Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2005

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. 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 dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Décembre 2005

Publication autorisée par le ministre responsable de Statistique Canada

Mai 2006

Nº 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

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

Vol. 31, N° 2, pp. 153-164
Statistique Canada, N° 12-001 au catalogue

Imputation hot deck pour le modèle de réponse

Wayne A. Fuller et Jae Kwang Kim [1]

Résumé

L'imputation hot deck est une procédure qui consiste à remplacer les réponses manquantes à certaines questions par des valeurs empruntées à d'autres répondants. L'un des modèles sur lesquels elle s'appuie est celui où l'on suppose que les probabilités de réponse sont égales dans les cellules d'imputation. Nous décrivons une version efficace de l'imputation hot deck pour le modèle de réponse dans les cellules et donnons un estimateur de la variance dont le traitement informatique est efficace. Nous détaillons une approximation de la procédure entièrement efficace dans laquelle un petit nombre de valeurs sont imputées pour chaque non-répondant. Nous illustrons les procédures d'estimation de la variance dans une étude de Monte Carlo.

Mots clés : Non-réponse, imputation fractionnaire; probabilité de réponse; estimation de la variance par rééchantillonnage.

1. Introduction

Dans les enquêtes par sondage, l'imputation est utilisée comme méthode de traitement de la non–réponse partielle. Dans le cas de l'imputation hot deck, les valeurs imputées sont des fonctions des répondants compris dans l'échantillon courant. Sande (1983) et Ford (1983) décrivent l'imputation hot deck. Kalton et Kasprzyk (1986), ainsi que Little et Rubin (2002) passent en revue diverses procédures d'imputation.

Dans l'une des versions de l'imputation hot deck, la valeur imputée est celle donnée par un répondant appartenant à la même cellule d'imputation, où les cellules d'imputation forment une subdivision exhaustive et disjointe de la population. Dans le cas de l'imputation hot deck aléatoire, des valeurs provenant de répondants appartenant à la même cellule d'imputation sont attribuées au hasard aux non-répondants. L'enregistrement qui fournit la valeur est appelé le *donneur* et celui dans lequel la valeur manque est appelé le *receveur*.

La variance est généralement plus grande pour l'estimateur imputé que pour l'échantillon complet, parce que la non-réponse réduit la taille de l'échantillon et que l'estimateur imputé peut contenir une composante due à l'imputation aléatoire. Rao et Shao (1992) ont proposé pour l'imputation hot-deck une méthode du jackknife ajusté où les unités de la première phase sont sélectionnées avec remise. Rao et Sitter (1995) discutent de la méthode d'estimation de la variance par le jackknife ajusté pour l'imputation par le ratio. Rao (1996) et Sitter (1997) utilisent la méthode du jackknife ajusté dans le cas de l'imputation par la régression. Shao, Chen et Chen (1998) appliquent la notion de Rao et Shao (1992) à la méthode des répliques répétées équilibrées (BRR). Shao et Steel (1999) proposent une estimation de la variance pour les données d'enquête avec imputation composite, où plus d'une méthode d'imputation est utilisée, et introduisent les fractions d'échantillonnage dans les expressions de la variance. Yung et Rao (2000) appliquent la méthode du jackknife ajusté à des estimateurs imputés construits en utilisant un échantillon stratifié a posteriori. Rubin (1987), ainsi que Rubin et Schenker (1986) proposent des méthodes d'imputation multiples. Tollefson et Fuller (1992), ainsi que Särndal (1992) proposent diverses méthodes d'imputation et les estimateurs correspondants de la variance. Kim et Fuller (2004) étudient l'utilisation de l'imputation fractionnée dans le cas du modèle où les observations dans une cellule d'imputation sont indépendantes et de même loi (iid).

Dans le présent article, nous examinons l'imputation hot deck pour une population subdivisée en cellules d'imputation. À la section 2, nous décrivons le modèle de réponse. À la section 3, nous introduisons l'imputation fractionnée entièrement efficace et présentons une méthode d'estimation de la variance pour l'estimateur par imputation, sous l'hypothèse que la probabilité de non-réponse est constante dans une cellule. À la section 4, nous proposons une modification de la méthode entièrement efficace avec utilisation d'un plus petit nombre de donneurs. À la section 5, nous donnons un exemple en vue d'illustrer la mise en œuvre de la méthode proposée. À la section 6, nous exposons les résultats d'une étude en simulation. Enfin, à la dernière section, nous résumons l'étude.

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011, États-Unis; Jae Kwang Kim, Department of Applied Statistics, Yonsei University, Séoul, 120-749, Corée.

2. Conditions de base

Considérons une population de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \ldots, N\}$. À chaque unité i de la population est associée une variable étudiée y_i et un vecteur $\mathbf{x}_i$ de données auxiliaires. L'ensemble de vecteurs, $(y_i, \mathbf{x}_i),\ i = 1, 2, \ldots, N,$ est noté F.

Soit A les indices des éléments d'un échantillon sélectionné d'après un ensemble de règles probabilistes appelées *mécanisme d'échantillonnage*. Soit θ_N la quantité d'intérêt dans la population et $\hat{\theta}$ un estimateur de θ_N, pour l'échantillon complet, linéaire en y et écrivons que

$$\hat{\theta} = \sum_{i \in A} w_i\, y_i. \tag{1}$$

Si w_i est l'inverse de la probabilité de sélection, alors $\hat{\theta}$ est sans biais pour le total de population.

Soit A_R et A_M les ensembles d'indices pour les répondants et les non-répondants dans l'échantillon, respectivement. Définissons la fonction indicateur de réponse

$$R_i = \begin{cases} 1 & \text{si } i \in A_R \\ 0 & \text{si } i \in A_M \end{cases} \tag{2}$$

et posons que $\mathbf{R} = \{(i, R_i); i \in A\}$. La loi de $\mathbf{R}$ est appelée *mécanisme de réponse*.

Supposons que la population finie U soit constituée de G cellules d'imputation, où l'ensemble d'éléments dans la cellule g est U_g. Soit n_g le nombre d'éléments de l'échantillon compris dans la cellule d'imputation g et soit $r_g, r_g > 0,$ le nombre de répondants dans la cellule d'imputation g. Supposons que nous ayons le modèle de réponse uniforme dans les cellules, où les r_g réponses dans une cellule sont équivalentes à un échantillon de Poisson tiré avec probabilités égales à partir des n_g éléments.

