

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée



MODÈLES DE DURÉE

Support de cours 2021-2022

Statistique des modèles non paramétriques

Frédéric PLANCHET

Version 4.6

Août 2021

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

SOMMAIRE

1. Introduction	3
2. Modèles de durée et processus ponctuels	3
2.1. Rappels sur les martingales	4
2.2. Application aux modèles de durée.....	5
3. Les estimateurs non paramétriques dans les modèles de durée	8
3.1. L'estimateur de Nelson-Aalen du taux de hasard cumulé	8
3.1.1. Présentation générale.....	8
3.1.2. Variance de l'estimateur de Nelson-Aalen.....	10
3.1.3. Un exemple.....	10
3.1.4. Propriétés asymptotiques	12
3.2. L'estimateur de Kaplan-Meier de la fonction de survie	12
3.2.1. Présentation générale.....	13
3.2.2. Comparaison avec l'estimateur de Harrington et Fleming.....	14
3.2.3. Exemples.....	15
3.2.4. Principales propriétés	16
3.2.5. Variance de l'estimateur de Kaplan Meier	17
3.2.6. Propriétés asymptotiques	18
3.2.7. Version discrétisée : lien avec l'approche paramétrique	19
4. Prise en compte de variables explicatives.....	20
4.1. Le modèle additif d'Aalen.....	20
4.2. Variante semi-paramétrique : le modèle de Lin et Ying	22
5. Comparaison d'échantillons : approche non paramétrique.....	23
5.1. Rappel : principe des tests de rang	23
5.2. Adaptation des tests de rang au cas censuré	23
5.2.1. Le test du log-rank.....	24
5.2.2. Le test de Gehan.....	25
5.2.3. Exemple : application aux données de Freireich.....	25
5.3. Approche par les processus ponctuels	27
6. Références.....	28

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

1. Introduction

On peut souhaiter, dans un certain nombre de situations, ne pas faire d'hypothèse *a priori* sur la forme de la loi de survie ; on cherche donc à estimer directement cette fonction, dans un espace de dimension infinie ; ce cadre d'estimation fonctionnelle est le domaine de l'estimation non paramétrique.

Sous réserve de disposer de données en quantités suffisantes, on peut alors obtenir des estimations fiables de la fonction de survie et des fonctionnelles associées.

Dans le contexte usuel d'un échantillon i.i.d. non censuré (T_1, \dots, T_n) , on dispose de l'estimateur empirique de la fonction de répartition $F_n(t) = \frac{1}{n} \sum_{i=1}^n I_{\{T_i \leq t\}}$. Cet estimateur possède un certain nombre de « bonnes propriétés » bien connues : il est sans biais, convergent et asymptotiquement gaussien. Plus précisément, la convergence est uniforme au sens presque sûr, et on a le « théorème central limite » suivant :

$$\sqrt{n}(F_n - F) \rightarrow W$$

où W est un processus gaussien centré de covariance $\rho(s, t) = F(s) \wedge F(t) - F(s)F(t)$. Ce résultat découle directement du théorème de Donsker dans le cas de la loi uniforme¹ et du fait que $F(T)$ suit une loi uniforme sur $[0, 1]$.

L'objectif de l'estimation empirique dans les modèles de durée est de rechercher un estimateur vérifiant des propriétés équivalentes en présence de censure. Pour ce faire, on commence par introduire la présentation des modèles de durée à partir de processus ponctuels, qui facilite ensuite l'obtention d'un certain nombre de résultats via les résultats limite sur les martingales.

Dans la suite on note F la fonction de répartition du modèle non censuré, G la fonction de répartition de la censure et $T = X \wedge C$ la variable censurée. On note également :

$$\begin{aligned} S_0(t) &= P(T > t, D = 0), \quad S_1(t) = P(T > t, D = 1) \text{ et} \\ S_c(t) &= S_0(t) \times S_1(t) = P(T > t) = (1 - F(t))(1 - G(t)). \end{aligned}$$

2. Modèles de durée et processus ponctuels

L'étude d'une durée de survie s'effectue en général en étudiant la loi de la variable X , associée à la fonction de survie S . On se propose ici de raisonner différemment et de considérer le processus ponctuel naturellement associé à X , $N(t)$, égal à 0 tant que l'événement n'a pas eu lieu, puis 1 après : $N(t) = I_{\{X \leq t\}}$. Lorsque l'on prend en compte la censure, on construit de même $N^1(t) = I_{\{T \leq t, D = 1\}}$ le processus des sorties non censurées².

¹ Le processus limite étant alors le pont brownien, processus gaussien centré de covariance $s \wedge t - st$.

² On reprend les notations du support sur les modèles paramétriques, avec X la variable non censurée, et le couple (T, D) en situation de censure droite.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

La présentation faite ici est heuristique et a pour ambition de faire comprendre les mécanismes en jeu. Le lecteur intéressé par la formalisation mathématique rigoureuse des outils évoqués pour se reporter à l'article de GILL [1980] ou à l'ouvrage de FLEMING et HARRINGTON [1991], ou encore pour une présentation en français à DACUNHA-CASTELLE et DUFLO [1983].

Cette approche fait largement appel à la théorie des martingales, dont les résultats essentiels sont rappelés ci-après.

2.1. Rappels sur les martingales

On dit qu'un processus (M_t) adapté à une filtration $(F_t)_{t \geq 0}$ est une martingale s'il est à trajectoire continues à droites avec des limites à gauche (càd-làg), et vérifie :

$$E(|M_t|) < \infty \quad \forall t \geq 0 \text{ et } E(M_t | F_s) = M_s \quad \forall s \leq t.$$

Une martingale peut être vue comme un processus d'erreurs, au sens où d'une part son espérance est constante (on pourra donc toujours supposer qu'elle est nulle) et d'autre part les incrémentés d'une martingale sont non corrélés :

$$\text{cov}(M_t - M_s, M_v - M_u) = 0, \quad 0 \leq s \leq t \leq u \leq v.$$

Si la condition de constance de l'espérance conditionnelle est affaiblie et que le processus est croissant en espérance conditionnelle au sens où $E(M_t | F_s) \geq M_s \quad \forall s \leq t$, on dit que M est une sous-martingale. Par l'inégalité de Jensen, si M est une martingale alors M^2 est une sous-martingale puisque $E(M_t^2 | F_s) \geq (E(M_t | F_s))^2 = M_s^2 \quad \forall s \leq t$.

Afin de poursuivre la formalisation, il est nécessaire d'introduire une nouvelle définition :

Définition : Un processus prévisible est une variable aléatoire mesurable définie sur l'espace produit $([0, +\infty] \times \Omega, \mathcal{P})$ muni de la tribu \mathcal{P} engendrée par les ensembles de la forme $[s, t] \times \Gamma$, avec $\Gamma \in \mathcal{F}_s$.

La tribu des événements prévisibles est engendrée par les processus adaptés à la filtration $(F_{t-})_{t \geq 0}$ avec $F_{t-} = \bigvee_{s < t} F_s$ et à trajectoires continues à gauche.

De manière intuitive, on peut dire qu'un processus prévisible est un processus dont la valeur en t est connue « juste avant » t . Ainsi un processus continu à gauche (et adapté) est prévisible du fait de la propriété de continuité.

Ces différents outils conduisent à la décomposition de Doob-Meyer d'un processus X càd-làg adapté³, qui exprime qu'un tel processus est la différence de deux sous-martingales (locales) si et seulement si il existe une unique décomposition de X sous la forme $X = A + M$ avec A un processus prévisible à variation bornée (au sens où

³ Voir par exemple DACUNHA-CASTELLE et DUFLO [1983].

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

$\int_0^t |dA_s| = \sup_D \sum |A_{t_i} - A_{t_{i-1}}| < \infty$ avec D l'ensemble des subdivisions de $[0, t]$) et M une martingale (locale) centrée.

On en déduit en particulier que si M est une martingale, M^2 possède un compensateur prévisible, que l'on note $\langle M \rangle$ (que l'on prendra garde de ne pas confondre en général avec la variation quadratique $[M]$).

2.2. Application aux modèles de durée

Rappelons la définition d'un processus ponctuel :

Définition : un processus ponctuel $(N(t), t \geq 0)$ est un processus à valeurs entières adapté à une filtration $(F_t)_{t \geq 0}$ tel que $N(0) = 0$, $N(t) < \infty$ presque sûrement et tel que les trajectoires soient continues à droite, constantes par morceaux et ne présentent que des sauts d'amplitude +1. En pratique on considérera souvent pour $(F_t)_{t \geq 0}$ la filtration naturelle associée à N , soit $F_t = \sigma(N(u), 0 \leq u \leq t) \vee N$ avec N les événements P -négligeables.

