République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

N° Attribué par la bibliothèque

Année univ.: 2019/2020

Estimateur à noyau de la fonction de risque et applications

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université de Saida - Dr Moulay Tahar

Discipline : MATHEMATIQUES

Spécialité : Analyse Stochastique

par

Hamdad **I**kram [1]

Sous la direction de

Dr R. Rouane

Soutenue le 14/09/2020 devant le jury composé de

S. Idrissi	Université Dr Tahar Moulay - Saïda	Président
R. Rouane	Université Dr Tahar Moulay - Saïda	Encadreur
S. Rahmani	Université Dr Tahar Moulay - Saïda	Examinateur
T. Djebbouri	Université Dr Tahar Moulay - Saïda	Examinateur

1. e-mail : hamdadikram@gmail.com

REMERCIEMENT

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.

Un remerciement trés particulier va à Ma directrice de mémoire **Dr. Rachida Rouane** pour son encadrement et son encouragement durant toute la période de la réalisation de ce travail. Je souhaite tout d'abord lui exprimer ma profonde gratitude.

Je voudrais aussi remercier chaleureusement chacun des membres du jury qui me font le grand honneur d'y participer.

Je remercie sincèrement Madame **S. Idrissi** pour l'honneur qu'il me fait en présidant ce jury.

Je remercie vivement monsieur **T. Djebbouri** et madame **S. Rahmani** pour la confiance dont ils me font preuve en faisant parties de ce jury.

je remercie toute ma famille.

Je remercie infiniment mon amie **Boubred Maroua** pour son aide et les services qu'il m'a rendus. Il m'a toujours encouragé et précieusement conseillé.

Je tiens à exprimer aussi ma reconnaissance à tous mes enseignants en particulier **Dr. Benziadi Fatima** et **Dr. Laouni Mimoun** .

Enfin, Je remercie trés amicalement tous mes amies et d'ailleurs de leurs sympathie et leurs aides de près ou de loin.

Dédicace

J'ai toujours pensée faire ou offrir quelque chose à mes parents en signe de reconnaissance pour tout ce qu'ils ont consenti comme efforts, rien que pour me voir réussir, et voilà l'occasion est venue.

Je tiens à dédier ce mémoire à :
Mes parents
Ma trés chère mère, qui a œuvré pour ma réussite, de part sont amour, sont soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude.

Mon père que dieu te garde pour moi. Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de vous.

Mon cher frère **Amine** et mes chères soeurs : **Soumia**, **Siham**, **Wahiba** qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

Toute mes amies pour leurs aide et supports.

Tous ceux qui j'ai connu durant mon cycle d'étude.

Toute la famille **Hamdad**.

Table des matières

Introduction générale

L'analyse des modèles de survie est une branche de la statistique qui a pris son développement depuis la deuxième guerre mondiale. La première méthode d'analyse de survie c'est la méthode actuarielle est apparue en 1912. Sa première utilisation fut dans le domaine médical. Ensuite, elle est devenue une branche indispensable dans divers domaines comme l'actuariat, les séismes etc...

Dans la fiabilité, la durée de survie est, par exemple, définie comme le temps qui sépare la mise en marche d'une machine de la panne de celle-ci. Elle est aussi utilisée dans d'autres domaines comme l'économie, les assurance etc...

L'analyse des donnéées durées de survie est l'étude du délai de la survenue de cet évènement.

Dans le domaine biomédical, on étudie ces durées dans le contexte des études longitudinales comme les enquêtes de cohorte (suivi de patients dans le temps) ou les essais thérapeutiques (tester l'efficacité d'un médicament). On estime la distribution des temps de survie dans le cas paramétrique et non paramétrique.

Une caractéristique importante de l'analyse de la survie est la présence des données censurées. Cette caractéristique, source de difficulté, a nécessité le développement de techniques alternatives à l'inférence usuelle. Les données censurées sont des observations pour lesquelles la valeur exacte d'un évènement n'est pas toujours connue. Cependant, nous disposons tout de même d'une information partielle permettant de fixer une borne inférieure (censure à droite) ou une borne supérieure (censure à gauche). Les raisons de cette censure peuvent être le fait que le patient soit toujours vivant ou non malade à la fin de l'étude, ou qu'il se soit retiré de l'étude pour des

raisons personnelles (immigration, mutation professionnelle).

Pour estimer les fonctions en analyse, on ne dispose généralement que d'un ensemble fini d'observations issu d'une même variable aléatoire. Pour construire des estimateurs de ces fonctions. Il existe deux approches trés complémentaires pour réaliser ces estimations : l'approche paramétrique et l'approche non-paramétrique.

L'approche paramétrique stipule l'appartenance de la loi de probabilité réelle des observations à une classe particulière de lois, qui dépendent d'un certain nombre (fini) de paramètres. L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres, ainsi que de l'obtention d'intervalles de confiance et de la construction de tests. L'inconvénient de la méthode paramétrique est l'inadéquation pouvant exister entre le phénomène étudié et le modèle retenu.

L'estimation non-paramétrique ne suppose aucun modèle, et la distribution est evaluée directement à partir de l'échantillon de données. Ce type d'estimation est préférable lorsqu'on ignore le type de la distribution ayant généré les données et que celui-ci est difficile a déterminer.

En 1983, Tanner et Wong [**?**] ont introduit une famille d'estimateurs de risques non paramétriques basés sur des données. Plusieurs de ces estimateurs ont été étudiés dans le cadre d'une vaste expérience de simulation. L'estimateur qui permet une fenêtre s'est avéré avoir une performance supérieure c'est à dire l'estimateur à noyau qui était fortement compatible avec des conditions suffisantes pour l'estimation de la fonction de risque.

Dans ce mémoire, nous présentons une étude sur les modèles de survie en abordant principalement l'estimation de la fonction de risque. La méthode d'estimation de cette fonction est de type non paramétrique et basée sur l'estimateur à noyau dans les cas i.i.d. sans censures et avec censures.

Ce document est partagé en trois chapitres :

Dans le premier chapitre, nous rappelons des préliminaires sur les modèles de survie. 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...)et un aperçu des modèles de survie paramétriques et non paramétriques. Nous présentons quelques exemples de données censurées avec les graphes des estimateurs correspondants.

Dans le chapitre 2, en première partie, nous donnons un rappel sur l'estimateur à noyau. Nous développons les résultats de H. Ramlau. Hansen [12] 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.

Dans la deuxième partie, nous développons les résultats de M. A. Tanner and W.H. Wong [16] 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 3, nous présentons des simulations numériques illustrant les comportements des estimateurs des fonctions de survie et de fonction de risque sur des données médicales censurées à droite.

Chapitre 1

Modèle de survie

Ce chapitre est consacré à une présentation générale sur les notions de base requises pour la compréhension des autres chapitres. Dans la première partie de ce chapitre, nous donnons les expressions mathématiques des fonctions d'intérêt en analyse de survie. Dans la deuxième partie, nous définissons la notion de censure et ces trois catégories : la censure à droite, la censure à gauche et la censure par intervalle. Dans la troisième partie, nous rappelons la construction de la vraisemblance. Enfin, nous exposons, trés brièvement, les principales méthodes d'estimation de la survie. Dans le cas d'observations censurées à droite, nous introduisons l'estimateur de Nelson-Aalen du risque cumulé et Kaplan-Meier de la fonction de survie et d'autres estimateurs de Fleming-Harrington et Breslow.

1.1 Concepts de base et notation

Une étude de survie est une étude :

Longitudinale (suivi des personnes au cours du temps).

Prospective (prise en compte des évènements survenant dans la durée de l'étude).

Cohorte : ensemble de sujets inclus dans une étude au même moment, et suivis dans des conditions standardisées pendant une durée prédéfinie.

Quelques définitions sont couramment utilisées dans les études de survie :

Évènement d'intérêt : évènement auquel on s'intéresse au cours de l'étude $\Longrightarrow$ décès, décès lié à un AVC, complication, rechute, disparition de symptômes. On utilisera l'analyse de survie dès qu'il y aura une notion de durée jusqu'à la survenue de l'évènement d'intérêt (qu'on nommera décès).

Durée de survie : délai entre la date d'origine et la date de survenue ou la date des dernières nouvelles.

Date d'origine : c'est la date correspondant au point de départ de la surveillance. Elle peut être différente pour chaque sujet selon les modalités d'inclusion du sujet. Dans certains cas la date d'origine peut être antérieure à l'inclusion dans l'étude cohorte historique.

Date de point : c'est la date choisie pour faire le bilan.

Date des dernières nouvelles : c'est la date la plus récente à laquelle on a recueilli des informations sur le patient, notamment la survenue ou non de l'évènement d'intérêt.

Perdu de vue : un sujet est perdu de vue lorsque sa surveillance est interrompue avant la date de point et que l'évènement d'intérêt ne s'est pas produit.

Placeboo : un médicament fictif donné à titre exprémental afin de vérifier les effets psychologiques d'une médication.

1.1.1 Description de la distribution des temps de survie

On appelle durée de vie une variable aléatoire T positive, généralement, la durée s'écoulant entre deux évènements.
Exemple d'évènements : mort, panne, sinistre, entrée en chômage, maladie.
Cinq fonctions équivalentes définissent la loi de la durée : Supposons que la durée de survie T soit une variable positive ou nulle, et absolument continue. Alors sa loi de probabilitée peut être définie par l'une des fonctions suivantes :

Définition 1.1.1. ***(Fonction de survie** S**)***

Pour des données continues la durée de vie T, c'est-à-dire la durée observée d'un individu dans un état initial, est une variable aléatoire définie sur $[0, +\infty[$ de fonction de répartition F. La fonction de survie est définie comme :

$$S(t) = \mathbb{P}\{T \geq t\}, t \geq 0$$

Pour t fixé c'est la probabilité de survivre jusqu'à l'instant t. C'est donc une fonction continue monotone non croissante telle que

$$S(0) = 1$$

et

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

Définition 1.1.2. ***(Fonction de répartition** F**)***

La fonction de répartition (f.r ou c.d.f en anglais pour "cumulative distribution function") est

$$F(t) = \mathbb{P}\{T < t\}, t \geq 0$$

Pour t fixé, c'est la probabilité de mourir avant l'instant t. F est une fonction croissante et continue à gauche en tout point de $\mathbb{R}$

$$F(0) = 0$$

et

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

Définition 1.1.3. ***(Densité de probabilité** f**)***

C'est une fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$

$$F(t) = \int f(s)ds$$

Si la fonction de répartition à une dérivée au point t alors

$$f(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t+dt)}{dt} = F'(t) = -S'(t)$$

Pour t fixé, la densité de probabilité caractérise la probabilité de mourir dans un petit intervalle de temps aprés l'instant t.

Définition 1.1.4. ***(Risque instantané** h **(ou taux de hasard))***

Le risque instantané (ou taux d'incidence), pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps aprés t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt | X \geq t)}{dt} = \frac{f(t)}{S(t)}$$

Définition 1.1.5. ***(Taux de hasard cumulé** H**)***

C'est l'intégrale du taux de hasard h :

$$H(t) = \int_0^t h(u)du = -\ln S(t)$$

On peut déduire la fonction de survie du taux de hasard cumulé grâce a la relation :

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

Toutes ces fonctions sont donc liées entre elles : la connaissance de $S(t)$ permet celle de $f(t)$, $h(t)$ et $H(t)$. De même, la connaissance de $h(t)$ permet celle de $H(t)$ donc de $S(t)$ et finalement de $f(t)$. En d'autres termes, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également définies. En particulier, un choix de spécification sur la fonction de risque instantané implique la sélection d'une certaine distribution des données de survie.
Et voici la relation entre ces différentes fonctions :

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

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

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

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

Exemple1 : loi exponentiel

T suit une loi exponentielle de paramètre $\gamma > 0$ si :

- $F(t) = 1 - \exp(-\gamma t)$
- $S(t) = \exp(-\gamma t)$
- $f(t) = \gamma \exp(-\gamma t)$
- $h(t) = \gamma$

Exemple2 : loi de Weibull

C'est une famille de lois indexée par deux paramètres généralisant la loi exponentielle.

