Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 89-98
Statistique Canada

La méthode du jackknife convient-elle à un échantillon à deux phases?

PHILLIP S. KOTT et DIANA M. STUKEL[1]

RÉSUMÉ

L'estimateur de variance jackknife présente des propriétés intéressantes quand on s'en sert avec les estimateurs lisses tirés d'échantillons stratifiés à plusieurs degrés. L'article que voici porte sur l'application de cet estimateur à un plan particulier d'échantillonnage à deux phases: on commence par constituer un échantillon aléatoire stratifié en grappes par tirage non exhaustif, puis on restratifie les éléments des grappes échantillonnées et on prélève des sous-échantillons aléatoires simples de chaque strate de la deuxième phase. Apparemment, l'estimateur jackknife donne des résultats raisonnables pour ce qui est d'estimer la variance d'un estimateur de «développement» commun, mais pas celle d'un autre. Les auteurs parlent de l'application de leurs résultats à des stratégies d'estimation plus complexes. Une étude de Monte Carlo étaye leurs principales constatations.

MOTS CLÉS: Stratifié; estimateur de développement repondéré; estimateur de développement double; asymptotique.

1. INTRODUCTION

Krewski et Rao (1981) et, après eux, Rao et Wu (1985) se sont penchés sur les propriétés de plan de sondage de l'estimateur de variance jackknife dans le cas d'une stratification à plusieurs degrés intégrant un échantillonnage non exhaustif au premier degré. Bien qu'assez généraux en soi, les résultats obtenus par ces chercheurs ne peuvent directement être appliqués à bon nombre de plans d'échantillonnage à plusieurs phases. On lira à ce sujet Wolter (1985; chapitre 4.5).

Nous examinerons ici un simple exemple d'échantillonnage à deux phases. Dans un premier temps, on prélève un échantillon aléatoire stratifié en grappes, par tirage non exhaustif. Les éléments des grappes échantillonnées subissent ensuite une nouvelle stratification, peut-être au moyen de renseignements recueillis à la première phase, et on construit de façon aléatoire un nouveau sous-échantillon simple stratifié, par tirage exhaustif.

Il est possible d'estimer un total sans information auxiliaire de deux façons. La première consiste à multiplier la valeur de chaque élément sous-échantillonné par le produit de ses facteurs de pondération à chaque phase (à savoir, l'inverse de la probabilité de sélection à la première et à la deuxième phase), puis d'en faire la somme. C'est *l'estimateur de développement double* que Särndal, Swensson et Wretman (1992, p. 347) appellent «estimateur π^*».

Si on parle beaucoup de l'estimateur de développement double dans les traités de statistique, dans la pratique, il est plus courant de recourir à *l'estimateur de développement repondéré*, surtout si on traite la non-réponse des comme une deuxième phase de l'échantillonnage, comme Oh et Scheuren (1983, p. 150) le font avec l'estimateur de la classe de pondération. Pour obtenir un estimateur de la taille de la population applicable aux strates de la deuxième phase d'échantillonnage, on additionne les facteurs de pondération de la première phase pour tous les éléments de la strate avant le sous-échantillonnage. Ensuite, on multiplie le résultat par la moyenne estimée de la strate de la deuxième phase à partir du sous-échantillon, ce qui donne le total estimé de la strate. Enfin, on fait la somme des totaux estimés pour les strates de la deuxième phase de l'échantillonnage, ce qui produit l'estimateur de développement repondéré de la population totale.

Dans le cas présent, nous nous intéresserons plus à un véritable échantillonnage à deux phases qu'à l'usage de la non-réponse comme phase d'échantillonnage artificielle supplémentaire. Le National Agricultural Statistics Service (NASS) recourt actuellement à l'estimateur de développement double dans ses enquêtes trimestrielles sur l'agriculture (ETA). On obtient un échantillon stratifié, ariolaire en grappe est de nombre en juin. Les exploitations agricoles identifiées en juin sont restratifiées d'après les réponses données le même mois, puis rééchantillonnées en vue d'un nouveau dénombrement en septembre, en décembre et en mars.

Le NASS se sert d'un plan d'échantillonnage à deux phases et de l'estimateur de développement repondéré dans le cadre de son enquête sur l'usage des produits agrochimiques à la ferme. On commence par identifier les exploitations qui produisent certaines cultures, puis on quantifie l'emploi de pesticides avec ces cultures.

Le présent article montre que si on peut utiliser l'estimateur jackknife pour estimer la variance de l'estimateur de développement repondéré dans certaines conditions, cette méthode ne s'avère pas très efficace, en général, pour estimer la variance de l'estimateur de développement double. À la partie 2, il est question de l'estimateur de développement repondéré et son erreur quadratique moyenne. À la partie 3, on verra que l'estimateur de

[1] Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030; Diana M. Stukel, Division des méthodes d'enquête des ménages, Statistique Canada, Ottawa, (Ontario), Canada, K1A OT6.

variance jackknife ne présente presque aucun biais à l'égard de l'estimateur de variance repondéré, tandis que la partie 4 expose les lacunes de l'estimateur jackknife en tant qu'estimateur de la variance de l'estimateur de développement double. À la partie 5, on trouvera une étude de simulation qui semble confirmer les grandes hypothèses des parties antérieures. La partie 6 aborde les applications de l'estimateur de développement repondéré et la partie 7 sert de conclusion. L'annexe résume le cadre asymptotique hypothétique utilisé comme point de départ et fournit des éléments de preuve.

2. L'ESTIMATEUR DE DÉVELOPPEMENT REPONDÉRÉ

2.1 L'estimateur

Soit $h(= 1, ..., H)$, les strates de la première phase d'un échantillon aléatoire en grappes stratifié, obtenu par tirage non exhaustif; n_h le nombre de grappes de la strate h échantillonnées et F_h l'ensemble des grappes. Soit $g(= 1, ..., G)$, la strate de la deuxième phase d'où on prélève par tirage exhaustif un sous-échantillon aléatoire simple stratifié. Un élément de la grappe prélevé p fois lors de la première phase donne p éléments distincts pour le sous-échantillon. Soit M_g le nombre d'éléments dans g avant le sous-échantillonnage et m_g le nombre d'éléments sous-échantillonnés dans g. Dans la pratique, les strates de la deuxième phase G sont rarement définies avant prélèvement de l'échantillon de la première phase.

Soit S_g l'ensemble des éléments dans g avant le sous-échantillonnage; s_g le jeu d'éléments sous-échantillonnés dans g; s, l'ensemble complet d'éléments sous-échantillonnés et $m = \sum_g m_g$ la taille du sous-échantillon. Enfin, soit y_i la valeur à laquelle on s'intéresse pour l'élément i et w_i le facteur de développement à la première phase de i (c'est-à-dire, la valeur inverse de la probabilité de sélection de la grappe renfermant i).

