REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
==========

UNIVERSITE MENTOURI – CONSTANTINE
FACULTE DES SCIENCES EXACTES
==========

DEPARTEMENT DE MATHEMATIQUES

N° d'ordre : .....
N° de série : .....

MEMOIRE PRESENTE POUR L'OBTENTION
DU
DIPLOME DE MAGISTERE
EN
MATHEMATIQUES

« Sur les processus empiriques : Application à la statistique des données censurées »

Par
Abderrahim Kitouni

OPTION
Probabilités et Statistique

Devant le jury :

Président	M.	Z. Mohdeb	Professeur	Université Mentouri
Rapporteur	M^{me}	F. Messaci	Professeur	Université Mentouri
Examinateur	M^{me}	S. Belaloui	M. C.	Université Mentouri
Examinateur	M^{me}	N. Nemouchi	M. C.	Université Mentouri

Soutenu le : .....

Remerciements

Je remercie en premier lieu Mme. Fatiha Messaci pour tout le temps qu'elle m'a consacré, ainsi que pour toute l'aide qu'elle a patiemment apportée à la construction de ce mémoire.

J'ai été très honoré que M. Zaher Mohdeb accepte la présidence du jury, et j'aimerais lui adresser de ce fait de vifs remerciments.

Enfin, je remercie sincèrement Mme. Nahima Nemouchi, et Mme. Soheir Belaloui pour avoir accepté de faire partie du jury, et pour le temps accordé à la lecture attentive du mémoire.

Table des matières

Introduction **ii**

1 Introduction aux processus empiriques **1**
1.1 Définitions 1
1.2 Théorème de Glivenko-Cantelli 2
1.3 Processus empirique uniforme 3
1.4 La convergence faible des processus 4
1.4.1 Définition 4
1.4.2 Les espaces $\mathcal{C}$ et $\mathcal{D}$ 5
1.4.3 Critère de convergence faible 6
1.5 Théorème de représentation de Skorokhod 8
1.6 Théorème de Donsker 9

2 La loi du logarithme itéré **11**
2.1 LIL classique 11
2.2 LIL pour les processus empiriques 12
2.3 Lois fonctionnelles du logarithme itéré 12

3 Le processus empirique de Kaplan-Meier **14**
3.1 Introduction aux données censurées 14
3.2 L'estimateur de Kaplan-Meier 16
3.3 La loi du logarithme itéré pour l'estimateur de Kaplan-Meier 17
3.4 Autres théorèmes limites pour le processus empirique de Kaplan-Meier . 18

4 Fonctionnelles locales : cas de la censure à droite **20**
4.1 Introduction 20
4.2 Lois fonctionnelles pour les processus basés sur des données censurées . . 24
4.3 Application : Estimation de la densité 39

5 Extension au cas de la censure double **43**
5.1 Présentation du modèle de Patilea et Rolin (2006) 43
5.2 Quelques résultats préliminaires 45
5.3 Étude de simulation 45

Introduction

La théorie des processus empiriques est l'un des outils les plus puissant de la statistique semi-paramétrique. Ses débuts remontent aux années 1930 avec des résultats tels que le théorème de Glivenko-Cantelli (Glivenko, 1933; Cantelli, 1933) ou encore le test de Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1939). Ces deux résultats ont en commun le fait qu'ils donnent la convergence de la quantité $\|\mathrm{F}_n - \mathrm{F}\|$ (la convergence presque sûre pour le premier et la loi asymptotique pour le second), où F est la fonction de répartition de la loi d'où l'échantillon est tiré, F_n est la fonction de répartition empirique et $\|.\|$ est la norme de la convergence uniforme. Ce que l'on désigne par processus empirique n'est rien de plus que la quantité $\mathrm{F}_n - \mathrm{F}$ normalisée par le facteur $\sqrt{n}$.

L'extension du résultat de Smirnov a donné ce qu'on appelle la loi du logarithme itéré (Chung, 1949; Deheuvels, 1991), et plus tard des lois limites fonctionnelles, globales ou locales, (Finkelstein, 1971; Deheuvels, 1992; Deheuvels et Mason, 1992; Deheuvels, 2000).

Pour une description approfondie des résultats classiques concernant les processus empiriques et leurs applications, nous renvoyons aux ouvrages de Shorack et Wellner (1986); Pollard (1984); van der Vaart et Wellner (1996); Dudley (1999). Pour un texte introductif le livre de Kosorok (2008) ou la thèse de Viallon (2006) sont un bon début.

Dans les années 1980, Stute (voir Stute, 1982a,b, 1986a,b) a été l'un des premiers statisticiens (voir également Csörgő, 1981; Deheuvels, 1974), à faire un usage systématique des méthodes de la théorie des processus empiriques dans l'étude des propriétés asymptotiques d'estimateurs fonctionnels non paramétriques. Ses travaux ont concerné principalement les estimateurs à noyau.

Depuis, de nombreux auteurs, parmi lesquels nous citons Deheuvels, J.H.J. Einmahl, U. Einmahl et Mason (voir par exemple Einmahl et Mason, 2000; Deheuvels et Mason, 2004; Einmahl et Mason, 2005), ont introduit des techniques nouvelles pour aborder ces problèmes, en utilisant les lois limites fonctionnelles, ou encore des variantes locales de la théorie des processus empiriques indexés par des ensembles ou par des fonctions.

Dans ce mémoire, nous allons exposer certains résultats de base de la théorie des processus empiriques, ainsi que quelque résultats analogues dans le cas des données censurées. En particulier, nous étudions le résultat de Deheuvels et Einmahl (1996) avec plus de détails. Enfin, nous nous intéressons aux perspectives de recherche dans la généralisation

de ce résultats au cas de la censure double. Ceci est possible grâce à la loi du logarithme itéré de Messaci et Nemouchi (2011).

Dans le premier chapitre, nous exposons les définitions et les résultats de base sur les processus empiriques. Le deuxième chapitre est consacré à la loi du logarithme itéré, un résultat classique concernant la convergence des sommes de variables aléatoires i.i.d., ainsi que ses généralisations pour les processus empiriques. Le troisième chapitre décrit les données censurées, avant de donner une définition du processus empirique pour le cas de la censure aléatoire à droite en utilisant l'estimateur de Kaplan et Meier (1958), ainsi que les résultats obtenus dans ce cas, en particulier par Földes et Rejtő (1981); Einmahl et Koning (1992). Dans le quatrième chapitre, nous étudions en détails un des résultats de la théorie des processus empiriques, en l'occurrence une loi fonctionnelle du logarithme itéré pour les incréments du processus empirique de Kaplan-Meier, résultat obtenu par Deheuvels et Einmahl (1996). Nous détaillons aussi l'utilisation de ce résultat pour montrer une vitesse de convergence de l'estimateur à noyau de la densité. Le dernier chapitre est, quant à lui, consacré à l'étude du modèle de censure double de Patilea et Rolin (2006). Nous présentons la loi du logarithme itéré de Messaci et Nemouchi (2011) pour ce dernier estimateur ainsi que les perspectives de recherche concernant l'extension du résultat de Deheuvels et Einmahl (1996) à ce modèle. Nous présentons en outre une étude de simulation pour explorer les propriétés de l'estimateur à noyau de la densité obtenu pour ce même modèle.

Chapitre 1

Introduction aux processus empiriques

1.1 Définitions

Intuitivement, un processus empirique est un processus aléatoire qui dépend d'un échantillon. Par exemple, la fonction de répartition empirique.

Plus précisément, si l'on considère un espace probabilisable $(\mathcal{X}, \mathscr{A})$, et $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ un échantillon i.i.d de loi de probabilité P_X, on définit la mesure empirique P_n par :

$$\mathrm{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathrm{X}_i},$$

où δ_x est la mesure de Dirac au point x.

Pour une famille $\mathcal{S}$ d'ensembles mesurables, on définit alors le processus empirique $\{\mathrm{P}_n(\mathrm{A}), \mathrm{A} \in \mathcal{S}\}$. Un processus empirique ainsi défini est dit *indexé par des ensembles*. Dans le cas réel, la fonction de répartition empirique peut s'écrire ainsi en prenant $\mathcal{S} = \{]-\infty, t], t \in \mathbb{R}\}$.

Pour une classe $\mathcal{F}$ de fonctions mesurables, on peut définir un processus empirique $\{\mathrm{P}_n f, f \in \mathcal{F}\}$ où

$$\mathrm{P}_n f = \int f d\mathrm{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(\mathrm{X}_i).$$

Un tel processus est dit *indexé par des fonctions*. Cette définition est plus générale car elle permet de décrire toutes les fonctions mesurables de l'échantillon. La définition précédente est obtenue en se limitant à des fonctions indicatrices.

De plus, pour peu que la classe $\mathcal{F}$ soit dénombrable (resp. ait la puissance du continu), le processus empirique peut être vu comme un processus "classique" à indices entiers (resp. réels).

Souvent, ce que l'on appelle processus empirique est $\{\sqrt{n}(\mathrm{P}_n(\mathrm{A}) - \mathrm{P}_\mathrm{X}(\mathrm{A})), \mathrm{A} \in \mathcal{S}\}$ ou bien $\{\sqrt{n}(\mathrm{P}_n f - \mathrm{P}_\mathrm{X} f), f \in \mathcal{F}\}$. C'est-à dire qu'au lieu de prendre la mesure empirique, on prend la différence entre cette dernière et la mesure de probabilité théorique.

Dans la suite, nous nous restreignons au cas réel, et nous appellons processus empirique un processus de la forme $\sqrt{n}(\mathrm{F}_n(x) - \mathrm{F}(x))$ où F_n est la fonction de répartition empirique ou un autre estimateur de la fonction de répartition. Par exemple, dans le cas des données censurées à droite, F_n est l'estimateur de Kaplan-Meier et on parle alors du *processus empirique de Kaplan-Meier*. Pour plus de détails concernant la théorie générale des processus empiriques, voir Shorack et Wellner (1986, chap. 26), Kosorok (2008) et van der Vaart et Wellner (1996).

1.2 Théorème de Glivenko-Cantelli

La loi forte des grands nombres permet de donner, en tout point de $\mathbb{R}$, la convergence presque sûre de la fonction de répartition empirique vers la fonction de répartition des observations. Le résultat suivant, aussi connu sous le nom de loi uniforme des grands nombres, est dû à Glivenko (1933) et Cantelli (1933) et donne une version uniforme sur $\mathbb{R}$ de cette convergence.

Théorème 1.1. *Soit* $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ *une suite de variables aléatoires indépendates de même loi de probabilité, de fonction de répartition* F. *Et soit* F_n *la fonction de répartition empirique définie par*

$$\mathrm{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,x]}(\mathrm{X}_i).$$

Alors, $\sup_{x\in\mathbb{R}} |\mathrm{F}_n(x) - \mathrm{F}(x)| \to 0$ *p.s.*

Démonstration. voir Billingsley (1986, Théorème 20.6) □

Ce résultat inspire la notion de classe de Glivenko-Cantelli, qui est une généralisation de cette convergence pour un processus empirique quelconque :

Définition 1.1. Soit $\mathrm{X}_1, \ldots, \mathrm{X}_n$ un échantillon de loi P_X, et soit P_n la mesure empirique correspondante. Une classe $\mathcal{F}$ de fonctions mesurables (resp. une famille $\mathcal{S}$ d'ensembles mesurables) est dite *de Glivenko-Cantelli* (GC) par rapport à P_X si :

$$\sup_{f\in\mathcal{F}} |\mathrm{P}_n f - \mathrm{P}_\mathrm{X} f| \to 0 \text{ p.s.}$$

(resp. $\sup_{\mathrm{A}\in\mathcal{S}} |\mathrm{P}_n(\mathrm{A}) - \mathrm{P}_\mathrm{X}(\mathrm{A})| \to 0$ p.s.).

Une *classe de Glivenko-Cantelli universelle* est une classe de Glivenko-Cantelli par rapport à toutes les mesures de probabilité.

Par exemple, la classe de fonctions $\{1_{]-\infty,t]}, t \in \mathbb{R}\}$ est une classe GC universelle.

1.3 Processus empirique uniforme

Un résultat important, qui est souvent utilisé pour simplifier les preuves, est la réduction au cas uniforme. Ceci consiste à introduire un processus empirique uniforme (c'est à dire un processus empirique basé sur un échantillon de loi uniforme sur $[0,1]$) qui est plus facile à étudier. L'extension au cas d'une loi arbitraire est souvent possible sans trop d'hypothèses, mais nous devons parfois supposer la continuité de la fonction de répartition.

Soit F une fonction de répartition, et soit F^{inv} son inverse généralisée définie pour $t \in [0,1]$ par :

$$\mathrm{F}^{inv}(t) = \inf\{x \in \mathbb{R} : \mathrm{F}(x) \geqslant t\}.$$

On a les propriétés suivantes :
- $\forall t \in [0,1] : \mathrm{F}\left(\mathrm{F}^{inv}(t)\right) \geqslant t$;
- $\forall x \in \mathbb{R} : \mathrm{F}^{inv}\left(\mathrm{F}(x)\right) \leqslant x$;
- $\mathrm{F}(x) \geqslant t \iff \mathrm{F}^{inv}(t) \leqslant x$;
- Si X suit la loi F, alors $\forall t \in [0,1] : \mathrm{P}(\mathrm{F}(\mathrm{X}) \leqslant t) \leqslant t$, et si t appartient à l'image de F alors $\mathrm{P}(\mathrm{F}(\mathrm{X}) \leqslant t) = t$. En particulier, si F est continue alors $\mathrm{F}(\mathrm{X})$ suit la loi uniforme sur $[0,1]$.

Théorème 1.2. *Soit ξ une variable aléatoire de loi uniforme sur $[0,1]$. Et soit P_{X} une loi de probabilité sur $\mathbb{R}$, de fonction de répartition F. On définit la variable aléatoire X par $\mathrm{X} = \mathrm{F}^{inv}(\xi)$.*

Alors X suit la loi P_{X}.

Démonstration. D'après les propriétés précédentes et la croissance de F, si $\xi \leqslant \mathrm{F}(x)$ alors $\mathrm{X} = \mathrm{F}^{inv}(\xi) \leqslant \mathrm{F}^{inv}(\mathrm{F}(x)) \leqslant x$. Et si $\mathrm{X} \leqslant x$, alors $\mathrm{F}(x) \geqslant \mathrm{F}(\mathrm{X}) = \mathrm{F}\left(\mathrm{F}^{inv}(\xi)\right) \geqslant \xi$.

Ce qui montre que les événements $\{\mathrm{X} \leqslant x\}$ et $\{\xi \leqslant \mathrm{F}(x)\}$ sont équivalents, et on a alors $\mathrm{P}(\mathrm{X} \leqslant x) = \mathrm{P}(\xi \leqslant \mathrm{F}(x)) = \mathrm{F}(x)$. □

Ce résultat élémentaire peut se généraliser au cas des processus empiriques : Considérons une suite $(\mathrm{U}_n)_{n\geqslant 1}$ de v.a. i.i.d. de loi uniforme sur $[0,1]$. Nous notons respectivement $\mathbb{U}_n$ et α_n la fonction de répartition empirique et le processus empirique associés

$$\alpha_n(t) = \sqrt{n}(\mathbb{U}_n(t) - t) \quad \text{et} \quad \mathbb{U}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\mathrm{U}_i \leqslant t\}}.$$

Théorème 1.3. *Les suites de processus $\{\mathrm{F}_n(t), t \in \mathbb{R}\}$ et $\{\mathbb{U}_n(\mathrm{F}(t)), t \in \mathbb{R}\}$ ont les mêmes lois de probabilité conjointes en n (c'est-à-dire que pour tout $k \in \mathbb{N}^*$ et pour tout $n_1, n_2, \ldots, n_k \in \mathbb{N}$, $(\mathrm{F}_{n_1}, \mathrm{F}_{n_2}, \ldots \mathrm{F}_{n_k})$ et $(\mathbb{U}_{n_1}(\mathrm{F}), \mathbb{U}_{n_2}(\mathrm{F}), \ldots \mathbb{U}_{n_k}(\mathrm{F})$ ont la même loi). De même pour les processus $\{\sqrt{n}(\mathrm{F}_n(t) - \mathrm{F}(t)), t \in \mathbb{R}\}$ et $\{\alpha_n(\mathrm{F}(t)), t \in \mathbb{R}\}$.*

Démonstration. voir Shorack et Wellner (1986, Théorème 2, page 4) □

Ceci nous permet de limiter le travail probabiliste au cas des échantillons de loi uniforme, et de généraliser ces résultats en insérant la fonction de répartition F dans le résultat (On a parfois besoin de supposer que F est continue pour faire cela).

Si F est continue, et si on choisit $\mathrm{U}_n = \mathrm{F}(\mathrm{X}_n)$ pour tout n dans le théorème précédent, alors non seulement on a égalité des distributions mais également l'égalité presque sûre. Ce résultat peut être généralisé sous des hypothèses assez faibles au cas où F n'est pas continue (voir par exemple Shorack et Wellner, 1986, p. 102).

1.4 La convergence faible des processus

1.4.1 Définition

Soit $(\mathrm{X}_n)_{n\in\mathbb{N}}$ une suite de v.a. réelles. On dit que (X_n) converge en loi vers une variable X, et on note $\mathrm{X}_n \xrightarrow{\mathscr{L}} \mathrm{X}$, si la suite des fonctions de répartition (F_n) de (X_n) converge simplement vers la fonction de répartition F de X en tout point de continuité de cette dernière.

Ceci est équivalent à dire que la suite des lois de probabilité (P_n) des variables (X_n) converge faiblement vers la loi de probabilité de X, ce qui veut dire que pour toute fonction f continue et bornée sur $\mathbb{R}$, on a :

$$\int_{\mathbb{R}} f \, d\mathrm{P}_n \longrightarrow \int_{\mathbb{R}} f \, d\mathrm{P}.$$

Contrairement à la première, cette définition a l'avantage de pouvoir être généralisée à un espace métrique quelconque.

Dans la suite, (S, d) désigne un espace métrique muni de sa tribu borélienne $\mathscr{S}$ (la tribu engendrée par les ouverts de S). Un tel espace est dit polonais s'il est séparable et complet.
Définition 1.2. On dit qu'une suite de mesures de probabilité (P_n) définies sur $(\mathrm{S}, \mathscr{S})$ converge faiblement vers une mesure de probabilité P (définie sur le même espace $(\mathrm{S}, \mathscr{S})$), et on note $\mathrm{P}_n \Rightarrow \mathrm{P}$, si pour toute fonction réelle f définie sur S continue et bornée, on a :

$$\int_{\mathrm{S}} f d\mathrm{P}_n \longrightarrow \int_{\mathrm{S}} f d\mathrm{P}.$$

Le théorème suivant donne d'autres définitions équivalentes de la convergence faible :
Théorème 1.4 (Portmanteau)**.** *Soient* $(\mathrm{P}_n, n \geqslant 1), \mathrm{P}$ *des mesures de probabilité sur* $(\mathrm{S}, \mathscr{S})$. *Les conditions suivantes sont équivalentes :*

1. $\mathrm{P}_n \Rightarrow \mathrm{P}$ *;*
2. $\int f \, d\mathrm{P}_n \longrightarrow \int f \, d\mathrm{P}$ *pour toute fonction réelle* f *bornée et uniformément continue ;*

3. $\limsup \mathrm{P}_n(\mathrm{F}) \leqslant \mathrm{P}(\mathrm{F})$ *pour tout fermé* F *;*
4. $\liminf \mathrm{P}_n(\mathrm{O}) \geqslant \mathrm{P}(\mathrm{O})$ *pour tout ouvert* O *;*
5. $\lim \mathrm{P}_n(\mathrm{A}) = \mathrm{P}(\mathrm{A})$ *pour tout* $\mathrm{A} \in \mathscr{S}$ *tel que* $\mathrm{P}(\mathrm{Fr}\,\mathrm{A}) = 0$ *(*$\mathrm{Fr}\,\mathrm{A}$ *désigne la frontière de l'ensemble* A*).*

Démonstration. cf. Billingsley (1968, Théorème 2.1) □

Si l'espace S n'est pas séparable, la tribu $\mathscr{S}_0$ engendrée par les boules ouvertes peut être plus petite que la tribu borélienne $\mathscr{S}$. Dudley (1966, 1967) a introduit une théorie de la convergence faible qui utilise uniquement les ensembles de $\mathscr{S}_0$ et les fonctions $\mathscr{S}_0$-mesurables.

1.4.2 Les espaces $\mathcal{C}$ et $\mathcal{D}$

Soit M une famille de fonctions réelles définies sur un ensemble d'indices T. Pour $k \geqslant 1$ et $t = (t_1, \ldots, t_k) \in \mathrm{T}^k$, on note $\pi_t = \pi_{t_1,\ldots,t_n}$ l'application projection de M dans $\mathbb{R}^k$ définie par

$$\pi_t(x) = (x(t_1), \ldots, x(t_k)) .$$

On appelle *ensemble fini-dimensionnel* toute partie de M de la forme $\pi_t(\mathrm{B})$ où $t \in \mathrm{T}^k$ et $\mathrm{B} \in \mathscr{B}(\mathbb{R}^k)$ pour $k \geqslant 1$. L'ensemble de $\mathscr{M}_0$ de tout les ensembles fini-dimensionnels est une algèbre. On note $\mathscr{M}$ la σ-algèbre engendrée par l'algèbre des ensembles fini-dimensionnels.

