Techniques d'enquête, décembre 2000
Vol. 26, N° 2, pp. 207-212
Statistique Canada, N° 12-001 au catalogue

Échantillonnage à deux phases pour l'estimation par quotient ou par régression avec sous-échantillonnage des non-répondants

Fabian C. Okafor et Hyunshik Lee [1]

Résumé

Cochran (1977, page 374) a proposé certains estimateurs par quotient ou par régression de la moyenne de population fondés sur la méthode de Hansen et Hurwitz (1946) consistant à sous-échantillonner les non-répondants en supposant que l'on connaît la moyenne de population de la variable auxiliaire. Le présent article décrit certains estimateurs par quotient ou par régression axés sur un échantillonnage double (à deux phases) applicables aux cas où l'on ne connaît pas la moyenne de population de la variable auxiliaire. On y compare aussi la performance de ces estimateurs à celle de l'estimateur proposé par Hansen et Hurwitz (1946).

Mots clés : Estimateur de Hansen et Hurwitz; coût d'enquête; fraction optimale d'échantillonnage.

1. Introduction

Très souvent, lors d'enquêtes visant des personnes, on ne réussit pas à recueillir les renseignements auprès de toutes les unités d'échantillonnage, même après plusieurs rappels. Or, une estimation calculée d'après des données incomplètes peut être trompeuse, surtout si les caractéristiques des répondants diffèrent de celles des non-répondants, auquel cas l'estimation risque d'être biaisée. Hansen et Hurwitz (1946) ont proposé une méthode de correction pour la non-réponse afin de résoudre le problème du biais. Cette méthode consiste à sélectionner un sous-échantillon de non-répondants afin d'obtenir une estimation pour la sous-population que ces derniers représentent.

En s'appuyant sur la méthode de Hansen et Hurwitz (1946), Cochran (1977) a proposé les estimateurs par quotient ou par régression de la moyenne de population de la variable étudiée pour lesquels les renseignements sur la variable auxiliaire proviennent de toutes les unités d'échantillonnage, alors que certaines unités n'ont pas toutes fourni les renseignements sur la variable étudiée. En outre, on connaît la moyenne de population de la variable auxiliaire. Ici, nous supposons que l'on ne connaît pas la moyenne de population de la variable auxiliaire. Par conséquent, nous utilisons la méthode d'échantillonnage à deux phases pour estimer d'abord la moyenne de la variable auxiliaire, puis la moyenne de la variable étudiée selon une méthode similaire à celle de Cochran (1977).

En pratique, on compense souvent la non-réponse par rajustement de la pondération (Oh et Scheuren 1983) ou par imputation (Kalton et Karsprzyk 1986). Les méthodes appliquées pour rajuster la pondération ou pour procéder à l'imputation visent à éliminer le biais dû à la non-réponse. Cependant, cette méthode se fonde sur des hypothèses insoutenables quant au mécanisme de réponse. Si le mécanisme hypothétique est erroné, les estimations résultantes risquent d'être fortement biaisées. De surcroît, il est difficile d'éliminer entièrement le biais s'il y a confusion de la non-réponse, en ce sens que la probabilité de réponse dépend de la variable étudiée. Rancourt, Lee et Särndal (1994) ont réussi à corriger partiellement cette situation. La méthode de sous-échantillonnage de Hansen et Hurwitz ne présente pas ce défaut, mais elle est plus coûteuse à cause du travail supplémentaire qu'exige le sous-échantillonnage des non-répondants. Néanmoins, si le biais est important, la méthode offre un moyen viable de résoudre le problème sans recourir à la réponse totale qui pourrait coûter fort cher.

À la section suivante, nous examinons les estimateurs par quotient ou par régression axés sur l'échantillonnage à deux phases. En général, on recoure à l'échantillonnage à deux phases lorsqu'il faut se servir de données auxiliaires pour améliorer la précision d'une estimation, mais que l'on ne connaît pas la loi de distribution de la population pour les variables auxiliaires. Nous nous servons de l'échantillon de la première phase pour estimer la distribution de population de la variable auxiliaire et de l'échantillon de la deuxième phase pour obtenir les renseignements nécessaires sur la variable étudiée. Nous calculons la fraction optimale d'échantillonnage pour les divers estimateurs, pour un coût prédéterminé. Enfin, nous comparons théoriquement et empiriquement les estimateurs proposés à l'estimateur de Hansen et Hurwitz.

