

GÉRARD DERZKO

EVE LECONTE

**Estimation non paramétrique pour des événements
répétés, tronqués et censurés aléatoirement à
droite - Applications biomédicales**

Journal de la société française de statistique, tome 145, n° 2 (2004),
p. 79-103

http://www.numdam.org/item?id=JSFS_2004__145_2_79_0

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ESTIMATION NON PARAMÉTRIQUE POUR DES ÉVÉNEMENTS RÉPÉTÉS TRONQUÉS ET CENSURÉS ALÉATOIREMENT À DROITE – APPLICATIONS BIOMÉDICALES

Gérard DERZKO * et Eve LECONTE **

RÉSUMÉ

Dans le domaine médical, il est fréquent d'observer des événements répétés, qui peuvent être interrompus aléatoirement par un événement terminal ou par une censure. Nous proposons une procédure d'estimation non paramétrique de l'incidence cumulée de ces événements, à un rang donné puis totale, tout d'abord uniquement en présence de censure à droite, puis en présence de censure à droite et de troncature. La méthode fournit également des estimateurs non paramétriques des prévalences associées à un événement de rang donné. L'emploi de ces procédures est illustré sur des cas d'école avec ex æquo, des données simulées et des essais thérapeutiques.

Mots clés : Événements répétés, événement terminal, estimation non paramétrique, censure, troncature, prévalence.

ABSTRACT

Repeated events, which can be either randomly truncated by a terminal event or censored, are commonly encountered in medical studies. We propose a nonparametric procedure to estimate the cumulative incidence of repeated events for a given rank and globally in the presence of right censoring without and with truncation. The method also provides nonparametric estimators of prevalences associated to an event of a given rank. These procedures are applied to very simple cases with ties observations, to simulated data and to biomedical trials.

Keywords : Repeated events, terminal event, nonparametric estimation, censoring, truncation, prevalence.

* SANOFI-SYNTHELABO Recherche, 371, rue du Professeur Joseph Blayac, 34184 Montpellier Cedex 4

e-mail : gerard.derzko@sanofi-synthelabo.com

** Groupe de Recherche en Économie Mathématique et Quantitative, Université des Sciences Sociales, 21, allée de Brienne, 31000 Toulouse et Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse cedex 4

e-mail : leconte@cict.fr

1. Introduction

Dans de nombreux domaines, en particulier en recherche clinique et en épidémiologie, il est fréquent d'observer des événements répétés chez un même patient. Ce peut par exemple être des crises (asthme, bronchite, épilepsie, . . .), des rechutes (progression cancéreuse), des infections, des accidents. On veut décrire comment ces événements apparaissent dans le temps, au moyen de ce que l'on appellera leur incidence, et aussi comparer ces courbes dans différents groupes. Le problème se complique du fait que l'observation statistique peut être interrompue de façon aléatoire : c'est le mécanisme de censure à droite. D'autre part, un événement terminal tel que le décès peut interrompre complètement le processus des répétitions, conduisant à une troncature des événements non mortels.

Dans cette situation, la méthode d'estimation non paramétrique usuelle consiste à ne considérer que la première occurrence du phénomène chez chaque patient, et à utiliser la procédure d'estimation de Kaplan et Meier (1958) basée sur le mécanisme de censure à droite d'Efron. Cette procédure est bien adaptée au cas du décès, qui est un événement unique, mais elle implique dans le cas des événements répétés une perte d'information qui peut être considérable, à savoir tous les temps d'événements qui surviennent ultérieurement. En pratique, du fait de la non disponibilité de méthodes d'estimation non paramétriques appropriées, l'expérimentateur arrête l'observation après la première occurrence, alors même que l'objet de l'étude est l'estimation de la fréquence globale de survenue en fonction du temps, sans considération particulière pour le premier événement chez un patient donné.

Dans les vingt dernières années, des modèles semi-paramétriques généralisant le modèle de Cox au cas des événements répétés ont été développés (voir Andersen et Gill, 1982, Prentice *et al.*, 1981 et Wei *et al.*, 1989). Comme dans le modèle de Cox usuel, l'accent est surtout mis dans ces modèles sur l'estimation des paramètres associés aux facteurs de risque plutôt que sur l'estimation de l'incidence cumulée des événements. D'autre part, les événements terminaux sont rarement pris en compte ; quand ils le sont (méthode de Wei *et al.*, 1989), ils sont considérés comme des censures vis-à-vis des événements récurrents, approche qui nous paraît incorrecte du fait de la dépendance entre les événements récurrents et l'événement terminal. Dans un cadre non paramétrique, Lawless et Nadeau (1995) ont proposé un estimateur convergent de l'incidence globale (qu'ils appellent « fonction moyenne cumulative ») des événements récurrents dans le cas où il n'y a pas d'événement terminal, mais seulement une censure aléatoire indépendante. Cook et Lawless (1997) ont étendu cet estimateur non paramétrique au cas où il y a des événements terminaux, qui peuvent être dépendants des événements récurrents, en présence de censure aléatoire indépendante des événements récurrents et terminaux ; les propriétés asymptotiques de cet estimateur ont été étudiées par Gosh et Lin (2000). Cet estimateur est convergent mais il ne fournit d'estimation que pour l'incidence globale des événements répétés. En effet, hormis les modèles de Cox généralisés, aucune des méthodes de la littérature ne fournit à notre connaissance d'estimateur de l'incidence des événements qui se produisent à un rang

donné dans la suite des événements récurrents. Or, l'incidence des événements rang par rang a un intérêt en soi, en particulier pour vérifier la cohérence des effets d'un traitement sur les différentes récurrences : en effet, dans certains cas, un phénomène de compensation peut exister, où un traitement retardera par exemple le premier événement d'un individu, mais aura l'effet inverse sur les événements ultérieurs. D'autre part, les estimations rang par rang des incidences cumulées permettent par différence le calcul de la prévalence associée à un événement de rang donné, qui présente un intérêt clinique certain. Récemment, Menjoge (2003) a proposé un estimateur de l'incidence globale construit comme une somme d'estimateurs des incidences d'événements aux différents rangs ; malheureusement, ces estimateurs ne sont pas convergents.

Nous montrons dans la section 2 que ces situations peuvent être modélisées à l'aide d'un processus de comptage censuré. Nous développons dans la section 3 une procédure non paramétrique d'estimation applicable au cas d'un processus censuré, sans événement terminal ; puis dans la section 4, nous proposons une procédure qui s'étend au cas d'une troncature aléatoire à droite des événements récurrents par un événement terminal. Ces procédures font appel à des algorithmes très simples, qui découlent des mécanismes de sélection pour les risques compétitifs présentés dans Derzko et Leconte (2004). Dans chacun des cas, avec et sans événement terminal, un exemple d'école est présenté, puis analysé, et des simulations sont effectuées pour des modèles simples et solubles analytiquement afin de permettre l'appréciation de la qualité des estimateurs présentés. Enfin, des essais thérapeutiques connus sont analysés par ces méthodes ainsi que par d'autres méthodes mentionnées ci-dessus. La section 5 conclut et propose des extensions des procédures proposées.

2. Modélisation

Soit $(X_{1p}, p \in \mathbb{N}^*)$ (l'intérêt de l'indice 1 apparaîtra plus loin) une suite croissante de variables aléatoires strictement positives, de distributions $F_{1p}(t) = \Pr(X_{1p} \leq t)$ telles que $\forall q < p, F_{1p}(t) < F_{1q}(t)$. Ces variables représentent les délais d'occurrence d'un événement non terminal pour un individu, depuis une origine commune des temps ; p indice le rang d'un événement pour l'individu.

À chaque X_{1p} peut être associé un processus de comptage défini par $N_p(t) = \mathbf{I}(X_{1p} \leq t)$, et $N(t) = \sum_{p=1}^{\infty} N_p(t)$ définit un processus global. $E(N(t))$ représente l'espérance du nombre d'événements, tous rangs confondus, qui se produisent avant la date t dans une unité statistique. Elle correspond à une « incidence cumulée » totale au temps t . Notons que cette incidence peut être plus grande que 1. La propriété suivante est immédiate :

$$E(N(t)) = \sum_{p=1}^{\infty} F_{1p}(t). \quad (1)$$

Elle permet d'estimer $E(N(t))$ dès lors qu'on dispose d'estimateurs pour les distributions $F_{1p}(t)$; cela s'applique en particulier aux estimateurs non paramétriques.

Un n -échantillon se présente sous la forme d'un ensemble de n séries indépendantes (individus) des variables $(X_{1p}, p \in \mathbb{N}^*)$. Chaque série ne comprend qu'un nombre fini de délais d'événements, variable d'une série à l'autre, du fait de la censure aléatoire à droite, qui vient masquer l'observation d'événements ultérieurs.

