Bootstrap pour un tirage à plusieurs degrés avec tirage sans remise des unités primaires

Guillaume Chauvet [1]

[1] *Ensai (Crest), Campus de Ker Lann, 35170 Bruz, France, chauvet@ensai.fr*

Résumé. L'échantillonnage à plusieurs degrés est utilisé de façon habituelle quand il n'existe pas de base de sondage, ou quand les unités sont fortement dispersées géographiquement (e.g. [2]). L'échantillonnage à plusieurs degrés introduit une dépendance complexe dans la sélection des unités finales, ce qui rend les résultats asymptotiques difficiles à démontrer. Dans ce travail, nous introduisons une procédure de couplage pour le tirage à plusieurs degrés et un tirage avec/sans remise des unités primaires. Quand la fraction de sondage du premier degré est faible, cette méthode est utilisée pour montrer la consistance d'une méthode de Bootstrap avec remise des unités primaires pour un sondage aléatoire simple sans remise des unités primaires, et la consistance de l'estimateur Bootstrap de variance pour une fonction lisse de moyennes.

Mots-clés. Bootstrap, couplage, estimation de variance, tirage sans remise.

Abstract. Multistage sampling is commonly used for household surveys when there exists no sampling frame, or when the population is scattered over a wide area (e.g. [2]). Multistage sampling usually introduces a complex dependence in the selection of the final units, which makes limiting results quite difficult to prove. In this work, we introduce a coupling method between multistage sampling and with/without replacement sampling at the first stage. When the first-stage sampling fraction is small, this method is used to prove consistency of a with-replacement Bootstrap for simple random without replacement sampling at the first stage, including consistency of Bootstrap variance estimators for smooth functions of means.

Keywords. Bootstrap, coupling, variance estimation, without replacement sampling.

1 Introduction

Nous considérons une population finie $U = \{1, \ldots, N\}$. Les unités sont regroupées au sein de N_I sous-populations non chevauchantes $u_1, \ldots, u_{N_I}$ appelées unités primaires (UP). Nous nous intéressons à l'estimation du total

$$Y = \sum_{k \in U} y_k = \sum_{u_i \in U_I} Y_i$$

pour une variable d'intérêt y, avec $Y_i = \sum_{k \in u_i} y_k$ le sous-total sur l'UP u_i. Nous notons $\hat{Y}_i$ un estimateur sans biais de Y_i, de variance $V_i = V(\hat{Y}_i)$, et nous notons $\hat{V}_i$ un estimateur

de cette variance. Pour étudier les propriétés asymptotiques des plans de sondage et des estimateurs, nous considérons le cadre asymptotique introduit par [5]. Nous supposons que la population U appartient à une suite croissante $\{U_t\}$ de populations finies de tailles N_t, et que le vecteur des valeurs $y_{Ut} = (y_{1t}, \ldots, y_{Nt})^\top$ appartient également à une suite $\{y_{Ut}\}$ de vecteurs de taille N_t. Pour simplifier, l'indice t est supprimé dans la suite, mais les limites sont prises quand $t \to \infty$.

Dans la population $U_I = \{u_1, \ldots, u_{N_I}\}$ d'UP, un échantillon de premier degré S_I est sélectionné selon un plan de sondage $p_I(\cdot)$. Si l'UP u_i est tirée dans S_I, un échantillon de second degré S_i est tiré dans u_i au moyen d'un plan de sondage $p_i(\cdot|S_I)$. Nous supposons que le second degré d'échantillonnage est indépendant du premier (propriété d'invariance). Nous supposons également que, conditionnellement au premier degré de tirage, les échantillons du second degré sont tirés indépendamment d'une UP à l'autre. En dehors de ces deux hypothèses, le plan de sondage utilisé au second degré est quelconque et peut différer d'une UP à l'autre ; il peut par exemple s'agir d'un recensement (ce qui conduit à un tirage par grappes), d'un tirage stratifié, ou d'un plan à plusieurs degrés.

Nous utiliserons les hypothèses suivantes :

H1: $N_I \underset{t\to\infty}{\longrightarrow} \infty$ et $n_I \underset{t\to\infty}{\longrightarrow} \infty$.

