Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Certaines contributions à la méthode du jackknife appliquée aux estimateurs sous échantillonnage à deux phases

par Patrick J. Farrell et Sarjinder Singh

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 63-74
Statistique Canada, N° 12-001-X au catalogue

Certaines contributions à la méthode du jackknife appliquée aux estimateurs sous échantillonnage à deux phases

Patrick J. Farrell et Sarjinder Singh [1]

Résumé

Dans le présent article, le problème de l'estimation de la variance de divers estimateurs de la moyenne de population sous échantillonnage à deux phases est traité par application de la méthode du jackknife aux poids calés en deux phases de Hidiroglou et Särndal (1995, 1998). Nous montrons que plusieurs estimateurs de la moyenne de population décrits dans la littérature sont des cas particuliers de la méthode élaborée ici, y compris ceux proposés par Rao et Sitter (1995) et par Sitter (1997). En nous inspirant de Raj (1965) et de Srivenkataramana et Tracy (1989), nous introduisons de nouveaux estimateurs de la moyenne de population et nous estimons leur variance par la méthode du jackknife proposée. Nous estimons également la variance des estimateurs en chaîne par le ratio et par la régression dus à Chand (1975) en utilisant le jackknife. Une étude par simulations nous permet d'évaluer l'efficacité des estimateurs jackknife proposés comparativement aux estimateurs de variance usuels.

Mots clés : Information auxiliaire ; calage ; estimation de la moyenne et de la variance ; jackknife ; échantillonnage à deux phases.

1. Introduction

Hidiroglou et Särndal (1995, 1998) ont fait remarquer que le recours à l'échantillonnage à deux phases pour estimer les caractéristiques d'une population finie est une technique puissante et rentable qui joue donc un rôle très important dans l'échantillonnage. L'échantillonnage à deux phases peut être décrit comme il suit. Considérons une population finie que nous désignons par $\Omega = \{1, 2, ..., i, ..., N\}$. Supposons que l'information sur une variable Z est disponible pour l'ensemble de la population ; autrement dit, les valeurs Z_i pour tout $i = 1, ..., N$ sont connues, ce qui implique que la moyenne de population, $\bar{Z}$, est également connue. Un échantillon probabiliste de première phase s_1, $s_1 \subset \Omega$, de taille m est tiré de la population avec les probabilités de sélection π_{1i}. Donc, les poids d'échantillonnage de première phase peuvent être définis comme $d_{1i} = 1/\pi_{1i}$. Supposons que, pour cet échantillon, l'information est recueillie sur une variable X, qui est ensuite appariée avec l'information sur Z pour chacune des m unités, ce qui fournit les données $\{(x_i, z_i) \mid i \in s_1\}$ pour $i = 1, ..., m$. Lorsque l'échantillon de première phase s_1 a été tiré, un échantillon de deuxième phase s_2, $s_2 \subset s_1 \subset \Omega$, de taille n est sélectionné à partir de s_1 avec les probabilités de sélection $\pi_{2i} = \pi_{i|s_1}$, ce qui permet de définir les poids d'échantillonnage de deuxième phase par $d_{2i} = 1/\pi_{2i}$. Dans l'échantillon de deuxième phase, l'information est maintenant recueillie sur une variable Y pour chaque unité sélectionnée. Cette information est liée à celle disponible antérieurement sur Z et X pour ces unités, ce qui fournit les données $\{(x_i, y_i, z_i) \mid i \in s_2\}$ pour $i = 1, ..., n$. Supposons que nous souhaitons estimer la moyenne de population $\bar{Y}$, ainsi que la variance de l'estimateur employé.

Soit $w_{1i}^o = d_{1i} / \sum_{i \in s_1} d_{1i}$ les poids de sondage originaux normalisés de première phase. L'estimateur habituel de la moyenne de population $\bar{X}$ est donné par

$$\hat{\bar{X}}_1^o = \sum_{i \in s_1} w_{1i}^o x_i$$

tandis qu'un estimateur calé de première phase de $\bar{X}$ est donné par

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^c x_i,$$

où les w_{1i}^c sont les poids calés de telle façon que la fonction de distance du khi-deux

$$D_1 = \sum_{i \in s_1} \{(w_{1i}^c - w_{1i}^o)^2 / (w_{1i}^o q_{1i})\} \tag{1.1}$$

est minimisée sous la contrainte

$$\sum_{i \in s_1} w_{1i}^c z_i = \bar{Z}. \tag{1.2}$$

Dans (1.1), les q_{1i} sont un ensemble de poids choisis de manière appropriée. La minimisation de (1.1) sous la contrainte (1.2) produit les poids calés de première phase

$$w_{1i}^c = w_{1i}^o + \left\{(q_{1i}\, w_{1i}^o\, z_i) \Big/ \left(\sum_{i \in s_1} q_{1i}\, w_{1i}^o\, z_i^2\right)\right\}\left(\bar{Z} - \sum_{i \in s_1} w_{1i}^o\, z_i\right).$$

Donc, un estimateur de première phase calé de $\bar{X}$ est donné par

1. Patrick J. Farrell, School of Mathematics and Statistics, Carleton University, 1125 promenade Colonel By, Ottawa (Ontario), Canada, K1S 5B6. Courriel : pfarrell@math.carleton.ca ; Sarjinder Singh, Department of Mathematics, Texas A&M University - Kingsville, Kingsville, Texas, É.-U., 78363. Courriel : sarjinder@yahoo.com.

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^o x_i + \hat{\beta}_1 \left(\bar{Z} - \sum_{i \in s_1} w_{1i}^o z_i \right),$$

où

$$\hat{\beta}_1 = \left(\sum_{i \in s_1} q_{1i} w_{1i}^o x_i z_i \right) \Big/ \left(\sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2 \right).$$

Maintenant, désignons par $w_{2i}^o = d_{1i} d_{2i} / \sum_{i \in s_2} d_{1i} d_{2i}$ les poids de sondage normalisés de deuxième phase. L'estimateur habituel de $\bar{Y}$ est donné par

$$\hat{\bar{Y}}_2^o = \sum_{i \in s_2} w_{2i}^o y_i.$$

Considérons l'estimateur de deuxième phase calé de $\bar{Y}$ de la forme

$$\hat{\bar{Y}}^c = \sum_{i \in s_2} w_{2i}^c y_i, \tag{1.3}$$

où les w_{2i}^c sont les poids calés de deuxième phase tels que la fonction de distance du khi-deux

$$D_2 = \sum_{i \in s_2} \{ (w_{2i}^c - w_{2i}^o)^2 / (w_{2i}^o q_{2i}) \} \tag{1.4}$$

est minimisée sous la contrainte de calage

$$\sum_{i \in s_2} w_{2i}^c x_i = \hat{\bar{X}}_1^c. \tag{1.5}$$

La minimisation de (1.4) sous la contrainte (1.5) donne les poids de deuxième phase calés

$$w_{2i}^c = w_{2i}^o + \left\{ (q_{2i} w_{2i}^o x_i) \Big/ \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2 \right) \right\} \left(\hat{\bar{X}}_1^c - \sum_{i \in s_2} w_{2i}^o x_i \right).$$

Donc, l'estimateur de deuxième phase calé de $\bar{Y}$ spécifié en (1.3) peut s'écrire

$$\hat{\bar{Y}}^c = \hat{\bar{Y}}_2^o + \hat{\beta}_2 (\hat{\bar{X}}_1^o - \hat{\bar{X}}_2^o) + \hat{\beta}_1 \hat{\beta}_2 (\bar{Z} - \hat{\bar{Z}}_1^o), \tag{1.6}$$

où $\hat{\bar{Z}}_1^o = \sum_{i \in s_1} w_{1i}^o z_i$, $\hat{\bar{X}}_1^o = \sum_{i \in s_1} w_{1i}^o x_i$, $\hat{\bar{X}}_2^o = \sum_{i \in s_2} w_{2i}^o x_i$, $\hat{\bar{Y}}_2^o = \sum_{i \in s_2} w_{2i}^o y_i$, et

$$\hat{\beta}_2 = \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i y_i \right) \Big/ \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2 \right).$$

Hidiroglou et Särndal (1995, 1998), ainsi que Singh (2000) ont étudié le problème de l'estimation de la variance de l'estimateur calé $\hat{\bar{Y}}^c$ donné par (1.6) en utilisant une approche fondée sur le plan de sondage. Dans un contexte plus général, Rao et Sitter (1995), ainsi que Sitter (1997) ont fait observer que, sous échantillonnage aléatoire simple sans remise (EASSR), une méthode jackknife peut être utilisée pour estimer les variances des estimateurs par le ratio et par la régression d'une moyenne de population. Ces auteurs ont également signalé que l'estimation de la variance par la méthode du jackknife est plus commode et efficace que par les méthodes classiques fondées sur les estimations des moments.