T suit une loi de Weibull de paramètre $\gamma > 0$ et $\alpha > 0$ si :

- $F(t) = 1 - \exp(-\gamma t^{\alpha})$
- $S(t) = \exp(-\gamma t^{\alpha})$
- $f(t) = \gamma\alpha t^{\alpha-1} \exp(-\gamma t^{\alpha})$
- $h(t) = \gamma\alpha t^{\alpha-1}$

Exemple3 : loi log-logistique

T suit une loi log-logistique de paramètre $\gamma > 0$ et $\alpha > 0$

- $F(t) = 1 - \left[\dfrac{1}{1 + (\gamma t)^{\alpha}} \right]$
- $S(t) = \dfrac{1}{1 + (\gamma t)^{\alpha}}$
- $f(t) = \dfrac{\gamma\alpha(\gamma t)^{\alpha-1}}{(1 + (\gamma t)^{\alpha})^2}$
- $h(t) = \dfrac{\gamma\alpha(\gamma t)^{\alpha-1}}{1 + (\gamma t)^{\alpha}}$

Définition 1.1.6. ***(Quantiles de la durée de survie)*** *Pour $0 < p < 1$, on définit le quantile t_p et la fonction $q(p)$ tq $p \in (0, 1)$ comme*

$$t_p \equiv q(p) = \inf\{t : F(t) \geq p\}$$

Quand $F(t)$ est strictement croissante et continue alors

$$t_p \equiv q(p) = F^{-1}(p), 0 < p < 1$$

Pour p fixé, le quantile t_p est le temps auquel une proportion p de la population à disparu.

Définition 1.1.7. ***(Moyenne et variance de la durée de survie)*** *Le temps moyen de survie $\mathbb{E}(T)$ ainsi que sa variance $Var(T)$ sont des quantités importantes :*

$$\mathbb{E}(T) = \int_0^\infty S(t)dt$$

$$Var(T) = 2\int_0^\infty tS(t)dt - \{\mathbb{E}(T)\}^2$$

La moyenne et la variance peuvent être déduites de n'importe laquelle des cinq fonctions ci-dessus (F, S, f, h, H), mais pas vice versa.

1.2 Les données incomplètes (Données censurées)

L'analyse des durées de vie pose des problèmes particulier dus au fait que les observations des durées de vie sont le plus souventcensurées. Une des caractéristiques des données de survie est l'existence d'observations incomplètes. Par exemple, dans les enquêtes épidémiologiques, les données sont souvent recueillies de façon incomplète. La censure fait partie un processus générant ce type de données. Elle doit être prises en compte dans l'écriture de la vraisemblance. Nous parlerons de donnée censurées lorsque la durée de survie n'est connue que lorsqu'elle est limitée par une durée limitée d'observation.

1.2.1 Censure

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour l'individu i, considérons

– Son temps de survie T_i
– Son temps de censure C_i
– La durée réellement observée X_i

Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'évènement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées, pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

1. La censure de type I

   Soit C une valeur fixée, au lieu d'observer les variables $T_1, ..., T_n$ qui nous intéressent, on n'observe T_i uniquement lorsque $T_i \leq C$, sinon on sait uniquement que $T_i > C$, on utilise la notation suivante :

   $$X_i = T_i \wedge C = min(T_i, C)$$

   Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Par exemple, on peut tester la durée de vie de N objet identiques (ampoules) sur un intervalle d'observation fixé $[0, u]$. En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps u sont sacrifiées).

2. La censure de type II

   Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $T(i)$ et $X(i)$ les statistiques d'ordre des variables T_i et X_i, la date de censure est donc $T(k)$ et on observe les variables suivantes

   $$X_{(1)} = T_{(1)}$$

.

.

.

$$X_{(k)} = T_{(k)}$$

$$X_{(k+1)} = T_{(k)}$$

.

.

.

$$X_{(n)} = T_{(k)}$$

3. La censure de type III (ou censure aléatoire de type I)

   Soient $C_1, ..., C_n$ des variables aléatoires i.i.d. On observe les variables $X_i = T_i \wedge C_i$. L'information disponible peut être résumée par :

   – La durée réellement observée X_i.

   – Un indicateur $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$.

   $\delta_i = 1$ si l'évènement est observé (d'où $X_i = T_i$). On observe les vraies durées ou les durées complètes.

   $\delta_i = 0$ si l'individu est censuré (d'où $X_i = C_i$). On observe des durées incomplètes (censurées).

La censure aléatoire est La plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par

a - La perte de vue : le patient quitte l'étude en cours et on ne le revoit plus(à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients perdus de vue.

b -L'arrêt ou le changement du traitement : les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.

c -La fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'évènement). Ce sont des patients (exclus-vivants). Les (perdus de vue) (et les exclusions) et les (exclus-vivants) correspondent à des observations censurées mais les deux mécanismes sont de nature différente(la censure peut être informative chez les perdus de vue).

Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'évènement avant que l'individu soit observé. On sait uniquement que la date de l'évènement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (X, δ) : comme pour la censure à droite, on suppose que la censure C est indépendante de T tel que

$$X_i = T_i \vee C = \max(T_i, C)$$

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

Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'évènement, la seule information disponible est qu'il a eu lieu entre deux dates connues. Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent

suivies par intermittence (pas en continu), on sait alors uniquement que l'évènement s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'évènement correspond au temps de la visite pour se ramener à de la censure à droite.

1.3 Fonction de vraisemblance

Considérons le cas d'une censure aléatoire droite C indépendante de la durée d'intérêt T. Supposons que les variables T et C ont pour densités respectives f et g et pour survies S et G. La distribution de T est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (X_i, i), où $X_i = min(T_i, C_i)$ est la durée observée et l'indicateur de censure $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}} g$. Ainsi, la contribution à la vraisemblance pour l'individu i est

$$\begin{aligned} L_i &= \mathbb{P}(X_i \in [t_i, t_i + dt], \delta_i = 1|\theta)^{\delta_i} \times \mathbb{P}(X_i \in [t_i, t_i + dt], \delta_i = 0|\theta)^{1-\delta_i} \\ &= \mathbb{P}(T_i \in [t_i, t_i + dt], C_i \geq T_i|\theta)^{\delta_i} \times \mathbb{P}(C_i \in [t_i, t_i + dt], C_i < T_i|\theta)^{1-\delta_i} \\ &= [f(t_i|\theta)G(t_i^{-1})]^{\delta_i} \times [g(t_i)S(t_i|\theta)]^{1-\delta_i} \end{aligned}$$

.

Par l'hypothèse (de censure non informative), le paramètre d'intérêt n'apparaît pas dans la loi de la censure (il existe des mécanismes indépendants et informatifs). La partie utile de la vraisemblance se réduit alors à

$$L = \prod_{i=1}^{n} f(t_i|\theta)^{\delta_i} S(t_i|\theta)^{1-\delta_i} \tag{1.1}$$

1.4 Erreur quadratique moyenne (MSE)

L'erreur quadratique moyenne d'un estimateur $\widehat{\theta}$ d'un paramètre θ de dimension 1 est une mesure caractérisant la précision de cet estimateur. Elle est plus souvent appelée erreur quadratique, moyenne (étant sous-entendu), elle est parfois appelée

aussi risque quadratique. Nous la noterons MSE (pour Mean Squared Error). L'erreur quadratique moyenne est définie par :

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

.

Si l'on compare deux estimateurs sans biais, le meilleur est bien sûr celui qui présente le MSE le plus faible.

1.5 Estimation sur les données de survie

Il existe de nombreuses méthodes d'estimation de la fonction de survie $S(t)$. Ces méthodes peuvent se regrouper en deux catégories : les méthodes paramétriques et les méthodes non paramétriques. Les méthodes paramétriques par définition utilisent des modèles spécifiant a priori la forme de la courbe de $S(t)$. Parmi les modèles paramétriques estimant $S(t)$, on peut citer : le modèle exponentiel, le modèle de Weibull, le modèle log-logistique et le modèle de Gompertz. Ces modèles sont utilisés en analyse de survie seulement dans le cas où l'on sait déjà quelle est la forme de la courbe de $S(t)$. Dans de nombreuses situations en médecine, on ne le sait pas et l'on préfère alors les méthodes non paramétriques, qui elles ne font pas d'hypothèse quant à la forme de la courbe de $S(t)$.

1.5.1 Estimation paramétrique de la fonction de survie

Soit l'indicateur :

$$\delta_i = \begin{cases} 1 & \text{si} T_i \leq C \\ 0 & \text{sinon} \end{cases}$$

Si les données ne comportaient pas de censures, la fonction de vraisemblance de ce modèle serait :

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i}.$$

Or avec la présence de données censurées indépendantes, il faut rajouter dans notre fonction de vraisemblance le cas des données censurées, on obtient :

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Distribution exponentielle

On a :

$$h(t) = \lambda$$

$$S(t) = \exp(-\lambda t).$$

La contribution à la vraisemblance peut s'écrire sous la forme :

$$L(t_i, \lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda t_i)^{1-\delta_i}$$

$$logL(t_i, \lambda) = \sum_{i=1}^{n} \delta_i log(\lambda) + \sum_{i=1}^{n} log(\exp(-\lambda t_i)).$$

La dérivée s'annule pour

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

Distribution de Weibull

On a :

$$f(t) = \gamma \lambda^{\gamma} t^{\gamma-1} \exp -(\lambda t)^{\gamma}$$

$$S(t) = \exp(-\lambda t)^{\gamma}.$$

La vraisemblance de ce modèle s'écrit :

$$L(t_i, \lambda, \gamma) = \prod_{i=1}^{n} (\gamma \lambda^{\gamma} t^{\gamma-1} \exp -(\lambda t)^{\gamma})^{\delta_i} (\exp(-\lambda t)^{\gamma})^{1-\delta_i}.$$

D'où l'on déduit la log-vraisemblance :

$$\log L(t_i, \lambda, \gamma) = \gamma \log \lambda \sum_{i=1}^{n} \delta_i + \log \gamma \sum_{i=1}^{n} \delta_i + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log t_i - \sum_{i=1}^{n} (\lambda t_i)^{\gamma}.$$

La dérivée s'annule pour

$$\lambda = \left(\frac{1}{\sum_{i=1}^{n} t_i^{\gamma}} \sum_{i=1}^{n} \delta_i) \right)^{\frac{1}{\gamma}}$$

$$\frac{1}{\gamma} = \frac{\sum_{i=1}^{n} t_i^{\gamma} \log t_i}{\sum_{i=1}^{n} t_i^{\gamma}} - \frac{1}{\sum_{i=1}^{n} t_i^{\gamma}} \sum_{i=1}^{n} \log t_i$$

1.5.2 Estimation non-paramétrique

Dans cette partie, nous définissons l'estimateur empirique en absence de censure et la méthode de Kaplan-Meier ces des méthodes non paramétriques d'estimation de la fonction $S(t)$ tel que L'estimateur de la fonction de survie le plus utilisé lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie est l'estimateur de Kaplan-Meier et on peut être intéressé par l'estimation d'autres fonctions qui caractérisent la distribution des temps d'évènements. Nous traiterons donc de l'estimation de la fonction de risque cumulée, avec l'estimateur de Nelson-Aalen.