En supposant le dénombrement de tous les éléments de l'échantillon de la première phase, pour estimer la population totale T, on recourrait à l'estimateur

$$t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i y_i. \tag{1}$$

Soit *l'estimateur de développement repondéré* de T,

$$t_2 = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} (M_g/m_g) w_i y_i}{\sum_{i \in s_g} (M_g/m_g) w_i} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} \right\}. \tag{2}$$

Une autre façon d'écrire t_2 est

$$t_2 = \sum_{g=1}^{G} \sum_{i \in s_g} a_i y_i = \sum_{i \in s} a_i y_i, \tag{3}$$

où

$$a_i = \left[\sum_{k \in S_g} w_k \Big/ \sum_{k \in s_g} w_k \right] w_i \text{ pour } i \in s_g$$

correspond au *poids pondéré* de l'élément i. L'équation (3) explique le nom donné à l'estimateur (estimateur de développement repondéré).

2.2 Erreur quadratique moyenne de l'estimateur (un peu de théorie)

En général, t_2 donne une estimation biaisée de T. Toutefois, sous de légères conditions, précisées en annexe, l'estimateur de T est cohérent avec le plan d'échantillonnage. En d'autres termes, $\text{plim}_{m \to \infty} (t_2 - T)/T = 0$ (Isaki et Fuller 1982). Dans notre article, on supposera simplement que m_g est élevé.

Notons que

$$E[(t_2 - T)^2] = E[(\{t_1 - T\} + \{t_2 - t_1\})^2]$$

$$\approx \text{Var}_1(t_1) + E_1\{E_2[(t_2 - t_1)^2]\},$$

où les indices de Var et de E indiquent la phase de l'échantillonnage. Étant donné la valeur élevée de m_g, $E_2[t_1(t_2 - t_1)] = t_1 E_2(t_2 - t_1) \approx 0$. En outre, $E(t_2 - T) = E_1[E_2(t_2 - T)] \approx 0$, et l'erreur quadratique moyenne de t_2 correspond en réalité à sa variance (asymptotique).

Puisqu'on a procédé à un échantillonnage non exhaustif à la première phase, $\text{Var}_1(t_1)$ peut en principe être estimé au moyen de l'estimateur suivant:

$$v_{L1} = \sum_{h=1}^{H} (n_h/[n_h - 1])$$

$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i y_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i y_i \right]^2 / n_h \right), \tag{4}$$

où U_{hj} correspond à l'ensemble des éléments de la grappe j tirée de la strate h à la première phase. L'indice L est utilisé pour des raisons historiques, pour indiquer qu'il y a «linéarisation», même s'il n'y a rien à linéariser dans le cas actuel. Notons que quand on effectue un deuxième échantillonnage, il s'avère généralement impossible de calculer v_{L1}, dans la pratique.

Par conséquent,

$$t_2 - t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i \left\{ \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} - \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i r_i}{\sum_{i \in s_g} w_i},$$

où

$$r_i = y_i - \sum_{k \in S_g} w_k y_k \Big/ \sum_{k \in S_g} w_k \text{ pour } i \in S_g.$$

Dans l'argumentation subséquente, il est capital de se rappeler que r_i a été défini afin que $\sum_{i \in S_g} w_i r_i = 0$ pour toutes les valeurs de g.

Si on poursuit,

$$t_2 - t_1 \approx \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i, \tag{5}$$

puisque $\sum_{i \in S_g} w_i \approx \sum_{i \in s_g} (M_g/m_g) w_i$ (voir l'équation (A1) de l'annexe). Il s'ensuit que

$$E_2[(t_2 - t_1)^2] \approx \text{Var}_2 \left\{ \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i \right\}$$

$$= \sum_{g=1}^{G} (M_g^2/[\{M_g - 1\} m_g])(1 - m_g/M_g)$$

$$* \left\{ \sum_{i \in S_g} (w_i r_i)^2 - \left(\sum_{i \in S_g} w_i r_i \right)^2 / M_g \right\}$$

$$\approx \sum_{g=1}^{G} ([M_g/m_g] - 1) \left\{ \sum_{i \in S_g} (w_i r_i)^2 \right\}. \tag{6}$$

Précisons que l'équation (6) *tient compte* des corrections pour la population finie qui résultent de l'échantillonnage de la deuxième phase.

3. L'ESTIMATEUR DE VARIANCE JACKKNIFE

3.1 L'estimateur de variance

Le moment est venu de parler de l'estimateur jackknife. Pour $j \in F_h$, soit la répétition $t_{(hj)2}$ de l'estimateur jackknife

$$t_{(hj)2} = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} \right\}, \tag{7}$$

où

$$w_{hji} = \begin{cases} w_i n_h/(n_h - 1) & \text{quand } i \in U_{hj'} \text{ et } j' \neq j \\ 0 & \text{quand } i \in U_{hj} \\ w_i & \text{quand } i \in U_{h'j'} \text{ et } h' \neq h. \end{cases}$$

De même, on peut dire que

$$t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} y_i.$$

Selon Rust (1985), l'*estimateur de variance jackknife* $v_{Jf} (f = 1 \text{ or } 2)$, se définit simplement par

$$v_{Jf} = \sum_{h=1}^{H} (n_h - 1)/n_h \sum_{j \in F_h} (t_{(hj)f} - t_f)^2. \tag{8}$$

Krewski et Rao (1981, équation (2.4)) dénotent cette forme $v_J^{(2)}$. On peut démontrer aisément que $v_{J1} = v_{L1}$.

3.2 Pourquoi l'estimateur fonctionne (un peu plus de théorie)

Nous verrons bientôt que v_{J2} estime presque sans biais la variance de l'estimateur de développement repondéré de l'équation (2). Rao et Shao (1992) parviennent indirectement à la même conclusion (notre équation (2) correspond à l'espérance de l'estimateur qu'ils présentent à la partie 3.3, pp. 818-819). Dans leurs travaux cependant, ces auteurs considèrent la non-réponse comme une phase d'échantillonnage supplémentaire où on recourt à l'échantillonnage de Poisson (Särndal et coll. 1992, p. 85) plutôt qu'à un échantillonnage aléatoire simple stratifié. Dans la démonstration de Rao et Shao, chaque élément prélevé à la première phase constitue en réalité une strate de deuxième phase. La quasi-absence de biais pour v_{J2} se résume donc au cas particulier d'un résultat signalé par Krewski et Rao (1981), (Rao et Shao (1992), p. 821).

