République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieure et de la recherche scientifique

N° Attribué par la bibliothèque

Année univ.: 2016/2017

L'application des processus empiriques à la statistique des données censurées

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université Dr Tahar Moulay - Saïda

Discipline : MATHÉMATIQUES

Spécialité : Analyse Stochastique, Statistique des Processus et Applications

par

Bellouz **K**erroum [1]

Sous la direction de

Dr. Fethi Madani

Soutenue le 25 Mai 2017 devant le jury composé de

Pr. A. Kandouci	Université Dr Tahar Moulay - Saïda	Président
Dr. F. Madani	Université Dr Tahar Moulay - Saïda	Rapporteur
Dr. N. Hachemi	Université Dr Tahar Moulay - Saïda	Examinatrice
Mme. W. Benzatout	Université Dr Tahar Moulay - Saïda	Examinatrice

1. e-mail : kerroum.belouz@gmail.com

Remerciements

Avant tout je remercie, le ***Dieu*** *tout puissant de je accordée la volonté et la patience pour accomplir ce modeste travail.*

Je tenie d'abord à remercier très chaleureusement mon encadreur de mémoire de fin d'études monsieur ***Dr.Fethi Madani****, pour ses précieux conseils et son orientation tout au long de mon recherche.*

Tous les membres de jury d'avoir participé à la commission des examinateurs en vue dune évaluation prompte et à sa juste valeur.

Je remercie le directeur du laboratoire LMSSA, ***Pr.A.Kandouci*** *. Merci aux autres membres permanents du laboratoire, toujours à l'écoute, prêts à nous aider en cas de besoin et à discuter autour d'un café !*

Je remercie mes chers parents qui m'ont indiqué le bon chemin à entreprendre et qui m'ont encouragé et soutenue tout au long de mon parcoursn quotidien.

Les conseils qu'elle j'ai prodigué, la patience, la confiance qu'elle j'ai témoignés ont été déterminants dans la réalisation de mon travail de recherche. Mon remerciements s'étendent également à tous mes enseignants durant les années des études.

Enfin, je dis un grand merci du fond du coeur à tous mes amis pour leur soutient moral.

Tous mes oncles et tantes, tous mes cousins et cousines.

Tous mes enseignants de département de mathématiques . Tous mes camarades de promotion 2016 /2017.

Dédicace

Je dédie ce modeste travail à mes chers parents.

Table des matières

Introduction **9**

1 ***Les processus empiriques et données censurées*** **13**

1.1 Sur les processus empiriques 13

1.1.1 Définitions 13

1.1.2 Le processus empirique uniforme 14

1.1.3 Le processus empirique généralisé 16

1.1.4 Quelques résultats mathématiques 17

1.2 Rappel sur l'analyse de survie 23

1.2.1 Introduction 23

1.2.2 Cas des données censurées 24

1.3 Le processus empirique de Kaplan-Meier 28

1.3.1 Definition 28

1.3.2 Quelques propriétés sur l'estimateur de Kaplan - Meier 29

1.3.3 La loi du logarithme itéré pour l'estimateur de Kaplan-Meier 31

2 ***Loi fonctionnelle uniforme du logarithme pour les incréments du processus empirique censuré*** **33**

2.1 Introduction 33

2.2 Quelques résultats principaux 37

2.2.1 Nouveaux résultats 38

2.3 Preuves 39

2.3.1 Quelques notations supplémentaires et résultats utiles . 39

2.3.2 Un résultat d'approximation utile 41

2.3.3 Lemmes préliminaires 43

2.4 Approximation et loi limite fonctionnelle 45

Conclusion **51**

Bibliographie **51**

Introduction

La théorie des processus empiriques joue un rôle principal en statistique, puisqu'elle conserne l'ensemble des résultas limites généraux se rapportant aux échantillon aléatoire . De ce fait, elle comporte d'innombrables applications $\grave{a}$ des problèmes particuliers.
En 1933, F. P. Cantelli[20]et Glivenko, V[31], A. N. Kolmogorov[10]et N. Smirnov[25] donnent des resultats en commun sur la convergence de la quantité $\|F_n - F\|$, où F est la fonction de répartition de la loi d'où l'échantllon est tiré, F_n est la fonction de répartition empirique et $\|.\|$est la convergence uniforme. Ce que l'on désigne par processus empirique n'est rien de plus que la quantité F_n - F normalisée par le facteur $\sqrt{n}$.
Un tel résultat de Smirnov a donnée ce qu'on appelle la loi du logarithme itéré(Kai-Lai Chung[21]. Paul Deheuvels[5]), et plus tard des lois limites fonctionnelles , globales ou locales, (Helen Finkelstein[27]. Paul Deheuvels et David Mason[3]).
La densité f(·) fournit une description naturelle de la loi de X, au sens qu'elle en permet une interprétation visuelle simple, en révélant directement les facteurs de concentration de cette distribution de probabilité. Nous nous intéressons dans cet mémoire à l'estimation de cette densité f(·) à partir d'un échantillon aléatoire $X_1, ..., X_n$, composé de répliques indépendantes et de même loi que X. L'estimation d'une densité de probabilité est un problème statistique ancien et classique.

Pour une description approfondie des résultats classiques concernant les processus empiriques, les et leurs applications, nous renvoyons aux ouvrages de Galen R. Shorack et Jon W. Wellner[11] ; Pollard, D[41] ; van der Vaart, A. et Wellner, J[23]. Pour un texte introductif la thèse de Vivian Viallon[39] sont un bon début.

Dans les années 1980, Stute (voir [14]et[15], [16]et[17]) a été l'un des premiers statisticiens (voir également [22]), à faire un usage systématique des méthodes de la théorie des processus empiriques dans l'étude des propriétés asymptotiques d'estimateurs fonctionnels non paramétriques.

Depuis, de nombreux auteurs, parmi lesquels nous citerons Paul Deheuvels et

John H. J. Einmahl[2] (voir par exemple Einmahl, U. et Mason, D[29]), 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.

Le point de départ de l'estimation non paramétrique de la fonction de répartition fut l'introduction de la fonction de répartition empirique qui se calcule sur la base de véritables observations de la variable d'intérêt.

En analyse de survie et en fiabilité, les variables aléatoires d'intérêt représentent une durée : temps qui s'écoule jusqu'à la réalisation d'un certain événement. Ce temps est appelé temps de défaillance, durée de vie ou durée de survie, et se caractérise par la présence d'observations incomplètes. Le cas d'incomplétude le plus courant et le plus étudié aussi est la censure à droite. Il y a censure à droite lorsque la durée de survie d'intérêt est supérieure à la durée observée.

C'est le cas par exemple, dans des études de fiabilité lorsque la panne d'un appareil ne permet pas de poursuivre l'observation de l'appareil objet de l'étude. La censure à droite n'est pas la seule censure que l'on peut rencontrer avec des données de survie, beaucoup d'autres mécanismes peuvent causer des censures diverses.

Un phénomène de censure à gauche (symétrique du précédent) peut aussi empêcher l'observation du phénomène d'intérêt pour lequel on saura seulement qu'il est inférieur à la valeur observée. Généralement, la censure à gauche s'accompagne de la censure à droite comme cela est le cas pour la censure mixte à laquelle nous nous intéressons dans cette mémoire.

Dans cette 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 Paul Deheuvels et John H. J. Einmahl[2].

Dans le même esprit, nous nous sommes particuliérement intéressés dans ce manuscrit à l'étude du principe d'invariance du processus empirique de $\alpha_n(.)$ dans le cadre général ; c'est-à dire, lorsque les marges ne sont pas forcément indépendantes.

Par la suite, dans le chapitre 1,nous définitions quelques notions et résultats de base sur les processus empiriques qui seront adoptées dans la suite de la mémoire et nous énonçons des résultats mathématiques non usuels en analyse de survie qui seront utilisés pour l'obtention des résultats de convergence

de chapitre suivant. Nous présentont les caractéristiques et les propriétés essentielles permettant la bonne compréhension du sujet. Nous rappellerons également les définitions et les références les plus importantes.

La troisième partie de ce chapitre définit les données censurées, ensuite donner une définition du processus empirique pour le cas de la censure aléatoire à droite en utilisant l'estimateur de Kaplan-Meire [7].

L'objet de notre étude dans le chapitre 2 est d'établir un principe d'invariance pour les processus empiriques de α_n. Nous commençons d'abord par étudier les résultats des processus empiriques, en l'utilisation d'une loi fonctionnelle du logarithme itéré pour les incréments du processus empirique de Kaplan-Meier, résultat obtenu par Kaplan-Meire [7].

Chapitre 1

Les processus empiriques et données censurées

1.1 Sur les processus empiriques

1.1.1 Définitions

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 ($\mathbf{X}$,$\mathbf{A}$), et X_1,X_2,...,X_n un échantillon i.i.d de loi de probabilité P_x, on définit la mesure empirique P_n par :

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

où δ_X est la mesure de Dirac au point x.

On définit alors le processus empirique pour une famille S d'ensembles mesurables par $\{P_n(A), A \in S\}$.
Dans le cas réel, la fonction de répartition empirique peut s'écrire ainsi en prenant $S = \{]-\infty; t], t \in \mathbb{R}\}$.
Pour une classe F de fonctions mesurables, on peut définir un processus em-

pirique $\{P_n f, f \in F\}$ où

$$P_n = \int f dP_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

On appelle processus empirique est $\{\sqrt{n}(P_n(A) - P_X(A)), A \in \mathcal{S}\}$ ou bien $\{\sqrt{n}(P_n f - P_X f), \in \mathcal{F}\}$. 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.

Remarque 1.1.1. *Dans le cas réel,nous appellons processus empirique un processus de la forme $\sqrt{n}(F_n(x) - F(x))$ où F_n est la fonction de répartition empirique ou un autre estimateur de la fonction de répartition.*

Example : 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.

1.1.2 Le processus empirique uniforme

La réduction au cas uniforme est souvent utilisé pour simplifier les preuves.Ceci consiste à introduire un processus empirique uniforme . Soit $U_1, U_2, ..., U_n$ des variables indépendantes et idem-distribuées [i.i.d.] de loi uniforme sur[0,1], Pour tout entier $n \geq 1$, notons $F_n := n^{-1}\sharp\{U_i \leq t : 1 \leq t \leq n\}$, où $\sharp$E désigne la cardinalité de l'ensemble E.

Donc la définition du processus empirique uniforme basé sur l'échantillon $U_1, ..., U_n$ est donnée par :

$$\alpha_n(t) = \sqrt{n}(F_n(t) - t)$$

. En suite, nous notons $\xi_n(h, t; s)$, pour

$$0 \leq t \leq 1, -1 \leq s \leq 1 \quad et \quad 0 < h \leq t \wedge (1 - t).$$

Les incréments du processus empirique uniforme, définis par :
$\xi_n(h, t; s) = \alpha_n(t + hs) - \alpha_n(t)$. Dans l'étude des incréments du processus empirique, la taille de la fenêtre h joue, comme on peut s'y attendre, un rôle central, et influe notamment trés largement sur les vitesses de convergence dans les lois limites . Dans ce qui suit, nous travaillerons principalement sur des incréments h $= h_n$, vérifiant les conditions données ci-dessous . Soit $\{h_n\}_{n\geq 1}$ une suite de réelle positive, Nous supposons que :
(H.1) $(i)\, 0 < h_n < 1\, et\, h_n \to 0, \quad (ii)\, h_n \downarrow, \quad (iii)\, nh_n \uparrow \infty$;
(H.2) $nh_n / loglogn \to \infty$;

(H.3) $nh_n/logn \to c \in (0,\infty)$;
(H.4) $nh_n/logn \to 0$;
(H.5) $nh_n/logn \to \infty$;
(H.6) $nh_n/log(1/(h_n\sqrt{n})) \to \infty$;
(H.7) $(log(1/h_n))/loglogn \to \infty$;
(H.8) $(log(1/h_n))/loglogn \to d \in [0,\infty)$;
(H.9) $(log(1/h_n\sqrt{n}))/loglogn \to d \in [-\infty,\infty]$;
(H.10) $(log(1/h_n))/logn \to 1$.
Nous distinguons quatre catégories d'incréments $\xi_n(h_n,t;s)$ qui dépendent de la vitesse à laquelle h_n tend vers 0. Les incréments sont dits :

1. **Incréments larges**, lorsque h_n vérifie les conditions :
- (H.1) et (H.1)(iii) $\Rightarrow h_n > 1/n$;
- (H.8)$\Rightarrow h_n = 1/(logn)^{d_n}, avec\, d_n \to d \in [0,\infty)$;
- (H.9) $\Rightarrow h_n > \frac{1}{\sqrt{n}logn}, avec\, d \in [-\infty,0]$.
2. **Incréments standards**, lorsque h_n vérifie les conditions :
- (H.1)-(H.5)-(H.7) $\Rightarrow \frac{logn}{n} < h_n < \frac{1}{(logn)^\alpha}, 0 < \alpha < 1$.
3. **Incréments intermédiaires**, lorsque h_n vérifie les conditions :
- (H.3) $\Rightarrow h_n = c_n logn/navec\, c_n \to c \in (0,\infty)$;
- (H.9) $\Rightarrow pour\, d = \infty : h_n < \frac{1}{\sqrt{n}(logn)^\alpha}, \alpha > 0$.
4. **Incréments petits**, lorsque h_n vérifie les conditions :
- (H.1)-(H.4)-(H.7) $\Rightarrow \frac{1}{n} < h_n < \frac{logn}{n}$.