2. Estimateurs par quotient ou par régression avec échantillonnage à deux phases

2.1 Renseignements généraux

Pour estimer la moyenne de population $\bar{X}$ de la variable auxiliaire, nous sélectionnons un échantillon de première

1. Fabian C. Okafor, Dept. of Statistics, University of Nigeria, Nsukka, Nigeria; Hyunshik Lee, anciennement Statistique Canada, maintenant Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, États-Unis.

phase de grande taille n' à partir de N unités de la population par échantillonnage aléatoire simple sans remise (EASSR). Puis, nous sélectionnons par EASSR un échantillon de deuxième phase de plus petite taille n à partir de n' et nous mesurons la caractéristique y sur cet échantillon. L'estimateur par quotient de la moyenne de y est $\bar{y}'_r = (\bar{y}/\bar{x})\bar{x}'$, où $\bar{x}'$ est la moyenne d'échantillon calculée pour n' unités, et où $\bar{y}$ et $\bar{x}$ sont obtenues d'après l'échantillon de deuxième phase s'il n'y a pas de non-réponse dans cet échantillon. Cependant, en cas de non-réponse dans l'échantillon de deuxième phase, nous pouvons utiliser un estimateur fondé uniquement sur les répondants ou sélectionner un sous-échantillon de non-répondants et reprendre contact avec eux. La première option est beaucoup moins coûteuse que la seconde, parce que recueillir les renseignements manquants auprès des non-répondants en les contactant de nouveau nécessite habituellement beaucoup plus d'efforts et de dépenses. Cependant, il se pourrait fort bien qu'en ce qui concerne la caractéristique étudiée, les non-répondants diffèrent des répondants au point que les résultats soient sérieusement biaisés. Le cas échéant, le sous-échantillonnage des non-répondants pourrait être souhaitable. Par conséquent, nous poursuivons l'idée du sous-échantillonnage de Hansen et Hurwitz dans le cas d'un échantillonnage à deux phases. Fondamentalement, les estimateurs que nous proposons ici constituent une version à échantillonnage à deux phases des estimateurs proposés par Cochran (1977, page 374), c'est-à-dire des estimateurs de $\bar{Y}$ par quotient ou par régression avec échantillonnage à deux phases, corrigés pour la non-réponse par la méthode de Hansen et Hurwitz (1946).

Supposons que les n' unités fournissent toutes des renseignements sur la variable auxiliaire x à la première étape d'échantillonnage. Mais posons que n_1 unités fournissent des renseignements sur y et que n_2 unités refusent de répondre lors de la deuxième étape. À partir des n_2 non-respondants, nous sélectionnons un échantillon aléatoire simple sans remise de m unités au taux inverse d'échantillonnage k, où $m = n_2/k$, $k > 1$. Cette fois-ci, les m unités répondent toutes. Ces conditions pourraient s'appliquer à une enquête-ménages où l'on se sert de la taille du ménage comme variable auxiliaire pour l'estimation, disons, des dépenses familiales. On pourrait obtenir des renseignements complets sur la taille de la famille durant l'établissement de la liste des ménages, mais faire face à une non-réponse en ce qui concerne les dépenses du ménage.

Dans l'exposé qui suit, nous supposons que l'ensemble de la population (représenté par A) est divisé en deux strates : une strate(représentée par A_1) de N_1 unités qui répondent lors de la première visite à la deuxième étape et une strate (représentée par A_2) de N_2 unités qui ne répondent pas lors de la première visite à la deuxième étape d'échantillonnage, mais qui répondent lors de la deuxième visite. Représentons les échantillons de premier et de deuxième phase par a' et a, respectivement, et posons que $a_1 = a \cap A_1$ et $a_2 = a \cap A_2$. Le sous-échantillon de a_2 sera représenté par a_{2m}. La somme sur l'ensemble des unités est un ensemble s que nous représentons par $\sum_s$.

En règle générale, les paramètres de population sont représentés par des lettres majuscules, sauf les lettres grecques, et les statistiques d'échantillon, par les lettres minuscules correspondantes.