Dernièrement, un certain nombre d'auteurs ont étudié l'emploi de procédures jackknife pour estimer les variances (voir Arnab et Singh 2006, Berger 2007, Berger et Skinner 2005, Chen et Shao 2001, et Kovar et Chen 1994). Fuller (1998), Kim, Navarro et Fuller (2000, 2006), Kim et Sitter (2003), ainsi que Kott et Stukel (1997) ont proposé une approche d'estimation de la variance dans le cas de l'échantillonnage à deux degrés. Fuller (1998), ainsi que Kim et Sitter (2003) se sont penchés sur l'estimateur par la régression. En particulier, ils ont considéré l'estimateur par la régression généralisée du total de population

$$\hat{Y}_{\text{DS}} = \sum_{i \in s_2} \alpha_i y_i$$

dû à Deville et Särndal (1992). Selon Kim et coll. (2000, 2006), pour chaque $k \in s_2$, l'estimateur jackknife du total de population est spécifié comme étant

$$\hat{Y}_{\text{Kim}} = \sum_{i \in s_2 \backslash k} \alpha_i^{(k)} y_i \tag{1.7}$$

et la distance du khi-deux entre les poids de sondage et les poids calés comme étant

$$D_{(k)} = (1/2) \sum_{i \in s_2 \backslash k} \{ (\alpha_i^{(k)} - w_i^{(k)} w_i^{*(k)})^2 / (w_i^{(k)} q_i^{(k)}) \}. \tag{1.8}$$

La minimisation de (1.8) sous la contrainte

$$\sum_{i \in s_2 \backslash k} \alpha_i^{(k)} x_i = \sum_{i \in s_1 \backslash k} w_i^{(k)} x_i$$

mène aux poids calés jackknife donnés par

$$\alpha_i^{(k)} = w_i^{(k)} w_i^{*(k)} + \left\{ (w_i^{(k)} q_i^{(k)} x_i) \Big/ \left(\sum_{i \in s_2 \backslash k} w_i^{(k)} q_i^{(k)} \right) \right\} \left\{ \sum_{i \in s_2 \backslash k} w_i^{(k)} x_i - \sum_{i \in s_2 \backslash k} w_i^{(k)} w_i^{*(k)} x_i \right\}.$$

Il semble que Kim et coll. (2006) ont rajusté ces poids de la façon suivante

$$\alpha_i^{(k)} = \begin{cases} \alpha_i^{(k)} & \text{si } k \in s_2 \\ w_i^{(k)} & \text{si } j \in (s_1 - s_2). \end{cases}$$

Pour un tel rajustement, l'estimateur donné par (1.7) est équivalent à celui de Rao et Sitter (1995).

Dans le présent article, nous considérons une nouvelle méthode jackknife pour estimer la variance de l'estimateur $\hat{\bar{Y}}^c$ sous échantillonnage à deux phases en suivant Hidiroglou et Särndal (1995, 1998). Comme l'estimateur de Kim et coll. (2006), nous montrons que celui présenté par Rao et Sitter (1995) est un cas particulier de la méthode proposée. Cependant, notre approche diffère de celle de Fuller (1998), de Kim et Sitter (2003) et de Kim et coll.

(2000, 2006) en ce sens que nous considérons le calage à la première ainsi qu'à la deuxième phase, ce qui permet d'élaborer la méthode des estimateurs par le ratio en chaîne et par la régression en chaîne. Nous évaluons aussi, dans une étude par simulations, l'efficacité des estimateurs jackknife de variance comparativement aux estimateurs de variance usuels.

2. Estimation de la variance par la méthode du jackknife

Dans la suite de l'exposé, nous supposons qu'un plan à un seul degré est employé aux deux phases du processus d'échantillonnage. Soit $\hat{\bar{Y}}^c(j)$ un estimateur calé de la moyenne de population, $\bar{Y}$, obtenu en retranchant la j^{e} unité de l'échantillon s_1 de m unités. Nous prouvons en annexe que l'estimateur jackknife de la moyenne de population sous échantillonnage à deux phases peut s'écrire

$$\hat{\bar{Y}}^c(j) = \begin{cases} \hat{\bar{Y}}_2^o(j) + \hat{\beta}_2(j)\,\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_2^o(j)\} \\ + \hat{\beta}_1(j)\,\hat{\beta}_2(j)\,\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} \quad \text{si } j \in s_2 \\ \\ \hat{\bar{Y}}_2^o + \hat{\beta}_2\,\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} \\ + \hat{\beta}_1(j)\,\hat{\beta}_2\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} \quad \text{si } j \in (s_1 - s_2) \end{cases} \tag{2.1}$$

où la quantité $\hat{\bar{Z}}_1^o(j) = \hat{\bar{Z}}_1^o + \{w_{1j}^o/(1 - w_{1j}^o)\}\{\hat{\bar{Z}}_1^o - z_j\}$, les termes $\hat{\bar{X}}_1^o(j)$, $\hat{\bar{X}}_2^o(j)$ et $\hat{\bar{Y}}_2^o(j)$ sont définis de manière analogues, $\hat{\beta}_1(j) = \hat{\beta}_1 + \{q_{1j}w_{1j}^o z_j(x_j - \hat{\beta}_1 z_j)\}/\{q_{1j}w_{1j}^o z_j^2 - \sum_{i\in s_1} q_{1i}w_{1i}^o z_i^2\}$ et $\hat{\beta}_2(j) = \hat{\beta}_2 + \{q_{2j}w_{2j}^o x_j\,(y_j - \hat{\beta}_2 x_j)\}/\{q_{2j}w_{2j}^o x_j^2 - \sum_{i\in s_1} q_{2i}w_{2i}^o x_i^2\}$. L'estimateur jackknife modifié de la variance de $\hat{\bar{Y}}^c$ est alors donné par

$$\hat{V}_{\text{JACK}}(\hat{\bar{Y}}^c) = \{(m-1)/m\} \sum_{j\in s_1} \{\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c\}^2. \tag{2.2}$$

Nous montrons en annexe que cet estimateur est convergent.

Notons que nous pouvons écrire

$$\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c = \begin{cases} \varepsilon_2(j) + \hat{\beta}_2\varepsilon_1(j) + \hat{\beta}_2(j)\,d_2(j) \\ + \hat{\beta}_2\delta_2(j) \quad \text{si } j \in s_2 \\ \\ \hat{\beta}_2\varepsilon_1(j) \quad \text{si } j \in (s_1 - s_2) \end{cases} \tag{2.3}$$

où les termes de (2.3) sont donnés par $\varepsilon_1(j) = \{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}$, $\varepsilon_2(j) = \{\hat{\bar{Y}}_2^o(j) - \hat{\bar{Y}}_2^o\} - \hat{\beta}_2(j)\{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_2^o\} - \hat{\beta}_1(j)\,\hat{\beta}_2(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}$, $d_2(j) = \{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_2^o(j)\}$ et $\delta_2(j) = \{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} - \hat{\beta}_1\{\bar{Z} - \hat{\bar{Z}}_1^o\}$. Le terme $\varepsilon_1(j)$ est analogue au terme d'erreur associé à la régression de la variable auxiliaire x_i sur z_i, pour $i \in s_1$, tandis que $\varepsilon_2(j)$ est analogue au terme d'erreur associé à la régression de la variable étudiée y_i sur x_i ainsi que z_i simultanément, pour $i \in s_2$. À condition que $j \in s_2$, le terme $d_2(j)$ reflète la différence entre les moyennes jackknife d'échantillon de première et de deuxième phase pour la variable X, tandis que $\delta_2(j)$ désigne une correction de $d_2(j)$ obtenue en utilisant l'information sur la variable auxiliaire Z.

En utilisant (2.3) dans (2.2), l'estimateur jackknife de la variance de l'estimateur $\hat{\bar{Y}}^c$ est donné par

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{\bar{Y}}^c) = {} & \{(m-1)/m\} \\ & \Big[\sum_{j\in s_2} \varepsilon_2^2(j) + \sum_{j\in s_2} \hat{\beta}_2^2(j)\,d_2^2(j) \\ & + \hat{\beta}_2^2 \sum_{j\in s_2} \delta_2(j)\,\{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2\hat{\beta}_2 \sum_{j\in s_2} \varepsilon_1(j)\,\varepsilon_2(j) \\ & + 2\hat{\beta}_2 \sum_{j\in s_2} \hat{\beta}_2(j)\,d_2(j)\{\varepsilon_1(j) + \delta_2(j)\} \\ & + \hat{\beta}_2^2 \sum_{j\in s_1} \varepsilon_1^2(j)\Big]. \end{aligned} \tag{2.4}$$

Notons que l'expression (2.4) est exacte. Elle peut être utilisée pour estimer la variance de plusieurs estimateurs décrits dans la littérature.

3. Cas particuliers

À la section suivante, nous démontrons que les estimateurs proposés par Rao et Sitter (1995), Sitter (1997), Raj (1965), Srivenkataramana et Tracy (1989), Chand (1975) et Ahmed (1997) peuvent être considérés comme des cas particuliers de la méthode proposée.

Cas 3.1 : Rao et Sitter (1995)

Si $\hat{\bar{X}}_1^c = \hat{\bar{X}}_1^o$ (aucun calage de première phase n'est effectué) et $q_{2i} = 1/x_i$, l'estimateur calé de $\bar{Y}$ devient

$$\hat{\bar{Y}}_r^c = \left(\sum_{i\in s_2} w_{2i}^o y_i\right)\left\{\left(\sum_{i\in s_2} w_{1i}^o x_i\right)\Big/\left(\sum_{i\in s_2} w_{2i}^o x_i\right)\right\}.$$

Si l'échantillon de première phase s_1 est sélectionné selon un plan EASSR tel que les poids de sondage de première phase sont donnés par $d_{1i} = N/m$ et que l'échantillon de deuxième phase s_2 est sélectionné à partir de s_1 par EASSR de façon que $d_{2i} = m/n$, l'estimateur calé de la moyenne de population devient

$$\hat{\bar{Y}}_{\text{RS}}^c = \bar{y}(\bar{x}'/\bar{x}), \tag{3.1}$$

où $\bar{y} = \sum_{i \in s_2} y_i/n$, $\bar{x} = \sum_{i \in s_2} x_i/n$ et $\bar{x}' = \sum_{i \in s_1} x_i/m$. Dans (2.1), le mécanisme jackknife devient

$$\hat{\bar{Y}}_{\text{RS}}^c(j) = \begin{cases} \dfrac{(n\bar{y} - y_j)(m\bar{x}' - x_j)}{(n\bar{x} - x_j)(m-1)} & \text{si } j \in s_2 \\ (\bar{y}/\bar{x})\dfrac{(m\bar{x}' - x_j)}{(m-1)} & \text{si } j \in (s_1 - s_2). \end{cases} \tag{3.2}$$