Soit $(\mathrm{X}_t, t \in \mathrm{T})$ un processus stochastique à valeurs réelles. Si toutes ses trajectoires sont dans M, il peut être vu comme un élément aléatoire de M. Pour pouvoir utiliser la théorie de la convergence faible citée plus haut, il faut munir M d'une métrique dont la tribu borélienne coïncide avec la tribu engendrée par les ensembles fini-dimensionnels.

Dans la suite, nous nous restreignons au cas où $\mathrm{T} = [0, 1]$ (car comme nous venons de le voir, on peut toujours se ramener à des variables uniformes sur $[0, 1]$), et l'ensemble M de fonctions est soit l'ensemble des fonctions continue, soit l'ensemble des fonctions "cadlag" (c'est-à-dire l'ensemble des fonctions continues à droite et ayant des limites à gauche en tout point).

On note par $\mathcal{C}$ l'ensemble des fonctions réelles définies et continues sur $[0, 1]$, et on définit la norme uniforme d'une fonction x de $\mathcal{C}$ par

$$\|x\| = \sup_{t\in[0,1]} |x(t)|.$$

Cette norme induit une distance sur $\mathcal{C}$ qui en fait un espace métrique séparable et complet. De plus, sa tribu borélienne que l'on note $\mathscr{C}$ est engendrée par les ensembles fini-dimensionnels (Shorack et Wellner, 1986, Chapitre 2).

Seulement, les processus que l'on manipule ne sont pas toujours à trajectoires continues. C'est ce qui nous amène à travailler dans un espace plus grand. Comme une fonction de répartition est toujours continue à droite, nous allons considérer l'espace des fonctions "cadlag" définies sur $[0,1]$, que l'on note $\mathcal{D}$.

Muni de la distance uniforme, $\mathcal{D}$ est un espace complet mais n'est pas séparable. De plus, la tribu engendrée par les ensembles fini-dimensionnels est strictement incluse dans la tribu borélienne de $(\mathcal{D}, \|.\|)$. Ceci fait que la norme uniforme est inappropriée pour l'étude de la convergence faible des processus à trajectoire dans $\mathcal{D}$.

Pour remédier à ce problème, Skorokhod (1956) a proposé la métrique définie par :

$$d(x,y) = \inf_{\lambda \in \Lambda} \max(\|x - y \circ \lambda\|, \|\lambda - \mathrm{I}\|),$$

où Λ est l'ensemble des bijections continues et strictement croissantes de $[0,1]$ dans lui-même et I est l'application identique.

Cette métrique induit une topologie séparable sur $\mathcal{D}$, qu'on appelle topologie de Skorokhod. Une suite (x_n) d'éléments de $\mathcal{D}$ converge pour cette topologie vers un élément x de $\mathcal{D}$, s'il existe une suite (λ_n) d'éléments de Λ telle que : $x_n \circ \lambda_n \longrightarrow x$, et $\lambda_n \longrightarrow \mathrm{I}$ uniformément sur $[0,1]$.

Il est clair qu'une suite qui converge uniformément converge pour cette topologie, et que la réciproque est fausse en général. Ces deux notions sont cependant équivalentes si la fonction limite est continue. En particulier, la topologie trace sur $\mathcal{C}$ de la topologie de Skorokhod coïncide avec la topologie uniforme.

Cependant, $\mathcal{D}$ n'est pas complet pour cette métrique. En effet, la suite $(1_{[\frac{1}{2}, \frac{1}{2}+\frac{1}{n}[})_n$ est une suite de Cauchy car $d(1_{[\frac{1}{2}, \frac{1}{2}+\frac{1}{n}[}, 1_{[\frac{1}{2}, \frac{1}{2}+\frac{1}{m}[}) = \left|\frac{1}{n} - \frac{1}{m}\right|$, mais n'est pas convergente. Mais on peut définir une distance équivalente (au sens qu'elle induit la même topologie) qui fait de $\mathcal{D}$ un espace complet. Ceci est justifié par le fait que la convergence faible des mesures de probabilité définies sur un espace S dépend uniquement de sa topologie, et non de la distance qui engendre cette topologie.

Un autre inconvénient de cette approche est qu'elle ne peut être utilisée pour les processus empiriques que si les observations sont des nombres réels. Une autre approche consiste à utiliser les notions de mesure extérieure et d'intégrale extérieure pour remédier à ce problème (voir par exemple Kosorok, 2008).

1.4.3 Critère de convergence faible

Un résultat important de la convergence faible, est que si $\mathrm{X}_n \xrightarrow{\mathscr{L}} \mathrm{X}$ alors pour toute fonction continue h on a $h(\mathrm{X}_n) \xrightarrow{\mathscr{L}} h(\mathrm{X})$. Une conséquence de ce résultat, dans l'espace $\mathcal{C}$ par exemple, est que la convergence faible de la loi de X_n entraîne la convergence faible des

lois fini-dimensionnelles (car les applications projections sont continues). Dans $\mathcal{D}$ muni de la topologie de Skorokhod, les applications projections ne sont pas toutes continues, mais le résultat suivant reste valable (voir Billingsley, 1968, Théorème 5.1).

L'inverse n'est pas vrai en général, considérons par exemple la suite $(\mathrm{P}_n)_{n\geqslant 0}$ de mesures de Dirac au points $x_n \in \mathcal{C}$ définis pour $n \geqslant 1$ par :

$$x_n(t) = \begin{cases} nt & \text{si } 0 \leqslant t \leqslant \frac{1}{n}, \\ 2 - nt & \text{si } \frac{1}{n} \leqslant t \leqslant \frac{2}{n}, \\ 0 & \text{si } \frac{2}{n} \leqslant t \leqslant 1, \end{cases}$$

et x_0 est la fonction identiquement nulle sur $[0, 1]$. Comme x_n ne converge pas uniformément (c'est-à-dire pour la topologie de $\mathcal{C}$) vers x_0, la suite P_n ne converge pas faiblement vers P_0. Il suffit pour s'en convaincre de prendre B la boule ouverte de $\mathcal{C}$ de centre 0 (la fonction identiquement nulle) et de rayon $\frac{1}{2}$, on a alors : $\mathrm{P}_0(\mathrm{Fr}\,\mathrm{B}) = 1$ (car P_0 est concentrée en 0) mais $\mathrm{P}_n(\mathrm{B}) = 0$ pour tout $n \geqslant 0$ (car $\|x_n - 0\| = 1$). Cependant, pour tout $k > 0$ et tout $(t_1, t_2, \ldots, t_k) \in [0, 1]^k$, le vecteur $(x_n(t_1), x_n(t_2), \ldots, x_n(t_k))$ converge vers le vecteur nul de $\mathbb{R}^k$ (ce qui implique la convergence des lois fini-dimensionnelles).

Toutefois, comme les lois fini-dimensionnelles permettent de caractériser une loi (deux lois ayant les mêmes lois fini-dimensionnelles sont nécessairement identiques), la convergence des lois fini-dimensionnelles peut être utilisée pour montrer l'unicité de la limite si elle existe.

Définition 1.3. Considérons une famille Π de mesures de probabilité sur $(\mathrm{S}, \mathscr{S})$. On dit que Π est relativement compacte si toute suite d'éléments de Π admet une sous suite qui converge faiblement vers une mesure de probabilité (qui n'appartient pas nécessairement à Π)

Si (P_n), la suite des lois de (X_n), est relativement compacte alors de toute sous suite (P_{n_k}) on peut extraire une sous suite $(\mathrm{P}_{n'_k})$ qui converge faiblement vers une mesure de probabilité Q. Si de plus les lois fini-dimensionnelles de P_n convergent vers les lois fini-dimensionnelles de P, alors nécessairement $\mathrm{P} = \mathrm{Q}$. En effet, la convergence de $\mathrm{P}_{n'_k}$ vers Q assure la convergence des lois fini-dimensionnelles respectives, ce qui veut dire que P et Q ont les même lois fini-dimensionnelles, d'où $\mathrm{P} = \mathrm{Q}$. Comme toute sous suite admet une sous suite qui converge vers P alors la suite (P_n) converge elle même vers P.

Remarquons que si $\mathrm{P}_n \Rightarrow \mathrm{P}$ alors nécessairement (P_n) est relativement compacte. Le fait d'imposer la compacité relative n'est pas vraiment restrictif.

En conclusion, pour montrer la convergence faible d'une suite de mesures de probabilité définies sur un espace de fonction, il suffit de montrer la convergence des lois fini-dimensionnelles et la compacité relative.

La question qui se pose maintenant est de trouver un critère assez simple à vérifier pour la compacité relative, et c'est là qu'intervient la notion de tension.

Définition 1.4. Soit (S, d) un espace métrique et $\mathscr{S}$ sa tribu borélienne.

Une mesure de probabilité P définie sur $\mathscr{S}$ est dite tendue si :

$$\forall \varepsilon > 0, \exists \mathrm{K} \subset \mathrm{S} \text{ compact} : \mathrm{P}(\mathrm{K}) \geqslant 1 - \varepsilon.$$

Une famille Π de mesures de probabilité définies sur $\mathscr{S}$ est dite tendue si :

$$\forall \varepsilon > 0, \exists \mathrm{K} \subset \mathrm{S} \text{ compact} : \inf_{\mathrm{P} \in \Pi} \mathrm{P}(\mathrm{K}) \geqslant 1 - \varepsilon.$$

Remarquons que si l'espace S est séparable, alors toutes les mesures de probabilité définies sur $\mathscr{S}$ sont tendues. On est maintenant en mesure d'énoncer le résultat suivant.
Théorème 1.5 (Prokhorov, 1956)**.** *Soit (S, d) un espace métrique et $\mathscr{S}$ sa tribu borélienne, et soit Π une famille de mesures de probabilité définies sur $\mathscr{S}$.*
– *Si Π est tendue, alors elle est relativement compacte.*
– *Si S est séparable et complet, et si Π est relativement compacte, alors elle est tendue.*

Démonstration. voir Billingsley (1968, théorèmes 6.1 et 6.2 p. 37) □

On peut donc énoncer le critère suivant :
Corollaire 1.1. *Soit $(\mathrm{P}_n)_{n \geqslant 0}$ une suite de mesures de probabilité, définies sur $(\mathrm{C}, \mathcal{C})$ ou $(\mathrm{D}, \mathcal{D})$, et P une mesure de probabilité. Si les lois finie-dimensionnelles de P_n convergent vers celles de P et si la suite (P_n) est tendue, alors $\mathrm{P}_n \Rightarrow \mathrm{P}$.*

Signalons le fait que dans la pratique, la tension est généralement prouvée non pas directement mais elle découle de conditions plus "manipulables". Pour des résultats spécifiques aux espaces $\mathcal{C}$ et $\mathcal{D}$ voir Billingsley (1968, Théorèmes 8.2, 8.3, 15.2 et 15.3).

1.5 Théorème de représentation de Skorokhod

Un autre résultat important de la théorie des processus empiriques est le théorème de représentation de Skorokhod.

Considérons une suite (F_n) de fonctions de répartition qui converge faiblement vers une fonction de répartition F, et ξ une v.a. de loi uniforme sur $[0, 1]$, et soit la suite (X_n^*) définie pour tout n par $\mathrm{X}_n^* = \mathrm{F}_n^{inv}(\xi)$ et $\mathrm{X}^* = \mathrm{F}^{inv}(\xi)$. Il est évident (d'après le Théorème 1.2) que X_n^* converge en loi vers X^*. De plus, la convergence faible de (F_n) entraîne que F_n^{inv} converge vers F^{inv} en tout point de continuité de cette dernière, ce qui assure que X_n^* converge presque sûrement vers X^*.

Ce résultat est un cas particulier de celui de Skorokhod (1956), mais illustre bien son utilité : Partant d'une suite de processus aléatoires on peut construire une suite de processus équivalents dont les trajectoires convergent presque sûrement. De là, on peut établir de nouveaux résultats pour cette suite et les généraliser (si possible) à la suite d'origine.

Théorème 1.6. *Soit* $(\mathrm{S}, d, \mathscr{S})$ *un espace Polonais, et soit* $(\mathrm{X}_n)_{n\geqslant 0}$ *une suite d'éléments aléatoires de* S *de lois respectives* $(\mathrm{P}_n)_{n\geqslant 0}$ *tels que* $\mathrm{X}_n \Rightarrow \mathrm{X}_0$.

Alors il existe un espace de probabilité $(\Omega^*, \mathscr{A}^*, \mathrm{P}^*)$ *et une suite* (X_n^*) *d'applications mesurables de* $(\Omega^*, \mathscr{A}^*, \mathrm{P}^*)$ *dans* $(\mathrm{S}, \mathscr{S})$ *qui induisent les lois de probabilité* P_n *(ce qui revient à dire, dans le cas de processus, que les processus* X_n *et* X_n^* *sont équivalents pour tout* n*) telles que :*

$$d(\mathrm{X}_n^*, \mathrm{X}_0^*) \longrightarrow 0 \text{ p.s.}$$

Démonstration. Billingsley (1971) a une preuve assez simple de ce résultat où il prend pour espace $(\Omega^*, \mathscr{A}^*, \mathrm{P}^*)$ l'intervalle de Lebesgue, c'est à dire l'intervalle $[0, 1]$ muni de sa tribu borélienne et de la mesure de Lebesgue. □

Ce résultat permet, par exemple, une preuve élégante du théorème de la fonction continue, ou une version du théorème de la convergence dominée pour la convergence en loi.

Plusieurs auteurs ont donné des extensions de ce résultat. Par exemple, Dudley (1968) a prouvé ce résultat sans l'hypothèse de complétude. Wichura (1970) et Fernandez (1974) ont montré une extension de ce résultat au cas où S n'est pas nécessairement séparable (dans ce cas, la définition de la convergence faible est donnée par rapport à la tribu engendrée par les boules ouvertes, qui coïncide avec la tribu borélienne dans le cas séparable), mais avec l'hypothèse que le support de P_0 soit séparable. Voir par exemple Dudley (1976) pour une preuve de ce résultat, aussi connu sous le nom du théorème de Skorokhod-Dudley-Wichura.

Par exemple, dans l'espace $\mathcal{D}$ des fonctions cadlag on peut utiliser soit la distance de Skorokhod (qui a été introduite pour ce résultat même) ainsi que sa tribu borélienne, soit la distance uniforme et la tribu engendrée par les boules ouvertes, ainsi que la définition de Dudley (1966) de la convergence faible (à condition, bien sûr, que le support de la loi limite soit séparable pour la distance uniforme).

1.6 Théorème de Donsker

Le théorème de la limite centrale donne, pour tout $t \in \mathbb{R}$, une convergence en loi de $\sqrt{n}(\mathrm{F}_n(t) - \mathrm{F}(t))$ vers une v.a. de loi normale centrée et de variance $\mathrm{F}(t)[1-\mathrm{F}(t)]$. Plus généralement, on peut montrer que pour tous $t_1, \ldots, t_k \in \mathbb{R}$, le vecteur aléatoire $\sqrt{n}(\mathrm{F}_n(t_1) - \mathrm{F}(t_1), \ldots, \mathrm{F}_n(t_k) - \mathrm{F}(t_k))$ converge en loi vers un vecteur Gaussien centré $(\mathrm{G}(t_1), \ldots, \mathrm{G}(t_k))$ avec

$$\mathrm{C}ov(\mathrm{G}(s), \mathrm{G}(t)) = \mathrm{F}(\min(s,t)) - \mathrm{F}(s)\mathrm{F}(t) \text{ pour } s, t \in \{t_1, \ldots, t_k\}.$$

Donsker (1952) a montré un résultat encore plus fort : en regardant la fonction de répartition empirique comme un élément aléatoire de l'espace $\mathcal{D}$ des fonctions "cadlag",

$\sqrt{n}(\mathrm{F}_n(x) - \mathrm{F}(x))$ converge faiblement vers un processus Gaussien.

Théorème 1.7. *Soit* $(\mathrm{X}_n)_{n\geqslant 0}$ *une suite de variables aléatoires i.i.d. de fonction de répartition* F*, et* $\mathrm{Y}_n = \sqrt{n}(\mathrm{F}_n - \mathrm{F})$ *le processus empirique associé.*

Alors : $\mathrm{Y}_n \Rightarrow \mathrm{Y}$ *où* Y *est un processus Gaussien centré de fonction de covariance* $\mathrm{E}(\mathrm{Y}(s)\mathrm{Y}(t)) = \mathrm{F}(s)(1 - \mathrm{F}(t))$ *pour* $s \leqslant t$*.*

Chapitre 2

La loi du logarithme itéré

La loi du logarithme itéré (ou LIL pour "Law of the Iterated Logarithm") est un des théorème limites important de la statistique, nous commençons donc par donner la version classique de ce résultat (pour la somme d'une suites de variables aléatoires i.i.d.) avant de passer aux processus empiriques.

2.1 LIL classique

On considère une suite (X_n) de v.a. i.i.d. centrées de variance 1, et $\mathrm{S}_n = \sum_{i=1}^n \mathrm{X}_i$. La loi forte des grands nombres donne une convergence presque sûre de $\frac{\mathrm{S}_n}{n}$ vers 0. Le théorème de la limite centrale, donne par contre la convergence en loi de la suite $\frac{\mathrm{S}_n}{\sqrt{n}}$ vers la loi normale $\mathcal{N}(0,1)$.

La loi du logarithme itéré, donne un résultat pour une suite "intermédiaire" : $\frac{\mathrm{S}_n}{\sqrt{n \log_2 n}}$ où $\log_2 n = \log\log n$.[1] D'où le nom de "logarithme itéré". Notons que cette suite converge en probabilité vers 0 (il suffit d'utiliser l'inégalité de Tchebychev), mais on n'a pas la convergence presque sûre :

Théorème 2.1 (Loi du logarithme itéré)**.** *Soient* (X_n) *une suite de v.a i.i.d. centrées de variance 1, et* $\mathrm{S}_n = \sum_{i=1}^n \mathrm{X}_i$*. Alors :*

$$\limsup_{n\to\infty} \frac{\mathrm{S}_n}{\sqrt{n \log_2 n}} = \sqrt{2} \quad \textit{p.s.}$$

Ce résultat, dû à Hartman et Wintner (1941) a été d'abord prouvé pour des v.a de Bernoulli par Khinchine (1924), puis par Kolmogorov (1929) pour des v.a. de loi normale. Une preuve plus moderne de ce résultat est donnée par de Acosta (1983).

1. Par souci de rigueur, on remplace parfois $\log$ par $\log_+$ défini par $\log_+ x = \log(\sup(x, e))$, ce qui ne change rien vu que tous les résultats donnés sont pour la limite quand $n \to \infty$

2.2 LIL pour les processus empiriques

Dans cette section, nous allons exposer des résultats qui seront, ainsi que leur généralisation au cas des données censurées, essentiels pour la suite. Considérons d'abord le cas du processus empirique uniforme, la généralisation au cas d'une loi continue quelconque étant immédiate.

Soient $\mathrm{U}_1, \ldots, \mathrm{U}_n$ des v.a. i.i.d. de loi $\mathcal{U}_{[0,1]}$ et soient la fonction de répartition empirique $\mathbb{U}(t) = \frac{1}{n}\sum 1_{\{\mathrm{U}_i \leqslant t\}}$ et le processus empirique $\alpha_n(t) = \sqrt{n}(\mathbb{U}(t) - t)$ associés.
Théorème 2.2 (Smirnov). *Soit $b_n = \sqrt{2\log_2 n}$. Alors :*

$$\limsup_{n\to\infty} \frac{\|\alpha_n\|}{b_n} = \limsup_{n\to\infty} \frac{\|n(\mathbb{U}_n - \mathrm{I})\|}{\sqrt{n}b_n} = \frac{1}{2} \quad p.s.$$

où $\|x\| = \sup_{t\in[0,1]} |x(t)|$ est la norme de la convergence uniforme.

Le résultat suivant, qui a été prouvé indépendamment par Chung (1949), est plus fort au sens qu'il implique le résultat précédent (voir Shorack et Wellner, 1986, Chapitre 13) :
Théorème 2.3. *Soit (λ_n) une suite croissante de nombres positifs, alors :*

$$\mathrm{P}\left(\limsup_{n\to\infty} \{\|\alpha_n\| \geqslant \lambda_n\}\right) = \begin{cases} 0 \textit{ si } \sum \frac{\lambda_n}{n}\exp(-2\lambda_n^2) < \infty \\ 1 \textit{ sinon} \end{cases}$$

2.3 Lois fonctionnelles du logarithme itéré

Un autre résultat très important est celui de Finkelstein (1971), il consiste à prouver une loi du logarithme itéré pour le processus empirique considéré comme un élément de $\mathcal{B}$, l'ensemble des fonctions réelles bornées, et pas seulement pour la norme uniforme de ce processus, comme c'est le cas pour les résultats précédents. Un tel résultat est parfois appelé loi fonctionnelle du logarithme itéré.

