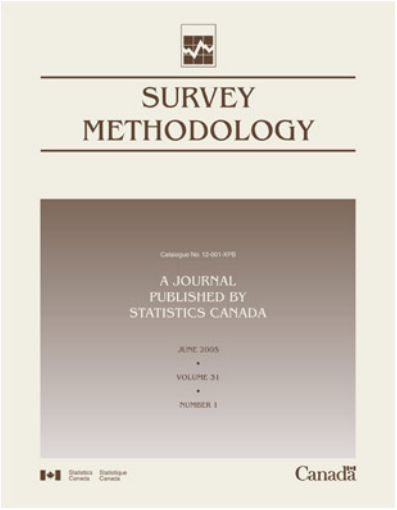
Catalogue no. 12-001-XIE

Survey Methodology

2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

2005

Published by authority of the Minister responsible for Statistics Canada

November 2005

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 1, pp. 81-90
Statistics Canada, Catalogue No. 12-001

Inference for Population Means Under Unweighted Imputation for Missing Survey Data

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

Abstract

In the presence of item nonresponse, unweighted imputation methods are often used in practice but they generally lead to biased estimators under uniform response within imputation classes. Following Skinner and Rao (2002), we propose a bias-adjusted estimator of a population mean under unweighted ratio imputation and random hot-deck imputation and derive linearization variance estimators. A small simulation study is conducted to study the performance of the methods in terms of bias and mean square error. Relative bias and relative stability of the variance estimators are also studied.

Key Words: Bias-adjusted estimator; Item nonresponse; Random hot-deck imputation; Ratio imputation.

1. Introduction

Item nonresponse occurs when a sampled unit fails to provide information on some variables of interest. Many surveys use imputation to handle item nonresponse but one should be aware of the difficulties when imputation is used. For example, the imputed values are commonly treated as if they are true values, and the variance estimates are computed using standard formulas. This can lead to serious underestimation of the true variance of the estimators when the proportion of missing values is not small. The relationships between variables may also be distorted.

Imputation methods can be classified into two broad classes: deterministic and stochastic. Deterministic methods include ratio or regression imputation and nearest neighbour imputation, using auxiliary variables observed on all the sampled units. For nearest neighbour imputation, a non-respondent item is assigned the respondent item value of the "nearest" respondent, where "nearest" is usually defined in terms of a distance function based on the auxiliary variables. Stochastic methods include random hot-deck imputation where the value assigned for a missing response is randomly selected from the set of respondents within an imputation cell.

In the presence of item nonresponse, weighted or unweighted imputation may be used. Weighted (deterministic or stochastic) imputation uses the sampling weights induced by the sampling design to select donors. However, weighted imputation is not feasible in practice when the sampling weights are not available at the imputation stage. Note that unweighted and weighted imputation methods lead to identical results for self-weighting designs (*i.e.*, designs with equal weights). Also, unweighted imputation methods are appealing to users.

Unweighted imputation generally leads to biased estimators under uniform response within imputation classes. Following the approach of Skinner and Rao (2002), we propose bias-adjusted estimators of population means under unweighted imputation and derive linearization variance estimators.

Let θ be a finite population parameter and $\hat{\theta}_I$ be its estimator based on the observed and imputed data respectively. Using the traditional two-phase approach: population → complete sample → sample with non-respondents, we have

$$E(\hat{\theta}_I) = E_p[E_r(\hat{\theta}_I)], \tag{1}$$

$$V(\hat{\theta}_I - \theta) = E_p[V_r(\hat{\theta}_I - \theta)] + V_p E_r[(\hat{\theta}_I - \theta)] \tag{2}$$

under deterministic imputation, where $E_r(.)$ and $V_r(.)$ denote respectively the expectation and the variance with respect to the response mechanism given the sample, and $E_p(.)$ and $V_p(.)$ denote respectively the expectation and the variance with respect to sampling under the given design. In the model-based approach (see section 2), we replace $E_r(.)$ and $V_r(.)$ by $\tilde{E}_m(.) = E_r E_m(.)$ and $\tilde{V}_m(.) = E_r V_m(.) + V_r E_m(.)$ respectively, where $E_m(.)$ and $V_m(.)$ denote respectively the expectation and the variance with respect to the imputation model.

