

Remerciements

Je tiens tout d'abord à remercier le Professeur Mourid T. pour m'avoir encadré ces deux années, je le remercie de tous ses conseils lors de la rédaction de ce mémoire. Son expérience m'a été tout à fait profitable.

J'adresse un remerciement particulier à M^r Dib et à M^r Boukhari pour toutes les connaissances dont ils nous ont enrichi et sans lesquelles ce travail n'aurait été possible.

Je tiens à exprimer ma reconnaissance aux membres du jury de m'avoir fait l'honneur de participer à ma soutenance.

Je dis un grand merci du fond du coeur à tous mes amis pour leur soutien moral.

Je me permets également de remercier mes parents, mes soeurs et mes frères pour leur soutien moral et leur encouragement tout au long de mes études. Je ne manquerai de remercier mon mari et mon fils pour leur compréhension et de leur soutien infaillible.

Je tiens enfin à remercier toutes les personnes non citées qui auraient contribué d'une manière ou d'une autre à la réalisation de ce travail.

Table des Matières

Remerciments	1
Introduction	5
1 Modèles de survie	8
1.1 Généralités	8
1.1.1 Fonctions associées aux distributions de survie	9
1.1.2 Moments et quantiles	11
1.1.3 Différentes formes de la fonction de risque	11
1.1.4 Censures et troncatures	14
1.2 Processus ponctuels associés	18
1.3 Estimateur de Nelson-Aalen	23
1.3.1 Introduction.	23
1.3.2 Construction de l'estimateur de Nelson-Aalen	24
1.3.3 Exemples	25
1.4 Consistance et loi asymptotique de l'estimateur de Nelson-Aalen	29
1.5 Estimateur de Kaplan-Meier	38
1.5.1 Estimateur de Fleming-Harington	39
1.5.2 Construction de l'estimateur de Kaplan-Meier	39
1.5.3 Exemples	42
1.5.4 Consistance et loi asymptotique de l'estimateur de Kaplan-Meier	44
1.6 Estimateurs des variances de \hat{S}_{KM} et \hat{S}_{FH}	49

2	Estimation de la densité des durées de survie	53
2.1	Données non censurées	53
2.1.1	Introduction	53
2.1.2	Comportement asymptotique de f_n	55
2.1.3	Choix optimaux du noyau K et de la fenêtre b_n	55
2.2	Données censurées	56
2.2.1	Introduction	56
2.2.2	Etude du biais de l'estimateur \hat{f}_n	59
2.2.3	Consistance et normalité asymptotique de \hat{f}_n	65
2.2.4	Erreur quadratique intégrée et fenêtre optimale	76
3	Estimation de la fonction de risque	85
3.1	Données non censurées	85
3.1.1	Introduction	85
3.1.2	Estimateur à noyau de la fonction de risque h	88
3.1.3	Moyenne et variance de l'estimateur à noyau \hat{h}_n	88
3.1.4	Estimation des dérivées de la fonction de risque	93
3.1.5	Propriétés asymptotiques	94
3.1.6	Normalité asymptotique	97
3.2	Données censurées	100
3.2.1	Introduction	100
3.2.2	Moyenne et variance de l'estimateur \hat{h}_n	100
4	Simulations	107
4.1	Introduction.	107
4.2	Lois usuelles en survie	108
4.3	Données simulées	108
4.3.1	Estimation de la fonction de survie S	108
4.3.2	Estimation du taux de hasard cumulé H	109
4.4	Données réelles	110
4.4.1	Données des ventilateurs.	110

4.4.2	Données PBC ([7] p. 359) (malades atteints de la cirrhose)	114
4.5	Estimation des fonctions de densité et de risque par la méthode du noyau	117
4.5.1	Estimation de la densité des durées T	118
4.5.2	Estimation de la fonction de risque	119
4.6	Conclusion.	120
	Conclusion	121
	Annexe	122
	Bibliographie	127

Introduction

L'analyse des données de survie s'intéresse à l'étude statistique des données qui proviennent des expériences sur des durées de vie ou durée de fonctionnement.

En fiabilité: durée de vie d'une lampe, durée de vie d'un matériel ...

En biologie: en culture de cellules les durées d'apparition de parasites ...

En médecine: durée de guérison d'un patient, durée de rémission d'un malade ...

En économie: durée de chômage

En éducation : durée d'apprentissage d'une langue ...

Dans plusieurs situations les observations sont "censurées". Une donnée est dite censurée si elle n'est pas observée. Il existe plusieurs mécanismes de censures. Particulièrement la censure aléatoire à droite: une v.a. de durée T est censurée à droite par une v.a. de censure C si on n'observe T que lorsque $T \leq C$. Dans les problèmes d'estimation de fonctions que nous allons aborder nous observons des données dites "censurées" (dites aussi incomplètes).

Une fonction essentielle en analyse de survie est la fonction de survie définie par $S(t) = P(T > t)$ qui exprime la probabilité que l'individu "survive" à l'instant t . Une autre fonction utile est la fonction de risque h (ou aussi taux de hasard, risque instantané ou intensité) qui exprime le taux de l'occurrence de "l'évènement d'intérêt" à l'instant t (panne, décès ...).

Dans ce mémoire nous présentons une étude sur les modèles de survie en abordant principalement l'estimation de la fonction de survie et la fonction de risque. Les méthodes d'estimation de ces fonctions sont de type non paramétriques et sont basées sur les estimateurs à noyau. Nous développons les résultats des articles suivants qui portent sur l'estimation de la fonction de survie et la fonction de risque dans les cas i.i.d. sans censures et avec censures:

1. Zhang Biao "Some asymptotic results for kernel estimation under random censor-

ship". *Bernoulli* 2 (2) 1996, p. 183-198.

2. Henrik Ramlau Hansen "Smoothing counting process intensities by means of kernel functions". *The Annals of Statistics* 1983 Vol 11 n 2 p. 453-466.

3. Tanner M. A. and Wong W.H. "The estimation of the hazard function from randomly censored data by the kernel method". *The Annals of Statistics* 1983 Vol 11 n 33 p. 989-993.

Dans le chapitre 1, nous rappelons des préliminaires sur les modèles de survie en se basant sur les références [5] et [10]. Nous introduisons les principales fonctions en analyse de survie: fonction de survie- taux de survie et les différentes formes du taux de risque ect... Nous donnons aussi les différents types de censure (censure à droite, censure à gauche, censure par intervalle, troncature ect...) Nous rappelons quelques notions et propriétés sur les processus ponctuels utilisés pour l'estimation des fonctions de survie et de risque.

Dans le cas d'observations censurées à droite, nous introduisons l'estimateur de Nelson-Aalen du risque cumulé et les estimateurs de Fleming-Harrington et Kaplan-Meier de la fonction de survie. Nous donnons leurs principales propriétés de convergence, lois limites et variances asymptotiques. Nous présentons quelques exemples de données censurées avec les graphes des estimateurs correspondants.

Dans le chapitre 2, en première partie, nous rappelons les résultats de l'estimation de la densité f des durées T dans le cas de données non censurées par la méthode du noyau en se basant essentiellement sur le livre de Bosq D. (cf. [4]). Dans la deuxième partie, nous développons les résultats de Biao Zhang sur l'estimation de la densité f des données T par la méthode du noyau dans le cas de la censure à droite. Cet estimateur est construit sur l'estimateur de Kaplan-Meier par lissage à l'aide d'un noyau. Nous donnons des résultats sur la consistance, convergence en moyenne quadratique, loi limite, forme de l'erreur quadratique intégrée asymptotique, choix de la fenêtre optimale ect...

Dans le chapitre 3, en première partie, nous développons les résultats de H. Ramlau Hansen sur l'estimation de la fonction du risque dans le cas de données non censurées par la méthode du noyau par lissage de l'estimateur de Nelson-Aalen. Nous donnons les expressions asymptotiques du biais et de la variance de cet estimateur ainsi qu'une loi limite et des convergences uniforme en moyenne quadratique.

Dans la deuxième partie, nous développons les résultats de M. A. Tanner and W.H.

Wong sur l'estimation de la fonction de risque h par la méthode du noyau dans le cas de la censure à droite. Nous donnons des résultats sur le biais et la variance asymptotique sous la condition de noyau "compatible".

Dans le chapitre 4, nous présentons des simulations numériques illustrant les comportements des estimateurs des fonctions de survie et fonction de risque et de la densité. En Annexe, nous donnons quelques énoncés des théorèmes utilisés dans ce mémoire.

Chapitre 1

Modèles de survie

1.1 Généralités

L'analyse de survie s'intéresse à l'étude statistique des données qui proviennent des expériences sur des durées de vie ou durée de fonctionnement. Comme exemples en médecine, en biologie, en économie..., on s'intéresse souvent à des données sous forme de durées:

1. Durée de rémission d'une leucémie aiguë.
2. Durée de séropositivité sans symptôme de patients infectés par le V.I.H.
3. Durée de fièvre chez un patient atteint de pneumonie.
4. Durée de chômage avant de retrouver un travail.
5. Durée d'apprentissage d'une langue.
6. Durée de développement de bactéries dans une culture de cellules.

Pour les trois premiers exemples on peut désigner "*l'évènement d'intérêt*" suivant:

1. Fin de rémission.
2. Début des symptômes chez un patient séropositif.
3. Fin de fièvre chez un patient atteint de pneumonie.

A partir des évènements précédents on définit la: *durée jusqu'à l'apparition de "l'évènement d'intérêt"* qui sera la variable T correspondante:

1. Temps écoulé avant la fin de la rémission.
2. Temps écoulé sans symptôme.
3. Temps écoulé avant la fin de la fièvre.
4. Temps écoulé avant de retrouver un travail.

La durée de survie T désigne le temps qui s'écoule depuis un instant initial (début du traitement, diagnostic, chômage,...) jusqu'à la survenue d'un évènement d'intérêt final (décès du patient, rechute, rémission, guérison, travail,...). On dit que l'individu survit au temps t si à cet instant l'évènement d'intérêt final n'a pas encore eu lieu.

La variable étudiée est appelée durée de survie et elle sera notée T .

1.1.1 Fonctions associées aux distributions de survie

Dans ce qui suit, on suppose qu'un individu subit une fois et une seule un certain type d'évènement. Les observations de la survenue (censuré ou non censuré) de cet évènement chez des individus constituent les échantillons pour identifier les modèles de survie.

Soit T la durée de survie supposée une variable positive. Sa loi de probabilité peut être définie par l'une des fonctions équivalentes suivantes:

1. Fonction de répartition

$$F(t) = \mathbf{P}(T \leq t) \tag{1.1}$$

Pour t fixé, c'est la probabilité que l'évènement survienne avant l'instant t .

2. Fonction de survie

$$S(t) = \mathbf{P}(T > t), \quad t \geq 0 \tag{1.2}$$

Pour t fixé $S(t)$ désigne la probabilité que l'individu survit jusqu'à l'instant t

On a: $\diamond \quad S(t) = 1 - F(t)$

- ◇ S est continue à droite, i.e $S(t^+) = S(t)$, pour $t > 0$.
- ◇ S est décroissante. $S(0) = 1$ et $\lim_{t \rightarrow \infty} S(t) = 0$.

Remarque 1.1 Si la loi de T est continue on a

$$\mathbf{P}(T > t) = \mathbf{P}(T \geq t) \quad \text{et} \quad \mathbf{P}(T < t) = \mathbf{P}(T \leq t)$$

3. Densité de probabilité

La densité de probabilité de T notée f est donnée par

$$f(t) = \lim_{\Delta \rightarrow 0} \frac{\mathbf{P}(t < T \leq t + \Delta)}{\Delta} \quad (1.3)$$

On a: $f(t) = F'(t) = -S'(t)$ si F est absolument continue. Pour t fixé, $f(t)\Delta$ s'interprète comme la probabilité que l'évènement survienne dans le petit intervalle de temps $]t, t + \Delta]$.

4. Fonction de risque (ou taux de hasard)

La fonction de risque est défini par:

$$\begin{aligned} h(t) &= \lim_{\Delta \rightarrow 0} \frac{\mathbf{P}(t < T \leq t + \Delta / T > t)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \frac{\mathbf{P}(t < T \leq t + \Delta)}{\mathbf{P}(T > t)} = \frac{f(t)}{S(t)} \end{aligned} \quad (1.4)$$

c'est le risque instantané que l'évènement survienne dans l'intervalle $]t, t + \Delta]$ sachant qu'il n'a pas eu lieu avant l'instant t (être en "fonctionnement" ou être "en vie" avant l'instant t).

5. Fonction de risque cumulé (Taux de hasard cumulé)

Il est défini par: (si F est continue)

$$H(t) = \int_0^t h(u) du \quad (1.5)$$

Si F est quelconque:

$$H(t) = \int_0^t \frac{dF(s)}{1 - F(s^-)}$$

Relation entre ces différentes fonctions

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)du\right\} \quad (1.6)$$

$$H(t) = -\ln(S(t)) \quad (1.7)$$

$$f(t) = h(t)S(t) = h(t)\exp\{-H(t)\}$$

$$F(t) = 1 - S(t) = 1 - \exp\{-H(t)\}$$

Remarque 1.2 Dans la pratique, ces cinq fonctions sont inconnues et il s'agit de les estimer en particulier la fonction de survie S et la fonction de risque cumulé H .

1.1.2 Moments et quantiles

On suppose que la f.d.r. de la v.a. T est continue et les paramètres de position et de dispersion: moyenne, médiane, étendue, écart-type, etc... peuvent être calculés.

1. Quantiles de la durée de survie

Pour $0 < p < 1$, le quantile d'ordre p est défini par

$$q_p = \inf\{t : F(t) \geq p\}$$

Si F est strictement croissante alors $q_p = F^{-1}(p)$.

2. Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbf{E}(T)$ ainsi que sa variance $\mathbf{V}(T)$ sont des quantités importantes et sont données par

$$\mathbf{E}(T) = \int_0^\infty S(t)dt, \quad \mathbf{V}(T) = 2 \int_0^\infty tS(t)dt - (\mathbf{E}(T))^2$$

1.1.3 Différentes formes de la fonction de risque

Dans cette partie nous donnons les principales formes de lois utilisées en données de survie. Les formes usuelles de la fonction de risque h sont les suivantes:

1. Constante
2. Monotone (croissante ou décroissante)
3. En forme de \cap

Risque constant

La seule distribution qui admette une fonction de risque h constante est la loi exponentielle.

la loi exponentielle $\mathcal{E}(\lambda)$

$$\begin{aligned}
 S(t, \lambda) &= e^{-\lambda t} & t \geq 0, \lambda > 0 \\
 f(t, \lambda) &= \lambda e^{-\lambda t} & t \geq 0 \\
 h(t) &= \lambda \\
 q_p &= -\frac{\ln(1-p)}{\lambda}, & 0 < p < 1, \mathbf{E}(T) = \frac{1}{\lambda}, \mathbf{V}(T) = \frac{1}{\lambda^2}
 \end{aligned}$$

La fonction de risque h est égale au paramètre λ .

Risque monotone

Il existe plusieurs distributions de durée de vie ayant une fonction de risque monotone

Lois de Weibull $W(\lambda, \alpha)$ Ce sont des lois qui généralisent la loi Exponentielle

$$\begin{aligned}
 S(t, \alpha, \lambda) &= \exp\{-(\lambda t)^\alpha\} & \lambda, \alpha > 0, t \geq 0 \\
 f(t, \alpha, \lambda) &= \alpha \lambda^\alpha t^{\alpha-1} \exp\{-(\lambda t)^\alpha\} \\
 h(t, \alpha, \lambda) &= \alpha \lambda^\alpha t^{\alpha-1} \\
 q_p &= \frac{1}{\lambda} (-\ln(1-p))^{\frac{1}{\alpha}} \\
 \mathbf{E}(T) &= \frac{1}{\lambda} \Gamma\left(1 + \frac{1}{\alpha}\right), & \mathbf{V}(T) = \left(\frac{1}{\lambda}\right)^2 \Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)
 \end{aligned}$$

Le paramètre α est un paramètre de forme de h et λ est un paramètre d'échelle.

1. Si $\alpha = 1$, on retrouve la loi Exponentielle $W(\lambda, 1) = \mathcal{E}(\lambda)$
2. Si $0 < \alpha < 1$, la fonction de risque h est décroissante de ∞ à 0.
3. Si $\alpha > 1$, la fonction de risque h est croissante de 0 à ∞ .

Lois Gamma $\Gamma(\lambda, \beta)$

$$\begin{aligned}
 f(t, \lambda, \beta) &= \lambda^\beta \frac{1}{\Gamma(\beta)} t^{\beta-1} e^{-\lambda t} & \lambda, \beta > 0, t \geq 0 \\
 F(t, \lambda, \beta) &= \frac{1}{\Gamma(\beta)} \int_0^{\lambda t} u^{\beta-1} e^{-u} du \\
 h(t, \lambda, \beta) &= \frac{f(t, \lambda, \beta)}{1 - F(t, \lambda, \beta)} \\
 \mathbf{E}(T) &= \frac{\beta}{\lambda}, \quad \mathbf{V}(T) = \frac{\beta}{\lambda^2}
 \end{aligned}$$

Le paramètre β est un paramètre de forme de h et λ est un paramètre d'échelle

1. Si $\beta = 1$, on retrouve la loi Exponentielle de paramètre λ : $\Gamma(\lambda, 1) = \mathcal{E}(\lambda)$
2. Si $\beta > 1$, la fonction de risque h est croissante de 0 à λ .
3. Si $0 < \beta < 1$, la fonction de risque h est décroissante de ∞ à $\frac{1}{\lambda}$.

Risque en \cap

Lois Log-Normales $LN(\mu, \sigma)$

$$\begin{aligned}
 S(t, \mu, \sigma) &= 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right) & \mu \in \mathbb{R}, \sigma > 0 \\
 f(t, \mu, \sigma) &= \frac{1}{\sigma t} \varphi\left(\frac{\ln t - \mu}{\sigma}\right) \\
 h(t, \mu, \sigma) &= \frac{f(t, \mu, \sigma)}{S(t, \mu, \sigma)} \\
 \mathbf{E}(T) &= e^{\mu + \frac{\sigma^2}{2}}, \quad \mathbf{V}(T) = e^{2\mu + \frac{\sigma^2}{2}} (e^{\sigma^2} - 1)
 \end{aligned}$$

où Φ est la fonction de répartition de la loi normale standard.

La fonction de risque h est en forme de \cap : elle croit de 0 à sa valeur maximum puis décroît vers 0.

Lois Log-Logistiques $LL(\lambda, \nu)$

$$\begin{aligned} S(t, \lambda, \nu) &= \frac{1}{1 + (\lambda t)^\nu} \quad \lambda, \nu > 0 \\ f(t, \lambda, \nu) &= (\lambda \nu) \lambda t^{\nu-1} (1 + (\lambda t)^\nu)^{-2} \\ h(t, \lambda, \nu) &= (\lambda \nu) \lambda t^{\nu-1} (1 + (\lambda t)^\nu)^{-1} \\ q_p &= \left(\frac{1}{\lambda}\right) \frac{p^{\frac{1}{\nu}}}{1-p}, \quad 0 < p < 1 \end{aligned}$$

Pour $0 < \nu \leq 1$ la moyenne n'existe pas. Pour $\nu > 1$

$$\mathbf{E}(T) = \frac{1}{\lambda} \Gamma\left(1 + \frac{1}{\nu}\right) \Gamma\left(1 - \frac{1}{\nu}\right)$$

La variance existe pour $\nu > 2$

$$\mathbf{V}(T) = \left(\frac{1}{\lambda}\right)^2 \left\{ \Gamma\left(1 + \frac{2}{\nu}\right) \Gamma\left(1 - \frac{2}{\nu}\right) - \Gamma^2\left(1 + \frac{1}{\nu}\right) \Gamma^2\left(1 - \frac{1}{\nu}\right) \right\}$$

Le paramètre ν est un paramètre de forme de h .

Pour $\nu > 1$ la fonction de risque h croit de 0 à sa valeur maximum puis décroît vers 0.

1.1.4 Censures et troncatures

Dans plusieurs situations les observations sont "censurées": Une donnée est dite censurée si elle n'est pas observée. Il existe plusieurs mécanismes de censures. Particulièrement la censure aléatoire à droite: une v.a. de durée T est censurée à droite par une v.a. de censure C si on n'arrive pas à observer T quand $T > C$ (ou on observe la v.a. T que lorsque $T \leq C$). Cela arrive particulièrement dans des données de survie où on n'a pas accès à toutes les observations.

Il existe trois types de censures avec des catégories de censure qu'on nomme censure à droite, censure à gauche ou censure par intervalle.

1. Censure de type I (*Censure fixe*)

Si le temps de censure est fixé par le chercheur comme étant la fin de l'étude.

Il y'a trois types:

- *Censure à droite fixe*: les variables T_1, T_2, \dots, T_n ne sont pas tous observées et on observe que les T_i qui sont inférieures ou égal à une durée fixée C (censure non aléatoire), sinon on observe C . Ainsi on observe $T_i \wedge C = X_i$
- *Censure à gauche fixe*: dans ce cas on observe $T_i \vee C = X_i$
- *Censure par intervalle fixe* $[c, C]$: dans ce cas on observe $(T_i \wedge C) \vee c = X_i$

2. Censure de type II (*Censure avec attente*)

Elle se caractérise par le fait que l'étude cesse dès que l'on a observé un nombre d'évènements prédéterminé par l'expérimentateur.

Par exemple dans le suivi de traitement de n patients, on observe leurs durées de survie jusqu'à ce qu'un nombre r d'entre eux soient décédés et on arrête l'étude à ce moment là. A partir de l'échantillon T_1, T_2, \dots, T_n et par l'utilisation de la statistique d'ordre $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$, nous aurons la date de censure égale à $T_{(r)}$ et donc on observe: $T_{(1)}, T_{(2)}, \dots, T_{(r)}$ et les variables $T_{(r+1)}, \dots, T_{(n)}$ sont censurées.

3. Censure de type III (*Censure aléatoire*)

Dans ce cas les v.a. de censure ne sont plus sous le contrôle du chercheur.

- *Censure à droite aléatoire* : Dans ce cas les censures C_i sont des variables aléatoires. Par exemple dans le suivi de traitement de n patients, nous associons à chaque patient i , non seulement son temps de guérison T_i mais aussi la censure C_i (soit "perdu de vue", soit on arrête l'étude et le patient est toujours pas guéri ect...). Dans ce cas on observe $T_i \wedge C_i = X_i$

De la même façon on définit les censures à gauche ou par intervalle.

En général les v.a. T_i sont i.i.d de f.d.r F les v.a C_i sont i.i.d. de f.d.r G et $(T_i)_i$ est indépendante de $(C_i)_i$. Pour le i ème individu, on associe: la durée T_i et la v.a. $D_i = \mathbf{1}_{\{T_i \leq C_i\}}$ qui vaut 1 si T_i est observée et 0 si T_i est censurée. Lorsqu'on ordonne les données (X_i, D_i) :

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}, \quad D_{(1)} \leq D_{(2)} \leq \dots \leq D_{(n)}$$

la v.a. $D_{(i)}$ est l'indicateur de censure (à droite) associé à $X_{(i)}$.

Dans les essais thérapeutiques, on peut citer certaines causes entraînant la censure aléatoire:

1. "Perdu de vue": le patient peut décider de se faire soigner ailleurs et on ne le revoit plus.
2. Arrêt du traitement: suite à des effets secondaires le traitement est arrêté.
3. Fin de l'étude: l'étude se termine et certains patients sont toujours vivants.

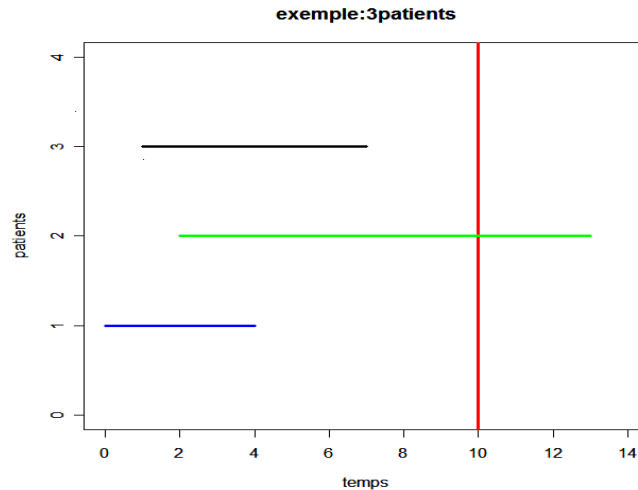


Figure 1

Dans la figure 1 on représente le suivi de trois patients. Le premier est entré au début de l'étude et est mort à $T_1 = 4$ (non censuré). Le deuxième est vivant jusqu'à la fin de l'étude au temps 10, il est donc censuré en $t = 10$. Quand au troisième, il a été perdu de vue à $t = 7$, donc il est censuré à $t = 7$.

Exemples

1. Censure à droite de type I

Dans l'apprentissage d'une langue par un groupe d'étudiants durant un stage de période fixée. On note T la durée d'apprentissage de cette langue. Pour certains étudiants nous allons observer leurs durées T_i d'apprentissage de la langue par contre pour d'autres leurs T_i ne seront pas observées car le stage est limité dans le temps.

2. Censure à droite de type III

Exemple A. En 1963, le Docteur Feireich a fait un essai thérapeutique pour comparer les durées de rémission (des rechutes), en semaines de 21 patients atteints de leucémie aiguë selon qu'ils ont reçu ou non un médicament appelé 6-mercaptopurine (6-MP), le groupe témoin a reçu un placebo. Les résultats obtenus sont les suivantes:

6 M-P	6	6	6	6 ⁺	7	9 ⁺	10	10 ⁺	11 ⁺	13	16
	17 ⁺	19 ⁺	20 ⁺	22	23	25 ⁺	32 ⁺	32 ⁺	34 ⁺	35 ⁺	
placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	

Les nombres avec + correspondent à des patients perdus de vue à la date considérée. Ils sont exclus "vivants" de l'étude et leur "durée de survie" est supérieure à celle indiquée. Ils sont **censurés** et correspond au type de censure III.

Dans l'analyse de survie on tient compte de toutes les observations censurées ou non. En effet dans les problèmes d'estimations statistiques si on élimine les observations censurés du groupe traité par le 6 M-P (12 patients) on perd de l'information puisque on ne tient pas compte des patients ayant des durées de rémission plus longues.

Exemple B. Notons T l'âge à laquelle une certaine maladie apparaît pour la première fois chez un individu. Après un examen médical on a reçu deux types de réponses:

- i) l'individu a déjà été malade mais l'âge exact de la première apparition n'a pas été retenu.
- ii) l'individu n'a jamais eu de maladie.

Dans le premier cas on n'a pas observé T mais on sait que T est inférieur à l'âge de l'individu lors de l'examen. Il s'agit d'une observation censurée à *gauche*.

Dans le deuxième cas on sait seulement que T est supérieur à l'âge de l'individu, donc on a une observation censurée à droite.

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

Les troncatures

Une variable aléatoire T est tronquée par un sous ensemble A fixé de \mathbb{R}^+ si, on observe T uniquement lorsque $T \notin A$. Plus précisément les points de l'échantillon "tronqués" sont tous dans A (dans les calculs de loi de v.a. tronquées il faut conditionner par $T \in A$). En général l'ensemble A peut être un ensemble aléatoire et on peut faire le lien avec les censures. Par exemple si C est une v.a. et on pose $A =]C, \infty[$ on dit qu'il y a une troncature à droite lorsque T n'est observée ie. si $T \in A =]C, \infty[$ et on retrouve la définition de la censure à droite. De même on dit qu'il y a troncature à gauche lorsque T n'est observée que si $T > C$.

Exemple

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

1.2 Processus ponctuels associés

Au milieu des années 1970, Aalen présente une théorie des martingales pour le processus de comptage qui offre un cadre unifié pour les méthodes statistiques de l'analyse de survie. Les processus de comptage sont utilisés dans les représentations intégrales des statistiques des

données censurées. La théorie fournit des formes simples et unifiées des estimateurs, des statistiques de test et des méthodes de régression. Ces méthodes permettent aussi d'obtenir des expressions simples de statistiques compliquées et des distributions asymptotiques des test et des estimateurs.

Dans l'analyse des durées de survie, chaque "individu" a une durée de vie T , de densité f , de fonction de répartition F et de fonction de survie $S = 1 - F$ et les v.a. T et la censure C sont supposées indépendantes. Pour chaque "individu" on lui associe un couple de processus ponctuel $((Y(t), N(t)), t \geq 0)$ donné par (où $X = T \wedge C$)

$$\begin{aligned}
 Y(t) &= \mathbf{1}_{\{X \geq t\}} = \mathbf{1}_{\{T \wedge C \geq t\}} \\
 &= \begin{cases} 1 & \text{si l'individu est en observation et à risque à l'instant } t \\ 0 & \text{sinon} \end{cases}
 \end{aligned}$$

Le processus $(Y(t), t \geq 0)$ est dit processus risque: il indique la présence du risque juste avant l'instant t (le sujet n'a pas encore subi l'évènement à t^-). Le processus $N(t)$ est défini par:

$$N(t) = \text{l'indicateur de l'"évènements d'intérêt" dans l'intervalle }]0, t]$$

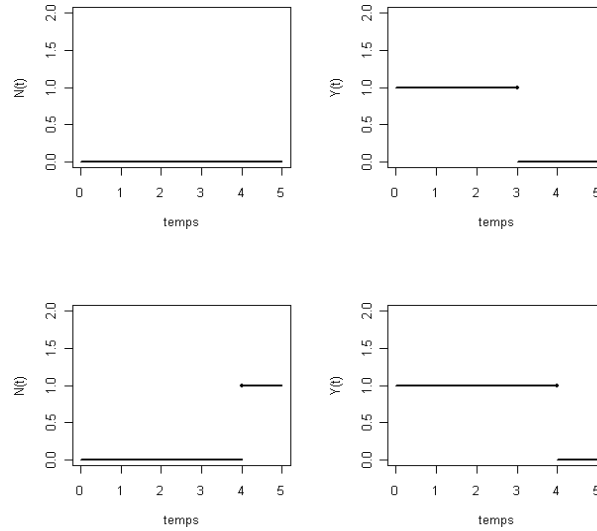
et nous avons avec l'indicatrice de censure $D = \mathbf{1}_{\{T \leq C\}}$

$$N(t) = \mathbf{1}_{\{X \leq t, D=1\}} = \mathbf{1}_{\{T \leq t, D=1\}} = \mathbf{1}_{\{T \leq t\}} \mathbf{1}_{\{D=1\}} = \mathbf{1}_{\{T \leq t\}} \mathbf{1}_{\{T \leq C\}} = D \mathbf{1}_{\{T \leq t\}}$$

Le processus $(N(t), t \geq 0)$ est continu à droite avec limite à gauche et $N(0) = 0$.

Le processus $(Y(t), t \geq 0)$ est continu à gauche avec limite à droite.

Représentation graphique des processus $(N(t), t \geq 0)$ et $(Y(t), t \geq 0)$



La figure 2 donne les représentations graphiques des processus $(N(t), t \geq 0)$ et $(Y(t), t \geq 0)$. Dans les deux premiers graphes la durée de vie T est censurée à $t = 3$ et dans les deux derniers l'évènement est survenu à $t = 4$.

Rappelons que la fonction de risque

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbf{P}(t < T \leq t + dt / T > t)}{dt}$$

et donc pour dt "petit" on peut écrire:

$$\begin{aligned} h(t)dt &\approx \mathbf{P}(t < T \leq t + dt / T > t) \\ &\approx \mathbf{P}(t < T \leq t + dt / T > t, C > t) \end{aligned}$$

Si on note $dN(t)$ l'accroissement du processus N sur $]t, t + dt]$, avec dt "petit" $dN(t)$ ne prend que 0 (pas d'évènement) ou 1 (survenue de l'évènement) et donc

$$\begin{aligned} \mathbf{P}(dN(t) = 1 / T > t, C > t) &= \mathbf{P}(t < T \leq t + dt / T > t, C > t) \\ &\approx h(t)dt \end{aligned}$$

$dN(t)$ est une v.a. de Bernoulli et sa moyenne est $\mathbf{E}(dN(t) / T > t, C > t) = h(t)dt$.

Si on note la tribu $\mathcal{F}_t = \sigma(N(s), \mathbf{1}_{\{X \leq t, D=0\}}, s \leq t)$ et la tribu $\mathcal{F}_{t-} = \sigma(N(s), \mathbf{1}_{\{X \leq t, D=0\}}, s < t)$ (évènements avant t). Alors conditionnellement à \mathcal{F}_{t-} la v.a. $dN(t)$ admet une loi de Bernoulli de paramètre $Y(t)h(t)dt$ (cf [7] p.25):

$$\mathbf{P}(dN(t) = 1 / \mathcal{F}_{t-}) = Y(t)h(t)dt$$

$$\mathbf{P}(dN(t) = 0 / \mathcal{F}_{t-}) = 1 - Y(t)h(t)dt$$

et

$$\mathbf{E}(dN(t) / \mathcal{F}_{t-}) = Y(t)h(t)dt$$

Le processus stochastique défini par

$$M(t) = N(t) - \int_0^t Y(u)h(u)du \quad (1.8)$$

est une martingale par rapport à \mathcal{F}_{t-} et $\int_0^t h(u)Y(u)du$ est le processus croissant de $N(t)$ (cf. corol. 1 Annexe). En effet pour la propriété de martingale on a:

$$\begin{aligned} \mathbf{E}(dM(t) / \mathcal{F}_{t-}) &= \mathbf{E}(dN(t) - Y(t)h(t)dt / \mathcal{F}_{t-}) \\ &= \mathbf{E}(dN(t) / \mathcal{F}_{t-}) - \mathbf{E}(Y(t)h(t)dt / \mathcal{F}_{t-}) \\ &= \mathbf{P}(dN(t) = 1 / \mathcal{F}_{t-}) - Y(t)h(t)dt \\ &= 0 \end{aligned}$$

Les processus

$$(\lambda(t) = Y(t)h(t), t \geq 0) \quad (1.9)$$

et

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t Y(u)h(u)du = \int_0^{t \wedge X} h(u)du = H(t \wedge X) \quad (1.10)$$

sont appelés respectivement processus intensité et processus intensité cumulée du processus $(N(t), t \geq 0)$. Le processus $(\Lambda(t), t \geq 0)$ est appelée "compensateur prévisible" de $N(t)$ (il est

déterminé par \mathcal{F}_{t-}). Enfin on a

$$\mathbf{E}(dN(t) | \mathcal{F}_{t-}) = \lambda(t)dt, \quad \mathbf{V}(dN(t) | \mathcal{F}_{t-}) = (\lambda(t)dt)(1 - \lambda(t)dt) \quad (1.11)$$

Exemple 1.1 Si la durée de survie T suit une loi exponentielle $\varepsilon(\theta)$, alors pour tout $t \geq 0$

$$\begin{aligned} f(t) &= \theta e^{-\theta t}, S(t) = \mathbf{P}(T \geq t) = \int_t^{\infty} \theta e^{-\theta u} du = e^{-\theta t} \\ h(t) &= \frac{f(t)}{S(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta, H(t) = \int_0^t h(u) du = \int_0^t \theta du = \theta t \end{aligned}$$

Le processus intensité cumulée est

$$\Lambda(t) = H(t \wedge T) = \theta(t \wedge T)$$

Dans l'exemple précédent la v.a. T est observée.

