

Bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases

Hiroshi Saigo¹

Résumé

L'échantillonnage à deux phases est un plan utile lorsque l'on ne dispose pas de variables auxiliaires a priori. L'estimation de la variance sous ce plan est toutefois compliquée, particulièrement si les fractions d'échantillonnage sont grandes. Le présent article décrit une méthode bootstrap simple pour l'échantillonnage aléatoire simple à deux phases sans remise à chaque phase avec fraction d'échantillonnage élevée. Elle est applicable à l'estimation des fonctions de répartition et des quantiles, puisqu'aucune remise à l'échelle n'est effectuée. La méthode peut être étendue à l'échantillonnage à deux phases stratifié en répétant indépendamment la procédure proposée dans diverses strates. L'estimation de la variance de certains estimateurs classiques, comme les estimateurs par le ratio et par la régression, est étudiée à titre d'exemple. Une étude par simulation est réalisée pour comparer la méthode proposée aux estimateurs de la variance existants pour l'estimation des fonctions de répartition et des quantiles.

Mots clés : Échantillonnage double; rééchantillonnage; estimation de la variance.

1. Introduction

L'échantillonnage à deux phases ou échantillonnage double est un outil puissant pour l'estimation efficace dans les sondages. Habituellement, on tire un grand échantillon de première phase où les variables auxiliaires, corrélées aux caractéristiques d'intérêt et relativement faciles à obtenir, sont observées. Puis, on sélectionne un petit sous-échantillon à partir de l'échantillon de première phase pour mesurer les caractéristiques d'intérêt qui sont plus difficiles à obtenir. À l'étape de l'estimation, les variables auxiliaires de la première phase sont utilisées pour obtenir un estimateur efficace.

Une formule explicite de la variance d'échantillon d'un estimateur peut être compliquée, voire même inexistante sous échantillonnage à deux phases. Par conséquent, les méthodes de rééchantillonnage, comme le jackknife et le bootstrap, sont séduisantes dans ces conditions. Rao et Sitter (1995) et Sitter (1997) ont étudié l'approche du jackknife avec suppression d'une unité pour les estimateurs par le ratio et par la régression sous échantillonnage à deux phases et constaté que la méthode produit des estimations de la variance convergentes par rapport au plan ayant des propriétés conditionnelles désirables sachant les variables auxiliaires.

Une faiblesse du jackknife avec suppression d'une unité est qu'il ne permet pas de traiter l'estimation des quantiles. De surcroît, l'intégration de la correction pour population finie dans l'estimation de la variance par le jackknife sous échantillonnage à deux phases n'est pas une question triviale (voir Lee et Kim 2002 et Berger et Rao 2006). Le bootstrap, par contre, élimine ces problèmes s'il est formulé convenablement.

Plusieurs méthodes bootstrap ont été proposées et étudiées pour l'échantillonnage à deux phases. Schreuder, Li et Scott (1987), Biemer et Atkinson (1993) et Sitter (1997) ont considéré des méthodes bootstrap similaires qui fournissent une estimation de la variance convergente lorsque les fractions d'échantillonnage sont négligeables. Rao et Sitter (1997) ont proposé un bootstrap avec rééchantillonnage pour les fractions d'échantillonnage élevées.

Un inconvénient de l'approche de rééchantillonnage est qu'elle ne permet pas de traiter l'estimation des fonctions de répartition ni des quantiles. Dans le présent article, nous proposons un bootstrap corrigé sur la moyenne pour l'échantillonnage à deux phases qui permet l'estimation des fonctions de répartition et des quantiles. La méthode est simple et englobe les méthodes existantes pour les fractions d'échantillonnage négligeables à titre de cas particuliers. Récemment, Kim, Navarro et Fuller (2006) ont étudié l'estimation de la variance par rééchantillonnage sans rééchantillonnage pour l'échantillonnage à deux phases dans un cadre plus généralisé que celui du présent article. Toutefois, notre méthode diffère en ce que la correction pour population finie y est intrinsèque.

La présentation de l'article est la suivante. La section 2 décrit le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La section 3 illustre le fonctionnement de la méthode proposée pour certains estimateurs classiques. La section 4 décrit l'exécution d'une simulation pour l'estimation des fonctions de répartition et des quantiles. La section 5 comprend la discussion d'autres applications du bootstrap avec moyenne ajustée. Enfin, les conclusions sont présentées à la section 6.

1. Hiroshi Saigo, Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku Tokyo 169-8050, Japon.

2. Bootstrap avec moyenne ajustée

Pour simplifier la notation, nous supposons ici qu'il n'existe qu'une seule strate. Pour étendre notre méthode à l'échantillonnage stratifié, il suffit de répéter la même procédure indépendamment dans diverses strates pour obtenir un échantillon bootstrap (voir Rao et Sitter 1997, pages 759 à 762).

Soit P l'ensemble d'étiquettes d'unité dans une population de taille N . Supposons que l'on sélectionne un échantillon aléatoire simple sans remise (EASSR) de taille n_{A+B} à partir de P et dénotons les étiquettes échantillonnées par $A+B$. La variable auxiliaire (le vecteur de variables auxiliaires) x_i est observée pour $i \in A+B$. Puis, nous tirons un EASSR de deuxième phase de taille $n_A < n_{A+B}$ à partir de $A+B$ et dénotons les étiquettes échantillonnées par A . La caractéristique (le vecteur de caractéristiques) y_i est mesurée pour $i \in A$. Soit $B = (A+B) - A$, $n_B = n_{A+B} - n_A$, $\mathbf{y}_A = \{y_i : i \in A\}$, $\mathbf{x}_A = \{x_i : i \in A\}$, et $\mathbf{x}_B = \{x_j : j \in B\}$. Nous supposons qu'un estimateur approximativement sans biais par rapport au plan du paramètre θ peut s'écrire sous la forme $\hat{\theta} = t(\mathbf{y}_A, \mathbf{x}_A, \mathbf{x}_B)$.

Sous la méthode proposée, nous construisons un échantillon bootstrap comme il suit.

1. Considérer A comme un EASSR de taille n_A tiré de P . Choisir n_A unités à partir de A par une méthode bootstrap appropriée pour un EASSR de taille n_A tiré de P . Dénoter les étiquettes échantillonnées par A^* .
2. Considérer B comme un EASSR de taille n_B tiré de $P-A$, sachant que A a été sélectionné. Choisir n_B unités à partir de B par une méthode bootstrap appropriée pour un EASSR de taille n_B tiré de $P-A$. Dénoter les étiquettes échantillonnées B^* .
3. Pour $j \in B^*$, définir l'ajustement de la moyenne comme étant \tilde{x}_j , où

$$\tilde{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_{A^*}) / (1 - f_A), \quad (1)$$

avec $\bar{x}_A = n_A^{-1} \sum_{i \in A} x_i$, $\bar{x}_{A^*} = n_A^{-1} \sum_{i \in A^*} x_i$, et $f_A = n_A / N$.