2.2 Estimateur par quotient avec échantillonnage à deux phases

Nous définissons l'estimateur par quotient avec échantillonnage à deux phases comme suit :

$$d^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' = r^*\bar{x}' \tag{2.1}$$

où $\bar{x}^*$ et $\bar{y}^*$ sont, respectivement, les estimateurs de $\bar{X}$ et $\bar{Y}$, de Hansen-Hurwitz qui sont donnés par

$$\bar{u}^* = w_1\bar{u}_1 + w_2\bar{u}_{2m}, \qquad u = x, y. \tag{2.2}$$

Conformément à la règle générale, nous définissons $W_j = N_j/N$ et $w_j = n_j/n$, $j = 1$ ou 2. Les statistiques d'échantillon calculées d'après a_{2m} sont marquées de l'indice « $2m$ », (par exemple, $\bar{u}_{2m} = (1/m)\sum_{a_{2m}} u_i$); celles calculées d'après a_1 sont marquées de l'indice « 1 », (par exemple, $\bar{u}_1 = (1/n_1)\sum_{a_1} u_i$), et celles calculées d'après l'échantillon de première phase a' sont marqués d'un signe prime (par exemple, $\bar{x}' = (1/n')\sum_{a'} x_i$).

Pour un grand échantillon, l'approximation de premier ordre de la variance de d^*, calculée selon la méthode de linéarisation par série de Taylor, est donnée par

$$V(d^*) \cong \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_r^2 + \frac{W_2(k-1)}{n} S_{2r}^2 \tag{2.3}$$

où

$$S_r^2 = S_y^2 + R^2 S_x^2 - 2RS_{xy},$$

$$S_{2r}^2 = S_{2y}^2 + R^2 S_{2x}^2 - 2RS_{2xy}, \tag{2.4}$$

R représente le rapport de population de $\bar{Y}$ à $\bar{X}$. S_u^2 et S_{2u}^2 représentent, respectivement, la variance de la variable u pour l'ensemble de la population et pour la strate de non-répondants. S_{xy} et S_{2xy} représentent, respectivement, la covariance pour l'ensemble de la population et pour la population de non-répondants.

Nous pouvons estimer la variance approximative de d^* par

$$v(d^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_r^2 + \frac{w_2(k-1)}{n}\hat{S}_{2r}^2 \tag{2.5}$$

où

$$\hat{S}_y^2 = \frac{1}{n-1}\left\{\sum_{a_1} y_i^2 + k\sum_{a_{2m}} y_i^2 - n\bar{y}^{*2} + w_2(k-1)\, s_{2my}^2\right\},$$

$$\hat{S}_r^2 = \frac{1}{n-1}\left\{\sum_{a_1} (y_i - r^* x_i)^2 + k\sum_{a_{2m}} (y_i - r^* x_i)^2\right\} \text{ et}$$

$$\hat{S}_{2r}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - r^* x_i)^2. \tag{2.6}$$

Il convient de souligner que $\hat{S}_y^2$ est un estimateur non biaisé de S_y^2. Il semble naturel d'utiliser $\hat{S}_r^2$ pour estimer S_r^2 puisque l'expression obtenue d'après $\hat{S}_r^2$ en remplaçant r^* par R est un estimateur convergent de S_r^2. Nous pouvons nous servir du même argument pour justifier l'emploi de $\hat{S}_{2r}^2$.

Nous obtenons un autre estimateur de $V(d^*)$ en remplaçant $\hat{S}_r^2$ par $\hat{S}_{2r}^2$, c'est-à-dire :

$$\tilde{S}_r^2 = \hat{S}_y^2 + r^{*2} s_x'^2 - 2r^* s_{xy}^* \quad \text{et}$$

$$\tilde{S}_{2r}^2 = s_{2my}^2 + r^{*2} s_{2x}^2 - 2r^* s_{2mxy}, \tag{2.7}$$

respectivement, dans (2.5), où

$$s_x'^2 = \frac{1}{n'-1}\sum_{a'} (x_i - \bar{x}')^2,$$

$$s_{2my}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - \bar{y}_{2m})^2,$$

$$s_{2x}^2 = \frac{1}{n_2-1}\sum_{a_2} (x_i - \bar{x}_2)^2,$$

$$s_{2mxy} = \frac{1}{m-1}\left(\sum_{a_{2m}} x_i\, y_i - m\bar{x}_{2m}\,\bar{y}_{2m}\right)$$

et où s_{xy}^* a la même forme que dans (2.9). La variance de cet estimateur de rechange sera vraisemblablement plus faible que celle de l'estimateur (2.5), puisque les estimateurs $s_x'^2$ et s_{2x}^2 sont fondés sur des échantillons plus grands et sont par conséquent plus précis.