Remarque 1.3 On peut définir les processus précédents si nous avons de plus une troncature à gauche U . Par exemple le processus risque devient:

$$Y(t) = \mathbf{1}_{\{U \leq t \leq T \wedge C\}}$$

Dans ce mémoire on se place dans le cadre d'un mécanisme de censure aléatoire à droite seulement.

Notations: Si dans l'étude nous avons n "individus" (machine, patients ...ect) indexés par $i = 1, \dots, n$, on note T_1, T_2, \dots, T_n les instants de survenue de "l'évènement d'intérêt". On observe $X_i = T_i \wedge C_i$ et $D_i = \mathbf{1}_{\{T_i \leq C_i\}}$. Donc on dispose des observations (X_i, D_i) , $i = 1, \dots, n$. A chaque "individu i " correspond le processus risque $Y_i(t)$ défini par

$$Y_i(t) = \mathbf{1}_{\{X_i \geq t\}} \quad (1.12)$$

et le processus ponctuel $N_i(t)$ indicateur de "survenue de l'évènement" dans $]0, t]$

$$N_i(t) = \mathbf{1}_{\{X_i \leq t, D_i=1\}} \quad (1.13)$$

On définit le processus risque $\bar{Y}_n(t)$ par

$$\bar{Y}_n(t) = \sum_{i=1}^n Y_i(t) \quad (1.14)$$

qui est le nombre d'"individus" à risque à l'instant t et en observation. De même on définit le processus $\bar{N}_n(t)$ par

$$\bar{N}_n(t) = \sum_{i=1}^n N_i(t) \quad (1.15)$$

qui est le nombre d'"événements" observés dans l'intervalle $]0, t]$. Le processus à sauts $\bar{Y}_n(t)$ est continu à gauche, les autres processus (intensité....ect...) sont définis de manière similaire.

1.3 Estimateur de Nelson-Aalen

1.3.1 Introduction.

Pour l'individu i , on associe le couple $(N_i(t), Y_i(t), t \geq 0)$ et $(\bar{Y}_n(t), \bar{N}_n(t), t \geq 0)$ définis précédemment. On définit le processus $J_n(t) = \mathbf{1}_{\{\bar{Y}_n(t) > 0\}}$ qui est l'indicateur de présence d'au moins un "individu" à risque. Dans cette partie on prend un taux de hasard $h_i(u) = h(u), \forall i = 1, \dots, n$. Le risque cumulé avec existence d'au moins un "individu" à risque sur $[0, t]$ est:

$$\tilde{H}_n(t) = \int_0^t J_n(u) h(u) du \quad (1.16)$$

On pose

$$\hat{H}_n(t) = \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{N}_n(u) \quad (1.17)$$

qui sera un estimateur de $\tilde{H}_n(t)$. Calculons leur différence :

$$\begin{aligned} \hat{H}_n(t) - \tilde{H}_n(t) &= \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} (d\bar{N}_n(u) - \bar{Y}_n(u) h(u) du) \\ &= \sum_{i=1}^n \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} dM_i(u) \end{aligned}$$

où la martingale

$$M_i(u) = N_i(u) - \int_0^u Y_i(s) h(s) ds \quad (1.18)$$

est d'espérance nulle et elle représente la différence entre le nombre d'évènements observés sur $[0, u]$ et son processus croissant (qui est le nombre d'évènements "prédit" par le modèle pour l'individu i , il est noté: $\Lambda_i(t)$ et est appelé "compensateur prévisible" de $N_i(t)$ puisqu'il est déterminé par \mathcal{F}_{t-} (cf. Coro. 4.1, Annexe).

On pose $\bar{M}_n(u) = \sum_{i=1}^n M_i(u)$

1.3.2 Construction de l'estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen $\hat{H}_{NA}(t)$ du taux de risque cumulé $H(t)$ a été proposé par Nelson en 1972 et ensuite par Aalen en 1978, (cf. [2]). Cet estimateur est fondé sur la remarque suivante pour Δ petit:

$$H(s + \Delta) - H(s) \approx h(s)\Delta \approx \mathbf{P}(s < T \leq s + \Delta \mid T > s) = \frac{\mathbf{P}(s < T \leq s + \Delta)}{\mathbf{P}(T > s)}$$

C'est la probabilité que l'évènement survient dans un petit intervalle $]s, s + \Delta]$ sachant qu'il y a un risque à l'instant s . Cette probabilité est approchée par la fréquence suivante :

$$\frac{[\bar{N}_n(s + \Delta) - \bar{N}_n(s)]}{\bar{Y}_n(s)} = \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} \quad (1.19)$$

où $d\bar{N}_n(s) = \bar{N}_n(s + \Delta) - \bar{N}_n(s)$ c'est le nombre "d'évènements" observés dans l'intervalle $]s, s + \Delta]$ et $\bar{Y}_n(s)$ est le nombre d'"individus" à risque à l'instant s . (cf. (1.14) et (1.15)).

En sommant les quantités (1.19) sur tous les sous intervalles formant $]0, t]$ et en faisons tendre Δ vers 0, de telle sorte que chaque intervalle contienne au plus un "évènement". Ainsi l'estimateur de Nelson-Aalen est défini par

$$\hat{H}_{NA}(t) = \int_0^t \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} \quad (1.20)$$

En effet:

Cet estimateur peut aussi s'écrire si on note t_1, t_2, \dots, t_n les instants ordonnés de survenue

de l'évènement" pour les n "individus" (processus à sauts):

$$\begin{aligned} \int_0^t \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} &= \sum_{\{i, t_i \leq t\}} \int_{t_{i-1}}^{t_i} \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} = \sum_{\{i, t_i \leq t\}} \frac{\bar{N}_n(t_i) - \bar{N}_n(t_{i-1})}{\bar{Y}_n(t_i)} \\ &= \sum_{\{i, t_i \leq t\}} \frac{\Delta \bar{N}_n(t_i)}{\bar{Y}_n(t_i)} = \hat{H}_{NA}(t) \end{aligned} \quad (1.21)$$

où sur $]t_{i-1}, t_i]$ le processus $\bar{Y}_n(s)$ est constant égal à $\bar{Y}_n(t_i)$ et

$$\Delta \bar{N}_n(t_i) = \bar{N}_n(t_i) - \bar{N}_n(t_{i-1})$$

est le nombre "d'évènements" observés dans l'intervalle $]t_{i-1}, t_i]$ i.e. en t_i . On peut directement construire un estimateur du taux de hasard ou risque instantané $h(t_i)$ à l'instant t_i en utilisant la remarque précédente. En effet on pose

$$\hat{h}(t_i) = \frac{m(t_i)}{r(t_i)} \quad (1.22)$$

où $m(t_i)$ désigne le nombre "d'évènements" observés jusqu'à l'instant t_i et $r(t_i)$ est le nombre "d'individus" à risque à l'instant t_i .

Ainsi on estime le taux de risque cumulé $H(t)$ en sommant les risques aux instants $t_i \leq t$ par

$$\hat{H}_{NA}(t) = \sum_{\{i, t_i \leq t\}} \hat{h}(t_i) = \sum_{\{i, t_i \leq t\}} \frac{m(t_i)}{r(t_i)} \quad (1.23)$$

1.3.3 Exemples

Exemple 1. Données des ventilateurs (Données de Aalen)

On étudie la durée de vie de 70 ventilateurs, en milliers d'heures de fonctionnement. La question qui se pose est de savoir si la fonction de risque h est décroissante dans le temps (qui signifie que le risque de panne est décroissant avec le temps et donc une bonne durée de fonctionnement

des ventilateurs). On a le tableau des données suivant:

4.5	4.6 ⁺	11.5	11.5	15.6 ⁺	16.0	16.6 ⁺	18.5 ⁺	18.5 ⁺	18.5 ⁺
18.5 ⁺	18.5 ⁺	20.3 ⁺	20.3 ⁺	20.3 ⁺	20.7	20.7	20.8	22.0 ⁺	30.0 ⁺
30.0 ⁺	30.0 ⁺	30.0 ⁺	31.0	32.0 ⁺	34.5	37.5 ⁺	37.5 ⁺	41.5 ⁺	41.5 ⁺
41.5 ⁺	41.5 ⁺	43.0 ⁺	43.0 ⁺	43.0 ⁺	43.0 ⁺	46.0	48.5 ⁺	48.5 ⁺	48.5 ⁺
48.5 ⁺	50.0 ⁺	50.0 ⁺	50.0 ⁺	61.0 ⁺	61.0	61.0 ⁺	61.0 ⁺	63.0 ⁺	64.5 ⁺
64.5 ⁺	67.0 ⁺	74.5 ⁺	78.0 ⁺	78.0 ⁺	81.0 ⁺	81.0 ⁺	82.0 ⁺	85.0 ⁺	85.0 ⁺
85.0 ⁺	87.5 ⁺	87.5	87.5 ⁺	94.0 ⁺	99.0 ⁺	101.0 ⁺	101.0 ⁺	101.0 ⁺	115.0 ⁺

(+ indique une donnée censurée). Si on note t_1, t_2, \dots, t_n avec $n = 70$, les instants de panne observés. On calcule $\hat{H}_{NA}(t)$, par exemple pour tout $t \geq t_n = 87500$ heures (plus grande valeur) par la formule (1.21):

$$\begin{aligned}
 \hat{H}_{NA}(t) &= \frac{\Delta \bar{N}(t_1)}{\bar{Y}(t_1)} + \dots + \frac{\Delta \bar{N}(t_n)}{\bar{Y}(t_n)} = \frac{\text{nombre de pannes en } t_1}{\text{nombre de ventilateurs à risque en } t_1} \\
 &+ \dots + \frac{\text{nombre de pannes en } t_{70}}{\text{nombre de ventilateurs à risque en } t_{70}} \\
 &= \frac{\text{nombre de pannes en 4.5}}{\text{nombre de ventilateurs à risque en 4.5}} \\
 &+ \dots + \frac{\text{nombre de pannes en 87.5}}{\text{nombre de ventilateurs à risque en 87.5}} \\
 &= \frac{1}{70} + \frac{0}{69} + \frac{2}{68} + \dots + \frac{1}{8} \\
 &= 0.3368
 \end{aligned}$$

Interprétation: On a 33% de taux de risque de tomber en panne pour ce lot de ventilateurs durant la période de fonctionnement (87500 heures).

Remarque 1.4 *Interprétation de l'estimateur de Nelson-Aalen*

1. L'estimateur $\hat{H}_{NA}(t)$ estime le nombre moyen de survenue de l'évènement (pannes, ect...) pour des "individus" à risque sur l'intervalle de temps $(0, t]$. Pour l'exemple de ventilateurs, c'est le nombre moyen de pannes attendu quand un ventilateur fonctionne pendant un intervalle

de temps $(0, t]$ où t est en milliers d'heures en faisant le protocole dit "réparation minimale": en remplaçant chaque fois qu'il tombe en panne par un autre ventilateur qui a le même taux de panne instantané (même caractéristique) que celui qui a été remplacé. Dans le tableau précédent on a un risque de 0.337 de pannes pendant une durée de 87500 heures de fonctionnement. (C'est utilisé en fiabilité).

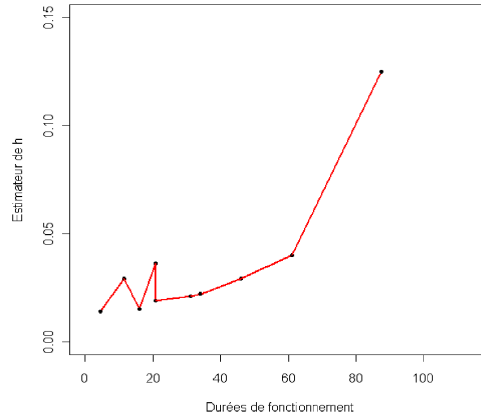
2. Le graphe de \hat{H}_{NA} permet une estimation de la fonction risque h par la tangente au graphe (car $h = H'$). Le problème de l'estimation de h est comparable à celui de l'estimation d'une densité à partir de la fonction de répartition empirique. L'estimateur $\hat{H}_{NA}(t)$ de H est une fonction en escalier et il faut donc la lisser pour obtenir un estimateur de h (cf chap. 3).

Représentation graphique de h .

Des estimations $\hat{h}_n(t_i) = \frac{\Delta \bar{N}(t_i)}{Y(t_i)} = \frac{m_i}{r_i}$ de $h(t_i)$, on peut construire un graphe de h par interpolation linéaire des points $(t_i, \hat{h}_n(t_i))$. On a le tableau suivant:

Durées t_i	m_i	r_i	$\hat{h}_n(t_i)$
4.5	1	70	$1/70 = 0.014$
11.5	2	68	$2/68 = 0.029$
16.0	1	65	$1/65 = 0.015$
20.7	2	55	$2/55 = 0.036$
20.8	1	53	$1/53 = 0.019$
31.0	1	47	$1/47 = 0.021$
34.5	1	45	$1/45 = 0.022$
46.0	1	34	$1/34 = 0.029$
61.0	1	25	$1/25 = 0.04$
87.5	1	8	$1/8 = 0.125$

On obtient le graphe suivant de h :



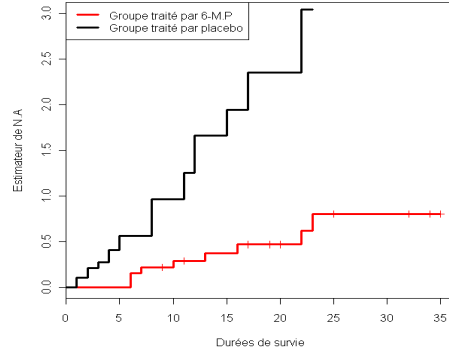
Le graphe donne une fonction de risque h croissante dans le temps

Exemple 2. Données de Freireich.

Les données de rémission de Freireich (rechutes des malades) pour le traitement 6-MP sont présentées dans le tableau suivant avec l'estimateur de Nelson-Aalen \hat{H}_{NA} :

semaines i de rechutes	Durées t_i	$\hat{h}_n(t_i) = \frac{\Delta \bar{N}(t_i)}{Y(t_i)}$	$\hat{H}_{NA}(t)$
1 – 2 – 3	6	3/21	0.143
5	7	1/17	0.201
7	10	1/15	0.268
10	13	1/12	0.351
11	16	1/11	0.442
15	22	1/7	0.585
16	23	1/6	0.752

Représentation graphique Le graphe suivant donne l'estimateur du taux de risque cumulé (Nelson- Aalen et l'empirique) pour les données de Freireich :



La courbe du traitement 6-MP est strictement inférieure à celle de l'effet Placebo d'où l'efficacité du traitement 6-MP.

1.4 Consistance et loi asymptotique de l'estimateur de Nelson-Aalen

Le théorème suivant donne la convergence en probabilité de \hat{H}_{NA} (F continue, ref [7])

Théorème 1.1 Pour $t \geq 0$ on a: $\left| \hat{H}_{NA}(t) - H(t) \right| \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$

Preuve: Dans le cas de présence d'au moins un individu à risque, le taux de hasard cumulé est

$$\tilde{H}_n(t) = \int_0^t J_n(s)h(s)ds$$

et son estimateur de Nelson-Aalen est

$$\hat{H}_{NA}(t) = \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{N}_n(u)$$

On a

$$\hat{H}_{NA}(t) - H(t) = \left(\hat{H}_{NA}(t) - \tilde{H}_n(t) \right) + \left(\tilde{H}_n(t) - H(t) \right)$$

D'autre part (cf 1.18)

$$\begin{aligned}\hat{H}_{NA}(t) - \tilde{H}_n(t) &= \int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{N}_n(s) - \int_0^t J_n(s)h(s)ds \\ &= \int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} (d\bar{N}_n(s) - \bar{Y}_n(s)h(s)ds) = \int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{M}_n(s)\end{aligned}$$

où $\bar{M}_n(s) = \sum_{i=1}^n M_i(s)$

Par l'inégalité de Tchebychev et le Th.2.4.4. p.72 de [7], on obtient

$$\begin{aligned}P\left(\left|\hat{H}_{NA}(t) - \tilde{H}_n(t)\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon^2} E\left(\left|\hat{H}_{NA}(t) - \tilde{H}_n(t)\right|^2\right) \\ &\leq \frac{1}{\varepsilon^2} E\left(\int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{M}_n(s)\right)^2 \\ &\leq \frac{1}{\varepsilon^2} \int_0^t E\left(\frac{J_n(s)}{\bar{Y}_n(s)}\right)^2 d\langle \bar{M}_n \rangle(s)\end{aligned}$$

On applique le théorème de Glivenko-Cantelli pour $\bar{Y}_n(t) = \sum_{i=1}^n \mathbf{1}_{\{X_i \geq t\}}$ et on a

$$\sup_{s \in [0, t]} \left| \frac{\bar{Y}_n(s)}{n} - (1 - F(s))(1 - G(s^-)) \right| \xrightarrow[n \rightarrow +\infty]{\mathbf{P}} 0$$

où F est a f.d.r. des durées de survie T_i , G la f.d.r. des C_i et S_X la fonction de survie du $(X_i = T_i \wedge C_i)$ donnée par

$$S_X(s) = (1 - F(s))(1 - G(s))$$

Par ailleurs $J_n(t) = \mathbf{1}_{\{\bar{Y}_n(t) > 0\}}$, on en déduit que

$$1 - J_n(t) = \mathbf{1}_{\{\bar{Y}_n(t) = 0\}} \xrightarrow[n \rightarrow \infty]{\mathbf{P}} 0$$

En effet

$$\mathbf{P}(\bar{Y}_n(u) = 0) = \mathbf{P}\left(\sum_{i=1}^n Y_i(t) = 0\right)$$

qui est la probabilité d'avoir 0 individus à risque sur $(0, t]$ parmi les n individus.

Or pour un individu

$$\begin{aligned}
\mathbf{P}(Y_i(t) = 1) &= \mathbf{P}(T_i \geq t, C_i \geq t) = \mathbf{P}(T_i \geq t)P(C_i \geq t) \\
&= (1 - F(t^-))(1 - G(t^-)) \\
&= (1 - F(t))(1 - G(t^-)) \quad \text{car } F \text{ est continue}
\end{aligned}$$

Comme

$$\begin{aligned}
\bar{Y}_n(t) &\hookrightarrow B(n, (1 - F(t))(1 - G(t^-))) \\
P(\bar{Y}_n(t) = 0) &= C_n^0 [1 - (1 - F(t))(1 - G(t^-))]^n \xrightarrow{n \rightarrow \infty} 0
\end{aligned}$$

Par suite

$$J_n(t) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 1$$

Par conséquent

$$n \frac{J_n(s)}{\bar{Y}_n(s)} \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} \frac{1}{(1 - F(s))(1 - G(s^-))} > 0$$

D'où

$$\frac{J_n(s)}{\bar{Y}_n(s)} \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0.$$

Par le théorème de la convergence monotone on déduit que

$$\frac{1}{\varepsilon^2} \int_0^t E \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right)^2 d \langle \bar{M}_n \rangle (s) \xrightarrow{n \rightarrow \infty} 0$$

D'où

$$\hat{H}_{NA}(t) - \tilde{H}_n(t) \xrightarrow{P}_{n \rightarrow \infty} 0$$

D'autre part, on a $(1 - J_n(s) \geq 0)$

$$\begin{aligned}
\mathbf{E} \left(\left| \tilde{H}_n(t) - H(t) \right| \right) &\leq \mathbf{E} \left(\int_0^t |(1 - J_n(s))h(s)| ds \right) = \int_0^t \mathbf{E}(1 - J_n(s))h(s) ds \\
&= \int_0^t \mathbf{P}(\bar{Y}_n(u) = 0) h(u) du \\
&= \int_0^t [1 - (1 - F(t))(1 - G(t^-))]^n h(u) du \xrightarrow{n \rightarrow \infty} 0
\end{aligned}$$

D'où

$$\tilde{H}_n(t) - H(t) \xrightarrow{P}_{n \rightarrow \infty} 0$$

D'où le théorème. ■

Le résultat suivant donne la convergence uniforme en probabilité de l'estimateur de \hat{H}_{NA} (cf [7])

Théorème 1.2 *Si pour $t > 0$, $F(t) < 1$ et $\bar{Y}_n(t) \xrightarrow{P}_{n \rightarrow \infty} \infty$ alors*

$$\sup_{0 \leq s \leq t} \left| \hat{H}_{NA}(s) - H(s) \right| \xrightarrow{P}_{n \rightarrow \infty} 0$$

Preuve: Rappelons que $\hat{H}_{NA}(s) = \int_0^s \frac{d\bar{N}(u)}{\bar{Y}_n(u)}$. Soit $s \in [0, t]$. On a:

$$\begin{aligned} \left| \int_0^s \frac{d\bar{N}(u)}{\bar{Y}_n(u)} - H(s) \right| &\leq \left| \int_0^s \frac{d\bar{N}(u)}{\bar{Y}_n(u)} - \int_0^s \mathbf{1}_{\{\bar{Y}_n(u) \geq 0\}} dH(u) \right| \\ &\leq \left| \int_0^s \frac{d\bar{N}(u)}{\bar{Y}_n(u)} - \int_0^s \mathbf{1}_{\{\bar{Y}_n(u) > 0\}} dH(u) \right| + \left| \int_0^s \mathbf{1}_{\{\bar{Y}_n(u) = 0\}} dH(u) \right| \\ &\leq \left| \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} (d\bar{N}(u) - \bar{Y}_n(u)h(u)du) \right| + \left| \int_0^s \mathbf{1}_{\{\bar{Y}_n(u) = 0\}} dH(u) \right| \\ &\leq \left| \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u) \right| + \mathbf{1}_{\{\bar{Y}_n(t) = 0\}} H(t) \end{aligned}$$

On a par hypothèse, $\forall t > 0$, $\bar{Y}_n(t) \xrightarrow{P}_{n \rightarrow \infty} \infty$, ceci entraîne que

$$\forall t > 0, \mathbf{1}_{\{\bar{Y}_n(t) = 0\}} \xrightarrow{P} 0$$

Pour monter que

$$\sup_{0 \leq s \leq t} \left| \int_0^s \frac{d\bar{N}_n(u)}{\bar{Y}_n(u)} - H(s) \right| \xrightarrow{P}_{n \rightarrow \infty} 0$$

montrons que

$$\sup_{0 \leq s \leq t} \left| \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}(u) \right| \xrightarrow{P}_{n \rightarrow \infty} 0$$

Par l'inégalité de Lengart, pour η et ε positifs nous avons

$$\begin{aligned} \mathbf{P} \left(\sup_{0 \leq s \leq t} \left\{ \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u) \right\}^2 \geq \varepsilon \right) &\leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\int_0^t \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} dH(u) \geq \eta \right) \\ &\leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\frac{H(t)}{\bar{Y}_n(t)} \geq \eta \right) \quad (\bar{Y}_n(\cdot) \searrow) \\ &= \frac{\eta}{\varepsilon} + \mathbf{P} \left(\bar{Y}_n(t) \leq \frac{H(t)}{\eta} \right) \end{aligned}$$

Or

$$\bar{Y}_n(t) \xrightarrow{P}_{n \rightarrow \infty} \infty \quad \text{donc} \quad \mathbf{P} \left(\bar{Y}_n(t) \leq \frac{H(t)}{\eta} \right) \rightarrow_{n \rightarrow \infty} 0$$

Comme ε et η sont arbitraires on en déduit

$$\mathbf{P} \left(\sup_{0 \leq s \leq t} \left\{ \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u) \right\}^2 \geq \varepsilon \right) \rightarrow_{n \rightarrow \infty} 0$$

et donc

$$\sup_{0 \leq s \leq t} \left\{ \int_0^s \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u) \right\} \xrightarrow{P}_{n \rightarrow \infty} 0$$

D'où le résultat. ■

Le théorème suivant donne la loi limite de l'estimateur $\hat{H}_{NA}(t)$ (cf [7]).

Théorème 1.3

1. L'estimateur $\hat{H}_{NA}(t)$ est un estimateur avec biais de $H(t)$.
2. Sous l'hypothèse $F(t) < 1$ nous avons

$$\sqrt{n} \left(\hat{H}_{NA}(t) - H(t) \right) \Longrightarrow_{n \rightarrow \infty} U(t)$$

où $U(t)$ est une martingale gaussienne centrée de variance

$$\mathbf{V}(U(t)) = \int_0^t \frac{h(s)}{[1 - F(s)][1 - G(s^-)]} ds$$

et G la f.d.r de C .

Preuve:

1. Le biais de $\hat{H}_{NA}(t)$ est donné grace à (1.8) et à la martingale

$$\bar{M}_n(u) = \sum_{i=1}^n M_i(u)$$

On a

$$\begin{aligned} \mathbf{E}(\hat{H}_{NA}(t)) &= \mathbf{E} \left(\int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{N}_n(u) \right) \\ &= \mathbf{E} \left(\int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} [d\bar{M}_n(u) + \bar{Y}_n(u)h(u)du] \right) \\ &= \mathbf{E} \left(\int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{M}_n(u) \right) + \mathbf{E} \left(\int_0^t J_n(u)h(u)du \right) \\ &: = E(M^*(t)) + \mathbf{E} \left(\int_0^t J_n(u)h(u)du \right) \\ &= E(M^*(0)) + \mathbf{E} \left(\int_0^t J_n(u)h(u)du \right) \quad (M^*(t) \text{ est une martingale}) \\ &= \int_0^t \mathbf{E}[J_n(u)]h(u)du \\ &= \int_0^t \mathbf{P}(\bar{Y}_n(u) > 0) h(u)du \\ &= \int_0^t h(u)du - \int_0^t \mathbf{P}(\bar{Y}_n(u) = 0) h(u)du \\ &= H(t) - \int_0^t \mathbf{P}(\bar{Y}_n(u) = 0) h(u)du \end{aligned}$$

Soit F la f.d.r des durées de survie T_i et G la f.d.r des C_i . On a vu au th1.1 que

$$\int_0^t \mathbf{P}(\bar{Y}_n(u) = 0) h(u)du \rightarrow_{n \rightarrow \infty} 0$$

D'où le résultat.

2. La fonction de survie S_X des X_i est

$$S_X(t) = (1 - F(t))(1 - G(t)).$$

Par le théorème de Glivenko-Cantelli pour $\bar{Y}_n(t) = \sum_{i=1}^n \mathbf{1}_{\{X_i \geq t\}}$ on a

$$\sup_{s \in [0, t]} \left| \frac{\bar{Y}_n(s)}{n} - (1 - F(s))(1 - G(s^-)) \right| \xrightarrow[n \rightarrow +\infty]{\mathbf{P}} 0$$

Par ailleurs $J_n(t) = \mathbf{1}_{\{\bar{Y}_n(t) > 0\}} \xrightarrow[n \rightarrow \infty]{\mathbf{P}} 1$ (cf. th1.1).

Or par (1.8) on a

$$dM(t) = dN(t) - Y(t)h(t)dt$$

et de

$$d \langle M \rangle_t = \text{var}(dM(t) / \mathcal{F}_{t-})$$

on obtient par (1.11)

$$\begin{aligned} d \langle M \rangle_t &= \text{var}(dN(t) - Y(t)h(t)dt / \mathcal{F}_{t-}) \\ &= \text{var}(dN(t) / \mathcal{F}_{t-}) = Y(t)h(t)dt \end{aligned}$$

Par conséquent pour

$$L_t = \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} dM_i(u)$$

son processus croissant est

$$\langle L \rangle_t = \int_0^t \frac{J_n^2(u)}{\bar{Y}_n^2(u)} d \langle M \rangle_u$$

D'une part,

$$\begin{aligned} \langle \sqrt{n} \sum_{i=1}^n \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} dM_i(u) \rangle_t &= \sum_{i=1}^n n \int_0^t \frac{J_n^2(u)}{\bar{Y}_n^2(u)} Y_i(u) h(u) du \\ &= \int_0^t n \frac{J_n^2(u)}{\bar{Y}_n^2(u)} \sum_{i=1}^n Y_i(u) h(u) du = \int_0^t n \frac{J_n^2(u)}{\bar{Y}_n^2(u)} \bar{Y}_n(u) h(u) du \\ &= \int_0^t n \frac{J_n(u)}{\bar{Y}_n(u)} h(u) du \end{aligned}$$

D'autre part, le processus

$$\sqrt{n} \left(\hat{H}_{NA}(t) - H(t) \right) = \sqrt{n} \sum_{i=1}^n \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} dM_i(u)$$

est une martingale (cf. th.4.3 Annexe) que l'on notera $Z_n(t)$. On applique le théorème central limite pour les martingales donné par:

Théorème de Rebolledo. Soit $\bar{M}_n = \sum_{i=1}^n M_i$ une suite de martingales où $M_i = N_i - A_i$ où N_i désigne un processus de comptage et A_i son compensateur. Considérons les processus

$$I_n(t) = \int_0^t f_n(s) d\bar{M}_n(s)$$

et pour $\varepsilon > 0$

$$I_{n,\varepsilon}(t) = \int_0^t f_n(s) \mathbf{1}_{\{|f_n(s)| > \varepsilon\}} d\bar{M}_n(s)$$

Supposons que f_n et f sont des processus \mathcal{F}_{s-} prévisibles et localement bornés tels que

$$\sup_s |f_n(s) - f(s)| \xrightarrow{n \rightarrow \infty} 0$$

Supposons également que les processus N_i, A_i, f_n sont bornés. Posons pour tout $t > 0$

$$\alpha(t) = \int_0^t f^2(s) ds < \infty$$

Si

$$\langle I_n \rangle_t \xrightarrow[n \rightarrow \infty]{P} \alpha(t)$$

pour tout $\varepsilon > 0$

$$\langle I_{n,\varepsilon} \rangle_t \xrightarrow[n \rightarrow \infty]{P} 0$$

Alors

$$(I_n(t), t > 0) \xRightarrow{n \rightarrow \infty} \left(\int_0^t f(s) dW(s), t > 0 \right)$$

où $\xRightarrow{\quad}$ désigne la convergence faible dans l'espace des fonctions continues à droite, ayant une limite à gauche muni de la topologie de Skorokhod et où W est un mouvement brownien.

Vérifions les conditions précédentes pour la martingale $Z_n(t) = \sqrt{n} \left(\hat{H}_{NA}(t) - H(t) \right)$.

1. On a

$$\langle Z_n \rangle_t = \int_0^t n \frac{J(u)}{Y(u)} h(u) du$$

et aussi par la preuve du th 1.1

$$\frac{\bar{Y}_n(u)}{n} \xrightarrow[n \rightarrow \infty]{P} (1 - F(u))(1 - G(u^-)), \quad J_n(u) \xrightarrow[n \rightarrow \infty]{P} 1$$

Donc

$$\langle Z_n \rangle_t = \int_0^t \frac{J_n(u)}{\frac{\bar{Y}_n(u)}{n}} h(u) du \xrightarrow[n \rightarrow \infty]{} \int_0^t \frac{h(u) du}{(1 - F(u))(1 - G(u^-))} = \alpha(t)$$

qui est déterministe. D'où la 1^{ère} condition.

2. Pour la 2^{ème} condition posons

$$Z_{n,\varepsilon}(t) = \int_0^t \sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)} \mathbf{1}_{\{|\sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)}| > \varepsilon\}} d\bar{M}_n(u)$$

Montrons que $\langle Z_{n,\varepsilon} \rangle_{t \rightarrow n \rightarrow \infty} 0$ en probabilité. On a

$$\begin{aligned} \langle Z_{n,\varepsilon} \rangle_t &= \int_0^t n \frac{J_n(u)}{\bar{Y}_n(u)^2} \mathbf{1}_{\{|\sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)}| > \varepsilon\}} d\bar{M}_n \rangle_u \\ &= \int_0^t n \frac{J_n(u)}{\bar{Y}_n(u)^2} \mathbf{1}_{\{|\sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)}| > \varepsilon\}} \bar{Y}_n(u) h(u) du \\ &= \int_0^t n \frac{J_n(u)}{\bar{Y}_n(u)} \mathbf{1}_{\{|\sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)}| > \varepsilon\}} h(u) du \xrightarrow[n \rightarrow \infty]{P} 0 \end{aligned}$$

car

$$n \frac{J_n(u)}{\bar{Y}_n(u)} \xrightarrow[n \rightarrow \infty]{P} \frac{1}{(1 - F(u))(1 - G(u^-))}$$

Donc

$$\sqrt{n} \frac{J_n(u)}{\bar{Y}_n(u)} = \frac{1}{\sqrt{n}} n \frac{J_n(u)}{\bar{Y}_n(u)} \xrightarrow[n \rightarrow \infty]{P} 0$$

D'où la 2^{ème} condition.