4. Soit $\mathbf{y}_{A^*} = \{y_i : i \in A^*\}$, $\mathbf{x}_{A^*} = \{x_i : i \in A^*\}$, et $\tilde{\mathbf{x}}_{B^*} = \{\tilde{x}_j : j \in B^*\}$. L'analogue bootstrap de $\hat{\theta}$ est alors donné par $\hat{\theta}^* = t(\mathbf{y}_{A^*}, \mathbf{x}_{A^*}, \tilde{\mathbf{x}}_{B^*})$.

Pour les méthodes bootstrap applicables à une population finie, voir Shao et Tu (1995, chapitre 6). Le bootstrap de Bernoulli (BBE) proposé par Funaoka, Saigo, Sitter et Toida (2006) convient pour notre méthode, pour une raison que nous mentionnerons plus loin. Pour obtenir un échantillon bootstrap A^* dans le BBE, nous procédons à un

remplacement aléatoire de chaque i dans A : garder le couple (x_i, y_i) dans l'échantillon bootstrap avec une probabilité $p = \{1 - (1 - n_A^{-1})^{-1} (1 - f_A)\}^{1/2}$ ou le remplacer par celui sélectionné aléatoirement à partir de A . Pour le cas où $p \notin [0, 1]$, voir Funaoka et coll. (2006).

Pour estimer la variance de $\hat{\theta}$, répéter les étapes 1 à 4 un grand nombre K de fois et utiliser

$$v_{\text{boot}}(\hat{\theta}) = K^{-1} \sum_{k=1}^K (\hat{\theta}_{(k)}^* - \hat{\theta}_{(\cdot)}^*)^2, \quad (2)$$

où $\hat{\theta}_{(k)}^*$ est la valeur de $\hat{\theta}^*$ dans le $k^{\text{ième}}$ échantillon bootstrap et $\hat{\theta}_{(\cdot)}^* = K^{-1} \sum_k \hat{\theta}_{(k)}^*$.

Quand f_A est négligeable, l'ajustement de la moyenne (1) est inutile. La méthode susmentionnée se réduit alors pour un grand n_A à celle de Schreuder et coll. (1987) et de Sitter (1997).

La méthode bootstrap proposée est motivée par les deux observations qui suivent. En premier lieu, posons que les plans d'échantillonnage I et II sont $[P \rightarrow A+B, A+B \rightarrow A]$ et $[P \rightarrow A, P-A \rightarrow B]$, respectivement, où \rightarrow signifie que « le deuxième membre est un EASSR provenant du premier membre ». Alors, I et II implémentent le plan de sondage identique. En fait, la probabilité d'échantillonnage attribué à un échantillon particulier $\{\mathbf{i} = (i_1, i_2, \dots, i_{n_A}) \in A, \mathbf{j} = (j_1, j_2, \dots, j_{n_B}) \in B\}$ dans I est $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_{A+B}} \times {}_{n_{A+B}} C_{n_A}]^{-1} = n_A! n_B! (N - n_{A+B})! / N!$, tandis qu'elle est $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_A} \times {}_{N-n_A} C_{n_B}]^{-1} = n_A! n_B! (N - n_{A+B})! / N!$ dans II. De toute évidence, la distribution d'échantillonnage d'un estimateur sous échantillonnage répété dépend du plan d'échantillonnage. Donc, il est commode de supposer que II est réalisé, même si I est employé.

En deuxième lieu, pour justifier l'ajustement de la moyenne (1), observons que la moyenne de x de l'ensemble $P-A$, ou l'espérance conditionnelle de \bar{x}_B sous échantillonnage répété sachant A , est $\bar{X}_{P-A} = (\bar{X} - f_A \bar{x}_A) / (1 - f_A)$. La valeur bootstrap de \bar{X}_{P-A} est donnée par $\bar{X}_{P-A^*} = (\bar{X} - f_{A^*} \bar{x}_{A^*}) / (1 - f_{A^*})$. Donc, l'équation (1) équivaut à $\tilde{x}_j = x_j - \bar{X}_{P-A} + \bar{X}_{P-A^*}$, un ajustement de la moyenne semblable à celui proposé par Rao et Shao (1992) dans le contexte de l'imputation hot deck sous mécanisme de réponse uniforme. Cet ajustement de la moyenne fait en sorte qu'existent les corrélations appropriées entre x dans A^* et x dans B^* nécessaires pour que l'estimation de la variance soit convergente lorsque les fractions d'échantillonnage sont élevées (voir Rao et Sitter 1997, page 760). Notons que la condition $n_A = n_{A^*}$ ou $f_A = f_{A^*}$ est essentielle à l'annulation de \bar{X} dans l'ajustement de la moyenne. Par conséquent, le bootstrap avec moyenne ajustée requiert une méthode bootstrap pour l'EASSR qui retient la taille originale d'échantillon, telle que le BBE.

Nous montrons à l'annexe A que la méthode bootstrap proposée produit une estimation de la variance convergente par rapport au plan pour la classe d'estimateurs étudiés par Rao et Sitter (1997). Puisqu'aucun rééchantillonnage n'est effectué, la méthode s'applique aussi à l'estimation des fonctions de répartition. Sous certaines conditions de régularité pour la fonction de répartition de population, elle produit des estimateurs de la variance convergent par rapport au plan pour les quantiles.

3. Illustrations

3.1 Estimateur par le ratio

En guise d'illustration, commençons par considérer l'estimateur par le ratio $\bar{y}_r = r_A \bar{x}_{A+B}$, où $r_A = \bar{y}_A / \bar{x}_A$, $w_A = n_A / n_{A+B}$, et $\bar{x}_{A+B} = w_A \bar{x}_A + (1 - w_A) \bar{x}_B$. Soit $\bar{y}_r^* = (\bar{y}_A / \bar{x}_A) \{w_A \bar{x}_{A^*} + (1 - w_A) \bar{x}_{B^*}\}$, l'analogie bootstrap de \bar{y}_r . En utilisant les résultats de l'annexe A avec $h(\bar{y}_A, \bar{x}_A, \bar{x}_B) = (\bar{y}_A / \bar{x}_A) \{w_A \bar{x}_A + (1 - w_A) \bar{x}_B\}$, nous pouvons approximer la variance de \bar{y}_r^* sous la méthode bootstrap proposée $V_*(\bar{y}_r^*)$ par

$$V_*(\bar{y}_r^*) \doteq (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 + 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right], \quad (3)$$

où $\hat{S}_{dA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)^2$, $\hat{S}_{dxA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)(x_i - \bar{x}_A)$, $\hat{S}_{xA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^2$, et $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B)^2$. Le deuxième membre de (3) peut être décrit comme un estimateur de la variance par « bootstrap-linéarisation ». Nous le dénotons par $v_{BL}(\bar{y}_r)$. Soulignons que $v_{BL}(\bar{y}_r)$ est presque identique à l'estimateur jackknife-linéarisation de la variance de Rao et Sitter (1995),

$$v_{JL}(\bar{y}_r) = (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 + 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \hat{S}_{xA+B}^2, \quad (4)$$

où $\hat{S}_{xA+B}^2 = (n_{A+B} - 1)^{-1} \sum_{i \in A+B} (x_i - \bar{x}_{A+B})^2$, qui concorde avec l'équation 4.8 de Demnati et Rao (2004), page 25. Puisqu'ils sont proches de $v_{JL}(\bar{y}_r)$, $V_*(\bar{y}_r)$, son approximation de Monte Carlo $v_{boot}(\bar{y}_r^*)$ et $v_{BL}(\bar{y}_r)$ devraient donner de bons résultats non seulement inconditionnellement, mais conditionnellement à $(\bar{x}_{A+B} / \bar{x}_A)$ également. Il est intéressant de souligner que la linéarisation de Taylor dans la dérivation de $v_{BL}(\bar{y}_r)$ est effectuée autour des

moyennes d'échantillon et non des moyennes de population (voir le commentaire fait par Demnati et Rao 2004, page 21).

