N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

ESTIMATION DE LA VARIANCE DANS UN ÉCHANTILLONNAGE À DEUX DEGRÉS

M.A. Hidiroglou et J.N.K. Rao[1]

RÉSUMÉ

On recourt souvent à un échantillonnage à deux degrés pour estimer un total ou une moyenne de population là où le coût unitaire de la collecte de données sur des variables auxiliaires $\boldsymbol{x}$ est bien inférieur au coût correspondant de mesure de la caractéristique d'intérêt. Au premier degré, on prélève un grand échantillon s_1 en appliquant un plan d'échantillonnage précis $p(s_1)$ et observe x pour les unités $i \in s_1$. Étant donné l'échantillon du premier degré s_1, on tire au second degré un échantillon s_2 de s_1 en appliquant un plan d'échantillonnage spécifié $\{p(s_2 \mid s_1)\}$ et observe (y, $\boldsymbol{x}$) pour les unités $i \in s_2$. Dans certains cas, les totaux de population de certains éléments de $\boldsymbol{x}$ peuvent aussi être connus. L'échantillonnage à deux degrés sert à la stratification en seconde étape (Neyman, 1938; Rao, 1973) ou aux deux étapes (Binder et coll., 2000), ainsi qu'à l'estimation par régression (Särndal et coll., 1992, chapitre 9; Hidiroglou et Särndal, 1998). Pour l'estimation de la variance, on fait appel à des estimateurs de Horvitz-Thompson (HT), mais on sait que, dans un échantillonnage à un degré, ceux-ci sont des plus instables et peuvent prendre des valeurs négatives lorsque les probabilités de sélection des unités sont inégales. En revanche, l'estimateur de variance de Sen-Yates-Grundy (SYG) est relativement stable et non négatif pour un certain nombre de plans d'échantillonnage inéquiprobables à tailles d'échantillon fixes. Dans le présent document, nous étendons l'application des estimateurs de variance SYG à l'échantillonnage à deux degrés en posant une taille d'échantillon fixe au premier degré, tout comme une taille d'échantillon fixe au second degré compte tenu de l'échantillon du premier degré. Nous appliquons les nouveaux estimateurs SYG aux plans d'échantillonnage à deux degrés comportant une stratification en seconde étape ou aux deux étapes. Nous concevons également des estimateurs de variance SYG des estimateurs par régression à deux degrés qui utilisent les données auxiliaires de première étape.

MOTS CLÉS : Estimateur d'extension double, estimateur par quotient, estimateur par régression, stratification.

1. INTRODUCTION

On recourt souvent à un échantillonnage à deux phases pour estimer un total ou une moyenne de population là où le coût unitaire de la collecte de données auxiliaires $\boldsymbol{x}$ est bien inférieur au coût correspondant de mesure de caractéristiques d'intérêt y. Dans un tel plan d'échantillonnage à deux phases, on prélève d'abord un grand échantillon s_1 de taille n_1 sur l'univers U selon des probabilités $\{p(s_1)\}$ et on observe $\boldsymbol{x}$ pour les unités d'échantillon $i \in s_1$. Étant donné l'échantillon s_1 à la première phase, on tire de celui-ci l'échantillon de seconde phase s_2 en application d'un plan d'échantillonnage spécifié et selon des probabilités conditionnelles $\{p(s_2|s_1)\}$ et on observe $(y, \boldsymbol{x})$ pour les unités $i \in s_2$. Dans certains cas, les totaux de population de certains éléments $\boldsymbol{x}_1$ de $\boldsymbol{x}$ peuvent aussi être connus.

Neyman (1938) a initialement proposé un échantillonnage à deux phases pour la stratification. Ainsi, on stratifie l'échantillon s_1 à la première phase prélevé par échantillonnage aléatoire simple en se reportant à une variable auxiliaire scalaire x observée pour les unités dans le contexte d'un tel échantillon à la première phase s_1 de taille n_1, $i \in s_1$: $s_1 = \bigcup_g s_{1g}$, où s_{1g} est l'échantillon à la première phase de taille aléatoire n_{1g} de la strate g. À la seconde phase, on prélève des échantillons aléatoires simples s_{2g} de taille fixe n_{2g} sur les échantillons à la première

[1] Mike Hidiroglou, directeur, Survey Methods Division, pièce D141, Methodology and Statistical Development Directorate, Cardiff Road, Newport, NP9 1XG, Royaume-Uni; J.N.K. Rao, École de mathématiques et de statistique, Ottawa (Ontario) K1S 5B6.

phase s_{1g} de taille aléatoire n_{1g}, $\sum_g n_{1g} = n_1$. En seconde étape, il y a tirage indépendant d'échantillons aléatoires simples s_{2g} de taille fixe n_{2g} sur les échantillons au premier degré s_{1g}. L'hypothèse de tailles fixes n_{2g} ne s'accorde toutefois pas avec la procédure d'échantillonnage, puisque n_{2g} a pour borne supérieure la variable aléatoire n_{1g} qui varie de 0 à $\min(n_1, N_g)$, où N_g est le nombre d'unités de population de la strate *g*. Rao (1973) a proposé une autre répartition d'échantillon qui nous évite les difficultés que présente la méthode de répartition des unités de seconde phase de Neyman. Dans cette nouvelle méthode, on sélectionne une fraction fixe ν_g des unités d'échantillon à la première phase, c'est-à-dire $n_{2g} = \nu_g n_{1g}$, $0 < \nu_g \leq 1$. Cochran (1977, chapitre 12) a étudié une estimation par quotient et régression pour le cas d'espèce d'un échantillonnage aléatoire simple aux deux phases.

Plus récemment, Särndal, Swensson et Wretman (1992, chapitre 9) ont admis un plan d'échantillonnage arbitraire à l'une et l'autre des étapes. Soit π_{1i} et π_{1ij} les probabilités d'inclusion de premier et de second ordre pour l'échantillon s_1 au premier degré et $\pi_{2i|s_1}$ et $\pi_{2ij|s_1}$ les probabilités correspondantes pour l'échantillon s_2 au second degré compte tenu de s_1. Un estimateur sans biais du total de population $Y = \sum_U y_i$ nous est donné par

$$\hat{Y}_2 = \sum_{s_2} \frac{y_i}{\pi_{1i}\,\pi_{2i|s}} = \sum_{s_2} \frac{\dot{y}_i}{\pi_{2i|s_1}} \ , \tag{1.1}$$

où $\dot{y}_i = y_i / \pi_{li}$ et $\sum_a$ désignent la sommation sur les unités $i \in a$. C'est ce qu'on appelle un estimateur de double extension par analogie avec l'estimateur d'extension (de Horvitz-Thompson (HT)) dans l'échantillonnage à une phase. Särndal et coll. (1992) ont établi un estimateur sans biais de la variance de $\hat{Y}_2$ sous la forme suivante :

$$\mathrm{v}_{\mathrm{HT}}\left(\hat{Y}_2\right) = \sum\sum_{s_2} \frac{\Delta_{1ij}}{\pi^*_{\mathrm{ij}}}\, \dot{\mathrm{y}}_{\mathrm{i}} \dot{\mathrm{y}}_{\mathrm{j}} + \sum\sum_{s_2} \frac{\Delta_{2ij|s_1}}{\pi_{2\mathrm{ij}|s_1}} \frac{\dot{\mathrm{y}}_{\mathrm{i}} \dot{\mathrm{y}}_{\mathrm{j}}}{\pi_{2\mathrm{i}|s_1}\pi_{2\mathrm{j}|s_1}} \ , \tag{1.2}$$

où $\pi^*_{ij} = \pi_{1ij}\,\pi_{2ij|s_1}$, $\Delta_{1ij} = \pi_{1\mathrm{ij}} - \pi_{1\mathrm{i}}\pi_{1\mathrm{j}}$ et $\Delta_{2ij|s_1} = \pi_{2\mathrm{ij}|s_1} - \pi_{2\mathrm{i}|s_1}\pi_{2\mathrm{j}|s_1}$.