L'imputation fractionnaire est une méthode consistant à utiliser plus d'un donneur par receveur. Kalton et Kish (1984) ont proposé l'imputation fractionnaire comme méthode d'imputation efficace. Elle a été discutée par Fay (1996). Soit d_{ij} le nombre de fois que y_i est utilisé comme donneur pour la valeur manquante y_j et définissons $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$. La loi de $\mathbf{d}$ est appelée *mécanisme d'imputation*. Soit w_{ij}^* le facteur appliqué au poids original de l'élément j quand y_i est utilisé pour cet élément. Pour l'élément $j,\ j \in A_M$,

$$Y_{Ij} = \sum_{i \in A_R} w_{ij}^*\, y_i \tag{3}$$

est la moyenne pondérée des valeurs pour les répondants. Le facteur w_{ij}^* est appelé *fraction d'imputation*, c'est-à-dire la fraction de la réponse manquante y_j que fournit le donneur i. Notons que $w_{ii}^* = 1$ pour $i \in A_R$ et $w_{ij}^* = 0$ pour $i \neq j, i, j \in A_R$. La somme des facteurs d'imputation pour une réponse manquante est contrainte d'être égale à 1,

$$\sum_{i \in A_R} w_{ij}^* = 1, \quad \forall j \in A. \tag{4}$$

Un estimateur ayant les valeurs imputées définies par (3) et un facteur $w_{ij}^* < 1$ est appelé estimateur *par imputation fractionnaire*.

Nous pouvons écrire un estimateur par imputation linéaire en y sous la forme

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} w_j w_{ij}^* \right) y_i \tag{5}$$

$$=: \sum_{i \in A_R} \alpha_i y_i, \tag{6}$$

où la notation A =: B signifie que la définition de B est telle qu'il soit égal à A. La somme des $w_{ij}^*\, w_j$ sur l'ensemble des receveurs pour lesquels i est un donneur (y compris pour lui–même), noté α_i, est le poids total appliqué au donneur i. Si une unité répondante i n'est pas utilisée comme donneur, sauf pour elle–même, alors $\alpha_i = w_i$.

3. Imputation fractionnaire entièrement efficace

Supposons que tous les éléments d'une cellule d'imputation aient la même probabilité de répondre et supposons que les réponses soient indépendantes. Alors, nous pouvons obtenir la loi globale d'un estimateur imputé sous le modèle de réponse en utilisant la structure de probabilité de l'échantillonnage à plusieurs phases, où le modèle de réponse est traité comme étant la deuxième phase du mécanisme d'échantillonnage.

Si les probabilités de réponse dans une cellule sont uniformes, alors un estimateur raisonnable du total est la somme pondérée des estimateurs par le ratio

$$\hat{\theta}_{\mathrm{FE}} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A_R \cap U_g} w_i}. \tag{7}$$

Dans le contexte de l'échantillonnage à deux phases, Kott et Stukel (1997) ont donné à l'estimateur (7) le nom d'estimateur avec facteur d'extension repondéré. L'estimateur (7) est dit entièrement efficace parce qu'il ne contient aucune variabilité due à la sélection aléatoire des donneurs. Si les w_i sont les mêmes pour tous les éléments d'une cellule, le ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i\, y_i \tag{8}$$

est une moyenne simple et, donc, sans biais pour la moyenne de cellule, sachant qu'il existe au moins un répondant dans la cellule. Si les w_i d'une cellule ne sont pas égaux, alors (8) présente un biais de ratio. Il est possible que le nombre d'éléments dans une cellule, n_g, soit positif et

que le nombre de répondants, r_g, soit nul. Quand cela se produit en pratique, les cellules sont regroupées.

Nous pouvons obtenir les propriétés de grand échantillon de l'estimateur pour une série de populations et d'échantillons. Supposons que la population soit composée de G_v cellules disjointes et exhaustives, où v est l'indice de la série. Supposons que la variance d'un estimateur de la moyenne pour l'échantillon complet soit $O(n_v^{-1})$, où n_v est la taille de l'échantillon sélectionné à partir de la v^{e} population. Supposons que les réponses sont indépendantes. Alors, sous des conditions de régularité, nous pouvons nous servir des procédures utilisées par Kim, Navarro et Fuller (2005) dans la preuve de leur théorème 2.1 pour montrer que l'estimateur (7) satisfait

$$\hat{\theta}_{\text{FE}_v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1)e_{iv} + o_p(n_v^{-1/2} N_v), \quad (9)$$

où $e_{iv} = y_{iv} - \overline{Y}_{gv}$, A_{gv} est l'ensemble d'indices d'échantillon dans la g^{e} cellule pour le v^{e} échantillon, $\overline{Y}_{gv}$ est la moyenne de population de la variable y dans la cellule gv de population F_v, π_{gv} est la probabilité qu'un élément dans la cellule gv réponde, et F_v représente la v^{e} population. En outre

$$V(\tilde{\theta}_{\text{FE}v} \mid F_v) = V(\hat{\theta}_v \mid F_v) + E\left\{ \sum_{g_v=1}^{Gv} \pi_{gv}^{-1}(1-\pi_{gv}) \sum_{i \in Agv} w_{iv}^2 e_{iv}^2 \mid F_v \right\}, \quad (10)$$

où

$$\tilde{\theta}_{\text{FE}v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1)e_{iv}.$$

Nous pouvons appliquer l'estimateur (7) en utilisant une imputation fractionnaire dans laquelle chaque unité répondante figurant dans une cellule d'imputation est utilisée comme donneur pour chaque non-répondant compris dans la cellule. Alors, l'estimateur (7) peut s'écrire sous la forme de l'estimateur par imputation fractionnaire

$$\hat{\theta}_{\text{FEFI}} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i, \quad (11)$$

où w_j w_{ij}^* est le poids du donneur i pour le receveur j, w_{ij}^* est la fraction d'imputation du donneur i pour le receveur j définie dans (3), et

$$w_{ij}^* = \begin{cases} \left(\sum_{s \in A_R \cap U_g} w_s\right)^{-1} w_i R_i & \text{si } R_j = 0 \\ 1 & \text{si } R_j = 1 \text{ et } i = j. \end{cases} \quad (12)$$

L'estimateur (11) avec w_{ij}^* donné par (12), qui est algébriquement équivalent à (7), est appelé *estimateur par imputation entièrement efficace* (FEFI pour *fully efficient fractionally imputed*). L'estimateur par imputation fractionnaire a l'avantage de permettre d'estimer directement des fonctions de y, telles que la fraction inférieure à un nombre donné, d'après l'ensemble de données imputées fractionnaires.

