INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2007-39

Bootstrap pour un tirage à plusieurs degrés avec échantillonnage à forte entropie à chaque degré

G. CHAUVET[1]

[1] CREST-ENSAI, Laboratoire de Statistique d'Enquête, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz, France. (chauvet@ensai.fr)

BOOTSTRAP POUR UN TIRAGE A PLUSIEURS DEGRÉS AVEC ÉCHANTILLONNAGE A FORTE ENTROPIE A CHAQUE DEGRÉ

Guillaume Chauvet[1]

Décembre 2007

Abstract

In this paper we propose a Bootstrap technique for multistage sampling with large entropy sampling designs at each stage. Our method enables easy variance estimates for both linear and non linear statistics. It consists in a correction of the method originally proposed by Gross (1980) for simple random sampling, and known not to be consistent for multistage sampling. The idea consists in building a pseudo-population of pseudo Primary Sampling Units, based on the original sample. The first stage drawing is performed, but the second stage drawing is corrected to get the usual variance estimate in the linear case. Two stage sampling with simple random sampling at each stage and self-weighted two stage sampling are covered by the proposed Bootstrap technique. The method is evaluated through a limited set of simulations.

Keywords : Bootstrap ; Entropy ; Linearization by means of the influence function ; Multistage Sampling ; Unequal probability sampling ; Variance estimation.

Résumé

Nous proposons dans cet article une méthode de Bootstrap permettant de traiter le cas d'un échantillonnage à plusieurs degrés, avec tirage à probabilités inégales à forte entropie à chaque degré. Cette méthode permet d'estimer facilement la précision de statistiques linéaires et non linéaires. Elle adapte la méthode proposée pour le sondage aléatoire simple par Gross (1980), connue pour ne pas être directement généralisable à un tirage à plusieurs degrés. L'idée consiste à construire à partir de l'échantillon une population constituée de pseudo unités primaires dans laquelle on reproduit le tirage du 1er degré, mais pour laquelle le tirage du 2nd degré est modulé afin de reproduire l'estimateur habituel sans biais de variance dans le cas linéaire. Le Bootstrap proposé couvre les cas particuliers importants du tirage avec sondage aléatoire simple à chaque degré et celui du tirage autopondéré. La méthode est évaluée à l'aide de simulations.

Mots Clés : Bootstrap ; Echantillonnage à plusieurs degrés ; Estimation de variance ; Tirage à probabilités inégales ; Entropie ; Linéarisation par la fonction d'influence.

[1] Guillaume Chauvet (chauvet@ensai.fr), Laboratoire de Statistique d'Enquête, CREST-ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz, France

1. INTRODUCTION

Contrairement à la plupart des techniques d'échantillonnage, le tirage à plusieurs degrés est utilisé non pour améliorer la précision de l'enquête, mais pour réduire les coûts. La concentration des unités échantillonnées permet d'éviter la constitution d'une base de sondage exhaustive, et limite les déplacements éventuels des enquêteurs. De nombreuses enquêtes à grande échelle utilisent un plan stratifié à plusieurs degrés. C'est le cas à l'Insee pour le tirage de l'échantillon maître d'unités primaires qui sert de base de sondage aux enquêtes sur les ménages. Afin de réduire la variance associée à l'échantillonnage du premier degré, les unités primaires y sont sélectionnées à probabilités proportionnelles à leur taille, ce qui complique l'estimation de variance.

La précision de ces enquêtes peut être estimée à l'aide de la linéarisation par la fonction d'influence (Deville 1999, Goga et al. 2007) en utilisant des logiciels tels que POULPE (Caron 1999). Nous proposons ici une méthode de calcul de précision par Bootstrap. Cette méthode a l'avantage de fournir des poids de rééchantillonnage qui peuvent être intégrés aux données d'enquête et utilisés a posteriori pour produire des estimations de précision pour une large catégorie de statistiques sans nécessiter de l'utilisateur aucune connaissance sur le plan de sondage d'origine.

Il existe une abondante littérature sur le Bootstrap en population finie dans le cas d'un échantillonnage à un seul degré (Gross 1980, Bickel et Freedman 1984, Chao et Lo 1984, Mac Carthy et Snowden 1985, Deville 1986, Rao et Wu 1988, Sitter 1992ab, Booth, Butler et Hall 1994, Bertail et Combris 1997, Chauvet et Deville 2007). Les résultats sont plus limités dans le cas d'un échantillonnage à plusieurs degrés. Une méthode simple et efficace consiste à assimiler le tirage du premier degré à un plan avec remise, si les probabilités d'inclusion correspondantes

sont faibles, et à ne bootstrapper que les unités primaires. Si ces probabilités ne sont pas négligeables, on peut utiliser les méthodes proposées par Rao et Wu (1988) et Sitter (1992a,b) mais elles ne sont applicables qu'avec un sondage aléatoire simple à chaque degré. La méthode que nous proposons ici est une correction d'une extension naïve de la méthode de Gross (1980). On peut la voir également comme une adaptation de la méthode de Bootstrap des unités primaires pour un taux de sondage non négligeable au 1er degré. Elle est applicable avec un échantillonnage à grande entropie à chaque degré, ce qui couvre les cas particuliers importants du sondage aléatoire simple à chaque degré et du plan autopondéré.