Il peut également arriver que la série des délais d'événements soit tronquée aléatoirement par la survenue d'un événement terminal, qui fait perdre définitivement au processus sa capacité de produire ultérieurement des récurrences. Ainsi, pour l'individu i , on a la série $X_{11i}, X_{12i}, \dots, X_{1n_i i}$, à laquelle on doit ajouter un dernier délai : le délai de censure ou le délai jusqu'à l'événement terminal selon les cas, qui prend le rang $n_i + 1$.

Nous nous proposons, moyennant l'adjonction d'un modèle de censure et de troncature à droite, de fournir des estimations non paramétriques convergentes des fonctions F_{1p} (ainsi que des distributions de censure et de troncature) et en conséquence une estimation de $E(N(t))$ et des prévalences associées à un événement de rang donné.

3. Processus censuré sans événement terminal

3.1. Méthode d'estimation non paramétrique

Nous présentons ici le cas général, mais le lecteur peut acquérir rapidement une idée intuitive de la méthode en se reportant d'emblée à la section suivante, où un cas d'école est traité en détail.

Dans cette section, la production d'événements récurrents est supposée sans fin, mais nous considérons que, pour chaque individu, la série de délais se termine nécessairement par un délai de censure, au-delà duquel aucun événement n'est plus observable. L'individu i fournit donc n_i délais d'événements et un délai de censure, de rang $n_i + 1$.

Notre but étant d'obtenir des estimations non paramétriques convergentes des fonctions F_{1p} , fonctions de répartition des variables aléatoires $(X_{1p}, p \in \mathbb{N}^*)$, nous allons donc travailler à rang p fixé. La censure est unique pour chaque individu et vient masquer la série des événements récurrents à un rang donné, variable selon les individus. Son délai correspond à une variable aléatoire X_0 . Au rang p , pour chaque individu de l'échantillon, ou bien la censure a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent X_{1p} est alors censuré par le délai de censure X_0 , ou bien la censure est associée à un rang supérieur à p et le délai X_{1p} n'est pas censuré.

Il paraît donc naturel d'imposer le mécanisme de censure suivant au rang p (notations classiques d'Efron) :

$$\begin{cases} T_p = \min(X_0, X_{1p}) \\ J_p = \mathbf{1}(X_{1p} \leq X_0) \end{cases} \quad (2)$$

où T_p désigne le délai observé au rang p , et J_p est l'indicatrice d'événement ($J_p = 1$ si on observe un événement et $J_p = 0$ si l'on observe une censure).

De plus, nous supposerons que les délais d'événements et de censure sont indépendants, quel que soit le rang associé à ces délais.

Les observations utiles pour l'estimation au rang p sont donc de 2 sortes : les délais X_{1p} non censurés et les délais de censure X_0 associés à des rangs $k \leq p$. Notons qu'au rang p , chaque individu contribue une fois et une seule, si bien que l'on peut résumer les observations pertinentes au rang p par les couples $\{(T_i, J_i), i = 1, \dots, n\}$ (pour ne pas alourdir les notations, l'indice p sera omis lorsque l'indice i intervient).

Introduisons les notations suivantes au rang p :

$$\begin{aligned} U_p(t) &= \Pr(T_p \leq t), \\ U_{0p}(t) &= \Pr(T_p \leq t, J_p = 0), \\ U_{1p}(t) &= \Pr(T_p \leq t, J_p = 1). \end{aligned} \quad (3)$$

On a d'autre part :

$$\begin{aligned} F_{1p}(t) &= \Pr(T_p \leq t \mid J_p = 1), \\ F_{0p}(t) &= \Pr(T_p \leq t \mid J_p = 0). \end{aligned}$$

Les fonctions U_p , U_{0p} et U_{1p} sont facilement estimables empiriquement; F_{1p} et F_{0p} sont les fonctions d'intérêt à estimer. L'indépendance des processus d'événements et de censure ainsi que le mécanisme de censure (2) impliquent alors :

$$\Pr(T_p > t) = \Pr(X_0 > t) \Pr(X_{1p} > t), \quad (4)$$

c'est-à-dire

$$1 - U_p(t) = (1 - F_{0p}(t)) (1 - F_{1p}(t)). \quad (5)$$

En notant R la variable aléatoire correspondant au rang associé à un délai de censure, on obtient la décomposition suivante de la fonction F_{0p} :

$$F_{0p}(t) = \sum_{k=1}^p \Pr(T_k \leq t, R = k \mid J_k = 0) = \sum_{k=1}^p I_{0k}(t). \quad (6)$$

Cette décomposition nous suggère d'utiliser, par commodité d'expression et pour pouvoir replacer notre problème d'estimation dans un cadre déjà étudié, la notion de mélange : les délais de censure qui interviennent au rang p peuvent être considérés comme issus d'un mélange de délais de censure associés à des rangs $k \leq p$.

La relation (5) peut s'écrire :

$$1 - U_{0p}(t) - U_{1p}(t) = (1 - \sum_{k=1}^p I_{0k}(t)) (1 - F_{1p}(t)). \quad (7)$$

Elle correspond à la relation (11) de Derzko et Leconte (2004), caractéristique du «mécanisme de mélange censuré», introduit pour des événements compétitifs en présence d'une censure aléatoire à droite. Les risques compétitifs peuvent en effet être modélisés en considérant un mélange d'événements dont l'observation peut être masquée du fait de la censure à droite. La situation est simplement inversée dans le présent article puisque le mélange est un mélange de censures de rang $k \leq p$, censuré par un événement, mais une démarche similaire peut s'appliquer. En conséquence, les estimations des quantités qui nous intéressent, à savoir les fonctions F_{1p} et F_{0p} , découlent d'un algorithme simple présenté dans Derzko et Leconte (2004). La démarche est rappelée ci-après.

En imposant la condition de non-informativité de F_{0p} pour U_{1p} et de F_{1p} pour U_{0p} (exhaustivité de F_{0p} pour U_{0p} et de F_{1p} pour U_{1p}) :

$$\left(\frac{\partial U_{0p}(t)}{\partial F_{1p}(t)} \right) \equiv \left(\frac{\partial U_{1p}(t)}{\partial F_{0p}(t)} \right) \equiv 0, \quad (8)$$

on en déduit le système différentiel suivant (voir la démonstration dans Derzko et Leconte, 2004) :

$$\begin{cases} dF_{0p}(t) = \frac{dU_{0p}(t)}{(1 - F_{1p}(t))} \\ dF_{1p}(t) = \frac{dU_{1p}(t)}{(1 - F_{0p}(t))}. \end{cases} \quad (9)$$

Ce système différentiel peut s'intégrer numériquement, et il serait ainsi possible de dériver des estimateurs de $F_{1p}(t)$ et de $F_{0p}(t)$ pour p donné, à partir des estimations empiriques des fonctions U_{0p} et U_{1p} . Nous choisissons plutôt, comme dans Derzko et Leconte (2004), d'obtenir nos estimateurs via un algorithme simple.

L'algorithme est le suivant au rang p , pour $1 \leq p \leq r$, où r est le rang maximal observé dans l'échantillon. On considère les n couples $\{(t_i, j_i), i = 1, \dots, n\}$ impliqués au rang p , délais d'événements de rang p et délais de censures de rangs $k \leq p$, ordonnés dans l'ordre croissant des délais t_i .

Conditions initiales : $t_0 = 0, \hat{F}_{0p}(0) = \hat{F}_{1p}(0) = 0$.

Pour $i = 1, \dots, n$,

$$\begin{cases} \hat{F}_{0p}(t_i) = \hat{F}_{0p}(t_{i-1}) + \frac{d\widehat{U}_{0p}(t_i)}{1 - \hat{F}_{1p}(t_{i-1})} \\ \hat{F}_{1p}(t_i) = \hat{F}_{1p}(t_{i-1}) + \frac{d\widehat{U}_{1p}(t_i)}{1 - \hat{F}_{0p}(t_{i-1})}, \end{cases} \quad (10)$$

avec les estimations empiriques suivantes :

$$d\widehat{U}_{0p}(t_i) = \frac{\mathbf{1}_{[j_i=0]}}{n}, \quad d\widehat{U}_{1p}(t_i) = \frac{\mathbf{1}_{[j_i=1]}}{n}. \quad (11)$$

Comme l'ont montré Derzko et Leconte (2004), l'estimateur de F_{1p} ainsi obtenu est l'estimateur de Kaplan-Meier au rang p , tout individu censuré à un rang inférieur ou égal à p étant considéré comme censuré pour le rang p :

$$\hat{F}_{1p}(t) = 1 - \prod_{i:j_i=1} \left(1 - \frac{\mathbf{1}_{[t_i \leq t]}}{n - i + 1}\right). \quad (12)$$

Les estimations des fonctions $F_{0p}(t) = \sum_{k=1}^p I_{0k}(t)$, lois de mélanges de délais de censures de rangs inférieurs ou égaux à p , ne correspondent pas à des distributions d'intérêt. En revanche, F_{0r} correspond à la fonction de répartition de la censure, que nous noterons plus simplement F_0 . En effet, au rang r , rang maximal observé, il n'y a plus que des délais de censure, si bien que la relation caractéristique (5) s'écrit :

$$1 - U_{0r}(t) = 1 - F_{0r}(t), \quad (13)$$

et donc l'estimateur empirique de U_{0r} constitue un estimateur non paramétrique de F_0 , puisque la censure est entièrement observable : dans le contexte d'événements répétés, les événements n'empêchent pas la censure de se produire.