Le processus de Poisson fournit un exemple de processus ponctuel ; le processus $N(t)$ introduit ci-dessus est un cas simple dans lequel le processus ne saute qu'une fois.

Les processus ponctuels sont à trajectoires positives et croissantes, donc à variation bornée, et on peut alors définir pour un processus adapté $X(t)$ l'intégrale $\int_0^t X(u) dN(u)$ comme une intégrale de Stieltjes, trajectoire par trajectoire. Par exemple, en présence de censure le processus d'événements non censurés $N^1(t) = 1_{\{T \leq t, D=1\}}$ peut s'écrire :

$$N^1(t) = \int_0^t C(u) dN(u)$$

avec $C(u) = 1_{[0, C]}(s)$. La censure agit donc comme un filtre. Comme un processus ponctuel est une sous-martingale (puisque il est croissant), on lui associe son compensateur prévisible, qui est donc un processus prévisible croissant, de sorte que la différence entre le processus ponctuel et son compensateur soit une martingale. De manière plus formelle on a le résultat suivant :

Proposition : Si un processus ponctuel $(N(t), t \geq 0)$ adapté à la filtration $(F_t)_{t \geq 0}$ est tel que $E[N(t)] < \infty$, alors il existe un unique processus croissant continu à droite Λ tel que $\Lambda(0) = 0$, $E[\Lambda(t)] < \infty$ et $M(t) = N(t) - \Lambda(t)$ est une martingale.

Lorsque Λ peut se mettre sous la forme $\Lambda(t) = \int_0^t \lambda(u) du$, le processus λ s'appelle l'intensité du processus ponctuel. Par exemple le compensateur d'un processus de Poisson

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

homogène est $\Lambda(t) = \lambda t$, ou, de manière équivalente, l'intensité d'un processus de Poisson homogène est constante égale à λ .

D'un point de vue heuristique, la décomposition $N(t) = \Lambda(t) + M(t)$ exprime que le processus N « oscille » autour de la tendance prévisible Λ de sorte que la différence entre le processus d'intérêt N et sa tendance soit assimilable à un résidu, dont on maîtrise les variations. L'équation $N(t) = \Lambda(t) + M(t)$ peut ainsi se lire comme « observations = modèle + terme d'erreur ». On a en particulier $E(N_t) = E(\Lambda_t)$.

On cherche maintenant à déterminer le compensateur prévisible du processus $N(t) = 1_{\{X \leq t\}}$.

On note $N(t-) = \lim_{u \uparrow t} N(u)$ la limite à gauche de $N(t)$ et on s'intéresse à la loi de la variable aléatoire $P(dN_t = 1 | N(t-))$, en ayant noté formellement $dN_t = N(t+dt) - N(t)$ avec dt « petit ». La variable aléatoire dN_t ne peut prendre que les valeurs 0 et 1. Par définition de la fonction de survie et de la fonction de hasard, on a :

$$P(dN_t = 1 | N(t-)) = h(t)dt \text{ avec la probabilité } S(t)$$

et

$$P(dN_t = 1 | N(t-)) = 0 \text{ avec la probabilité } 1 - S(t).$$

En effet, si $N(t-) = 1$, la sortie s'est déjà produite et le processus ne peut plus sauter. Cet événement se produit avec la probabilité $1 - S(t)$. Le processus N ne peut sauter entre t et $t+dt$ que si $N(t-) = 0$ (événement de probabilité $S(t)$) et la probabilité de saut est $h(t)dt$. On pose alors $\lambda(t) = h(t)1_{\{X \geq t\}}$, produit de la fonction de hasard en t et de l'indicatrice de présence juste avant t , $Y(t) = 1_{\{X \geq t\}}$. Le processus $\lambda(t)$ est prévisible et $Y(t) = 1$ est équivalent à $N(t-) = 0$. Donc $P(dN_t = 1 | N(t-)) = \lambda(t)dt$, ou encore de manière équivalente $E(dN_t | N(t-)) = \lambda(t)dt$. Les remarques ci-dessus impliquent que :

$$M(t) = N(t) - \int_0^t \lambda(u)du = N(t) - \int_0^t h(u)Y(u)du = N(t) - H(t \wedge X)$$

est une martingale centrée puisque $E(dM_t | N(t-)) = 0$ et que l'intensité de processus N peut se calculer selon :

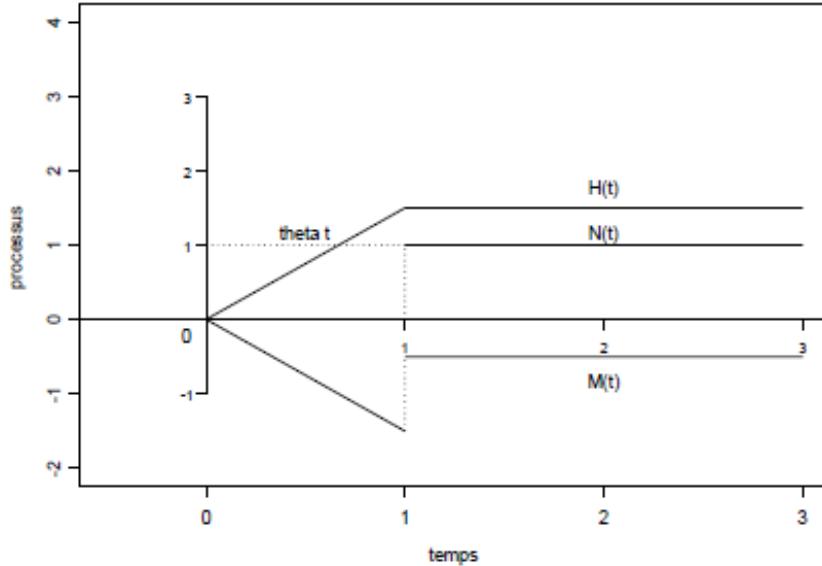
$$\lambda(t) = \lim_{u \rightarrow 0^+} \frac{1}{u} P[N(t+u) - N(t) = 1 | F_{t-}]$$

Le processus $\lambda(t)$ est donc l'intensité de processus $N(t)$, qui est aléatoire. Conditionnellement au « passé immédiat », l'accroissement de $N(t)$ entre t et $t+dt$ suit donc une loi de Bernoulli de paramètre $\lambda(t)dt$.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Modèles de durée

À titre d'illustration, on trouve, dans le cas d'une loi exponentielle les allures suivantes de N , M et H :



On peut montrer de même que le compensateur prévisible du processus d'évènements non censurés $N^1(t) = 1_{\{T \leq t, D=1\}}$ s'écrit :

$$\Lambda^1(t) = \int_0^t R(u) h(u) du,$$

avec $R(t) = 1_{\{T \geq t\}}$ l'indicatrice de présence à risque avant t (i.e. la fonction valant 1 si l'individu n'est ni mort ni censuré ; on rappelle en effet que comme $T = X \wedge C$, $\{T \geq t\} = \{X \geq t, C \geq t\}$). On est donc passé du modèle statistique où l'on se donnait le couple (T, D) comme informations observées au modèle composé de (N^1, R) .

Dans le cas d'une population, dont on suppose que tous les individus ont la même fonction de hasard h , on associe à chaque membre de la population un processus d'évènement non censuré $N_i^1(t) = 1_{\{T_i \leq t, D_i=1\}}$ ainsi que l'indicatrice de présence sous risque, comptabilisant les individus ni morts ni censurés $R_i(t) = 1_{\{T_i \geq t\}}$ et on construit les processus agrégés

$$\bar{R}(t) = \sum_{i=1}^n R_i(t) \text{ et } \bar{N}^1(t) = \sum_{i=1}^n N_i^1(t).$$

Ils comptabilisent respectivement l'effectif sous risque

et le nombre d'évènements survenus non censurés.

On se trouve donc en présence d'un modèle à « intensité multiplicative » (AALEN [1978]), en ce sens que le processus de comptage \bar{N}^1 possède une intensité qui se met sous la forme :

$$\lambda(t) = \bar{R}(t)h(t)$$

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

ressources-actuarielles.net

Modèles de durée

avec \bar{R} un processus observable (prévisible) et h la fonction de hasard, inconnue et à estimer. Ces processus vont permettre d'introduire simplement les estimateurs non paramétriques usuels.

3. Les estimateurs non paramétriques dans les modèles de durée

On notera en préambule que la distribution peut être, comme on l'a vu, caractérisée par différentes fonctions : fonction de hasard, fonction de hasard cumulée, fonction de répartition, densité... Il est évident que l'estimation de la fonction de hasard est du même degré de complexité que l'estimation de la densité ; on se tournera donc de manière privilégiée vers l'estimation empirique du hasard cumulé ou de la fonction de survie, *a priori* plus simple. L'estimation de la fonction de hasard nécessitera alors de régulariser l'estimateur de la fonction de hasard cumulée, qui sera en général discontinu. Ces aspects ne sont pas abordés ici⁴. Les deux estimateurs principaux dans ce contexte sont l'estimateur de Nelson-Aalen du taux de hasard cumulé et l'estimateur de Kaplan-Meier de la fonction de survie.