La papier est organisé de la façon suivante. La section 2 introduit nos notations sur l'échantillonnage à deux degrés. La méthode du Bootstrap des unités primaires est brièvement rappelée en section 3. La section 4 donne une méthode générale de Bootstrap pour un tirage à deux degrés, qui se généralise facilement à un tirage à trois degrés et plus. La section 5 propose une simplification de l'algorithme. Les propriétés du Bootstrap sont évaluées en section 6 à l'aide de quelques simulations. Le détail des preuves est donné en annexe.

2. ECHANTILLONNAGE A DEUX DEGRÉS

Afin d'alléger les notations, nous nous plaçons dans le cas d'un échantillonnage à deux degrés avec une seule strate de tirage. Les résultats présentés s'étendent de façon immédiate au cas d'un nombre fini de strates. Le cas d'un tirage stratifié où le nombre de strates devient grand ne sera pas considéré ici. On suppose que le plan à deux degrés vérifie les hypothèses classiques d'invariance et d'indépendance (voir Särndal et al., 1992, page 134).

La population U_{GR} est constituée de M unités primaires (UP). Chaque UP u_i contient N_i unités secondaires (US). On note π_{Ii} la probabilité d'inclusion de u_i pour le plan de sondage p_I

utilisé au 1er degré, c'est-à-dire la probabilité de sélectionner u_i dans l'échantillon S_I d'UP. Soit $m = \sum_{u_i \in U_{GR}} \pi_{Ii}$ la taille de S_I. On note également π_{Iij} la probabilité de sélectionner conjointement u_i et u_j dans S_I et $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$. Si l'UP u_i est sélectionnée dans S_I, on note S_i l'échantillon d'US sélectionné dans u_i selon un plan de sondage p_i, $\pi_{k/i}$ la probabilité de sélectionner l'US k appartenant à u_i dans S_i, $\pi_{kl/i}$ la probabilité de sélectionner les US k et l appartenant à u_i dans S_i, et $\Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i}$. La taille de l'échantillon S_i est notée $n_i = \sum_{k \in u_i} \pi_{k/i}$. On notera $U = \bigcup_{i=1}^{M} u_i$ l'ensemble des US. L'échantillon final S d'US est donné par la réunion des S_i.

L'information relevée sur chaque individu est résumée par une variable z_k éventuellement vectorielle. En adoptant la notation introduite par Deville (1999), nous notons $M = \sum_{k \in U} \delta_{z_k}$ la mesure qui place une masse unité sur chaque point z_k de la population. Nous supposons que les valeurs prises par la variable z sont distinctes, quitte à introduire un rang arbitraire pour chaque unité de la population qui constitue alors une des composantes de z. Par abus de notation, nous écrirons seulement δ_k au lieu de δ_{z_k}. On note encore $\hat{M} = \sum_{k \in S} \frac{\delta_k}{\pi_k}$ la mesure qui affecte la masse $1/\pi_k$ à chaque unité z_k de S et 0 partout ailleurs. Nous nous intéressons à un paramètre d'intérêt qui peut s'exprimer comme une fonctionnelle $\theta = \theta(M)$ de la mesure M. $\theta(M)$ est estimé par $\theta(\hat{M})$ selon un principe de substitution. L'estimateur $\theta(\hat{M})$ sera appelé estimateur par substitution de $\theta(M)$.

Soit y la variable d'intérêt, prenant la valeur y_k sur l'unité k de U. Le total $t_y(U) = \sum_{k \in U} y_k$ de la variable y sur U peut être estimé sans biais par son π-estimateur

$$\hat{t}_{y\pi}(S) = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{u_i \in S_I} \sum_{k \in S_i} \frac{y_k}{\pi_{Ii}\pi_{k/i}}.$$

Le total $t_y(u_i)$ de la variable y sur u_i est estimé sans biais par $\hat{t}_{y\pi}(S_i) = \sum_{k \in S_i} \frac{y_k}{\pi_{k/i}}$. On a donc encore la relation :

$$\hat{t}_{y\pi}(S) = \sum_{u_i \in S_I} \frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}}.$$

La variance de $\hat{t}_{y\pi}(S)$ (voir par exemple Tillé, 2001) est égale à

$$\begin{aligned} V\left(\hat{t}_{y\pi}(S)\right) &= V\left(\sum_{u_i \in S_I} \frac{t_y(u_i)}{\pi_{Ii}}\right) + \sum_{u_i \in U_{GR}} \frac{V\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i \in U_{GR}} \sum_{u_j \in U_{GR}} \frac{t_{yi}}{\pi_{Ii}} \frac{t_{yj}}{\pi_{Ij}} \Delta_{Iij}}_{V_{UP}} + \underbrace{\sum_{u_i \in U_{GR}} \frac{V\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}}}_{V_{US}} \end{aligned} \qquad (2.1)$$

avec $V\left(\hat{t}_{y\pi}(S_i)\right) = \sum_{k \in u_i} \sum_{l \in u_i} \frac{y_k}{\pi_{k/i}} \frac{y_l}{\pi_{l/i}} \Delta_{kl/i}$. La variance se décompose en deux termes V_{UP} et V_{US}, respectivement associés aux premier et second degré de tirage. Si les probabilités d'inclusion π_{Iij} et $\pi_{kl/i}$ sont toutes strictement positives, un estimateur sans biais de variance est donné par