Par «strate de deuxième phase», nous entendons les classes de repondération de Rao et Shao (1992). On présume que les éléments d'une classe donnée présentent la même probabilité inconnue de réponse ou de sélection. L'échantillonnage de Poisson équivaut à un échantillonnage aléatoire simple stratifié, *conditionnellement à* la taille du sous-échantillon obtenu à l'intérieur de la classe de repondération. Dans leurs travaux, Rao et Shao (1992) utilisent une approche *inconditionnelle*.

Revenant au problème qui nous intéresse, on remarque que

$$t_{(hj)2} - t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} \left\{ \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} - \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} r_{hji}}{\sum_{i \in s_g} w_{hji}} \right\},$$

où

$$r_{hji} = y_i - \sum_{k \in S_g} w_{hjk} y_k \Big/ \sum_{k \in S_g} w_{hjk} \quad \text{pour} \quad i \in s_g.$$

Sous réserve de légères conditions (voir les équations (A2) et (A3) de l'annexe), on obtient l'équation que voici, analogue à l'équation (5):

$$t_{(hj)2} \approx t_{(hj)1} + \sum_{g=1}^{G} (M_g/m_g) \sum_{i \in s_g} w_{hji} r_{hji} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + [M_g/m_g] c_i r_{hji}), \quad (9)$$

où c_i est une variable indicatrice égale à 1 quand i fait partie du sous-échantillon et a la valeur nulle dans les autres cas.

Poursuivons,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + \{[M_g/m_g] c_i - 1\} r_{hji}) = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_{hji}, \quad (10)$$

où $z_{hji} = y_i + \{[M_g/m_g] c_i - 1\} r_{hji}$. Une fois encore, puisque la valeur de m_g est toujours élevée, on peut raisonnablement supposer que $r_{hji} \approx r_i$ (voir l'équation (A4) de l'annexe). Par conséquent,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_i,$$

où $z_i = y_i + \{[M_g/m_g] c_i - 1\} r_i$. Pour les mêmes raisons, $t_2 \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_i z_i$. Puisque t_2 est linéaire dans z_i,

$$v_{J2} \approx v_{L1}(\sum \sum w_i z_i) = \sum_{h=1}^{H} (n_h/[n_h - 1]) * \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i z_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right]^2 / n_h \right). \quad (11)$$

Soit $e_i = M_g/m_g$, le facteur de pondération à la deuxième phase pour $i \in S_g$. On constate que c_i est une variable aléatoire, $E(c_i) = m_g/M_g$ et que $E(c_i c_k) = (m_g/M_g)(m_g - 1)/(M_g - 1)$ pour $i, k \in S_g$, $i \neq k$.

Il s'ensuit que

$$E_2\left[\left(\sum_{i \in U_{hj}} w_i z_i \right)^2 \right] \approx \left(\sum_{i \in U_{hj}} w_i y_i \right)^2 + \sum_{i \in U_{hj}} (e_i - 1)(w_i r_i)^2 - \sum_{g=1}^{G} \sum_{\substack{i, k \in S_g \cap U_{hj} \\ i \neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (12)$$

Pareillement, en appelant F_h^* l'ensemble des éléments venant des grappes tirées de la strate h du premier degré avant le sous-échantillonnage, on obtient

$$E_2\left[\left(\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right)^2 \right] = E_2\left[\left(\sum_{i \in F_h^*} w_i z_i \right)^2 \right] \approx (\sum_{i \in F_h^*} w_i y_i)^2 + \sum_{i \in F_h^*} (e_i - 1)(w_i r_i)^2 - \sum_{g=1}^{G} \sum_{\substack{i, k \in S_g \cap F_h^* \\ i \neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (13)$$

Dans l'annexe, on suppose que le dernier terme des équations (12) et (13) est négligeable, sous réserve de légères conditions. Par conséquent,

$$E_2(v_{J2}) \approx v_{J1} + \sum_{h=1}^{H} \sum_{i \in F_h^*} (e_i - 1)(w_i r_i)^2 = v_{J1} + \sum_{g=1}^{G} \sum_{i \in S_g} ([M_g/m_g] - 1)(w_i r_i)^2 \approx v_{L1} + E_2[(t_2 - t_1)^2], \quad (14)$$

qui, à son tour, implique que v_{J2} donne une estimation presque non biaisée de $E[(t_2 - T)^2]$.

4. L'ESTIMATEUR DE DÉVELOPPEMENT DOUBLE

L'estimateur de développement double est une solution de rechange à t_2, et se présente comme suit:

$$t_3 = \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i y_i. \quad (15)$$

La répétition jackknife de t_3 ne se définit pas clairement. Une simple possibilité serait

$$t_{(hj)3} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{hji} (M_g/m_g) y_i. \quad (16)$$

Une autre, qui se rapproche peut-être davantage d'une véritable «répétition», est

$$t_{(hj)3}^* = \sum_{g=1}^{G} \sum_{i \in s_g} w_{hji} (M_{ghj}/m_{ghj}) y_i, \quad (17)$$

où M_{ghj} représente le nombre d'éléments de l'échantillon de la première phase (plus exactement d'une grappe de l'échantillon de la première phase) qu'on trouve dans S_g mais *pas* dans U_{hj}. Parallèlement, m_{ghj} correspond au nombre d'éléments de l'échantillon de la deuxième phase qu'on trouve dans s_g mais *pas* dans U_{hj}. À partir de contre-exemples, nous verrons l'annexe, qu'aucune variante de la répétition ne donne d'estimateur de variance jackknife (v_{J3} de l'équation (8)) non biaisé de façon asymptotique, en général.

5. ÉTUDE DE SIMULATION DE MONTE CARLO

5.1 Conception de l'étude

Les résultats qu'on a pu examiner jusqu'ici sont asymptotiques. Nous avons effectué une étude de simulation de Monte Carlo afin d'évaluer la précision de l'estimateur jackknife en tant qu'estimateur de la variance de l'estimateur de développement repondéré dans un univers fini. Parallèlement, nous avons évalué la précision des deux estimateurs jackknife proposés pour l'estimateur de développement double à la partie 4.