3.1. L'estimateur de Nelson-Aalen⁵ du taux de hasard cumulé

On rappelle que la fonction de hasard cumulée est définie, dans le cas général, par $H(t) = \int_0^t \frac{dS(u)}{S(u-)}$, expression qui conduit à l'expression classique dans le cas d'un modèle continu $H(t) = \int_0^t h(u)du$ où $h(t) = -\frac{d}{dt} \ln S(t)$.

3.1.1. Présentation générale

Le fait que $M(t) = \bar{N}^1(t) - \int_0^t \bar{R}(u)h(u)du$ soit une martingale centrée suggère de proposer

$\bar{N}^1(t)$ comme estimateur de $\int_0^t \bar{R}(u)h(u)du$. Mais alors le processus $\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} dM(u)$ est

également une martingale et on a par construction de M :

$$\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} dM(u) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) - \int_0^t h(u)du = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) - H(t)$$

pour autant que t soit tel que $\bar{R}(t) > 0$. Ainsi $\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u)$ est un estimateur naturel de H . Cet estimateur s'appelle l'estimateur de Nelson-Aalen. Il a été proposé

⁴ Le lecteur intéressé pourra consulter DROESBEKE et al. [1989].

⁵ L'étude originale de Nelson-Aalen porte sur la durée de fonctionnement de ventilateurs.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Modèles de durée

initialement par NELSON [1972]. On peut en donner une autre justification, en remarquant que la fonction de hasard cumulé vérifie, par construction :

$$H(u+du) - H(u) \approx h(u)du$$

et $h(u)du = P(\text{sortie entre } u \text{ et } u+du \mid \text{en vie en } u)$; un estimateur naturel de cette quantité est donc $\frac{\bar{N}^1(u+du) - \bar{N}^1(u)}{\bar{R}(u)} = \frac{d\bar{N}^1(u)}{\bar{R}(u)}$ si $\bar{R}(u) > 0$, de sorte qu'en sommant sur un découpage de $[0, t]$ suffisamment fin pour chaque subdivision contienne au plus un saut on obtient :

$$\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u),$$

ce qui est bien l'expression précédente. Comme les processus considérés ici sont purement à sauts on peut, en notant $\Delta\bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$, mettre cette expression sous la forme :

$$\hat{H}(t) = \sum_{\{i/T_i \leq t\}} \frac{\Delta\bar{N}^1(T_i)}{\bar{R}(T_i)}$$

En posant $d(t) = \Delta\bar{N}(t)$ le nombre de décès en t et $r(t) = \bar{R}(t)$ l'effectif sous risque juste avant t , on peut ainsi réécrire l'équation ci-dessus sous la forme intuitive suivante :

$$\hat{H}(t) = \sum_{\{i/T_i \leq t\}} \frac{d(T_i)}{r(T_i)} = \sum_{\{i/T_i \leq t\}} \frac{d_i}{r_i} = \sum_{T_i \leq t} \frac{d_i}{n-i+1},$$

la seconde égalité n'étant vrai que s'il n'y a pas d'*ex-æquo*. La fonction \hat{H} est continue à droite. On peut vérifier que cet estimateur est biaisé et sous-estime en moyenne la fonction de hasard cumulée. En effet,

$$\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} (dM(u) + \bar{R}(u)h(u)du).$$

Comme M est une martingale, il vient en prenant l'espérance des deux membres de l'équation ci-dessus $E[\hat{H}(t)] = \int_0^t E[1_{\{\bar{R}(u)>0\}}]h(u)du$. Mais :

$$E[1_{\{\bar{R}(u)>0\}}] = P[\bar{R}(u) > 0] = 1 - P[\bar{R}(u) = 0].$$

On en déduit finalement :

$$E[\hat{H}(t)] = \int_0^t h(u)du - \int_0^t P[\bar{R}(u) = 0]h(u)du = H(t) - \int_0^t P[\bar{R}(u) = 0]h(u)du$$

ce qui implique que $E[\hat{H}(t)] \leq H(t)$: l'estimateur de Nelson-Aalen a bien tendance à sous-estimer la fonction de hasard cumulée du modèle.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

3.1.2. Variance de l'estimateur de Nelson-Aalen

Il résulte de l'approximation effectuée à la section précédente que l'accroissement du processus $\bar{N}^1(t)$ entre t et $t+u$ suit approximativement une loi de Poisson de paramètre $\int_t^{t+u} \bar{R}(s)h(s)ds \approx \bar{R}(t)h(t)u$. En effet, on avait vu que conditionnellement au « passé immédiat », l'accroissement de $N^1(t)$ entre t et $t+dt$ suit donc une loi de Bernouilli de paramètre $q = h(t)\bar{R}(t)dt$. Comme q est petit, on peut utiliser, en choisissant $dt = \frac{u}{n}$, l'inégalité de Le Cam⁶ pour en déduire que la somme sur les différents individus conduit donc à une loi de Poisson de paramètre $\lambda = h(t)\bar{R}(t) \times u$. On en déduit donc que,

conditionnellement à $\bar{R}(t)$, $V\left(\frac{\bar{N}^1(t+u) - \bar{N}^1(t)}{\bar{R}(t)}\right) \approx \frac{h(t)u}{\bar{R}(t)}$; or on a vu à la section précédente que $h(t)u$ pouvait être estimé par $\frac{\bar{N}^1(t+u) - \bar{N}^1(t)}{\bar{R}(t)}$, d'où l'estimateur de la variance $\hat{V}\left(\frac{\bar{N}^1(t+u) - \bar{N}^1(t)}{\bar{R}(t)}\right) \approx \frac{\bar{N}^1(t+u) - \bar{N}^1(t)}{\bar{R}(t)^2}$, qui conduit finalement à proposer comme estimateur de la variance de \hat{H} :

$$\hat{V}(\hat{H}(t)) = \sum_{\{i | T_i \leq t\}} \frac{\Delta \bar{N}^1(T_i)}{\bar{R}(T_i)^2}$$

qui peut s'écrire avec les notations simplifiées, en l'absence d'*ex aequo* :

$$\hat{V}(\hat{H}(t)) = \sum_{\{i | T_i \leq t\}} \frac{d(T_i)}{(n-i+1)^2}.$$

On peut observer que cet estimateur est obtenu en utilisant comme estimateur de la variance d'un accroissement de la fonction de hasard cumulée $\frac{\Delta \bar{N}^1(T_i)}{\bar{R}(T_i)^2}$

3.1.3. Un exemple

Freireich, en 1963, a fait un essai thérapeutique pour comparer les durées de rémission, en semaines, de patients atteints de leucémie selon qu'ils ont reçu ou non un médicament appelé 6 M-P ; le groupe témoin a reçu un placebo. Les résultats obtenus sont les suivants⁷ :

6 M-P : 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+.

⁶ https://fr.wikipedia.org/wiki/In%C3%A9galit%C3%A9_de_Le_Cam

⁷ Durée de rémission, en semaines.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

ressources-actuarielles.net

Modèles de durée

Placebo : 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

Les nombres suivis du signe + correspondent à des données censurées. L'application des formules ci-dessus à ces données conduit à :

Tab. 1. Calcul de l'estimateur (groupe 6 M-P)

Rechutes	t_i	r_i	d_i	$\frac{d_i}{r_i}$	$\hat{H}(t)$	$\frac{d_i}{r_i^2}$	$\sigma^2(\hat{H}(t))$	$\sigma(\hat{H}(t))$
1-2-3	6	21	3	0,143	0,143	0,007	0,007	0,082
5	7	17	1	0,059	0,202	0,003	0,010	0,101
7	10	15	1	0,067	0,268	0,004	0,008	0,089
10	13	12	1	0,083	0,352	0,007	0,011	0,107
11	16	11	1	0,091	0,443	0,008	0,015	0,123
15	22	7	1	0,143	0,585	0,020	0,029	0,169
16	23	6	1	0,167	0,752	0,028	0,048	0,220

pour le groupe traité avec 6 M-P et pour le groupe traité avec le placebo on obtient :

Tab. 2. Calcul de l'estimateur (groupe placebo)