3.2 Estimateur par la régression

Nous considérons maintenant l'estimateur par la régression. L'estimateur de la moyenne de population est $\bar{y}_{lr} = \bar{y}_A + b_A(\bar{x}_{A+B} - \bar{x}_A) = \bar{y}_A + (1 - w_A) b_A(\bar{x}_B - \bar{x}_A)$, où $b_A = \hat{S}_{xyA} / \hat{S}_{xA}^2$ avec $\hat{S}_{xyA} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)(y_i - \bar{y}_A)$. Soit $\bar{y}_{lr}^* = \bar{y}_{A^*} + (1 - w_A) b_{A^*}(\bar{x}_{B^*} - \bar{x}_{A^*})$. En utilisant les résultats de l'annexe A (voir aussi l'annexe B), nous avons

$$V_*(\bar{y}_{lr}^*) \doteq \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right] + z_A^2 \frac{(1 - f_A)}{n_A} m_{22} + 2z_A \frac{(1 - f_A)}{n_A} m_{12} + 2z_A \frac{(1 - f_{A+B})}{n_{A+B}} b_A m_{21} + 4z_A^2 \frac{(1 - f_A)}{n_A} a_A b_A \bar{x}_A \hat{S}_{xA}^2, \quad (5)$$

où $z_A = n_A(\bar{x}_{A+B} - \bar{x}_A) / \{(n_A - 1) \hat{S}_{xA}^2\}$, $m_{pq} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^p e_i^q$, $e_i = y_i - \bar{y}_A - b_A(x_i - \bar{x}_A)$, et $a_A = \bar{y}_A - b_A \bar{x}_A$. Nous appelons le deuxième membre de (5) un estimateur bootstrap-linéarisation de la variance de \bar{y}_{lr} et le dénotons par $v_{BL}(\bar{y}_{lr})$. L'estimateur jackknife-linéarisation de la variance de \bar{y}_{lr} (Sitter 1997, page 781) est

$$v_{JL}(\bar{y}_{lr}) = \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \hat{S}_{xA+B}^2 + \frac{z_A^2}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A)^2 e_i^2}{(1 - c_i)^2} + \frac{2z_A}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A) e_i^2}{(1 - c_i)} + \frac{2z_A b_A}{n_A(n_{A+B} - 1)} \sum_{i \in A} \frac{(x_i - \bar{x}_A)(x_i - \bar{x}_{A+B}) e_i}{(1 - c_i)}, \quad (6)$$

où $c_i = n_A^{-1} + (x_i - \bar{x}_A)^2 / \{(n_A - 1) \hat{S}_{xA}^2\}$, les valeurs d'effet de levier. Partant de (5) et (6), $v_{boot}(\bar{y}_{lr})$, $v_{BL}(\bar{y}_{lr})$ et $v_{JL}(\bar{y}_{lr})$ donnent des résultats similaires à condition que $f_{A+B} \doteq 0$, que n_A soit suffisamment grand pour que tous les c_i soient presque nuls et que le dernier terme du deuxième membre de (5) soit négligeable.

3.3 Estimation des fonctions de répartition

À titre d'exemple, prenons l'estimateur du maximum de vraisemblance pseudo-empirique calé sur un modèle (ME)

sous échantillonnage à deux phases proposé par Wu et Luan (2003) et défini par

$$\hat{F}_{ME}(t) = \sum_{i \in A} \hat{p}_i I(y_i \leq t), \quad (7)$$

où \hat{p}_i maximise la fonction de pseudo-vraisemblance $\hat{l}(p) = \sum_A (N/n_A) \log p_i$ sous les contraintes a) $\sum_A p_i = 1$ ($0 < p_i < 1$); et b) $\sum_A p_i g_i = n_{A+B}^{-1} \sum_{A+B} g_i$ où $g_i = g(x_i, t) = P(y \leq t | x_i)$ sous un certain modèle de travail. Par exemple, nous pouvons supposer que $\log(g_i/(1-g_i)) = x_i' \theta$ avec une fonction de variance $V(g) = g(1-g)$. Chen, Sitter et Wu (2002) ont montré un algorithme simple pour le calcul de \hat{p}_i . Il peut être démontré (voir Wu et Luan 2003) que, sous l'échantillonnage à deux phases considéré dans le présent article,

$$\hat{F}_{ME}(t) = n_A^{-1} \sum_{i \in A} I(y_i \leq t) + \left\{ n_{A+B}^{-1} \sum_{i \in A+B} g_i - n_A^{-1} \sum_{i \in A} g_i \right\} \beta + o_p(n_A^{-1/2}),$$

où $\beta = \sum_P (g_i - \bar{g}) I(y \leq t) / \sum_P (g_i - \bar{g})^2$ avec $\bar{g} = N^{-1} \sum_P g_i$. Notons que cette équation n'est pas utilisée dans l'estimation, mais elle montre que la variance de $\hat{F}_{ME}(t)$ peut être estimée par le bootstrap avec moyenne ajustée, puisque $\hat{F}_{ME}(t)$ est approximé par un estimateur de type régression.

3.4 Estimation des quantiles

L'estimation des quantiles peut être obtenue directement en inversant $\hat{F}(t)$ par $\hat{F}^{-1}(\alpha) = \inf \{t : \hat{F}(t) \geq \alpha\}$ pour un certain $\alpha \in (0, 1)$. Par exemple, si on utilise (7), alors une estimation des quantiles est donnée par $y_{(k)}$, où $y_{(k)}$ est la statistique de $k^{\text{ième}}$ ordre de y telle que $\sum_{i=1}^{k-1} \hat{p}_{(i)} < \alpha$ et $\sum_{i=1}^k \hat{p}_{(i)} \geq \alpha$ (Chen et Wu 2002). Sous certaines conditions spécifiées dans Chen et Wu (2002), une représentation de type Bahadur de $\hat{F}_{ME}^{-1}(\alpha)$ peut être établie. Donc, l'estimateur de la variance par le bootstrap avec moyenne ajustée pour $\hat{F}_{ME}^{-1}(\alpha)$ est convergent par rapport au plan. Notons qu'il n'existe aucune forme explicite de l'estimateur de la variance pour $\hat{F}_{ME}^{-1}(\alpha)$, mais qu'on peut appliquer un estimateur de la variance convergent basé sur l'estimation d'intervalle de Woodruff (Woodruff 1952).

