N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

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

2003

Statistique Canada Statistics Canada

Canada

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

INFÉRENCE POUR DES TOTAUX DANS LE CAS D'ÉCHANTILLONNAGE PAR GRAPPES EN PRÉSENCE D'IMPUTATION PAR LA MOYENNE POUR DONNÉES D'ENQUÊTE MANQUANTES

D. Haziza and J.N.K. Rao[1]

RÉSUMÉ

Deux cadres de travail distincts sont utilisés pour l'estimation de la variance en présence de données d'enquête manquantes: (i) basé sur le plan de sondage; (ii) basé sur un modèle. Dans le cas de (i), un mécanisme de réponse uniforme est supposé à l'intérieur des classes d'imputation alors que (ii) utilise l'hypothèse plus faible de réponse ignorable mais suppose un modèle de population (ou d'imputation). Dans cet article, nous montrons comment adapter ces cadres de travail dans le cas de l'échantillonnage à deux degrés. L'estimation de la variance est effectuée en utilisant deux méthodes dans le cas du cadre de travail basé sur un modèle: (i) La méthode de Särndal (1992) et (ii) la méthode de Shao-Steel basée sur l'approche renversée de Fay (1991). Une étude de simulation est effectuée afin d'évaluer la robustesse des deux méthodes en terme de biais relatif des estimateurs de variance lorsque le modèle est mal spécifié.

MOTS CLÉS: Cadre de travail basé sur le plan de sondage, cadre de travail basé sur un modèle, corrélation intra-grape, estimation de la variance.

1. INTRODUCTION

L'échantillonnage à degrés multiples est fréquemment utilise dans les enquêtes, particulièrement lorsque l'échantillonnage direct des éléments n'est pas praticable (ou impossible). Dans cet article, nous étudions le cas de l'échantillonnage par grappes à deux degrés. La notation suivante sera utilisée: Soit une population comprenant N grappes disjointes, U_i de taille $M_i, i=1,...,N$. Soit $K=\sum_{i=1}^{N} M_i$ le nombre total d'unités ultimes (éléments) dans la population. De plus, soit y_{ij} la valeur d'une variable d'intérêt y pour le j^e élément dans la i^e grappe, $i=1,...,N;\ j=1,...,M_i;$ et soit Y_i le total pour la i^e grappe. L'objectif est d'estimer le total dans la population $Y=\sum_{i=1}^{N} Y_i=\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$, en tirant un échantillon selon un plan de sondage à deux degrés: au premier degré, un échantillon aléatoire de grappes, s, de taille n, est tiré selon un plan de sondage $p(s)$ à partir de la population de grappes. Au deuxième degré, un échantillon aléatoire d'éléments, s_i, de taille $m_i (i=1,...,n)$ est tiré selon un plan de sondage $p_i(s_i)$ si la i^e grappe est tirée. En l'absence de non-réponse, un estimateur de Y, denoté $\hat{Y}$, est défini selon

$$\hat{Y}=\sum_{i\in s}\sum_{j\in s_i} w_{ij} y_{ij}, \qquad (1)$$

où $w_{ij}=(\pi_i \pi_{j|i})^{-1}, \pi_i=P(i\in s)$ est la probabilité d'inclusion de la grappe i dans s, et $\pi_{j|i}$ est la probabilité d'inclusion conditionnelle de l'élément j appartenant à la grappe i dans s_i, $i=1,...,N;\ j=1,...,M_i$. Il est bien

[1] D. Haziza, Household Survey Methods Division, Statistics Canada, Ottawa, ON, Canada, K1T 0T6, David.Haziza@statcan.ca, J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada, K1S 5B6

connu que l'estimateur $\hat{Y}$ en (1) est sans biais pour *Y* par rapport au plan de sondage; c'est-à-dire, $E_p\left(\hat{Y}\right)=Y$, où $E_p(.)$ dénote l'espérance par rapport au plan de sondage. Dans le cas de l'échantillonnage à deux degrés, notons que $E_p(.)=E_1E_2\left(.|s\right)$, où $E_1(.)$ et $E_2(.)$ dénotent respectivement les espérances par rapport au premier et au deuxième degrés.

En présence de non-réponse à la variable *y*, il n'est pas possible de calculer l'estimateur en (1) puisque certaines valeurs de *y* sont manquantes. Nous imputons pour les valeurs manquantes à la variable *y* et définissons un estimateur imputé de *Y*, dénoté $\hat{Y}_I$, selon

$$\hat{Y}_I=\sum_{i\in s}\left[\sum_{j\in s_i}w_{ij}a_{ij}y_{ij}+\sum_{j\in s_i}w_{ij}\left(1-a_{ij}\right)y_{ij}^*\right], \tag{2}$$

où y_{ij}^* dénote la valeur imputée pour la valeur manquante y_{ij} et a_{ij} est un indicateur de réponse tel que $a_{ij}=1$ si l'élément j dans la grappe *i* répond à la variable *y* et $a_{ij}=0$ sinon. Par souci de simplicité, nous considérons une classe d'imputation unique, mais les résultats peuvent être généralisés au cas de classes multiples. Nous considérons l'imputation par la moyenne qui utilise les valeurs imputées

$$y_{ij}^*=\bar{y}_r=\frac{\sum_{i\in s}\sum_{j\in s_i}w_{ij}a_{ij}y_{ij}}{\sum_{i\in s}\sum_{j\in s_i}w_{ij}a_{ij}}. \tag{3}$$