$$\begin{aligned}\hat{V}\left(\hat{t}_{y\pi}(S)\right) &= \hat{V}\left(\sum_{u_i\in S_I}\frac{t_y(u_i)}{\pi_{Ii}}\right) + \sum_{u_i\in U_{GR}}\frac{\hat{V}\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i\in S_I}\sum_{u_j\in S_I}\frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}}\frac{\hat{t}_{y\pi}(S_j)}{\pi_{Ij}}\frac{\Delta_{Iij}}{\pi_{Iij}}}_{\hat{V}_A} + \underbrace{\sum_{u_i\in S_I}\frac{\hat{V}\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}}}_{\hat{V}_B}\end{aligned} \qquad (2.2)$$

avec $\hat{V}\left(\hat{t}_{y\pi}(S_i)\right)=\sum_{k\in S_i}\sum_{l\in S_i}\frac{y_k}{\pi_{k/i}}\frac{y_l}{\pi_{l/i}}\frac{\Delta_{kl/i}}{\pi_{kl/i}}$. Il est important de noter que $\hat{V}\left(\hat{t}_{y\pi}(S)\right)$ n'est pas un estimateur sans biais terme à terme de $V\left(\hat{t}_{y\pi}\right)$. En particulier, V_{US} et $\hat{V}_B$ ne sont pas du même ordre de grandeur.

Pour l'estimateur $\hat{V}\left(\hat{t}_{y\pi}(S)\right)$, la difficulté réside dans le calcul des probabilités d'inclusion doubles, aussi bien au premier degré qu'au second. Hajek (1964) offre une solution dans le cas d'un tirage de taille fixe à entropie maximale, appelé tirage réjectif. Il donne une approximation de variance, qui conduit à un estimateur asymptotiquement sans biais. Ce résultat est étendu par Berger (1998) aux plans de sondage proches de l'entropie maximale tels que l'échantillonnage de Rao-Sampford (Rao 1965, Sampford 1967) et le tirage successif (Hajek, 1964). Dans le cas d'un échantillonnage à plusieurs degrés, on peut utiliser (voir Tillé, 2001) :

$$\begin{aligned}\hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S)\right) &= \hat{V}_{HAJ}\left(\sum_{u_i\in S_I}\frac{t_y(u_i)}{\pi_{Ii}}\right) + \sum_{u_i\in U_{GR}}\frac{\hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i\in S_I}\sum_{u_j\in S_I}\frac{c_{Ii}}{\pi_{Ii}}\left(\hat{t}_{y\pi}(S_i)-\hat{\tilde{t}}_{yi}\right)^2}_{\hat{V}_A^{HAJ}} + \underbrace{\sum_{u_i\in S_I}\frac{\hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}}}_{\hat{V}_B^{HAJ}}\end{aligned} \qquad (2.3)$$

où $\hat{\tilde{t}}_{yi}=\pi_{Ii}\frac{\sum_{u_j\in S_I}c_{Ij}\frac{\hat{t}_{y\pi}(S_j)}{\pi_{Ij}}}{\sum_{u_j\in S_I}c_{Ij}}$ et $c_{Ij}=1-\pi_{Ij}$, avec

$$\hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right) = \sum_{k\in S_i} \frac{c_{k/i}}{\pi_{k/i}^2}\left(y_k - \tilde{y}_k\right)^2,$$

où $\tilde{y}_k = \pi_{k/i} \dfrac{\sum_{l\in S_i} c_{l/i} \dfrac{y_l}{\pi_{l/i}}}{\sum_{l\in S_i} c_{l/i}}$ et $c_{l/i} = 1 - \pi_{l/i}$.

$\hat{V}_A$ et $\hat{V}_B$ peuvent être remplacés respectivement par $\hat{V}_A^{HAJ}$ et $\hat{V}_B^{HAJ}$ asymptotiquement sans biais, et l'estimateur de Hajek présenté en (2.3) donne une estimation de variance asymptotiquement sans biais de (2.1) si l'échantillonnage implique un plan à forte entropie à chaque degré. L'asymptotique est ici celle proposée par Hajek (1964), en supposant que $d_I = \sum_{u_i\in U_{GR}} \pi_{Ii}(1-\pi_{Ii}) \to \infty$ et $d_{k/i} = \sum_{k\in u_i} \pi_{k/i}(1-\pi_{k/i}) \to \infty$ pour chaque UP u_i.

Le tirage réjectif a été étudié par Chen, Dempster et Liu (1994) et Deville (2000). Pour un inventaire très complet des différents estimateurs de variance proposés pour un tirage à forte entropie, voir Brewer et Donadio (2003) et Matei et Tillé (2005). En particulier, l'estimateur proposé en (2.3) peut être modifié en prenant $c_{Ij} = \dfrac{m}{m-1}\left(1-\pi_{Ij}\right)$ et $c_{l/i} = \dfrac{n_i}{n_i - 1}\left(1-\pi_{l/i}\right)$, ce qui permet de restituer l'estimateur sans biais habituel dans le cas d'un sondage aléatoire simple à chaque degré.

3. LE BOOTSTRAP DES UNITÉS PRIMAIRES

Une méthode simple de Bootstrap, décrite dans l'algorithme 1, consiste à rééchantillonner les unités primaires sans reproduire de second degré de tirage (voir par exemple Rao et al. 1992).

Algorithme 1 : Bootstrap des unités primaires pour le tirage à deux degrés

Etape 1 : Sélectionner à probabilités égales et avec remise un échantillon S_I^* d'unités primaires dans S_i, de taille $m-1$.

Etape 2 : Calculer $\hat{\theta}^* = \theta(\hat{M}^*)$ où $\hat{M}^* = \sum_{u_i \in S_I^*} \sum_{k \in S_i} \frac{\delta_k}{\pi_{Ii}\pi_{k/i}}$.