En fixant $\hat{R} = \bar{y}/\bar{x}$, la différence entre (3.2) et (3.1) peut s'écrire

$$\hat{\bar{Y}}_{\text{RS}}^c(j) - \hat{\bar{Y}}_{\text{RS}}^c =$$

$$\begin{cases} -\hat{R}\dfrac{(x_j - \bar{x}')}{(m-1)} - \dfrac{\bar{x}'(j)}{\bar{x}(j)}\dfrac{(y_j - \hat{R}x_j)}{(n-1)} & \text{si } j \in s_2 \\ -\hat{R}\dfrac{(x_j - \bar{x}')}{(m-1)} & \text{si } j \in (s_1 - s_2). \end{cases} \tag{3.3}$$

L'expression (3.3) est exactement la même que celle donnée par Rao et Sitter (1995). En supposant que $\bar{x}'(j)/\bar{x}(j) \approx \bar{x}'/\bar{x}$, l'estimateur jackknife approximatif de la variance est donné par

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{RS}}^c) \approx & \left(\frac{\bar{x}'}{\bar{x}}\right)^2 \sum_{i \in s_2} \frac{(y_i - \hat{R}x_i)^2}{n(n-1)} \\ & + 2\left(\frac{\bar{x}'}{\bar{x}}\right)\hat{R}\sum_{j \in s_2}\frac{(x_j - \bar{x}')(y_j - \hat{R}x_j)}{n-1} \\ & + \hat{R}^2\sum_{j \in s_1}\frac{(x_j - \bar{x}')^2}{m(m-1)}. \end{aligned}$$

Donc, l'estimateur de Rao et Sitter (1995) est un cas particulier de la méthode du jackknife proposée.

Cas 3.2 : Sitter (1997)

Dans le cas 3.1, si nous considérons $q_{2i} = 1$, l'estimateur calé sous EASSR devient

$$\hat{\bar{Y}}_{lr}^c = \bar{y} + b^*(\bar{x}' - \bar{x}), \tag{3.4}$$

où $b^* = \sum_{i \in s_2} x_i y_i / \sum_{i \in s_2} x_i^2$ désigne un estimateur du coefficient de régression β qui est légèrement différent de celui envisagé par Sitter (1997). Le mécanisme du jackknife prend la forme

$$\hat{\bar{Y}}_{lr}^c(j) =$$

$$\begin{cases} \dfrac{n\bar{y} - y_j}{n-1} + \left\{ b^* + \dfrac{x_j(y_j - b^* x_j)}{\sum_{i \in s_2} x_i^2 - x_j^2} \right\} \\ \qquad \left\{ \dfrac{m\bar{x}' - x_j}{m-1} - \dfrac{n\bar{x} - x_j}{n-1} \right\} & \text{si } j \in s_2 \\ \bar{y} + b^* \left\{ \dfrac{m\bar{x}' - x_j}{m-1} - \bar{x} \right\} & \text{si } j \in (s_1 - s_2). \end{cases} \tag{3.5}$$

Si nous posons que $d_j^* = (y_j - \bar{y}) - b^*(x_j - \bar{x})$, $a_j^* = x_j\{\bar{x}(j) - \bar{x}'(j)\}/K$ et $k_j^* = x_j^2/K$, où $K = (n-1)s_x^2 + n\bar{x}^2$, la différence entre (3.5) et (3.4) peut s'écrire sous la forme

$$\hat{\bar{Y}}_{lr}^c(j) - \hat{\bar{Y}}_{lr}^c =$$

$$\begin{cases} -b^*\dfrac{(x_j - \bar{x}')}{(m-1)} - \dfrac{\dfrac{d_j^*}{(n-1)}}{\left[1 + \dfrac{a_j^*}{(1 - k_j^*)}\right]} & \text{si } j \in s_2 \\ -b^*\dfrac{(x_j - \bar{x}')}{(m-1)} & \text{si } j \in (s_1 - s_2) \end{cases}$$

qui est semblable à l'expression donnée par Sitter (1997).

Cas 3.3 : Raj (1965)

Afin d'examiner ce cas, nous supposons que l'échantillon initial s_1 de taille m est sélectionné avec remise avec les probabilités p_i proportionnelles à z_i, $i = 1, 2, ..., N$. L'information sur la variable auxiliaire X est recueillie sur cet échantillon de première phase, s_1. L'échantillon de deuxième phase, dont la taille est fixée à n, est un sous-échantillon de s_1 sélectionné sans remise avec probabilités égales. C'est pour cet échantillon s_2 que l'information sur Y est recueillie. Sous ce plan d'échantillonnage, $d_{1i} = 1/\pi_{1i} = 1/(mp_i)$ et $d_{2i} = m/n$. Donc, $w_{1i}^o = (1/p_i)/\sum_{i \in s_1}(1/p_i)$ et $w_{2i}^o = (1/p_i)/\sum_{i \in s_2}(1/p_i)$. Notons aussi que, pour ce plan, $\hat{\bar{X}}_1^c = \hat{\bar{X}}_1^o$; donc, aucun calage de première phase n'est effectué. Si $q_{2i} = 1/x_i$, l'estimateur calé $\hat{\bar{Y}}^c$ devient

$$\hat{\bar{Y}}_{\text{Raj}}^c = \hat{\bar{Y}}_2^o(\hat{\bar{X}}_1^o / \hat{\bar{X}}_2^o), \tag{3.6}$$

où $\hat{\bar{Y}}_2^o = \sum_{i \in s_2}(y_i/p_i)/\sum_{i \in s_2}(1/p_i)$, $\hat{\bar{X}}_2^o = \sum_{i \in s_2}(x_i/p_i)/\sum_{i \in s_2}(1/p_i)$ et $\hat{\bar{X}}_1^o = \sum_{i \in s_1}(x_i/p_i)/\sum_{i \in s_1}(1/p_i)$. Donc, alternativement, $\hat{\bar{Y}}_{\text{Raj}}^c = \{\sum_{i \in s_2}(y_i/p_i)\sum_{i \in s_1}(x_i/p_i)\}/\{\sum_{i \in s_2}(x_i/p_i)\sum_{i \in s_1}(1/p_i)\}$.

Sous le plan d'échantillonnage décrit ci-dessus, l'estimateur jackknife de la moyenne de population est

$$\hat{\bar{Y}}_{\text{Raj}}^c(j) = \begin{cases} \hat{\bar{Y}}_2^o(j)\dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o(j)} & \text{si } j \in s_2 \\ \hat{\bar{Y}}_2^o\dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o} & \text{si } j \in (s_1 - s_2) \end{cases} \tag{3.7}$$

où

$$\hat{\bar{Y}}_2^o(j) =$$

$$\frac{\sum_{i \in s_2}(y_i/p_i)}{\sum_{i \in s_2}(1/p_i)} + \frac{(1/p_j)\Big/\sum_{i \in s_2}(1/p_i)}{1 - \dfrac{(1/p_j)}{\sum_{i \in s_2}(1/p_i)}}\left\{\frac{\sum_{i \in s_2}(y_i/p_i)}{\sum_{i \in s_2}(1/p_i)} - y_j\right\},$$

et $\hat{\hat{X}}_2^o(j)$ et $\hat{\hat{X}}_1^o(j)$ sont définis de manière analogue. Si $\hat{R} = \hat{\hat{Y}}_2^o / \hat{\hat{X}}_2^o$ et $w_{2j}^o = (1/p_j)/\sum_{i\in s_2}(1/p_i)$, la différence entre (3.7) et (3.6) peut facilement s'écrire sous la forme

$$\hat{\hat{Y}}_{\text{Raj}}^c(j) - \hat{\hat{Y}}_{\text{Raj}}^c = \begin{cases} -w_{2j}^o \dfrac{\hat{\hat{X}}_1^o(j)}{\hat{\hat{X}}_2^o(j)}(y_j - \hat{R}x_j) \\ \qquad + \hat{R}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\} & \text{si } j\in s_2 \\ \hat{R}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\} & \text{si } j\in(s_1 - s_2). \end{cases}$$

Donc, l'estimateur jackknife de la variance de l'estimateur $\hat{\hat{Y}}_{\text{Raj}}^c$ est donné par

$$\hat{V}_{\text{JACK}}(\hat{\hat{Y}}_{\text{Raj}}^c) = \frac{m-1}{m}\Bigg[\sum_{j\in s_2}(w_{2j}^o)^2 \left(\frac{\hat{\hat{X}}_1^o(j)}{\hat{\hat{X}}_2^o(j)}\right)^2 (y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}^2 - 2\hat{R}\sum_{j\in s_2} w_{2j}^o \frac{\hat{\hat{X}}_1^o(j)}{\hat{\hat{X}}_2^o(j)}(y_j - \hat{R}x_j)\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}\Bigg].$$

À l'instar de Rao et Sitter (1995), si nous supposons que $\hat{\hat{X}}_1^o(j)/\hat{\hat{X}}_2^o(j) \approx \hat{\hat{X}}_1^o/\hat{\hat{X}}_2^o$, l'estimateur jackknife de la variance de $\hat{\hat{Y}}_{\text{Raj}}^c$ prend la forme

$$\hat{V}_{\text{JACK}}(\hat{\hat{Y}}_{\text{Raj}}^c) \approx \frac{m-1}{m}\Bigg[\{\hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o\}^2 \sum_{j\in s_2}(w_{2j}^o)^2 (y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}^2 - 2\hat{R}\{\hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o\}\sum_{j\in s_2} w_{2j}^o (y_j - \hat{R}x_j)\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}\Bigg].$$

Cas 3.4 : Srivenkataramana et Tracy (1989)

Afin d'examiner ce cas, comme dans Raj (1965), nous supposons que l'échantillon initial s_1 de taille m est sélectionné avec remise avec probabilités proportionnelles à z_i. Cependant, le sous-échantillon, s_2, de n unités est maintenant sélectionné avec remise avec probabilités proportionnelles à x_i / z_i. Par conséquent, $w_{1i}^o = (1/z_i)/\sum_{i\in s_1}(1/z_i)$ et $w_{2i}^o = (1/x_i)/\sum_{i\in s_2}(1/x_i)$. Comme dans Raj (1965), nous n'effectuons aucun calage de première phase ; donc $\hat{\hat{X}}_1^c = \hat{\hat{X}}_1^o$. D'où, si $q_{2i} = 1/x_i$, l'estimateur calé $\hat{\hat{Y}}^c$ est

$$\hat{\hat{Y}}_{\text{ST}}^c = \hat{\hat{Y}}_2^o(\hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o), \tag{3.8}$$

où $\hat{\hat{Y}}_2^o = \sum_{i\in s_2}(y_i/x_i)/\sum_{i\in s_2}(1/x_i)$, $\hat{\hat{X}}_2^o = n/\sum_{i\in s_2}(1/x_i)$ et $\hat{\hat{X}}_1^o = \sum_{i\in s_1}(x_i/z_i)/\sum_{i\in s_1}(1/z_i)$. Donc, alternativement, $\hat{\hat{Y}}_{\text{ST}}^c = \{\sum_{i\in s_2}(y_i/x_i)\sum_{i\in s_1}(x_i/z_i)\}/\{n\sum_{i\in s_1}(1/z_i)\}$.

