S_n(x) + u_n)} \quad .$$

Soit $x < T_H := \sup\{t : P(T \wedge C \leq t) < 1\}$.
3) Montrer que sous (**H1**), (**H4**), (**H5**) et (**H6**) (rappelés ci joint), nous avons

$$\lambda_n(x) \overset{p.co.}{\longrightarrow} \lambda(x).$$

4) En partant de l'estimateur de Nelson Aalen Λ_n, proposez un estimateur à noyau $g_n(x)$ pour $\lambda(x)$ (inspirez vous de la construction de f_n). En admettant que pour $\theta < T_H$, $\sup_{t\leq\theta} |\Lambda_n(t) - \int_0^t \lambda(x)dx| \overset{p.co.}{\longrightarrow} 0$, donner un résultat de convergence presque complète pour $g_n(x)$ en précisant les hypothèses de validité du résultat.

RAPPELS

H1 f est continue au point x.

H4 $h_n \to 0$ et $nh_n^2/\log n \to \infty$.

H5 K est une densité continue à droite, à variation bornée sur $\mathbb{R}$ et telle que $\exists M > 0$, $\forall u \in \mathbb{R}$, $|u| \geq M \Rightarrow K(u) = 0$.

H6 K est bornée

Chapter 5

Introduction à l'estimation non paramétrique pour des données censurées à droite ou (et) à gauche

5.1 Estimation de la fonction de survie

A notre connaissance, Turnbull (1974) a été le premier à s'intéresser à l'estimation non paramétrique de la fonction de survie d'une variable d'intérêt T, de fonction de survie S, dans un modèle que nous nommons de censure double, c'est à dire que l'observation consiste en $(X = \max(\min(T, D), G), A)$, où D et G sont des variables de censure et

$$A = \begin{cases} 0 & \text{si } G \leq T \leq D, \\ 1 & \text{si } D < T, \\ 2 & \text{si } T < G, \end{cases}$$

avec $P(G \leq D) = 1$. Dans ce modèle T est observée ssi $T \in [G, D]$ et une donnée censurée l'est soit à droite soit à gauche mais pas les deux à la fois.
IL a utilisé le critère de self consistance pour construire un estimateur pour S. Ce dernier est donc donné implicitement et se calcule numériquement par l'algorithme EM.
Chang et Yang (1987) ont prouvé la consistance forte de cet estimateur malgré la difficulté inhérente au fait que nous ne disposons pas d'une écriture

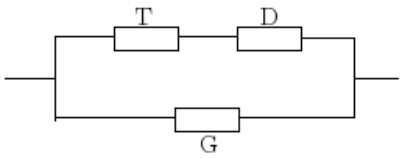

Figure 5.1: Model I

explicite de l'estimateur en question. Ensuite, Chang (1990) a montré sa convergence en loi (en tant que processus).
Dans Turnbull (1976), l'estimation de la fonction de survie a été considérée dans un modèle de censure par intervalle accompagnée (ou non) d'une troncature. L'idée de self consistance y a encore été utilisée et un test de comparaison de deux populations a été proposé dans le cas où au moins un des deux échantillons observés est incomplet (dans le sens précédent).
Mais dans certaines situations, ces deux derniers modèles ne s'appliquent pas. Ainsi, en fiabilité prenons l'exemple d'un système formé de trois composants A_1, A_2 et A_3 avec A_1 et A_2 placés en série et A_3 est placé en parallèle avec le système en série (voir figure 5.1).
Le but est d'étudier le temps de fonctionnement du composant A_2, noté T mais l'observation consiste en $(X = \max(\min(T, D), G), A)$, où

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

et D (resp. G) est le temps de fonctionnement du composant A_1 (resp. A_3). Ici, il n'est pas raisonnable d'imposer l'hypothèse $G \leq D$ p.s. Le but est d'estimer la fonction de survie S de T sur la base d'un échantillon $(X_i = \max(\min(T_i, D_i), G_i), A_i)$ de v.a. i.i.d. de même loi que (X, A) en supposant l'indépendance des variables latentes T, D et G (que nous appelons modèle de censure mixte). Ce problème a été solutionné par Patilea et Rolin (2006). En procédant d'abord comme dans un modèle de censure à gauche, pour estimer la fonction de répartition de $\min(T, D)$, puis comme dans un modèle de censure à droite, ils ont proposé d'estimer S comme suit.
Notons par $\{Z_j, 1 \leq j \leq M\}$ les valeurs distinctes de $\{X_i, 1 \leq i \leq n\}$, rangées dans l'ordre croissant. Posons