H2: Il existe $\delta > 0$ et une constante C_1 telle que $N_I^{-1} \sum_{u_i \in U_I} E|\hat{Y}_i|^{2+\delta} < C_1$.

2 Sondage aléatoire simple sans remise des UP

Nous considérons le cas où l'échantillon S_I est sélectionné dans U_I par sondage aléatoire simple sans remise (SI) de taille n_I, que nous notons $S_I \sim SI(U_I; n_I)$. L'estimateur de Narain-Horvitz-Thompson est défini par

$$\hat{Y} \;=\; \frac{N_I}{n_I} \sum_{u_i \in U_I} I_i \hat{Y}_i = \frac{N_I}{n_I} \sum_{u_i \in S_I} \hat{Y}_i, \tag{1}$$

avec I_i l'indicatrice d'appartenance à l'échantillon S_I pour l'UP u_i. Nous pouvons réécrire cet estimateur sous la forme

$$\hat{Y} = N_I \bar{Z} \quad \text{avec} \quad \bar{Z} = \frac{1}{n_I} \sum_{j=1}^{n_I} Z_j, \tag{2}$$

où l'échantillon S_I est obtenu en tirant n_I fois sans remise une UP dans U_I, et où Z_j désigne l'estimateur du total pour l'UP sélectionnée lors du j-ème tirage. La variance de

$\hat{Y}$ s'écrit

$$V(\hat{Y}) = \frac{N_I^2}{n_I}\left\{(1-f_I)S_{Y,U_I}^2 + \frac{1}{N_I}\sum_{u_i\in U_I} V_i\right\}. \tag{3}$$

avec $S_{Y,U_I}^2 = (N_I-1)^{-1}\sum_{u_i\in U_I}(Y_i-\mu_Y)^2$ la dispersion des sous-totaux Y_i, et $\mu_Y = N_I^{-1}Y$. Sous les hypothèses (H1) et (H2)

$$E(\bar{Z}-\mu_Y)^2 \underset{t\to\infty}{\longrightarrow} 0 \tag{4}$$

et $\hat{Y}$ est faiblement consistant pour Y.

3 Sondage aléatoire simple avec remise des UP

Nous considérons maintenant le cas d'un échantillon S_I^{WR} tiré dans U_I par sondage aléatoire simple avec remise de taille n_I, que nous notons $S_I^{WR} \sim SIR(U_I; n_I)$. Soit W_i le nombre de sélections de l'UP u_i dans S_I^{WR}, et S_I^d de taille n_I^d l'ensemble d'UP distinctes associées à S_I^{WR}. A chaque fois $j = 1,\ldots,W_i$ que l'unité u_i est tirée dans S_I^{WR}, un échantillon de second degré $S_{i[j]}$ est tiré dans u_i. Le total Y est estimé sans biais par l'estimateur de Hansen-Hurwitz

$$\hat{Y}_{WR} = \sum_{u_i\in S_I^d}\frac{1}{E(W_i)}\sum_{j=1}^{W_i}\hat{Y}_{i[j]} = \frac{N_I}{n_I}\sum_{u_i\in S_I^d}\sum_{j=1}^{W_i}\hat{Y}_{i[j]} \tag{5}$$

où $\hat{Y}_{i[j]}$ désigne un estimateur sans biais de Y_i calculé sur $S_{i[j]}$. Nous pouvons réécrire l'estimateur de Hansen-Hurwitz sous la forme

$$\hat{Y}_{WR} = N_I\bar{X} \quad \text{avec} \quad \bar{X} = \frac{1}{n_I}\sum_{j=1}^{n_I} X_j, \tag{6}$$

où l'échantillon S_I^{WR} d'UP est obtenu en tirant n_I fois avec remise une UP dans U_I, et où X_j désigne l'estimateur du total sur l'UP sélectionnée lors du j-ème tirage.

La variance de $\hat{Y}_{WR}$ s'écrit

$$V\left(\hat{Y}_{WR}\right) = \frac{N_I^2}{n_I}\left\{\frac{N_I-1}{N_I}S_{Y,U_I}^2 + \frac{1}{N_I}\sum_{u_i\in U_I} V_i\right\}. \tag{7}$$

Sous les hypothèses (H1) et (H2)

$$E(\bar{X}-\mu_Y)^2 \underset{t\to\infty}{\longrightarrow} 0 \tag{8}$$

et $\hat{Y}_{WR}$ est faiblement consistant pour Y.