Afin d'examiner l'estimation de la variance par rééchantillonnage, posons qu'un estimateur de la variance par rééchantillonnage pour l'échantillon complet est donné par

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k(\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

où $\hat{\theta}^{(k)}$ est la k^{e} estimation de θ_N d'après les observations incluses dans la k^{e} réplique, L est le nombre de répliques, et c_k est un facteur associé à la réplique k déterminé par la méthode de rééchantillonnage. Pour une discussion de la répétition des échantillons d'enquête, voir Krewski et Rao (1981), ainsi que Rao, Wu et Yue (1992). Si l'estimateur original $\hat{\theta}$ est un estimateur linéaire de la forme (1), la k^{e} estimation répétée de $\hat{\theta}$ peut s'écrire

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \quad (14)$$

où $w_i^{(k)}$ est le poids de rééchantillonnage de la i^{e} unité de la k^{e} réplique.

Nous proposons pour l'estimateur $\hat{\theta}_{\text{FEFI}}$ la réplique

$$\hat{\theta}_{\text{FEFI}}^{(k)} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} =: \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i. \quad (15)$$

Si nous utilisons la réplique (15), nous pouvons écrire l'estimateur de la variance par rééchantillonnage sous la forme

$$\hat{V}_{\text{FEFI}} = \sum_{k=1}^{L} c_k(\hat{\theta}_{\text{FEFI}}^{(k)} - \hat{\theta}_{\text{FEFI}})^2. \quad (16)$$

Les répliques données par (15) peuvent être calculées en deux étapes. Premièrement, nous créons la réplique habituelle en définissant les poids $w_i^{(k)}$ pour chaque élément. Deuxièmement, pour un non-répondant, nous utilisons comme fraction d'imputation par rééchantillonnage du donneur i au receveur j

$$w_{ij}^{*(k)} = \frac{w_i^{(k)}}{\sum_{s \in A_R \cap U_g} w_s^{(k)}}.$$

Notons que la somme des poids de rééchantillonnage fractionnaire des enregistrements donneurs pour chaque receveur est égale au poids de rééchantillonnage de chaque unité dans un échantillon complet.

La méthode proposée est étroitement associée à l'estimateur de la variance de Rao et Shao (1992). Voir aussi Yung et Rao (2000). Toutefois, l'utilisation de l'imputation fractionnaire simplifie beaucoup l'estimation de la variance. Dans la création des répliques, seuls les poids appliqués aux valeurs imputées changent. Il n'est pas nécessaire de recalculer les valeurs imputées et, une fois qu'ils sont calculés, les poids des répliques peuvent être utilisés pour n'importe quelle fonction lisse du vecteur y. En outre, les répliques fractionnaires rendent l'estimateur (16) approprié pour un vecteur de variables y.

Nous pouvons utiliser le théorème 3.1 de Kim, Navarro et Fuller (2005) pour montrer que, étant donné une méthode de production de répliques de l'échantillon complet convergente,

$$\hat{V}_{\text{FEFI}} = V(\tilde{\theta}_{\text{FE}v} \mid F_v) - N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) e_{iv}^2 + o_p(n_v^{-1}), \qquad (17)$$

où $\tilde{\theta}_{\text{FE}v}$ est défini dans (10), et où la loi a trait aux mécanismes d'échantillonnage et de réponse.

Si l'on peut ignorer la correction pour population finie, l'estimateur (16) est convergent pour $V\{\hat{\theta}_{\text{FE}}\}$. Si la taille d'échantillon est grande comparativement à N, alors un estimateur de

$$N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv})\, e_{iv}^2$$

devrait être ajouté à (16).

La méthode d'imputation et d'estimation de la variance décrite pour le modèle de réponse produit aussi des estimateurs convergents pour le modèle de moyenne de cellule. Sous ce modèle, les éléments contenus dans une cellule de la population finie sont une réalisation de variables aléatoires indépendantes et de même loi. La méthode d'imputation fondée sur le modèle de réponse n'est pas nécessairement entièrement efficace pour la moyenne de population sous le modèle de moyenne de cellule, mais on peut montrer que l'estimateur de la moyenne et l'estimateur de la variance de la moyenne estimée sont convergents.

4. Approximations de la méthode entièrement efficace

Aux sections précédentes, nous avons construit l'estimateur $\hat{\theta}_{\text{FEFI}}$ de façon à ce que la variance due à l'imputation soit nulle. L'application de la méthode d'imputation fractionnaire, telle qu'elle est décrite en (11), pourrait nécessiter l'utilisation d'un grand nombre de donneurs pour chaque receveur. Par conséquent, nous décrivons une procédure comportant un nombre fixe de donneurs par receveur qui est entièrement efficace pour le total général, mais qui n'est pas forcément entièrement efficace pour les sous-populations. La méthode consiste à affecter des donneurs pour produire une variance faible des valeurs imputées entre receveurs et à modifier la pondération des donneurs pour arriver à l'efficacité complète pour le total.

Supposons que M donneurs soient affectés à chaque receveur. Nous proposons d'affecter les donneurs aux receveurs de façon à approximer la distribution de tous les répondants dans la cellule. L'une des méthodes de sélection possibles consiste à tirer un échantillon stratifié pour chaque receveur. Une autre consiste à recourir à l'échantillonnage systématique avec probabilités proportionnelles aux poids pour sélectionner les donneurs pour chaque receveur. Les fractions initiales w_{ij0}^* sont affectées aux valeurs données. Dans le cas de l'échantillonnage systématique avec poids égaux, la fraction initiale w_{ij0}^* est M^{-1}.

Après avoir affecté les donneurs, nous corrigeons les fractions initiales, w_{ij0}^*, de sorte que la somme des poids donne l'estimateur entièrement efficace de la moyenne de y et que la fonction de distribution cumulative estimée d'après les poids soit une approximation de l'estimateur entièrement efficace de la fonction de distribution cumulative. La modification de la pondération par la régression a été proposée par Fuller (1984, 2003). Chen, Rao et Sitter (2000) discutent d'une méthode d'imputation efficace où l'on modifie les valeurs imputées plutôt que les poids. Soit $\mathbf{z}_{gj} = (z_{gj1}, z_{gj2}, \ldots, z_{gj\alpha})$ un vecteur défini par

$$\begin{aligned} z_{gj1} &= y_j \\ z_{gj2} &= 1 \quad \text{si } y_j \le L_2 \\ &= 0 \quad \text{autrement} \\ &\vdots \\ z_{gj\alpha} &= 1 \quad \text{si } L_{\alpha-1} < y_j \le L_\alpha \\ &= 0 \quad \text{autrement,} \end{aligned}$$

où $L_2, L_3, \ldots, L_\alpha$ divisent la fourchette de valeurs observées de y dans la cellule g en $\alpha - 1$ sections. Le nombre de sections que l'on peut utiliser dépend du nombre et du type d'observations dans la cellule, du nombre de receveurs et du nombre de donneurs par receveur. Si le nombre de donneurs par receveur est grand, il est possible d'ajuster l'ensemble de poids pour chaque receveur de façon à ce que la somme des w_{ij}^* sur i soit égale à l'unité pour chaque j et que la somme des $w_{ij}^* y_i$ sur i soit l'estimateur entièrement efficace pour chaque j. Dans la plupart des cas, les poids sont ajustés de sorte que la somme des w_{ij}^* sur i soit égale à l'unité pour chaque j et que les moyennes de cellule des valeurs imputées soient égales à l'estimateur entièrement efficace.