Les formules (1.1) et (1.2) peuvent recevoir une forme plus compacte :

$$\hat{Y}_2 = \sum_{s_2} \frac{y_i}{\pi^*_i} \tag{1.3}$$

et

$$\mathrm{v}_{\mathrm{HT}}\left(\hat{Y}_2\right) = \sum\sum_{s_2} \frac{\Delta^*_{ij}}{\pi^*_{ij}}\, y_i y_j \ , \tag{1.4}$$

où $\pi^*_i = \pi_{1i}\,\pi_{2i|s_1}$ et $\Delta^*_{ij} = \pi^*_{ij} - \pi^*_i \pi^*_j$ (Särndal et coll., 1992, p. 347).

L'estimateur de variance (1.4) (ou son équivalent (1.2)) est de la même forme que l'estimateur de variance HT dans un échantillonnage à une phase. On sait que, dans le cas d'un plan d'échantillonnage général à une phase où les probabilités d'inclusion sont inégales, ce dernier estimateur est hautement instable et peut prendre des valeurs négatives (voir Rao et Singh, 1973, et Cochran, 1977, chapitre 10a). En revanche, l'estimateur de variance de Sen-Yates-Grundy (SYG) est relativement plus stable. Il est donc utile de concevoir des estimateurs de variance SYG pour l'échantillonnage à deux phases.

Särndal et coll. (1992) ont élargi l'estimateur sans biais (1.1) de manière à intégrer des données auxiliaires $\boldsymbol{x}$ recueillies à la première phase, et ce, à l'aide d'estimateurs par régression généralisée (GREG). Ils ont aussi obtenu un estimateur de variance en linéarisation de Taylor sous la forme (1.2). Les estimateurs GREG étalonnent les estimateurs à la première phase de totaux $\boldsymbol{x}$. En d'autres termes, l'estimateur GREG de *Y* est de la forme $\sum_{s_2} w_i\, y_i$ avec des valeurs de pondération w_i respectant la relation $\sum_{s_2} w_i\, \boldsymbol{x}_i = \sum_{s_1} d_{1i}\, \boldsymbol{x}_i$. Hidiroglou et Särndal (1998) ont proposé des estimateurs GREG en étalonnage sur des estimateurs à la première phase, lesquels sont étalonnés par les

totaux connus $\boldsymbol{x}_1$. En d'autres termes, $\sum_{s_2} w_i\ \boldsymbol{x}_i = \sum_{s_1} w_{1i}\ \boldsymbol{x}_i$ et $\sum_{s_2} w_{1i}\ \boldsymbol{x}_{1i} = \sum_U\ \boldsymbol{x}_{1i}$. Ces auteurs ont enfin obtenu un estimateur de variance en linéarisation de la forme (1.2) (voir aussi Estevao et Särndal, 2002).

Binder, Babyak, Brodeur, Hidiroglou et Jocelyn (2000) ont simplifié l'estimateur de variance HT (1.2) là où un échantillon aléatoire simple stratifié à la première phase est restratifié de seconde phase au moyen de données auxiliaires $\boldsymbol{x}$ recueillies en première étape et où on tire sans remise des échantillons aléatoires simples de strates de seconde phase en vue de l'observation de *y*. Kott et Stukel (1997) ont étudié un échantillonnage semblable à deux phases sauf que, dans ce cas, l'échantillon à la première phase est un échantillon en grappes avec stratification et remise. Ils ont proposé un estimateur d'extension repondéré qui diffère en général de l'estimateur d'extension double, ce qui leur a donné un estimateur de variance jackknife. Ils ont aussi démontré que la méthode jackknife proposée n'est pas efficace pour l'estimateur d'extension double. Lee et Kim (2002) ont également étudié un plan d'échantillonnage à deux phases qui y ressemble sauf que l'échantillon à la première phase est un échantillon en grappes stratifié, dont les grappes sont constituées par échantillonnage aléatoire simple et sans remise. Les estimateurs d'extension double et d'extension-repondération sont d'une conception identique. Lee et Kim (2002) ont conçu un estimateur de variance jackknife qui tient compte des taux d'échantillonnage aux deux phases.

Bien que toutes les données auxiliaires disponibles servent à l'estimation du total, l'estimateur de variance repose habituellement sur des données d'échantillon de seconde phase. On peut raisonnablement se demander si les données auxiliaires disponibles pour les éléments exclus de l'échantillon de seconde phase pourraient servir plus largement à l'estimation de variance et, si oui, s'il conviendrait de le faire. Dorfman (1994), Rao et Sitter (1995), Sitter (1997) et Axelson (1998) ont proposé de mettre les données auxiliaires à la première phase au service de l'estimation de variance.

Notre document est ainsi structuré : à la section 2, nous élaborons un estimateur de variance SYG de $\hat{Y}_2$; à la section 3, nous appliquons ce résultat au plan d'échantillonnage à deux degrés de Binder et coll. (2000) et obtenons un estimateur de variance non négatif qui diffère de l'estimateur de variance HT de ces auteurs; à la section 4, nous obtenons un estimateur SYG pour l'estimateur par régression à deux phases du total, où on exploite les données auxiliaires de première étape.

2. ESTIMATEUR DE VARIANCE DU TYPE SYG

L'estimateur de $\hat{Y}_2$ est conditionnellement sans biais pour l'estimateur au premier degré $\hat{Y}_1 = \sum_{s_1} \dot{y}_i$ compte tenu de l'échantillon s_1 à la première phase, où $\dot{y}_i = y_i/\pi_{1i} = d_{1i}y_i$, c'est-à-dire $E\left(\hat{Y}_2 \mid s_1\right) = \hat{Y}_1$. Il est donc inconditionnellement sans biais pour le total $Y = \sum_U y_i$. La variance de $\hat{Y}_2$ nous est donnée par

$$\begin{aligned} V\left(\hat{Y}_2\right) &= E\left[V\left(\hat{Y}_2 \mid s_1\right)\right] + V\left[E\left(\hat{Y}_2 \mid s_1\right)\right] \\ &= E\left[V\left(\hat{Y}_2 \mid s_1\right)\right] + V\left(\hat{Y}_1\right). \end{aligned} \tag{2.1}$$

Nous pouvons estimer la variance conditionnelle $V\left(\hat{Y}_2 \mid s_1\right)$ en (2.1) en utilisant l'estimateur de variance SYG à condition que la taille d'échantillon de seconde phase soit fixe pour un s_1 donné (voir Rao, 1979). L'estimateur de variance SYG est tiré de

$$\mathrm{v}\left(\hat{Y}_2 \mid s_1\right) = \sum_{i<j\in s_2}\sum \frac{\left(\pi_{2i|s_1}\pi_{2j|s_1} - \pi_{2ij|s_1}\right)}{\pi_{2ij|s_1}}\left(\frac{\dot{y}_i}{\pi_{2i|s_1}} - \frac{\dot{y}_j}{\pi_{2j|s_1}}\right)^2. \tag{2.2}$$

Il est conditionnellement sans biais pour $V\left(\hat{Y}_2 \mid s_1\right)$ et l'est donc inconditionnellement pour $E\left[V\left(\hat{Y}_2 \mid s_1\right)\right]$.