Fay (1991) proposed a different approach obtained by reversing the order of sampling and response: population → census with nonrespondents → sample with non-respondents. Fay's approach facilities variance estimation, as explained below. Using this approach, we have

$$E(\hat{\theta}_I) = E_r[E_p(\hat{\theta}_I)], \tag{3}$$

and

$$V(\hat{\theta}_I - \theta) = E_r[V_p(\hat{\theta}_I - \theta)] + V_r[E_p(\hat{\theta}_I - \theta)], \tag{4}$$

1. David Haziza, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

see Shao and Steel (1999). Note that the inner expectation and variance in (4) are with respect to sampling, conditional on the response. An estimator of the overall variance $V(\hat{\theta}_I - \theta)$ in (4) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{\theta}_I - \theta)$ conditional on the response indicators, and v_2 is an estimator of $V_r E_p(\hat{\theta}_I - \theta)$. The estimator v_1 does not depend on the response mechanism or the assumed model, and hence v_1 is valid under either the design-based approach or the model-based approach (see section 2).

In the case of stochastic imputation, $V_p(\hat{\theta}_I - \theta)$ in (4) may be written as

$$V_p(\hat{\theta}_I - \theta) = V_p[E_*(\hat{\theta}_I - \theta)] + E_p[V_*(\hat{\theta}_I - \theta)], \quad (5)$$

where the inner expectation and variance, E_* and V_*, denote respectively the expectation and the variance with respect to the imputation scheme given the sample with respondents and non-respondents. An estimator of $V_p(\hat{\theta}_I - \theta)$, denote v_1^*, is then given by $v_1^* = v_1 + v_*$ where v_1 is an estimator of $V_p E_*(\hat{\theta}_I - \theta)$ and v_* an estimator of $E_p V_*(\hat{\theta}_I - \theta)$. Also, in the case of stochastic imputation we replace $E_p(.)$ by $E_p E_*(.)$ in (4) and the formula for v_2 is the same as in the case of deterministic imputation provided $E_*(\hat{\theta}_I)$ agrees with the imputed estimator for the deterministic case. Hence, an estimator of the deterministic case. Hence, an estimator of the overall variance $V(\hat{\theta}_I - \theta)$ is given by $v_t = v_1 + v_* + v_2$.

We set out our basic framework and assumptions in section 2. In section 3, we study both weighted and unweighted ratio imputation. We show that the imputed estimator under unweighted imputation is asymptotically biased, and propose a bias-adjusted estimator. The estimator under unweighted imputation is asymptotically biased, and propose a bias-adjusted estimator. The estimator under weighted imputation and the bais-adjusted estimator under unweighted imputation are shown to be robust in the sense of validity under both the design-based and model-based approaches. We also derive linearization variance estimators of the imputed estimators in section 3. We consider the case of random hot-deck imputation in section 4. A small simulation is conducted in section 5 to compare the performances of the imputed estimators in terms of bias and mean square error. Relative bias and relative stability of the variance estimators are also studied.

2. Framework and Assumptions

Let P be a finite population of possibly unknown size N. The objective is to estimate the population mean $\bar{Y} = 1/N\sum_p y_i$ when imputation has been used to compensate for nonresponse. For brevity, $\sum_A$ will be used for $\sum_{i\in A}$, where $A \subseteq P$. Suppose a probability sample, s, of size n is selected according to a specified design $p(s)$ from P. Let s_r be the set of respondents of size r and let s_m be the set of nonrespondents of size $m; r + m = n$.

Imputation is often done by first dividing the population into J nonoverlapping imputation classes and then imputing sample nonrespondents within each imputation class using sample respondents within the same class as donors, independently across the J imputation classes. For simplicity, we assume that $J = 1$; the extension to $J > 1$ imputation classes is straightforward.