Un estimateur non paramétrique de l'incidence cumulée totale $E(N(t))$, espérance du nombre total d'événements répétés qui ont eu lieu avant t tous rangs confondus, s'obtient par sommation sur p (jusqu'au rang maximal $r - 1$ où on observe des événements) des estimateurs non paramétriques des incidences $F_p(t)$ à chaque rang (cf (1)) :

$$E(\widehat{N}(t)) = \sum_{p=1}^{r-1} \hat{F}_{1p}(t). \quad (14)$$

Comme les estimateurs à chaque rang p sont convergents (en tant qu'estimateurs de Kaplan-Meier), leur somme l'est également.

Les fonctions d'incidence peuvent aussi être interprétées dans le cadre d'un système à plusieurs états successifs ; l'estimation non paramétrique des fonctions de prévalence, définies ci-dessous, est également utile dans un tel cadre, qui peut être décrit de la façon suivante. En l'absence d'événement terminal, les individus d'une population initialement dans un état 0 passent aléatoirement au cours du temps dans l'état 1, dès l'instant où ils produisent le premier événement récurrent, avec la probabilité $F_{11}(t)$, puis de l'état 1 à l'état 2, dès l'instant où ils produisent le deuxième événement, avec la probabilité (non conditionnelle) $F_{12}(t)$. La prévalence $\pi_p(t)$ de l'état p dans ce système se définit comme la probabilité de cet état au temps t , c'est-à-dire :

$$\pi_p(t) = F_{1p}(t) - F_{1,p+1}(t), \quad 0 < p < r,$$

$$\pi_0(t) = 1 - F_{11}(t).$$

Ces prévalences sont donc estimables non paramétriquement à partir des estimations non paramétriques des fonctions $F_{1p}(t)$.

3.2. Structure des données : exemple d'école

On considère les $n = 6$ réalisations de la série des variables X_{1p} suivantes, représentées sur 6 lignes, comportant au plus $r - 1 = 2$ événements, et terminées par la censure (qui seule peut avoir le rang $r = 3$), présente à la fin de chaque série; les délais d'arrivées des événements et de la censure ont été pris tous distincts (et même entiers) pour une illustration simple; la présence d'ex æquo est un problème très général qui sera traité plus loin. Le symbole 0 désigne l'origine du temps, @ désigne un événement, et x désigne la censure; les délais d'arrivées de l'ensemble des individus (T) ont été ensuite interclassés, et on a associé à chacun son rang (R) à l'intérieur de sa série, et sa nature (J), événement ($J=1$) ou censure ($J=0$) :

```

0-----x
0-----@-----x
0-----@-----@---x
0-----@-----x
0--@-----@---x
0-----@-----x

```

Ensemble des individus :

```

0--@---x---@---@---x---@---@---x---x---@---x---x

```

T	1	2	3	4	5	6	7	8	9	10	11	12
R	1	1	1	2	3	1	1	2	1	2	3	2
J	1	0	1	1	0	1	1	0	0	1	0	0

Lorsque les délais sont distincts, ce qui se produit dans le cas de données simulées obtenues avec une grande précision, la totalité de l'information peut être disposée dans un tableau comportant autant de lignes que de délais et 3 colonnes, qui correspondent à la transposition des 3 lignes T, R et J ci-dessus. L'information peut aussi être stockée sous la forme du tableau suivant :

Temps	Événement			Censure		
	Rang 1	Rang 2	Rang 3	Rang 1	Rang 2	Rang 3
1	1	0	0	0	0	0
2	0	0	0	1	0	0
3	1	0	0	0	0	0
4	0	1	0	0	0	0
5	0	0	0	0	0	1
6	1	0	0	0	0	0
7	1	0	0	0	0	0
8	0	0	0	0	1	0
9	0	0	0	1	0	0
10	0	1	0	0	0	0
11	0	0	0	0	0	1
12	0	0	0	0	1	0

L'indicatrice d'événements de rang 3 est insérée pour la forme, car elle est naturellement nulle, toute série se terminant nécessairement par la censure. L'estimation non paramétrique présentée à la section précédente équivaut à la procédure suivante, d'une grande simplicité : construisons le tableau ci-dessous, obtenu à partir du tableau précédent en remplaçant les zéros par des 1 à droite de tout 1 dans les colonnes correspondant à la censure, de façon à prendre en compte le fait qu'une censure à un rang k va censurer tous les événements de rangs supérieurs ou égaux à k ; on obtient :

Temps	Événement			Censure		
	Rang 1	Rang 2	Rang 3	Rang 1	Rang 2	Rang 3
1	1	0	0	0	0	0
2	0	0	0	1	1	1
3	1	0	0	0	0	0
4	0	1	0	0	0	0
5	0	0	0	0	0	1
6	1	0	0	0	0	0
7	1	0	0	0	0	0
8	0	0	0	0	1	1
9	0	0	0	1	1	1
10	0	1	0	0	0	0
11	0	0	0	0	0	1
12	0	0	0	0	1	1

Du fait qu'à un rang donné, soit on observe un événement, soit on a observé une censure à ce rang ou à un rang inférieur, la somme des 1 obtenus pour les événements et pour la censure à un rang donné est égale à $n = 6$. L'estimation non paramétrique de l'incidence cumulée des événements au rang p s'obtient en appliquant la procédure de Kaplan-Meier à ce rang ; la procédure de Kaplan-Meier est donc appliquée séparément à chacun des 3 tableaux standard suivants, où la colonne J est une indicatrice :

Rang = 1	Rang = 2	Rang = 3
$\left(\begin{array}{c c} T & J \\ \hline 1 & 1 \\ 2 & 0 \\ 3 & 1 \\ 6 & 1 \\ 7 & 1 \\ 9 & 0 \end{array} \right)$	$\left(\begin{array}{c c} T & J \\ \hline 2 & 0 \\ 4 & 1 \\ 8 & 0 \\ 9 & 0 \\ 10 & 1 \\ 12 & 0 \end{array} \right)$	$\left(\begin{array}{c c} T & J \\ \hline 2 & 0 \\ 5 & 0 \\ 8 & 0 \\ 9 & 0 \\ 11 & 0 \\ 12 & 0 \end{array} \right)$

On observe qu'au rang 3, il n'y a aucune compétition puisque seule la censure est observée, et elle l'est complètement ; la loi de celle-ci peut donc être estimée

par la procédure empirique au rang 3. On obtient sur cet exemple les résultats suivants :

t_i	$E(\widehat{N}(t_i))$	$\widehat{F}_{11}(t_i)$	$\widehat{F}_{12}(t_i)$	$\widehat{F}_0(t_i)$
0	0.000	0.000	0.000	0.000
1	0.167	0.167	0.000	0.000
2	0.167	0.167	0.000	0.167
3	0.375	0.375	0.000	0.167
4	0.575	0.375	0.200	0.167
5	0.575	0.375	0.200	0.333
6	0.783	0.583	0.200	0.333
7	0.992	0.792	0.200	0.333
8	0.992	0.792	0.200	0.500
9	0.992	0.792	0.200	0.667
10	1.392	0.792	0.600	0.667
11	1.392	0.792	0.600	0.833
12	1.392	0.792	0.600	1.000

3.3. Simulations

Nous avons simulé des statistiques d'ordre d'un échantillon de variables de loi uniforme sur $(0, 1)$, avec une censure aléatoire à droite. Pour cela, nous générons $s = r - 1$ réalisations indépendantes issues de la loi uniforme sur $(0, 1)$, puis nous les ordonnons. Nous leur appliquons ensuite une censure, en prenant le minimum de chaque réalisation avec un même délai de censure généré aléatoirement. La loi de la p -ème statistique d'ordre de la loi uniforme sur $(0, 1)$ est une loi bêta de paramètres p et $s - p + 1$, de densité :

$$f_p(t) = s C_{s-1}^{p-1} t^{p-1} (1-t)^{s-p} \mathbf{I}_{[0 \leq t \leq 1]}.$$

On peut en déduire l'expression de l'incidence cumulée $E(N(t))$:

$$dE(N(t)) = \sum_{p=1}^s f_p(t) dt = s dt, \text{ d'où } E(N(t)) = st \text{ pour } t \in (0, 1).$$

Le processus ainsi généré peut être qualifié de processus binomial homogène, étant donné que le nombre de sauts de ce processus sur un intervalle de longueur t ($0 < t < 1$) suit une loi binomiale de paramètres s (nombre de statistiques d'ordre) et t . En comparaison au classique processus de Poisson, l'intérêt de ce modèle est que ce n'est pas un processus de renouvellement : en effet, les espacements entre deux récurrences sont identiquement distribués (de loi bêta de paramètres 1 et s), mais ils ne sont pas indépendants.