Si nous passons au second terme de (2.1), nous obtenons

$$V\left(\hat{Y}_1\right)=\sum_{i<j\in U}\sum \left(\pi_{1i}\,\pi_{1j}-\pi_{1ij}\right)\left(\dot{y}_i-\dot{y}_j\right)^2 \tag{2.3}$$

à condition que la taille d'échantillon à la première phase soit fixe. Si les y_i étaient connus pour tous les $i\in s_1$, l'estimateur de variance SYG de $V\left(\hat{Y}_1\right)$ serait donné par

$$\mathrm{v}\left(\hat{Y}_1\right)=\sum_{i<j\in s_1}\sum \frac{\pi_{1i}\,\pi_{1j}-\pi_{1ij}}{\pi_{1ij}}\left(\dot{y}_i-\dot{y}_j\right)^2 . \tag{2.4}$$

Mais comme y_i est connu seulement pour $i\in s_2$, nous estimons (2.4) à l'aide de l'échantillon s_2 de seconde phase et obtenons

$$\mathrm{v}_2\left(\hat{Y}_1\right)=\sum_{i<j\in s_2}\sum \frac{\pi_{1i}\,\pi_{1j}-\pi_{1ij}}{\pi_{1ij}\pi_{2ij|s_1}}\left(\dot{y}_i-\dot{y}_j\right)^2 . \tag{2.5}$$

L'estimateur de variance (2.5) est sans biais pour $V\left(\hat{Y}_1\right)$. Il s'ensuit de (2.1) qu'un estimateur sans biais du type SYG pour $V\left(\hat{Y}_1\right)$ est donné par

$$\mathrm{v}_{\mathrm{SYG}}\left(\hat{Y}_2\right)=\mathrm{v}\left(\hat{Y}_2\middle|s_1\right)+\mathrm{v}_2\left(\hat{Y}_1\right), \tag{2.6}$$

où $\mathrm{v}\left(\hat{Y}_2\middle|s_1\right)$ et $\mathrm{v}_2\left(\hat{Y}_1\right)$ sont respectivement donnés en (2.2) et (2.5).

Par analogie avec l'estimateur SYG à une phase, Chaudhuri (1994) a présenté un estimateur SYG de $\mathrm{V}\left(\hat{Y}_2\right)$, mais sa formule pour $\mathrm{v}\left(\hat{Y}_2\middle|s_1\right)$ paraît peu juste, puisqu'elle emploie $\left(\dot{y}_i-\dot{y}_j\right)^2$ au lieu du bon terme $\left(\dot{y}_i/\pi_{2i|s_1}-\dot{y}_j/\pi_{2j|s_1}\right)^2$ en (2.2).

L'estimateur de variance du type HT, $\mathrm{v}_{\mathrm{HT}}\left(\hat{Y}_2\right)$, vaut pour les plans d'échantillonnage à taille fixe ou non contrairement à l'estimateur du type SYG (2.6). Il reste que l'estimateur SYG demeure valide pour un grand nombre de plans d'échantillonnage à deux phases qui sont d'un usage courant et, par analogie avec les estimateurs à une phase, il devrait être plus stable que l'estimateur HT et demeurer non négatif pour plusieurs plans bien connus d'échantillonnage avec probabilités proportionnelles à la taille (PPT). Rao et Singh (1973) ont livré d'abondantes indications empiriques au sujet de la supériorité d'un estimateur SYG par rapport à un estimateur HT en cas d'échantillonnage à une phase.

3. ÉCHANTILLONNAGE À DEUX PHASES POUR LA STRATIFICATION

3.1 Cadre général

Dans cette section, nous évaluons l'estimateur de variance SYG (2.6) dans le cas d'un échantillonnage à deux phases pour la stratification. À la première phase, on prélève un grand échantillon s_1 de taille n_1 selon un plan d'échantillonnage spécifié où les probabilités marginales d'inclusion du 1er et 2e ordre sont respectivement π_{1i} et π_{1ij}. Au moyen de données auxiliaires recueillies à peu de frais sur les unités $i\in s_1$, on divise l'échantillon s_1 à la première phase en $G(s_1)$ strates, ce qui est désigné par $s_{1g}\ (g=1,\ldots,G(s_1))$ et comporte m_{1g} éléments dans la strate g, ($\sum_g m_{1g}=n_1$). À la seconde phase, on tire indépendamment un échantillon probabiliste s_{2g} de taille m_{2g}

de s_{1g} sur g et observe la caractéristique d'intérêt y. Il convient de noter que le nombre de strates $G(s_1)$ de seconde phase et les tailles d'échantillon m_{1g} et m_{2g} dépendent de s_1, bien que $G(s_1)$ puisse être prédéterminé, c'est-à-dire que $G(s_1) \approx G$. Pour simplifier la notation, nous supprimons la dépendance à l'égard de s_1.

Pour $\pi_{2ij|s_1} = \pi_{2i|s_1}\pi_{2j|s_1}$ si $i \in s_{1g}$ et $j \in s_{1\ell}$ $(g \neq \ell)$, $\mathrm{v}\left(\hat{Y}_2 \mid s_1\right)$ se ramène à

$$\mathrm{v}\left(\hat{Y}_2 \mid s_1\right) = \sum_{g=1}^{G} \sum_{i<j \in s_{2g}} \sum \Delta_{2ij|s_{1g}} \left(\frac{\dot{y}_i}{\pi_{2i|s_{1g}}} - \frac{\dot{y}_j}{\pi_{2j|s_{1g}}} \right)^2, \tag{3.1}$$

où

$$\Delta_{2ij|s_{1g}} = \frac{\pi_{2i|s_{1g}}\pi_{2j|s_{1g}} - \pi_{2ij|s_{1g}}}{\pi_{2ij|s_{1g}}}. \tag{3.2}$$

L'expression (3.1) vaut pour un échantillonnage général à la seconde phase à l'intérieur des strates pour des probabilités conditionnelles d'inclusion $\pi_{2i|s_{1g}}$ et $\pi_{2ij|s_{1g}}$, à condition que $\sum_{s_1} \pi_{2i|s_{1g}}$ soit fixe pour un s_1 donné. Dans le cas d'espèce d'un échantillonnage aléatoire simple dans des strates de seconde phase, nous avons $\pi_{2i|s_{1g}} = m_{2g}/m_{1g}$ et $\pi_{2ij|s_{1g}} = m_{2g}(m_{2g}-1)/\left[m_{1g}(m_{1g}-1)\right]$ et (3.1) se réduit à

$$\mathrm{v}\left(\hat{Y}_2 \mid s_1\right) = \sum_{g=1}^{G} m_{1g}^2 \left(\frac{1-f_{2g}}{m_{2g}} \right) \left(\frac{1}{m_{2g}-1} \right) \sum\sum_{i<j \in s_{2g}} \left(\dot{y}_i - \dot{y}_j \right)^2, \tag{3.3}$$

où $f_{2g} = m_{2g}/m_{1g}$.