Rechutes	t_i	r_i	d_i	$\frac{d_i}{r_i}$	$\hat{H}(t)$	$\frac{d_i}{r_i^2}$	$\sigma^2(\hat{H}(t))$	$\sigma(\hat{H}(t))$
1-2	1	21	2	0,095	0,095	0,005	0,005	0,067
3-4	2	19	2	0,105	0,201	0,006	0,010	0,100
5	3	17	1	0,059	0,259	0,003	0,014	0,116
6-7	4	16	2	0,125	0,384	0,008	0,021	0,146
8-9	5	14	2	0,143	0,527	0,010	0,032	0,178
10-11-12-13	8	12	4	0,333	0,861	0,028	0,059	0,244
14-15	11	8	2	0,250	1,111	0,031	0,091	0,301
16-17	12	6	2	0,333	1,444	0,056	0,146	0,382
18	15	4	1	0,250	1,694	0,063	0,209	0,457
19	17	3	1	0,333	2,027	0,111	0,320	0,565
20	22	2	1	0,500	2,527	0,250	0,570	0,755
21	23	1	1	1,000	3,527	1,000	1,570	1,253

On constate notamment que le taux de hasard cumulé du groupe traité est sensiblement inférieur à celui du groupe non traité, ce qui laisse supposer une certaine efficacité du traitement. Ce point sera repris *infra*.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

ressources-actuarielles.net

Modèles de durée

3.1.4. Propriétés asymptotiques

L'estimateur de Nelson-Aalen est asymptotiquement gaussien ; plus précisément on a le résultat suivant :

Proposition : si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sqrt{n}(\hat{H} - H) \rightarrow W_H$$

avec W_H un processus gaussien centré de covariance $\rho(s, t) = \int_0^{s \wedge t} \frac{dS_c(u)}{S_c(u)^2}$ avec $S_c(t) = (1 - F(t))(1 - G(t))$ et $S_1(t) = P(T > t, D = 1)$.

3.2. L'estimateur de Kaplan-Meier de la fonction de survie

On peut remarquer que l'estimateur de Nelson-Aalen du taux de hasard cumulé conduit à un estimateur naturel de la fonction de survie, en exploitant la relation $S(t) = \exp(-H(t))$; on peut ainsi proposer comme estimateur de la fonction de survie

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

Cet estimateur est l'estimateur de Harrington et Fleming ; sa variance peut être obtenue par la méthode Delta qui, sous des conditions raisonnables de régularité de la fonction f permet d'écrire que $V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X)$. En effet, si $X = \mu + \sigma Z$ avec σ petit et Z centrée réduite, on remarque que pour une fonction $x \rightarrow f(x)$ suffisamment régulière, en effectuant le développement limité $f(\mu + h) \approx f(\mu) + h \frac{df}{dx}(\mu)$, on trouve que $V(f(X)) \approx V\left(f(\mu) + \sigma Z \frac{df}{dx}(\mu)\right) = \sigma^2 \frac{df}{dx}(\mu)^2$. En prenant ici $f(x) = e^{-x}$, on trouve que $V(\hat{S}) \approx e^{-2E(\hat{H})} V(\hat{H}) \approx \hat{S}^2 V(\hat{H})$, ce qui conduit à l'estimateur de la variance :

$$\hat{V}(\hat{S}_{HF}(t)) = \exp\left(-2 \sum_{\{i | t_i \leq t\}} \frac{d(t_i)}{n-i+1}\right) \sum_{\{i | t_i \leq t\}} \frac{d(t_i)}{(n-i+1)^2} = \hat{S}_{HF}(t)^2 \times \sum_{\{i | t_i \leq t\}} \frac{d(t_i)}{(n-i+1)^2}.$$

Comme on a montré que $E[\hat{H}_{NA}(t)] \leq H(t)$ et que la fonction $g(x) = e^{-x}$ est convexe, on en déduit que :

$$E(\hat{S}_{HF}(t)) = E(g(\hat{H}_{NA}(t))) \geq g(E(\hat{H}_{NA}(t))) = \exp(-E(\hat{H}_{NA}(t))) \geq \exp(-H(t)) = S(t),$$

en d'autres termes l'estimateur de Harrington-Fleming de la fonction de survie présente un biais de surestimation.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

Toutefois, cet estimateur peut être amélioré, ce qui amène à introduire l'estimateur de Kaplan-Meier.

3.2.1. Présentation générale

L'estimateur de Kaplan-Meier (KAPLAN et MEIER [1958]) peut être introduit via les processus ponctuels, en remarquant que la fonction de survie de base du modèle est l'unique solution de l'équation intégrale suivante :

$$S(t) = 1 - \int_0^t S(u-) h(u) du .$$

L'équation ci-dessus exprime simplement le fait que la somme des survivants en t et des individus sortis avant t est constante. Lorsque la fonction de survie est continue, la démonstration est immédiate en effectuant le changement de variable $v = \ln S(u)$, $dv = -h(u) du$.

En remplaçant $h(u) du$ par son estimateur $\frac{d\bar{N}^1(u)}{\bar{R}(u)}$ introduit à la section précédente on peut proposer un estimateur de la fonction de survie en cherchant une solution à l'équation :

$$\hat{S}(t) = 1 - \int_0^t \hat{S}(u-) \frac{d\bar{N}^1(u)}{\bar{R}(u)} .$$

On peut montrer (cf. GILL [1992]) qu'il existe une unique solution à cette équation et on obtient alors l'estimateur de Kaplan-Meier de la fonction de survie. Si l'existence n'est pas simple à prouver, l'unicité découle directement de la remarque que si deux estimateurs sont solutions de l'équation ci-dessus alors :

$$\hat{S}_1(t) - \hat{S}_2(t) = \sum_{T_i \leq t} [\hat{S}_1(T_i-) - \hat{S}_2(T_i-)] \frac{d_i}{r_i}$$

et comme $\hat{S}_1(0) - \hat{S}_2(0) = 0$, par récurrence $\hat{S}_1(t) - \hat{S}_2(t) = 0$ pour tout t . Cet estimateur peut s'exprimer à l'aide de l'estimateur de Nelson-Aalen de la manière suivante :

$$\hat{S}(t) = \prod_{s \leq t} (1 - \Delta \hat{H}(s))$$

où $\Delta \hat{H}(s) = \hat{H}(s) - \hat{H}(s-)$. On peut toutefois proposer une construction explicite plus intuitive de cet estimateur, décrite *infra*.

La construction heuristique de l'estimateur de Kaplan-Meier s'appuie sur la remarque suivante : la probabilité de survivre au-delà de $t > s$ peut s'écrire :

$$S(t) = P(T > t | T > s) P(T > s) = P(T > t | T > s) S(s) .$$

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

On peut renouveler l'opération, ce qui fait apparaître des produits de termes en $P(T > t | T > s)$; si on choisit comme instants de conditionnement les instants où se produit un événement (sortie ou censure), on se ramène à estimer des probabilités de la forme :

$$p_i = P(T > T_{(i)} | T > T_{(i-1)})$$

p_i est la probabilité de survivre sur l'intervalle $[T_{(i-1)}, T_{(i)}]$ sachant qu'on était vivant à l'instant $T_{(i-1)}$. Un estimateur naturel de $q_i = 1 - p_i$ est $\hat{q}_i = \frac{d_i}{r_i} = \frac{d_i}{n-i+1}$. On observe alors qu'à l'instant $T_{(i)}$, et en l'absence d'*ex aequo*, si $D_{(i)} = 1$ alors il y a sortie par décès donc $d_i = 1$, et dans le cas contraire l'observation est censurée et $d_i = 0$. L'estimateur de Kaplan-Meier s'écrit donc finalement :

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

En pratique cependant on est confronté à la présence d'*ex aequo*; on suppose alors par convention que les observations non censurées précèdent toujours les observations censurées. On obtient l'expression suivante de l'estimateur :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

Remarque n°1 : on travaille ici avec la version continue à droite de la fonction de survie ; certains auteurs utilisent la version continue à gauche. Dans ce cas, les expressions ci-dessus restent valables en remplaçant le terme $T_{(i)} \leq t$ par $T_{(i)} < t$.

Remarque n°2 : dans le cas où il y a des arrivées en cours de période (troncatures gauches), l'expression $\hat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$ reste valable en tenant compte dans le calcul de r_i ; là encore, la fonction de survie ne saute qu'au moment des sorties non censurées.

3.2.2. Comparaison avec l'estimateur de Harrington et Fleming

Les deux estimateurs s'écrivent respectivement, après transformation par le logarithme

$$\ln \hat{S}_{KM}(t) = \sum_{T_{(i)} \leq t} \ln \left(1 - \frac{d_i}{r_i}\right) \text{ et } \ln \hat{S}_{HF}(t) = - \sum_{T_{(i)} \leq t} \frac{d_i}{r_i} \text{ et donc}$$

$$\ln \hat{S}_{KM}(t) - \ln \hat{S}_{HF}(t) = \sum_{T_{(i)} \leq t} \left(\ln \left(1 - \frac{d_i}{r_i}\right) + \frac{d_i}{r_i} \right).$$

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

On vérifie aisément que la fonction $f(x) = \ln(1-x) + x$ est toujours négative et donc $\hat{S}_{KM}(t) \leq \hat{S}_{HF}(t)$.

