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

Article

Application de la méthode des répliques des différences successives pour estimer les variances

par Stephen Ash

Juin 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 53-67
Statistique Canada, N° 12-001-X au catalogue

Application de la méthode des répliques des différences successives pour estimer les variances

Stephen Ash[1]

Résumé

Fay et Train (1995) présentent une méthode qu'ils nomment *successive difference replication*, c.-à-d. répliques des différences successives, qui peut être utilisée pour estimer la variance d'un total estimé au moyen d'un échantillon aléatoire systématique tiré d'une liste ordonnée. L'estimateur prend la forme générale d'un estimateur de variance par rééchantillonnage, où les facteurs de rééchantillonnage sont construits de manière à imiter l'estimateur par différences successives. Cet estimateur est une modification de celui présenté dans Wolter (1985). Le présent article étend la méthodologie en expliquant l'effet de l'attribution des lignes de matrice sur l'estimateur de variance, en montrant comment un jeu réduit de répliques mène à un estimateur raisonnable et en établissant les conditions pour que la méthode des répliques des différences successives soit équivalente à l'estimateur par différences successives.

Mots-clés : Différences successives; répliques des différences successives; échantillonnage aléatoire systématique.

1 Introduction

Fay et Train (1995) présentent une méthode qu'ils nomment *successive difference replication* (SDR), c.-à-d. répliques des différences successives, qui peut être utilisée pour estimer la variance d'un total estimé au moyen d'un échantillon aléatoire systématique tiré d'une liste ordonnée. L'estimateur prend la forme générale de l'estimateur de variance par rééchantillonnage, où les facteurs de rééchantillonnage sont construits de manière à imiter l'estimateur par différences successives (SD).

L'article décrit l'établissement et l'utilisation de nouveaux concepts en vue de mieux comprendre la méthodologie proposée au départ par Fay et Train (1995), ci-après appelés F et T. Ces nouveaux concepts aident à expliquer l'effet de l'attribution des lignes de matrice sur l'estimateur de variance, à montrer comment un jeu réduit de répliques mène à un estimateur raisonnable, et à établir les conditions pour que la méthode des répliques des différences successives soit équivalente à l'estimateur par différences successives. Nous espérons qu'en étant mieux comprise, la SDR semblera moins mystérieuse et sera donc plus facile à utiliser par toute personne qui souhaite estimer les variances dans le cas d'un échantillonnage aléatoire systématique.

L'article débute par un examen de l'estimateur SD et de la façon dont il convient à l'estimation de la variance des échantillons aléatoires systématiques. La partie principale est consacrée à l'introduction de deux théorèmes qui fournissent les conditions pour que l'estimateur SDR soit équivalent à l'estimateur SD. L'article se termine par la présentation d'exemples empiriques en vue d'examiner l'effet de différents schémas d'affectation des lignes de matrice et de montrer qu'il est approprié d'utiliser un ensemble réduit de répliques.

1. Stephen Ash, U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233. Courriel : stephen.eliot.ash@census.gov.

Dans la suite de l'exposé, l'abréviation *sys* sera utilisée pour désigner l'échantillonnage aléatoire systématique à partir d'une liste ordonnée. Nous utilisons l'abréviation *sys* parce que l'on peut montrer que l'échantillonnage systématique à partir d'une liste non ordonnée ou d'une liste ordonnée aléatoirement est équivalent à l'échantillonnage aléatoire simple (Madow et Madow 1944). Pour les besoins de notre discussion, nous nous concentrons uniquement sur la sélection équiprobabiliste et sur les méthodes de sélection d'un échantillon dans une seule dimension. Le lecteur trouvera d'excellents résumés de l'échantillonnage *sys* et de l'estimation des variances sous *sys* dans Iachan (1982), Wolter (1985, chapitre 7), Murthy et Rao (1988), et Bellhouse (1988).

1.1 Revue de la méthode des différences successives

Wolter (1984; estimateur 2) donne un estimateur par différences successives de la variance d'une moyenne estimée $(\bar{y})$ sous un plan de sondage *sys* de la forme

$$\hat{v}_{\mathrm{SD1}}(\hat{\bar{y}}) = (1 - f)\frac{1}{2n(n-1)}\sum_{k=2}^{n}(y_k - y_{k-1})^2,$$

où y_k est la variable d'intérêt, k indice les unités de l'échantillon ordonné, et $f = n / N$ est la fraction d'échantillonnage. La statistique d'intérêt est Y ou le total de y_k sur l'univers d'intérêt, et $\hat{Y}$ est un estimateur de Y. Soit N et n la taille de l'univers et de l'échantillon, respectivement. La moyenne de y_k et son estimateur sont définis comme étant $\bar{y} = Y/N$ et $\hat{\bar{y}}$, respectivement. Nous définissons aussi l'estimateur du total Y comme étant $\hat{Y} = \sum_{k=1}^{n} \breve{y}_k$, où la variable d'intérêt pondérée par des poids égaux est $\breve{y}_k = (N/n)\, y_k$; pour des poids de sondage inégaux w_k, elle est définie comme étant $\breve{y}_k = w_k y_k$. L'estimateur $\hat{v}_{\mathrm{SD1}}(\hat{\bar{y}})$ a été décrit par Yates (1953; pages 229 à 231) et recommandé par Wolter (1984). Murthy et Rao (1988, équation 32) donnent un aperçu des raisons pour lesquelles l'estimateur fonctionne. La version abrégée est que, puisque sous échantillonnage *sys* une seule unité est sélectionnée dans chaque strate implicite, la solution de l'estimateur SD consiste à fusionner les strates implicites adjacentes. Avec deux unités, nous pouvons estimer la variance d'une strate implicite. Après fusion des strates implicites, la moyenne est calculée sur toutes les paires possibles, puis multipliée par n, le nombre de strates implicites, pour donner la variance de toutes les strates implicites.

F et T donnent un estimateur de variance SD d'un total sous échantillonnage *sys* de la forme

$$\hat{v}_{\mathrm{SD1}}(\hat{Y}) = (1 - f)\frac{n}{2(n-1)}\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2.$$

Wolter (1985, équation 7.7.4) définit le même estimateur où $w_k = (np_k)^{-1}$ et p_k est la probabilité de sélection avec remise de l'unité k. F et T définissent un deuxième estimateur SD

$$\hat{v}_{\mathrm{SD2}}(\hat{Y}) = \frac{1}{2}(1 - f)\left[\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2 + (\breve{y}_n - \breve{y}_1)^2\right],$$

qui est « circulaire » en ce sens qu'il inclut une différence au carré supplémentaire qui relie les première et dernière unités de la liste triée.