Utilisant les valeurs imputées (3) dans (2), l'estimateur imputé $\hat{Y}_I$ devient

$$\hat{Y}_I=\frac{\sum_{i\in s}\sum_{j\in s_i}w_{ij}}{\sum_{i\in s}\sum_{j\in s_i}w_{ij}a_{ij}}\sum_{i\in s}\sum_{j\in s_i}w_{ij}a_{ij}y_{ij}. \tag{4}$$

Afin d'étudier les propriétés (par exemple, biais et variance) de l'estimateur imputé $\hat{Y}_I$, deux cadres de travail distincts ont été utilisés dans la littérature: (i) Cadre de travail basé sur le plan de sondage; voir Rao (1990), Rao et Sitter (1995) et Shao et Steel (1999); (ii) Cadre de travail basé sur un modèle; voir Särndal (1992), Deville et Särndal (1994) et Shao et Steel (1999). Le cadre de travail habituel basé sur le plan de sondage suppose :

Hypothèse BP: $E_r\left(a_{ij}\right)=p_1$ (réponse uniforme) et $E_r\left(a_{ij}a_{i'j'}\right)=E_r\left(a_{ij}\right)E_r\left(a_{i'j'}\right)=p_1^2$ sauf pour $i=i'$ et $j=j'$ (indépendance des statuts de réponse), où $E_r(.)$ dénote l'espérance par rapport au mécanisme de réponse.

Le cadre de travail habituel basé sur un modèle suppose :

Hypothèse BM: Le mécanisme de réponse est ignorable ou non-confondu; c'est-à-dire que le fait qu'une unité réponde ou non ne dépend pas de la variable à imputer mais peut dépendre de variables indépendantes du modèle d'imputation. Dans le cas de l'imputation par la moyenne, le modèle est donné par

$$E_m\left(y_{ij}\right)=\mu, V_m\left(y_{ij}\right)=\sigma^2, Cov_m\left(y_{ij},y_{i'j'}\right)=0 \text{ si } (ij)\neq(i'j'), \tag{5}$$

où $E_m(.), V_m(.)$ et $Cov_m(.)$ dénotent respectivement l'espérance, la variance et la covariance par rapport au modèle d'imputation (5).

Les classes d'imputation sont habituellement définies de manière à ce que les hypothèses BP ou BM soient approximativement valides. Le mécanisme de réponse dans le cas de l'hypothèse BM est beaucoup plus faible que le mécanisme de réponse uniforme dans le cas de l'hypothèse BP, mais dans ce cas les inférences dépendent du modèle d'imputation. Notons que l'estimateur imputé (4) est robuste dans la mesure où il est approximativement sans biais sous l'hypothèse BP ainsi que sous l'hypothèse BM.

Les hypothèses BP et BM pourraient ne pas être valides dans le cas de l'échantillonnage à deux degrés puisque les corrélations intra-grappe ne sont pas prises en compte. C'est pourquoi, dans la section 2, nous proposons des hypothèses plus réalistes puisqu'elles reflètent la dépendance à l'intérieur des grappes, et montrons que l'estimateur imputé (4) reste valide. La section 3 compare les variances conditionnelles, étant donné l'échantillon, afin d'étudier

l'effet des corrélations intra-grappe. Dans la section 4, nous développons des estimateurs de variance dans le cas du cadre de travail basé sur un modèle, à l'aide de la méthode de Särndal et de la méthode de Shao-Steel basée sur l'approche renversée de Fay (1991). Dans la section 5, une étude de simulation est effectuée afin de comparer les deux méthodes d'estimation de la variance lorsque le modèle d'imputation est mal spécifié. Finalement, dans la section 6, nous discutons brièvement de certaines généralisations possibles.

2. CADRES DE TRAVAIL POUR L'ÉCHANTILLONNAGE À DEUX DEGRÉS

Dans cette section, nous proposons des hypothèses plus faibles correspondant au deux cadres de travail et qui reflètent les corrélations intra-grappe. Le cadre de travail basé sur le plan de sondage suppose maintenant :

Hypothèse BPG: $E_r\left(a_{ij}\right)=p_1$ et $E_r\left(a_{ij}a_{ij'}\right)=p_2$, pour $j\neq j'$. Notons qu'en général, $E_r\left(a_{ij}a_{ij'}\right)\neq E_r\left(a_{ij}\right)E_r\left(a_{ij'}\right)=p_1^2$; c'est-à-dire que les statuts de réponses entre deux unités dans la même grappe ne sont pas indépendants.

Le cadre de travail basé sur un modèle suppose maintenant :

Hypothèse BMG: Le mécanisme de réponse est ignorable ou non-confondu; c'est-à-dire que le fait qu'une unité réponde ou non ne dépend pas de la variable à imputer mais peut dépendre de variables indépendantes du modèle d'imputation. Dans le cas de l'imputation par la moyenne, le modèle est celui du modèle ANOVA à un facteur aléatoire donné par

$$m: y_{ij}=\mu+\alpha_i+\varepsilon_{ij}, \tag{6}$$

où μ est la moyenne globale, α_i est l'effet aléatoire pour la i^e grappe et ε_{ij} représente l'erreur résiduelle. Nous supposons que

(i) $E_m\left(\alpha_i\right)=E_m\left(\varepsilon_{ij}\right)=0$,

(ii) $Cov_m\left(\varepsilon_{ij},\varepsilon_{i'j'}\right)=0$ sauf pour $i=i'$ et $j=j'$, $Cov_m\left(\alpha_i,\alpha_{i'}\right)=0 \quad \forall i\neq i', Cov_m\left(\alpha_i,\varepsilon_{i'j'}\right)=0 \ \forall i,i'$ et j',