3.2.3. Exemples

3.2.3.1. Données de Freireich

On reprend les données de Freireich utilisée en 3.1.3 ci-dessus, et on s'intéresse à la comparaison des résultats obtenus par les méthodes de Kaplan-Meier et de Nelson-Aalen ; on trouve que :

Tab. 3. Calcul de l'estimateur de Kaplan-Meier

Rechutes	t_i	r_i	d_i	d_i/r_i	$\hat{H}_{NA}(t)$	$\hat{S}_{KM}(t)$	$-\ln \hat{S}_{KM}(t)$
1-2-3	6	21	3	0,143	0,143	0,857	0,154
5	7	17	1	0,059	0,202	0,807	0,215
7	10	15	1	0,067	0,268	0,753	0,284
10	13	12	1	0,083	0,352	0,690	0,371
11	16	11	1	0,091	0,443	0,627	0,466
15	22	7	1	0,143	0,585	0,538	0,620
16	23	6	1	0,167	0,752	0,448	0,803

On constate que le taux de hasard cumulé obtenu avec Kaplan-Meier est supérieur au taux de hasard cumulé issu de l'estimateur de Nelson-Aalen. Si on calcule l'estimateur de Harrington et Fleming de la fonction de survie $\hat{S}(t) = \exp(-\hat{H}_{NA}(t))$, on constate de même qu'il est systématiquement supérieur à l'estimateur de Kaplan-Meier. Au-delà des aspects strictement statistiques, des considérations prudentielles pourraient donc orienter vers le choix d'un estimateur ou d'un autre.

3.2.3.2. Autre exemple

Sur 10 patients atteints de cancer des bronches on a observé les durées de survie suivantes, exprimées en mois⁸: 1 / 3 / 4+ / 5 / 7+ / 8 / 9 / 10+ / 11 / 13+. L'estimateur de Kaplan-Meier de la fonction de survie $S(t)$ se calcule de la manière suivante :

⁸ Le signe + indique une observation censurée

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

Tab. 4. Calcul de l'estimateur

t_i	r_i	d_i	Survie	Intervalle
0	10	0	100,0%	[0 1[
1	10	1	90,0%	[1 3[
3	9	1	80,0%	[3 5[
5	7	1	68,6%	[5 8[
8	5	1	54,9%	[8 9[
9	4	1	41,1%	[9 11[
11	2	1	20,6%	

3.2.4. Principales propriétés

L'estimateur de Kaplan-Meier possède un certain nombre de « bonnes propriétés » qui en font la généralisation naturelle de l'estimateur empirique de la fonction de répartition en présence de censure : il est convergent⁹, asymptotiquement gaussien, cohérent et est également un estimateur du maximum de vraisemblance généralisé. Toutefois, cet estimateur est biaisé positivement. L'estimateur de Kaplan-Meier est l'unique estimateur cohérent de la fonction de survie (voir DROESBEKE et al. [1989] pour la démonstration de cette propriété).

La notion de « maximum de vraisemblance » doit être adaptée au contexte non paramétrique de la manière suivante¹⁰ :

Définition : soit Φ est une famille de probabilités sur \mathbb{R}^n (avec la tribu borélienne) non dominée ; $\forall x \in \mathbb{R}^n$, et $P_1, P_2 \in \Phi$, on pose $l(x, P_1, P_2) = \frac{dP_1}{d(P_1 + P_2)}(x)$; on dit alors que \hat{P} est GMLE pour P si $l(x, \hat{P}, P) \geq l(x, P, \hat{P})$.

On peut alors montrer que l'estimateur \hat{S} est GMLE pour S , pour autant que les lois de la durée de vie non censurée et de la censure soient diffuses, et à condition que la famille Φ contienne les lois de probabilité chargeant les points (T_i, D_i) . Les autres propriétés sont détaillées ci-après.

⁹ Pour autant que la fonction de survie et la distribution des censures n'aient pas de discontinuités communes.

¹⁰ On verra en 3.2.7 le lien avec le maximum de vraisemblance dans un contexte paramétrique.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

3.2.5. Variance de l'estimateur de Kaplan Meier

On propose ici une justification heuristique d'un estimateur de la variance de l'estimateur de Kaplan-Meier, l'estimateur de Greenwood.

L'expression $\hat{S}(t) = \prod_{T_i \leq t} (1 - \Delta \hat{H}(T_i)) = \prod_{T_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$ permet d'écrire¹¹ :

$$\ln(\hat{S}(t)) = \sum_{T_{(i)} \leq t} \ln\left(1 - \frac{d_i}{r_i}\right) = \sum_{T_{(i)} \leq t} \ln(1 - \hat{q}_i).$$

Avec l'indépendance des variables $\ln(1 - \hat{q}_i)$, comme la loi de $r_i \hat{p}_i$ est binomiale de paramètres (r_i, p_i) , on a par la méthode delta, $V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X)$:

$$V(\ln \hat{S}(t)) \approx V(\hat{p}_i) \left[\frac{d}{dp} \ln(\hat{p}_i) \right]^2 = \frac{\hat{q}_i(1 - \hat{q}_i)}{r_i} \frac{1}{(1 - \hat{q}_i)^2} = \frac{\hat{q}_i}{r_i(1 - \hat{q}_i)}$$

ce qui conduit à proposer comme estimateur de la variance de $\ln \hat{S}(t)$:

$$\hat{V}(\ln \hat{S}(t)) = \sum_{T_{(i)} \leq t} \frac{\hat{q}_i}{r_i(1 - \hat{q}_i)} = \sum_{T_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)}$$

En appliquant de nouveau la méthode delta avec pour f la fonction logarithme, on obtient finalement :

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \gamma(t)^2$$

avec $\gamma(t) = \sqrt{\sum_{T_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)}}$ Cet estimateur est l'estimateur de Greenwood. Il est

convergeant pour la variance asymptotique de l'estimateur de Kaplan-Meier. Il permet avec la normalité asymptotique¹² de l'estimateur de Kaplan-Meier de calculer des intervalles de confiance (asymptotiques) dont les bornes sont, pour la valeur de la survie en $T_{(i)}$:

$$S_i \times \left(1 \pm u_{\frac{1-\alpha}{2}} \gamma(T_{(i)})\right) = S_i \times \left(1 \pm u_{\frac{1-\alpha}{2}} \sqrt{\frac{d_1}{r_1(r_1 - d_1)} + \frac{d_2}{r_2(r_2 - d_2)} + \dots + \frac{d_i}{r_i(r_i - d_i)}}\right)$$

On construit de la sorte des intervalles ponctuels, à t fixé.

¹¹ Cette formule fournit un estimateur de la fonction de hasard cumulé appelé estimateur de Breslow de H .

¹² Voir 3.2.6.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

Remarque

Pour estimer la variance de $\ln(1 - \hat{q}_i)$, on aurait pu également utiliser l'approximation de la loi binomiale par une loi de Poisson, comme à la section 3.1.2, ce qui conduit à

$$V(\ln \hat{p}_i) \approx V(\hat{p}_i) \left[\frac{d}{dp} \ln(\hat{p}_i) \right]^2 = \frac{\hat{q}_i}{r_i} \frac{1}{(1 - \hat{q}_i)^2} = \frac{d_i}{r_i^2 \left(1 - \frac{d_i}{r_i}\right)^2} = \frac{d_i}{(r_i - d_i)^2},$$

soit une expression un peu différente de celle utilisée pour la formule de Greenwood.

On peut alors chercher à construire des bandes de confiance pour la fonction de survie. Nair propose ainsi en 1984 (cf. KLEIN et MOESCHBERGER [2005]) des bandes de confiance linéaires de la forme :

$$\hat{S}(t)(1 \pm c_\alpha(a(t_m), a(t_M))\gamma(t))$$

avec $a(t) = \frac{n \times \gamma(t)^2}{1 + n \times \gamma(t)^2}$ et où les coefficients de confiance $c_\alpha(x_1, x_2)$ sont tabulés (ils sont fournis en annexe de KLEIN et MOESCHBERGER [2005]).