D'autre part, on définit la fonction de quantile empirique uniforme par la formule

$$F_n^{-1}(t) = \inf\{s : F_n(t) \geq t\}\, pour\, 0 \leq t \leq 1.$$

Et pour

$$0 \leq t \leq 1, -1 \leq s \leq 1 \quad et \quad 0 < h < t \wedge (1-t)$$

, les incréments du processus empirique de quantiles uniforme sont définis par

$$\xi_n(h,t;s) = \beta_n(t+hs) - \beta_n(t).$$

Le processus empirique uniforme, le processus empirique des quantiles uniforme, ainsi que leurs incréments respectifs, ont fait l'objet de recherches approfondies (voir par exemple les livres de [22], et [11]).
Ces processus interviennent dans le traitement statistique des échantillons réels, ainsi que dans les procédures d'inférences non-paramétriques correspondantes. On trouvera une littérature abondante concernant l'étude de ces processus.

La connaissance des lois limites et des propriétés asymptotiques du processus empirique uniforme permet d'établir les lois asymptotiques de plusieurs fonctionnelles basées sur ce dernier. La question qu'on peut se poser est avec

quelle vitesse le processus empirique converge-t-il vers sa loi limite ? C'est l'étude du principe d'invariance du processus empirique uniforme $\alpha_n(.)$.

1.1.3 Le processus empirique généralisé

L'étude des processus empiriques devient plus délicate dans $\mathbb{R}^d, d \geq 1$ par rapport au cas où les observations prennent leurs valeurs dans $\mathbb{R}$, par exemple à cause de la transformation de quantile, et la vitesse d'approximation du processus $\alpha_n(.)$ par un processus gaussien dans le principe d'invariance n'est pas optimale.

Un autre « type » de processus empirique nous sera trés utile pour l'étude des estimateurs à noyau. Il s'agit du processus empirique généralisé.

Nous présentons la définition correspondante dans le cas multidimensionnel, en nous limitant au cadre strict des applications que nous allons développer. Dans le cas des processus empirique généralisé,on peut étudié des variables aléatoires i.i.d. à valeurs dans un espace mesurable plus général que $\mathbb{R}$ (comme, par exemple, $\mathbb{R}^d$) repose sur la notion de processus empirique généralisé.

Soit $(\mathbf{X}_i, \mathbf{Y}_i)_{i\geq 1}$, une suite de couples i.i.d. de variables aléatoires, à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^q$,respectivement, avec $q, d \geq 1$. Soit par ailleurs $\mathcal{L}$, un ensemble de fonctions boréliennes réelles, définies sur$\mathbb{R}^{d+q}$.Pour tout $n \geq 1$, définissons le processus empirique d'ordre n, associé à $(\mathbf{X}_i, \mathbf{Y}_i)_{1\leq i\leq n}$ et indexé par $\mathcal{L}$,par :

$$\Phi_n(g) = \sqrt{n}\sum_{i=1}^{n}\{g(\mathbf{X}_i, \mathbf{Y}_i - \mathbb{E}(g(\mathbf{X}_i, \mathbf{Y}_i))\}, pour l \in \mathcal{L}$$

Les ensembles $\mathcal{L}$ correspondants sont dits former des classes de Glivenko-Cantelli (resp., de [41]). Dans le cas du théoréme de Glivenko-Cantelli, on exprime ainsi, par définition, que

$$\sqrt{n}\sup_{l\in\mathcal{L}}|\Phi_n(l)| \to 0$$

Dans des perspectives d'application à la statistique non paramétrique, ont été apportées dans les travaux successifs de [47], [30], pour des familles $\mathcal{L} = \mathcal{L}_n$ de fonctions appropriées, dépendant de $n \geq 1$, et présentant toutes la propriéeté de posséder un nombre de recouvrement polynomial.

L'étude du processus $\Phi_n(l) : l \in \mathcal{L}$, pour des choix appropriées de $\mathcal{L} = \mathcal{L}_n$, [47], [30] (voir également [42]), ont obtenu des résultats portant sur la consistence uniforme d'estimateurs non paramétriques de la densitée f_x de $\mathbf{X}$, ainsi que d'estimateurs de la la fonction de régression généralisée, $\mathbb{E}(\psi(\mathbf{Y})|\mathbf{X} = x)$, de $\psi(\mathbf{Y})$ sachant $\mathbf{X} = x$. Ici, ψ désigne une fonction spécifiée, mesurable à valeurs réelles, vérifiant des hypothéses additionnelles convenables.

1.1.4 Quelques résultats mathématiques

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 [31] et [20] et donne une version uniforme sur $\mathbb{R}$ de cette convergence.

Théorème 1.1. Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires indépendantes de même loi de probabilité, de fonction de répartition F. Et soit F_n la fonction de répartition empirique définie par

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

Alors, $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 p.s.$
Démonstration. voir [19] (Théorème 20.6)
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 $\{\mathbb{I}_{]-\infty,t]}, t \in \mathbb{R}\}$ est une classe GC universelle.

Les espaces $(\mathcal{C}[0,1],\mathbb{R})$ et $(\mathcal{D}[0,1],\mathbb{R})$

Notons $(\mathcal{C}[0,1], E)$ l'espace des fonctions continues définies sur l'intervalle [0,1] à valeurs dans l'espace E et $(\mathcal{D}[0,1], E)$ l'espace des fonctions continues à droite et limites à gauche (càdlàg) définies sur [0,1] à valeurs dans E.
Soit M une famille de fonctions réelles définies sur un ensemble d'indices T. Pour $k \geq 1$ et $t = (t_1, ..., t_k) \in T_k$, on note $\pi_t = \pi_{t_1,...t_n}$ l'application projection de M dans $\mathbb{R}^k$ définie par

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

Soit $(X_t, t \in T)$ 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.
Par la suite, nous nous restreignons au cas où 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.

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 $(D, \|.\|)$. 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, [24] a proposé la métrique définie par :

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

où Λ est l'ensemble des bijections continues et strictement croissantes de [0,1] dans luimême 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 : $x_n o\lambda_n \to x$, et $\lambda_n \to I$ 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.

La convergence faible des processus

Définition Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. réelles. On dit que (X_n) converge en loi vers une variable X, et on note $X_n \xrightarrow{\mathcal{L}} X$, 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 :

$$\int_{\mathbb{R}} f dP_n \to \int_{\mathbb{R}} f dP$$

L'avantage de cette définition c'est de pouvoir être généralisée à un espace métrique quelconque.
On suppose que (S,d) désigne un espace métrique muni de sa tribu borélienne $\mathcal{S}$ (la tribu engendrée par les ouverts de S).
Théorème 1.4 (Portmanteau). Soient $(P_n, n \geq 1)$, P des mesures de probabilité sur (S,$\mathcal{S}$). Les conditions suivantes sont équivalentes :
1. $P_n \Rightarrow P$;
2. $\int f dP_n \to \int f dP$ f dP pour toute fonction réelle f bornée et uniformément continue ;
3. $\limsup P_n(F) \leq P(F)$ pour tout fermé F ;
4. $\liminf P_n(O) \geq P(O)$ pour tout ouvert O ;
5. $\lim P_n(A) = P(A)$ pour tout A $\in \mathcal{S}$ tel que P(Fr A) = 0 (Fr A désigne la frontière de l'ensemble A).
Démonstration. cf. [18] (Théorème 2.1)
Si l'espace S n'est pas séparable, la tribu $\mathcal{S}_0$ engendrée par les boules ouvertes peut être plus petite que la tribu borélienne $\mathcal{S}$. Dudley ([6], [33]) a introduit une théorie de la convergence faible qui utilise uniquement les ensembles de $\mathcal{S}_0$ et les fonctions $\mathcal{S}_0$- mesurables.

Critère de convergence faible

Un résultat important de la convergence faible, est que si $X_n \xrightarrow{\mathcal{L}} X$ alors pour toute fonction continue g on a $g(X_n) \to g(X)$.
Dans l'espace $\mathcal{C}$ la convergence faible de la loi de X_n entraîne la convergence faible des lois fini-dimensionnelles (car les applications projections sont continues).
Dans l'éspace $\mathcal{D}$ muni de la topologie de Skorokhod, les applications projections ne sont pas toutes continues.
L'inverse n'est pas vrai en général, considérons par exemple la suite $(P_n)_{n\geq 0}$ de mesures de Dirac au points $x_n \in \mathcal{C}$ définis pour $n \geq 1$ par :

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

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 : $P_0(FrB) = 1$ (car P_0 est concentrée en 0) mais $P_n(B) = 0$ pour tout $n \geq 0$ (car $\|x_n - 0\| = 1$). Cependant, pour tout k > 0 et tout $(t_1, t_2, ..., t_k) \in [0,1]^k$, le vecteur $(x_n(t_1), x_n(t_2), ..., x_n(t_k))$ converge vers le vecteur nul de $\mathbb{R}^k$ (ce qui implique la convergence des lois fini-dimensionnelles).

Définition 1.3. Considérons une famille $\prod$ de mesures de probabilité sur (S,$\mathcal{S}$). On dit que $\prod$ est relativement compacte si toute suite d'éléments de $\prod$ admet une sous suite qui converge faiblement vers une mesure de probabilité (qui n'appartient pas nécessairement à $\prod$)
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 $(P_{n'_k})$ qui converge faiblement vers une mesure de probabilité Q. Si de plus les lois fini-dimensionnelles de P_n convergent vers les lois finidimensionnelles de P, alors nécessairement P = Q. En effet, la convergence de $P_{n'_k}$ 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 P = 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 $P_n \Rightarrow P$ 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 dune suite de mesures de probabilité définies sur un espace de fonction, il suffit de montrer la convergence des lois finidimensionnelles et la compacité relative.
Définition 1.4. Soit (S,d) un espace métrique et $\mathcal{S}$ sa tribu borélienne.
Une mesure de probabilité P définie sur $\mathcal{S}$ est dite tendue si :

$$\forall \varepsilon > 0, \exists K \subset S \, compact : P(K) \geq 1 - \varepsilon.$$

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

$$\forall \varepsilon > 0, \exists K \subset S \, compact : \inf_{P \in \prod} P(K) \geq 1 - \varepsilon.$$

Remarquons que si l'espace S est séparable, alors toutes les mesures de probabilité définies sur $\mathcal{S}$ sont tendues. On est maintenant en mesure d'énoncer le résultat suivant.

Théorème 1.5 ([44]). Soit (S, d) un espace métrique et $\mathcal{S}$ sa tribu borélienne, et soit $\prod$ une famille de mesures de probabilité définies sur $\mathcal{S}$.
- Si $\prod$ est tendue, alors elle est relativement compacte.
- Si S est séparable et complet, et si $\prod$ est relativement compacte, alors elle est tendue.