2.3 Estimateur par régression avec échantillonnage à deux phases

Nous définissons l'estimateur par régression comme suit :

$$t^* = \bar{y}^* + \hat{\beta}^*(\bar{x}' - \bar{x}^*) \tag{2.8}$$

où $\hat{\beta}^*$ est un estimateur de $\beta = S_{xy}/S_x^2$. Il pourrait exister plusieurs solutions pour $\hat{\beta}^*$, mais un choix naturel semble être $\hat{\beta}^* = s_{xy}^*/s_x^{*2}$, où

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{a_1} x_i\, y_i + k\sum_{a_{2m}} x_i\, y_i - n\overline{xy}^*\right)$$

et

$$s_x^{*2} = \frac{1}{n-1}\left(\sum_{a_1} x_i^2 + k\sum_{a_{2m}} x_i^2 - n\overline{xx}^*\right). \tag{2.9}$$

Il est facile de montrer que s_{xy}^* et s_x^{*2} sont des estimateurs non biaisés de S_{xy} et S_x^2, respectivement. La variance approximative de t^* est donnée par

$$V(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_l^2 + \frac{W_2(k-1)}{n} S_{2l}^2 \tag{2.10}$$

où S_l^2 et S_{2l}^2 sont calculés d'après (2.4) en remplaçant R par β.

Pour estimer $V(t^*)$, nous pouvons appliquer la formule suivante :

$$v(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right) \hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \hat{S}_l^2 + \frac{w_2(k-1)}{n} \hat{S}_{2l}^2 \tag{2.11}$$

où

$$\hat{S}_l^2 = \frac{1}{n-1}\left\{\sum_{a_1} (y_i - y_i^*)^2 + k\sum_{a_{2m}} (y_i - y_i^*)^2\right\},$$

$$\hat{S}_{2l}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - y_i^*)^2,$$

et

$$y_i^* = \bar{y}_i^* - \hat{\beta}^*(x_i - \bar{x}^*). \tag{2.12}$$

Comme pour (2.7), nous obtenons un estimateur un peu meilleur de $V(t^*)$ en utilisant :

$$\tilde{S}_l^2 = \hat{S}_y^2 + \hat{\beta}^{*2} s_x'^2 - 2\hat{\beta}^* s_{xy}^*$$

et

$$\tilde{S}_{2l} = s_{2my}^2 + \hat{\beta}^{*2} s_{2x}^2 - 2\hat{\beta}^* s_{2mxy}. \tag{2.13}$$

3. Choix des fractions d'échantillonnage

Nous allons maintenant déterminer les valeurs optimales de k, n, et n' qui réduisent au minimum la variance des estimateurs proposés pour un coût particulier ou qui réduisent au minimum le coût pour une variance particulière.

Considérons pour d^* la fonction de coût donnée par

$$C = c'n' + cn + c_1n_1 + c_2m \qquad (3.1)$$

où les c représentent les coûts unitaires définis comme suit :

c': coût unitaire associé à l'échantillon de première phase, a';

c: coût unitaire associé au premier effort de collecte de renseignements sur y auprès de l'échantillon de deuxième phase, a;

c_1: coût unitaire du traitement des renseignements sur y fournis par les répondants lors du premier effort de collecte auprès de a_1;

c_2: coût unitaire associé au sous-échantillon a_{2m} de a_2.