Etape 3 : Répéter B fois les étapes 1 et 2 pour obtenir $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$. $V(\hat{\theta})$ est estimée par $\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_\cdot^*\right)^2$, où $\hat{\theta}_\cdot^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

Dans le cas de l'estimation d'un total, on a $\theta = t_y$ et en notant $\hat{\theta}^* = \hat{t}_{y\pi}^*$ on peut montrer que

$$V\left(\hat{t}_{y\pi}^* \middle| S\right) = \frac{m}{m-1}\sum_{u_i \in S_I}\left(\frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}} - \frac{\hat{t}_{y\pi}(S)}{m}\right)^2.$$

On obtient une estimation sans biais de variance si les unités primaires de l'échantillon S_I ont été sélectionnées avec remise (Särndal et al., 1992, page 151). En pratique, la variance sera surestimée si un plan de sondage sans remise plus efficace que le plan avec remise est utilisé pour la sélection des unités primaires, ce qui est le cas pour le plan réjectif (Qualité, 2007).

4. MÉTHODE GÉNÉRALE DE BOOTSTRAP

Dans le cas d'un sondage aléatoire simple à un seul degré, Gross (1980) propose une méthode simple et intuitive de calcul de précision par Bootstrap. Chaque individu de l'échantillon est dupliqué p fois pour créer une pseudopopulation U^*, p désignant l'inverse du taux de sondage. La variance est alors estimée par des tirages répétés dans U^*, selon le plan de sondage d'origine.

L'extension de cette méthode à p non entier a été étudié par Bickel et Freedman (1984), Chao et Lo (1984) et Booth et al. (1994). L'extension de Booth, Butler et Hall a été généralisée au cas d'un tirage à probabilités inégales et à forte entropie par Chauvet et Deville (2007).

L'algorithme 2 donne une méthode générale de Bootstrap pour l'échantillonnage à deux degrés, valide pour des probabilités d'inclusion quelconques au premier degré. Cette méthode présente des similarités avec le Bernoulli Bootstrap de Funaoka et al. (2006). La consistance de la méthode dans le cas linéaire est établie en Annexe 1. Nous supposons dans cet algorithme que les inverses de probabilités d'inclusion du 2nd degré sont approximativement entières, ce qui induit un biais conditionnel de l'ordre de

$$\sum_{u_i \in S_I} \frac{1}{\pi_{Ii}} \sum_{k \in S_i} \left(\left[\frac{1}{\pi_{k/i}} \right] - \frac{1}{\pi_{k/i}} \right) \frac{1-\pi_{k/i}}{\pi_{k/i}^2} (y_k - \tilde{y}_k)^2$$

qui est limité si les probabilités $\pi_{k/i}$ sont faibles. On note ici $[x]$ pour l'entier le plus proche de x, et $[x-1/2]$ donne la partie entière de x. Notons que l'étape 2 de l'algorithme peut poser problème si le plan de sondage p_I est de taille fixe et que la somme des probabilités α_{Ii} n'est pas entière. L'algorithme nécessite en fait seulement que la procédure utilisée à l'étape 2 respecte les probabilités d'inclusion α_{Ii} ; on peut donc par exemple utiliser alternativement un tirage poissonien.

Cet algorithme peut être facilement généralisé au cas d'un plan de sondage à r degrés. Avec r étapes de duplications, on obtient une pseudopopulation U^* image de la population d'origine. La variance est estimée par applications répétées du plan de sondage dans U^*, le tirage étant modulé à chaque degré $d \geq 2$ selon le même procédé que dans l'algorithme 1, en fonction des

probabilités d'inclusion du degré $d-1$. On suppose là encore que les probabilités d'inclusion de chaque degré $d \geq 2$ restent faibles.

Algorithme 2 : Bootstrap général pour le tirage à deux degrés

Etape 1 : Soit $u_i \in S_I$. Chaque unité k de S_i est dupliquée $[1/\pi_{k/i}]$ fois pour créer une pseudo UP que l'on note u_i^*.

Etape 2 : Chaque couple (S_i, u_i^*) est dupliqué $[1/\pi_{Ii} - 1/2]$ fois. Soit $\alpha_{Ii} = 1/\pi_{Ii} - \lfloor 1/\pi_{Ii} \rfloor$. On complète les couples déjà dupliqués par un échantillon sélectionné dans $\{(S_i, u_i^*), u_i \in S_I\}$ selon le plan p_I, avec les probabilités d'inclusion $\alpha_{Ii}, u_i \in S_I$. On obtient ainsi une population U_{GR}^* de pseudo UP.

Etape 3 : On tire un échantillon S_I^* dans U_{GR}^* selon le plan p_I, avec les probabilités π_{Ii}.

Etape 4 : Soit $(S_i, u_i^*) \in S_I^*$. On tire un échantillon S_i^{**} de pseudo US dans u_i^* selon le plan de sondage du 2nd degré d'origine. On prend $S_i^* = S_i^{**}$ avec une probabilité π_{Ii}, et $S_i^* = S_i$ avec une probabilité $1 - \pi_{Ii}$.

La même procédure est appliquée pour chaque couple $(S_i, u_i^*) \in S_I^*$. Le rééchantillon S_* est donné par la réunion des S_i^*.

