Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 169-175
Statistique Canada, N° 12-001-XPB au catalogue

Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés

Fumio Funaoka, Hiroshi Saigo, Randy R. Sitter et Tsutom Toida [1]

Résumé

Nous proposons dans cet article une méthode de bootstrap de type Bernoulli facilement applicable à des plans stratifiés à plusieurs degrés où les fractions de sondage sont grandes, à condition qu'un échantillonnage aléatoire simple sans remise soit utilisé à chaque degré. La méthode fournit un ensemble de poids de rééchantillonnage qui donnent des estimations convergentes de la variance pour les estimateurs lisses ainsi que non lisses. La force de la méthode tient à sa simplicité. Elle peut être étendue facilement à n'importe quel nombre de degrés d'échantillonnage sans trop de complications. L'idée principale est de garder ou de remplacer une unité d'échantillonnage à chaque degré d'échantillonnage en utilisant des probabilités prédéterminées pour construire l'échantillon bootstrap. Nous présentons une étude par simulation limitée afin d'évaluer les propriétés de la méthode et, à titre d'illustration, nous appliquons cette dernière à l'Enquête nationale sur les prix menée en 1997 au Japon.

Mots clés : Enquête complexe; linéarisation; quantiles; rééchantillonnage; stratification.

1. Introduction

De nombreuses enquêtes à grande échelle sont réalisées selon un plan d'échantillonnage stratifié à plusieurs degrés. Or, lorsqu'on utilise ce genre de plan, l'estimation de la variance peut être analytiquement complexe, voire même impossible. En outre, pour les ensembles de données à grande diffusion, les formes particulières des estimateurs dont l'utilisateur pourrait souhaiter se servir pour obtenir les estimations de la variance sont inconnues. Par conséquent, des méthodes de rééchantillonnage sont souvent utilisées pour produire un ensemble de poids de rééchantillonnage qui peuvent être fournis avec l'ensemble de données et utilisés en vue d'estimer la variance pour une grande gamme d'estimateurs possibles. Le bootstrap est particulièrement utile, puisqu'il permet de traiter des statistiques d'échantillon lisses ainsi que non lisses sous des plans d'échantillonnage à plusieurs degrés. Un sommaire de plusieurs méthodes du bootstrap pour l'échantillonnage en population finie peut être consulté dans Shao et Tu (1995, pages 232–282) (voir aussi, Gross 1980; Bickel et Freedman 1984; McCarthy et Snowden 1985; Rao et Wu 1988; Kovar, Rao et Wu 1988; Sitter 1992a, b; Booth, Butler et Hall 1994; Shao et Sitter 1996).

Si la fraction de sondage de premier degré est faible, diverses méthodes du bootstrap existent pour traiter l'échantillonnage de premier degré comme s'il avait eu lieu avec remise afin d'estimer la variance. Dans le cas où les fractions de sondage de premier degré ne sont pas négligeables, un moins grand nombre de résultats sont disponibles. Pour le « bootstrappage » sous échantillonnage à deux degrés avec échantillonnage aléatoire simple (EAS) à chaque degré, voir Sitter (1992a, 1992b) et pour celui avec probabilités inégales, voir Rao et Wu (1988). Cependant, si les fractions de sondage de premier degré ne sont pas négligeables, aucune méthode du bootstrap simple n'existe pour trois degrés ou plus d'échantillonnage. Dans le présent article, nous proposons une nouvelle méthode du bootstrap qui permet de traiter facilement les cas pour lesquels un échantillonnage aléatoire simple (EAS) est utilisé à chaque degré. Nous l'appelons bootstrap de type Bernoulli (EBB) à cause de sa ressemblance à l'échantillonnage à partir d'une loi de Bernoulli. Nous utilisons les données de l'Enquête nationale sur les prix (ENP) du Japon pour l'illustrer.

Le plan de l'article est le suivant. À la section 2, nous présentons la notation pour l'échantillonnage stratifié à trois degrés. À la section 3, nous décrivons deux types d'EBB. À la section 4, nous étudions les propriétés de la méthode par simulation. À la section 5, nous décrivons le plan d'échantillonnage de l'ENP de 1997 et illustrons l'application de l'EBB aux données de l'ENP. Enfin, à la section 6, nous présentons nos conclusions.