Démonstration. voir [24] (théorèmes 6.1 et 6.2 p. 37)

Représentation de Skorokhod

On considére une suite (F_n) de fonctions de répartition qui converge faiblement vers une fonction de répartition F, et ε une v.a. de loi uniforme sur [0, 1], et soit la suite (X_n) définie pour tout n par $X_n = F_n^{-1}(\varepsilon)$ et $X = F^{-1}(\varepsilon)$. Il est évident (d'après le Théorème de Glivenko-Cantelli)que X_n converge en loi vers X. De plus, la convergence faible de (F_n) entraîne que F_n^{-1} converge

vers F^{-1} 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 [12], 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 (S, d,$\mathcal{S}$) un espace Polonais(s'il est séparable et complet), et soit $(X_n)_n \geq 0$ une suite d'éléments aléatoires de S de lois respectives $(P_n)_n \geq 0$ tels que $X_n \Rightarrow X_0$.
Alors il existe un espace de probabilité $(\Omega, \mathcal{A}, P)$ et une suite (X_n) d'applications mesurables de $(\Omega, \mathcal{A}, P)$ dans $(S, \mathcal{S})$ 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 :

$$d(X_n, X_0) \rightarrow 0\, p.s.$$

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

Théorèmes limite

Le théorème de la limite centrale donne, pour tout $t \in \mathbb{R}$, une convergence en loi de $\sqrt{n}(F_n(t) - F(t))$ vers une v.a. de loi normale centrée et de variance F(t)[1 - F(t)]. Plus généralement,on peut montrer que pour tous $t_1, ..., t_k \in \mathbb{R}$, le vecteur aléatoire $\sqrt{n}(F_n(t_1) - F(t1), ..., F_n(t_k) - F(t_k))$ converge en loi vers un vecteur Gaussien centré $(G(t_1), ..., G(t_1))$ avec :

$$Cov(G(s), G(t)) = F(min(s,t)) - F(s)F(t)\, pour\, s, t \in \{t_1, ..., t_k\}$$

[4] 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", $\sqrt{n}(F_n(x) - F(x))$ converge faiblement vers un processus Gaussien.
Théorème 1.7. Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires i.i.d. de fonction de répartition F, et $Y_n = \sqrt{n}(F_n - F)$ le processus empirique associé.
Alors : $Y_n \Rightarrow Y$ où Y est un processus Gaussien centré de fonction de covariance E(Y(s)Y(t)) =F(s)(1 - F(t)) pour $s \leq t$.

La loi du logarithme itéré

La loi du logarithme itéré (ou LIL pour "Lawof 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.
Théorème 1. (Loi du logarithme itéré). Soient (X_n) une suite de v.a i.i.d. centrées de variance 1, et $S_n = \sum_{i=1}^{n} X_i$ alors :

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{nlog_2 n}}$$

où log_2n = log log n.
Ce résultat, dû à [44] a été d'abord prouvé pour des v.a de Bernoulli par [8], puis par [9] pour des v.a. de loi normale.

1).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 $U_1, ..., U_n$ des v.a. i.i.d. de loi $\mathcal{U}_{[0,1]}$ et soient la fonction de répartition empirique $U(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$ et le processus empirique $\alpha_n(t) = \sqrt{n}(U(t) - t)$ associés.
Théorème 2. (Smirnov[25]). Soit $b_n = \sqrt{2log_2 n}$. Alors :

$$\limsup_{n\to\infty} \frac{\|\alpha_n\|}{b_n} = \limsup_{n\to\infty} \frac{\|(U_n - I)\|}{\sqrt{nb_n}} = \frac{1}{2} p.s$$

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

2).Lois fonctionnelles du logarithme itéré
Lois fonctionnelles du logarithme itéré de [27] est un résultat très important, 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.
Définition 2.1. Soient $(X_n)_{n\geq 0}$ une suite d'éléments aléatoires dun 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 $P(\Omega_0) = 1$ tel que pour tout $w \in \Omega_0$:
1. Toute suite $n^{'}$ de nombres entiers admet une sous suite $n^{"}$ telle que $X_{n^{"}}(w)$ converge dans (S,d) ;
2. Toute les valeurs d'adhérence de $X_n(w)$ appartiennent à H ;
3. Pour tout $h \in H$, il existe une suite $n^{'} = n_{h,w}$ telle que $X_{n'}(w)$ 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 :
$\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1 \leq i \leq n}$ intervalles disjoints de I :

$$\sum_{i=1}^{n} (y_i - x_i) < \delta \Rightarrow \sum_{i=1}^{n} |h(y_i) - h(x_i)| < \varepsilon$$

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 $\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1 \leq i \leq n}$ intervalles disjoints de I :

$$\sum_{i=1}^{n} \mu(]x_i, y_i[) < \delta \Rightarrow \sum_{i=1}^{n} |h(y_i) - h(x_i)| < \varepsilon$$

Théorème 3. Soit $b_n = \sqrt{2log_2 n}$. Alors la suite $\{\frac{\alpha_n}{b_n}\}$ est presque sûrement relativement compacte dans B([0,1]) avec pour ensemble limite l'ensemble H ci-dessus.

Théorème 4. Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires i.i.d. ayant une fonction de répartition F définie et continue sur un intervalle [a,b], et soit an le processus empirique associé. Alors la suite $\{\frac{a_n}{b_n}\}$ est presque sûrement relativement compacte dans B([0,1]) avec pour ensemble limite l'ensemble H_F des fonctions f définies sur [a,b] et vérifiant :
- f(a) = f(b) = 0,
- f est absolument continue par rapport à F,
- $\int_a^b (\frac{df}{dF})^2 \leq 1$ où $\frac{df}{dF}$ est la dérivée de f par rapport à F.

1.2 Rappel sur l'analyse de survie

1.2.1 Introduction

L'analyse de survie est une spécialité importante des biostatistiques qui consiste à étudier des durées. L'étude de ces durées peut, par exemple, permettre de comparer les temps jusqu'à la guérison, la rechute ou encore le décès de différents patients. Dans un essai clinique, l'efficacité d'un nouveau traitement peut être déterminée en évaluant si la durée de vie moyenne des patients a été rallongée après la prise du traitement. Dans les études épidémiologiques, l'analyse de survie permet d'évaluer si un ou plusieurs facteurs de risque sont liés à la durée de vie. Ceci peut amener à limiter l'exposition à des facteurs de risque environnementaux, lorsque cela est possible, afin

d'allonger la durée de vie. On désigne communément l'évènement auquel on s'intéresse par le terme générique de décès et la durée est appelée durée de vie. L'analyse de survie peut également permettre d'établir des prédictions sur les temps de survie des patients. Ceci peut se faire de manière non paramétrique, par exemple avec l'estimateur de Kaplan-Meier de la survie [7], ou de manière paramétrique ou semi-paramétrique par l'utilisation de modèles pronostiques. Ces prédictions ont une importance capitale dans le domaine de la santé. Ils permettent aux patients d'être informés sur leur état de santé futur (pronostic du médecin) à partir de leur état de santé actuel (diagnostic établi par le médecin). Parfois, le traitement du patient peut être adapté en fonction de ces prédictions. Par exemple, en oncologie, les effets secondaires de certains traitements sont lourds, amenant à réserver ces thérapies aux patients dont la probabilité de survivre après un temps donné est faible.
La particularité de l'analyse de survie, hormis le fait d'étudier des variables aléatoires positives, réside dans la présence de censure : certaines durées ne sont pas entièrement observées. Par exemple, lors d'une étude, si un patient ne se rend plus aux rendez-vous fixés avec le médecin, il sera désigné comme "perdu de vue". Après la dernière visite à laquelle il s'est rendu, aucune information supplémentaire n'est disponible pour ce patient. Aussi, l'étude peut se terminer alors que des patients n'ont pas encore subi l'évènement d'intérêt. Ces patients sont appelés les "exclus vivants"(aprés une date d'observation le patient est décé). Dans ces deux situations, on n'observe pas le temps de survie du patient, la seule information disponible est que le patient a survécu jusqu'à une date connue (la dernière visite ou la date de fin de l'étude). Les estimateurs usuels de la statistique ne peuvent pas être utilisés en présence de censure, comme nous allons le voir dans la Section 1.3 avec l'exemple de l'estimateur de la survie empirique qui est introduit par l'estimateur de Kaplan-Meier.

1.2.2 Cas des données censurées

Une caractéristique importante de données de survie est la présence de données censurées. Cette caractéristique, source de dificulté, a nécessité le développement de techniques alternatives à l'inférence usuelle.
Les données censurées sont des observations ne correspondant pas à de vraies valeurs de la variable d'intérêt mais seulement une estimation, inferieure ou

superieure, c'est-a-dire une information grossiére, du type $X \geq c$ ou $X \leq c$.

La particularité de ces données, c'est qu'à la fin de la période d'observation , l'événement d'intérêt (correspondant à l'état du patient en deux éventualités (vivant ou décédé)ne sera probablement pas survenu pour tous les patients. Le mot "censure" est ici d'usage statistique ; il n'a pas grand chose a voir avec une commission de controle, qui aurait decide de tronquer les donnees pour que le public n'en ait pas connaissance, encore que, comme nous le verrons, ceci se produise dans certains exemples. Il aurait certainement ete preferable de parler de "donnees tronquees", mais le mot "censure" est consacre par l'usage, et c'est lui que nous emploierons.

On peut ranger les situations ou l'on trouve des données censurées en deux classes :

- Celles ou les données reélles existent, mais n'ont pas ete utilisees ;

- Celles ou les données reélles n'existent pas.

Cependant, on dispose tout de même d'une information partielle permettant de fixer une borne inférieure (censure àdroite) ou une borne supérieure (censure à gauche). Les raisons de cette censure peuvent etre le fait que le patient soit toujours vivant ou non malade à la fin de l'étude ou qu'il se soit retiré de l'étude pour des raisons personnelles (immigration, mutation professionnelle ;.....etc.). La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour l'individu i ; considérons :

- son temps de survie $\mathbf{X}_i$

- son temps de censure $\mathbf{C}_i$

- la durée réellement observée $\mathbf{T}_i$.

Les type des données censurées :

Il existe trois catégories de censures qu'on nomme censure à droite, censure à gauche et censure par intervalle (lorsqu'on connait la borne supérieure et la borne inférieure d'un évènement). Il existe différents types de censures à l'intérieur de ces trois catégories :

A)- Censure de type I :

Etant donné un nombre positif fixé c et un n-échantillon $\mathbf{X}_1.....\mathbf{X}_n$ on observe :

$$\mathbf{T}_i = \mathbf{X}_i \wedge \mathbf{X}(r) et \delta_i = \mathbb{I}_{\{\mathbf{X}_i = \mathbf{T}_i\}}$$

Telque :$\mathbf{X}_i \wedge \mathbf{X}(r)$ représente le minimum $(\mathbf{X}_i, \mathbf{X}(r))$ Le temps de censure est fixé par le chercheur comme étant la fin de l'étude.

B)- Censure de type II :

Soit i tel qu'à chaque i=1,...,n est associé un couple de variables aléatoires non nul $\mathbf{X}_i, \mathbf{C}_i$ ou seul le minimum est observé c'est-à-dire qu'on observe :

$$\mathbf{T}_i = \mathbf{X}_i \wedge \mathbf{C}_i et \delta_i = \mathbb{I}_{\{\mathbf{X}_i = \mathbf{C}_i\}}$$

Où δ_i est un indicateur de censure tel que :

$$\delta_i = \begin{cases} 1 & \text{si } \mathbf{X}_i \leq \mathbf{C}_i\,; \\ 0 & \text{si } \mathbf{X}_i > \mathbf{C}_i. \end{cases}$$

Où $\mathbf{X}_i$ est l'instant de l'événement. $\mathbf{C}_i$ est l'instant de censure.

C)- Censure de type III :
Etant donné un entier positif r fixé, un n-échantillon $\mathbf{X}_1$.....$\mathbf{X}_i$ d'une variable aléatoire positive $\mathbf{X}$ et les statistiques d'ordre $\mathbf{X}(1)$,.....,$\mathbf{X}(n)$ on observe :

$$\mathbf{T}_i = \mathbf{X}_i \wedge \mathbf{X}(r) et \delta_i = \mathbb{I}_{\{\mathbf{X}_i = \mathbf{T}_i\}}$$

autrement dit, ce genre de censure se caractérise par le fait que l'étude cesse aussitôt qu'a eu lieu un nombre d'événements prédéterminés par lexpérimentateur.