$D_{kj} = \sum_{i=1}^{n} 1_{\{X_i = Z_j, A_i = k\}}$ et $N_j = \sum_{i=1}^{n} 1_{\{X_i \leq Z_j\}}$.
L'estimateur produit limite de Patilea et Rolin (2006), est donné par

$$S_n(t) = \prod_{j/Z_j \leq t} \{1 - D_{0j}/(U_{j-1} - N_{j-1})\} \text{ où } U_{j-1} = n \prod_{j \leq l \leq M} \{1 - D_{2l}/N_l\}. \tag{5.1}$$

Remarquons que nous retrouvons l'EKM si $G \equiv 0$, ce qui correspond au cas de la seule censure à droite (à faire en exercice). De plus, cet estimateur à l'avantage, par rapport à celui de Turnbull (1974), d'être explicitement donné. Remarquons, que contrairement au modèle de Turnbull (1974), une même donnée peut être censurée à droite et à gauche en même temps (si $A = 2$, on peut avoir $T > D$ et $T < L$).
Sous une hypothèse liant leurs supports, Patilea et Rolin (2006) ont prouvé la convergence presque sûre uniforme de S_n sur tout $\mathbb{R}$. En étant un peu plus exigeant sur les relations liant les supports de ces variables et en imposant la continuité de leurs fonctions de survie, Messaci et Nemouchi (2011) ont précisé le taux de convergence qui est aussi de l'ordre de $\sqrt{(\log\log n)/n}$.
Kitouni et al. (2015) ont donné un taux de convergence presque complète de l'ordre de $\sqrt{(\log n)/n}$ en se passant de l'hypothèse de continuité. Autrement dit, ils ont amélioré le mode de convergence en allégeant les hypothèses mais au prix de la détérioration du taux.

5.2 Estimation de la fonction de densité et du taux de hasard

Les estimateurs à noyau de la densité et du taux de hasard, dans un modèle de censure double, ont été étudiés dans Ren (1997). Il y a été prouvé leurs consistances fortes (uniformément) et leurs normalités asymptotiques. Quant au cas du modèle de censure mixte, Kitouni et al. (2015) ont proposé des estimateurs à noyaux pour les fonctions de densité et de taux de hasard pour lesquels ils ont donné des taux de convergence presque complète, travail complété par Boukeloua (2015) qui a prouvé des taux de convergence en moyenne quadratique pour ces estimateurs. Puis, dans Boukeloua et Messaci (2016) a été introduite l'idée de censure générale qui a permis d'unifier l'étude, de la normalité asymptotique d'estimateurs à noyau des fonctions de densité et du taux de hasard, pour plusieurs cas de censure. Le résultat obtenu a été donc appliqué à trois cas connus de censure. Le premier, à savoir la censure à

droite très fréquente dans la pratique, a permis d'améliorer légèrement (en allégeant un peu les hypothèses) le résultat établi en 1986 par Mielniczuk. Le second cas, consacré à la censure double a également permis d'améliorer, en termes d'hypothèses moins contraignantes, un résultat déjà connu et dû à Ren en 1997. Le dernier cas, qui concerne la censure mixte, est un résultat établi pour la première fois dans le cadre de cet article.
Par ailleurs est née l'idée de proposer l'estimation de la fonction de régression, pour le modèle de censure mixte, sur laquelle nous donnons un petit aperçu ci dessous.