Les conditions du Th. de Rebolledo sont vérifiées par conséquent:

$$(Z_n(t), t > 0) \xrightarrow[n \rightarrow \infty]{} \left(\int_0^t f(s) dW(s), t > 0 \right)$$

avec $\alpha(t) = \int_0^t f^2(s) ds$.

D'où pour tout $t > 0$

$$Z_n(t) \xrightarrow{n \rightarrow \infty} Z(t) = \int_0^t f(s) dW(s) \hookrightarrow N(0, \alpha(t))$$

D'où le résultat du théorème. ■

1.5 Estimateur de Kaplan-Meier

La fonction de survie S est *inconnue* et elle caractérise la loi de T . On va construire un estimateur en utilisant les observations

$$(X_i, D_i) = (T_i \wedge C_i, \mathbf{1}_{\{T_i \leq C_i\}}) \quad i = 1, \dots, n$$

Si les données ne sont pas censurées, $S(t)$ peut être estimée par la f.d.r. empirique

$$S_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{T_i > t\}} = 1 - F_n(t) \quad (1.24)$$

où $F_n(t)$ est la f.d.r. empirique des T_i . Si on dispose d'observations (T_i) avec présence de censures, estimer S seulement par les données $(T_i)_{i=1, \dots, n}$ non censurées ($D_i = 1$) fournit une estimation avec biais. En effet si on définit un estimateur \tilde{S}_n

$$\tilde{S}_n(t) = \frac{\sum_{i=1}^n \mathbf{1}_{\{T_i > t, D_i = 1\}}}{\sum_{i=1}^n \mathbf{1}_{\{D_i = 1\}}} \quad (1.25)$$

alors par la loi forte des grands nombres on a

$$\tilde{S}_n(t) = \frac{\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{T_i > t, D_i = 1\}}}{\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{D_i = 1\}}} \xrightarrow{n \rightarrow +\infty} \frac{\mathbf{P}(T_i > t, D_i = 1)}{\mathbf{P}(D_i = 1)} \neq S(t)$$

car

$$\begin{aligned}
\frac{\mathbf{P}(T_i > t, D_i = 1)}{\mathbf{P}(D_i = 1)} &= \frac{\int_t^\infty (1 - G(u)) f(u) du}{\mathbf{P}(T_i \leq C_i)} = \frac{\int_t^\infty (1 - G(u)) f(u) du}{\int_0^\infty (1 - G(u)) f(u) du} \\
&= \frac{\int_t^\infty f(u) du - \int_t^\infty G(u) f(u) du}{1 - \int_0^\infty G(u) f(u) du} \neq \int_t^\infty f(u) du = S(t)
\end{aligned}$$

Ainsi dans le cas de données censurées, l'estimation de la fonction de survie S nécessite des outils spécifiques. Il existe plusieurs méthodes non paramétriques parmi elles la plus connue est celle de *Kaplan-Meier*. D'abord on définit l'estimateur de *Fleming-Harrington* qui se déduit immédiatement de l'estimation du taux cumulé.

1.5.1 Estimateur de Fleming-Harrington

Par la relation $S = \exp(-H)$ et en estimant H par l'estimateur de *Nelson-Aalen* \hat{H}_{NA} on obtient l'estimateur de Fleming-Harrington de S donné par

$$\hat{S}_{FH}(t) = \exp(-\hat{H}_{NA}(t)) \quad (1.26)$$

Il peut s'écrire aussi par (??)

$$\hat{S}_{FH}(t) = \exp\left(-\sum_{\{i, t_i \leq t\}} \frac{m(t_i)}{r(t_i)}\right) \quad (1.27)$$

1.5.2 Construction de l'estimateur de Kaplan-Meier

Kaplan et Meier (1958), (cf. [12]) ont proposé un estimateur de S nommé aussi estimateur produit-limite. Il repose sur l'idée suivante: un "individu" est en vie après l'instant t , c'est être en vie juste avant l'instant t et ne pas "mourir" en t . Cette idée se traduit par les relations

suivantes:

$$\begin{aligned}
 S(t) &= \mathbf{P}(T > t) \\
 &= \mathbf{P}(T > t/T > t - 1)\mathbf{P}(T > t - 1) \\
 &= \dots \\
 &= \mathbf{P}(T > t/T > t - 1) \dots \mathbf{P}(T > 1/T > 0)\mathbf{P}(T > 0)
 \end{aligned}$$

Si l'on choisit les instants de conditionnement où il se produit un évènement t_i (mort, panne ou censure ...) on aura à estimer des quantités de la forme:

$$\mathbf{P}(T > t_{(i)} / T > t_{(i-1)}) = p_i$$

où les $t_{(i)} < t_{(i-1)}$ et p_i est la probabilité de survivre pendant l'intervalle de temps $I_i =]t_{(i-1)}, t_{(i)}]$ sachant qu'on était "vivant" au début de cet intervalle.

Notons comme précédemment, r_i le nombre des sujets qui sont "vivants" (donc à risque) juste avant l'instant $t_{(i)}$ et m_i le nombre d'"évènements" à l'instant $t_{(i)}$. Or $q_i = 1 - p_i$ est la probabilité de la survenue de l'évènement" durant I_i . Un estimateur naturel de q_i est la fréquence

$$\hat{q}_i = \frac{m_i}{r_i} \tag{1.28}$$

Si on suppose qu'il n'y ait pas d'ex-aequo (plusieurs pannes en $t_{(i)}$): si $D_i = 1$, c'est qu'il y a eu un "évènement" en $t_{(i)}$ et donc $m_i = 1$, si $D_i = 0$, c'est qu'il y a eu une censure en $t_{(i)}$ et donc $m_i = 0$.

Par suite

$$\begin{aligned}
 \hat{p}_i &= 1 - \hat{q}_i = 1 - \frac{1}{r_i} \text{ si } D_i = 1 \\
 &= 1 \text{ sinon}
 \end{aligned}$$

On a $r_i = n - (i - 1)$ (car il y'a eu $i - 1$ "évènements" ou censures avant $t_{(i)}$ et il y'a n

individus dans l'étude). L'estimateur de *Kaplan-Meier* dans ce cas est

$$\hat{S}_{KM}(t) = \prod_{X_{(i)} \leq t} \left(1 - \frac{1}{n - i + 1}\right)^{D_{(i)}} \quad (1.29)$$

Remarque 1.5 1. *L'estimateur de Kaplan-Meier est une fonction en escaliers qui fait des sauts à chaque instant t_i . La valeur du saut dépend du nombre d'évènements au temps t_i et aussi du nombre de censures à ce temps là.*

2. *Les estimateurs $\hat{S}_{KM}(t)$ et $\hat{S}_{FH}(t)$ sont nuls dès que $\bar{Y}_n(t) = 0$.*

3. *Traitement des ex-aequo: Dans le cas d'existence des ex-aequo on n'a plus m_i égale à 1 en $t_{(i)}$ mais au nombre d'"évènements" en $t_{(i)}$ et aussi on n'a plus $r_i = n - (i - 1)$. Dans ce cas on doit garder r_i et m_i et l'estimateur de Kaplan-Meier devient*

$$\hat{S}_{KM}(t) = \prod_{X_{(i)} \leq t} \left(1 - \frac{m_i}{r_i}\right)^{D_{(i)}} \quad (1.30)$$

4. *L'estimateur de Kaplan-Meier est aussi l'estimateur du maximum de vraisemblance non paramétrique de la survie dans l'espace des fonctions de répartition (de dimension infinie) (cf. [6])*

5. *Les estimateurs de Kaplan-Meier et de Neslon-Aalen sont continus à droite par construction (cf. Figure 3 et Figure 4).*

1.5.3 Exemples

Exemple 1

Dans l'exemple des données de Freireich, l'estimateur (1.24) pour le groupe traité par un placebo (pas de censure) donne le tableau suivant :

Semaine i	Nombre de rémissions à la semaine i	$\hat{S}_{\text{placebo}}(t)$
0	21	1
1	19	$19/21 = 0.90$
2	17	$17/21 = 0.81$
3	16	0.76
4	14	0.66
5	12	0.57
8	8	0.38
11	6	0.28
12	4	0.19
15	3	0.14
17	2	0.09
22	1	0.05
23	0	0

Exemple 2

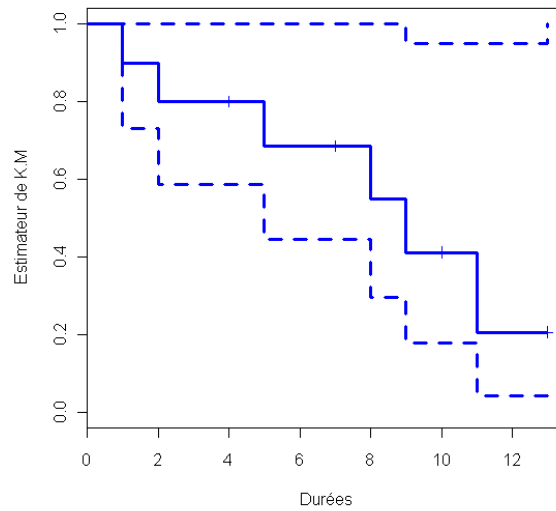
On observe les durées de vie de 10 diodes exprimées en mois:

1 3 4⁺ 5 7⁺ 8 9 10⁺ 11 13⁺

L'estimateur de *Kaplan-Meier* de S dans cet exemple se calcule par la formule (1.29) :

Temps	r_i	m_i	$\hat{S}_{KM}(t_i)$	Intervalle
0	10	0	1	$[0, 1[$
1	10	1	$(1 - 1/10) \hat{S}(0) = 0.900$	$[1, 3[$
3	9	1	$(1 - 1/9) \hat{S}(1) = 0.800$	$[3, 5[$
5	7	1	$(1 - 1/7) \hat{S}(3) = 0.686$	$[5, 8[$
8	5	1	$(1 - 1/5) \hat{S}(5) = 0.589$	$[8, 9[$
9	4	1	$(1 - 1/4) \hat{S}(8) = 0.411$	$[9, 11[$
11	2	1	$(1 - 1/2) \hat{S}(9) = 0.206$	$[11, \infty[$

Représentation graphique le graphe suivant présente l'estimation de la fonction de survie par Kaplan-Meier (1.29) et les intervalles de confiance à un niveau de 95%.



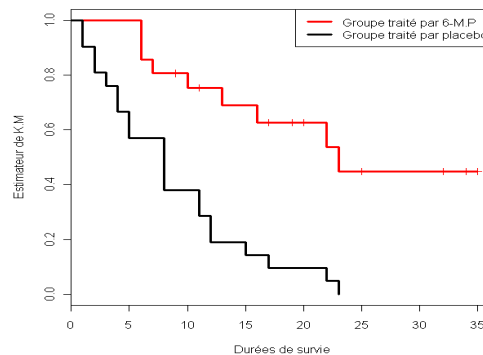
par exemple $\hat{S}_{KM}(10) = 0.4$ ie on peut dire que 40% des diodes ont une durée de fonctionnement supérieure à 10 mois.

Exemple 3

L'estimateur de *Kaplan-Meier* (1.30) de la fonction de survie S du groupe de 21 malades traité par le traitement 6-MP (existence d'ex-aequo) donne le tableau suivant :

Temps t_i	r_i	m_i	$\hat{S}_{6-MP}(t_i)$	Intervalle
0	21	0	1	$[0, 6[$
6	21	3	$(1 - 3/20) * 1 = 0.857$	$[6, 7[$
7	17	1	$(1 - 1/17) * 0,857 = 0.807$	$[7, 10[$
10	15	1	$(1 - 1/15) * 0.807 = 0.753$	$[10, 13[$
13	12	1	$(1 - 1/12) * 0.753 = 0.690$	$[13, 16[$
16	11	1	$(1 - 1/11) * 0.690 = 0.627$	$[16, 22[$
22	7	1	$(1 - 1/7) * 0.627 = 0.538$	$[22, 23[$
23	6	1	$(1 - 1/6) * 0.538 = 0.448$	$[23, \infty[$

Représentations graphiques Le graphe suivant présente les estimateurs de la fonction de survie par Kaplan-Meier (1.30) et par (1.24) (pour les deux traitements (cf. ex. 1 et ex. 3))



Nous remarquons que la courbe de survie 6 MP est nettement supérieur à celle du Placebo. D'où l'efficacité du traitement 6MP.

1.5.4 Consistance et loi asymptotique de l'estimateur de Kaplan-Meier

Le résultat suivant donne la consistance uniforme de l'estimateur \hat{S}_{KM} de *Kaplan-Meier*.

Théorème 1.4 Si pour $t > 0$, $F(t) < 1$ et $\bar{Y}_n(t) \xrightarrow{P_{n \rightarrow \infty}} \infty$ alors

$$\sup_{0 \leq s \leq t} \left| \hat{S}_{KM}(s) - S(s) \right| \xrightarrow{P_{n \rightarrow \infty}} 0$$

Preuve: L'estimateur de Kaplan-Meier vérifie $\hat{S}_{KM}(s) = 1 - \hat{F}_n(s)$ où \hat{F}_n est définie par:

$$\hat{F}_n(t) = \int_0^t \hat{S}_{KM}(s^-) d\hat{H}_{NA}(s) \quad (\text{cf. chap. 1 p. 55})$$

Il suffit de montrer que

$$\sup_{0 \leq s \leq t} \left| \hat{F}_n(s) - F(s) \right| \xrightarrow{P_{n \rightarrow \infty}} 0.$$

Soit $t > 0$, $F(t) < 1$ et $\bar{Y}_n(t) \xrightarrow{P_{n \rightarrow \infty}} \infty$, le th.3.2.3 p. 97 de [7], donne

$$\mathbf{P} \left(\frac{\hat{F}_n(s) - F(s)}{1 - F(s)} = Z_n(s), s \in [0, t] \right) \rightarrow_{n \rightarrow \infty} 1$$

où

$$Z_n(s) = \int_0^s \frac{1 - \hat{F}_n(u^-)}{1 - F(u)} \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u)$$

Il suffit maintenant de montrer que

$$\sup_{0 \leq s \leq t} |Z_n(s)| \xrightarrow{\mathbf{P}_{n \rightarrow \infty}} 0$$

pour avoir

$$\sup_{0 \leq s \leq t} \left| \hat{F}_n(s) - F(s) \right| \xrightarrow{P_{n \rightarrow \infty}} 0$$

Par l'inégalité de Lenggart, pour η et ε positifs nous avons

$$\begin{aligned}
\mathbf{P} \left(\sup_{0 \leq s \leq t} |Z_n(s)|^2 \geq \varepsilon \right) &\leq \mathbf{P} \left(\sup_{0 \leq s \leq t} \left| \int_0^s \frac{1 - \hat{F}_n(u^-)}{1 - F(u)} \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} d\bar{M}_n(u) \right|^2 \geq \varepsilon \right) \\
&\leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\int_0^t \left(\frac{1 - \hat{F}_n(u^-)}{1 - F(u)} \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} \right)^2 d \langle \bar{M}_n \rangle_u \geq \eta \right) \\
&\leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\int_0^s \left(\frac{1 - \hat{F}_n(u^-)}{1 - F(u)} \right)^2 \frac{\mathbf{1}_{\{\bar{Y}_n(u) > 0\}}}{\bar{Y}_n(u)} dH(u) \geq \eta \right) \\
&\leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\frac{1}{(1 - F(t))^2} \frac{1}{\bar{Y}_n(t)} H(t) \geq \eta \right) \\
&= \frac{\eta}{\varepsilon} + \mathbf{P} \left(\bar{Y}_n(t) \leq \frac{H(t)}{\eta(1 - F(t))^2} \right)
\end{aligned}$$

Or

$$\bar{Y}_n(t) \xrightarrow[n \rightarrow \infty]{P} \infty$$

donc

$$\mathbf{P} \left(\bar{Y}_n(t) \leq \frac{H(t)}{\eta(1 - F(t))^2} \right) \xrightarrow[n \rightarrow \infty]{} 0$$

Comme ε et η sont arbitraires on en déduit

$$\mathbf{P} \left(\sup_{0 \leq s \leq t} |Z_n(s)|^2 \geq \varepsilon \right) \xrightarrow[n \rightarrow \infty]{} 0$$

Par conséquent

$$\sup_{0 \leq s \leq t} \left| \hat{F}_n(s) - F(s) \right| \xrightarrow[n \rightarrow \infty]{P} 0$$

D'où le résultat. ■

Nous avons une loi limite de l'estimateur de Flemming-Harrington \hat{S}_{FH}

Théorème 1.5 *L'estimateur \hat{S}_{FH} possède un biais.*

Si nous supposons que:

1. *pour tout $s \in [0, t]$*

$$n \int_0^s \frac{J_n(u)}{\bar{Y}_n(u)} h(u) du \xrightarrow[n \rightarrow \infty]{P} \sigma^2(s)$$

2. pour tout $\varepsilon > 0$

$$n \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} h(u) \mathbf{1}_{\{\sqrt{n} \left| \frac{J_n(u)}{\bar{Y}_n(u)} \right| > \varepsilon\}} du \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

3.

$$\sqrt{n} \int_0^t (1 - J_n(u)) h(u) du \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

alors pour tout $t > 0$

$$\sqrt{n} \left(\hat{S}_{FH}(t) - S(t) \right) \Longrightarrow_{n \rightarrow \infty} -U(t)S(t)$$

où U est une martingale gaussienne centrée

Preuve: Rappelons que $\hat{S}_{FH}(t) = \exp\left(-\hat{H}_{NA}(t)\right)$ où $\hat{H}_{NA}(t) = \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{N}_n(u)$

A). Calcul du biais: posons $\tilde{S}_n(t) = \exp\left(-\tilde{H}_n(t)\right)$ où $\tilde{H}_n(t) = \int_0^t J_n(u) h(u) du$

Nous avons pour $t \in [0, \tau[$ et $\tau > 0$ (en appliquant le th. 4.7 Annexe)

$$\begin{aligned} \frac{\hat{S}_{FH}(t)}{\tilde{S}_n(t)} - 1 &= - \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} d\left(\hat{H}_{NA} - \tilde{H}_n\right)(u) \\ &= - \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} d\left(\hat{H}_{NA} - \tilde{H}_n\right)(u) \\ &= - \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} d\left(\sum_{i=1}^n \int_0^u \frac{J_n(s)}{\bar{Y}_n(s)} dM_i(s)\right) \\ &= - \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} \frac{J_n(u)}{\bar{Y}_n(u)} d\left(\sum_{i=1}^n M_i(u)\right) \\ &= - \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{M}_n(u) \end{aligned} \tag{1.30}$$

où $d\bar{M}_n(t) = d\bar{N}_n(t) - \bar{Y}_n(t)h(u)du$. Puisque $\frac{\hat{S}_{FH}(t)}{\tilde{S}_n(t)} - 1$ est une martingale locale de carré intégrable sur $[0, \tau[$ nulle en $t = 0$ (cf. [13]), On en déduit que

$$\mathbf{E} \left(\frac{\hat{S}_{FH}(t)}{\tilde{S}_n(t)} \right) = 1 \text{ pour tout } t \in [0, \tau[$$

On a

$$S(t) \leq \tilde{S}_n(t)$$

Donc

$$\frac{\hat{S}_{FH}(t)}{\tilde{S}(t)} \leq \frac{\hat{S}_{FH}(t)}{S(t)}$$

et alors

$$\mathbf{E} \left(\frac{\hat{S}_{FH}(t)}{\tilde{S}_n(t)} \right) \leq \mathbf{E} \left(\frac{\hat{S}_{FH}(t)}{S(t)} \right)$$

D'où: $1 \leq \frac{\mathbf{E}(\hat{S}_{FH}(t))}{S(t)}$ ou encore $\mathbf{E}(\hat{S}_{FH}(t)) \geq S(t)$.

Par suite le biais est: $\mathbf{E}(\hat{S}_{FH}(t)) - S(t) \geq 0$

B). Pour la convergence en loi vérifions les conditions du Th. de Robolledo:

par la condition **1** on a

$$\left\langle \sqrt{n} \left(\frac{\hat{S}_{FH}}{\tilde{S}_n} - 1 \right) \right\rangle_t = n \int_0^t \frac{\hat{S}_{FH}(u^-)^2}{\tilde{S}_n(u)^2} \frac{J_n(u)}{\bar{Y}_n(u)} h(u) du \xrightarrow[n \rightarrow \infty]{P} \sigma^2(s)$$

Par la preuve du Th.1.1 on en déduit $\frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} \xrightarrow[n \rightarrow \infty]{P} 1$. D'où la 1^{ère} condition du Th. de Robolledo.

Pour la 2^{ème} condition du Th. de Robolledo, la condition **2** et de façon analogue à la preuve du Th1.3 (2^{ème} condition) on trouve que

$$n \int_0^t \frac{\hat{S}_{FH}(u^-)^2}{\tilde{S}_n(u)^2} \frac{J(u)}{\bar{Y}_n(u)} \mathbf{1}_{\left\{ \sqrt{n} \left| \frac{J(u)}{\bar{Y}_n(u)} \right| > \varepsilon \right\}} h(u) du \rightarrow_{n \rightarrow \infty} 0$$

Donc pour chaque $t > 0$

$$\sqrt{n} \int_0^t \frac{\hat{S}_{FH}(u^-)}{\tilde{S}_n(u)} \frac{J(u)}{\bar{Y}_n(u)} d\bar{M}_n(u) \implies U(t)$$

où $U(t) \hookrightarrow N(0, \sigma^2(t))$.

Enfin

$$\sqrt{n} \left(\frac{\hat{S}_{FH}(t)}{\tilde{S}_n(t)} - 1 \right) \implies -U(t)$$

Le fait que $S(u) \leq \tilde{S}_n(u)$, pour $u \in [0, s[$ et la condition **3** implique

$$\begin{aligned} \sqrt{n} \left| \frac{S(t)}{\tilde{S}_n(t)} - 1 \right| &\leq \sqrt{n} \int_0^t \frac{S(u)}{\tilde{S}_n(u)} d(H - \tilde{H}_n)(u) \\ &\leq \sqrt{n} \int_0^t (1 - J(u)) h(u) du \xrightarrow[n \rightarrow \infty]{P} 0 \end{aligned}$$

comme $\tilde{S}_n(t) \xrightarrow[n \rightarrow \infty]{} S(t)$ on en déduit que

$$\sqrt{n} (\tilde{S}_n(t) - S(t)) \xrightarrow[n \rightarrow \infty]{P} 0$$

Il s'ensuit que

$$\begin{aligned} \sqrt{n} (\hat{S}_{FH}(t) - S(t)) &= \sqrt{n} (\hat{S}_{FH}(t) - \tilde{S}_n(t)) + \sqrt{n} (\tilde{S}_n(t) - S(t)) \\ &= \frac{\sqrt{n} (\hat{S}_{FH}(t) - \tilde{S}_n(t))}{\tilde{S}_n(t)} \tilde{S}_n(t) + \sqrt{n} (\tilde{S}_n(t) - S(t)) \\ &\Rightarrow_{n \rightarrow \infty} -U(t)S(t) \end{aligned}$$

D'où le résultat du théorème. ■

1.6 Estimateurs des variances de \hat{S}_{KM} et \hat{S}_{FH}

On suppose que chaque individu i à un taux de hasard h_i et on considère les observations

$$(N_i(t), Y_i(t), \quad t > 0) \quad i = 1, \dots, n$$

Soit \hat{h}_i l'estimateur de h_i défini par (1.22). Ces estimateurs sont des estimateurs du maximum de vraisemblance et la théorie asymptotique de ces estimateurs donne:

$$\text{cov}(\hat{h}_i, \hat{h}_j) \approx \frac{1}{I^{ij}(h)}$$

où $h = (h_1, h_2, \dots, h_k)$ et $I^{ij}(h)$ est l'élément (i, j) de la matrice d'information de Fisher $I(h)$ qui peut être estimée par $I(\hat{h})$ avec $\hat{h} = (\hat{h}_1, \hat{h}_2, \dots, \hat{h}_k)$

$$I(\hat{h}) = - \frac{\partial^2 \ell}{\partial h_i \partial h_j} \Big|_{\hat{h}} = \begin{cases} \frac{r_i}{\hat{h}_i(1-\hat{h}_i)} & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

où ℓ est le log de la vraisemblance. Ainsi

$$\text{cov}(\hat{h}_i, \hat{h}_j) = \begin{cases} \frac{\hat{h}_i(1-\hat{h}_i)}{r_i} & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad (1.31)$$

et les estimateurs $(\hat{h}_i)_{i=1,\dots,k}$ sont asymptotiquement indépendants.

Pour estimer la variance des estimateurs \hat{S}_{KM} et \hat{S}_{FH} , on utilise la formule suivante: pour une fonction différentiable g on a:

$$g(Y) \approx g(\mu) - g'(\mu)(Y - \mu)$$

avec $\mu = \mathbf{E}(Y)$ et en prenant la variance des deux côtés on obtient une approximation d'ordre 1

$$\mathbf{V}(g(Y)) \approx [g'(\mu)]^2 \mathbf{V}(Y)$$

La delta-méthode consiste à remplacer μ par Y et on a

$$\mathbf{V}(g(Y)) \approx [g'(Y)]^2 \mathbf{V}(Y) \quad (1.32)$$

D'où un estimateur de la variance $\mathbf{V}(g(Y))$ donné par $[g'(Y)]^2 \mathbf{V}(Y)$

D'où pour l'estimateur de Kaplan Meier, on applique la dernière formule pour $g(t) = \exp t$:

$$\begin{aligned} \mathbf{V}(\hat{S}_{KM}(t)) &= \mathbf{V}(\exp\{\ln \hat{S}_{KM}(t)\}) \\ &\approx \left\{ \exp(\ln \hat{S}_{KM}(t)) \right\}^2 \mathbf{V}(\ln \hat{S}_{KM}(t)) \\ &= [\hat{S}_{KM}(t)]^2 \mathbf{V} \left(\ln \prod_{\{i, t_i < t\}} (1 - \hat{h}_i) \right) \\ &= [\hat{S}_{KM}(t)]^2 \mathbf{V} \left(\sum_{\{i, t_i < t\}} \ln(1 - \hat{h}_i) \right) \end{aligned}$$

Par l'indépendance asymptotique des \hat{h}_i on obtient:

$$\mathbf{V}(\hat{S}_{KM}(t)) \approx [\hat{S}_{KM}(t)]^2 \sum_{\{i, t_i < t\}} \mathbf{V}(\ln(\mathbf{1} - \hat{h}_i))$$

En appliquant la delta-méthode une deuxième fois pour $g(t) = \ln t$, on obtient

$$\begin{aligned} \mathbf{V}(\hat{S}_{KM}(t)) &\approx [\hat{S}_{KM}(t)]^2 \sum_{\{i, t_i < t\}} \frac{1}{(\mathbf{1} - \hat{h}_i)^2} \mathbf{V}(\hat{h}_i) \\ &= [\hat{S}_{KM}(t)]^2 \sum_{\{i, t_i < t\}} \frac{1}{(\mathbf{1} - \hat{h}_i)^2} \frac{\hat{h}_i(1 - \hat{h}_i)}{r_i} \\ \mathbf{V}(\hat{S}_{KM}(t)) &\approx [\hat{S}_{KM}(t)]^2 \sum_{\{i, t_i < t\}} \frac{m_i}{r_i(r_i - m_i)} \end{aligned} \quad (1.33)$$

Pour l'estimateur de Fleming-Harrington on a:

$$\begin{aligned} \mathbf{V}(\hat{S}_{FH}(t)) &= \mathbf{V}(\exp\{-\hat{H}_{NA}(t)\}) \\ &\approx (\exp\{-\hat{H}(t)\})^2 \mathbf{V}(\hat{H}(t)) \\ &= [\hat{S}_{FH}(t)]^2 \mathbf{V}\left(\sum_{\{i, t_i < t\}} \hat{h}_i\right) \\ &\approx [\hat{S}_{FH}(t)]^2 \sum_{\{i, t_i < t\}} \mathbf{V}(\hat{h}_i) \\ &= [\hat{S}_{FH}(t)]^2 \sum_{\{i, t_i < t\}} \frac{\hat{h}_i(1 - \hat{h}_i)}{r_i} \\ \mathbf{V}(\hat{S}_{FH}(t)) &\approx [\hat{S}_{FH}(t)]^2 \sum_{\{i, t_i < t\}} \frac{m_i(r_i - m_i)}{r_i^3} \end{aligned} \quad (1.34)$$

Ainsi un estimateur de la variance asymptotique des estimateurs $\hat{S}_{KM}(t)$ et $\hat{S}_{FH}(t)$ est de la forme

$$\mathbf{V}(\hat{S}_n(t)) \approx [\hat{S}_n(t)]^2 \mathbf{V}^2(t)$$

où

$$\hat{S}_n(t) = \begin{cases} \hat{S}_{KM}(t) & \text{pour Kaplan-Meier} \\ \hat{S}_{FH}(t) & \text{pour Fleming-Harrington} \end{cases}$$

et

$$\mathbf{V}^2(t) = \begin{cases} \sum_{\{i, t_i < t\}} \frac{m_i}{r_i(r_i - m_i)} & \text{pour Kaplan-Meier} \\ \sum_{\{i, t_i < t\}} \frac{m_i(r_i - m_i)}{r_i^3} & \text{pour Fleming-Harrington} \end{cases}$$

Il est connu que les expressions précédentes de $\mathbf{V}^2(t)$ sous-estiment la variance des deux estimateurs en particulier lorsque $S(t)$ est proche de 0 ou 1. Aalen et Johansen (1978) (cf. [2]), ont suggéré pour $\mathbf{V}^2(t)$

$$\mathbf{V}^2(t) = \sum_{\{i, t_i < t\}} \frac{m_i}{r_i^2} \quad (1.35)$$

et donnent les formules finales

$$\hat{\mathbf{V}}(\hat{S}_{KM}(t)) = \left(\prod_{\{i, t_i < t\}} \left(1 - \frac{m_i}{r_i}\right) \right)^2 \sum_{\{i, t_i < t\}} \frac{m_i}{r_i^2} \quad (1.36)$$

et

$$\hat{\mathbf{V}}(\hat{S}_{FH}(t)) = \left(\exp\left\{-2 \sum_{\{i, t_i < t\}} \frac{m_i}{r_i}\right\} \right)^2 \sum_{\{i, t_i < t\}} \frac{m_i}{r_i^2} \quad (1.37)$$

Dans les études de simulations on présente souvent les graphiques des variances $\hat{\mathbf{V}}(\hat{S}_{KM}(t))$ et $\hat{\mathbf{V}}(\hat{S}_{FH}(t))$ ce qui permet de les comparer. On a $\hat{\mathbf{V}}(\hat{S}_{KM}(t)) \approx \hat{\mathbf{V}}(\hat{S}_{FH}(t))$ pour n assez grand puisque $\exp\{-x\} \approx 1 - x$ pour x petit.

Chapitre 2

Estimation de la densité des durées de survie

2.1 Données non censurées

2.1.1 Introduction

Dans cette partie nous rappelons quelques propriétés de l'estimateur à noyau de la densité sans démonstrations. Soit T_1, T_2, \dots, T_n des durées de survie de f.d.r. F et soit $V_{x_0} =]t_0 - \frac{b_n}{2}, t_0 + \frac{b_n}{2}]$ un voisinage de t_0 , où $(b_n)_{n \geq 0}$ est suite de paramètre positifs appelée fenêtre.

Un estimateur de la densité f au point t_0 est donné par

$$\begin{aligned} f_n(t_0) &= \frac{F_n(t_0 + \frac{b_n}{2}) - F_n(t_0 - \frac{b_n}{2})}{b_n} \\ &= \frac{\text{nombre d'évènements dans }]t_0 - \frac{b_n}{2}, t_0 + \frac{b_n}{2}]}{b_n} \end{aligned}$$

où F_n est la fonction de répartition empirique. On peut aussi écrire

$$\begin{aligned} f_n(t_0) &= \frac{1}{nb_n} \sum_{i=1}^n \mathbf{1}_{]t_0 - \frac{b_n}{2}, t_0 + \frac{b_n}{2}]}(T_i) \\ &= \frac{1}{nb_n} \sum_{i=1}^n \mathbf{1}_{]-\frac{1}{2}, \frac{1}{2}]}(\frac{t_0 - T_i}{b_n}) \end{aligned}$$

Cette écriture fait apparaître un noyau de la forme

$$K(\cdot) = \mathbf{1}_{]-\frac{1}{2}, \frac{1}{2}] }(\cdot)$$

D'où la définition d'un estimateur à noyau de f , pour $t \in]-\frac{1}{2}, \frac{1}{2}]$

$$f_n(t) = \frac{1}{nb_n} \sum_{i=1}^n K\left(\frac{t - T_i}{b_n}\right)$$

où le noyau K est une application: $\mathbb{R} \rightarrow \mathbb{R}^+$, borné, d'intégrale égale à 1 et symétrique et $b_n \rightarrow 0$.

Exemples de noyaux

Noyau gaussien

$$\forall u \in \mathbb{R}, K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Noyau uniforme

$$\forall u \in \mathbb{R}, K(u) = \frac{1}{2} \mathbf{1}_{|u| \leq 1}$$

Noyau d'Epanechnikov

$$\forall u \in \mathbb{R}, K(u) = \frac{3}{4} (1 - u^2) \mathbf{1}_{|u| \leq 1}$$

Noyau triangulaire

$$\forall u \in \mathbb{R}, K(u) = (1 - |u|) \mathbf{1}_{|u| \leq 1}$$

Noyau quadratique

$$\forall u \in \mathbb{R}, K(u) = \frac{15}{16} (1 - u^2)^2 \mathbf{1}_{|u| \leq 1}$$

Un noyau K est dit de Parzen-Rosemblatt si K est symétrique et si $\lim_{|u| \rightarrow \infty} |u| |K(u)| = 0$.

Si on pose $K_{b_n}(y) = \frac{1}{b_n} K\left(\frac{y}{b_n}\right)$, l'estimateur à noyau f_n s'écrit

$$f_n(t) = \int_{-\infty}^{+\infty} K_{b_n}(t - s) d\mu_n(s) = (K_{b_n} * \mu_n)(t)$$

où $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{T_i}$. L'étude de f_n repose en particulier sur le lemme de Bochner (cf. [4]).