Estimateur de Kaplan-Meier de la fonction de survie

Présentation heuristique :

La fonction de survie est donc définie comme :

$$S(t) = \mathbb{P}[T > t] = 1 - F(t), t \geq 0.$$

Soient 2 durées t_1 et t_2 telles que $t_2 > t_1$, alors :

$$\mathbb{P}[T > t_2] = \mathbb{P}[T > t_2 \ et \ T > t_1].$$

Puisque pour survivre aprés t_2 il faut naturellement avoir déjà survécu au moins pendant une durée t_1. On utilise ensuite le théorème des probabilités conditionnelles et il vient :

$$\mathbb{P}[T > t_2] = \underbrace{\mathbb{P}[T > t_2 | T > t_1]}_{(a)} \times \underbrace{\mathbb{P}[T > t_1]}_{(b)}.$$

Où

(a) peut être estimé par $1 - \frac{dt_2}{nt_2}$ où dt_2 est le nombre d'individus ayant connu l'évènement en t_2. nt_2 le nombre d'individus qui auraient pu connaître l'évènement en question entre t_1 exclu et t_2. En d'autres termes nt_2 est le nombre d'individus à risque au temps t_2. Si le temps était vraiment continu on devrait toujours avoir $d = 1$. En pratique la périodicité de collecte des données dissimule cette continuité et on observe couramment des valeurs de d supérieures à l'unité traduisant le fait que la discrétisation du temps imposée par le mode de collecte fait que plusieurs individus connaissent l'évènement au même instant t. Lorsqu'existent des durées censurées entre t_1 inclus et t_2 exclus, la convention retenue est de ne pas prendre en compte les individus concernés dans le calcul de nt_2 (nombre d'individus à risque en t_2). Ainsi, si nt_1 et nt_2 sont respectivement les nombres d'individus à risque en t_1 et t_2, dt_1 le nombre d'individus ayant connu l'évènement en t_1, $c[t_1, t_2[$ le nombre d'individus censurés entre les deux dates, on a

$$nt_2 = nt_1 - dt_1 - c[t_1, t_2[$$

(b) est par définition $S(t_1)$. On a donc

$$\hat{S}(t_2) = (1 - \frac{dt_2}{nt_2}) \times \hat{S}(t_1).$$

L'équation précédente donne une récurrence permettant de calculer $\hat{S}(t_1)$ pour tout temps d'évènement t observé, sachant qu'initialement

$$\hat{S}(0) = 1$$

$$\hat{S}(t) = \prod_{i|t_i \leq t} 1 - \frac{d_i}{n_i} \tag{1.2}$$

où

– t_i représente le temps de suivi depuis l'inclusion dans l'étude pour chaque patient i.

– n_i est le nombre de sujets à risque de présenter l'évènement étudié à l'instant t_i, i.e. le nombre de patients n'ayant pas encore subi l'évènement ni la censure juste avant t_i.

– d_i est le nombre de décès au temps t_i.

$\hat{S}(t)$ une fonction en escalier décroissante,continue à droite, qui ne saute qu'aux instants de morts réelles. On peut montrer que sous des conditions assez faibles, l'estimateur de Kaplan-Meier $\hat{S}(t)$ à asymptotiquement une distribution normale centrée sur $S(t)$. Une de ces conditions est que **la censure soit non informative** relativement à l'évènement étudié : une façon de comprendre cette condition est que la probabilité de connaître l'évènement étudié à un temps t quelconque est la même pour les individus censurés et les individus non censurés.

Exemple 1 :

On ne considère pour l'instant que des données complètes c.à.d non censurées relatives à des temps de réalisation d'un évènement mesurés en jours et observé sur

10 individus.

$$6, 19, 32, 42, 42, 43, 94, 105, 105, 120$$

Représentation des données dans le tableau I :

t_i	d_i	n_i	$1 - \frac{d_i}{n_i}$	$\hat{S}(t_i)$	$\hat{F}(t_i) = 1 - \hat{S}(t_i)$
6	1	10	$1 - \frac{1}{10} = 0.90$	0.90	0.10
19	1	9	$1 - \frac{1}{9} = 0.889$	0.80	0.20
32	1	8	$1 - \frac{1}{8} = 0.875$	0.70	0.30
42	2	7	$1 - \frac{2}{7} = 0.7143$	0.50	0.50
43	1	5	$1 - \frac{1}{5} = 0.80$	0.40	0.60
94	1	4	$1 - \frac{1}{4} = 0.75$	0.30	0.70
105	2	3	$1 - \frac{2}{3} = 0.330$	0.10	0.90
120	1	1	$1 - 1 = 0$	0	1

tableau I

Exemple 2 :

On introduit des données censurées. Dans ce cas la fonction de survie n'est estimée que pour les temps observés mais il faut naturellement ajuster le nombre d'individus à risque. La règle est que pour une durée donnée t_i on ne comptabilise dans les individus risqués que ceux qui ont une date d'évènement égale ou supérieure à t_i ou une durée de censure supérieure à t_i (au passage on notera qu'une convention est que si, pour un individu quelconque, les survenues de l'évènement et de la censure sont concomitantes alors on le considère comme non censuré. En d'autres termes, on impose que la réalisation de l'évènement précède la censure). Dans la liste ci-dessous relatives à 19 durées mesurées en jours, les données censurées sont signalées par l'exposant + :

$6, 19, 32, 42, 42, 43+, 94, 126+, 169+, 207, 211+, 227+, 253, 255+, 270+, 310+, 316+, 335+, 346+$

On obtient alors

t_i	d_i	n_i	$1-\frac{d_i}{n_i}$	$\hat{S}(t_i)$	$\hat{F}(t_i)=1-\hat{S}(t_i)$
6	1	19	0.947	0.947	0.053
19	1	18	0.944	0.895	0.105
32	1	17	0.941	0.842	0.158
42	2	16	0.875	0.737	0.263
94	1	13	0.923	0.680	0.320
207	1	10	0.90	0.612	0.388
253	1	7	0.957	0.525	0.475

tableau II

Estimateur de Nelson-Aalen de la fonction de risque cumulé

Présentation heuristique :

Définition 1.5.1. *Nelson (1972) et Aalen (1978) ont proposé un estimateur de la fonction de risque cumulée $H(t)$. En raison de la relation $S(t) = exp(-H(t))$on peut estimer le hasard intégré par $-\ln \hat{S}_{KM}(t)$ soit :*

$$H_n^{(1)}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{X_{(i)} \leq t\}} . \ln(1 - \frac{d_{(i)}}{n-i+1}) \tag{1.3}$$

Si $\hat{S}_{KM}(t) \neq 0$

Autre estimateur : l'estimateur de Nelson-Aalen (plus simple, meilleures propriétés de convergence)

Rappel : *Si T admet une densité f par rapport à la mesure de Lebesgue,*

$$H(t) = \int_0^t \frac{f(u)}{S(u)} du.$$

En fait, même si la distribution de T n'admet pas de dérivée en tout point de $\mathbb{R}+$, on peut définir par :

$$H(t) = -\int_0^t \frac{S(du)}{S(u^-)}$$

Dans le cas de données censurées à droite, on peut écrire :

$$H(t) = -\int_0^t \frac{F^{(1)}(du)}{F(u^-)}.$$

Où

$$F(u) = \mathbb{P}(T > u).$$

Et

$$F^{(1)}(u) = \mathbb{P}(T > u, \delta = 1).$$

Donc

$$F^{(1)}(du) = G(u^-)S(du)$$

$$F(u^-) = G(u^-)S(u^-)$$

Remarque 1.5.1. *Si f et λ existent en tout point réel positif, on retrouve ainsi que le hasard intégré est la primitive de λ. En remplaçant les deux fonctions $F^{(1)}$ et F par leurs équivalents empiriques, on obtient alors l'estimateur de Nelson-Aalen, défini en tout t tel que $\mathbb{P}(T > t) > 0$*

$$\hat{H}_n^{(2)}(t) = -\int_0^t \frac{F^{(1)}(du)}{F(u^-)}.$$

Avec

$$F_n(u) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)} > u\}}$$

$$F_n^{(1)}(u) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{T_i > u, \delta_i = 1\}}$$

Il s'écrit également

$$H_n^{(2)}(t) = \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{\{T_i \leq t\}}}{\sum_{j=1}^{n} \mathbb{1}_{\{T_j \geq T_i\}}} \tag{1.4}$$

L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut à chaque instant de mort t_i La plus simple écriture de l'estimateur est :

$$\hat{H}(t) = \sum_{T_{(i)} \leq t} \frac{d_i}{n_i} \tag{1.5}$$

Exemple :

On observe les durées de vie de 10 diodes exprimées en mois (+ si censurées) :

$1, 2, 4+, 5, 7+, 8, 9, 10+, 11, 13+$

$Obs = \{(1,1), (2,1), (4,0), (5,1), (7,0), (8,1), (9,1), (10,0), (11,1), (13,0)\}$

$$\hat{H}_n(t) = \sum_{T_{(i)} \leq t} \frac{d_i}{n_i}$$

$$\hat{H}_n(0) = 0$$

t_i	d_i	n_i	$\frac{d_i}{n_i}$	$\hat{H}_n(t)$
1	1	10	$\frac{1}{10}$	$\frac{1}{10} + \hat{H}_n(0) = 0.1$
2	1	9	$\frac{1}{9}$	$\frac{1}{9} + \hat{H}_n(1) = 0.21$
5	1	7	$\frac{1}{7}$	$\frac{1}{7} + \hat{H}_n(2) = 0.35$
8	1	5	$\frac{1}{5}$	$\frac{1}{5} + \hat{H}_n(5) = 0.55$
9	1	4	$\frac{1}{4}$	$\frac{1}{4} + \hat{H}_n(8) = 0.80$
11	1	2	$\frac{1}{2}$	$\frac{1}{2} + \hat{H}_n(9) = 1.3$

L'estimateur de Breslow

Définition 1.5.2. *Les fonctions de hasard cumulé et de survie sont liées par :*

$$H(t) = -lnS(t)$$

l'estimateur de Kaplan-Meier de $S(t)$ induit donc un estimateur non paramétrique de $H(t)$ appelé l'estimateur de Breslow :

$$\hat{H}_B(t) = -\ln \hat{S}_{KM}(t) \tag{1.6}$$

Remarque 1.5.2. *En tout point l'estimateur de Nelson-Aalen est inférieur à celui de Breslow : $\hat{H}_{NA}(t) < \hat{H}_B(t)$*

L'estimateur d'Harrington-Fleming

Définition 1.5.3. *La fonction de survie et le hasard cumulé sont liés par :*

$$S(t) = \exp\{H(t)\}$$

l'estimateur de Nelson-Aalen du hasard cumulé induit donc un estimateur non paramétrique de la survie :

$$\hat{S}_{HF}(t) = \exp\{-H_{NA}(t)\}. \tag{1.7}$$

*Appelé **l'estimateur d'Harrington-Fleming**.*

Chapitre 2

Estimateur à noyau de la fonction de risque

Dans ce chapitre, nous avons 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 [12] qui propose une méthodologie, citée par KLein et Moeschberger (2003), permettant de construire un estimateur $\hat{h}_n$ de la fonction de risque par lissage. Dans le cas des données censurées, nous avons donné les formes exactes et asymptotiques de la moyenne et la variance de l'estimateur à noyau de la fonction de risque h en développant les résultats de l'article de Martin Tanner et Wing Hung Wong(1983)[**?**].