Censure à droite :

Une durée de survie est dite censurée à droite si l'individu n'a pas connu l'événement d'intérêt à sa dernière visite. La censure à droite est l'exemple le plus fréquent d'observation incomplète en analyse de survie, Formellement, la durée de survie d'un évènement est définie par le couple $(\mathbf{X}; \delta)$ où :

$$\mathbf{X} = inf(\mathbf{T}; \mathbf{C})$$

et

$$\delta = \begin{cases} 1 & \text{si } \mathbf{T} \leq \mathbf{C}\,; \\ 0 & \text{si } \mathbf{T} > \mathbf{C}. \end{cases}$$

Censure à gauche :

Une durée de survie est dite censurée à gauche si l'individu a déjà connu l'évènement d'intérêt avant l'entrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple $(\mathbf{X}; \delta)$ où :

$$\mathbf{X} = max(\mathbf{T}; \mathbf{C})$$

et

$$\delta = \begin{cases} 1 & \text{si } \mathbf{T} > \mathbf{C}\,; \\ 0 & \text{si } \mathbf{T} \leq \mathbf{C}. \end{cases}$$

Remarque : Trés peu de travaux s'intéressent à la seule censure à gauche car beaucoup moins fréquente et constitue un phénoméne symétrique à celui

de la censure à droite. Certains auteurs ont proposé de renverser l'échelle de temps.
Dans un même échantillon peuvent être présentes des données censurées à droite et d'autres censurées gauche, comme c'est le cas dans ce qui suit.

Censure par intervalle :

Une situation plus générale de la censure se produit lorsque la durée de survie n'est pas connue mais on sait seulement qu'il appartient à un certain intervalle.
Dans ce cas 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 aprés que l'évènement d'intérêt se soit produit.
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 ;+∞[et [0 ; a] respectivement, ce qui permet de considérer ce modèle comme étant plus générique.

Censure mixte :

Il y a censure mixte lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Dans le modèle II décrit dans l'article de [46], au lieu d'observer un échantillon de la variable d'intérêt T, on observe un échantillon du couple (Z,A) avec Z = min(max(T,L),R) et

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

où L et R sont des variables de censure et A est l'indicateur de censure. Un exemple de ce modèle est donné par un système formé par trois composants, dont deux sont placés en parallèle (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en série avec ce système en parallèle. L'analyse de survie a connu un développement important dans la seconde moitié du vingtiéme siécle aprés que [7] aient introduit leur célébre estimateur de la fonction de survie pour des données censurées à droite. Estimateur qui généralise le complément à un de la fonction de répartition empirique et que nous rappelons ci dessous.

1.3 Le processus empirique de Kaplan-Meier

En presence de censure, les durees complétes $T_1, ..., T_n$ ne sont pas toutes observees. Ainsi, l'estimateur usuel de la fonction de repartition empirique de T,

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T_i \leq t}, 0 \leq t \leq T,$$

ne peut pas être evalue. Sans données censurées, cet estimateur sans biais de $F(t) = P(T \leq t)$ peut être utilise. Une solution pour estimer F pourrait etre de considerer les temps censures comme des decés et d'utiliser l'estimateur

$$1 - \frac{1}{n}\sum_{i=1}^{n} Y_i(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq t}, 0 \leq t \leq T.$$

Cet estimateur de la fonction de repartition de T est biaise si l'echantillon comporte de la censure. Son esperance au temps t vaut $P(T \leq t, C \leq t)$, menant a une sous-estimation de $P(T \leq t)$. Kaplan et Meier (1958) ont introduit un estimateur convergent de F qui tient compte de la censure.

1.3.1 Definition

L'estimateur de Kaplan-Meier de F est encore appelé Produit limite car il s'obtient comme la limite d'un produit.
L'estimateur de Kaplan-Meier de F, noté $\widehat{F}$, est défini par :

$$\widehat{\mathbf{F}}_n(t) = 1 - \prod_{i:Z_i \geq t} \left(\frac{\mathbf{N}_n(Z_i) - 1}{\mathbf{N}_n(Z_i)} \right) pour t \in \mathbb{R}$$

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

Pour les valeurs t supérieurs à la plus grande observation t_{max}, cet estimateur n'est pas bien défini. Si t_{max} correspond à un véritable temps de mort, alors l'estimateur s'annule après ce point. Si t_{max} est censuré alors $\widehat{F}(t)$ ne tendra pas vers zéro à l'infini alors que c'est un estimateur d'une fonction de survie.

Plusieurs suggestions sont proposées pour traiter cette ambiguïté.
Efron(1967) à proposé d'estimer F(t) par 0 pour $t > t_{max}$. Gill(1980) a, quant à lui, proposé d'estimer F(t) par $\widehat{F}_n(t_{max})$ pour t > tmax. Ces deux estimateurs ont les mêmes propriétés asymptotiques et convergent vers la

vraie fonction de survie.
On définit le processus empirique de Kaplan-Meier par :

$$\widehat{a}_n(t) = \sqrt{n}(\widehat{F}_n(t) - F(t)).$$

On peut ensuite, comme pour le processus empirique usuel, définir les incréments du processus empirique de Kaplan-Meier par :

$$\widehat{\eta}_n(h, t; s) = \widehat{a}_n(t + hs) - \widehat{a}_n(t),$$

pour $h > 0, 0 \leq s \leq 1$ et $t \in \mathbb{R}$. [2] et [47] ont obtenu des lois limites fonctionnelles associées à η_n. Nous établirons une extension de leurs résultats dans le chapitre 2. Dans ce chapitre, nous généralisons le résultat, obtenu par [42] pour le processus empirique usuel, aux incréments du processus de Kaplan-Meier $\widehat{\eta}_n$. Nous en déduisons des lois uniformes du logarithme pour des estimateurs de la densité.

Cet estimateur a des propriétés assez similaires à la fonction de répartition empirique, par exemple la convergence uniforme presque sûre ([34] ; [37], la normalité asymptotique ([40] ; [48]), et la loi du logarithme itéré ([38]). Ceci justifie que l'on s'intéresse à généraliser la théorie des processus empiriques au cas des données censurées.

1.3.2 Quelques propriétés sur l'estimateur de Kaplan - Meier

En analyse de survie, $\widehat{F}_n$ joue pour les données incomplétes le même rôle que la fonction de répartition empirique pour les données classique.

a)Biais et convergence

L'estimateur de Kaplan - Meier est légèrement biaisé : en général,

$$\mathbb{E}(\widehat{F}_n[t]) < F(t);$$

où $\mathbb{E}$ désigne l'espérance.
Il est en revanche convergent (consistent estimator) :

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|\widehat{F}_n[t] - F[t]| \geq \epsilon) = 0$$

Il est donc asymptotiquement non biaisé :

$$\lim_{n\to\infty} \mathbb{E}(\widehat{F}_n[t]) = F(t)$$

b)Auto-cohérence

En l'absence de censure, un estimateur de S(t) est

$$\widetilde{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(t_i > t)$$

En présence de censure, on peut encore écrire

$$\widetilde{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} (\delta_i \mathbb{I}[t_i > t] + [1-\delta_i]\mathbb{I}[t_i > t]),$$

mais la valeur de $\mathbb{I}(t_i > t)$ n'est pas connue pour les données censurées. Si un estimateur . $S_n(t)$ de S(t) est connu, on peut estimer l'espérance de $\mathbb{I}(t_i > t)$ sachant que $\delta_i = 0$ et $t_i > s_i$: on a

$$\mathbb{E}(\mathbb{I}[t_i > t]|\delta_i = 0 \quad et \quad t_i > s_i) = \frac{\mathbb{P}(T_i > t)}{\mathbb{P}(T_i > s_i)},$$

donc

$$\widetilde{E}(\mathbb{I}[t_i > t]|\delta_i = 0 \quad et \quad t_i > s_i) = \frac{\widetilde{S}_n(t)}{\widetilde{S}_n(s_i)}.$$

L'estimateur de Kaplan-Meier présente la propriété d'être auto-cohérent, c.-à-d. que

$$\widehat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\delta_i \mathbb{I}[t_i > t] + [1-\delta_i]\frac{\widehat{S}_n(t)}{\widehat{S}_n(s_i)} \right)$$

Si l'on part d'une fonction de survie arbitraire (mais compatible avec les contraintes sur une fonction de survie)$\widetilde{S}_n^{(0)}$, on peut calculer itérativement une estimation $\widetilde{S}_n^{(k)}$ par

$$\widetilde{S}_n^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\delta_i \mathbb{I}[t_i > t] + [1-\delta_i]\frac{\widetilde{S}_n^{(k-1)}(t)}{\widetilde{S}_n^{(k-1)}(s_i)} \right)$$

. et l'on a $\lim_{k\to\infty} \widetilde{S}_n^{(k)} = \widehat{S}_n$.

c)Limitations de l'estimateur de Kaplan - Meier

L'estimateur de Kaplan - Meier est discontinu. Pour certaines applications, il est nécessaire de le lisser en le convoluant avec un noyau.

Il ne prend pas en compte les incertitudes sur les valeurs t_i et les censures s_i. Elles sont négligeables en statistiques médicales (la date de décès ou de sortie

de l'échantillon d'un patient est connue précisément), mais souvent cruciales en astrophysique.

Ces incertitudes s'ajoutant à la dispersion intrinsèque de la loi de distribution (qu'on peut caractériser par l'écart-type autour de la moyenne), la dispersion apparente estimée à partir de l'estimateur de Kaplan - Meier surestime la dispersion intrinsèque.

Par ailleurs, le seuil de censure est souvent arbitraire en astrophysique : on peut ainsi considérer qu'une source n'est pas détectée si son flux est inférieur à 2, 3 ou 5 fois le bruit .

Des simulations, de type bootstrap par exemple, peuvent permettre de modéliser ces phénomènes et de corriger leurs effets.

1.3.3 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.
En posant $F_n(z) = 0$ pour $z > Z_{(n)}$. Pour toute fonction de répartition L, on note par $T_L = max\{t : L(t) < 1\}$ le point terminal du support de L.
Théorème : ([47]). On suppose que F et G sont continues, et que $T_F < T_G$. Alors,

$$\mathbb{P}\left(\sup_{-\infty<u<+\infty} |F_n(u) - F(u)| = 0\left(\sqrt{\frac{\log_2 n}{n}}\right)\right) = 1$$

. La condition $T_F < T_G$ pouvant paraître restrictive, on peut citer le théorème autrement :

Corollaire : ([38]). On suppose que F et G sont continues, et on considère T tel que G(T) > 0. Alors,

$$\mathbb{P}\left(\sup_{-\infty<u\leq T^*} |F_n(u) - F(u)| = 0\left(\sqrt{\frac{\log_2 n}{n}}\right)\right) = 1,$$

où $T^* = \min\{T, T_F\}$.

L'estimateur de Kaplan-Meier est un estimateur qui permet de tenir compte des censures, il est consistant et asymptotiquement gaussien mais présente l'inconvénient d'être biaisé et par nature discontinu.

Chapitre 2

Loi fonctionnelle uniforme du logarithme pour les incréments du processus empirique censuré

2.1 Introduction

Dans ce chapitre, nous allons étudier un résultat intéressant dans le cadre d'un modèle de censure à droite. Nous présentons une loi limite fonctionnelle uniforme pour les incréments du processus empirique de Kaplan-Meier [7]. Ces lois constituent une extension du théorème de Deheuvels et Einmahl [47] ; en effet, nous montrons que leur résultat principal (qui constitue une version, adaptée à la convergence uniforme sur un intervalle, d'un théorème de Deheuvels et Einmahl [2] écrit dans le cadre de la convergence ponctuelle des estimateurs) reste valable uniformément en fonction du choix de la fenêtre $h \in [h'_n, h"_n]$, où h'_n et $h"_n$ vérifient tous deux les conditions (H.1- 2-3) rappelées ci-dessous), avec $0 < h'_n \leq h"_n < \infty$.
Nous déduisons de ce théorème des résultats de convergence uniforme, relativement à la localisation et à la taille de fenêtre h, pour certaines fonctionnelles du processus empirique de Kaplan-Meier.

On considère Y,Y_1, Y_2, . . . une suite de variables aléatoires positives i.i.d.

représentant les durées de vie.