Soit $\overline{\mathbf{z}}_{\mathrm{FE},g}$ l'estimateur entièrement efficace pour la cellule g. Si nous utilisons des procédures de régression, les w_{ij}^* modifiés pour donner la moyenne de cellule entièrement efficace de $\mathbf{z}$, sont

$$w_{ij}^* = w_{ij0}^* + (\overline{\mathbf{z}}_{\mathrm{FE},g} - \overline{\mathbf{z}}_g^*)\,\mathbf{S}_{\mathbf{zz}g}^{-1} w_{ij0}^* (\overline{\mathbf{z}}_{g[i]j} - \overline{\mathbf{z}}_{g\cdot j})', \quad (18)$$

où

$$\mathbf{S}_{\mathbf{zz}g} = \sum_{j\in A_{Lg}} b_j \sum_{i\in A_{Rg}} w_{ij0}^* (\mathbf{z}_{g[i]j} - \overline{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \overline{\mathbf{z}}_{g\cdot j}) d_{ij},$$

$$\overline{\mathbf{z}}_{g\cdot j} = \sum_{i\in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},$$

$$\overline{\mathbf{z}}_g^* = \sum_{j\in A_{Lg}} b_j \sum_{i\in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},$$

$$b_j = \left(\sum_{s\in A_{Lg}} w_s\right)^{-1} w_j,$$

A_{Lg} est l'ensemble d'indices des receveurs dans la cellule g, $\mathbf{z}_{g[i]j} = \mathbf{z}_{gi}$ est la valeur imputée d'après le donneur i au receveur j, et $\overline{\mathbf{z}}_{g\cdot j}$ est la moyenne pondérée des valeurs imputées pour le receveur j en utilisant le poids initial w_{ij0}^*.

Pour estimer la variance, nous créons des répliques de sorte que les poids appliqués aux donneurs reflètent l'effet de la suppression d'un élément sur l'estimateur entièrement efficace. Nous utilisons les mots « suppression » et « supprimer » pour identifier l'élément choisi pour la modification principale du poids pour l'estimation de la variance par rééchantillonnage.

Soit $w_i^{(k)}$ le poids attribué à l'élément i pour la k^{e} réplique pour l'estimation de la variance de l'estimateur pour l'échantillon complet. Alors, la réplique pour la moyenne entièrement efficace de y pour la cellule g est

$$\overline{\mathbf{z}}_g^{(k)} = \left[\sum_{i\in A_{Rg}} w_i^{(k)}\right]^{-1} \sum_{i\in A_{Rg}} w_i^{(k)} \mathbf{z}_i. \quad (19)$$

Les fractions de rééchantillonnage sont attribuées aux donneurs dans la cellule g de sorte que l'estimation de la moyenne de cellule par rééchantillonnage soit $\overline{\mathbf{z}}_g^{(k)}$. Nous assignons les poids fractionnaires initiaux $w_{ij0}^{*(k)}$, où $w_{ij0}^{*(k)}$ est faible, mais positif, si i est une unité supprimée pour la réplique k. Nous calculons les poids fractionnaires finaux $w_{ij}^{*(k)}$ selon la procédure (18) en remplaçant $\overline{\mathbf{z}}_{\mathrm{FE},g}$ par $\overline{\mathbf{z}}_g^{(k)}$ et w_{ij0}^* par $w_{ij0}^{*(k)}$. La procédure simule l'effet de la suppression d'un seul élément sur l'estimateur entièrement efficace.

5. Un exemple artificiel

Nous présentons ici un exemple fondé sur des données artificielles afin d'illustrer l'application de la méthode proposée. Supposons que nous observions deux variables d'intérêt, x et y, dans un échantillon de taille $n = 10$ obtenu par échantillonnage aléatoire simple. La variable x est une variable nominale comptant trois catégories, disons 1, 2 et 3, et la variable y est une variable continue. Il y a non–réponse partielle pour les deux variables et il existe un ensemble de cellules d'imputation pour chaque variable. Le tableau 5.1 donne les observations sur l'échantillon, où la non–réponse est représentée par M. Nous utilisons un poids unitaire pour simplifier la présentation. Nous divisons par dix pour obtenir les poids pour la moyenne.

Tableau 5.1
Un ensemble de données illustratif

Observation	Poids	Cellule pour x	Cellule pour y	x	y
1	1	1	1	1	7
2	1	1	1	2	M
3	1	1	2	3	M
4	1	1	1	M	14
5	1	1	2	1	3
6	1	2	1	2	15
7	1	2	2	3	8
8	1	2	1	3	9
9	1	2	2	2	2
10	1	2	1	M	M

Comme la variable x est une variable nominale à trois catégories, l'utilisation de trois fractions pour l'imputation fractionnaire donne des estimateurs entièrement efficaces pour la distribution de la variable x. Donc, dans le tableau 5.2, les poids pour les trois valeurs imputées de x pour la quatrième observation sont les fractions pour les trois catégories dans la cellule 1 pour x.

Si l'on utilise un sous–ensemble de donneurs pour chaque receveur, nous suggérons une méthode contrôlée de sélection des donneurs, telle que l'échantillonnage systématique. Dans notre exemple simple, nous pourrions facilement utiliser l'imputation fractionnaire avec les quatre réponses y dans la cellule 1, mais afin d'illustrer l'ajustement par la régression, nous n'en utilisons que trois. Voir le tableau 5.2.

Dans la situation où les réponses à deux questions manquent, plusieurs approches sont possibles, y compris la définition d'un troisième ensemble de cellules d'imputation pour ce genre de cas. Étant donné la petite taille de l'échantillon dans notre illustration, nous imputons sous l'hypothèse que x et y sont indépendantes dans les cellules. Donc, nous imputons quatre valeurs pour l'observation 10. Pour chacune des deux valeurs possibles de x, nous imputons deux valeurs possibles de y. Nous choisissons l'une des paires de valeurs de y imputées de façon qu'elles soient inférieures à la moyenne des réponses et l'autre, de façon à ce qu'elle soit plus grande que la moyenne. Voir les valeurs imputées pour l'observation 10 au tableau 5.2.