2.1 Construction de l'estimateur pour des données non censurées

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

$$\hat{f}_n(t) = \frac{1}{b_n} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{b_n} \right) dF_n(s) \tag{2.1}$$

.

Où F_n est la fonction de répartition empirique des T_i, b_n est la fenêtre avec $b_n \to 0$ et K un noyau de Parzen -Rosemblatt de support $[-1, 1]$.

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

$$\hat{h}_n(t) = \frac{1}{b_n} \sum_{i=1}^{n} \frac{1}{n-i+1} K\left(\frac{t - T_{(i)}}{b_n} \right). \tag{2.2}$$

Où $T_{(1)}, T_{(2)}, ..., T_{(n)}$ est la statistique d'ordre des T_i.
Si on pose $\overline{N}_n(t) = nF_n(t) = \sum_{i=1}^{n} \mathbb{1}_{T_i \leq t}$ et $\overline{Y}_n(t) = n - \overline{N}_n(t^-) = \sum_{i=1}^{n} \mathbf{1}_{T_i \geq t}$, alors 2.1 peut s'écrire

$$\hat{h}_n(t) = \frac{1}{b_n} \int_0^{\infty} K\left(\frac{t-s}{b_n} \right) d\hat{H}_{NA}(s). \tag{2.3}$$

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

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

Dans l'étude de l'estimation 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 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.

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

– Le processus $N_i(t)$ indicateur de "survenue de l'évènement" $N_i(t) = \mathbb{1}_{\{X_i \leq t, D_i = 1\}}$.

– 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 : $Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}$.

– Le processus $\bar{Y}_n(t) = \sum_{i=1}^{n} Y_i(t)$: le nombre d'individus à risque à l'instant t.

Soit $H(t) = \int_0^t h(t)dt$. 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 $J_n(s) = \mathbb{1}_{\{\bar{Y}_n(s)>0\}}$. 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 $\Big(\ (\hat{H}_n - \tilde{H}_n)(t), t \geq 0 \ \Big)$ est une martin-

gale de carré intégrable et de variation $< \hat{H}_n - \tilde{H}_n >_t = \int_0^t \frac{J(s)}{\bar{Y}_n(s)} h(s)ds$ (d'aprés le théorème qui s'écrit au dessous). On note l'erreur quadratique par

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

On pose

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

.

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

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

$\hat{\eta}_n(t)$ est un estimateur sans biais de $\eta_n(t)$. En effet :

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

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

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

Par suite

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

2.1.1 L'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 $T_1, T_2, ..., T_n$, donc

$$\hat{h}_n(t) = \frac{\frac{1}{b_n}\sum_i K(\frac{t-T_{(i)}}{b_n})}{\bar{Y}_n(T_{(i)}} \tag{2.4}$$

2.1.2 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 $t \in [b_n, 1-b_n]$

$$\tilde{h}_n(t) = \frac{1}{b_n}\int_0^1 K\left(\frac{t-s}{b_n}\right) d\tilde{H}_n(s) = \int_{-1}^1 K(u)h(t-b_nu)J(t-b_nu)du. \tag{2.5}$$

Proposition 2.1.1. *Posons* $K_{b_n}(s) = \frac{1}{b_n}K(\frac{s}{b_n})$ *et* $j_n(s) = \mathbb{E}(J_n(s))$ *alors pour tout* $t \geq 0$

$$\mathbb{E}\left(\hat{h}_n(t)\right) = \mathbb{E}\left(\tilde{h}_n(t)\right) = (K_{b_n} * hj_n)(t). \tag{2.6}$$

Et

$$\sigma_n^2(t) = \mathbb{E}\left(\hat{h}_n(t) - \tilde{h}_n(t)\right)^2 = \frac{1}{b_n}\int_{-1}^1 K^2(u)h(t-b_nu)\mathbb{E}\left(\frac{J_n(t-b_nu)}{\bar{Y}_n(t-b_nu)}\right)du. \tag{2.7}$$

Proposition 2.1.2. *Pour tout* $t \in [b_n, 1-b_n]$ *un estimateur sans biais de* $\sigma^2(t)$ *est*

$$\sigma_n^2(t) = \frac{1}{b_n^2}\int_0^1 K^2\left(\frac{t-s}{b_n}\right)\left(\frac{J_n(s)}{\bar{Y}_n^2(s)}\right)d\bar{N}_n(s). \tag{2.8}$$

Preuve de proposition : On a

$$\hat{h}_n(t) - \tilde{h}_n(t) = \frac{1}{b_n}\int_0^1 K\left(\frac{t-s}{b_n}\right) d\left(\hat{H}_{NA} - \tilde{H}_n\right)(s). \tag{2.9}$$

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

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

D'où

$$\begin{aligned} \mathbb{E}(\hat{h}_n(t)) &= \mathbb{E}(\tilde{h}_n(t)) = \mathbb{E}\left(\int_{-1}^{1} K(u)h(t-b_nu)J_n(t-b_nu)du \right) \\ &= \int_{-1}^{1} K(u)h(t-b_nu)\mathbb{E}(J_n(t-b_nu))du \\ &= \int_{-1}^{1} K(u)h(t-b_nu)j_n(t-b_nu))du = (K_{b_n} * hj_n)(t). \end{aligned} \tag{2.10}$$

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

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

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

$$\begin{aligned} \hat{\sigma}_n^2(t) &= \frac{1}{b_n^2}\int_{-1}^{1} K^2(\frac{t-s}{b_n})\left(\frac{J_n(s)}{\bar{Y}_n(s)} \right)^2 d\bar{N}_n(s) \\ &= \frac{1}{b_n^2}\int_{-1}^{1} K^2(\frac{t-s}{b_n})\frac{J_n(s)}{\bar{Y}_n(s)}h(s)ds \\ &+ \frac{1}{b_n^2}\int_{-1}^{1} K^2(\frac{t-s}{b_n})\frac{J_n(s)}{\bar{Y}_n(s)^2}d\bar{M}_n(s). \end{aligned} \tag{2.11}$$

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

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

D'où le résultat.

Si on note le premier terme de l'équation 2.11 par

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

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

Donc de l'équation 2.11 on a

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

D'où le résultat.

Le résultat suivant donne une expression asymptotique de $\mathbb{E}(\hat{\sigma}_n^2(t) - \sigma_n^{*2}(t))^2$. De même l'équation 2.6 permet de déduire que l'estimateur à noyau de la fonction de risque est asymptotiquement sans biais.

Propriété 2.1.1. ***(a)*** *Si h est continue au point t et si $j_n(s) = \mathbb{E}J_n(s) \to_{n\to\infty} 1$ uniformément dans un voisinage de $t > 0$, alors*

$$\mathbb{E}(\hat{h}_n(t)) \to_{n\to\infty} h(t) \tag{2.12}$$

Propriété 2.1.2. ***(b)*** *Si* $n\mathbb{E}\left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) \to \frac{1}{\tau(s)}$ *uniformément dans un voisinage de* t *et* h *et* τ *sont continues au point* $t > 0$*, alors quand* $b_n \to 0$

$$\sigma_n^2(t) = \mathbb{E}\left(\hat{h}_n(t) - \tilde{h}_n(t) \right)^2 = \frac{1}{nb_n}\frac{h(t)}{\tau(t)}\int_{-1}^{1} K^2(u)du + o\left(\frac{1}{nb_n} \right) \tag{2.13}$$

Preuve : (a) Par la relation 2.6 et $j_n(s) \to_{n\to\infty} 1$ uniformement 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

$$\begin{aligned} \mathbb{E}(\hat{h}_n(t)) &= (K_{b_n} * hj_n)(t) = \frac{1}{b_n}\int K(\frac{t-u}{b_n})h(u)j_n(u)du \\ &= \int_{-1}^{1} K(v)h(t-vb_n)j_n(t-vb_n)dv \to h(t)\int_{-1}^{1} K(v)dv = h(t) \end{aligned} \tag{2.14}$$

(b) On a par la relation 2.7

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

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

$$\int_{-1}^{1} K^2(u)h(t-b_nu)\left(n\mathbb{E}(\frac{J_n(t-b_nu)}{\bar{Y}_n(t-b_nu)}) - \frac{1}{\tau(t-b_nu)} \right)du \to_{n\to\infty} 0.$$

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

Pour le deuxième on a

$$\int_{-1}^{1} K^2(u)h(t-b_nu)\frac{1}{\tau(t-b_nu)}du \to_{n\to\infty} \frac{h(t)}{\tau(t)}\int_{-1}^{1} K^2(u)du.$$

Ce qui termine la démonstration.

Remarque 2.1.1. *La relation 2.13 montre que si* $nb_n \to \infty$ *alors*

$$\mathbb{E}\left(\ \hat{h}_n(t) - \tilde{h}_n(t)\ \right)^2 \to_{n\to\infty} 0.$$

Et par suite

$$\hat{h}_n(t) - \tilde{h}_n(t) \to^P_{n\to\infty} 0$$

2.1.3 Propriétés asymptotiques

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

Propriété 2.1.3. *Si* $n\mathbb{E}\left(\ \frac{J_n(s)}{\bar{Y}_n(s)}\ \right) \to_{n\to\infty} \frac{1}{\tau(s)}$ *uniformément dans un voisinage de* t *et* h *et* τ *sont continues en* $t > 0$, *alors*

$$\mathbb{E}\left(\ \hat{h}_n(t) - h(t)\ \right)^2 \to_{n\to\infty} 0$$

Quand $b_n \to 0$ *et* $nb_n \to \infty$

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

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

Par la proposition 2.2 on a

$$\mathbb{E}\left(\ \hat{h}_n(t) - \tilde{h}_n(t)\ \right)^2 \to_{n\to\infty} 0.$$

Montrons que

$$\mathbb{E}\left(\tilde{h}_n(t) - h(t) \right)^2 \to_{n\to\infty} 0.$$

Comme $J_n(u) \to^P 1$ uniformément et la convergence dominée donne

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

Par suite : $|\tilde{h}_n(t) - h(t)| \leq C$ où $C > 0$. Donc par $\mathbb{E}(\hat{h}_n(t)) = \mathbb{E}(\tilde{h}_n(t))$ (proposition 2.1) et par $\mathbb{E}(\hat{h}_n(t)) \to h(t)$ (proposition 2.2)

$$\begin{aligned} \mathbb{E}\left(\tilde{h}_n(t) - h(t) \right)^2 &\leq C\mathbb{E}\left(\tilde{h}_n(t) - h(t) \right) = C\left(\mathbb{E}(\tilde{h}_n(t)) - h(t) \right) \\ &\leq C\left| \mathbb{E}(\tilde{h}_n(t)) - h(t) \right| \to_{n\to\infty} 0. \end{aligned}$$

D'où le résultat de la proposition.

Pour la convergence uniforme dans $\mathcal{L}^2$ 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 2.1.1. *Supposons que*

(i) $J_n \to_{n\to\infty} 1$ *en probabilité et uniformément sur* $[0, 1]$.

(ii) h *est continue sur* $[0, 1]$.