2. Échantillonnage stratifié à trois degrés

Dans l'échantillonnage aléatoire stratifié, la population finie, constituée de N unités primaires d'échantillonnage (UPE) est fractionnée en H *strates* non chevauchantes contenant $N_1, N_2, \ldots, N_H$ UPE, respectivement; donc, $\sum_{h=1}^{H} N_h = N$. Un échantillon aléatoire simple sans remise (EASSR) d'UPE est tiré indépendamment dans chaque strate. Les tailles d'échantillon dans chaque strate sont dénotées par $n_1, n_2, \ldots, n_H$, et la taille totale de

1. F. Funaoka, professeur, Faculty of Economics, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano, 390-8621, Japon; H. Saigo, professeur, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050, Japon; R.R. Sitter, professeur, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada; T. Toida, professeur associé, Faculty of Social and Information Studies, Gunma University, 2-4 Aramakicho, Maebashi, Gunma 371-8510, Japon.

l'échantillon d'UPE est $n = \sum_{h=1}^{H} n_h$. Au deuxième degré, un échantillon de m_{hi} unités secondaires d'échantillonnage (USE) est sélectionné à partir de l'UPE i de taille M_{hi} dans la strate h par EASSR. Au troisième degré, un échantillon de l_{hij} unités finales d'échantillonnage (UFE) est sélectionné à partir de l'USE ij de taille L_{hij} dans la strate h par EASSR. Un vecteur de mesures de certaines caractéristiques des unités est représenté par $\mathbf{y}_{hijk} = (y_{1hijk}, y_{2hijk}, y_{\tau hijk})^T$, où les indices inférieurs $hijk$ sont l'étiquette de strate, l'étiquette d'UPE, l'étiquette d'USE et l'étiquette d'UFE, respectivement. Le paramètre de population d'intérêt $\theta = \theta(\mathrm{S})$, où $\mathrm{S} = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, N_h; j = 1, \ldots, M_{hi}; k = 1, \ldots, L_{hij}\}$, est habituellement estimé par $\hat{\theta} = \hat{\theta}(\mathrm{s})$, où $\mathrm{s} = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, n_h; j = 1, \ldots, m_{hi}; k = 1, \ldots, l_{hij}\}$. Le vecteur des totaux de population est dénoté $\mathbf{Y} = (Y_1, \ldots, Y_\tau)^T$. Ici, son estimateur sans biais est :

$$\hat{\mathbf{Y}} = \sum_{h=1}^{H} \hat{\mathbf{Y}}_{\mathbf{h}} = \sum_{h=1}^{H} (N_h / n_h) \sum_{i=1}^{n_h} \hat{\mathbf{Y}}_{\mathbf{hi}},$$

où $\hat{\mathbf{Y}}_{\mathbf{hi}} = (M_{hi} / m_{hi}) \sum_{j=1}^{m_{hi}} \hat{\mathbf{Y}}_{\mathbf{hij}}$ et $\hat{\mathbf{Y}}_{\mathbf{hij}} = (L_{hij} / l_{hij}) \sum_{k=1}^{l_{hij}} \mathbf{y}_{\mathbf{hijk}}$, ce qui peut s'écrire sous la forme $\hat{\mathbf{Y}} = \sum_{hijk} w_{hij} \mathbf{y}_{\mathbf{hijk}}$, où $w_{hij} = (N_h / n_h)(M_{hi} / m_{hi})(L_{hij} / l_{hij})$.

Pour $\tau = 1$, une estimation sans biais de $\mathrm{Var}(\hat{Y})$ est $v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h)$, où

$$v(\hat{Y}_h) = \frac{N_h^2 (1 - f_{1h}) s_h^2}{n_h} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2 (1 - f_{2hi}) s_{hi}^2}{m_{hi}} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{L_{hij}^2 (1 - f_{3hij}) s_{hij}^2}{l_{hij}}$$

avec $\bar{\hat{Y}}_{h\cdot} = n_h^{-1} \sum_i \hat{Y}_{hi}$, $\bar{\hat{Y}}_{hi\cdot} = m_{hi}^{-1} \sum_j \hat{Y}_{hij}$, $\bar{y}_{hij\cdot} = l_{hij}^{-1} \sum_k y_{hijk}$, $f_{1h} = n_h / N_h$, $f_{2hi} = m_{hi} / M_{hi}$, $f_{3hij} = l_{hij} / L_{hij}$, $s_h^2 = \sum_i (\hat{Y}_{hi} - \bar{\hat{Y}}_{h\cdot})^2 / (n_h - 1)$, $s_{hi}^2 = \sum_j (\hat{Y}_{hij} - \bar{\hat{Y}}_{hi\cdot})^2 / (m_{hi} - 1)$, et $s_{hij}^2 = \sum_k (y_{hijk} - \bar{y}_{hij\cdot})^2 / (l_{hij} - 1)$ (Särndal, Swensson et Wretman 1992, pages 148–149).