La figure 1 montre les estimations des incidences cumulées rang par rang, de l'incidence cumulée totale, de l'incidence cumulée de la censure, ainsi que les courbes de prévalences des états, pour $n = 1000$ individus, 4 récurrences maximum par individu ($r = 5$), avec une censure de loi uniforme sur $(0, 1)$.

On constate que les estimateurs donnent d'excellents résultats. On peut aussi noter que dans ce modèle, où le nombre de récurrences est fixé à 4, la prévalence de l'état 4 est égale à l'incidence cumulée des événements de rang 4 (l'état 4 est un état absorbant).

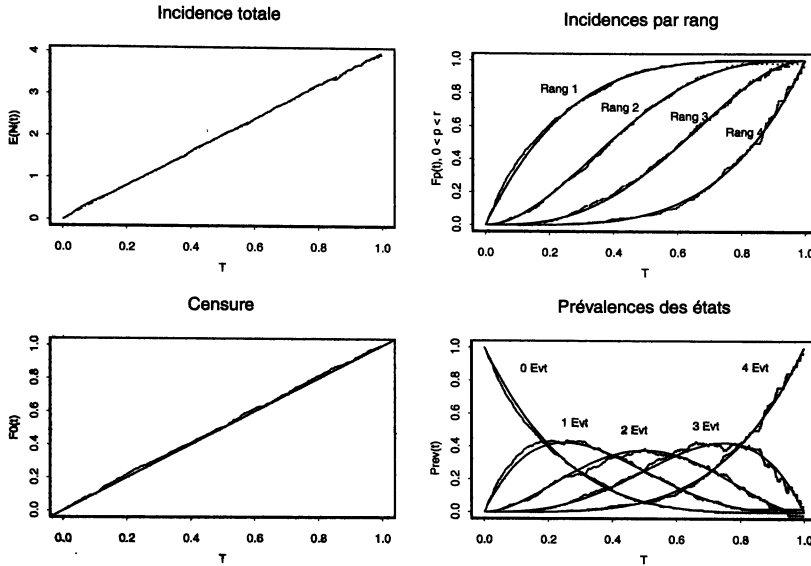


FIG 1. — Estimations des incidences cumulées et des prévalences pour un processus binomial homogène. (Les courbes les plus lisses sont les courbes théoriques. Pour les courbes des prévalences, les nombres d'événements correspondant à chacun des états ont été indiqués.)

3.4. Application biomédicale

Nous avons repris la classique étude des cancers de la vessie de Byar (1980), dont les données (*bladder*) sont disponibles en exemple dans le logiciel *S-Plus*, en considérant les deux bras placebo et thiotepa. Quand ils entrent dans l'essai, les 86 patients présentent des tumeurs de la vessie superficielles qui leur sont enlevées, puis ils sont randomisés dans les deux groupes de chimiothérapie (placebo ou thiotepa). Beaucoup de patients présentent des récurrences multiples de tumeurs au cours de l'étude, qui leur sont enlevées à chaque visite : 47 patients ont présenté une seule récurrence, 29 deux récurrences, 22 trois récurrences et 14 patients trois récurrences ou plus. Très peu de patients ont eu plus de 4 récurrences, donc seules les 4 premières récurrences ont été considérées.

Les estimations des incidences cumulées rang par rang et totale sont fournies dans la figure 2. On peut noter que, pour les 4 récurrences, les courbes des patients traités sont plus basses que celles des patients non traités.

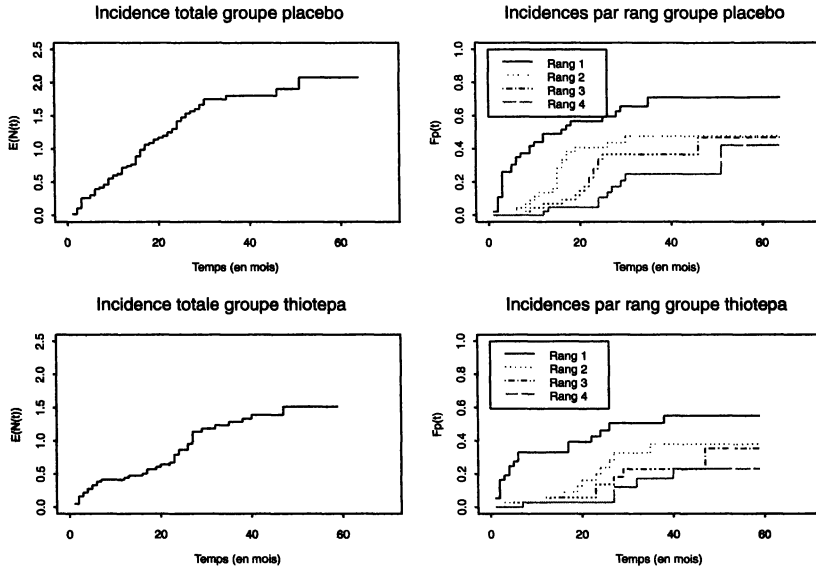


FIG 2. — Estimations des incidences cumulées totales et par rang pour l'essai Bladder.

4. Processus censuré tronqué par un événement terminal

4.1. Méthode d'estimation non paramétrique

Là encore, le lecteur pourra se reporter d'emblée au cas d'école traité dans la section suivante pour avoir une idée intuitive de la méthode.

Considérons à présent que les séries des délais des individus se terminent nécessairement soit par un *délai de perte* (la perte désigne l'événement terminal), au-delà duquel aucun événement ne peut plus se produire, soit par un *délai de censure*, au-delà duquel aucun événement ou perte n'est plus observable; nous attacherons à ces délais, qu'il s'agisse d'une perte ou d'une censure, un rang, celui du dernier événement d'intérêt observé augmenté d'une unité. L'individu i fournit donc n_i délais d'événements et un délai de censure ou un délai de perte, de rang $n_i + 1$.

Notre but est d'obtenir des estimations non paramétriques convergentes des incidences cumulées des événements rang par rang, en présence de l'événement terminal. Nous allons donc travailler à rang p fixé. En plus du délai de censure X_0 existe maintenant un délai de perte X_2 . La censure est unique pour chaque individu et vient masquer la série des événements récurrents à un rang donné, variable selon les individus, à moins que ce ne soit le délai de perte qui intervienne, en venant tronquer la série des événements récurrents à un rang donné, variable lui aussi selon les individus. Au rang p , pour chaque individu de l'échantillon, ou bien la censure a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent X_{1p} est alors censuré par le délai de censure X_0 , ou

bien la perte a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent X_{1p} est alors tronqué par le délai de perte X_2 , ou bien la censure ou la perte, dont l'une seule des deux peut être observée, est associée à un rang supérieur à p et le délai X_{1p} n'est ni censuré ni tronqué.

Nous allons supposer, parce que cette hypothèse paraît raisonnable dans beaucoup d'études, que les délais de survenue des événements ainsi que les délais de survenue des pertes sont indépendants du délai de censure. Par contre, il n'y a pas de raison de supposer l'indépendance du délai de perte et du délai jusqu'à l'événement d'intérêt. On peut donc modéliser par commodité d'expression la compétition qui se joue entre la perte et l'événement de rang p par un mélange (voir Derzko et Leconte, 2004). La variable résultante sera notée X_{12p} et sa fonction de répartition F_{12p} . Ces remarques conduisent à considérer le mécanisme suivant de censure et de troncature au rang p :

$$\begin{cases} T_p = \min(X_0, X_{12p}) \\ J_p = 0 \text{ si } X_0 < X_{12p} \\ J_p = 1 \text{ si } X_{12p} \leq X_0 \text{ et } X_{12p} = X_{1p} \\ J_p = 2 \text{ si } X_{12p} \leq X_0 \text{ et } X_{12p} = X_2 \end{cases} \quad (15)$$

où T_p désigne le délai observé au rang p , et J_p est le type de ce délai ($J_p = 1$ si on observe un événement, $J_p = 2$ si on observe une perte et $J_p = 0$ si on observe une censure).

