JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

CATHERINE HUBER-CAROL

# Durées de survie tronquées et censurées

*Journal de la société statistique de Paris*, tome 135, n° 4 (1994), p. 3-23

<http://www.numdam.org/item?id=JSFS_1994__135_4_3_0>

I

COMMUNICATION

# DURÉES DE SURVIE TRONQUÉES ET CENSURÉES

par Catherine HUBER-CAROL
*Professeur, Université René Descartes*[1]

## 1. Introduction

Le problème des données manquantes, incomplètes ou erronées est très vaste et a suscité beaucoup d'intérêt parmi les statisticiens ces dernières années. L'attitude vis-à-vis de ce type de données a longtemps été soit de les éliminer, soit de minimiser le mauvais impact qu'elles pourraient avoir sur des procédures statistiques adaptées à des données complètes. Dans le domaine des durées de survie, les données sont souvent incomplètes à cause de deux phénomènes distincts : la censure et la troncature.

Dans leur acception la plus générale, ces deux notions ont la signification suivante. Premièrement, et c'est la troncature, on n'observe $X$ que si elle appartient à un sous-ensemble $B$ de ses valeurs possibles. On dit que $X$ est tronquée par $B$. Deuxièmement, et il s'agit là de la censure, même dans le cas où $X$ appartient à $B$, on n'observe pas $X$ complètement ; on sait seulement de cette variable qu'elle appartient à un sous-ensemble $A$ de $B$. On dit qu'elle est censurée par $A$. Dans le cas où un $n$-échantillon de $X$, soit $X_1, ..., X_n$, est concerné, à chacun des $X_i$ sont associés deux ensembles $B_i$ et $A_i$, le premier qui tronque $X_i$ et le second qui le censure. La plupart du temps, $X_i$ est une variable réelle positive, et $A_i$ et $B_i$ sont des demi-droites du type $(-\infty, c_i)$ ou $(c_i + \infty)$, ce qui correspond à des censures ou troncatures dites gauches ou droites, et c'est donc ce cas qui sera illustré par des exemples au paragraphe suivant. Pour des censures et troncatures par intervalles, on pourra voir Turnbull (1976), Groeneboom (1992), Frydman (1994) et Alioum et Commenges (1994).

Par ailleurs, la modélisation la plus couramment adoptée pour les durées de survie est de type semi-paramétrique, car elle a la souplesse du non paramétrique en incluant dans le modèle une (ou plusieurs) fonction inconnue, tout en permettant de disposer de *paramètres* pour pouvoir finalement *interpréter* les résultats, par exemple l'effet diagnostique ou l'effet pronostique d'un facteur. Le plus utilisé de ces modèles est le *modèle de Cox*, mais il en existe d'autres comme le modèle de survie accéléré, ou des variantes du modèle de Cox utilisant une approche bayésienne.

---

1. UFR Biomédicale, 45, rue des Saints-Pères, 75270 PARIS Cedex 06. Conférence prononcée le 16 mars 1993.

Le plan de l'exposé est le suivant. Au paragraphe 2, la différence entre les effets de la troncature et ceux de la censure est explicitée et des exemples de censures aléatoires droites et gauches sont donnés. Au paragraphe 3, on présente un exemple motivant de durées tronquées à droite, le temps d'incubation du SIDA pour des patients transfusés. Au paragraphe 4, on traite l'estimation non paramétrique de la fonction de survie pour des données continues tronquées à gauche, qui est l'équivalent de l'estimateur de Kaplan-Meier pour des données censurées à droite. L'estimation est appliquée à l'exemple d'une loi uniforme. Enfin au paragraphe 5, un modèle de régression pour des durées discrètes tronquées à droite est donné et appliqué à l'exemple du SIDA acquis par transfusion.

Un ouvrage récent qui fait autorité sur le sujet est le livre de Andersen, Borgan, Gill, Keiding (1992) cité en référence. On peut aussi trouver une revue sur le sujet en français dans le livre de la série Economica *Analyse statistique des durées de vie* : « modélisation des données censurées » (1989), ou encore, dans un registre plus proche des applications médicales, dans le livre *Analyse statistique des données de survie* de C. Hill et coll. (1990), dans la série « Statistique en biologie et en médecine ». On peut citer aussi *Survival Analysis: State of the Art*, écrit par J. D. Klein et P. K. Goel (1991).

## 2. Exemples de censure et troncature droites ou gauches

### a) Censure aléatoire droite ou gauche

Une durée de vie aléatoire $X$ est dite censurée par une variable aléatoire de censure $C$ si on observe parfois $C$ au lieu de $X$. L'information donnée par $C$ sur $X$ est :

$X > C$ s'il y a *censure droite*
$X < C$ s'il y a *censure gauche*

*Exemple de censure droite :*

Un exemple classique de censure droite est celui où l'étude porte sur la durée de survie $X$ de patients atteints d'une certaine maladie. Pour les patients perdus de vue au bout du temps $C$ alors qu'ils étaient *encore vivants*, $C$ censure $X$ à droite puisque, pour eux, *$X$ est inconnue mais supérieure à $C$* : $X > C$.

*Exemple de censure gauche :*

Un ethnologue étudie la durée d'apprentissage d'une tâche. Cette durée est une variable aléatoire $X$ et $C$ est l'âge de l'enfant. Pour les enfants qui *savent déjà* accomplir la tâche, *$C$ censure $X$ à gauche* car pour eux *$X$ est inconnu mais inférieur à $C$* : $X < C$.

D'ailleurs cet exemple comporte aussi des censures droites. En effet, les enfants qui ne savent pas encore accomplir la tâche en question lors du départ de l'ethnologue sont censurés à droite par la durée d'apprentissage $C'$ observée par l'ethnologue, car $X > C'$.

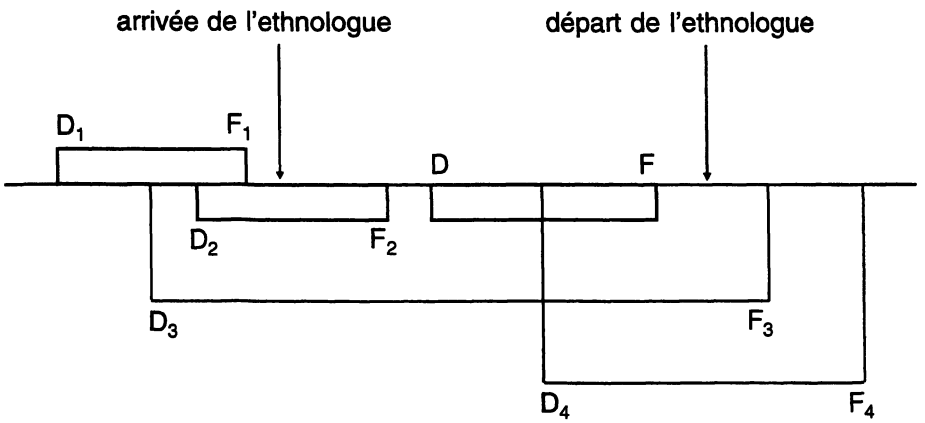



Figure 1. Exemple de censures droite et gauche

Dans la figure 1, $D_i$ est le début de l'apprentissage, $F_i$ la fin, pour le sujet *i*.

| | |
|---|---|
| $D_1F_1$ | est censuré *à gauche* par l'âge C de l'enfant. |
| $D_2F_2$ | bien qu'étant de trois types différents, ils sont tous les trois |
| $D_3F_3$ | censurés *à droite* : le premier par l'âge de l'enfant, le second par |
| $D_4F_4$ | la durée de séjour de l'ethnologue et le troisième par la durée d'aprentissage observée par l'ethnologue. |
| *DF* | n'est pas censuré. |

*Censure par intervalle*

Si, au lieu de $X$, on observe $C_1 < C_2$ tels que $C_1 < X < C_2$ ($X$ non observé), il y a censure par intervalle. En particulier, la censure gauche peut être considérée comme une censure par un intervalle tel que $C_1 = -\infty$, et la censure droite par un intervalle tel que $C_2 = +\infty$.