Nous nous sommes servis des données de l'Enquête sur la population active (EPA) canadienne de décembre 1990 pour la province de Terre-Neuve. De cette population finie, nous avons tiré des échantillons répétitifs. L'EPA est la plus vaste enquête-ménage par sondage poursuivie en permanence par Statistique Canada. Les données sur le marché du travail sont recueillies mensuellement grâce à un plan d'échantillonnage complexe à degrés multiples, comportant plusieurs niveaux de stratification. On trouvera plus de précisions sur ce plan d'échantillonnage avant la modification qu'il a subie en 1991 dans Singh, Drew, Gambino et Mayda (1990), ainsi que dans Stukel et Boyer (1992). En bref, les provinces sont stratifiées en «régions économiques», vastes régions à structure économique analogue; Terre-Neuve en compte quatre. Les régions économiques sont subdivisées en strates de niveau inférieur. À Terre-Neuve, le niveau de stratification le plus bas donnait 45 strates comprenant chacune moins de six grappes ou *unités primaires d'échantillonnage* (UPE), ce qui était insuffisant pour l'échantillonnage dans le cadre de la simulation. On a donc regroupé les 45 strates en 18, comprenant chacune 6 à 18 UPE. Les régions économiques ont été préservées lors du regroupement des strates, tout comme on a maintenu les régions métropolitaines de recensement de St. John's et de Cornerbrook.

Dans le cadre de l'étude de Monte Carlo, on a prélevé $R = 4\,000$ échantillons de la «population» de Terre-Neuve (composée de 9 152 individus), selon le plan d'échantillonnage à deux phases que voici: on a d'abord tiré deux UPE de chaque strate de la première phase par échantillonnage aléatoire simple (EAS) *non exhaustif* (NE). On a ainsi obtenu au total 36 UPE. Tous les ménages des UPE sélectionnées à la première phase (et les personnes composant ces ménages) ont été retenus, ce qui a donné un échantillon en grappes exhaustif à la première phase. À la deuxième phase, tous les éléments précédemment sélectionnés (les sujets, en comptant chaque personne choisie deux fois dans une UPE comme deux sujets distincts) ont été restratifiés en cinq groupes d'âge (< = 14, 15-24, 25-44, 45-64, > = 65) et les éléments de l'échantillon de la deuxième phase (lire les sujets) ont été prélevées par EAS *exhaustif* dans chacune des cinq strates de la deuxième phase.

Nous avons varié la taille de l'échantillon des strates à la deuxième phase en prenant m_g = 5, 10, 20, et 50, de manière à obtenir des échantillons au deuxième degré de taille m = 25, 50, 100 et 250. Quand le nombre de sujets échantillonnés à la première phase faisant partie d'une strate à la deuxième phase était inférieur à la valeur m_g désirée, notre intention était d'établir $m_g = M_g$, mais le cas ne s'est jamais présenté.

Une populaire règle heuristique applicable à «l'estimateur par le quotient distinct» comme l'estimateur de développement repondéré de l'équation (2) est que chaque strate de la deuxième phase comporte au moins 20 éléments (lire, à ce sujet, Särndal, Swensson et Wretman 1992, p. 270). Notre but, en attribuant les valeurs 5 et 10 à m_g, était de vérifier l'utilité d'une telle règle.

Nous avons envisagé deux paramètres intéressants: T_y, soit le nombre total de personnes occupées, et T_y / T_z le taux d'emploi. Dans le cas présent, $T_y = \sum_{i \in U} y_i$, ou $y_i = 1$ quand le sujet i a un emploi et a la valeur nulle dans les autres cas. De même, $T_z = \sum_{i \in U} z_i$, ou $z_i = 1$ quand le sujet i fait partie de la population active (c.-à-d. travaille ou chôme) et a la valeur nulle dans les autres cas. Pour chacun des $R = 4\,000$ échantillons, nous avons calculé l'estimateur de développement repondéré (EER) t_2 de l'équation (2), l'estimateur de développement double (EED) t_3 de l'équation (15) et l'estimateur de développement intégral de la première phase (EEIPD) t_1 de l'équation (1). Quoique ces estimateurs soient définis en fonction d'un total (nombre de personnes occupées), il est facile d'en étendre l'application à un rapport de totaux (au taux d'emploi, par exemple).

Pour chacun des $R = 4\,000$ échantillons de la deuxième phase, nous avons calculé la variance jackknife qui correspondait à l'estimateur de développement repondéré et à l'estimateur de développement double de l'équation (8), pour $f = 2$ et $f = 3$ respectivement. En ce qui concerne l'estimateur de développement double, nous avons testé les répétitions décrites aux équations (16) et (17), que nous appellerons respectivement variantes 1 et 2.

Nous avons aussi établi l'estimateur de variance jackknife qui correspondait à l'estimateur intégral de la première phase pour chacun des $R = 4\,000$ échantillons de la première phase, aux fins de comparaison. On obtient cet estimateur avec l'équation (8), quand $f = 1$.

Nous avons étudié diverses propriétés de fréquence des estimateurs précités et de l'estimateur de variance jackknife qui y correspond. Ces propriétés apparaissent ci-dessous. Pour plus de simplicité, elles ne sont exprimées qu'en fonction du nombre total estimatif de personnes occupées.

Le biais relatif en pour cent du nombre estimé de personnes occupées par rapport à la population globale est estimé par

$$\mathrm{BRP}(t^*) = \{[E_M(t^*)/T_y] - 1\} \times 100, \qquad (18)$$

où

$$E_M(t^*) = (1/4\ 000) \sum_{r=1}^{4\,000} t_r^*$$

représente l'espérance de Monte Carlo de l'estimateur ponctuel t^* applicable aux 4 000 échantillons. La valeur t^* peut correspondre à t_1, t_2, ou t_3, alors que t_r^* est la valeur t^* de l'échantillon r.

Le biais relatif en pour cent de l'estimateur de variance jackknife par rapport à l'erreur quadratique moyenne réelle est estimé par

$$\text{BRP}[v_{Jf}(t^*)] =$$

$$(\{E_M[v_{Jf}(t^*)] - \text{EQM}_{\text{vraie}}\}/\text{EQM}_{\text{vraie}}) \times 100, \qquad (19)$$

où

$$E_M[v_{Jf}(t^*)] = (1/4\ 000)\sum_{r=1}^{4\ 000} v_{Jfr}(t^*),$$

$$\text{EQM}_{\text{vraie}} = (1/4\ 000)\sum_{r=1}^{4\ 000} (t_r^* - T_y)^2,$$

et $v_{Jfr}(t^*)$ est la valeur de $v_{Jf}(t^*)$ de l'échantillon r.

Le coefficient de variation (en pour cent) de l'estimateur de variance jackknife par rapport à l'EQM/réelle est estimé par:

$$\text{CV}[v_{Jf}(t^*)] =$$

$$(\{(1/4\ 000)\sum [v_{Jfr}(t^*) - \text{EQM}_{\text{vraie}}]^2\}^{1/2}/\text{EQM}_{\text{vraie}}) \times 100; \qquad (20)$$

bref, la racine de l'erreur quadratique moyenne estimée de l'estimateur de variance, divisée par l'EQM réelle estimée et exprimée en pourcentage.