(iii) $V_m\left(\alpha_i\right)=\sigma_\alpha^2\forall i,\ V_m\left(\varepsilon_{ij}\right)=\sigma_\varepsilon^2\forall i,j.$

À partir de (i)-(iii), nous obtenons

$$Cov_m\left(y_{ij},y_{i'j'}\right)=\begin{cases}\sigma_\alpha^2 & \text{si } i=i' \text{ et } j\neq j'\\ \sigma_\alpha^2+\sigma_\varepsilon^2 & \text{si } i=i' \text{ et } j=j'\\ 0 & \text{si } i\neq i'\end{cases}$$

Encore une fois, l'estimateur imputé (4) est robuste dans la mesure où il est approximativement sans biais sous l'hypothèse BPG ainsi que sous l'hypothèse BMG.

3. COMPARAISONS DES VARIANCES CONDITIONNELLES

L'erreur totale, $\hat{Y}_I-Y$, peut être décomposée comme suit:

$$\hat{Y}_I-Y=\left(\hat{Y}-Y\right)+\left(\hat{Y}_I-\hat{Y}\right). \tag{7}$$

Le terme $\hat{Y}-Y$ en (7) est appelé l'erreur due à l'échantillonnage alors que le terme $\hat{Y}_I-\hat{Y}$ est appelé l'erreur due à la non-réponse et à l'imputation. Puisque l'erreur due à l'échantillonnage ne dépend pas de la non-réponse et de méthode d'imputation, nous mettons l'emphase sur la composante de non-réponse $\hat{Y}_I-\hat{Y}$ afin d'étudier la variance conditionnelle de $\hat{Y}_I-\hat{Y}$, dans le cas du cadre de travail basé sur le plan de sondage ainsi que celui basé sur un modèle, étant donné l'échantillon total d'éléments $\tilde{s}$.

3.1 Cadre de travail basé sur le plan de sondage

Dans cette section, nous étudions la variance conditionnelle $V_r\left(\hat{Y}_I - \hat{Y}|\tilde{s}\right)$, sous l'hypothèse BP et l'hypothèse BPG. Premièrement, après linéarisation en série de Taylor, on peut facilement montrer que, sous l'hypothèse BP,

$$V_r^{DB}\left(\hat{Y}_I - \hat{Y}|\tilde{s}\right) \approx \frac{p_1(1-p_1)}{p_1^2}\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2\left(y_{ij}-\bar{y}\right)^2, \tag{8}$$

où $\bar{y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij} \Big/ \sum_{i\in s}\sum_{j\in s_i} w_{ij}$. De manière similaire, sous l'hypothèse BPG,

$$V_r^{DBC}\left(\hat{Y}_I - \hat{Y}|\tilde{s}\right) \approx \frac{p_1(1-p_1)}{p_1^2}\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2\left(y_{ij}-\bar{y}\right)^2 + \frac{p_2-p_1^2}{p_1^2}\sum_{i\in s}\sum_{j\in s_i}\sum_{\substack{j'\in s_i \\ j\neq j'}} w_{ij}\left(y_{ij}-\bar{y}\right)w_{ij'}\left(y_{ij'}-\bar{y}\right). \tag{9}$$

Soit $\tilde{R}_{DB} = V_r^{DBC}\left(\hat{Y}_I - Y|\tilde{s}\right)\Big/V_r^{DB}\left(\hat{Y}_I - Y|\tilde{s}\right)$. On a donc

$$\tilde{R}_{DB} = 1+\rho_p\left(\tilde{c}_{DB}-1\right), \tag{10}$$

où

$$\rho_p = \frac{Cov_r\left(a_{ij},a_{ij'}\right)}{\sqrt{V_r\left(a_{ij}\right)V_r\left(a_{ij'}\right)}} = \frac{p_2-p_1^2}{p_1(1-p_1)}$$

est la corrélation intra-grappe pour les indicateurs de réponse sous l'hypothèse BPG et

$$\tilde{c}_{DB} = \frac{\sum_{i\in s}\left(\sum_{j\in s_i} w_{ij}\left(y_{ij}-\bar{y}\right)\right)^2}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2\left(y_{ij}-\bar{y}\right)^2}.$$

Plusieurs points peuvent être soulevés en rapport avec (10). Premièrement, si $\rho_p = 0$ alors $\tilde{R}_{DB} = 1$, comme prévu. Si ρ_p est grand alors $\tilde{R}_{DB}$ pourrait être substantiel. Donc, dans ce cas, utiliser l'hypothèse BP au lieu de l'hypothèse BPG pourrait mener à une sous-estimation importante de la variance de l'estimateur imputé $\hat{Y}_I$. Par l'inégalité de Cauchy-Schwarz, $\left(\sum_{j\in s_i} b_{ij}c_{ij}\right)^2 \le \sum_{j\in s_i} b_{ij}^2\sum_{j\in s_i} c_{ij}^2$, avec $b_{ij}=1$ et $c_{ij} = w_{ij}\left(y_{ij}-\bar{y}\right)$, on a

$$\left(\sum_{j\in s_i} w_{ij}\left(y_{ij}-\bar{y}\right)\right)^2 \le m_i\sum_{j\in s_i} w_{ij}^2\left(y_{ij}-\bar{y}\right)^2,$$

ce qui implique que

$$\tilde{c}_{DB} \le \frac{\sum_{i\in s}\sum_{j\in s_i} m_i w_{ij}^2\left(y_{ij}-\bar{y}\right)^2}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2\left(y_{ij}-\bar{y}\right)^2} \equiv \tilde{d}_{DB}.$$