Nous supposerons disposer de copies aléatoires, indépendantes et de même loi, (Y_i, C_i), i = 1, 2, . . . , du couple aléatoire générique (Y,C). Ici, Y désigne la variable aléatoire d'intérêt, supposée positive, et représentant une durée de survie, et C la variable aléatoire, supposée positive elle aussi, représentant un temps de censure. Nous supposerons de plus disposer, sous l'hypothése générale d'indépendance entre Y et C, de l'échantillon observé (Z_i, δ_i), i = 1, . . . , n, $n \geq 1$, où

$$Z_i := Y_i \wedge C_i \quad et \quad \delta_i := \mathbb{I}_{\{Y_i \leq C_i\}},$$

avec Z ayant pour fonction de répartition

$$H(.) := P(T \leq .) = 1 - (1 - F(.))(1 - G(.)),$$

et dans ce cas, pour le i^{me} patient, Y_i désigne la durée de vie du patient (nonobservée), et C_i, sa durée d'hospitalisation (observée).
Ici, nous désignons par

$$F(x) = \mathbb{P}(Y \leq x), G(x) = \mathbb{P}(C \leq x) et H(x) = \mathbb{P}(Z \leq x),$$

les versions continues à droite des fonctions de répartition de X, Y et Z. Notons que, dans la suite du présent exposé, F, G et H ne seront pas nécessairement supposées continues.

En présence du censure, la fonction de répartition empirique de la variable X n'est plus valable car elle dépend de variables aléatoires parmi Y,Y_1, Y_2, . . . qui ne sont pas observées. Afin d'estimer la loi de Y, il a été donc nécessaire de construire un estimateur de la fonction de répartition en présence de données censurées, qui puisse avoir des propriétés analogues à celle de la fonction de répartition empirique classique. En 1958, Kaplan et Meier ont introduit les estimateurs non paramétiques du maximum de vraisemblance de F(.) et G(.) .

Nous utilisons les mêmes notations que précédemment : Soient $X_1, ..., X_n$ les durées d'intérêt, indépendantes et de fonction de répartition F, et indépendamment d'elles, soient $C_1, ..., C_n$ les durées de censure, indépendantes et de fonction de répartition G. Et on observe les couples $(Z_1, \delta_1), ..., (Z_n, \delta_n)$ avec $Z_i = min\{X_i, C_i\}$ et $\delta_i = \mathbb{I}\{X_i \leq C_i\}$. Pour toute fonction R, on note :

$$R_+(x) = lim_{\varepsilon \downarrow 0} R(x + \varepsilon) \quad et \quad R_-(x) = lim_{\varepsilon \downarrow 0} R(x - \varepsilon);$$

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

$$T_L = sup\{t : L(t) < 1\} \quad et \quad L_-(\infty) = lim_{x \to \infty} L(x).$$

On suppose que $F_-(\infty) = 1$, mais on accepte pour la distribution de Y que $G_-(\infty) = 1 - \mathbb{P}(Y = 1) \leq 1$. En particulier, si $\mathbb{P}(\mathrm{Y} = 1) = 1$, on revient aucasnoncensuré.

Posons $\Theta = min\{T_F, T_G\}$. On suppose que $\Theta > 0 (si \Theta = 0$, alors F_n est presque sûrement dégénérée).0. Pour $0 \leq z \leq \Theta$, les estimateurs de Kaplan-Meier (voir [7]) de F et de G sont définis par :

$$\widehat{F}_n(z) = 1 - \prod_{\substack{i:Z_i \leq z \\ 1 \leq i \leq n}} \left(\frac{N_n(Z_i) - 1}{N_n(Z_i)} \right)^{\delta_i},$$

$$\widehat{G}_n(z) = 1 - \prod_{\substack{i:Z_i \leq z \\ 1 \leq i \leq n}} \left(\frac{N_n(Z_i) - 1}{N_n(Z_i)} \right)^{1-\delta_i},$$

où $N_n(x) = \sum_{i=1}^n \mathbb{I}_{\{Z_i \geq x\}}$.
Par suite, on supposera que h > 0 est tel que $h'_n \leq h \leq h"_n$. Ici, $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$ sont deux suites de constantes réelles positives, vérifiant $0 < h'_n \leq h"_n < \infty$.

Nous supposerons que $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$ sont telles que, pour chacun des choix $h_n = h'_n$ et $h_n = h"_n$, la suite $\{h_n\}_{n\geq 1}$ vérifie les hypothéses (H.1-2-3) ci-dessous.

(H.1) $h_n \downarrow 0 \quad et \quad nh_n \uparrow \infty \quad lorsque \quad n \uparrow \infty$;
(H.2) $log(1/h_n)/loglogn \to \infty \quad lorsque \quad n \to \infty$;
(H.3) $nh_n/logn \to \infty \quad lorsque \quad n \to \infty$.
Pour définir un estimateur à noyau f_n de la densité de survie, on se donne une fenêtre h > 0 et on introduit un noyau K, fonction mesurable réelle de variable réelle, vérifiant les hypothéses suivantes :
(K1) K est une fonction à variation bornée sur $\mathbb{R}$.
(K2) Il existe une constante $0 < T < \infty$, telle que $K(u) = 0$ pour $|u| \leq T/2$.
(K3) $\int_{-\infty}^{+\infty} K(u)du = 1$.
Nous travaillerons sous les hypothéses de régularité suivantes, portant sur F et G. Soient des constantes a, a', b et b' telles que $0 < a' < a < b < b' < \Theta$. Soit, de plus, $H^{(1)}(b) = \mathbb{P}(Z \leq b, \delta = 1)$.
On suppose que
(F1) F(0)=G(0)=0 ;
(F2) F et G sont continues sur [a', b'] ;
(F3) f=(d/dx)F(x) existe, est continue et strictement positive sur [a', b'] ;
(F4) $h" \leq [(b' - b) \wedge (1 - H^{(1)}(b))]$, pour tout $n \geq 1$.
Pour tout $x \in \mathbb{R}$, l'estimateur à noyau $\widehat{f}_{n,h}$(x) de f(x) = (d/dx)F(x), associé au noyau K et à la fenêtre h (voir Deheuvels et Einmahl [2], Deheuvels et

Einmahlest [47]), est défini par :

$$\widehat{f}_{n,h}(x) = \int_{-\infty}^{+\infty} \frac{1}{h_n} K\left(\frac{t-x}{h_n}\right) d\widehat{F}_n(t) \qquad (1),$$

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.

Pour tout $x \in \mathbb{R}$, on pose

$$\widehat{\mathbb{E}}\widehat{f}_{n,h}(x) = \int_{-\infty}^{+\infty} \frac{1}{h_n} K\left(\frac{t-x}{h_n}\right) dF_n(t) \qquad (2).$$

Remarque : On remarque que dans le cas non censuré,$\widehat{\mathbb{E}}f_n(x) = \mathbb{E}f_n(x)$, où $\mathbb{E}$ désigne l'espéranceusuelle.
Cependant, $\widehat{\mathbb{E}}f_n(x)$ et $\mathbb{E}f_n(x)$ sont généralement distinctes dans le cas censuré strict (correspondant à $\mathbb{P}(C < T_F) > 0$).

L'objectf de Deheuvels et Einmahl [2] 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 [3] 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.
On note $\{\widehat{a}_n(t) : t \in \mathbb{R}\}$, le processus empirique de Kaplan-Meier basé sur les observations $\{X_i : 1 \leq i \leq n\}$. Pour $n \geq 1$ et $x \in \mathbb{R}$, il est défini par :

$$\widehat{a}_n(x) = \sqrt{n}(\widehat{F}_n(x) - F(x)). \qquad (3)$$

Ensuite,pour $0 \leq t \leq \Theta$. Nous introduisons ensuite les incréments de ce processus, en posant

$$\widehat{\eta}_n(h,t;s) = \widehat{a}_n(t+hs) - \widehat{a}_n(t), s \in \mathbb{R} \qquad (4),$$

pour $0 \leq t \leq \Theta, 0 \leq s \leq 1\, et\, 0 \leq h \leq \Theta - t$. Les définitions (1), (2) et (3), compte tenu des hypothéses (K1-2-3), permettent d'écrire

$$\begin{aligned}
\widehat{f}_{n,h}(x) - \widehat{\mathbb{E}}\widehat{f}_{n,h}(x) &= h^{-1} \int_{-T/2}^{T/2} K(u) d\{\widehat{F}_n(x+hu) - \widehat{F}_n(x) - F_n(x+hu) + F_n(x)\} \\
&= -h^{-1} \int_{-T/2}^{T/2} \{\widehat{F}_n(x+hu) - \widehat{F}_n(x) - F_n(x+hu) + F_n(x)\} dK(u) \\
&= -h^{-1} n^{-1/2} \int_{-T/2}^{T/2} \widehat{\eta}_n(h,x,u) dK(u).
\end{aligned}$$

On désigne par $\psi(.)$ une fonction spécifique qui est continue et positive sur [a',b']
On note $\psi_n(.)$ un estimateur de $\psi(.)$ tel que quand $n \to \infty$, on a

$(C1)$ $\sup_{a \leq x \leq b} |\psi_n(x)/\psi(x) - 1| \to 0$ presque sûrement.

Soit $(\mathcal{B}[0,1], \mathcal{U})$ et $(\mathcal{AC}[0,1], \mathcal{U})$ l'ensemble des fonctions g bornées et l'ensemble des fonctions g absolument continues sur [0,1], munis, tous deux, de la topologie uniforme $\mathcal{U}$,définie par la norme uniforme $\|g\| = \sup_{0 \leq t \leq 1} |\acute{g}(t)|$. Pour tout $g \in \mathcal{B}[0,1]$, nous posons

$$|g|_H = \begin{cases} \{\int_0^1 \acute{g}(t)^2 dt\}^{1/2} & \text{si } g \in \mathcal{AC}(0,1] \text{ et g(0)=0}\,; \\ \infty & \text{si non.} \end{cases}$$

Pour chaque $g \in \mathcal{AC}[0,1], \acute{g} = (d/dt)g.$ représente la dérivée de Lebesgue de g.
Posons les ensembles des fonctions suiventes :

$$\mathbb{S}_\omega = \{g \in \mathcal{AC}[0,1] : g(0) = 0, |g|_H^2 \leq \omega\};$$

Pour tout $w > 0$, notons que $\mathbb{S} := \mathbb{S}_1$ est l'ensemble de Strassen [26]. Par ailleurs $\mathbb{S}_w = w^{\frac{1}{2}}\mathbb{S}$.

2.2 Quelques résultats principaux

Aprés avoir introduit les définitions, notations et hypothèses nécessaires, nous pouvons maintenant faire une brève synthèse des principaux résultats de convergence pour les incréments du processus de Kaplan-Meier, définis.
Soit une suite de constantes positives $\{h_n\}_{n \geq 1}$ on définit, pour $x_0 \in (a,b)$ et $n \geq 1$, le sous-ensemble de fonctions $\widehat{\xi}_n$ de (ψ_n) définit par :

$$\mathbf{E}(\widehat{\xi}_n(\psi_n)) = \left\{ \frac{\widehat{\eta}_n(h, x_0; I)}{\sqrt{2h_n log_2 n}} \times \left(\psi_n(x_0) \times \frac{1 - G(t)}{f(x_0)} \right)^{\frac{1}{2}} \right\} \subseteq \mathcal{B}[0,1]$$

où, d'une manière générale $log_2(u) := log_+ log_+(u)$ et $log_+(u) = log(u \vee e)$, pour $u \in \mathbb{R}$. On définit également le sous-ensemble de fonctions.

$$\widehat{\mathcal{L}}_n(\psi_n) = \left\{ \frac{\widehat{\eta}_n(h,x;I)}{\sqrt{2h_n log(1/h_n)}} \times \left(\psi_n(x) \times \frac{1-G(t)}{f(x)} \right)^{\frac{1}{2}} : a \leq x \leq b \right\} \subseteq \mathcal{B}[0,1].$$

Deheuvels et Einmahl [47] ont montré que l'on peut obtenir des résultats similaires (mais avec des constantes de normalisation différentes), pour la convergence uniforme en $x \in [a,b]$ de $\widehat{f}_{n,h_n}(x)$.

Théorème 2.1 Soit $\{h'_n\}_{n\geq 1}$ une suite de constantes positives vérifiant les conditions (H1-H3), Sous les conditions (F1-F4) et (C1), la suite $(\widehat{\mathcal{L}}_n(\psi_n))_{n\geq 1}$ est presque sûrement relativement compacte dans B[0,1] (muni de la topologie de la convergence uniforme), avec comme ensemble limite l'ensemble $\mathbb{S}_M$, où $M = sup_{a\leq x\leq b}\psi(x)$.