(iii) $n\eta_n(1) = n\int_0^1 \mathbb{E}\left(\frac{J_n(s)}{\bar{Y}_n(s)} \right) h(s)ds$ *est borné quand* $n \to \infty$

(iv) *le noyau* K *est à variations bornées.*

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

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

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

$$\mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) \to_{n\to\infty} 0. \tag{2.15}$$

Et

$$\mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \tilde{h}(t) - h(t) \right|^2 \right) \to_{n\to\infty} 0. \tag{2.16}$$

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

$$\left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \leq \frac{2}{b_n} Var(K) \sup_{s\in[0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|.$$

Où $Var(K)$ 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 la relation(2.15) il suffit de montrer que

$$\frac{1}{b_n^2} \mathbb{E}\left(\sup_{s\in[0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2 \right) \to_{n\to\infty} 0.$$

Si on applique l'inégalité de Doob (pour une martingale (M_t) :

$$\mathbb{E}(\sup_{s\leq t} M_s^2) \leq 4 \sup_{s\leq t} \mathbb{E}(M_s^2)$$

à la martingale $(\hat{H}_n(s) - \tilde{H}_n(s))$ et on obtient

$$\begin{aligned} \mathbb{E}\left(\sup_{s\in[0,1]} |\hat{H}_n(s) - \tilde{H}_n(s)|^2 \right) &\leq 4 \sup_{s\in[0,1]} \mathbb{E}\left(\hat{H}_n(s) - \tilde{H}_n(s) \right)^2 \\ &= 4 \sup_{s\in[0,1]} \eta_n(s) = 4\eta_n(1). \end{aligned}$$

Par suite

$$\begin{aligned} \mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) &\leq \frac{4}{b_n^2} C\mathbb{E}\left(\sup_{s\in[0,1]} \left| \hat{H}_n(s) - \tilde{H}_n(s) \right|^2 \right) \\ &\leq \frac{16}{nb_n^2} Cn\eta_n(1). \end{aligned}$$

Comme $\eta_n(1)$ est borné pour n assez grand et $nb_n^2 \to \infty$, on en déduit que

$$\mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \hat{h}_n(t) - \tilde{h}(t) \right|^2 \right) \to_{n\to\infty} 0.$$

Montrons maintenant la relation 2.16.

Dans la preuve de la proposition 2.4 on a montré qu'il existe une constante $C_n(t) > 0$ telle que $\left| \tilde{h}_n(t) - h(t) \right| \leq C_n(t)$ (où $C_n(t)$ est choisi petit pour n assez grand pour chaque t). Posons $C_n = \sup_{t\in[z_0,z_1]} C_n(t)$

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

D'où

$$\mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \tilde{h}_n(t) - h(t) \right|^2 \right) \leq C_n \mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \tilde{h}_n(t) - h(t) \right| \right) \to_{n\to\infty} 0.$$

Car

$$\sup_{t\in[z_0,z_1]} \left| \tilde{h}_n(t) - h(t) \right| \leq \sup_{t\in[z_0,z_1]} C_n \to_{n\to\infty} 0.$$

Et alors par la convergence dominée on a

$$\mathbb{E}\left(\sup_{t\in[z_0,z_1]} \left| \hat{h}_n(t) - h(t) \right|^2 \right) \to_{n\to\infty} 0.$$

D'où le résultat.

2.2 Données censurées

2.2.1 Introduction

Dans cette partie, nous allons donné les formes exactes et asymptotiques de la moyenne et la variance de l'estimateur à noyau de la fonction de risque h en développant les résultats de l'article de **Martin Tanner** et **Wing Hung Wong** (1983) [**?**].

Soient les durées de survie $T_1, T_2, ..., T_n$ de f.d.r F et de densité f et $C_1, C_2, ..., C_n$ les v.a de censures de f.d.r F_C. On suppose que les T_i sont indépendantes des C_i pour tout $i = 1, .., n$ et on observe $X_i = T_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$. 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

$$\hat{h}_n(t) = \sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} K_{b_n}(t - X_{(j)}) = \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-R_i+1} K_{b_n}(t - X_i).$$

Où $X_{(1)}, X_{(2)}, ..., X_{(n)}$ 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) et Watson et Leadbetter (1964).

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 $t \to \infty$ et $\int K(t)dt = 1$.

2.2.2 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 $\hat{h}_n$, on a besoin du lemme suivant

Lemme 2.2.1. *Pour tout j on a*

$$\mathbb{E}(\delta_{(j)}/X_{(j)} = y) = m(y).$$

Et pour tout $r < s$, $y < z$, on a

$$\mathbb{E}(\delta_{(r)}\delta_{(s)}/X_{(r)} = y, X_{(s)} = z) = m(y)m(z)$$

Le théorème suivant donne la moyenne et la variance de $\hat{h}_n(t)$

Théorème 2.2.1. *On a*

$$\mathbb{E}\left(\hat{h}_n(t) \right) = \int (1 - F_X^n(y))h(y)K_{b_n}(t-y)dy$$

Et

$$\begin{aligned} Var\left(\hat{h}_n(t) \right) &= \int I_n(F_X(y))h(y)K_{b_n}^2(t-y)dy + 2\int\int_{y\le z}\{F_X^n(z) \\ &- F_X^n(y)F_X^n(z) - \frac{1-F_X(y)}{F_X(z)F_X(y)}[F_X^n(z) - F_X^n(y)]\}h(y)h(z)K_{b_n}(t-y)K_{b_n}(t-z)dydz \end{aligned}$$

Où $I_n(F_X) = \sum_{k=0}^{n}(n-k)^{-1}C_n^k F_X^k(1-F)^{n-k}$

Preuve de lemme : Soit $y > 0$ et j fixé, comme $T_{(j)}$ sont indépendantes des $C_{(j)}$ on a ($\delta_i = \mathbb{1}_{T_i \le C_i}$) :

$$\begin{aligned}
\mathbb{E}(\delta_{(j)}/X_{(j)}=y) &= \mathbb{P}(\delta_{(j)}=1/X_{(j)}=y)\\
&= \mathbb{P}(T_j\leq C_j/X_{(j)}=y)\\
&= \frac{\mathbb{P}(T_j\leq C_j, X_{(j)}=y)}{\mathbb{P}(X_{(j)}=y)}\\
&= \frac{\mathbb{P}(T_{(j)}\leq C_{(j)}, T_{(j)}\wedge C_{(j)}=y)}{\mathbb{P}(X_{(j)}=y)}\\
&= \frac{\mathbb{P}(C_{(j)}\geq y, T_{(j)}=y)}{\mathbb{P}(X_{(j)}=y)} = \frac{\mathbb{P}(C_{(j)}\geq y)\mathbb{P}(T_{(j)}=y)}{\mathbb{P}(X_{(j)}=y)}\\
&= \frac{(1-F_C(y))f(y)dy}{f_X(y)dy} = \frac{(1-F_C(y))f(y)}{f_X(y)} = m(y)
\end{aligned}$$

Soit $r<s$, $y<z$, comme $T_{(i)}$ sont indépendantes des C_i pour $i=r,s$ on a

$$\begin{aligned}
\mathbb{E}(\delta_{(r)}\delta_{(s)}/X_{(r)}=y, X_{(s)}=z) &= \mathbb{P}(\delta_{(r)}=1,\delta_{(s)}=1/X_{(r)}=y, X_{(s)}=z)\\
&= \mathbb{P}(T_{(r)}\leq C_{(r)}, T_{(s)}\leq C_{(s)}/X_{(r)}=y, X_{(s)}=z)\\
&= \frac{\mathbb{P}(T_{(r)}\leq C_{(r)}, T_{(s)}\leq C_{(s)}, X_{(r)}=y, X_{(s)}=z)}{\mathbb{P}(X_{(r)}=y, X_{(s)}=z)}\\
&= \frac{\mathbb{P}(y\leq C_{(r)}, z\leq C_{(s)}, T_{(r)}=y, T_{(s)}=z)}{\mathbb{P}(X_{(r)}=y, X_{(s)}=z)}\\
&= \frac{\mathbb{P}(y\leq C_{(r)}, T_{(r)}=y)\mathbb{P}(z\leq C_{(s)}, T_{(s)}=z)}{\mathbb{P}(X_{(r)}=y, X_{(s)}=z)}\\
&= \frac{\mathbb{P}(y\leq C_{(r)})\mathbb{P}(T_{(r)}=y)}{\mathbb{P}(X_{(r)}=y)}\frac{\mathbb{P}(z\leq C_{(s)})\mathbb{P}(T_{(s)}=z)}{\mathbb{P}(X_{(s)}=z)}\\
&= m(y)m(z)
\end{aligned}$$

D'où les résultats du lemme.

Preuve de théorème : 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

$$\begin{aligned}
\mathbb{E}\left(\ \hat{h}_n(t)\ \right) &= \int \mathbb{E}\left(\ \hat{h}_n(t)/X_{(j)}=y\ \right) f_{X_{(j)}}(y)dy \\
&= \int \mathbb{E}\left(\ \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-i+1} K_{b_n}(t-X_{(i)})/X_{(j)}=y\ \right) f_{X_{(j)}}(y)dy \\
&= \sum_{i=1}^{n} \frac{1}{n-i+1} \int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})/X_{(j)}=y) f_{X_{(j)}}(y)dy \\
&= \sum_{i=1,i\neq j}^{n} \frac{1}{n-i+1} \int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})/X_{(j)}=y) f_{X_{(j)}}(y)dy \\
&+ \frac{1}{n-j+1} \int \mathbb{E}(\delta_{(j)} K_{b_n}(t-X_{(j)})/X_{(j)}=y) f_{X_{(j)}}(y)dy \\
&= \sum_{i=1,i\neq j}^{n} \frac{1}{n-i+1} \int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})/X_{(j)}=y) f_{X_{(j)}}(y)dy \\
&+ \frac{1}{n-j+1} \int m(y) K_{b_n}(t-y) f_{X_{(j)}}(y)dy \\
&= \sum_{i=1}^{n} \frac{1}{n-i+1} \int m(y) K_{b_n}(t-y) f_{X_{(i)}}(y)dy \\
&= \sum_{i=1}^{n} \frac{1}{n-i+1} \int m(y) \frac{n!}{(i-1)!(n-i)!} F_X^{i-1}(y)(1-F_X(y))^{n-i} f_X(y) K_{b_n}(t-y)dy \\
&= \int \left[\sum_{i=1}^{n} \frac{1}{n-i+1} \frac{n!}{(i-1)!(n-i)!} F_X^{i-1}(y)(1-F_X(y))^{n-i+1} \right] \\
&\quad \frac{f_X(y)m(y)}{(1-F_X(y))} K_{b_n}(t-y)dy \\
&= \int \left[\sum_{i=1}^{n} C_n^{i-1} F_X^{i-1}(y)(1-F_X(y))^{n-i+1} \right] \frac{f(y)(1-F_C(y))}{(1-F_X(y))} K_{b_n}(t-y)dy \\
&= \int \left[\sum_{j=0}^{n-1} C_n^{i} F_X^{i}(y)(1-F_X(y))^{n-j} \right] \frac{f(y)(1-F_C(y))}{(1-F_X(y))} K_{b_n}(t-y)dy \\
&= \int (1-F_X^n(y)) h(y) K_{b_n}(t-y)dy
\end{aligned}$$

Où dans le passage à la 6ème égalité on a utilisé pour $i \neq j$

$$\begin{aligned}
\int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})/X_{(j)}=y) f_{X_{(j)}}(y)dy &= \int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})) \\
&= \int \mathbb{E}(\delta_{(i)} K_{b_n}(t-X_{(i)})/X_{(i)}=y) f_{X_{(i)}}(y)dy \\
&= \int m(y) K_{b_n}(t-y) f_{X_{(i)}}(y)dy
\end{aligned}$$

Le calcul de la variance $Var\left(\hat{h}_n(t) \right)$ est plus long mais se fait de façon similaire.

