Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 342 (2006) 275–278
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Estimation sous biais de sélection et avec fonction de poids inconnue

Agathe Guilloux

LSTA, université Pierre et Marie Curie, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 3 février 2005 ; accepté après révision le 15 novembre 2005

Présenté par Paul Deheuvels

Résumé

Nous considérons le problème de l'estimation de la fonction de répartition G d'une variable aléatoire (v.a.) positive X à partir de l'observation d'une v.a. biaisée Y de fonction de répartition $F_w = \int w(x)\,\mathrm{d}G(x)/\mu_w$, où w est une fonction de poids inconnue. En supposant de plus que l'échantillon issu de la fonction de répartition F_w est censuré à droite, nous construisons un estimateur $\widehat{G}$ de la fonction de répartition G pour lequel on énonce un théorème de consistance forte et de convergence faible. ***Pour citer cet article : A. Guilloux, C. R. Acad. Sci. Paris, Ser. I 342 (2006).***

Abstract

Distribution estimation from biased data with unknown weighting function. We consider the problem of estimating the cumulative distribution function (cdf) G of a non-negative random variable (r.v.) X from the observation of a biased r.v. Y with cdf $F_w = \int w(x)\,\mathrm{d}G(x)/\mu_w$, where w is an unknown weighting function. We assume moreover that the random sample with common cdf F_w is right-censored. We construct an estimator $\widehat{G}$ for the cdf G and state its strong consistency and weak convergence. ***To cite this article: A. Guilloux, C. R. Acad. Sci. Paris, Ser. I 342 (2006).***

1. Introduction

Considérons une population d'individus $i \in I$, dans laquelle on note σ_i l'instant de naissance et X_i la durée de vie de l'individu i. Soit G la fonction de répartition de la v.a. X. Supposons que l'on ne peut observer que les individus vivants à un instant t_0 fixé d'échantillonnage. L'individu i peut donc entrer dans l'échantillon si $\sigma_i < t_0$ et $\sigma_i + X_i > t_0$ (ce qui assure qu'il est né avant t_0 et qu'il meurt après t_0). On peut alors montrer que les v.a. ς et Y, qui représentent, respectivement, l'instant de naissance et la durée de vie pour les individus vivants à l'instant t_0, souffrent d'un biais de sélection. On a, en effet, pour tout s et t :

$$\mathbb{P}(\varsigma \leqslant s, Y \leqslant t) = \mathbb{P}(\sigma \leqslant s, X \leqslant t | \sigma < t_0, \sigma + X > t_0) \neq \mathbb{P}(\sigma \leqslant s, X \leqslant t).$$

Adresse e-mail : aguillou@ccr.jussieu.fr (A. Guilloux).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.11.016

Plus précisément, supposons que le processus ponctuel $\eta = \sum_{i\in I} \delta_{\sigma_i}$ (où δ_a est la masse de Dirac en a), formé à partir des instants de naissance dans la population I, est poissonnien non-homogène d'intensité φ. On peut alors montrer, voir en particulier Lund [8], que Y a pour fonction de répartition F_w donnée, pour tout $t > 0$, par :

$$F_w(t) = \frac{\int_0^t w(x)\,\mathrm{d}G(x)}{\int_0^\infty w(x)\,\mathrm{d}G(x)}, \quad \text{où } w(t) = \int_0^t \varphi(t_0 - \sigma)\,\mathrm{d}\sigma, \text{ pour tout } t > 0, \tag{1}$$

et $\mu_w = \int_0^\infty w(x)\,\mathrm{d}G(x) < \infty$. La fonction F_w est souvent appelée version *biaisée* de la fonction G et l'on dit alors que la v.a. X souffre d'un biais de sélection.

Depuis Fisher [5], de nombreux auteurs ont étudié ce problème, en particulier Patil et Rao [9], Gill et al. [6] et Efromovitch [4]. Ces derniers ont fait l'hypothèse que la fonction de poids w est connue. Notre but est ici de montrer que, dans le modèle d'échantillonnage décrit ci-dessus, on peut se passer de cette hypothèse. Dans ce cadre et dans le cas où la fonction de biais w est inconnue, nous construisons un estimateur $\widehat{G}$ de la fonction de répartition G à partir d'un n-échantillon de couples $(\varsigma_1, Y_1), \ldots, (\varsigma_n, Y_n)$, où les durées de vie $Y_1, \ldots, Y_n$ sont censurées à droite par des v.a. positives $C_1, \ldots, C_n$. Nous énonçons également un théorème de consistance uniforme pour l'estimateur $\widehat{G}$ et de convergence faible pour le processus $\sqrt{n}(\widehat{G} - G)$.