**b) Troncature**

*Troncature gauche*

On dit qu'il y a troncature *gauche* lorsque la variable d'intérêt $X$ n'est *observable que si elle est supérieure à T*. $T$ est alors la variable aléatoire de troncature gauche :

$$X \text{ n'est observée que si } \quad X > T.$$

*Troncature droite*

On dit qu'il y a troncature *droite* lorsque $X$ n'est *observable que si elle est inférieure à T*. $T$ est alors la variable aléatoire de troncature droite :

$$X \text{ n'est observée que si } \quad X < T.$$

Plus généralement, il y a troncature si l'observation de la variable d'intérêt $X$ n'a lieu que conditionnellement à un événement $B$.

**Exemples**

1) Durée de vie après la retraite : on étudie la durée de vie après la retraite de sujets qui entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite. Un sujet n'est donc observé que si sa durée de vie après la retraite excède le délai entre sa prise de retraite et l'instant de l'enquête. La durée de vie après la retraite est donc tronquée à gauche par ce délai. Elle peut aussi être censurée à droite si la fin de l'enquête a lieu alors que le sujet est toujours vivant (Florence Curt, mémoire de DEA de Statistique et Santé).

2) Durée d'induction du SIDA pour des transfusés : cet exemple est examiné en détail au paragraphe suivant. Il conduit, au contraire du précédent, à des durées tronquées à droite et censurées à gauche.

## 3. Le problème relatif au SIDA acquis par transfusion

La variable d'intérêt est ici la durée d'induction $T$ de la maladie, durée qui s'écoule entre la date d'infection $Y$ et la date $(Y + T)$ de déclaration de la maladie. On suppose que l'observation a lieu entre deux dates fixes 0 et $b$.

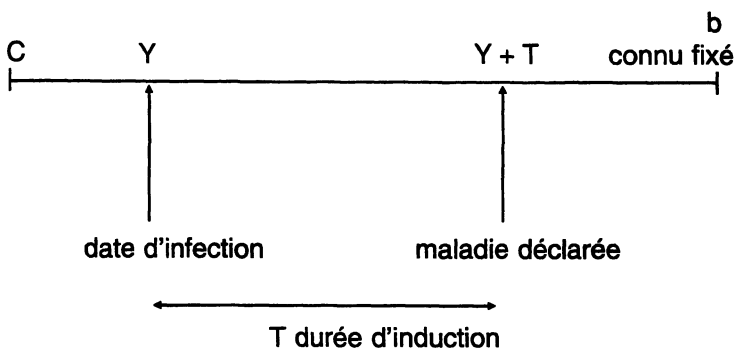



Figure 2. Schéma correspondant au SIDA

On fait l'hypothèse que $Y$ et $T$ sont indépendantes, de lois diffuses, et on note $F$ la fonction de répartition de $T$ et $G$ celle de $Y$ et $f$ et $g$ les densités correspondantes :

$$F(t) = P(T \leq t), \text{ densité } f(t),$$

$$G(y) = P(Y \leq y), \text{ densité } g(y).$$

Ne sont observés que les couples $(Y, T)$ tels que $0 \leq Y + T \leq b$.

Il y a donc troncature puisqu'on n'observe $T$ que conditionnellement à l'événement $B = \{0 \leq Y + T \leq b\}$ de probabilité

$$\int_0^b g(y)\, F(b - y)\, dy$$

Conditionnellement à cet événement, la densité au point $(y, t)$ vaut :

$$\varphi(y, t) = \frac{g(y) f(t)}{\int_0^b g(u)\, F(b-u)\, du} \cdot 1_{\{0 \le y \le b-t;\ 0 \le t\}}$$

Conditionnellement à $Y = y$, la densité au point $t$ vaut :

$$\varphi^*(t/y) = f(t)/F(b-y)$$

**Identification de la loi $\mathcal{L}(T)$ de la durée d'induction $T$**

La limitation $\{T \le b\}$ ne permet pas d'identifier autre chose que $\mathcal{L}\ (T / T \le b)$, par exemple :

$$Q(t) = F(t)/F(b)$$

ou $$q(t) = [dQ(t)/dt] \cdot 1_{\{0 \le t \le b\}}$$

ou encore $$k(t) = [f(t)/F(t)] \cdot 1_{\{0 \le t \le b\}}$$

Cette fonction $k$ n'est pas un taux de hasard, mais le devient si on va à reculons dans le temps : on appellera *$k$ le taux de hasard rétro de $T$.* C'est $k$, et non le taux de hasard usuel $h = f/(1-F)$, qui est intéressant car, lorsqu'on a observé $Y$, *$T$ ne pourra être observé que s'il est inférieur ou égal à $b - Y$.*

*Renversement du temps*

Remplaçons la variable d'intérêt $T$ par la variable

$$X = b - T$$

qui lui est équivalente, on constatera alors que le taux de hasard usuel $h$ de la variable $X$ est identique au taux de hasard rétro $k$ de $T$. En effet

$$P(t - dt < T \le t / T \le t, Y = y, 0 \le Y + T \le b) = P(t - dt < T \le t / T \le t) = k(t)\, dt$$

dès que $t \le b - y$.

Si $h$ est le taux instantané de mort de $X$, ou taux de hasard de $X$,

$$\begin{aligned} h(x)\, dx &= P(x \le X < x + dx / X \ge x) \\ &= P(x \le b - T < x + dx / b - T \ge x) \\ &= P(b - x - dx \le T \le b - x / T \le b - x) \\ &= k(b - x)\, dx. \end{aligned}$$

Si l'on remplace maintenant le couple initial $(T, Y)$ par le couple $(X, Y)$, on voit que $X$ est distribué sur $[0\ ;\ b]$ et que $X$ est tronquée à gauche par $Y$. En effet, on n'observe $(X, Y)$ que si

$$0 \le Y \le X \le b.$$

On peut alors se poser les trois questions suivantes :

1) Sous quelles conditions la loi de la variable d'intérêt $X = b - T$, $Y$ étant nuisible,

$$\mathcal{L}(X \,.\, 1_{\{X \leq b\}}) \text{ est-elle identifiable ?}$$

2) Quel estimateur construire pour cette loi, ou plus précisément pour sa fonction de répartition, notée $F$ dans la suite ?

3) Quelles sont les propriétés de cet estimateur ?

## 4. Estimation de la survie pour des durées tronquées à gauche

Les résultats de ce paragraphe sont dûs à Woodroofe (1985).

### a) Identifiabilité d'une loi tronquée à gauche sur [0 ; ∞[

Soient $X$ et $Y$ deux variables aléatoires positives, indépendantes, de fonctions de répartition respectives $F$ et $G$. L'intérieur du support convexe de $F$ est $]a_F \;\; b_F[$ :

$$a_F = \inf \{x : F(x) > 0\}$$

$$b_F = \inf \{x : F(x) > 0\}$$

$X$ est la variable d'intérêt et $Y$ tronque $X$ à gauche, c'est-à-dire qu'au lieu d'observer le couple $(X, Y)$, on observe en réalité un couple $(X^*, Y^*)$ tel que :

$$\mathcal{L}(X^*, Y^*) = R^* = \mathcal{L}\left((X, Y) / X > Y\right)$$

On note les marginales de $R^*$ respectivement $F^*$ et $G^*$. La probabilité d'observer effectivement le couple $(X, Y)$ n'est pas égale à 1 mais à

$$\alpha = \int G \, df = \int (1 - F^-) \, dG$$

Si l'on définit les deux ensembles de lois

$$\mathcal{M} = \{(F, G) : F(0) = G(0) = 0, \;\; \alpha > 0\}$$

$$\mathcal{M}_0 = \{(F, G) \in \mathcal{M} \;;\; a_G \leq a_F \;;\; b_G \leq b_F\}$$

ainsi que la fonctionnelle

$$T(F, G) = R^* \text{ de marginales } F^* \text{ et } C^*,$$

$R^*$ n'est pas égal au produit $F^* \otimes G^*$, et notre problème consiste à obtenir la première marginale $F^*$ de la fonctionnelle inverse $T^{-1}(R^*)$.