Avant d'aller plus loin, donnons quelques définitions
Définition 2.1. Soient $(\mathrm{X}_n)_{n\geqslant 0}$ une suite d'éléments aléatoires d'un espace métrique (S, d) définies sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathcal{P})$. On dit que (X_n) est presque sûrement relativement compacte dans (S, d) avec pour ensemble limite H, s'il existe $\Omega_0 \in \mathcal{A}$ avec $\mathrm{P}(\Omega_0) = 1$ tel que pour tout $\omega \in \Omega_0$:

1. toute suite n' de nombres entiers admet une sous suite n'' telle que $\mathrm{X}_{n''}(\omega)$ converge dans (S, d) ;
2. toute les valeurs d'adhérence de $\mathrm{X}_n(\omega)$ appartiennent à H ;
3. pour tout $h \in \mathrm{H}$, il existe une suite $n' = n_{h,\omega}$ telle que $\mathrm{X}_{n'}(\omega)$ converge vers h.

Définition 2.2. Soit h une fonction définie sur un intervalle I et à valeurs réelles. On dit que h est absolument continue si $\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1 \leqslant i \leqslant \mathrm{N}}$ intervalles disjoints de I :

$$\sum_{i=1}^{\mathrm{N}} (y_i - x_i) < \delta \implies \sum_{i=1}^{\mathrm{N}} |h(y_i) - h(x_i)| < \varepsilon.$$

On dit que h est absolument continue par rapport à une mesure μ (ou une fonction de répartition en sous entendant que c'est par rapport à la mesure de probabilité liée à cette fonction), si $\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1 \leqslant i \leqslant \mathrm{N}}$ intervalles disjoints de I :

$$\sum_{i=1}^{\mathrm{N}} \mu(]x_i, y_i[) < \delta \implies \sum_{i=1}^{\mathrm{N}} |h(y_i) - h(x_i)| < \varepsilon.$$

Considérons l'ensemble

$$\mathscr{H} = \left\{ \begin{array}{c} h : h \text{ est absolument continue sur } [0,1] \text{ avec} \\ h(0) = h(1) = 0 \quad \text{et} \quad \displaystyle\int_0^1 [h'(t)]^2 \, dt \leqslant 1 \end{array} \right\},$$

où h' est la dérivée au sens de Lebesgue de h (c'est à dire que $h(t) = \int_0^t h'(u)\, du$, les intégrales étant ici au sens de Lebesgue).

On donne d'abord le théorème pour le cas uniforme.

Théorème 2.4. *Soit $b_n = \sqrt{2 \log_2 n}$. Alors la suite $\left\{ \frac{\alpha_n}{b_n} \right\}$ est presque sûrement relativement compacte dans $\mathcal{B}([0,1])$ avec pour ensemble limite l'ensemble $\mathscr{H}$ ci-dessus.*

Il peut se généraliser au cas de variables aléatoires de lois continues quelconques :

Théorème 2.5. *Soit $(\mathrm{X}_n)_{n \geqslant 1}$ une suite de variables aléatoires i.i.d. ayant une fonction de répartition F définie et continue sur un intervalle $[a,b]$, et soit a_n le processus empirique associé. Alors la suite $\left\{ \frac{a_n}{b_n} \right\}$ est presque sûrement relativement compacte dans $\mathcal{B}([0,1])$ avec pour ensemble limite l'ensemble $\mathscr{H}_{\mathrm{F}}$ des fonctions f définies sur $[a,b]$ et vérifiant :*

- $f(a) = f(b) = 0$,
- *f est absolument continue par rapport à* F,
- $\int_a^b (df/d\mathrm{F})^2 \, d\mathrm{F} \leqslant 1$ *où $df/d\mathrm{F}$ est la dérivée de f par rapport à* F.

Chapitre 3

Le processus empirique de Kaplan-Meier

3.1 Introduction aux données censurées

Dans de nombreuses applications statistiques, on est amené à étudier une variable de durée, c'est-à-dire un délai séparant un instant initial de l'instant où un événement d'intérêt est observé. C'est souvent le cas pour les applications médicales liées à l'étude des durées de survie. Pour cette raison, et bien que le champ d'application des méthodes que nous allons exposer ne se limite pas à la médecine ou à la biologie, nous utilisons la terminologie associées aux durées de survie. Celle-ci est la plus commune dans la littérature statistique, puisque c'est précisément dans le domaine médical que les avancées méthodologiques liées à ces variables ont été développées en premier.

Nous parlons de données censurées lorsque la durée de survie n'est connue que lorsqu'elle est dans les limites des durées d'observation. Ces limites pouvant être imposée par le type d'observation (par exemple, la durée d'hospitalisation d'un malade) ou par des événements fortuits (comme un accident ou une migration du patient au cours de son suivi). Par exemple, dans le cas dit de censure à droite, seul le minimum entre la durée de survie et une durée limite supérieure d'observation est connu, ainsi que l'indicateur exprimant si la durée de survie a été censurée ou non. Ce minimum constitue, en quelque sorte, une durée de participation qui est observée en ne donnant qu'une information partielle sur la vraie durée de survie. Il existe plusieurs autres catégories de modèles de censure, obtenus par des variations du principe de la censure à droite décrit ci-dessus. Parmi ceux-ci, mentionnons les suivants, en y incluant le modèle de censure à droite.

Censure à droite Il y a censure à droite lorsque la durée de survie est supérieure à la durée de participation. Un exemple typique est celui où l'événement considéré est le décès d'un patient malade, et la durée d'observation est une durée totale d'hospitalisation. On peut aussi observer ce genre de phénomène dans des études de fiabilité quand la panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant. L'expérimentateur peut également fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

Pour ce type de censure, tout ce que l'on sait est que la vraie durée est supérieure à la durée observée.

Censure à gauche Il y a censure à gauche lorsque la durée de survie est inférieure à la durée observée.

Supposons par exemple que nous étudions la fiabilité d'un certain composant électronique qui est branché en parallèle avec un ou plusieurs autres composants. Une panne de ce composant n'entraîne pas nécessairement l'arrêt du système : le système peut continuer à fonctionner, quoique de façon aberrante, jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). La durée observée pour ce composant est alors censurée à gauche.

Un autre exemple est que l'on s'intéresse à l'age à partir duquel une personne commence à accomplir une certaine tâche. Certaines personnes peuvent ne pas se rappeler, et donner juste une valeur supérieure (le cas extrême est que l'on prenne le début de l'étude comme observation). Cette donnée est donc censurée à gauche.

Dans ce dernier exemple, ainsi que dans beaucoup de cas, on trouve des données censurées à gauche dans un même échantillon que des données censurées à droite, ce qui conduit à la définition suivante.

Censure double (ou mixte) On dit qu'on a une censure double si on a des données censurées à droite et des données censurées à gauche dans le même échantillon. C'est l'objet d'étude du chapitre 5

Censure par intervalle Dans le cas de la censure par intervalle, on observe à la fois une borne inférieure et une borne supérieure de la durée d'intérêt. Ceci arrive dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. On a aussi pour ce genre d'expériences des données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de représenter les données censurées à droite ou à gauche par des intervalles du type

$[a, +\infty[$ et $[0, a]$ respectivement, ce qui permet de considérer ce modèle comme étant plus générique. Turnbull (1976) présente ce genre de censure avec plus de détails.

Les trois catégories de censure décrits ci-dessus peuvent se décliner en fonction du mode ou mécanisme de censure. On obtient alors les types suivants :

Censure de type I L'expérimentateur fixe une date (non aléatoire) de fin d'expérience. La durée de participation maximale est alors fixée (non aléatoire) et vaut, pour chaque observation, la différence entre la date de fin d'expérience, et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quant à lui, aléatoire. Ce modèle est souvent utilisé dans les études épidémiologiques.

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

Censure aléatoire C'est typiquement ce modèle qui est utilisé pour les essais thérapeutique. Dans ce type d'expérience, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'événements observés et la durée totale de l'expérience sont aléatoires.

Dans ce mémoire, nous nous restreignons à l'étude de la censure aléatoire, principalement pour les données censurées à droite. Dans le chapitre 5, nous présentons quelques perspectives de recherche pour les données doublement censurées.

3.2 L'estimateur de Kaplan-Meier

Soient $\mathrm{X}_1, \ldots, \mathrm{X}_n$ un échantillon représentant les durée d'intérêt (ces variables sont donc supposées positives), de fonction de répartition F, et $\mathrm{C}_1, \ldots, \mathrm{C}_n$ un échantillon représentant les temps de censure, que l'on suppose indépendants des durées d'intérêt, de fonction de répartition G. Dans le modèle de censure aléatoire à droite, on observe non pas la durée d'intérêt X_i mais plutôt la plus petite des deux valeurs $\mathrm{Z}_i = \min(\mathrm{X}_i, \mathrm{C}_i)$, ainsi que l'indicateur de censure δ_i qui vaut 1 si la durée d'intérêt est observée, et 0 si elle est censurée, i.e. $\delta_i = 1_{\{\mathrm{X}_i \leqslant \mathrm{C}_i\}}$.

Dans ce genre de données, qui sont souvent des durées de survie ou des données de fiabilité, la fonction de répartition F est estimée par l'estimateur introduit par Kaplan et

Meier (1958), donné pour $z < \mathrm{Z}_{(n)} = \max\{\mathrm{Z}_1, \ldots, \mathrm{Z}_n\}$ par

$$\mathrm{F}_n(z) = 1 - \prod_{i:\mathrm{Z}_i \leqslant z} \left(\frac{\mathrm{N}_n(\mathrm{Z}_i) - 1}{\mathrm{N}_n(\mathrm{Z}_i)} \right),$$

où $\mathrm{N}_n(x) = \sum_{i=1}^n 1_{\{\mathrm{Z}_i \geqslant x\}}$. Pour $z \geqslant \mathrm{Z}_{(n)}$, il y a plusieurs conventions pour définir $\mathrm{F}_n(z)$: Soit on le définit par $\mathrm{F}_n(\mathrm{Z}_{(n)})$, ce qui fait que F_n peut ne pas être une fonction de répartition si $\mathrm{Z}_{(n)}$ est une donnée censurée, soit on le définit par 0, soit on le laisse non défini.

Cet estimateur a des propriétés assez similaires à la fonction de répartition empirique, par exemple la convergence uniforme presque sûre (Winter *et al.*, 1978; Stute et Wang, 1993), la normalité asymptotique (Breslow et Crowley, 1974; Gill, 1983), et la loi du logarithme itéré (Földes et Rejtő, 1981). Ceci justifie que l'on s'intéresse à généraliser la théorie des processus empiriques au cas des données censurées.

Aussi, on définit *le processus empirique de Kaplan-Meier* par $a_n(t) = \sqrt{n}(\mathrm{F}_n(t) - \mathrm{F}(t))$.

3.3 La loi du logarithme itéré pour l'estimateur de Kaplan-Meier

Le résultat suivant est une loi du logarithme itéré pour l'estimateur de Kaplan-Meier de la fonction de répartition, qui est d'une certaine manière similaire au théorème de Chung-Smirnov.

On définit l'estimateur de Kaplan-Meier en utilisant la deuxième convention plus haut, c'est à dire en posant $\mathrm{F}_n(z) = 0$ pour $z > \mathrm{Z}_{(n)}$. Pour toute fonction de répartition L, on note par $\mathrm{T}_\mathrm{L} = \max\{t : \mathrm{L}(t) < 1\}$ le point terminal du support de L.

Théorème 3.1 (Földes et Rejtő, 1981)**.** *On suppose que* F *et* G *sont continues, et que* $\mathrm{T}_\mathrm{F} < \mathrm{T}_\mathrm{G}$. *Alors,*

$$\mathrm{P}\left(\sup_{-\infty<u<+\infty} |\mathrm{F}_n(u) - \mathrm{F}(u)| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}} \right) \right) = 1.$$

La condition $\mathrm{T}_\mathrm{F} < \mathrm{T}_\mathrm{G}$ pouvant paraître restrictive, on peut citer le théorème autrement :

Corollaire 3.1 (Földes et Rejtő, 1981)**.** *On suppose que* F *et* G *sont continues, et on considère* T *tel que* $\mathrm{G}(\mathrm{T}) > 0$. *Alors,*

$$\mathrm{P}\left(\sup_{-\infty<u\leqslant \mathrm{T}^*} |\mathrm{F}_n(u) - \mathrm{F}(u)| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}} \right) \right) = 1,$$

où $\mathrm{T}^* = \min\{\mathrm{T}, \mathrm{T}_\mathrm{F}\}$.

3.4 Autres théorèmes limites pour le processus empirique de Kaplan-Meier

Dans le modèle de censure aléatoire à droite que nous venons de voir, un rôle dominant est joué par les fonctions H et $\mathrm{H}^{(1)}$ avec

$$\mathrm{H}(t) = \mathrm{P}(\mathrm{Z}_i \leqslant t) = 1 - (1 - \mathrm{F}(t))(1 - \mathrm{G}(t)),$$

$$\mathrm{H}^{(1)}(t) = \mathrm{P}(\mathrm{Z}_i \leqslant t, \delta_i = 1) = \int_0^t (1 - \mathrm{G}(s))\, d\mathrm{F}(s),$$

ainsi que les versions empiriques

$$\mathrm{H}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{[0,t]}(\mathrm{Z}_i),$$

$$\mathrm{H}_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i 1_{[0,t]}(\mathrm{Z}_i).$$

La méthode classique pour traiter ce genre de données consistait à considérer $n\mathrm{H}_n^{(1)}$ comme un processus de comptage (voir par exemple Aalen, 1976; Fleming et Harrington, 1991). Son compensateur est donné par

$$n \int_0^t (1 - \mathrm{H}_n(s_-))\, d\Lambda(s),$$

où $\Lambda(t) = -log(1-\mathrm{F}(t))$ est la fonction de hasard de X_i. La martingale de base est définie par

$$\mathrm{M}_n(t) = n \left(\mathrm{H}_n^{(1)}(t) - \int_0^t (1 - \mathrm{H}_n(s_-))\ d\Lambda(s) \right).$$

C'est une martingale par rapport à la filtration $\mathscr{F}_x = \sigma\{\mathrm{Z}_i 1_{\{\mathrm{Z}_i \leqslant x\}}, \delta_i 1_{\{\mathrm{Z}_i \leqslant x\}}\}$. La martingale de base joue un rôle fondamental car plusieurs processus intéressants peuvent s'écrire comme des intégrales stochastiques par rapport à cette martingale, c'est à dire sous la forme

$$\mathrm{Q}_n(t) = \int_0^t \mathrm{L}_n(s)\, d\mathrm{M}_n(s).$$

Un processus particulièrement intéressant qui s'écrit sous cette forme est le processus empirique de Kaplan-Meier divisé par $1 - \mathrm{F}$ qui s'écrit

$$\begin{aligned} \Pi_n(t) &= \frac{a_n(t)}{1 - \mathrm{F}(t)} = \frac{\sqrt{n}\,(\mathrm{F}_n(t) - \mathrm{F}(t))}{1 - \mathrm{F}(t)} \\ &= \frac{1}{\sqrt{n}} \int_0^t \frac{1 - \mathrm{F}_n(s_-)}{1 - \mathrm{F}(s)} \frac{1}{1 - \mathrm{H}_n(s_-)} 1_{\{\mathrm{H}_n(s_-) < 1\}}\, d\mathrm{M}_n(s) \end{aligned}$$

Einmahl et Koning (1992) étudient une version pondérée du processus $\mathbb{Q}_n$ et donnent des conditions nécessaires et suffisantes pour obtenir des résultats analogues au cas des données complètes, à savoir le théorème de Chibisov (1964) et O'Reilly (1974), le théorème de Glivenko-Cantelli tel que formulé par Lai (1974) et Wellner (1977) ainsi que la loi fonctionnelle du logarithme itéré de James (1975).

Ces théorèmes, donnés dans un cadre général, s'appliquent au processus Π_n donné plus haut, et peuvent même s'appliquer (à quelque modifications près) au processus empirique de Kaplan-Meier lui même.

Ces résultats, ainsi que d'autres résultats récents, n'utilisent cependant pas les méthodes de martingales (bien qu'ils utilisent ce processus ainsi que la martingale de base).

Bien entendu, les résultats valables pour le processus pondéré restent valables pour le processus originel (il suffit de prendre la fonction poids identiquement égale à 1).

Chapitre 4

Sur le comportement limite des fonctionnelles locales des processus empiriques dans un modèle de censure à droite

4.1 Introduction

Dans ce chapitre, nous allons étudier un résultat intéressant de la théorie des processus empiriques en la présence de données censurées à droite. Il s'agit d'une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique de Kaplan-Meier, aussi appelés processus empirique local.

Un tel résultat peut être utilisé pour donner des vitesses de convergence pour certains estimateurs à noyau. Nous nous intéressons tout particulièrement à l'estimateur à noyau de la densité, bien qu'il ne soit pas le seul estimateur auquel ce résultat s'applique.

L'estimateur à noyau de la densité est donné par :

$$f_n(z) = \int_{-\infty}^{+\infty} \frac{1}{h_n} \mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}_n(t),$$

où h_n (appelée fenêtre) est une suite de nombres strictement positifs et K (appelé noyau) est une fonction définie de $\mathbb{R}$ dans $\mathbb{R}$, sur lesquels des conditions sont imposées plus loin. Cet estimateur est une généralisation de l'estimateur de Parzen-Rosenblatt (Parzen, 1962; Rosenblatt, 1956) au cas des données censurées à droite.

Les résultats de ce chapitre sont dûs à Deheuvels et Einmahl (1996). Remarquons qu'ils impliquent la convergencence ponctuelle d'estimateurs de la densité, résultat qui a été

d'abord étendu au cas uniforme sur un compact dans Deheuvels (2000) ; ensuite dans Viallon (2006) ce dernier a été aussi généralisé à une convergence uniforme par rapport à une fenêtre appartenant à un compact bien spécifié.

Notations Nous utilisons les mêmes notations que précédemment : Soient $\mathrm{X}_1, \ldots, \mathrm{X}_n$ les durées d'intérêt, indépendantes et de fonction de répartition F, et indépendamment d'elles, soient $\mathrm{C}_1, \ldots \mathrm{C}_n$ les durées de censure, indépendantes et de fonction de répartition G. Et on observe les couples $(\mathrm{Z}_1, \delta_1), \ldots, (\mathrm{Z}_n, \delta_n)$ avec $\mathrm{Z}_i = \min\{\mathrm{X}_i, \mathrm{C}_i\}$ et $\delta_i = 1_{\{\mathrm{X}_i \leqslant \mathrm{C}_i\}}$.

Pour toute fonction R, on note :

$$\mathrm{R}_+(x) = \lim_{\varepsilon \downarrow 0} \mathrm{R}(x+\varepsilon) \quad \text{et} \quad \mathrm{R}_-(x) = \lim_{\varepsilon \downarrow 0} \mathrm{R}(x-\varepsilon),$$

si ces limites existent, et pour toute fonction de répartition L, on note :

$$\mathrm{T}_{\mathrm{L}} = \sup\{t : \mathrm{L}(t) < 1\} \quad \text{et} \quad \mathrm{L}_-(\infty) = \lim_{x \to \infty} \mathrm{L}(x).$$

On suppose que $\mathrm{F}_-(\infty) = 1$, mais on accepte pour la distribution de Y que $\mathrm{G}_-(\infty) = 1 - \mathrm{P}(\mathrm{Y} = \infty) \leqslant 1$. En particulier, si $\mathrm{P}(\mathrm{Y} = \infty) = 1$, on revient au cas non censuré.

Posons $\Theta = \min\{\mathrm{T}_{\mathrm{F}}, \mathrm{T}_{\mathrm{G}}\}$. On suppose que $\Theta > 0$ (si $\Theta = 0$, alors F_n est presque sûrement dégénérée).

On note f_+ et g_+ (resp. f_- et g_-) les dérivées à droite (resp. à gauche) de F et G (resp. de F_- et G_-) si elles existent.

Remarquons que l'existence de f_+ et f_- (resp. g_+ et g_-) en un point donné ne garantit pas la continuité de F (resp. de G) en ce point. Si F (resp. G) est continue, alors f_+ et f_- (resp. g_+ et g_-) ne sont que les dérivées à droite et à gauche de F (resp. G).

Si $\mathcal{J}$ est un ensemble, on note $\mathbb{B}(\mathcal{J})$ l'ensemble des fonctions réelles bornées définies sur $\mathcal{J}$, muni de la norme de la convergence uniforme.

Hypothèses et résultats La fenêtre h_n doit vérifier les conditions suivantes :

H1 $h_n \downarrow 0$; $\quad nh_n \uparrow \infty$.

H2 $\dfrac{nh_n}{\log_2 n} \to \infty$ où $\log_2 u = \log\log u$.