2. Description des observations et de la censure

Considérons maintenant la population J des individus vivants à l'instant t_0. Soit un individu j de cette population J, né en ς_j et de durée de vie Y_j. La v.a. positive $t_0 - \varsigma_j$ représente alors son âge à l'instant t_0 d'échantillonnage et peut donc être appelée « temps de récurrence arrière ». On peut montrer, toujours en supposant que le processus $\eta = \sum_{i\in I} \delta_{\sigma_i}$ est poissonnien non-homogène d'intensité φ, que, pour tout $ut \geqslant 0$:

$$\mathbb{P}(t_0 - \varsigma \leqslant t) = \mathbb{P}(t_0 - \sigma \leqslant t \,|\, \sigma < t_0, \sigma + X > t_0) = \frac{1}{\mu_w} \int_0^{t\wedge t_0} \varphi(t_0 - s)\overline{G}(s)\,\mathrm{d}s. \tag{2}$$

Ainsi la loi des v.a. ς et Y, voir les Éqs. (1) et (2), détermine de manière unique les fonctions G et w. Ceci assure l'identifiabilité du problème proposé.

La durée de vie après l'échantillonnage de l'individu j est donnée par la v.a. positive $\varsigma_j + y_j - t_0$ qui peut alors être nommée « temps de récurrence avant ». Comme les individus entrent dans l'échantillon au temps t_0, il est naturel de supposer que seul le temps de récurrence avant peut être censuré ou, de façon équivalente, qu'un individu ne peut être censuré qu'à partir du moment où il est dans l'échantillon. On peut alors supposer que les durées de vie des individus de la population J sont censurées à droite par une v.a. C de fonction de répartition H de telle sorte que l'on n'observe pas des réalisations de la v.a. Y mais de la v.a. Z donnée par :

$$Z = t_0 - \varsigma + (\varsigma + Y - t_0) \wedge C$$

où $x \wedge y$ est le minimum entre x et y. On suppose également que l'indicatrice $I(\{\varsigma + Y - t_0 \leqslant C\})$ est observable, où $I(A)$ est l'indicatrice de l'événement A.

Ce modèle de censure a été étudié par Winter et Földes [11] et Asgharian et al. [3], notamment, et défendu vivement par Asgharian [2]. Il faut noter ici que ces derniers auteurs ont étudié le cas particulier où, dans l'expression de la fonction de répartition F_w, la fonction de poids est donnée par $w(t) = t$ pour tout $t > 0$, ce cas particulier est appelé *biais de longueur*. Ceci correspond, dans notre formulation, à supposer que le processus φ est poissonnien homogène, c'est-à-dire d'intensité φ constante.

3. Estimation de la fonction de répartition

Comme décrit dans la partie précédente, nous travaillons maintenant avec n individus vivants à l'instant t_0 pour lesquels on observe les triplets indépendants :

$$\big(\varsigma_j, Z_j, I\big(\{\varsigma_j + Y_j - t_0 \leqslant C_j\}\big)\big) \quad \text{pour } j = 1, \ldots, n.$$

On peut alors construire les processus M et O définis, pour $t > 0$, par :

$$M(t) = \sum_{j=1}^{n} I\big(\{Y_j \leqslant t, \varsigma_j + Y_j - t_0 \leqslant C_j\}\big) \quad \text{et} \quad O(t) = \sum_{j=1}^{n} I\big(\{t_0 - \varsigma_j \leqslant t \leqslant Y_j, C_j + t_0 - \varsigma_j \geqslant t\}\big).$$

Pour tout $t > 0$, la v.a. $M(t)$ donne, à chaque âge t, le nombre d'individus de la population J morts avant cet âge t. Le processus M est donc un processus de comptage classique. La v.a. $O(t)$ donne le nombre d'individus de J ni morts, ni censurés à l'âge t.

On remarquera que la différence entre le processus O défini plus haut et le processus *nombre-à-risque*, habituellement utilisé en statistique des durées de vie, réside dans la première indicatrice $I(\{t_0 - \varsigma_j \leqslant t\})$ qui assure que l'individu j est dans l'échantillon à l'âge t. Une conséquence de la définition du processus O est qu'il n'est pas décroissant, comme l'est le processus *nombre à risque*.

A l'aide des processus M et O, on peut construire un estimateur $\widehat{G}$, de type produit-limite, de la fonction de répartition G, défini, pour tout $t > 0$, par :

$$\widehat{G}(t) = 1 - \prod_{s \leqslant t} \big(1 - I\big(\{O(s) > 0\}\big)\, \mathrm{d}M(s)/O(s)\big),$$

où $\prod$ est le produit-intégral et est défini, pour X fonction càdlàg à variations localement bornées, par :

$$\prod_{s \leqslant t} \big(1 + \mathrm{d}X(s)\big) = \lim_{\max|u_i - u_{i-1}| \to 0} \prod_{1 \leqslant i \leqslant n} \big(1 + X(u_i) - X(u_{i-1})\big)$$

où $0 = u_0 < \cdots < u_i < \cdots < u_n = t$ est une partition de l'intervalle $[0, t]$, cf. Andersen et al. [1] pour plus de détails.