5.2 Résultats de l'étude

Le tableau 1A indique le biais relatif estimé en pour cent des trois estimations ponctuelles du nombre total de personnes occupées selon l'équation (18). Le tableau 1B en fait autant mais pour le taux d'emploi. Tous les biais ont une valeur absolue inférieure à 1 %.

Tableau 1A
Biais relatif en pour cent des estimations ponctuelles du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	0,14	-0,3	-0,29	-0,56
EED	–	0,16	-0,01	0,03	0,115
EIPD	0,04	–	–	–	–

Tableau 1B
Biais relatif en pour cent des estimations ponctuelles du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	-0,09	-0,31	-0,19	-0,26
EED	–	-0,08	-0,27	-0,12	-0,13
EIPD	-0,09	–	–	–	–

EER - Estimation de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)

Ni l'estimation de Monte Carlo de l'erreur quadratique moyenne (c.-à-d. la valeur $\text{EQM}_{\text{vraie}}$ ni les coefficients de variation correspondants, obtenus grâce à l'estimateur de développement double ou à sa version repondérée, n'apparaissent dans les tableaux car l'article porte principalement sur *l'estimation* de l'erreur quadratique moyenne. L'erreur quadratique moyenne (et les coefficients de variation) qui dérive de l'application des deux estimateurs est comparable, peu importe la taille de l'échantillon (l'écart relatif entre les coefficients de variation correspond à peu près à la moitié de l'écart relatif entre les erreurs quadratiques moyennes). L'estimateur de développement repondéré s'avère légèrement plus efficace lorsqu'il s'agit d'estimer le nombre total de personnes occupées (à savoir, quand $m_g = 5$, l'erreur quadratique moyenne de l'estimateur de développement double augmente de 17 %). Quand on estime le taux d'emploi, l'écart entre l'erreur quadratique moyenne des deux méthodes est inférieur à 1 %. On ne sera guère surpris d'apprendre que l'erreur quadratique moyenne des estimateurs augmente à mesure que la taille de l'échantillon de la deuxième phase diminue.

Tableau 2A
Biais relatif en pour cent de l'estimateur de variance jackknife du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	-0,99	-2,51	-5,81	-5,13
EED (Variante 1)	–	46,35	68,24	78,18	86,22
EED (Variante 2)	–	101,59	278,44	654,99	1997,51
EIPD	0,94	–	–	–	–

Tableau 2B
Biais relatif en pour cent de l'estimateur de variance jackknife du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	-3,53	-3,45	-7,09	-6,55
EED (Variante 1)	–	-2,46	-1,53	-5,21	-7,41
EED (Variante 2)	–	-0,36	4,91	9,09	30,46
EIPD	2,08	–	–	–	–

EER - Estimateur de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)
La variante 1 utilise la répétition de l'estimateur jackknife de l'équation (16).
La variante 2 utilise la répétition de l'estimateur jackknife de l'équation (17).

Le tableau 2A présente le biais relatif estimé en pour cent de l'estimateur de variance jackknife pour le nombre

total de personnes occupées selon l'équation (19), tandis que le tableau 2B en fait autant pour le taux d'emploi. Commençons par examiner le premier. On se rend compte que la variance de l'estimateur intégral de la première phase est presque totalement dépourvue de biais (0,94 %). L'estimateur jackknife donne de bons résultats avec l'estimateur de développement repondéré alors que la variance ne présente qu'un léger biais négatif, toujours inférieur à −6 %. Ce biais a tendance à devenir plus négatif (même si ce n'est pas de façon uniforme) à mesure que la taille de l'échantillon de la deuxième phase diminue.

Les deux variantes de l'estimateur jackknife de l'estimateur de développement double, en revanche, donnent de piètres résultats, avec un fort biais positif pour la variance, allant de 46,35 % à 1997,51 %! La deuxième variante est pire que la première, mais les deux, se comportent d'une manière absolument inacceptable.

Le tableau 2B reprend l'analyse pour l'estimation par quotient du taux d'emploi. Les résultats ont de quoi surprendre. En effet, tous les estimateurs de variance se comportent raisonnablement bien, sauf la variante 2 de l'estimateur de développement double quand $m_g = 5$. Outre ce cas, où il atteint 30,46 %, le biais est inférieur à 10 % en valeur absolue.

Dans l'ensemble, les tableaux 2A et 2B appuient fortement l'usage de l'estimateur de variance jackknife avec l'estimateur de développement repondéré, même avec un très petit échantillon de la deuxième phase. Par contre, cet estimateur échoue lamentablement avec l'estimateur de développement double quand on estime les totaux. Il arrive cependant que la variante 1 donne des résultats acceptables, selon l'estimateur et les données.

Bien que la majorité des études insistent sur le *biais* des estimateurs de variance, il vaut la peine d'examiner le *coefficient de variation* des estimateurs de variance pour établir la stabilité des estimations de la variance. Le coefficient de variation estimé (en pour cent) se rapportant au nombre total de personnes occupées et au taux d'emploi apparaît respectivement aux tableaux 3A et 3B. L'expression sous la racine carrée du numérateur à l'équation (20) donne l'EQM de la variance, composée de la valeur quadratique du biais de la variance et de la variance de la variance. Les tableaux 3A et 3B ne présentent pas les valeurs correspondantes des entrées des tableaux 2A et 2B (signalées par un *) pour lesquelles le biais de la variance est trop élevé (supérieur à 20 %, par exemple), car il est clair que ces valeurs seront elles aussi trop élevées. Au tableau 3A, les coefficients de variation estimés associés à l'estimateur de développement repondéré fluctuent entre 46,86 % et 53,42 %, ce qui est caractéristique aux estimateurs de la variance. Des coefficients de variation aussi importants ont été relevés dans d'autres études de simulation sur la variance, notamment celle de Kovačević et Yung (1997). En l'occurrence, on remarquera que les coefficients de variation estimés des estimateurs intégraux de la première phase se situent dans la même fourchette de valeurs. En réalité, ils dépassent légèrement ceux des estimateurs de la deuxième phase.

Tableau 3A
Coefficient de variation de la variance jackknife du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	51,33	49,3	46,86	53,42
EED (Variante 1)	–	*	*	*	*
EED (Variante 2)	–	*	*	*	*
EIPD	56,71	–	–	–	–

Tableau 3B
Coefficient de variation de la variance jackknife du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	59,28	65,66	74,26	103,06
EED (Variant 1)	–	59,24	66,16	72,89	99,1
EED (Variant 2)	–	60,94	73,2	92,71	*
EIPD	78,42	–	–	–	–

EER - Estimateur de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)
La variante 1 utilise la répétition de l'estimateur jackknife de l'équation (16).
La variante 2 utilise la répétition de l'estimateur jackknife de l'équation (17).