Nous exprimons l'estimateur SD2 de manière plus générale sous une forme quadratique $\breve{\mathbf{y}}'\mathbf{C}\breve{\mathbf{y}}$, où $\breve{\mathbf{y}}' = [\breve{y}_1\breve{y}_2 \ldots \breve{y}_n]$ est défini comme le vecteur d'observations pondérées de dimension $n \times 1$ et $\mathbf{C}$ est une matrice carrée dont tous les éléments de la diagonale principale valent 2, tous les éléments de la diagonale supérieure et de la diagonale inférieure valent -1, et l'élément inférieur gauche et l'élément supérieur droit valent -1. Ici, les diagonales supérieures sont définies comme étant les diagonales adjacentes à la diagonale principale, excepté dans le cas d'une matrice de dimensions 2×2.

2 Répliques des différences successives

2.1 Définition de la méthode des répliques des différences successives

F et T présentent une méthode qu'ils nomment *successive difference replication* (SDR), c.-à-d. répliques des différences successives, qui permet d'estimer la variance sous échantillonnage *sys* en imitant $\hat{v}_{\mathrm{SD2}}(\hat{Y})$, ce qui signifie que l'estimateur SDR est équivalent ou quasi équivalent à $\hat{v}_{\mathrm{SD2}}(\hat{Y})$. Nous montrons comment la méthode SDR peut être appliquée pour produire les facteurs et les poids de rééchantillonnage pour un estimateur de variance par rééchantillonnage général qui est équivalent à l'estimateur SD2. Avant de définir l'estimateur SDR dans le premier théorème, nous établissons certains termes et fournissons un lemme qui est utilisé dans le théorème.

Un schéma d'attribution de lignes, ou plus simplement schéma AL, correspond à l'attribution de deux lignes d'une matrice à chaque unité de l'échantillon. Nous désignons habituellement la paire de lignes par (a_i, b_i) pour l'unité i. Une boucle connectée est un schéma AL qui ne répète aucune des lignes, c.-à-d. $a_i \neq a_j$ et $b_i \neq b_j$ pour tous i et j dans la boucle connectée, et qui est circulaire, c.-à-d. $b_i = a_{i+1}$ pour tout $i < n$ et $b_n = a_1$. Un exemple de boucle connectée pour trois observations est (1,2), (2,3), (3,1).

Une matrice de décalage $\mathbf{S}$ peut être utilisée pour déplacer les lignes ou les colonnes d'une matrice. Nous expliquons le processus de déplacement des lignes, qui est similaire au processus de déplacement des colonnes. Une matrice de décalage est une matrice carrée dont tous les éléments valent 0, à l'exception d'une valeur 1 unique dans chaque colonne. Si nous voulons déplacer la ligne p jusqu'à la ligne q, nous plaçons une valeur 1 dans la q^{e} ligne de la p^{e} colonne et des 0 ailleurs. Nous insistons sur le fait que l'ordre est important lorsqu'on applique une matrice de décalage à une autre matrice. L'application de $\mathbf{S}$ à une autre matrice carrée $\mathbf{A}$ sous la forme $\mathbf{AS}$ déplace les colonnes de $\mathbf{A}$, mais sous la forme $\mathbf{SA}$, elle déplace les lignes de $\mathbf{A}$.

Lemme : Soit $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_c$ les matrices de décalage, alors $\mathrm{bloc}(\mathbf{S}_1'\mathbf{S}_1, \mathbf{S}_2'\mathbf{S}_2, \ldots, \mathbf{S}_C'\mathbf{S}_C) = \mathbf{I}$.

Preuve. Nous commençons par définir une matrice diagonale par blocs générale $\mathbf{A}$ qui est formée par les matrices carrées $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ comme

$$\mathbf{A} = \mathrm{bloc}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_C \end{bmatrix}.$$

On peut montrer que, si $\mathbf{A}$ et $\mathbf{B}$ sont toutes deux des matrices diagonales par blocs et que les matrices carrées $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ ont les mêmes dimensions que $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_C$, respectivement, alors $\mathbf{AB} = \text{bloc}(\mathbf{A}_1\mathbf{B}_1, \mathbf{A}_2\mathbf{B}_2, \ldots, \mathbf{A}_C\mathbf{B}_C)$. Pour une matrice de décalage donnée, nous savons aussi que $\mathbf{S}'\mathbf{S} = \mathbf{I}$, puisque le décalage d'une ligne vers le bas d'une matrice de décalage est $\mathbf{I}$. Le lemme découle des deux éléments qui précèdent.

Nous définissons aussi une matrice de décalage d'une ligne comme étant une matrice de décalage qui décale toutes les lignes d'une autre matrice d'une ligne vers le bas et transfère la dernière ligne à la première ligne, ou qui décale toutes les lignes d'une autre matrice d'une ligne vers le haut et transfère la première ligne à la dernière ligne. Si $\mathbf{S}_D$ est une matrice de décalage d'une ligne qui déplace les lignes vers le bas, tous les éléments de la diagonale supérieure et l'élément inférieur gauche de la matrice ont une valeur de 1, par exemple $\mathbf{S}_1$. De même, si $\mathbf{S}_U$ est une matrice de décalage d'une ligne qui déplace les lignes vers le haut, tous les éléments de la diagonal inférieure et l'élément supérieur droit de la matrice ont une valeur de 1, par exemple la matrice $\mathbf{S}_2$ subséquemment définie. Notons la propriété que $\mathbf{S}_D = \mathbf{S}'_U$ et $\mathbf{S}_U = \mathbf{S}'_D$; donc, $\mathbf{S}_U + \mathbf{S}'_U = \mathbf{S}_D + \mathbf{S}'_D$. Nous présentons maintenant le théorème principal de l'article qui établit les conditions sous lesquelles l'estimateur SDR est équivalent à l'estimateur SD2.

Théorème 1 : Soit n la taille d'un échantillon *sys* donné et $\breve{\mathbf{y}}' = [\breve{y}_1 \breve{y}_2 \ldots \breve{y}_n]$, le vecteur d'observations pondérées de dimension $n \times 1$, où l'ordre des observations reflète l'ordre de tirage de l'échantillon *sys*.

(a) Choisir une matrice de Hadamard d'ordre k $(\mathbf{HH}' = k\,\mathbf{I})$, où $n \le k$.

(b) Choisir un schéma d'attribution de lignes (AL) qui assigne deux lignes (a_i, b_i) à chaque unité i de l'échantillon. Poser que le schéma AL définit C boucles connectées c contenant chacune m_c unités.