Théorème 2.2 Soit $\{h'_n\}_{n\geq 1}$ une suite de constantes positives vérifiant les conditions (H1-H3). On suppose vérifiées les conditions (F1-F4), (K1-K3) et (C1). Alors, avec probabilité 1,

$$\lim_{n\to\infty} \left\{ \frac{nh_n}{2log(1/h_n)} \right\}^{1/2} \sup_{a\leq x\leq b} \pm(\widehat{f}_{n,h}(x) - \widehat{\mathbb{E}}\widehat{f}_{n,h}(x)) \left\{ \psi_n(x) \times \frac{1-G(t)}{f(x)} \right\}^{\frac{1}{2}}$$

$$= \sup_{a\leq x\leq b} \{\psi(x)\}^{1/2} \left\{ \int_{-\infty}^{\infty} K^2(u)du \right\}^{\frac{1}{2}}.$$

2.2.1 Nouveaux résultats

Dans cette partie, nous présentons des généralisations nouvelles des théorèmes 2.1 et 2.2. Soit $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$ désignent des suites de constantes, vérifiant les conditions (H.1-2-3) et telles que $0 < h'_n \leq h"_n < \infty, n. \geq 1$. Dans le même esprit, nous établissons un théorème de convergence pour l'estimateur du taux de mortalité (ou de panne).

Introduisons le sous-ensemble de fonctions de B[0, 1] dèfini par

$$\widehat{\mathcal{K}}_n(\psi_n) = \left\{ \frac{\widehat{\eta}_n(h,x;I)}{\sqrt{2h_n log(1/h_n)}} \times \left(\psi_n(x) \times \frac{1-G(t)}{f(x)} \right)^{\frac{1}{2}} : a \leq x \leq b, \right\} \subseteq \mathcal{B}[0,1].$$

Théorème 2.3 Soit $\{h'_n\}_{n\geq 1}$, $\{h"_n\}_{n\geq 1}$ deus suites de constantes positives vérifiant les conditions (H1-H3). avec $0 < h'_n \leq h"_n < \infty$ On suppose vérifiées les conditions (F1-F4), (K1-K3) et (C1). Alors, avec probabilité 1,

$$\lim_{n\to\infty} \sup_{h_n\in[h'_n,h"_n]} \left\{ \frac{nh_n}{2log(1/h_n)} \right\}^{1/2} \sup_{a\leq x\leq b} \pm(\widehat{f}_{n,h}(x) - \widehat{\mathbb{E}}\widehat{f}_{n,h}(x)) \left\{ \psi_n(x) \times \frac{1-G(t)}{f(x)} \right\}^{\frac{1}{2}}$$

$$= \sup_{a\leq x\leq b} \{\psi(x)\}^{1/2} \left\{ \int_{-\infty}^{\infty} K^2(u)du \right\}^{\frac{1}{2}}.$$

Pour définit le résultat suivant, nous nous plaçons sous les hypothèses de régularité (F1-F4). Celles-ci nous permettent de définir le taux de mortalité, ou taux de panne d'instantané λ associé à F. Ce dernier est défini, pour tout $x < T_F$, par

$$\lambda(x) = \frac{f(x)}{1 - F(x)}.$$

L'estimateur non-paramétrique $\widehat{\lambda}_{n,h}$ de λ, associé à l'estimateur de Kaplan-Meier, est défini par

$$\widehat{\lambda}_{n,h}(x) = \frac{\widehat{f}_{n,h}(x)}{1 - \widehat{F}_{n,h}(x)}.$$

où $\widehat{f}_{n,h}$ et $\widehat{F}_n$ sont les estimateurs respectifs de f et F.
En suite, nous établirons le théorème suivant pour décrire la convergence de $\widehat{\lambda}_{n,h}(x)$.
Théorème 2.4 Soit $\{h'_n\}_{n\geq 1}$, $\{h"_n\}_{n\geq 1}$ deus suites de constantes positives vérifiant les conditions (H1-H3). avec $0 < h'_n \leq h"_n < \infty$ On suppose vérifiées les conditions (F1-F4), (K1-K3) et (C1). Alors, avec probabilité 1,

$$\lim_{n\to\infty} \sup_{h_n\in[h'_n,h"_n]} \left\{ \frac{nh_n}{2log(1/h_n)} \right\}^{1/2} \sup_{a\leq x\leq b} \pm \left(\widehat{\lambda}_{n,h}(x) - \frac{\widehat{\mathbb{E}}\widehat{f}_{n,h}(x)}{1 - F(x)} \right)$$

$$\times \left\{ \psi_n(x) \times \frac{1 - G(t)}{f(x)} \right\}^{\frac{1}{2}} = \sup_{a\leq x\leq b} \{\psi(x)\}^{\frac{1}{2}} \left\{ \int_{-\infty}^{\infty} K^2(u)du \right\}^{\frac{1}{2}}.$$

2.3 Preuves

Cette patrie est consacrée aux démonstrations des théorèmes 2.2, 2.3, 2.4

2.3.1 Quelques notations supplémentaires et résultats utiles

Dans ce qui suit, on utilisera pour la suite les notations suivantes :
Soit H la fonction de répartition de Z notée H(x) = H(x+), pour $x \in \mathbb{R}$, se décompose de la manière suivante :

H(x) = 1 - (1 - F(x))(1 - G(x)) = $H^{(1)}(x) + H^{(0)}(x)$,
où

$$H^{(1)}(x) = \mathbb{P}(Z \leq x, \delta = 1) = \int_0^x (1 - G_-(t))dF(t) = H_+^{(1)}(x)$$

et

$$H^{(0)}(x) = \mathbb{P}(Z \leq x, \delta = 0) = \int_0^x (1 - F-(t))dG(t) = H_+^{(0)}(x)$$

on note

$$p = \mathbb{P}(\delta = 1) = \int_0^x (1 - G_-(t))dF(t) = H_-^{(1)}(\infty) = 1 - H_-^{(0)}(\infty)$$

On suppose que 0 < p < 1. On peut donc définir les fonctions de quantile associées respectivement à $H^{(1)}$ et $H^{(0)}$ comme suit. Soient les fonctions $Q^{(1)}$ et $Q^{(0)}$ définies par :

$$Q^{(1)}(j) = \inf\{x : H^{(1)}(x) \geq j\}\, pour\, 0 < j < p$$

$$Q^{(0)}(j) = \inf\{x : H^{(0)}(x) \geq j\}\, pour\, 0 < j < 1 - p$$

On définit les fonctions de répartitions empiriques correspondant à H, $H^{(1)}, H^{(0)}$ sont données par

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq x\}} = H_n^{(1)}(x) + H_n^{(0)}(x)$$

où

$$H_n^{(1)}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{I}_{\{Z_i \leq x\}},$$

et

$$H_n^{(0)}(x) = \frac{1}{n}\sum_{i=1}^{n} (1 - \delta_i)\mathbb{I}_{\{Z_i \leq x\}}.$$

La fonction empirique cumulée de taux de panne est définie par

$$\Lambda_n(x) = \int_0^x \frac{1}{1 - H_{n-}(u)} dH_n^{(1)}(u) = \Lambda_{n+}(x)\, pour\, x \geq 0.$$

Les estimateurs non-paramétriques de Kaplan-Meier, $\widehat{F}_n$ et $\widehat{G}_n$, de F et G, respectivement, basés sur l'échantillon $\{(Z_i, d_i) : 1 \leq i \leq n\}$ peuvent s'exprimer de la maniére suivante (voir, par exemple, [17] p.295). On a, pour tout

$x \geq 0$,

$$\begin{aligned}
\widehat{F}_n(x) &= 1 - \prod_{\substack{i:Z_i,n\leq x \\ 1\leq i\leq n}} \left(1 - \frac{\delta_{i,n}}{n-i+1}\right) \\
&= \int_0^x (1 - \widehat{F}_{n-}(u)) d\Lambda_n(u) \\
&= \int_0^x \frac{1 - \widehat{F}_{n-}(u)}{1 - H_{n-}(u)} dH_n^{(1)}(u) \\
&= \int_0^x \frac{1}{1 - \widehat{G}_{n-}(u)} dH_n^{(1)}(u).
\end{aligned}$$

et de même façon pour

$$\begin{aligned}
\widehat{G}_n(x) &= 1 - \prod_{\substack{i:Z_i,n\leq x \\ 1\leq i\leq n}} \left(1 - \frac{\delta_{i,n}}{n-i+1}\right) \\
&= \int_0^x \frac{1}{1 - \widehat{F}_{n-}(u)} dH_n^{(1)}(u).
\end{aligned}$$

Dans le même esprit, on va voir maintenant un résultat d'approximation utile, en premier lieu, que l'on peut décomposer $\widehat{a}_n$(x) de la façon suivante : On désigne par :

$$\widehat{a}_n(x) = \sqrt{n}(\widehat{F}_n(x) - F(x))$$

et

$$\widehat{b}_n(x) = \sqrt{n}(\widehat{G}_n(x) - G(x))$$

les processus empiriques de Kaplan-Meier associés à F et G.

On considère le processus empirique suivant :

$$\mathcal{W}_n^i(x) = \sqrt{n}(H_n^i(x) - H^i(x)), \textit{pour } i = 0, 1 \textit{ et } x \in \mathbb{R}.$$

2.3.2 Un résultat d'approximation utile

Pour tout $x \in \mathbb{R}$, nous avons la relation :

$$\begin{aligned}
\widehat{a}_n(x) &= \sqrt{n}(\widehat{F}_n(x) - F(x)) \\
&= \sqrt{n}\left\{\int_0^x d\widehat{F}_n(u) - \int_0^x dF(u)\right\} \\
&= \left\{\int_0^x \frac{1}{1 - \widehat{G}_{n-}(u)} d\mathcal{W}_{n,1}(u) + \int_0^x \frac{\beta_{n-}(u)}{1 - \widehat{G}_{n-}(u)} dF(u)\right\} \qquad (*) \\
&= \widehat{a}_n'(x) + \widehat{a}_n''(x).
\end{aligned}$$

La formule (*) nous permet d'écrire que

$$\widehat{\eta}(h,t,s) = \{\widehat{a}'_n(t+hs) - \widehat{a}'_n(t)\} - \{\widehat{a}"_n(t+hs) - \widehat{a}"_n(t)\}.$$

Le lemme suivant a été obtenu par Deheuvels et Einmahl [47].
Lemme .1.. On suppose f(·) uniformément bornée sur [0,R], 0 < R < Θ. Alors il existe une constante $C(R) < \infty$, telle que, presque sûrement, pour tout $n \geq 1$, on a uniformément en $0 \leq s \leq t \leq R$,

$$|a"_n(t) - a"_n(s)| = \left| \int_s^t \frac{\widehat{b}_{n-}(u)}{1-\widehat{G}_{n-}(u)} dF(u) \right| \leq C(R)(log_2(n))^{1/2} \times |t-s|.$$

Preuve.
Soit c(R) = $sup_{0\leq u\leq R}$ |f(u)|, pour tout $n \geq 1$

$$\begin{aligned} \left| \int_s^t \frac{\widehat{b}_{n-}(u)}{1-G_{n-}(u)} dF(u) \right| &\leq \frac{1}{1-G_{n-}(R)} \times \sup_{0\leq u\leq} |\widehat{b}_{n-}(u)| \times \{F(t)-F(s)\} \\ &\leq \frac{c(R)}{1-G(R)} \times C(R)(log_2(n))^{1/2} \times |t-s| \\ &= C(R)(log_2(n))^{1/2} \times |t-s|. \end{aligned}$$

Pour prendre la mesure de l'amplitude des oscillations du processus $\widehat{a}'_n$, il sera commode d'introduire la quantité $\mathcal{A}_{n,1}$, définie par

$$\begin{aligned} \mathcal{A}_{n,1}(s,t) &= \widehat{a}'_n(s) - \widehat{a}'_n(t) - \frac{1}{1-G_-(s)} \int_s^t d\mathcal{W}_n^{(1)}(u) \\ &= \int_s^t \left(\frac{1}{1-\widehat{G}_{n-}(u)} - \frac{1}{1-G_{n-}(u)} \right) d\{\mathcal{W}_n^{(1)}(u) - \mathcal{W}_n^{(1)}(s)\} \\ &= \left(\frac{1}{1-\widehat{G}_{n-}(t)} - \frac{1}{1-G_{n-}(s)} \right) \{\mathcal{W}_n^{(1)}(t) - \mathcal{W}_n^{(1)}(s) \\ &- \int_s^t \{\mathcal{W}_n^{(1)}(u) - \mathcal{W}_n^{(1)}(s)\} d\left\{ \frac{1}{1-G_{n-}(t)} \right\}. \end{aligned}$$

Sur un espace de probabilité convenablement élargi $(\Omega, \mathcal{A}, \mathbb{P})$, il est possible de définir $\{Y_n : n \geq 1\}$ et $\{C_n : n \geq 1\}$, conjointement à une suite i.i.d. $\{U_n : n \geq 1\}$ de variables aléatoires, de loi uniforme sur (0,1), telle que les propriétées suivantes soient vérifiées. Pour $n \geq 1$ et $s \in \mathbb{R}$, soit

$$U_n(s) = \frac{1}{n} \sum_i^n \mathbb{I}_{\{U_i \leq s\}}, \; et$$

$$\alpha_n(s) = \sqrt{n}(U_n(s) - s).$$

On a, presque sûrement,

$$H_n^{(1)}(x) = U_n(H^{(1)}(x)) \, pour \, 0 < H^{(1)}(x) < p,$$

et

$$H_n^{(0)}(x) = U_n(H^{(0)}(x) + p) - U_n(p) \, pour \, 0 < H^{(0)}(x) < 1 - p.$$

où $p = \mathbb{P}(\delta = 1)$ et $U_n(\Delta)$ est la fonction de répartition empirique définie.