Sous le plan d'échantillonnage décrit ci-dessus, l'estimateur jackknife de la moyenne de population est

$$\hat{\hat{Y}}_{\text{ST}}^c(j) = \begin{cases} \hat{\hat{Y}}_2^o(j)\{\hat{\hat{X}}_1^o(j)/\hat{\hat{X}}_2^o(j)\} & \text{si } j\in s_2 \\ \hat{\hat{Y}}_2^o\{\hat{\hat{X}}_1^o(j)/\hat{\hat{X}}_2^o\} & \text{si } j\in(s_1 - s_2) \end{cases} \tag{3.9}$$

où

$$\hat{\hat{Y}}_2^o(j) = \frac{\sum_{i\in s_2}(y_i/x_i)}{\sum_{i\in s_2}(1/x_i)} + \frac{1}{x_j\sum_{i\in s_2}(1/x_i) - 1}\left\{\frac{\sum_{i\in s_2}(y_i/x_i)}{\sum_{i\in s_2}(1/x_i)} - y_j\right\}.$$

Les termes $\hat{\hat{X}}_2^o(j)$ et $\hat{\hat{X}}_1^o(j)$ sont définis similairement ; autrement dit

$$\hat{\hat{X}}_2^o(j) = \frac{n}{\sum_{i\in s_2}(1/x_i)} + \frac{1}{x_j\sum_{i\in s_2}(1/x_i) - 1}\left\{\frac{n}{\sum_{i\in s_2}(1/x_i)} - x_j\right\},$$

tandis que $\hat{\hat{X}}_1^o(j)$ peut s'écrire

$$\hat{\hat{X}}_1^o(j) = \frac{\sum_{i\in s_1}(x_i/z_i)}{\sum_{i\in s_1}(1/z_i)} + \frac{1}{x_j\sum_{i\in s_1}(1/x_i) - 1}\left\{\frac{\sum_{i\in s_1}(y_i/x_i)}{\sum_{i\in s_1}(1/x_i)} - x_j\right\}.$$

Si $\hat{R} = \sum_{i\in s_2}(y_i/x_i)/n$ et $w_{2j}^o = (1/x_j)/\sum_{i\in s_2}(1/x_i)$, la différence entre (3.9) et (3.8) est donnée par

$$\hat{\hat{Y}}_{\text{ST}}^c(j) - \hat{\hat{Y}}_{\text{ST}}^c = \begin{cases} -w_{2j}^o \dfrac{\hat{\hat{X}}_1^o(j)}{\hat{\hat{X}}_2^o(j)}(y_j - \hat{R}x_j) + \hat{R}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\} & \text{si } j\in s_2 \\ \hat{R}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\} & \text{si } j\in(s_1 - s_2). \end{cases}$$

Suivant Rao et Sitter (1995), si nous supposons que $\hat{\hat{X}}_1^o(j)/\hat{\hat{X}}_2^o(j) \approx \hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o$, l'estimateur jackknife de la variance de $\hat{\hat{Y}}_{\text{ST}}^c$ prend la forme

$$\hat{V}_{\text{JACK}}(\hat{\hat{Y}}_{\text{ST}}^c) \approx \frac{m-1}{m}\Bigg[\{\hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o\}^2 \sum_{j\in s_2}(w_{2j}^o)^2 (y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}^2 - 2\hat{R}\{\hat{\hat{X}}_1^o / \hat{\hat{X}}_2^o\}\sum_{j\in s_2} w_{2j}^o (y_j - \hat{R}x_j)\{\hat{\hat{X}}_1^o(j) - \hat{\hat{X}}_1^o\}\Bigg].$$

Cas 3.5 : Chand (1975)

Afin d'examiner ce cas, l'échantillon de première phase s_1 de taille m est sélectionné par EASSR et les variables auxiliaires Z et X sont toutes deux observées sur les unités choisies. Le sous-échantillon, s_2, de n unités est également sélectionné par EASSR. Manifestement, $d_{1i} = N/m$ et $d_{2i} = m/n$, de sorte que $w_{1i}^o = 1/m$ et $w_{2i}^o = 1/n$. Si $q_{1i} = 1/z_i$ et $q_{2i} = 1/x_i$, l'estimateur calé $\hat{\bar{Y}}^c$ devient

$$\hat{\bar{Y}}_{\text{Ch}}^c = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}'), \tag{3.10}$$

où

$$\bar{y} = \sum_{i\in s_2} y_i/n,\ \bar{x} = \sum_{i\in s_2} x_i/n,\ \bar{x}' = \sum_{i\in s_1} x_i/m$$

et $\bar{z}' = \sum_{i\in s_1} z_i/m$. L'estimateur jackknife de $\bar{Y}$ est

$$\hat{\bar{Y}}_{\text{Ch}}^c(j) = \begin{cases} \bar{y}(j)\dfrac{\bar{x}'(j)}{\bar{x}(j)}\dfrac{\bar{Z}}{\bar{z}'(j)} & \text{si } j\in s_2 \\[2ex] \bar{y}(j)\dfrac{\bar{x}'(j)}{\bar{x}}\dfrac{\bar{Z}}{\bar{z}'(j)} & \text{si } j\in(s_1-s_2) \end{cases} \tag{3.11}$$

où $\bar{y}(j) = (n\bar{y} - y_j)/(n-1)$, $\bar{x}(j) = (n\bar{x} - x_j)/(n-1)$, $\bar{x}'(j) = (m\bar{x}' - x_j)/(m-1)$ et enfin $\bar{z}'(j) = (m\bar{z}' - z_j)/(m-1)$. Si nous posons que $\hat{R}_1 = \bar{x}'/\bar{z}'$ (un estimateur de $R_1 = \bar{X}/\bar{Z}$) et que $\hat{R}_2 = \bar{y}/\bar{x}$ (un estimateur de $R_2 = \bar{Y}/\bar{X}$), et similairement que $\hat{R}_1(j) = \bar{x}'(j)/\bar{z}'(j)$ et $\hat{R}_2(j) = \bar{y}(j)/\bar{x}(j)$, la différence entre (3.11) et (3.10) peut s'écrire

$$\hat{\bar{Y}}_{\text{Ch}}^c(j) - \hat{\bar{Y}}_{\text{Ch}}^c = \begin{cases} \varepsilon_2(j) + \hat{R}_2\varepsilon_1(j) + \hat{R}_2(j)\,d_2(j) + \hat{R}_2\delta_2(j) & \text{si } j\in s_2 \\[2ex] \hat{R}_2\varepsilon_1(j) & \text{si } j\in(s_1-s_2) \end{cases} \tag{3.12}$$

où nous pouvons écrire dans (3.12) que $\varepsilon_2(j) = \{\bar{y}(j) - \bar{y}\} - \hat{R}_2(j)\{\bar{x}(j) - \bar{x}\} - \hat{R}_1(j)\hat{R}_2(j)\{\bar{z}'(j) - \bar{Z}\}$, $d_2(j) = \{\bar{x}'(j) - \bar{x}'\}$, $\delta_2(j) = \{\bar{x}(j) - \bar{x}'(j)\} - \hat{R}_1(j)\{\bar{Z} - \bar{z}'(j)\} - \hat{R}_1\{\bar{Z} - \bar{z}'\}$ et, enfin, que le terme $\varepsilon_1(j) = \{\bar{x}'(j) - \bar{x}'\} - \hat{R}_1(j)\{\bar{z}'(j) - \bar{Z}\}$. Donc, l'estimateur jackknife de la variance de l'estimateur $\hat{\bar{Y}}_{\text{Ch}}^c$ est donné par

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Ch}}^c) = {} & \\ \{(m-1)/m\}\Bigg[& \sum_{j\in s_2}\varepsilon_2^2(j) + \sum_{j\in s_2}\hat{R}_2^2(j)\,d_2^2(j) \\ & + \hat{R}_2^2\sum_{j\in s_2}\delta_2(j)\{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2\hat{R}_2\sum_{j\in s_2}\varepsilon_1(j)\varepsilon_2(j) \\ & + 2\hat{R}_2\sum_{j\in s_2}\hat{R}_2(j)\,d_2(j)\{\varepsilon_1(j) + \delta_2(j)\} \\ & + \hat{R}_2^2\sum_{j\in s_1}\varepsilon_1^2(j)\Bigg]. \end{aligned}$$

Cas 3.6 : Ahmed (1997)