Lemme 2.1 (i) Soit K un noyau de Parzen-Rosemblatt et g une fonction intégrable. Alors en tout point $t \in \mathbb{R}$ où g est continue on a

$$\lim_{n \rightarrow 0} (K_{b_n} * g)(t) = g(t)$$

(ii) Si g est uniformément continue, alors

$$\lim_{n \rightarrow 0} \sup_t |(K_{b_n} * g)(t) - g(t)| = 0$$

2.1.2 Comportement asymptotique de f_n

On pose C_t l'ensemble des fonctions continues en un point $t \in \mathbb{R}$.

Théorème 2.1 Si f_n est un estimateur à noyau et si le noyau K est de Parzen-Rosemblatt, alors

$$(b_n \rightarrow 0 \text{ et } nb_n \rightarrow \infty) \iff \left[\forall t \in \mathbb{R}, \forall f \in C_t, \mathbf{E} [f_n(t) - f(t)]^2 \rightarrow_{n \rightarrow \infty} 0 \right]$$

Théorème 2.2 Soit K un noyau positif et à variations bornées et f uniformément continue, on a

$$\left(b_n \rightarrow 0 \text{ et } \forall \gamma > 0, \sum_{n \geq 0} e^{-\gamma n b_n^2} < \infty \right) \iff \sup_t |f_n(t) - f(t)| \rightarrow_{n \rightarrow \infty} 0 \text{ p.co.}$$

Théorème 2.3 On a la loi limite gaussienne:

$$\frac{f_n(t) - \mathbf{E}f_n(t)}{\sigma(f_n(t))} \implies_{n \rightarrow \infty} N(0, 1)$$

2.1.3 Choix optimaux du noyau K et de la fenêtre b_n

Les résultats de convergence de l'estimateur à noyau f_n nécessitent des conditions sur K et b_n , d'où se pose le problème de choix optimaux de (K, b_n) .

On pose

$$\begin{aligned} \Delta_n(t) &= \mathbf{E} [f_n(t) - f(t)]^2 \\ &= \mathbf{E} [f_n(t) - \mathbf{E}f_n(t) + \mathbf{E}f_n(t) - f(t)]^2 \\ &= [\mathbf{E}f_n(t) - f(t)]^2 + \mathbf{E} [f_n(t) - \mathbf{E}f_n(t)]^2 \end{aligned}$$

On suppose que le support du noyau K est $[-1, 1]$

Le théorème suivant donne le comportement asymptotique du biais et de la variance de f_n

Théorème 2.4 *Si f est de classe C^2 et K est un noyau de Parzen-Roseblatt. Alors si $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$ on a*

1.

$$\mathbf{E}f_n(t) - f(t) = \frac{b_n^2}{2} f''(t) \int_{-1}^1 u^2 K(u) du + \mathcal{O}(b_n^2)$$

2.

$$\mathbf{E} [f_n(t) - \mathbf{E}f_n(t)]^2 = \frac{1}{nb_n} f(t) \int_{-1}^1 K^2(u) du + \mathcal{O}\left(\frac{1}{nb_n}\right)$$

3.

$$\Delta_n(t) = \frac{b_n^4}{4} f''^2(t) \left[\int_{-1}^1 u^2 K(u) du \right]^2 + \frac{1}{nb_n} f(t) \int_{-1}^1 K^2(u) du + \mathcal{O}\left(b_n^4 + \frac{1}{nb_n}\right)$$

Le choix de la fenêtre optimale b_n^* se fait ici si K est donnée.

Théorème 2.5 *Sous les conditions du théorème précédent, la fenêtre optimale est*

$$b_n^* = \left(\frac{f(t)}{f''^2(t) \left[\int_{-1}^1 u^2 K(u) du \right]^2 \int_{-1}^1 K^2(u) du} \right)^{1/5} n^{-1/5}$$

Pour les démonstrations des théorèmes précédents (cf. [4]).

2.2 Données censurées

2.2.1 Introduction

Le but de cette partie est l'estimation de la densité des durées de survie T en développant l'article de Zhang [21] dans le cadre de données censurées.

Soient T_1, T_2, \dots, T_n une suite de v.a. positives i.i.d. de fonction de répartition F représentant les durées de survie et C_1, C_2, \dots, C_n est la suite de v.a i.i.d. représentant les censures de f.d.r. G . On suppose que la suite (T_i) est indépendante de la suite (C_i) et on observe

$$X_i = T_i \wedge C_i, \quad D_i = 1_{\{T_i \leq C_i\}} \quad i = 1, \dots, n \quad (2.1)$$

Dans le chapitre 1, l'estimateur de Kaplan-Meier de la fonction de survie S est donné par

$$\hat{S}_{KM}(t) = \prod_{t^{(i)} \leq t} \left(1 - \frac{1}{n - i + 1}\right)^{D_{(i)}}$$

il peut aussi s'écrire

$$\hat{S}_{KM}(t) = \begin{cases} \prod_{X_{(i)} \leq t} \left(1 - \frac{D_{(i)}}{n - i + 1}\right) & \text{si } t \leq X_{(n)} \\ 1 - \hat{F}_n(X_{(n)}) & \text{si } t > X_{(n)} \end{cases} \quad (2.2)$$

où $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ est la statistique d'ordre de l'échantillon X_1, X_2, \dots, X_n et $D_{(i)}$ est la valeur de D_i (l'indicateur de censure à droite) associée à $X_{(i)}$. (Pour $t > X_{(n)}$ on a posé $1 - \hat{F}_n(t) = 1 - \hat{F}_n(X_{(n)})$ puisque il n'y a plus d'observations après t)

Nous reprenons les notations du chapitre 1. On suppose que F admet une densité f par rapport à la mesure de Lebesgue qu'on se propose d'estimer en utilisant les observations (X_i, D_i) , $i = 1, \dots, n$. En se basant sur l'estimateur de Kaplan-Meier, Blum et Susarla (1980) ont proposé un estimateur de la densité f par la méthode du noyau donné par

$$f_n(t) = \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) d\hat{F}_n(s) \quad (2.3)$$

où \hat{F}_n est une fonction empirique, $(b_n)_{n \geq 1}$ est la fenêtre avec $b_n \rightarrow_{n \rightarrow \infty} 0$ et K un noyau de support $[-1, 1]$. Cet estimateur f_n a été étudié en particulier par Blum et Susarla (1980), Földes et al. (1981) et Mielniczuk (1986). Dans ce chapitre on étudie l'article de Zhang qui concerne les propriétés asymptotiques de f_n en utilisant la théorie des martingales pour les processus continus.

Nous utilisons les notations du chap.1 suivantes:

Soit

$$S(t) = 1 - F(t) \quad , \quad S_C(t) = 1 - G(t)$$

$$\pi(t) = \mathbf{P}(X_1 \geq t) = S(t)S_C(t^-)$$

$$h(t) = \frac{f(t)}{S(t)} \quad , \quad H(t) = \int_0^t h(s) ds$$

Où h et H sont respectivement le taux de hasard et le taux de hasard cumulé.

On considère les processus ponctuels suivants définis sur $[0, \infty[$:

- Le processus $N_i(t)$ indicateur de "survenue de l'évènement" $N_i(t) = \mathbf{1}_{\{X_i \leq t, D_i=1\}}$
- Le processus $\bar{N}_n(t) = \sum_{i=1}^n N_i(t)$: le nombre d'"évènements" observés dans l'intervalle $]0, t]$
- Le processus risque: $Y_i(t) = \mathbf{1}_{\{X_i \geq t\}}$
- Le processus $\bar{Y}_n(t) = \sum_{i=1}^n Y_i(t)$: le nombre d'"individus" à risque à l'instant t

On note aussi les processus suivants:

$$\begin{aligned}
\Lambda(t) &= \int_0^t \bar{Y}_n(s) dH(s) \quad \text{le processus intensité cumulée} \\
\bar{M}_n(t) &= \bar{N}_n(t) - \Lambda(t) = \bar{N}_n(t) - \int_0^t \bar{Y}_n(s) h(s) ds \\
\hat{H}_{NA} &= \int_0^t \frac{d\bar{N}_n(u)}{\bar{Y}_n(u)} \\
\tilde{H}_n(t) &= \int_0^t J_n(u) h(u) du \quad , \quad J_n(t) = \mathbf{1}_{\{\bar{Y}_n(t) > 0\}} \\
\tilde{f}_n(t) &= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) d\tilde{H}_n(s)
\end{aligned} \tag{2.4}$$

Le processus $(M(t), t \geq 0)$ est une \mathcal{F}_t -martingale où

$$\mathcal{F}_t = \sigma(\mathbf{1}_{\{X_i \leq s, D_i=1\}}, \mathbf{1}_{\{X_i \leq s, D_i=0\}}; 0 \leq s \leq t, i = 1, \dots, n)$$

$(M(t), t \geq 0)$ est aussi une martingale locale de carré intégrable (cf. Coro. 4.2, Annexe).

On suppose que:

- **C1:** Le noyau K est borné sur $[-1, 1]$ et satisfait les conditions suivantes

$$\int_{-1}^1 K(t) dt = 1, \quad \int_{-1}^1 tK(t) dt = 0, \quad \int_{-1}^1 t^2 K(t) dt = k_2 \neq 0$$

2.2.2 Etude du biais de l'estimateur \hat{f}_n

On note dorénavant l'estimateur f_n dans (2.3) par \hat{f}_n

Dans le cas général où F est une f.d.r. quelconque, la fonction de survie $S(t) = \exp\{-H(t)\}$ est l'unique solution de l'équation intégrale suivante:

$$S(t) = 1 - F(t) = 1 - \int_0^t S(s^-) dH(s)$$

et l'estimateur de Kaplan-Meier Vérifie

$$\hat{S}_{KM}(t) = 1 - \int_0^t \hat{S}_{KM}(s^-) d\hat{H}_{NA}(s) =: 1 - \hat{F}_n(t)$$

où encore

$$d\hat{S}_{KM}(t) = -\hat{S}_{KM}(t^-) d\hat{H}_{NA}(t)$$

En effet

$$\begin{aligned} \hat{S}_{KM}(t) - \hat{S}_{KM}(t^-) &= \Delta \hat{S}_{KM}(t) \\ &= -\hat{S}_{KM}(t^-) \left[\hat{H}_{NA}(t) - \hat{H}_{NA}(t^-) \right] \\ &= : -\hat{S}_{KM}(t^-) \Delta \hat{H}_{NA}(t) \\ &= -\hat{S}_{KM}(t^-) \frac{\Delta \bar{N}_n(t)}{\bar{Y}_n(t)} \text{ par construction de } \hat{H}_{NA} \text{ (cf chap.1)} \end{aligned}$$

D'où

$$\hat{S}_{KM}(t) = \hat{S}_{KM}(t^-) \left(1 - \frac{\Delta \bar{N}_n(t)}{\bar{Y}_n(t)} \right)$$

où $\Delta \bar{N}(t) = \bar{N}(t) - \bar{N}(t^-)$. Comme \hat{H}_{NA} présente des sauts seulement aux points t_i , on obtient

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{\Delta \bar{N}(t_i)}{\bar{Y}(t_i)} \right) = \prod_{t_i \leq t} \left(1 - \Delta \hat{H}_{NA}(t_i) \right) \quad (2.5)$$

Ainsi de l'équation $\hat{S}_{KM}(t) = 1 - \hat{F}_n(t)$, l'estimateur à noyau (2.3) peut s'écrire -en changeant

de notation- comme suit:

$$f_n(t) = \hat{f}_n(t) = \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) d\hat{H}_{NA}(s) \quad (2.6)$$

car $d\hat{F}_n(t) = -d\hat{S}_{KM}(t)$.

Dans chap. 1 nous avons montré que $\hat{H}_{NA}(t) - \tilde{H}_n(t) = \int_0^t \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{M}_n(u)$. Par suite de (2.4) et (2.6) on a

$$\begin{aligned} \hat{f}_n(t) - \tilde{f}_n(t) &= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) d(\hat{H}_{NA} - \tilde{H}_n)(s) \\ &= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{M}_n(s) \\ &= \sum_{i=1}^n \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) \frac{J_n(s)}{\bar{Y}_n(s)} dM_i(s) \\ &= \sum_{i=1}^n \int_0^\infty L_i(s) dM_i(s) \end{aligned} \quad (2.7)$$

Comme $(L_i(s))$ sont des processus localement bornés (sur $[0, t]$) et \mathcal{F}_s -prévisibles, et $M_i(s)$ est une martingale locale de carré intégrable (cf. Corol. 2, Annexe), alors par le Th. 2.4.5 de [7] $(\hat{f}_n(t) - \tilde{f}_n(t), t \geq 0)$ est une martingale locale de carré intégrable.

On note

$$f_n^*(t) = \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) f(s) ds = \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) S(s) dH(s)$$

On a le résultat suivant

Théorème 2.6 *Supposons que f est continue.*

(i) Si $b_n \xrightarrow[n \rightarrow \infty]{} 0$ alors

$$f_n^*(t) - f(t) = \int_{-1}^1 K(u) [f(t - b_n u) - f(t)] du \xrightarrow[n \rightarrow \infty]{} 0$$

(ii) Pour tout $t \geq 0$ on a

$$\mathbf{E}\hat{f}_n(t) = \mathbf{E}\tilde{f}_n(t) = \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) J_n(s)\right) dH(s)$$

(iii) Si la fonction $\pi(\cdot) = \mathbf{P}(X_1 \geq \cdot)$ est strictement positive dans un voisinage d'un point t , alors on a

$$\begin{aligned} \mathbf{E}(\hat{f}_n(t) - f_n^*(t)) &= \mathbf{E}(\tilde{f}_n(t) - f_n^*(t)) \\ &= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \frac{\hat{S}_{KM}(\tau_{\bar{Y}})S(s)}{S(\tau_{\bar{Y}})}\right) dH(s) \end{aligned}$$

où $\tau_{\bar{Y}} = \inf \{s : \bar{Y}(s) = 0\}$

(iv) Sous les mêmes conditions de (iii), on a, pour n assez grand

$$\left| \mathbf{E}(\hat{f}_n(t) - f_n^*(t)) \right| \leq e^{-n\pi(t+b_n)} \int_{-1}^1 |K(u)| h(t - b_n u) du$$

ie $\left| \mathbf{E}(\hat{f}_n(t) - f_n^*(t)) \right| \rightarrow_{n \rightarrow \infty} 0$ avec une vitesse exponentielle.

(v) Si les conditions de (iii) sont vérifiées, f de classe C^2 et $b_n \rightarrow 0$ alors

$$\mathbf{E}(\hat{f}_n(t) - f(t)) = \frac{1}{2} b_n^2 f''(t) k_2 + o(b_n^2) + o\left(\frac{1}{n}\right)$$

où k_2 est donné par **C1**.

Preuve: (i) On a

$$\begin{aligned} f_n^*(t) - f(t) &= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) f(s) ds - f(t) \\ &= \int_{-1}^1 K(u) f(t - b_n u) du - \int_{-1}^1 K(u) f(t) du \\ &= \int_{-1}^1 K(u) [f(t - b_n u) - f(t)] du \end{aligned}$$

Par la continuité de f on a $|f(t - b_n u) - f(t)| \rightarrow_{n \rightarrow \infty} 0$ et comme K est bornée, on a par le théorème de convergence dominée:

$$\begin{aligned} |f_n^*(t) - f(t)| &\leq \int_{-1}^1 |K(u)| |f(t - b_n u) - f(t)| du \\ &\leq \sup_{u \in [-1, 1]} |K(u)| \int_{-1}^1 |f(t - b_n u) - f(t)| du \rightarrow_{n \rightarrow \infty} 0 \end{aligned}$$

et par suite $f_n^*(t) \xrightarrow{n \rightarrow \infty} f(t)$

On peut aussi avoir le résultat par le lemme de Bochner: on a

$$f_n^*(t) = (K_{b_n} * f)(t)$$

Comme f est continue et $b_n \rightarrow 0$, alors

$$(K_{b_n} * f)(t) \xrightarrow{n \rightarrow \infty} f(t)$$

donc

$$f_n^*(t) - f(t) \xrightarrow{n \rightarrow \infty} 0$$

(ii) On par (2.7)

$$\hat{f}_n(t) - \tilde{f}_n(t) = \sum_{i=1}^n \int_0^\infty L_i(s) dM_i(s)$$

et en notant que $\langle M_n \rangle_t = \int_0^t \bar{Y}_n(s) h(s) ds$ (cf. chap 1, th1.3), on a

$$\begin{aligned} \mathbf{E} \left[\hat{f}_n(t) - \tilde{f}_n(t) \right]^2 &= \frac{1}{b_n^2} \mathbf{E} \left[\int_0^t K^2 \left(\frac{t-s}{b_n} \right) \hat{S}_{KM}^2(s^-) \frac{J_n(s)}{\bar{Y}_n^2(s)} d \langle M_n \rangle_s \right] \\ &= \frac{1}{b_n} \int_0^{t/b_n} K^2(u) \mathbf{E} \left[\hat{S}_{KM}^2[(t-b_n u)^-] \frac{J_n(t-b_n u)}{\bar{Y}_n(t-b_n u)} \right] h(t-b_n u) du \\ &\leq \frac{1}{b_n} \int_0^1 K^2(u) h(t-b_n u) du < \infty \end{aligned}$$

où nous avons utilisé que

$$\mathbf{E} \left[\hat{S}_{KM}^2[(t-b_n u)^-] \frac{J_n(t-b_n u)}{\bar{Y}_n(t-b_n u)} \right] \leq 1$$

car $\hat{S}_{KM} \leq 1$, $J_n \leq 1$, $\bar{Y}_n \geq 1$. Par le Th. 2.4.5 de [7] p. 73, on a pour n assez grand

$$\mathbf{E}(\hat{f}_n(t) - \tilde{f}_n(t)) = 0$$

Ce qui donne $\mathbf{E}(\hat{f}_n(t)) = \mathbf{E}(\tilde{f}_n(t))$ et on a

$$\begin{aligned}
\mathbf{E}(\hat{f}_n(t)) &= \mathbf{E}(\tilde{f}_n(t)) = \mathbf{E}\left(\frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM} d\tilde{H}_n(s)\right) \\
&= \mathbf{E}\left(\frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) J_n(s) h(s) ds\right) \\
&= \mathbf{E}\left(\frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \hat{S}_{KM}(s^-) J_n(s) dH(s)\right) \\
&= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) J_n(s)\right) dH(s) \quad (\text{th. de Fubini})
\end{aligned}$$

(iii) et (iv).

En utilisant (ii) et le lemme 3.2.1 de [7] p. 99, on obtient du fait que

$$J_n(s) = 1 - \mathbf{1}_{\{\tilde{Y}_n(s)=0\}}$$

les égalités suivantes:

$$\begin{aligned}
\mathbf{E}\left[\hat{f}_n(t) - f_n^*(t)\right] &= \mathbf{E}\left[\tilde{f}_n(t) - f_n^*(t)\right] \\
&= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) J_n(s)\right) dH(s) \\
&\quad - \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) S(s) dH(s) \\
&= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-)\right) dH(s) \\
&\quad - \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) \mathbf{1}_{\{\tilde{Y}_n(s)=0\}}\right) dH(s) \\
&\quad - \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) S(s) dH(s) \\
&= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) - S(s^-)\right) dH(s) \quad (S(s) = S(s^-)) \\
&\quad - \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) \mathbf{1}_{\{\tilde{Y}_n(s)=0\}}\right) dH(s) \quad (\text{car } S \text{ continue}) \\
&= \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\mathbf{1}_{\{\tau_{\tilde{Y}} < s^-\}} \frac{\hat{S}_{KM}(\tau_{\tilde{Y}}) [S(\tau_{\tilde{Y}}) - S(s^-)]}{S(\tau_{\tilde{Y}})}\right) dH(s) \\
&\quad - \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) \mathbf{E}\left(\hat{S}_{KM}(s^-) \mathbf{1}_{\{\tau_{\tilde{Y}} < s\}}\right) dH(s)
\end{aligned}$$

D'où

$$\begin{aligned}
\mathbf{E} \left[\hat{f}_n(t) - f_n^*(t) \right] &= \frac{1}{b_n} \int_0^\infty K \left(\frac{t-s}{b_n} \right) \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \frac{\hat{S}_{KM}(\tau_{\bar{Y}}) [S(\tau_{\bar{Y}}) - S(s^-)]}{S(\tau_{\bar{Y}})} \right. \\
&\quad \left. - \hat{S}_{KM}(s^-) \mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \right) dH(s) \\
&= \frac{1}{b_n} \int_0^\infty K \left(\frac{t-s}{b_n} \right) \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \left(\hat{S}_{KM}(\tau_{\bar{Y}}) - \hat{S}_{KM}(s^-) \right) \right) dH(s) \\
&\quad - \frac{1}{b_n} \int_0^\infty K \left(\frac{t-s}{b_n} \right) \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \frac{\hat{S}_{KM}(\tau_{\bar{Y}}) S(s)}{S(\tau_{\bar{Y}})} \right) dH(s) \\
&= -\frac{1}{b_n} \int_0^\infty K \left(\frac{t-s}{b_n} \right) \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \frac{\hat{S}_{KM}(\tau_{\bar{Y}}) S(s)}{S(\tau_{\bar{Y}})} \right) dH(s)
\end{aligned}$$

$$\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} (\hat{S}_{KM}(\tau_{\bar{Y}}) - \hat{S}_{KM}(s^-)) = 0 \text{ car } \hat{S}_{KM}(\tau_{\bar{Y}}) = \hat{S}_{KM}(s^-) \forall \tau_{\bar{Y}} < s$$

Par suite pour n assez grand on a

$$\begin{aligned}
\left| \mathbf{E} \left[\hat{f}_n(t) - f_n^*(t) \right] \right| &\leq \frac{1}{b_n} \int_0^\infty \left| K \left(\frac{t-s}{b_n} \right) \right| \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \frac{\hat{S}_{KM}(\tau_{\bar{Y}}) S(s)}{S(\tau_{\bar{Y}})} \right) dH(s) \\
&\leq \frac{1}{b_n} \int_0^\infty \left| K \left(\frac{t-s}{b_n} \right) \right| \mathbf{E} \left(\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \right) dH(s) \quad \left(\frac{S(s)}{S(\tau_{\bar{Y}})} \leq 1, \hat{S}_{KM} \leq 1 \right) \\
&\leq \frac{1}{b_n} \int_0^\infty \left| K \left(\frac{t-s}{b_n} \right) \right| \mathbf{E} \left(\mathbf{1}_{\{\bar{Y}(s)=0\}} \right) h(s) ds \\
&= \int_{-1}^1 |K(u)| [1 - \pi(t - b_n u)]^n h(t - b_n u) du \\
&\leq [1 - \pi(t + b_n)]^n \int_{-1}^1 |K(u)| h(t - b_n u) du \quad (\text{car } \pi \searrow) \\
&\leq e^{-n\pi(t+b_n)} \int_{-1}^1 |K(u)| h(t - b_n u) du \quad (\text{car } 1 - x \leq e^{-x})
\end{aligned}$$

car par la preuve du Th. 1.1 on a

$$\mathbf{E} \left(\mathbf{1}_{\{\bar{Y}_n(s)=0\}} \right) = P(\bar{Y}_n(s) = 0) = [(1 - (1 - F(s))(1 - G(s^-)))]^n = [1 - \pi(s)]^n$$

(v). On sait que $\mathbf{E} \hat{f}_n(t) = (K_{b_n} * f)(t)$ et aussi $f_n^*(t) = (K_{b_n} * f)(t)$ où $K_{b_n}(\cdot) = \frac{1}{b_n} K(\frac{\cdot}{b_n})$.

On a par (i) $f_n^*(t) \rightarrow_{n \rightarrow \infty} f(t)$ et par (iv) on a

$$n \left| \mathbf{E} \left[\hat{f}_n(t) - f_n^*(t) \right] \right| \leq C.n e^{-n\pi(t+b_n)} \rightarrow_{n \rightarrow \infty} 0 \text{ où } C > 0$$

Nous avons

$$\begin{aligned}\mathbf{E}\hat{f}_n(t) - f(t) &= \mathbf{E}(\hat{f}_n(t) - f_n^*(t)) + (f_n^*(t) - f(t)) \\ &= o\left(\frac{1}{n}\right) + \int_{-1}^1 K(u) [f(t - b_n u) - f(t)] du\end{aligned}$$

Un développement de Taylor donne

$$f(t - b_n u) - f(t) = -b_n u f'(t) + \frac{b_n^2}{2} u^2 f''(t) + o(b_n^2)$$

et alors par **C1**

$$\begin{aligned}\int_{-1}^1 K(u) [f(t - b_n u) - f(t)] du &= -b_n f'(t) \int_{-1}^1 u K(u) du \\ &\quad + \frac{b_n^2}{2} f''(t) \int_{-1}^1 u^2 K(u) du + o(b_n^2) \\ &= \frac{b_n^2}{2} f''(t) k_2 + o(b_n^2)\end{aligned}$$

et par suite

$$\mathbf{E}\hat{f}_n(t) - f(t) = \frac{1}{2} b_n^2 f''(t) k_2 + o(b_n^2) + o\left(\frac{1}{n}\right)$$

Ce qui achève la démonstration. ■

Remarque 2.1 le point (v) montre que l'estimateur \hat{f}_n est asymptotiquement sans biais.

2.2.3 Consistance et normalité asymptotique de \hat{f}_n

Le théorème suivant donne la convergence en probabilité de l'estimateur \hat{f}_n .

Théorème 2.7 Si f est continue en t et π est positive au voisinage de t , $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$ alors

$$\hat{f}_n(t) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} f(t)$$

Preuve: Puisque $\pi > 0$ sur un voisinage V_t d'un point $t > 0$, f est continue et bornée, S continue et $S > 0$ sur V_t , alors $h = \frac{f}{S}$ est aussi continue et bornée dans V_t .

On a

$$\left| \hat{f}_n(t) - f(t) \right| \leq \left| \hat{f}_n(t) - \tilde{f}_n(t) \right| + \left| \tilde{f}_n(t) - f_n^*(t) \right| + |f_n^*(t) - f(t)|$$

Par le théorème précédent (i) on a:

$$|f_n^*(t) - f(t)| \xrightarrow{b_n \rightarrow n \rightarrow \infty} 0$$

Reste à montrer que

$$\left| \hat{f}_n(t) - \tilde{f}_n(t) \right| \xrightarrow[n \rightarrow \infty]{P} 0 \text{ et } \left| \tilde{f}_n(t) - f_n^* \right| \xrightarrow[n \rightarrow \infty]{P} 0 \quad (2.8)$$

Par (2.7) et le corollaire 3.4.1 de [7] p. 113 et l'inégalité de Lengart: on a pour $\varepsilon, \eta > 0$ et n assez grand

$$\begin{aligned} \mathbf{P} \left(\left| \hat{f}_n(t) - \tilde{f}_n(t) \right| > \varepsilon \right) &= \mathbf{P} \left(\left| \frac{1}{b_n} \int_0^\infty K \left(\frac{t-u}{b_n} \right) \hat{S}_{KM}(u^-) \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{M}_n(u) \right| > \varepsilon \right) \\ &\leq \mathbf{P} \left(\sup_{t-b_n \leq s \leq t+b_n} \left| \frac{1}{b_n} \int_{t-b_n}^s K \left(\frac{t-u}{b_n} \right) \hat{S}_{KM}(u^-) \frac{J_n(u)}{\bar{Y}_n(u)} d\bar{M}_n(u) \right| > \varepsilon \right) \\ &\leq \frac{\eta}{\varepsilon^2} + \mathbf{P} \left(\frac{1}{b_n^2} \int_{t-b_n}^{t+b_n} K^2 \left(\frac{t-u}{b_n} \right) \hat{S}_{KM}^2(u^-) \frac{J_n(u)}{\bar{Y}_n^2(u)} d < \bar{M}_n >_{u \geq \eta} \right) \\ &= \frac{\eta}{\varepsilon^2} + \mathbf{P} \left(\frac{1}{b_n} \int_{-1}^1 K^2(v) \hat{S}_{KM}^2[(t-b_nv)^-] \frac{J_n(t-b_nv)}{\bar{Y}_n(t-b_nv)} h(t-b_nv) dv \geq \eta \right) \\ &\leq \frac{\eta}{\varepsilon^2} + \mathbf{P} \left(\frac{1}{nb_n} \frac{1}{n^{-1}\bar{Y}_n(t+b_n)} \sup_{s \in [t-b_n, t+b_n]} h(s) \int_{-1}^1 K^2(v) dv \geq \eta \right) \\ &\leq \frac{\eta}{\varepsilon^2} + \mathbf{P} \left(\frac{1}{n^{-1}\bar{Y}_n(t+b_n)} \geq C.\eta nb_n \right) \end{aligned}$$

Comme $n^{-1}\bar{Y}(t+b_n) \xrightarrow[n \rightarrow \infty]{P} (1-F(t))(1-G(t^-))$ (cf dem Th 1.1, chap 1) et $nb_n \rightarrow \infty$, la probabilité ci-dessous tend vers 0. Or ε et η sont arbitraires et donc on a

$$\left| \hat{f}_n(t) - \tilde{f}_n(t) \right| \xrightarrow[n \rightarrow \infty]{P} 0$$

D'autre part, on a

$$\begin{aligned}
\left| \tilde{f}_n(t) - f_n^*(t) \right| &= \left| \frac{1}{b_n} \int_0^\infty K\left(\frac{t-s}{b_n}\right) [\hat{S}_{KM}(s^-)J(s) - S(s)]dH(s) \right| \\
&\leq \int_{-1}^1 |K(u)| \mathbf{1}_{\{\bar{Y}_n(t-ub_n) > 0\}} \left| \hat{S}_{KM}((t-ub_n)^-) - S(t-ub_n) \right| h(t-ub_n) du \\
&\quad + \int_{-1}^1 |K(u)| \mathbf{1}_{\{\bar{Y}_n(t-ub_n) = 0\}} f(t-ub_n) du \\
&\leq \left[\sup_{s \in [t-b_n, t+b_n]} h(s) \sup_{s \in [t-b_n, t+b_n]} \left| \hat{S}_{KM}(s) - S(s) \right| \right. \\
&\quad \left. + \mathbf{1}_{\{\bar{Y}_n(t-ub_n) = 0\}} \sup_{s \in [t-b_n, t+b_n]} f(s) \right] \int_{-1}^1 |K(u)| du
\end{aligned} \tag{2.9}$$

Comme $\mathbf{1}_{\{\bar{Y}_n(t-ub_n) = 0\}} \xrightarrow{\mathbf{P}} 0$ (cf dem. Th. 1.1) et par le th. 3.4.2 de [7] p. 115 on a

$$\sup_{s \in [t-b_n, t+b_n]} \left| \hat{S}_{KM}(s) - S(s) \right| \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

On en déduit que

$$\left| \tilde{f}_n(t) - f_n^*(t) \right| \xrightarrow{P}_{n \rightarrow \infty} 0$$

Doù le résultat du théorème. ■

Le théorème suivant donne la normalité asymptotique de \hat{f}_n .

Théorème 2.8 *Supposons que f est continue en un point t et π est continue et strictement positive.*

Si $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$, alors

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f_n^*(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

où

$$\sigma^2(t) = \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du \tag{2.10}$$

Preuve: On décompose:

$$\sqrt{nb_n}(\hat{f}_n(t) - f_n^*(t)) = \sqrt{nb_n}(\hat{f}_n(t) - \tilde{f}_n(t)) + \sqrt{nb_n}(\tilde{f}_n(t) - f_n^*(t))$$

Montrons que le premier terme à une loi limite et le deuxième terme tend vers 0 en proba.