4. Simulation

4.1 Population et échantillonnage

Nous avons réalisé une étude par simulation afin d'examiner l'estimation de la variance par le bootstrap avec moyenne ajustée pour les estimateurs de la section 3. Nous présentons ici les résultats pour l'estimation des fonctions de répartition et des quantiles. Les résultats pour les estimateurs

par le ratio et par la régression peuvent être obtenus auprès de l'auteur sur demande.

Pour commencer, nous avons généré la variable auxiliaire x pour une population finie P de taille $N = 2000$ en utilisant une loi Gamma(1, 1). La variable dépendante y a ensuite été générée au moyen de $y_i = x_i + \sqrt{x_i} v_i$, où $v_i \sim N(0, 0.5^2)$. Un EASSR $A+B$ de taille $n_{A+B} = 800$ a été sélectionné à partir de la population, puis un EASSR A de taille $n_A = 200$ a été sélectionné à partir de $A+B$. La population est demeurée fixe au cours des exécutions de la simulation, puisque nous nous concentrons sur les propriétés de l'échantillonnage répété par rapport au plan.

4.2 Estimation des fonctions de répartition

Pour l'estimation des fonctions de répartition, nous avons pris $\hat{F}_{ME}(t)$ comme exemple. D'autres estimateurs, comme ceux de Chambers et Dunstan (1986) et de Rao, Kovar et Mantel (1990) peuvent être traités de la même façon quand un estimateur approximativement sans biais par rapport au plan. Nous avons supposé que le modèle de travail pour g dans $\hat{F}_{ME}(t)$ était le logit avec variance binomiale. L'estimateur bootstrap de la variance $v_{boot}(\hat{F}_{ME}(t))$ a été calculé avec $K = 200$. Nous avons utilisé le BBE pour construire un échantillon bootstrap. Le nombre total de simulations était $M = 5000$, tandis que l'EQM réelle de $\hat{F}_{ME}(t)$ à un temps t donné a été estimée sur 50 000 exécutions.

Nous avons comparé $v_{boot}(\hat{F}_{ME}(t))$ à trois estimateurs de la variance : l'estimateur analytique de Wu et Luan (2003), le jackknife avec suppression d'une unité standard et le jackknife avec suppression d'une unité et une correction pour population finie *ad hoc*. L'estimateur de Wu et Luan (2003) est

$$v_a(\hat{F}_{ME}(t)) = (n_{A+B}^{-1} - N^{-1}) \hat{S}_I^2 + (n_A^{-1} - n_{A+B}^{-1}) \hat{S}_D^2,$$

où les deux composantes \hat{S}^2 sont estimées respectivement par

$$\hat{S}^2 = s^2 + \left[\frac{1}{n_{A+B}(n_{A+B}-1)} \sum_{j>i:i, j \in A+B} u_{ij} - \frac{1}{n_A(n_A-1)} \sum_{j>i:i, j \in A} u_{ij} \right] \hat{\beta}_F,$$

où $s^2 = \{n_A(n_A-1)\}^{-1} \sum_{i<j:i, j \in A} v_{ij}$, et $\hat{\beta}_F = \sum_{i<j:i, j \in A} u_{ij} v_{ij} / \sum_{i<j:i, j \in A} u_{ij}^2$ avec u_{ij} et v_{ij} spécifiés comme suit : Pour \hat{S}_I^2 , $v_{ij} = (I_i - I_j)^2$ et $u_{ij} = (\hat{g}_i - \hat{g}_j)^2$ avec $I_i = I(y_i \leq t)$ et $\hat{g}_i = \hat{g}(x_i, t)$ estimé en A ; pour \hat{S}_D^2 , $v_{ij} = (\hat{D}_i - \hat{D}_j)^2$ et $u_{ij} = \hat{g}_i(1 - \hat{g}_i) + \hat{g}_j(1 - \hat{g}_j)$ avec $\hat{D}_i = I_i - \hat{g}_i$, $\hat{\beta} = \sum_{i \in A} I_i (\hat{g}_i - \bar{g}_A) / \sum_{i \in A} (\hat{g}_i - \bar{g}_A)^2$ et $\bar{g}_A = n_A^{-1} \sum_{i \in A} \hat{g}_i$.

La formule du jackknife avec suppression d'une unité standard est donnée par

$$v_J(\hat{\theta}) = \frac{(n_{A+B} - 1)}{n_{A+B}} \sum_{j \in A+B} (\hat{\theta}_{(-j)} - \hat{\theta}_{(\cdot)})^2,$$

où $\hat{\theta} = \hat{F}_{ME}(t)$, $\hat{\theta}_{(-j)}$ est la j° pseudo-estimation par le jackknife et $\hat{\theta}_{(\cdot)} = n_{A+B}^{-1} \sum_{j \in A+B} \hat{\theta}_{(-j)}$. Notons que, pour $j \in A$, y_j et x_j sont toutes deux éliminées de l'échantillon, tandis que pour $j \in B$, seul x_j est éliminée (voir Rao et Sitter 1995 et Sitter 1997). La formule avec correction pour population finie *ad hoc* est $v_{Jfpc}(\hat{F}_{ME}(t)) = (1 - f_{A+B})v_J(\hat{F}_{ME}(t))$.

Le tableau 1 présente le biais relatif (%biais) et le coefficient de variation (CV) des quatre estimateurs de la variance pour $\hat{F}_{ME}(t_\alpha)$ ($\alpha = 0,10, 0,25, 0,50, 0,75, 0,90$), où $F(t_\alpha) = \alpha$. Ici, %Biais et CV ont été calculés sous la forme %Biais = $100 \times (M^{-1} \sum_{m=1}^M v^{(m)} - EQM) / EQM$ et $CV = [M^{-1} \sum_{m=1}^M (v^{(m)} - EQM)^2]^{1/2} / EQM$, respectivement, où $v^{(m)}$ est une estimation de la variance dans la m° exécution de la simulation. Le tableau 1 démontre que $v_J(\hat{F}_{ME}(t))$ présente un biais par excès, puisque les fractions d'échantillonnage ne sont pas négligeables, que $v_{Jfpc}(\hat{F}_{ME}(t))$ présente un biais par défaut puisque le facteur d'ajustement *ad hoc* $(1 - f_{A+B})$ est trop faible, et que $v_a(\hat{F}_{ME}(t))$ et $v_{boot}(\hat{F}_{ME}(t))$ sont tous deux approximativement sans biais, quoique le dernier soit un peu plus instable, ce qui est typique d'une méthode de rééchantillonnage.