Ces conditions sont utilisées pour les théorèmes 4.1 et 4.2 ci dessous. Un exemple de fenêtre vérifiant ces hypothèses est de la forme $n^{-\alpha}$ avec $0 < \alpha < 1$. La condition (H1) peut être affaiblie en :

H1' Il existe une suite h_n^* vérifiant (H1) et telle que :

$$0 < \liminf_{n\to\infty} \left(\frac{h_n}{h_n^*}\right) \leqslant \limsup_{n\to\infty} \left(\frac{h_n}{h_n^*}\right) < \infty.$$

Les hypothèses sur le noyau K sont :

K1 K est à variation bornée

K2 K est à support compact

K3 $\int_{-\infty}^{+\infty} \mathrm{K}(u)\, du = 1$.

Remarquons que la seule condition qui puisse être considérée comme restrictive est (K2). Parmi les noyaux les plus utilisés, seul le noyau Gaussien ne vérifie pas cette condition. De plus, ces hypothèses sont uniquement utilisées pour le Théorème 4.2.

L'article de Deheuvels et Einmahl (1996) est structuré comme suit : Dans un premier temps, ils donnent un résultat analogue (mais plus simple, vu qu'il ne considère que le cas réel) au résultat de Deheuvels et Mason (1994) dans le cas des données censurée à droite. Puis, ils appliquent ce théorème pour obtenir la vitesse de convergence de l'estimateur à noyau de la densité. L'avantage de cette approche réside dans le fait que les résultats ainsi obtenus peuvent être utilisés pour donner les vitesses de convergence d'autres types d'estimateurs.

Pour $z \in]0, \Theta[$ fixé, on considère la suite de fonctions aléatoires définie par :

$$\xi_n(u) = \frac{1}{b_n}(a_n(z + h_n u) - a_n(z)) \quad \text{pour } -\mathrm{M} \leqslant u \leqslant \mathrm{M}.$$

où M est une constante donnée, $b_n = \sqrt{2h_n \log_2 n}$, et a_n est le processus empirique de Kaplan-Meier défini par :

$$a_n(x) = \sqrt{n}(\mathrm{F}_n(x) - \mathrm{F}(x)).$$

Nous allons maintenant énoncer le résultat principal, objet de notre étude. Il donne le comportement asymptotique de la suite (ξ_n) :

Théorème 4.1 (Deheuvels et Einmahl, 1996, Théorème 1.2)**.** *Soit $z \in]0, \Theta[$ fixé. On suppose que F est continue dans un voisinage de z et que $f_+(z)$ et $f_-(z)$ existent. Alors, sous les hypothèses (H1) et (H2), la suite $(\xi_n)_{n\geqslant 1}$ est presque sûrement relativement compacte dans $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$ avec pour ensemble limite l'ensemble des fonctions $h \in \mathbb{B}([-\mathrm{M}, \mathrm{M}])$ de la forme*

$$h(u) = \int_0^u \psi(s)\, ds; \qquad -\mathrm{M} \leqslant u \leqslant \mathrm{M},$$

avec

$$\frac{1 - \mathrm{G}_-(z)}{f_-(z)} \int_{-\mathrm{M}}^0 \psi^2(s)\, ds + \frac{1 - \mathrm{G}(z)}{f_+(z)} \int_0^{\mathrm{M}} \psi^2(s)\, ds \leqslant 1.$$

Remarque 4.1. Si $f_-(z) = 0$, on remplace la condition précédente par :

$$\int_{-\mathrm{M}}^{0} \psi^2(s)\, ds = 0 \quad ; \quad \int_{0}^{\mathrm{M}} \psi^2(s)\, ds \leqslant \frac{f_+(z)}{1-\mathrm{G}(z)};$$

et de même pour $f_+(z) = 0$.

Si F est dérivable en z et G continue en z, la condition précédente devient

$$h(u) = \int_0^u \psi(s)\, ds; \qquad -\mathrm{M} \leqslant u \leqslant \mathrm{M},$$

avec

$$\int_{-\mathrm{M}}^{\mathrm{M}} \psi^2(s)\, ds \leqslant \frac{f(z)}{1-\mathrm{G}(z)}.$$

La preuve de ce théorème se déduit d'une succession de lemmes à la section 4.2.

Nous allons maintenant énoncer le résultat pour l'estimateur à noyau de la densité. Posons

$$\mathbb{E}f_n(z) = \int_{-\infty}^{+\infty} \frac{1}{h_n}\mathrm{K}\left(\frac{t-z}{h_n}\right)\, d\mathrm{F}(t).$$

En général $\mathbb{E}f_n(z) \neq \mathrm{E}(f_n(z))$, sauf dans le cas non censuré.

Théorème 4.2 (Deheuvels et Einmahl, 1996, Théorème 1.1)**.** *Soit $z \in]0, \Theta[$ fixé. On suppose que F est continue dans un voisinage de z et admet des dérivées à droite et à gauche en z. Alors sous les hypothèses (H1),(H2), (K1), (K2) et (K3) on a :*

$$\limsup_{n\to\infty} \pm\sqrt{\frac{nh_n}{2\log_2 n}}(f_n(z) - \mathbb{E}f_n(z)) = \left(\frac{f_-(z)}{1-\mathrm{G}_-(z)}\int_{-\infty}^{0} \mathrm{K}^2(t)\, dt + \frac{f_+(z)}{1-\mathrm{G}(z)}\int_0^{+\infty} \mathrm{K}^2(t)\, dt\right)^{1/2} \quad p.s.$$

La preuve de ce théorème est reportée à la section 4.3.

Remarque 4.2. Si on suppose que F est dérivable en z, et G continue en z, alors la relation précédente devient :

$$\limsup_{n\to\infty} \pm\sqrt{\frac{nh_n}{2\log_2 n}}(f_n(z) - \mathbb{E}f_n(z)) = \left(\frac{f(z)}{1-\mathrm{G}(z)}\int_{-\infty}^{+\infty} \mathrm{K}^2(t)\, dt\right)^{1/2} \quad \text{p.s}$$

On suppose F éventuellement discontinue en z, mais que $f_+(z)$ et $f_-(z)$ existent, alors, sous (H1), (K1) et (K2) on a :

$$\lim_{n\to\infty} \mathbb{E}f_n(z) = \frac{1}{2}(f_+(z) + f_-(z)) + \frac{1}{2}(f_+(z) - f_-(z))\int_0^{+\infty} (\mathrm{K}(t) - \mathrm{K}(-t))\, dt,$$

qui se réduit à $f(z)$ si $f_-(z) = f_+(z) = f(z)$.

4.2 Lois fonctionnelles pour les processus basés sur des données censurées

En plus des notations précédentes, on utilisera pour la suite les notations suivantes : Soit H la fonction de répartition de Z

$$\mathrm{H}(x) = \mathrm{P}(\mathrm{Z} \leqslant x) = 1 - (1 - \mathrm{F}(x))(1 - \mathrm{G}(x)),$$

elle est décomposée en somme de $\mathrm{H}^{(1)}$ et $\mathrm{H}^{(0)}$ avec :

$$\mathrm{H}^{(1)}(x) = \mathrm{P}(\mathrm{Z} \leqslant x, \delta = 1) = \int_0^x (1 - \mathrm{G}_-(t))\, d\mathrm{F}(t), \tag{4.1}$$

$$\mathrm{H}^{(0)}(x) = \mathrm{P}(\mathrm{Z} \leqslant x, \delta = 0) = \int_0^x (1 - \mathrm{F}(t))\, d\mathrm{G}(t),$$

Posons :

$$p = \mathrm{P}(\delta = 1) = \int_0^{+\infty} (1 - \mathrm{G}_-(t))\, d\mathrm{F}(t) = \mathrm{H}_-^{(1)}(\infty).$$

Les hypothèses excluent le cas $p = 0$, par contre le cas $p = 1$ est possible et correspond au cas non censuré. Ce dernier a été traité de manière similaire par Deheuvels et Mason (1994), on peut l'exclure sans perte de généralité. On suppose donc que $0 < p < 1$.

Soient les fonctions $\mathrm{Q}^{(1)}$ et $\mathrm{Q}^{(0)}$ définies par :

$$\begin{aligned} \mathrm{Q}^{(1)}(s) &= \inf\{x : \mathrm{H}^{(1)}(x) \geqslant s\} \quad \text{pour } 0 < s < p, \\ \mathrm{Q}^{(0)}(s) &= \inf\{x : \mathrm{H}^{(0)}(x) \geqslant s\} \quad \text{pour } 0 < s < 1 - p. \end{aligned}$$

Ce sont les inverses généralisées (ou fonctions quantile) de $\mathrm{H}^{(1)}$ et $\mathrm{H}^{(0)}$ respectivement. Ces définitions impliquent :

$$\begin{aligned} \mathrm{Q}^{(1)}(s) \leqslant x &\iff s \leqslant \mathrm{H}^{(1)}(x) \quad \text{pour } 0 < s < p \\ \mathrm{Q}^{(0)}(s) \leqslant x &\iff s \leqslant \mathrm{H}^{(0)}(x) \quad \text{pour } 0 < s < 1 - p \end{aligned}$$

Les fonctions de répartitions empiriques correspondant à H, $\mathrm{H}^{(1)}$, $\mathrm{H}^{(0)}$ sont données par

$$\begin{aligned} \mathrm{H}_n(x) &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{\mathrm{Z}_i \leqslant x\}}, \\ \mathrm{H}_n^{(1)}(x) &= \frac{1}{n} \sum_{i=1}^{n} \delta_i 1_{\{\mathrm{Z}_i \leqslant x\}}, \\ \mathrm{H}_n^{(0)}(x) &= \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) 1_{\{\mathrm{Z}_i \leqslant x\}}. \end{aligned}$$

Réduction au cas uniforme Comme on l'a fait dans le cas des données complètes (section 1.3), on peut construire une suite de variables aléatoires de loi uniforme sur $[0,1]$, et telle que la suite des observations (Z_i, δ_i) puisse s'écrire en fonction de cette suite. Notons $\mathrm{Z} = \mathrm{Z}_1$ et $\delta = \delta_1$.

Lemme 4.1. *Sur un espace de probabilité suffisamment riche* [1]*, on peut définir une variable aléatoire* U *de loi uniforme sur* $[0,1]$*, telle que presque sûrement :*

$$\delta = 1_{\{0<\mathrm{U}<p\}} = 1 - 1_{\{p<\mathrm{U}<1\}};$$

$$\mathrm{Z} = \begin{cases} \mathrm{Q}^{(1)}(\mathrm{U}) & \textit{si } 0 < \mathrm{U} < p, \\ \mathrm{Q}^{(0)}(\mathrm{U}-p) & \textit{si } p < \mathrm{U} < 1. \end{cases}$$

Démonstration. Si les fonctions $\mathrm{H}^{(1)}$ et $\mathrm{H}^{(0)}$ sont continues, on peut poser

$$\mathrm{U} = \delta \mathrm{H}^{(1)}(\mathrm{Z}) + (1-\delta)(p + \mathrm{H}^{(0)}(\mathrm{Z})).$$

Alors la variable U suit la loi uniforme sur $[0,1]$ et vérifie la relation du théorème.

Si $\mathrm{H}^{(1)}$ et $\mathrm{H}^{(0)}$ sont quelconques, on peut remarquer que la fonction de répartition de la variable Z conditionnellement à $\{\delta = 1\}$ est $\frac{1}{p}\mathrm{H}^{(1)}$, et que conditionnellement à $\{\delta = 0\}$, la fonction de répartition est $\frac{1}{1-p}\mathrm{H}^{(0)}$. Leurs inverses généralisées sont donc $\mathrm{Q}^{(1)}(ps)$ et $\mathrm{Q}^{(0)}((1-p)s)$ respectivement.

Ainsi, le théorème 1.2 implique que pour une variable aléatoire V de loi uniforme sur $[0,1]$, la variable définie par $\mathrm{Q}^{(1)}(p\mathrm{V})$ (resp. $\mathrm{Q}^{(0)}((1-p)\mathrm{V})$) suit la même loi que Z conditionnellement à $\{\delta = 1\}$ (resp. $\{\delta = 0\}$).

L'égalité en distribution plus haut peut être renforcée en égalité presque sûre et on aura alors

$$\mathrm{Z} = \delta \mathrm{Q}^{(1)}(p\mathrm{V}_1) + (1-\delta)\mathrm{Q}^{(0)}((1-p)\mathrm{V}_2),$$

pour V_1 et V_2 de loi uniforme sur $[0,1]$. En posant

$$\mathrm{U} = \delta p \mathrm{V}_1 + (1-\delta)(p + (1-p)\mathrm{V}_2),$$

on peut facilement voir que U vérifie les relations du théorème. □

D'après ce lemme, on peut supposer que la suite originelle est définie sur un espace de probabilité tel qu'il existe une suite $\{\mathrm{U}_n, n \geqslant 1\}$ de variables aléatoires indépendantes, de même loi uniforme sur $[0,1]$ et vérifiant pour tout i :

$$\delta_i = 1_{\{0<\mathrm{U}_i<p\}} = 1 - 1_{\{p<\mathrm{U}_i<1\}};$$

1. Ceci peut s'exprimer par l'existence d'une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$ (ou de n'importe quelle autre loi continue).

$$\mathrm{Z} = \begin{cases} \mathrm{Q}^{(1)}(\mathrm{U}_i) & \text{si } 0 < \mathrm{U}_i < p, \\ \mathrm{Q}^{(0)}(\mathrm{U}_i - p) & \text{si } p < \mathrm{U}_i < 1. \end{cases}$$

Considérons maintenant la fonction de répartition empirique

$$\mathbb{U}_n(s) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\mathrm{U}_i \leqslant s\}},$$

et le processus empirique correspondant

$$\alpha_n(s) = \sqrt{n} \left(\mathbb{U}_n(s) - s \right).$$

D'après ce qui précède, on peut réécrire $\mathrm{H}_n^{(1)}$ et $\mathrm{H}_n^{(0)}$ comme suit :

$$\begin{aligned} &\mathrm{H}_n^{(1)}(x) = \mathbb{U}_n(\mathrm{H}^{(1)}(x)) && \text{si } 0 < \mathrm{H}^{(1)}(x) < p, \\ &\mathrm{H}_n^{(0)}(x) = \mathbb{U}_n(\mathrm{H}^{(0)}(x) + p) - \mathbb{U}_n(p) && \text{si } 0 < \mathrm{H}^{(0)}(x) < 1 - p. \end{aligned}$$

En effet, si $0 < \mathrm{H}^{(1)}(x) < p$ on a pour tout x :

$$\begin{aligned} \mathrm{H}_n^{(1)}(x) &= \frac{1}{n} \sum_{i=1}^{n} \delta_i 1_{\{\mathrm{Z}_i \leqslant x\}} \\ &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{0 < \mathrm{U}_i < p, \mathrm{Z}_i \leqslant x\}} \\ &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{0 < \mathrm{U}_i < p, \mathrm{Q}^{(1)}(\mathrm{U}_i) \leqslant x\}} \\ &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{0 < \mathrm{U}_i < p, \mathrm{U}_i \leqslant \mathrm{H}^{(1)}(x)\}} \\ &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{0 < \mathrm{U}_i \leqslant \mathrm{H}^{(1)}(x)\}} \\ &= \mathbb{U}_n(\mathrm{H}^{(1)}(x)). \end{aligned}$$

Et si $0 < \mathrm{H}^{(0)}(x) < 1-p$:

$$\begin{aligned}
\mathrm{H}_n^{(0)}(x) &= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)1_{\{\mathrm{Z}_i\leqslant x\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}1_{\{\mathrm{U}_i>p,\mathrm{Z}_i\leqslant x\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}1_{\{\mathrm{U}_i>p,\mathrm{Q}^{(0)}(\mathrm{U}_i-p)\leqslant x\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}1_{\{\mathrm{U}_i>p,\mathrm{U}_i-p\leqslant \mathrm{H}^{(0)}(x)\}} \\
&= \frac{1}{n}\sum_{i=1}^{n}1_{\{\mathrm{U}_i\leqslant p+\mathrm{H}^{(0)}(x)\}} - \frac{1}{n}\sum_{i=1}^{n}1_{\{\mathrm{U}_i\leqslant p\}} \\
&= \mathbb{U}_n(p+\mathrm{H}^{(0)}(x)) - \mathbb{U}_n(p).
\end{aligned}$$

Soient $s_1 \in]0,p[$, $s_0 \in]p,1[$ et $\mathrm{M}_1 > 0$ fixés, et définissons les suites de fonctions aléatoires $\left(f_n^{(1)}\right)_{n\geqslant 1}$ et $\left(f_n^{(2)}\right)_{n\geqslant 1}$ par :

$$f_n^{(j)}(u) = \frac{1}{b_n}\left(\alpha_n(s_j+h_nu)-\alpha_n(s_j)\right) \qquad \text{pour } -\mathrm{M}_1\leqslant u\leqslant \mathrm{M}_1,$$

où $b_n = \sqrt{2h_n\log_2 n}$, et les suites $\left(w_n^{(1)}\right)$ et $\left(w_n^{(0)}\right)$ par :

$$\begin{aligned}
w_n^{(1)} &= \frac{1}{\sqrt{2\log_2 n}}\alpha_n(s_1), \\
w_n^{(0)} &= \frac{1}{\sqrt{2\log_2 n}}\left(\alpha_n(s_0)-\alpha_n(p)\right).
\end{aligned}$$

Le lemme suivant, qui est un cas particulier du Théorème 1.2 de Deheuvels et Mason (1994), est le point de départ de ce travail.

Lemme 4.2. *Sous (H1) et (H2), la suite* $\left\{\left(w_n^{(1)}, w_n^{(0)}, f_n^{(1)}, f_n^{(0)}\right), n\geqslant 1\right\}$ *est presque sûrement relativement compacte dans* $\mathbb{R}^2\times\mathcal{B}([-\mathrm{M},\mathrm{M}])$ *avec pour ensemble limite l'ensembles de tous les quadruplets* $\left(w^{(1)}, w^{(0)}, f^{(1)}, f^{(0)}\right)$ *vérifiant :*

$$w^{(1)} = \int_0^{s_1}\Psi(s)\,ds \quad ; \quad w^{(0)} = \int_p^{s_0}\Psi(s)\,ds$$

$$f^{(j)}(u) = \int_0^{u}\Psi_j(s)\,ds \quad ; \quad j=0,1 \quad ; \quad -\mathrm{M}_1\leqslant u\leqslant \mathrm{M}_1;$$

avec

$$\int_0^1\Psi(s)\,ds = 0 \quad ; \quad \int_0^1\Psi^2(s)\,ds + \int_{-\mathrm{M}_1}^{\mathrm{M}_1}\left(\Psi_1^2(s)+\Psi_0^2(s)\right)\,ds \leqslant 1.$$

Soit $\mathrm{M} > 0$ fixé, $\left(\gamma_n^{(1)}\right)$ et $\left(\gamma_n^{(0)}\right)$ deux suites de fonctions définies sur $[-\mathrm{M}, \mathrm{M}]$ et telles que :

$$\lim_{n\to\infty} \sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\gamma_n^{(j)}(u) - \gamma_j u\right| = 0 \quad ; \quad j = 0, 1;$$

où γ_1 et γ_0 sont deux constantes positives. Posons pour $j = 0, 1$ et $u \in [-\mathrm{M}, \mathrm{M}]$

$$g_n^{(j)} = f_n^{(j)}\left(\gamma_n^{(j)}(u)\right) = \frac{1}{b_n}\left(\alpha_n\left(s_j + h_n\gamma_n^{(j)}(u)\right) - \alpha_n(s_j)\right).$$

Nous allons faire usage d'une version modifiée du lemme précédent :

Lemme 4.3. *Sous (H1) et (H2), la suite* $(w_n^{(1)}, w_n^{(0)}, g_n^{(1)}, g_n^{(0)})$ *est presque sûrement relativement compacte dans* $\mathbb{R}^2 \times \mathbb{B}^2([-\mathrm{M}, \mathrm{M}])$ *avec pour ensemble limite l'ensemble de tous les quadruplets* $(w^{(1)}, w^{(0)}, g^{(1)}, g^{(0)})$ *tels que :*

$$w^{(1)} = \int_0^{s_1} \phi(s)\, ds \quad ; \quad w^{(0)} = \int_p^{s_0} \phi(s)\, ds;$$

$$g^{(j)}(u) = \int_0^u \phi_j(s)\, ds \quad ; \quad j = 0, 1 \quad ; \quad -\mathrm{M} \leqslant u \leqslant \mathrm{M};$$

avec

$$\int_0^1 \phi(s)\, ds = 0 \quad ; \quad \int_0^1 \phi^2(s)\, ds + \int_{-\mathrm{M}}^{\mathrm{M}} \frac{1}{\gamma_1}\phi_1^2(s) + \frac{1}{\gamma_0}\phi_1^2(s)\, ds \leqslant 1.$$