Une borne supérieure pour $\tilde{R}_{DB}$ est alors donnée par

$$\tilde{R}_{DB} \le 1+\rho_p\left(\tilde{d}_{DB}-1\right) \tag{11}$$

si $\rho_p \ge 0$. Dans le cas particulier de tailles égales, $m_i = m$, (11) devient

$$\tilde{R}_{DB} \le 1+\rho_p(m-1). \tag{12}$$

L'expression (12) suggère que le ratio $\tilde{R}_{DB}$ augmente à mesure que le nombre d'éléments tirés dans chaque grappe, *m*, augmente étant donné ρ_p, ou, à mesure que ρ_p augmente étant donné *m*.

3.2 Cadre de travail basé sur un modèle

Dans cette section, nous étudions la variance conditionnelle $V_m\left(\hat{Y}_I - \hat{Y}|\tilde{s}\right)$, sous l'hypothèse BM et l'hypothèse BMG. Premièrement, en notant que

$$\hat{Y}_I - \hat{Y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}d_{ij}y_{ij}$$

avec

$$d_{ij} = \frac{\sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}} a_{ij} - 1,$$

on peut facilement montrer que sous l'hypothèse BM

$$V_m^{MB}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right) = \sigma^2 \sum_{i\in s}\sum_{j\in s_i} w_{ij}^2 d_{ij}^2. \tag{13}$$

De manière similaire, on peut facilement montrer que sous l'hypothèse BMG

$$V_m^{MBC}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right) = \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right)\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2 d_{ij}^2 + \sigma_\alpha^2 \sum_{i\in s}\sum_{j\in s_i}\sum_{\substack{j'\in s_i \\ j\neq j'}} w_{ij}w_{ij'}d_{ij}d_{ij'}. \tag{14}$$

En notant que $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$, le ratio $\tilde{R}_{MB} = V_m^{MBC}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right) \Big/ V_m^{MB}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right)$ est égal à

$$\tilde{R}_{MB} = 1 + \rho_m\left(\tilde{c}_{MB} - 1\right), \tag{15}$$

où

$$\rho_m = \frac{Cov_m\left(y_{ij}, y_{ij'}\right)}{\sqrt{V_m\left(y_{ij}\right)V_m\left(y_{ij'}\right)}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

représente la corrélation intra-grappe pour les valeurs de *y* sous l'hypothèse BMG et

$$\tilde{c}_{MB} = \frac{\sum_{i\in s}\left(\sum_{j\in s_i} w_{ij}d_{ij}\right)^2}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2 d_{ij}^2}.$$

Encore une fois, l'inégalité de Cauchy-Schwarz donne

$$\left(\sum_{j\in s_i} w_{ij}d_{ij}\right)^2 \leq m_i \sum_{j\in s_i} w_{ij}^2 d_{ij}^2,$$

ce qui implique que

$$\tilde{c}_{MB} \leq \frac{\sum_{i\in s}\sum_{j\in s_i} m_i w_{ij}^2 d_{ij}^2}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}^2 d_{ij}^2} \equiv \tilde{d}_{MB}.$$

Une borne supérieure pour $\tilde{R}_{MB}$ est alors donnée par

$$\tilde{R}_{MB} \leq 1 + \rho_m\left(\tilde{d}_{MB} - 1\right) \tag{16}$$

si $\rho_m \geq 0$. Dans le cas particulier $m_i = m$, l'expression (16) devient

$$\tilde{R}_{MB} \leq 1 + \rho_m\left(m - 1\right). \tag{17}$$

L'expression (17) implique que le ratio $\tilde{R}_{MB}$ augmente à mesure que *m* augmente étant donné ρ_m ou à mesure que ρ_m augmente étant donné *m*.

4. ESTIMATION DE LA VARIANCE DANS LE CADRE DE TRAVAIL BASÉ SUR UN MODÈLE

Traditionnellement, les chercheurs ont utilisé la configuration échantillonnage-réponse suivante (approche deux-phases) pour l'estimation de la variance :

Population $\rightarrow$ échantillon complet $\rightarrow$ échantillon avec non-répondants.

Dans le cas de l'hypothèse BM, Särndal (1992) a utilisé la décomposition suivante de la variance totale :

$$V\left(\hat{Y} - Y\right) \equiv V_{tot} = E_m E_p E_r\left(\hat{Y}_I - Y\right)^2 = V_{sam} + V_{imp} + 2V_{mix}, \tag{18}$$

où $V_{sam} = E_m V_p\left(\hat{Y} - Y\middle|\tilde{s}\right), V_{imp} = E_p E_r V_m\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right)$ et $V_{mix} = E_m E_p\left[\left(\hat{Y} - Y\right)\left(E_r\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right)\right)\right]$.

Dans le cas de la méthode de Särndal, un estimateur de la variance totale $V\left(\hat{Y}_I - Y\right)$ est donné par $v_t = v_{sam} + v_{imp} + 2v_{mix}$, où v_{sam} est un estimateur de V_{sam}, v_{imp} est un estimateur de V_{imp} et v_{mix} est un estimateur de V_{mix}.