4 Une procédure de couplage pour un tirage à plusieurs degrés et un tirage avec/sans remise des UP

La procédure est décrite dans l'Algorithme 1. Conditionnellement à n_I^d, l'échantillon S_I^d obtenu à l'étape 1 est par symétrie un sondage aléatoire simple de taille n_I^d dans U_I, ce qui implique que $S_I^d \cup S_I^c$ est obtenu par sondage aléatoire simple sans remise de taille n_I dans U_I. Cette procédure conduit donc à un échantillon S_I sélectionné selon un plan de sondage SI.

Algorithme 1 Une procédure de couplage entre un sondage aléatoire simple sans remise d'unités primaires, et un sondage aléatoire simple avec remise d'unités primaires

1. Tirer l'échantillon $S_I^{WR} \sim SIR(U_I; n_I)$. Soit S_I^d de taille (aléatoire) n_I^d l'ensemble des UP distinctes de S_I^{WR}.
2. Tirer un échantillon complémentaire $S_I^c \sim SI(U_I \setminus S_I^d; n_I - n_I^d)$ et prendre $S_I = S_I^d \cup S_I^c$.
3. Pour chaque $u_i \in S_I^d$:
   - Chaque fois $j = 1, \ldots, W_i$ que l'unité u_i est tirée dans S_I^{WR}, sélectionner un échantillon de second degré $S_{i[j]}$ avec un estimateur associé $\hat{Y}_{i[j]}$ pour $\hat{Y}_{WR}$.
   - Prendre $S_i = S_{i[1]}$ et $\hat{Y}_i = \hat{Y}_{i[1]}$ pour $\hat{Y}$.
4. Pour chaque $u_i \in S_I^c$, sélectionner un échantillon de second degré S_i avec l'estimateur associé $\hat{Y}_i$ pour $\hat{Y}$.

Proposition 1 *Si les échantillons S_I^{WR} et S_I sont sélectionnés selon l'algorithme 1, alors*

$$\frac{E(\hat{Y}_{WR} - \hat{Y})^2}{V(\hat{Y}_{WR})} \leq \frac{n_I - 1}{N_I - 1}. \tag{9}$$

5 Bootstrap avec remise des unités primaires

Nous considérons la méthode de Bootstrap avec remise des UP décrite par exemple dans [6]. En utilisant la notation introduite dans l'équation (2), soit $(Z_1, \ldots, Z_{n_I})^\top$ l'échantillon d'estimateurs sous un plan de sondage SI pour les UP. Soit $(Z_1^*, \ldots, Z_m^*)^\top$ obtenu en échantillonnant m fois indépendamment et avec remise dans $(Z_1, \ldots, Z_{n_I})^\top$. De façon analogue, en utilisant la notation introduite dans l'équation (6), soit $(X_1, \ldots, X_{n_I})^\top$ l'échantillon

d'estimateurs sous un plan de sondage SIR pour les UP, et $(X_1^*, \ldots, X_m^*)^\top$ obtenu en échantillonnant m fois indépendamment et avec remise dans $(X_1, \ldots, X_{n_I})^\top$.

Nous démontrons tout d'abord la consistance de la procédure de Bootstrap. Nous notons

$$\bar{Z}_m^* = \frac{1}{m}\sum_{j=1}^{m} Z_j^*, \qquad s_Z^{*2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(Z_j^* - \bar{Z}_m^*\right)^2,$$

et

$$\bar{X}_m^* = \frac{1}{m}\sum_{j=1}^{m} X_j^*, \qquad s_X^{*2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(X_j^* - \bar{X}_m^*\right)^2.$$

Theorème 1 *On suppose que* $m \underset{t\to\infty}{\longrightarrow} \infty$. *Alors la quantité pivotale Bootstrap*

$$\frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*}$$

converge en loi vers une normale centrée réduite.