Si on les examine un à un, on se rend compte que les coefficients de variation de la variance du taux d'emploi estimé qui apparaissent au tableau 3B sont plus élevés que les coefficients correspondants du tableau 3A. D'autre part, tous les estimateurs se remarquent par une hausse appréciable du coefficient de variation correspondant quand la taille de l'échantillon de la deuxième phase diminue. L'effet est plus prononcé pour les estimateurs par quotient que pour les estimateurs du total. Les coefficients de variation très importants de la colonne $m_g = 5$ aux deux tableaux ne surprendra personne puisque la taille globale de l'échantillon de la deuxième phase (25) est en fait inférieure au nombre d'UPE prélevées à la première phase de l'échantillonnage (36). Le nombre de membres de la population active échantillonnés (c.-à-d. au dénominateur) constitue d'ailleurs un meilleur dénombrement de l'échantillon pour l'estimateur par quotient. Cette valeur varie d'un échantillon à l'autre et est souvent considérablement inférieure à 25.

6. EXTENSION DE L'ESTIMATEUR DE DÉVELOPPEMENT REPONDÉRÉ

6.1 L'estimateur de développement repondéré

Élaborer un estimateur de variance linéarisé pour l'estimateur de développement repondéré de l'équation (2)

ne s'avère pas très difficile. Supposons cependant que le plan d'échantillonnage compte plus de deux phases ou qu'on désire estimer le quotient de deux totaux. Quoi qu'elle demeure réalisable en pareil cas, la linéarisation gagne de plus en plus en difficulté. Il n'en va pas autant avec l'estimateur jackknife.

On peut aisément généraliser les résultats de la partie 3 pour un échantillonnage à p-phases par induction. La lettre h désigne toujours les strates de la première phase, mais la lettre g correspond désormais à celles de la phase p-ième représente le jeu d'éléments de l'échantillon de la S_g phase de la strate g, alors que s_g est le sous-échantillon de la p-ième phase de g. On remplace la valeur w_i de l'équation (2) par a_i de (3), pour l'estimateur de la (p-1)-ième phase. De même, on calcule la valeur $t_{(hj)2}$ de l'estimateur jackknife avec a_{hji} de la (p-1)-ième phase, au lieu de w_{hji}.

Remplacer l'échantillon en grappes stratifié prélevé à la première phase par un échantillon stratifié à plusieurs phases s'avère aussi assez simple (nous laissons au lecteur le soin de le faire). On obtient encore les résultats de la partie 3 pourvu que l'échantillon à plusieurs phases soit toujours prélevé par tirage non exhaustif à la première phase.

Enfin, il n'est pas difficile d'étendre les résultats de la partie 3 à des estimateurs plus complexes. Soit $\boldsymbol{U}_2$, un vecteur des estimateurs de l'équation (2) adoptant la forme t_2. L'erreur quadratique moyenne d'un estimateur quelconque $\Theta = g(\boldsymbol{U}_2)$, où g est une fonction continue, peut être estimée presque sans biais grâce à l'estimateur jackknife, chaque fois qu'on peut en faire autant pour les éléments de $\boldsymbol{U}_2$. Cette remarque respecte les preuves données dans les ouvrages. Ainsi, Rao et Wu (1985) examinent le plan asymptotique où toutes les valeurs n_h sont bornées, tandis que Wolter (1985; chapitre 4.5) analyse le cas où n_h augmente considérablement de façon arbitraire.

6.2 Régression à la deuxième phase

On peut généraliser l'estimateur t_2 par l'estimateur de régression:

$$t_{2\text{reg}} = \sum_{i\in S} w_i \boldsymbol{x}_i \left(\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' y_i \right), \quad (21)$$

où S représente l'échantillon original; $\boldsymbol{x}_i$, un vecteur ligne; d_i, une grandeur scalaire et où il existe un vecteur ligne $\boldsymbol{\gamma}$ tel que $d_i \boldsymbol{\gamma} \boldsymbol{x}_i' = 1$ pour toutes les valeurs de i. Dans la pratique, d_i est habituellement égal à 1 pour toutes les valeurs de i. Une exception survient fréquemment quand $\boldsymbol{x}_i = x_i$ et $d_i = 1/x_i$. Dans l'équation (2), $d_i = 1$ pour toutes les valeurs de i, et $\boldsymbol{x}_i$ correspond à un vecteur G de valeur 1 à la g-ième position mais de valeur nulle ailleurs, pour $i \in S_g$.

Soit

$$r_i = y_i - \boldsymbol{x}_i \left(\sum_{i\in S} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i\in S} w_i d_i \boldsymbol{x}_i' y_i \right).$$

La répétition $t_{2\text{reg}(hj)}$ a une forme identique à $t_{2\text{reg}}$, mais w_{hji} est remplacé par w_i. De même, r_{hji} a la même forme que r_i, si ce n'est que w_{hji} se substitue à w_i. Remarquons que e_i ne change pas dans $t_{2\text{reg}}$ et $t_{2\text{reg}(hj)}$.

Puisqu'il n'y a pas eu modification du plan d'échantillonnage, l'équation (6) ne change pas, si ce n'est que désormais $(\sum_{i\in S_g} w_i r_i)^2$ est non négatif au lieu d'être strictement égal à zéro. L'intéressé pourra s'assurer que les équations (10) à (13) gardent leur forme actuelle. Dans l'équation (14), on note que, si biais de l'estimateur jackknife il y a, celui-ci tend (approximativement) à la hausse. Bref, il s'agit d'un estimateur *conservateur* de la variance. Encore une fois, le lecteur est prié de se reporter à l'annexe (équations (A6) à (A9)) pour se faire une meilleure idée des hypothèses asymptotiques.

Le biais de l'estimateur jackknife disparaît quand $\sum_{i\in S_g} w_i r_i = 0$ pour toutes les valeurs de g. Pareille situation survient lorsqu'il existe G vecteurs de rangée $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_G$ de sorte que $d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' = 1$ quand $i \in S_g$ et 0 prend la valeur nulle dans les autres cas (puisque $\sum_{i\in s} w_i r_i = \sum_{i\in s} d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' w_i r_i = \boldsymbol{\gamma}_g \sum_{i\in s} w_i d_i \boldsymbol{x}_i' r_i = \boldsymbol{\gamma}_g \{\sum_{i\in s} w_i d_i \boldsymbol{x}_i' (y_i - \boldsymbol{x}_i [\sum_{i\in s} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i]^{-1} \sum_{i\in s} w_i d_i \boldsymbol{x}_i' y_i)\} = 0$). L'existence de $\boldsymbol{\gamma}_g$ quand $d_i = 1$, signifie qu'un membre de $\boldsymbol{x}_i$ est une variable indicatrice égale à 1 lorsque $i \in S_g$ et a la valeur nulle dans les autres cas, ou qu'un membre de la transformation linéaire de $\boldsymbol{x}_i$ est cette variable indicatrice.