Les observations utiles pour l'estimation au rang p sont donc de 3 sortes : les délais X_{1p} non censurés et non tronqués, les délais de perte X_2 associés à des rangs $k \leq p$ et les délais de censure X_0 associés à des rangs $k \leq p$. Notons qu'au rang p , chaque individu contribue une fois et une seule, si bien que l'on peut résumer les observations pertinentes au rang p par les couples $\{(T_i, J_i), i = 1, \dots, n\}$ (l'indice p sera omis lorsque l'indice i intervient).

En plus des fonctions U_p , U_{0p} , U_{1p} et F_{0p} déjà définies dans le cas sans perte (voir section 3.1), nous avons besoin des fonctions suivantes, définies pour tout rang p compris entre 1 et r , rang maximal observé dans l'échantillon :

$$\begin{aligned} U_{2p}(t) &= \Pr(T_p \leq t, J_p = 2), \\ I_{1p}(t) &= \Pr(T_p \leq t, J_p = 1 \mid J_p = 1 \text{ ou } 2), \\ I_{2p}(t) &= \Pr(T_p \leq t, J_p = 2 \mid J_p = 1 \text{ ou } 2), \\ F_{12p}(t) &= \Pr(T_p \leq t \mid J_p = 1 \text{ ou } 2) = I_{1p}(t) + I_{2p}(t). \end{aligned} \quad (16)$$

L'indépendance de X_{0p} et de X_{12p} pour tout p ainsi que la relation (15) conduisent, pour les mêmes raisons qu'à la section 3.1, à

$$\Pr(T_p > t) = \Pr(X_{12p} > t) \Pr(X_0 > t), \quad (17)$$

c'est-à-dire, pour tout rang p , $1 \leq p \leq r$:

$$\begin{aligned}
 1 - U_p(t) &= 1 - U_{0p}(t) - U_{1p}(t) - U_{2p}(t) \\
 &= (1 - F_{0p}(t)) (1 - F_{12p}(t)) \\
 &= (1 - F_{0p}(t)) (1 - I_{1p}(t) - I_{2p}(t)).
 \end{aligned} \tag{18}$$

En notant R la variable aléatoire correspondant au rang associé aux délais de censure ou de perte, la décomposition (6) de la fonction F_{0p} proposée dans la section sans perte est toujours valable.

Remarquons qu'une décomposition similaire au rang p s'applique à la fonction I_{2p} :

$$I_{2p}(t) = \sum_{k=1}^p \Pr(T_k \leq t, R = k, J_k = 2 \mid J_k = 1 \text{ ou } 2). \tag{19}$$

Les délais de perte qui interviennent au rang p peuvent donc être considérés comme issus d'un mélange de délais de perte associés à des rangs $k \leq p$.

Nous avons donc au rang p un mécanisme de mélange censuré : un mélange constitué de l'événement d'intérêt de rang p et de délais de pertes de rangs $\leq p$ est censuré par minimum avec un mélange de délais de censures de rangs $\leq p$. Nous pouvons alors appliquer l'algorithme correspondant à ce cas proposé dans Derzko et Leconte (2004).

Le processus satisfait, dans les conditions habituelles d'exhaustivité de F_{0p} pour U_{0p} , de I_{1p} pour U_{1p} et de I_{2p} pour U_{2p} , les équations différentielles suivantes :

$$\begin{aligned}
 dU_{0p}(t) &= (1 - I_{1p}(t) - I_{2p}(t)) dF_{0p}(t), \\
 dU_{1p}(t) &= (1 - F_{0p}(t)) dI_{1p}(t), \\
 dU_{2p}(t) &= (1 - F_{0p}(t)) dI_{2p}(t).
 \end{aligned} \tag{20}$$

À partir des estimations empiriques de dU_{0p} , dU_{1p} et dU_{2p} , on peut donc déduire des estimations de I_{1p} , I_{2p} et F_{0p} par l'algorithme suivant au rang p , pour tout p compris entre 1 et r , rang maximal observé dans l'échantillon. On considère les n couples $\{(t_i, j_i), i = 1, \dots, n\}$ impliqués au rang p , délais d'événements de rang p et délais de pertes et de censures de rang $k \leq p$, ordonnés dans l'ordre croissant des délais t_i .

Conditions initiales : $t_0 = 0, \hat{F}_{0p}(0) = \hat{I}_{1p}(0) = \hat{I}_{2p}(0) = \hat{F}_{12p}(0) = 0$.

Pour $i = 1, \dots, n$,

$$\begin{aligned}\hat{I}_{1p}(t_i) &= \hat{I}_{1p}(t_{i-1}) + \frac{d\widehat{U}_{1p}(t_i)}{(1 - \widehat{F}_{0p}(t_{i-1}))} \\ \hat{I}_{2p}(t_i) &= \hat{I}_{2p}(t_{i-1}) + \frac{d\widehat{U}_{2p}(t_i)}{(1 - \widehat{F}_{0p}(t_{i-1}))} \\ \widehat{F}_{0p}(t_i) &= \widehat{F}_{0p}(t_{i-1}) + \frac{d\widehat{U}_{0p}(t_i)}{(1 - \widehat{F}_{12p}(t_{i-1}))} \\ \widehat{F}_{12p}(t_i) &= \hat{I}_{1p}(t_i) + \hat{I}_{2p}(t_i)\end{aligned}\tag{21}$$

avec les estimations empiriques suivantes :

$$d\widehat{U}_{0p}(t_i) = \frac{\mathbf{I}_{[j_i=0]}}{n}, \quad d\widehat{U}_{1p}(t_i) = \frac{\mathbf{I}_{[j_i=1]}}{n}, \quad d\widehat{U}_{2p}(t_i) = \frac{\mathbf{I}_{[j_i=2]}}{n}.$$

En se référant à Derzko et Leconte (2004), on peut montrer facilement que l'estimateur obtenu pour F_{12p} est l'estimateur de Kaplan-Meier :

$$\widehat{F}_{12p}^{KM}(t) = 1 - \prod_{i:j_i \in \{1,2\}} \left(1 - \frac{\mathbf{I}_{[t_i \leq t]}}{n - i + 1} \right).\tag{22}$$

L'estimateur obtenu pour I_{1p} est l'estimateur de Prentice (Kalbfleisch et Prentice, 1980) :

$$\hat{I}_{1p}^{Pr}(t) = \sum_{i:j_i=1} \frac{\mathbf{I}_{[t_i \leq t]}}{n - i + 1} \left(1 - \widehat{F}_{12p}^{KM}(t_i) \right).$$

Les fonctions F_{0p} et I_{2p} , lois de mélanges de délais de censure et de mélanges de délais de pertes de rangs inférieurs ou égaux à p , ne correspondent pas à des distributions d'intérêt ; en revanche, les estimateurs de F_{0r} et de I_{2r} constituent des estimateurs non paramétriques de F_0 , fonction de répartition de la censure et de F_2 , fonction de répartition de la perte. En effet, il est aisé de constater qu'au rang r , il ne peut plus y avoir d'événement récurrent, et donc la relation caractéristique (18) s'écrit :

$$\begin{aligned}1 - U_{0r}(t) - U_{2r}(t) &= (1 - F_{0r}(t))(1 - I_{2r}(t)) \\ &= (1 - F_0(t)) (1 - F_2(t)).\end{aligned}$$

On voit donc que la censure et la perte se censurent mutuellement, et la procédure d'estimation de F_0 et de F_2 est donc la procédure habituelle de Kaplan-Meier, appliquée à l'échantillon de l'ensemble des pertes et des censures :

$$\hat{F}_0^{KM}(t) = 1 - \prod_{i:j_i=0} \left(1 - \frac{\mathbf{1}_{[t_i \leq t]}}{n - \ell(i, r) + 1} \right),$$

$$\hat{F}_2^{KM}(t) = 1 - \prod_{i:j_i=2} \left(1 - \frac{\mathbf{1}_{[t_i \leq t]}}{n - \ell(i, r) + 1} \right),$$

où $\ell(i, r)$ désigne le rang de l'observation i dans l'échantillon constitué de l'ensemble des pertes et des censures à tous les rangs.