Remarque 2.2.1. *Dans le théorème 2.2.1, la partie dominante de* $\mathbb{E}\left(\hat{h}_n(t) \right)$ *se comporte quand* $n \to \infty$ *comme la convolution* $(h * K_{b_n})(t)$*. En effet*

$$\begin{aligned}
\mathbb{E}(\hat{h}_n(t)) &= \int (1 - F_X^n(y))h(y)K_{b_n}(t-y)dy \\
&= \int h(y)K_{b_n}(t-y)dy - \int F_X^n(y)h(y)K_{b_n}(t-y)dy \\
&= (h * K_{b_n})(t) - \int F_X^n(y)h(y)K_{b_n}(t-y)dy \\
&\approx (h * K_{b_n})(t).
\end{aligned}$$

Et le 2ème 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*. Donc* $\mathbb{E}\left(\hat{h}_n(t) \right) \simeq h(t)$*. 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 2.2.1. *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 2.2.2. *Soit $b_n \to_{n\to\infty} 0$ et $nb_n \to_{n\to\infty} \infty$. Nous avons*

1. *Si K est compatible avec F, alors $\mathbb{E}(\hat{h}_n(t)) \to_{n\to\infty} h(t)$*
2. *Si K est compatible avec F et F_C, alors*

$$\mathrm{Var}\left(\ \hat{h}_n(t)\ \right) = \frac{1}{nb_n}\left(\ \int K^2(u)du\ \right) h(t)(1-F_X(t))^{-1} + o\left(\ \frac{1}{nb_n}\ \right).$$

Preuve : Pour la partie 1, nous avons par le théorème 2.2.1 :

$$\mathbb{E}\left(\ \hat{h}_n(t)\ \right) - h(t) = \int [(1-F_X^n(y))h(y) - h(t)]K_{b_n}(t-y)dy.$$

En utilisant que K est compatible avec F, on obtient pour $M > 0$

$$\begin{aligned}
\left|\ \mathbb{E}\left(\ \hat{h}_n(t)\ \right) - h(t)\ \right| &\leq \int |[(1-F_X^n(y))h(y) - h(t)]K_{b_n}(t-y)|dy \\
&\leq \int_{|y-t|>M} |[(1-F_X^n(y))h(y) - h(t)]K_{b_n}(t-y)|dy \\
&+ \int_{|y-t|\leq M} |[(1-F_X^n(y))h(y) - h(t)]K_{b_n}(t-y)|dy \\
&\leq \int_{|y-t|>M} \left|\ [(1-F_X^n(y))h(y) - h(t)](1-F_X(y))\frac{K_{b_n}(t-y)}{(1-F_X(y))}\ \right| dy \\
&+ \int_{|y-t|\leq M} |[(1-F_X^n(y))h(y) - h(t)]K_{b_n}(t-y)|dy \\
&\leq G_M \int_{|y-t|>M} |[(1-F_X^n(y))h(y) - h(t)](1-F_X(y))|dy \\
&+ \int_{|y-t|\leq M} |h(y)-h(t)|K_{b_n}(t-y)dy + \int_{|y-t|\leq M} F_X^n(y)h(y)K_{b_n}(t-y)dy \\
&\leq G_M \int_{|y-t|>M} |[(1-F_X^n(y))h(y) - h(t)]|dy \\
&+ \int_{|y-t|\leq M} |h(y)-h(t)|K_{b_n}(t-y)dy + \int_{|y-t|\leq M} F_X^n(y)h(y)K_{b_n}(t-y)dy \\
&= (1)+(2)+(3).
\end{aligned}$$

Chacun des trois termes tend vers 0. En effet :

– Pour le terme (1), en choisissant M assez grand on a par le théorème de convergence dominée

$$\int_{|y-t|>M} |[(1-F_X^n(y))h(y)-h(t)]|dy \leq \epsilon.$$

– Pour le terme (2), en posant $z=\frac{t-y}{b_n}$ et en appliquant le th. de C. D on a :

$$\int_{|y-t|\leq M} |h(y)-h(t)|K_{b_n}(t-y)dy = \int_{|z|\leq \frac{M}{b_n}} |h(t-zb_n)-h(t)|K(z)dz \to_{n\to\infty} 0.$$

– Pour le terme (3), en posant $z=\frac{t-y}{b_n}$ et en appliquant le th. de C. D on a :

$$\int_{|y-t|\leq M} F_X^n(y)h(y)K_{b_n}(t-y)dy = \int_{|z|\leq \frac{M}{b_n}} F_X^n(t-zb_n)h(t-zb_n)K(z)dz \to_{n\to\infty} 0.$$

$\Diamond$ Pour la deuxième partie du théorème, le lemme de Watson.Leadbetter [**?**] montre que pour $F(y)<1$

$$nI_n(F) \to_{n\to\infty} (1-F(y))^{-1}$$

$\Diamond$ Pour obtenir le premier terme de $Var\Big(\ \hat{h}_n(t)\ \Big)$ on applique le lemme de Bochner [**?**] pour le noyau $K^2/(\int K^2(t)dt)$ et on a

$$nb_n\Big(\ \int K^2(t)dt\ \Big)^{-1} \int I_n(F_X(y))h(y)K_{b_n}^2(t-y)dy \to_{n\to\infty} h(t)(1-F(x))^{-1}.$$

C'est aussi le terme dominant dans l'expression $nb_n\left(\int K^2(u)du\right)^{-1} Var\Big(\ \hat{h}_T(t)\ \Big)$.

$\Diamond$Le deuxième terme converge vers 0 à cause de $F_X^n \to_{n\to\infty} 0$ par les mêmes arguments que dans Watson. Leadbetter.

D'où le résultat du théorème.

Remarque 2.2.2. *1. Comme conséquense immédiate du théorème on a* $Var\left(\hat{h}_n(t) \right) \to_{n\to\infty} 0$ *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 2.2.2. On a :*

$$\left(\hat{h}_n(t) - \mathbb{E}(\hat{h}_n(t)) \right) Var^{1/2}\left(\hat{h}_n(t) \right) \to \mathcal{N}(0,1).$$

Chapitre 3

Application

L'objet de ce chapitre est de réanalyser un jeu de données. Pour illustrer notre travail, nous présentons des exemples d'applications simples qui a été également utilisé pour détailler les travaux de [16] [**?**]. Cette application repose sur des données réelles en utilisant des données "gastricXelox" et des données simulées dans le cas non censurés et censurés. Nous utilisons le logiciel R. Nous présentons les graphes des estimateurs de Kaplan-Meier et estimateur à noyau pour des données réelles. Nous donnons aussi les graphes des estimateurs à noyau de la fonction de risque.

3.1 Données réelles

Dans cette partie, nous faisons de l'analyse de survie sur des données réelles. Dans le package **R**, il existe une bibliothèque **"muhaz"** pour estimer et tracer fonctions de risque non paramétrique. Ce package doit être téléchargé et installé dans **R**.

3.1.1 Données "gastricXelox"

Il s'agit d'un essai clinique de phase II (échantillon unique) sur la chimiothérapie Xeloda et xaliplatine (XELOX) (indiqué dans le traitement du cancer colorectal métastatique) administrée avant la chirurgie à 48 patients atteints d'un cancer gastrique

avancé avec métastase ganglionnaire para-aortique. La survie sans progression, qui correspond au temps écoulé entre l'entrée dans l'essai clinique et la progression ou le décès, selon la première éventualité, est un critère de survie important. Les données sont dans le jeu de données "Xelox gastrique" dans le paquet "asaur", un échantillon des observations (pour les patients 23 à 27) sont les suivants :

```
> library (asaur)
> gastricXelox[23 :27,]
```

	timeWeeks	delta
23	42	1
24	43	1
25	43	0
26	46	1
27	48	0

Tout d'abord, divisons le temps en intervalles égaux de largeur 5 mois, et observons le nombre d'évènements (progression ou décès) d_i et le nombre de patients à risque à chaque intervalle, n_i, l'estimation du danger pour cet intervalle est $h_i = \frac{d_i}{n_i}$. L'estimation du risque à l'aide de cette méthode peut être obtenue à l'aide de la fonction **"pehaz"** :

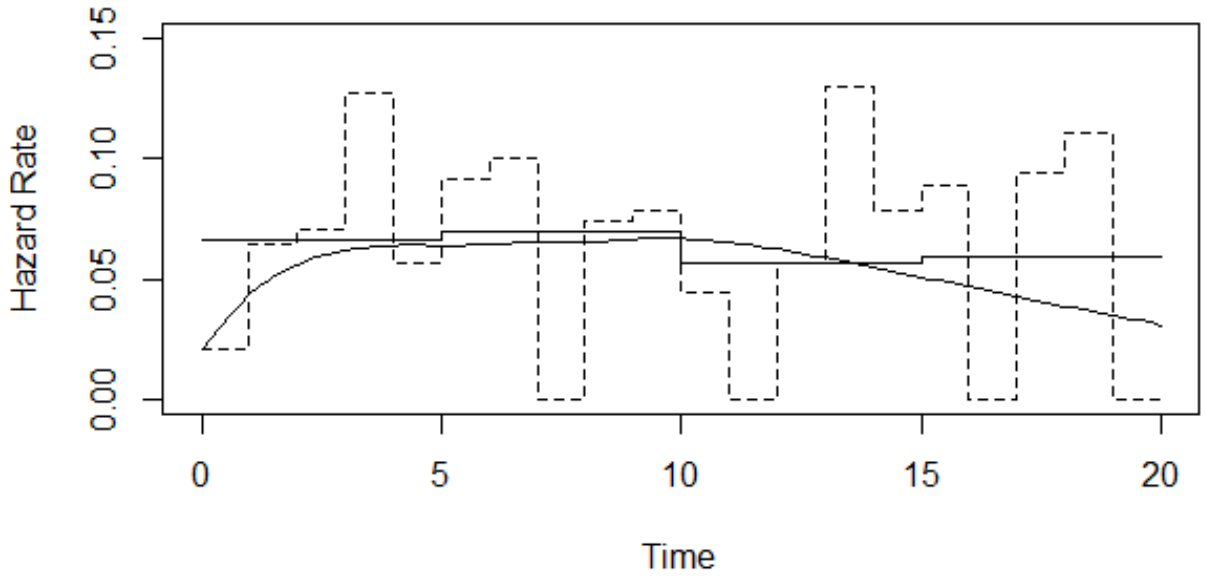
```
result.pe5 <- pehaz(timeMonths, delta, width=5, max.time=20)
plot(result.pe5, ylim=c(0,0.15), col="black")
```

Dans la même figure, nous présentons également la fonction escalier pour des intervalles d'un mois :

```
result.pe1 <- pehaz(timeMonths, delta, width=1, max.time=20)
lines(result.pe1)
```

Nous pouvons calculer une estimation régulière du fonction de risque en utilisant le code suivant :

```
result.smooth <- muhaz(timeMonths, delta, bw.smooth=20, b.cor="left", max.time=20)
```



lines(result.smooth))