Fay (1991) a utilisé une configuration différente en renversant l'ordre de l'échantillonnage et de réponse (nous l'appellerons l'approche renversée) qui peut être décrite comme suit:

Population $\rightarrow$ recensement avec non-répondants $\rightarrow$ échantillon avec non-répondants.

Dans ce cas (Shao and Steel, 1999),

$$V\left(\hat{Y}_I - Y\right) = E_r E_m V_p\left(\hat{Y}_I - Y|\mathbf{a}\right) + E_r V_m E_p\left(\hat{Y}_I - Y|\mathbf{a}\right), \tag{19}$$

où **a** est le vecteur des indicateurs de réponse a_{ij} en notant que $V_r E_m E_p\left(\hat{Y}_I - Y|\mathbf{a}\right) = 0$ puisque l'estimateur imputé $\hat{Y}_I$ est sans biais sous l'hypothèse BM ou l'hypothèse BMG. Sous l'approche renversée, un estimateur de la variance totale $V\left(\hat{Y}_I - Y\right)$ est donné par $v_t = v_1 + v_2$, où v_1 est un estimateur de $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ étant donné le vecteur des indicateurs de réponse **a**, et v_2 est un estimateur de $E_r V_m E_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$. L'estimateur v_1 ne dépend pas du mécanisme de réponse et/ou du modèle d'imputation. C'est pourquoi, v_1 reste valide indépendamment des hypothèses établies à propos du mécanisme de réponse et/ou du modèle d'imputation. Puisque la composante $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ représente simplement une variance due à l'échantillonnage, son estimation peut être facilement implantée à l'aide de n'importe quelle méthode standard d'estimation de la variance : linéarisation en série de Taylor, Jackknife et Bootstrap. Dans le contexte de l'échantillonnage à deux degrés, la deuxième composante v_2 de v_t est typiquement négligeable par rapport à v_1 si la fraction de sondage globale est négligeable.

Dans cet article, nous mettons l'emphase sur l'échantillonnage aléatoire simple sans remise au premier et au deuxième degrés, ce qui implique que $w_{ij} = \frac{N}{n}\frac{M_i}{m_i}$. En l'absence de non-réponse, un estimateur de la variance de $\hat{Y}$ est donné par

$$v\left(\hat{Y}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s_Y^2 + \frac{N}{n}\sum_{i\in s} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{iy}^2, \tag{20}$$

où $s_Y^2 = \frac{1}{n-1}\sum_{i\in s}\left(\hat{Y}_i - \frac{\hat{Y}}{N}\right)^2$ avec $\hat{Y}_i = \frac{M_i}{m_i}\sum_{j\in s_i} y_{ij}$ et $s_{iy}^2 = \frac{1}{m_i - 1}\sum_{j\in s_i}\left(y_{ij} - \bar{y}_i\right)^2$ avec $\bar{y}_i = \hat{Y}_i / M_i$. Nous utiliserons (20) afin d'estimer $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ dans la section 4.2.

4.1 La méthode de Särndal

Dans cette section, nous développons un estimateur de la variance sous l'hypothèse BMG et échantillonnage aléatoire simple sans remise au premier et au deuxième degrés. L'estimation de V_{sam}, V_{imp} and V_{mix} est effectuée comme suit:

(i) Soit v_p l'estimateur de variance obtenue en l'absence de non-réponse et soit v_I l'estimateur de variance naïf de $\hat{Y}_I$ obtenu à partir de (20) en traitant les valeurs imputées comme si elles avaient été observées. Il est bien connu que pour plusieurs méthodes d'imputation (en particulier pour les méthodes déterministes), v_I sous-estime V_{sam}. Afin de compenser pour cette sous-estimation, nous évaluons l'espérance suivante :

$$E_m\left(v_p - v_I|\tilde{s}\right) \equiv V_{dif}.$$

Ensuite, nous déterminons un estimateur sans biais sous le modèle de V_{dif}, dénoté v_{dif}. Cela requière habituellement l'estimation de certains paramètres du modèle *m*. Alors

$$v_{sam} = v_I + v_{dif}$$

est sans biais sous le modèle pour V_{sam}.

(ii) Ensuite, nous déterminons un estimateur sans biais sous le modèle de V_{imp}, dénoté v_{imp}, c'est-à-dire, $E_m\left(v_{imp}|\tilde{s}\right) = V_{imp}$. Encore une fois, cela pourrait requérir l'estimation de certains paramètres du modèle *m*.

(iii) Finalement, nous déterminons un estimateur sans biais sous le modèle de V_{mix}, dénoté v_{mix}, c'est-à-dire, $E_m\left(v_{mix}|\tilde{s}\right) = V_{mix}$. Encore une fois, cela pourrait requérir l'estimation de certains paramètres du modèle *m*.