Etape 5 : Les étapes 3 et 4 sont répétées C fois, pour obtenir les rééchantillons $S_*^1, \cdots, S_*^C$. Soit $v = \frac{1}{C-1} \sum_{c=1}^{C} \left(\hat{\theta}_c^* - \hat{\theta}_\cdot^*\right)^2$, où $\hat{\theta}_\cdot^* = \frac{1}{C} \sum_{c=1}^{C} \hat{\theta}_c^*$.

Etape 6 : Les étapes 2 à 5 sont répétées B fois, pour obtenir $v_1, \cdots, v_B$. $V(\hat{\theta})$ est estimée par $\hat{V}_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} v_b$.

5. MÉTHODE SIMPLIFIÉE DE BOOTSTRAP

L'étape 2 de l'algorithme 2 est nécessaire si les probabilités d'inclusion au 1er degré sont fortes, ce qui peut être courant en pratique, voir Funoaka et al. (2007, page 155). Dans le cas contraire, on peut introduire une simplification présentée dans l'algorithme 3, qui limite le volume de calcul car elle évite la double étape de randomisation nécessitée par l'algorithme 2. Cet algorithme simplifié permet d'intégrer aux données d'enquête les B variables de poids Bootstrap associées au rééchantillonnage, facilitant la production ultérieure d'indicateurs de précision par les utilisateurs. Notons que si les probabilités d'inclusion du 1er degré sont négligées, l'algorithme sélectionne systématiquement à l'étape 4 l'échantillon S_i d'origine : la méthode est alors très similaire au Bootstrap des unités primaires. Notre algorithme apparaît donc comme une correction de cette méthode, permettant de tenir compte du tirage sans remise des UP.

5. SIMULATIONS

Nous générons une population selon le même procédé que dans Funaoka et al. (2006). La moyenne β_i de la variable auxiliaire dans l'Unité Primaire u_i est générée selon une loi normale $N(\beta, \sigma^2)$ pour $i = 1, \cdots, M$. Puis la valeur x_{ik} de la variable auxiliaire x sur l'Unité Secondaire k de l'UP u_i est générée selon le modèle

$$x_{ik} = \beta_i + \varepsilon_{ik} \left(k = 1, \cdots, N_i \, ; i = 1, \cdots, M\right)$$

Algorithme 3 : Bootstrap simplifié pour le tirage à deux degrés

Etape 1 : Soit $u_i \in S_I$. Chaque unité k de S_i est dupliquée $\lfloor 1/\pi_{k/i} \rfloor$ fois pour créer une pseudo UP que l'on note u_i^*.

Etape 2 : Chaque couple (S_i, u_i^*) est dupliqué $[1/\pi_{Ii}]$ fois pour obtenir U_{GR}^*.

Etape 3 : On tire un échantillon S_I^* dans U_{GR}^* selon le plan de sondage p_I.

Etape 4 : Soit $(S_i, u_i^*) \in S_I^*$. On tire un échantillon S_i^{**} de pseudo US dans u_i^* selon le plan de sondage p_i. On prend $S_i^* = S_i^{**}$ avec une probabilité π_{Ii}, et $S_i^* = S_i$ avec une probabilité $1 - \pi_{Ii}$.

La même procédure est appliquée pour chaque couple $(S_i, u_i^*) \in S_I^*$. Le rééchantillon S_* est donné par la réunion des S_i^*.

Etape 5 : Les étapes 3 et 4 sont répétées B fois pour obtenir les rééchantillons $S_*^1, \cdots, S_*^B$.

$V(\hat{\theta})$ est estimée par $\hat{V}_{boot}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}(S_*^b) - \hat{\theta}_*^m \right)^2$, où $\hat{\theta}(S_*^b)$ donne la valeur de l'estimateur sur le rééchantillon S_*^b et $\hat{\theta}_*^m = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}(S_*^b)$.

où $\varepsilon_{ik} \approx N(0, (1-\rho)\sigma^2/\rho)$. On utilise ici $\beta = 100$, $\sigma = 10$, $\rho = 0.1$. La valeur y_{ik} (respectivement z_{ik}) de la variable d'intérêt y (respectivement z) sur l'US k de l'UP u_i est générée selon le modèle

$$y_{ik} = a + bx_{ik} + e_{ik} \ (k = 1, \cdots, N_i \ ; i = 1, \cdots, M)$$

où e_{ik} suit une loi normale $N(0, \sigma_y^2)$ avec $\sigma_y = 25$ (respectivement, e_{ik} suit une loi normale $N(0, \sigma_z^2)$ avec $\sigma_z = 40$). On utilise un plan à deux degrés autopondéré, avec un tirage réjectif

de taille m à probabilités proportionnelles à x au 1er degré, et un sondage aléatoire simple au 2nd degré de taille n_0 pour chaque UP.

On génère une population de taille $M = 310$, comptant $M_1 = 250$ unités primaires de taille 240, $M_2 = 100$ unités primaires de taille 300 et $M_3 = 60$ unités primaires de taille 400. On tire un échantillon d'unités primaires de taille $m = 15$ (respectivement $m = 75$), et un échantillon de taille $n_0 = 20$. On considère les estimateurs des totaux t_y et t_z des variables y et z, le ratio t_y / t_z et le coefficient de corrélation des variables y et z. On compare la méthode de Bootstrap proposée (notée BWO) avec la méthode du Bootstrap des Unités Primaires (notée BPU).