(c) Choisir les $m = n$ lignes de $\mathbf{H}$ correspondant au schéma AL pour créer la matrice $\mathbf{M}$ de dimensions $m \times k$. L'ordre des lignes de $\mathbf{M}$ doit correspondre à la première ligne du schéma AL. Par exemple, la première ligne de $\mathbf{M}$ doit être la ligne $a_{i=1}$ de $\mathbf{H}$, la deuxième ligne doit être la ligne $a_{i=2}$ de $\mathbf{H}$, *etc*. Ensuite, définir la matrice de décalage de dimensions $m \times m$ comme étant $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_C)$, où les matrices de décalage d'une ligne $\mathbf{S}_c$ de dimensions $m_c \times m_c$ sont définies en vue d'identifier la position de la deuxième ligne b_i du schéma AL dans $\mathbf{M}$. En général, chaque matrice de décalage $\mathbf{S}_c$ sera une matrice de décalage vers le haut, une matrice de décalage vers le bas ou une matrice de décalage de dimensions 2×2 (voir la matrice $\mathbf{S}_4$ subséquemment définie).

Définir l'estimateur du total r pour chaque réplique comme $\hat{Y}_r = \sum_{i=1}^{n} f_{i,r}\breve{y}_i$, où la matrice des facteurs de rééchantillonnage est $\mathbf{F} = \mathbf{1}_m\mathbf{1}'_k + \left(2^{-3/2}\mathbf{I}_m - 2^{-3/2}\mathbf{S}\right)\mathbf{M}$ et les valeurs individuelles dans la matrice sont définies pour chaque unité i (lignes de $\mathbf{F}$) de la réplique r (colonnes de $\mathbf{F}$) comme étant $f_{i,r} = 1 + 2^{-3/2}h_{a_i,r} - 2^{-3/2}h_{b_i,r}$. $\mathbf{I}_m$ est une matrice identité de dimensions $m \times m$ et $\mathbf{1}_m$ est un vecteur de dimension $m \times 1$ de 1. Alors, l'estimateur de variance SDR $\hat{v}_{\text{SDR}}\left(\hat{Y}\right) = (1 - f)\,4/k \sum_{r=1}^{m}\left(\hat{Y}_r - \hat{Y}\right)^2$ est équivalent à la somme des C différents estimateurs SD2.

Preuve. L'estimateur SDR peut s'écrire en notation matricielle sous la forme

$$(1-f)\frac{4}{k}\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m\mathbf{1}_k'+\left(2^{-3/2}\mathbf{I}_m-2^{-3/2}\mathbf{S}\right)\mathbf{M}\right)-\breve{\mathbf{y}}'\mathbf{1}_m\mathbf{1}_k'\right)\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m\mathbf{1}_k'+\left(2^{-3/2}\mathbf{I}_m-2^{-3/2}\mathbf{S}\right)\mathbf{M}\right)-\breve{\mathbf{y}}'\mathbf{1}_m\mathbf{1}_k'\right)'$$

$$=(1-f)\frac{4}{k}\left(2^{-3/2}\right)^2\breve{\mathbf{y}}'(\mathbf{I}_m-\mathbf{S})\mathbf{M}\mathbf{M}'(\mathbf{I}_m-\mathbf{S})'\breve{\mathbf{y}}$$

Comme $\{\text{lignes de } \mathbf{M}\} \subseteq \{\text{lignes de } \mathbf{H}\}$, on peut montrer que $\mathbf{M}\mathbf{M}' = k\,\mathbf{I}$. Partant de ce résultat, la variance devient

$$(1-f)\frac{1}{2k}\breve{\mathbf{y}}'(\mathbf{I}_m-\mathbf{S})(k\mathbf{I}_m)(\mathbf{I}_m-\mathbf{S})'\breve{\mathbf{y}}=\frac{1}{2}(1-f)\breve{\mathbf{y}}'(\mathbf{I}_m-\mathbf{S})(\mathbf{I}_m-\mathbf{S})'\breve{\mathbf{y}}$$

$$=\frac{1}{2}(1-f)\breve{\mathbf{y}}'(2\mathbf{I}_m-\mathbf{S}-\mathbf{S}')\breve{\mathbf{y}}$$

La dernière ligne découle du lemme et a une valeur constante pour tout choix de $\mathbf{H}$. En notant la structure diagonale par blocs de $\mathbf{S}$, nous pouvons écrire l'estimateur sous la forme

$$\frac{1}{2}(1-f)\sum_{c=1}^{C}\breve{\mathbf{y}}_c'(2\mathbf{I}_m-\mathbf{S}_c-\mathbf{S}_c')\breve{\mathbf{y}}_c,$$

où $\breve{\mathbf{y}}_c$ correspond au vecteur des observations pondérées dans la boucle connectée c, qui est un résultat de la partition du vecteur d'observations pondérées pour donner $\breve{\mathbf{y}}' = [\breve{\mathbf{y}}_{c=1}\breve{\mathbf{y}}_{c=2}\ldots\breve{\mathbf{y}}_{c=C}]$. Le choix du schéma AL ne modifie pas le résultat, puisque nous savons que $2\mathbf{I}_m-\mathbf{S}_c-\mathbf{S}_c'$ est constant pour une matrice de décalage d'une ligne vers le haut ou vers le bas $\mathbf{S}_c$.

Note 1 : Le théorème 1 définit l'estimateur SDR en fonctions des facteurs de rééchantillonnage, mais nous pouvons aussi l'exprimer en fonction des poids de rééchantillonnage sous la forme

$$(1-f)\frac{4}{k}\mathbf{y}'(\mathbf{W}-\mathbf{1}_m\mathbf{1}_k')(\mathbf{W}-\mathbf{1}_m\mathbf{1}_k')'\mathbf{y}.$$

Ici, $\mathbf{W}$ est la matrice de dimensions $m \times k$ des poids de rééchantillonnage définie comme étant $\mathbf{W} = \mathbf{w} * \mathbf{F}$, où $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ est le vecteur de poids de sondage pour les n unités de l'échantillon et l'opérateur $*$ multiplie les éléments du vecteur $\mathbf{w}$ par chacune des colonnes de $\mathbf{F}$, c.-à-d. que, si $W_{i,r}$ et w_i sont des entrées de $\mathbf{W}$ et $\mathbf{w}$, respectivement, les entrées de $\mathbf{W}$ sont définies comme étant $W_{i,r} = w_i \times f_{i,r}$.

Note 2 : Huang et Bell (2009) définissent similairement l'estimateur SDR sous une forme quadratique et l'utilisent pour établir certaines propriétés générales de l'estimateur quand y_k est i.i.d.(μ, σ^2). Nous souhaitons interpréter la façon dont l'estimateur SDR fonctionne et la qualité de son fonctionnement. Définir la forme quadratique avec des matrices de décalage et des boucles connectées permet de mieux comprendre les attributions de lignes et l'efficacité de l'estimateur.