Démonstration. Soit $\mathrm{M}_1 > 0$ vérifiant $\left|\gamma_n^{(j)}\right| \leqslant \mathrm{M}_1,\ \forall u, j$ pour n suffisamment grand. D'après le Lemme 4.2, de toute sous suite $(w_{n_k}^{(1)}, w_{n_k}^{(0)}, f_{n_k}^{(1)}, f_{n_k}^{(0)})$, on peut extraire une sous suite convergente, c'est-à-dire qu'il existe une sous suite (n'_k) de (n_k) et $(w^{(1)}, w^{(0)}, f^{(1)}, f^{(0)})$ vérifiant :

$$\sum_{j=0,1}\left(\left|w_{n'_k}^{(j)} - w^{(j)}\right| + \sup_{v\in[-\mathrm{M}_1,\mathrm{M}_1]}\left|f_{n'_k}^{(j)}(v) - f^{(j)}(v)\right|\right) \longrightarrow 0,$$

d'où :

$$\sum_{j=0,1}\left(\left|w_{n'_k}^{(j)} - w^{(j)}\right| + \sup_{v\in[-\mathrm{M}_1,\mathrm{M}_1]}\left|f_{n'_k}^{(j)}(\gamma_{n'_k}(u)) - f^{(j)}(\gamma_{n'_k}(u))\right|\right) \longrightarrow 0.$$

De plus, les fonctions $f^{(j)}, j = 0, 1$ sont uniformément équicontinues. En effet, d'après l'inégalité de Schwarz :

$$\begin{aligned}
\left|f^{(j)}(v') - f^{(j)}(v'')\right| &= \left|\int_{v'}^{v''} \psi_j(s)ds\right| \\
&\leqslant |v' - v''|^{1/2}\left|\int_{v'}^{v''} \psi_j^2(s)ds\right|^{1/2} \\
&\leqslant |v' - v''|^{\frac{1}{2}}, \qquad \text{car } \left|\int_{v'}^{v''} \psi_j^2(s)ds\right| \leqslant 1;
\end{aligned}$$

il en résulte que : $\forall v', v'' : -\mathrm{M} \leqslant v', v'' \leqslant \mathrm{M}$

$$\sum_{j=0,1} \left(\left| w_{n'_k}^{(j)} - w^{(j)} \right| + \sup_{u \in [-\mathrm{M},\mathrm{M}]} \left| g_{n'_k}^{(j)}(u) - f^{(j)}(\gamma_j u) \right| \right) \underset{k\to\infty}{\longrightarrow} 0,$$

car :

$$\begin{aligned} \left| g_n^{(j)}(u) - f^{(j)}(\gamma_j u) \right| &= \left| f_n^{(j)}(\gamma_n^{(j)}(u)) - f^{(j)}(\gamma_n^{(j)}(u)) + f^{(j)}(\gamma_n^{(j)}(u)) - f^{(j)}(\gamma_j u) \right| \\ &\leqslant \left| f_n^{(j)}(\gamma_n^{(j)}(u)) - f^{(j)}(\gamma_n^{(j)}(u)) \right| + \left| f^{(j)}(\gamma_n^{(j)}(u)) - f^{(j)}(\gamma_j u) \right| \\ &\leqslant \left| f_n^{(j)}(\gamma_n^{(j)}(u)) - f^{(j)}(\gamma_n^{(j)}(u)) \right| + \left| \gamma_n^{(j)}(u) - \gamma_j(u) \right|^{\frac{1}{2}}. \end{aligned}$$

En posant :

$$\begin{gathered} \phi(u) = \psi(u) \\ \begin{array}{ll} g^{(j)} = f^{(j)}(\gamma_j u) & \\ \phi_j(u) = \gamma_j \psi_j(\gamma_j u) & j = 0, 1; -\mathrm{M} < u < \mathrm{M} \end{array} \end{gathered}$$

on obtient :

$$\begin{aligned} \int_0^u \phi_j(s) ds &= \int_0^u \gamma_j \psi_j(\gamma_j s) ds = \int_0^{\gamma_j u} \psi_j(s) ds \\ &= f^{(j)}(\gamma_j u) = g^{(j)}(u); \end{aligned}$$

$$\int_{-\mathrm{M}}^{\mathrm{M}} \left(\frac{\phi_1^2(s)}{\gamma_1} + \frac{\phi_0^2(s)}{\gamma_0} \right) ds = \int_{-\mathrm{M}}^{\mathrm{M}} \left(\gamma_1 \psi_1^2(s) + \gamma_0 \psi_0^2(s) \right) ds;$$

$$\begin{aligned} \int_{-\mathrm{M}}^{\mathrm{M}} \left(\frac{\phi_1^2(s)}{\gamma_1} + \frac{\phi_0^2(s)}{\gamma_0} \right) ds &= \int_{-\gamma_1 \mathrm{M}}^{\gamma_1 \mathrm{M}} \psi_1^2(s) ds + \int_{-\gamma_0 \mathrm{M}}^{\gamma_0 \mathrm{M}} \psi_0^2(s) ds \\ &\leqslant \int_{-\mathrm{M}_1}^{\mathrm{M}_1} \left(\psi_1^2(s) + \psi_0^2(s) \right) ds; \end{aligned}$$

$|\gamma_j \mathrm{M}| \leqslant \mathrm{M}_1$ car $\lim_{n\to\infty} \gamma_n^{(j)}(\mathrm{M}) = \gamma_j \mathrm{M}$ et $\left| \gamma_n^{(j)}(\mathrm{M}) \right| \leqslant \mathrm{M}_1, \forall j = 0, 1.$

Ainsi, la suite $\left\{ \left(w_n^{(1)}, w_n^{(0)}, g_n^{(1)}, g_n^{(0)} \right), n \geqslant 1 \right\}$ est presque sûrement relativement compacte et l'ensemble de ses valeurs d'adhérence est inclus dans l'ensemble défini dans le lemme. L'autre inclusion se démontre de la même manière. □

Considérons $z_1, z_0 \in]0, \Theta[$ fixés et posons :

$$s_1 = \mathrm{H}^{(1)}(z_1) \in]0, p[\quad \text{et} \quad s_0 = p + \mathrm{H}^{(0)}(z_0) \in]p, 1[.$$

Pour $\mathrm{M} > 0$ fixé, on considère les suites $\left(k_n^{(1)} \right)$ et $\left(k_n^{(0)} \right)$ de fonctions aléatoires définies sur $[-\mathrm{M}, \mathrm{M}]$ par :

$$k_n^{(j)}(u) = \frac{\sqrt{n}}{b_n} \left(\mathrm{H}_n^{(j)}(z_j + h_n u) - \mathrm{H}^{(j)}(z_j + h_n u) - \mathrm{H}_n^{(j)}(z_j) + \mathrm{H}^{(j)}(z_j) \right),$$

et les suites de variables aléatoires $\left(v_n^{(0)}\right)$, $\left(v_n^{(1)}\right)$ définies par :

$$v_n^{(j)} = \frac{\sqrt{n}}{\sqrt{2\log_2 n}} \left(\mathrm{H}_n^{(j)}(z_j) - \mathrm{H}^{(j)}(z_j)\right) \text{ pour } j = 0, 1.$$

D'après (H1), il existe $n_0 \geqslant 1$ tel que pour tout $n \geqslant n_0$ et tout $j = 0, 1$, $-\mathrm{M} \leqslant u \leqslant \mathrm{M}$, on ait $z_j + h_n u \in]0, \Theta[$. Ceci assure que $\left(v_n^{(1)}, v_n^{(0)}, k_n^{(1)}, k_n^{(0)}\right) \in \mathbb{R}^2 \times \mathbb{B}^2([-\mathrm{M}, \mathrm{M}])$ est bien définie. Le théorème suivant décrit le comportement asymptotique de cette suite.

Théorème 4.3 (Deheuvels et Einmahl, 1996, Théorème 2.1)**.** *On suppose que* F *est continue et* G *dérivable en* z_0 *avec* $g(z_0) = \mathrm{G}'(z_0) > 0$. *On suppose aussi que* G *est continue et* F *dérivable en* z_1 *avec* $f(z_1) = \mathrm{F}'(z_1) > 0$.

Alors sous les hypothèses (H1) et (H2), la suite $\left\{\left(v_n^{(1)}, v_n^{(0)}, k_n^{(1)}, k_n^{(0)}\right) : n \geqslant 1\right\}$ *est presque sûrement relativement compacte dans* $\mathbb{R}^2 \times \mathbb{B}^2([-\mathrm{M}, \mathrm{M}])$ *avec pour ensemble limite l'ensemble des quadruplets* $\left(v^{(1)}, v^{(0)}, k^{(1)}, k^{(0)}\right)$ *vérifiant :*

$$v^{(1)} = \int_0^{\mathrm{H}^{(1)}(z_1)} \phi(s)\, ds \quad ; \quad v^{(0)} = \int_p^{p+\mathrm{H}^{(0)}(z_0)} \phi(s)\, ds;$$

$$k^{(j)}(u) = \int_0^u \phi_j(s) ds;$$

pour $j = 0, 1$ *et* $-\mathrm{M} \leqslant u \leqslant \mathrm{M}$, *avec :*

$$\int_0^1 \phi(s) ds = 0;$$

$$\int_0^1 \phi^2(s) ds + \int_{-\mathrm{M}}^{\mathrm{M}} \left(\frac{\phi_1^2(s)}{f(z_1)(1-\mathrm{G}(z_1))} + \frac{\phi_0^2(s)}{g(z_0)(1-\mathrm{F}(z_0))}\right) ds \leqslant 1.$$

Démonstration. On a :

$$\begin{aligned}
k_n^{(1)}(u) &= \frac{1}{b_n}\left(\sqrt{n}\left(\mathbb{U}_n\left(\mathrm{H}^{(1)}(z_1 + h_n u)\right) - \mathrm{H}^{(1)}(z_1 + h_n u)\right) - \sqrt{n}\left(\mathbb{U}_n\left(\mathrm{H}^{(1)}(z_1)\right) - \mathrm{H}^{(1)}(z_1)\right)\right) \\
&= \frac{1}{b_n}\left(\alpha_n\left(\mathrm{H}^{(1)}(z_1 + h_n u)\right) - \alpha_n\left(\mathrm{H}^{(1)}(z_1)\right)\right) \\
&= \frac{1}{b_n}\left(\alpha_n\left(\mathrm{H}^{(1)}(z_1) + h_n \frac{1}{h_n}\left(\mathrm{H}^{(1)}(z_1 + h_n u) - \mathrm{H}^{(1)}(z_1)\right)\right) - \alpha_n\left(\mathrm{H}^{(1)}(z_1)\right)\right) \\
&= f_n^{(1)}\left(\frac{1}{h_n}\left(\mathrm{H}^{(1)}(z_1 + h_n u) - \mathrm{H}^{(1)}(z_1)\right)\right);
\end{aligned}$$

et de la même manière : $k_n^{(0)}(u) = f_n^{(0)}\left(\frac{1}{h_n}\left(\mathrm{H}^{(0)}(z_0 + h_n u) - \mathrm{H}^{(0)}(z_0)\right)\right)$. On a aussi pour $v_n^{(1)}$ et $v_n^{(0)}$:

$$\begin{aligned} v_n^{(1)} &= \frac{1}{\sqrt{2\log_2 n}}\sqrt{n}\left(\mathbb{U}_n\left(\mathrm{H}^{(1)}(z_1)\right) - \mathrm{H}^{(1)}(z_1)\right) \\ &= \frac{1}{\sqrt{2\log_2 n}}\alpha_n\left(\mathrm{H}^{(1)}(z_1)\right) = w_n^{(1)}; \end{aligned}$$

et $v_n^{(0)} = w_n^{(0)}$. De plus, si on considère les fonctions :

$$\gamma_n^{(j)}(u) = \frac{1}{h_n}\left(\mathrm{H}^{(j)}(z_j + h_n u) - \mathrm{H}^{(j)}(z_j)\right),$$

pour $j = 0, 1$ et $u \in [-\mathrm{M}, \mathrm{M}]$, alors en vertu de (4.1) :

$$\begin{aligned} &\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{H}^{(1)}(z_1 + hu) - \mathrm{H}^{(1)}(z_1) - f(z_1)(1 - \mathrm{G}(z_1))hu\right| \\ &\quad= \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_{z_1}^{z_1+hu} (1 - \mathrm{G}_-(t))\, d\mathrm{F}(t) - f(z_1)(1 - \mathrm{G}(z_1))hu\right| \\ &\quad\leqslant \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_{z_1}^{z_1+hu} (1 - \mathrm{G}_-(t))\, d\mathrm{F}(t) - \int_{z_1}^{z_1+hu} (1 - \mathrm{G}(z_1))\, d\mathrm{F}(t)\right| \\ &\qquad+ \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|(1 - \mathrm{G}(z_1))(\mathrm{F}(z_1 + hu) - \mathrm{F}(z_1)) - (1 - \mathrm{G}(z_1))f(z_1)hu\right| \\ &\quad\leqslant \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_{z_1}^{z_1+hu} (\mathrm{G}(z_1) - \mathrm{G}_-(t))\, d\mathrm{F}(t)\right| \\ &\qquad+ (1 - \mathrm{G}(z_1)) \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} |\mathrm{F}(z_1 + hu) - \mathrm{F}(z_1) - f(z_1)hu| \\ &\quad\leqslant \sup_{z_1 - h\mathrm{M}\leqslant t\leqslant z_1 + h\mathrm{M}} |\mathrm{G}(z_1) - \mathrm{G}_-(t)| \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} |\mathrm{F}(z_1 + hu) - \mathrm{F}(z_1)| \\ &\qquad+ (1 - \mathrm{G}(z_1)) \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{F}(z_1 + hu) - \mathrm{F}(z_1) - f(z_1)hu\right|. \end{aligned}$$

Or, puisque $\sup_{z_1 - h\mathrm{M}\leqslant t\leqslant z_1 + h\mathrm{M}} |\mathrm{G}(z_1) - \mathrm{G}_-(t)|$ tend vers 0 quand h tend vers 0, $\frac{1}{h}\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} |\mathrm{F}(z_1 + hu) - \mathrm{F}(z_1)|$ est bornée, et $\mathrm{F}(z_1 + hu) = \mathrm{F}(z_1) + f(z_1)hu + o(h)$, on obtient :

$$\frac{1}{h}\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{H}^{(1)}(z_1 + hu) - \mathrm{H}^{(1)}(z_1) - f(z_1)(1 - \mathrm{G}(z_1))hu\right| \longrightarrow 0,$$

quand $h \to 0$, ce qui montre que

$$\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\gamma_n^{(1)}(u) - \gamma_1 u\right| \longrightarrow 0,$$

avec $\gamma_1 = f(z_1)(1 - \mathrm{G}(z_1))$. De la même manière, on peut montrer que

$$\sup_{-\mathrm{M} \leqslant u \leqslant \mathrm{M}} \left| \gamma_n^{(0)}(u) - \gamma_0 u \right| \longrightarrow 0,$$

avec $\gamma_0 = g(z_0)(1 - \mathrm{F}(z_0))$, d'où le résultat en appliquant le Lemme 4.3.

□

Remarque 4.3. Comme il apparaît dans la preuve ci-dessus, la continuité de F en z_0 et l'existence et la positivité de la dérivée de G en z_0 ne sont nécessaires que pour la compacité relative de la suite $\{k_n^{(0)}\}$. Si on veut uniquement la compacité relative de la suite $\{k_n^{(1)}\}$ dans $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$ (ce qui est le cas dans la suite de ce travail) avec pour ensemble limite l'ensemble des fonctions $k^{(1)} \in \mathbb{B}([-\mathrm{M}, \mathrm{M}])$ de la forme

$$k^{(1)}(u) = \int_0^u \phi_1(s)\, ds \text{ avec } \int_{-\mathrm{M}}^{\mathrm{M}} \phi_1^2(s)\, ds \leqslant f(z_1)(1 - \mathrm{G}(z_1)),$$

alors il suffit que G soit continue en z_1, et que F soit dérivable en z_1.

La deuxième partie de ce chapitre consiste à trouver une relation entre ξ_n et $k_n^{(1)}$, ce qui nous permettra de généraliser les résultats du Théorème 4.3 et de la remarque qui le suit au processus ξ_n, objet de notre étude.

Comme nous n'avons plus besoin de $k_n^{(0)}$ pour la suite, nous notons $z = z_1$. Nous supposons aussi, pour simplifier, que F est continue dans un voisinage de z, et dérivable en z, et que G est continue en z.

Rappelons que la martingale de base est définie par :

$$\mathrm{M}_n(x) = n \left(\mathrm{H}_n^{(1)}(x) - \int_0^x (1 - \mathrm{H}_n(s_-))\, d\Lambda(s) \right), \tag{4.2}$$

où Λ est la mesure de hasard de X (voir par exemple Gu et Lai, 1990).

$\mathrm{H}^{(1)}$ peut être réécrite comme suit :

$$\begin{aligned} \mathrm{H}^{(1)}(x) &= \int_0^x (1 - \mathrm{G}_-(s))(1 - \mathrm{F}(s)) \frac{d\mathrm{F}(s)}{1 - \mathrm{F}(s)} \\ &= \int_0^x (1 - \mathrm{H}_-(s)) d\Lambda(s); \end{aligned}$$

d'où

$$\frac{\mathrm{M}_n(x)}{\sqrt{n}} = \sqrt{n} \left(\mathrm{H}_n^{(1)}(x) - \mathrm{H}^{(1)}(x) \right) + \sqrt{n} \int_0^x (\mathrm{H}_{n-}(s) - \mathrm{H}_-(s))\, d\Lambda(s). \tag{4.3}$$

Pour $z \in]0, \Theta[$ fixé, considérons l'accroissement :

$$\mu_n(u) = \frac{1}{b_n \sqrt{n}} \left(\mathrm{M}_n(z + h_n u) - \mathrm{M}_n(z) \right).$$

Lemme 4.4. *On suppose que* F *est continue dans un voisinage de* z*, et dérivable en* z*, et que* G *est continue en* z*. Alors, sous les hypothèse (H1) et (H2), on a :*

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} |\mu_n(u) - k_n^{(1)}(u)| = \mathrm{O}(\sqrt{h_n}) \to 0 \; p.s$$

Démonstration. On a d'après (4.3)

$$\begin{aligned}
\mu_n(u) &= \frac{1}{b_n\sqrt{n}} \left(\mathrm{M}_n(z+h_nu) - \mathrm{M}_n(z)\right) \\
&= \frac{1}{b_n}\left(\sqrt{n}\left(\mathrm{H}_n^{(1)}(z+h_nu) - \mathrm{H}^{(1)}(z+h_nu)\right) + \sqrt{n}\int_0^z \left(\mathrm{H}_{n-}^{(1)}(s) - \mathrm{H}_-^{(1)}(s)\right) d\Lambda(s)\right) \\
&\quad - \frac{1}{b_n}\left(\sqrt{n}\left(\mathrm{H}_n^{(1)}(z) - \mathrm{H}^{(1)}(z)\right) + \sqrt{n}\int_0^z \left(\mathrm{H}_{n-}^{(1)}(s) - \mathrm{H}_-^{(1)}(s)\right) d\Lambda(s)\right) \\
&= \frac{\sqrt{n}}{b_n}\left(\mathrm{H}_n^{(1)}(z+h_nu) - \mathrm{H}_n^{(1)}(z) - \mathrm{H}^{(1)}(z+h_nu) + \mathrm{H}^{(1)}(z)\right) \\
&\quad + \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu} \left(\mathrm{H}_{n-}^{(1)}(s) - \mathrm{H}_-^{(1)}(s)\right) d\Lambda(s) \\
&= k_n^{(1)}(u) + \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu} \left(\mathrm{H}_{n-}^{(1)}(s) - \mathrm{H}_-^{(1)}(s)\right) d\Lambda(s).
\end{aligned}$$

En appliquant la loi du logarithme itéré de Chung-Smirnov au processus empirique continu à gauche [2] $\sqrt{n}\left(\mathrm{H}_{n-}(t) - \mathrm{H}_-(t)\right)$ on obtient

$$\lim_{n\to\infty} \sup_{x\in[-\mathrm{M},\mathrm{M}]} \frac{|\sqrt{n}(\mathrm{H}_{n-}(t) - \mathrm{H}_-(t))|}{\sqrt{2\log_2 n}} = \frac{1}{2}.$$

Donc, pour n assez grand,

$$\sup_{x\in[-\mathrm{M},\mathrm{M}]} |\sqrt{n}\left(\mathrm{H}_{n-}(t) - \mathrm{H}_-(t)\right)| \leqslant \sqrt{\log_2 n},$$

d'où

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} |\mu(u) - k_n^{(1)}(u)| \leqslant \frac{1}{b_n}\sqrt{\log_2(n)}\left(\Lambda(z+h_n\mathrm{M}) - \Lambda(z-h_n\mathrm{M})\right),$$

ce qui donne que

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\mu_n(u) - k_n^{(1)}(u)\right| = \mathrm{O}\left(\sqrt{h_n}\right).$$

□

2. Ceci est équivalent à la définition de la fonction de répartition par $\mathrm{P}(\mathrm{X} < t)$ et ne change rien aux théorèmes que nous avons vus précédemment.