5.3 Estimation de la fonction de régression

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^n$, constituant la variable explicative et Y la variable réponse supposée positive. Soient R et G deux v.a. positives et bornées telles que G est une censure à gauche opérant sur $\min(Y, D)$.
Nous voulons estimer la fonction de régression $r(x) = E(Y/X = x)$ sur la base de l'échantillon de v.a. i.i.d.$(X_i, Z_i = \max(\min(Y_i, D_i), G_i), A_i)$, où

$$A_i = \begin{cases} 0 & \text{si } G_i < Y_i < D_i, \\ 1 & \text{si } G_i < D_i \leq Y_i, \\ 2 & \text{si } \min(Y_i, D_i) \leq G_i. \end{cases}$$

Dans Messaci (2010), sont proposés des estimateurs à poids donnés par

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_n(Z_i)F_n(Z_i)} \quad \left(\frac{0}{0} := 0\right),$$

où
• $W_{n,i}(x)$ est une fonction poids ($= \frac{K((x-X_i)/h_n)}{\sum_{i=1}^{n} K((x-X_i)/h_n)}$ pour l'estimateur à noyau par exemple),
• S_n est l'estimateur de la fonction de survie de D,
• F_n est l'estimateur de la fonction de répartition de G (qu'on peut déduire de l'EKM par inversion du temps : modèle de censure à gauche).
Les résultats de Kohler et al. (2002), concernant la seule censure à droite, ont été généralisés aussi bien pour les estimateurs à poids que pour les estimateurs des moindres carrés (voir Kebabi et al. (2011)).

Le travail de Guessoum et Ould-Said (2008), concernant la seule censure à droite, a été étendu puisque Kebabi et Messaci (2012) ont montré la convergence presque complète uniforme de l'estimateur à noyau de la fonction de régression quant la variable réponse est soumise à une censure mixte.

EXERCICE 5.1
Montrer que l'estimateur de Patilea et Rolin donné à l équation (5.1) généralise l'EKM.

EXERCICE 5.2
Trouver une condition nécessaire et suffisante pour que l'estimateur de Patilea et Rolin donné à l équation (5.1) s'annule.

EXERCICE 5.3
Pour toute variable aléatoire réelle (v.a.r.) X, on note par F_X (resp. S_X) sa fonction de répartition (resp. survie). Dans notre modèle la variable d'intérêt T est censurée à gauche par la v.a.r. positive G et la v.a.r. $\max(T, G)$ est elle même censurée à droite par la v.a.r. positive D, c'est à dire que nous observons le couple $(Y = \min(\max(T, G), D), A)$, où

$$A = \begin{cases} 0 & \text{si } G < T \leq D, \\ 1 & \text{si } D \leq \max(T, G), \\ 2 & \text{si } T \leq G \leq D. \end{cases}$$

Nous supposons que les variables T, D et G sont indépendantes. Posons $Z = \max(T, G)$ et notons pour tout $k \in \{0, 1, 2\}$, $H_k(t) = P(Y \leq t, A = k)$.
1) Pour quelles valeurs de k a t'on des données censurées, justifier.
2) Ecrire F_Y en fonctions des sous lois H_k.
3) Montrer que $dH_0(t) = S_D(t_-)F_G(t_-)dF_T(t)$ où t_- désigne la limite à gauche de t.
4) Trouver des relations analogues pour $dH_1(t)$ et $dH_2(t)$.
5) Posons $H_{0,2} = H_0 + H_2$, montrer que $dH_{0,2}(t) = S_D(t_-)dF_Z(t)$.
6) Ecrire $d\Gamma(t)$ en fonction de dH_0, S_D et F_Z où $d\Gamma(t) = \frac{dF_T(t)}{F_T(t_-)}$. Par quoi peut on estimer S_D et F_Z ? en déduire un estimateur pour $d\Gamma(t)$ et enfin pour $F(t)$ (il n'est pas demandé de les écrire explicitement mais d'expliquer seulement les idées à utiliser).

Chapter 6

Tests de comparaison de populations basés sur des données censurées

Tous les chapitres précédents ont traité de la problématique de l'estimation, qui est un des deux volets de la statistique. Nous allons finaliser ce document en introduisant quelques tests non paramétriques, qui rentrent dans le cadre du second volet. Le temps alloué à ce cours ne permet pas d'aller plus loin mais nous espérons pouvoir donner une idée sur des tests non paramétriques basés sur des données censurées à droite.
Nous allons aborder le problème de comparaison de k ($k \geq 2$) populations en se basant sur leurs taux de hasard λ_i ($1 \leq i \leq k$), pour cela testons l'hypothèse suivante :
$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t)$, pour $t \leq \tau$,
contre
H_1 : Au moins un des $\lambda_i(t)$ est différent pour $t \leq \tau$,
où τ est le temps maximum de l'étude.
Les échantillons recueillis pour mener le test peuvent comporter des données censurées à droite (et même tronquées à gauche) pour chacune des k populations.
Soient $t_1 < t_2 < \ldots < t_D$ les différents temps de morts (autrement dit des données complètes) dans l' échantillon global obtenu par concaténation de tous les échantillons observés. Nous notons
d_{ij} : le nombre de données complètes dans le j ème échantillon au temps t_i,
Y_{ij} : le nombre d'individus à risque dans le j ème échantillon au temps t_i,