3. Bootstrap de type Bernoulli proposé

Afin de traiter la question de l'échantillonnage à plusieurs degrés dans une strate, nous proposons un bootstrap à plusieurs degrés. Pour simplifier les idées, nous commençons par introduire une version simple, dont l'application présente certaines limites. Puis, nous décrivons une forme plus générale qui permet d'éviter ces difficultés.

EBB abrégé

Étape I. Pour chaque UPE de l'échantillon, hi, dans la strate $h, h = 1, \ldots, H$: a) la garder dans l'échantillon bootstrap avec la probabilité

$$p_h = \sqrt{1 - \frac{(1 - f_{1h})}{(1 - n_h^{-1})}}; \qquad (3.1)$$

ou b) la remplacer par une autre sélectionnée au hasard parmi les n_h UPE. Si l'opiton est a), passer à l'étape II.

Étape II. Pour chaque USE hij dans l'UPE hi de la strate h retenue à l'étape I : c) la garder dans l'échantillon bootstrap avec la probabilité

$$q_{hi} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}}; \qquad (3.2)$$

ou d) la remplacer par une autre sélectionnée au hasard parmi les m_{hi} USE dans l'UPE hi de la strate h. Si l'option est c), passer à l'étape III.

Étape III. Pour chaque UFE $hijk$ dans l'USE hij dans l'UPE hi de la strate h : e) la garder dans l'échantillon bootstrap avec la probabilité

$$r_{hij} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{f_{2hi}}{q_{hi}^{-1}} \frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}}; \qquad (3.3)$$

ou f) la remplacer par une autre sélectionnée au hasard parmi les l_{hij} UFE dans l'USE hij dans l'UPE hi de la strate h.

Soit K_{hij}^* le nombre de fois que l'unité $hijk$ figure dans la réplique bootstrap; alors, l'estimation du total par le bootstrap est $\hat{\mathbf{Y}}^* = \sum_{hijk} w_{hij}^* \mathbf{y}_{\mathbf{hijk}}$, où $w_{hij}^* = K_{hij}^* w_{hij}$, et l'estimation de $V(\hat{\theta})$ par le bootstrap est $v_B(\hat{\theta}) = V_*(\hat{\theta}^*)$, où $\hat{\theta}^* = \theta(\hat{\mathbf{Y}}^*)$ et V_* représente la variance sous la procédure de rééchantillonnage. Habituellement, l'estimation de la variance par le bootstrap est obtenue par simulation de Monte Carlo. Autrement dit, on répète les étapes I à III un grand nombre de fois, B, pour obtenir $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ et on utilise

$$v_B(\hat{\theta}) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}_{(\cdot)}^*)^2 / B,$$

où $\bar{\theta}_{(\cdot)}^* = \sum_{b=1}^{B} \hat{\theta}^* / B$. Dans la plupart des cas, il est possible de remplacer $\bar{\theta}_{(\cdot)}^*$ par $\hat{\theta}$. Cela permet au méthodologiste d'enquête de créer un ensemble de poids de rééchantillonnage w_{hij}^* pour chaque réplique bootstrap et de les inclure dans les fichiers de données à grande diffusion.

Il est clair que l'EBB abrégé n'est applicable que si $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; h, i, j$. Par exemple, il est nécessaire que $f_{1h} \geq n_h^{-1}$. Pour traiter les cas arbitraires $n_h, m_{hi}, l_{hij} \geq 2$, nous pouvons modifier chaque étape et changer p_h, q_{hi}, r_{hij} en conséquence.

EBB général

Étape I′. Tirer $(n_h - 1)$ UPE par EAS avec remise parmi les n_h UPE de l'échantillon, $h = 1, \ldots, H$.

Dénoter l'ensemble candidat par $\{\tilde{\text{UPE}}_{hi}: i = 1, 2, \ldots, n_h - 1\}$. Pour chaque UPE i dans l'échantillon de la strate h : a) la garder dans l'échantillon bootstrap avec la probabilité

$$p_h = 1 - \frac{1}{2}\frac{(1 - f_{1h})}{(1 - n_h^{-1})}; \qquad (3.4)$$

ou b) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{UPE}}_{hi}: i = 1, 2, \ldots, n_h - 1\}$. Si l'option est a), passer à l'étape II′.