Tableau 1 Estimation de la variance pour l'EMV pseudo-empirique $\hat{F}_{ME}(t_\alpha)$

Estimateur		α				
		0,10	0,25	0,50	0,75	0,90
$v_{boot}(\hat{F}_{ME}(t_\alpha))$	%Biais	0,27	-0,22	0,64	0,83	2,73
	CV	0,19	0,14	0,14	0,15	0,24
$v_a(\hat{F}_{ME}(t_\alpha))$	%Biais	-2,29	-2,03	-0,47	-1,95	-3,26
	CV	0,17	0,11	0,09	0,11	0,19
$v_J(\hat{F}_{ME}(t_\alpha))$	%Biais	14,24	17,29	22,98	23,80	24,97
	CV	0,24	0,21	0,25	0,27	0,36
$v_{Jfpc}(\hat{F}_{ME}(t_\alpha))$	%Biais	-31,45	-29,63	-26,21	-25,72	-25,02
	CV	0,33	0,30	0,27	0,27	0,30

En nous inspirant de Royall et Cumberland (1981a, 1981b), nous avons ordonné les $M = 5\,000$ échantillons simulés sur les valeurs de $\bar{x}_{A+B} - \bar{x}_A$, nous les avons classés en vingt groupes consécutifs de $G = 250$ dans chacun desquels l'EQM conditionnelle (EQM_c) simulée et la moyenne conditionnelle de $v(E_c(v))$ ont été calculées. La

figure 1 montre la représentation graphique d'EQM_c et d' $E_c(v)$ en fonction des moyennes de groupe de $\bar{x}_{A+B} - \bar{x}_A$ pour $t_{0,10}$ et $t_{0,90}$. On constate que $v_a(\hat{F}_{ME}(t))$ et $v_{boot}(\hat{F}_{ME}(t))$ ont tous deux le même comportement conditionnellement à $\bar{x}_{A+B} - \bar{x}_A$. Les estimateurs jackknife de la variance, $v_J(\hat{F}_{ME}(t))$ et $v_{Jfpc}(\hat{F}_{ME}(t))$, quoique biaisés, suivent une tendance de l'EQM_c.

4.3 Estimation des quantiles

Par inversion directe de $\hat{F}_{ME}(t)$, nous estimons le quantile α . Pour obtenir \hat{p}_i pour $\hat{F}_{ME}(t)$, nous avons fixé t à la valeur \hat{t}_α , où $\hat{t}_\alpha = \inf \{t: n_A^{-1} \sum_A I(y_i \leq t) \geq \alpha\}$, un estimateur utilisant uniquement $\{y_i: i \in A\}$. Pour l'estimation de la variance, nous avons créé $K = 1\,000$ échantillons bootstrap. En vue de comparaison, nous avons également calculé l'estimateur de la variance de Woodruff (Woodruff 1952 et Shao et Tu 1995, page 238),

$$v_W(\hat{F}_{ME}^{-1}(\alpha)) = \left[\frac{\hat{F}_{ME}^{-1}(\alpha + \zeta_{1-\kappa/2} \hat{\sigma}_{\hat{F}}) - \hat{F}_{ME}^{-1}(\alpha - \zeta_{1-\kappa/2} \hat{\sigma}_{\hat{F}})}{2\zeta_{1-\kappa/2}} \right]^2,$$

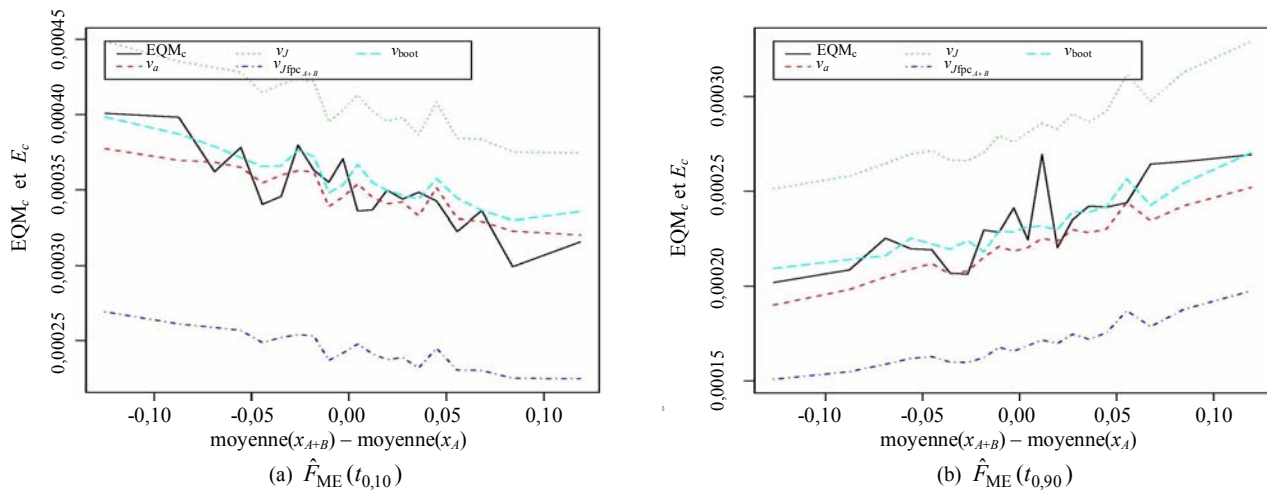
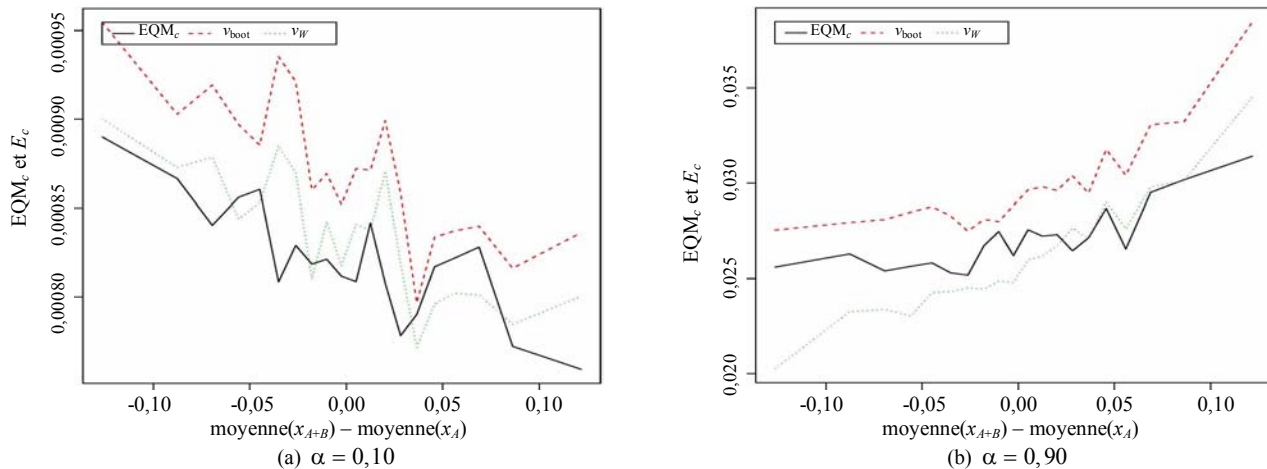
où $\hat{\sigma}_{\hat{F}}^2 = v(\hat{F}_{ME}(t))$ avec $t = \hat{F}_{ME}^{-1}(\alpha)$ et $\zeta_{1-\kappa/2}$ est le $(1 - \kappa/2)$ quantile de $N(0, 1)$. Soit $\kappa = 0,05$, quoique le meilleur choix de κ soit inconnu. Les mesures de performance, %Biais et CV, ont été calculées sur $M = 5\,000$ exécutions, tandis que l'EQM réelle a été estimée sur 50 000 exécutions de la simulation.