Ces formules peuvent être utilisées pour construire des intervalles de confiance pour les taux conditionnels de sortie $\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$; en effet, on en déduit de $\hat{S}(x) = \prod_{T_i \leq x} \left(1 - \frac{d_i}{r_i}\right)$

que $1 - \hat{q}(x) = \prod_{x < T_i \leq x+1} \left(1 - \frac{d_i}{r_i}\right)$ et donc :

$$\hat{V}(\hat{q}(x)) = (1 - \hat{q}(x))^2 \sum_{x < T_i \leq x+1} \frac{d_i}{r_i(r_i - d_i)}$$

d'où immédiatement l'expression d'un intervalle de confiance asymptotique :

$$\hat{q}_{\pm}(x) = 1 - (1 - \hat{q}(x)) \times \sqrt{1 \pm u_{1-\frac{\alpha}{2}} \times \sum_{x < T_i \leq x+1} \frac{d_i}{r_i(r_i - d_i)}}.$$

3.2.6. Propriétés asymptotiques

L'estimateur de Kaplan-Meier est asymptotiquement gaussien ; précisément on a le résultat suivant :

Proposition : si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sqrt{n}(\hat{S} - S) \rightarrow W_s$$

avec W_s un processus gaussien centré de covariance :

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

$$\rho(s, t) = S(s)S(t) \int_0^{s \wedge t} \frac{dF(u)}{(1-F(u))^2(1-G(u))}.$$

En particulier lorsque le modèle n'est pas censuré (i.e. $G(u)=0$) on retrouve le résultat classique rappelé en introduction. L'intérêt de résultats de convergence au niveau du processus lui-même plutôt que pour un instant fixé est que l'on peut en déduire des bandes de confiance asymptotique pour l'estimateur de Kaplan-Meier.

On peut trouver dans GILL [1980] une démonstration de la normalité asymptotique de \hat{S}_{KM} , fondée sur la théorie des processus ponctuels. En notant $F = 1 - S$ et $\hat{F} = 1 - \hat{S}_{KM}$, la bande de confiance qu'il obtient s'écrit :

$$\liminf_{n \rightarrow \infty} P \left\{ \sup_{s \in [0, t]} \left| \frac{\hat{F}(s) - F(s)}{1 - \hat{F}(s)} \right| \leq \frac{\sqrt{\hat{V}(t)}}{1 - \hat{F}(t)} x \right\} \geq \sum_{k=-\infty}^{\infty} (-1)^k [\Phi((2k+1)x) - \Phi((2k-1)x)]$$

où $\hat{V}(t) = \hat{S}_{KM}^2 \int_0^t \frac{d\bar{N}^1(u)}{\bar{R}(u)(\bar{R}(u) - \Delta\bar{N}^1(u))}$ estime la variance du processus gaussien limite W_S

.

3.2.7. Version discrétisée : lien avec l'approche paramétrique

Le calcul de l'estimateur de Kaplan-Meier implique que l'on dispose des données individuelles avec les dates précises de survenance des évènements ; en pratique, autre que sur des populations importantes le calcul peut être lourd, cette information n'est pas toujours accessible. On souhaite alors utiliser cette démarche pour des données regroupées par période, par exemple en fixant comme unité de temps le mois et en comptabilisant des sorties d'incapacité mois par mois. C'est la démarche suivie par le BCAC pour l'élaboration des lois de maintien¹³ du décret de 1996.

Formellement, si on considère les instants $t_1 < \dots < t_N$ auxquels se produisent les sorties (par exemple les âges entiers de décès) et que l'on dispose d'un échantillon de taille n pour lequel on a observé une séquence (r_i, d_i) d'effectifs sous risque et de décès aux dates $t_1 < \dots < t_N$, on peut remarquer que le nombre de sorties D_i sur l'intervalle $[t_i, t_{i+1}[$ suit une loi binomiale de paramètres (r_i, h_i) ; h_i désigne ici le taux de hasard à la date t_i (homogène à un q_x).

Les sorties dans les intervalles $[t_i, t_{i+1}[$ étant indépendantes les unes des autres, on trouve donc que la vraisemblance de ce modèle s'écrit :

$$L = \prod_{i=1}^N C_{r_i}^{d_i} h_i^{d_i} (1-h_i)^{r_i-d_i}.$$

¹³ <http://www.ressources-actuarielles.net/bcac>.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

ressources-actuarielles.net

Modèles de durée

La log-vraisemblance s'écrit donc :

$$\ln(L) = \sum_{i=1}^N \left[C_{r_i}^{d_i} + d_i \ln(h_i) + (r_i - d_i) \ln(1-h_i) \right]$$

et les conditions du premier ordre $\frac{\partial}{\partial h_i} \ln L = 0$ conduisent aux estimateurs :

$$\hat{h}_i = \frac{d_i}{r_i}.$$

On retrouve donc l'estimateur présenté en 3.2.1 ci-dessus. Pour que cette démarche soit pertinente, il convient de s'assurer que la discréttisation ne génère pas de biais important sur l'estimation des taux de sortie : par exemple dans le cas de l'arrêt de travail, il est connu que les sorties sont très importantes au cours du premier mois (en pratique environ 50 % des arrêts de travail durent moins de 30 jours). Si donc on adopte un pas mensuel, on prend mal en compte le rythme élevé des sorties au cours de la première période ; il conviendrait donc ici de choisir un pas de discréttisation petit. Plus généralement, le raisonnement ci-dessus est pertinent pour autant que la longueur de chaque intervalle considéré soit « petite » au regard de la vitesse de variation de la fonction de survie.

4. Prise en compte de variables explicatives

Lorsque la population étudiée est hétérogène, il est important de prendre en compte les spécificités de chaque sous-groupe. En supposant que l'hétérogénéité est la conséquence d'un mélange de sous-populations caractérisées chacune par des variables observables, on s'intéresse ici à des modélisations de la fonction de hasard intégrant l'effet des variables explicatives. Cette question a déjà été abordée dans un contexte paramétrique et semi-paramétrique (modèle de Cox), on s'intéresse ici au cas non paramétrique.

Ce chapitre est inspiré de MARTINUSSEN et SCHEIKE [2006] auquel le lecteur pourra se reporter pour les démonstrations. Il est également précisé que la mise en pratique des modèles présentés ici peut être effectuée à l'aide du package *timereg* du logiciel R, développé par ces auteurs ou en utilisant le package *survival*.

4.1. Le modèle additif d'Aalen

La fonction de hasard est supposée s'écrire :

$$h(t) = X^T(t)\beta(t)$$

avec $X^T(t) = (X_1(t), \dots, X_p(t))$ un vecteur de variables explicatives (prévisible) et $\beta(t)$ un processus p-dimensionnel localement intégrable. On peut de manière équivalente dire que l'intensité du modèle de comptage sous-jacent s'écrit :

$$\lambda(t) = R(t)X^T(t)\beta(t).$$

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

On dispose d'un ensemble d'observations $(N_i^1(t), R_i(t), X^i(t))_{1 \leq i \leq n}$ et on cherche à estimer le vecteur $\beta(t)$; en pratique on va être en mesure de construire aisément un estimateur de $B(t) = \int_0^t \beta(u) du$ en s'appuyant sur les remarques qui suivent.

On note pour alléger les formules $\lambda(t) = (\lambda_1(t), \dots, \lambda_n(t))^T$ et $N^1(t) = (N_1^1(t), \dots, N_n^1(t))^T$, puis $X(t) = (R_1(t)X^1(t), \dots, R_n(t)X^n(t))^T$ qui est une matrice de taille $n \times p$. Avec ces notations on a en désignant par $\Lambda(t) = \int_0^t \lambda(u) du$ le processus vectoriel de taille n des intensités cumulées, $M(t) = N^1(t) - \Lambda(t)$ est une martingale. En observant alors que :

$$dN^1(t) = X(t)\beta(t)dt + dM(t) = X(t)dB(t) + dM(t)$$

comme le terme en $dM(t)$ est centré et que les incrément de la martingale sont non corrélés, on peut chercher à estimer les incrément $dB(t)$ par des techniques classiques de régression linéaire. Pour cela on pose :

$$X^-(t) = (X^T(t)X(t))^{-1}X^T(t),$$

si $X^T(t)X(t)$ est inversible et 0 sinon. $X^-(t)$ s'appelle l'inverse généralisé de X , qui est une matrice de taille $p \times n$ vérifiant $X^-(t)X(t) = J(t)I_p$ avec $J(t)$ qui vaut 1 si l'inverse existe, et 0 sinon. En pratique lorsque $X(t)$ est de plein rang $X^T(t)X(t)$ est inversible et on a alors simplement $X^-(t)X(t) = I_p$. Il est alors naturel de proposer comme estimateur de B le processus :

$$\hat{B}(t) = \int_0^t X^-(u)dN^1(u).$$

Le fait que $\hat{B}(t) = \int_0^t J(s)dB(s) + \int_0^t X^-(s)dM(s)$ assure en effet que \hat{B} estime B essentiellement sans biais et on peut de plus montrer sous certaines conditions techniques peu restrictives que $\sqrt{n}(\hat{B} - B)$ converge en loi en tant que processus vers un processus gaussien centré dont on peut de plus calculer la fonction de covariance.