Dans le cas de l'estimation du total, la vraie variance est calculée. Dans le cas de l'estimation du ratio et du coefficient de corrélation, la vraie variance est approchée à l'aide de $10\,000$ simulations. Pour chacune des méthodes de Bootstrap, on prélève $2\,000$ échantillons indépendants pour chacun desquels $B = 400$ rééchantillons Bootstrap sont prélevés. Les intervalles de confiance sont déterminés à l'aide de la méthode des percentiles.

Quelques points importants sont résumés ci-dessous :

- Les deux méthodes donnent des résultats analogues sur les taux de couverture.
- Avec un faible taux de sondage des unités primaires, les deux méthodes donnent un biais négatif pour l'estimation de variance. Conformément à la théorie, le biais est de l'ordre de $1 - 1/m = 7\,\%$ pour l'estimation d'un total dans le cas de la méthode de type BWO.

- Avec un fort taux de sondage des unités primaires, la méthode BUP conduit à une surestimation de variance dans le cas de l'estimation d'un total quand la méthode BWO donne un biais très faible. Pour l'estimation de statistiques homogènes, les deux méthodes donnent un biais analogue dans le cas du ratio, et la méthode BUP est plus faiblement biaisée pour le coefficient de corrélation.
- La méthode de Bootstrap proposée conduit à une estimation de variance plus stable.

6. CONCLUSION

Nous proposons dans cet article une méthode de Bootstrap d'estimation de précision dans le cas d'un tirage multidegrés avec un taux de sondage non négligeable des unités primaires. Notre méthode est applicable dans le cas d'un tirage à entropie forte à chaque degré, ce qui couvre les cas particuliers importants du sondage aléatoire simple à chaque degré et du tirage autopondéré. La méthode proposée conduit à une estimation de variance asymptotiquement sans biais dans le cas de l'estimation d'un total et plus stable qu'avec le Bootstrap des unités primaires.

Tab. 1 : Taux de couverture, biais relatif et stabilité relative pour deux méthodes de Bootstrap dans le cas d'un tirage à 2 degrés autopondéré avec sélection de m=15 unités primaires

	Taux d'erreur unilatéral							
	5 %			10 %				
Méthode	L [a]	U [b]	L+U	L	U	L+U	Biais	Stabilité
Total de y								
BUP	7.65	6.25	13.90	13.10	10.05	23.15	-4.6	0.37
BWO	7.65	6.25	13.90	13.20	10.15	23.35	-7.1	0.36
Total de z								
BUP	8.05	6.45	14.50	12.10	11.45	23.55	-5.1	0.36
BWO	8.00	6.40	14.40	12.20	11.55	23.75	-7.2	0.35
Ratio								
BUP	6.85	7.25	14.10	12.30	12.65	24.95	-5.5	0.38
BWO	6.90	7.40	14.30	12.30	12.65	24.95	-5.8	0.36
Corrélation								
BUP	6.15	7.00	13.15	11.25	12.75	24.00	-3.6	0.39
BWO	6.25	7.05	13.30	11.20	12.75	23.95	-3.7	0.38

Note de lecture

(a) % des IC pour lesquels le paramètre se situe en-dessous de l'IC (niveau théorique : 5 %)

(b) % des IC pour lesquels le paramètre se situe au-dessus de l'IC (niveau théorique : 5 %)

(c) Biais rel. = $100\times\left(E\left(\hat{V}\right)/V\left(\hat{\theta}\right)-1\right)$

(d) Stabilité = $\sqrt{V\left(\hat{V}\right)}\Big/V\left(\hat{\theta}\right)$

Tab. 2 : Taux de couverture, biais relatif et stabilité relative pour deux méthodes de Bootstrap dans le cas d'un tirage à 2 degrés autopondéré avec sélection de m = 75 unités primaires

	Taux d'erreur unilatéral							
	5 %			10 %				
Méthode	L [a]	U [b]	L+U	L	U	L+U	Biais	Stabilité
Total de y								
BUP	3.85	3.80	7.65	8.55	8.15	16.70	+16.1	0.20
BWO	4.00	3.90	7.90	8.65	8.35	17.00	-2.1	0.15
Total de z								
BUP	4.65	4.10	8.75	10.10	8.60	18.70	+15.0	0.20
BWO	4.75	4.10	8.85	10.10	8.65	18.75	+0.4	0.16
Ratio								
BUP	6.40	4.55	10.95	12.10	9.30	21.40	+0.9	0.19
BWO	6.30	4.55	10.85	12.05	9.30	21.35	-0.9	0.15
Corrélation								
BUP	5.10	4.85	9.95	10.15	10.95	21.10	+0.3	0.19
BWO	5.00	4.85	9.85	10.20	10.90	21.10	-3.7	0.15

Note de lecture

(a) % des IC pour lesquels le paramètre se situe en-dessous de l'IC (niveau théorique : 5 %)

(b) % des IC pour lesquels le paramètre se situe au-dessus de l'IC (niveau théorique : 5 %)

(c) Biais rel. = $100\times\left(E\left(\hat{V}\right)/V\left(\hat{\theta}\right)-1\right)$

(d) Stabilité = $\sqrt{V\left(\hat{V}\right)}\Big/V\left(\hat{\theta}\right)$

ANNEXE

Dans cette annexe, nous démontrons la consistance de la méthode générale de Bootstrap proposée dans l'algorithme 2 dans le cas de l'estimation de précision pour un estimateur de total. Avec une démonstration analogue à celle de Chauvet et Deville (2007) pour un tirage à un seul degré, la consistance peut ensuite être établie sous de faibles hypothèses pour l'estimateur par substitution $\theta(\hat{M})$ d'une statistique $\theta(M)$ Fréchet Différentiable à l'aide de la technique de linéarisation selon la fonction d'influence (Deville, 1999).

