Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

INFERENCE FOR TOTALS IN CLUSTER SAMPLING UNDER MEAN IMPUTATION FOR MISSING DATA

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

ABSTRACT

Two different frameworks are used for variance estimation under imputation for missing survey data: (i) design-based; (ii) model-based. Under (i), uniform response is assumed within imputation classes, while (ii) uses the weaker assumption of ignorable response but assumes a population (or imputation) model. In this paper, we show how to adapt these frameworks to the case of two-stage sampling. Variance estimation is performed using two methods using the model-based framework: (i) the method of Särndal (1992) and (ii) the method of Shao-Steel based on the reverse approach of Fay (1991). A simulation study is conducted to evaluate the robustness of the two methods in terms of relative bias of the variance estimators when the imputation model is misspecified.

KEYWORDS: Design-Based Framework; Intracluster Correlation; Model-Based Framework; Variance Estimation.

1. INTRODUCTION

Multi-stage sampling is often used in surveys, especially when direct element sampling is impractical (or impossible). In this paper, we confine to the case of two-stage cluster sampling. We adopt the following notation: Let a finite population consisting of *N* nonoverlapping clusters, U_i of size $M_i, i=1,...,N$. Let $K=\sum_{i=1}^{N} M_i$ denote the total number of ultimate units (elements) in the population. Further, let y_{ij} be the value of a variable of interest *y* for the j^{th} element in the i^{th} cluster, $i=1,...,N;\ j=1,...,M_i$ and Y_i be the *i*-th cluster total. The objective is to estimate the population total $Y=\sum_{i=1}^{N} Y_i=\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$, by selecting a sample according to a two-stage design: at the first stage, a random sample of clusters, *s*, of size *n*, is selected according to a given design $p(s)$ from the population of clusters. At the second stage, a random sample of elements, s_i, of size $m_i\,(i=1,...,n)$ is selected according to a given design $p_i(s_i)$ if the *i*-th cluster is sampled. Under complete response to item *y*, an estimator of *Y*, denoted $\hat{Y}$, is given by

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

where $w_{ij}=\left(\pi_i\pi_{j|i}\right)^{-1}, \pi_i=P(i\in s)$ is the inclusion probability of cluster *i* in *s*, and $\pi_{j|i}$ is the conditional probability of inclusion of element *j* belonging to cluster *i* in s_i, $i=1,...,N;\ j=1,...,M_i$. It is well-known that the estimator $\hat{Y}$ given in (1) is design-unbiased for *Y*; that is, $E_p(\hat{Y})=Y$, where $E_p(.)$ denotes the expectation with respect to the sampling design. In the case of two-stage sampling, note that $E_p(.)=E_1E_2(.|s)$, where $E_1(.)$ and $E_2(.)$ denote respectively the expectation with respect to the first and second stages.

[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

In the case of nonresponse to item *y*, it is not possible to compute the estimator given in (1) since some *y*-values are missing. We impute for the missing *y*-values and define an imputed estimator of *Y*, denoted $\hat{Y}_I$, as

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

where y_{ij}^* denotes the imputed value for missing y_{ij} and a_{ij} is a response indicator such that $a_{ij} = 1$ if element j in cluster i responds to item y and $a_{ij} = 0$ otherwise. For simplicity, we assume a single imputation class, but the results extend to multiple classes. We consider mean imputation that uses the imputed values

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

Using the imputed values (3) in (2), the imputed estimator $\hat{Y}_I$ reduces to

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

To study the properties (e.g., bias and variance) of the imputed estimator $\hat{Y}_I$, two distinct frameworks have been used in the literature: (i) Design-based framework; see Rao (1990), Rao and Sitter (1995) and Shao and Steel (1999); (ii) Model-based framework ; see Särndal (1992), Deville and Särndal (1994) and Shao and Steel (1999). The customary design-based is based on:

Assumption DB: $E_r(a_{ij}) = p_1$ (uniform response) and $E_r(a_{ij} a_{i'j'}) = E_r(a_{ij}) E_r(a_{i'j'}) = p_1^2$ except for $i = i'$ and $j = j'$ (independence of the response statuses), where $E_r(.)$ denotes the expectation with respect to the response mechanism.

The customary model-based framework is based on:

Assumption MB: The response mechanism is ignorable or unconfounded in the sense that whether or not a unit responds does not depend on the variable being imputed but may depend on the covariates in the assumed imputation model. For mean imputation, the model is given by

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

where $E_m(.), V_m(.)$ and $Cov_m(.)$ denote respectively the expectation, variance and covariance with respect to the imputation model (5).