Un estimateur de l'incidence cumulée totale des événements d'intérêt, tous rangs confondus, s'obtient en sommant les estimateurs des incidences à chaque rang p :

$$E(\widehat{N}(t)) = \sum_{p=1}^{r-1} \hat{I}_{1p}(t).$$

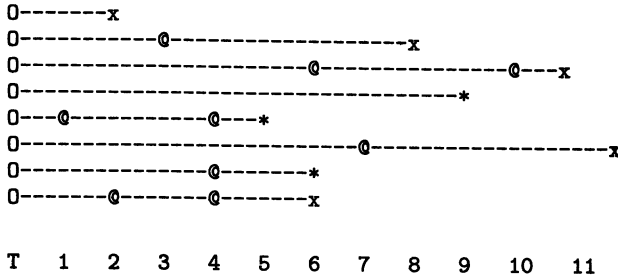
Pour ce qui est de la prévalence associée aux événements de rang p , en présence d'événements terminaux, les individus qui sont entrés dans l'état p avec une probabilité $I_{1p}(t)$ peuvent à l'instant t soit produire un $(p+1)$ -ème événement récurrent, avec la probabilité $I_{1,p+1}(t)$ qui les fait passer dans l'état suivant $p+1$, soit produire un événement terminal de rang $p+1$ avec la probabilité $\Pr(T \leq t, R = p+1, J = 2 \mid J = 1 \text{ ou } 2) = I_{2,p+1}(t) - I_{2p}(t)$. D'où :

$$\begin{aligned} \pi_p(t) &= I_{1p}(t) - I_{1,p+1}(t) - (I_{2,p+1}(t) - I_{2p}(t)) \\ &= F_{12p}(t) - F_{12,p+1}(t), \quad 0 < p < r \\ \pi_0(t) &= 1 - F_{12,1}(t). \end{aligned} \tag{23}$$

4.2. Structure des données : exemple d'école

On considère les $n = 8$ réalisations de la série des variables X_{1p} suivantes, représentées sur 8 lignes, comportant au plus $r - 1 = 2$ événements, et terminées soit par la censure, soit par une perte (qui seules peuvent avoir le rang $r = 3$), présentes à la fin de chaque série; les délais d'arrivées des événements, de la perte et/ou de la censure ne sont pas tous distincts, pour illustrer le traitement des ex æquo; le symbole * désigne la perte ($J = 2$) :

ESTIMATION NON PARAMÉTRIQUE – APPLICATIONS BIOMÉDICALES



Aux délais d'arrivées $T = 2$, $T = 4$ et $T = 6$, se produisent 2 ou 3 événements, censures ou pertes, qui de plus correspondent à des rangs distincts ; la totalité de l'information peut être disposée dans un tableau unique, comportant autant de lignes que de délais d'arrivées éventuellement égaux et 3 colonnes, à condition d'imposer un *ordre d'arrivée* pour les délais d'arrivées égaux. Par exemple, on peut convenir qu'un événement récurrent de rang j ou une perte de rang j se produisent avant une censure de rang j et qu'un événement, une perte ou une censure de rang j se produisent après un événement, une perte ou une censure de rang $j' < j$. On obtient ainsi le tableau suivant :

T	1	2	2	3	4	4	4	5	6	6	6	7	8	9	10	11	12
R	1	1	1	1	1	2	2	3	1	2	3	1	2	1	2	3	2
J	1	1	0	1	1	1	1	2	1	2	0	1	0	2	1	0	0

L'information peut aussi être mise sous la forme du tableau suivant, où tous les temps en première colonne sont distincts, et dans lequel ce ne sont plus des indicatrices (0/1) qui apparaissent comme dans le cas sans ex æquo, mais des fréquences absolues :

Temps	Événement			Perte			Censure		
	R 1	R 2	R 3	R 1	R 2	R 3	R 1	R 2	R 3
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	0	0
4	1	2	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0
6	1	0	0	0	1	0	0	0	1
7	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	1	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	1	0

ESTIMATION NON PARAMÉTRIQUE – APPLICATIONS BIOMÉDICALES

L'estimation non paramétrique équivaut à la procédure suivante : construisons le tableau suivant, obtenu à partir du tableau précédent en cumulant (séparément) vers la droite les fréquences qui apparaissent dans les colonnes correspondant à la perte et les colonnes correspondant à la censure :

Temps	Événement			Perte			Censure		
	R 1	R 2	R 3	R 1	R 2	R 3	R 1	R 2	R 3
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	1	1
3	1	0	0	0	0	0	0	0	0
4	1	2	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0
6	1	0	0	0	1	1	0	0	1
7	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	1
9	0	0	0	1	1	1	0	0	0
10	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	1	1

Remarquons que la somme des fréquences obtenues pour les événements, les pertes et les censures à un rang donné est égale à $n=8$, ce qui traduit le fait qu'à un rang donné, soit on observe un événement, soit on a observé une censure à ce rang ou à un rang inférieur, soit on a observé une perte à ce rang ou à un rang inférieur. L'estimation non paramétrique de la distribution des événements au rang p s'obtient en appliquant la procédure de Prentice à ce rang, c'est-à-dire séparément à chacun des 3 tableaux standard suivants :

Rang = 1				Rang = 2				Rang = 3			
$\left(\begin{array}{c c c c} T & E & P & C \\ \hline 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 \\ 3 & 1 & 0 & 0 \\ 4 & 1 & 0 & 0 \\ 6 & 1 & 0 & 0 \\ 7 & 1 & 0 & 0 \\ 9 & 0 & 1 & 0 \end{array} \right)$	$\left(\begin{array}{c c c c} T & E & P & C \\ \hline 2 & 0 & 0 & 1 \\ 4 & 2 & 0 & 0 \\ 6 & 0 & 1 & 0 \\ 8 & 0 & 0 & 1 \\ 9 & 0 & 1 & 0 \\ 10 & 1 & 0 & 0 \\ 12 & 0 & 0 & 1 \end{array} \right)$	$\left(\begin{array}{c c c c} T & E & P & C \\ \hline 2 & 0 & 0 & 1 \\ 5 & 0 & 1 & 0 \\ 6 & 0 & 1 & 1 \\ 8 & 0 & 0 & 1 \\ 9 & 0 & 1 & 0 \\ 11 & 0 & 0 & 1 \\ 12 & 0 & 0 & 1 \end{array} \right)$									

Dans ces tableaux, on a dédié une colonne (E) aux événements, une colonne (P) à la perte, et une colonne (C) à la censure, afin de permettre le stockage de fréquences, et non plus d'indicatrices. L'estimation non paramétrique utilise l'hypothèse d'indépendance entre la censure et l'événement d'une part, la censure et la perte d'autre part, mais ne suppose pas l'indépendance entre l'événement et la perte (modèle de mélange). La colonne des événements de rang 3 est remplie de fréquences nulles, puisqu'un événement a au plus le rang 2, si bien que le tableau de rang 3 se réduit à une compétition par minimum

entre perte et censure, ce qui permet d'estimer les distributions de la perte et de la censure en appliquant la procédure de Kaplan-Meier au rang 3.

Des programmes S-Plus calculant les estimations des incidences cumulées (avec ou sans événement terminal) sont disponibles sur la page Web de E. Leconte (<http://www.univ-tlse1.fr/GREMAQ/Statistique/Eveweb>).

Pour illustrer les résultats en présence de différents types d'événements, les données ont été modifiées ainsi : les événements au temps 2 et au temps 7 ont été classés d'un type différent des autres ; dans le tableau de résultats suivant, on trouve donc en colonne 1 les temps (distincts) d'observation, ordonnés de façon croissante, en colonne 2 l'incidence totale du premier type d'événements, en colonne 3 l'incidence totale du deuxième type d'événement, en colonne 4 l'incidence de la perte, et enfin celle de la censure en colonne 5.

t_i	$E(\widehat{N}_1(t_i))$	$E(\widehat{N}_2(t_i))$	$\widehat{F}_2(t_i)$	$\widehat{F}_0(t_i)$
1	0.125	0.000	0.000	0.000
2	0.125	0.125	0.000	0.125
3	0.275	0.125	0.000	0.125
4	0.711	0.125	0.000	0.125
5	0.711	0.125	0.143	0.125
6	0.861	0.125	0.286	0.300
7	0.861	0.275	0.286	0.300
8	0.861	0.275	0.286	0.475
9	0.861	0.275	0.524	0.475
10	1.051	0.275	0.524	0.475
11	1.051	0.275	0.524	0.738
12	1.051	0.275	0.524	1.000

4.3. Simulations

Nous simulons des statistiques d'ordre d'un échantillon de variables de loi uniforme sur $(0, 1)$ avec une censure à droite et une probabilité de perte constante à chaque rang. Nous pouvons qualifier le processus correspondant de binomial homogène (voir section 3.3) avec perte constante. Pour cela, comme dans le cas sans perte, nous générons pour chaque individu $s = r - 1$ réalisations de variables aléatoires indépendantes uniformes sur $(0, 1)$ que nous ordonnons. A chacune de ces variables est associée une variable de Bernoulli de paramètre μ correspondant à la probabilité qu'il n'y ait pas de perte. La série des s récurrences d'un individu est stoppée dès la première perte, qui prend alors le rang de l'événement précédent augmenté d'une unité. A cela s'ajoute une censure en compétition par minimum avec chaque terme de la série, y compris avec le délai de perte.