Comme le processus O n'est pas décroissant, on peut avoir pour un $t_1 > 0$ $\mathrm{d}M(t_1) = 1$ (une mort observée en t_1) et $O(t_1) = 1$. Dans ce cas, l'estimateur $\widehat{G}$ vaut 0 pour tous les $t \geqslant t_1$. Pour palier ce problème, nous introduisons, en suivant Winter et Földes [11], l'estimateur $\widehat{G}_n$ défini pour tout $t \geqslant 0$ par :

$$\widehat{G}_n(t) = 1 - \prod_{s \leqslant t} \left(1 - \frac{I(\{O(s) > 0\})\, \mathrm{d}M(s)}{O(s) + n\theta_n}\right) \tag{3}$$

où $(\theta_n)_{n \geqslant 1}$ est une suite d'entiers positifs qui vérifie $\lim_{n \to \infty} n\theta_n = 0$. Cet estimateur explicite a été introduit par Winter et Földes [11] dans le cas du biais de longueur et d'une censure déterministe. Asgharian et al. [3] ont considéré, toujours dans le cas du biais de longueur, un estimateur non explicite de la fonction de répartition.

On montre que, parmi les individus vivants à t_0, on a, pour tout $t > 0$:

$$\mathbb{E}\big(M(t)\big) = \frac{n}{\mu_w}\left(\int_0^t w(x)\, \mathrm{d}G(x) - \int_0^t \int_0^x w(x - c)\, \mathrm{d}H(c)\, \mathrm{d}G(x)\right) \quad \text{et}$$

$$\mathbb{E}\big(O(t)\big) = \frac{n}{\mu_w}\left(\big(1 - G(t)\big)w(t) - \big(1 - G(t)\big)\int_0^t w(t - c)\, \mathrm{d}H(c)\right).$$

Ce résultat et le fait que le processus $M - \int O(x)\, \mathrm{d}G(x)/(1 - G(x))$ est une martingale locale de carré intégrable permet de montrer le théorème suivant. On utilise alors, principalement, l'inégalité de Lenglart [7] pour montrer la consistance uniforme et le théorème de la limite centrale pour les martingales de Rebolledo pour la convergence faible [10].

Théorème 3.1. *Notons* $\tau = \sup\{t > 0,\ (1 - G(t))(1 - H(t)) > 0\}$, *on montre la convergence, pour tout* $\tau' \leqslant \tau$, *quand n tend vers l'infini* :

$$\sup_{t \leqslant \tau'} \big|\widehat{G}_n(t) - G(t)\big| \xrightarrow{\mathbb{P}} 0.$$

De plus, pour tout $\tau' \leqslant \tau$, *on montre la convergence faible suivante dans l'espace* $\mathbb{D}[0, \tau']$ *des fonctions càdlàg sur* $[0, \tau']$:

$$\sqrt{n}\big(\widehat{G}_n - G\big) \xrightarrow{\mathcal{D}} L,$$

où L est un processus gaussien centré de fonction de variance donnée, pour tout $(s,t) \in [0,\tau']^2$, par :

$$\big\langle L(s), L(t)\big\rangle = \big(1-G(s)\big)\big(1-G(t)\big) \int_0^{s\wedge t} \frac{\mu_w \,\mathrm{d}G(x)}{(1-G(x))^2 (w(x) - \int_0^x w(x-c)\,\mathrm{d}H(c))}.$$

On remarque que la variance limite du processus $\sqrt{n}(\widehat{G}-G)$ est proche de celle du processus classique de Kaplan–Meier, si ce n'est qu'ici on obtient $\mu_w^{-1}(w(x) - \int_0^x w(x-c)\,\mathrm{d}H(c))$ au lieu de $1-H(x)$.

Références

[1] P.K. Andersen, O. Borgan, R.D. Gill, N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, 1993.
[2] M. Asgharian, Biased sampling with right censoring: a note on Sun, Cui & Tiwari, Canad. J. Statist. 31 (2002) 349–350.
[3] M. Asgharian, C.E. M'Lan, D.B. Wolfson, Length-biased sampling with right-censoring: an unconditional approach, J. Amer. Statist. Assoc. 97 (2002) 201–209.
[4] S. Efromovitch, Density estimation for biased data, Ann. Statist. 32 (2004) 1137–1161.
[5] R.A. Fisher, The effect of methods of ascertainment upon the estimation of frequencies, Ann. Eugen. 6 (1934) 13–25.
[6] R.D. Gill, Y. Vardi, J.A. Wellner, Large sample theory of empirical distributions in biased sampling model, Ann. Statist. 16 (1988) 1069–1172.
[7] E. Lenglart, Relation de domination entre deux processus, Ann. Inst. H. Poincaré 13 (1977) 171–179.
[8] J. Lund, Sampling bias in population studies – how to use the Lexis diagram, Scand. J. Statist. 27 (2000) 589–604.
[9] G.P. Patil, C.R. Rao, The weighted distributions: a survey of their applications, in: P.R. Krishnaiah (Ed.), Applications of Statistics, North-Holland Publishing Co., Amsterdam, 1977, pp. 383–405.
[10] R. Rebolledo, Central limit theorems for local martingales, Z. Wahrsch. Verw. Gebiete 51 (1980) 269–286.
[11] B.B. Winter, A. Földes, A product-limit estimator for use with length-biased data, Canad. J. Statist. 16 (1988) 337–355.