Imputation classes are usually chosen to make the assumption DB or MB approximately valid. The response mechanism in assumption MB is much weaker than the uniform response in assumption DB, but inferences depend on the assumed imputation model. Note that the imputed estimator (4) is robust in the sense that it is approximately unbiased under assumption DB as well as assumption MB.

Assumptions DB and MB may not be tenable in the case of two-stage sampling because the within cluster correlations are not taken into account. Therefore, in section 2, we propose more realistic assumptions that reflect dependence within a cluster and show that the imputed estimator (4) remains valid. Section 3 compares the conditional variances, given the sample, to study the effect of within-cluster correlations. In section 4, we derive variance estimators under the model-based framework, using Särndal's method and the Shao-Steel's method based on the reverse approach of Fay (1991). In section 5, a simulation study is conducted to compare the two variance estimation methods when the imputation model is misspecified. Finally, in section 6, we briefly discuss some possible extensions.

2. FRAMEWORKS FOR TWO-STAGE SAMPLING

In this section we propose weaker assumptions, corresponding to the two frameworks, that reflect within-cluster correlations. The design-based framework is now based on

Assumption DBC: $E_r\left(a_{ij}\right)=p_1$ and $E_r\left(a_{ij}a_{ij'}\right)=p_2$, for $j \neq j'$. Note that, in general, $E_r\left(a_{ij}a_{ij'}\right) \neq E_r\left(a_{ij}\right)E_r\left(a_{ij'}\right)=p_1^2$; that is, the response statuses between two units in the same cluster are not independent.

The model-based framework is now based on

Assumption MBC: The response mechanism is ignorable or unconfounded in the sense that whether or not a unit responds does not depend on the variable being imputed but may depend on the covariates in the assumed imputation model. For mean imputation, the model is the well known one-way ANOVA model with random effects given by

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

where μ is the general mean, α_i is i-th cluster random effect and ε_{ij} is the residual error. We assume that

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

(ii) $Cov_m\left(\varepsilon_{ij},\varepsilon_{i'j'}\right)=0$ except for $i=i'$ and $j=j'$, $Cov_m\left(\alpha_i,\alpha_{i'}\right)=0 \quad \forall i\neq i', Cov_m\left(\alpha_i,\varepsilon_{i'j'}\right)=0 \;\forall i,i'$ and j',

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

From (i)-(iii), we get

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

Once again, the imputed estimator (4) is robust in the sense that it is approximately unbiased under assumption DBC as well as assumption MBC.

3. COMPARISON OF CONDITIONAL VARIANCES

The total error, $\hat{Y}_I-Y$, may be decomposed as

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

The term $\hat{Y}-Y$ in (7) is called the sampling error, whereas the term $\hat{Y}_I-\hat{Y}$ is called the error due nonresponse and imputation. Since the sampling error does not depend on nonresponse and on the imputation method, we focus on the nonresponse component $\hat{Y}_I-\hat{Y}$ to study the conditional variance of $\hat{Y}_I-\hat{Y}$, under the design-based and the model-based frameworks, given the full sample of elements $\tilde{s}$.

3.1 Design-based framework

In this section, we study the conditional variance $V_r\left(\hat{Y}_I-\hat{Y}\middle|\tilde{s}\right)$, under assumption DB and assumption DBC. First, using Taylor linearization, it is easily seen that under assumption DB,

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

where $\bar{y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij} \Big/ \sum_{i\in s}\sum_{j\in s_i} w_{ij}$. Similarly, under assumption DBC,

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

Let $\tilde{R}_{DB} = V_r^{DBC}\left(\hat{Y}_I - Y\middle|\tilde{s}\right)\Big/V_r^{DB}\left(\hat{Y}_I - Y\middle|\tilde{s}\right)$. Then, we have

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

where

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

is the intracluster correlation of the response indicators under assumption DBC and

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

Several points may be noted from (10). First, if $\rho_p = 0$ then $\tilde{R}_{DB} = 1$, as expected. If ρ_p is large then $\tilde{R}_{DB}$ may be substantial. Hence, in this case, using assumption DB instead of assumption DBC may lead to severe underestimation of the variance of the imputed estimator $\hat{Y}_I$. Using the Cauchy-Schwarz inequality, $\left(\sum_{j\in s_i} b_{ij}c_{ij}\right)^2 \le \sum_{j\in s_i} b_{ij}^2 \sum_{j\in s_i} c_{ij}^2$, with $b_{ij} = 1$ and $c_{ij} = w_{ij}\left(y_{ij}-\bar{y}\right)$, we have

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

which implies that

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

It follows that an upper bound for $\tilde{R}_{DB}$ is given by

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

if $\rho_p \ge 0$. In the particular case of equal subsample sizes, $m_i = m$, (11) becomes

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

Expression (12) suggests that the ratio $\tilde{R}_{DB}$ increases, for fixed ρ_p as the number of elements selected in each cluster, *m*, increases or as ρ_p increases for fixed *m*.