The usual imputed estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_I = \frac{1}{\sum_s w_i}\left[\sum_{s_r} w_i y_i + \sum_{s_m} w_i y_i^*\right], \quad (6)$$

where w_i is the sampling (or design) weight attached to unit i and y_i^* denotes the value imputed for missing y_i. We use the Horvitz-Thompson weight $w_i = 1/\pi_i$, where π_i is the probability of including unit *i* in the sample.

We consider two approaches: (i) design-based and (ii) model-based. Under the design-based approach, we assume a uniform response mechanism within classes so that the following assumption holds:

Assumption DB: Within an imputation cell, the response probability for a given variable of interest is constant and the response statuses for different units are independent.

Under the model-based approach, the following assumption holds:

Assumption MB: Within an imputation cell the response mechanism is ignorable or unconfounded in the sense that the response status of a unit does not depend on the variable being imputed but may depend on covariates used for imputation. In this case, an imputation model is assumed.

The imputation classes are chosen to make the assumption DB or MB hold approximately. The response mechanism in assumption MB is much weaker than the uniform response in assumption DB, but inferences depend on the assumed imputation model. Under ratio imputation, the imputation model used is the "ratio model" given by

$$E_m(y_i) = \beta z_i, V_m(y_i) = \sigma^2 z_i, \mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j, \quad (7)$$

where β and σ^2 are unknown parameters, z_i is an auxiliary variable available for all $i \in s$. Under random hot-deck imputation, the imputation model used is given by

$$E_m(y_i) = \mu, V_m(y_i) = \sigma^2, \mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j. \quad (8)$$

3. Ratio Imputation

In this section, we study the properties of the imputed estimator (6) under both weighted and unweighted ratio imputation. We also derive linearization variance estimators. We study point estimation in section 3.1 under weighted and unweighted ratio imputation, and corresponding variance estimation in section 3.2.

3.1 Estimation of a Mean

3.1.1 Weighted Ratio Imputation

Weighted ratio imputation uses $y_i^* = \hat{R}_r\ z_i$ for missing y_i, where $\hat{R}_r = \bar{y}_r / \bar{z}_r$ and $(\bar{y}_r,\ \bar{z}_r) = \sum_{s_r} w_i (y_i,\ z_i) / \sum_{s_r} w_i$ are the weighted means of respondents for variables y and z respectively. Using the y_i^*'s, the imputed estimator (6) reduces to

$$\bar{y}_{\mathrm{IR}} = \hat{R}_r\ \bar{z}, \tag{9}$$

where $\bar{z} = \sum_s w_i z_i / \sum_s w_i$. It is easy to verify that $\bar{y}_{\mathrm{IR}}$ is approximately unbiased for $\bar{Y}$ under both the design-based and the model-based approaches, (Särndal 1992). Hence $\bar{y}_{\mathrm{IR}}$ is robust in the sense of validity under both approaches.

3.1.2 Unweighted Ratio Imputation

Unweighted ratio imputation uses $y_i^* = \hat{R}_r^{\mathrm{un}} z_i$ for missing y_i, where $\hat{R}_r^{\mathrm{un}} = \bar{y}_r^{\mathrm{un}} / \bar{z}_r^{\mathrm{un}}$ and $(\bar{y}_r^{\mathrm{un}},\ \bar{z}_r^{\mathrm{un}}) = \sum_{s_r} (y_i,\ z_i)/r$ are the unweighted means of respondents for variables y and z respectively. Using the y_i^*'s, the imputed estimator (6) reduces to

$$\bar{y}_{\mathrm{IR}} = \frac{1}{\sum_s w_i} \left[\sum_{s_r} w_i\ y_i + \hat{R}_r^{\mathrm{un}} \sum_{s_m} w_i\ z_i \right], \tag{10}$$