Considérons le même plan d'échantillonnage que pour le cas 3.5. Plutôt que $q_{1i} = 1/z_i$ et $q_{2i} = 1/x_i$, comme dans Chand (1975), nous posons que $q_{1i} = q_{2i} = 1$ et $q_{2i} = 1/x_i$, de sorte que l'estimateur calé se réduit à

$$\hat{\bar{Y}}_{\text{Chlr}}^c = \bar{y} + b_2^*(\bar{x}' - \bar{x}) + b_1^* b_2^*(\bar{Z} - \bar{z}'), \tag{3.13}$$

où $b_2^* = \sum_{i\in s_2} x_i y_i / \sum_{i\in s_2} x_i^2$ et $b_1^* = \sum_{i\in s_1} x_i z_i / \sum_{i\in s_1} z_i^2$. Notons que (3.13) est un estimateur de type par régression en chaîne semblable à celui d'Ahmed (1997). En posant que $b_2^*(j) = b_2^* + \{x_j(y_j - b_2^* x_j)/(x_j^2 - \sum_{i\in s_2} x_i^2)\}$ et $b_1^*(j) = b_1^* + \{z_j(x_j - b_1^* z_j)/(z_j^2 - \sum_{i\in s_1} z_i^2)\}$, après l'application du jackknife, l'estimateur $\hat{\bar{Y}}_{\text{Chlr}}^c$ devient

$$\hat{\bar{Y}}_{\text{Chlr}}^c(j) = \begin{cases} \bar{y}(j) + b_2^*(j)\{\bar{x}'(j) - \bar{x}(j)\} \\ \quad + b_1^*(j)b_2^*(j)\{\bar{Z} - \bar{z}'(j)\} & \text{si } j\in s_2 \\ \bar{y} + b_2^*\{\bar{x}'(j) - \bar{x}\} \\ \quad + b_1^*(j)b_2^*\{\bar{Z} - \bar{z}'(j)\} & \text{si } j\in(s_1-s_2). \end{cases} \tag{3.14}$$

La différence entre (3.14) et (3.13) peut s'exprimer

$$\hat{\bar{Y}}_{\text{Chlr}}^c(j) - \hat{\bar{Y}}_{\text{Chlr}}^c = \begin{cases} \varepsilon_2(j) + b_2^*\varepsilon_1(j) + b_2^*(j)\,d_2(j) + b_2^*\delta_2(j) & \text{si } j\in s_2 \\ b_2^*\varepsilon_1(j) & \text{si } j\in(s_1-s_2) \end{cases} \tag{3.15}$$

où nous pouvons écrire dans (3.15) que $\varepsilon_2(j) = \{\bar{y}(j) - \bar{y}\} - b_2^*(j)\{\bar{x}(j) - \bar{x}\} - b_1^*(j)\,b_2^*(j)\{\bar{z}'(j) - \bar{Z}\}$, $d_2(j) = \{\bar{x}'(j) - \bar{x}'\}$, $\delta_2(j) = \{\bar{x}(j) - \bar{x}'(j)\} - b_1^*(j)\{\bar{Z} - \bar{z}'(j)\} - b_1^*\{\bar{Z} - \bar{z}'\}$ et, enfin, que le terme $\varepsilon_1(j) = \{\bar{x}'(j) - \bar{x}'\} - b_1^*(j)\{\bar{z}'(j) - \bar{Z}\}$. Donc, l'estimateur jackknife de la variance de l'estimateur $\hat{\bar{Y}}_{\text{Chlr}}^c$ est donné par

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Chlr}}^c) = {} & \\ \{(m-1)/m\}\Bigg[& \sum_{j\in s_2}\varepsilon_2^2(j) + \sum_{j\in s_2}\{b_2^*(j)\}^2 d_2^2(j) \\ & + \{b_2^*(j)\}^2\sum_{j\in s_2}\delta_2(j)\{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2b_2^*\sum_{j\in s_2}\varepsilon_1(j)\varepsilon_2(j) \\ & + 2b_2^*\sum_{j\in s_2}b_2^*(j)\,d_2(j)\{\varepsilon_1(j) + \delta_2(j)\} \\ & + \{b_2^*\}^2\sum_{j\in s_1}\varepsilon_1^2(j)\Bigg]. \end{aligned}$$

4. Étude par simulations

À la présente section, nous présentons les résultats d'études par simulations conçues pour examiner les propriétés de la méthode du jackknife proposée lorsqu'elle est utilisée pour estimer la variance de quatre des estimateurs à deux phases de la moyenne de population présentés à la section 3. Plus précisément, nous considérons l'estimateur de type ratio de Rao et Sitter (1995), l'estimateur de type régression de Sitter (1997), l'estimateur de type ratio en chaîne de Chand (1975) et l'estimateur de type régression en chaîne d'Ahmed (1997). Pour commencer, nous décrivons et présentons les résultats des simulations qui ont été effectuées pour les estimateurs de Sitter et Rao (1995) et de Sitter (1997). Suivent une discussion et un résumé de simulations semblables pour les estimateurs de Chand (1975) et d'Ahmed (1997). Puisque, contrairement au cas des estimateurs par le ratio et par la régression, de l'information complète sur une deuxième variable auxiliaire Z est requise pour l'ensemble de la population afin d'appliquer les deux estimateurs en chaîne, les simulations qui ont été effectuées pour ces estimateurs sont un peu plus compliquées que celles exécutées pour les estimateurs par le ratio et par la régression.

4.1 Étude par simulations : estimateurs de Rao et Sitter (1995) et de Sitter (1997)

Pour le premier ensemble de simulations, nous supposons que l'échantillon de première phase de m unités est sélectionné dans une population de N unités et que seule la variable auxiliaire X est mesurée. Dans l'échantillon de première phase de m unités, nous sélectionnons ensuite un échantillon de deuxième phase de n unités par EASSR dans lequel sont mesurées la variable étudiée, Y, ainsi que la variable auxiliaire, X.

Nous commençons par créer une population de N unités constituée de paires $(X_i,\ Y_i)$ en utilisant le modèle

$$Y_i = \beta X_i + \sqrt{X_i^g}\,\varepsilon_i,$$

avec $\beta = 10$. Au départ, nous fixons que $g = 0$ et $N = 500$. Pour chaque $i,\ i = 1, ..., N,$ nous générons X_i à partir d'une loi gamma dont le paramètre de forme est égal à 3,1 et le paramètre d'échelle est égal à 1, et ε_i suit une loi normale standard. Parmi la population résultante de paires $(X_i,\ Y_i),$ nous sélectionnons 1 000 échantillons de première phase de $m = 100$ unités, et dans chacun de ces échantillons, nous sélectionnons 10 000 échantillons de deuxième phase de $n = 20$ unités.

Sous le plan d'échantillonnage utilisé ici, Rao et Sitter (1995) ont proposé l'estimateur par le ratio

$$\hat{\bar{Y}}_{\mathrm{RS}}^c = \bar{y}(\bar{x}'/\bar{x}), \tag{4.1}$$

dont la variance approximative est

$$V(\hat{\bar{Y}}_{\mathrm{RS}}^c) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2,$$

où

$$S_d^2 = (N-1)^{-1}\sum_{i=1}^{N}[(Y_i - \bar{Y}) - R(X_i - \bar{X})]^2$$

et

$$S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2,$$

avec $\bar{Y} = \sum_{i=1}^N Y_i/N,\ \bar{X} = \sum_{i=1}^N X_i/N$ et $R = \bar{Y}/\bar{X}$. Pour le t^{e} échantillon de deuxième phase ($t = 1, \ldots, 10\,000$) tiré du k^{e} échantillon de première phase ($k = 1, \ldots, 1\,000$), nous calculons l'estimateur usuel de variance

$$\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] = \left(\frac{1}{n} - \frac{1}{m}\right)s_{d(t|k)}^2 + \left(\frac{1}{m} - \frac{1}{N}\right)s_{y(t|k)}^2, \tag{4.2}$$

où les variances d'échantillon sont

$$s_{d(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

et

$$s_{y(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_{i(t|k)} - \bar{y}_{(t|k)})^2$$

avec $\bar{y}_{(t|k)} = \sum_{i=1}^n y_{i(t|k)}/n$ et $\bar{x}_{(t|k)} = \sum_{i=1}^n x_{i(t|k)}/n$. En outre, $r_{(t|k)} = \bar{y}_{(t|k)}/\bar{x}_{(t|k)}$. Nous calculons aussi l'estimateur jackknife de variance

$$\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] = \frac{m-1}{m}\sum_{j=1}^{m}\left[\bar{y}_{(t|k)}(j)\frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)} - \bar{y}_{(t|k)}\frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}}\right]^2 \tag{4.3}$$

et le ratio des variances estimées

$$\mathrm{RV}(t|k) = \hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)]/\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)].$$

Puis, nous calculons la moyenne de $\mathrm{RV}(t|k)$ sur tous les k et t, laquelle est donnée par

$$\mathrm{RV} = \frac{1}{10\,000\,000}\sum_{k=1}^{1\,000}\sum_{t=1}^{10\,000}\mathrm{RV}(t|k).$$

Nous déterminons aussi les estimations empiriques des biais dans (4.2) et (4.3) en calculant

$$\mathrm{EBU} = \frac{1}{10\,000\,000}\sum_{k=1}^{1\,000}\sum_{t=1}^{10\,000}\{\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^c)\},$$

et

$$\mathrm{EBJ} = \frac{1}{10\,000\,000}\sum_{k=1}^{1\,000}\sum_{t=1}^{10\,000}\{\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^c)\}.$$

Notons que l'estimateur donné par (4.2) est sans biais. Enfin, nous calculons l'efficacité relative de l'estimateur usuel de variance par rapport à l'estimateur jackknife par

$$\mathrm{ER} = \left(\frac{\sum_{k=1}^{1\,000}\sum_{t=1}^{10\,000}\{\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^c)\}^2}{\sum_{k=1}^{1\,000}\sum_{t=1}^{10\,000}\{\hat{V}_{\mathrm{JACK}}[\hat{\bar{Y}}_{\mathrm{RS}}^c(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^c)\}^2}\right).$$

En nous servant de la même population générée de taille $N = 500$, nous répétons la simulation, mais en utilisant cette fois $m = 400$ et $n = 80$. Puis, nous créons quatre populations supplémentaires de taille $N = 500$ en utilisant $g = 0{,}5,\ 1{,}0,\ 1{,}5$ et $2{,}0$. Pour chacune de ces quatre populations, nous répétons les deux simulations décrites plus haut où, dans la première simulation, $m = 100$ et $n = 20$ et, dans la deuxième simulation, $m = 400$ et $n = 80$. Enfin, pour étudier l'effet de la taille de la population, nous répétons toutes les simulations basées sur les diverses valeurs de g, m et n quand $N = 500$ pour trois valeurs supplémentaires de N, à savoir 5 000, 50 000 et 500 000. Les résultats obtenus pour RV, EBU, EBJ et ER pour chacune de ces simulations sont présentés au tableau 1.