3.2 Model-based framework

In this section, we study the conditional variance $V_m\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right)$, under assumption MB and assumption MBC. First, noting that

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

with

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

it is easily seen that under assumption MB

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

Similarly, it is easily seen that under assumption MBC

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

Noting that $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$, the ratio $\tilde{R}_{MB} = V_m^{MBC}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right) \big/ V_m^{MB}\left(\hat{Y}_I - \hat{Y}\middle|\tilde{s}\right)$ is equal to

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

where

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

is the intracluster correlation of the *y*-values under assumption MBC and

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

Once again, using Cauchy-Schwarz inequality, we have

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

which implies that

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

It follows that an upper bound for $\tilde{R}_{MB}$ is given by

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

if $\rho_m \geq 0.$ In the particular case $m_i = m$, the expression (16) becomes

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

It follows from (17) that the ratio $\tilde{R}_{MB}$ increases as *m* increases for fixed ρ_m or as ρ_m increases for fixed *m*.

4. VARIANCE ESTIMATION UNDER THE MODEL-BASED FRAMEWORK

Traditionally, researchers have used the following sample-response path (two-phase approach) for variance estimation:

Population → complete sample → sample with nonrespondents.

Under assumption MB, Särndal (1992) used the following decompostion of the total variance :

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

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

Under Särndal's method, an estimator of the total variance $V\left(\hat{Y}_I - Y\right)$ is given by $v_t = v_{sam} + v_{imp} + 2v_{mix}$, where v_{sam} is an estimator of V_{sam}, v_{imp} is an estimator of V_{imp} and v_{mix} is an estimator of V_{mix}.

Fay (1991) used a different approach by reversing the order of sampling and response (we will call it the reverse approach) that can be depicted as:

Population → census with nonrespondents → sample with nonrespondents.

In this case (Shao and Steel, 1999),

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

where **a** is the vector of response indicators a_{ij} and noting that $V_r E_m E_p\left(\hat{Y}_I - Y|\mathbf{a}\right) = 0$ since the imputed estimator $\hat{Y}_I$ is unbiased under either assumption MB or assumption MBC. Under the reverse approach, an estimator of the overall variance $V\left(\hat{Y}_I - Y\right)$ is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ conditional on the vector of response indicators **a**, and v_2 is an estimator of $E_r V_m E_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$. The estimator v_1 does not depend on the response mechanism and/or the imputation model. As a result, v_1 is valid regardless of the assumptions made on the response mechanism and/or the imputation model. Since the component $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ represents only sampling variability, its estimation may be readily implemented using any standard variance estimation method: Taylor linearization, Jackknife and Bootstrap. In the context of two-stage sampling, the second component v_2 of v_t is typically negligible relative to v_1 if the overall sampling fraction is small.

In this paper, we focus on the case of simple random sampling without replacement at the first and second stages, so that $w_{ij} = \frac{N}{n}\frac{M_i}{m_i}$. Under complete response, an estimator of variance of $\hat{Y}$ is given by

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

where $s_Y^2 = \frac{1}{n-1}\sum_{i\in s}\left(\hat{Y}_i - \frac{\hat{Y}}{N}\right)^2$ with $\hat{Y}_i = \frac{M_i}{m_i}\sum_{j\in s_i} y_{ij}$ and $s_{iy}^2 = \frac{1}{m_i - 1}\sum_{j\in s_i}\left(y_{ij} - \bar{y}_i\right)^2$ with $\bar{y}_i = \hat{Y}_i / M_i$. We will use (20) to estimate $V_p\left(\hat{Y}_I - Y|\mathbf{a}\right)$ in section 4.2.

4.1 Särndal's method

In this section, we derive an estimator of the variance under assumption MBC and simple random sampling without replacement at the first and second stages. The estimation of V_{sam}, V_{imp} and V_{mix} may be performed as follows:

(i) Let v_p be the variance estimator under full response and let v_I be the "naïve" variance estimator of $\hat{Y}_I$ obtained from (20) by treating the imputed values as if they were observed. It is well known that for several imputation methods (in particular, for the deterministic methods), v_I underestimates V_{sam}. To compensate for this underestimation, evaluate the following expectation:

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

Then, determine a model unbiased estimator of V_{dif}, denoted v_{dif}. This will usually require the estimation of certain parameters of the imputation model *m*. Then,

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

is model unbiased for V_{sam}.