Par (2.7) on peut écrire en posant:

$$\begin{aligned}
U_n(t) & : \equiv \sqrt{nb_n} \left(\hat{f}_n(t) - \tilde{f}_n(t) \right) \\
& = \sqrt{nb_n^{-1}} \int_0^\infty K \left(\frac{t-s}{b_n} \right) \hat{S}_{KM}(s^-) \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{M}_n(s) \\
& = : \int_0^\infty L_n(t, s) d\bar{M}_n(s)
\end{aligned} \tag{2.11}$$

Le processus croissant de $(U_n(t), t \geq 0)$ est donné par (cf Th. 2.4.3 de [7] p. 70)

$$\begin{aligned}
\langle U_n \rangle_t & = \int_0^\infty L_n^2(t, s) d \langle \bar{M} \rangle_s \\
& = \int_0^\infty L_n^2(t, s) \bar{Y}_n(s) dH(s) \\
& = \int_{-1}^1 nK^2(u) \hat{S}_{KM}^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\bar{Y}_n(t-b_nu)} h(t-b_nu) du
\end{aligned}$$

Posons

$$\begin{aligned}
A_1(t) & : = \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\frac{1}{n}\bar{Y}_n(t-b_nu)} h(t) du \\
A_2(t) & : = \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\pi(t-b_nu)} h(t) du \\
A_3(t) & : = \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\pi(t)} h(t) du \\
A_4(t) & : = \int_{-1}^1 K^2(u) S^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\pi(t)} h(t) du \\
A_5(t) & : = \int_{-1}^1 K^2(u) S^2(t) \frac{J_n(t-b_nu)}{\pi(t)} h(t) du
\end{aligned}$$

En utilisant $\hat{S}_{KM}(t) \leq 1$ et $J_n(t) \leq 1$, on a

$$\begin{aligned}
|\langle U_n \rangle_t - A_1(t)| & = \left| \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t-b_nu)^-) \frac{J_n(t-b_nu)}{\frac{1}{n}\bar{Y}_n(t-b_nu)} [h(t-b_nu) - h(t)] du \right| \\
& \leq \frac{1}{\frac{1}{n}\bar{Y}_n(t+b_n)} \sup_{s \in [t-b_n, t+b_n]} |h(s) - h(t)| \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.12}$$

De même on majore la différence:

$$\begin{aligned}
|A_1(t) - A_2(t)| &= \left| \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t - b_n u)^-) J_n(t - b_n u) h(t) \right. \\
&\quad \left. \left(\frac{1}{\frac{1}{n} \bar{Y}_n(t - b_n u)} - \frac{1}{\pi(t - b_n u)} \right) du \right| \\
&\leq \frac{h(t)}{\pi(t + b_n) n^{-1} \bar{Y}_n(t + b_n)} \\
&\quad \sup_{s \in [t - b_n, t + b_n]} \left| \frac{1}{n} \bar{Y}_n(s) - \pi(s) \right| \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.13}$$

De même on majore la différence

$$\begin{aligned}
|A_2(t) - A_3(t)| &= \left| \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t - b_n u)^-) J_n(t - b_n u) \right. \\
&\quad \left. h(t) \left[\frac{1}{\pi(t - b_n u)} - \frac{1}{\pi(t)} \right] du \right| \\
&\leq \frac{h(t)}{\pi(t) \pi(t + b_n)} \sup_{s \in [t - b_n, t + b_n]} |\pi(t) - \pi(s)| \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.14}$$

De même on a

$$\begin{aligned}
|A_3(t) - A_4(t)| &= \left| \int_{-1}^1 K^2(u) \left[\hat{S}_{KM}^2((t - b_n u)^-) - \right. \right. \\
&\quad \left. \left. S^2((t - b_n u)^-) \right] \frac{J_n(t - b_n u)}{\pi(t)} h(t) du \right| \\
&\leq 2 \frac{h(t)}{\pi(t)} \sup_{s \in [t - b_n, t + b_n]} \left| \hat{S}_{KM}(s) - S(t) \right| \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.15}$$

Aussi on a

$$\begin{aligned}
|A_4(t) - A_5(t)| &= \left| \int_{-1}^1 K^2(u) \left[S^2((t - b_n u)^-) - S^2(t) \right] \frac{J_n(t - b_n u)}{\pi(t)} h(t) du \right| \\
&\leq 2 \frac{h(t)}{\pi(t)} \sup_{s \in [t - b_n, t + b_n]} |S(t) - S(s)| \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.16}$$

Enfin, en remarquant que $\pi(t) = S_C(t)S(t)$ et que $S(t) = f(t)/h(t)$ on a

$$\begin{aligned}
|A_5(t) - \sigma^2(t)| &= \left| \frac{h(t)S^2(t)}{\pi(t)} \int_{-1}^1 K^2(u) J_n(t - b_n u) du - \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du \right| \\
&= \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) [1 - J_n(t - b_n u)] du \\
&\leq \frac{f(t)}{S_C(t^-)} \mathbf{1}_{\{\bar{Y}_n(t+b_n)=0\}} \int_{-1}^1 K^2(u) du
\end{aligned} \tag{2.17}$$

En utilisant les majorations ((2.12) à (2.17)) on obtient

$$\begin{aligned}
|\langle U_n \rangle_t - \sigma^2(t)| &\leq |\langle U_n, U_n \rangle_t - A_1(t)| + |A_1(t) - A_2(t)| + |A_2(t) - A_3(t)| \\
&\quad + |A_3(t) - A_4(t)| + |A_4(t) - A_5(t)| + |A_5(t) - \sigma^2(t)| \\
&\leq \frac{1}{\frac{1}{n}\bar{Y}_n(t+b_n)} \sup_{s \in [t-b_n, t+b_n]} |h(s) - h(t)| \int_{-1}^1 K^2(u) du \\
&\quad + \frac{h(t)}{\pi(t+b_n)n^{-1}\bar{Y}_n(t+b_n)} \sup_{s \in [t-b_n, t+b_n]} \left| \frac{1}{n}\bar{Y}(s) - \pi(s) \right| \int_{-1}^1 K^2(u) du \\
&\quad + \frac{h(t)}{\pi(t)\pi(t+b_n)} \sup_{s \in [t-b_n, t+b_n]} |\pi(s) - \pi(t)| \int_{-1}^1 K^2(u) du \\
&\quad + 2 \frac{h(t)}{\pi(t)} \sup_{s \in [t-b_n, t+b_n]} \left| \hat{S}_{KM}(s) - S(t) \right| \int_{-1}^1 K^2(u) du \\
&\quad + 2 \frac{h(t)}{\pi(t)} \sup_{s \in [t-b_n, t+b_n]} |S(s) - S(t)| \int_{-1}^1 K^2(u) du \\
&\quad + \frac{f(t)}{S_C(t^-)} \mathbf{1}_{\{\bar{Y}_n(t+b_n)=0\}} \int_{-1}^1 K^2(u) du
\end{aligned}$$

Chacun des majorants tend vers 0 en proba. En effet:

◇ Pour le premier, on utilise: $\frac{1}{n}\bar{Y}(t+b_n) \xrightarrow{P_{n \rightarrow \infty}} (1-F(t))(1-G(t^-))$ uniformément, la continuité de h et la condition **C1**.

◇ Pour le deuxième, on utilise $\frac{1}{n}\bar{Y}(t+b_n) \xrightarrow{P_{n \rightarrow \infty}} (1-F(t))(1-G(t^-))$ uniformément, la continuité de π et **C1**.

◇ Pour le troisième, on utilise la continuité de π et **C1**.

◇ Pour le quatrième, on utilise le Th. 3.4.2 FH p.115 et **C1**.

◇ Pour le cinquième, on utilise la continuité de S et **C1**.

◇ Pour le sixième, on utilise $\bar{Y}_n(t+b_n) \xrightarrow{P_{n \rightarrow \infty}} \infty$ (car $\frac{1}{n}\bar{Y}_n(t+b_n) \xrightarrow{P} (1-F(t))(1-G(t^-))$)

et **C1**.

Par suite

$$\langle U_n \rangle_t \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} \sigma^2(t) \quad (2.18)$$

Pour établir la loi limite on utilise le th. de Robolledo (cf. Annexe):

On pose

$$U_{n,\varepsilon}(t) := \int_0^\infty L_n(t, s) \mathbf{1}_{\{|L_n(t,s)| \geq \varepsilon\}} d\bar{M}_n(s)$$

où L_n est donnée par (2.11). Par le théorème 2.4.3 de [7], p. 70 on a

$$\begin{aligned} \langle U_{n,\varepsilon} \rangle_t &= \int_0^\infty L_n^2(t, s) \mathbf{1}_{\{|L_n(t,s)| \geq \varepsilon\}} \bar{Y}_n(s) dH(s) \\ &= \int_{-1}^1 K^2(u) \hat{S}_{KM}^2((t - b_n u)^-) \frac{J_n(t - b_n u)}{n^{-1} \bar{Y}_n(t - b_n u)} h(t - b_n u) \mathbf{1}_{A(n,u)} du \\ &\leq \frac{1}{n^{-1} \bar{Y}_n(t + b_n)} \left[\sup_{s \in [t - b_n, t + b_n]} h(s) \right] \mathbf{1}_{B(n)} \int_{-1}^1 K^2(u) du \end{aligned}$$

où

$$A(n, u) = \left\{ \left| K(u) \hat{S}_{KM}((t - b_n u)^-) \frac{J_n(t - b_n u)}{n^{-1} \bar{Y}_n(t - b_n u)} \right| \geq \varepsilon \sqrt{nb_n} \right\}$$

et

$$B(n) = \left\{ \frac{1}{n} \bar{Y}_n(t + b_n) \sup_{u \in [-1, 1]} |K(u)| \geq \varepsilon \sqrt{nb_n} \right\}$$

Le théorème de Glivenko-Cantelli implique que

$$\frac{1}{n} \bar{Y}_n(t + b_n) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} S(t) (1 - G(t))$$

et comme $nb_n \rightarrow \infty$ on en déduit que

$$\mathbf{1}_{B(n)} \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

Par suite

$$\langle U_{n,\varepsilon} \rangle_t \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0 \quad (2.19)$$

Par conséquent, de (2.18) et (2.19) et par le théorème de Rebolledo (Th. 5.1.1 de [7] p. 204)

on a

$$U_n(t) \equiv \sqrt{nb_n} \left(\hat{f}_n(t) - \tilde{f}_n(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, \sigma^2(t)) \quad (2.20)$$

Montrons maintenant que

$$\sqrt{nb_n} \left(\tilde{f}_n(t) - f_n^*(t) \right) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0 \quad (2.21)$$

Par (2.9) on a

$$\begin{aligned} \sqrt{nb_n} \left| \tilde{f}_n(t) - f_n^*(t) \right| &\leq \sqrt{nb_n} \left[\sup_{s \in [t-b_n, t+b_n]} h(s) \sup_{s \in [t-b_n, t+b_n]} \left| \hat{S}_{KM}(s) - S(s) \right| \right] \int_{-1}^1 |K(u)| du \\ &\quad + \sqrt{nb_n} \mathbf{1}_{\{\bar{Y}(t+ub_n)=0\}} \left[\sup_{s \in [t-b_n, t+b_n]} f(s) \right] \int_{-1}^1 |K(u)| du \end{aligned} \quad (2.22)$$

◇ Pour le deuxième terme, nous avons du fait que $1 - x \leq e^{-x}$:

$$\mathbf{E} \left(\sqrt{nb_n} \mathbf{1}_{\{\bar{Y}(t+ub_n)=0\}} \right) = \sqrt{nb_n} (1 - \pi(t + b_n))^n \leq \sqrt{nb_n} e^{-n\pi(t+b_n)}$$

puisque π est positive au voisinage de t . Donc

$$\sqrt{nb_n} \mathbf{1}_{\{\bar{Y}(t+ub_n)=0\}} \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

Par conséquent

$$\sqrt{nb_n} \mathbf{1}_{\{\bar{Y}(t+ub_n)=0\}} \left[\sup_{s \in [t-b_n, t+b_n]} f(s) \right] \int_{-1}^1 |K(u)| \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

◇ Pour le premier terme, par le théorème 6.3.1 de [7] p. 235, on a: pour $\varepsilon > 0$

$$\sqrt{n} \left(\hat{S}_{KM}(\cdot) - S(\cdot) \right) \Longrightarrow_{n \rightarrow \infty} S(\cdot) W(v(\cdot))$$

sur $D[t - \varepsilon, t + \varepsilon]$ où $v(t) \equiv \int_0^t [\pi(s)]^{-1} dH(s)$ et W est un processus de Wiener et $D[t - \varepsilon, t + \varepsilon]$ est l'espace de Skorokhod (l'espace des fonctions cadlag.)

Ainsi par la continuité de la fonction sup

$$\sup_{s \in [t-\varepsilon, t+\varepsilon]} \sqrt{n} \left| \hat{S}_{KM}(\cdot) - S(s) \right| \Longrightarrow_{n \rightarrow \infty} \sup_{s \in [t-\varepsilon, t+\varepsilon]} S(s) |W(v(s))|$$

De la propriété du scaling du mouvement brownien on a :

$$S(s)W(v(s)) = W(S^2(s)v(s))$$

et que

$$\sup_{s \in [t-\varepsilon, t+\varepsilon]} S(s) |W(v(s))| = \sup_{s \in [t-\varepsilon, t+\varepsilon]} |W^*(s)|$$

où W^* est un autre mouvement brownien. On sait que

$$\sup_{s \in [0, t]} |W^*(s)| \stackrel{\mathcal{L}}{=} \sqrt{t} |W^*(1)|$$

et donc bornée en probabilité. On en déduit que pour t fixé et $n \rightarrow \infty$

$$\sup_{s \in [t-\varepsilon, t+\varepsilon]} \sqrt{n} \left| \hat{S}_{KM}(s) - S(s) \right| = \mathcal{O}_P(1)$$

et par suite si $n \rightarrow \infty$ et $b_n \rightarrow 0$

$$\sqrt{nb_n} \sup_{s \in [t-b_n, t+b_n]} |S_n(s) - S(s)| = \mathcal{O}_p(\sqrt{b_n})$$

Donc de (2.22) on conclut que

$$\sqrt{nb_n} \left| \tilde{f}_n(t) - f_n^*(t) \right| \xrightarrow[n \rightarrow \infty]{P} 0$$

Enfin avec (2.20) et (2.21) on obtient

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f_n^*(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

Ce qui termine la démonstration. ■

Le corollaire suivant est une conséquence du Th. précédent.

Corollaire 2.1 *On suppose que: f continue en t et satisfait*

$$|f(t+b) - f(t)| \leq C |b|^\alpha$$

où b est proche de 0, $\alpha \in [0, 1]$ et la constante C dépend que de t .

π est continue et positive en t .

Si $b_n \rightarrow 0$, $nb_n \rightarrow \infty$, et $nb_n^{1+2\alpha} \rightarrow 0$ alors

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

où $\sigma^2(t)$ est donnée par (2.10)

Preuve: On a

$$\sqrt{nb_n}(\hat{f}_n(t) - f(t)) = \sqrt{nb_n}(\hat{f}_n(t) - f_n^*(t)) + \sqrt{nb_n}(f_n^*(t) - f(t))$$

Or

$$\begin{aligned} \sqrt{nb_n} |f_n^*(t) - f(t)| &= \sqrt{nb_n} \left| \int_{-1}^1 K(u)(f(t - b_n u) - f(t)) du \right| \\ &\leq \sqrt{nb_n} \int_{-1}^1 |K(u)| |f(t - b_n u) - f(t)| du \\ &\leq C \sqrt{nb_n^{1+2\alpha}} \int_{-1}^1 |u|^\alpha |K(u)| du \rightarrow_{n \rightarrow \infty} 0 \text{ car } nb_n^{1+2\alpha} \rightarrow_{n \rightarrow \infty} 0 \end{aligned}$$

et d'après le théorème précédent on a

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f_n^*(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

D'où le résultat. ■

Remarque 2.2 La condition $nb_n^{1+2\alpha} \rightarrow_{n \rightarrow \infty} 0$ pour $\alpha \in [0, 1]$ exige que la fenêtre b_n tend vers 0 plus rapidement que $n^{-1/1+2\alpha}$.

Le corollaire suivant donne la normalité asymptotique de \hat{f}_n pour b_n de l'ordre $n^{-1/5}$ (fenêtre optimale en estimation fonctionnelle).

Corollaire 2.2 Supposons que f est de classe C^2 et que $b_n = O(n^{-1/5})$.

Alors sous les conditions du théorème 2.8, nous avons

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2 \right) \Rightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

où $\sigma^2(t)$ et k_2 son donnés respectivement par (2.10) et **C1**.

Preuve: On a

$$\begin{aligned} \sqrt{nb_n}(\hat{f}_n(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2) &= \sqrt{nb_n}(\hat{f}_n(t) - f_n^*(t)) + \\ &\quad \sqrt{nb_n}(f_n^*(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2) \end{aligned}$$

Par le th. **2.8** on a

$$\sqrt{nb_n}(\hat{f}_n(t) - f_n^*(t)) \Rightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

Il reste à montrer que

$$\sqrt{nb_n}(f_n^*(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

Un développement de Taylor donne

$$\begin{aligned} \sqrt{nb_n} \left| f_n^*(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2 \right| &= \sqrt{nb_n} \left| \int_{-1}^1 K(u)(f(t - b_n u) - f(t))du - \frac{1}{2}b_n^2 f''(t)k_2 \right| \\ &= \sqrt{nb_n} \left| b_n f'(t) \int_{-1}^1 u K(u) du + \frac{b_n^2}{2} f''(t) \int_{-1}^1 u^2 K(u) du \right. \\ &\quad \left. + O(b_n^2) - \frac{1}{2}b_n^2 f''(t)k_2 \right| \\ &= \sqrt{nb_n} \left| \frac{b_n^2}{2} f''(t)k_2 + O(b_n^2) - \frac{1}{2}b_n^2 f''(t)k_2 \right| \\ &= \sqrt{nb_n} O(b_n^2) \\ &= O(\sqrt{nb_n} b_n^2) = O(1) \quad \text{car } b_n = O(n^{-1/5}) \end{aligned}$$

Par suite

$$\sqrt{nb_n} \left(\hat{f}_n(t) - f(t) - \frac{1}{2}b_n^2 f''(t)k_2 \right) \Rightarrow_{n \rightarrow \infty} N(0, \sigma^2(t))$$

D'où le résultat. ■

2.2.4 Erreur quadratique intégrée et fenêtre optimale

L'erreur quadratique intégrée permet de choisir une fenêtre optimale. Soit τ tel que $\pi(\tau) > 0$, et on définit, pour $\varepsilon > 0$ l'erreur quadratique intégrée de \hat{f}_n sur l'intervalle $[0, \tau - \varepsilon]$ par:

$$\mathbf{EQI}(\hat{f}_n) = \mathbf{E} \left(\int_0^{\tau-\varepsilon} [\hat{f}_n(t) - f(t)]^2 dt \right) \quad (2.23)$$

Le théorème suivant résume quelques propriétés de $\mathbf{EQI}(\hat{f}_n)$.

Théorème 2.9 *Supposons que f est continue sur $[0, \tau]$ où τ est tel que $\pi(\tau) > 0$.*

(i) *Si $b_n \rightarrow 0$ on a*

$$\mathbf{EQI}(\hat{f}_n) = \int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)]^2 dt + \int_0^{\tau-\varepsilon} \mathbf{E} [\hat{f}_n(t) - f_n^*(t)]^2 dt + o\left(\frac{1}{n}\right) \quad (2.24)$$

(ii) *Si f est de classe C^2 sur $[0, \tau]$, alors si $b_n \rightarrow 0$ le premier terme de (2.24) s'écrit*

$$\int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)]^2 dt = \frac{k_2^2}{4} b_n^4 \int_0^{\tau-\varepsilon} [f''(t)]^2 dt + o(b_n^4)$$

où k_2 est donné par **C1**.

(iii) *Si π est continue sur $[0, \tau]$, alors si $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$, le deuxième terme de (2.24) s'écrit*

$$\int_0^{\tau-\varepsilon} \mathbf{E} [\hat{f}_n(t) - f_n^*(t)]^2 dt = \frac{1}{nb_n} \int_{-1}^1 K^2(t) dt \int_0^{\tau-\varepsilon} \frac{f(t)}{S_C(t^-)} dt + o\left(\frac{1}{nb_n}\right)$$

(iv) *Si f est de classe C^2 sur $[0, \tau]$ et π est continue sur $[0, \tau]$, alors si $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$, on a*

$$\mathbf{EQI}(\hat{f}_n) = \frac{k_2^2}{4} b_n^4 \int_0^{\tau-\varepsilon} [f''(t)]^2 dt + \frac{1}{nb_n} \int_{-1}^1 K^2(t) dt \int_0^{\tau-\varepsilon} \frac{f(t)}{S_C(t^-)} dt + o(b_n^4) + o\left(\frac{1}{nb_n}\right) \quad (2.25)$$

Preuve: (i) On développe en ajoutant et retranchant f_n^* dans (2.23)

$$\begin{aligned} \mathbf{EQI}(\hat{f}_n) &= \int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)]^2 dt + \int_0^{\tau-\varepsilon} \mathbf{E} [\hat{f}_n(t) - f_n^*(t)]^2 dt \\ &\quad + 2 \int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)] [\mathbf{E} \hat{f}_n(t) - f_n^*(t)] dt \end{aligned}$$

Pour le troisième terme, d'après (i) et (iv) du **th. 2.6** on a

$$\begin{aligned}
& \left| \int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)] [\mathbf{E} \hat{f}_n(t) - f_n^*(t)] dt \right| \\
& \leq \int_0^{\tau-\varepsilon} \left[\int_{-1}^1 |K(u)| |(f(t - b_n u) - f(t))| du \right] \left[e^{-n\pi(t+b_n)} \int_{-1}^1 |K(u)| h(t - b_n u) du \right] dt \\
& \leq e^{-n\pi(\tau-\varepsilon+b_n)} \left[\sup_{t \in [0, \tau-\varepsilon]} \sup_{u \in [-1, 1]} |(f(t - b_n u) - f(t))| \right] \left[\int_{-1}^1 |K(u)| du \right]^2 H(\tau - \varepsilon + b_n) \\
& = o\left(\frac{1}{n}\right)
\end{aligned}$$

Par suite

$$\mathbf{EQI}(\hat{f}_n) = \int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)]^2 dt + \int_0^{\tau-\varepsilon} \mathbf{E} [\hat{f}_n(t) - f_n^*(t)]^2 dt + o\left(\frac{1}{n}\right)$$

D'où le résultat.

(ii) On a par définition de f_n^* et un développement de Taylor:

$$\begin{aligned}
\int_0^{\tau-\varepsilon} [f_n^*(t) - f(t)]^2 dt &= \int_0^{\tau-\varepsilon} \left[\int_{-1}^1 K(u) (f(t - b_n u) - f(t)) du \right]^2 dt \\
&= \int_0^{\tau-\varepsilon} \left[b_n f'(t) \int_{-1}^1 u K(u) du + \frac{b_n^2}{2} f''(t) \int_{-1}^1 u^2 K(u) du + o(b_n^2) \right]^2 dt \\
&= \int_0^{\tau-\varepsilon} \left[\frac{b_n^2}{2} f''(t) k_2 + o(b_n^2) \right]^2 dt \\
&= \frac{k_2^2}{4} b_n^4 \int_0^{\tau-\varepsilon} [f''(t)]^2 dt + o(b_n^4)
\end{aligned}$$

D'où le résultat.

(iii) On a

$$\begin{aligned}
\mathbf{E} [\hat{f}_n(t) - f_n^*(t)]^2 &= \mathbf{E} [\hat{f}_n(t) - \tilde{f}_n(t)]^2 + \mathbf{E} [\tilde{f}_n(t) - f_n^*(t)]^2 + \\
&\quad 2\mathbf{E} [\hat{f}_n(t) - \tilde{f}_n(t)] [\tilde{f}_n(t) - f_n^*(t)]
\end{aligned} \tag{2.26}$$

Par la preuve du **th. 2.6 (ii)** on a montré que

$$\mathbf{E} [\hat{f}_n(t) - \tilde{f}_n(t)]^2 = \frac{1}{b_n} \int_{-1}^1 K^2(u) \mathbf{E} \left[\hat{S}_{KM}^2[(t - b_n u)^-] \frac{J_n(t - b_n u)}{\tilde{Y}_n(t - b_n u)} \right] h(t - b_n u) du$$

et du fait que

$$\frac{f(t)}{S_C(t^-)} = \frac{h(t)S^2(t)}{\pi(t)}$$

et par le lemme 3.2.1 de [7], p. 99, on en déduit pour n assez grand

$$\begin{aligned}
& \sup_{t \in [0, \tau - \varepsilon]} \left| nb_n \mathbf{E} \left[\hat{f}_n(t) - \tilde{f}_n(t) \right]^2 - \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du \right| \\
&= \sup_{t \in [0, \tau - \varepsilon]} \left| n \int_{-1}^1 K^2(u) \mathbf{E} \left[\hat{S}_{KM}^2[(t - b_n u)^-] \frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right] h(t - b_n u) du - \frac{h(t)S^2(t)}{\pi(t)} \int_{-1}^1 K^2(u) du \right| \\
&= \sup_{t \in [0, \tau - \varepsilon]} \left| \int_{-1}^1 K^2(u) \mathbf{E} \left[\hat{S}_{KM}^2[(t - b_n u)^-] \left(\frac{J_n(t - b_n u)}{\frac{1}{n} \bar{Y}_n(t - b_n u)} - \frac{1}{\pi(t - b_n u)} \right) \right] h(t - b_n u) du \right. \\
&\quad + \int_{-1}^1 K^2(u) \mathbf{E} \hat{S}_{KM}^2[(t - b_n u)^-] \left(\frac{1}{\pi(t - b_n u)} - \frac{1}{\pi(t)} \right) h(t - b_n u) du \\
&\quad + \int_{-1}^1 K^2(u) \frac{\mathbf{E} \hat{S}_{KM}^2[(t - b_n u)^-]}{\pi(t)} [h(t - b_n u) - h(t)] du \\
&\quad + \int_{-1}^1 K^2(u) \frac{\mathbf{E} \left[\hat{S}_{KM}^2[(t - b_n u)^-] - S^2[(t - b_n u)^-] \right]}{\pi(t)} h(t) du \\
&\quad \left. + \int_{-1}^1 K^2(u) \frac{S^2[(t - b_n u)^-] - S^2(t)}{\pi(t)} h(t) du \right| \\
&\leq \left[\sup_{s \in [0, \tau]} \mathbf{E} \left| \frac{J_n(s)}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| \right] \left[\sup_{s \in [0, \tau]} h(s) \right] \int_{-1}^1 K^2(u) du \\
&\quad + \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} \left| \frac{1}{\pi(t - b_n u)} - \frac{1}{\pi(t)} \right| \right] \left[\sup_{s \in [0, \tau]} h(s) \right] \int_{-1}^1 K^2(u) du \\
&\quad + \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} |h(t - b_n u) - h(t)| \right] \frac{1}{\pi(t)} \int_{-1}^1 K^2(u) du \\
&\quad + 2 \left[\sup_{s \in [0, \tau]} \mathbf{E} \left(\hat{S}_{KM}(s^-) - S(s^-) \right) \right] \frac{h(t)}{\pi(t)} \int_{-1}^1 K^2(u) du \\
&\quad + 2 \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} |S[(t - b_n u)^-] - S(t)| \right] \frac{h(t)}{\pi(t)} \int_{-1}^1 K^2(u) du
\end{aligned}$$

D'où

$$\begin{aligned}
& \sup_{t \in [0, \tau - \varepsilon]} \left| nb_n \mathbf{E} \left[\hat{f}_n(t) - \tilde{f}_n(t) \right]^2 - \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du \right| \\
\leq & \left[\sup_{s \in [0, \tau]} \mathbf{E} \left| \frac{J_n(s)}{n^{-1} \bar{Y}_n(s)} - \frac{J_n(s)}{\pi(s)} \right| + \frac{[1 - \pi(\tau)]^n}{\pi(\tau)} \right] \left[\sup_{s \in [0, \tau]} h(s) \right] \int_{-1}^1 K^2(u) du \quad (2.1) \\
& + \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} \left| \frac{1}{\pi(t - b_n u)} - \frac{1}{\pi(t)} \right| \right] \left[\sup_{s \in [0, \tau]} h(s) \right] \int_{-1}^1 K^2(u) du \\
& + \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} |h(t - b_n u) - h(t)| \right] \frac{1}{\pi(t)} \int_{-1}^1 K^2(u) du \\
& + 2 [1 - \pi(\tau)]^n \frac{h(t)}{\pi(t)} \int_{-1}^1 K^2(u) du \\
& + 2 \left[\sup_{t \in [0, \tau - \varepsilon]} \sup_{u \in [-1, 1]} |S[(t - b_n u)^-] - S(t)| \right] \frac{h(t)}{\pi(t)} \int_{-1}^1 K^2(u) du \quad (2.27)
\end{aligned}$$

(où on a utilisé $|\hat{S}_{KM}^2(t) - S^2(t)| = |(\hat{S}_{KM}(t) - S(t))(\hat{S}_{KM}(t) + S(t))| \leq 2 |(\hat{S}_{KM}(t) - S(t))|$)

Puisque $\bar{Y}_n(s) \leftrightarrow B(n, \pi(s))$ pour chaque s fixé dans $[0, \tau]$ et comme $J_n(s) = 1_{\bar{Y}_n(s) > 0}$

$$\begin{aligned}
\mathbf{E} \left| \frac{J_n(s)}{n^{-1} \bar{Y}_n(s)} - \frac{J_n(s)}{\pi(s)} \right| &= \mathbf{E} \mathbf{E} \left(\left| \frac{J_n(s)}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| / \bar{Y}_n(s) > 0 \right) \\
&= \mathbf{E} \left(1_{\bar{Y}_n(s) > 0} \mathbf{E} \left(\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| / \bar{Y}_n(s) > 0 \right) \right) \\
&= \mathbf{E} \left(1_{\bar{Y}_n(s) > 0} \mathbf{E} \left(\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| \right) \right) \\
&= \mathbf{E}(J_n(s)) \mathbf{E} \left(\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| \right) \\
&= [1 - (1 - \pi(s))^n] \mathbf{E} \left[\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| / \bar{Y}_n(s) > 0 \right]
\end{aligned}$$

car $\mathbf{E} \left(\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| / \bar{Y}_n(s) > 0 \right) = \mathbf{E} \left(\left| \frac{1}{n^{-1} \bar{Y}_n(s)} - \frac{1}{\pi(s)} \right| \right)$, et par un lemme de Aalen (cf.[1], p. 18) on a

$$\lim_{n \rightarrow \infty} \sup_{s \in [0, \tau]} \mathbf{E} \left| \frac{J_n(s)}{n^{-1} \bar{Y}_n(s)} - \frac{J_n(s)}{\pi(s)} \right| = 0 \quad (2.28)$$

ce qui donne la convergence vers 0 du 1er terme de (2.27) et les trois autres termes convergent vers 0 par $b_n \rightarrow 0$ et les conditions sur les fonction π , h et S .

Nous arrivons à :

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, \tau - \varepsilon]} \left| nb_n \mathbf{E} \left[\hat{f}_n(t) - \tilde{f}_n(t) \right]^2 - \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du \right| = 0$$

D'où

$$\mathbf{E} \left[\hat{f}_n(t) - \tilde{f}_n(t) \right]^2 = \frac{1}{nb_n} \frac{f(t)}{S_C(t^-)} \int_{-1}^1 K^2(u) du + o\left(\frac{1}{nb_n}\right) \quad (2.29)$$

uniformément pour $t \in [0, \tau - \varepsilon]$.

Posons

$$B(s) = \mathbf{E} \left[\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \hat{S}_{KM}(\tau_{\bar{Y}}) [S(\tau_{\bar{Y}}) - S(s)] / S(\tau_{\bar{Y}}) \right]$$

où $\tau_{\bar{Y}} = \inf \{s, \bar{Y}(s) = 0\}$. Alors par l'équation **(2.13)** de [7] p. 104, on a pour $s \in [0, \tau]$ et du fait $1 - \Delta H(u) = \frac{1-F(s)}{1-F(s^-)}$ (cf [7] p.105) (aussi $1 - \Delta H(u) = 1$ car F continue sur $[0, \tau]$)

$$\begin{aligned} n \mathbf{E} \left[\hat{S}_{KM}(s) - S(s) - B(s) \right]^2 &= n S^2(s) \int_0^s \mathbf{E} \left[\frac{\hat{S}_{KM}^2(u^-) J_n(u)}{S^2(u) \bar{Y}_n(u)} \right] (1 - \Delta H(u)) h(u) du \\ &= S^2(s) \int_0^s \mathbf{E} \left[\hat{S}_{KM}^2(u^-) \left(\frac{J_n(u)}{n^{-1} \bar{Y}_n(u)} - \frac{J_n(u)}{\pi(u)} \right) \right] \frac{h(u)}{S^2(u)} du \\ &\quad + S^2(s) \int_0^s \mathbf{E} \left[\hat{S}_{KM}^2(u^-) \frac{J_n(u)}{\pi(u)} \right] \frac{h(u)}{S^2(u)} du \\ &\leq \sup_{u \in [0, \tau]} \mathbf{E} \left| \frac{J_n(u)}{n^{-1} \bar{Y}_n(u)} - \frac{J_n(u)}{\pi(u)} \right| \cdot \frac{\tau}{S^2(\tau)} \left[\sup_{u \in [0, \tau]} h(u) \right] \\ &\quad + \frac{\tau}{S^2(\tau) \pi(\tau)} \left[\sup_{u \in [0, \tau]} h(u) \right] \end{aligned}$$

car $S^2(s) \mathbf{E}(\hat{S}_{KM}^2(u^-) J_n(u)) \leq 1$ et S et π sont décroissantes.

Comme h , S et π sont bornées on a

$$\frac{\tau}{S^2(\tau) \pi(\tau)} \left[\sup_{u \in [0, \tau]} h(u) \right]$$

est borné. Donc dans (2.29) pour le 1er terme on applique (2.28) et on déduit que

$$\sup_{s \in [0, \tau]} \mathbf{E} \left[\hat{S}_{KM}(s) - S(s) - B(s) \right]^2 = \mathcal{O}\left(\frac{1}{n}\right) \quad (2.30)$$

De plus par l'équation (2.4) et le lemme 3.2.1 de [7], p. 99 on obtient pour n assez grand (en utilisant $(a+b)^2 \leq 2a^2 + 2b^2$ et l'inégalité de Cauchy-Schwarz)

$$\begin{aligned}
& \sup_{t \in [0, \tau - \varepsilon]} \mathbf{E} \left[\tilde{f}_n(t) - f_n^*(t) \right]^2 \\
&= \sup_{t \in [0, \tau - \varepsilon]} \left\{ \mathbf{E} \left[\int_{-1}^1 K(u) \mathbf{1}_{\{\bar{Y}_n(t - b_n u) > 0\}} \left[\hat{S}_{KM}((t - b_n u)^-) - S((t - b_n u)^-) \right] h(t - b_n u) du \right. \right. \\
&\quad \left. \left. + \int_{-1}^1 K(u) \mathbf{1}_{\{\bar{Y}_n(t - b_n u) = 0\}} f(t - b_n u) du \right]^2 \right\} \\
&\leq \sup_{t \in [0, \tau - \varepsilon]} \left\{ 4 \int_{-1}^1 K^2(u) \mathbf{1}_{\{\bar{Y}_n(t - b_n u) > 0\}} \mathbf{E} \left[\hat{S}_{KM}((t - b_n u)^-) - S((t - b_n u)^-) \right]^2 h^2(t - b_n u) du \right. \\
&\quad \left. + 4 \int_{-1}^1 K^2(u) \mathbf{1}_{\{\bar{Y}_n(t - b_n u) = 0\}} f^2(t - b_n u) du \right\} \\
&\leq \sup_{t \in [0, \tau - \varepsilon]} 8 \int_{-1}^1 K^2(u) \mathbf{E} \left[\hat{S}_{KM}((t - b_n u)^-) - S((t - b_n u)^-) - B((t - b_n u)^-) \right]^2 h^2(t - b_n u) du \\
&\quad + \sup_{t \in [0, \tau - \varepsilon]} 8 \int_{-1}^1 K^2(u) \mathbf{E} [B^2((t - b_n u)^-)] h^2(t - b_n u) du + \\
&\quad 4 \int_{-1}^1 K^2(u) \mathbf{E} \mathbf{1}_{\{\bar{Y}_n(t - b_n u) = 0\}} f^2(t - b_n u) du \\
&\leq \left[16 \sup_{s \in [0, \tau]} h^2(s) \right] \left\{ \sup_{s \in [0, \tau]} \mathbf{E} [S_n(s) - S(s) - B(s)]^2 \right\} \int_{-1}^1 K^2(u) du \\
&\quad + \left[\sup_{s \in [0, \tau]} \mathbf{P} (\bar{Y}(s) = 0) \right] \left[16 \sup_{s \in [0, \tau]} h^2(s) + 4 \sup_{s \in [0, \tau]} f^2(s) \right] \int_{-1}^1 K^2(u) du \\
&\leq \left[8 \sup_{s \in [0, \tau]} h^2(s) \right] \left\{ \sup_{s \in [0, \tau]} \mathbf{E} [S_n(s) - S(s) - B(s)]^2 \right\} \int_{-1}^1 K^2(u) du \\
&\quad + e^{-n\pi(\tau)} \left[8 \sup_{s \in [0, \tau]} h^2(s) + 4 \sup_{s \in [0, \tau]} f^2(s) \right] \int_{-1}^1 K^2(u) du
\end{aligned}$$

où on a utilisé

$$\begin{aligned}
B(s) &= \mathbf{E} \left[\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \hat{S}_{KM}(\tau_{\bar{Y}}) [S(\tau_{\bar{Y}}) - S(s)] / S(\tau_{\bar{Y}}) \right] \leq \mathbf{E} \left[\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} S_n(\tau_{\bar{Y}}) (1 - S(s) / S(\tau_{\bar{Y}})) \right] \\
&\leq \mathbf{E} \left[\mathbf{1}_{\{\tau_{\bar{Y}} < s\}} \right] = \mathbf{P} (\bar{Y}_n(s) = 0)
\end{aligned}$$

et $P(\bar{Y}_n(s) = 0) = [(1 - (1 - F(s))(1 - G(s^-))]^n = [1 - \pi(s)]^n \leq e^{-n\pi(s)}$ (cf. preuve du **th.2.6**).