Le tableau 2 résume les résultats pour l'estimation des quantiles. Il démontre que le bootstrap avec moyenne ajustée produit un biais par excès dans l'estimation de $V(\hat{F}_{ME}^{-1}(\alpha))$, mais un biais négligeable dans l'estimateur de la variance de Woodruff.

Tableau 2 Estimation de la variance pour les quantiles

Estimateur		α				
		0,10	0,25	0,50	0,75	0,90
$v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$	%Biais	6,27	14,32	10,05	10,02	10,28
	CV	0,53	0,53	0,51	0,52	0,61
$v_W(\hat{F}_{ME}^{-1}(\alpha))$	%Biais	1,64	3,75	2,92	0,70	-3,67
	CV	0,50	0,45	0,45	0,46	0,52

La figure 2 montre les propriétés conditionnelles de $v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$ et de $v_W(\hat{F}_{ME}^{-1}(\alpha))$ pour $\alpha = 0,10, 0,90$. Nous voyons que $v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$ et $v_W(\hat{F}_{ME}^{-1}(\alpha))$ suivent tous deux l'EQM_c de la même façon, quoique le premier possède uniformément un biais par excès.

Figure 1 EQM_c et $E_c(v)$ pour $\hat{F}_{ME}(t_\alpha)$ Figure 2 EQM_c et $E_c(v)$ pour l'estimation des quantiles

5. Remarques supplémentaires

5.1 Échantillonnage à deux phases stratifié

Supposons qu'une population doit être stratifiée en H strates, mais qu'on ne dispose d'aucune information pour la stratification. Une solution possible dans cette situation est de commencer par obtenir un EASSR de taille n' à partir de la population, d'observer les variables auxiliaires, y compris celles pour la stratification, de stratifier l'échantillon en H strates et, dans chaque strate, de tirer un EASSR de taille n_h à partir de n'_h unités appartenant à la strate h dans l'échantillon. Voir, par exemple, Cochran (1977, section 12.2) pour les détails.

Soit N_h la taille de la strate h dans la population. Sous la condition $n'_h > 0$, l'échantillonnage de première phase dans la strate h décrit plus haut est équivalent à l'échantillonnage aléatoire simple sans remise de taille n'_h dans la strate h réalisé de façon indépendante dans chacune des strates.

Donc, sachant n'_h ($h = 1, \dots, H$), le bootstrap avec moyenne ajustée peut être appliqué indépendamment dans diverses strates pour obtenir un échantillon bootstrap. Quand N_h est inconnu, comme cela est habituellement le cas pour l'échantillonnage à deux phases stratifié, on peut utiliser un estimateur sans biais $\hat{N}_h = N(n'_h/n')$ dans le bootstrap avec moyenne ajustée. Dans ce cas, la fraction d'échantillonnage n'/N est habituellement utilisée dans toutes les strates.

Notons toutefois que la présente discussion est légitime pour l'estimation sachant les tailles d'échantillon de première phase. La variance due à la variable n'_h peut être grande. Pour l'estimation non conditionnelle de la variance, voir Kim et coll. (2006).

5.2 Non-réponse

Le commentaire qui précède s'applique aux données d'enquête imputées sous le mécanisme de réponse univoque. Supposons qu'une population est stratifiée en S_h ($h = 1, \dots, H$) où l'échantillonnage aléatoire simple est

réalisé indépendamment. Un échantillon est divisé en classes d'imputation C_l ($l=1, \dots, L$) dans chacune desquelles on suppose que le taux de réponse est uniforme et on procède à l'imputation. Une classe d'imputation peut recouper les strates. Nous supposons aussi que la classe d'imputation à laquelle appartient une unité échantillonnée est identifiée correctement avant l'imputation. Dénotons les nombres d'unités échantillonnées et de répondants dans $S_h \cap C_l$ comme étant n_{hl} et r_{hl} , respectivement. Alors, on voit que, sachant n_{hl} et r_{hl} , le plan correspondant dans $S_h \cap C_l$ est le même que celui discuté dans le présent article si nous considérons les n_{hl} unités et les r_{hl} répondants comme étant $A+B$ et A , respectivement. Par conséquent, le bootstrap avec moyenne ajustée peut être exécuté indépendamment dans différents $S_h \cap C_l$ ($h=1, \dots, H; l=1, \dots, L$). La taille de $S_h \cap C_l$, dénotée par N_{hl} , peut être estimée par $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Notons qu'il s'agit d'une méthode bootstrap conditionnée sur le nombre de répondants.

6. Conclusion

Dans le présent article, nous avons proposé le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La méthode requiert un simple ajustement de la moyenne et permet de traiter l'estimation des fonctions de répartition et de quantiles, car elle ne nécessite pas de rééchantillonnage. Le développement en série de Taylor montre que la méthode a de bonnes propriétés conditionnelles pour les estimateurs par le ratio et par la régression. Une étude par simulation démontre qu'elle a aussi des propriétés conditionnelles similaires lors de l'estimation des fonctions de répartition et des quantiles. Une extension à l'échantillonnage à deux phases stratifiées est simple. Conditionnellement aux tailles d'échantillon de première phase, la méthode permet de traiter l'échantillonnage à deux phases stratifié et l'imputation sous le mécanisme de réponse uniforme. Nous étudions à l'heure actuelle une extension de la méthode proposée à des plans d'échantillonnage multiphasés plus généralisés.

Remerciements

Cette étude a été financée par une bourse de la Société japonaise de promotion de la science. L'auteur remercie le professeur Randy R. Sitter, le rédacteur en chef, le rédacteur adjoint et deux examinateurs de leurs commentaires et suggestions utiles.

Annexe A

Dans la présente annexe, nous montrons que la méthode bootstrap proposée fournit des estimations de la variance convergentes pour une classe d'estimateurs considérés par Rao et Sitter (1997). Nous utilisons les mêmes conditions que dans Rao et Sitter (1997) avec une notation légèrement différente. Pour simplifier, nous supposons qu'il n'existe qu'une seule strate, mais une extension à l'échantillonnage à deux phases stratifié est simple.