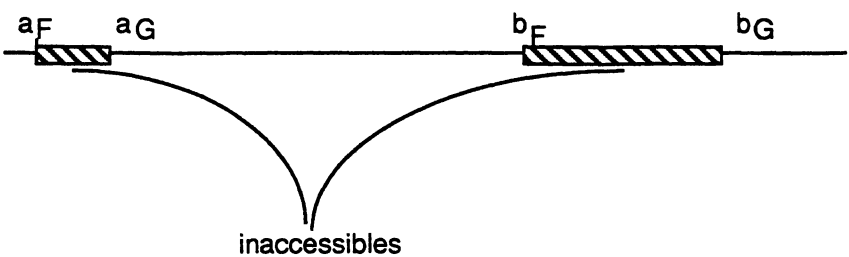



Figure 3. Supports de $F$ et $G$

La partie du support convexe de $F$ qui est hachurée sur la figure 3 est rendue inacessible par la troncature gauche de $X$ par $Y$. C'est pourquoi on doit se restreindre à $\mathcal{M}_0$

$$a_F \geq a_G \qquad (1)$$
$$b_F \geq b_G \qquad (2)$$

Dans notre exemple, $b_F = b_G = b$, donc (2) est automatiquement vérifiée.

**Lemme**

Soit $S$ l'application de $\mathcal{M}$ dans $\mathcal{M}_0$ qui à $(F, G)$ fait correspondre $(F_0, G_0)$ tels que

$$F_0 = \mathcal{L}(X/X \geq a_G),$$
$$G_0 = \mathcal{L}(Y/Y \leq b_F).$$

Alors $T(F_0, G_0) = T(F, G)$.

*Démonstration*

Si $Y \leq X$ alors $X > a_G$ et $Y \leq b_F$ presque sûrement. Comme $\mathcal{M}_0 \subset \mathcal{M}$,

$$T(\mathcal{M}_0) = T(\mathcal{M}).$$

**Théorème 1** (Inversion de $T$ dans $\mathcal{M}_0$) (Woodroofe)

Soit $R^* \in T(\mathcal{M})$, de marginales $F^*$ et $G^*$. Alors il existe un couple unique $(F, G)$ dans $\mathcal{M}_0$ tel que $T(F, G) = R^*$. Ce couple est déterminé par les équations suivantes :

1) Taux de hasard cumul $H$ de $F$

$$H = \int \frac{dF}{1 - F^-} = \int \frac{dF^*}{G^* - F^{*-}}$$

2) Taux de hasard rétro cumulé $K$ de $G$

$$K = \int \frac{dG}{G} = \int \frac{dG^*}{G^* - F^{*-}}$$

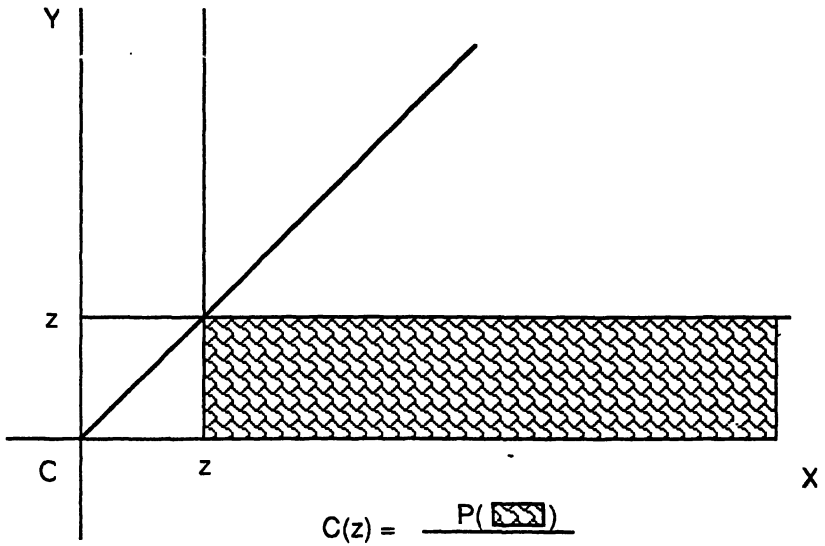
*Démonstration*

La probabilité d'être « à risque » en $z$ est $C(z) = G^*(z) - F^*(z^-)$. Or on peut aussi écrire $C(z)$ comme

$$\begin{aligned} C(z) &= \left(1 - F^*(z^-)\right) - \left(1 - G^*(z)\right) \\ &= P^*(X \geq z) - P^*(Y \geq z) \\ &= P^*(Y \leq z, X \geq z). \end{aligned}$$

Donc $C = G(1 - F^-)/\alpha$ où $\alpha$ est la probabilité que $(X, Y)$ soit observée. Comme $dF^* = G\, dF/\alpha$ et $dG^* = (1 - F^-)\, dG/\alpha$, on en déduit que

$$\int dF^*/C = \int dF/(1 - F^-) = H \quad \text{taux de hasard cumulé de } F,$$
$$\int dG^*/C = \int dG/G \qquad = K \quad \text{taux rétro cumulé de } G.$$




Le taux de hasard cumulé $H$ détermine complètement la f.r. $F$ par l'équation suivante, dans laquelle $H_c$ représente la partie continue, et $\Delta H = H^+ - H^-$ les sauts de $H$ :

$$1 - F(x) = e^{-H_c(x)} \prod (1 - \Delta H).$$

**Remarque 1**

L'inversion n'utilise que les marginales $F^*$ et $G^*$. On peut donc reconstituer $R^*$ à partir de ses marginales.

**Remarque 2**

La probabilité $\alpha$ d'observation conditionnellement à la troncature peut être calculée à partir de $F^*$ et $G^*$.

**Remarque 3**

$T$ est continue pour la convergence en loi en tout point $(F, G)$ pourvu que $F$ et $G$ n'aient aucune discontinuité commune.

Par contre $T^{-1}$ n'est pas continue.

En voici un contre-exemple : soient $F$ et $G$ continues, de support $[0\ \infty[$, et soient $G_n = (G + \delta_n)/2$. Alors $G_n$ ne tend pas vers $G$ quand $n$ tend vers l'infini, alors que $T(F, G_n)$ tend vers $T(F, G)$ :

$$R_n^*(x, y) = \frac{\int F\big(\max(x, z)\, dG_n(z)\big)}{\int (1 - F)\, dG_n}$$

$$= \frac{\frac{1}{2}\left\{\int F\big(\max(x, z)\, dG(z)\big) + F(n)\, 1_{n<x}\right\}}{\frac{1}{2}\left\{\int (1 - F)\, dG + \big(1 - F(n)\big)\right\}}$$

Au numérateur comme au dénominateur, le deuxième terme tend vers 0 quand $n$ tend vers l'infini, à $x$ fixé.

**b) Estimation d'une loi tronquée à gauche**

Soit un $n$-échantillon $(X^*, Y^*)_i$, $i$ variant de 1 à $n$. La version empirique du théorème d'inversion donne :

Si $F_n^*$ est l'empirique de $F^*$, c'est-à-dire $F_n^*(x) = (\sum 1_{X_i \le x})/n$, et si, de même, $G_n^*$ est l'empirique correspondant à $G^*$, l'empirique de $C$ est

$$C_n = G_n^* - F_n^{*-}$$

Alors l'empirique du taux de hasard cumulé de $F$ sera

$$H_n(z) = \int_0^z \frac{dF_n^*}{C_n} = \sum_{i\,:\,x_i \le z} \left[\frac{1}{nC_n(x_i)}\right]$$

**Définition de l'estimateur**

Si l'on ordonne les $x_i$ en $x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(k)}$, si l'on note $n_i$ le nombre $\nu(i)$

$$\nu(i) = \#\{j : x_j^* = x_{(i)}\}$$

et si de plus $C_{ni}$ désigne $C_{ni} = C_n(x_i) = \#\{j : x_j^* \ge x_i\,;\ y_j^* \le x_i\}$, l'estimateur non paramétrique de $F$ est

$$F_n(z) = 1 - \prod_{x(i) \le z} \left[1 - \frac{n_i}{nC_n}\right]$$

**Remarque**

Dans le cas censuré à droite, on remarque que c'est le même procédé qui est utilisé : écrire la fonction de survie que l'on cherche à estimer comme une fonctionnelle de quantités qui ont des empiriques naturelles étant donnée la structure des données. En effet, la fonction de survie $S$ peut s'écrire :

$$S(t) = \exp\left[\int_0^t \frac{dST_1}{ST_1 + ST_0} + \sum \mathrm{Log}\left[\frac{ST_0(u^+) + ST_1(u^+)}{ST_0(u^-) + ST_1(u^-)}\right]\right]$$

où, si $X$ est la variable censurée à droite et $D$ l'indicateur de décès, la notation (mnémotechnique) est la suivante :

$$ST_0(s) = P(X \le s,\ D = 0) \quad \text{et} \quad ST_1(s) = P(X \le s,\ D = 1)$$

et, ces fonctions de répartition sous-stochastiques ayant des empiriques évidentes, l'estimateur de Kaplan-Meier peut s'écrire de la même façon les vraies $ST_0$ et $ST_1$ étant remplacées par leurs empiriques.