Par le fait que f et h sont bornées et par **(2.30)** on conclut que

$$\sup_{t \in [0, \tau - \varepsilon]} \mathbf{E} \left[\tilde{f}_n(t) - f_n^*(t) \right]^2 = \mathcal{O}\left(\frac{1}{n}\right) \quad (2.31)$$

Finalement par (2.29) et (2.31) et par l'inégalité de Cauchy-Schwarz, on a

$$\begin{aligned} & \sup_{t \in [0, \tau - \varepsilon]} \mathbf{E} \left[\left(\hat{f}_n(t) - \tilde{f}_n(t) \right) \left(\tilde{f}_n(t) - f_n^*(t) \right) \right] \\ & \leq \sup_{t \in [0, \tau - \varepsilon]} \mathbf{E}^{1/2} \left(\hat{f}_n(t) - \tilde{f}_n(t) \right)^2 \mathbf{E}^{1/2} \left(\tilde{f}_n(t) - f_n^*(t) \right)^2 \\ & = \mathcal{O}\left(\frac{1}{n\sqrt{b_n}}\right) \end{aligned} \quad (2.32)$$

Par conséquent, par (2.26) et en utilisant (2.29), (2.31) et (??), on arrive

$$\begin{aligned} \int_0^{\tau - \varepsilon} \mathbf{E} \left[\hat{f}_n(t) - f_n^*(t) \right]^2 dt &= \frac{1}{nb_n} \int_{-1}^1 K^2(t) dt \int_0^{\tau - \varepsilon} \frac{f(t)}{S_C(t^-)} dt + o\left(\frac{1}{nb_n}\right) + \\ &+ \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}\left(\frac{1}{n\sqrt{b_n}}\right) \end{aligned}$$

comme

$$\mathcal{O}\left(\frac{1}{n}\right) = \mathcal{O}\left(\frac{1}{n\sqrt{b_n}}\right) = o\left(\frac{1}{nb_n}\right)$$

(iv) En injectant **(ii)** et **(iii)** dans **(i)** on obtient **(iv)** en remarquant que $o\left(\frac{1}{n}\right) = o\left(\frac{1}{nb_n}\right)$.

Ce qui termine la démonstration. ■

Le résultat du **th. 2.9** permet un choix optimal de la fenêtre b_n en minimisant par rapport à b_n l'erreur quadratique intégrée de \hat{f}_n ce qui revient à minimiser les deux termes principaux de (2.25).

On a le résultat suivant:

Corollaire 2.3 *La fenêtre optimale asymptotique est*

$$b_n^* = \frac{1}{n^{1/5}} \frac{1}{k_2^{2/5}} \left[\int_{-1}^1 K^2(t) dt \int_0^{\tau - \varepsilon} \frac{f(t)}{S(t^-)} dt \right]^{1/5} \left[\int_0^{\tau - \varepsilon} [f''(t)]^2 dt \right]^{-1/5} = \mathcal{O}\left(\frac{1}{n^{1/5}}\right) \quad (2.33)$$

Remarque 2.3 *Si on injecte l'expression de la fenêtre optimale b_n^* dans (2.25) on obtient*

l'expression optimale de l' $\mathbf{EQI}(\hat{f}_n)$

$$\begin{aligned}\mathbf{EQI}(\hat{f}_n) &= \frac{5 k_2^{2/5}}{4 n^{4/5}} \left[\int_{-1}^1 K^2(t) dt \int_0^{\tau-\varepsilon} \frac{f(t)}{S(t^-)} dt \right]^{4/5} \left[\int_0^{\tau-\varepsilon} [f''(t)]^2 dt \right]^{1/5} + o(n^{-4/5}) \\ &= \mathcal{O}\left(\frac{1}{n^{4/5}}\right) + o\left(\frac{1}{n^{4/5}}\right) = \mathcal{O}\left(\frac{1}{n^{4/5}}\right)\end{aligned}$$

Chapitre 3

Estimation de la fonction de risque

3.1 Données non censurées

3.1.1 Introduction

Dans cette partie nous allons étudier un estimateur de la fonction de risque dans le cas de données non censurées, par la méthode du noyau en développant l'article de Henrik Ramlau-Hansen (1983) [9].

Dans cet article Ramlau-Hansen propose une méthodologie, citée par Klein et Moeschberger (2003), permettant de construire un estimateur \hat{h}_n de la fonction de risque par lissage. Ce lissage s'opère grâce à la convolution par un noyau et utilisation d'une fenêtre b_n .

Soit T_1, T_2, \dots, T_n des v.a i.i.d. de densité f et de fonction de répartition F . L'estimateur à noyau de la densité proposé par Roseblatt (1956) est défini par

$$\hat{f}_n(t) = \frac{1}{b_n} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{b_n}\right) dF_n(s) \quad (3.1)$$

où F_n est la f.d.r. empirique des T_i , b_n est la fenêtre avec $b_n \rightarrow_{n \rightarrow \infty} 0$ et K un noyau de Parzen-Roseblatt de support $[-1, 1]$ (ie K bornée, intégrable, symétrique, $\int_{-1}^1 K d\mu = 1$ (μ une mesure sur \mathbb{R} et $\lim_{|x| \rightarrow +\infty} |x| |K(x)| = 0$)).

Rappelons que la fonction de risque $h(t) = \frac{f(t)}{1-F(t)}$ avec $t \geq 0$ et $F(t) < 1$. Watson et Leadbetter (1964) (cf. [19]) ont étudié un estimateur à noyau de la fonction de risque h donné

par

$$\hat{h}_n(t) = \frac{1}{n-i+1} \sum_{i=1}^n \frac{1}{b_n} K\left(\frac{t-T_{(i)}}{b_n}\right) \quad (3.2)$$

où $T_{(1)}, T_{(2)}, \dots, T_{(n)}$ est la statistique d'ordre des T_i .

Si on pose $\bar{N}_n(t) = nF_n(t) = \sum_{i=1}^n \mathbf{1}_{\{T_i \leq t\}}$ et $\bar{Y}_n(t) = n - \bar{N}_n(t^-) = \sum_{i=1}^n \mathbf{1}_{\{T_i \geq t\}}$, alors **(3.2)** peut s'écrire

$$\hat{h}_n(t) = \frac{1}{b_n} \int_0^{+\infty} K\left(\frac{t-s}{b_n}\right) d\hat{H}_{NA}(s) \quad (3.3)$$

où $\hat{H}_{NA}(t) = \int_0^t \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)}$ est l'estimateur de Nelson-Aalen du taux de hasard cumulé. En effet

$$\begin{aligned} \hat{h}_n(t) &= \frac{1}{b_n} \int_0^{+\infty} K\left(\frac{t-s}{b_n}\right) d\hat{H}_{NA}(s) \\ &= \frac{1}{b_n} \int_0^{+\infty} K\left(\frac{t-s}{b_n}\right) \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} \\ &= \frac{1}{b_n} \int_0^{+\infty} K\left(\frac{t-s}{b_n}\right) \frac{ndF_n(s)}{\bar{Y}_n(s)} \\ &= \frac{1}{b_n} \sum_{i=1}^n K\left(\frac{t-T_{(i)}}{b_n}\right) \frac{1}{\bar{Y}_n(T_{(i)})} \\ &= \frac{1}{b_n} \sum_{i=1}^n K\left(\frac{t-T_{(i)}}{b_n}\right) \frac{1}{n - \bar{N}_n(T_{(i)})} \\ &= \frac{1}{b_n} \sum_{i=1}^n K\left(\frac{t-T_{(i)}}{b_n}\right) \frac{1}{n-i+1} \end{aligned}$$

L'estimateur **(3.2)** a été aussi étudié par Rice et Roseblatt (1976) et Yandell (1981). Dans l'étude de l'estimation de la densité ou du taux de hasard par la méthode du noyau, la plupart des auteurs considèrent des observations i.i.d. mais non censurées. En réalité, en épidémiologie, démographie, études de survie, criminologie et autres domaines les observations sont souvent censurées (ou au moins une partie). Donc les résultats avec des variables i.i.d. ne peuvent pas s'appliquer. En utilisant les processus ponctuels continus et les intégrales stochastiques, Aalen (1978) montre qu'il est possible de traiter une telle situation.

Soit $H(t) = \int_0^t h(s)ds$. Rappelons que l'estimateur de Nelson-Aalen de H s'écrit:

$$\hat{H}_{NA}(t) = \int_0^t \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)}$$

et que

$$d\bar{N}_n(t) = h(t)\bar{Y}_n(t)dt + d\bar{M}_n(t)$$

Si $\bar{Y}_n(t) = 0$ on modifie H par \tilde{H}_n avec

$$\tilde{H}_n(t) = \int_0^t J_n(s)h(s)ds$$

où $J_n(s) = \mathbf{1}_{\{\bar{Y}_n(s) > 0\}}$. Un estimateur naturel de \tilde{H}_n est

$$\hat{H}_n(t) = \int_0^t J_n(s) \frac{d\bar{N}_n(s)}{\bar{Y}_n(s)}$$

Par la théorie des intégrales stochastiques, $\left((\hat{H}_n - \tilde{H}_n)(t), t \geq 0 \right)$ est une martingale de carré intégrable et de variation $\langle \hat{H}_n - \tilde{H}_n \rangle_t = \int_0^t \frac{J(s)}{\bar{Y}_n^2(s)} h(s)ds$ (cf chap. 1 **Th. 1.3**). On note l'erreur quadratique par

$$\eta_n(t) = \mathbf{E} \left(\hat{H}_n(t) - \tilde{H}_n(t) \right)^2 = \mathbf{E} \left(\int_0^t \frac{J(s)}{\bar{Y}_n^2(s)} h(s)ds \right)$$

On pose

$$\hat{\eta}_n(t) = \int_0^t J(s) \frac{d\bar{N}_n(s)}{\bar{Y}_n^2(s)}$$

En remplaçant $d\bar{N}_n(t)$ par $h(t)\bar{Y}_n(t)dt + d\bar{M}_n(t)$, on obtient

$$\hat{\eta}_n(t) = \int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} h(s)ds + \int_0^t \frac{J(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s)$$

Calculons $\mathbf{E}(\hat{\eta}_n(t))$:

$$\begin{aligned}\mathbf{E}(\hat{\eta}_n(t)) &= \mathbf{E} \left[\int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds + \int_0^t \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right] \\ &= \mathbf{E} \left(\int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds \right) + \mathbf{E} \left(\int_0^t \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right)\end{aligned}$$

Comme $\bar{M}_n(s)$ est une martingale centrée, alors

$$\mathbf{E} \left(\int_0^t \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right) = 0$$

Par suite

$$\mathbf{E}(\hat{\eta}_n(t)) = \mathbf{E} \left(\int_0^t \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds \right) = \eta_n(t)$$

Donc $\hat{\eta}_n(t)$ est un estimateur sans biais de $\eta_n(t)$

3.1.2 Estimateur à noyau de la fonction de risque h

Soit K une fonction bornée d'intégrale 1 et b_n une fenêtre positive. L'estimateur à noyau correspondant au taux de hasard h est donné par

$$\hat{h}_n(t) = \frac{1}{b_n} \int_0^1 K \left(\frac{t-s}{b_n} \right) d\hat{H}_{NA}(s)$$

Les instants de sauts du processus N sont $T_{(1)}, T_{(2)}, \dots, T_{(n)}$, donc

$$\hat{h}_n(t) = \frac{\frac{1}{b_n} \sum_i K \left(\frac{t-T_{(i)}}{b_n} \right)}{\bar{Y}_n(T_{(i)})} \quad (3.4)$$

3.1.3 Moyenne et variance de l'estimateur à noyau \hat{h}_n

Pour simplifier les calculs, on suppose que le support du noyau K est $[-1, 1]$ et que $0 < b_n < 1/2$.

On pose pour $t \in [b_n, 1 - b_n]$

$$\tilde{h}_n(t) = \frac{1}{b_n} \int_0^1 K\left(\frac{t-s}{b_n}\right) d\tilde{H}_n(s) = \int_{-1}^1 K(u)h(t-b_nu)J(t-b_nu)du \quad (3.5)$$

Proposition 1 Posons $K_{b_n}(s) = \frac{1}{b_n}K\left(\frac{s}{b_n}\right)$ et $j_n(s) = \mathbf{E}(J_n(s))$ alors pour tout $t \geq 0$

$$\mathbf{E}\left(\hat{h}_n(t)\right) = \mathbf{E}\left(\tilde{h}_n(t)\right) = (K_{b_n} * hj_n)(t) \quad (3.6)$$

et

$$\sigma_n^2(t) := \mathbf{E}\left(\hat{h}_n(t) - \tilde{h}_n(t)\right)^2 = \frac{1}{b_n} \int_{-1}^1 K^2(u)h(t-b_nu)\mathbf{E}\left(\frac{J_n(t-b_nu)}{\bar{Y}_n(t-b_nu)}\right) du \quad (3.7)$$

proposition 3.1 Pour tout $t \in [b_n, 1 - b_n]$ un estimateur sans biais de $\sigma^2(t)$ est

$$\hat{\sigma}_n^2(t) = \frac{1}{b_n^2} \int_{-1}^1 K^2\left(\frac{t-s}{b_n}\right) \left(\frac{J_n(s)}{\bar{Y}_n^2(s)}\right) d\bar{N}_n(s) \quad (3.8)$$

Preuve: On a

$$\hat{h}_n(t) - \tilde{h}_n(t) = \frac{1}{b_n} \int_0^1 K\left(\frac{t-s}{b_n}\right) d\left(\hat{H}_{NA} - \tilde{H}_n\right)(s) \quad (3.9)$$

Comme $(\hat{H}_{NA} - \tilde{H}_n)$ est une martingale, $(\hat{h}_n(t) - \tilde{h}_n(t))$ est aussi une martingale d'espérance

$$\mathbf{E}\left(\hat{h}_n(t) - \tilde{h}_n(t)\right) = \mathbf{E}\left(\hat{h}_n(0) - \tilde{h}_n(0)\right) = 0.$$

D'où

$$\begin{aligned} \mathbf{E}\left(\hat{h}_n(t)\right) &= \mathbf{E}\left(\tilde{h}_n(t)\right) = \mathbf{E}\left(\int_{-1}^1 K(u)h(t-b_nu)J_n(t-b_nu)du\right) \\ &= \int_{-1}^1 K(u)h(t-b_nu)\mathbf{E}(J_n(t-b_nu)) du \\ &= \int_{-1}^1 K(u)h(t-b_nu)j_n(t-b_nu)du = (K_{b_n} * hj_n)(t) \end{aligned}$$

Pour $\sigma^2(t)$ on a

$$\begin{aligned}
\sigma^2(t) &= \mathbf{E} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right)^2 = \frac{1}{b_n^2} \mathbf{E} \left(\int_0^1 K \left(\frac{t-s}{b_n} \right) d \left(\hat{H}_{NA} - \tilde{H}_n \right) (s) \right)^2 \\
&= \frac{1}{b_n^2} \mathbf{E} \left(\int_0^1 K^2 \left(\frac{t-s}{b_n} \right) d \langle \hat{H}_n - \tilde{H}_n \rangle_s \right) \\
&= \frac{1}{b_n^2} \mathbf{E} \left(\int_0^1 K^2 \left(\frac{t-s}{b_n} \right) h(s) \frac{J_n(s)}{\bar{Y}_n(s)} ds \right) \\
&= \frac{1}{b_n} \int_0^1 K^2(u) h(t - b_n u) \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) ds
\end{aligned}$$

Montrons que $\hat{\sigma}_n^2(t)$ est un estimateur sans biais de $\sigma^2(t)$.

En utilisant: $d\bar{N}_n(s) = \bar{Y}_n(s)h(s)ds + d\bar{M}_n(s)$ on a

$$\begin{aligned}
\hat{\sigma}_n^2(t) &= \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right)^2 d\bar{N}_n(s) \\
&= \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds + \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \quad (3.10)
\end{aligned}$$

Calculons $\mathbf{E}(\hat{\sigma}_n^2(t))$

$$\begin{aligned}
\mathbf{E}(\hat{\sigma}_n^2(t)) &= \mathbf{E} \left(\frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds \right) + \mathbf{E} \left(\frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right) \\
&= \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) h(s) \mathbf{E} \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) ds + \frac{1}{b_n^2} \mathbf{E} \left(\int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right) \\
&= \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) h(s) \mathbf{E} \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) ds \\
&= \frac{1}{b_n} \int_{-1}^1 K^2(u) h(t - b_n u) \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) ds \\
&= \sigma_n^2(t)
\end{aligned}$$

D'où le résultat.

Si on note le premier terme de l'équation (3.10) par

$$\sigma_n^{*2}(t) = \frac{1}{b_n^2} \int_{-1}^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} h(s) ds$$

on obtient $\mathbf{E}(\hat{\sigma}_n^2(t)) = \mathbf{E}(\sigma_n^{*2}(t))$.

Donc de (3.10) on a

$$\begin{aligned}
\mathbf{E} (\hat{\sigma}_n^2(t) - \sigma_n^{*2}(t))^2 &= \mathbf{E} \left(\frac{1}{b_n^2} \int_{-1}^1 K^2\left(\frac{t-s}{b_n}\right) \frac{J(s)}{\bar{Y}_n^2(s)} d\bar{M}_n(s) \right)^2 \\
&= \frac{1}{b_n^4} \mathbf{E} \left(\int_{-1}^1 K^4\left(\frac{t-s}{b_n}\right) \frac{J(s)}{\bar{Y}_n^4(s)} d < \bar{M}_n > (s) \right) \\
&= \frac{1}{b_n^3} \int_{-1}^1 K^4(u) h(t - b_n u) \mathbf{E} \left(\frac{J(t - b_n u)}{\bar{Y}_n^3(t - b_n u)} \right) du
\end{aligned}$$

D'où le résultat. ■

Le résultat suivant donne une expression asymptotique de $\mathbf{E} (\hat{\sigma}_n^2(t) - \sigma_n^{*2}(t))^2$.

De même l'équation (3.6) permet de déduire que l'estimateur à noyau de la fonction de risque est asymptotiquement sans biais.

Proposition 2 (a) *Si h est continue au point t et si $j_n(s) = \mathbf{E} J_n(s) \rightarrow_{n \rightarrow \infty} 1$ uniformément dans un voisinage de $t > 0$, alors*

$$\mathbf{E} \hat{h}_n(t) \rightarrow_{n \rightarrow \infty} h(t) \quad (3.11)$$

proposition 3.2 (b) *Si $n \mathbf{E} \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) \rightarrow \frac{1}{\tau(s)}$ uniformément dans un voisinage de t et h et τ sont continues au point $t > 0$, alors quand $b_n \rightarrow 0$*

$$\sigma_n^2(t) = \mathbf{E} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right)^2 = \frac{1}{nb_n} \frac{h(t)}{\tau(t)} \int_{-1}^1 K^2(u) du + o \left(\frac{1}{nb_n} \right) \quad (3.12)$$

Preuve: (a) Par la relation (3.6) et $j_n(s) \rightarrow_{n \rightarrow \infty} 1$ uniformément sur un voisinage de t ce qui donne une limite continue et avec h continue on a $|K(v)h(t - vb_n)j_n(t - vb_n)| \leq |K(v)|C$ où $C > 0$ et le noyau K est intégrable. Par suite, de la convergence dominée on obtient

$$\begin{aligned}
\mathbf{E} \hat{h}_n(t) &= (K_{b_n} * h j_n)(t) = \frac{1}{b_n} \int K\left(\frac{t-u}{b_n}\right) h(u) j_n(u) du \\
&= \int_{-1}^1 K(v) h(t - vb_n) j_n(t - vb_n) dv \rightarrow h(t) \int_{-1}^1 K(v) dv = h(t)
\end{aligned}$$

(b) On a par (3.7)

$$\begin{aligned}
\sigma_n^2(t) &= \mathbf{E} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right)^2 = \frac{1}{b_n} \int_{-1}^1 K^2(u) h(t - b_n u) \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) du \\
&= \frac{1}{nb_n} \int_{-1}^1 K^2(u) h(t - b_n u) n \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) du \\
&= \frac{1}{nb_n} \int_{-1}^1 K^2(u) h(t - b_n u) \left(n \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) - \frac{1}{\tau(t - b_n u)} \right) du + \\
&\quad \frac{1}{nb_n} \int_{-1}^1 K^2(u) h(t - b_n u) \frac{1}{\tau(t - b_n u)} du \\
&= o \left(\frac{1}{nb_n} \right) + \frac{1}{nb_n} \int_{-1}^1 K^2(u) \frac{h(t)}{\tau(t)} du
\end{aligned}$$

où dans la première intégrale, comme $n \mathbf{E} \left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) \rightarrow_{n \rightarrow \infty} \frac{1}{\tau(s)}$ uniformément dans un voisinage de t et h et τ sont continues en t et par la convergence dominée:

$$\int_{-1}^1 K^2(u) h(t - b_n u) \left(n \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) - \frac{1}{\tau(t - b_n u)} \right) du \rightarrow_{n \rightarrow \infty} 0$$

et on obtient $o \left(\frac{1}{nb_n} \right)$.

Pour le deuxième on a

$$\int_{-1}^1 K^2(u) h(t - b_n u) \frac{1}{\tau(t - b_n u)} du \rightarrow_{n \rightarrow \infty} \frac{h(t)}{\tau(t)} \int_{-1}^1 K^2(u) du$$

Ce qui termine la démonstration. ■

Remarque 3.1 La relation (3.12) montre que si $nb_n \rightarrow \infty$ alors

$$\mathbf{E} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right)^2 \rightarrow_{n \rightarrow \infty} 0$$

et par suite

$$\hat{h}_n(t) - \tilde{h}_n(t) \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

3.1.4 Estimation des dérivées de la fonction de risque

L'estimateur à noyau (3.3) peut s'écrire

$$\hat{h}_n = K_{b_n} * \hat{H}_{NA} \quad (3.13)$$

Pour estimer la dérivée d'ordre k de la fonction de risque h , il est naturel de considérer la dérivée d'ordre k de \hat{h}_n . Si on suppose que K est absolument continu à l'ordre k c'est à dire que $K', K'', \dots, K^{(k-1)}$ existent et sont absolument continus et donc on peut dériver \hat{h}_n jusqu'à l'ordre k . L'estimateur correspondant sera

$$\hat{h}_n^{(k)} = \frac{1}{b_n^{k+1}} \int_0^1 K^{(k)}\left(\frac{t-s}{b_n}\right) d\hat{H}_{NA}(s)$$

Cet estimateur a les mêmes propriétés que \hat{h}_n et sont résumés dans le résultat suivant:

Proposition 3

(a) On suppose que h et $j = \mathbf{E}J$ sont de classe $C^{(k)}$ sur $[0, 1]$. Alors

$$\mathbf{E}(\hat{h}_n^{(k)}(t)) = \mathbf{E}(\tilde{h}_n^{(k)}(t)) = K_{b_n}^{(k)} * (hj)^{(k)}(t) \quad (3.14)$$

où $\tilde{h}_n^{(k)}(t) = (K_{b_n}^{(k)} * hj)(t)$

De plus pour $t \in [b_n, 1 - b_n]$, on a

$$\begin{aligned} \sigma_{n,k}^2(t) &= \mathbf{E} \left(\hat{h}_n^{(k)}(t) - \tilde{h}_n^{(k)}(t) \right)^2 \\ &= \frac{1}{b_n^{2k+1}} \int_{-1}^1 K^{(k)}(u)^2 h(t - b_n u) \mathbf{E} \left(\frac{J_n(t - b_n u)}{\bar{Y}_n(t - b_n u)} \right) du \end{aligned}$$

et cette quantité est estimée par

$$\hat{\sigma}_{n,k}^2(t) = \frac{1}{b_n^{2k+1}} \int_0^1 K^{(k)}\left(\frac{t-s}{b_n}\right)^2 \frac{J_n(s)}{\bar{Y}_n^2(s)} d\bar{N}(s)$$

(b) Sous les mêmes conditions de la proposition 3.3, on a

$$\begin{aligned}\sigma_{n,k}^2(t) &= \mathbf{E}(\hat{h}_n^{(k)}(t) - \tilde{h}_n^{(k)}(t))^2 \\ &= \frac{1}{b_n^{2k+1}} \frac{h(t)}{\tau(t)} \int_{-1}^1 K^{(k)}(u)^2 du + o((b_n^{2k+1})^{-1})\end{aligned}$$

3.1.5 Propriétés asymptotiques

Le résultat suivant donne la convergence ponctuelle dans L^2 de \hat{h}_n .

Proposition 4 Si $n\mathbf{E}\left(\frac{J_n(s)}{Y_n(s)}\right) \rightarrow_{n \rightarrow \infty} \frac{1}{\tau(s)}$ uniformément dans un voisinage de t et h et τ sont continues en $t > 0$, alors

$$\mathbf{E}\left(\hat{h}_n(t) - h(t)\right)^2 \rightarrow_{n \rightarrow \infty} 0$$

quand $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$

Preuve: Par la décomposition de $\mathbf{E}\left(\hat{h}_n(t) - h(t)\right)^2$ on a

$$\mathbf{E}\left(\hat{h}_n(t) - h(t)\right)^2 \leq 2\mathbf{E}\left(\hat{h}_n(t) - \tilde{h}_n(t)\right)^2 + 2\mathbf{E}\left(\tilde{h}_n(t) - h(t)\right)^2$$

Par la proposition 3.2 on a

$$\mathbf{E}\left(\hat{h}_n(t) - \tilde{h}_n(t)\right)^2 \rightarrow_{n \rightarrow \infty} 0$$

Montrons que

$$\mathbf{E}\left(\tilde{h}_n(t) - h(t)\right)^2 \rightarrow_{n \rightarrow \infty} 0$$

Comme $J_n(u) \xrightarrow{P} 1$ uniformément et la convergence dominée donne

$$\tilde{h}_n(t) - h(t) = \int_{-1}^1 K(u) [h(t - b_n u) J_n(t - b_n u) - h(t)] du \xrightarrow{P} 0$$

par suite elle est bornée : $|\tilde{h}_n(t) - h(t)| \leq C$ où $C > 0$. Donc par $\mathbf{E}\left(\hat{h}_n(t)\right) = \mathbf{E}\left(\tilde{h}_n(t)\right)$ (cf. Prop. 3.1) et par $\mathbf{E}\left(\hat{h}_n(t)\right) \rightarrow h(t)$ (cf. Prop 3.2 a)

$$\begin{aligned}\mathbf{E}\left(\tilde{h}_n(t) - h(t)\right)^2 &\leq C\mathbf{E}\left(\tilde{h}_n(t) - h(t)\right) = C\left(\mathbf{E}\tilde{h}_n(t) - h(t)\right) \\ &\leq C\left|\mathbf{E}\hat{h}_n(t) - h(t)\right| \rightarrow_{n \rightarrow \infty} 0\end{aligned}$$

D'où le résultat de la proposition. ■

Pour la convergence uniforme dans L^2 de \hat{h}_n nous considérons un intervalle fixé $[z_0, z_1]$ avec $0 < z_0 < z_1 < 1$. On a le résultat suivant:

Théorème 3.1 *Supposons que*

- (i) $J_n \xrightarrow{n \rightarrow \infty} 1$ en probabilité et uniformément sur $[0, 1]$.
- (ii) h est continue sur $[0, 1]$.
- (iii) $n\eta_n(1) := n \int_0^1 \mathbf{E} \left(\frac{J_n(s)}{Y_n(s)} \right) h(s) ds$ est borné quand $n \rightarrow \infty$.
- (iv) le noyau K est à variations bornées.

Alors si $b_n \rightarrow 0$ et $nb_n^2 \rightarrow \infty$ on a:

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - h(t) \right|^2 \right) \xrightarrow{n \rightarrow \infty} 0$$

Preuve: On décompose

$$\left| \hat{h}_n(t) - h(t) \right|^2 \leq 2 \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 + 2 \left| \tilde{h}(t) - h(t) \right|^2$$

Par suite

$$\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - h(t) \right|^2 \leq \sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 + \sup_{t \in [z_0, z_1]} \left| \tilde{h}(t) - h(t) \right|^2$$

Il suffit de prouver que

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) \xrightarrow{n \rightarrow \infty} 0 \quad (3.15)$$

et

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} \left| \tilde{h}(t) - h(t) \right|^2 \right) \xrightarrow{n \rightarrow \infty} 0 \quad (3.16)$$

Par définition on a

$$\hat{h}_n(t) - \tilde{h}(t) = \frac{1}{b_n} \int_0^1 K \left(\frac{t-s}{b_n} \right) d(\hat{H}_n - \tilde{H}_n)(s)$$

Comme K est à variations bornées, une intégration par parties donne

$$\left| \hat{h}_n(t) - \tilde{h}(t) \right| \leq \frac{2}{b_n} V(K) \sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|$$

où $V(K)$ est la variation totale de K .

Par suite

$$\left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \leq \frac{4}{b_n^2} C \sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2$$

D'où

$$\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \leq \frac{4C}{b_n^2} \sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2$$

Pour prouver (3.15) il suffit de montrer que

$$\frac{1}{b_n^2} \mathbf{E} \left(\sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2 \right) \rightarrow_{n \rightarrow \infty} 0$$

Si on applique l'inégalité de Doob (pour une martingale (M_t) : $E(\sup_{s \leq t} M_s^2) \leq 4 \sup_{s \leq t} E M_s^2$) à la martingale $(\hat{H}_n(s) - \tilde{H}_n(s))$ et on obtient

$$\begin{aligned} \mathbf{E} \left(\sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2 \right) &\leq 4 \sup_{s \in [0,1]} \mathbf{E} \left(\hat{H}_n(s) - \tilde{H}_n(s) \right)^2 \\ &= 4 \sup_{s \in [0,1]} \eta_n(s) = 4\eta_n(1) \end{aligned}$$

Par suite

$$\begin{aligned} \mathbf{E} \left(\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) &\leq \frac{4}{b_n^2} C \mathbf{E} \left(\sup_{s \in [0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2 \right) \\ &\leq \frac{16}{nb_n^2} C n \eta_n(1) \end{aligned}$$

Comme $n\eta_n(1)$ est borné pour n assez grand et $nb_n^2 \rightarrow \infty$ on en déduit que

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) \rightarrow_{n \rightarrow \infty} 0$$

Montrons maintenant (3.16)

Dans la preuve de la prop. **3.4** on a montré qu'il existe une constante $C_n(t) > 0$ telle que $|\tilde{h}_n(t) - h(t)| \leq C_n(t)$ (où $C_n(t)$ est choisi petit pour n assez grand pour chaque t). Posons $C_n = \sup_{t \in [z_0, z_1]} C_n(t)$

$$\sup_{t \in [z_0, z_1]} |\tilde{h}_n(t) - h(t)|^2 \leq C_n \sup_{t \in [z_0, z_1]} |\tilde{h}_n(t) - h(t)|$$

D'où

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} |\tilde{h}_n(t) - h(t)|^2 \right) \leq C_n \mathbf{E} \left(\sup_{t \in [z_0, z_1]} |\tilde{h}_n(t) - h(t)| \right) \xrightarrow{n \rightarrow \infty} 0$$

car

$$\sup_{t \in [z_0, z_1]} |\tilde{h}_n(t) - h(t)| \leq \sup_{t \in [z_0, z_1]} C_n(t) \xrightarrow{n \rightarrow \infty} 0$$

et alors par la convergence dominée on a

$$\mathbf{E} \left(\sup_{t \in [z_0, z_1]} |\hat{h}_n(t) - h(t)|^2 \right) \xrightarrow{n \rightarrow \infty} 0$$

D'où le résultat. ■

3.1.6 Normalité asymptotique

Parzen (1962) a montré que l'estimateur à noyau de la densité est asymptotiquement normal quand le nombre d'observations tend vers l'infini. Dans cette section on va étudier la normalité asymptotique de l'estimateur \hat{h}_n en utilisant les processus de comptage. Associons à un processus de comptage \bar{N}_n sa martingale correspondante

$$\bar{M}_n(t) = \bar{N}_n(t) - \int_0^t \lambda_n(s) ds$$

où $\lambda_n(s) = \bar{Y}_n(s)h(s)$ est la suite de processus intensité. Soit L_n la suite de processus prévisible vérifiant

$$\mathbf{E} \int_0^1 L_n^2(s) \lambda_n(s) ds < \infty$$

Posons

$$\tilde{M}_n(t) = \int_0^t L_n(s) d\bar{M}_n(s)$$

On admet le résultat auxiliaire suivant donnant une loi limite pour des martingales (conséquence du Th. de Robolledo cf. Annexe Th. 4.11.)

proposition 3.3 *Supposons que les conditions suivants sont vérifiées*

$$(i) \forall \varepsilon > 0, \int_0^1 L_n^2(s) \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} \lambda_n(s) ds \xrightarrow{P} 0$$

$$(ii) \int_0^1 L_n^2(s) \lambda_n(s) ds \xrightarrow{P} 1$$

Alors

$$\tilde{M}_n(1) \Longrightarrow_{n \rightarrow \infty} N(0, 1)$$

Nous utilisons la proposition précédente pour établir la loi asymptotique de \hat{h}_n . On introduit les notations suivantes. Par (3.9) on a

$$\begin{aligned} \hat{h}_n(t) &= \tilde{h}_n(t) + \frac{1}{b_n} \int_0^1 K\left(\frac{t-s}{b_n}\right) d\left(\hat{H}_{NA} - \tilde{H}_n\right)(s) \\ &= \tilde{h}_n(t) + \frac{1}{b_n} \int_0^1 K\left(\frac{t-s}{b_n}\right) \frac{J_n(s)}{\bar{Y}_n(s)} d\bar{M}_n(s) \end{aligned}$$

Pour t fixé on a

$$\sqrt{nb_n} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right) = \int_0^1 L_n(s) d\bar{M}_n(s)$$

où

$$L_n(s) := \sqrt{\frac{n}{b_n}} K\left(\frac{t-s}{b_n}\right) \frac{J_n(s)}{\bar{Y}_n(s)}$$

On a le théorème suivant

Théorème 3.2 *Supposons que:*

$$(i) n \frac{J_n(s)}{\bar{Y}_n(s)} \xrightarrow{P} 1/\tau(s) \text{ uniformément sur un voisinage de } t.$$

(ii) h et τ sont continues au point t .