7. CONCLUSION

Notre article avait principalement pour but de montrer qu'un simple estimateur de variance jackknife peut être presque dépourvu de biais lorsque la méthode d'estimation s'articule sur un échantillonnage à deux phases, pourvu qu'on recoure à un estimateur de développement repondéré plutôt qu'à un estimateur de développement double. L'application pratique des résultats théoriques de l'estimateur de développement repondéré dépendra du contexte puisque ces résultats reposent sur une argumentation asymptotique. L'étude de simulation de Monte Carlo que nous avons effectuée donne néanmoins à penser que l'estimateur jackknife a son utilité pour estimer la variance de l'estimateur de développement repondéré, même en présence de strates étonnamment peu importantes à l'échantillonnage de la deuxième phase, c'est-à-dire de strates qui ne comptent que 5 ou 10 éléments.

ANNEXE

Cohérence de l'estimateur de développement repondéré au niveau du plan d'échantillonnage

Pour vérifier la cohérence théorique de t_2 dans l'équation (2), on suppose simplement que le plan d'échantillonnage et la population de y_i sont tels que

$$\left\{ \sum_{g=1}^{G} (M_g/m_g) \sum_{i\in s_g} w_i y_i / T \right\} - 1 = O_p(1/\sqrt{m}),$$

et que, pour *tout* échantillon de première phase,

$$\left(\sum_{k\in S_g} w_k \Big/ \sum_{k\in s_g} w_k\right)(m_g/M_g) - 1 = O_p(1/\sqrt{m}) \qquad \text{(A1)}$$

pour toutes les valeurs de g. Ces hypothèses justifient l'équation (5) dans le corps du texte.

L'analyse présume que G est borné et que chaque valeur de m_g présente le même degré asymptotique que m. La chose n'est réalisable que lorsqu'on définit S_g *après* prélèvement de l'échantillon de la première phase. Sans cela, M_g équivaudrait à une variable aléatoire, si bien qu'on ne pourrait garantir une valeur minimale pour m_g, pour tous les échantillons envisageables de la première phase. En principe, on suppose qu'un mécanisme permet de déterminer S_g et les fractions de l'échantillonnage de la deuxième phase, compte tenu d'un échantillon quelconque de la première phase. Les valeurs exactes de G et de m_g, en revanche, peuvent être arrêtées avant le prélèvement de l'échantillon de la première phase, sans que cela soit toutefois une obligation.

Remarque au sujet du cadre asymptotique

Nous avons montré que l'estimateur jackknife intègre une composante permettant d'estimer la variance à la deuxième phase $E_2[(t_2 - t_1)^2])$ sans introduire de biais asymptotique, *quel que soit* l'échantillon de la première phase (voir l'équation (14)). Par voie de conséquence, cette composante permet d'estimer la variance moyenne à la deuxième phase (donc non conditionnelle) de tous les échantillons possibles de la première phase $E_1\{E_2[(t_2 - t_1)^2]\}$, sans introduire de biais asymptotique.

Nous nous sommes éloignés du cadre décrit ci-dessus dans le travail empirique afin que les résultats soient plus faciles à résumer. Plus précisément, nous avons défini S_g au préalable et laissé M_g varier. Advenant le cas où l'échantillon de la première phase donnerait une valeur M_g inférieure à la valeur m_g désirée (50, par exemple) à la strate de la deuxième phase, nous avions l'intention de retenir tous les sujets de S_g pour constituer l'échantillon de la deuxième phase. La présence de cette strate g de la deuxième phase n'augmenterait donc pas l'erreur quadratique moyenne (ou biais) de t_2 et les hypothèses asymptotiques sur m_g s'avéreraient superflues. Ainsi qu'on a pu le voir, M_g n'a jamais obtenu une valeur inférieure à 50 dans la simulation. Quoi qu'il en soit, on disposait d'une règle applicable aux fractions de l'échantillonnage de la deuxième phase, pour tous les échantillons de la première phase.

Répétitions de l'estimateur jackknife

Deux cadres asymptotiques distincts (au moins) s'appliquent à l'échantillon de la première phase. Le premier comprend un nombre arbitrairement élevé de strates à la première phase, la taille de chacune étant limitée; bref, pour chacune d'elles, $1/n_h = O(1)$ tandis que $1/H = O(1/m)$. Dans le deuxième cas, toutes les strates de la première phase sont arbitrairement importantes, soit $1/n_h = O(1/m)$. On suppose que chaque grappe renferme O(1) éléments dans les deux cas, bref que chaque grappe est bornée.

Puisque m_g est du même ordre asymptotique que m, il est raisonnable de penser que dans l'un ou l'autre cas, pour un échantillon donné de la première phase,

$$\sum_{i\in S_g} w_{hji} \Big/ \sum_{i\in S_g} w_i - 1 = O_p(1/m), \qquad \text{(A2)}$$

$$\sum_{i\in s_g} w_{hji} \Big/ \sum_{i\in s_g} w_i - 1 = O_p(1/m), \qquad \text{(A3)}$$

ce dont on peut se servir pour dériver l'équation (9). De même, on présume que pour tout échantillon de la première phase

$$\sum_{i\in S_g} w_{hji}y_i \Big/ \sum_{i\in S_g} w_i y_i - 1 = O_p(1/m), \qquad \text{(A4)}$$

ce qui donne $r_{hji} - r_i = O_p(1/m)$.

Équations (12), (13) et (14)

Le nombre d'éléments dans chaque grappe étant limité, par B par exemple, le troisième terme de l'équation (12) compte au plus GB^2 termes, un nombre fini.

Chaque terme est d'ordre $1/m_g$ (plus exactement, la probabilité qu'un terme soit asymptotiquement d'ordre supérieur à $1/m_g$ est égale à zéro). Par conséquent, on peut négliger la deuxième ligne de l'équation (12), sur le plan asymptotique.

L'équation (14) se vérifie chaque fois que $1/n_h = O(1)$, car si n_h est inférieur à C (par exemple), le troisième terme de droite de l'équation (13) correspond à la somme d'un maximum de $G(BC)^2$ termes, un nombre fini. Cette fois encore, chaque terme est d'ordre $1/m_g$. On peut donc ignorer la deuxième ligne de l'équation (13) sur un plan asymptotique.