Puisqu'on ne connaît pas la valeur de n_1 tant qu'on n'a pas fait la première collecte de données, on se sert du coût prévu pour la minimisation. Le coût prévu est donné par

$$E(C) = C^* = c'n' + \left(c + c_1W_1 + \frac{c_2W_2}{k}\right)n. \qquad (3.2)$$

Nous nous servons du multiplicateur de Lagrange pour recalculer les valeurs optimales de k, n, et n' qui réduisent au minimum la variance de d^* pour un coût prévu fixe C^*. Les valeurs optimales ainsi obtenues sont :

$$k_o = \sqrt{\frac{c_2(S_r^2 - W_2S_{2r}^2)}{S_{2r}^2(c + c_1W_1)}}$$

$$n_o = \frac{C^*\sqrt{A}}{D\sqrt{G}} \quad \text{et} \quad n_o' = \frac{C^*\sqrt{S_y^2 - S_r^2}}{D\sqrt{c'}} \qquad (3.3)$$

où

$$A = S_r^2 + W_2(k_o - 1)S_{2r}^2,$$

$$G = c + c_1W_1 + \frac{c_2W_2}{k_o} \quad \text{et}$$

$$D = \sqrt{(S_y^2 - S_r^2)c'} + \sqrt{AG}.$$

Si nous supposons que $\gamma = c_2/(c + c_1W_1)$, $\delta = S_r^2/S_{2r}^2$, et $\xi = S_y^2/S_r^2$, alors nous obtenons

$$k_o = \sqrt{\gamma(\delta - W_2)},$$

$$n_o = \frac{C^*\sqrt{1 + W_2(k_o - 1)/\delta}}{\sqrt{Gc'(\xi - 1)} + G\sqrt{1 + W_2(k_o - 1)/\delta}} \quad \text{et}$$

$$n_o' = \frac{C^*\sqrt{\xi - 1}}{c'\sqrt{\xi - 1} + \sqrt{Gc'\{1 + W_2(k_o - 1)/\delta\}}}. \qquad (3.4)$$

Les valeurs optimales de n_o et n_o' sont proportionnelles au coût prévu, C^*. Pour obtenir les valeurs optimales de k, n et n' qui réduisent au minimum la valeur de $V(t^*)$, nous remplaçons simplement dans l'expression (3.3) susmentionnée S_r^2 et S_{2r}^2 par S_l^2 et S_{2l}^2, respectivement. Le tableau 1 montre les valeurs optimales de k_o, n_o, et n_o' pour des paramètres donnés.

Tableau 1
Valeur optimale de k_o, n_o, et n_o'

C^*	c'	C	c_1	c_2	δ	ξ	W_2	γ	k_o	G	n_o	n_o'
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92	382
200	0,1	0,5	1	2	1	4	0,3	1,67	1,08	1,76	81	580
200	0,1	0,5	1	2	2	2	0,3	1,67	1,68	1,56	104	389
200	0,1	0,5	1	2	2	4	0,3	1,67	1,68	1,56	91	590
200	0,1	0,5	1	4	1	2	0,3	3,33	1,52	1,99	83	345
200	0,1	0,5	1	4	1	4	0,3	3,33	1,52	1,99	74	531
200	0,1	0,5	1	4	2	2	0,3	3,33	2,38	1,70	96	361
200	0,1	0,5	1	4	2	4	0,3	3,33	2,38	1,70	85	553
200	0,5	0,5	1	2	1	2	0,3	1,67	1,08	1,76	85	250
200	0,5	0,5	1	2	1	4	0,3	1,67	1,08	1,76	72	366
200	0,5	0,5	1	2	2	2	0,3	1,67	1,68	1,56	96	255
200	0,5	0,5	1	2	2	4	0,3	1,67	1,68	1,56	80	372
200	0,5	0,5	1	4	1	2	0,3	3,33	1,52	1,99	78	228
200	0,5	0,5	1	4	1	4	0,3	3,33	1,52	1,99	67	338
200	0,5	0,5	1	4	2	2	0,3	3,33	2,38	1,70	89	238
200	0,5	0,5	1	4	2	4	0,3	3,33	2,38	1,70	76	351

4. Comparaison des estimateurs

À la présente section, nous comparons théoriquement la performance des estimateurs proposés à celle des estimateurs de Hansen et Hurwitz (1946), d'abord sans tenir compte du coût, puis en tenant compte de celui-ci.