$j = 1, \ldots, k$ et $i = 1, \ldots, D$.
Soient $d_i = \sum_{j=1}^{k} d_{ij}$ et $Y_i = \sum_{j=1}^{k} Y_{ij}$ respectivement le nombre de données complètes et le nombre d'individus à risque dans l'échantillon global au temps $t_i; i = 1, \ldots, D$.
Le test proposé est basé sur des comparaisons pondérées de l'estimateur du taux de hasard de la j ème population à celui de la population globale et utilise l'estimateur de Nelson-Aalan.
Si l'hypothèse nulle est vraie alors le taux de hasard espéré dans la j ème population devrait pouvoir être estimé par le taux de hasard observé dans l'échantillon global, c'est à dire $\frac{d_i}{Y_i}$.
En utilisant les données du j ème échantillon, l'estimateur du taux de hasard est $\frac{d_{ij}}{Y_{ij}}$.
Soit, $W_j(t)$ une fonction positive de poids avec la propriété que $W_j(t_i) = 0$ lorsque $Y_{ij} = 0$ ($\frac{0}{0} := 0$).
Le test proposé est basé sur la statistique

$$U_j(\tau) = \sum_{i=1}^{D} W_j(t_i)\Big\{\frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i}\Big\}, \quad j = 1, \ldots, k. \tag{6.1}$$

Si toutes les valeurs $U_j(\tau)$ sont proches de zéro alors il est peu probable que l'hypothèse nulle soit fausse, tandis que, si une des $U_j(\tau)$ est loin de zéro alors nous pouvons admettre que la j ème population a un taux de hasard différent de celui attendu sous H_0.
Quoique ce test peut théoriquement être mené pour différentes fonctions de poids, dans la pratique les poids les plus fréquemment utilisés sont :

$$W_j(t_i) = Y_{ij}W(t_i),$$

$W(t_i)$ est la contribution de chaque groupe, et avec ce choix il vient

$$U_j(\tau) = \sum_{i=1}^{D} W(t_i)\Big\{d_{ij} - Y_{ij}\Big(\frac{d_i}{Y_i}\Big)\Big\}, \quad j = 1, \ldots, k. \tag{6.2}$$

La variance de $U_j(\tau)$ est donnée par

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\Big(1 - \frac{Y_{ij}}{Y_i}\Big)\Big(\frac{Y_i - d_i}{Y_i - 1}\Big)d_i, \quad j = 1, \ldots, k, \tag{6.3}$$

et la covariance de $(U_j(\tau), U_g(\tau))$ est

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}Y_{ig}}{Y_i^2}\Big(\frac{Y_i - d_i}{Y_i - 1}\Big)d_i, \quad \text{pour } g \neq j. \tag{6.4}$$

Les termes $\frac{Y_{ij}}{Y_i}\Big(1-\frac{Y_{ij}}{Y_i}\Big)d_i$ et $-\frac{Y_{ij}Y_{ig}}{Y_i^2}$ proviennent de la variance et de la covariance d'une variable aléatoire multinomiale de paramètres $(d_i; p_j = \frac{Y_{ij}}{Y_i})$; $j = 1, \ldots, k$.
Les composantes du vecteur $(U_1(\tau), \ldots, U_k(\tau))$ sont linéairement dépendantes car $\sum_{j=1}^{D} U_j(\tau) = 0$.
Le test statistique est construit en choisissant $k-1$ variables parmi les k variables U_j, les $k-1$ premières par exemple.
La statistique du test est donnée par la forme quadratique suivante

$$\chi^2 = (U_1(\tau), \ldots, U_{k-1}(\tau))\Sigma^{-1}(U_1(\tau), \ldots, U_{k-1}(\tau))'. \tag{6.5}$$