Une utilisation de la fonction de risque consiste à obtenir une estimation régulière de la fonction de survie, en utilisant la relation $S(t) = e^{-\int_0^t h(u)du}$.

haz <- result.smooth$haz.est

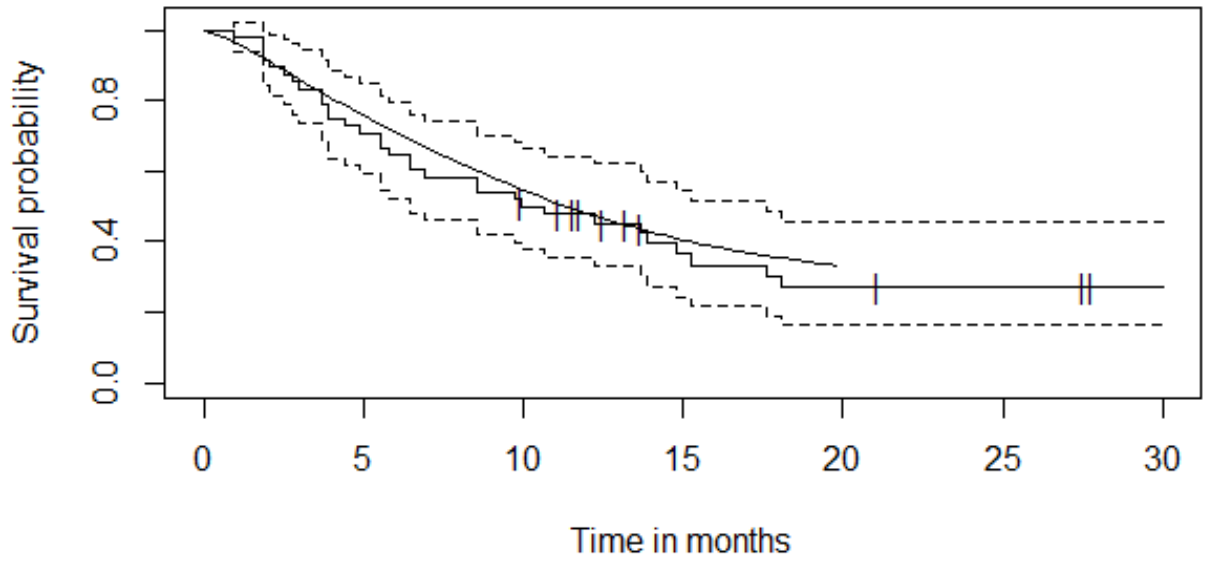
times <- result.smooth$est.grid

surv <- exp(-cumsum(haz[1 :(length(haz)-1)]*diff(times))) Nous pouvons comparer cet estimateur de survie par l'estimation de Kaplan-Meier comme suit :

result.km <- survfit(Surv(timeMonths, delta) 1, conf.type="none")

plot(result.km, conf.int=T, mark="|", xlab="Time in months", xlim=c(0,30), ylab="Survival probability")

lines(surv times[1 :(length(times) - 1)])



La fonction de risque lissée suit assez bien la courbe de survie. Seuls les 30 premiers mois sont affichés ici, car la procédure de lissage ne produit pas des estimations au-delà du dernier moment de risque. Alors que certaines applications spécialisées peuvent nécessiter une estimation de la courbe de survie lisse, la plupart des études publiées sur les données de survie préfèrent rapporter l'estimation de la fonction de survie de Kaplan-Meier. Cette estimation a la propriété théorique d'être l'estimation du maximum de vraisemblance de la fonction de survie. De plus, le trace de la fonction de survie est un affichage visuel efficace des données, en ce qu'il montre quand les risques et les temps de censure se sont produits.

3.2 Données simulées

Dans cette partie, nous utilisons l'estimateur à noyau pour estimer la fonction de risque. Nous choisissons une loi pour T : la loi de weibull $\mathcal{W}(1, 1.5)$ avec une taille d'échantillon $n = 50, 100$.

3.2.1 Cas non censurées

Les échantillons de tailles $n = 50$:

```
noy=function(x,t,bn)
i=1 :10 ; ksp=1/(10-i+1) ;
p=ksp[i]*dnorm((t-x[i])/bn)
noy=(1/bn)*sum(p)
x=rweibull(50,0.5,2)
s=seq(0,5,by=0.01)
n=vector(length=length(s))
for(i in
1 :length(s))n[i]=noy(x,s[i],0.1)
par(mfrow=c(2,2))
plot(s,n,col="red",main="bn=0.1",type="s")
curve(dweibull(x,0.5,2), add=T)
x=rweibull(50,0.5,2)
s=seq(0,5,by=0.01)
n=vector(length=length(s))
for(i in
1 :length(s))n[i]=noy(x,s[i],0.4)
plot(s,n,col="red",main="bn=0.4",type="s")
curve(dweibull(x,0.5,2), add=T)
x=rweibull(100,0.5,2)
s=seq(0,5,by=0.01)
```

```
n=vector(length=length(s))
for(i in_ 1 :length(s))n[i]=noy(x,s[i],0.1)
plot(s,n,col="red",main="bn=0.1",type="s")
curve(dweibull(x,0.5,2), add=T)
x=rweibull(100,0.5,2)
s=seq(0,5,by=0.01)
n=vector(length=length(s))
for(i in
1 :length(s))n[i]=noy(x,s[i],0.4)
plot(s,n,col="red",main="bn=0.4",type="s")
cudrve(dweibull(x,0.5,2), add=T)
```

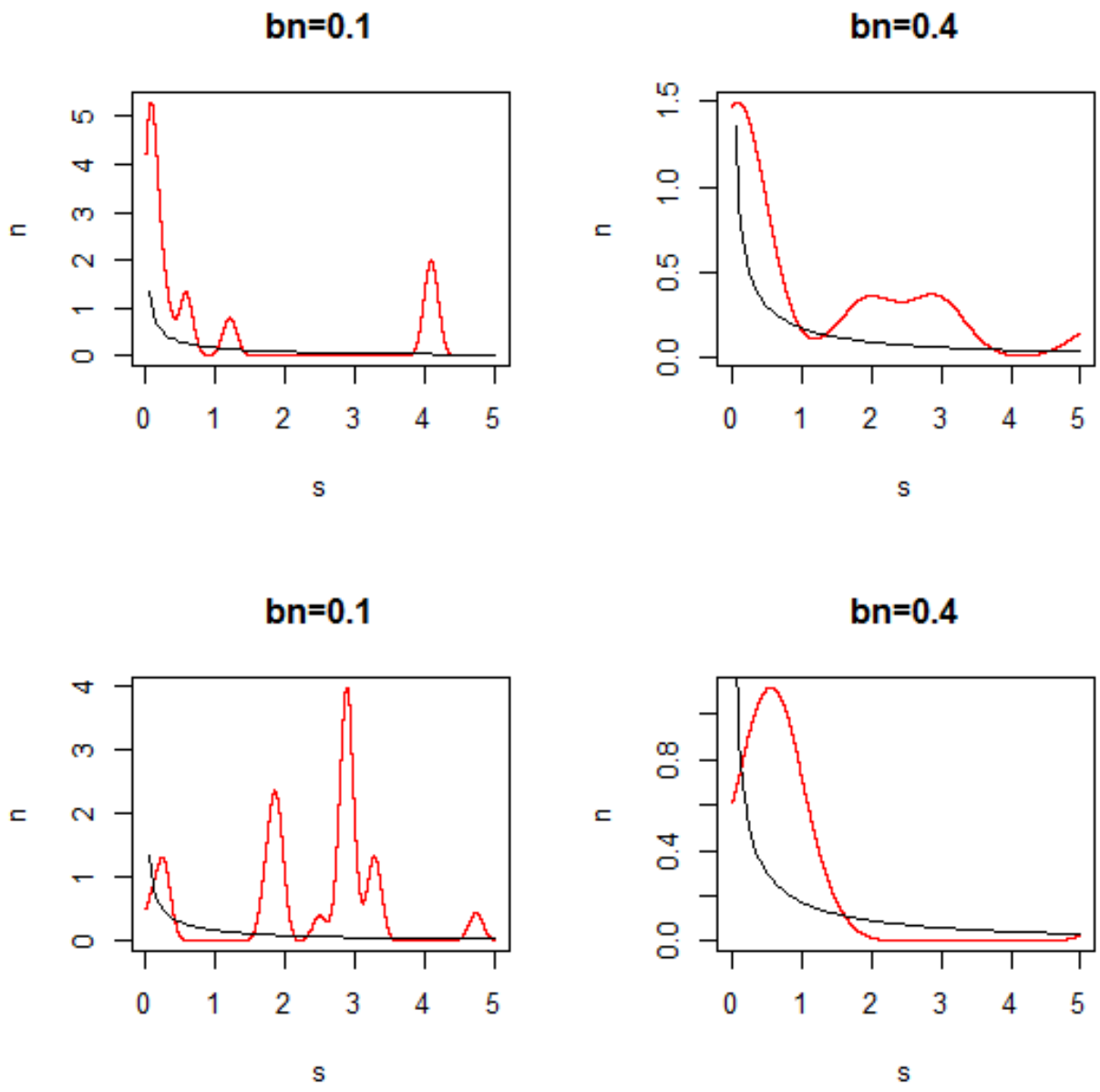


FIGURE 3.1 – Hasard estimé non censure ($n = 50, 100$)

3.2.2 Cas censurées

L'échantillon de taille $n = 50, 100, b_n = 0.1$

```
b=muhaz(x,d,bw.method="local",b.cor="both")
plot(b,col="red")
curve(dweibull(x,0.5,2)/(1-pweibull(x,0.5,2)),add=T)
legend("topright",legend = c("vraie","estimée"), col=1 :2,
lty = 1, lwd =2)
noy=function(x,d,t,bn)
i=1 :length(x) ; ksp=d[i]/(length(x)-i+1) ;
p=ksp[i]*dnorm((t-x[i])/bn)
noy=(1/bn)*sum(p)

d=sample(c(0,1),100, rep=T,prob=c(0.1,0.9))
x=rweibull(100,0.5,2)
s=seq(0,5,by=0.001)
n=vector(length=length(s))
for(i in
1 :length(s))n[i]=noy(x,d,s[i],0.1)
plot(s,n,col="red",type="s")
curve(dweibull(x,0.5,2)/(1-pweibull(x,0.5,2)),add=T)
library(muhaz)
b=muhaz(x,d,bw.method="local",b.cor="both")
plot(b,col="red")
curve(dweibull(x,0.5,2)/(1-pweibull(x,0.5,2)),add=T)
legend("topright",legend = c("vraie","estimée"), col=1 :2,
lty = 1, lwd =2)
```