Finalement, un estimateur de V_{tot}, dénoté v_{tot}, est donné par

$$v_{tot} = v_I + v_{dif} + v_{imp} + 2v_{mix}. \tag{21}$$

Nous donnons maintenant les expressions pour les différentes composantes dans (21). Premièrement, notons que la composante v_I est obtenue en remplaçant s_Y^2 et s_{iy}^2 par s_{IY}^2 et s_{Iiy}^2 dans (20), où s_{IY}^2 et s_{Iiy}^2 sont calculés de la même manière que s_Y^2 et s_{iy}^2 en traitant les valeurs imputées comme si elles avaient été observées. Après de longs mais élémentaires calculs algébriques, nous obtenons :

$$v_{dif} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n}\sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2\left\{\left[m_i\left(\hat{\sigma}_\alpha^2+\hat{\sigma}_\varepsilon^2\right)-r_i^2A_i\right]-s_M^2\left[\frac{\sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2 r_i^2h_i}{\left(\sum_{i\in s}\frac{M_i}{m_i}r_i\right)^2}\right]\right.$$

$$\left.-\frac{2}{n-1}\left[\frac{1}{\sum_{i\in s}\frac{M_i}{m_i}r_i}\left[\sum_{i\in s}\left(M_i-\hat{\bar{M}}\right)\left(\frac{M_i}{m_i}\right)^2 r_i^2h_i\right]-\frac{\sum_{i\in s}\left(M_i-\hat{\bar{M}}\right)\left(\frac{M_i}{m_i}\right)^2 r_i}{\sum_{i\in s}\frac{M_i}{m_i}r_i}\sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2 r_i^2h_i\right]\right\}$$

$$+\frac{N}{n}\sum_{i\in s}M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\left\{\left[1-\frac{r_i-1}{m_i-1}\right]\hat{\sigma}_\varepsilon^2-\left(1-\frac{r_i}{m_i}\right)\left(\frac{r_i-1}{m_i-1}\right)\left(\frac{r_i}{r_i-1}\right)A_i\right\}, \tag{22}$$

où r_i est le nombre de répondants à la variable *y* dans s_i,

$$A_i = h_i + \frac{\sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2 r_i^2h_i}{\left(\sum_{i\in s}\frac{M_i}{m_i}r_i\right)^2} - 2\frac{\frac{M_i}{m_i}r_i}{\sum_{i\in s}\frac{M_i}{m_i}r_i}h_i,$$

$h_i = \hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2/r_i$ et $s_M^2 = \frac{1}{n-1}\sum_{i\in s}\left(M_i-\hat{\bar{M}}\right)^2$ avec $\hat{\bar{M}} = \sum_{i\in s}M_i/n$. De plus,

$$v_{imp} = \left(\frac{N}{n}\right)^2\sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2\left\{\left[\left(\hat{p}^{-1}-1\right)^2 r_i + o_i\right]\hat{\sigma}_\varepsilon^2 + B_i^2\hat{\sigma}_\alpha^2\right\} \tag{23}$$

où o_i est le nombre de non-répondants à la variable *y* dans s_i, $B_i = \hat{p}^{-1}r_i - m_i$ avec $\hat{p} = \sum_{i\in s}\sum_{j\in s_i}w_{ij}a_{ij}\Big/\sum_{i\in s}\sum_{j\in s_i}w_{ij}$. Finalement,

$$v_{mix} = \left(\frac{N}{n}\right)^2 \sum_{i\in s}\left(\frac{M_i}{m_i}\right)^2 \left\{B_i\left(m_i\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2\right)\right\} - \frac{N}{n}\sum_{i\in s}\frac{M_i}{m_i}\left\{B_i\left(M_i\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2\right)\right\} \tag{24}$$

Notons que $\hat{\sigma}_\alpha^2$ et $\hat{\sigma}_\varepsilon^2$ sont des estimateurs de σ_α^2 et σ_ε^2 respectivement qui peuvent être obtenus à l'aide de méthodes disponibles telles que le méthodes ML, REML ou MINQUE. De plus, notons que les estimateurs de variance sous l'hypothèse BM sont obtenus en posant $\hat{\sigma}_\alpha^2 = 0$ dans (22)-(24).

4.2 La méthode de Shao-Steel

Afin d'utiliser la méthode de Shao-Steel, nous exprimons d'emblée $\hat{Y}_I$ comme $\hat{Y}_I = \hat{K}\hat{R}_a$, où $\hat{K} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}$ et $\hat{R}_a = \hat{Y}_a / \hat{K}_a$, avec $\hat{Y}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}y_{ij}$ et $\hat{K}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}$. Dénotons maintenant l'estimateur de la variance de $\hat{Y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij}$ basé sur l'échantillon total par $v(y)$. Alors, on peut montrer en utilisant la linéarisation en série de Taylor que v_1 est égal à

$$v_1 = v\left(\hat{\xi}\right), \tag{25}$$

où

$$\hat{\xi}_{ij} = a_{ij}y_{ij} + \left(1 - a_{ij}\right)\hat{R}_a + \frac{\left(\hat{K} - \hat{K}_a\right)}{\hat{K}_a}a_{ij}\left(y_{ij} - \hat{R}_a\right). \tag{26}$$

Dans le cas de l'échantillonnage aléatoire simple sans remise au premier et deuxième de degrés, la composante v_1 est obtenue à partir de (20) en remplaçant y_{ij} par $\hat{\xi}_{ij}$, où $\hat{\xi}_{ij}$ est donné par (26). Afin d'obtenir v_2, notons que $E_p\left(\hat{Y}_I - Y\middle|\mathbf{a}\right) = \sum_{i\in U}\sum_{j\in U_i} b_{ij}y_{ij}$, où $b_{ij} = \left(K/K_a\right)a_{ij} - 1$ avec $K_a = \sum_{i\in U} K_{ai}$ et $K_{ai} = \sum_{j\in U_i} a_{ij}$. De plus, on peut facilement montrer que

$$E_r V_m E_p\left(\hat{Y}_I - Y\right) \approx \sigma_\varepsilon^2 K\left(\frac{K}{E_r\left(K_a\right)} - 1\right) + \sigma_\alpha^2\left[\frac{K^2}{E_r\left(K_a^2\right)}\sum_{i\in U}E_r\left(K_{ai}^2\right) - 2\frac{K}{E_r\left(K_a\right)}\sum_{i\in U}M_i E_r\left(K_{ai}\right) + \sum_{i\in U}M_i^2\right]. \tag{27}$$