Nous raisonnons conditionnellement à S et à la pseudopopulation U_{GR}^{*} obtenue à l'étape 2. Soit a_i le nombre de fois où le couple (S_i, u_i^{*}) apparaît dans U_{GR}^{*}. Notons que $E(a_i|S) = 1/\pi_{Ii}$. Soit y une variable quelconque. L'estimateur bootstrappé du total est égal à

$$\hat{t}_y^{*} = \sum_{(S_i,u_i^{*})\in S_I^{*}} \frac{\hat{t}_i^{*}}{\pi_{Ii}},$$

avec

$$\hat{t}_i^{*} = \varepsilon_i \hat{t}_{y\pi}(S_i^{**}) + (1-\varepsilon_i)\hat{t}_{y\pi}(S_i),$$

en utilisant les notations de l'algorithme 1, où ε_i suit une loi de Bernoulli de paramètre π_{Ii}. Pour alléger les notations, nous écrirons simplement

$$\hat{t}_y^{*} = \sum_{u_i^{*}\in S_I^{*}} \frac{\hat{t}_i^{*}}{\pi_{Ii}}.$$

A l'aide de la formule de décomposition de la variance, on a :

$$V\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*\right) = \underbrace{V\left(E\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) \middle| S, U_{GR}^*, S_I^*\right)}_{V_1} + \underbrace{E\left(V\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) \middle| S, U_{GR}^*, S_I^*\right)}_{V_2}$$

On a $E\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) = \hat{t}_{y\pi}(S_i)$, ce qui implique que $V_1 = 0$, et

$$E\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) = \hat{t}_{y\pi}(S_i). \tag{2.4}$$

On a également

$$\begin{aligned} V\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) &= \varepsilon_i^2 V\left(\hat{t}_{y\pi}(S_i^{**}) \middle| S, U_{GR}^*, S_I^*, \varepsilon_i\right) \\ &= \varepsilon_i^2 V\left(\hat{t}_{y\pi}(S_i^{**}) \middle| S, U_{GR}^*, S_I^*\right) \end{aligned}$$

car S_i^{**} et ε_i sont indépendants. On en déduit que :

$$\begin{aligned} V_2 &= E\left(\varepsilon_i^2 V(\hat{t}_{y\pi}(S_i^{**}) \middle| S, U_{GR}^*, S_I^*) \middle| S, U_{GR}^*, S_I^*\right) \\ &= \varepsilon_i^2 V\left(\hat{t}_{y\pi}(S_i^{**}) \middle| S, U_{GR}^*, S_I^*\right) \end{aligned} \tag{2.5}$$

En utilisant à nouveau la formule de décomposition de la variance, on obtient :

$$\begin{aligned} V\left(\hat{t}_y^* \middle| S, U_{GR}^*\right) &= V\left(E\left(\hat{t}_y^* \middle| S, U_{GR}^*, S_I^*\right) \middle| S, U_{GR}^*\right) + E\left(V\left(\hat{t}_y^* \middle| S, U_{GR}^*, S_I^*\right) \middle| S, U_{GR}^*\right) \\ &= V\left(\sum_{u_i^* \in S_I^*} \frac{E\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*\right)}{\pi_{Ii}} \middle| S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*\right)}{\pi_{Ii}^2} \middle| S, U_{GR}^*\right) \\ &= V\left(\sum_{u_i^* \in S_I^*} \frac{E\left(\hat{t}_i^* \middle| S, U_{GR}^*, S_I^*\right)}{\pi_{Ii}} \middle| S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V\left(\hat{t}_i^* \middle| S, U_{GR}^*\right)}{\pi_{Ii}^2} \middle| S, U_{GR}^*\right) \end{aligned}$$

par invariance, d'où

$$\begin{aligned} V\left(\hat{t}_y^* \middle| S, U_{GR}^*\right) &= V\left(\sum_{u_i^* \in S_I^*} \frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}} \middle| S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V\left(\hat{t}_{y\pi}\left(S_i^{**}\right) \middle| S, U_{GR}^*\right)}{\pi_{Ii}} \middle| S, U_{GR}^*\right) \\ &= V\left(\sum_{u_i^* \in S_I^*} \frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}} \middle| S, U_{GR}^*\right) + \sum_{u_i^* \in U_{GR}^*} V\left(\hat{t}_{y\pi}\left(S_i^{**}\right) \middle| S, U_{GR}^*\right) \\ &= \underbrace{V\left(\sum_{u_i^* \in S_I^*} \frac{\hat{t}_{y\pi}(S_i)}{\pi_{Ii}} \middle| S, U_{GR}^*\right)}_{V_3} + \underbrace{\sum_{u_i \in S_I} a_i V\left(\hat{t}_{y\pi}\left(S_i^{**}\right) \middle| S, U_{GR}^*\right)}_{V_4} \end{aligned}$$