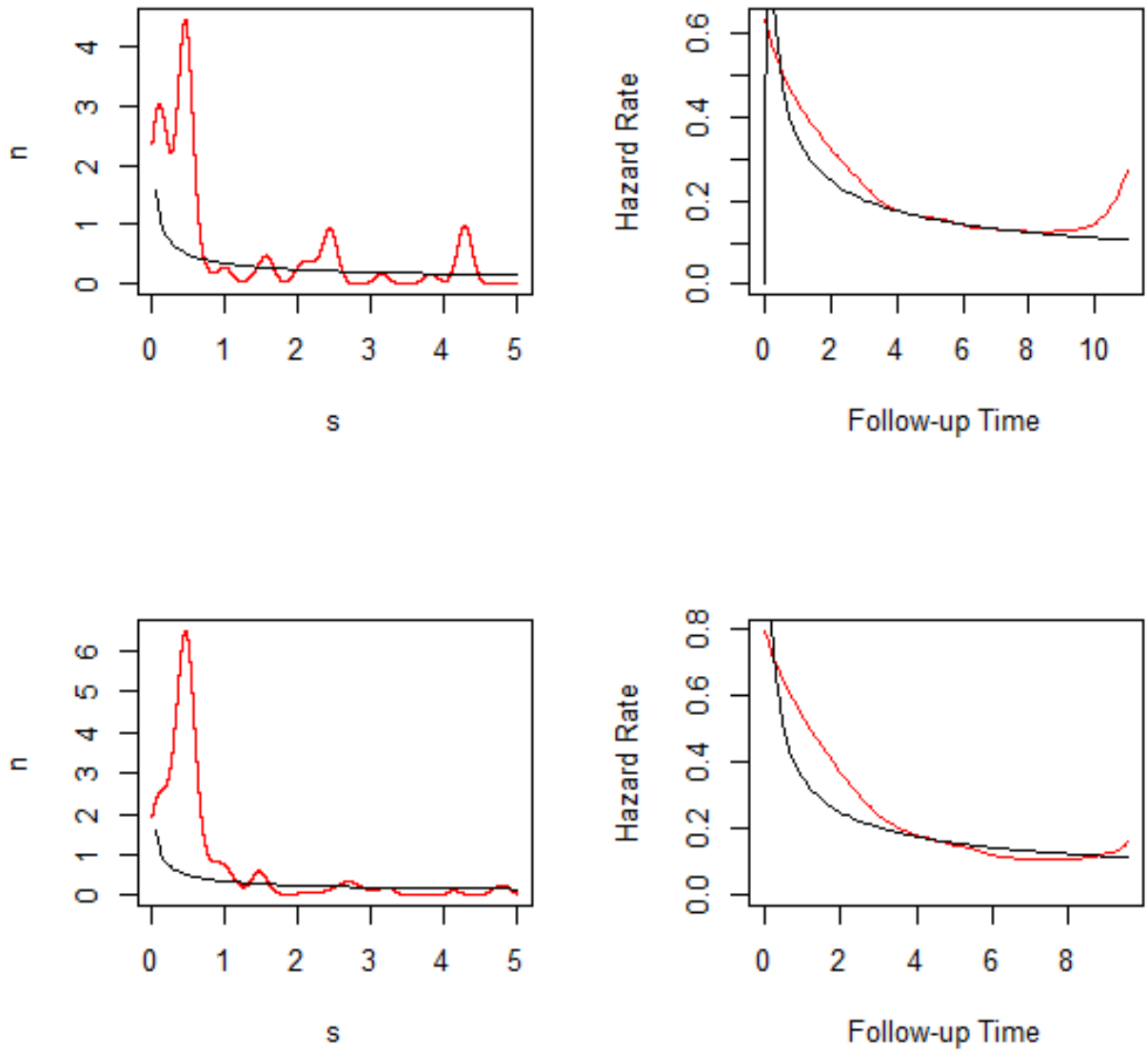


FIGURE 3.2 – Hasard estimé avec censure ($n = 50, 100, b_n = 0.1$)

On fait les même commandes pour l'échantillon de taille $n = 50, 100, b_n = 0.4$ et ces résultats représentent par le graphe 3.2.2.

Discussion :

Nous remarquons que la qualité de l'estimation par noyau dépend : le choix du noyau et du choix de la fenêtre : si elle est trop petite, le bruit sera élevée et l'estimation trop irrégulière. À l'inverse, le biais sera élevé, l'estimation sera trop lisse.

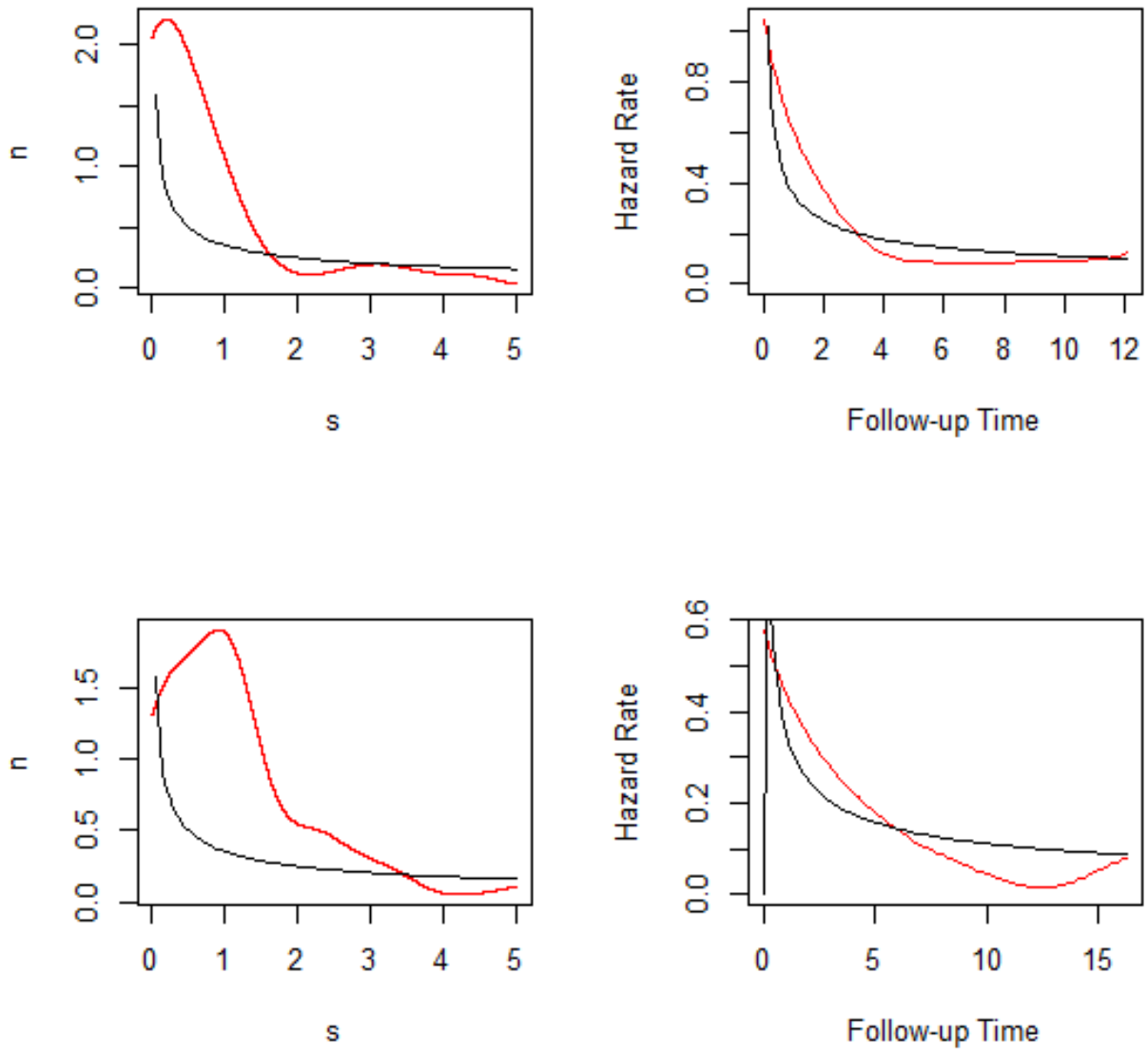


FIGURE 3.3 – Hasard estimé avec censure ($n = 50, 100, b_n = 0.4$)

Conclusion

Au cours de notre mémoire de fin d'études, 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. Puis, nous avons traité les techniques de l'estimation non paramétrique de la fonction de risque. Nous avons utilisé la méthode de noyau (Parzen-Rosenblatt) pour estimer la fonction de risque dans les deux cas non censures et censures. Cette méthode est particulièrement utilisée pour évaluer la performance de l'estimateur.

Pour bien présenter l'utilité des méthodes développées au cours de ce travail, nous les avons appliquées à un jeu de données réels médicales et simulés pour illustrer les comportements des estimateurs étudiés dans ce mémoire.

Bibliographie

[1] Aalen, O. O. Statistical inference for afamily of counting processes. (Thèse de doctorat). University of California, Berkeley (1975).

[2] Aalen, O. O. "Nonparametric inference in connection with multiple decrement models". Scand. J. Statist., 3, 15-27 (1976).

[3] Aalen, O. Nonparametric inference for a family of counting processes. The Annals of Statistics, 6(4), 701-726 (1978).

[4] Andersen P.K., Borgan O., Gill R. D. and Keiding N. Statistical Models Based on Counting Processes. Springer-Verlag. New York (1993).

[5] André Berchtold -R-bases : Les notions de base de R 2009-2014.

[6] Catherine Huber, Analyse des durées de survie : http :www.biomedicale.univ-paris5.fr/survie/enseign/survie-sansi.pdf

[7] Catherine Huber, Cours de Modélisation Biostatistique en Splus : http :www.biomedicale.univ-paris5.fr/survie/enseign/coursstat-avec-splus.pdf

[8] G. Colletaz. Modèles de survie. Notes de cours. Master 2 ESA (Novembre 2012). http ://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Survie-Sas.pdf.

[9] Henrik Ramlau Hansen "Smoothing counting process intensities by means of kernel func- tions". The Annals of Statictics 1983 Vol 11 n 2 p. 453-466.

[10] J. D. Fermanian. Modèles de durées. Cours ENSAE. http ://www.crest.fr/ck nder/user les/ les/Pageperso/fermania/JDF-duree3.pdf.

[11] Kaplan, E. L. and Meier, P. "Non-parametric estimation for incomplete observations".J.Amer. Statist. Assoc., 53,457-481 (1958).

[12] Martin A. Tanner. A Note on the Variable Kernel Estimator of the Hazard Function from Randomly Censored Data. The Annals of Statistics Vol. 11, No. 3 (Septembre 1983).

[13] Michel Fioc. Analyse de survie (Michel.Fioc@iap.fr,www2.iap.fr/users/ oc/enseignement/analyse de survie/)

[14] Nathalie Graffeo. Méthodes d'analyse de la survie nette : utilisation des tables de mortalité, test de comparaison et détection d'agrégats spatiaux. Aix-Marseille Université(thèse 2014). http :// http ://www.lsta.upmc.fr/psp/Cours Survie 1.pdf

[15] S. Ambapour, S. M. Ibara. Survie des enfants et pauvreté au Congo : application d'un modèle de durée. BAMSI B.P. 13734 Brazzaville(2012).

[16] Ségolen Geffray. Analyse des durées de vie avec le logiciel R. http ://iml.univ-mrs.fr/ reboul/R-survie.pdf

[17] Tanner M. A. and Wong W.H. "The estimation of the hasard function from randomly censored data by the kernel method". The Annals of Statictics 1983 Vol 11 n 33 p. 989- 993.

[18] Virginie Ehrlacher. Introduction aux problèmes d'évolution et espaces de Bochner. ENPC, IMI, Cours Problèmes d'évolution(07/04/2020).

[19] Watson, G. S. and Leadbetter, M. R. Hazard Analysis I. Biometrika, 51, 175-184 (1964 a).

[20] Watson, G. S. and Leadbetter, M. R. Hazard Analysis II. Sankhya Ser. A, 26, 101-116 (1964 b).

[21] Zhang Biao "Some asymptotic results for kernel estimation under random censorship". Bernoulli, p. 183-198 (2 (2) 1996).