La composante v_2 est alors obtenue en substituant des estimateurs pour remplacer les quantités inconnues dans (27), ce qui mène à

$$v_2 \approx \hat{\sigma}_\varepsilon^2\hat{K}\left(\frac{\hat{K}}{\hat{K}_a} - 1\right) + \hat{\sigma}_\alpha^2\left\{\frac{\hat{K}^2}{\hat{K}_a^2}\sum_{i\in s} w_i\left[\hat{K}_{ai}^2 - \hat{V}_2\left(\hat{K}_{ai}\right)\right] - 2\frac{\hat{K}}{\hat{K}_a}\sum_{i\in s} w_i M_i\hat{K}_{ai} + \sum_{i\in s} w_i M_i^2\right\}, \tag{28}$$

où $w_i = \frac{1}{\pi_i}$, $\hat{K}_{ai} = \sum_{j\in s_i} w_{j|i}a_{ij}$ et $w_{j|i}$ est le poids de sondage de l'élément j au deuxième degré étant donné que la grappe i a été tirée au premier degré. Notant que $V_2\left(\hat{K}_{ai}\right) = E_2\left(\hat{K}_{ai}^2\right) - K_{ai}^2$, une estimation de K_{ai}^2 est donnée par $\hat{K}_{ai}^2 - \hat{V}_2\left(\hat{K}_{ai}\right)$, où $\hat{V}_2\left(\hat{K}_{ai}\right)$ est une estimation de $V_2\left(\hat{K}_{ai}\right)$. La somme de (25) et (28) donne v_t. Notons que les estimateurs de variance dans le cas de l'hypothèse BM sont obtenus en posant $\hat{\sigma}_\alpha^2 = 0$ dans (28). De plus, notons que le ratio v_1 / v_2 est $O(n/K)$, où K est le nombre total d'éléments. Donc, la composante v_2 pourrait être négligeable par rapport à v_1 même lorsque le taux de sondage du premier degré, n/N, est grand. Il s'ensuit que lorsque n/K est négligeable (comme c'est souvent le cas dans des plans à degrés multiples), le calcul de v_2 peut être omis.

5. ÉTUDE DE SIMULATION

Nous avons effectué une petite étude de simulation afin d'étudier la performance relative des estimateurs de variance obtenus par la méthode de Särndal et celle de Shao-Steel. Nous avons généré plusieurs populations comprenant $N = 120$ grappes avec $M_i = M = 5, 20$ éléments dans la i^e grappe et de telle manière que la corrélation intra-grappe $\rho_m = 0.05, 0.1, 0.2$. Pour cela, nous avons d'abord généré des variables aléatoires i.i.d.

$$y_i \sim N\left(\mu, \sigma_\alpha^2\right), i = 1,..., N,$$

avec $\mu = 200$ et $\sigma_\alpha^2 = 100.$ Ensuite, nous avons généré des variables aléatoires i.i.d.

$$\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right), j = 1,..., M_i,$$

avec $\sigma_\varepsilon^2 = \dfrac{1-\rho_m}{\rho_m}\sigma_\alpha^2$ pour $\rho_m \neq 0$. Les valeurs de *y* sont alors générées en posant

$$y_{ij} = y_i + \varepsilon_{ij}.$$

Dans chaque population, nous avons tiré $R = 5000$ échantillons de grappes, de taille *n*, selon un plan d'échantillonnage aléatoire simple sans remise, avec $n/N = 0.1, 0.5.$ Afin de simplifier la discussion, nous considérons l'échantillonnage par grappe à un seul degré, c'est-à-dire, $m_i = M_i$. Dans chaque échantillon de grappe, nous avons assigné une probabilité p_i à la grappe $i, i = 1,..., n,$ où p_i a été générée d'une loi beta. Ensuite, les indicateurs de réponse a_{ij} ont été générés d'une loi de Bernoulli avec paramètre p_i . Le taux global de réponse a été fixé à 0.65.

Dans chaque échantillon, l'estimateur imputé, $\hat{Y}_I,$ donné en (2), a été calculé et sa variance a été estimée par les méthodes de Särndal et de Shao-Steel. Pour chaque méthode, les estimateurs de variance ont été calculés dans le cas des hypothèses BMG et BM. L'estimateur imputé $\hat{Y}_I$ était approximativement sans biais dans tous les scénarios avec un biais relatif Monte Carlo plus petit que 0.3%.

Comme mesure du biais d'un estimateur de variance, $v\left(\hat{Y}_I\right),$ nous avons utilisé le biais relatif

$$RB\left(v\left(\hat{Y}_I\right)\right) = \frac{E\left[v\left(\hat{Y}_I\right)\right] - MSE\left(\hat{Y}_I\right)}{MSE\left(\hat{Y}_I\right)} \times 100,$$

où $MSE\left(\hat{Y}_I\right)$ dénote l'erreur quadratique moyenne de $\hat{Y}_I$. Les tableaux 1 et 2 exhibent le biais relatif pour *n*/*N* = 0.1 et 0.5, respectivement.