Si $b_n \rightarrow 0$ et $nb_n \rightarrow \infty$ alors

$$\sqrt{nb_n} \left(\hat{h}_n(t) - \tilde{h}_n(t) \right) \Longrightarrow_{n \rightarrow \infty} N(0, v(t))$$

$$\text{où } v(t) = h(t)/\tau(t) \int_{-1}^1 K^2(u) du$$

Preuve: Vérifions les deux conditions de la proposition précédente.

Pour la première condition on a

$$\begin{aligned}
\int_0^1 L_n^2(s) \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} \lambda_n(s) ds &= \int_0^1 \frac{n}{b_n} K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n^2(s)} \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} \bar{Y}_n(s) h(s) ds \\
&= \frac{n}{b_n} \int_0^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} h(s) ds \\
&= \int_0^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} h(s) ds
\end{aligned}$$

Comme

$$\{|L_n(s)| > \varepsilon\} = \left\{ \left| K \left(\frac{t-s}{b_n} \right) \frac{nJ_n(s)}{\bar{Y}_n(s)} \right| > \varepsilon \sqrt{nb_n} \right\} \subset \left\{ C \frac{nJ_n(s)}{\bar{Y}_n(s)} > \varepsilon \sqrt{nb_n} \right\} \xrightarrow{n \rightarrow \infty} \emptyset$$

car $\frac{nJ_n(s)}{\bar{Y}_n(s)} \xrightarrow{P} 1/\tau$ uniformément sur un voisinage de t , $nb_n \rightarrow \infty$ et K est borné d'où

$$\mathbf{1}_{\{|L_n(s)| > \varepsilon\}} \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

uniformément sur un voisinage de t . Par la convergence dominée

$$\int_0^1 K^2 \left(\frac{t-s}{b_n} \right) \frac{J_n(s)}{\bar{Y}_n(s)} \mathbf{1}_{\{|L_n(s)| > \varepsilon\}} h(s) ds \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

D'où la première condition.

Pour la deuxième condition on utilise (i) et la convergence dominée pour avoir asymptotiquement

$$\begin{aligned}
\int_0^1 L_n^2(s) \lambda_n(s) ds &= \int_0^1 \frac{1}{b_n} K^2 \left(\frac{t-s}{b_n} \right) \frac{nJ_n(s)}{\bar{Y}_n^2(s)} \bar{Y}_n(s) h(s) ds \\
&= \int_{-1}^1 K^2(u) \frac{nJ_n(t-b_n u)}{\bar{Y}_n(t-b_n u)} h(t-b_n u) du \\
&\xrightarrow{\mathbf{P}}_{n \rightarrow \infty} \frac{h(t)}{\tau(t)} \int_{-1}^1 K^2(u) du
\end{aligned}$$

Finalement les deux conditions sont vérifiées et on a le résultat. ■

3.2 Données censurées

3.2.1 Introduction

Dans cette partie nous allons donner les formes exactes et asymptotiques de la moyenne et la variance de l'estimateur à noyau de la fonction de risque h dans le cas de données censurées, en développant les résultats de l'article de Martin Tanner et Wing Hung Wong (1983) [16].

Soient les durées de survie T_1, T_2, \dots, T_n de f.d.r. F et de densité f et C_1, C_2, \dots, C_n les v.a. de censures de f.d.r. F_C . On suppose que les T_i sont indépendantes des C_i pour tout $i = 1, \dots, n$ et on observe $X_i = T_i \wedge C_i$ et $D_i = \mathbf{1}_{\{T_i \leq C_i\}}$. On note par f_X la densité des X_i et par F_X leur f.d.r.

Rappelons que l'estimateur à noyaux de la fonction de risque h des T_n est donnée par

$$\hat{h}_n(t) = \sum_{j=1}^n \frac{D_{(j)}}{n-j+1} K_{b_n}(t - X_{(j)}) = \sum_{i=1}^n \frac{D_{(i)}}{n-R_i+1} K_{b_n}(t - X_i)$$

où $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ est la statistique d'ordre et R_i est le rang de X_i . Cet estimateur est apparu dans les articles de Ramlau-Hansen (1983) [9] et Watson et Leadbetter (1964) (cf. [20]).

Nous supposons que :

1. la fonction de risque h est continue et que $0 < F(y) < 1$.
2. le noyau K est symétrique positif et $K(t) = o(t^{-1})$ quand $t \rightarrow \infty$ et $\int k(t)dt = 1$.

3.2.2 Moyenne et variance de l'estimateur \hat{h}_n

On pose $m(y) = f(y)[1 - F_C(y)]/f_X(y)$ avec $f_X(y) > 0$. Pour le calcul de la moyenne et la variance de $\hat{h}_n(t)$, on a besoin du lemme suivant

Lemme 3.1 *Pour tout j on a*

$$\mathbf{E}(D_{(j)} / X_{(j)} = y) = m(y)$$

et pour tout $r < s$, $y < z$, on a

$$\mathbf{E}(D_{(r)}D_{(s)} / X_{(r)} = y, X_{(s)} = z) = m(y)m(z)$$

Preuve: Soit $y > 0$ et j fixé, comme $T_{(j)}$ sont indépendantes des $C_{(j)}$ on a ($D_i = \mathbf{1}_{\{T_i \leq C_i\}}$) :

$$\begin{aligned}
\mathbf{E}(D_{(j)} / X_{(j)} = y) &= \mathbf{P}(D_{(j)} = 1 / X_{(j)} = y) = \mathbf{P}(T_{(j)} \leq C_{(j)} / X_{(j)} = y) \\
&= \frac{\mathbf{P}(T_{(j)} \leq C_{(j)}, X_{(j)} = y)}{\mathbf{P}(X_{(j)} = y)} = \frac{\mathbf{P}(T_{(j)} \leq C_{(j)}, T_{(j)} \wedge C_{(j)} = y)}{\mathbf{P}(X_{(j)} = y)} \\
&= \frac{\mathbf{P}(C_{(j)} \geq y, T_{(j)} = y)}{\mathbf{P}(X_{(j)} = y)} = \frac{\mathbf{P}(C_{(j)} \geq y) \mathbf{P}(T_{(j)} = y)}{\mathbf{P}(X_{(j)} = y)} \\
&= \frac{(1 - F_C(y))f(y)dy}{f_X(y)dy} = \frac{(1 - F_C(y))f(y)}{f_X(y)} = m(y)
\end{aligned}$$

Soit $r < s$, $y < z$, comme T_i sont indépendantes des C_i pour $i = r, s$ on a

$$\begin{aligned}
\mathbf{E}(D_{(r)}D_{(s)} / X_{(r)} = y, X_{(s)} = z) &= \mathbf{P}(D_{(r)} = 1, D_{(s)} = 1 / X_{(r)} = y, X_{(s)} = z) \\
&= \mathbf{P}(T_{(r)} \leq C_{(r)}, T_{(s)} \leq C_{(s)} / X_{(r)} = y, X_{(s)} = z) \\
&= \frac{\mathbf{P}(T_{(r)} \leq C_{(r)}, T_{(s)} \leq C_{(s)}, X_{(r)} = y, X_{(s)} = z)}{\mathbf{P}(X_{(r)} = y, X_{(s)} = z)} \\
&= \frac{\mathbf{P}(y \leq C_{(r)}, z \leq C_{(s)}, T_{(r)} = y, T_{(s)} = z)}{\mathbf{P}(X_{(r)} = y, X_{(s)} = z)} \\
&= \frac{\mathbf{P}(y \leq C_{(r)}, T_{(r)} = y) \mathbf{P}(z \leq C_{(s)}, T_{(s)} = z)}{\mathbf{P}(X_{(r)} = y, X_{(s)} = z)} \\
&= \frac{\mathbf{P}(y \leq C_{(r)})\mathbf{P}(T_{(r)} = y)}{\mathbf{P}(X_{(r)} = y)} \frac{\mathbf{P}(z \leq C_{(s)})\mathbf{P}(T_{(s)} = z)}{\mathbf{P}(X_{(s)} = z)} \\
&= m(y)m(z)
\end{aligned}$$

D'où les résultats du lemme. ■

Le théorème suivant donne la moyenne et la variance de $\hat{h}_n(t)$

Théorème 3.3 On a

$$\mathbf{E}(\hat{h}_n(t)) = \int (1 - F_X^n(y)) h(y) K_{b_n}(t - y) dy$$

et

$$\begin{aligned} \mathbf{V}(\hat{h}_n(t)) &= \int I_n(F_X(y)) h(y) K_{b_n}^2(t-y) dy + 2 \int \int_{y \leq z} \{F_X^n(z) - F_X^n(y) F_X^n(z) \\ &\quad - \frac{1 - F_X(y)}{F_X(z) - F_X(y)} [F_X^n(z) - F_X^n(y)]\} h(y) h(z) K_{b_n}(t-y) K_{b_n}(t-z) dy dz \end{aligned}$$

où $I_n(F_X) = \sum_{k=0}^n (n-k)^{-1} C_n^k F_X^k (1-F)^{n-k}$

Preuve: Nous savons que la densité de $X_{(j)}$ est:

$$f_{X_{(j)}}(y) = \frac{n!}{(j-1)!(n-j)!} F_X^{j-1}(y) (1-F_X(y))^{n-j} f_X(y)$$

On a

$$\begin{aligned} \mathbf{E}(\hat{h}_n(t)) &= \int \mathbf{E}(\hat{h}_n(t)/X_{(j)} = y) f_{X_{(j)}}(y) dy \\ &= \int \mathbf{E}\left(\sum_{i=1}^n \frac{D_{(i)}}{n-i+1} K_{b_n}(t-X_{(i)}) / X_{(j)} = y\right) f_{X_{(j)}}(y) dy \\ &= \sum_{i=1}^n \frac{1}{n-i+1} \int \mathbf{E}(D_{(i)} K_{b_n}(t-X_{(i)}) / X_{(j)} = y) f_{X_{(j)}}(y) dy \\ &= \sum_{i=1, i \neq j}^n \frac{1}{n-i+1} \int \mathbf{E}(D_{(i)} K_{b_n}(t-X_{(i)}) / X_{(j)} = y) f_{X_{(j)}}(y) dy \\ &\quad + \frac{1}{n-j+1} \int \mathbf{E}(D_{(j)} K_{b_n}(t-X_{(j)}) / X_{(j)} = y) f_{X_{(j)}}(y) dy \\ &= \sum_{i=1, i \neq j}^n \frac{1}{n-i+1} \int \mathbf{E}(D_{(i)} K_{b_n}(t-X_{(i)}) / X_{(j)} = y) f_{X_{(j)}}(y) dy \\ &\quad + \frac{1}{n-j+1} \int m(y) K_{b_n}(t-y) f_{X_{(j)}}(y) dy \quad (\text{lemme 3.1}) \\ &= \sum_{i=1}^n \frac{1}{n-i+1} \int m(y) K_{b_n}(t-y) f_{X_{(i)}}(y) dy \\ &= \sum_{i=1}^n \frac{1}{n-i+1} \int m(y) \frac{n!}{(i-1)!(n-i)!} F_X^{i-1}(y) (1-F_X(y))^{n-i} f_X(y) K_{b_n}(t-y) dy \end{aligned}$$

$$\begin{aligned}
\mathbf{E} \left(\hat{h}_n(t) \right) &= \int \left[\sum_{i=1}^n \frac{1}{(n-i+1)} \frac{n!}{(i-1)!(n-i)!} F_X^{i-1}(y) (1-F_X(y))^{n-i+1} \right] \frac{f_X(y)m(y)}{(1-F_X(y))} K_{b_n}(t-y) dy \\
&= \int \left[\sum_{i=1}^n C_n^{i-1} F_X^{i-1}(y) (1-F_X(y))^{n-i+1} \right] \frac{f(y)(1-F_C(y))}{(1-F_X(y))} K_{b_n}(t-y) dy \\
&= \int \left[\sum_{j=0}^{n-1} C_n^j F_X^j(y) (1-F_X(y))^{n-j} \right] \frac{f(y)(1-F_C(y))}{(1-F_X(y))} K_{b_n}(t-y) dy \\
&= \int (1-F_X^n(y)) h(y) K_{b_n}(t-y) dy
\end{aligned}$$

où dans le passage à la 6^{ème} égalité on a utilisé pour $i \neq j$

$$\begin{aligned}
\int \mathbf{E} (D_{(i)} K_{b_n}(t - X_{(i)}) / X_{(j)} = y) f_{X_{(j)}}(y) dy &= \mathbf{E} (D_{(i)} K_{b_n}(t - X_{(i)})) \\
&= \int \mathbf{E} (D_{(i)} K_{b_n}(t - X_{(i)}) / X_{(i)} = y) f_{X_{(i)}}(y) dy \\
&= \int m(y) K_{b_n}(t - y) f_{X_{(i)}}(y) dy
\end{aligned}$$

Le calcul de la variance $\mathbf{V} \left(\hat{h}_n(t) \right)$ est plus long mais se fait de façon similaire. ■

Remarque 3.2 Dans le th 3.3, la partie dominante de $\mathbf{E} \left(\hat{h}_n(t) \right)$ se comporte quand $n \rightarrow \infty$ comme la convolution $(h * K_{b_n})(t)$. En effet

$$\begin{aligned}
\mathbf{E} \left(\hat{h}_n(t) \right) &= \int (1 - F_X^n(y)) h(y) K_{b_n}(t - y) dy \\
&= \int h(y) K_{b_n}(t - y) dy - \int F_X^n(y) h(y) K_{b_n}(t - y) dy \\
&= (h * K_{b_n})(t) - \int F_X^n(y) h(y) K_{b_n}(t - y) dy \\
&\approx (h * K_{b_n})(t)
\end{aligned}$$

et le 2^{ème} terme est négligeable car $F_X^n(y) = 0$ pour n assez grand ($0 < F(y) < 1$). Par le lemme de Bochner $h(t)$ peut être approximée par

$$h(t) \approx (h * K_{b_n})(t) = \int h(y) K_{b_n}(t - y) dy$$

et qui est une moyenne pondérée de h . et donc $\mathbf{E} \left(\hat{h}_n(t) \right) \simeq h(t)$. Pour une bonne approximation

de $h(t)$ (au point t) il est nécessaire que le poids K_{b_n} donne des valeurs assez faibles pour des y "loin" de t . Comme

$$h(t) = f(t)(1 - F(t))^{-1}$$

cela revient à donner des conditions sur le comportement des queues de K_{b_n} et de $1 - F$.

D'où la définition suivante:

Définition 3.1 *Le noyau K est dit compatible avec F si $\forall M > 0, \exists G_M$ et (b_n) un paramètre assez petit tel que*

$$\sup_{|y-x|>M} \left| \frac{K_{b_n}(y-x)}{(1-F(y))} \right| \leq G_M$$

On a le résultat suivant:

Théorème 3.4 *Soit $b_n \rightarrow_{n \rightarrow \infty} 0$ et $nb_n \rightarrow_{n \rightarrow \infty} \infty$. Nous avons*

1. *Si K est compatible avec F , alors $\mathbf{E}(\hat{h}_n(t)) \rightarrow_{n \rightarrow \infty} h(t)$*
2. *Si K est compatible avec F et F_C , alors*

$$\mathbf{V}(\hat{h}_n(t)) = \frac{1}{nb_n} \left(\int K^2(u) du \right) h(t)(1 - F_X(t))^{-1} + o\left(\frac{1}{nb_n}\right)$$

Preuve: Pour la partie 1, nous avons par le th. 3.3:

$$\mathbf{E}(\hat{h}_n(t)) - h(t) = \int [(1 - F_X^n(y)) h(y) - h(t)] K_{b_n}(t-y) dy$$

En utilisant que K est compatible avec F , on obtient pour $M > 0$

$$\begin{aligned}
\left| \mathbf{E} \left(\hat{h}_n(t) \right) - h(t) \right| &\leq \int |[(1 - F_X^n(y)) h(y) - h(t)] K_{b_n}(t - y)| dy \\
&\leq \int_{|y-t|>M} |[(1 - F_X^n(y)) h(y) - h(t)] K_{b_n}(t - y)| dy + \\
&\quad \int_{|y-t|\leq M} |[(1 - F_X^n(y)) h(y) - h(t)] K_{b_n}(t - y)| dy \\
&\leq \int_{|y-t|>M} \left| [(1 - F_X^n(y)) h(y) - h(t)] (1 - F_X(y)) \frac{K_{b_n}(t - y)}{(1 - F_X(y))} \right| dy + \\
&\quad \int_{|y-t|\leq M} |[(1 - F_X^n(y)) h(y) - h(t)] K_{b_n}(t - y)| dy \\
&\leq G_M \int_{|y-t|>M} |[(1 - F_X^n(y)) h(y) - h(t)] (1 - F_X(y))| dy \\
&\quad + \int_{|y-t|\leq M} |h(y) - h(t)| K_{b_n}(t - y) dy + \int_{|y-t|\leq M} F_X^n(y) h(y) K_{b_n}(t - y) dy \\
&\leq G_M \int_{|y-t|>M} |[(1 - F_X^n(y)) h(y) - h(t)]| dy \\
&\quad + \int_{|y-t|\leq M} |h(y) - h(t)| K_{b_n}(t - y) dy + \int_{|y-t|\leq M} F_X^n(y) h(y) K_{b_n}(t - y) dy \\
&= (1) + (2) + (3)
\end{aligned}$$

Chacun des trois termes tend vers 0. En effet:

◇ Pour le terme (1), en choisissant M assez grand on a par le th. de convergence dominée

$$\int_{|y-t|>M} |[(1 - F_X^n(y)) h(y) - h(t)]| dy \leq \varepsilon$$

◇ Pour le terme (2), en posant $z = \frac{t-y}{b_n}$ et en appliquant le th. de C. D. on a:

$$\int_{|y-t|\leq M} |h(y) - h(t)| K_{b_n}(t - y) dy = \int_{|z|\leq \frac{M}{b_n}} |h(t - zb_n) - h(t)| K(z) dz \xrightarrow{n \rightarrow \infty} 0$$

◇ Pour le terme (3) en posant $z = \frac{t-y}{b_n}$ et en appliquant le th. de C. D. on a:

$$\int_{|y-t|\leq M} F_X^n(y) h(y) K_{b_n}(t - y) dy = \int_{|z|\leq \frac{M}{b_n}} F_X^n(t - zb_n) h(t - zb_n) K(z) dz \xrightarrow{n \rightarrow \infty} 0$$

Pour la deuxième partie du théorème, le lemme 6 de W.L (cf.[20]) montre que pour $F(y) < 1$

$$nI_n(F) \rightarrow_{n \rightarrow \infty} (1 - F(y))^{-1}$$

◇ Pour obtenir le premier terme de $\mathbf{V}(\hat{h}_n(t))$ on applique le lemme de Bochner pour le noyau $K^2 / (\int K^2(t)dt)$ et on a

$$nb_n \left(\int K^2(t)dt \right)^{-1} \int I_n(F_X(y))h(y)K_{b_n}^2(t-y)dy \rightarrow_{n \rightarrow \infty} h(t)(1 - F(x))^{-1}$$

C'est aussi le terme dominant dans l'expression $nb_n (\int K^2(u)du)^{-1} \mathbf{V}(\hat{h}_T(t))$.

◇ Le deuxième terme converge vers 0 à cause de $F_X^n \rightarrow_{n \rightarrow \infty} 0$ par les mêmes arguments que dans W.L (cf.[20]).

D'où le résultat du théorème. ■

Remarque 3.3 1. Comme conséquence immédiate du théorème on a $\mathbf{V}(\hat{h}_n(t)) \rightarrow_{n \rightarrow \infty} 0$ donc $\hat{h}_n(t)$ est un estimateur convergent en probabilité pour $h(t)$.

2. Pour la loi limite de l'estimateur $\hat{h}_n(t)$, sous les conditions du th. 3.4. on a : $(\hat{h}_n(t) - E(\hat{h}_n(t))) \mathbf{V}^{1/2}(\hat{h}_n(t)) \rightarrow N(0,1)$ (cf. [16] Th. 3)

Chapitre 4

Simulations

4.1 Introduction.

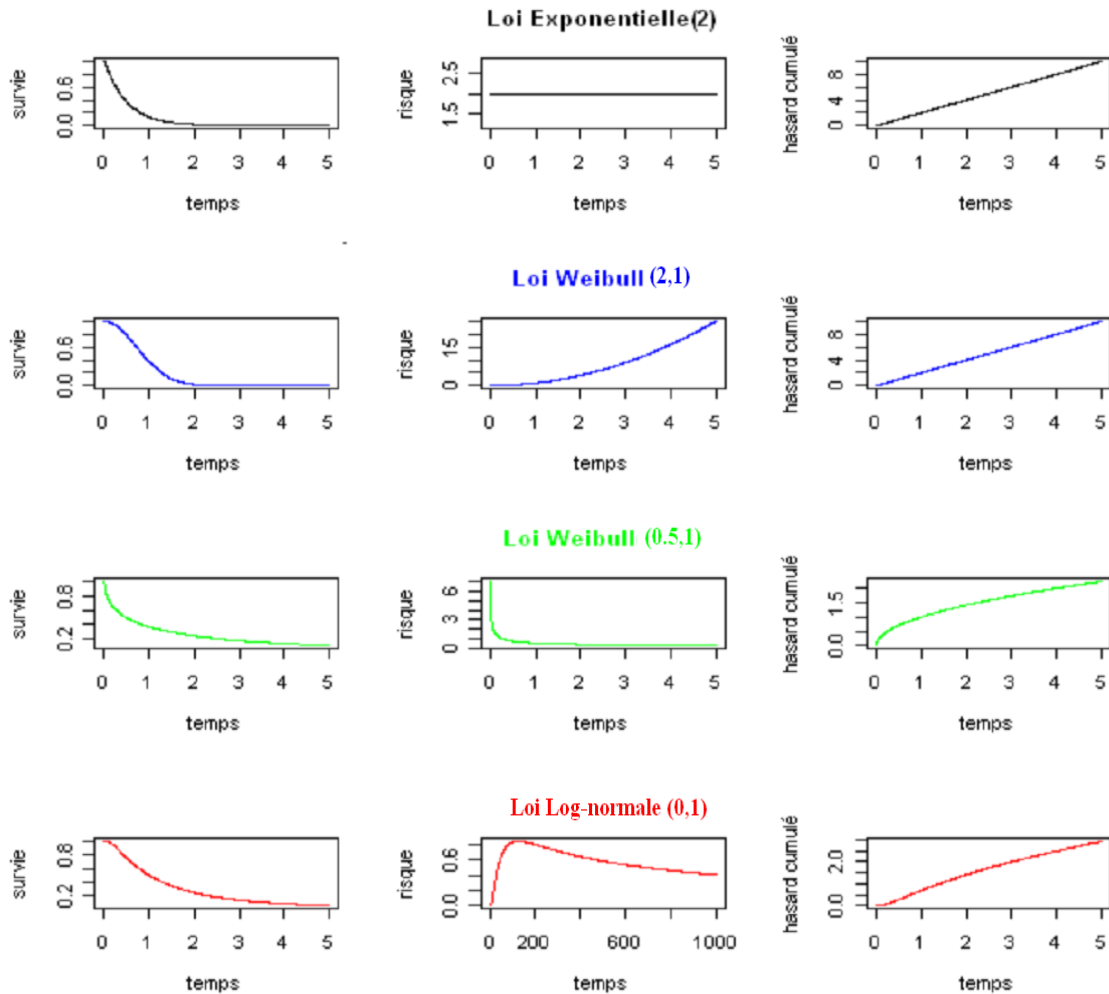
Dans cette partie nous illustrons la performance des estimateurs étudiés dans ce mémoire par des simulations numériques sur des données simulées et sur des données réelles (données de ventilateurs (chap. 1) et données sur des patients atteints de la maladie cirrrose de vésicule biliaire ([7] Appendice D p. 359)). Nous utilisons les logiciels R et Minitab. Nous donnons quelques exemples de lois utilisées en analyse de survie et leurs fonctions associées.

Nous présentons les graphes des estimateurs de Kaplan-Meier, Fleming-Harington et de Nelson Aalen pour des données simulées et réelles. Nous utilisons un test d'ajustement (test d'alignement) sur les données réelles pour choisir dans un ensemble de quatre lois usuelles utilisées en analyse de survie.

Nous donnons aussi les graphes des estimateurs à noyau de la fonction de densité de la durée de survie et la fonction de risque.

4.2 Lois usuelles en survie

Nous présentons les graphes de quelques lois usuelles en analyse de survie avec leurs fonctions associées: fonction de survie, fonction de risque et taux de risque cumulée (cf chap. 1)



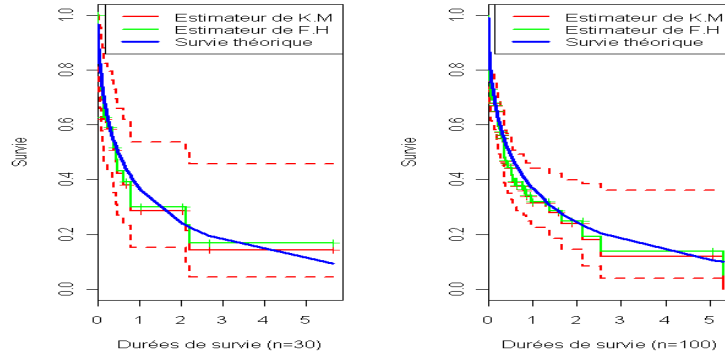
4.3 Données simulées

4.3.1 Estimation de la fonction de survie S

Nous avons simulé un échantillon de taille 800 données qui sont censurées à droite: (X_i, D_i) avec $X_i = T_i \wedge C_i$ où $T_i \hookrightarrow W(0.5, 1)$ et $C_i \hookrightarrow \mathcal{E}(1/2)$, $n = 30, 100$ obs. La figure suivante donne

l'estimateur de Kaplan-Meier \hat{S}_{KM} , l'estimateur de Fleming-Harrington \hat{S}_{FH} de la fonction de survie S avec des intervalles de confiance à 95% et le graphe de la fonction de survie $S = S_T$ théorique. Rappelons que \hat{S}_{KM} et \hat{S}_{FH} sont défini par:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{1}{n - i + 1}\right)^{D(i)}, \quad \hat{S}_{FH}(t) = \exp\left(- \sum_{\{i, t_i \leq t\}} \frac{m(t_i)}{r(t_i)}\right)$$

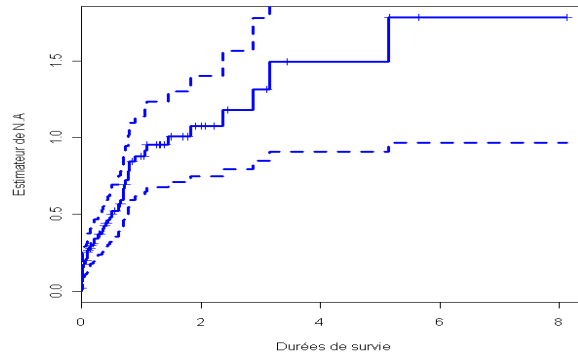


Les courbes des estimateurs \hat{S}_{KM} et \hat{S}_{FH} sont très "proches" et estime correctement la survie S puisque n est très grand.

4.3.2 Estimation du taux de hasard cumulé H

Pour un échantillon de 100 données censurées précédentes, nous traçons le graphe suivant de l'estimateur de Nelson-Aalen \hat{H}_{NA} du taux de hasard cumulé H avec des intervalles de confiance à 95%. Rappelons que:

$$\hat{H}_{NA}(t) = \sum_{\{i, t_i \leq t\}} \frac{m(t_i)}{r(t_i)}$$



Interprétation : pour t fixé, $\hat{H}_{NA}(t)$ représente la somme des risques de subir l'évènement dans l'intervalle $]0, t]$.

4.4 Données réelles

Dans cette partie nous ferons de l'analyse de survie sur des données réelles.

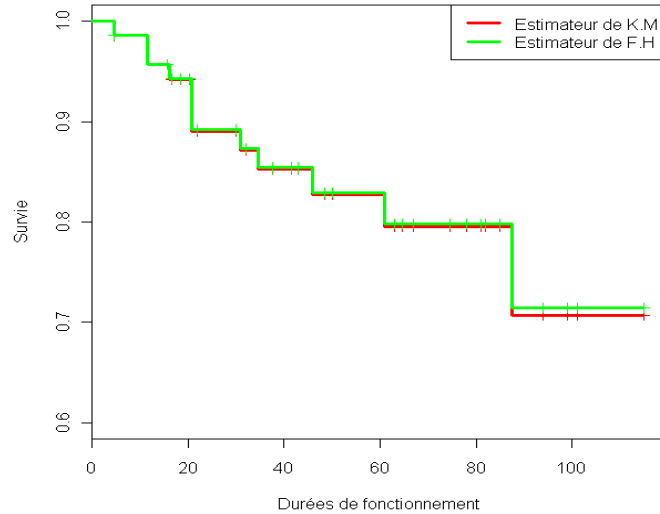
4.4.1 Données des ventilateurs.

Le tableau suivant donne la durées de fonctionnement de 70 ventilateurs (Les nombres avec + correspondent à des données censurées) (cf. [3])

4.5	4.6 ⁺	11.5	11.5	15.6 ⁺	16.0	16.6 ⁺	18.5 ⁺	18.5 ⁺	18.5 ⁺
18.5 ⁺	18.5 ⁺	20.3 ⁺	20.3 ⁺	20.3 ⁺	20.7	20.7	20.8	22.0 ⁺	30.0 ⁺
30.0 ⁺	30.0 ⁺	30.0 ⁺	31.0	32.0 ⁺	34.5	37.5 ⁺	37.5 ⁺	41.5 ⁺	41.5 ⁺
41.5 ⁺	41.5 ⁺	43.0 ⁺	43.0 ⁺	43.0 ⁺	43.0 ⁺	46.0	48.5 ⁺	48.5 ⁺	48.5 ⁺
48.5 ⁺	50.0 ⁺	50.0 ⁺	50.0 ⁺	61.0 ⁺	61.0	61.0 ⁺	61.0 ⁺	63.0 ⁺	64.5 ⁺
64.5 ⁺	67.0 ⁺	74.5 ⁺	78.0 ⁺	78.0 ⁺	81.0 ⁺	81.0 ⁺	82.0 ⁺	85.0 ⁺	85.0 ⁺
85.0 ⁺	87.5 ⁺	87.5	87.5 ⁺	94.0 ⁺	99.0 ⁺	101.0 ⁺	101.0 ⁺	101.0 ⁺	115.0 ⁺

Estimation de la fonction de survie S

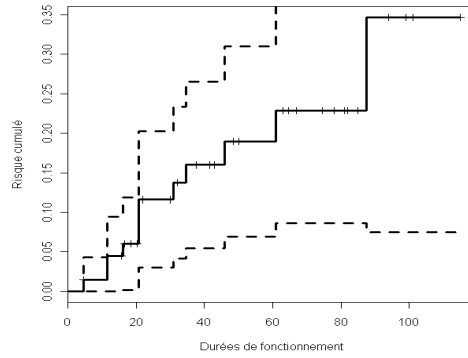
le graphe suivant donne l'estimateur de Kaplan-Meier \hat{S}_{KM} et l'estimateur de Fleming-Harrington \hat{S}_{FH} de la fonction de survie S des données de ventilateurs



Par exemple $\hat{S}_{KM}(60 \times 10^3) = 0.8$ ce qui donne 80% des ventilateurs vont fonctionner au delà de 60×10^3 heures. Les courbes des estimateurs \hat{S}_{KM} et $\hat{S}_{F.H}$ sont là aussi très "proches" ($n = 70$).