Dans ce tableau, les résultats pour l'efficacité relative (ER) donnent à penser que, quand la taille de population N tend vers l'infini (comme l'ont envisagé Rao et Sitter 1995), l'estimateur jackknife de variance demeure plus efficace que l'estimateur sans biais usuel de variance. Nous constatons aussi que, pour de très grandes valeurs de N, les valeurs de RV tendent vers 1. Cependant, si nous prenons les cas où $N = 500$, quand la taille de population est relativement faible, non seulement les valeurs de RV sont appréciablement plus faibles que 1, mais l'estimateur jackknife de variance semble présenter un biais significatif. En outre, l'estimateur jackknife paraît nettement moins efficace que l'estimateur sans biais usuel de variance surtout quand m et n sont grands. Il convient de souligner ici que Rao et Sitter (1995) et Sitter (1997) déclarent que la façon de fixer la valeur des facteurs de correction pour la population finie dans l'estimateur jackknife de variance sous échantillonnage à deux phases n'est pas claire. Il s'agit peut-être d'un domaine où de futurs travaux de recherche seraient fructueux, car il semble que, si la taille de population est faible, il pourrait être préférable d'ajuster les facteurs de correction pour population finie au lieu d'appliquer directement la méthode du jackknife selon l'approche proposée ici. Notons que Kim et coll. (2006) ont intégré un facteur de correction pour population finie dans un cas particulier.

Tableau 1
Comparaison des estimateurs jackknife et usuel de la variance de l'estimateur par le ratio de la moyenne de population quand β = 10 et que la variable auxiliaire *X* suit une loi gamma avec un paramètre de forme de 3,1 et un paramètre d'échelle de 1

N	*m*	*n*	*g*	RV	EBU	EBJ	ER
500	100	20	0,0	0,801	0,006	0,542	1,521
			0,5	0,800	0,010	0,579	1,310
			1,0	0,805	-0,071	0,561	1,267
			1,5	0,816	-0,358	0,575	1,149
			2,0	0,840	-0,720	1,777	0,935
5 000	100	20	0,0	0,979	-0,028	0,042	4,015
			0,5	0,976	0,007	0,096	3,709
			1,0	0,965	0,023	0,172	3,210
			1,5	0,936	-0,073	0,337	1,308
			2,0	0,916	-1,103	0,493	0,967
50 000	100	20	0,0	1,001	-0,002	0,003	6,241
			0,5	0,998	0,107	0,126	4,936
			1,0	0,981	0,101	0,196	2,965
			1,5	0,937	-0,211	0,167	1,558
			2,0	0,924	-0,355	0,940	1,005
500 000	100	20	0,0	1,001	-0,057	-0,054	4,730
			0,5	0,999	0,014	0,024	4,669
			1,0	0,993	0,185	0,229	3,223
			1,5	0,940	-0,235	0,122	1,420
			2,0	0,907	-1,054	0,530	1,009
500	400	80	0,0	0,214	0,000	0,520	0,002
			0,5	0,237	-0,001	0,523	0,002
			1,0	0,320	0,000	0,544	0,006
			1,5	0,530	-0,001	0,616	0,066
			2,0	0,733	-0,012	1,091	0,452
5 000	400	80	0,0	0,919	-0,003	0,061	2,687
			0,5	0,920	-0,001	0,064	2,505
			1,0	0,922	0,003	0,077	2,058
			1,5	0,930	-0,028	0,077	1,372
			2,0	0,940	-0,089	0,184	1,088
50 000	400	80	0,0	0,991	-0,008	-0,001	4,550
			0,5	0,991	0,004	0,012	5,276
			1,0	0,991	0,000	0,009	4,163
			1,5	0,980	-0,024	-0,001	1,777
			2,0	0,967	-0,171	-0,040	1,099
500 000	400	80	0,0	1,000	0,009	0,009	5,501
			0,5	0,999	0,001	0,001	5,180
			1,0	0,993	-0,001	0,006	3,852
			1,5	0,992	-0,022	-0,018	1,809
			2,0	0,971	-0,179	-0,079	1,136

Nous considérons aussi l'estimateur par la régression de Sitter (1997) et répétons entièrement l'étude par simulations exécutée en utilisant l'estimateur par le ratio (4.1). En particulier, au lieu de (4.1), nous utilisons l'estimateur

$$\hat{\bar{Y}}_S^c = \bar{y} + b^*(\bar{x}' - \bar{x}) \tag{4.4}$$

dont la variance approximative est

$$V(\hat{\bar{Y}}_S^c) = (n^{-1} - m^{-1}) S_d^2 + (m^{-1} - N^{-1}) S_y^2, \tag{4.5}$$

où

$$S_d^2 = (N-1)^{-1} \sum_{i=1}^{N} [(Y_i - \bar{Y}) - \beta_{\text{POP}}(X_i - \bar{X})]^2$$

avec

$$\beta_{\text{POP}} = \sum_{i=1}^{N} X_i Y_i \Big/ \sum_{i=1}^{N} X_i^2.$$

Pour chaque combinaison distincte de N, g, m et n utilisée dans l'étude par simulations, nous calculons

$$\hat{V}[(\hat{\bar{Y}}_S^c(t|k)] = (n^{-1} - m^{-1}) s_{d(t|k)}^2 + (m^{-1} - N^{-1}) s_{y(t|k)}^2, \tag{4.6}$$

pour le t^{e} échantillon de deuxième phase provenant du k^{e} échantillon de première phase, où la variance d'échantillon est

$$s_{d(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^{n} [(y_{i(t|k)} - \bar{y}_{(t|k)}) - b_{(t|k)}^*(x_{i(t|k)} - \bar{x}_{(t|k)})]^2.$$

Nous calculons aussi l'estimateur jackknife de variance

$$\hat{V}_{\text{JACK}}[(\hat{\bar{Y}}_S^c(t|k)] = \frac{m-1}{m} \sum_{j=1}^{m} [\bar{y}_{(t|k)}(j) + b_{(t|k)}^*(j)\{\bar{x}'_{(t|k)}(j) - \bar{x}_{(t|k)}(j)\} - \{\bar{y} + b^*(\bar{x}' - \bar{x})\}]^2. \tag{4.7}$$

Pour chaque combinaison distincte de N, g, m et n, nous utilisons les équations (4.5) à (4.7) pour calculer les valeurs de RV, EBU, EBJ et ER correspondant à celles présentées dans le tableau 1 pour l'estimateur (4.1). Les résultats sont extrêmement semblables à ceux obtenus pour l'estimateur par le ratio.

4.2 Étude par simulations : estimateurs de Chand (1975) et d'Ahmed (1997)

Pour le deuxième ensemble de simulations, nous supposons que, quand l'échantillon de première phase de m unités est sélectionné dans la population de taille N, l'information sur deux variables auxiliaires X et Z est recueillie. Quand l'échantillon de deuxième phase de taille n est sélectionné dans l'échantillon de première phase, la variable étudiée Y est mesurée, ainsi que les deux variables auxiliaires X et Z. Nous supposons aussi que la variable auxiliaire Z est connue pour l'ensemble de la population.

Nous commençons par créer une population de $N = 500$ unités d'observation (X_i, Z_i, Y_i) en utilisant

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \varepsilon_i,$$

avec $\beta_1 = 3{,}5$ et $\beta_2 = 2{,}5$. Pour chaque i, $i = 1, \ldots, N$, nous tirons X_i d'une loi gamma dont le paramètre de forme est égal à 2,2 et le paramètre d'échelle, à 1, Z_i d'une loi gamma dont le paramètre de forme est égal à 0,1 et le paramètre d'échelle, à 1, et ε_i d'une loi normale standard. Parmi la population résultante d'observations (X_i, Z_i, Y_i), nous sélectionnons 1 000 échantillons de première phase de $m = 100$ unités et, dans chacun de ces échantillons, nous sélectionnons 10 000 échantillons de deuxième phase de $n = 20$ unités.