où Σ est la matrice carrée d'ordre $k-1$ formée des éléments $\hat{\sigma}_{jg}$ correspondant au choix des $k-1$ variables.
Sous l'hypothèse nulle, cette statistique se distribue asymptotiquement selon la loi du χ^2 à $k-1$ degrés de liberté (cf le chapitre V du livre de Andersen et al. 1993).
Pour $k = 2$ la statistique du test s'écrit

$$N = \frac{\sum_{i=1}^{D} W(t_i)\big[d_{i1} - Y_{i1}\big(\frac{d_i}{Y_i}\big)\big]}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{i1}}{Y_i}\big(1 - \frac{Y_{i1}}{Y_i}\big)\big(\frac{Y_i - d_i}{Y_i - 1}\big)d_i}}, \tag{6.6}$$

qui suit asymptotiquement une loi normale standard sous H_0.
Selon le choix de la fonction poids, nous aboutissons à l'un ou l'autre des tests suivants (parmi d'autres).

6.0.1 Test de Gehan(1965)

C'est une généralisation du test de Wilcoxon-Mann-Whitney au cas des données censurées à droite, il est donné par le choix de la fonction poids $W(t_i) = Y_i$ et il peut se généraliser comme suit

6.0.2 Test de Tarone et Ware(1977)

Il s'obtient pour $W(t_i) = g(Y_i)$ où g est une fonction fixée (par exemple $g(y) = \sqrt{y}$).
Pour $W = 1$, nous obtenons le test de Log-Rang, c'est le test le plus utilisé dans les packages statistiques et se prête à être étendu à différents cas de censure et peut aussi se généraliser comme suit.

6.0.3 Tests de Fleming et Harrington

Soit $\widehat{S}(t)$ l'estimateur EKM (de Kaplan-Meier) basé sur l' échantillon global, alors la fonction des poids proposée par Fleming et Harrington est donnée par

$$W_{p,q}(t_i) = \widehat{S}(t_{i-1})^p[1 - \widehat{S}(t_{i-1})]^q. \tag{6.7}$$

avec $p \geq 0$, $q \geq 0$.
Dans cette fonction de poids la fonction de survie au temps précédent de mort est utilisé comme un poids pour assurer que ces poids soient connus juste avant le temps où on a fait la comparaison.

- dans le cas où $p = q = 0$ nous retrouvons le test de log-rang.

- Le cas $q = 0$ et $p > 0$ donne plus de poids aux résultats qui apparaissent tôt dans le temps.

- Par contre, le cas $p = 0$ et $q > 0$ donne plus de poids aux données qui apparaissent tard dans le temps.

Ces tests permettent une certaine souplesse en fonction des données dont nous disposons et cela par un choix adéquat de p et q .

6.0.4 Application sur données réelles

Le temps d'infection dû à deux emplacements différents du tube utilisé pour la dialyse chez des patients souffrant d'insuffisance rénale, est donné ci dessous.

Premier emplacement

- observations complètes =[1.5 3.5 4.5 4.5 5.5 8.5 8.5 9.5 10.5 11.5 15.5 16.5 18.5 23.5 26.5];
- observations censurées =[2.5 2.5 3.5 3.5 3.5 4.5 5.5 6.5 6.5 7.5 7.5 7.5 7.5 8.5 9.5 10.5 11.5 12.5 12.5 13.5 14.5 14.5 21.5 21.5 22.5 22.5 25.5 27.5];

Deuxième emplacement

- observations complètes =[0.5 0.5 0.5 0.5 0.5 0.5 2.5 2.5 3.5 6.5 15.5];
- observations censurées =[0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.5 1.5 1.5 1.5 2.5 2.5 2.5 2.5 2.5 3.5 3.5 3.5 3.5 3.5 4.5 4.5 4.5 5.5 5.5 5.5 5.5 5.5 6.5 7.5 7.5 7.5 8.5 8.5 8.5 9.5 9.5 10.5 10.5 10.5 11.5 11.5 12.5 12.5 12.5 12.5 14.5 14.5 16.5 16.5 18.5 19.5 19.5 19.5 20.5 22.5 24.5 25.5 26.5 26.5 28.5];