Pour un échantillon de grande taille, il n'est habituellement pas pratique d'utiliser une matrice $\mathbf{H}$ où $n < k$. Le deuxième théorème offre un moyen d'utiliser $\mathbf{H}$ en prenant $k < n$ pour produire une plus grande matrice de Hadamard $\tilde{\mathbf{H}}$ où $k \geq n$ qui résultera en un estimateur SDR équivalent à l'estimateur SD2. Le deuxième théorème étoffe et clarifie aussi les instructions données par F et T pour le cas où $n > k$. Dans leurs instructions, F et T utilisent le mot cycle pour désigner chaque tranche de $m_d \leq k$ unités de l'échantillon. Le théorème 2 n'impose pas de contraintes sur le schéma AL, mais suit à part cela les conditions établies par F et T.

Théorème 2 : Soit n la taille d'un échantillon sys donné.

(a) Choisir une matrice de Hadamard $\mathbf{H}_A$ d'ordre k_A, où $n > k_A$.

(b) Choisir un schéma AL qui assigne les lignes de $\mathbf{H}_A$ à l'échantillon. En gardant l'ordre original, répartir les n unités de l'échantillon en D cycles. Chaque cycle d comprend $m_d \leq k_A$ unités. Dans chaque cycle, le schéma AL définit une ou plusieurs boucles connectées.

(c) Choisir une matrice de Hadamard semi-normale $\mathbf{H}_B$ d'ordre k_B et l'utiliser pour définir une plus grande matrice de Hadamard $\tilde{\mathbf{H}}$ d'ordre $\tilde{k}$ générée à partir de la matrice $\mathbf{H}_A$ originale. Cela peut se faire en appliquant une construction de Welsch à $\mathbf{H}_A$, c.-à-d. $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$.

(d) Choisir les $m = \sum_{d=1}^{D} m_d$ lignes de $\tilde{\mathbf{H}}$ qui correspondent au schéma AL pour créer la matrice $\tilde{\mathbf{M}}$ de dimensions $m \times \tilde{k}$. L'ordre des lignes de $\tilde{\mathbf{M}}$ doit correspondre à la première ligne du schéma AL. Ensuite, définir la matrice de décalage de dimensions $m \times m$ comme étant $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_D)$ où les matrices $\mathbf{S}_d$ de dimensions $m_d \times m_d$ identifient la position de la deuxième ligne b_i du schéma AL dans $\tilde{\mathbf{M}}$.

Dans ces conditions, l'estimateur SDR est défini comme

$$\hat{v}_{\text{SDR}}\left(\hat{Y}\right) = (1 - f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}\left(\hat{Y}_r - \hat{Y}\right)^2$$

et est équivalent à la somme d'au moins D estimateurs SD2.

Preuve. Le résultat découle de l'application du théorème 1. La valeur particulière de D découle du fait que chacun des D cycles peut posséder une ou plusieurs boucles connectées, de manière à avoir un total d'au moins D boucles connectées.

Exemple 1 : Soit $n = 14$ et choisissons la matrice de Hadamard non normale $\mathbf{H}_A = \mathbf{H}_{4b}$ d'ordre $k_A = 4$. Le nombre de cycles est $D = 4$ et le schéma AL dans chaque cycle est donné dans la deuxième colonne du tableau 2.1 pour chaque unité. Définissons $\tilde{\mathbf{H}}$ d'ordre $\tilde{k} = 16$ en utilisant une construction de Welsh de la matrice de Hadamard normale originale comme il suit

$$\mathbf{H}_{16} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4b} = \begin{bmatrix} \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} \end{bmatrix}$$

où

$$\mathbf{H}_{4a} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \text{ et } \mathbf{H}_{4b} = \begin{bmatrix} 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix}.$$

En utilisant $\mathbf{H}_{16}$, nous pouvons calculer les facteurs de rééchantillonnage pour 16 répliques comme au tableau 2.1. En notation matricielle, $\tilde{\mathbf{M}}$ englobe toutes les lignes de $\tilde{\mathbf{H}} = \mathbf{H}_{16}$ sauf les lignes 13 et 16. Les lignes de $\tilde{\mathbf{M}}$ sont ordonnées par a_i, la première ligne assignée dans le schéma AL. La matrice de décalage est définie comme $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4)$, où les matrices de décalage correspondant à chaque cycle sont

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \mathbf{S}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_4 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Tableau 2.1
Matrice des facteurs de rééchantillonnage $(f_{i,r})$ pour l'exemple 1

Unité	AL	AL		Réplique															
#	$\mathbf{H}_A = \mathbf{H}_{4b}$	$\tilde{\mathbf{H}} = \mathbf{H}_{16}$	Cycle	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	(1,2)	(1,2)		1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0
2	(2,3)	(2,3)	1	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7
3	(3,4)	(3,4)		1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3
4	(4,1)	(4,1)		1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0
5	(1,3)	(5,7)		1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3
6	(3,1)	(7,5)	2	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7
7	(2,4)	(6,8)		0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0
8	(4,2)	(8,6)		1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0
9	(1,4)	(9,12)		1,0	0,3	1,7	1,0	1,0	0,3	1,7	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0
10	(4,3)	(12,11)	3	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3
11	(3,2)	(11,10)		1,7	1,0	1,0	0,3	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7
12	(2,1)	(10,9)		0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0
13	(2,3)	(14,15)	4	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7
14	(3,2)	(15,14)		1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3

Étant donné les facteurs de rééchantillonnage du tableau 2.1, l'estimateur SDR est équivalent à la somme de cinq estimateurs SD2 différents, un pour chaque boucle connectée du schéma AL, c.-à-d.

$$(1-f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}\left(\hat{Y}_r-\hat{Y}\right)^2=\frac{1}{2}(1-f)\begin{bmatrix}\sum_{i=2}^{4}(y_i-y_{i-1})^2+(y_4-y_1)^2+2(y_6-y_5)^2\\+2(y_8-y_7)^2+\sum_{i=10}^{12}(y_i-y_{i-1})^2+(y_{12}-y_9)^2\\+2(y_{13}-y_{13})^2\end{bmatrix}. \qquad (2.1)$$

Il convient de souligner quelques éléments concernant l'exemple 1. Premièrement, le nombre de répliques nécessaires est supérieur à la taille de l'échantillon. Cela se produit lorsque m_d n'est pas constant dans tous les cycles. Le quatrième cycle ne comprend que deux unités d'échantillon, mais nous avons dû utiliser quatre répliques de chaque $\mathbf{H}_{4b}$ parce qu'au moins un des cycles utilisait quatre lignes.