Par l'identité de Lagrange

$$\sum_{i<j=1}^{m} \sum \left(z_i - z_j \right)^2 = m \sum_{i=1}^{m} \left(z_i - \bar{z} \right)^2, \tag{3.4}$$

l'expression (3.3) se ramène à

$$\mathrm{v}\left(\hat{Y}_2 \mid s_1\right) = \sum_{g=1}^{G} \left(\frac{1-f_{2g}}{m_{2g}} \right) m_{1g}^2 \left(\frac{1}{m_{2g}-1} \right) \hat{S}_{2g\dot{y}}^2, \tag{3.5}$$

où $\hat{S}_{2g\dot{y}}^2$ est le carré en moyenne d'échantillon des valeurs pondérées à la première phase $\dot{y}_i = y_i/\pi_i$ pour $i \in s_{2g}$. Le second élément de l'estimateur de variance HT (1.2) pour un échantillonnage aléatoire simple dans des strates à la seconde phase concorde avec (3.5); voir la formule (9.4.8) de Särndal et coll. (1992), p. 352. L'élément $\mathrm{v}_2\left(\hat{Y}_1\right)$ de l'estimateur de variance SYG (2.6) dans un tel échantillonnage aléatoire simple se réduit à

$$\begin{aligned} \mathrm{v}_2\left(\hat{Y}_1\right) &= \sum_{g=1}^{G} \frac{m_{1g}(m_{1g}-1)}{m_{2g}(m_{2g}-1)} \sum\sum_{i<j \in s_{2g}} \Delta_{1ij} \left(\dot{y}_i - \dot{y}_j \right)^2 + \sum_{g<\ell=1}^{G} \sum \frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}} \sum_{i \in s_{2g}} \sum_{j \in s_{2\ell}} \Delta_{1ij} \left(\dot{y}_i - \dot{y}_j \right)^2 \\ &=: \mathrm{v}_2^{(1)}\left(\hat{Y}_1\right) + \mathrm{v}_2^{(2)}\left(\hat{Y}_1\right), \end{aligned} \tag{3.6}$$

où

$$\Delta_{1ij} = \frac{\pi_{1i}\pi_{1j} - \pi_{1ij}}{\pi_{1ij}}. \tag{3.7}$$

Une nouvelle réduction est impossible pour les plans généraux d'échantillonnage à la première phase où les probabilités d'inclusion sont π_{1i} et π_{1ij}.

Exemple 3.1 : Si on prélève l'échantillon à la première phase s_1 de taille n_1 par échantillonnage aléatoire simple sur une population *U* de taille *N*, $\pi_{1i} = n_1/N$, $\pi_{1ij} = n_1(n_1-1)/[N(N-1)]$ et $\Delta_{1ij} = (1-f_1)/(n_1-1)$. L'estimateur de seconde phase de $\hat{Y}_2$ se ramène à $N\sum_g w_{1g}\,\bar{y}_{2g}$, où $\bar{y}_{2g} = m_2^{-1}\sum_{s_{2g}} y_i$. Par l'identité de Lagrange (3.4) et les valeurs précitées de π_{1i} et Δ_{1ij}, le premier terme du côté droit de (3.6) se réduit à

$$\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right) = \frac{N^2(1-f_1)}{n_1}\sum_{g=1}^{G} w_{1g}\,\frac{(m_{1g}-1)}{n_1-1}\,\hat{S}_{2gy}^2, \tag{3.8}$$

où $\hat{S}_{2gy}^2$ est le carré en moyenne d'échantillon de y_i pour $i \in s_2$. Ajoutons que le second terme du côté droit de (3.6) se réduit à

$$\begin{aligned}\mathrm{v}_2^{(2)}\left(\hat{Y}_1\right) &= \frac{N^2(1-f_1)}{n_1^2(n_1-1)}\left[\frac{1}{2}\sum_{g=1}^{G}\sum_{\ell=1}^{G}\frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}}\sum_{i\in s_{2g}}\sum_{i\in s_{2\ell}}(y_i-y_j)^2 - \sum_{g=1}^{G} m_{1g}^2\,\frac{m_{2g}-1}{m_{2g}}\hat{S}_{2gy}^2\right] \\ &= \frac{N^2(1-f_1)}{n_1}\left[\sum_{g=1}^{G}\frac{(n_1-m_{1g})}{n_1-1}\,w_{1g}(m_{2g}-1)\hat{S}_{2gy}^2 + \frac{n_1}{n_1-1}\sum_{g=1}^{G} w_{1g}\left(\bar{y}_{2g}-\bar{y}_{2a}\right)^2\right],\end{aligned} \tag{3.9}$$

où $\bar{y}_{2a} = \hat{Y}_2/N = \sum_{g=1}^{G} w_{1g}\bar{y}_{2g}$. La somme de (3.8) et (3.9) donne $\mathrm{v}_2\left(\hat{Y}_1\right)$ sous la forme suivante :

$$\mathrm{v}_2\left(\hat{Y}_1\right) = \sum_{g=1}^{G}(1-\delta_g)\,w_{1g}\hat{S}_{2gy}^2 + \frac{n_1}{n_1-1}\sum_{g=1}^{G} w_{1g}\left(\bar{y}_{2g}-\bar{y}_{2a}\right)^2, \tag{3.10}$$

où $\delta_g = \dfrac{1}{m_{2g}}\dfrac{n_1-m_{1g}}{n_1-1}$.

Särndal et coll. (1992) ont simplifié le premier élément du côté droit de l'estimateur de variance HT (1.2) dans le cas particulier d'un échantillonnage aléatoire simple à la première phase (sans donner de détails) pour ainsi obtenir leur (9.4.12) à la p. 353. Cette formule s'accorde avec notre $\mathrm{v}_2\left(\hat{Y}_1\right)$ en (3.10).

3.2 Échantillonnage stratifié à deux phases

Posons que la population *U* se divise en *H* strates U_h et qu'il y a N_h éléments dans la h^{e} strate ($\sum_{h=1}^{H} N_h = N$). À la première phase, nous tirons des échantillons simples s_{1h} indépendamment des strates U_h à la première phase et observons une variable scalaire *x* pour $i \in s_{1h}$, $h = 1,...,H$, où la taille de s_{1h} est n_{1h} $\left(\sum_{h=1}^{H} n_{1h} = n_1\right)$. Nous redivisons l'échantillon de première phase $s_1 = \bigcup_{h=1}^{H} s_{1h}$ en *G* strates $\tilde{s}_{1g}$ de taille m_{1g} $\left(\sum_{h=1}^{G} m_{1g} = n_1\right)$ au moyen de la variable auxiliaire *x* observée en première étape. Nous tirons ensuite indépendamment des échantillons aléatoires simples s_{2g} de taille m_{2g} des strates de seconde phase $\tilde{s}_{1g}$ *(g=1, …, G).*

Pour le plan d'échantillonnage mentionné, $\pi_{1i} = \frac{n_{1h}}{N_h}$ si $i \in s_{1h}$ et, pour $i \neq j$,

$$\pi_{1ij} = \begin{cases} \frac{n_{1h}(n_{1h}-1)}{N_h(N_h-1)} \text{if i} \neq j \in s_{1h} \\ \frac{n_{1h}}{N_h}\frac{n_{1k}}{N_k} \text{ if i} \in s_{1h};\ j \in s_{1k};\ h \neq k. \end{cases} \tag{3.11}$$

L'estimateur à deux phases $\hat{Y}_2$ se ramène à

$$\hat{Y}_2 = \sum_{h=1}^{H} \frac{N_h}{n_{1h}} \sum_{g=1}^{G} \frac{m_{1g}}{m_{2g}} \sum_{i \in s_{2gh}} y_i , \tag{3.12}$$

où $s_{2gh} = s_{1h} \bigcap s_{2g}$. À noter que certains des s_{2gh} peuvent être vides, auquel cas nous fixons $\sum_{i \in s_{2gh}} y_i$ à 0 en (3.12).