La figure suivante montre les courbes des EKM calculés à partir des deux échantillons précédents. Les tests de Log-Rang et de Gehan donnent respectivement une p-valeur de 0.112 et 0.964, ce qui permet d'accepter l'hypothèse de l'équivalence des deux méthodes d'emplacement au niveau de signification $\alpha = .05$ tandis que les deux courbes dans la figure 6.1 s'éloignent considérablement pour t assez grand.
En faisant la comparaison par le test de Fleming et Harrington pour $p = 1$ et $q = 1$ ou $p = 0.5$ et $q = 0.5$ ou $p = 0.5$ et $q = 2$, alors nous obtenons un rejet significatif au même seuil.
Tous les résultats donnés dans cet exemple de données réelles sont repris du chapitre 7 du livre de Klein et Moeschberger (1997).

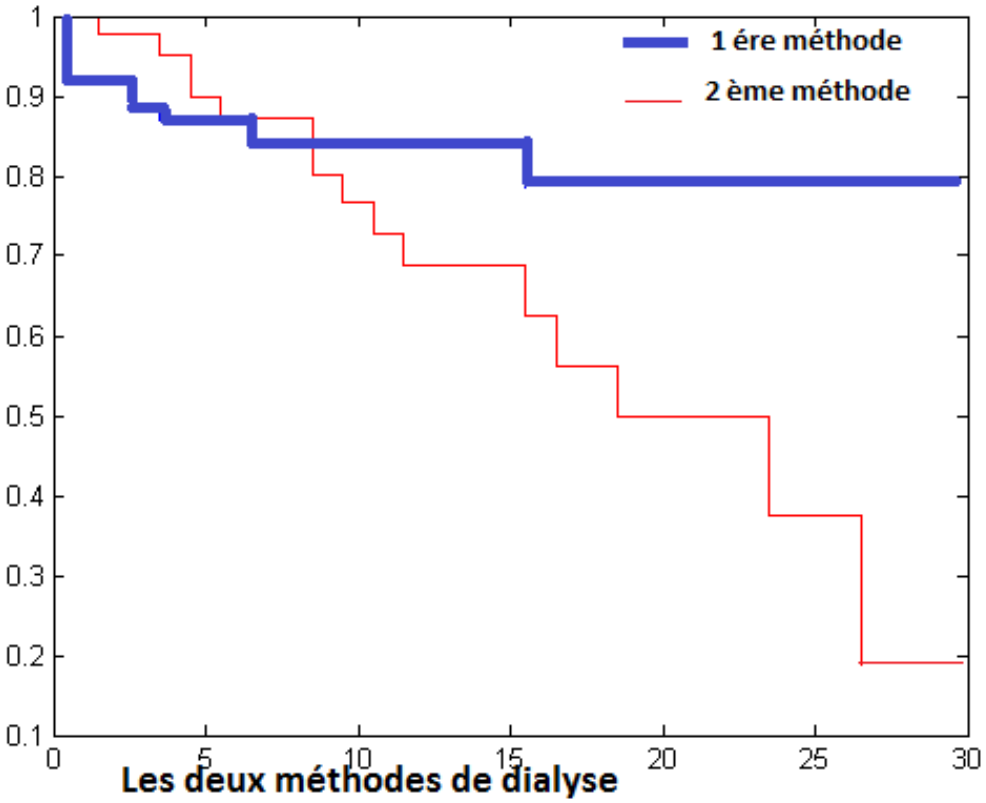

Figure 6.1: Les deux courbes de survie

EXERCICE 6.1

Nous voulons comparer l'efficacité de l'insémination artificielle chez deux races de bovins, Pie Noire et Pie rouge, afin de détecter une éventuelle différence entre les deux races. A cette fin, nous avons utilisé deux échantillons de vaches élevées dans le territoire de la Daira de Teleghma, wilaya de Mila (cf Dib (2012)). La variable d'intérêt est le nombre de fois que l'on pratique l'insémination artificielle pour que la vache soit en gestation. Les deux vecteurs 'echantillon1' et 'echantillon2' donnent les valeurs de notre variable d'intérêt chacun pour une race donnée. Par ailleurs, les éléments des vecteurs respectifs 'delta1' et 'delta2' prennent la valeur 0 si l'observation associée est censurée à droite, et la valeur 1 dans le cas d'une observation complète. Ici la censure correspond soit à la vente, soit à la mort d'une vache sur laquelle les tentatives d'insémination ont échoué. Signalons le fait que nous avons choisi des échantillons comparables à tout point de vu et en particulier à l'aptitude à la fertilité. Voici le tableau des données recueillies où $x+$ indique que x est censurée à droite.