Lemme 4.5. *Sous les hypothèses du Lemme 4.4, la suite (μ_n) est presque sûrement relativement compacte dans $\mathbb{B}([-\mathrm{M},\mathrm{M}])$, avec pour ensemble limite l'ensemble de tous les $\mu \in \mathbb{B}([-\mathrm{M},\mathrm{M}])$ satisfaisant :*

$$\mu(u) = \int_0^u \Phi(s)ds \qquad -\mathrm{M} \leqslant u \leqslant \mathrm{M},$$

avec

$$\int_{-\mathrm{M}}^{\mathrm{M}} \Phi^2(s)ds \leqslant f(z)(1-\mathrm{G}(z)).$$

Démonstration. D'après la remarque 4.3, les hypothèses du lemme permettent de conclure que $\{k_n^{(1)}\}$ est presque sûrement relativement compacte dans $\mathcal{B}([-\mathrm{M},\mathrm{M}])$ avec pour ensemble limite l'ensemble des éléments $\mu \in \mathcal{B}([-\mathrm{M},\mathrm{M}])$ vérifiant :

$$\mu(u) = \int_0^u \phi(s)\,ds \text{ pour } u \in [-\mathrm{M},\mathrm{M}],$$

avec

$$\int_{-\mathrm{M}}^{\mathrm{M}} \phi^2(s)\,ds \leqslant f(z_1)(1-\mathrm{G}(z_1)).$$

D'après le lemme précédent, il en est de même pour $\{\mu_n\}_{n\geqslant 1}$. □

Nous allons maintenant utiliser la représentation intégrale suivante du processus empirique en utilisant la martingale de base (4.2) (voir aussi la section 3.4). On a, pour tout $x \leqslant \mathrm{Z}_{(n)} = \max\{\mathrm{Z}_1, \mathrm{Z}_2, \ldots \mathrm{Z}_n\}$

$$\begin{aligned}\Pi_n(x) &= \frac{\sqrt{n}(\mathrm{F}_n(x)-\mathrm{F}(x))}{1-\mathrm{F}(x)} \\ &= \frac{1}{\sqrt{n}}\int_0^x \frac{d\mathrm{M}_n(t)}{1-\mathrm{H}_-(t)} \\ &+ \frac{1}{\sqrt{n}}\int_0^x \left(\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)}\frac{1}{1-\mathrm{H}_{n-}(t)} - \frac{1}{1-\mathrm{H}_-(t)}\right) d\mathrm{M}_n(x),\end{aligned}$$

d'après l'équation de Duhamel (voir Gill et Johansen, 1990). On peut alors décomposer $\Pi_n(x)$ en somme $\Pi_{n,1}(x) + \Pi_{n,2}(x)$ avec

$$\Pi_{n,1} = \frac{1}{\sqrt{n}}\int_0^x \frac{d\mathrm{M}_n(t)}{1-\mathrm{H}_-(t)},$$

et

$$\Pi_{n,2} = \frac{1}{\sqrt{n}}\int_0^x \left(\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)}\frac{1}{1-\mathrm{H}_{n-}(t)} - \frac{1}{1-\mathrm{H}_-(t)}\right) d\mathrm{M}_n(x).$$

Considérons maintenant l'accroissement :

$$\pi_n(u) = \frac{1}{b_n}\left(\Pi_n(z+h_nu) - \Pi_n(z)\right).$$

Lemme 4.6. *Sous les hypothèses du Lemme 4.4, on a :*

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\pi_n(u)-\frac{\mu_n(u)}{1-\mathrm{H}(z)}\right|\to 0 \quad p.s$$

Démonstration. Comme $\mathrm{Z}_n=\max\{\mathrm{Z}_1,\ldots,\mathrm{Z}_n\}\to\Theta$ p.s. alors pour tout $\theta<\Theta$, il existe presque sûrement un rang n_θ à partir duquel on a la représentation intégrale ci-dessus pour tout $x\leqslant\theta$ et tout $n\geqslant n_\theta$. On peut donc supposer sans perte de généralité que $n\geqslant n_\theta$ avec $\theta>z+h_n\mathrm{M}$. On peut alors poser

$$\begin{aligned}\pi_{n,1}(u)&=\frac{1}{b_n}(\Pi_{n,1}(z+h_nu)-\Pi_{n,1}(z))\\&=\frac{1}{b_n\sqrt{n}}\int_z^{z+h_nu}\frac{d\mathrm{M}_n(t)}{1-\mathrm{H}_-(t)};\end{aligned}$$

et

$$\begin{aligned}\pi_{n,2}(u)&=\frac{1}{b_n}(\Pi_{n,2}(z+h_nu)-\Pi_{n,2}(z))\\&=\frac{1}{b_n\sqrt{n}}\int_z^{z+h_nu}\left(\left(\frac{1-\mathrm{F}_n(t_-)}{1-\mathrm{F}(t_-)}\right)\frac{1}{1-\mathrm{H}_{n-}(t)}-\frac{1}{1-\mathrm{H}_-(t)}\right)d\mathrm{M}_n(t).\end{aligned}$$

En intégrant par parties dans la définition de $\pi_{n,1}$, on obtient (en utilisant le fait que $d\mathrm{M}_n(t)=d(\mathrm{M}_n(t)-\mathrm{M}_n(z))$)

$$\begin{aligned}\pi_{n,1}(u)=&\frac{1}{b_n\sqrt{n}}\frac{\mathrm{M}_n(z+h_nu)-\mathrm{M}_n(z)}{1-\mathrm{H}(z+h_nu)}\\&-\frac{1}{b_n\sqrt{n}}\int_z^{z+h_nu}(\mathrm{M}_n(t)-\mathrm{M}_n(z))\,d\left(\frac{1}{1-\mathrm{H}(t)}\right)\\=&\frac{1}{b_n\sqrt{n}}\frac{\mathrm{M}_n(z+h_nu)-\mathrm{M}_n(z)}{1-\mathrm{H}(z+h_nu)}\\&-\frac{1}{b_n\sqrt{n}}\int_0^{u}(\mathrm{M}_n(z+h_nv)-\mathrm{M}_n(z))\,d\left(\frac{1}{1-\mathrm{H}(z+h_nv)}\right)\\=&\frac{\mu_n(u)}{1-\mathrm{H}(z+h_nu)}-\int_0^u\mu_n(v)\,d\left(\frac{1}{1-\mathrm{H}(z+h_nv)}\right);\end{aligned}$$

d'où

$$\begin{aligned}\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\pi_{n,1}(u)-\frac{\mu_n(u)}{1-\mathrm{H}(z)}\right|\leqslant&\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\mu_n(u)\left(\frac{1}{1-\mathrm{H}(z+h_nu)}-\frac{1}{1-\mathrm{H}(z)}\right)\right|\\&+\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\int_0^u\mu_n(u)\,d\left(\frac{1}{1-\mathrm{H}(z+h_nv)}\right)\right|.\end{aligned}$$

Or d'après le lemme précédent, la suite $\{\mu_n\}$ est relativement compacte dans $\mathcal{B}([-\mathrm{M},\mathrm{M}])$, donc elle est bornée

$$\exists \mathrm{C}>0, \forall n \geqslant 1 : \sup_{u\in[-\mathrm{M},\mathrm{M}]} |\mu_n(u)| \leqslant \mathrm{C};$$

ce qui donne

$$\begin{aligned}\sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\pi_{n,1}(u) - \frac{\mu_n(u)}{1-\mathrm{H}(z)}\right| \leqslant & \mathrm{C} \sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\frac{1}{1-\mathrm{H}(z+h_n u)} - \frac{1}{1-\mathrm{H}(z)}\right| \\ & + \mathrm{C} \sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\frac{1}{1-\mathrm{H}(z+h_n u)} - \frac{1}{1-\mathrm{H}(z)}\right| \\ \leqslant & 2\mathrm{C}\left(\frac{1}{1-\mathrm{H}(z+h_n\mathrm{M})} - \frac{1}{1-\mathrm{H}(z-h_n\mathrm{M})}\right) \\ \underset{n\to\infty}{\longrightarrow} & 0 \quad \text{p.s.}\end{aligned}$$

car H est continue en z.

Il reste à montrer que

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} |\pi_{n,2}(u)| \underset{n\to\infty}{\longrightarrow} 0 \quad \text{p.s.}$$

D'autre part, d'après la définition de M_u :

$$\begin{aligned}\pi_{n,2}(u) = & \frac{\sqrt{n}}{b_n}\int_z^{z+h_n u}\left(\left(\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)}\right)\frac{1}{1-\mathrm{H}_n(t)} - \frac{1}{1-\mathrm{H}(t)}\right) d\mathrm{H}_n^{(1)}(t) \\ & + \frac{\sqrt{n}}{b_n}\int_z^{z+h_n u}\left(\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)} - \frac{1-\mathrm{H}_{n-}(t)}{1-\mathrm{H}_{n-}(t')}\right) d\Lambda(t).\end{aligned}$$

Comme $h_n \to 0$, et en utilisant la continuité locale de F, on a pour n suffisamment grand $\mathrm{F}_-(t) = \mathrm{F}(t)$ pour tout $t \in [z-h_n\mathrm{M}, z+h_n\mathrm{M}]$. D'où, en utilisant la loi du logarithme itéré de Földes et Rejtő (1981)

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]} \left|\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)} - 1\right| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}}\right) \quad \text{p.s.}$$

De la même manière, la loi du logarithme itéré de Chung-Smirnov donne

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]} \left|\frac{1-\mathrm{H}_{n-}(t)}{1-\mathrm{H}_-(t)} - 1\right| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}}\right) \quad \text{p.s.}$$

On obtient alors pour un certain $\mathrm{A}_n = \mathrm{O}\left(\sqrt{\log_2 n}\right)$ que :

$$\begin{aligned}\sup_{u\in[-\mathrm{M},\mathrm{M}]} |\pi_{n,2}(u)| = & \frac{\mathrm{A}_n}{b_n}\left(\mathrm{H}_n^{(1)}(z+h_n\mathrm{M}) - \mathrm{H}_n^{(1)}(z-h_n\mathrm{M})\right) \\ & + \frac{\mathrm{A}_n}{b_n}\left(\Lambda(z+h_n\mathrm{M}) - \Lambda(z-h_n\mathrm{M})\right).\end{aligned}$$

De plus, sachant que

$$k_n^{(1)}(u) = \frac{\sqrt{n}}{b_n}\left(\mathrm{H}_n^{(1)}(z+h_nu) - \mathrm{H}^{(1)}(z+h_nu) - \mathrm{H}_n^{(1)}(z) + \mathrm{H}^{(1)}(z)\right),$$

on trouve

$$\begin{aligned}\frac{1}{b_n}(\mathrm{H}_n^{(1)}(z+h_n\mathrm{M}) - &\mathrm{H}_n^{(1)}(z-h_n\mathrm{M})) \\ &= \frac{1}{\sqrt{n}}(k_n^{(1)}(\mathrm{M}) - k_n^{(1)}(-\mathrm{M})) \\ &\quad + \frac{1}{b_n}\left(\mathrm{H}^{(1)}(z+h_n\mathrm{M}) - \mathrm{H}^{(1)}(z-h_n\mathrm{M})\right) \\ &= \mathrm{O}\left(\frac{1}{\sqrt{n}}\right) + \mathrm{O}\left(\sqrt{\frac{h_n}{\log_2 n}}\right),\end{aligned}$$

car la suite $\left\{k_n^{(1)}(\mathrm{M}) - k_n^{(1)}(-\mathrm{M})\right\}$ est bornée d'après le théorème 4.3. Sachant (d'après l'hypothèse (H2)) que $\frac{nh_n}{\log_2 n} \longrightarrow \infty$, on a $\frac{1}{\sqrt{n}} = \mathrm{O}\left(\sqrt{\frac{h_n}{\log_2 n}}\right)$ d'où

$$\frac{1}{b_n}\left(\mathrm{H}^{(1)}(z+h_n\mathrm{M}) - \mathrm{H}^{(1)}(z-h_n\mathrm{M})\right) = \mathrm{O}\left(\sqrt{\frac{h_n}{\log_2 n}}\right).$$

D'autre part, sachant que $\Lambda(s) = -\log(1-\mathrm{F}(s)$ et que $b_n = \sqrt{2h_n \log_2 n}$, la dérivabilité de F assure que

$$\frac{1}{b_n}\left(\Lambda(z+h_n\mathrm{M}) - \Lambda(z-h_n\mathrm{M})\right) = \mathrm{O}\left(\sqrt{\frac{h_n}{\log_2 n}}\right).$$

En substituant ces deux derniers résultats, on obtient

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} |\pi_{n,2}(u)| = \mathrm{O}\left(\sqrt{h_n}\right) \longrightarrow 0 \quad p.s$$

ce qui termine la preuve du lemme. □

Lemme 4.7. *Sous les hypothèses du Lemme 4.4, on a :*

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]} \left|\xi_n(u) - \frac{k_n^{(1)}(u)}{1-\mathrm{G}(z)}\right| \to 0 \quad p.s$$

Démonstration. On a :

$$\begin{aligned}
\xi_n(u) &= \frac{1}{b_n}\left(a_n(z+h_n) - a_n(z)\right) \\
&= \frac{\sqrt{n}}{b_n}\left(\mathrm{F}_n(z+h_nu) - \mathrm{F}(z+h_nu) - \mathrm{F}_n(z) + \mathrm{F}(z)\right) \\
&= \frac{1}{b_n}(1-\mathrm{F}(z+h_nu))\frac{\sqrt{n}\left(\mathrm{F}_n(z+h_nu) - \mathrm{F}(z+h_nu)\right)}{1-\mathrm{F}(z+h_nu)} \\
&\quad - \frac{1}{b_n}(1-\mathrm{F}(z))\frac{\sqrt{n}\left(\mathrm{F}_n(z) - \mathrm{F}(z)\right)}{1-\mathrm{F}(z)} \\
&= \frac{1}{b_n}\left((1-\mathrm{F}(z+h_nu))\,\Pi_n(z+h_nu) - (1-\mathrm{F}(z))\Pi_n(z)\right) \\
&= \frac{1}{b_n}(1-\mathrm{F}(z))(\Pi_n(z+h_nu) - \Pi_n(z)) \\
&\quad - (1-\mathrm{F}(z))\Pi_n(z+h_nu) + (1-\mathrm{F}(z+h_nu))\Pi_n(z+h_nu) \\
&= (1-\mathrm{F}(z))\pi_n(u) + \frac{1}{b_n}(\mathrm{F}(z) - \mathrm{F}(z+h_nu))\Pi_n(z+h_nu).
\end{aligned}$$

De plus

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\Pi(z+h_nu)| = \sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\frac{1}{\sqrt{n}}\int_0^{z+h_nu}\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)}\frac{1}{1-\mathrm{H}_{n-}(t)}\,d\mathrm{M}_n(t)\right|;$$

et comme on a d'après la loi du logarithme itéré de Földes et Rejtő (1981)

$$\sup_{u\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\left|\frac{1-\mathrm{F}_{n-}(t)}{1-\mathrm{F}(t)} - 1\right| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}}\right);$$

on en déduit que

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\Pi(z+h_nu)| = \mathrm{O}\left(\sqrt{\log_2 n}\right).$$

D'autre part, la dérivabilité de F en z et le fait que $h_n \longrightarrow 0$ donnent

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\mathrm{F}(z) - \mathrm{F}(z+h_nu)| = \mathrm{O}(h_n).$$

Ce qui fait qu'au final on ait

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\xi_n(u) - (1-\mathrm{F}(z))\pi_n(u)| = \mathrm{O}(\sqrt{h_n}) \longrightarrow 0 \quad \text{p.s}$$

Le résultat du lemme découle enfin des lemmes 4.4 et 4.6 sachant que $\frac{1-\mathrm{F}(z)}{1-\mathrm{H}(z)} = \frac{1}{1-\mathrm{G}(z)}$. □

Preuve du Théorème 4.1. Si on suppose que F est dérivable en z et G continue en z, alors, sous les hypothèses (H1) et (H2), et par le Théorème 4.3, la suite $\left\{\frac{k_n^{(1)}}{1-\mathrm{G}(z)} : n \geqslant 1\right\}$ est presque sûrement relativement compacte avec pour ensemble limite l'ensemble des fonctions $h \in \mathbb{B}([-\mathrm{M}, \mathrm{M}])$ de la forme :

$$h(u) = \int_0^u \frac{\phi(s)}{1-\mathrm{G}(z)} ds \qquad -\mathrm{M} \leqslant u \leqslant \mathrm{M},$$

avec

$$\int_{-\mathrm{M}}^{\mathrm{M}} \phi^2(s) ds \leqslant f(z)(1-\mathrm{G}(z)).$$

Le résultat en découle en posant $\psi(s) = \frac{\phi(s)}{1-\mathrm{G}(z)}$ et en appliquant le Lemme 4.7. □

4.3 Application : Estimation de la densité

Le but du Théorème 4.1 est de donner des résultats pour quelques estimateurs qui dépendent localement du processus empirique. L'exemple que nous allons traiter est l'estimateur à noyau de la densité, mais la méthode que nous exposerons peut être utilisée pour les autres estimateurs du même type (par exemple l'estimateur du taux de hasard).

Considérons une fonctionnelle Γ définie et continue sur un ensemble fermé $\mathscr{S}$ de $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$ et satisfaisant la condition $\xi_n \in \mathscr{S}, \forall n \geqslant 1$. On définit alors la statistique $\mathrm{T}_n = \Gamma(\xi_n)$.

Théorème 4.4 (Deheuvels et Einmahl, 1996, Théorème 3.1)**.** *Soit $z \in]0, \Theta[$ et $\mathrm{M} > 0$ fixés. On suppose que F est continue dans un voisinage de z et dérivable en z, et que G est continue en z. Alors, sous (H1) et (H2), la suite $(\mathrm{T}_n)_{n\geqslant 1}$ est presque sûrement relativement compacte dans $\mathbb{R}$, avec pour ensemble limite l'intervalle*

$$\left[\inf_{h\in\mathbb{L}_{\mathrm{M}}} \Gamma(h), \sup_{h\in\mathbb{L}_{\mathrm{M}}} \Gamma(h)\right],$$

où $\mathbb{L}_{\mathrm{M}}$ est l'ensemble limite de (ξ_n) (voir le théorème 4.1).

Démonstration. Le fait que T_n soit presque sûrement relativement compacte dans $\mathbb{R}$ avec pour ensemble limite $\Gamma(\mathbb{L}_{\mathrm{M}})$ découle directement du Théorème 4.1.

L'ensemble $\mathbb{L}_{\mathrm{M}}$ est connexe dans $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$. En effet, l'ensemble des fonctions ψ vérifiant $\int_{-\mathrm{M}}^{\mathrm{M}} \psi^2(s)\, ds \leqslant \frac{f(z)}{1-\mathrm{G}(z)}$ est une boule fermée de $\mathrm{L}^2([-\mathrm{M}, \mathrm{M}])$, elle est donc connexe. L'opérateur qui à à toute fonction $\psi \in \mathrm{L}^2([-\mathrm{M}, \mathrm{M}])$ associe la fonction Ψ définie pour $u \in [-\mathrm{M}, \mathrm{M}]$ par $\Psi(u) = \int_0^u \psi(s)\, ds$ étant continu, $\mathbb{L}_{\mathrm{M}}$ est aussi connexe.

De plus, si $\Psi \in \mathbb{L}_{\mathrm{M}}$, alors il existe $\psi \in \mathrm{L}^2([-\mathrm{M},\mathrm{M}])$ tel que $\Psi(u) = \int_0^u \psi(s)\,ds$, et on a alors, pour tout $x_1, x_2 \in [-\mathrm{M},\mathrm{M}]$:

$$\begin{aligned}
|\Psi(x_2) - \Psi(x_1)| &= \left| \int_{x_1}^{x_2} \psi(s)\,ds \right| \\
&\leqslant |x_2 - x_1|^{1/2} \left(\int_{-\mathrm{M}}^{\mathrm{M}} \psi^2(s)\,ds \right)^{1/2} \\
&\leqslant \left(\frac{f(z)}{1-\mathrm{G}(z)} \right)^{1/2} |x_2 - x_1|^{1/2};
\end{aligned}$$

ce qui montre que $\mathbb{L}_{\mathrm{M}}$ est uniformément équicontinu. Et comme $\mathbb{L}_{\mathrm{M}}$ est borné (c'est l'image par un opérateur borné d'une boule), il est compact d'après le théorème d'Arzelà-Ascoli.

Son image par l'application continue Γ est donc un intervalle fermé. □

En choisissant la fonctionnelle Γ, on peut montrer des taux de convergences pour quelques estimateurs qui dépendent localement du processus empirique, la preuve ci dessous est un exemple pour l'estimateur à noyau de la densité.