Étape II′. Pour l'unité hi retenue à l'étape **I′**, tirer $(m_{hi} - 1)$ USE par EAS avec remise parmi les m_{hi} USE dans l'UPE hi. Dénoter l'ensemble candidat par $\{\tilde{\text{USE}}_{hij}: j = 1, 2, \ldots, m_{hi} - 1\}$. Pour chaque USE hij dans l'UPE hi retenue à l'étape **I′** : c) la garder dans l'échantillon bootstrap avec la probabilité

$$q_{hi} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}; \qquad (3.5)$$

ou d) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{USE}}_{hij}: j = 1, 2, \ldots, m_{hi} - 1\}$. Si l'option est c), passer à l'étape **III′**.

Étape III′. Pour l'unité hij retenue à l'étape **II′**, tirer $l_{hij} - 1$ UFE par EAS avec remise parmi les l_{hij} UFE dans l'USE hij dans l'UPE hi. Dénoter l'ensemble candidat par $\{\tilde{\text{UFE}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$. Pour chaque UFE $hijk$ dans l'USE hij dans l'UPE hi : e) la garder dans l'échantillon bootstrap avec la probabilité

$$r_{hij} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{f_{2hi}}{q_{hi}^{-1}}\frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}; \qquad (3.6)$$

ou f) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{UFE}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$.

Il est facile de voir que $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; n_h, m_{hi}, l_{hij} \geq 2$.

La raison justifiant la sélection aléatoire d'un ensemble candidat dans l'EBB général est la suivante. Pour fixer les idées, considérons un EASSR à un degré dans une seule strate. Soit $\bar{y}^*$ une moyenne d'échantillon bootstrap sous l'EBB abrégé avec une probabilité arbitraire $p \in [0, 1]$. Alors, nous pouvons montrer que $V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})\, s^2(1 - p^2)$, où $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$. Notons que $V_*(\bar{y}^*)$ est monotone décroissante par rapport à p dans l'intervalle $[0, 1]$. Donc, $\min_{p \in [0,1]} V_*(\bar{y}^*) = 0$ et $\max_{p \in [0,1]} V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})s^2$. Si $f_1 < n^{-1}$, puis $\max_p V_*(\bar{y}^*) < v(\bar{y})$. La notion clé de l'EBB général est que nous pouvons rendre $\max_p V_*(\bar{y}^*)$ plus grand que $v(\bar{y})$ en introduisant une variation supplémentaire dans le remplacement des unités grâce à la sélection aléatoire d'un ensemble candidat.

Nous pouvons montrer que l'EBB abrégé et l'EBB général produisent une estimation convergente de la variance pour les fonctions lisses des totaux de population estimés. En outre, sous des conditions de régularité appropriées pour la fonction de répartition de la population, ils produisent aussi une estimation convergente de la variance pour les quantiles d'échantillon. Qui plus est, dans les deux méthodes EBB, la taille des répliques est égale à celle de l'échantillon original, propriété qui peut être désirable lorsque l'on a affaire à des données d'enquête imputées (voir Saigo, Shao et Sitter 2001).

Il n'est pas difficile d'étendre l'approche de l'EBB à des plans comportant plus de trois degrés. Par exemple, pour un plan stratifié à quatre degrés, une UFE au quatrième degré dans la strate h est retenue avec la probabilité

$$\sqrt{1 - p_h^{-1} f_{1h} q_{hi}^{-1} f_{2hi} r_{hij}^{-1} f_{3hij} (1 - g_{hijk}^{-1})^{-1} (1 - f_{4hijk})}$$

ou remplacée dans l'EBB abrégé, où g_{hijk} est la taille d'échantillon de quatrième degré et f_{4hijk} est la fraction de sondage de quatrième degré. Les extensions ultérieures sont analogues.

L'EBB général randomise un ensemble candidat simplement pour remédier à l'infaisabilité de l'EBB abrégé. Ce concept présente des similarités avec le bootstrap approximativement bayésien de Rubin et Schenker (1986).