Si nous passons à l'estimation de la variance, l'élément $\mathrm{v}\left(\hat{Y}_2 \mid s_1\right)$ est donné par (3.5) avec $\dot{y}_i = y_i\left(N_h / n_{1h}\right)$ si $i \in s_{1h}$. Pour évaluer $\mathrm{v}_2\left(\hat{Y}_1\right)$ en (3.6), nous avons besoin des valeurs Δ_{1ij}. Par (3.11), nous obtenons

$$\Delta_{1ij} = \begin{cases} \frac{1-f_{1h}}{n_{1h}-1} \ si \ \ i,j \in \tilde{s}_{2h} \\ =0 \quad si\ i \in \tilde{s}_{2h},\ j \in \tilde{s}_{2k}, h \neq k \ , \end{cases} \tag{3.13}$$

où $\tilde{s}_{2h} = \bigcup_g s_{2gh}$ et $f_{1h} = n_{1h} / N_h$. Si nous substituons les valeurs mentionnées de Δ_{1ij} en (3.6), le premier élément $\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right)$ se réduit à

$$\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right) = \sum_{g=1}^{G} \frac{m_{1g}(m_{1g}-1)}{m_{2g}(m_{2g}-1)} \sum_{h \in A_g} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} \sum_{i<j \in s_{2gh}} \sum \left(y_i - y_j\right)^2 \ , \tag{3.14}$$

où A_g est l'ensemble de strates h de première phase avec au moins deux unités dans s_{2gh}; le reste des strates de première phase ne contribue pas à $\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right)$. Par l'identité de Lagrange (3.4), l'expression (3.14) se ramène à

$$\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right) = \sum_{g=1}^{G} \frac{m_{1g}(m_{1g}-1)}{m_{2g}(m_{2g}-1)} \sum_{h \in A_g} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} m_{2gh}\left(m_{2gh}-1\right)\hat{S}_{2ghy}^2 , \tag{3.15}$$

où m_{2gh} est le nombre d'unités dans s_{2gh} et où $\hat{S}_{2ghy}^2$ est le carré en moyenne d'échantillon des valeurs y_i pour $i \in s_{2gh}$.

Nous pouvons alors exprimer $\mathrm{v}_2^{(2)}\left(\hat{Y}_1\right)$ sous la forme suivante :

$$\mathrm{v}_2^{(2)}\left(\hat{Y}_1\right) = \sum_{g<\ell}^{G} \sum \frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}} \sum_{h \in U_{2g\ell}} \left(\frac{N_h}{n_{1h}}\right)^2 \frac{1-f_{1h}}{n_{1h}-1} \left\{ \sum_{i \in s_{2gh}} \sum_{j \in s_{2\ell h}} \left(y_i - y_j\right)^2 \right\} , \tag{3.16}$$

où $U_{2g\ell}$ est l'ensemble de strates *h* de première phase avec au moins une unité tant dans s_{2gh} que dans $s_{2h\ell}$. Il est impossible de simplifier (3.16) davantage à moins que $m_{2gh} \geq 2$ pour tout (*gh*). L'estimateur de variance du type SYG, $\mathrm{v}\left(\hat{Y}_2\right)$, nous est maintenant donné par la somme de (3.5), (3.15) et (3.16), et il est toujours non négatif.

Considérons maintenant le cas d'espèce $m_{2gh} \geq 2$ pour tout (*gh*), auquel cas $\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right)$ nous est donné par (3.15) avec $\sum_{A_g}$ changé en $\sum_{h=1}^{H}$. Nous pouvons écrire $\mathrm{v}_2^{(2)}\left(\hat{Y}_1\right)$ sous la forme suivante :

$$\begin{aligned} v_2^{(2)}\left(\hat{Y}_1\right) &= \frac{1}{2}\sum_{g=1}^{G}\sum_{\ell=1}^{G}\frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}}\sum_{i\in s_{2gh}}\sum_{j\in s_{2\ell h}}\left(\dot{y}_i-\dot{y}_j\right)^2-\sum_{g=1}^{G}\left(\frac{m_{1g}}{m_{2g}}\right)^2\Delta_{lij}\sum_{i<j\in s_{2g}}\sum\left(\dot{y}_i-\dot{y}_j\right)^2 \\ &= I-II. \end{aligned} \tag{3.17}$$

À la suite des étapes de l'obtention de (3.15), nous avons

$$\text{-II} = -\sum_{g=1}^{G}\left(\frac{m_{1g}}{m_{2g}}\right)^2\sum_{h=1}^{H}\left(\frac{N_h}{n_{1h}}\right)^2\frac{1-f_{1h}}{n_{1h}-1}m_{2gh}\left(m_{2gh}-1\right)\hat{S}_{2ghy}^2 . \tag{3.18}$$

En combinant (3.15) et (3.18), nous obtenons

$$\mathrm{v}_2^{(1)}\left(\hat{Y}_1\right)\text{-II} = \sum_{g=1}^{G}\left(\frac{m_{1g}}{m_{2g}}\right)^2\frac{1-f_{2g}}{m_{2g}-1}\sum_{h=1}^{H}\left(\frac{N_h}{n_{1h}}\right)^2\frac{1-f_{1h}}{n_{1h}-1}m_{2gh}\left(m_{2gh}-1\right)\hat{S}_{2ghy}^2 . \tag{3.19}$$

Pour le terme *I* en (3.17), nous pouvons écrire

$$\begin{aligned} I &= \sum_{h=1}^{H}\frac{N_h^2}{n_{1h}^2}\frac{1-f_{1h}}{n_{1h}-1}\left\{\frac{1}{2}\sum_{g=1}^{G}\sum_{\ell=1}^{G}\frac{m_{1g}m_{1\ell}}{m_{2g}m_{2\ell}}\sum_{i\in s_{2gh}}\sum_{j\in s_{2\ell h}}\left(\dot{y}_i-\dot{y}_j\right)^2\right\} \\ &= \sum_{h=1}^{H}\frac{N_h^2}{n_{1h}^2}\frac{1-f_{1h}}{n_{1h}-1}\hat{n}_{1h}\sum_{g=1}^{G}\frac{m_{1g}}{m_{2g}}\sum_{i=1}^{m_{2gh}}\left(y_i-\bar{y}_{ah}\right)^2 , \end{aligned} \tag{3.20}$$

où

$$\hat{n}_{1h} = \sum_{g=1}^{G}\frac{m_{1g}}{m_{2g}}m_{2gh} \quad \text{et} \quad \bar{y}_{ah} = \hat{n}_{1h}^{-1}\left(\sum_{g=1}^{G}\frac{m_{1g}}{m_{2g}}\sum_{k=1}^{m_{2gh}}y_k\right). \tag{3.21}$$

L'estimateur de variance $\mathrm{v}\left(\hat{Y}_2\right)$ vient maintenant de la somme de (3.50), (3.19) et (3.20).