2.3.3 Lemmes préliminaires

A partir de la définition du processus empirique uniforme et de la définition de $\mathcal{W}_n$, j(·), j = 0, 1, on pose,

$$\begin{aligned} w_{n,1}(h) &= \sup_{\substack{s,t\in I \\ |t-s|\leq h}} |\mathcal{W}_{n,1}(t) - \mathcal{W}_{n,1}(s)|. \\ &= \sup_{\substack{s,t\in I \\ |t-s|\leq h}} |\alpha_n(H_1(t)) - \alpha_n(H_1(s))|, \; h > 0 \end{aligned}$$

et

$$\widehat{w}_{n,1} = \sup_{h\in\mathcal{W}_n} \frac{w_{n,1}(h)}{\Gamma(h)}. \qquad avec \; \Gamma(h) = \sqrt{2hlog_+(1/h)}$$

Nous pouvons maintenant énoncer un lemme permettant d'évaluer le comportement uniforme asymptotique de $\mathcal{A}_{n,1}(s,t)$.

Lemme .2. Sous les conditions (F1-F2-F4), il existe une suite de constantes C_n, telle que $C_n \to 0$ lorsque $n \to \infty$, et une constante C' > 0, de telle sorte que, presque sûrement pour tout n suffisamment grand,

$$\sup_{h\in\mathcal{W}_n} \sup_{\substack{s,t\in I \\ |t-s|\leq h}} \frac{|\mathcal{A}_{n,1}(s,t)|}{\Gamma(h)} \leq \widehat{w}_{n,1} \times \left\{ C'\sqrt{\frac{log_2(n)}{n}} + C_n \right\}.$$

Preuve.

D'après la loi du logarithme itéré de Földes et Rejtő [38], il existe une constante $C' \in \mathbb{R}$, telle que, presque sûrement pour tout n suffisamment grand,

$$\sup_{a\leq t\leq b'} \left| \frac{1}{1 - G_{n-}(t)} - \frac{1}{1 - \widehat{G}_-(t)} \right| \leq \frac{C'}{3}\sqrt{\frac{log_2(n)}{n}}.$$

Et d'après la condition de continuité (F2), portant sur G, nous pouvons écrire, au vu de (F1-4), que

$$C_n := 2 \sup_{h\in[h'_n,h"_n]} \sup_{\substack{a\leq s,t\leq b' \\ |t-s|\leq h}} \left| \frac{1}{1 - G_-(t)} - \frac{1}{1 - G_-(s)} \right| \to 0, \quad lorsque \, n \to \infty.$$

On combine la définition de $\widehat{w}_{n,1}$ avec les observations de lemme, pour obtenir les relations suivantes lorsque $n \to \infty$,

$$\sup_{h\in[h'_n,h"_n]} \sup_{\substack{s,t\in I\\ |t-s|\leq h}} \Gamma(h)^{-1} \left| \left(\frac{1}{1-\widehat{G}_-(t)} - \frac{1}{1-G_-(s)} \right) \{\mathcal{W}_{n,1}(t) - \mathcal{W}_{n,1}(s)\} \right|$$

$$\leq \widehat{w}_{n,1} \sup_{t\in[h'_n,h"_n]} \sup_{\substack{a\leq s,t\leq b'\\ |t-s|\leq h}} \left| \frac{1}{1-\widehat{G}_{n-}(t)} - \frac{1}{1-G_-(t)} + \frac{1}{1-G_-(t)} - \frac{1}{1-G_-(s)} \right|$$

$$\leq \widehat{w}_{n,1} \times \left\{ \frac{2C'}{3}\sqrt{\frac{log_2(n)}{n}} + \frac{C_n}{2} \right\}.$$

Considérons maintenant la quantité $\widehat{\eta}_n^{(1)}(h,t;s)$, définie par
Pour tout choix de $h \geq 0$ et $t \in \mathbb{R}$, posons

$$\begin{aligned} \widehat{\eta}_n^{(1)}(h,t;s) &= \frac{1}{1-G_-(t)}\{\mathcal{W}_{n,1}(t+hs) - \mathcal{W}_{n,1}(t)\} \\ &= \frac{1}{1-G_-(t)}\{a_n(H_1(t+hs)) - a_n(H_1(t))\}, \quad s \in \mathbb{R}. \end{aligned}$$

On déduit des lemmes 1 et 2 le résultat d'approximation suivant.
Lemme .3. Lorsque $n \to \infty$, Sous les conditions (F1-F4), il existe une suite de constantes C_n, telle que $C_n \to 0$ lorsque $n \to \infty$, et des constantes C' et C = C(b), telles que, presque sûrement pour tout n suffisamment grand,

$$\sup_{h\in[h'_n,h"_n]} \sup_{a\leq t\leq b} \Gamma^{-1}||\widehat{\eta}_n(h;t;I) - \widehat{\eta}_{n,1}(h;t;I)||$$
$$\leq C\sqrt{\frac{h"_n log_2(n)}{2log(1/h'_n}} + \widehat{w}_{n,1}\left(C'\sqrt{\frac{log_2(n)}{n}} + C_n \right).$$

Preuve.

D'après les définitions précédante, on observe que

$$\begin{aligned} \sup_{h\in[h'_n,h"_n]} \sup_{a\leq t\leq b} \Gamma^{-1}||\widehat{\eta}_n(h;t;I) - \widehat{\eta}_{n,1}(h;t;I)|| &\leq \sup_{h\in[h'_n,h"_n]} \sup_{\substack{a\leq t\leq b\\ s\in[0,1]}} |a"_n(t+hs) - a"_n(t)| \\ &+ \sup_{h\in[h'_n,h"_n]} \sup_{\substack{a\leq t\leq b\\ s\in[0,1]}} |\mathcal{A}_{n,1}(t+sh,t)|. \end{aligned}$$

La conclusion du cette lemme est alors obtenue, en faisant un usage combinré des lemmes 1 et 2

2.4 Approximation et loi limite fonctionnelle

L'objet de ce paragraphe est d'approximer la fonction d'increment du processus empirique de Kaplan-Meier par une fonction d'incrément spécifique du processus empirique uniforme , et cela en vue d'appliquer une nouvelle loi limite fonctionnelle.

Preuve du théorème 2.2

Posons, pour h > 0,

$$w_n(h) = \sup_{\substack{0\leq s,t\leq 1\\ |t-s|\leq h}} |\alpha_n(t) - \alpha_n(s)|$$

(*) Soit deux suites de constantes positives, $\{h'\}_{n\geq 1}$ et $\{h"\}_{n\geq 1}$, vérifiant les conditions (H.1-2-3), avec $0 < h' \leq h" < \infty$. Pour tout $\sigma > 0$, on a, avec probabilité un,

$$\limsup_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} w_n(\sigma h) = \sigma^{1/2}.$$

Nous rappelons la notation utilisée pour les incréments du processus empirique. On pose

$$\xi_n(h,t,s) = \alpha_n(t+hs) - \alpha_n(t).$$

(**) Soit deux suites de constantes positives, $\{h'\}_{n\geq 1}$ et $\{h"\}_{n\geq 1}$, vérifiant les conditions (H1-H3), avec $0 < h' \leq h" < \infty$. Pour tout couple de constantes réelles (c_1, c_2) tel que $0 \leq c1 < c_2 \leq 1 - h"_n$, et tout $\sigma > 0$, on a presque sûrement,

$$\lim_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \sup_{c_1\leq t\leq c_2} \inf_{g\in\mathbb{S}_\sigma} \left\| \frac{\widehat{\beta}_n(\sigma h,t,.)}{\sqrt{2hlog(1/2)}} - g \right\| = 0,$$

et

$$\forall g \in \mathbb{S}_\sigma \lim_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \inf_{c_1\leq t\leq c_2} \left\| \frac{\widehat{\beta}_n(h,t,.)}{\sqrt{2hlog(1/2)}} - g \right\| = 0.$$

En combinant (*) et le lemme 3, on obtient le résultat suivant
Lemme .4. Soit deux suites de constantes positives, $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$, vérifiant les conditions (H1-H3), avec $0 < h'_n \leq h"_n < \infty$. on a, avec probabilité 1,

$$\lim_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \sup_{a\leq t\leq b} ||\widehat{\eta}_n(h,t,I) - \widehat{\eta}_n^{(1)}(h,t,I)|| = 0.$$

Preuve.
Soit

$$\delta = \sup_{u\in I} \frac{dH^{(1)}(u)}{du} = \sup_{u\in I} f(u)(1-G(u)).$$

On a, uniformément en $a \leq s,t \leq b$,

$$|H^{(1)}(t) - H^{(1)}(s)| \leq \delta|t-s|.$$

Ainsi, on a, à partir d'un certain rang, pour $h \in [h'_n, h"_n]$,

$$w_n^{(1)}(h) \leq w_n(\delta h),$$

et donc, en rappelant la définition de $w_n^{(1)}$, on a, presque sûrement,

$$\limsup_{n\to\infty} w_n^{(1)} \leq \delta^{1/2}.$$

Enfin, d'aprés les hypothéses (H.1) et (H.2) sur les suites $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$,

$$\sqrt{\frac{h"_n log_2(n)}{2log(1/h'_n)}} = o(h"_n^{1/2}) \to 0, \qquad lorsque \quad n\to\infty.$$

Soit $N \geq 1$ fixé. Considérons maintenant la discrétisation de l'intervalle [a,b] suivante.
Pour $1 \leq i \leq N$, soit $t_{i,N} = a+(i-1)(b-a)N^{-1}$ et $\sigma_{i,N} = f(t_{i,N})(1-G(t_{i,N}))$. On pose

$$\begin{aligned}\widehat{\eta}_{n,N}^{(1)}(h,t,s) &= \tfrac{1}{1-G(t)}\left\{a_n\left(H^{(1)}(t) + shf(t_{i,N})(1-G(t_{i,N}))\right) - a_n\left(H^{(1)}(t)\right)\right\} \\ &= \tfrac{1}{1-G(t)}\xi_n(\sigma_{i,N}h, H^{(1)}(t); s).\end{aligned}$$

Le lemme suivant montre que l'on peut approcher $\widehat{\eta}_n^{(1)}$ par $\widehat{\eta}_{n,N}^{(1)}$ avec un ordre suffisant pour nos besoins.
Lemme .5. Soit deux suites de constantes, $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$, vérifiant les conditions (H1-H3), avec $0 < h'_n \geq h"_n < \infty$. pour tout $\varepsilon > 0$, il existe presque sûrement un $N_0 = N_0(\varepsilon) < \infty$, tel que, pour tout $N \geq N_0$,

$$\limsup_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \sup_{a\leq t\leq b} \|\widehat{\eta}_{n,N}^{(1)}(h,t,I) - \widehat{\eta}_n^{(1)}(h,t,I)\| \leq \varepsilon$$

Preuve.

$$\gamma_N = \max\left(\sup_{h\in[h'_n,h"_n]} \sup_{t_{i,N}\leq t\leq t_{i+1,N}+h} |f(t)(1-G(t)) - f(t_{i,N})(1-G(t_{i,N}))|\right).$$