Un inconvénient de l'EBB général comparativement à l'EBB abrégé est que le premier nécessite, en moyenne, $\sum_h \{(n_h - 1) + p_h \sum_i (m_{hi} - 1) + p_h \sum_i q_{hi} \sum_j (l_{hij} - 1)\}$ générations de nombres aléatoires de plus que le second, où $p_h, q_{hi},$ et r_{hij} sont donnés par (3.4), (3.5) et (3.6), respectivement, ce qui peut demander beaucoup de temps lorsque les tailles d'échantillon et(ou) le nombre de strates sont grands. Afin de réduire les générations de nombres aléatoires dans l'EBB général, on peut créer un ensemble candidat en supprimant aléatoirement une unité de l'échantillon original et utiliser

$$p_h = (n_h + 1/2) - \sqrt{(n_h + 1/2)^2 - n_h(1 + f_{1h})}, \qquad (3.7)$$

$$q_{hi} = (m_{hi} + 1/2) - \sqrt{(m_{hi} + 1/2)^2 - f_{h1} p_h^{-1} m_{hi}(1 + f_{2hi})}, \qquad (3.8)$$

$$r_{hij} = (l_{hij} + 1/2) - \sqrt{(l_{hij} + 1/2)^2 - f_{h1} p_h^{-1} f_{2hi} q_{hi}^{-1} l_{hij}(1 + f_{3hij})}, \qquad (3.9)$$

au lieu des trois équations susmentionnées. On peut montrer que $p_h, q_{hi}, r_{hij} \in [0, 1]$. La preuve de cette version modifiée de l'EBB général est similaire.

4. Une étude par simulation

À la présente section, nous décrivons l'exécution de simulations limitées pour étudier l'EBB dans le cas de l'estimation par le ratio et de l'estimation par quantile. Pour simplifier, nous considérons un EASSR à deux degrés et nous nous limitons à une seule strate.

4.1 Description générale de la simulation

Une population finie unistratifiée est générée selon la procédure qui suit et est maintenue fixe pour toutes les exécutions de la simulation afin d'observer les propriétés fondées sur le plan de sondage de l'EBB. Premièrement, la moyenne des variables auxiliaires dans la grappe i est générée par $\mu_i \sim N(\mu, \sigma^2)$ pour $i = 1, 2, \ldots, N$. Puis, la variable auxiliaire x_{ik} de l'unité k dans la grappe i est générée par

$$x_{ik} = \mu_i + \varepsilon_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \quad (4.1)$$

où $\varepsilon_{ik} \sim N(0, (1-\rho)\sigma^2/\rho)$. La variable cible y_{ik} de l'unité k dans la grappe i est obtenue par

$$y_{ik} = a + bx_{ik} + e_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \quad (4.2)$$

où $e_{ik} \sim N(0, \sigma^2/4)$. Les valeurs des paramètres sont fixées à $\mu = 100, \sigma = 10, \rho = 0{,}1(0{,}3), a = 0$ et $b = 1$, et l'EASSR à deux degrés est utilisé tout au long de l'étude par simulation.

4.2 Estimation par le ratio

Soit $N = 50, n = 15, M_i = 20$ et $m_i = 3$, for $i = 1, \ldots, n$. Considérons l'estimateur par le ratio du total de population, Y,

$$\hat{Y}_R = \hat{R}\, X,$$

où $X = \sum_{i=1}^{N} \sum_{k=1}^{M_i} x_{ik}$ est le total de population des x, $\hat{R} = \hat{Y}/\hat{X}$, $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{Y}_{hi}$, $\hat{X} = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{X}_{hi}$, $\hat{Y}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{Y}_{hik}$ et $\hat{X}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{X}_{hik}$.

Aux fins de comparaison, nous considérons un certain nombre d'estimateurs de la variance utilisables dans ce simple contexte :

1) L'estimateur classique de la variance est dénoté

$$v_0(\hat{Y}_R) = N^2 \frac{1-f_1}{n} \frac{\sum_i (\hat{Y}_i - \hat{R}\hat{X}_i)^2}{n-1} + \frac{N}{n} \sum_i \frac{M_i^2 (1-f_{2i}) s_{d'2i}^2}{m_i}, \quad (4.3)$$

où $f_1 = n/N$, $f_{2i} = m_i/M_i$ et

$$s_{d'2i}^2 = \sum_j (y_{ij} - \hat{R}x_{ij})^2 / (m_i - 1).$$

2) L'estimateur par le jackknife avec suppression d'une UPE à la fois corrigé pour la fraction de sondage de premier degré, parfois utilisé même s'il n'est pas entièrement correct, est dénoté

$$v_{cj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2, \quad (4.4)$$

où $\hat{Y}_{R(i)}$ est l'estimateur recalculé après l'élimination de la i^{e} UPE et $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n$.