Pour rendre l'exemple plus intéressant, nous avons choisi une matrice de Hadamard non normale $\mathbf{H}_{4b}$ pour $\mathbf{H}_A$. Cette matrice de Hadamard non normale a été construite en partant de la matrice de Hadamard normale $\mathbf{H}_{4a}$ et en inversant la procédure décrite par Hedayat et Wallis (1978) pour trouver une matrice de Hadamard normale. Ici nous avons simplement changé le signe de tous les éléments de la deuxième ligne, puis nous avons changé le signe de tous les éléments de la deuxième colonne.

Si nous avions utilisé la matrice de Hadamard normale $\mathbf{H}_{4a}$ pour $\mathbf{H}_A$ ainsi que $\mathbf{H}_B$, les facteurs de rééchantillonnage pour les répliques 1, 5, 9 et 13 auraient tous été égaux à 1,0. Nous disons qu'une réplique est « morte » quand chaque élément reçoit une valeur de 1,0 et que l'estimation basée sur la réplique est donc égale à l'estimation originale. Dans l'estimateur SDR, les répliques mortes sont tout à fait valables et dues simplement à la façon dont les facteurs de rééchantillonnage sont répartis par la matrice de Hadamard. En cas de réplique morte, de nombreuses valeurs 1,0 se trouvent dans celle-ci, et la composition des autres répliques est plus mélangée, avec des valeurs de 1,7 et de 0,3. Cependant, toutes les répliques, même les répliques mortes, sont nécessaires pour l'estimation.

La valeur réelle du théorème 2 tient au fait qu'il permet de comprendre la prescription originale de F et T pour l'estimateur SDR quand $n > k$. Dans F et T, le schéma AL est appliqué de manière répétée aux $m = k - 1$ lignes de $\mathbf{H}_A$ (en sautant la première ligne de $\mathbf{H}_A$), où $\mathbf{H}_A$ est choisie comme une matrice de Hadamard normale. Les répliques sont ensuite formées en utilisant les k_A colonnes de $\mathbf{H}_A$. Si nous appliquons le cadre plus vaste du théorème 2, nous dirions qu'ils ont utilisé implicitement une matrice normale $\mathbf{H}_B$, qui donne $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$ et n'inclut que les k_A premières répliques dans l'estimateur de variance. Puisqu'un sous-ensemble des répliques nécessaires pour que l'estimateur SDR soit équivalent à l'estimateur SD2 est utilisé, nous disons que l'estimateur résultant est une approximation de l'estimateur SD2.

Exemple 1 (suite) : Si nous utilisons seulement les quatre premières répliques du tableau 2.1, l'estimateur SDR sera équivalent à (2.1) plus le terme de reste R qui est défini comme

$$R = \begin{bmatrix} (y_1 - y_2)(y_8 - y_7) + (y_1 - y_2)(y_{11} - y_{12}) + (y_1 - y_2)(-y_{13} + y_{14}) \\ +(y_8 - y_7)(y_{11} - y_{12}) + (y_8 - y_7)(y_{14} - y_{13}) + (y_{11} - y_{12})(y_{14} - y_{13}) \\ +(y_4 - y_3)(y_8 - y_7) + (y_4 - y_3)(y_{10} - y_9) + (y_8 - y_7)(y_{10} - y_9) \\ +(y_1 - y_4)(y_5 - y_6) + (y_1 - y_4)(y_9 - y_{12}) + (y_5 - y_6)(y_9 - y_{12}) \\ +(y_2 - y_3)(y_5 - y_6) + (y_2 - y_3)(y_{10} - y_{11}) + (y_2 - y_3)(y_{13} - y_{14}) \\ +(y_5 - y_6)(y_{10} - y_{11}) + (y_5 - y_6)(y_{13} - y_{14}) + (y_{10} - y_{11})(y_{13} - y_{14}) \end{bmatrix}$$

Notons que R comprend le même nombre de termes positifs et négatifs, qui ne s'annulent pas exactement, mais qui font que la valeur de R est habituellement proche de zéro. De même, utiliser les répliques 1 à $q \times k_A$, où $q = 1, 2, \ldots, k_B$, donne un reste R comprenant un nombre égal de termes positifs et de termes négatifs. Ce n'est qu'en utilisant toutes les répliques de $\tilde{\mathbf{H}}$ que le terme de reste R est nul.

Exemple 2 : La taille de l'échantillon mensuel de la Current Population Survey (CPS) est de $n = 72\,000$ ménages par mois (U.S. Census Bureau 2006). La CPS est réalisée selon un plan de sondage à deux degrés comprenant la sélection d'un échantillon de premier degré formé d'unités primaires d'échantillonnage (UPE), qui sont habituellement des comtés ou des groupes de comtés, puis le tirage de l'échantillon de deuxième degré de ménages à partir de l'échantillon d'UPE. Certaines UPE, généralement les régions métropolitaines, sont sélectionnées avec certitude, c.-à-d. que leur probabilité de sélection au premier degré est 1,0. Dans le cas des UPE sélectionnées avec certitude, l'échantillon sys peut être traité comme le plan de sondage de premier degré dans l'estimation de la variance, c.-à-d. que la méthode SDR est appliquée pour produire les répliques. Dans le cas des UPE sélectionnées sans certitude, la méthode des répliques équilibrées répétées (BRR pour *Balanced Repeated Replication*) [McCarthy 1966] est appliquée pour produire les répliques. Environ 75 % de l'échantillon ou 54 000 unités sont comprises dans les UPE autoreprésentatives, auxquelles est appliquée la méthode SDR.

L'application de la méthode SDR à la CPS comprend l'utilisation d'une matrice de Hadamard d'ordre $k = 160$ dont sont exclues deux lignes, c.-à-d. que $m = 158$. Les poids de rééchantillonnage sont produits pour 160 répliques. Même s'il peut sembler qu'il s'agit d'une conclusion logique du présent article, nous ne suggérons pas que l'on utilise pour la CPS une matrice de Hadamard d'ordre $k = 54\,000$ ni que l'on produise 54 000 jeux de poids de rééchantillonnage. Cela donnerait en effet un nombre irraisonnable de répliques. Nous sommes plutôt d'avis que le sous-ensemble de 160 répliques utilisé pour la CPS est grand et fournit par conséquent une approximation raisonnable de l'estimateur SD2. Plus loin, dans les exemples empiriques, nous examinons l'effet de l'utilisation d'un jeu réduit de répliques.

2.2 Attribution de lignes quand $n > k$

Jusqu'ici, nous avons supposé qu'un schéma AL était donné et nous n'avons pas discuté de la façon de générer ce schéma pour un échantillon particulier, où $n > k$. À la présente section, nous examinons deux schémas AL et formulons certains commentaires au sujet de l'attribution de lignes en général. Le premier schéma AL est similaire à celui décrit par Sukasih et Jang (2003) et est destiné à être utilisé quand $k < n$ et avec le théorème 2.