En utilisant l'expression asymptotique de la variance pour un tirage réjectif donnée par Hajek (1964), on obtient

$$\begin{aligned} V_3 &= (1+o_p(1))\sum_{u_i^*\in U_{GR}^*}\frac{\pi_{Ii}(1-\pi_{Ii})}{\pi_{Ii}^2}\left(\hat{t}_{y\pi}(S_i)-\tilde{t}_{yi}^*\right)^2 \\ &= (1+o_p(1))\sum_{u_i\in S_I}\frac{a_i\pi_{Ii}(1-\pi_{Ii})}{\pi_{Ii}^2}\left(\hat{t}_{y\pi}(S_i)-\tilde{t}_{yi}^*\right)^2 \end{aligned}$$

avec

$$\begin{aligned} \tilde{t}_{yi}^* &= \pi_{Ii}\frac{\sum_{u_j^*\in U_{GR}^*}\pi_{Ij}(1-\pi_{Ij})\dfrac{\hat{t}_{y\pi}(S_j)}{\pi_{Ij}}}{\sum_{u_j^*\in U_{GR}^*}\pi_{Ij}(1-\pi_{Ij})} \\ &= \pi_{Ii}\frac{\sum_{u_j^*\in S_I}a_j\pi_{Ij}(1-\pi_{Ij})\dfrac{\hat{t}_{y\pi}(S_j)}{\pi_{Ij}}}{\sum_{u_j^*\in S_I}a_j\pi_{Ij}(1-\pi_{Ij})} \end{aligned}$$

Et le théorême 1 de Deville (1999) implique que

$$E\left(V_3|S\right)\to\sum_{u_i\in S_I}\frac{(1-\pi_{Ii})}{\pi_{Ii}^2}\left(\hat{t}_{y\pi}(S_i)-\tilde{t}_{yi}\right)^2=\hat{V}_A^{HAJ}.$$

En utilisant à nouveau l'approximation de variance de Hajek (1964), on a

$$\begin{aligned} V\left(\hat{t}_{y\pi}\left(S_i^{**}\right)\middle|S,U_{GR}^*\right) &= (1+o_p(1))\sum_{k\in u_i^*}\frac{\pi_{k/i}(1-\pi_{k/i})}{\pi_{k/i}^2}\left(y_k-\tilde{y}_k^*\right)^2 \\ &= (1+o_p(1))\sum_{k\in S_i}\frac{(1-\pi_{k/i})}{\pi_{k/i}^2}\left(y_k-\tilde{y}_k\right)^2 \\ &\to \hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right) \end{aligned}$$

avec

$$\tilde{y}_k^* = \pi_{k/i} \frac{\sum_{l \in u_i^*} \pi_{l/i}(1-\pi_{l/i}) \frac{y_l}{\pi_{l/i}}}{\sum_{l \in u_i^*} \pi_{l/i}(1-\pi_{l/i})}$$

$$= \pi_{k/i} \frac{\sum_{l \in u_i^*} \pi_{l/i}(1-\pi_{l/i}) \frac{y_l}{\pi_{l/i}}}{\sum_{l \in u_i^*} \pi_{l/i}(1-\pi_{l/i})} = \tilde{y}_k .$$

Comme $E\left(\sum_{u_i \in S_I} a_i \hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right) \middle| S \right) = \sum_{u_i \in S_I} \frac{\hat{V}_{HAJ}\left(\hat{t}_{y\pi}(S_i)\right)}{\pi_{Ii}} = \hat{V}_B^{HAJ}$, on en déduit que V_4 est asymptotiquement équivalent à $\hat{V}_B^{HAJ}$, d'où le résultat.

RÉFÉRENCES

Berger, Y. (1998). Rate of convergence for asymptotic variance for the Horbitz-Thompson estimator, *Journal of Statistical Planning and Inference*, 74, 149-168.

Bertail, P., and Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique*, 46, 49-83.

Bickel, P.J., and Freedman, D.A. (1994). Asymptotic normality and the Bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Booth, J.G., and Butler, R.W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Brewer, K.R.W., and Donadio, M.E. (2003). The High Entropy Variance of the Horvitz-Thompson Estimator. *Survey Methodology*, 29, 189-196.

Caron, N. (1999). Le logiciel POULPE aspects méthodologiques. *Actes des Journées de Méthodologie Statistisque*, Insee, 84, 173-200.

Chao, M.-T., and Lo, S.-H. (1985). A Bootstrap method for finite population. *Sankhya Series A*, 47, 399-405.

Chauvet, G., and Deville, J-C. (2007). Bootstrap for unequal probability sampling, submitted.

Chen, X.-H. and Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Deville, J.-C. (1987). Réplications d'échantillons, Demi-échantillons, Jackknife, Bootstrap. *Les Sondages*, Paris : Economica.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-204.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. *Document de travail non publié*.

Funaoka, F., Saigo, H., Sitter, R.R., and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.

Goga, C., and Deville, J.-C., and Ruiz-Gazen, A. (2007). Composite estimation and linearization method for two-sample survey data. *Document de travail*.

Gross, S.T. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Matei, A., and Tillé, Y., (2005). Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size. *Journal of Official Statistics*, 21, 543-570.

McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics,* Serie 2, 95, Public Health Service Publication, 85-1369, Washington, DC : U.S. Government Printing Office .

Qualité, L. (2007). A comparison of conditional poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference,* doi: 10.1016/j.jspi.2007.04.027

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association,* 3, 173-180.

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

Rao, J.N.K., and Wu, C.F.J., and Yue, K. (1992). Quelques travaux récents sur les methods de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquêtes,* 18, 225-234.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 494-513.

Särndal, C-E., and Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling* New-York : Springer-Verlag.

Sitter, R.R. (1992a). Comparing three Bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Sitter, R.R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Tillé, Y. (2001). Théorie des sondages: échantillonnage et estimation en populations finies. Paris : Dunod.