Considérons une classe d'estimateurs, $\theta = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$, d'un paramètre de population $\theta = h(\bar{Y}, \bar{X}, \bar{X})$, où \bar{Y} et \bar{X} sont les moyennes de population des vecteurs y et x , c'est-à-dire $\bar{Y} = N^{-1} \sum_{i \in P} y_i$ et $\bar{X} = N^{-1} \sum_{i \in P} x_i$. Ici, x est observé dans l'échantillon de première phase $A+B$, tandis que y est mesuré uniquement dans l'échantillon de deuxième phase A . Les moyennes d'échantillon (\bar{y}_A, \bar{x}_A) et \bar{x}_B sont calculées dans A et B , respectivement, c'est-à-dire $\bar{y}_A = n_A^{-1} \sum_{i \in A} y_i$, $\bar{x}_A = n_A^{-1} \sum_{i \in A} x_i$, et $\bar{x}_B = n_B^{-1} \sum_{i \in B} x_i$.

Par un développement en série de Taylor, nous obtenons

$$\hat{\theta} = \theta + \nabla h'(\Delta \bar{y}_A, \Delta \bar{x}_A, \Delta \bar{x}_B)' + o_p(n_A^{-1/2}),$$

où ∇h est le vecteur de gradients de h évalué à $(\bar{Y}, \bar{X}, \bar{X})$, $\Delta \bar{y}_A = \bar{y}_A - \bar{Y}$, $\Delta \bar{x}_A = \bar{x}_A - \bar{X}$, $\Delta \bar{x}_B = \bar{x}_B - \bar{X}$, et $'$ dénote une matrice transposée (voir l'équation 33.7 de Rao et Sitter 1997, page 757 et les conditions requises). Alors, la variance de $\hat{\theta} = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$ est approximée par

$$V(\hat{\theta}) \doteq \nabla h' \sum_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'} \nabla h,$$

où $\sum_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}$ est la matrice de variance-covariance de $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ sous échantillonnage à deux phases répété. Comme A et B sont des EASSR de taille n_A et n_B tirés de la population P , respectivement, nous voyons que $\sum_{(\bar{y}_A, \bar{x}_A)'}$ = $(1-f_A)S_{(y', x')^2}^2/n_A$ et $\sum_{\bar{x}_B}$ = $(1-f_B)S_x^2/n_B$, où $S_u^2 = (N-1)^{-1} \sum_{i \in P} (u_i - \bar{U})(u_i - \bar{U})'$ est la variance de population de $u = (y', x)'$ ou x et $f_B = n_B/N$. Pour $\text{Cov}(\bar{y}_A, \bar{x}_B)$, soit E_A et $E_{B|A}$ les espérances pour la sélection d'un EASSR A à partir de P et pour le choix d'un EASSR B à partir de $P-A$ sachant A , respectivement. Notons que $E_{B|A}(x_B) = (\bar{X} - f_A \bar{x}_A)/(1-f_A)$. Donc, nous avons

$$\begin{aligned} \text{Cov}(\bar{y}_A, \bar{x}_B) &= E(\bar{y}_A \bar{x}_B') - E(\bar{y}_A) E(\bar{x}_B') \\ &= E_A(\bar{y}_A E_{B|A}(\bar{x}_B)) - \bar{Y} \bar{X}' \\ &= -S_{yx}/N, \end{aligned}$$

où $S_{yx} = (N-1)^{-1} \sum_{i \in P} (y - \bar{Y})(x - \bar{X})'$. De même, $\text{Cov}(\bar{x}_A, \bar{x}_B) = -S_x^2/N$.

Maintenant, considérons un développement en série de Taylor de $\hat{\theta}^* = h(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{x}_{B^*})$ avec $\bar{x}_{B^*} = \bar{x}_{B^*} + f_A(\bar{x}_A - \bar{x}_{A^*})/(1 - f_A)$, l'analogue bootstrap de $\hat{\theta} = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$. Soit E_* et V_* l'espérance et la variance sous la procédure bootstrap proposée, respectivement. Pour commencer, observons que $E_*(\bar{y}_{A^*}) = \bar{y}_A$, $E_*(\bar{x}_{A^*}) = \bar{x}_A$ et

$$\begin{aligned} E_*(\bar{x}_{B^*}) &= E_{*A^*}(E_{*B^*|A^*}(\bar{x}_{B^*})) \\ &= E_{*A^*}(\bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*})/(1 - f_A)) \\ &= \bar{x}_B, \end{aligned}$$

où E_{*A^*} et $E_{*B^*|A^*}$ sont, respectivement, l'espérance par rapport à l'échantillonnage A^* et l'espérance conditionnelle par rapport à l'échantillonnage B^* sachant A^* sous la méthode bootstrap proposée. Alors, $\hat{\theta}^* = h(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{x}_{B^*})$ est approximé par

$$\hat{\theta}^* = \hat{\theta} + \nabla h^*(\Delta \bar{y}'_{A^*}, \Delta \bar{x}'_{A^*}, \Delta \bar{x}'_{B^*})' + o_p(n_A^{-1/2}),$$

où ∇h^* est le gradient de h évalué à $(\bar{y}_A, \bar{x}_A, \bar{x}_B)$, $\Delta \bar{y}'_{A^*} = \bar{y}_{A^*} - \bar{y}_A$, $\Delta \bar{x}'_{A^*} = \bar{x}_{A^*} - \bar{x}_A$, et $\Delta \bar{x}'_{B^*} = \bar{x}_{B^*} - \bar{x}_B$ (voir l'équation 33.A.1 de Rao et Sitter 1997, page 767 et les conditions requises connexes). Par conséquent, $V_*(\hat{\theta}^*)$ est approximé par

$$V_*(\hat{\theta}^*) \doteq \nabla h^* \sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{x}'_{B^*})'} \nabla h^*,$$

où $\sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{x}'_{B^*})'}$ est la matrice de variance-covariance de $(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{x}_{B^*})'$ sous l'échantillonnage bootstrap proposé.

L'estimation convergente de la variance sous la méthode proposée est prouvée en montrant que ∇h^* et $\sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{x}'_{B^*})'}$ sont convergents pour ∇h et $\sum_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}$, respectivement. La convergence de ∇h^* pour ∇h découle de la convergence de $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ pour $(\bar{Y}, \bar{X}, \bar{X})$ et de la continuité de h .