AL1 : Ce schéma AL attribue une paire de lignes a_i et b_i à chaque tranche de m_d unités de l'échantillon, que nous appelons cycle d, où $m_d \leq k$. Après $m_d - 1$ cycles, le schéma AL est répété jusqu'à ce qu'une paire de lignes ait été attribuée à chacune des unités de l'échantillon.

Étape 1 : Trier l'échantillon dans l'ordre dans lequel il était trié avant la sélection de l'échantillon.

Étape 2 : Initialiser le numéro du cycle par $d = 1$ et le nombre de boucles connectées par $c = 1$.

Étape 3 : Commencer l'AL au début d'un cycle ou d'une boucle connectée en prenant $a_1 = c$.

Étape 4 : Répéter le schéma AL suivant : $b_i = \mathrm{mod}(a_i + d, k)$ et $a_i = b_i$ jusqu'à ce que chacune des m_d lignes du cycle ait été utilisée ou que l'AL devienne une boucle connectée. Ici, la fonction modulo ou $\mathrm{mod}(a, b)$ est définie comme étant le reste de la division de a par b. Si les m_d lignes du cycle ont toutes été utilisées, commencer un nouveau cycle : poser que $d = d + 1$ et retourner à l'étape 3. Sinon (fin d'une boucle connectée, mais non la fin d'un cycle), commencer une nouvelle boucle connectée : poser que $c = c + 1$ et retourner à l'étape 3.

Étape 5 : À la fin de $d = m_d - 1$ cycles, recommencer au premier cycle – retourner à l'étape 2.

Le schéma AL1 possède les caractéristiques suivantes :

- Chacun des cycles $d = 1, 2, \ldots, m_d - 1$ de l'AL attribue m_d paires de lignes. Cela crée un total de $m_d (m_d - 1)$ paires de lignes.
- Le schéma d'AL se répète après $m_d - 1$ cycles. F et T suggèrent de redémarrer l'AL après 10 cycles. Nous recommandons d'utiliser chacun des $m_d - 1$ cycles avant de redémarrer l'AL.
- Les valeurs de a_i et b_i sont toujours espacées de c unités.
- Au milieu de la séquence, le schéma se répète en ordre inverse. Si m est un nombre pair, les cycles avant et après le $(m_d + 1)/2^{\mathrm{e}}$ cycle se répètent en ordre inverse.

Le schéma AL1 diffère de du schéma AL de Sukasih et Jang (2003), en ce sens que nous ne suggérons pas de sauter la ligne 1 ni de répéter le schéma AL après 10 cycles et nous n'exigeons pas que $k - 1$ soit un nombre premier. Premièrement, une ligne dont tous les éléments valent 1 peut paraître étrange, mais cela ne pose pas de problème. Comme dans le cas d'une colonne dont tous les éléments valent 1 dans $\mathbf{M}$, ce qui donne une réplique morte, une ligne ne contenant que des 1 n'aura d'effet que sur la distribution des facteurs de rééchantillonnage. Une unité i à laquelle a été attribuée la ligne 1 (soit $a_i = 1$ ou $b_i = 1$) possédera un plus grand nombre de facteurs de rééchantillonnage valant 1,0 qu'autrement. Cela n'est pas incorrect; il s'agit simplement de la façon dont les facteurs de rééchantillonnage sont distribués par $\mathbf{H}_A$. La deuxième différence est que nous suggérons de répéter l'attribution après m cycles, c'est-à-dire au moment où le schéma se répète, plutôt qu'après un nombre fixé de 10 cycles. Enfin, nous n'exigeons pas que $k - 1$ soit un nombre premier, mais notons que si $m_d = k - 1$ et que $k - 1$ est un nombre premier, il est garanti que chaque cycle ne possédera qu'une seule boucle connectée.

Nous fournissons un deuxième schéma AL plus facile à mettre en œuvre, appelé AL2, que nous comparons au schéma AL1 dans les exemples empiriques.

AL2 : Pas de mélange des attributions de lignes. Répéter la même AL simple toutes les m_d unités, c.-à-d. $(1,2),(2,3),\ldots,(m_d,1)$.

3 Exemples empiriques

Les exemples empiriques servent à examiner les questions suivantes :

Q1. Dans quelle mesure l'estimateur SDR donne-t-il de bons résultats quand on se sert d'un sous-ensemble de toutes les répliques nécessaires pour que l'estimateur SDR soit équivalent à l'estimateur SD?

Q2. Quel schéma d'attribution de lignes est le meilleur, AL1 ou AL2?

Q3. Devrions-nous utiliser un plus grand nombre ou un moins grand nombre de boucles connectées?

Pour répondre à ces questions, nous avons appliqué l'estimateur de variance SDR à plusieurs populations. Pour chaque population, nous avons sélectionné un échantillon *sys* de taille $n = 64$. Le tableau 3.1 décrit les trois estimateurs SDR étudiés.

Tableau 3.1
Estimateurs SDR pour les exemples empiriques

Estimateur	k_A	$\mathbf{H}_A$	k_B	$\mathbf{H}_B$
1	4	$\mathbf{H}_{4a}$	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$
2	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	4	$\mathbf{H}_{4a}$
3	64	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	1	1

Sous cette construction, les estimateurs SDR comprenaient $k_B = 1, 4$ ou 16 cycles, mais tous utilisaient la même matrice $\tilde{\mathbf{H}} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$, qui est la matrice de Hadamard normale d'ordre $\tilde{k} = 64$. Pour les trois estimateurs du tableau 3.1, nous avons également fait varier le schéma d'attribution de lignes (AL1 ou AL2), ainsi que le nombre de répliques utilisées par chaque estimateur, soit 16, 32, 48 ou 64. Tant avec AL1 qu'avec AL2, il n'existe qu'une seule boucle connectée par cycle, de sorte que le nombre de boucles connectées que possédaient les estimateurs 1, 2 et 3 était $k_B = 16, 4$ et 1, respectivement. En annexe, les résultats pour les estimateurs SDR sont résumés au tableau A1, tandis que le tableau A2 donne les résultats pour les estimateurs de variance SD1, SD2 et *eassr* appliqués aux fins de comparaison.

Jeux de données utilisés. Les populations « A » sont empruntées à l'exemple empirique de Wolter (1984). Pour les populations A1 à A7, nous avons généré 400 populations finies de taille $N = 64\,000$. Pour chaque population, il existait $b = 100$ échantillons possibles de taille $n = 64$. Les échantillons sont désignés par l'indice $i = 1, 2, \ldots, b = 100$ et, dans chaque échantillon, les unités sont désignées par l'indice $j = 1, 2, \ldots, n = 64$. Le tableau 3.2 résume comment la variable d'intérêt μ_{ij} est générée pour chacune des populations « A ».