La preuve est basée sur le fait qu'en tirant les échantillons S_I et S_I^{WR} selon l'algorithme 1, et en utilisant ensuite les mêmes poids de rééchantillonnage pour constituer $\bar{Z}_m^*$ et $\bar{X}_m^*$, les statistiques pivotales $\dfrac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*}$ et $\dfrac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s_X^*}$ sont proches. Le théorème découle alors du Théorème 2.1 de [1].

Nous considérons maintenant le cas où $y_k = (y_{1k}, \ldots, y_{qk})^\top$ est multivarié, et désigne la valeur prise sur l'unité k par un vecteur d'intérêt y de taille q. Nous nous intéressons à un paramètre $\theta = f(\mu_Y)$, où $f : \mathbb{R}^q \longrightarrow \mathbb{R}$ est supposée différentiable sur $\mathbb{R}^q$ avec des dérivées partielles bornées. Nous supposons que $f'(\mu_Y) \neq 0$. Avec un plan de sondage SI pour les UP, l'estimateur plug-in de θ est $\hat{\theta} = f(\bar{Z})$. Avec un plan de sondage SIR pour les UP, l'estimateur plug-in de θ est $\hat{\theta}_{WR} = f(\bar{X})$.

Proposition 2 *On suppose que les échantillons* S_I^{WR} *et* S_I *sont sélectionnés selon l'Algorithme 1. On suppose que les hypothèses (H1) et (H2) sont vérifiées. On suppose que* $f_I \underset{t\to\infty}{\longrightarrow} 0$. *Alors :*

$$E(\|\bar{Z} - \bar{X}\|^2) = o(n_I^{-1}), \tag{10}$$

$$E(\hat{\theta} - \hat{\theta}_{WR})^2 = o(n_I^{-1}) \tag{11}$$

avec $\|\cdot\|$ *la norme euclidienne.*

Proposition 3 *On suppose que les échantillons S_I^{WR} et S_I sont sélectionnés selon l'Algorithme 1. On suppose que les hypothèses (H1) et (H2) sont vérifiées. On suppose que $f_I \underset{t\to\infty}{\longrightarrow} 0$ et que $m \underset{t\to\infty}{\longrightarrow} \infty$. Alors :*

$$E(\|\bar{Z}^* - \bar{X}^*\|^2) = o(m^{-1}) + o(n_I^{-1}), \tag{12}$$

$$E(\hat{\theta}^* - \hat{\theta}^*_{WR})^2 = o(m^{-1}) + o(n_I^{-1}). \tag{13}$$

Si le Bootstrap avec remise donne une estimation de variance consistante dans le cas d'un plan de sondage SIR pour les UP, nous avons $\dfrac{V_{\{X\}}(\hat{\theta}^*_{WR})}{V(\hat{\theta}_{WR})} \underset{Pr}{\longrightarrow} 1$ en notant $\underset{Pr}{\longrightarrow}$ la convergence en probabilité. En utilisant les Propositions 2 et 3, on montre alors que le Bootstrap avec remise est alors également consistant dans le cas d'un plan de sondage SI pour les UP.

Bibliographie

[1] Bickel, P. J. et Freedman, D.A. (1981), *Some asymptotic theory for the bootstrap*, The Annals of Statistics, 9, 1196-1217.

[2] Ezzati, T.M. et Hoffman, K. et Judkins, D.R. et Massey, J.T. et Moore, T.F. (1992), *Sample design: Third National Health and Nutrition Examination Survey*, Vital and Health Statistics, 2, 113.

[3] Fuller, W.A. (2009), *Sampling Statistics*, New-York, Wiley.

[4] Hajek, J. (1960), *Limiting distributions in simple random sampling from a finite population*, Publications of the Mathematics Institute of the Hungarian Academy of Science, 5, 361-74.

[5] Isaki, C.T. et Fuller, W.A. (1982), *Survey design under the regression superpopulation model*, Journal of the American Statistical Association, 77, 89-96.

[6] Rao, J.N.K et Wu, C.F.J. (1988), *Resampling inference with complex survey data*, Journal of the American Statistical Association, 83, 231-241.

[7] Särndal, C.-E. et Swensson, B. et Wretman, J.H. (1992), *Model Assisted Survey Sampling*, New-York, Springer-Verlag.