Tableau 5.2
Poids fractionnaires pour les moyennes

Observation	Poids	Donneur pour y	Cellule pour x	Cellule pour y	x	y
1	1,0000		1	1	1	7
2	0,2886	1	1	1	2	7
2	0,3960	6	1	1	2	15
2	0,3154	8	1	1	2	9
3	0,3333	5	1	2	3	3
3	0,3333	7	1	2	3	8
3	0,3334	9	1	2	3	2
4	0,5000		1	1	1	14
4	0,2500		1	1	2	14
4	0,2500		1	1	3	14
5	1,0000		1	2	1	3
6	1,0000		2	1	2	15
7	1,0000		2	2	3	8
8	1,0000		2	1	3	9
9	1,0000		2	2	2	2
10	0,2247	8	2	1	2	9
10	0,2753	4	2	1	2	14
10	0,2095	1	2	1	3	7
10	0,2905	6	2	1	3	15

Nous attribuons une fraction initiale égale à un tiers aux trois valeurs imputées pour les observations 3 et 4, et une fraction initiale égale à un quart aux quatre valeurs imputées pour l'observation 10. Puis, nous ajustons les poids fractionnaires en utilisant la méthode de régression de l'équation (18) pour donner la moyenne par imputation fractionnaire entièrement efficace (FEFI) de y comme estimateur, où l'estimateur entièrement efficace de la moyenne de y est

$$\bar{y}_{\mathrm{FE}} = \sum_{g=1}^{2} \frac{n_g}{n} \bar{y}_{Rg} = 8{,}4833.$$

Nous contraignons les poids pour l'observation 10 de sorte que les fractions estimées pour les deux catégories de x soient les fractions de cellule. Alors, comme la moyenne pondérée de la variable nominale est contrôlée pour chaque individu, le vecteur $\mathbf{z}$ contient uniquement la variable y. Le tableau 5.2 donne les poids fractionnaires finaux calculés sous pondération par la régression.

Un analyste peut utiliser l'ensemble de données du tableau 5.2 et tout programme informatique pour échantillon complet pour calculer des estimations des fonctions de y et x, telles que la moyenne de y pour les catégories de x. L'ensemble de données fractionnaires est entièrement efficace pour toute fonction de la variable x et est également entièrement efficace pour la moyenne de la variable y.

Pour l'estimation de la variance par le jackknife, nous répétons le calcul des poids pour chaque réplique. Les estimations répétées des moyennes de cellule de y sont données au tableau 5.3 et les estimations répétées des fractions pour les catégories de x sont données au tableau 5.4. Nous utilisons les valeurs des tableaux 5.3 et 5.4 comme totaux de contrôle $\bar{\mathbf{z}}_{\mathrm{FE},g}$ dans la pondération par la régression. Nous prenons $w_{ij0}^{*(k)} = 3^{-1}$ comme valeur initiale des fractions de rééchantillonnage pour l'observation 2 et $w_{ij0}^{*(k)} = 4^{-1}$ pour l'observation 10.

Le tableau 5.5 contient les poids jackknife pour l'ensemble de données obtenu par imputation fractionnaire du tableau 5.2. Les poids de rééchantillonnage sont utilisés de la même façon que les répliques pour un échantillon complet. Ils conviennent, avec les mises en garde de la section suivante, pour toute statistique pour laquelle le jackknife avec échantillon complet est approprié. Donc, la procédure est particulièrement séduisante pour un ensemble de données d'usage général, car l'analyste ne doit effectuer aucun calcul supplémentaire.

Nous obtenons l'estimateur entièrement efficace de la moyenne de y en considérant que les répondants représentent la deuxième phase d'un échantillon à deux phases. Un estimateur de variance pour échantillon à deux phases peut s'écrire

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2} \frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{\mathrm{FE}})^2 + \sum_{g=1}^{2}\left(\frac{n_g}{n}\right)^2 \frac{1}{r_g} s_{Rg}^2 = 3{,}043,$$

où s_{Rg}^2 est la variance d'échantillon intracellulaire pour la cellule g. Si nous utilisons les poids de rééchantillonnage du tableau 5.5, l'estimation de la variance par rééchantillonnage pour la moyenne de y est

$$\hat{V}_{\mathrm{JK}}(\bar{y}_{\mathrm{FI}}) = \sum_{k=1}^{10} 0{,}9\,(\bar{y}_{\mathrm{FI}}^{(k)} - \bar{y}_{\mathrm{FI}})^2 = 3{,}078.$$

La différence entre l'estimateur de la variance linéarisé et l'estimateur de la variance par le jackknife est

$$\sum_{g=1}^{2}\left(\frac{r_g}{r_g - 1}\frac{n-1}{n} - 1\right) s_{Rg}^2.$$

Donc, l'estimateur de la variance par le jackknife surestime légèrement la variance réelle dans notre exemple.

Tableau 5.3
Répliques jackknife de la moyenne de cellule de la variable y

Cellule	Réplique									
	1	2	3	4	5	6	7	8	9	10
1	12,67	11,25	11,25	10,33	11,25	10,00	11,25	12,00	11,25	11,25
2	4,33	4,33	4,33	4,33	5,00	4,33	2,50	4,33	5,50	4,33

Tableau 5.4
Répliques jackknife de la moyenne de cellule des variables nominales de la variable *x*

Cellule	Niveau de *x*	Réplique									
		1	2	3	4	5	6	7	8	9	10
	1	0,33	0,67	0,67	0,50	0,33	0,50	0,50	0,50	0,50	0,50
1	2	0,33	0,00	0,33	0,25	0,33	0,25	0,25	0,25	0,25	0,25
	3	0,33	0,33	0,00	0,25	0,33	0,25	0,25	0,25	0,25	0,25
	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	2	0,50	0,50	0,50	0,50	0,50	0,33	0,67	0,67	0,33	0,50
	3	0,50	0,50	0,50	0,50	0,50	0,67	0,33	0,33	0,67	0,50