4.1 Sans tenir compte du coût

La variance de l'estimateur de Hansen-Hurwitz est donnée par

$$V(\bar{y}^*) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k-1)}{n}S_{2y}^2 \qquad (4.1)$$

où $\bar{y}^*$ est défini comme dans (2.2).

$$V(\bar{y}^*) - V(d^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)(2RS_{xy} - R^2S_x^2) + \frac{W_2(k-1)}{n}(2RS_{2xy} - R^2S_{2x}^2). \quad (4.2)$$

La différence est positive (autrement dit, d^* est plus efficace que $\bar{y}^*$) si $R < 2\beta$ et $R < 2\beta_2$, où $\beta_2 = S_{2xy}/S_{2x}^2$. Par ailleurs, nous avons

$$V(\bar{y}^*) - V(t^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{S_{xy}^2}{S_x^2} + \frac{W_2(k-1)}{n}\beta S_{2x}^2(2\beta_2 - \beta). \quad (4.3)$$

Par conséquent, t^* est plus efficace que l'estimateur de Hansen-Hurwitz si la différence (4.3) est positive. Cette condition s'observe notamment si $\beta_2 \geq \beta/2$, avec $\beta \geq 0$. Comme les conditions dont nous discutons ici sont suffisantes, d^* ou t^* peut être plus efficace que $\bar{y}^*$ dans des conditions moins rigoureuses.

4.2 En tenant compte du coût

Nous allons maintenant comparer les estimateurs proposés à l'estimateur de Hansen-Hurwitz ($\bar{y}^*$) en nous servant de la fonction de coût donnée à la section 3.

Pour l'estimateur $\bar{y}^*$, si nous sélectionnons un échantillon aléatoire simple (sans recourir à l'échantillonnage à deux phases) pour y, nous pouvons calculer la taille optimale d'échantillon pour un coût prévu,

$$C^* = \left(c + c_1W_1 + \frac{c_2W_2}{k}\right)n,$$

semblable à celui donné par (3.2) selon la même technique (c'est-à-dire, le multiplicateur de Lagrange) que celle utilisée à la section 3, comme suit :

$$n_{o\mathrm{HH}} = \frac{C^*}{c + c_1W_1 + c_2W_2/k_{o\mathrm{HH}}}$$

et

$$k_{o\mathrm{HH}} = \sqrt{\frac{c_2(S_y^2 - W_2S_{2y}^2)}{S_{2y}^2(c + c_1W_1)}}. \quad (4.4)$$

Alors, la variance optimale de l'estimateur de Hansen-Hurwitz devient

$$V(\bar{y}^*) = \left(\frac{1}{n_{o\mathrm{HH}}} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k_{o\mathrm{HH}} - 1)}{n_{o\mathrm{HH}}}S_{2y}^2. \quad (4.5)$$

Si nous comparons cette expression à celle de $V(d^*)$ avec les valeurs optimales de k, n, et n' données par (3.3), alors la condition voulant que d^* soit plus précis que $\bar{y}^*$ s'écrit

$$2\rho - Rh > \frac{1}{1-\theta_1}\times\left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2Q_{\mathrm{HHy}}) - hQ_x(2\beta_2 - R)\right\} \quad (4.6)$$

où

$$h = \frac{S_x}{S_y},\ \theta_1 = \frac{n_o}{n_o'},\ \theta_2 = \frac{n_o}{n_{o\mathrm{HH}}},\ Q_{\mathrm{HHy}} = \frac{W_2(k_{o\mathrm{HH}} - 1)S_{2y}^2}{S_y^2},$$

$$Q_u = \frac{W_2(k_o - 1)S_{2u}^2}{S_u^2},\quad u = x,\ y,$$

et où ρ représente le coefficient de corrélation de x et y.

Nous pouvons procéder à une comparaison similaire entre $\bar{y}^*$ et t^*. Autrement dit, t^* est plus efficace que $\bar{y}^*$ si

$$2\rho - \beta h > \frac{1}{1-\theta_1}\times\left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2Q_{\mathrm{HHy}}) - hQ_x(2\beta_2 - \beta)\right\}. \quad (4.7)$$