Tableau 3.2
Description des populations artificielles de Wolter

Population	Description	n	b	μ_{ij}	e_{ij}
A1	Aléatoire	20	50	0	e_{ij} iid $N(0,100)$
A2	Tendance linéaire	20	50	$i + (j-1)k$	e_{ij} iid $N(0,100)$
A3	Effets de stratification	20	50	j	e_{ij} iid $N(0,100)$
A4	Effets de stratification	20	50	$j + 10$	$e_{ij} = \begin{cases} \varepsilon_{ij,} & \text{si } \varepsilon_{ij} \geq -(j+10) \\ -(j+10), & \text{autrement.} \end{cases}$ ε_{ij} iid $N(0,100), \rho = 0,8$
A5	Autocorrélée	20	50	0	$e_{ij} = \rho e_{i-1,j} + \varepsilon_{ij}$ $e_{i1} \sim N\left(0,100/\left(1-\rho^2\right)\right)$ ε_{ij} iid $N(0,100), \rho = 0,8$
A6	Autocorrélée	20	50	0	comme A5 avec $\rho = 0,4$
A7	Périodique	20	50	$20\sin\{2\pi/50[i+(j-1)k]\}$	e_{ij} iid $N(0,100)$

Mesures d'évaluation. Nous avons évalué les divers estimateurs de variance au moyen des trois mesures utilisées par Wolter, à savoir le biais relatif prévu (ERB pour *expected relative bias*), l'erreur quadratique moyenne relative (RMSE pour *relative mean squared error*) et le ratio de couverture. La première mesure, ERB, que nous avons utilisée pour examiner l'exactitude des estimateurs, est définie pour un estimateur donné θ comme $\mathrm{ERB}(\hat{v}_\theta) = E_m\left(E_p\left(\hat{v}_\theta - v\right)\right)/E_m(v)$. Dans notre notation, E_p et E_m désignent les espérances sous le plan et sous le modèle, respectivement. Pour examiner la variance des estimateurs, nous avons également mesuré la RMSE, qui est définie comme étant $\mathrm{RMSE}(\hat{v}_\theta) = E_m\left(E_p\left(\hat{v}_\theta - v\right)^2\right)/E_m(v)$. Nous avons calculé le ratio de couverture sous forme du pourcentage de fois que le vrai total de population était compris dans l'intervalle de confiance produit en utilisant l'estimation, c.-à-d. $\left(\hat{Y} - z_\alpha\sqrt{\hat{v}_\alpha}, \hat{Y} + z_\alpha\sqrt{\hat{v}_\alpha}\right)$. Ici, z_α est la valeur tirée d'une distribution normale qui a été choisie pour produire les intervalles de confiance à 95 %.

Résultats. En ce qui concerne Q1, les colonnes 4 à 7 du tableau A1 montrent que l'augmentation du nombre de répliques n'a qu'un effet minime sur le biais. Ce n'est que pour la population à tendance linéaire (A2) que l'estimateur SDR avec quatre boucles connectées présente une tendance cohérente de réduction du biais à mesure qu'augmente le nombre de répliques. Les autres combinaisons de population et d'estimateur ne révèlent aucune tendance décroissante ni croissante significatives lorsque le nombre de répliques augmente. Cette constatation représente un résultat positif, parce qu'elle indique que la réduction du jeu de répliques n'accroît pas le biais. Comme prévu, les RMSE dans les colonnes 8 à 11 du tableau A1 augmentent à mesure que le nombre de répliques diminue, mais curieusement, l'accroissement est relativement faible. De même, les intervalles de confiance présentés dans les colonnes 12 à 15 s'améliorent parallèlement à l'augmentation du nombre de répliques, sauf dans le cas des populations A2 et A7.

En ce qui concerne la question Q2, la comparaison des schémas AL1 et AL2 indique que l'estimateur SDR avec quatre boucles connectées produit habituellement de plus faibles biais (colonnes 4 à 7 du tableau A1) et variances (colonnes 8 à 11 du tableau A1) avec AL1 qu'avec AL2. Dans le cas de 16 boucles connectées, les biais et les variances sont similaires pour AL1 et AL2. Ces résultats laissent entendre que le biais et la variance sont tous deux améliorés, mais que l'effet est réduit à mesure que la taille des boucles connectées diminue.

En ce qui concerne Q3, les biais présentés dans les colonnes 4 à 7 diminuent lorsque le nombre de boucles connectées augmente. Fait exception la population périodique (A7). Lorsque les RMSE des estimateurs SD1 et SD2 ne sont pas similaires, comme dans le cas de la population à tendance linéaire (A2), l'augmentation du nombre de boucles connectées réduit également la RMSE. Ce résultat n'est pas étonnant. L'estimateur comprenant une seule grande boucle connectée est équivalent à l'estimateur SD2, de sorte qu'il peut présenter des biais et RMSE plus importants en raison du terme $(\hat{y}_1 - \hat{y}_{64})^2$. Dans l'autre sens, un plus grand nombre de boucles connectées réduit effectivement l'effet du terme $(\hat{y}_1 - \hat{y}_{64})^2$, de sorte que l'estimateur agit davantage comme l'estimateur SD1, dont le biais et la variance sont généralement plus faibles que ceux de l'estimateur SD2.

4 Conclusion

Le présent article décrit les conditions pour que l'estimateur SDR soit équivalent à l'estimateur SD2, et montre de quelle façon ils sont équivalents quand la taille de l'échantillon est plus petite ou plus grande que la matrice de Hadamard choisie. Lorsqu'une matrice de Hadamard $\mathbf{H}_A$ plus petite est utilisée et que les répliques sont tirées uniquement de $\mathbf{H}_A$, l'article montre comment le jeu réduit de répliques produit une approximation raisonnable de l'estimateur SD2. Les exemples empiriques indiquent qu'utiliser un jeu réduit de répliques est raisonnable, puisque la réduction du nombre de répliques n'accroît pas le biais des estimations. En outre, nous voyons que l'utilisation d'un grand nombre de boucles connectées réduit l'effet du carré de la différence entre la première et la dernière unité dans l'échantillon. Puisque le biais et la RMSE de l'estimateur SD1 sont généralement plus grands que ceux de l'estimateur SD2, les estimateurs SDR utilisant un plus grand nombre plutôt qu'un plus petit nombre de boucles connectées donneront des biais et RMSE plus faibles que les estimateurs SDR.

Remerciements

L'auteur remercie David Hornick et Brian Dumbacher de leur révision de la première ébauche du manuscrit, ainsi que les examinateurs et le rédacteur de leurs commentaires qui lui ont permis d'améliorer et de clarifier l'article.