Tableau 5.5
Poids jackknife pour l'imputation fractionnaire

Obs.	Réplique									
	1	2	3	4	5	6	7	8	9	10
1	0	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111
2	0,1664	0	0,3206	0,4205	0,3206	0,4563	0,3206	0,2392	0,3206	0,2724
2	0,6559	0	0,4400	0,3002	0,4400	0,2500	0,4400	0,5540	0,4400	0,5075
2	0,2888	0	0,3505	0,3904	0,3505	0,4048	0,3505	0,3179	0,3505	0,3312
3	0,3706	0,3706	0	0,3706	0,3226	0,3706	0,5018	0,3706	0,2867	0,3706
3	0,3697	0,3697	0	0,3697	0,5018	0,3697	0,0090	0,3697	0,6004	0,3697
3	0,3708	0,3708	0	0,3708	0,2867	0,3708	0,6003	0,3708	0,2240	0,3708
4	0,3703	0,7407	0,7407	0	0,3703	0,5556	0,5556	0,5556	0,5556	0,5556
4	0,3704	0	0,3704	0	0,3704	0,2777	0,2777	0,2777	0,2777	0,2777
4	0,3704	0,3704	0	0	0,3704	0,2778	0,2778	0,2778	0,2778	0,2778
5	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111	1,1111	1,1111	1,1111
6	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111	1,1111	1,1111
7	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1.,111	1,1111	1,1111
8	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111
9	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111
10	0,1624	0,2777	0,2777	0,3061	0,2777	0,2286	0,3474	0,3013	0,1520	0
10	0,3931	0,2778	0,2778	0,2494	0,2778	0,1417	0,3934	0,4395	0,2185	0
10	0,0932	0,2778	0,2778	0,3231	0,2778	0,4400	0,1483	0,0746	0,3171	0
10	0,4623	0,2778	0,2778	0,2324	0,2778	0,3008	0,2220	0,2957	0,4235	0

6. Études par simulation

6.1 Paramètres d'intérêt

Pour étudier les propriétés de la méthode d'imputation, nous avons réalisé une étude de Monte-Carlo. L'échantillon est stratifié, avec deux éléments par strate et deux cellules d'imputation, où les cellules recoupent les strates. La cellule 1 comprend 20 % de la population des strates 1 à 25 et 80 % de la population des strates 26 à 50. La probabilité de réponse est 0,7 pour la cellule 1 et 0,5 pour la cellule 2. Nous examinons deux variables. La variable D est toujours observée et définit une sous-population. La probabilité que $D = 1$ est de 0,25 pour la cellule 1 et de 0,40 pour la cellule 2. La variable y est sujette à la non–réponse avec probabilités de réponse dans les cellules constantes. La variable D est indépendante de y et de la probabilité de réponse. La variable y suit une loi normale, où les paramètres pour une population de 50 strates sont donnés au tableau 5.1. Dans le modèle de génération des données du tableau 6.1, il n'existe aucun effet de strate. Les paramètres d'intérêt sont : $\theta_1 =$ moyenne de $y, \theta_2 =$ moyenne de y pour $D = 1, \theta_3 =$ fraction de Y inférieure à deux, $\theta_4 =$ fraction de Y inférieure à un.

Tableau 6.1
Ensemble de paramètres A

		Cellule 1		Cellule 2	
Strates	Poids de l'élément	Moyenne	Variance	Moyenne	Variance
1 à 25	0,01	0,4	0,36	1,6	0,36
26 à 50	0,01	0,4	0,36	1,6	0,36

6.2 Méthodes d'estimation

Dans la simulation, nous avons utilisé $M = 5$ et $M = 3$ donneurs par receveur. Nous avons sélectionné des échantillons systématiques à titre de donneurs pour chaque receveur. Si le nombre de répondants dans la cellule est inférieur à M, chaque répondant est utilisé comme donneur pour chaque receveur et les w_{ij}^* sont proportionnels au poids w_i original des répondants. Si le nombre de répondants dans la cellule est supérieur à M, nous classons les donneurs par taille et les numérotons de 1 à r_g. Puis, nous plaçons les donneurs dans l'ordre , 3, 5, $\ldots, r_g, r_{g-1}, r_{g-3}, \ldots, 2$ pour les

valeurs impaires de r_g et dans l'ordre $1, 3, 5, \ldots, r_{g-1}, r_g$, $r_{g-2}, \ldots, 2$ pour les valeurs paires de r_g. Ensuite, nous calculons les sommes cumulées des poids et sélectionnons m_g échantillons systématiques de taille M, où $m_g = n_g - r_g$. Les sommes cumulées sont normalisées de sorte que la somme générale soit égale à l'unité, le nombre R_{Ng}, compris entre 0 et $0, 2m_g$, est sélectionné aléatoirement et les m_g échantillons sont les échantillons systématiques de taille M définis par le donneur associé à $R_{Ng} + 0,2(s-1) + (t-1)m_g^{-1}$, $s = 1, 2, 3, 4, 5$ pour les receveurs $t = 1, 2, \ldots, m_g$. Pour chaque donneur, la fraction d'imputation initiale est $w_{ij}^* = M^{-1}$.

Les fractions d'imputation initiale sont modifiées en utilisant la méthode de régression (18). Nous avons ordonné les donneurs dans une cellule du plus petit au plus grand et formé la somme cumulée des poids. Soit

$$S_{g,wt} = \sum_{i=1}^{t} w_{[i]}, i \in A_{Rg}, \qquad (20)$$

où $w_{[i]}, i = 1, 2, \ldots, r_g$ est le poids de $y_{g,(i)}$ et les $y_{g,(1)} \leq \ldots \leq y_{g,(n)}$ sont les valeurs ordonnées de y dans la cellule g. Pour définir les bornes des groupes qu'il convient d'utiliser pour créer des fonctions indicateurs, posons que t_{*s} est le t pour lequel

$$\max \{S_{g,wt} : S_{g,wt} \leq 0,2sS_{gw}\}$$

pour $s = 1, 2, 3, 4$, où S_{gw} est le total des poids pour les donneurs compris dans la cellule g. Définissons

$$\begin{aligned} z_{gi,s+1} &= 1 \quad \text{si } y_i \leq y_{g,(t_{*s})} \text{ et } i \in A_{Rg} \\ &= 0 \quad \text{autrement} \end{aligned} \qquad (21)$$

pour $s = 1, 2, 3, 4$ et soit $\mathbf{z}_{gj} = (y_{gj1}, z_{gj2}, \ldots, z_{gj5})$. L'estimateur par imputation, modifié par la régression, de la moyenne pour chacune des cinq variables comprises dans le vecteur z est l'estimateur entièrement efficace de la moyenne pertinente.

L'estimateur entièrement efficace (FE) avec unité supprimée pour la réplique k de la moyenne de cellule de $\mathbf{z}$ est défini en (19). Le poids fractionnaire initial du donneur k à l'élément j est fixé à $w_{kj0}^{*(k)} = 0,01 w_{kj}^*$. Ce poids initial assure que le poids final soit faible, mais permet l'ajustement par la régression. Les poids finaux $w_{ij0}^{*(k)}$ sont calculés par la procédure de régression (18) en utilisant le poids initial $w_{ij}^{*(k)}$.