**Exemple 1**

Soit un couple uniforme sur [0 ; 1], c'est-à-dire $F = G = U_{[0;1]}$. Les couples $(x, y)$ sont ordonnés par valeurs croissantes des $x$, et $p_k = F_n(x_{(k)}) - F_n(x_{(k-1)})$. Alors, le calcul de $F_n$ peut se faire à partir du tableau suivant (extrait de l'article de Biometrika 75, 3, p. 515-523) :

| $x_{(k)}$ | $y$ | $C_n(x_{(k)})$ | $F_n(x_{(k)})$ | $p_k$ |
|---|---|---|---|---|
| .3156 | .0672 | 6 | .1667 | .1667 |
| .3597 | .0136 | 5 | .3333 | .1667 |
| .4017 | .0816 | 4 | .5000 | .1667 |
| .4970 | .0816 | 5 | .6000 | .1000 |
| .5068 | .2559 | 4 | .7000 | .1000 |
| .6586 | .1113 | 4 | .7750 | .0750 |
| .7719 | .5820 | 4 | .83125 | .05625 |
| .7897 | .4106 | 3 | .88750 | .05625 |
| .8707 | .0592 | 2 | .94375 | .05625 |
| .9441 | .7175 | 1 | 1.0000 | .05625 |

Les deux estimateurs de $\mu$, moyenne de $X$, que sont la moyenne empirique et l'estimateur du maximum de vraisemblance valent respectivement

$$\bar{x} = 0{,}611\,6 \quad \text{et} \quad \widehat{\mu} = 0{,}519\,2$$

Dans ce cas,
$$F^*(x) = (x^2/2)\cdot 2 = x^2,$$
$$G^*(x) = \left[1-(1-y)^2\right] = y(2-y),$$
$$R^*(x,y) = \left[x^2-(x-y)^2\right]\cdot 1_{x \geq y}.$$

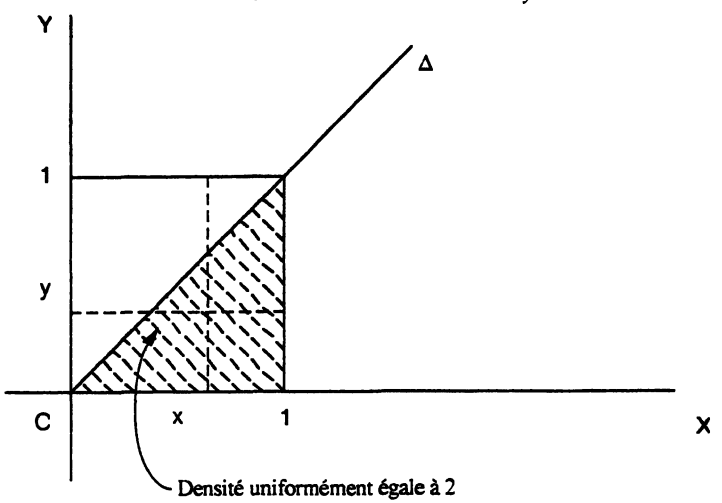



Figure 5. Exemple d'un couple d'uniformes sur [0 ; 1]

$F^*$, de densité $f^*(x) = 2x \cdot 1_{[0\,;\,1]}(x)$, a pour espérance

$$EX^* = \int 2x^2 \cdot 1_{[0\,;\,1]}(x)\, dx = 2/3,$$

alors que $F$, de densité $f^*(x) = 1_{[0\,;\,1]}(x)$, a pour espérance $EX + 1/2$. Dans l'exemple considéré, sur 10 observations, $\int x\, dF_n(x) = 0{,}519\,2$ est proche de cette valeur alors que la moyenne empirique est assez éloignée de $EX$.

**Exemple 2 : problème du SIDA**

$Y$ est la date de transfusion, $T$ est la durée d'incubation, et donc $Y + T$ est la date d'établissement de la séropositivité.

$$F_T(t) = P(T \leq t) = P(b - T \geq b - t) = 1 - P(X < b - t) = 1 - F_X^-(b - T),$$

ce qui donne pour estimateur de la fonction de répartition de $T$ :

$$F_{n,T}(t) = 1 - \left[1 - \prod_{\{i\,:\,x_{(i)} < b - t\}} \left(1 - \frac{n_i}{c_{ni}}\right)\right]$$

$$= \prod_{\{j\,:\,t_{(j)} > t\}} \left[1 - \frac{n_j}{c_{nj}}\right]$$

où $$n_j = \#\,\{i : t_i = t_{(j)}\}$$

et $$c_{nj} = \#\,\{i : t_i \geq t_{(j)}\ ;\ y_i < t_{(j)}\}$$

En effet, $x_{(i)} = b - t_{(m-i+1)}$ et $\{x_{(i)} : x_{(i)} < b - t\} \equiv \{t_{(j)} : t_{(j)} > t\}$.

Intuitivement, on pourrait dire que l'estimateur de la fonction de répartition est fondé sur un produit de probabilités conditionnelles analogue à celui qui définit l'estimateur de Kaplan-Meier, mais avec renversement du temps :

$$P(T \leq t_{(i)}) = P(T \leq t_{(i)}/T \leq t_{(i+1)}) \cdot P(T \leq t_{(i+1)}/T \leq t_{(i+2)}) \ldots P(T \leq t_{(m)})$$

avec pour estimateurs naturels de chacun des facteurs :

$$\tilde{P}(T \leq t_{(j-1)}/T \leq t_{(j)}) = 1 - \tilde{P}(T = t_{(j)}/T \leq t_{(j)})$$

$$= \frac{\text{nbre des « tués » en } t_{(j)}}{\text{nbre des « à risque » en } t_{(j)}}$$

Les « à risque » en $t_{(j)}$ doivent vérifier simultanément les deux conditions suivantes :

1) avoir sauté en $t_{(j)}$ ou avant : $T \leq t_{(j)}$,

2) avoir la possibilité de sauter en $t_{(j)}$ : $Y < b - t_{(j)}$.

Notons $$t_{(1)} < t_{(2)} < \ldots < t_{(j)} < \ldots < t_{(m)}$$

les $m$ valeurs distinctes, ordonnées, des temps d'induction

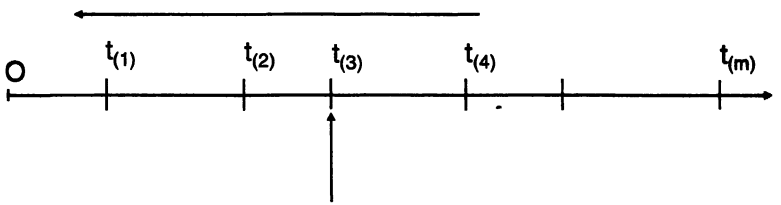
$$T_i,\ i = 1, 2, \ldots, n\ \ (m \leq n),$$

et $n_j = \sum_{i=1}^{n} 1_{T_i = t_{(j)}}$, effectif des sujets $i$ tels que $T_i = t_{(j)}$.

$C_{nj} = \sum_{i=1}^{n} 1_{T_i \le t_{(j)} \le x^* - x_i}$, effectif des sujets « à risque » en $t_{(j)}$.