Supposons d'autre part, que chaque rapport $1/n_h$ soit égal à $O(1/m)$. On présumera que le plan d'échantillonnage et la population sont tels que, pour un échantillon quelconque à la première phase,

$$A_h = \sum_{i\in F_h^*} w_i(e_i c_i - 1)r_i \Big/ \sum_{i\in F_h^*} w_i y_i = O_p(1/\sqrt{m}) \qquad \text{(A5)}$$

pour toutes les valeurs de h. Il s'agit d'une hypothèse raisonnable puisque, conditionnellement à l'échantillon de la première phase, le dénominateur de A_h représente le total d'un domaine – soit la somme de $w_i y_i$ pour les éléments de F_h^*. Par conséquent, il correspond à $O(m)$ (sans perte de généralité, on peut supposer que chaque w_i est égal à O(1)). Le numérateur de A_h indique l'écart entre l'estimateur de développement (somme des éléments $w_i e_i c_i r_i$ dans F_h^*) d'un échantillon aléatoire simple stratifié et sa cible (la somme des éléments $w_i r_i$ dans F_h^*). L'équation (A.5) repose sur la modeste hypothèse que le plan d'échantillonnage et la population donnent une différence de $O_p(\sqrt{m})$ pour tous les échantillons envisageables de la première phase.

En vertu de l'hypothèse (A5), $\sum_{i\in F_h^*} w_i z_i = \sum_{F_h^*} w_i y_i(1 + A_h)$ équivaut à peu près à $\sum_{i\in F_h^*} w_i y_i$, si bien

que $E_2[(\sum_{i\in F_h^*} w_i z_i)^2]/n_h \approx (\sum_{i\in F_h^*} w_i y_i)^2/n_h$. L'équation (14) provient de cette quasi-égalité et sur les équations (11) et (12) n_h étant élevé, $n_h/(n_h - 1) \approx 1$).

Contre-exemples de l'estimateur jackknife de l'estimateur de développement double

Comme contre-exemple de la forme répétée de l'équation (16), prenons le cas où chaque grappe ne renferme qu'un élément, $H = G = 1$, et où y_i est toujours égal à 1. Dans ce cas, $t_3 = T$, et t_3 n'a pas de variance. Malheureusement, $t_{(1j)3} = T[n_1/(n_1 - 1)](m - 1)/m$ quand $j \in s$ et $Tn_1/(n_1 - 1)$ dans les autres cas. Donc, $(t_{(1j)3} - T)/T = O_p(1/m)$. Le rapport v_{J3}/T^2 qui dérive de $t_{(1j)3}$ serait lui aussi égal à $O(1/m)$, puisqu'il s'agit de la somme des termes n_1 d'ordre $O(1/m^2)$.

Bien que v_{J3}/T^2 corresponde à $O(1/m)$, v_{J3} ne se rapproche pas assez de zéro pour nous être utile. En effet, si y_i était toujours égal à $N(1,1)$, la variance relative de t_3 serait $1/m$, qui correspond aussi à $O(1/m)$. Pour que v_{J3} soit presque égal à zéro, v_{J3}/T^2 devrait donc être inférieur à $O(1/m)$. Cela n'étant pas le cas, l'estimateur de variance jackknife est loin de ne pas être biaisé.

Comme contre-exemple de la forme répétée de l'équation (17), examinons le cas où chaque grappe renferme de nouveau un seul élément et où y_i est égal à un, mais où $H = m$, $G = 1$, la population de h est toujours égale à N_0, $n_h = 2$ pour toutes les valeurs de h, et $M_1 = 2m$. Il s'ensuit que $T = t_3 = mN_0$, si bien que t_3 ne présente pas de variance. La répétition $t^*_{(hj)3}$ peut donc prendre quatre valeurs. Si $hj \in s$ et $hj' \in s (j \neq j')$, alors, $t^*_{(hj)3} = [(m/2)(2m - 1)/(m - 1)]N_0$. Si $hj \in s$ et $hj' \notin s$, alors, $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/(m - 1)]N_0$. Si $hj \notin s$ et $hj' \in s$, alors, $t^*_{(hj)3} = [(m/2)(2m - 1)/m]N_0$. Si $hj \notin s$ et $hj' \notin s$, alors, $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/m]N_0$. Dans aucun de ces cas, $(t^*_{(hj)3} - T)/T = O_p(1/m)$, de sorte que l'estimateur de variance jackknife ne peut être presque dépourvu de biais.

Estimateur de régression de la deuxième phase

Pour étayer l'argumentation sur l'estimateur de régression de l'équation (21), supposons que le plan d'échantillonnage et la population soient tels qu'on obtient confirmation des relations asymptotiques que voici. En premier lieu,

$$\sum_{i\in S} w_i \boldsymbol{x}_i (\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i)^{-1} d_i \boldsymbol{x}_i' - 1 = O_p(1/\sqrt{m}), \quad \text{(A6)}$$

qui est une généralisation de l'équation (A1). De même, les équations (A2) et (A3) peuvent être généralisées pour donner

$$\sum_{i\in S_g} w_{hji} d_i q_i \Big/ \sum_{i\in S_g} w_i d_i q_i - 1 = O_p(1/m), \quad \text{(A7)}$$

et

$$\sum_{i\in s_g} w_{hji} e_i d_i q_i \Big/ \sum_{i\in s_g} w_i e_i d_i q_i - 1 = O_p(1/m) \quad \text{(A8)}$$

pour toutes les valeurs de q_i, où q_i est un élément de la matrice $\boldsymbol{x}_i' \boldsymbol{x}_i$. Enfin, l'équation (A4) se généralise pour devenir

$$\sum_{i\in S_g} w_{hji} d_i p_i \Big/ \sum_{i\in S_g} w_i d_i p_i - 1 = O_p(1/m) \quad \text{(A9)}$$

pour toutes les valeurs de p_i, où p_i représente un élément de la matrice $\boldsymbol{x}_i' y_i$.

BIBLIOGRAPHIE

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOVAČEVIĆ, M.S., et YUNG, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique. *Techniques d'enquête*, 23, 47-59.

KREWSKI, D., et RAO, J.N.K. (1981), Inferences from stratified samples: properties of linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. *Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies*, (Éds. W.G. Madow, I. Olkin, et D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4, 811-822.

RAO, J.N.K., et WU, C.F.J. (1985). Inferences from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada:* 1984-1990. N° 71-526 au catalogue, Statistique Canada.

STUKEL, D.M., et BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Direction de la méthodologie, document de travail, SSMD, 92-009E, Statistique Canada.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.