3) Un estimateur par le jackknife pondéré extérieurement (voir Folsom, Bayless et Shah 1971) qui comprend une correction pour les deux degrés d'échantillonnage peut être dérivé sous la forme

$$v_{ewj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2 + f_1 \sum_i (1-f_{2i}) \frac{m_i - 1}{m_i} \sum_j (\hat{Y}_{R(ij)} - \hat{Y}_{R(\cdot\cdot)})^2, \quad (4.5)$$

où $\hat{Y}_{R(i)}$ est la i^{e} pseudo valeur jackknife obtenue par suppression de l'UPE i, $\hat{Y}_{R(ij)}$ est la ij^{e} pseudo valeur jackknife obtenue par suppression de l'unité j dans l'UPE i, $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n$ et $\hat{Y}_{R(\cdot\cdot)} = \sum_j \hat{Y}_{R(ij)}/m_i$.

4) Il existe aussi un estimateur de la variance assisté par modèle (voir Särndal, Swensson et Wretman (1992), équation (8.10.6)) de la forme

$$v_{ma}(\hat{Y}_R) = (X/\hat{X})^2 v_0(\hat{Y}_R). \quad (4.6)$$

Nous utilisons $B = 100$ répliques bootstrap dans chacune des $S = 1\,000$ exécutions de la simulation. Nous obtenons une approximation des EQM réelles d'après 10 000 exécutions de la simulation et nous utilisons les estimations de Monte Carlo du biais relatif en pourcentage et du coefficient de variation des divers estimateurs de la variance, ainsi que les probabilités empiriques de couverture des intervalles de confiance à 90 %, comme mesures de leur performance relative.

Nous voyons au tableau 1 que $v_{\text{EBB}}, v_0, v_{ewj}$ et v_{ma} donnent des résultats comparables et bons, excepté que le coefficient de variation (cv) des méthodes de rééchantillonnage est un peu plus élevé que celui des méthodes sans rééchantillonnage, ce qui est typique. Le jackknife avec suppression d'une UPE à la fois donne des résultats médiocres.

Tableau 1
Comparaison des estimateurs de la variance pour $\hat{Y}_r$

ρ		Biais en %	CV	Couverture (90 %)
0,1	v_0	−1,70	0,28	89,2
	v_{EBB}	−0,62	0,33	88,9
	v_{ewj}	−0,33	0,30	89,4
	v_{cj}	−26,55	0,39	80,5
	v_{ma}	−0,39	0,30	89,4
0,3	v_0	−0,67	0,28	86,6
	v_{EBB}	−1,63	0,33	86,5
	v_{ewj}	−0,74	0,29	86,5
	v_{cj}	−26,85	0,39	80,2
	v_{ma}	−0,87	0,29	86,4

Afin d'étudier les propriétés conditionnelles, nous avons ordonné les 1 000 exécutions de la simulation selon $X/\hat{X}$ et réparti les exécutions en 20 groupes de taille égale. Pour chaque groupe, nous avons calculé la moyenne de chaque estimateur de la variance. La figure 1 représente ces moyennes groupées pour chaque estimateur de la variance (sauf v_{cj} puisqu'il présente un biais négatif important) en fonction de la moyenne groupée $X/\hat{X}$, pour $\rho = 0,3$. L'EQM réelle est incluse dans le tracé également. Le graphique est semblable à celui utilisé par Royall et Cumberland (1981a, 1981b). Nous voyons que v_{EBB} suit l'EQM réelle, en grande partie comme v_{ewj} et v_{ma}, tandis que v_0 ne le fait pas. Donc, l'EBB semble avoir une propriété conditionnelle désirable.

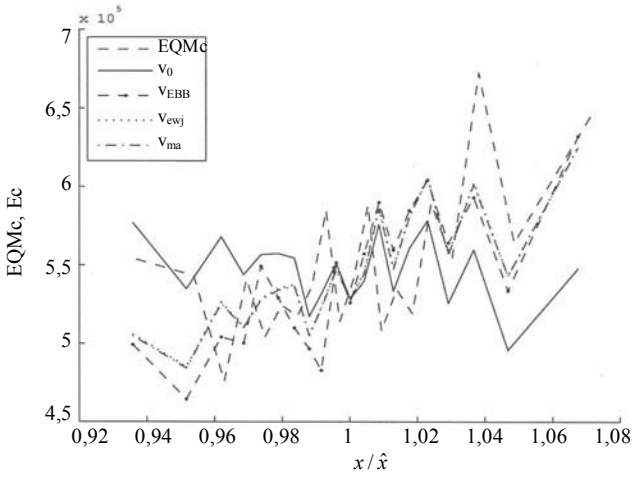

Figure 1. EQMc et Ec(v) pour l'estimation par le ratio.