La convergence de $\sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{x}'_{B^*})'}$ peut être montrée comme il suit. Pour commencer, puisque nous utilisons une méthode bootstrap appropriée pour l'échantillonnage aléatoire simple sans remise dans le sous-échantillonnage A^* , nous avons $\sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*})'} = (1 - f_A) \hat{S}_{(y', x')A}^2 / n_A$, où $\hat{S}_{uA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (\mathbf{u}_i - \bar{\mathbf{u}}_A) (\mathbf{u}_i - \bar{\mathbf{u}}_A)'$ avec $\mathbf{u} = (y', x')'$. Deuxièmement, parce que

1. $\sum_{\bar{x}_{B^*}} = E_{*A^*}(V_{*B^*|A^*}(\bar{x}_{B^*})) + V_{*A^*}(E_{*B^*|A^*}(\bar{x}_{B^*}))$, où V_{*A^*} et $V_{*B^*|A^*}$ sont, respectivement, la variance par rapport à l'échantillonnage A^* et la variance conditionnelle par rapport à l'échantillonnage B^* sachant A^* ,
2. $V_{*B^*|A^*}(\bar{x}_{B^*}) = (1 - f_{B|A}) \hat{S}_{xB}^2 / n_B$, où $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B) (x_i - \bar{x}_B)'$ et $f_{B|A} = n_B / (N - n_A)$, et
3. $E_{*B^*|A^*}(\bar{x}_{B^*}) = \bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*}) / (1 - f_A)$, nous avons $\sum_{\bar{x}_{B^*}} = (1 - f_{B|A}) \hat{S}_{xB}^2 / n_B + f_A \hat{S}_{xA}^2 / (N - n_A)$. Puisque \hat{S}_{xA}^2 et \hat{S}_{xB}^2 sont tous deux convergents pour S_x^2 , $\sum_{\bar{x}_{B^*}}$ est convergent pour $\sum_{\bar{x}_B} = (1 - f_B) S_x^2 / n_B$. Enfin, nous calculons $\text{Cov}_*(\bar{y}_{A^*}, \bar{x}_{B^*})$ et $\text{Cov}_*(\bar{x}_{A^*}, \bar{x}_{B^*})$. Pour la première, nous avons

$$\begin{aligned} \text{Cov}_*(\bar{y}_{A^*}, \bar{x}_{B^*}) &= E_*(\bar{y}_{A^*} \bar{x}_{B^*}') - E_*(\bar{y}_{A^*}) E_*(\bar{x}_{B^*}') \\ &= E_{*A^*}(\bar{y}_{A^*} E_{*B^*|A^*}(\bar{x}_{B^*}')) - \bar{y}_A \bar{x}_B' \\ &= E_{*A^*}(\bar{y}_{A^*} \{ \bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*}) / (1 - f_A) \}) - \bar{y}_A \bar{x}_B' \\ &= -\hat{S}_{yxA} / N, \end{aligned}$$

où $\hat{S}_{yxA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - \bar{y}_A) (x_i - \bar{x}_A)'$. De même, $\text{Cov}_*(\bar{x}_{A^*}, \bar{x}_{B^*}) = -\hat{S}_{xA}^2 / N$. Ceci complète la preuve de la convergence de $\sum_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{x}'_{B^*})'}$ pour $\sum_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}$.

Annexe B

Dans cette annexe nous dérivons $v_{BL}(\bar{y}_{lr})$. Sous le bootstrap avec moyenne ajustée,

$$\begin{aligned} \bar{y}_{lr}^* &= \bar{y}_A \\ &+ (1 - w_A) b_A \left\{ -\frac{(\bar{x}_{A^*} - \bar{x}_A)}{(1 - f_A)} + (\bar{x}_{B^*} - \bar{x}_B) + (\bar{x}_B - \bar{x}_A) \right\}. \end{aligned}$$

Définissons

$$\begin{aligned} \hat{\xi}_{pq}^* &= n_A^{-1} \sum_{i \in A^*} x_i^p y_i^q, \\ \hat{\xi}^* &= [\hat{\xi}_{10}^*, \hat{\xi}_{01}^*, \hat{\xi}_{11}^*, \hat{\xi}_{20}^*, \bar{x}_B]' \end{aligned}$$

et

$$\xi = [\bar{x}_A, \bar{y}_A, n_A^{-1} \sum_{i \in A} x_i y_i, n_A^{-1} \sum_{i \in A} x_i^2, \bar{x}_B]' = E_*(\hat{\xi}^*).$$

Notons que $b_A = (\hat{\xi}_{11}^* - \hat{\xi}_{10}^* \hat{\xi}_{01}^*) / (\hat{\xi}_{20}^* - \hat{\xi}_{10}^{*2})$. Soit $\bar{y}_{lr}^* = h(\hat{\xi}^*)$. Cette expression est légèrement différente de celle de l'annexe A, mais nous pouvons exploiter le sous-échantillonnage indépendant de A^* et B^* . Alors, par linéarisation de Taylor de $\bar{y}_{lr}^* = h(\hat{\xi}^*)$ autour de ξ , nous obtenons $\bar{y}_{lr}^* \doteq \bar{y}_{lr} + \nabla h^*(\hat{\xi}^* - \xi)$ et $V_*(\bar{y}_{lr}^*) \doteq \nabla h^* \sum_{\hat{\xi}^*} \nabla h^*$, où

$$\begin{aligned} \nabla h^* &= [-b_A(1 - w_A) / (1 - f_A) - z_A(\bar{y}_A - 2b_A \bar{x}_A), 1 - z_A \bar{x}_A, \\ &z_A, -z_A b_A, b_A(1 - w_A)]' \end{aligned}$$

et $\sum_{\hat{\xi}^*} = [v_{ij}]$ avec

$$\begin{aligned} v_{11} &= c_A \hat{S}_{xA}^2, \\ v_{21} &= c_A \hat{S}_{xyA}, \\ v_{22} &= c_A \hat{S}_{yA}^2, \\ v_{31} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(x_i - \bar{x}_A), \\ v_{32} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(y_i - \bar{y}_A), \\ v_{33} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})^2, \\ v_{41} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i - \bar{x}_A), \\ v_{42} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(y_i - \bar{y}_A), \end{aligned}$$

$$v_{43} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i y_i - \xi_{11}),$$

$$v_{44} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})^2,$$

$$v_{51} = v_{52} = v_{53} = v_{54} = 0,$$

$$v_{55} = \{n_B^{-1} - (N - n_A)^{-1}\} \hat{S}_{xB}^2,$$

$v_{ij} = v_{ji}$, et $c_A = (1 - f_A)/n_A$. En réécrivant les moments par rapport à l'origine sous forme de moments centrés, en notant que $y_i - \bar{y}_A = b_A(x_i - \bar{x}_A) + e_i$ et en utilisant les propriétés de e_i en tant que résidus des moindres carrés, nous obtenons le deuxième membre de (5) après certains calculs algébriques.

Bibliographie

- Berger, Y.G., et Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 68, 531-547.
- Biemer, P.P., et Atkinson, D. (1993). Estimation de l'erreur systématique de mesure par la prédiction modéliste. *Techniques d'enquête*, 19, 137-146.
- Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.
- Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.
- Cochran, W.G. (1977). *Sampling Techniques*. 3^{ième} Édition. New York : John Wiley & Sons, Inc.
- Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.
- Funaoka, F., Saigo, H., Sitter, R.R. et Toida, T. (2006). Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*. 32, 169-175.
- Kim, J.-K., Navarro, A. et Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Lee, H., et Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.
- Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.
- Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rao, J.N.K., et Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. Dans *Survey Measurement and Process Quality: Wiley Series in Probability and Statistics*. (Éds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York. 753-768.
- Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Schreuder, H.T., Li, H.G. et Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.
- Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag : New York.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Woodruff, R.S. (1952). Confidence intervals for median and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., et Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.