Preuve du Théorème 4.2. En vertu de l'hypothèse (K2), il existe un réel positif M tel que $\forall u, |u| \geqslant \frac{\mathrm{M}}{2} \implies \mathrm{K}(u) = 0$. Donc,

$$\begin{aligned}
f_n(z) - \mathbb{E}f_n(z) &= \int_{-\infty}^{+\infty} \frac{1}{h_n} \mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}_n(t) - \int_{-\infty}^{+\infty} \frac{1}{h_n} \mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}(t) \\
&= \frac{1}{h_n} \int_{-\infty}^{+\infty} \mathrm{K}(u)\, d(\mathrm{F}_n(z+h_n u) - \mathrm{F}(z+h_n u)) \\
&= \frac{1}{h_n} \int_{-\mathrm{M}}^{\mathrm{M}} \mathrm{K}(u)\, d(\mathrm{F}_n(z+h_n u) - \mathrm{F}_n(z) - \mathrm{F}(z+h_n u) + \mathrm{F}(z)) \qquad (4.4) \\
&= \frac{-1}{h_n} \int_{-\mathrm{M}}^{\mathrm{M}} (\mathrm{F}_n(z+h_n u) - \mathrm{F}_n(z) - \mathrm{F}(z+h_n u) + \mathrm{F}(z))\, d\mathrm{K}(u) \qquad (4.5) \\
&= \frac{-b_n}{h_n\sqrt{n}} \int_{-\mathrm{M}}^{\mathrm{M}} \xi_n(u)\, d\mathrm{K}(u),
\end{aligned}$$

où (4.4) est due au fait que $\mathrm{F}_n(z)$ et $\mathrm{F}(z)$ sont constants, et (4.5) est obtenue en utilisant l'intégration par parties (ce qui est possible grâce à l'hypothèse (K1)).

Définissons la fonctionnelle Γ en posant pour toute fonction $h \in \mathbb{B}([-\mathrm{M},\mathrm{M}])$ à variation bornée :

$$\Gamma(h) = -\int_{-\mathrm{M}}^{\mathrm{M}} h(u)\, d\mathrm{K}(u),$$

et, du du fait que $b_n = \sqrt{2h_n \log_2 n}$ on a alors :

$$\mathrm{T}_n = \Gamma(\xi_n) = \frac{h_n\sqrt{n}}{b_n}(f_n(z) - \mathbb{E}f_n(z)) = \sqrt{\frac{nh_n}{2\log_2 n}}(f_n(z) - \mathbb{E}f_n(z)).$$

Le Théorème 4.4 s'applique et l'ensemble limite de la suite T_n est l'intervalle

$$\left[\inf_{h\in\mathbb{L}_\mathrm{M}} \Gamma(h), \sup_{h\in\mathbb{L}_\mathrm{M}} \Gamma(h)\right].$$

Il reste à calculer $\sup_{h\in\mathbb{L}_\mathrm{M}} \pm\Gamma(h)$ (car $\inf_x f(x) = -\sup_x -f(x)$). Or par les hypothèses (K1) et (K2), une intégration par parties permet d'écrire

$$\begin{aligned}\sup_{h\in\mathbb{L}_\mathrm{M}} \pm\Gamma(h) &= \sup_{h\in\mathbb{L}_\mathrm{M}} \pm\int_{-\mathrm{M}}^{\mathrm{M}} \mathrm{K}(u)\, dh(u) \\ &= \sup_{h\in\mathbb{L}_\mathrm{M}} \left\{\pm\int_{-\mathrm{M}}^{\mathrm{M}} \mathrm{K}(u)\Psi(u)\, du : h(u) = \int_0^u \Psi(s)\, ds\right\}.\end{aligned}$$

Pour simplifier le calcul de cette expression, on pose :

$$c_- = \frac{1-\mathrm{G}_-(z)}{f_-(z)} \quad ; \quad c_+ = \frac{1-\mathrm{G}(z)}{f_+(z)};$$

$$\mathrm{K}^*(t) = \begin{cases} \frac{1}{\sqrt{c_-}}\mathrm{K}(t) & \text{si } t<0, \\ \frac{1}{\sqrt{c_+}}\mathrm{K}(t) & \text{sinon;}\end{cases}$$

$$\Psi^*(t) = \begin{cases} \sqrt{c_-}\Psi(t) & \text{si } t<0, \\ \sqrt{c_+}\Psi(t) & \text{sinon;}\end{cases}$$

ce qui donne que $\mathrm{K}(t)\Psi(t) = \mathrm{K}^*(t)\Psi^*(t)$ et $h \in \mathbb{L}_\mathrm{M} \iff \int_{-\mathrm{M}}^{\mathrm{M}} (\Psi^*(u))^2\, du \leqslant 1$. Enfin :

$$\begin{aligned}\sup_{h\in\mathbb{L}_\mathrm{M}} \pm\Gamma(h) &= \sup\left\{\pm\int_{-\mathrm{M}}^{\mathrm{M}} \mathrm{K}^*(u)\Psi^*(u)\, du : \int_{-\mathrm{M}}^{\mathrm{M}} (\Psi^*(u))^2\, du \leqslant 1\right\} \\ &\leqslant \left(\int_{-\mathrm{M}}^{\mathrm{M}} (\mathrm{K}^*(u))^2\, du\right)^{1/2}.\end{aligned}$$

d'après l'inégalité de Schwarz. D'autre part, le choix particulier de

$$\Psi^*(u) = \frac{\mathrm{K}^*(u)}{\left(\int_{-\mathrm{M}}^{\mathrm{M}} (\mathrm{K}^*(t))^2\right)^{1/2}}$$

montre l'égalité $\sup_{h\in\mathbb{L}_{\mathrm{M}}} \pm\Gamma(h) = \left(\int_{-\mathrm{M}}^{\mathrm{M}} (\mathrm{K}^*(u))^2 \, du\right)^{1/2}$. D'où

$$\sup_{h\in\mathbb{L}_{\mathrm{M}}} \pm\Gamma(h) = \left(\frac{f_-(z)}{1-\mathrm{G}_-(z)}\int_{-\mathrm{M}}^{0} \mathrm{K}^2(u)\, du + \frac{f_+(z)}{1-\mathrm{G}(z)}\int_{0}^{\mathrm{M}} \mathrm{K}^2(u)\, du\right)^{1/2}.$$

Le résultat visé en découle en rappelant que pour toute suite (x_n), $\limsup x_n$ n'est autre que la borne supérieure de l'ensemble des valeurs d'adhérence de (x_n). □

Chapitre 5

Extension au cas de la censure double

Plusieurs modèles non paramétriques ont été proposés pour l'étude de la censure double. Par exemple, le modèle de Turnbull (1974) est le plus utilisé, et plusieurs travaux sont basés sur ce modèle. Cependant, bien que la définition de ce modèle soit plus intuitive, l'estimateur qui est proposé n'est pas pour autant facile à utiliser, car il est donné par une équation intégrale dont la solution n'est pas connue explicitement.

D'autres modèles (qui englobent parfois la censure par intervalles) ont été proposés par Peto (1973), Samuelsen (1989) ou encore Huang (1999). Dans des rapports techniques, Patilea et Rolin (2001, 2004) discutent les avantages et les inconvénients de ces modèles, et en proposent d'autres.

Dans la suite nous nous intéressons au modèle de Patilea et Rolin (2006) (qui reprend d'ailleurs le modèle de Patilea et Rolin (2001)), et nous discutons de la possibilité d'étendre les résultats du chapitre 4 à ce modèle. Nous terminons par une étude de simulation pour voir la performance de l'estimateur de la densité obtenu pour ce dernier modèle.

5.1 Présentation du modèle de Patilea et Rolin (2006)

Considérons trois variables aléatoires positives indépendantes X, L et R de fonctions de répartition respectives F, $\mathrm{F_L}$ et $\mathrm{F_R}$, et de fonctions de survie [1] respectives S, $\mathrm{S_L}$ et $\mathrm{S_R}$, où X représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Dans le modèle I de Patilea et Rolin (2006), au lieu d'observer un échantillon de X on observe un échantillon du couple (Z, A) où $\mathrm{Z} = \max(\min(\mathrm{X}, \mathrm{R}), \mathrm{L})$

1. Si F est la fonction de répartition d'une variable aléatoire X, alors sa fonction de survie est $\mathrm{S} = 1 - \mathrm{F}$.

et

$$\mathrm{A} = \begin{cases} 0 & \text{si } \mathrm{L} < \mathrm{X} \leqslant \mathrm{R}, \\ 1 & \text{si } \mathrm{L} < \mathrm{R} < \mathrm{X}, \\ 2 & \text{si } \min(\mathrm{X}, \mathrm{R}) \leqslant \mathrm{L}. \end{cases}$$

Ce modèle considère la censure à droite et la censure à gauche comme deux phénomènes qui agissent indépendamment l'un de l'autre mais que l'un peut censurer l'autre. Un exemple de ce modèle est donné par un système formé par trois composants, dont deux sont placés en série (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en parallèle avec ce système en série.

Le modèle II proposé par les mêmes auteurs est similaire, mais le rôle de la censure à droite et à gauche est inversé. On observe un échantillon du couple (Z, A) où $\mathrm{Z} = \min(\max(\mathrm{X}, \mathrm{L}), \mathrm{R})$ et

$$\mathrm{A} = \begin{cases} 0 & \text{si } \mathrm{L} < \mathrm{X} < \mathrm{R}, \\ 1 & \text{si } \mathrm{R} < \max(\mathrm{X}, \mathrm{L}), \\ 2 & \text{si } \mathrm{X} \leqslant \mathrm{L} \leqslant \mathrm{R}. \end{cases}$$

Le traitement des deux modèles étant très similaire, nous nous contentons de parler du premier. Considérons H la fonction de répartition de Z, elle se décompose en $\sum_{k=0}^{2} \mathrm{H}^{(k)}(t)$ où

$$\mathrm{H}^{(k)}(t) = \mathrm{P}(\mathrm{Z} \leqslant t, \mathrm{A} = k), \quad \text{pour } k = 0, 1, 2.$$

Ces fonctions peuvent s'écrire :

$$\begin{aligned} \mathrm{H}^{(0)}(t) &= \int_0^t \mathrm{F}_{\mathrm{L}-}(t)\mathrm{S}_{\mathrm{R}-}(t)\, d\mathrm{F}(t), \\ \mathrm{H}^{(1)}(t) &= \int_0^t \mathrm{F}_{\mathrm{L}-}(t)\mathrm{S}(t)\, d\mathrm{F}_{\mathrm{R}}(t), \\ \mathrm{H}^{(2)}(t) &= \int_0^t \{1 - \mathrm{S}(t)\mathrm{S}_{\mathrm{R}}(t)\}\, d\mathrm{F}_{\mathrm{L}}(t), \end{aligned}$$

et c'est à partir de ces équations que l'estimateur est obtenu.

L'idée est de considérer dans un premier temps $\mathrm{S} = \min(\mathrm{X}, \mathrm{R})$ et L dans un modèle de censure à gauche (c'est-à-dire que l'on considère une donnée complète si $\mathrm{A} = 0$ ou $\mathrm{A} = 1$ et censurée à gauche si $\mathrm{A} = 2$), et d'estimer la fonction de répartition F_{S}, puis utiliser la fonction de répartition ainsi estimée au lieu de la fonction de répartition empirique de S pour estimer la fonction de répartition de la variable d'intérêt X en considérant un modèle de censure à droite.

L'estimateur de la fonction de survie S ainsi obtenu, en remplaçant à la fin les fonctions $\mathrm{H}^{(0)}$, $\mathrm{H}^{(1)}$ et $\mathrm{H}^{(2)}$ par leurs estimateurs empiriques, obtenus à partir d'un échantillon

$(\mathrm{Z}_i, \mathrm{A}_i)_{1\leqslant i\leqslant n}$, est donné par :

$$1 - \mathrm{F}_n(\mathrm{Z}'_j) = \prod_{1\leqslant l\leqslant j} \left\{1 - \frac{\mathrm{D}_{0l}}{\mathrm{U}_{l-1} - \mathrm{N}_{l-1}}\right\},$$

où $(\mathrm{Z}'_j)_{1\leqslant j\leqslant \mathrm{M}}$ sont les valeurs distinctes des Z_i prises dans l'ordre croissant, et

$$\mathrm{D}_{kj} = \sum_{1\leqslant i\leqslant n} 1_{\{\mathrm{Z}_i=\mathrm{Z}'_j, \mathrm{A}_i=k\}},$$

$$\mathrm{N}_j = \sum_{1\leqslant i\leqslant n} 1_{\{\mathrm{Z}_i\leqslant \mathrm{Z}'_j\}},$$

$$\mathrm{U}_{j-1} = n \prod_{j\leqslant l\leqslant \mathrm{M}} \left\{1 - \frac{\mathrm{D}_{2l}}{\mathrm{N}_l}\right\},$$

pour $0 \leqslant l \leqslant 2$ et $1 \leqslant j \leqslant \mathrm{M}$.

Soulignons le fait que si $\mathrm{L} \equiv 0$ (pas de censure à gauche), $1 - \mathrm{F}_n$ se réduit à l'estimateur de Kaplan-Meier qui lui-même se réduit au complément à 1 de la fonction de répartition empirique si $\mathrm{R} \equiv \infty$.

L'estimateur à noyau de la densité peut être défini en fonction de cet estimateur de la même manière que pour les données complètes ou censurées à droite, autrement dit il s'écrit

$$f_n(z) = \int_{-\infty}^{+\infty} \frac{1}{h_n} \mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}_n(t).$$

5.2 Quelques résultats préliminaires

Une étude attentive du chapitre précédant met en évidence le fait que les principaux outils utilisés par Deheuvels et Einmahl (1996) pour arriver à leurs résultats sont une loi fonctionnelle du logarithme itéré due à Deheuvels et Mason (1994) (Lemme 4.2) ainsi qu'une loi du logarithme itéré pour l'estimateur de Kaplan-Meier (Corollaire 3.1). Le premier est donné dans un cadre bien plus général et semble tout à fait s'appliquer au cas du modèle de censure mixte de Patilea et Rolin (2006). Le second a été étendu par Messaci et Nemouchi (2011) à l'estimateur F_n et s'énonce comme suit. Rappelons que pour toute v.a. U de fonction de répartition F, $\mathrm{I}_\mathrm{U} = \inf\{t : \mathrm{F}(t) > 0\}$ et $\mathrm{T}_\mathrm{U} = \sup\{t : \mathrm{F}(t) < 1\}$.

Théorème 5.1. *On suppose que* F, F_L *et* F_R *sont continues et que* $\mathrm{I}_\mathrm{L} < \mathrm{I}_\mathrm{X} \leqslant \mathrm{I}_\mathrm{R}$ *et* $\mathrm{T}_\mathrm{X} < \mathrm{T}_\mathrm{R}$. *Alors,*

$$\mathrm{P}\left(\sup_{u\in\mathbb{R}} |\mathrm{F}_n(u) - \mathrm{F}(u)| = \mathrm{O}\left(\sqrt{\frac{\log_2 n}{n}}\right)\right) = 1.$$

Ceci montre qu'il est tout à fait envisageable d'étendre les résultats du chapitre précédent à l'estimateur de Patilea et Rolin (2006), travail qui fera l'objet d'une recherche ultérieure. Nous nous contentons dans le paragraphe suivant de mener une étude de simulation afin d'évaluer les performances de l'estimateur à noyau de la densité, obtenu dans le cadre de la censure double.

5.3 Étude de simulation

On va maintenant mener une étude de simulation pour évaluer les performances de l'estimateur à noyau de la densité dans ce cadre.

Pour l'estimation de la densité, il est raisonnable de supposer qu'il n'y pas d'ex-æquo. Ainsi, l'estimateur de Patilea et Rolin (2006) devient :

$$1 - \mathrm{F}_n(\mathrm{Z}_j) = \prod_{1 \leqslant k \leqslant j} \left\{ 1 - \frac{1_{\{\mathrm{A}_k = 0\}}}{\mathrm{U}_{k-1} - k + 1} \right\},$$

avec

$$\mathrm{U}_{j-1} = n \prod_{j \leqslant k \leqslant n} \left\{ 1 - \frac{1_{\{\mathrm{A}_k = 2\}}}{k} \right\}.$$

En substituant dans la définition de f_n donnée ci-dessus nous obtenons :

$$f_n(z) = \sum_{1 \leqslant j \leqslant n} \frac{1_{\{\mathrm{A}_j = 0\}} \mathrm{F}_n(\mathrm{Z}_{j-1})}{\mathrm{U}_{j-1} - j + 1} \frac{1}{h_n} \mathrm{K}\left(\frac{\mathrm{Z}_j - z}{h_n} \right).$$

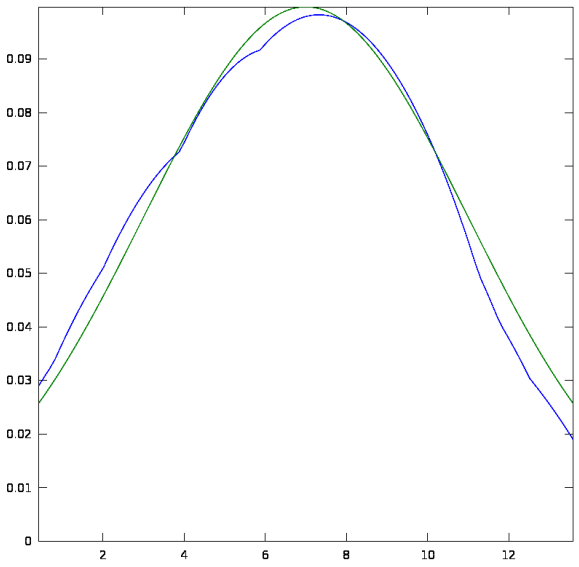
C'est cet estimateur que nous allons étudier pour deux modèles, d'abord un modèle normal puis un modèle Weibull.

Dans le premier modèle, la variable X suit une loi normale de moyenne 7 et d'écart type 4, R suit une loi normale de moyenne 11 et d'écart type 4, et L suit une loi normale de moyenne 3 et d'écart type 4. Le taux de censure n'est pas choisi à l'avance mais est calculé à partir de l'échantillon. Bien que la loi normale ne soit pas positive, ce qui est une condition pour le modèle de Patilea et Rolin (2006), on obtient de bons résultats, ce qui suggère que la condition de positivité n'est pas nécessaire.

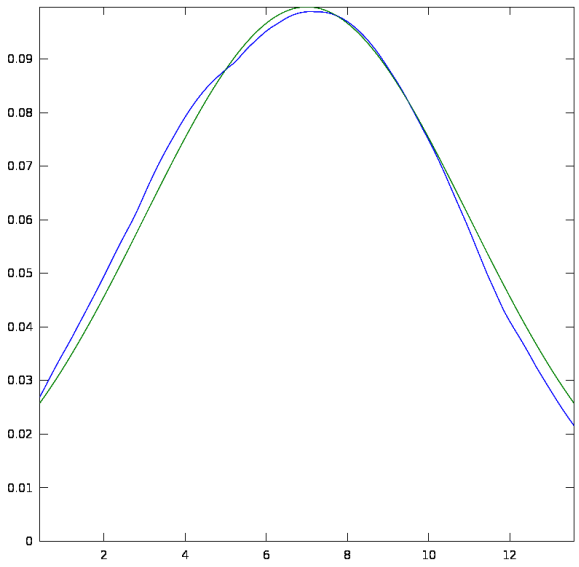
Les graphes suivants ont été réalisés en utilisant le noyau parabolique donné par

$$\mathrm{K}(x) = \frac{3}{4}(1 - x^2) 1_{\{|x| < 1\}}.$$

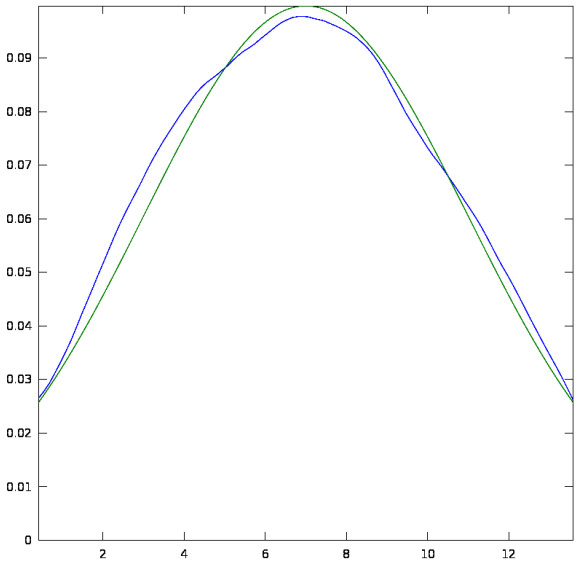
La densité théorique est en vert, et la densité estimée en bleu.