4.3 Estimation par quantile

Pour l'estimation par quantile, nous posons que $N = 100, n = 30, M_i = 100$ et $m_i = 10$, pour $i = 1, \ldots, n$. Nous utilisons $B = 500$ répliques bootstrap dans chacune des $S = 5\,000$ exécutions de la simulation. Nous obtenons une approximation des EQM réelles au moyen de 50 000 exécutions de la simulation. Seuls les résultats pour v_{EBB} et v_{ewj} quand $\rho = 0,1$ sont résumés au tableau 2, parce que ceux obtenus quand $\rho = 0,3$ sont similaires. Nous voyons que la méthode de l'EBB donne d'assez bons résultats, avec un léger biais par excès, tandis que la méthode du jackknife pondérée extérieurement produit un biais important, à cause de son absence de convergence dans l'estimation de la variance pour les quantiles.

Tableau 2
Propriétés de v_{EBB} et v_{ewj} pour les quantiles 0,10, 0,25, 0,50, 0,75 et 0,90

	v_{BBE}			v_{ewj}		
Quantile	Biais en %	CV	Couverture (90 %)	Biais en %	CV	Couverture (90 %)
0,10	8,40	0,51	87,7	51,87	1,93	81,3
0,25	6,21	0,42	88,2	21,19	1,28	83,3
0,50	2,53	0,37	87,4	14,27	1,00	83,0
0,75	6,23	0,42	87,8	28,07	1,33	83,4
0,90	6,32	0,50	88,0	54,47	2,05	80,3

5. Application à l'Enquête nationale sur les prix menée en 1997 au Japon

L'objectif de l'ENP est d'analyser la formation des prix des principaux biens de consommation, comme les aliments, les vêtements et les appareils électroménagers. L'estimation par quantile joue un rôle essentiel dans cette analyse, et de nombreuses estimations par quantile fondées sur plusieurs stratifications a posteriori sont incluses dans les rapports de l'ENP.

L'échantillonnage stratifié à plusieurs degrés utilisé dans l'ENP de 1997 se résume comme suit :

Stratification. Les municipalités forment les UPE et sont réparties en 537 strates, d'abord en fonction des préfectures et des sphères économiques que constitue chaque municipalité, puis d'après la taille de leur population.

Échantillonnage de premier degré. Ces UPE sont sélectionnées par EASSR indépendamment dans chaque strate. Le tableau 3 donne un aperçu des fractions de sondage de premier degré.

Échantillonnage de deuxième degré. Dans une municipalité sélectionnée, tous les grands points de vente sont dénombrés. Autrement dit, un échantillonnage en grappes à un degré est utilisé pour ces points de vente. Pour les petits points de vente, par contre, une municipalité échantillonnée est subdivisée en régions d'enquête (USE), chacune constituée d'environ 100 points de vente. Un échantillonnage systématique est utilisé pour échantillonner les régions d'enquête. Les fractions de sondage au deuxième degré sont comprises entre 0,1 et 1,0.

Échantillonnage de troisième degré. Dans chaque région d'enquête sélectionnée, 40 points de vente (UFE) sont choisis par échantillonnage systématique ordonné en fonction du type de point de vente et du chiffre de ventes annuel déclaré lors du Recensement du commerce de 1994.

Tableau 3
Fractions de sondage de premier degré dans l'ENP de 1997

Catégorie de région	Taille de la population	N^bre d'UPE	Fractions de sondage	Taille de l'échantillon
Villes	≥ 100 000	221	1/1	221
Villes	50 000 – 99 999	220	2/3	179
Villes	< 50 000	224	1/3	80
Petites villes et villages	≥ 40 000	32	1/5	4
Petites villes et villages	< 40 000	2 536	1/15	187

À proprement parler, il n'existe aucune formule de variance valide pour les données de l'ENP, parce que celles-ci comportent un échantillonnage systématique. Cependant, pour estimer la variance, nous supposons que l'échantillonnage systématique peut être approximé par l'EASSR. Même sous cette condition simplifiée, il n'existe

aucune expression analytique explicite de la variance pour les quantiles d'échantillon. En fait, aucune estimation de la variance n'est associée aux estimations des quantiles de prix dans le rapport de l'ENP, tandis que les prix moyens sont publiés avec l'estimation de leur variance.