Échantillon1= [1 + 2 + 2 + 2 + 1 1 1 1 3 2]

et

échantillon2= [5 2 1 + 2 + 1 1 1 1 4 + 2].

1) Tracer les courbes des estimateurs des fonctions de survie du nombre de fois que l'on pratique l'insémination pour arriver à une gestation, pour les deux races de bovins.

2) Tester l'hypothèse de réaction identique, à l'insémination pratiquée, pour les deux races.

Bibliography

Andersen P., Borgan O., Gill R. and Keiding N. (1993). *Statistical Models Based on Conting Processes* . Spring Series in Statistics. Springer.

Beran R. (1981), *Nonparametric regression with randomly censored survival data*. Technical report university of California, Berkeley.

Blum, J. R. and Susarla, V. (1980), Maximal deviation theory of density and failure rate function estimates based on censored data. *In Multivariate Analysis* **5** (P.R. Krishnaiah ed.), 213–222.

Bitouzé D. and Massart B., (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. **35** (6), 735–763.

Boukeloua M. (2013). *Etude des estimateurs de la fonction de répartition et de densité dans un modèle de censure*. Mémoire de Master, Département de Mathématiques, Université Constantine 1.

Boukeloua, M. (2015). Rates of mean square convergence of density and failure rate estimators under twice censoring. *Statistics & Probability Letters*, **106**, 121–128.

Boukeloua, M. and Messaci F. (2016). Asymptotic normality of kernel estimators based upon incomplete data. *Journal of Nonparametric Statistics*, **28** (3), 469-486, DOI: 10.1080/10485252.2016.1164312.

Breslow N. and Crowly J. (1974). A large Sample Study of Life Table and Product-limit Estimates under Random Censorship. *Ann. Statist*, **2** (3), 437–453.

Carbonez, A., Györfi, L., van der Meulen, E.C. (1995). Partitioning-estimates of a regression function under random censoring. *Statist. Decisions*, **13**, 21–37.

Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *The Annals of Statistics*, **18** (1), 391–404.

Chang, M. N., and Yang G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *The Annals of Statistics*, **15** (4), 1536–1547.

Cocozza-Thivent C. (1997). *Processus stochastiques et fiabilité des systèmes*. Springer-Verlag Berlin Heidelberg.

Devroye L. and Györfi L. (1983). Distribution-free exponential bounds on the L1 error of partitioning estimates of a regression function. In: Konecny, F., Mogyordi, J., Wertz, W. (Eds.), Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics. Akadémian Kiadö, Budapest, Hungary, pp. 67–76.

Devroye L. and Krzyżak A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimates. *J.Statist. Plann. Inference*. **23**, 71–82.

Devroye L., Györfi L., Krzyżak A. and Lugos, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.*. **22**, 1371–1385.

Dib K. (2012). *Tests d'hypothèses dans un modèle de censure*. Mémoire de Magistère. Université Mentouri, Constantine.

Diehl, S. and Stute W. (1988). Kernel density and hazard function estimation in the presence of censoring. *Journal of Multivariate Analysis*, **25**, 299–310.

Donoho D. L., Johnstone, I. M., Kerkyacharian G. and Picard D. (1995). Wavelet shrinkage: asymptopia (with discussion)? J. Roy. Statist. Soc., Ser. B **57** (2), 301–370.

Doukhan P. and Léon J. (1990). Déviation quadratique d'estimateurs de densité par projection orthogonale. *Compt. Rend. Acad. Sci. Paris* A **310**, 424–430.

Efron B. (1967). The Two sample problem with censored data. *Proc. Fifth Berkely Symp. Math. Statist. Probab.* **4**, 831–853.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis. Theory and Practice.* Springer Series in Statistics. New York.

Fleming T. and Harrington D. (1991). *Counting Processes & Survival Analysis.* Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics Section. Wiley & Sons.