C'est l'effectif des sujets tels que $T_i \le t_{(j)}$ et $Y_i \le b - t_{(j)}$.

$$\tilde{F}_{n,T}(t) = \prod_{t_{(j)} < t} \left[1 - \frac{n_j}{C_{nj}}\right]$$



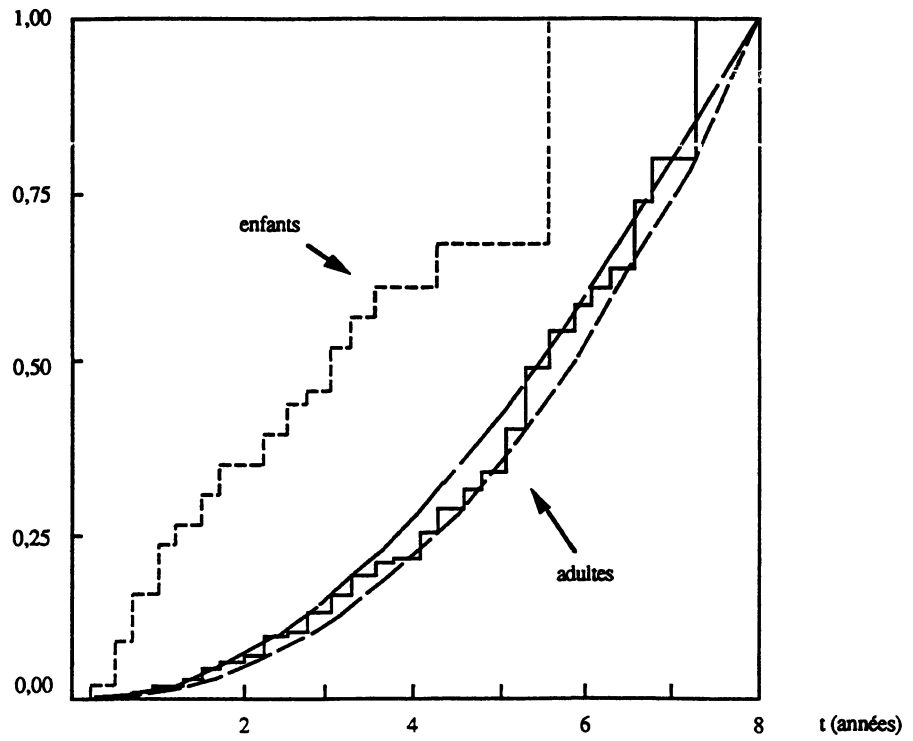




Figure 7. Estimateurs non-paramétriques de $F_T^{\flat}(t) = P(T \le t / T \le b)$ pour les enfants (-----) et les adultes (——) comparés à ceux $M - V$ pour un modèle de Weibull.

*Tableau 1.* Données de Lagakos, Barraj et de Gruttola : 258 adultes et 37 enfants (*) ayant eu le SIDA transmis par transfusion sanguine.

| *Y* date d'infection | *T* (* pour les enfants) durée d'incubation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,00 | 5 | | | | | | | | | |
| 0,25 | 6,75 | | | | | | | | | |
| 0,75 | 5(2) | 7,25 | | | | | | | | |
| 1,00 | 4,25 | 5,5* | 5,75 | 6,25 | 6,5 | | | | | |
| 1,25 | 4 | 4,25 | 4,75 | 5,75 | | | | | | |
| 1,50 | 2,25* | 2,75 | 3,75 | 5 | 5,5 | 6,5 | | | | |
| 1,75 | 2,75 | 3 | 5,25(2) | | | | | | | |
| 2,00 | 2,25 | 3 | 4 | 4,5 | 4,75 | 5 | 5,25(2) | 6 | | |
| 2,25 | 3 | 3* | 5,5 | | | | | | | |
| 2,50 | 2,25(4) | 2,5 | 2,75 | 3 | 3,25(2) | 4(3) | | | | |
| 2,75 | 1* | 1,25 | 1,5 | 2,5 | 3(2) | 3,25 | 3,75 | 4,5(2) | 5(2) | 5,25(5) |
| 3,00 | 1,75* | 2 | 3,25 | 3,5 | 3,75 | 4(2) | 4,25(3) | 4,75(3) | 5 | |
| 3,25 | 1,25 | 1,75 | 2(2) | 2,75 | 3(2) | 3,5(2) | 4,25 | 4,5 | | |
| 3,50 | 0,75* | 1,25 | 2,25(2) | 2,5 | 2,75(2) | 3 | 3,25 | 3,5(2) | 4(2) | 4,25 |
| | 4,5(2) | | | | | | | | | |
| 3,75 | 0,75* | 1* | 1,25 | 1,75(2) | 2 | 2,75 | 2,75* | 3(3) | 3* | 3,5* |
| | 4 | 4,25(2) | 4,25* | | | | | | | |
| 4,00 | 1 | 1* | 1,5(2) | 2 | 2,25 | 2,75 | 3,5 | 3,75(2) | 4 | |
| 4,25 | 1,25 | 1,5(2) | 1,75* | 2(3) | 2,25 | 2,5(3) | 3 | 3,5(2) | | |
| 4,50 | 1 | 1,5(3) | 1,75 | 2,25(2) | 2,5(4) | 2,75(4) | 3(3) | 3,25(2) | 3,25* | |
| 4,75 | 1 | 1* | 1,5(3) | 1,75(2) | 2 | 2,25* | 2,75 | 3(2) | 3,25(6) | |
| 5,00 | 0,5 | 0,5* | 0,75* | 1,5(2) | 1,5* | 1,75 | 2 | 2,25(3) | 2,5(2) | |
| | 2,5* | 3(3) | | | | | | | | |
| 5,25 | 0,25(2) | 0,25* | 0,75(3) | 1(2) | 1* | 1,25(2) | 1,5(4) | 1,5* | 2,25(2) | |
| | 2,5(2) | 2,75 | | | | | | | | |
| 5,50 | 0,5* | 1(3) | 1,25(2) | 1,5* | 1,75 | 2 | 2,25(2) | 2,5 | 2,5* | |
| 5,75 | 0,25 | 0,75 | 1 | 1,5 | 1,75* | 2(2) | 1,25 | | | |
| 6,00 | 0,5 | 0,5* | 0,75(3) | 1(3) | 1,25(2) | 1,25* | 1,5(2) | 1,75(3) | 2 | |
| 6,25 | 0,5* | 0,75 | 1 | 1,25 | 1,25* | 1,75(2) | | | | |
| 6,50 | 0,25(2) | 0,75 | 0,75* | 1 | 1,25 | 1,5 | | | | |
| 6,75 | 0,5* | 0,75(3) | 0,75* | 1 | 1,25(3) | | | | | |
| 7,00 | 0,75 | 0,75* | | | | | | | | |
| 7,25 | 0,25 | 0,25* | | | | | | | | |

Si l'on compare les estimateurs non-paramétriques de la figure 7 avec les estimateurs paramétriques fondés sur un modèle de Weibull : $F(t) = 1 - \exp\left(-(\theta t)^r\right)$ où l'on estime les deux paramètres par maximum de vraisemblance, on voit que le maximum de vraisemblance est très plat et n'a donc aucune robustesse. En effet

Log $V = -271{,}0$ correspond à $r = 2{,}2$ et $q = 0{,}1$ soit une médiane de 8 ans et demi

et Log $V = -270{,}1$ correspond à $r = 2{,}1$ et $q = 0{,}004$ soit à une médiane de 210 ans !

### c) Propriétés de l'estimateur

- **Consistance**

Soient $F$ et $G$ continues, $F$, $G \in \mathcal{M}_0$, $H$ le taux de hasard cumulé correspondant à $F$, et $\varphi$ une fonction mesurable sur $[0\ \infty[$ appartenant à $L^1(dH)$.

**Lemme :** calcul explicite du biais

Si on note $C = G(1-F)/\alpha$, pour tout $n \geq 1$ :

$$E_n\left\{\int_0^\infty \varphi\, dH_n\right\} = \int_0^\infty \varphi\, dH - \int_0^\infty \varphi(1-C)^n\, dH$$

En particulier, le biais sur l'estimateur $H_n$ de $H$ au point $x$ vaut

$$\text{Biais}_n\left\{H_n(x)\right\} = -\int_0^x \varphi(1-C)^n\, dH$$

**Exemple :**

Soient $F$ et $G$ les fonctions de répartition de deux lois uniformes sur $[0\ ;\ 1]$.

$$dH = dF/(1-F) = dx/(1-x)\ ;\quad H(x) = -\text{Log}\,(1-x)\ ;\quad h(x) = 1/(1-x).$$

$$C(z) = 2z(1-z)\,1_{[0\ ;\ 1]}(z).$$

Le biais en $n$ sur $H_n(x)$ vaut :

$$\text{Biais}_n\left(H_n(x)\right) = -\int_0^x \left(1-2z(1-z)\right)^n \frac{dz}{1-z}$$

Soit $x$ petit, $x \approx an^{-\beta}$, avec $a > 0$ et $\beta > 0$. Si $\beta = 1$, $\left(1-2z(1-z)\right)^n \approx e^{-2a}$ et le biais est proportionnel à $H(x)$.