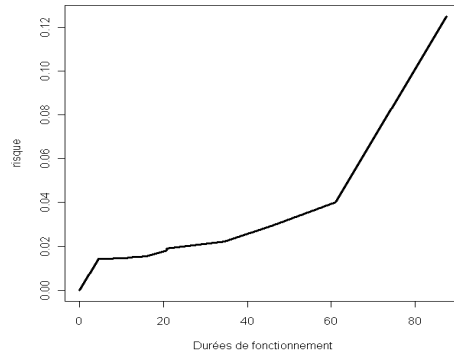
Estimation du taux de hasard cumulé H

le graphe suivant donne l'estimateur de Nelson-Aalen \hat{H}_{NA} du taux de hasard cumulé H des données des ventilateurs et des intervalles de confiance à 95%.



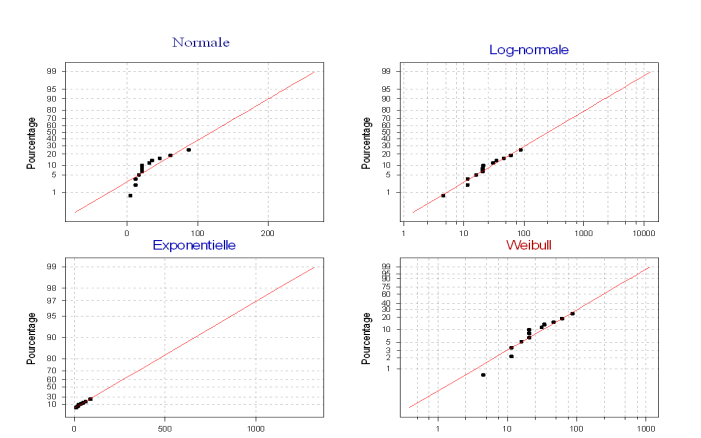
Par exemple $\hat{H}_{NA}(60) = 0.18 = 18\%$, on a 18% de risque que le lot des ventilateurs tombe en panne pendant une durée de fonctionnement de $60 \cdot 10^3$ heures.

Le graphe suivant donne l'estimateur $\hat{h}_n(t_i) = \frac{m(t_i)}{r(t_i)}$ de la fonction risque h des données des ventilateurs:

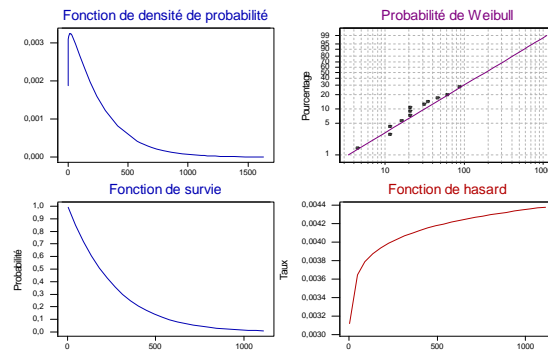


Le graphe de h étant croissant à partir de $t = 20 \times 10^3$ heures nous pouvons conclure que le risque de tomber en panne augmente sensiblement ce qui est normal avec l'usure du matériel.

Les graphes suivants permettent de tester l'ajustement des données par rapport à quelques lois usuelles (test d'alignement) (loi normale, loi exponentielle, loi log-normale, loi de Weibull):

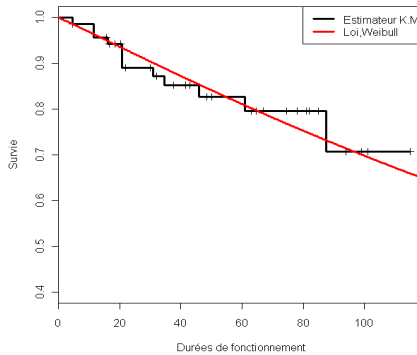


Nous remarquons que les données des ventilateurs s'ajustent correctement aux lois exponentielle, Weibull et Log normale. Cependant au vu du graphe de l'estimateur de la fonction de risque des données des ventilateurs (risque croissant) nous retenons la loi de Weibull estimée ($W(1.058, 262.968)$), (le paramètre de forme ici est $\alpha = 1.058$). La loi exponentielle et la loi Log-normale sont exclues à cause de la forme de leur fonction de risque. Les graphiques des fonctions: f , S et h de la loi retenue sont les suivants :



Le graphe suivant donne l'estimateur de Kaplan-Meier et la courbe de la fonction S de la loi $W(1.058, 262.968)$. Nous remarquons un bon ajustement des deux courbes ce qui confirme

le choix précédent de la loi de Weibull).



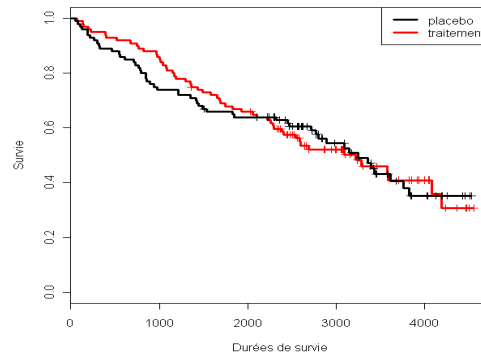
Dans ce cas nous disposons de peu de données non censurées (12 obser.), ce qui donne une estimation peu précise de h en comparaisant avec le graphe de \hat{h}_n précédent.

4.4.2 Données PBC ([7] p. 359) (malades atteints de la cirrhose)

Nous prenons les données présentées en Appendice D de [7] p. 359 de malades atteints de la cirrhose. les patients sont regrouper en deux groupes chacun de 100 malades l'un reçoit la D-penicillamine et l'autre un placebo. Le but étant de comparer la survie dans les deux groupes et de là évaluer l'efficacité biologique du traitement testé par rapport à son "effet placebo".

Estimation de la fonction de survie S

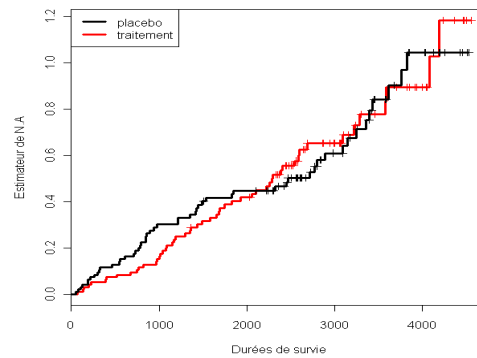
Le graphe suivant donne l'estimateur de Kaplan-Meier \hat{S}_{KM} de la fonction de survie S de deux groupes.



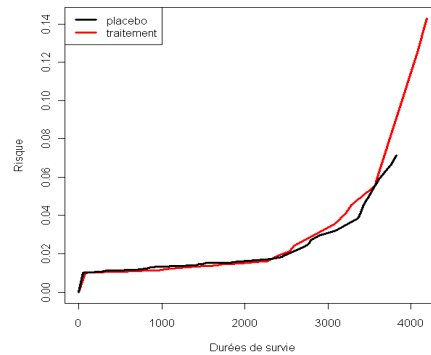
Nous remarquons que les deux courbes de Kaplan-Meier représentant les deux groupes sont "proches".

Estimation du taux de hasard cumulé H

Le graphe suivant donne l'estimateur de Nelson-Aalen \hat{H}_{NA} du taux de hasard cumulé H des deux groupes des données PBC.



Le graphe suivant donne l'estimateur $\hat{h}_n(t_i) = \frac{m(t_i)}{r(t_i)}$ de la fonction risque h des deux groupes de données PCB.

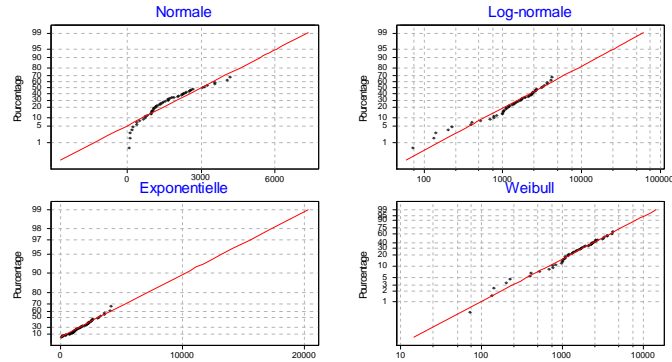


L'estimateur \hat{h}_n de h des deux groupes est une courbe croissante ce qui s'interprète par un risque croissant de décès par cette maladie. Les deux graphes de \hat{h}_n sont "proches" avec un écart quadratique moyen: $e = 0.0005$.

A partir des graphes précédents, nous pouvons conclure que le traitement et l'effet placebo ont la même efficacité. D'après le test "Log-Rank" (cf [7] p. 8) on retient l'égalité des deux

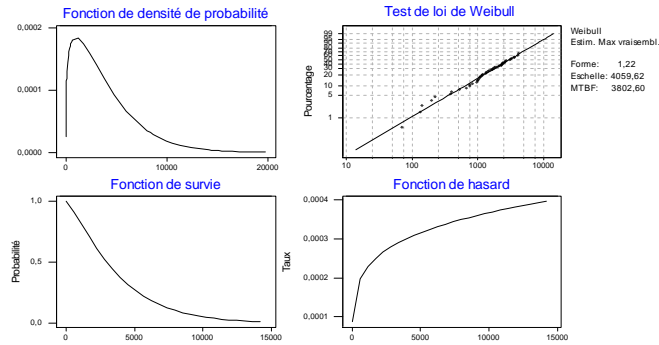
courbes de survie et par suite on aboutit à la même efficacité.

Test sur la loi des données PBC (groupe traité)

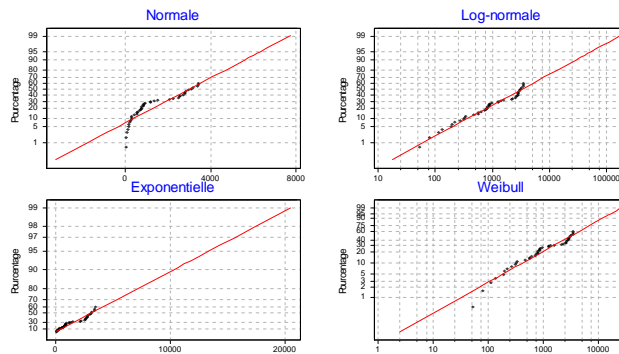


Vu la forme du risque naif on opte pour la loi de weibull

Les caractéristiques de la loi de Weibull

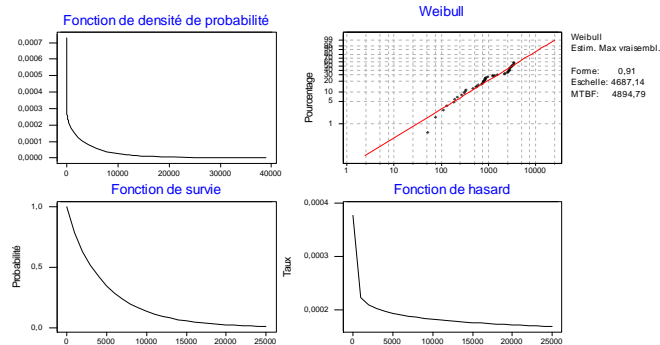


Test sur la loi des données PBC (groupe non traité)

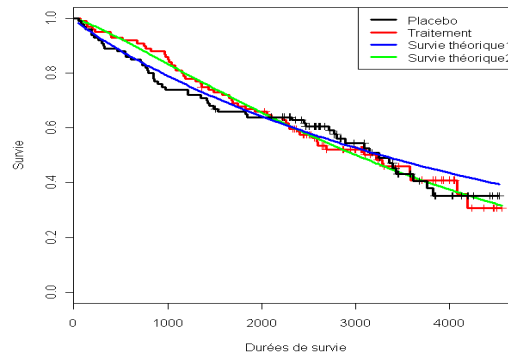


On choisie la loi Weibull

Les caractéristiques de la loi de Weibull



Graphes de KM des deux groupes la survie de $W(1.22, 4059.62)$ (théorique1) et la survie de $W(0.91, 4687.14)$ (théorique2)



4.5 Estimation des fonctions de densité et de risque par la méthode du noyau

Dans cette partie nous utilisons l'estimateur à noyau (Gaussien) pour estimer la densité de durée de survie T et la fonction de risque associée avec une fenêtre optimale $b_n^* = C.n^{-1/5}$ où C est une constante.

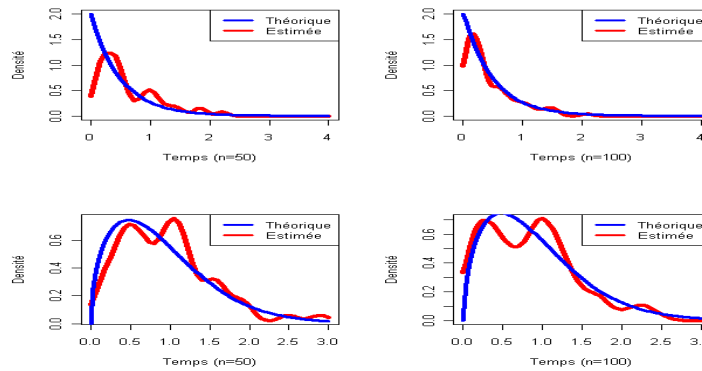
Nous choisissons deux lois pour T : la loi exponentielle $\mathcal{E}(2)$ et la loi de weibull $W(1; 1.5)$ avec une taille d'échantillon $n = 50, 100$

4.5.1 Estimation de la densité des durées T

Cas non censurées

a) $T \hookrightarrow \mathcal{E}(2)$, $n = 50, 100$, K Gaussien, $b_n = 0.09$

b) $T \hookrightarrow W(1.5, 1)$, $n = 50, 100$, K Gaussien, $b_n = 0.13$



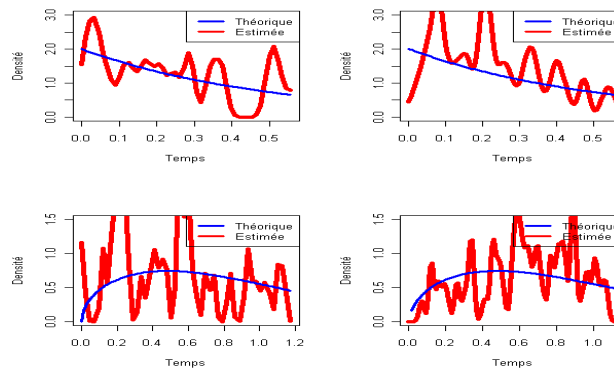
$$e_1 = 0.07, e_2 = 0.018, e_3 = 0.007, e_4 = 0.009$$

Nous remarquons une bonne estimation de la densité de T (sauf en début d'intervalle)

Cas censurées

a) $T, C \hookrightarrow \mathcal{E}(2)$, $n = 50, 100$, $b_n = 0.014$ et K noyau Gaussien

b) $T, C \hookrightarrow W(1.5, 1)$, $n = 50, 100$, $b_n = 0.013$, K noyau Gaussien



Nous remarquons une oscillation de l'estimateur \hat{f}_n autour de la fonction de densité théorique

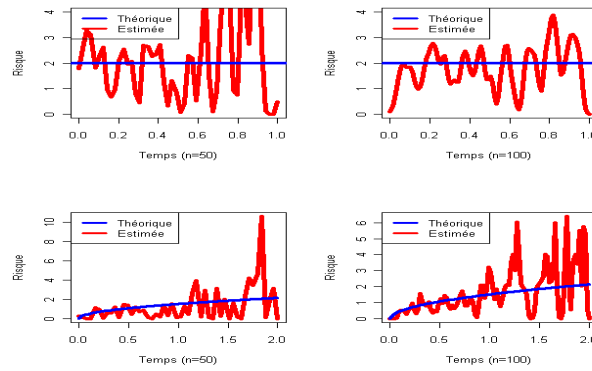
avec des erreurs quadratique moyenne: $e_1 = 0.72, e_2 = 0.61, e_3 = 0.9, e_4 = 0.53$

4.5.2 Estimation de la fonction de risque

Cas non censurées

a) $T \hookrightarrow \mathcal{E}(2), n = 50, 100, b_n = 0.014$

b) $T \hookrightarrow W(1.5, 1), n = 50, 100, b_n = 0.014$

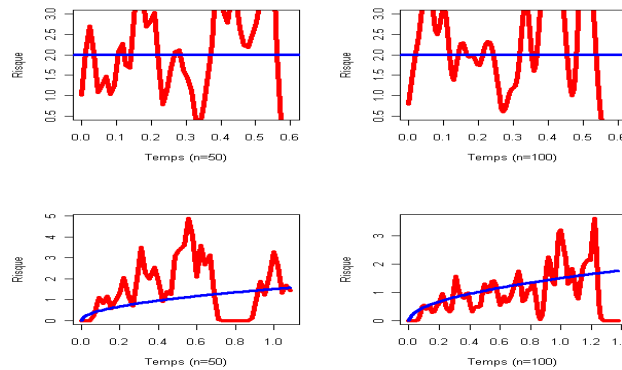


Nous remarquons une oscillation de l'estimateur \hat{h}_n autour de la fonction de risque théorique avec des erreurs quadratique moyenne: $e_1 = 0.9, e_2 = 0.82, e_3 = 0.95, e_4 = 0.88$

Cas censurées:

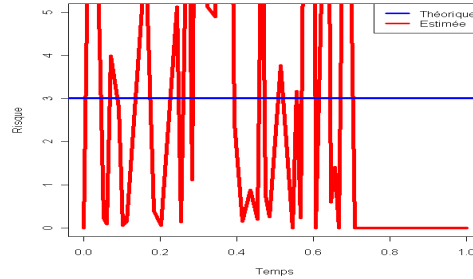
a) $T, C \hookrightarrow \mathcal{E}(2), n = 50, 100, b_n = 0.012$

b) $T, C \hookrightarrow W(1, 1.5), n = 50, 100, b_n = 0.014$



Nous remarquons une oscillation de l'estimateur \hat{h}_n autour de la fonction de risque théorique avec des erreurs quadratique moyenne: $e_1 = 0.99, e_2 = 0.94, e_3 = 0.82, e_4 = 0.97$

c) $T \hookrightarrow \mathcal{E}(3), C \hookrightarrow W(0.5, 1), n = 100, K$ noyau Epanchnicov $b_n = 0.003$



4.6 Conclusion.

Dans les études de simulations présentées dans cette partie, nous remarquons que les estimateurs étudiés présentent des comportements très satisfaisants (convergence, calcul d'erreurs..) pour le cas des données simulées et pour les données réelles.

Conclusion

Dans ce mémoire nous avons étudié l'analyse des données de survie et nous sommes intéressé au cas de censure aléatoire à droite avec l'hypothèse d'indépendance des durées de survie et les variables de censure.

Dans le chapitre un, nous avons étudié différentes méthodes d'estimation de la fonction de survie et le taux de risque cumulé en particulier les estimateurs de Fleming-Harrington , Kaplan-Meier et Nelson Aalen.

Dans le chapitre deux, en première partie nous avons étudié un estimateur à noyau de la densité des durées de survie dans le cas de données non censurées (cf [4]). Dans le cas de données censurées, la théorie des martingales et des processus ponctuels joue un rôle très important dans l'étude des estimateurs tels que Kaplan- Meier et Nelson Aalen et les estimateurs à noyau (cf. [21]).

Dans le chapitre trois, nous avons étudié un estimateur à noyau de la fonction de risque en donnant le comportement de sa moyenne et sa variance dans les deux cas: non censurées et censurées (cf. [9] et [16]).

Nous terminons par des simulations numériques pour illustrer les comportements des différents estimateurs étudiés dans ce mémoire.

Annexe

Dans cette partie nous rappelons des résultats que nous avons utilisés dans ce mémoire (cf [7] chap. 1 et 2).

Théorème 4.1 (Décomposition de Doob-Meyer) *Soit X une sous martingale positive continue à droite adaptée à la filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$. Il existe une \mathcal{F} -martingale continue à droite M et un processus prédictible croissant continu à droite A tel que $\mathbf{E}(A(t)) < \infty$ et*

$$X(t) = M(t) + A(t)$$

pour tout $t \geq 0$. Si $A(0) = 0$ presque sûrement et $X = M' + A'$ est une autre décomposition avec $A'(0) = 0$, alors pour tout $t \geq 0$

$$\mathbf{P}(M(t) \neq M'(t)) = 0 = \mathbf{P}(A(t) \neq A'(t))$$

Corollaire 4.1 *Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$, avec $\mathbf{E}(N(t)) < \infty$ pour tout t . Alors il existe un unique processus continu à droite A tel que $A(0) = 0$ presque sûrement, $\mathbf{E}(A(t)) < \infty$ et $M = N - A$ est une \mathcal{F} -martingale continue à droite.*

Théorème 4.2 (Décomposition de Doob-Meyer) *Soit $X = (X_t)_{t \geq 0}$ une sous martingale locale positive continue à droite adaptée à la filtration $\{\mathcal{F}_t, t \geq 0\}$ et soit $(\tau_n)_{n \geq 0}$ la suite de temps d'arrêts associée. Il existe un unique processus prédictible croissant continu à droite A tel que $A(0) = 0$ presque sûrement, $\mathbf{P}(A(t) < \infty) = 1$ pour tout t et $X - A$ est une martingale locale continue à droite. Pour tout t , $A(t)$ est la limite presque sûre de $A_n(t) \equiv X(t \wedge \tau_n)$.*

Corollaire 4.2 *Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$.*

Alors il existe un unique processus continu à droite croissant A tel que $A(0) = 0$ presque sûrement, $A(t) < \infty$ presque sûrement pour tout t et $M = N - A$ est une \mathcal{F} -martingale locale. De plus, A est localement borné et M est localement de carré intégrable.

Théorème 4.3 Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$ tel que $\mathbf{E}(N(t)) < \infty$ pour tout t . Soit $M = N - A$ est une \mathcal{F} -martingale où A est un processus \mathcal{F} -prédictible tel que $A(0) = 0$ et Soit H un processus \mathcal{F} -prédictible. Alors le processus L donné par

$$L(t) = \int_0^t H(u) dM(u)$$

est une \mathcal{F} -martingale.

Théorème 4.4 (Inégalité maximale de Doob). Soit $X = (X_t)_{t \geq 0}$ une martingale (ou sous martingale positive). Posons $X_t^* = \sup_{s \leq t} |X_s|$ alors:

1. $\forall p \geq 1, \forall \lambda > 0, \lambda^p \mathbf{P}(X_t^* \geq \lambda) \leq \sup_{s \leq t} \mathbf{E}(|X_s|^p).$
2. $\forall p > 1, \|X_t^*\|_{L^p} \leq \frac{p}{p-1} \sup_{s \leq t} \|X_s\|_{L^p}$

Théorème 4.5 (Th. 2.4.3 [7], p. 70). Si pour $i = 1, 2$, H_i est un processus \mathcal{F} -prédictible borné, N_i est un processus ponctuel borné et si la martingales $M_i = N_i - A_i$ vérifie $\mathbf{E}(M_i^2(t)) < \infty$ pour tout t où A_i est l'unique compensateur de N_i , alors

$$\left\langle \int H_1 dM_1, \int H_2 dM_2 \right\rangle = \int H_1 H_2 d \langle M_1, M_2 \rangle$$

et le processus

$$\int H_1 dM_1 \int H_2 dM_2 - \int H_1 H_2 d \langle M_1, M_2 \rangle$$

est une martingale locale sur $[0, +\infty[$.

Théorème 4.6 (Th. 2.4.5 [7], p. 73). Considérons deux processus $U_l \equiv \sum_{i=1}^n \int H_{i,l} dM_i$, $l = 1, 2$. Soit $\{\mathcal{F}_t, t \geq 0\}$ une filtration continue à droite telle que pour chaque i et l , $H_{i,l}$ est un processus \mathcal{F} -prédictible localement borné et $M_i = N_i - A_i$ est une martingale locale de carré intégrable. Alors:

1. (U_l) est une martingale locale de carré intégrable.

Si $\int H_{i,l} d \langle M_i \rangle < \infty$ pour tout i, l ,

2. $\mathbf{E}(U_i(t)) = 0$,
3. $\mathbf{E}(U_1(t)U_2(t)) = \mathbf{E} \sum_{i=1}^n \sum_{j=1}^n \int_0^t H_{i,1} H_{j,2} d < M_i; M_j > \text{ et}$
4. U_i est une martingale sur $[0, t]$.

Lemme 4.1 (lemme 3.2.1 [7], p. 99). Si $S(t) > 0$ alors $\forall T$ t.a. on a:

$$\begin{aligned} \mathbf{E}(\hat{S}_{KM}(t) - S(t)) &= \mathbf{E} \left[\mathbf{1}_{\{T < t\}} \frac{\hat{S}_n(T) (S(T) - S(t))}{S(T)} \right] \\ &= : \mathbf{E}(B(t)) \geq 0 \end{aligned}$$

où \hat{S}_{KM} désigne l'estimateur de Kaplan-Meier et S la fonction de survie.

Rappelons cette définition (cf Annexe [14]).

Définition 4.1 Soit X un processus cadlag nul en 0 et à variations bornées. On appelle produit intégrale (ou produit limite)

$$\prod_{]s,t]} (1 + dX) := \lim_{\delta \rightarrow 0} \prod_{i=1}^n [1 + X(]t_{i-1}, t_i])]$$

qui est indépendante du pas δ défini par $\delta = \max_i |t_i - t_{i-1}|$.

Propriété. Si le processus $(X(t), t \geq 0)$ est continu alors $\prod_{]0,t]} (1 + dX) = \exp(X(t))$

Théorème 4.7 Soient $Y(t) = \prod_{]0,t]} (1 + dX)$ et $Y'(t) = \prod_{]0,t]} (1 + dX')$, alors

$$Y(t) - Y'(t) = \int_{s \in [0,t]} \prod_{]0,s[} (1 + dX) (X(ds) - X'(ds)) \prod_{]s,t]} (1 + dX')$$

Si $Y'(t)$ est régulier, alors

$$\begin{aligned} \frac{Y(t)}{Y'(t)} - 1 &= \int_{s \in [0,t]} \prod_{]0,s[} (1 + dX) (X(ds) - X'(ds)) \left[\prod_{]s,t]} (1 + dX') \right]^{-1} \\ &= \int_0^t \frac{Y(t)(s^-)}{Y'(s)} (X(ds) - X'(ds)) \end{aligned}$$

Théorème 4.8 (*Th. 3.2.3 [7], p. 97*). Si $S(t) > 0$ alors

$$\frac{\hat{S}_{KM}(t)}{S(t)} = 1 - \int_0^t \frac{\hat{S}_{KM}(s^-)}{S(s)} \left(\frac{d\bar{N}_n(s)}{\bar{Y}_n(s)} - dH(s) \right)$$

Théorème 4.9 (*Inégalité de Lengart. cf [7], p. 113*). Soit X un processus continu à droite et Y un processus prédictible tel que $Y(0) = 0$. Si pour tout temps d'arrêt borné T , $\mathbf{E}(|X(T)|) \leq \mathbf{E}(Y(T))$ alors pour tout temps d'arrêt T et tout couple (ε, η) positifs

$$\mathbf{P}(\sup_{t \leq T} |X(t)| \geq \varepsilon) \leq \frac{\eta}{\varepsilon} + \mathbf{P}(Y(T) \geq \eta)$$

Equation: (cf (2.13) [7], p. 104). On a avec $B(t)$ défini dans le lemme 1.

$$\mathbf{E} \left(\hat{S}_{KM}(t) - S(t) - B(t) \right)^2 = S^2(t) \int_0^t \mathbf{E} \left(\frac{\hat{S}_{KM}^2(s^-)}{S^2(s)} \frac{\mathbf{1}_{\{\bar{Y}_n(s) > 0\}}}{\bar{Y}_n(s)} \right) (1 - \Delta H(s)) dH(s)$$

Corollaire 4.3 (*Coro. 3.4.1 de [7], p. 113*). Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel et $M = N - A$ la martingale locale de carré intégrable associée. Si H est un processus adapté continu à gauche possédant des limites à droite (ou plus généralement prédictible et localement borné), alors, pour tout temps d'arrêt T tel que $\mathbf{P}(T < \infty) = 1$ et tout $\varepsilon, \eta > 0$,

$$\mathbf{P} \left(\sup_{\mathbf{t} \leq \mathbf{T}} \left\{ \int_0^{\mathbf{t}} H(s) dM(s) \right\}^2 \geq \varepsilon \right) \leq \frac{\eta}{\varepsilon} + \mathbf{P} \left(\int_0^{\mathbf{T}} H^2(s) d \langle M \rangle (s) \geq \eta \right)$$

Les processus \bar{N}_n, \bar{Y}_n sont introduit en chap. 1 et \hat{F}_n en chap. 2.

Théorème 4.10 (*Th.3.4.2. de [7], p. 115*). Si pour tout $t > 0$, $\bar{Y}_n \xrightarrow{\mathbf{P}} \infty$, alors

$$\sup_{0 \leq s \leq t} \left| \int_0^s \frac{d\bar{N}_n(u)}{\bar{Y}_n(u)} - H(u) \right| \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

et

$$\sup_{0 \leq s \leq t} \left| \hat{F}_n(s) - F(s) \right| \xrightarrow{\mathbf{P}}_{n \rightarrow \infty} 0$$

Théorème 4.11 (*Théorème de Rebolledo*) (*Th. 5.1.1 de [7], p. 204*). Soit $\bar{M}_n = \sum_{i=1}^n M_i$ où M_i est une suite de martingales telle que $M_i = N_i - A_i$ où N_i désigne un processus de comptage et A_i son compensateur. Considérons la suite des processus $I_n(t) = \int_0^t f_n(s) d\bar{M}_n(s)$ et

pour $\varepsilon > 0$, $I_{n,\varepsilon}(t) = \int_0^t f_n(s) \mathbf{1}_{\{|f_n(s)| > \varepsilon\}} d\bar{M}_n(s)$. Supposons que f_n et f sont des processus \mathcal{F}_s -prévisibles et localement bornés tels que $\sup_s |f_n(s) - f(s)| \rightarrow_{n \rightarrow \infty} 0$. Supposons également que les processus N_i, A_i, f_n sont bornés. Posons pour tout $t > 0$, $\alpha(t) = \int_0^t f^2(s) ds < \infty$. Si

1. $\langle I_n \rangle_t \xrightarrow{P}_{n \rightarrow \infty} \alpha(t)$
2. Pour tout $\varepsilon > 0$, $\langle I_{n,\varepsilon} \rangle_t \xrightarrow{P}_{n \rightarrow \infty} 0$

Alors

$$(I_n(t), t > 0) \xRightarrow[n \rightarrow \infty]{} \left(\int_0^t f(s) dW(s), t > 0 \right)$$

où $\xRightarrow{\quad}$ désigne la convergence faible dans l'espace des fonctions continues à droite, ayant une limite à gauche muni de la topologie de Skohorod et W est un mouvement brownien.

Théorème 4.12 (Th. 6.3.1. de [7] p. 235). On a

$$\sqrt{n} \left(\hat{F}_n(\cdot) - F(\cdot) \right) \xRightarrow{n \rightarrow \infty} (1 - F(\cdot)) W(v(\cdot))$$

où $v(t) = \int_0^t \pi^{-1}(s) dH(s)$, pour tout $s \in [0, t]$ avec et $\pi^{-1}(s) = S(t)S_C(t)$

$$\sqrt{n} \left(\frac{\hat{F}_n(\cdot) - F(\cdot)}{1 - \hat{F}_n(\cdot)} \right) \xRightarrow{n \rightarrow \infty} W(v(\cdot))$$

Bibliographie

- [1] Aalen, O. O. (1976). "Nonparametric inference in connection with multiple decrement models". *Scand. J. Statist.*, 3, 15-27.
- [2] Aalen, O. O. (1978). "Nonparametric inference for a family of counting processes". *Ann. Statist.*, 6, 701-726.
- [3] Andersen P.K., Borgan O., Gill R. D. and Keiding N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag. New York.
- [4] Bosq, D. et Lecoutre, J. (1987). *Théorie de l'estimation fonctionnelle*. Economica, Paris.
- [5] Dupuy Jean-François . Cours de statistique avancée. *Université de La Rochelle* (2010)..
- [6] Fermanian Jean David. Modèles de durées. *Cours ENSEE 3ème année*.
- [7] Fleming, T.R. & D.P. Harrington. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons.
- [8] Hajek, J. (1968). "Asymptotic normality of simple linear statistics under alternatives". *Ann. Math. Statist.* 39 325-346.
- [9] Henrik Ramlau Hansen "Smoothing counting process intensities by means of kernel functions". *The Annals of Statistics* 1983 Vol 11 n 2 p. 453-466.
- [10] Hubert C. Modèles pour des durées de survie. *Cours de Statistiques, Université Paris 5*.
- [11] Kallenberg Olav (1997). *Foundations of Modern Probability*. Springer-Verlag. New York.

- [12] Kaplan, E. L. and Meier, P. (1958). "Non-parametric estimation for incomplete observations". *J. Amer. Statist. Assoc.*, 53, 457-481.
- [13] Lopez Olivier. *Réduction de dimension en présence de données censurées*. Thèse de Doctorat, Université de Rennes 1.
- [14] Lorino M. Tristan. *Modèles statistiques pour des données de survie corellées*. Thèse de Doctorat, Université Paris 7 (2002).
- [15] Rabhi Yassir. *Modèles de survie avec un point de rupture*. Thèse de Doctorat, Université du Québec à Montréal (2006).
- [16] Tanner M. A. and Wong W.H. "The estimation of the hazard function from randomly censored data by the kernel method". *The Annals of Statistics 1983 Vol 11 n 33 p. 989-993*.
- [17] Therneau T. M. and Grambsch P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag. New York.
- [18] Viallon Vivian. *Processus empiriques, estimation non paramétrique et données censurées*. *Thèse de Doctorat, Université Paris 6 (2006)*.
- [19] Watson, G. S. and Leadbetter, M. R. (1964 a). Hazard Analysis I. *Biometrika*, 51, 175-184.
- [20] Watson, G. S. and Leadbetter, M. R. (1964 b). Hazard Analysis II. *Sankhya Ser. A*, 26, 101-116.
- [21] Zhang Biao "Some asymptotic results for kernel estimation under random censorship". *Bernoulli 2 (2) 1996, p. 183-198*.