6.3 Résultats de l'étude de Monte-Carlo

Les résultats de Monte-Carlo pour les 5 000 échantillons générés par les paramètres du tableau 6.1 sont donnés aux tableaux 6.2 et 6.3. Nous présentons les résultats pour l'échantillon complet, pour l'imputation fractionnaire avec cinq donneurs, pour l'imputation fractionnaire avec trois donneurs et pour l'imputation multiple (MI) en utilisant le bootstrap bayésien approximatif (ABB) de Rubin et Schenker (1986) avec $M = 5$ et avec $M = 3$. Les procédures FI et MI sont toutes deux sans biais pour les quatre paramètres du tableau 6.2. La dernière colonne de ce tableau donne la variance de Monte-Carlo de l'estimateur divisée par la variance de Monte-Carlo de la procédure FI avec $M = 5$, exprimée en pourcentage. La procédure FI est de 5 % à 10 % plus efficace que la procédure MI avec $M = 5$ et de 9 % à 13 % plus efficace que la procédure MI avec $M = 3$.

Sous le modèle, la moyenne des valeurs observées n'est pas le meilleur estimateur de la moyenne de domaine. Dans cet exemple, l'estimateur FI est presque aussi efficace que l'estimateur pour l'échantillon complet. L'effet d'un nombre plus petit d'observations est compensé par l'utilisation d'un meilleur estimateur de la moyenne pour le domaine. Sous le modèle, l'indicateur de domaine est indépendant des valeurs de y, sachant la cellule. Par conséquent, il est efficace d'utiliser toutes les valeurs contenues dans la cellule comme donneurs, plutôt que simplement les répondants dans le domaine.

Les propriétés des estimateurs de la variance sont données au tableau 6.3. La colonne intitulée « moyenne relative » donne la moyenne estimée de Monte-Carlo des variances estimées divisées par la variance estimée de Monte-Carlo, où cette dernière est donnée au tableau 6.2. Les deux méthodes d'estimation de la variance semblent être quasiment sans biais pour la variance de la moyenne. La variance relative de l'estimateur de variance MI pour $M = 5$ est égale à près de deux fois celle de l'estimateur de variance FI pour $M = 5$. Pour $M = 3$, l'estimateur de variance MI vaut plus de trois fois l'estimateur de variance FI. La variance de l'estimateur de variance MI est grande, parce que la variance due aux observations manquantes est estimée avec quatre degrés de liberté pour $M = 5$ et avec deux degrés de liberté pour $M = 3$.

L'estimateur de variance MI de la moyenne de domaine est gravement biaisé. Cette propriété a été reconnue pour la première fois par Fay (1991, 1992) et étudiée par Meng (1994), ainsi que par Wang et Robins (1998). L'estimateur de variance FI pour la moyenne de domaine présente aussi un biais positif, quoique nettement plus faible que celui de MI. Nous pouvons réduire le biais dans l'estimateur de variance FI en augmentant M, mais le biais de MI dépend peu de M.

Tous les estimateurs de variance de $\hat{\theta}_4$ présentent un léger biais négatif. Nous pensons que l'estimateur FI est légèrement biaisé pour $\hat{\theta}_4$ parce que, bien que nous utilisions le vecteur z, les poids sont légèrement lissés par la procédure de régression. Il est connu que l'imputation multiple (MI) donne lieu à un biais de petit échantillon. Voir Kim (2002).

Tableau 6.2
Moyenne et variance des estimateurs ponctuels sous les conditions A (5 000 échantillons de taille 100)

Paramètre	Scénario d'imputation	Moyenne	Variance	Variance relative à FI (%)
Moyenne	Échantillon complet	1,00	0,00570	67
(θ_1)	FI(3)	1,00	0,00849	100
	ABB(3)	1,00	0,00926	109
	FI(5)	1,00	0,00849	100
	ABB(5)	1,00	0,00903	106
Moyenne de domaine	Échantillon complet	1,14	0,02020	99
(θ_2)	FI(3)	1,14	0,02050	100
	ABB(3)	1,14	0,02230	109
	FI(5)	1,14	0,02040	100
	ABB(5)	1,14	0,02170	106
$\Pr(Y < 2)$	Échantillon complet	0,87	0,00104	51
(θ_3)	FI(3)	0,87	0,00202	100
	ABB(3)	0,87	0,00228	113
	FI(5)	0,87	0,00202	100
	ABB(5)	0,87	0,00223	110
$\Pr(Y < 1)$	Échantillon complet	0,50	0,00208	66
(θ_4)	FI(3)	0,50	0,00313	100
	ABB(3)	0,50	0,00342	109
	FI(5)	0,50	0,00313	100
	ABB(5)	0,50	0,00329	105

Tableau 6.3
Moyenne relative, statistique *t* et variance relative pour les estimateurs de variance sous les conditions A (5 000 échantillons de taille 100)

Paramètre	Méthode	Moyenne relative (%)**	Statistique *t**	Variance relative (%)
Moyenne	FI(3)	100,1	0,05	5,66
(θ_1)	ABB(3)	99,6	–0,19	19,25
	FI(5)	100,1	0,03	5,65
	ABB(5)	98,2	–0,89	9,95
Moyenne de domaine	FI(3)	115,9	7,54	13,88
(θ_2)	ABB(3)	127,9	12,72	28,88
	FI(5)	106,6	3,14	11,62
	ABB(5)	128,4	13,43	20,03
$\Pr(Y < 2)$	FI(3)	103,9	1,86	13,90
(θ_3)	ABB(3)	100,8	0,36	48,42
	FI(5)	101,7	0,82	12,07
	ABB(5)	98,5	–0,67	25,10
$\Pr(Y < 1)$	FI(3)	98,5	–0,75	4,67
(θ_4)	ABB(3)	96,3	–1,80	18,51
	FI(5)	97,6	–1,20	4,45
	ABB(5)	96,7	–1,65	10,17

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.
** Moyenne de Monte-Carlo des estimations de variance divisée par la variance de Monte-Carlo des estimations, en pourcentage.

Dans un deuxième ensemble de paramètres, noté *C*, les moyennes étaient les suivantes :

Cellule 1 des strates 1 à 25; $\mu = 0{,}4$

Cellule 1 des strates 26 à 50; $\mu = 3{,}0$

Cellule 2 des strates 1 à 25; $\mu = 1{,}6$

Cellule 2 des strates 26 à 50; $\mu = 2{,}2$.

Tous les autres paramètres sont les mêmes que dans l'ensemble de paramètres A. Les propriétés des estimateurs sont données au tableau 6.4. L'imputation fractionnaire (FI) et l'imputation multiple (MI) produisent toutes deux des estimations sans biais des moyennes et de la moyenne de domaine. Comme pour l'ensemble de paramètres A, la procédure FI est de 8 % à 12 % plus efficace que la procédure MI pour $M = 5$ et de 14 % à 16 % plus efficace pour $M = 3$.