Pour tout $1 \leq i \leq N, h \in [h'_n, h"_n], t \in [t_{i,N}, t_{i+1,N}]$ et $s \in [0,1]$, il existe un $t_1 \in [t_{i,N}, t_{i+1,N+h}]$, tel que, à partir d'un certain rang,

$$|H^{(1)}(t+hs) - \{H^{(1)}(t) + hsf(t_{i,N})(1-G(t_{i,N}))\}|$$

$$\leq hs|f(t_1)(1-G(t_1) - f(t_{i,N})(1-G(t_{i,N}))| \leq \gamma_N h.$$

Cette derniére inégalité implique que, presque sûrement,

$$\limsup_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \sup_{a\leq t\leq b} \|\widehat{\eta}^{(1)}_{n,N}(h,t,I) - \widehat{\eta}^{(1)}_n(h,t,I)\|$$

$$\leq \sup_{a\leq t\leq b}\left\{\frac{1}{1-G(t)}\right\} \limsup_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} w_n(\gamma_N h) = \left\{\frac{1}{1-G(b)}\right\}\gamma_N^{1/2}.$$

Maintenant, on note R une fonction continue et positive sur [a,b], on pose

$$R_{i,N} = \sup_{t_{i,N}\leq t\leq t_{i+1,N}+h} \frac{R(t)}{1-G(t)}$$

Définissons, de plus,

$$\begin{aligned} \mathcal{M}^{1/2} &= \sup_{a\leq t\leq b}\left\{\tfrac{R(t)}{1-G(t)}\right\}\{f(t)(1-G(t))\}^{1/2}; \\ &= \sup_{a\leq t\leq b} R(t)\left\{\tfrac{f(t)}{(1-G(t))}\right\}^{1/2}. \end{aligned}$$

Lemme .6. Soit deux suites de constantes, $\{h'_n\}_{n\geq 1}$ et $\{h"_n\}_{n\geq 1}$, vérifiant les conditions (H1-H3), avec $0 < h'_n \leq h"_n < \infty$. on a, presque sûrement,

$$\lim_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \sup_{a\leq t\leq b} \inf_{g\in\mathbb{S}_{\mathcal{M}}} \left\| R(t)\frac{\widehat{\eta}^{(1)}_{n,N}(h,t,.)}{\sqrt{2hlog(1/2)}} - g\right\| = 0,$$

et

$$\forall g \in \mathbb{S}_{\mathcal{M}} \lim_{n\to\infty} \sup_{h\in[h'_n,h"_n]} \Gamma_h^{-1} \inf_{a\leq t\leq b} \left\| R(t)\frac{\widehat{\eta}^{(1)}_{n,N}(h,t,.)}{\sqrt{2hlog(1/2)}} - g\right\| = 0.$$

Preuve

Soit $\varepsilon > 0$. Pour tout $1 \leq i \leq N$ fixé, on pose $c_1 = H^{(1)}(t_{i,N})$ et $c_2 = H^{(1)}(t_{i+1,N})$. la fonction H(1) est strictement croissante sur [a,b]. On a donc,

$$0 \leq c_1 < c_2 \leq H^{(1)}(b) \leq 1 - h"_n.$$

Soit $\rho = max_{1\leq i\leq N} R_{i,N}$. on a, pour tout $h \in [h'_n, h"_n]$ et tout $t_{i,N} \leq t \leq t_{i+1,N}$,

$$\begin{aligned}\widehat{\mathcal{K}}^{(1)}_{n,N,i}(h,t,I) &= R(t)\Gamma_h^{(-1)}\widehat{\eta}^{(1)}_{n,N}(h,t,I);\\ &= \frac{R(t)}{1-G(t)}\Gamma_h^{(-1)}\xi_n(\sigma_{i,N}h, H^{(1)}(t), I).\end{aligned}$$

Ainsi, portant sur le processus empirique « classique » (i.e. dans le cas non censuré), il existe presque sûrement un $n_0 = n(\varepsilon, i, N)$ tel que, pour tout $n \geq n_0$,

$$\sup_{h\in[h'_n,h"_n]}\sup_{c_1\leq u\leq c_2}\inf_{g\in\mathbb{S}_{\sigma_{i,N}}}\left\|\frac{\xi_n(\sigma_{i,N}h,u,.)}{\sqrt{2hlog(1/h)}}-g\right\| \leq \varepsilon/(2\rho).$$

On a donc, presque sûrement,

$$\sup_{h\in[h'_n,h"_n]}\sup_{t_{i,N}\leq t\leq t_{i+1,N}}\inf_{g\in\mathbb{S}_{\mathcal{M}_{i,N,1}}}\left\|\widehat{\mathcal{K}}^{(1)}_{n,N,i}(h,t,.)-g\right\| \leq \varepsilon/2.$$

Comme $\bigcup_{i=1}^{N}\mathbb{S}_{\mathcal{M}_{i,N,1}} \subseteq \mathbb{S}_{M_{N,1}}$, on en conclut que, presque sûrement,

$$\sup_{h\in[h'_n,h"_n]}\sup_{a\leq t\leq b}\inf_{g\in\mathbb{S}_{\mathcal{M}_{,N,1}}}\left\|\widehat{\mathcal{K}}^{(1)}_{n,N,}(h,t,.)-g\right\| \leq \varepsilon/2.$$

Et comme conclusion de la preuve du théorème **2.3** on obtient que le choix de la fonction de R est définit par :

$$R(t) = \{\psi(t)(1-G(t))/f(t)\}^{1/2},$$

conduit à la relation

$$\sup_{a\leq t\leq b} R(t) = \sup_{a\leq t\leq b}(\psi(t))^{1/2}.$$

Preuve du théorème 2.3

Soit$(E,\mathcal{T})$ un ensemble E de fonctions, muni d'une topologie métrisable $\mathcal{T}$. La proposition suivante usagée par (cf. Deheuvels et Einmahl [47]).
Proposition. Soit Y_n une suite de fonctions de $(E,\mathcal{T})$, presque sûrement relativement compacte, et ayant comme ensemble limite E. Soit de plus Υ : $E \to R$ une fonctionnelle $\mathcal{T}$ -continue. Alors on a

$$\lim_{n\to\infty}\{\sup_{g\in Y_n}\Upsilon(g)\} = \sup_{g\in Y}\Upsilon(g).$$

Demonstration. Omise.□

En notant de plus que, pour tout $\sigma > 0$,

$$\begin{aligned}\sup_{g\in\mathbb{S}_\sigma}\left\{-\int_{-T/2}^{T/2} g(u)dK(u)\right\} &= \sup_{g\in\mathbb{S}_\sigma}\left\{-\int_{-T/2}^{T/2} \acute{g}(u)dK(u)\right\} \\ &= \left\{\sigma\int_{-T/2}^{T/2} K^2(u)\right\}^{1/2}\end{aligned}$$

la conclusion du théorème **2.3** s'obtient par des arguments analytiques de routine.

Preuve du théorème 2.4

Onpose

$$\psi_{n,1} = \psi\{(1-\widehat{F}_n)^2(1-F)^2\}.$$

On constate alors que

$$\widehat{f}_{n,h} - \widehat{\mathbb{E}}\widehat{f}_{n,h}(x))\left\{\psi_n\frac{1-G(x)}{f(x)}\right\} = \left(\frac{\widehat{f}_{n,h}}{1-\widehat{F}_n} - \frac{\widehat{\mathbb{E}}\widehat{f}_{n,h}(x)}{1-\widehat{F}_n}\right)\left\{\psi_{n,h}(x)\frac{1-H(x)}{\lambda(x)}\right\}^{1/2}.$$

Le théorème **2.4**, le théorème **2.5** s'obtient comme conséquence directe de la loi du logarithme itéré pour le processus $\widehat{a}_n$ (voir Földes et Rejtõ [38]),

Conclusion

Les méthodes classiques d'analyse de survie supposent l'indépendance des temps de survenue de l'événement d'intérêt ; or cette hypothèse ne peut plus être raisonnablement posée lors de l'étude de données de survie groupées. Les modèles robustes, qu'ils soient de type marginal ou de type mixte, permettent alors de traiter ces données hétérogènes.

Dans ce mémoire on s'intérese particulièrement au résultat de Deheuvels et Einmahl [2], qui est une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique dans le cadre des données censurés à droite.

Dans le 1^r temps, nous exposons les résultats de la théorie des processus empiriques, ainsi nous intéressons au processus empirique dans le cas de données censurées à droite. Dans ce modèle, on n'observe pas les données d'intérêt, et donc il suffit d'estimer la fonction de répartition empirique par l'estimateur de Kaplan- Meier.

Pour ce modèle, nous avons présentez quelques résultats préliminaires, et on s'est intéressés particulièrement à une loi fonctionnelle du logarithme itéré pour les accroissements du processus de Kaplan-Meier.

Bibliographie

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

[2] **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.

[3] **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.

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

[5] **Paul Deheuvels**, *Laws of the iterated logarithm for density estimators.*,In G. Roussas, éditeur :Non parametric Functional estimators and Related Topics, NATO adv.Sci.Ins. C, page 19-29. Kluwer Academic, 1991.

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

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

[8] **A. Khinchine**, *Über einen Satz der Wahrscheinlichkeitsrechnung.* Fundamenta Mathematica, 6 :9-20, 1924.

[9] **A. Kolmogorov**, *Über das Gesetz des iterierten Logarithmus.* Mathematische Annalen, 101 :126-135, 1929.

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

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

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

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

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

[15] **W. Stute**, *The oscillation behavior of empirical processes.* Ann. Prob., 10 :86-107, 1982b.

[16] **W. Stute**, *Conditional empirical processes.* Ann. Prob., 14(2) :638-647, 1986a.

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

[18] **Patrick Billingsley**, *Convergence of Probability Measures.* Wiley, New York, 1968.

[19] **Patrick Billingsley**, *Probability and Measure.* John Wiley et Sons, 2e édition, 1986.

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

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

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

[23] **van der Vaart, A. et Wellner, J.**, *Weak convergence and empirical processes.* Sprin- ger, New York. (1996).

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

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

[26] **Strassen, V**, *An invariance principle for the law of the iterated logarithm.* Z.Wahrsch.Gebiete, 3, 221-226. 1964

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

[28] **Deheuvels, P**, *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).

[29] **Einmahl, U. et Mason, D**, *An empirical process approach to the uniform consis- tency of kernel type estimators. Journ.* Theoretic. Probab., 13, 1-13. (2000).

[30] **Einmahl, U. et Mason, D. M**, *Uniform in bandwidth consistency of kernel-type function estimators.* Ann. Stat., 33(3), 1380-1403. (2005).

[31] **Glivenko, V**, *Sulla determinazione empirica delle leggi di probabilita.* Giorn. Ist. Ital. Attuari, 4, 92-99.(1933).

[32] **Donsker, M**, *An invariance principle for certain probability theorems.* Mem. Amer. Math. Soc., 6. (1951).

[33] **Dudley, R**, *Uniform central limit theorems.* Cambridge University Press. (1999).

[34] **B. B. Winter, A. Foldes et L. Rejto**, *Glivenko-Cantelli theorems for the product limit estimate.* Problems of Control and Information Theory, 7 :213-225, 1978.

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

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

[37] **W. Stute et J.-L.Wang**, *The strong lawunder random censorship.* The Annals of Statistics, 21(3) :1591-1607, 1993.

[38] **Földes, A. et Rejtő, L**, *A LIL type result for the product-limit estimator.* Z. Wahrsch. Verw. Gebiete, 56, 75-86, 1981.

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

[40] **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.

[41] **Pollard, D**, *Convergence of stochastic processes.* Springer-Verlag, N.Y, (1984).

[42] **Varron, D**, *Uniformity in h in the functional limit law for the increments of the empirical process indexed by functions.* (Uniformité en h dans la loi fonctionnelle limite uniforme pour les accroissements du processus empirique indexé par des fonctions). C. R., Math., Acad. Sci. Paris, 340(6), 453-456, (2005).

[43] **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.

[44] **Hartman, P. et Wintner, A**, *On the law of the iterated logarithm.* Amer. J. Math. 63, 169-176,(1941).

[45] **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.

[46] **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.

[47] **Deheuvels, P. et Einmahl, J**, *Functional limit laws for the increments of kaplan- meier product-limit processes and applications*. Ann. Prob, 28(7), 1301-1335, 2000.

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

[49] **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.