Annexe

Tableau A1
Résultats des simulations de l'estimateur SDR

Population	k_A	AL	Biais relatif prévu selon le nombre de répliques				Erreur quadratique moyenne relative				Ratio de couverture			
			16	32	48	64	16	32	48	64	16	32	48	64
A1	4	1	0,010	0,009	0,009	0,009	0,176	0,091	0,066	0,054	93	94	94	94
		2	0,010	0,010	0,010	0,009	0,176	0,095	0,064	0,048	92	94	94	95
	16	1	0,009	0,008	0,010	0,009	0,141	0,080	0,059	0,048	93	94	94	95
		2	0,009	0,010	0,010	0,009	0,194	0,096	0,065	0,049	92	94	94	95
	64	1 ou 2	0,009	0,009	0,010	0,009	0,194	0,096	0,064	0,049	92	94	94	94
A2	4	1	-0,696	-0,840	-0,888	-0,907	0,485	0,706	0,789	0,823	62	45	38	35
		2	-0,538	-0,768	-0,845	-0,883	0,290	0,590	0,714	0,780	77	54	45	39
	16	1	0,113	-0,270	-0,500	-0,615	0,013	0,073	0,250	0,378	100	97	80	100
		2	1,302	0,152	-0,231	-0,423	1,695	0,023	0,054	0,179	100	100	99	100
	64	1 ou 2	1,302	1,379	1,404	1,417	1,695	1,901	1,972	2,008	100	100	100	100
A3	4	1	0,049	0,031	0,025	0,021	0,195	0,095	0,068	0,054	93	94	94	95
		2	0,070	0,040	0,030	0,025	0,222	0,103	0,067	0,050	93	94	94	95
	16	1	0,155	0,105	0,075	0,060	0,207	0,106	0,070	0,055	95	95	95	95
		2	0,314	0,163	0,112	0,086	0,374	0,144	0,085	0,061	96	95	95	95
	64	1 ou 2	0,314	0,324	0,327	0,327	0,374	0,245	0,199	0,176	96	97	97	97
A4	4	1	0,040	0,023	0,017	0,014	0,192	0,104	0,077	0,063	93	94	94	94
		2	0,060	0,030	0,021	0,017	0,217	0,110	0,075	0,058	93	94	94	95
	16	1	0,144	0,095	0,066	0,052	0,208	0,109	0,077	0,063	95	95	95	95
		2	0,291	0,146	0,098	0,075	0,357	0,144	0,090	0,067	96	95	95	95
	64	1 ou 2	0,291	0,299	0,303	0,305	0,357	0,232	0,191	0,170	96	97	97	97
A5	4	1	0,063	0,063	0,063	0,065	0,192	0,106	0,076	0,063	94	94	95	95
		2	0,068	0,066	0,066	0,065	0,217	0,111	0,075	0,057	93	94	95	95
	16	1	0,063	0,063	0,063	0,065	0,161	0,093	0,068	0,057	94	95	95	95
		2	0,065	0,067	0,066	0,066	0,214	0,111	0,075	0,056	93	94	95	95
	64	1 ou 2	0,065	0,066	0,066	0,065	0,214	0,110	0,074	0,056	93	94	95	95
A6	4	1	0,093	0,092	0,093	0,094	0,211	0,117	0,088	0,072	94	95	95	95
		2	0,092	0,096	0,095	0,094	0,229	0,120	0,086	0,067	94	95	95	95
	16	1	0,099	0,095	0,094	0,094	0,185	0,107	0,080	0,067	94	95	95	95
		2	0,093	0,094	0,094	0,093	0,226	0,117	0,085	0,067	94	95	95	95
	64	1 ou 2	0,093	0,096	0,095	0,095	0,226	0,118	0,084	0,066	94	95	95	95
A7	4	1	0,105	0,069	0,112	0,253	0,219	0,106	0,091	0,143	94	95	95	97
		2	0,004	0,004	0,073	0,310	0,187	0,098	0,079	0,175	92	94	95	97
	16	1	0,177	0,168	0,462	0,847	0,229	0,137	0,351	0,828	95	96	98	99
		2	0,002	0,003	0,027	1,248	0,187	0,097	0,065	1,689	92	94	95	100
	64	1 ou 2	0,002	0,003	0,030	0,115	0,187	0,097	0,065	0,062	92	94	95	96

Tableau A2
Résultats des simulations des méthodes comparatives

Population	Biais relatif prévu selon le nombre de répliques			Erreur quadratique moyenne relative			Ratio de couverture		
	SD1	SD2	EASSR	SD1	SD2	EASSR	SD1	SD2	EASSR
A1	0,009	0,009	-0,001	0,049	0,049	0,032	94	94	97
A2	-0,960	1,417	25,317	0,921	2,008	640,916	23	100	100
A3	0,015	0,327	3,462	0,049	0,176	12,203	94	97	100
A4	0,006	0,305	3,284	0,057	0,170	11,109	94	97	100
A5	0,064	0,065	0,055	0,056	0,056	0,039	95	95	97
A6	0,093	0,095	0,084	0,065	0,066	0,046	95	95	98
A7	0,112	0,115	20,641	0,063	0,062	427,141	96	96	100

Bibliographie

Bellhouse, D.R. (1988). Systematic sampling. Extrait de *Handbook of Statistics*, 6, 125-145.

Fay, R.E., et Train, G.F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics*, American Statistical Association, 154-159.

Hedayat, A., et Wallis, W.D. (1978). Hadamard matrices and their applications. *The Annuals of Statistics*, 6, 1184-1238.

Huang, E.T., et Bell, W.R. (2009). A simulation study of the distribution of Fay's successive difference replication variance estimator. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 5294-5308.

Iachan, R. (1982). Systematic sampling: A critical review. *International Statistical Review*, 50, 293-303.

Madow, W.G., et Madow, L.H. (1944). On the theory of systematic sampling. *Annuals of Mathematical Statistics*, 15, 1-14.

McCarthy, P.J. (1966). Pseudo-replication: Half-samples. *Review of the International Statistical Institute*, 37, 239-264.

Murthy. M.N., et Rao, T.J. (1988). Systematic sampling with illustrative examples. Extrait de *Handbook of Statistics*, 6, 147-185.

Sukasih, A.S., et Jang, D. (2003). Monte Carlo study on the successive difference replication method for non-linear statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3608-3612.

Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 781-790.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag.

Yates, F. (1953). *Sampling Methods for Censuses and Surveys*, 2nd Edition, Hafner Publishing Company, New York, NY.

U.S. Census Bureau (2006). Technical Paper 66, "Design and Methodology: Current Population Survey," Octobre 2006.