**Théorème :**

Si $F$ et $G$ sont continues, $(F, G)$ dans $\mathcal{M}$ et $(F_0, G_0)$ les distributions de $X_1$ et $Y_1$ conditionnelles à $X_1 \geq a_G$ et $Y_1 \leq b_F$ $\left((F_0, G_0) \text{ dans } \mathcal{M}_0\right)$, alors

$$\sup_{x>0} \left|F_n(x) - F_0(x)\right| \text{——} P_n \longrightarrow 0$$

quand $n \longrightarrow \infty$

$$\sup_{y>0} \left|G_n(y) - G_0(y)\right| \text{——} P_n \longrightarrow 0$$

- **Normalité asymptotique**

**Théorème**

On considère les deux processus empiriques, sur $[0\ \infty[$,

$$Z_n^*(t) = \sqrt{n}\,\left[F_n^*(t) - F^*(t)\right]$$

$$Z_n'^*(t) = \sqrt{n}\,\left[G_n^*(t) - G^*(t)\right]$$

Alors $(Z_n^*,\ Z_n'^*)$ converge en loi vers le processus gaussien $(Z^*,\ Z'^*)$ à trajectoires continues et de fonction de covariance :

$$r_{xx}(s,t) = F^*(s) - F^*(s)\,F^*(t) \qquad 0 \le s \le t < \infty$$
$$r_{xy}(s,t) = R^*(s) - F^*(s)\,G^*(t) \qquad 0 \le s \le t < \infty$$
$$r_{yy}(s,t) = G^*(s) - G^*(s)\,G^*(t) \qquad 0 \le s \le t < \infty$$

**Théorème 1** (convergence vers un processus gaussien sur un compact) :

Soient $F$ et $G$ deux fonctions de répartition continues, appartenant à $\mathcal{M}_0$, et $[a\,;\,b]$ un compact strictement contenu dans $]a_G\,;\,b_F[$, c'est-à-dire que $a_G < a < b < b_F$, alors $W_{a,n}$ converge en loi vers $W^n$ sur $[a\,;\,b]$.

**Théorème 2** (convergence vers un processus gaussien) :

Si aux hypothèses du théorème 1 on ajoute la condition (*), alors

$$Z^*(a)/C(a) \longrightarrow 0 \quad \text{et} \quad W_1^n(t) \longrightarrow \int_{a_F}^{t} \frac{Z^*\,dG^* - Z'^*\,dF^*}{C^2}$$

quand $a \downarrow a_F$ et $a_F < t < b_F$.

**Remarques :**

1) Lorsque $a_F = a_G = 0$, la condition (*) devient $\int_{[0\ \infty]} G^{-1}\,dF < \infty$ et on a des difficultés au voisinage de 0. Par exemple, si $X$ et $Y$ sont deux exponentielles indépendantes $\mathcal{E}(a)$ et $\mathcal{E}(b)$, où $a$ et $b$ sont deux réels strictement positifs, cette condition n'est pas réalisée. En effet

$$\int_0^\infty \frac{dF}{G} = \int_0^\infty \frac{ae^{-au}}{1-e^{-bu}}\,du$$

et la fonction à intégrer est équivalente à $(a/b)(1/u)$ au voisinage de 0, et la condition (*) n'est pas remplie.

2) On peut avoir $F_n(x_i) = 1$ dès que le nombre des sujets « à risque », $nC_n(x_i)$, est égal à celui de ceux qui sautent en $x_i$, soit $n(x_i)$. On peut se demander si la condition (*) pourrait être affaiblie.

## 5. Modèle de régression pour des durées discrètes tronquées

Les résultats de ce paragraphe sont dûs à Gross S. et Huber C. (1992).

L'exemple illustratif est celui de la durée d'induction du SIDA induit par transfusion, qui donne des durées tronquées à droite.

À l'instant $E$ le sujet entre dans l'étude et la date $S$ de transfusion est alors connue. Cela est représenté sur la figure 8 ci-dessous. Le nombre total $N$ de sujets potentiels, qui, à l'instant final $t_f$ ont été infectés par transfusion est inconnu, puisque certains des sujets peuvent ne pas encore avoir développé de symptômes à cette date, d'où la troncature. Finalement, les données ont été reportées de trois mois en trois mois.

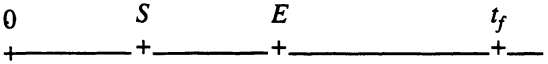



Figure 8

Dans cet exemple la durée d'induction $X = E - S$ est la variable d'intérêt, et $T = t_f - S$ est la troncature droite. L'instant zéro a été choisi arbitrairement pour inclure toutes les dates de transfusion des sujets observés. La seule covariable considérée vaut $Z = 0$ or 1 pour les adultes et les enfants respectivement.

De manière plus générale, on suppose qu'un vecteur de dimension $p$, $Z_S(t)$, est observé pour tout sujet non tronqué. Pour simplifier, on laisse tomber l'indice $s$ qui représente le sujet.

On commence par supposer que les composantes de $Z = (Z_1, Z_2, ..., Z_p)$ ne dépendent pas du temps et ne prennent qu'un nombre fini de valeurs. On ordonne les valeurs de $Z$, et on les appelle $l = 1, 2, ..., L$. Les observations peuvent alors être représentées par $L$ tables de dimension 2 représentant le croisement de $X$ et de $T$ pour chaque valeur fixée de $l$, $l \le l \le L$. On fait l'hypothèse habituelle que *la durée de vie X et la durée de troncature T sont indépendantes conditionnellement à Z.* Comme toutes les variables ne prennent qu'un nombre fini de valeurs, on peut résumer les données par les effectifs suivants :

$$n(i, j, l) = \#(s : X_S = i,\ T_S = j,\ Z_S = l) \qquad i, j \in I,\ l \le l \le L.$$

Comme seuls les $n(i, j, l)$ tels que $i \le j$ sont effectivement observés, les seules valeurs de la table $(l)$ sont $n(i, j, l)\ I[i \le j]$, dont la somme est $n^l$ ; le nombre total de sujets potentiels $N^l = \sum_{i,j} n(i, j, l)$ de la table $(l)$ n'est pas observé. Le nombre total non plus $N = \sum N^l$ n'est pas observé, seulement $n = \sum n^l$ l'est.

*Notation :* dans la table $(l)$, on notera le nombre des sujets de durée de troncature $j$, qui sont à risque à l'instant $i$, c'est-à-dire qui « meurent » en $i$ ou avant :

$$R(i, j, l) = \sum_{i' \le i} n(i', j, l) \qquad \text{pour} \qquad l \le i \le j \le k,\quad l \le l \le L,$$

et soit $h(i, l) = P(X = i \mid X \le i, T = j, Z = l) \qquad l \le i \le j \le k$ et $l \le l \le L$

le rétro-hasard à l'instant $i$ pour les sujets dont la covariable vaut $l$.

La vraisemblance conditionnelle qu'on considère est la suivante :

1) Conditionner par rapport à la covariance $Z$ donne le produit de $L$ vraisemblances, une par table.