Les hypothèses requises pour l'estimation de variance MI ne sont pas satisfaites pour l'ensemble de paramètres C. Par conséquent, la variance MI estimée est fortement biaisée pour tous les paramètres. Voir le tableau 6.5. Pour $M = 5$, le biais dans la variance MI estimée est d'environ 17 % pour la variance de la moyenne globale et de près de 50 % pour la moyenne de domaine. Le biais de la variance MI de la moyenne est plus faible pour une variable binomiale que pour une variable continue, parce que l'effet de stratification est plus faible dans le premier cas.

Les propriétés des variances estimées pour la procédure FI sont semblables à celles obtenues pour l'ensemble de paramètres A. La variance de la moyenne de domaine présente un biais positif d'environ 23 % pour $M = 3$ et d'environ 6 % pour $M = 5$.

La variance de l'estimation de la variance MI est de 2,4 à 3,5 fois plus élevée que la variance de l'estimation de la variance FI pour $M = 5$ et de 3 à 7 fois plus élevée pour $M = 3$, ce qui démontre la supériorité nette de l'estimateur de variance FI pour cette configuration.

Tableau 6.4
Moyenne et variance des estimateurs ponctuels sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Scénario d'imputation	Moyenne	Variance	Variance relative à FI (%)
Moyenne	Échantillon complet	2,10	0,00500	48
(θ_1)	FI(3)	2,10	0,01050	100
	ABB(3)	2,10	0,01220	116
	FI(5)	2,10	0,01050	100
	ABB(5)	2,10	0,01150	110
Moyenne de domaine	Échantillon complet		0,02530	102
(θ_2)	FI(3)	2,01	0,02510	101
	ABB(3)	2,01	0,02850	115
	FI(5)	2,01	0,02480	100
	ABB(5)	2,01	0,02710	109
$\Pr(Y < 2)$	Échantillon complet		0,00127	45
(θ_3)	FI(3)	0,45	0,00281	100
	ABB(3)	0,45	0,00322	115
	FI(5)	0,45	0,00280	100
	ABB(5)	0,45	0,00314	112
$\Pr(Y < 1)$	Échantillon complet		0,00107	54
(θ_4)	FI(3)	0,15	0,00199	100
	ABB(3)	0,15	0,00226	114
	FI(5)	0,15	0,00199	100
	ABB(5)	0,15	0,00214	108

Tableau 6.5
Moyenne relative, statistique *t* et variance relative pour les estimateurs de variance sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Méthode	Moyenne relative (%)	Statistique *t**	Variance relative (%)
Moyenne	FI(3)	100,9	0,41	6,42
(θ_1)	ABB(3)	116,7	7,31	40,14
	FI(5)	100,8	0,39	6,42
	ABB(5)	117,1	7,99	22,29
Moyenne de domaine	FI(3)	122,7	10,78	16,23
(θ_2)	ABB(3)	144,4	19,79	46,05
	FI(5)	106,1	2,95	11,95
	ABB(5)	148,7	22,51	32,49
$\Pr(Y < 2)$	FI(3)	104,4	2,18	6,63
(θ_3)	ABB(3)	114,7	6,54	42,32
	FI(5)	101,8	0,89	6,42
	ABB(5)	112,1	5,74	20,67
$\Pr(Y < 1)$	FI(3)	102,3	1,13	11,08
(θ_4)	ABB(3)	101,3	0,58	39,14
	FI(5)	99,9	–0,04	10,05
	ABB(5)	102,2	1,04	23,60

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.

7. Résumé

Dans l'imputation fractionnaire, plusieurs donneurs sont utilisés pour chaque valeur manquante et une fraction du poids du non-répondant est attribuée à chaque donneur. Si l'on utilise tous les donneurs, la procédure est entièrement efficace, sous le modèle, pour toutes les fonctions d'un vecteur **y**. Nous montrons que l'utilisation de l'imputation fractionnaire avec un petit nombre d'imputations par non-répondant peut donner un estimateur entièrement efficace de la moyenne. Les estimations d'autres paramètres, comme les estimations de la distribution cumulative sont presque entièrement efficaces.

L'imputation fractionnaire permet de construire des répliques d'usage général pour l'estimation de la variance. Il est possible d'utiliser un seul ensemble de répliques pour estimer la variance dans le cas de variables imputées, de variables observées sur l'ensemble des répondants et, sous les hypothèses du modèle, pour des fonctions de deux types de variables. Les répliques donnent des estimations des variances des moyennes de domaine dont le biais est nettement plus faible que celui des estimations par imputation multiple. Le biais tend vers zéro quand la valeur de M augmente et, dans la simulation, est modéré pour $M = 5$. L'estimateur de la variance par rééchantillonnage est facile à appliquer au moyen d'un logiciel de rééchantillonnage, tel que Wesvar.

L'imputation fractionnaire avec un nombre fixe de donneurs par receveur est un peu plus efficace pour la moyenne que l'imputation multiple avec le même nombre de donneurs. L'imputation fractionnaire donne des estimations de variance dont le biais est plus faible et dont la variance est nettement plus faible que les estimateurs par imputation multiple avec le même nombre d'imputations.

8. Remerciements

La présente étude a été financée partiellement aux termes d'un sous-contrat entre Westat et la Iowa State University en vertu du contrat n° ED-99-CO-0109 établi entre Westat et le Department of Education, ainsi que du contrat de coopération 13-3AEU-0-80064 conclu entre la Iowa State University, le U.S. National Agricultural Statistics Service et le U.S. Bureau of the Census. Nous remercions Jean Opsomer et Damiao Da Silva de leurs commentaires constructifs.

Bibliographie

Chen, J., Rao, J.N.K. et Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Ford, B.M. (1983). An overview of hot-deck procedures. Dans *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

Fuller, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-130.

Fuller, W.A. (1996). *Introduction to Statistical Time Series*. 2ième édition. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2003). Estimation for multiple phase samples. Dans *Analysis of Survey Data*, (Éds. R.L. Chambers et C.J. Shinner). Wiley, Chichister, England, 307-322.

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

Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics Part A – Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, 89, 470-477.

Kim, J.K., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Navarro, A. et Fuller, W.A. (2005). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, à paraître.

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

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2ième édition. New York: John Wiley & Sons, Inc.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-573.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

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

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with applications to imputation for missing data *Biometrika*, 82, 453-460.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Rubin, D.B., et Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.

Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., Chen, Y. et Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

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

Tollefson, M., et Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.

Wang, N., et Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.