À la présente section, nous appliquons l'EBB abrégé aux données de l'ENP, en supposant que l'échantillonnage systématique peut être approximé par l'EASSR. Certaines strates ne contiennent qu'une seule UPE. En outre, $f_{1h} < n_h^{-1}$ dans certaines strates. Les strates de ce genre sont intégrées à des strates adjacentes de sorte que p_h, donnée par (3.1), soit comprise dans l'intervalle $[0, 1]$. Après regroupement, il existe plus de 280 strates. Nous supposons que l'effet de la reformation des strates est négligeable.

Après reformation des strates, nous employons l'EBB abrégé dans les strates composées de grandes villes. Par ailleurs, nous utilisons le bootstrap avec remise (Shao et Tu 1995, page 247) où la taille des répliques est $(n_h - 1)$ dans les strates composées de petites villes et de villages, où les fractions de sondage de premier degré sont faibles. Les estimations par quantile et leurs erreurs-types pour certains produits vendus par les petits points de vente sont présentées au tableau 4. Notons que les prix d'un produit donné sont discrets. Cependant, nous appliquons le bootstrap comme s'ils étaient continus. Cette approximation devrait être acceptable pour de nombreux produits, mais non pour ceux qui sont très bon marché, puisque, dans ce cas, un pourcentage élevé d'observations est concentré sur un prix particulier et l'erreur-type estimée peut être nulle.

6. Conclusion

Le bootstrap est utile pour estimer les variances dans le cas des enquêtes complexes, particulièrement lorsque l'estimation par quantile est importante. Nous avons proposé deux méthodes du bootstrap de type Bernoulli qui permettent de traiter facilement les plans EASSR stratifiés à plusieurs degrés où les fractions de sondage sont grandes : l'EBB abrégé et l'EBB général. Dans les deux méthodes, une unité d'échantillonnage à un degré donné est soit retenue, soit remplacée avec une probabilité prédéterminée, afin de construire un échantillon bootstrap. L'EBB général a l'avantage de permettre le traitement de toute combinaison de tailles d'échantillon ≥ 2, mais il nécessite plus de générations de nombres aléatoires que l'EBB abrégé. À titre d'illustration, nous avons appliqué l'EBB abrégé aux données de l'Enquête nationale sur les prix menée en 1997 au Japon.

Remerciements

Les travaux du deuxième auteur ont été financés par la Japan Statistical Association. Ceux du troisième auteur ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient le Bureau de la statistique, le ministère de la Gestion publique, des Affaires intérieures, des Postes et des Télécommunications, ainsi que le ministère de l'Économie, du Commerce et de l'Industrie du Japon d'avoir fourni les données de l'ENP de 1997.

Tableau 4
Quantiles d'échantillon (erreurs-types) de certains produits pour les petits points de vente dans l'ENP

Produit	p	0,10	0,25	0,5	0,75	0,90
Riz (5kg)[a]	Quantile d'échantillon	239,4	255,2	278,3	299,1	315,0
(10 yens)	(erreur-type)	(0,24)	(0,53)	(0,21)	(0,02)	(0,61)
Café instantané (1 flacon)[b]	Quantile d'échantillon	714	788	859	893	914
(yen)	(erreur-type)	(0,13)	(0,40)	(0,00)	(2,68)	(1,43)
Bière (24 cannettes)[c]	Quantile d'échantillon	467,3	500,0	536,8	549,4	549,4
(10 yens)	(erreur-type)	(1,01)	(0,64)	(0,82)	(0,00)	(0,00)
PC[d]	Quantile d'échantillon	248,8	260,4	299,3	346,5	375,9
(1 000 yens)	(erreur-type)	(2,03)	(0,35)	(3,25)	(7,17)	(1,48)

Marques spécifiées : [a]Koshihikari; [b]Nescafe Gold Blend, 100g; [c]Sapporo (Nama) Black Label, 350ml; [d]NEC PC9821 NW133/D14.

Bibliographie

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statististics*, 12, 470-482.

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

Folsom, R.E., Bayless, D.L. et Shah, B.V. (1971). Jackknifing for variance components in complex sample survey designs. *Proceedings of the Social Statistics Section*, American Statistical Association, 36-39.

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

Kovar, J.G., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statististics*, 16, Supplément, 25-45.

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

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.

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.

Rubin, D.B., et Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statististical Association*, 81, 366-374.

Saigo, H., Shao, J. et Sitter, R.R. (2001). Bootstrap à demi-échantillon répété et répliques équilibrées répétées en cas d'imputation aléatoire de données. *Techniques d'enquête*, 27, 209-218.

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

Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.

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

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statististics*, 20, 135-154.