En notant $f_p(t)$ la densité de la loi béta de paramètres p et $s - p + 1$ (voir section 3.3), nous avons :

$$\begin{aligned} dE(N(t)) &= \sum_{p=1}^s \mu^p f_p(t) dt \\ &= s \mu ((1-t) + t \mu)^{s-1} dt, \end{aligned}$$

$$\text{d'où } E(N(t)) = \frac{\mu}{1-\mu} \left(1 - (1-t(1-\mu))^s \right).$$

Pour l'incidence cumulée de l'événement terminal, nous avons :

$$\begin{aligned} dF_2(t) &= (1-\mu) \sum_{p=1}^s \mu^{p-1} f_p(t) dt \\ &= s(1-\mu) ((1-t) + t \mu)^{s-1} dt \end{aligned}$$

$$\text{d'où } F_2(t) = 1 - (1-t(1-\mu))^s.$$

La prévalence de l'état p , pour $0 < p < s$, vaut (en notant $F_p(t)$ la fonction de répartition de la loi béta de paramètres p et $s - p + 1$) :

$$\begin{aligned} \pi_p(t) &= \mu^p F_p(t) - \mu^{p+1} F_{p+1}(t) \\ &- (1-\mu) \left(\sum_{m=1}^{p+1} \mu^{m-1} F_m(t) - \sum_{m=1}^p \mu^{m-1} F_m(t) \right) \\ &= \mu^p \left(F_p(t) - F_{p+1}(t) \right). \end{aligned}$$

Pour $p = 0$, on a $\pi_0(t) = 1 - F_1(t)$ et pour $p = s$, $\pi_s(t) = \mu^s F_s(t)$ (dans ce modèle, l'état récurrent s est absorbant car le nombre de récurrences est limité à s).

Les courbes estimées sont sur la figure 3, pour $n = 1000$ individus, 3 récurrences maximum par individu ($r = 4$), avec une censure de loi uniforme sur $(0, 1)$ et une probabilité de perte $1 - \mu = 0.2$. On constate que les estimateurs donnent là encore d'excellents résultats. Comme élément de comparaison de notre méthode avec des propositions précédentes, l'estimateur de l'incidence cumulée globale de Cook et Lawless (1997) a également été appliqué à ces données. Bien que le calcul de leur estimateur soit formellement différent du nôtre, la courbe obtenue est pratiquement confondue avec celle de notre estimateur de l'indidence totale et nous ne l'avons donc pas représentée. Notons cependant que la méthode de ces auteurs ne permet pas l'estimation d'incidences par rang, ni donc l'estimation des prévalences.

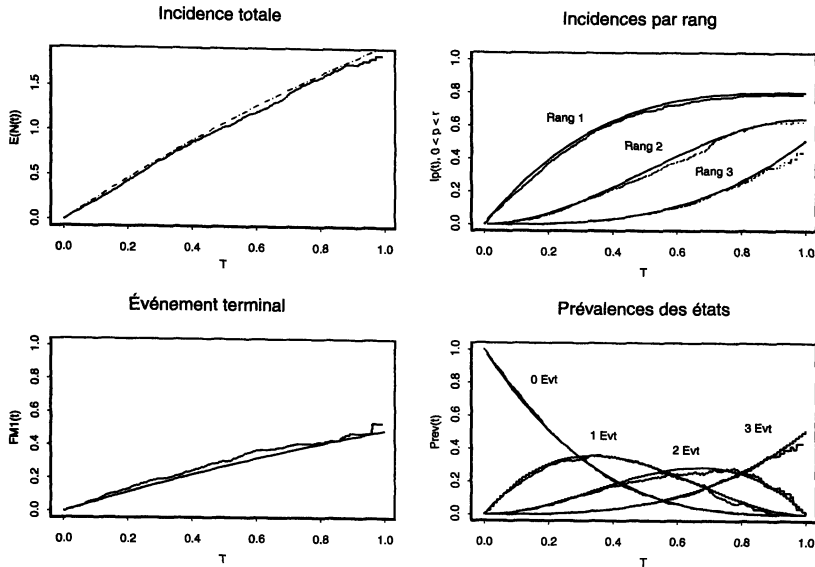


FIG 3. — Estimations des incidences cumulées et des prévalences pour un processus binomial homogène censuré avec perte. (Les courbes les plus lisses sont les courbes théoriques. Pour les courbes des prévalences, les nombres d'événements correspondant à chacun des états ont été indiqués.)

4.4. Application biomédicale

Nous ré-analisons l'étude CAPRIE (Clopidogrel versus Aspirine in Patients at Risk of Ischaemic Events) du CAPRIE Steering Committee (1996), un essai thérapeutique destiné à montrer la supériorité du clopidogrel sur l'aspirine dans la prévention des événements ischémiques. Environ 20 000 patients ont été suivis de 1 à 3 ans, et jusqu'à 3 récurrences ont été observées. L'analyse selon le protocole a porté sur un critère composite, à savoir le premier événement ischémique, fatal ou non. Les courbes de survie par groupe de traitement pour ce critère ont été estimées par la méthode de Kaplan-Meier et on obtient à l'aide du test du logrank une différence significative, qui atteint en différence relative 8 % à 3 ans en faveur du clopidogrel.

Nous reprenons cette étude en considérant tous les événements récurrents. Nous présentons à la figure 4 les courbes d'incidence cumulée par groupe des patients de la strate PAD (Peripheral Arterial Disease); il s'agit des patients inclus dans l'étude avec une artériopathie des membres inférieurs, qui les prédispose à produire des événements cardio-vasculaires (CV) graves, dont des infarctus du myocarde, des infarctus cérébraux ou encore à devoir être amputés du membre inférieur atteint par la maladie; tous ces événements peuvent être éventuellement fatals. Dans cet exemple sont considérés comme événements récurrents les événements cardio-vasculaires non fatals, les événements cardio-vasculaires fatals constituant quant à eux des événements terminaux que nous avons regroupé sous le terme « décès toutes causes ». Il y a environ



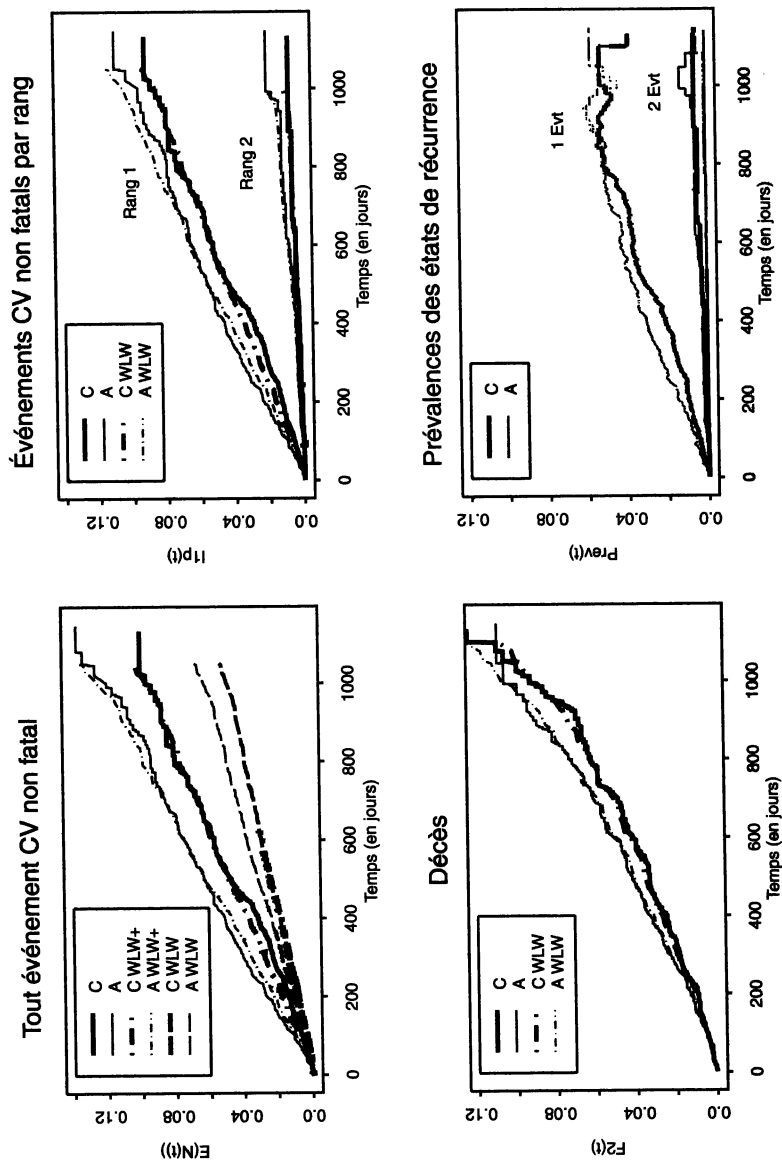


FIG 4. — Estimations des incidences cumulées et des prévalences pour l'étude CAPRIE (A = aspirine, C = clopidogrel; les courbes WLW+ sont obtenues en sommant les courbes WLW des rangs 1 et 2).