Les tableaux 1 et 2 indiquent que la méthode de Särndal et la méthode de Shao-Steel performent bien lorsque la corrélation intra-grappe est prise en compte (c'est-à-dire, lorsque les estimateurs de variance sont développés sous l'hypothèse BMG) puisque la valeur absolue du biais relatif des estimateurs de variance est plus petite que 7% dans tous les cas. Notons la composante v_{mix} dans le cas de la méthode de Särndal est égal à 0 pour ce plan de sondage. Lorsque la corrélation intra-grappe n'est pas prise en compte, (c'est-à-dire, lorsque les estimateurs de variance sont développés sous l'hypothèse BM), la méthode de Särndal mène à une sous estimation de la vraie variance. La sous-estimation augmente à mesure que le coefficient de corrélation intra-grappe ρ_M et/ou la taille de la grappe M_i augmentent. Par exemple, lorsque *n*/*N* = 0.1 et $\rho_m = 0.1,$ le biais relatif est égal à -9.2% pour $M_i = 5$ et -30.1% pour $M_i = 20.$ De plus, lorsque *n*/*N* = 0.1 et $M_i = 20,$ le biais relatif est égal à -17.6% si $\rho_m = 0.05$ et -39.8% si $\rho_m = 0.2.$ Notons que le taux de sondage *n*/*N* ne semble pas avoir un impact significatif sur le biais relatif des estimateurs de variance. De plus, il est intéressant de noter que la méthode de Särndal ne performe pas bien lorsque le modèle d'imputation est mal spécifié en grande partie à cause de la composante v_{dif} qui tend à être sévèrement biaisée dans ces cas. La méthode de Shao-Steel, quant à elle, performe relativement bien lorsque le modèle d'imputation est mal spécifié puisque la valeur absolue du biais relatif des estimateurs de variance est plus petite que 5% dans tous les cas. Ce résultat n'est pas surprenant puisque la première composante, $v_1,$ de la variance totale

est robuste dans la mesure où elle ne dépend pas des hypothèses faites à propos du modèle d'imputation et puisque la deuxième composante, v_2, est négligeable par rapport à v_1 lorsque n/K est négligeable. Notons que, même lorsque $n/N = 0.5$ et $M_i = 5$, on a $n/K = 0.1$, ce qui peut être considéré comme petit. En résumé, la méthode Shao-Steel semble plus robuste que la méthode de Särndal lorsque les hypothèses du modèle sont erronées.

Tableau 1
Biais relatif (%) des estimateurs de variance avec $n/N = 0.1$

	$\rho_m = 0.05$		$\rho_m = 0.1$		$\rho_m = 0.2$	
	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$
Méthode de Särndal (sous MBC)	6.5	6.7	6.3	3.3	1.6	2.3
Méthode de Shao-Steel (sous MBC)	-4.0	0.2	-1.2	-1.6	-3.2	-1.6
Méthode de Särndal (sous MB)	-6.2	-17.6	-9.2	-30.1	-19.6	-39.8
Méthode de Shao-Steel (sous MB)	-4.7	-0.2	-1.5	-1.8	-3.6	-2.1

Tableau 2
Biais relatif (%) des estimateurs de variance avec $n/N = 0.5$

	$\rho_m = 0.05$		$\rho_m = 0.1$		$\rho_m = 0.2$	
	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$
Méthode de Särndal (sous MBC)	1.3	1.4	2.6	3.3	2.4	3.2
Méthode de Shao-Steel (sous MBC)	0.2	0.3	1.2	0.2	0.6	1.5
Méthode de Särndal (sous MB)	-4.9	-18.0	-8.8	-27.1	-15.7	-38.1
Méthode de Shao-Steel (sous MB)	-0.4	-1.6	-1.5	-2.1	-2.2	-2.6

6. GÉNÉRALISATIONS

Dans cet article, nous avons étudié le cas de l'imputation par la moyenne dans le contexte de l'échantillonnage à deux degrés. Les résultats peuvent être généralisés au cas de l'échantillonnage à deux degrés et l'imputation par régression. Par exemple, supposons qu'un vecteur de q variables auxiliaires $\mathbf{z} = (z_1,...,z_q)$ est disponible pour tous les éléments dans l'échantillon (répondants et non-répondants). Dans le cas de l'hypothèse BMG, le modèle d'imputation est donné

$$y_{ij} = \mathbf{z}'_{ij}\boldsymbol{\beta} + \alpha_i + \varepsilon_{ij},$$

où $\boldsymbol{\beta}$ est un q-vecteur d'effets fixes et α_i et ε_{ij} sont définis comme précédemment. L'imputation par régression utilise $y^*_{ij} = \mathbf{z}'_{ij}\hat{\boldsymbol{\beta}}$ ou $y^*_{ij} = \mathbf{z}'_{ij}\hat{\boldsymbol{\beta}} + \hat{\alpha}_i$, où $\hat{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés ordinaire ou pondéré de $\boldsymbol{\beta}$ et $\hat{\alpha}_i$ est le prédicteur pour les effets aléatoires.

RÉFÉRENCES

Deville, J. C. et Särndal, C. E. (1994), "Variance estimation for the regression imputed Horvitz-Thompson estimator", *Journal of Official Statistics*, 10, 381-394.

Fay, R. E. (1991), "A design-based perspective on missing data variance", *Proceedings of the 1991 Annual Research Conference, US Bureau of the census*, 429-440.

Rao, J. N. K. (1990) "Variance estimation under imputation for missing data", Technical report, Ottawa, Canada: Statistics Canada.

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

Särndal, C. E. (1992), “Methods for estimating the precision of survey estimates when imputation has been used”, *Survey Methodology*, 18, 241-252.

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.