2) Conditionner par rapport à la variable de troncature $T$ donne le produit des vraisemblances de $k$ lignes observées, conditionnellement au total de cette ligne. Cette probabilité conditionnelle élémentaire $V_{jl}$ vaut, pour la ligne $(T = j)$ de la table $(l)$, le produit des binomiales suivantes :

$$V_{jl} = \prod_{|i : i \le j|} \text{Bin}\big(R(i, j, l),\ h(i, l)\big).$$

La vraisemblance conditionnelle globale $V$ est donc donnée par

$$V = \prod_{l, j} V_{jl} = \prod_{l, i} \left[h(i, l)\, M_i^l \left(1 - h(i, l)\right) R_i^l - M_i^l\right]$$

$$\left\{ \prod_{l,\,(i, j : j \ge i)} R(i, j, l)! / \big(n(i, j, l)!\ \big(R(i, j, l) - n(i, j, l)\big)!\big) \right\}$$

où $R_i^l$ est le total, pour la table $(l)$, des sujets à risque à l'instant $i$,

$$R_i^l = \sum_{\{j : j \geq i\}} R(i,j,l) = \sum_{\{i',j : i' \leq i \leq j\}} n(i',j,l)$$

et $M_i^l$ est le nombre des « décès » à l'instant $i$, pour la table $(l)$ :

$$M_i^l = \sum_{\{j : j \geq i\}} n(i,j,l).$$

Le logarithme de la vraisemblance $L = \ln V$ est donc égal à

$$L = \sum_{i,l} \left\{ M_i^l \ln \left[ h(i,l) / \big(1 - h(i,l)\big) \right] + R_i^l \ln \left[ 1 - h(i,l) \right] + \right.$$
$$\left. \sum_{(j : j \geq i)} \ln R(i,j,l)! / \big( n(i,j,l)! \, \big( R(i,j,l) - n(i,j,l) \big)! \big) \right] \right\}$$

Plusieurs fonctions de lien sont possibles pour définir un modèle fondé sur $h(i,l)$. Le plus naturel d'entre eux est le logistique, suggéré par la forme de $L$. Comme le dit Efron (1988), la fonction de lien complémentaire en log-log,

$$g\big(h(i,l)\big) = \log \big( -\log \big( 1 - h(i,l) \big) \big),$$

peut aussi être utilisée, mais elle ne donne en pratique pas beaucoup de différence. Brookmeyer et Liao (1990) signalent que l'utilisation de cette fonction de lien donne un modèle de type Lehmann.

Commençons par choisir le modèle logistique le plus simple pour le rétro-hasard $h$

$$\ln \left[ \big( h(i,l) / \big( 1 - h(i,l) \big) \big) \right] = \mu + \gamma_i + \theta_l$$

avec les contraintes usuelles pour les constantes $\sum_i \gamma_i = \sum_l \theta_l = 0$.

Alors les estimateurs du maximum de vraisemblance conduisent aux équations suivantes, où $M_i = \sum_l M_i^l$, est le nombre total des durées observées à l'instant $i$, et $M^l = \sum_i M_i^l$, le même total dans la table $(l)$ :

$$h(i,l) = \exp(\mu + \gamma_i + \theta_l) / \big( 1 + \exp(\mu + \gamma_i + \theta_l) \big)$$
$$\sum_{l,i} R_i^l h(i,l) = n$$
$$\sum_l \left\{ R_i^l h(i,l) - R_k^l h(k,l) \right\} = M_i - M_k \qquad \text{for } i = 1, ..., k-l$$
$$\sum_i \left\{ R_i^l h(i,l) - R_i^L h(i,L) \right\} = M^l - M^L \qquad \text{for } l = 1, ..., L-l$$

Les solutions $h(i,l)$ de ces équations peuvent être obtenues par utilisation de n'importe quel logiciel standard (maximum de vraisemblance) pour la régression logistique, comme BMDP procédure *LR* par exemple, en reconfigurant les données au préalable pour obtenir les $kL$ couples $\big($nombre total (à risque) $= R_i^l$, nombre de « succès » $= M_i^l\big)$.

**Exemple** (données de SIDA de Lagakos et al (1988))

Dans le tableau ci-dessous, nous présentons le nombre total à risque $R_i^l$ et le nombre $M_i^l$ de sujets présentant les symptômes de la maladie à l'instant $i$ pour les deux groupes d'âge 0, 1, adultes et enfants. À partir du retro-hasard estimé $h(i, l)$, qu'on a calculé grâce à la procédure *LR* de BMDP, on a obtenu les estimateurs des fonctions de répartition $F(.\,, l)$ en utilisant la relation

$$F(i, l) = \big(1 - h(i + 1, l)\big) \ldots \big(1 - h(k, l)\big).$$

Les résultats obtenus par les deux modèles choisis sont présentés en face de ceux obtenus par Lagakos et al' qui utilisaient un ajustement non paramétrique.

Le modèle 1 est le modèle (4.1.2) dans sa forme la plus proche du non-paramétrique, car $\gamma_i$ y est complètement non spécifié, et $\theta_l = b_l$, $l = 0,1$. Dans ce modèle, les paramètres nuisibles donnent 29 degrés de liberté, conduisant à un test d'adéquation du chi-deux de 18.94 pour 27 degrés de liberté, soit un degré de signification de .872. Le modèle confirme donc l'absence d'interaction entre l'âge et la date, et donne pour estimateur de l'effet de l'âge : $b = -.38173$ avec pour écart-type S.E. = .1054 et donc pour coefficient standardisé 3.62.

*Tableau 2.* Estimated cumulative distributions for induction times of adults and children*.

| | Adults | | | | | Children | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| time | $M_i^0$ | $R_i^0$ | Non- | Model 1 | Model 2 | $M_i^1$ | $R_i^1$ | Non- | Model 1 | Model 2 |
| *period* | | | *Parametric* | | | | | *Parametric* | | |
| 1 | 6 | 6 | .0000 | .0000 | .0000 | 2 | 2 | .0000 | .0000 | .0000 |
| 2 | 2 | 7. | .0030 | .0021 | .0016 | 5 | 7 | .0243 | .0444 | .0335 |
| 3 | 13 | 20 | .0042 | .0046 | .0062 | 6 | 13 | .0858 | .0700 | .0773 |
| 4 | 15 | 35 | .0119 | .0126 | .0136 | 5 | 17 | .1579 | .1261 | .1192 |
| 5 | 16 | 50 | .0209 | .0225 | .0232 | 2 | 18 | .2237 | .1724 | .1589 |
| 6 | 23 | 66 | .0307 | .0321 | .0347 | 3 | 19 | .2516 | .2068 | .1954 |
| 7 | 13 | 73 | .0471 | .0485 | .0479 | 3 | 21 | .2988 | .2562 | .2300 |
| 8 | 14 | 82 | .0573 | .0600 | .0628 | 0 | 19 | .3486 | .2843 | .2633 |
| 9 | 20 | 87 | .0692 | .0708 | .0794 | 2 | 19 | .3486 | .3082 | .2957 |
| 10 | 15 | 93 | .0898 | .0915 | .0978 | 2 | 20 | .3896 | .3503 | .3588 |
| 11 | 14 | 97 | .1071 | .1096 | .1180 | 1 | 18 | .4329 | .3825 | .3899 |
| 12 | 21 | 100 | .1251 | .1276 | .1400 | 2 | 17 | .4584 | .4119 | .4210 |
| 13 | 13 | 100 | .1584 | .1618 | .1640 | 1 | 14 | .5195 | .4633 | .4521 |
| 14 | 8 | 91 | .1820 | .1861 | .1900 | 1 | 13 | .5594 | .4958 | .4833 |
| 15 | 5 | 76 | .1996 | .2051 | .2183 | 0 | 12 | .6061 | .5193 | .5147 |
| 16 | 11 | 74 | .2137 | .2184 | .2488 | 0 | 11 | .6061 | .5351 | .5463 |
| 17 | 9 | 73 | .2510 | .2535 | .2817 | 1 | 11 | .6061 | .5751 | .5783 |
| 18 | .6 | 68 | .2829 | .2905 | .3171 | 0 | 5 | .6667 | .6142 | .6106 |
| 19 | .5 | 58 | .3140 | .3176 | .3553 | 0 | 4 | .6667 | .6409 | .6765 |
| 20 | 8 | 55 | .3436 | .3465 | .3962 | 0 | 4 | .6667 | .6681 | .7102 |
| 21 | .9 | 51 | .4021 | .4030 | .4402 | 0 | 3 | .6667 | .7189 | .7443 |
| 22 | 4 | 39 | .4883 | .4863 | .4873 | 1 | 3 | .6667 | .7882 | .7790 |
| 23 | .2 | 29 | .5441 | .5548 | .5377 | 0 | 3 | 1.00 | .8399 | .8143 |
| 24 | .1 | 28 | .5844 | .5939 | .5916 | 0 | 2 | 1.00 | .8674 | .8501 |
| 25 | 1 | 18 | .6060 | .6151 | .6492 | 0 | 2 | 1.00 | .8819 | .8866 |
| 26 | 2 | 16 | .6417 | .6493 | .7108 | 0 | 2 | 1.00 | .9048 | .9237 |
| 27 | .1 | 12 | .7333 | .7359 | .7728 | 0 | 1 | 1.00 | .9610 | .9615 |
| 28 | 0 | 5 | .7333 | .7359 | .8463 | 0 | 0 | 1.00 | 1.000 | 1.000 |
| 29 | 1 | .5 | .8000 | .8000 | .9207 | 0 | 0 | 1.00 | 1.000 | 1.000 |

* DATA as reported by Lagakos et al (1988).