L'estimateur de variance HT de Binder et coll. (2000), qui est tiré de (1.2), diffère de notre $\mathrm{v}\left(\hat{Y}_2\right)$, mais $\mathrm{v}\left(\hat{Y}_2 \mid s_1\right)$, qui est tiré de (3.5), est identique à la formule de ces auteurs. L'expression de Binder et coll. qui correspond à (3.19) est donnée par

$$\sum_{g=1}^{G}\left(\frac{m_{1g}}{m_{2g}}\right)^2\frac{1-f_{2g}}{m_{2g}-1}\sum_{h=1}^{L}\left(\frac{N_h}{n_{1h}}\right)^2\frac{1-f_{1h}}{n_{1h}-1}m_{2g}\left\{\left(m_{2gh}-1\right)S_{2ghy}^2+m_{2gh}\left(1-\frac{m_{2gh}}{m_{2g}}\right)\bar{y}_{2gh}^2\right\}, \tag{3.19)*}$$

où $\bar{y}_{2gh}$ est la moyenne de *y* pour s_{2gh}. L'expression de Binder et coll. qui correspond à (3.20) est donnée par

$$\sum_{h=1}^{H} \frac{N_h^2}{n_{1h}^2} \frac{1-f_{1h}}{n_{1h}-1} \left[\sum_{g=1}^{G} \left(\frac{m_{1g}^2}{m_{2g}} \right) \frac{1-f_{2g}}{m_{2g}-1} \frac{m_{2gh}}{m_{2g}} \sum_{i=1}^{m_{2gh}} \left(y_i - \bar{y}_{ah} \right)^2 + \hat{n}_{1h} \left(\frac{\hat{n}_{1h}}{n_{1h}} - 1 \right) \bar{y}_{ah}^2 \right]. \qquad (3.20)^*$$

L'estimateur de variance de Binder et coll. (2000) est alors la somme de (3.5), (3.19)* et (3.20)*. À noter que le terme $(\hat{n}_{1h} / n_{1h}) - 1$ peut être soit positif, soit négatif.

4. ESTIMATEUR DE VARIANCE DE L'ESTIMATEUR PAR RÉGRESSION AVEC LES DONNÉES AUXILIAIRES DE LA PREMIÈRE PHASE

Dans un échantillonnage à deux phases, des données auxiliaires peuvent être puisées à diverses sources. Considérons le cas où de telles données sont disponibles dans la base de sondage *U* et dans l'échantillon s_1 à la première phase. Les données auxiliaires de *U* sont désignées par $\boldsymbol{x}_{1i}$ et les données auxiliaires de l'échantillon s_1, par $\boldsymbol{x}_{2i}$. Le vecteur de données auxiliaires $\boldsymbol{x}_i = (\boldsymbol{x}'_{1i}, \boldsymbol{x}'_{2i})'$ contient les données tant de *U* que de s_1. On recueille des données $(y_i, \boldsymbol{x}'_i)$ par l'échantillon de seconde phase s_2. L'estimateur par régression $\hat{Y}_{2,REG}$ du total *Y* avec des données auxiliaires des deux phases est donné par

$$\hat{Y}_{2,REG} = \hat{Y}_2 + \left(\boldsymbol{X}_{1,1} - \hat{\boldsymbol{X}}_{1,1} \right)' \hat{\boldsymbol{B}}_1 + \left(\hat{\boldsymbol{X}}_1 - \hat{\boldsymbol{X}}_2 \right)' \hat{\boldsymbol{B}}_2 . \qquad (4.1)$$

Dans (4.1), $\hat{Y}_2 = \sum_{s_2} y_i / \pi_i^*$, $\boldsymbol{X}_{1,1} = \sum_U \boldsymbol{x}_{1i}$ est la somme des données auxiliaires $\boldsymbol{x}_{1i}$ disponibles dans la base de sondage *U;* $\hat{\boldsymbol{X}}_{1,1} = \sum_{s_1} \boldsymbol{x}_{1i} / \pi_{1i}$, $\hat{\boldsymbol{X}}_1 = \sum_{s_1} \boldsymbol{x}_i / \pi_{1i}$ et $\hat{\boldsymbol{X}}_2 = \sum_{s_2} \boldsymbol{x}_i / \pi_i^*$. Les vecteurs de régression $\hat{\boldsymbol{B}}_1$ et $\hat{\boldsymbol{B}}_2$ sont estimés par

$$\hat{\boldsymbol{B}}_1 = \left(\sum_{s_2} \boldsymbol{x}_{i1} \, \boldsymbol{x}'_{i1} / \lambda_{1i} \pi_i^* \right)^{-1} \sum_{s_2} \boldsymbol{x}_{1i} \, y_i / \lambda_{1i} \, \pi_i^*$$

et

$$\hat{\boldsymbol{B}}_2 = \left(\sum_{s_2} \boldsymbol{x}_i \, \boldsymbol{x}'_i / \lambda_i \, \pi_i^* \right)^{-1} \sum_{s_2} \left(\boldsymbol{x}_i \, y_i / \lambda_i \, \pi_i^* \right).$$

Les constantes connues λ_{1i} et λ_i sont des facteurs qui livrent des formes différentes de l'estimateur par régression du total. Ainsi, si les données auxiliaires x_i sont connues seulement pour $i \in s_1$ et que λ_i est proportionnel à x_i, (4.1) se ramène à l'estimateur par quotient à deux phases $\hat{Y}_{2,RAT} = \hat{Y}_2 \left(\hat{X}_1 / \hat{X}_2 \right)$.

Nous pouvons obtenir la variance estimée pour $\hat{Y}_{2,REG}$ en procédant d'abord à sa linéarisation. C'est ainsi que la différence entre $\hat{Y}_{2,REG}$ et Y se présente sous la forme

$$\hat{Y}_{2,REG} - Y = \left(\hat{Y}_2 - Y \right) + \left(\boldsymbol{X}_{1,1} - \hat{\boldsymbol{X}}_{1,1} \right)' \hat{\boldsymbol{B}}_1 + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_2 \right)' \hat{\boldsymbol{B}}_2 + \left(\hat{\boldsymbol{X}}_1 - \boldsymbol{X} \right)' \hat{\boldsymbol{B}}_2 \quad . \qquad (4.2)$$