(ii) Then, determine a model unbiased estimator of V_{imp}, denoted v_{imp}, i.e., $E_m\left(v_{imp}|\tilde{s}\right) = V_{imp}$. Again, this may require the estimation of unknown parameters of the imputation model *m*.

(iii) Finally, determine a model unbiased estimator of V_{mix}, denoted v_{mix}, i.e., $E_m\left(v_{mix}|\tilde{s}\right) = V_{mix}$. Again, this may require the estimation of unknown parameters of the imputation model *m*.

Finally, an estimator of V_{tot}, denoted by v_{tot}, is given by

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

We now give the expressions of the different components in (21). First, note that the component v_I is obtained by replacing s_Y^2 and s_{iy}^2 by s_{IY}^2 and s_{Iiy}^2 in (20), where s_{IY}^2 and s_{Iiy}^2 are computed the same way as s_Y^2 and s_{iy}^2 by treating the imputed values as if they were true values. After tedious but straightforward algebra we obtain:

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

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

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

where r_i is the number of respondents to item y in s_i,

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

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

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

where o_i is the number of nonrespondents to item y in s_i, $B_i = \hat{p}^{-1}r_i - m_i$ with $\hat{p} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij} \Big/ \sum_{i\in s}\sum_{j\in s_i} w_{ij}$. Finally,

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

Note that $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ are estimators of σ_α^2 and σ_ε^2 respectively that can be obtained using available methods such as ML, REML or MINQUE methods. Also, note that the variance estimators under assumption MB are obtained by setting $\hat{\sigma}_\alpha^2 = 0$ in (22)-(24).

4.2 Shao-Steel's method

To apply the Shao-Steel method, we first express $\hat{Y}_I$ as $\hat{Y}_I = \hat{K}\hat{R}_a$, where $\hat{K} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}$ and $\hat{R}_a = \hat{Y}_a / \hat{K}_a$, with $\hat{Y}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}y_{ij}$ and $\hat{K}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}$. Now, denote the estimator of the variance of $\hat{Y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij}$ based on the full sample as $v(y)$. Then, one can show, using Taylor linearization, that v_1 reduces to

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

where

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

In the case of simple random sampling without replacement at the first and second stages, the component v_1 is obtained from (20) by replacing y_{ij} by $\hat{\xi}_{ij}$, where $\hat{\xi}_{ij}$ is given by (26) . To obtain v_2, note that $E_p(\hat{Y}_I - Y|\mathbf{a}) = \sum_{i \in U} \sum_{j \in U_i} b_{ij} y_{ij}$, where $b_{ij} = (K/K_a) a_{ij} - 1$ with $K_a = \sum_{i \in U} K_{ai}$ and $K_{ai} = \sum_{j \in U_i} a_{ij}$. Also, it is easy to show that

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

The component v_2 is then obtained by substituting estimators for the unknown quantities in (27), which leads to

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

where $w_i = \frac{1}{\pi_i}$, $\hat{K}_{ai} = \sum_{j \in s_i} w_{j|i} a_{ij}$ and $w_{j|i}$ is the survey weight of element *j* at the second stage given that cluster *i* has been selected in the first stage sample. Noting that $V_2(\hat{K}_{ai}) = E_2(\hat{K}_{ai}^2) - K_{ai}^2$, an estimate of K_{ai}^2 is given by $\hat{K}_{ai}^2 - \hat{V}_2(\hat{K}_{ai})$, where $\hat{V}_2(\hat{K}_{ai})$ is an estimate of $V_2(\hat{K}_{ai})$. The sum of (25) and (28) gives v_t . Note that the variance estimators under assumption MB are obtained by setting $\hat{\sigma}_\alpha^2 = 0$ in (28). Also, note that the ratio v_1 / v_2 is $O(n/K)$, where *K* is the total number of ultimate units. Hence, the component v_2 may be negligible with respect to v_1 even when the first-stage sampling fraction n/N is large. As a result, when *n*/*K* is negligible (as it is frequently the case in multistage designs), the computation of v_2 may be omitted.