La différence entre les adultes et les enfants est ainsi confirmée.

Ayant observé une certaine régularité dans le comportement de $\gamma_i$ pour le modèle 1, on a essayé une forme paramétrique pour la fonction du temps $i$ qu'est $\gamma_i$ : $\gamma_i = a(1/i)^{1/2}$, ce qui donne le mocèle 2 dans le tableau ci-dessus. Le test d'ajustement du chi-deux vaut 47.54 pour 55 degrés de liberté, soit un degré de signification de 0.753 ; les valeurs estimées des paramètres sont :

$\mu$ = 3.6940 (estimated standard error = 0.2223, coeff./SE = 16,61),

$a$ = – 6.6861 (estimated standard error = 0.6494, coeff./SE = – 10.30),

$b$ = .76611 (estimated standard error = 0.2093, coeff./SE = 3.661).

À nouveau, pas d'interaction entre l'âge et le temps. Le coefficient standardisé $b$ pour l'âge est remarquablement proche de la valeur trouvée pour le modèle 1, ce qui suggère un test de même significativité pour l'hypothèse que l'âge n'a aucun effet sur la durée.

De manière générale, cette approche permet d'analyser la dépendance par rapport à des covariables d'intérêt en présence de troncature.

## RÉFÉRENCES

ALIOUM A. (1994) *Méthodes statistiques pour données tronquées et censurées*, thèse, Université de Bordeaux II.

ALIOUM A. et COMMENGES D. (1994) *A Proportional Hazards Model for Arbitrarily Censored and Truncated Data*, manuscrit Université de Bordeaux II.

ANDERSEN P.K., BORCH-JOHANSEN K., DECKERT T., GREEN A., HOUGAARD P., KEIDING N. and KREINER S. (1985) "A Cox Regression Model for Relative Mortality and its Application to Diabetes Mellitus Survival Data", *Biometrics*, 4, 921-932.

ANDERSEN P.K. and GILL R.D. (1982) "Cox's Regression Model for Counting Process : a Large Sample Study", *Ann. Statist.*, 4, 1100-1120.

ANDERSEN P.K., BORGAN O., GILL R.D., KEIDING N. (1992) *Statistical Models Based on Counting Processes*, Springer Verlag.

ARJAS E. and HAARA P. (1984) "A Marked Point Process Approach to Censored Failure Time Data with Complicated Covariates", *Scand. J. Statist.*, 193-209.

ARJAS E. (1985) "Stanford Heart Transplantation Data Revisited : a Real Time Approach", *in Modern Statistical Methods in Chronic Disease Epidemiology*, Wiley, New York.

ARJAS E. and HAARA P. (1987) "A Logistic Regression Model for Hazard : Asymptotic Results", *Scand. J. Statist.*, 1-18.

ARJAS E. and HAARA P. (1988) "A Note on the Asymptotic Normality in the Cox's Regression Model", *Ann. Statist.*, 1133-1140.

BECKER and MELBYE (1991) "Use of a Log-linear Model to Compute the Survival Curve from Interval-censored Data, with Application to Data on Tests for HIV Positivity", *Aust. J. Statist.*

BROOKMEYER R. and LIAO J. (1990) "The Analysis of Delays in Disease Reporting : Methods and Results for Acquired Immunodeficiency Syndrome", *Amer. J. Epid.*, 2, 355-365.

CNAAN A. and RYAN L. (1989) "Survival Analysis in Natural History Studies of Disease", *Statist. Med.*, 1255-1268.

DE GRUTTOLA V. and LAGAKOS S.W. (1989) "Analysis of Doubly-Censored Survival Data with Application to AIDS", *Biometrics*, 1-11.

EFRON B. (1988) "Logistic Regression, Survival Analysis and the Kaplan Meier Curve", *J. Amer. Statist. Assoc.*, 402, 414-426.

FRYDMAN H. (1994) "A Note on Non-parametric Estimation of the Distribution Function from Interval Censored and Truncated Observations", *JRSS* B 56, 71-74.

GROENEBOOM P. and WELLNER J.A. (1992) *Information Bounds and Non-parametric Maximum Likelihood Estimation*, Bukhaüser Verlag.

GROSS S. et HUBER C. (1992) "Regression Models for Truncated Data", *Scand. J. in Statistics*, n° 3, 193-213.

HILL C., COM-NOUGUÉ C., KRAMAR A., MOREAU T. et coll. (1990) *Analyse statistique des données de survie*, Flammarion éd.

HUBER C. (1989) *Analyse Statistique des durées de vie*, Economica, Droesbeke, Fichet, Tassi, éditeurs.

HUBER C. (1989) *Théorie de la robustesse*, Springer Verlag, Lecture Notes 1215, 1-128.

KALBFLEISH J.D. and LAWLESS J.F. (1989) "Inference Based on Retrospective Ascertainment : an Analysis of the Data on Transfusion Related AIDS", *J. Amer. Statist. Assoc.*, 406, 360-372.

KEIDING N. (1991) "Age-specific Incidence and Prevalence – a Statistical Perspective", *J. Roy. Statist. Soc. A.* (to appear with discussion).

KEIDING N., BAYER T. and WATT-BOOLSEN S. (1987) "Confirmatory Analysis of Survival Data Using Left Truncation of the Life Times of Primary Survivors", *Statist. Med.*, 939-944.

KEIDING N. and GILL R.D. (1990) "Random Truncation Models and Markow Processes", *Ann. Statist.*, 2, 582-602.

KLEIN J.P. and GOEL P.K. (1991) *Survival Analysis: State of the Art*, NATO ASI Series, serie E: *Applied Sciences*, vol. 211.

LAGAKOS S.W., BARRAJ L.M. and DE GRUTTOLA V. (1988) "Nonparametric Analysis of Truncated Survival Data, with Application to AIDS", *Biometrika*, 3, 515-523.

LAI T.L. and YING Z. (1989) *Rank Regression Methods for Left Truncated and Right Censored Data*, Technical report n° 8, Department of statistics, Stanford University.

LAI T.L. and YING Z. (1991) "Estimating a Distribution Function with Truncated and Censored Data", *Ann. Statist.*, 1, 417-442.

LAIRD N. and OLIVIER D. (1981) "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques", *J. Amer. Statist. Assoc.*, 374, 231-240.

LUI K.J., LAWRENCE D.N., MORGAN W.M., PETERMAN T.A., HAVERKOS H.H. & BREGMAN D.J. (1986) "A Model – based Approach for Estimating the Mean Incubation Period of Transfusion-associated Acquired Immunodeficiency Syndrome", *Proc. Nat. Acad. Sci.*, 2913-7.

MCCULLAGH P. and NELDER J. (1989) *Generalized Linear Models*, London, Chapman and Hall.

MEDLEY G.H., ANDERSON R.M., COX D.R. & BILLARD L. (1987) "Incubation Period of AIDS in Patients Infected Via Blood Transfusion", *Nature*, 719-21.

PRENTICE R.L. and GLOECKLER L.A. (1978) "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data", *Biometrics*, 57-67.

TSAI W.Y., JEWELL N.P. and WANG M.C. (1987) "A Note on the Productivity Estimator under Right Censoring and Left Truncation", *Biometrika*, 883-886.

TSUI K.L., JEWELL N.P. and WU C.F.J. (1988) "A Nonparametric Approach to the Truncated Regression Problem", *J. Amer. Statist. Assoc.*, 403, 786-792.

TURNBULL B.W. (1976) "The Empirical Distribution Function with Arbitrary Grouped Censored and Truncated Data", *J. Roy. Statist. Soc. Ser. B.*, 290-295.

WANG M.-C. (1991) "Nonparametric Estimation from Cross-sectional Survival Data", *J. Amer. Statist. Assoc.*, 413, 130-143.

WANG M.-C., JEWELL N.P. and TSAI W.Y. (1986) "Asymptotic Properties of the Product Limit Estimate under Random Truncation", *Ann. Statist.*, 4, 1597-1605.

WOODROOFE M. (1985) "Estimating a Distribution Function with Truncated Data", *Ann. Statist.*, 1, 163-177.