La version linéarisée de $\hat{Y}_{2,REG} - Y$ est donnée par

$$\begin{aligned} \hat{Y}_{\ell,2REG} &= \left(\hat{Y}_2 - Y \right) + \left(\boldsymbol{X}_{1,1} - \hat{\boldsymbol{X}}_{1,1} \right)' \boldsymbol{B}_1 + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_2 \right)' \boldsymbol{B}_2 + \left(\hat{\boldsymbol{X}}_1 - \boldsymbol{X} \right)' \boldsymbol{B}_2 \\ &= \left(\sum_{s_1} \frac{e_{1i}}{\pi_{1i}} - \sum_U e_{1i} \right) + \left(\sum_{s_2} \frac{e_{2i}}{\pi_i^*} - \sum_{s_1} \frac{e_{2i}}{\pi_{1i}} \right) , \end{aligned} \qquad (4.3)$$

où $e_{1i} = y_i - \boldsymbol{x}_{1i}' \boldsymbol{B}_1$, $e_{2i} = y_i - \boldsymbol{x}_i' \boldsymbol{B}_2$ avec $\boldsymbol{B}_1 = \left(\sum_U \boldsymbol{x}_{i1} \boldsymbol{x}_{i1}' / \lambda_{1i} \right)^{-1} \sum_U \boldsymbol{x}_{1i} y_i / \lambda_{1i}$ et $\boldsymbol{B}_2 = \left(\sum_U \boldsymbol{x}_i \boldsymbol{x}_i' / \lambda_i \right)^{-1} \sum_U (\boldsymbol{x}_i y_i / \lambda_i)$. Par (4.3), la variance de population de $\hat{Y}_{\ell,2REG}$ est de la forme (2.1). En d'autres termes,

$$\begin{aligned} V\left(\hat{Y}_{\ell,2REG}\right) &= E\left[V\left(\hat{Y}_{\ell,2REG} \mid s_1\right)\right] + V\left[E\left(\hat{Y}_{\ell,2REG} \mid s_1\right)\right] \\ &= E\left[\sum\sum_{i<j\in s_1} \left(\pi_{2i|s_1}\, \pi_{2j|s_1} - \pi_{2ij|s_1}\right)\left(\frac{e_{2i}}{\pi_i^*} - \frac{e_{2j}}{\pi_j^*}\right)^2 \right] + \sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{e_{1i}}{\pi_{1i}} - \frac{e_{1j}}{\pi_{1j}}\right)^2 . \end{aligned} \tag{4.4}$$

Il s'ensuit de (4.4) que le premier élément de $V\left(\hat{Y}_{\ell,2REG}\right)$ s'estime par

$$\mathrm{v}\left(\hat{Y}_{\ell,2REG} \middle| s_1\right) = \sum\sum_{i<j\in s_2} \frac{\pi_{2i|s_1}\, \pi_{2j|s_1} - \pi_{2ij|s_1}}{\pi_{2ij|s_1}} \left(\frac{\hat{e}_{2i}}{\pi_i^*} - \frac{\hat{e}_{2j}}{\pi_j^*}\right)^2 , \tag{4.5}$$

où $e_{2i} = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{B}}_2$. Nous procédons comme dans Axelson (1998) pour estimer le second élément de (4.4). Dans ce cas, nous substituons $e_{1i} = e_{2i} + d_i$, où $d_i = e_{1i} - e_{2i} = \mathbf{x}_i' \mathbf{B}_2 - \mathbf{x}_{1i}' \mathbf{B}_1$, au second terme de (4.4) pour ainsi obtenir

$$\begin{aligned} V\left[E\left(\hat{Y}_{\ell,2REG} \middle| s_1\right)\right] &= \sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{e_{2i} + d_i}{\pi_{1i}} - \frac{e_{2j} + d_j}{\pi_{1j}}\right)^2 \\ &= \sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{e_{2i}}{\pi_{1i}} - \frac{e_{2j}}{\pi_{1j}}\right)^2 + \sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}}\right)^2 \\ &\quad + 2\sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}}\right)\left(\frac{e_{2i}}{\pi_{1i}} - \frac{e_{2j}}{\pi_{1j}}\right) . \end{aligned} \tag{4.6}$$

Comme $d_i = \mathbf{x}_i' \mathbf{B}_2 - \mathbf{x}_{1i}' \mathbf{B}_1$ est disponible pour toutes les unités de l'échantillon de première phase s_1, un estimateur sans biais de $\sum\sum_{i<j\in U} \left(\pi_{1i}\, \pi_{1j} - \pi_{1ij}\right)\left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}}\right)^2$ est $\sum\sum_{i<j\in s_1} \frac{\pi_{1i}\, \pi_{1j} - \pi_{1ij}}{\pi_{1ij}} \left(\frac{d_i}{\pi_{1i}} - \frac{d_j}{\pi_{1j}}\right)^2$. Si nous substituons $\hat{d}_i = \mathbf{x}_i' \hat{\mathbf{B}}_2 - \mathbf{x}_{1i}' \hat{\mathbf{B}}_1$ dans (4.6), un estimateur de variance de $V\left[E\left(\hat{Y}_{\ell,2REG} \middle| s_1\right)\right]$ est

$$\begin{aligned} \mathrm{v}_2\left(\hat{Y}_{\ell,2REG}\right) &= \sum\sum_{i<j\in s_2} \frac{\pi_{1i}\, \pi_{1j} - \pi_{1ij}}{\pi_{ij}^*} \left(\frac{\hat{e}_{2i}}{\pi_{1i}} - \frac{\hat{e}_{2j}}{\pi_{1j}}\right)^2 + \sum\sum_{i<j\in s_1} \frac{\pi_{1i}\, \pi_{1j} - \pi_{1ij}}{\pi_{1ij}} \left(\frac{\hat{d}_i}{\pi_{1i}} - \frac{\hat{d}_j}{\pi_{1j}}\right)^2 \\ &\quad + 2\sum\sum_{i<j\in s_2} \frac{\pi_{1i}\, \pi_{1j} - \pi_{1ij}}{\pi_{ij}^*} \left(\frac{\hat{d}_i}{\pi_{1i}} - \frac{\hat{d}_j}{\pi_{1j}}\right)\left(\frac{\hat{e}_{2i}}{\pi_{1i}} - \frac{\hat{e}_{2j}}{\pi_{1j}}\right) . \end{aligned} \tag{4.7}$$

Un estimateur du type SYG en linéarisation de $V\left(\hat{Y}_{2,REG}\right)$ nous est maintenant donné par $\mathrm{v}_{\mathrm{SYG}}\left(\hat{Y}_{2,REG}\right) \doteq \mathrm{v}\left(\hat{Y}_{\ell,2REG} \middle| s_1\right) + \mathrm{v}_2\left(\hat{Y}_{\ell,2REG}\right)$, où $\mathrm{v}\left(\hat{Y}_{\ell,2REG} \middle| s_1\right)$ et $\mathrm{v}_2\left(\hat{Y}_{\ell,2REG}\right)$ viennent respectivement de (4.5) et (4.7).

Comme le signale Axelson (1998), les formes corrigées « g » de cet estimateur de variance sont nombreuses. Pour des facteurs « g » à la première et à la seconde phase $g_{1i} = 1 + \left(\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_{1,1}\right)' \left(\sum_{s_1} \frac{\boldsymbol{x}_{1i}\, \boldsymbol{x}_{1i}'}{\pi_{1i}\lambda_{1i}}\right)^{-1} \frac{\boldsymbol{x}_{1i}}{\lambda_{1i}}$ et $g_{2i} = 1 + \left(\hat{\boldsymbol{X}}_1 - \hat{\boldsymbol{X}}_2\right)' \left(\sum_{s_2} \frac{\boldsymbol{x}_i\, \boldsymbol{x}_i'}{\pi_i^* \lambda_1}\right)^{-1} \frac{\boldsymbol{x}_i}{\lambda_1}$, une forme corrigée « g » serait issue du remplacement de $\hat{d}_i$ par $\tilde{d}_i = g_{1i}\, \hat{d}_i$ et de $\hat{e}_{2i}$ par $\tilde{e}_{2i} = g_{2i}\, \hat{e}_{2i}$.