Suivant Chand (1975), un estimateur par le ratio en chaîne sous échantillonnage à deux phases est donné par

$$\hat{\bar{Y}}_{\text{Ch}}^c = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}')$$

dont la variance approximative est

$$V(\hat{\bar{Y}}_{\text{Ch}}^c) = (n^{-1} - m^{-1})\, S_{d_2}^2 + (m^{-1} - N^{-1})\, S_{d_1}^2, \tag{4.8}$$

où

$$S_{d_2}^2 = (N-1)^{-1} \sum_{i=1}^{N} [(Y_i - \bar{Y}) - R_2(X_i - \bar{X})]^2$$

et

$$S_{d_1}^2 = (N-1)^{-1} \sum_{i=1}^{N} [(Y_i - \bar{Y}) - R_1(Z_i - \bar{Z})]^2$$

avec

$$\bar{Y} = \sum_{i=1}^{N} Y_i / N,\ \bar{X} = \sum_{i=1}^{N} X_i / N,\ \bar{Z} = \sum_{i=1}^{N} Z_i / N,$$

$R_1 = \bar{Y}/\bar{Z}$ et $R_2 = \bar{Y}/\bar{X}$. Dans l'étude par simulations, nous calculons

$$\hat{V}[(\hat{\bar{Y}}_{\text{Ch}}^c(t|k)] = (n^{-1} - m^{-1})\, s_{d_2(t|k)}^2 + (m^{-1} - N^{-1})\, s_{d_1(t|k)}^2 \tag{4.9}$$

pour le t^{e} échantillon de deuxième phase tiré du k^{e} échantillon de première phase, où les variances d'échantillon sont

$$s_{d_2(t|k)}^2 = (n-1)^{-1} \sum_{i=1}^{n} [(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{2(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

avec

$$r_{2(t|k)} = \bar{y}_{(t|k)} / \bar{x}_{(t|k)}$$

et

$$s^2_{d_1(t|k)} = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{1(t|k)}(z_{i(t|k)} - \bar{z}_{(t|k)})]^2$$

avec $r_{1(t|k)} = \bar{y}_{(t|k)} / \bar{z}_{(t|k)}$. Nous calculons aussi l'estimateur jackknife de variance

$$\hat{V}_{\text{JACK}}[(\hat{\bar{Y}}^c_{\text{Ch}}(t|k)] = \frac{m-1}{m}\sum_{j=1}^{m}\left[\bar{y}_{(t|k)}(j)\frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)}\frac{\bar{Z}}{\bar{z}'_{(t|k)}(j)} - \bar{y}_{(t|k)}\frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}}\frac{\bar{Z}}{\bar{z}'_{(t|k)}}\right]^2. \quad (4.10)$$

En nous servant de la même population générée de taille $N = 500$, nous répétons la simulation ; toutefois, cette fois-ci, nous utilisons $m = 400$ et $n = 80$. Puis, nous créons trois populations supplémentaires de taille $N = 500$ en utilisant $\beta_1 = 0{,}5$ avec $\beta_2 = 0{,}5$, $\beta_1 = 3{,}5$ avec $\beta_2 = 0{,}5$, et $\beta_1 = 0{,}5$ avec $\beta_2 = 2{,}5$. Pour chacune de ces trois populations, nous répétons les deux simulations décrites plus haut, où, dans la première simulation, $m = 100$ avec $n = 20$ et, dans la deuxième simulation, $m = 400$ et $n = 80$. Enfin, pour étudier l'effet de la taille de la population, nous répétons toutes les simulations fondées sur les diverses valeurs de β_1, β_2, m et n quand $N = 500$ pour trois valeurs supplémentaires de N, à savoir 5 000, 50 000 et 500 000. Pour chaque combinaison distincte de N, β_1, β_2, m et n, nous utilisons les équations (4.8) à (4.10) pour calculer les valeurs de RV, EBU, EBJ et ER analogues à celles présentées au tableau 1 pour l'estimateur (4.1). Les résultats sont donnés dans le tableau 2.

En général, les résultats fondés sur les résultats du tableau 2 sont semblables à ceux obtenus pour les estimateurs fondés sur (4.1) et (4.4). En particulier, l'estimateur jackknife de variance est plus efficace que l'estimateur usuel quand la taille de population est suffisamment grande. Toutefois, il convient aussi de souligner que cette efficacité semble être reliée à l'ordre de grandeur des coefficients de régression β_1 et β_2 ; autrement dit, l'estimateur jackknife paraît être relativement plus efficace pour les cas où le coefficient associé à la variable auxiliaire X est grand comparativement au coefficient analogue lié à Z.

Tableau 2
Comparaison des estimateurs jackknife et usuel de la variance de l'estimateur par le ratio en chaîne de la moyenne de population où la variable auxiliaire *X* suit une loi gamma dont le paramètre de forme est 2,2 et le paramètre d'échelle, 1, et la variable auxiliaire *Z* suit une loi gamma dont le paramètre de forme est 0,1 et le paramètre d'échelle, 1

m	n	β_1	β_2	N	RV	EBU	EBJ	ER
100	20	3,5	2,5	500	0,769	0,000	0,027	1,063
				5 000	0,831	-0,012	0,020	2,282
				50 000	0,818	-0,006	0,028	1,785
				500 000	0,852	0,001	0,036	1,993
100	20	0,5	0,5	500	0,911	-0,001	0,004	0,791
				5 000	0,943	-0,001	0,002	0,888
				50 000	0,948	0,000	0,003	0,896
				500 000	0,946	0,000	0,003	0,899
100	20	3,5	0,5	500	0,845	-0,001	0,015	1,674
				5 000	0,932	-0,011	0,000	3,632
				50 000	0,947	-0,005	0,004	3,221
				500 000	0,947	0,000	0,010	3,637
100	20	0,5	2,5	500	0,866	-0,001	0,009	0,668
				5 000	0,858	-0,003	0,008	0,775
				50 000	0,855	-0,001	0,010	0,670
				500 000	0,855	0,000	0,012	0,697
400	80	3,5	2,5	500	0,540	0,000	0,013	0,044
				5 000	0,780	-0,001	0,009	1,346
				50 000	0,819	0,000	0,008	1,878
				500 000	0,810	-0,001	0,006	1,953
400	80	0,5	0,5	500	0,817	0,000	0,003	0,254
				5 000	0,956	0,000	0,000	0,885
				50 000	0,973	0,000	0,001	0,946
				500 000	0,973	0,000	0,000	0,963
400	80	3,5	0,5	500	0,579	0,000	0,010	0,041
				5 000	0,907	-0,001	0,003	3,158
				50 000	0,954	0,000	0,002	3,845
				500 000	0,950	-0,001	0,001	4,853
400	80	0,5	2,5	500	0,787	0,000	0,004	0,222
				5 000	0,862	0,000	0,002	0,570
				50 000	0,873	0,000	0,003	0,698
				500 000	0,875	0,000	0,002	0,595

Enfin, nous exécutons une étude par simulations analogue en utilisant l'estimateur par la régression d'Ahmed (1997). Cependant, nous créons les populations en utilisant $\beta_1 = 10$ avec $\beta_2 = 0{,}5$, $\beta_1 = 100$ avec $\beta_2 = 0{,}5$, $\beta_1 = 0{,}5$ avec $\beta_2 = 10$, et $\beta_1 = 10$ avec $\beta_2 = 10$. Comme auparavant, quand nous avons examiné les estimateurs de Rao et Sitter (1995), de Sitter (1997) et de Chand (1975), à condition que la population soit suffisamment grande, l'estimateur jackknife de variance semble être plus efficace que l'estimateur usuel.

5. Conclusion et discussion

Dans le présent article, nous étudions le problème de l'estimation de la variance de divers estimateurs de la moyenne de population sous échantillonnage à deux phases en appliquant la méthode du jackknife aux poids calés en deux phases bien connus de Hidiroglou et Särndal (1995, 1998). Des études par simulations fondées sur les estimateurs par le ratio, par la régression et de type en chaîne donnent à penser qu'à condition que la taille de population soit suffisamment grande et que les échantillons de première et de deuxième phases soient relativement petits, l'estimateur jackknife de variance est plus efficace que l'estimateur usuel de variance, quel que soit l'estimateur de la moyenne de population pris en considération. Pour les petites populations, il pourrait être avantageux d'ajuster les facteurs de correction pour population finie au lieu d'appliquer directement la méthode du jackknife. Il s'agit d'un domaine dans lequel les travaux de recherche pourraient se poursuivre.

Remerciements

Les travaux présentés ont été réalisés pendant que Sarjinder Singh était boursier de recherche postdoctorale à l'Université Carleton. Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont permis d'améliorer considérablement le manuscrit. La présente étude a été financée par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Annexe

Calcul de l'estimateur jackknife (2.1)

Dans cette partie de l'annexe, nous prouvons (2.1) pour l'estimateur jackknife de la moyenne de population sous échantillonnage à deux phases. Premièrement, notons que $\hat{\beta}_1(j) = \hat{\beta}_1 + t_{1j}\, e_{1j}$ et $\hat{\beta}_2(j) = \hat{\beta}_2 + t_{2j}\, e_{2j}$, où $t_{1j} = q_{1j}\, w^o_{1j}\, z_j / (q_{1j}\, w^o_{1j}\, z_j^2 - \sum_{i \in s_1} q_{1i}\, w^o_{1i}\, z_i^2)$, $e_{1j} = x_j - \hat{\beta}_1 z_j$, $t_{2j} = q_{2j} w^o_{2j} x_j / (q_{2j} w^o_{2j} x_j^2 - \sum_{i \in s_1} q_{2i} w^o_{2i} x_i^2)$ et $e_{2j} = y_j - \hat{\beta}_2\, x_j$. Nous avons aussi $\hat{\bar{Z}}{}^o_1(j) = \hat{\bar{Z}}{}^o_1 + h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j)$, $\hat{\bar{X}}{}^o_1(j) = \hat{\bar{X}}{}^o_1 + h_{1j}(\hat{\bar{X}}{}^o_1 - x_j)$, $\hat{\bar{X}}{}^o_2(j) = \hat{\bar{X}}{}^o_2 + h_{2j}(\hat{\bar{X}}{}^o_2 - x_j)$ et $\hat{\bar{Y}}{}^o_2(j) = \hat{\bar{Y}}{}^o_2 + h_{2j}(\hat{\bar{Y}}{}^o_2 - y_j)$, où $h_{1j} = w^o_{1j}/(1 - w^o_{1j})$ et $h_{2j} = w^o_{2j}/(1 - w^o_{2j})$.