Le calcul de l'estimateur $\hat{B}(t) = \int_0^t X^-(u)dN^1(u)$ se ramène à des calculs de sommes discrètes aux instants de saut du processus $N^1(t)$. De manière plus précise on a $\hat{B}(t)$ qui est un vecteur de taille p et :

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

ressources-actuarielles.net

Modèles de durée

$$\hat{B}_j(t) = \sum_i \int_0^t X_{ji}^-(u) dN_i^1(u)$$

Mais $N_i^1(t)$ saute au plus une fois à l'instant T_i et l'incrément à cet instant est de 1 (s'il y a saut). On en déduit l'expression suivante :

$$\hat{B}_j(t) = \sum_{T_i \leq t} X_{ji}^-(T_i) \times D_i.$$

Le calcul nécessite donc la détermination de $X^-(T_i) = (X^T(T_i) X(T_i))^{-1} X^T(T_i)$ pour toutes les sorties non censurées.

4.2. Variante semi-paramétrique : le modèle de Lin et Ying

Dans les situations d'assurance, les variables explicatives sont en général constantes au cours du temps (typiquement elles sont associées à une caractéristique telle que le sexe, la CSP, le niveau du contrat, etc.).

Cela se traduit par la constance des variables $X_j(t)$. Ce cas particulier conduit à un modèle semi-paramétrique, et les méthodes décrites ci-dessus sont légèrement modifiées. Parmi ces modèles on peut notamment mentionner le modèle de LIN et YING [1994], dans lequel la fonction de hasard est supposée de la forme :

$$h(t | Z = z) = h_0(t) + \gamma^T z.$$

LIN et YING [1994] et KLEIN et MOESCHBERGER [2005] montrent qu'à partir de la décomposition martingale du processus de Poisson, l'estimation des coefficients du modèle est :

$$\gamma = A^{-1}B,$$

$$\text{où } A = \sum_{i=1}^D \sum_{j \in R_i} (z_j - \bar{z}_i)^T (z_j - \bar{z}_i), \quad B = \sum_{i=1}^n d_i (z_i - \bar{z}_i) \text{ et } \bar{z}_i = \frac{1}{R_i} \sum_{j \in R_i} z_j.$$

La significativité globale du modèle peut être appréciée à partir de la statistique de Wald qui suit une distribution du Khi-deux à p degrés de libertés (p étant la dimension de Z représentant les variables explicatives du modèle) sous l'hypothèse $H_0: \gamma = 0$, soit :

$$\chi_w^2 = \gamma^T V^{-1} \gamma,$$

où $V = A^{-1} C A^{-1}$ avec $C = \sum_{i=1}^n d_i (z_i - \bar{z}_i)^T (z_i - \bar{z}_i)$. Dans le cas du test de significativité d'un paramètre, on teste l'hypothèse de nullité chaque paramètre γ_j (avec $j = 1, \dots, p$ et $\gamma = (\gamma_1, \dots, \gamma_p)$), et on considère donc $H_0: \gamma_j = 0$, soit $\chi_{W_j}^2 = \gamma_j^2 / V_{jj}$.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Modèles de durée

5. Comparaison d'échantillons : approche non paramétrique

On se place dans la situation où on souhaite comparer les durées de vie respectives de deux échantillons indépendants. Plus précisément, on dispose de deux échantillons indépendants, éventuellement censurés, et on souhaite tester l'hypothèse nulle d'égalité des fonctions de survie dans les deux échantillons. En l'absence de censure, on dispose des classiques tests de rang (test de Wilcoxon, test de Savage), que l'on va adapter à la présence de censure.

5.1. Rappel : principe des tests de rang¹⁴

On dispose donc de deux séries d'observations, E_1 et E_2 , de tailles respectives n_1 et n_2 ; on note $n = n_1 + n_2$; on range la séquence des valeurs observées (x_1, \dots, x_n) par ordre croissant :

$$x_{(1)} < \dots < x_{(n)}.$$

Le principe d'une statistique linéaire de rang est d'attribuer une pondération (un score) α_i à l'observation $x_{(i)}$ de rang i dans le classement commun des deux échantillons. On construit alors deux statistiques :

$$R_1 = \sum_{i \in E_1} \alpha_i \text{ et } R_2 = \sum_{i \in E_2} \alpha_i.$$

Comme $R_1 + R_2 = \sum_{i=1}^n \alpha_i$, qui est connue et déterministe, il est indifférent de travailler sur

l'une ou l'autre des statistiques ; en pratique on retient celle associée à l'échantillon le plus petit. En choisissant $\alpha_i = i$, on obtient le test de Wilcoxon ; le test de Savage est quant à

$$\text{lui associé au choix } \alpha_i = 1 - \sum_{j=1}^i \frac{1}{n-j+1}.$$

Enfin, le choix d'un test plutôt que d'un autre peut être guidé par la forme de l'alternative, en retenant le test (localement) le plus puissant pour une alternative donnée.

5.2. Adaptation des tests de rang au cas censuré¹⁵

L'adaptation des tests précédents au cas censuré conduit à introduire la suite ordonnée des instants de décès observés (non censurés) dans l'échantillon commun, que l'on notera $t_1 < \dots < t_N$. À chaque instant t_i on désigne par d_{ij} le nombre de décès et r_{ij} l'effectif sous risque dans le groupe j . L'effectif sous risque est calculé avant les sorties en t_i , de sorte

¹⁴ Pour des développements sur le sujet se reporter à CAPÉRAÀ et VAN CUTSEM [1988].

¹⁵ Voir par exemple HILL et al. [1996] pour de plus amples développements.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Modèles de durée

que les vivants après t_i sont en nombre $r_{ij} - d_{ij}$. On peut synthétiser cela dans le tableau ci-après :

Mise en œuvre d'un test de rang

	Décès en t_i	Survivants après t_i	Total
Groupe n°1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Groupe n°2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
Ensemble	d_i	$r_i - d_i$	r_i

Sous l'hypothèse nulle d'égalité des distributions de survie dans les deux groupes, à chaque instant on doit avoir égalité des proportions de décès dans les deux groupes, ce qui a pour conséquence l'indépendance des lignes et des colonnes dans le tableau ci-dessus. On est donc dans le cas d'un tableau de contingence à marges fixées, et alors la variable aléatoire

d_{ij} est distribuée selon une loi hypergéométrique¹⁶ $H\left(r_i, d_i, \frac{r_{ij}}{r_i}\right)$ (puisque on compte le

nombre de décès dans le groupe n° j choisis parmi les d_i décès totaux, la probabilité d'appartenance au groupe n° j étant $p = \frac{r_{ij}}{r_i}$ et la taille de la population étant r_i). On en

conclut que l'espérance et la variance de d_{ij} : $E(d_{ij}) = d_i \frac{r_{ij}}{r_i}$ et $V(d_{ij}) = d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$. Ces

observations conduisent à construire des statistiques fondées sur des sommes pondérées des $d_{ij} - E(d_{ij})$, qui sont asymptotiquement gaussiennes. En notant (w_i) les pondérations retenues, on utilise finalement des statistiques de la forme :

$$\phi_j = \frac{\left[\sum_{i=1}^N w_i \left(d_{ij} - d_i \frac{r_{ij}}{r_i} \right) \right]^2}{\sum_{i=1}^N w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}}$$

qui suit asymptotiquement un $\chi^2(1)$. Dans la suite on notera $\sigma^2 = \sum_{i=1}^N w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$.

5.2.1. Le test du log-rank

Le choix le plus simple que l'on puisse imaginer pour les pondérations est $w_i = 1$, il conduit au test dit du « log-rank ». Dans ce cas le numérateur de la statistique de test ϕ_j est le carré

¹⁶ On rappelle que la loi hypergéométrique $H(n, k, p)$ est la loi du nombre de boules noires lors d'un tirage avec remise de k boules dans une urne contenant n boules et les boules noires étant en proportion p .

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Modèles de durée

de la différence entre le nombre de décès observés et le nombre de décès théoriques, sous l'hypothèse nulle :

$$\phi_j = \frac{(D_j^{th} - D_j^{obs})^2}{\sigma^2}.$$

Ce test généralise au cas de données censurées le test de Savage. On peut noter que sous l'hypothèse nulle $D_1^{obs} + D_2^{obs} = D_1^{th} + D_2^{th}$, en d'autres termes la valeur de la statistique de test ne dépend pas du groupe sur laquelle on l'évalue. La forme de la statistique suggère la formule approchée suivante :

$$\phi = \frac{(D_1^{th} - D_1^{obs})^2}{D_1^{th}} + \frac{(D_2^{th} - D_2^{obs})^2}{D_2^{th}}$$

dont on peut montrer qu'elle est inférieure à celle du log-rank (cf. PETO et PETO [1972]). Sa forme évoque celle d'un Khi-2 d'ajustement usuel. Le test du log-rank est le test le plus couramment employé.