3 300 patients dans chaque groupe. On constate que la supériorité du clopidogrel sur l'aspirine s'exprime à la fois vis-à-vis des événements terminaux et des événements récurrents, globalement et rang par rang. Les patients ayant reçu le clopidogrel ont eu au plus deux récurrences; quelques patients ayant reçu l'aspirine en ont eu trois.

Nous avons appliqué aux mêmes données le modèle de Cox généralisé de Wei *et al.* (noté WLW). Seules les deux premières récurrences ont pu être considérées, puisque le groupe ayant reçu le clopidogrel a eu deux récurrences au plus. Le décès a été traité comme une strate à part, et les événements récurrents tronqués sont considérés comme censurés. Les risques relatifs estimés sont les suivants : 1,23 (degré de signification du test de Wald : $p = 0,03$) pour la première récurrence, 2,02 ($p = 0,02$) pour la deuxième récurrence, et 1,17 ($p = 0,14$) pour le décès, toujours en faveur du clopidogrel. Nous constatons sur les courbes de la figure 4 (en haut à droite) que les estimations des incidences rang par rang obtenues sont assez proches de nos estimations non paramétriques. De même, pour le décès, les courbes sont très proches. En ce qui concerne l'incidence globale, les estimations tous rangs confondus ont été calculées comme ce qui est fait classiquement par la méthode WLW en considérant un modèle de Cox avec paramètre unique pour prendre en compte l'effet du traitement sur les deux récurrences. Les courbes correspondantes sont les plus basses de la figure 4 en haut à gauche. Elles sont en quelque sorte une moyenne des courbes des rangs 1 et 2. Or, comme le montre la formule (1), la façon correcte d'obtenir une courbe d'incidence globale est de sommer les courbes des différents rangs, ce que nous avons également fait en sommant les courbes des rangs 1 et 2 de la méthode WLW. Nous constatons alors sur les courbes de la figure 4 en haut à gauche que les estimations obtenues sont assez proches de nos estimations non paramétriques. Sur ce même graphe, correspondant à tous les événements cardio-vasculaires non fatals, nos estimations montrent cependant une tendance : l'écart entre les deux groupes se resserre après 450 jours. Cette différence précoce est gommée par le modèle WLW, du fait de l'hypothèse de proportionnalité. Il se peut cependant que la tendance révélée par notre méthode soit réelle et que l'évolution ultérieure moins différenciée des courbes provienne des arrêts de traitement : en effet, les patients arrêtant le clopidogrel sont mis sous aspirine, mais l'analyse est faite en intention de traiter¹. Cela mériterait une analyse plus fine. Nous avons également appliqué la méthode de Cook et Lawless (1997) de calcul de l'incidence globale aux données CAPRIE. Comme pour les simulations, les courbes d'incidence globale obtenues sont extrêmement proches des nôtres, à un point tel que nous ne les avons pas représentées. Rappelons que cette dernière méthode ne comporte pas d'estimation des incidences rang par rang, et exclut donc le calcul d'estimation des prévalences.

5. Conclusion

Les situations expérimentales décrites dans cet article sont très fréquemment rencontrées, entre autres, dans le domaine de la recherche médicale. Pourtant,

1. On parle d'analyse en intention de traiter lorsque des patients ayant abandonné le traitement que leur a affecté la randomisation restent considérés pour l'analyse dans leur groupe de traitement d'origine, quels que soient les traitements réellement administrés par la suite. C'est ce qui est fait habituellement dans les études cliniques de phase III.

les méthodes d'estimation utilisées en pratique ignorent une partie de l'information disponible dans l'échantillon, ou modélisent de façon inappropriée les événements terminaux comme s'il s'agissait de censures à droite indépendantes. Nous proposons des procédures d'estimation convergentes et très simples de mise en œuvre, qui révèlent, sur des simulations de modèles pour lesquels la solution exacte est connue, une remarquable efficacité. Le calcul fait usage uniquement d'estimateurs non paramétriques bien connus appliqués à des modèles appropriés.

On peut aussi noter que les algorithmes présentés, que ce soit avec ou sans événement terminal, fournissent *ipso facto* des estimateurs non paramétriques pour les densités, les risques instantanés et cumulés. De plus, ces algorithmes s'appliquent de façon similaire lorsque les estimations empiriques de la formule (11) sont remplacées par des lisseurs discrets, par exemple l'histogramme lissé à noyau (Deheuvels *et al.*, 2004) : les estimateurs des incidences cumulés obtenus sont alors eux aussi lissés.

Il faut souligner l'intérêt d'une méthode d'estimation qui fournit non seulement l'incidence totale, mais aussi les incidences par rang ; en effet, cela permet d'estimer les fonctions de prévalences des états dans le système multi-états que l'on peut associer d'une façon standard en identifiant l'incidence cumulée des événements de rang p avec l'intensité de transition cumulée de l'état $p - 1$ vers l'état p . Les systèmes multi-états introduits ici sont les plus simples dans la mesure où nous n'avons traité que le cas où les événements récurrents sont d'un seul type.

Un autre intérêt d'avoir une méthode non paramétrique disponible est de permettre la validation de modèles plus restrictifs. En particulier, l'hypothèse de proportionnalité des risques des modèles de Cox généralisés peut être graphiquement validée en comparant les estimations des incidences par groupe de ces modèles à nos estimations non paramétriques. Par exemple, les données CAPRIE semblent assez bien se prêter à une analyse par un modèle de risques proportionnels.

La méthode d'estimation proposée s'étend de façon immédiate à tout mélange d'événements récurrents de différents types, permettant l'estimation non paramétrique de l'incidence d'un type particulier d'événement, comme nous l'avons illustré dans le cas d'école avec perte. Si les événements terminaux sont également catégorisés, on sait que le modèle de mélange censuré permet d'estimer l'incidence d'une de leurs catégories. Cela permet par exemple d'estimer l'incidence des infarctus mortels ou non mortels et récurrents, en présence d'autres événements mortels ou non mortels et récurrents (par exemple des hospitalisations ou des décès pour une cause non cardio-vasculaire).

Références

- ANDERSEN P.K. et GILL R.D. (1982). Cox's regression model for counting processes : a large sample study, *The Annals of Statistics* **10**, 1100–1120.
- BYAR D. P. (1980). The Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumors : Comparisons of Placebo, Pyridoxine, and Topical Thiotepa. In : Pavone-Macaluso M., Smith P. H., Edsmyn F., eds *Bladder Tumors and Other Topics in Urological Oncology*. New-York, NY : Plenum, 363–370.
- CAPRIE STEERING COMMITTEE (1996). A randomised, blinded trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE), *Lancet* **348**(9038), 1329–1339.
- COOK R. J. et LAWLESS J. F. (1997). Marginal analysis of recurrent events and a terminating event, *Statistics in Medicine* **16**, 911–924.
- DEHEUVELS P., DERZKO G. et DEVROYE L. (2004). Kernel Smoothed Histograms. In : Antonov, V., Huber, C., Nikulin, M., Polischok, V., eds *LAD'2004*. Saint-Petersbourg, Russie : **2**, 68–82.
- DERZKO G. et LECONTE E. (2004). Estimation non paramétrique d'incidences d'événements en compétition avec censure à droite, *Journal de la Société Française de Statistique* **145**, 1, 47–69 .
- GOSH D. et LIN D. Y. (2000). Nonparametric analysis of recurrent events and deaths, *Biometrics* **56**, 554–562.
- KALBFLEISCH J. D. et PRENTICE R. L. (1980). *The statistical analysis of failure time data*. John Wiley & Sons, New York.
- KAPLAN E. L. et MEIER P. (1958). Nonparametric estimation from incomplete observations, *JASA* **53**, 457–481.
- LAWLESS J. F. et NADEAU C. (1995). Some simple robust methods for the analysis of recurrent events, *Technometrics* **37**, 158–168.
- MENJOGE S. S. (2003). On estimation of frequency data with censored observations, *Pharmaceutical Statistics* **2**, 191–197.
- PRENTICE R. L., WILLIAMS B. J. et PETERSON A.V. (1981). On the regression analysis of multivariate failure time data, *Biometrics* **68**, 373–79.
- WEI L. J., LIN Y. et WEISSFELD L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *J. Am. Statist. Assoc.* **84**, 1065–1073.