En utilisant ces résultats, pour $j \in s_2$, nous avons

$$\begin{aligned}
\hat{\bar{Y}}{}^c(j) = {} & \hat{\bar{Y}}{}^o_2 + \hat{\beta}_2(\hat{\bar{X}}{}^o_1 - \hat{\bar{X}}{}^o_2) + \hat{\beta}_1\, \hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}{}^o_1) \\
& + h_{2j}(\hat{\bar{Y}} - y_j) + t_{2j}\, e_{2j}(\hat{\bar{X}}{}^o_1 - \hat{\bar{X}}{}^o_2) \\
& + \hat{\beta}_2 \{h_{1j}(\hat{\bar{X}}{}^o_1 - x_j) - h_{2j}(\hat{\bar{X}}{}^o_2 - x_j)\} \\
& + t_{1j}\, e_{1j}\, \hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}{}^o_1) - t_{1j}\, e_{1j}\, \hat{\beta}_2\, h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j) \\
& + \hat{\beta}_1\, t_{2j}\, e_{2j}(\bar{Z} - \hat{\bar{Z}}{}^o_1) - t_{2j}\, e_{2j}\, \hat{\beta}_1\, h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j) \\
& - \hat{\beta}_1\, \hat{\beta}_2\, h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j).
\end{aligned}$$

De même, pour $j \in (s_1 - s_2)$, nous avons

$$\begin{aligned}
\hat{\bar{Y}}{}^c(j) = {} & \hat{\bar{Y}}{}^o_2 + \hat{\beta}_2(\hat{\bar{X}}{}^o_1 - \hat{\bar{X}}{}^o_2) + \hat{\beta}_1\, \hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}{}^o_1) \\
& + \hat{\beta}_2\, h_{1j}(\hat{\bar{X}}{}^o_1 - x_j) + t_{1j}\, e_{1j}\, \hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}{}^o_1) \\
& - t_{1j}\, e_{1j}\, \hat{\beta}_2\, h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j) \\
& + \hat{\beta}_1\, \hat{\beta}_2 \{(\bar{Z} - \hat{\bar{Z}}{}^o_1) - h_{1j}(\hat{\bar{Z}}{}^o_1 - z_j)\}.
\end{aligned}$$

Donc, pour $j \in s_2$,

$$\begin{aligned}
\hat{\bar{Y}}{}^c(j) - \hat{\bar{Y}}{}^c = {} & \{\hat{\bar{Y}}{}^o_2(j) - \hat{\bar{Y}}{}^o_2\} - \hat{\beta}_2(j)\{\hat{\bar{X}}{}^o_2(j) - \hat{\bar{X}}{}^o_2\} \\
& - \hat{\beta}_1(j)\hat{\beta}_2(j)\{\hat{\bar{Z}}{}^o_1(j) - \bar{Z}\} \\
& + \hat{\beta}_2[\{\hat{\bar{X}}{}^o_1(j) - \hat{\bar{X}}{}^o_1\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}{}^o_1(j) - \bar{Z}\}] \\
& + \hat{\beta}_2(j)\{\hat{\bar{X}}{}^o_1(j) - \hat{\bar{X}}{}^o_2\} \\
& + \hat{\beta}_2[\{\hat{\bar{X}}{}^o_2(j) - \hat{\bar{X}}{}^o_1(j)\} \\
& \qquad - \hat{\beta}_1(j)\{\bar{Z} - \hat{\bar{Z}}{}^o_1(j)\} - \hat{\beta}_1\{\bar{Z} - \hat{\bar{Z}}{}^o_1\}]
\end{aligned}$$

et pour $j \in (s_1 - s_2)$,

$$\hat{\bar{Y}}{}^c(j) - \hat{\bar{Y}}{}^c = \hat{\beta}_2[\{\hat{\bar{X}}{}^o_1(j) - \hat{\bar{X}}{}^o_1\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}{}^o_1(j) - \bar{Z}\}],$$

ce qui prouve (2.1).

Convergence de l'estimateur de variance (2.2)

Dans cette partie de l'annexe, nous prouvons que l'estimateur $\hat{V}_{\text{JACK}}(\hat{\bar{Y}}{}^c)$ donné par (2.2) est convergent. Premièrement, notons que la variance de l'estimateur $\hat{\bar{Y}}{}^c$ défini par (1.13) peut être approximée par :

$$\begin{aligned}
V(\hat{\bar{Y}}{}^c) \approx {} & V(\hat{\bar{Y}}{}^o_2) + \beta_2^2[V(\hat{\bar{X}}{}^o_1) + V(\hat{\bar{X}}{}^o_2) - 2\mathrm{Cov}(\hat{\bar{X}}{}^o_1, \hat{\bar{X}}{}^o_2)] \\
& + \beta_1^2\, \beta_2^2\, V(\hat{\bar{Z}}{}^o_1) \\
& + 2\beta_2[\mathrm{Cov}(\hat{\bar{Y}}{}^o_2, \hat{\bar{X}}{}^o_1) - \mathrm{Cov}(\hat{\bar{Y}}{}^o_2, \hat{\bar{X}}{}^o_2)] \\
& - 2\beta_1\, \beta_2 \mathrm{Cov}(\hat{\bar{Y}}{}^o_2, \hat{\bar{Z}}{}^o_1) \\
& - 2\beta_1\, \beta_2^2[\mathrm{Cov}(\hat{\bar{X}}{}^o_1, \hat{\bar{Z}}{}^o_1) - \mathrm{Cov}(\hat{\bar{X}}{}^o_2, \hat{\bar{Z}}{}^o_1)].
\end{aligned}$$

Si nous supposons que $\hat{\beta}_1(j) \approx \beta_1, \hat{\beta}_2(j) \approx \beta_2,$ et comme Rao et Sitter (1995), que $\bar{x}_n(j)/\bar{x}_r(j) \approx \bar{x}_n/\bar{x}_r,$ il est assez simple de montrer que

$$\begin{aligned}\sum_{j\in s}[\hat{\hat{Y}}^c(j)-\hat{\hat{Y}}^c]^2 \approx & \sum_{j\in s_2}[\hat{\hat{Y}}_2^o(j)-\hat{\hat{Y}}_2^o]^2+\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\hat{X}}_2^o(j)-\hat{\hat{X}}_2^o]^2 \\ & + 2\hat{\beta}_2\sum_{j\in s_2}[\hat{\hat{Y}}_2^o(j)-\hat{\hat{Y}}_2^o][\hat{\hat{X}}_1^o(j)-\hat{\hat{X}}_1^o] \\ & - 2\hat{\beta}_2\sum_{j\in s_2}[\hat{\hat{Y}}_2^o(j)-\hat{\hat{Y}}_2^o][\hat{\hat{X}}_2^o(j)-\hat{\hat{X}}_2^o] \\ & - 2\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\hat{X}}_1^o(j)-\hat{\hat{X}}_1^o][\hat{\hat{X}}_2^o(j)-\hat{\hat{X}}_2^o] \\ & - 2\hat{\beta}_1\hat{\beta}_2\sum_{j\in s_2}[\hat{\hat{Y}}_2^o(j)-\hat{\hat{Y}}_2^o][\hat{\hat{Z}}_1^o(j)-\hat{\hat{Z}}_1^o] \\ & - 2\hat{\beta}_1\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\hat{X}}_2^o(j)-\hat{\hat{X}}_2^o][\hat{\hat{Z}}_1^o(j)-\hat{\hat{Z}}_1^o] \\ & + \hat{\beta}_2^2\sum_{j\in s}[\hat{\hat{X}}_1^o(j)-\hat{\hat{X}}_1^o]^2 \\ & + \hat{\beta}_1^2\sum_{j\in s}[\hat{\hat{Z}}_1^o(j)-\hat{\hat{Z}}_1^o]^2 \\ & - 2\hat{\beta}_1\hat{\beta}_2^2\sum_{j\in s}[\hat{\hat{X}}_1^o(j)-\hat{\hat{X}}_1^o][\hat{\hat{Z}}_1^o(j)-\hat{\hat{Z}}_1^o].\end{aligned}$$

Puisque les dix termes du deuxième membre de cette équation pour $\sum_{j\in s}[\hat{\hat{Y}}^c(j)-\hat{\hat{Y}}^c]^2$ sont les estimateurs convergents des dix termes analogues de l'équation susmentionnée pour $V(\hat{\hat{Y}}^c),$ nous pouvons conclure que l'estimateur jackknife de variance (2.2) est convergent.

Bibliographie

Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Communincations in Statistics, Theory and Methods*, 26(9), 2247-2254.

Arnab, R., et Singh, S. (2006). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics and Probability Letters*, 76, 513-519.

Berger, Y. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.

Berger, Y., et Skinner, C. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society*, Séries B, 67, 79-89.

Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables*. Thèse de doctorat, Iowa State Unversity, Ames, Iowa, É.-U.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 260-269.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 117-132.

Hidiroglou, M.A., et Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. II, 873-878.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Kim, J.K., Navarro, A. et Fuller, W.A. (2000). Variance estimation for 2000 Census coverage estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515-520.

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.

Kim, J.K., et Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kott, P.S., et Stukel, D. (1997). La méthode du jackknife convient-elle à un échantillon à deux degrés ? *Techniques d'enquête*, 23, 89-98.

Kovar, J., et Chen, E. (1994). Méthode du jackknife pour l'estimation de la variance en présence de données imputés. *Techniques d'enquête*, 20, 47-55.

Raj, D. (1965). On sampling over two occasions with probability proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-60.

Singh, S. (2000). Estimation of variance of regression estimator in two phase sampling. *Calcutta Statistical Association Bulletin*, 50, 49-63.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Srivenkataramana, T., et Tracy, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample surveys. *Biometrika*, 76, 818-821.