5.2.2. Le test de Gehan

Gehan (GEHAN E.A. [1965]) propose de retenir $w_i = r_i$, ce qui conduit à pondérer plus fortement les décès les plus précoces. Ce test généralise au cas de données censurées le test de Wilcoxon. La statistique de test n'admet pas d'expression simplifiée comme dans le cas du log-rank. Il présente l'inconvénient de dépendre assez fortement de la distribution de la censure.

5.2.3. Exemple : application aux données de Freireich

On reprend ici les deux groupes du protocole utilisé par Freireich. Les calculs des statistiques de test peuvent être menés à partir du tableau suivant :

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

Tab. 5. Calculs préparatoires

Durées	6-MP		Placebo		n_i	d_i	$E(d_{i2})$	$V(d_{i2})$
	n_{i1}	d_{i1}	n_{i2}	d_{i2}				
1	21	0	21	2	42	2	1,00	0,49
2	21	0	19	2	40	2	0,95	0,49
3	21	0	17	1	38	1	0,45	0,25
4	21	0	16	2	37	2	0,86	0,48
5	21	0	14	2	35	2	0,80	0,47
6	21	3	12	0	33	3	1,09	0,65
7	17	1	12	0	29	1	0,41	0,24
8	16	0	12	4	28	4	1,71	0,87
10	15	1	8	0	23	1	0,35	0,23
11	13	0	8	2	21	2	0,76	0,45
12	12	0	6	2	18	2	0,67	0,42
13	12	1	4	0	16	1	0,25	0,19
15	11	0	4	1	15	1	0,27	0,20
16	11	1	3	0	14	1	0,21	0,17
17	10	0	3	1	13	1	0,23	0,18
22	7	1	2	1	9	2	0,44	0,30
23	6	1	1	1	7	2	0,29	0,20

On obtient les résultats résumés ci-après :

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty]}(T_x)$$

Tab. 6. Statistiques de test

Durées	log-rank			Gehan		
	Pondération	Coefficient	Variance	Pondération	Coefficient	Variance
1	1,00	1,00	0,49	42	42,00	860,49
2	1,00	1,05	0,49	40	42,00	777,54
3	1,00	0,55	0,25	38	21,00	357,00
4	1,00	1,14	0,48	37	42,00	653,33
5	1,00	1,20	0,47	35	42,00	570,71
6	1,00	-1,09	0,65	33	-36,00	708,75
7	1,00	-0,41	0,24	29	-12,00	204,00
8	1,00	2,29	0,87	28	64,00	682,67
10	1,00	-0,35	0,23	23	-8,00	120,00
11	1,00	1,24	0,45	21	26,00	197,60
12	1,00	1,33	0,42	18	24,00	135,53
13	1,00	-0,25	0,19	16	-4,00	48,00
15	1,00	0,73	0,20	15	11,00	44,00
16	1,00	-0,21	0,17	14	-3,00	33,00
17	1,00	0,77	0,18	13	10,00	30,00
22	1,00	0,56	0,30	9	5,00	24,50
23	1,00	0,71	0,20	7	5,00	10,00
	105,07	6,26		73441,00	5457,11	
			$\varphi_2 = 16,79$			$\varphi_2 = 13,46$

On trouve dans les deux cas des p-valeurs très faibles, ce qui confirme le comportement différent des deux groupes, qui avait déjà été mis en évidence lors de l'étude des fonctions de risque cumulées respectives.

5.3. Approche par les processus ponctuels

De la même manière que les estimateurs du hasard cumulé ou de la fonction de survie peuvent être obtenus de manière « naturelle » dans le cadre des processus ponctuels, ce formalisme peut s'appliquer aux tests présentés ci-dessus. Cette méthode est détaillée dans GILL [1980].

On se place donc dans la situation où deux groupes sont observés, et on dispose donc des deux processus d'évènements non censurés $\bar{N}_1^1(t)$ et $\bar{N}_2^1(t)$. On fait l'hypothèse que les deux processus ne sautent pas en même temps (ce qui traduit l'orthogonalité des martingales M_1 et M_2 , $\langle M_1, M_2 \rangle = 0$). L'idée est, pour un processus K prévisible positif de considérer le processus :

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

ressources-actuarielles.net

Modèles de durée

$$\Delta(t) = \int_0^t K(u) \frac{d\bar{N}_1^1(u)}{\bar{R}_1(u)} - \int_0^t K(u) \frac{d\bar{N}_2^1(u)}{\bar{R}_2(u)}$$

Le processus $M(t) = \int_0^t K(u) \frac{dM_1(u)}{\bar{R}_1(u)} - \int_0^t K(u) \frac{dM_2(u)}{\bar{R}_2(u)}$ est une martingale et vérifie de plus :

$$M(t) = \Delta(t) - \int_0^t K(u)(h_1(u) - h_2(u)) du.$$

Enfin, sous l'hypothèse nulle d'identité de la loi sous-jacente des deux populations, $M(t) = \Delta(t)$.

Les tests classiques s'obtiennent alors en spécifiant le processus K . Ainsi $K(u) = R_1(u)R_2(u)$ conduit à la statistique de Wilcoxon-Gehan et $K(u) = \frac{R_1(u)R_2(u)}{R_1(u) + R_2(u)}$ à la statistique du log-rank.

Les résultats généraux sur les processus ponctuels permettent d'obtenir la loi limite de $\Delta(t)$ sous l'hypothèse nulle ; plus précisément, on montre que $\Delta(t)$ converge en loi vers une loi normale centrée de variance $\sigma^2(t)$; un estimateur convergent de la variance est donné par la variation quadratique de la martingale $\Delta(t)$:

$$\langle \Delta, \Delta \rangle_t = \int_0^t \left[\frac{K(u)}{R_1(u)} \right]^2 d\bar{N}_1^1(u) + \int_0^t \left[\frac{K(u)}{R_2(u)} \right]^2 d\bar{N}_2^1(u).$$

6. Références

- AALEN O. [1978] « Non-parametric inference for a family of counting processes ». *Ann. Stat.* 6, 710-726.
- BORGAN O. [2014] « [Kaplan-Meier Estimator](#) », Wiley StatsRef: Statistics Reference Online, doi: 10.1002/9781118445112.stat06033
- CAPÉRAÀ P., VAN CUTSEM B. [1988] Méthodes et modèles en statistique non paramétrique, Presses de l'Université Laval, Paris : Dunod.
- DACUNHA-CASTELLE D., DUFLO M. [1983] Probabilités et Statistiques. Vol. 1 et 2, Paris : Masson.
- DROESBEKE J.J., FICHET B., TASSI P. [1989] Analyse statistique des durées de vie , Paris : Economica.
- GEHAN E.A. [1965] « A generalized Wilcoxon test for comparing arbitrarily singly-censored samples ». *Biommetrika*, 41, 361-372.
- GILL R.D. [1980] « Censoring and stochastic Integrals ». *Mathematical Centre Tracts*, n°124, Amsterdam : Mathematische Centrum.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} I_{[t; \infty[}(T_x)$$

Modèles de durée

GILL R.D. [1992] « [Lecture on Survival Analysis](#) », Proceedings of the Ecole d'Eté de Probabilités de Saint Flour XXII

FLEMING T.R., HARRINGTON D.P. [1991] *Counting processes and survival analysis*, Wiley Series in Probability and Mathematical Statistics. New-York : Wiley.

HILL C., COM-NOUGUÉ C. [1996] *Analyse statistique des données de survie*, Médecine-Sciences, Paris : Flammarion.

KAPLAN E.L., MEIER P. [1958] « Non-parametric estimation from incomplete observations ». *Journal of the American Statistical Association*, 53, 457-481.

KLEIN J. P., MOESCHBERGER M. L. [2005] « Survival Analysis – Techniques for Censored and Truncated Data », Springer, 2nd edition.

LIN D. Y., YING Z. [1994] « Semiparametric analysis of the additive risk model », *Biometrika*, n. 81.

MARTINUSSEN T., SCHEIKE T. [2006] *Dynamic regression models for survival data*, New-York: Springer.

NAIR V. N. [1984] « Confidence Bands for Survival Functions with Censored Data: A Comparative Study », *Technometrics* 14: 945-965.

NELSON W.B. [1972] « Theory and applications of hazard plotting for censored data ». *Technometrics*, 14, 945-965.

PETO R., PETO J. [1972] « Asymptotically efficient rank invariant test procedures (with discussion) ». *J. R. Stat. Soc. A*, 135, 185-207.