Comparaison de procédures d'échantillonnage équilibré

Guillaume Chauvet [1] & David Haziza [2] & Éric Lesage [3]

[1] *Laboratoire de Statistique d'enquête, CREST(ENSAI), Campus de Ker Lann, 35 170 BRUZ, France, guillaume.chauvet@ensai.fr*

[2] *Département de mathématiques et de statistique, Université de Montréal, Québec, H3C 3J7, Canada, david.haziza@umontreal.ca*

[3] *Laboratoire de Statistique d'enquête, CREST(ENSAI), Campus de Ker Lann, 35 170 BRUZ, eric.lesage@ensai.fr*

Résumé.
Les procédures d'échantillonnage équilibré ont fait l'objet d'un intérêt marqué ces dernières années. Parmi les nombreuses méthodes conçues pour sélectionner des échantillons équilibrés ou approximativement équilibrés, nous nous intéressons à la méthode du Cube (Deville et Tillé, 2004) et au tirage réjectif (Fuller, 2009). Nous comparons les propriétés de plusieurs estimateurs associés à la procédure d'échantillonnage réjective de Fuller (2009). En particulier, nous montrons que la méthode d'estimation proposée par Fuller (2009) peut conduire à des biais importants pour des échantillons de petite taille, lorsque la relation entre la variable d'intérêt et les variables auxiliaires n'est pas linéaire. Une étude par simulation qui compare les différents estimateurs en terme de biais et d'efficacité illustre nos résultats.

Mots-clés. méthode du Cube, inférence sous le plan de sondage, estimateur greg, probabilité d'inclusion, estimation Monte Carlo, sondage réjectif.

Abstract.
Balanced sampling has received some attention in recent years. There exists a number of procedures leading to a balanced or approximately balanced sample, including the Cube method (Deville and Tillé, 2004) and rejective sampling (Fuller, 2009). In this paper, we examine the properties of several estimation procedures under the rejective sampling procedure of Fuller (2009). In particular, we argue that the estimation procedure advocated by Fuller (2009) may lead to large biases for finite sample sizes if the relationship between the variable of interest and the auxiliary variables used in the construction of the regression estimator is not linear. The results of an extensive simulation study that compares different estimators in terms of relative bias and efficiency support our findings.

Keywords. Cube algorithm, design-based inference, greg estimator, inclusion probability, Monte Carlo approximation, rejective sampling.

Les procédures d'échantillonnage équilibré ont fait l'objet d'un intérêt marqué ces dernières années : Deville et Tillé (2004), Chauvet et Tillé (2006), Fuller (2009) et, Legg et Yu (2010). L'utilisation d'un plan de sondage équilibré vise à éviter la sélection d'échantillons très mal répartis qu'il serait difficile de redresser à l'étape de l'estimation.
Soit une population finie P de taille N. On s'intéresse à l'estimation du total $t_y = \sum_{i \in P} y_i$, où y est une variable d'intérêt. On suppose qu'on dispose, avant échantillonnage, d'un vecteur de variables auxiliaires, $\boldsymbol{z}$, dont les valeurs $\boldsymbol{z}_i$ sont connues pour toutes les unités $i \in P$. Les variables $\boldsymbol{z}$ sont appelées variables d'échantillonnage. Soit $\boldsymbol{Z}$ la matrice dont la i^{eme} ligne est le vecteur $\boldsymbol{z}_i^\top$.
Un échantillon $s \subset P$ est dit $\Psi - \boldsymbol{z}$ équilibré si

$$\hat{\mathbf{t}}_{\mathbf{z}}^{\psi} \equiv \sum_{i \in s} \psi_i^{-1} \mathbf{z}_i = \sum_{i \in P} \mathbf{z}_i \equiv \mathbf{t}_{\mathbf{z}}, \tag{1}$$

où $0 < \psi_i < 1$ pout tout $i \in P$. Un plan de sondage vérifiant la condition (1) est qualifié de plan de sondage équilibré. Il existe plusieurs procédures d'échantillonnage menant à un plan de sondage équilibré ou approximativement équilibré. On s'intéresse dans cet article à la méthode du Cube (Deville et Tillé, 2004) et au tirage réjectif (Hajek, 1981 ; Fuller, 2009).

Soit π_i la probabilité d'inclusion de l'unité i résultant de la procédure d'échantillonnage. Dans le cas de la méthode du Cube, les probabilités d'inclusion, $\pi_1, \ldots, \pi_N$ sont fixées avant l'échantillonnage et la procédure de tirage respecte bien ces probabilités d'inclusion. Toutefois, cette procédure ne sélectionne pas forcement un échantillon qui satisfasse la contrainte (1).

Dans le cas du tirage réjectif, un procédure d'échantillonnage initiale (avec des probabilités d'inclusion notées p_i) est répétée jusqu'à obtenir un échantillon qui vérifie la contrainte d'équilibrage approchée :

$$(\hat{\mathbf{t}}_{\mathbf{z}}^{p} - \mathbf{t}_{\mathbf{z}})^\top V_b(\hat{\mathbf{t}}_{\mathbf{z}}^{p})^{-1}(\hat{\mathbf{t}}_{\mathbf{z}}^{p} - \mathbf{t}_{\mathbf{z}}) \leq \gamma, \tag{2}$$

où $\gamma > 0$ est une tolérance d'équilibrage, $\hat{\mathbf{t}}_{\mathbf{z}}^{p} = \sum_{i \in S_b} p_i^{-1} \mathbf{z}_i$, s_b désigne un échantillon sélectionné suivant le tirage initial et $V_b(.)$ est la variance calculée suivant le plan de sondage initial. A la différence de la méthode du Cube, les π_i sont en général inconnues pour la méthode réjective mais la contrainte d'équilibrage (2) est parfaitement contrôlée.

L'objet de cet article est de comparer les propriétés (biais et variance) de plusieurs estimateurs pour un tirage réjectif d'une part et un tirage effectué avec la méthode du Cube d'autre part. Nous adoptons l'approche par randomisation ("design based") pour l'estimation et l'inférence.

Bibliographie

[1] Chauvet, G., and Tillé, Y. (2006). A fast algorithm for balanced sampling. Computation Statistics, 21, 53-61.
[2] Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : The cube method. Biometrika, 91, 893-912.
[3] Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. Biometrika, 96, 933-944.
[4] Legg, J.C. and Yu, C.L. (2010). A comparison of sample set restriction procedures. Survey Methodology, 36, 69-79.