Földes A., Rejtő L. and Winter B.B. (1980). Strong consistency properties of nonparametric estimators for randomly censored data I: the product-limit estimator. *Period. Math. Hungar.* **11**, 233–250.

Földes A., Rejtő L. and Winter B.B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data II: the estimation of density and failure rate. *Period. Math. Hungar.* **12**, 15–29.

Földes A. and Rejtő L. (1981). A LIL type result for the product limit estimator. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.* **56**, 75–86.

Guessoum Z. and Ould-Said E. (2008). On nonparametric estimation of the regression function under random censorship model. *Statistics & Decisions* **26** (3), 159–177.

Guessoum Z. and Ould-Said E. (2010). Kernel regression uniform rate estimation for censored data under α mixing conditions *Electronic Journal of Statistics* **4**, 117–132.

Kagba N.S. (2004). *On kernel density estimation for censored data* (Ph.D. thesis), University of California, San Diego.

Kalbfleisch D. and Prentice R. L. (1980). *The Statistical Analysis of Failure Time Data.*Wiley & Sons.

Kaplan E. L. and Meier P. (1958). Nomparametric estimation from incomplete observations. *Jasa,* 457–481.

Kebabi K., Laroussi I. and Messaci F. (2011). Least squares estimators of the regression function with twice censored data, *Statist. Probab. Lett.* **81**, 1588–1593.

Kebabi K. and Messaci F. (2012). Rate of the almost complete convergence of a kernel regression estimate with twice censored data. *Statist. Probab. Lett.* **82**, 1908–1913.

Kerkyacharian G. and Picard D. (1992). Density estimation in Besov Spaces. *Statistics and Probability Letters* **13**, 15–24.

Kiefer J. 1(961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math.* **11**, 649–660.

Kitouni A., Boukeloua M. and Messaci F. (2015). Rate of strong consistency for nonparametric estimators based on twice censored data. *Statistics & Probability Letters*, **96**, 255–261.

Klein J. P. and Moeschberger M. L. (1997). *Survival Analysis. Techniques for Censored and Truncated Data.* Springer.

Kohler M., Krzyżak A. (2001). Nonparametric regression estimation using penalized least squares, *IEEE Trans. Inform. Theory.* **47**, 3054–3058.

Kohler M., Màth é K. and Pint ér M. (2002). Prediction from randomly right censored data. *J. Multivariate Anal.* **80**, 73–100.

Lee E. T. (1992). *Statistical Methods for Survival Data Analysis.* John Wiley & Sons. Inc.

Messaci F. (2010). Local averaging estimates of the regression function with twice censored data. *Statist. Probab. Lett.* **80**, 1508–1511.

Messaci F. and Nemouchi N. (2011). A law of the iterated logarithm for the product limit estimator with doubly censored data. *Statist. Probab. Lett.* **81**, 1241–1244.

Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *The Annals of Statistics*, **14** (2), 766–773.

Nadaraya, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* , **9**, 157–159.

Parzen E. (1962). On estimating of a probability density and mode. *Annals of Mathematicla Statistics*, **33**, 1065–1076.

Patilea V. and Rolin J. M. (2006). Product-limit estimators of the survival function with twice censored data. *Ann. Statist.* **34**, No 2, 925–938.

Peterson Jr. (1977). Expressing the Kaplan–Meyer estimator as a function, of empirical subsurvival functions. *J. Amer. Statist. Assoc.* . **72**, 854–858. MR 57 10903

Prakasa Rao B. L. S. (1983). *Nonparametric Functional Estimation.* Academic Press, New York.

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

Rosenblatt M. (1956). Remarks on some nonparametric estimates of a density function (1956). *Annals of Mathematical Statistics*, **27**, 832–837.

Shorack R. G. and Wellner W. J. (1986) . *Empirical processes with applications to statistics.* John Wiley Sons.

Stute W. and Wang J. L. (1993). The strong law under random censorship *Ann. Statist.* **21**, 3, 1591-1607.

Turnbull B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69**, 169–173. MR0381120.

Turnbull B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society.* Series B (Methodological), **38** (3), 290–295.

Walter, G. G. (1992). Approximation of the Delta Function by Wavelets. *J. Approx. Theory.* **71**, 329–343.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhya A* , **26**, 359–372.