Exemple 4.1 : Posons qu'il y a échantillonnage aléatoire simple sans remise aux deux phases et qu'un terme d'ordonnée à l'origine est inclus dans la régression. L'expression (4.7) devient

$$v_{SYG}\left(\hat{Y}_{2,\mathrm{Re}\,g}\right) \doteq N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1}\right)\hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\left(\hat{S}_{2e_2}^2 + 2\,\hat{S}_{2e_2,d}\right) + \left(\frac{1}{n_1} - \frac{1}{N}\right)\hat{S}_{1d}^2\right], \tag{4.8}$$

où

$\hat{S}_{1d}^2 = (1/(n-1))\sum_{s_1}\left(\hat{d}_i - \bar{\hat{d}}_1\right)^2$, $\hat{S}_{2e_2}^2 = (1/(n_2-1))\sum_{s_2}\left(\hat{e}_{2i} - \bar{\hat{e}}_2\right)^2$, $\hat{S}_{2e_2,d} = (1/(n_2-1))\sum_{s_2}\left(\hat{e}_{2i} - \bar{\hat{e}}_2\right)\left(\hat{d}_i - \bar{\hat{d}}_2\right)$

avec $\bar{\hat{d}}_1 = (1/n_1)\sum_{s_1}\hat{d}_i$; $\bar{\hat{d}}_2 = (1/n_2)\sum_{s_2}\hat{d}_i$; $\bar{\hat{e}}_2 = (1/n_2)\sum_{s_2}\hat{e}_{2i}$; $e_{1i} = \left(y_i - \bar{y}_{s_2}\right) - \left(\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_{1s_2}\right)'\hat{\boldsymbol{B}}_1$ et $e_{2i} = \left(y_i - \bar{y}_{s_2}\right) - \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{s_2}\right)'\hat{\boldsymbol{B}}_2$. Et $\hat{\boldsymbol{B}}_1$ et $\hat{\boldsymbol{B}}_2$ sont les estimateurs de régression par les moindres carrés corrigés pour l'ordonnée à l'origine.

La version correspondante $\tilde{v}_{SYG}\left(\hat{Y}_{2,\mathrm{Re}\,g}\right)$ qui vient du remplacement de $\tilde{d}_i = g_{1i}\, \hat{d}_i$ par $\hat{d}_i$ et de $\tilde{e}_{2i} = g_{2i}\, \hat{e}_{2i}$ par $\hat{e}_{2i}$ dans $v_{SYG}\left(\hat{Y}_{2,\mathrm{Re}\,g}\right)$ est

$$\tilde{v}_{SYG}\left(\hat{Y}_{2,\mathrm{Re}\,g}\right) \doteq N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1}\right)\tilde{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\left(\tilde{S}_{2e_2}^2 + 2\,\tilde{S}_{2e_2,d}\right) + \left(\frac{1}{n_1} - \frac{1}{N}\right)\tilde{S}_{1d}^2\right], \tag{4.9}$$

où

$\tilde{S}_{1d}^2 = (1/(n_1-1))\sum_{s_1}\left(\tilde{d}_i - \bar{\tilde{d}}_1\right)^2$, $\tilde{S}_{2e_2}^2 = (1/(n_2-1))\sum_{s_2}\left(\tilde{e}_{2i} - \bar{\tilde{e}}_2\right)^2$, $\tilde{S}_{2e_2,d} = (1/(n_2-1))\sum_{s_2}\left(\tilde{e}_{2i} - \bar{\tilde{e}}_2\right)\left(\tilde{d}_i - \bar{\tilde{d}}_2\right)$

avec $\bar{\tilde{d}}_1 = (1/n_1)\sum_{s_1}\tilde{d}_i$ $\bar{\tilde{d}}_2 = (1/n_2)\sum_{s_2}\tilde{d}_i$ et $\bar{\tilde{e}}_2 = (1/n_2)\sum_{s_2}\tilde{e}_{2i}$.

Si on dispose de données auxiliaires scalaires *x* dans l'échantillon de la première phase, l'expression (4.8) se ramène à celle que propose Dorfman (1994). Cette expression est donnée par

$$\mathrm{v}_{\mathrm{SYG}}\left(\hat{Y}_{2,REG}\right) = N^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1}\right)\hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\hat{S}_{2e_2}^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\hat{S}_{1x}^2\,\hat{B}_2^2\right], \tag{4.10}$$

où $\hat{S}_{1x}^2 = (1/(n_1-1))\sum_{s_1}\left(x_i - \bar{x}_1\right)^2$.

En incorporant les termes « g » à (4.10), on obtient

$$\tilde{\mathrm{v}}_{SYG}\left(\hat{Y}_{2,\mathrm{Re}\,g}\right) = N^2\left\{\left(\frac{1}{n_2}-\frac{1}{N}\right)\tilde{S}^2_{2e_2}+\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{S}^2_{1x}\,\hat{B}^2_2+2\left(\frac{1}{n_1}-\frac{1}{N}\right)\tilde{S}_{2\,e,x}\,\hat{B}_2\right\}, \tag{4.11}$$

où $g_{2i} = 1+\frac{n_2}{n_2-1}\frac{\bar{x}_1-\bar{x}_2}{\hat{S}^2_{2x}}\left(x_i-\bar{x}_2\right)$.

RÉFÉRENCES

Axelson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling-a modified approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 85-89.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A., et Jocelyn, W. (2000). Variance Estimation for Two-phase Stratified Sampling. *Canadian Journal of Statistics*, Vol. 28, No 4, 751-764.

Chaudhuri, A., et Roy, D. (1994. Model assisted survey-sampling strategy in two phases. *Metrika*, 41, 355-362.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed., New York : John Wiley.

Dorfman, A.H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association*, 89, 137-140.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**, 11-20, Vol. 24, 11-20.

Kott, P.S. et Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology, 23*, 81-90.

Lee, H. et Kim, J.K. (2002). Jackknife Variance Estimation for Two-Phase Samples: An Application to the Study on Personnel Need for Special Education (SPeNSE). *Proceedings of the Section on Survey Research Methods*, 2002 Joint Statistical Meetings in New York.

Neyman, J. (1938). Contribution To The Theory of Sampling Human Populations. *Journal of the American Statistical Association*, 33, 101-116.

Rao, J.N.K. (1973). Double Sampling for Stratification and Analytical Surveys", *Biometrika*, **60**, 125-133.

Rao, J.N.K. (1979), "On Deriving Mean Square Errors and their Nonnegative Unbiased Estimators", *Journal of the Statistical Association*, **17**, 125-136.

Rao, J.N.K. et Singh, M.P (1973),"On the Choice of Estimator in Survey Sampling", *Australian J. Stat.*, **15**, 95-104.

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

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