4.3 Comparaison empirique des estimateurs proposés

Nous nous servons d'une population générée artificiellement pour comparer l'efficacité relative des estimateurs d^* et t^* par rapport à $\bar{y}^*$. Les paramètres de la population sont les suivants :

$$R = 1{,}92,\ \beta = 1{,}52,\ \rho = 0{,}85,\ R_2 = 1{,}88,\ \beta_2 = 1{,}47,$$

$$\rho_2 = 0{,}83,\ N = 1\ 000,\ N_2 = 302,\ S_x^2 = 766{,}54,$$

$$S_y^2 = 2\ 426{,}82,\ S_{xy} = 1\ 164{,}08,\ S_{2x}^2 = 433{,}63,$$

$$S_{2y}^2 = 1\ 350{,}05,\ \text{et}\ S_{2xy} = 638{,}32.$$

Les efficacités relatives de d^* et t^* sont présentées au tableau 2. Notons que R diffère nettement de β, autrement

dit que la droite de régression ne passe pas l'origine. Dans le cas de cette population, l'estimateur par régression t^* est plus efficace que l'estimateur par quotient d^*. Nous constatons aussi que la taille initiale optimale d'échantillon, n'_o, est plus grande pour t^* que pour l'estimateur d^*. Pour la taille optimale d'échantillon de deuxième phase, n_o, nous observons l'inverse, parce qu'on peut obtenir un estimateur par régression plus précis avec un plus petit échantillon de deuxième phase, si bien qu'il est possible d'allouer davantage à l'échantillon de première phase. Enfin, le taux d'échantillonnage inverse optimal k_o est pratiquement le même pour les deux estimateurs.

Si la droite de régression passe par l'origine, l'avantage de t^* sur d^* disparaît, comme le prédit et le confirme une autre comparaison empirique que nous ne présentons pas ici.

Tableau 2
Efficacité relative de d^* et t^* par rapport à $\bar{y}^*$ ($C^* = 200,\ c = 0{,}5,\ c_1 = 1$)

c'	c_2	$k_{o\mathrm{HH}}$	$n_{o\mathrm{HH}}$	k_o	n_o	n'_o	Efficacité
				d^*			
0.1	2	1,58	127	1,46	92	514	1,85
0.1	4	2,23	115	2,06	85	477	1,91
0.3	2	1,58	127	1,46	78	250	1,23
0.3	4	2,23	115	2,06	73	234	1,32
				t^*			
0.1	2	1,58	127	1,47	89	563	2,11
0.1	4	2,23	115	2,08	83	523	2,19
0.3	2	1,58	127	1,47	74	269	1,34
0.3	4	2,23	115	2,08	70	253	1,45

5. Conclusions

Nous proposons des estimateurs par quotient ou par régression fondés sur une méthode d'échantillonnage à deux phases pour tenir compte de la non-réponse au sujet de la variable principale quand on ne connaît pas la moyenne de population de la variable auxiliaire. Nous éliminons le biais éventuellement important dû à la non-réponse par sous-échantillonnage des non-répondants, conformément à la méthode de Hansen et Hurwitz (1946). Nous calculons les tailles optimales d'échantillon pour un ensemble donné de coûts unitaires, puis nous comparons théoriquement et empiriquement la performance de nos estimateurs à celle de l'estimateur de Hansen et Hurwitz.

S'il existe une relation linéaire prononcée entre la variable principale et la variable auxiliaire et que l'on peut recueillir à faible coût les données sur la variable auxiliaire auprès d'un échantillon de grande taille, nos estimateurs donnent de nettement meilleurs résultats que l'estimateur de Hansen et Hurwitz. Notre méthode pourrait être utile s'il existe un biais important dû à la non-réponse que l'on ne peut corriger par rajustement de la pondération ni par imputation.

Remerciements

Nous remercions les examinateurs et les rédacteurs en chef adjoints de leurs commentaires qui nous ont aidés à améliorer notre article.

Bibliographie

Cochran, W.G. (1977). *Sampling Techniques*. 3e Édition. New York : John Wiley & Sons, Inc.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Oh, H.L., et Scheuren, F.J. (1983). Chapter 13. Weighting adjustment for unit non-response. Dans *Incomplete Data in Sample Surveys*, (Éds., I. Olkin, W.G. Madow, et D.B. Rubin). Theory and Bibliographies, 2, 143-184. New York : Academic Press.

Rancourt, E., Lee, H. et Särndal, C.-E. (1994). Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme avec confusion. *Techniques d'enquête*, 20, 143-153.