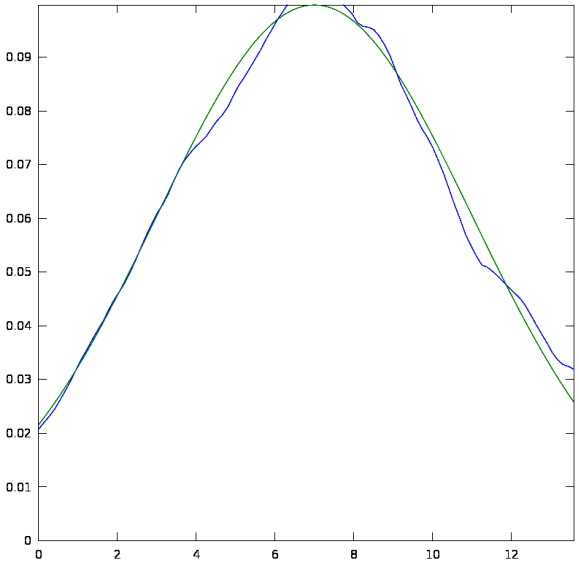
$n = 10$
20% de censure à droite
20% de censure à gauche



$n = 50$
14% de censure à droite
18% de censure à gauche



$n = 100$
15% de censure à droite
23% de censure à gauche

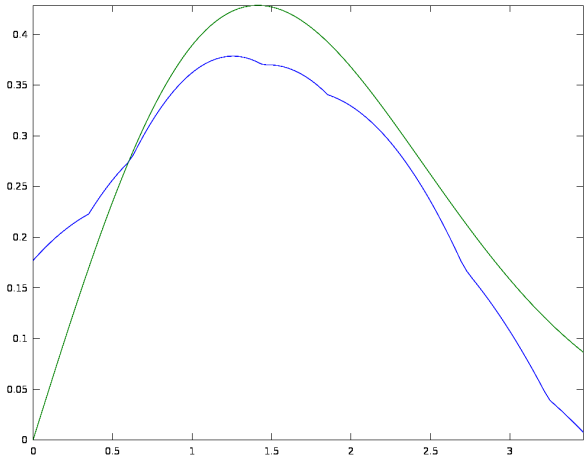


$n = 300$
21% de censure à droite
27% de censure à gauche

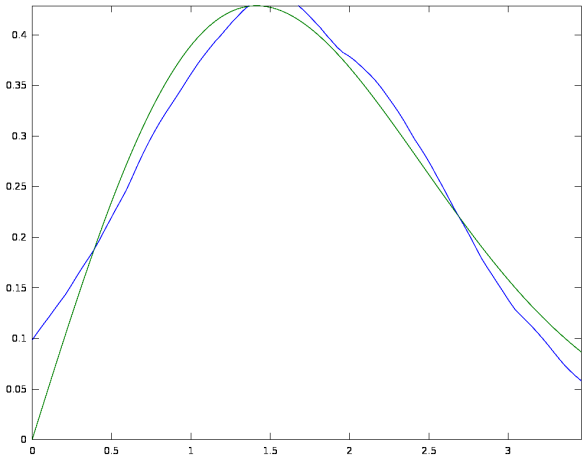
Dans le second modèle, X suit une loi de Weibull de paramètres $\alpha = 2, \beta = 2$, R suit une loi de Weibull de paramètres $\alpha = 3, \beta = 2$, et L suit une loi de Weibull de paramètres $\alpha = 1, \beta = 2$.

La loi de Weibull est définie par sa densité, donnée pour $x > 0$ par :

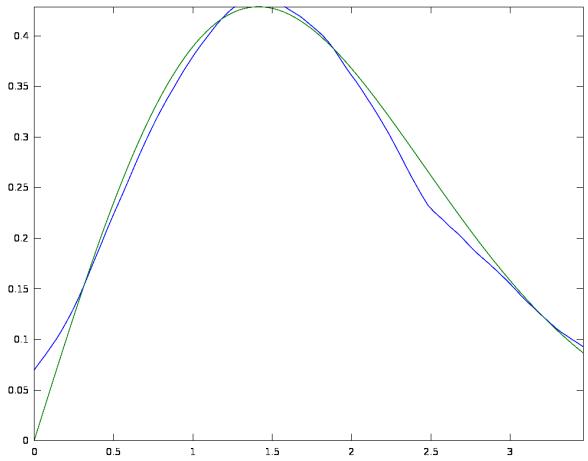
$$\frac{\beta}{\alpha^\beta} x^{\beta-1} \exp(-(\frac{x}{\alpha})^\beta)$$



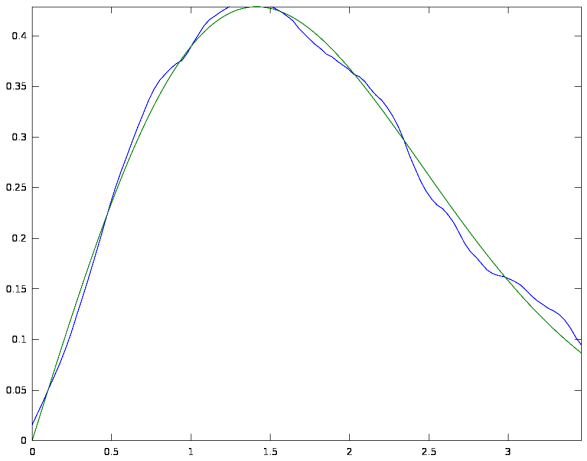
$n = 10$
30% de censure à droite
10% de censure à droite



$n = 50$
16% de censure à droite
20% de censure à gauche



$n = 100$
21% de censure à droite
20% de censure à gauche



$n = 300$
22% de censure à droite
28% de censure à gauche

Conclusion Les graphes obtenus pour le modèle gaussien montrent une bonne performance de l'estimateur de la densité et ce même pour les petites tailles de l'échantillon. De plus, la qualité de l'estimation s'améliore quand la taille de l'échantillon augmente, ce qui est tout à fait prévisible.

Quant au modèle de Weibull, bien que la qualité de l'estimation ne soit pas bonne pour $n = 10$, les graphes montrent une amélioration tout à fait satisfaisante quand la taille de l'échantillon augmente.

Notons aussi que le taux de censure se situe autour de 40%, ce qui nous semble appréciable.

Bibliographie

Odd Aalen : *Statistical theory for a family of counting processes*. Institute of Mathematical Statistics, Univ. of Copenhagen, 1976.

Patrick Billingsley : *Convergence of Probability Measures*. Wiley, New York, 1968.

Patrick Billingsley : *Weak Convergence of Measures : Applications in Probability*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, Pennsylvania, 1971.

Patrick Billingsley : *Probability and Measure*. John Wiley & Sons, 2e édition, 1986.

N. Breslow et J. Crowley : A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453, 1974.

F. P. Cantelli : Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424, 1933.

D. M. Chibisov : Some theorems on the limiting behavior of empirical distribution functions. *Selected Transl. Math. Statist. Prob.*, 6:147–156, 1964.

Kai-Lai Chung : An estimate concerning the Kolmogoroff limit distribution. *Transactions of the American Mathematical Society*, 67(1):36–50, September 1949.

P. Csörgő, M. et Révész : *Strong Approximation in Probability and Statistics*. Acedemic Press, New York, 1981.

Alejandro de Acosta : A new proof of the Hartman-Wintner law of the iterated logarithm. *The Annals of Probability*, 11(2):270–276, May 1983.

Paul Deheuvels : Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre des estimateurs de la densité. *C. R. Acad. Sci., Paris, Sér. A*, 278:1217–1220, 1974.

Paul Deheuvels : Laws of the iterated logarithm for density estimators. *In* G. Roussas, éditeur : *Nonparametric Functional estimators and Related Topics*, NATO adv.Sci.Ins. C, pages 19–29. Kluwer Academic, 1991.

Paul Deheuvels : Functional laws of the iterated logarithm for large increments of empirical and quantile processes. *Stochastic Processes Appl.*, 43(1):133–163, 1992.

Paul Deheuvels : Chung type functional laws of the iterated logarithm for tail empirical processes. *Ann. I. H. P. B*, 36:583–616, 2000.

Paul Deheuvels et John H. J. Einmahl : On the strong limiting behavior of local functionals of empirical processes based upon censored data. *The Annals of Probability*, 24 (1):504–525, 1996.

Paul Deheuvels et David Mason : Functional laws of the iterated logarithm for the increments of empirical and quantile processes. *Ann. Prob.*, 20:1248–1287, 1992.

Paul Deheuvels et David Mason : General confidence bounds for nonparametric functional estimators. *Stat. Inf. for Stoch. Proc.*, 7:225–277, 2004.

Paul Deheuvels et David M. Mason : Functional laws of the iterated logarithm for local empirical processes indexed by sets. *The Annals of Probability*, 22(3):1619–1661, 1994.

M. D. Donsker : Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 23:277–281, 1952.

R. M. Dudley : Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois Journal of Mathematics*, 10:109–126, 1966.

R. M. Dudley : measures on non-separable metric spaces. *Illinois Journal of Mathematics*, 11(3):449–453, 1967.

R. M. Dudley : Distances of probability measures and random variables. *The Annals of Mathematical Statistics*, 39:1563–1572, 1968.

R. M. Dudley : *Probabilities and metrics : Convergence of laws on metric spaces, with a view to statistical testing*, volume 45 de *Lecture Notes Series*. Matematish Institut, Aarhus Universitet, 1976.

R. M. Dudley : *Uniform central limit theorems*. Cambridge University Press, 1999.

John H. J. Einmahl et A. J. Koning : Limit theorems for a general weighted process under random censoring. *Canadian Journal of Statistics*, 20, 1992.

Uwe Einmahl et David Mason : An empirical process approach to the uniform consistency of kernel type estimators. *Journ. Theoretic. Probab.*, 13:1–13, 2000.

Uwe Einmahl et David M. Mason : Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

Pedro J. Fernandez : Almost surely convergent versions of sequences which converge weakly. *Bulletin of the Brazilian Mathematical Society*, 5:51–61, 1974.

Helen Finkelstein : The law of the iterated logarithm for empirical distribution. *The Annals of Mathematical Statistics*, 42(2):607–615, 1971.

Thomas R. Fleming et David P. Harrington : *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

A. Földes et L. Rejtő : A LIL type result for the product limit estimator. *Probability Theory and Related Fields*, 56(1):75–86, 1981.

Richard Gill : Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*, 11(1):49–58, 1983.

Richard D. Gill et Soren Johansen : A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555, 1990.

V. Glivenko : Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99, 1933.

Ming Gao Gu et Tze Leung Lai : Functional laws of the iterated logarithm for the product limit-estimator of a distribution function under random censorship or truncation. *The Annals of Probability*, 18(1):160–189, 1990.

Philip Hartman et Aurel Wintner : On the law of the iterated logarithm. *American Journal of Mathematics*, 63(1):169–176, January 1941.

J. Huang : Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, 9:501–519, 1999.

B. R. James : A functional law of the iterated logarithm for weighted empirical distributions. *The Annals of Probability*, 3:762–772, 1975.

E. L. Kaplan et P. Meier : Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

A. Khinchine : Über einen Satz der Wahrscheinlichkeitsrechnung. *Fundamenta Mathematica*, 6:9–20, 1924.

A. Kolmogorov : Über das Gesetz des iterierten Logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

A. N. Kolmogorov : Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

Michael R. Kosorok : *Introduction to Empirical Processes and Semiparametric Inference*. Springer Verlag, New York, 2008.

T. L. Lai : Convergence rates in the strong law of large numbers for random variables taking values in Banach spaces. *Bull. Inst. Math. Acad. Sinica*, 2:67–85, 1974.

Fatiha Messaci et Nahima Nemouchi : A law of the iterated logarithm for the product limit estimator with doubly censored data. *Statistics & Probability Letters*, 81(8):1241–1244, 2011.

N. E. O'Reilly : On the weak convergence of empirical processes in sup-norm metrics. *The Annals of Probability*, 2:642–651, 1974.

Emanuel Parzen : On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

Valentin Patilea et Jean-Marie Rolin : Product limit estimators of the survival function for doubly censored data. Rapport technique, Institut de Statistique, Université Catholique de Louvain, July 2001.

Valentin Patilea et Jean-Marie Rolin : Product-limit estimators of the survival function with left or right censored data. Rapport technique, Institut de Statistique, Université Catholique de Louvain, 2004.

Valentin Patilea et Jean-Marie Rolin : Product limit estimators of the survival function with twice censored data. *The Annals of Statistics*, 34(2):925–938, 2006.

Richard Peto : Experimental survival curves for interval censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(1):86–91, 1973.

D. Pollard : *Convergence of stochastic processes*. Springer-Verlag, New York, 1984.

Yu. V. Prokhorov : Convergence of random processes and limit theorems in probability theory. *Theory of Probability and its Applications*, 1(2):157–214, 1956.

Murray Rosenblatt : Remarks on some nonparametric estimates of density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

S. O. Samuelsen : Asymptotic theory for non-parametric estimators from doubly censored data. *Scandinavian Journal of Statistics*, 16:1–21, 1989.

Galen R. Shorack et Jon W. Wellner : *Empirical processes with applications to statistics*. John Wiley & Sons, 1986.

A. V. Skorokhod : Limit theorems for stochastic processes. *Theory of Probability and its Applications*, 1(3):261–290, 1956.

N. Smirnov : Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique [Matematicheskii Sbornik]*, 6(48)(1):3–26, 1939.

W. Stute : The law of the iterated logarithm for kernel density estimators. *Ann. Prob.*, 10:414–422, 1982a.

W. Stute : The oscillation behavior of empirical processes. *Ann. Prob.*, 10:86–107, 1982b.

W. Stute : Conditional empirical processes. *Ann. Prob.*, 14(2):638–647, 1986a.

W. Stute : On almost sure convergence of conditional empirical distribution functions. *Ann. Prob.*, 14(3):891–901, 1986b.

W. Stute et J.-L. Wang : The strong law under random censorship. *The Annals of Statistics*, 21(3):1591–1607, 1993.

Bruce W. Turnbull : Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345):169–173, March 1974.

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

A. W. van der Vaart et Jon A. Wellner : *Weak Convergence and Empirical Processes : With Applications in Statistics*. Springer Verlag, New York, 1996.

Vivian Viallon : *Processus empiriques, estimation non paramétrique et données censurées*. Thèse de doctorat, Université Paris 6, 2006.

Jon A. Wellner : A Glivenko-Cantelli theorem and strong law of large numbers for functions of order statistics. *The Annals of Statistics*, pages 473–780, 1977.

M. J. Wichura : On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *The Annals of Mathematical Statistics*, 41:284–291, 1970.

B. B. Winter, A. Földes et L. Rejtő : Glivenko-Cantelli theorems for the product limit estimate. *Problems of Control and Information Theory*, 7:213–225, 1978.

ملخّص

في هذه الرسالة، نقدم بعض النتائج الأساسية لنظرية العمليات التجريبية، كما نقدم بعض النتائج المماثلة في حالة المعطيات المحجوبة [1] . نهتمّ بشكل خاص بنتيجة Deheuvels and Einmahl (1996) ، وهو قانون دالّي للوغارتم المكرّر بالنسبة لتزايدات العملية التجريبية في حالة المعطيات المحجوبة من اليمين. و في الأخير، نناقش آفاق البحث في تعميم هذه النتيجة إلى قضية الحجب المزدوج.

إذا أخذنا عيّنة $(X_i)_{1\leq i\leq n}$ ذات دالّة توزيع F، وإذا كانت F_n دالة التوزيع التجريبيّة فإن العمليّة التجريبيّة تُعرّف بـ $a_n(t) = \sqrt{n}(F_n(t) - F(t))$. نقدم أوّلا بعض النتائج النظرية للعمليات التجريبية، وقانون اللوغارتم المكرّر لهذه العملية. ثمّ نركّز على عملية تجريبية في حالة المعطيات المحجوبة. في هذا النموذج، لا يمكن أن نشاهد بدقّة المعطيات التي نريد، ولكن نلا حظ عوضًا عنها كلّا من $Z_i = \min(X_i, C_i)$ و$\delta_i = 1_{\{X_i\leq C_i\}}$ حيث $(C_i)_{1\leq i\leq n}$ هي الأزمنة الحاجبة. ويكفي في هذه الحال أن نعوّض دالّة التوزيع التجريبية بمقدِّر Kaplan-Meier . نقدم بعض النتائج الأولية لهذا النموذج، ونهتم بشكل خاص بقانون دالّي للّوغارتم المكرّر لتزايدات عملية Kaplan-Meier ، أي بالعمليّة $\xi_n(u) = \frac{1}{b_n}\left(a_n(z+hu) - a_n(z)\right)$ حيث $b_n = \sqrt{2h\log\log n}$، ونقتصر في ذلك على الحالة أين يكون $h = h_n$ متتالية لا تتعلّق إلا بـ n. علينا أيضا أن تنظر في تطبيق هذه النتيجة لإعطاء سرعة تقارب مقدر نواة من الكثّافة. أخيرا، يتم إجراء دراسة محاكاة لاستكشاف خصائص امتدادا لهذا مقدر لحالة بيانات رقابة مزدوجة، على أساس نموذج Patilea and Rolin (2006) ، لتوجيه الحدس حول إمكانية تمديد نتيجة Deheuvels-Einmahl لهذا النموذج.

العبارات الرئيسية عملية تجريبية؛ معطيات محجوبة؛ قانون دالّي للّوغارتم المكرر؛ حجب مزدوج

[1] آثرت استعمال هذا المصطلح بدلًا من "مراقبة" الذي يستعمله البعض لأنّ هذا الأخير لا يؤدّي المعنى الإحصائي لكلمة "censoring" .

Abstract

In this work, we present some basic results of the theory of empirical processes, and some analogous results in the case of censored data. We are particularly interested in the result of Deheuvels and Einmahl (1996), which is a functional law of the iterated logarithm for the increments of the empirical process in the setting of right-censored data. Finally, we discuss the possibility of extending this result to the setting of twice censoring.

Given a sample $(X_i)_{1\leq i\leq n}$ with distribution function F, the empirical process is defined by $a_n(t) = \sqrt{n}(F_n(t) - F(t))$ where F_n is the empirical distribution function. First, we present some results of the theory of empirical processes, and some laws of iterated logarithm for this process. Then we look at the empirical process in the case of right censored data. In this model, we cannot observe the data of interest, but rather $Z_i = \min(X_i, C_i)$ and $\delta_i = 1_{\{X_i\leq C_i\}}$ where $(C_i)_{1\leq i\leq n}$ are censoring time. It is enough to replace the empirical distribution function by the Kaplan-Meier estimator. For this model, we present some preliminary results, and we are particularly interested in a functional law of iterated logarithm for the increments of the Kaplan-Meier empirical process, that is, for $\xi_n(u) = \frac{1}{b_n}(a_n(z + hu) - a_n(z))$ with $b_n = \sqrt{2h \log\log n}$. We consider only the case where $h = h_n$ is a deterministic sequence which only depends on n, we also consider the application of this result to give a rate of convergence for the kernel density estimator. Finally, a simulation study is conducted to explore the properties of this estimator in the setting of twice censored data, based on the model of Patilea and Rolin (2006), with the goal of investigating the possibility of extending the result of Deheuvels and Einmahl (1996) to this model.

Keywords Empirical process; Censored data; Functional law of the iterated logarithm; Twice censoring

Résumé

Dans ce mémoire, nous exposons certains résultats de base de la théorie des processus empiriques, ainsi que quelque résultats analogues dans le cas des données censurées. Nous nous intéressons particulièrement au résultat de Deheuvels et Einmahl (1996), qui est une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique dans le cadre des données censurés à droite. Enfin, nous abordons les perspectives de recherche dans la généralisation de ce résultat au cas de la censure double.

Étant donné un échantillon $(X_i)_{1\leq i\leq n}$ de fonction de répartition F, le processus empirique est définit par $a_n(t) = \sqrt{n}(F_n(t) - F(t))$ où F_n est la fonction de répartition empirique. Nous exposons dans un premier temps les résultats de la théorie des processus empiriques, ainsi que des loi du logarithme itéré pour ce processus. Ensuite, nous nous intéressons au processus empirique dans le cas de données censurées à droite. Dans ce modèle, on n'observe pas les données d'intérêt, mais plutôt $Z_i = \min(X_i, C_i)$ et $\delta_i = 1_{\{X_i\leq C_i\}}$ où $(C_i)_{1\leq i\leq n}$ sont des temps de censure, et il suffit alors de remplacer la fonction de répartition empirique par l'estimateur de Kaplan-Meier. Pour ce modèle, nous exposons quelque résultats préliminaires, et nous nous intéressons particulièrement à une loi fonctionnelle du logarithme itéré pour les accroissements du processus de Kaplan-Meier, c'est-à-dire pour $\xi_n(u) = \frac{1}{b_n}\left(a_n(z + hu) - a_n(z)\right)$ avec $b_n = \sqrt{2h \log\log n}$, nous considérons uniquement le cas où $h = h_n$ est une suite déterministe qui ne dépend que de n. Nous considérons aussi l'application de ce résultat pour donner une vitesse de convergence de l'estimateur à noyau de la densité. Enfin, une étude de simulation est menée pour explorer les propriétés d'une extension de cet estimateur au cas des données doublement censurées, d'après le modèle de Patilea et Rolin (2006), afin de guider l'intuition sur la possibilité d'étendre le résultat de Deheuvels et Einmahl (1996) à ce modèle.

Mots-clés Processus empirique ; Données censurées ; Loi fonctionnelle du logarithme itéré ; Censure double