5. SIMULATION STUDY

We conducted a limited simulation study on the relative performance of the variance estimators obtained by Särndal's method and the Shao-Steel's method. We generated several populations of $N = 120$ clusters with $M_i = M = 5, 20$ elements in cluster *i* such that the intracluster correlation $\rho_m = 0.05, 0.1, 0.2$. To do this, we first generated i.i.d. random variables

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

with $\mu = 200$ and $\sigma_\alpha^2 = 100$. Then, we generated i.i.d. random variables

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

with $\sigma_\varepsilon^2 = \frac{1 - \rho_m}{\rho_m} \sigma_\alpha^2$ for $\rho_m \neq 0$. The *y*-values are then generated by letting

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

From each generated population, we selected $R = 5000$ samples of clusters of size *n* according simple random sampling without replacement, where $n/N = 0.1, 0.5.$ To simplify the discussion, we consider single-stage cluster sampling, i.e., $m_i = M_i$. In each sample of cluster, we assigned a probability p_i to cluster $i, i = 1, ..., n,$ where p_i was generated from a beta distribution. Then, response indicators a_{ij} were generated from a Bernoulli distribution with parameter p_i. The overall response rate was set to 0.65.

From each sample, the imputed estimator $\hat{Y}_I$, given by (2), was computed and its variance was estimated using both Särndal's method and the Shao-Steel's method. For each method, the variance estimators were calculated under assumption MBC and assumption MB. The imputed estimator $\hat{Y}_I$ was approximately unbiased for all the scenarios with a Monte Carlo relative bias less than 0.3%.

As a measure of the bias of a variance estimator, $v\left(\hat{Y}_I\right)$, we used the relative bias

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

where $MSE\left(\hat{Y}_I\right)$ denotes the mean squared error of $\hat{Y}_I$. Table 1 and 2 report the relative bias values for *n*/*N* = 0.1 and 0.5, respectively.

From Table 1 and Table 2, we note that both Särndal's and Shao-Steel’s method perform well when the intracluster correlation is taken into account (i.e., when the variance estimators are derived under assumption MBC) since the absolute value of the relative bias of the variance estimators is less than 7% in all cases. Note that the component v_{mix} in Särndal's method is equal to 0 under this particular design. When the intracluster correlation is not taken into account (i.e., when the variance estimators are derived under assumption MB), Särndal's method leads to underestimation of the true variance. The underestimation increases as the intracluster correlation coefficient ρ_M and/or the cluster size M_i increases. For example, when *n*/*N* = 0.1 and $\rho_m = 0.1,$ the relative bias is -9.2% for $M_i = 5$ and -30.1% for $M_i = 20.$ Also, when *n*/*N* = 0.1 and $M_i = 20,$ the relative bias is -17.6% if $\rho_m = 0.05$ and -39.8% if $\rho_m = 0.2.$ Note that the sampling fraction *n*/*N* does not seem to have a significant impact on the relative bias of the variance estimators. Also, it is interesting to note that Särndal's method performs poorly when the imputation model is incorrectly specified mainly because the component v_{dif} tends to be severely biased in those cases. The Shao-Steel’s method, on the other hand, performs fairly well when the imputation model is incorrectly specified as the absolute value of the relative bias is less than 5% in all cases. This result is not surprising since the first component, v_1, of the total variance is robust in the sense that it does not depend on the assumptions made about the imputation model and since the second component v_2, is negligible with respect to v_1 when *n*/*K* is negligible. Note that, even when *n*/*N* = 0.5 and $M_i = 5$, we have *n*/*K* = 0.1, which may be considered as small. In summary, the Shao-Steel’s method seems to be robust unlike Särndal's method, when the assumptions about the model are violated.

Table 1
Relative bias (%) of the variance estimators with $n/N = 0.1$

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

Table 2
Relative bias (%) of the variance estimators with $n/N = 0.5$

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

6. EXTENSIONS

In this paper, we have investigated the case of mean imputation in two-stage sampling. The results obtained may be extended to the case of two-stage sampling and regression imputation. For example, suppose that a vector of q auxiliary variables $\mathbf{z} = \left(z_1,..., z_q\right)$ is observed on all the sampled elements (respondents and nonrespondents). In the case of assumption MBC, the imputation model is given by

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

where $\boldsymbol{\beta}$ is an unknown q-vector of fixed effects parameters and α_i and ε_{ij} are as before. Regression imputation uses $y^*_{ij} = \mathbf{z}'_{\mathbf{ij}}\hat{\boldsymbol{\beta}}$ or $y^*_{ij} = \mathbf{z}'_{\mathbf{ij}}\hat{\boldsymbol{\beta}} + \hat{\alpha}_i$, where $\hat{\boldsymbol{\beta}}$ is the ordinary or weighted least squares estimators of $\boldsymbol{\beta}$ and $\hat{\alpha}_i$ is the predictor of the random effect.

REFERENCES

Deville, J. C. and 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. and Steel, P. (1999), “Variance Estimation for Survey Data With composite Imputation and Nonnegligible Sampling Fractions”, *Journal of the American Statistical Association*, 94, 254-265.