where $\hat{R}_r^{\mathrm{un}} = \bar{y}_r^{\mathrm{un}} / \bar{z}_r^{\mathrm{un}}$. Under the ratio model (7) and assumption MB, the imputed estimator (10) is approximately unbiased for $\bar{Y}$, *i.e.*, $E_r E_p E_m (\bar{y}_{\mathrm{IR}}) \approx E_m(\bar{Y})$. However, it is biased under uniform response (assumption DB). We have $E_p E_r (\bar{y}_{\mathrm{IR}}) \approx p\bar{Y} + (1-p)\bar{Y}_\pi / \bar{Z}_\pi \bar{Z}$, where $(\bar{Y}_\pi,\ \bar{Z}_\pi) = \sum_p \pi_i (y_i, z_i) / \sum_p \pi_i$. Hence, the relative bias of $\bar{y}_{\mathrm{IR}}$, $\mathrm{RB}(\bar{y}_{\mathrm{IR}}) = (E_p E_r (\bar{y}_{\mathrm{IR}}) - \bar{Y}) / \bar{Y}$, is given by

$$\mathrm{RB}(\bar{y}_{\mathrm{IR}}) \approx (1-p) \left[\frac{\bar{Z}}{\bar{Z}_\pi} \frac{\bar{Y}_\pi}{\bar{Y}} - 1 \right] \tag{11}$$

$$\approx (1-p)\ \frac{\bar{Z}}{\bar{Z}_\pi}\ C_\pi\ [C_y \rho_{\pi y} - C_z \rho_{\pi z}], \tag{12}$$

where $\bar{Z} = 1/N \sum_p z_i$, $\rho_{\pi y}$ and $\rho_{\pi z}$ are the finite population correlation coefficients between the variables π and y and π and z respectively, C_π, C_z and C_y are respectively the coefficients of variation of π, z and y, and p is the probability or response to y. The bias is nonzero generally. It vanishes in the full response case (*i.e.*, $p = 1$) or if

$$C_\pi [C_y \rho_{\pi y} - C_z \rho_{\pi z}] = 0, \tag{13}$$

which is satisfied when $C_\pi = 0$ (the case when the design is self-weighting) or when

$$\frac{\rho_{\pi y}}{\rho_{\pi z}} = \frac{C_z}{C_y}. \tag{14}$$

We further explore the relative bias (11) for three cases. First, we consider unweighted mean imputation, $y_i^* = \bar{y}_r^{\mathrm{un}}$, which is a special case of unweighted ratio imputation with $z_i = 1$. Assume that a size variable x is available for all the units in the population and that the sample s is selected according to a probability proportional to size (PPS) sampling without replacement design, using x as the size, such that $\pi_i = n x_i / X$, where $X = \sum_P x_i$. For example, one may use the well-known Sampford method (Sampford 1967). Noting that $\rho_{\pi y} = \rho_{xy}$, $\bar{Z} / \bar{Z}_\pi = 1$ and $C_\pi = C_x$, the expression (12) for the relative bias may be written as

$$\mathrm{RB}(\bar{y}_{\mathrm{IR}}) \approx (1-p)\, C_x C_y \rho_{xy}. \tag{15}$$

Two particular cases of (15) are of interest. First, if x and y are uncorrelated, the bias of the imputed estimator vanishes. The case of weakly correlated x and y (*i.e.*, $\rho_{xy} \approx 0$) may occur in surveys with multiple characteristics y (Rao 1966). Second, if $y_i \propto x_i$, the relative bias (15) reduces to $(1-p)\, C_x^2$ which decreases with C_x. Note that, since $C_x = C_\pi$, the sampling design approaches a self-weighting design as C_x decreases.

Consider next the more general case of unweighted ratio imputation based on z_i, $i \in s$, and PPS sampling based on x_i, $i \in s$. In this case, the relative bias (11) is zero if and only if

$$\frac{\rho_{xy}}{\rho_{xz}} = \frac{C_y}{C_z},$$

provided $p < 1$ and $C_\pi \neq 0$. If $C_y = C_z$, then the relative bias (11) is zero if and only if $\rho_{xy} = \rho_{xz}$.

Finally, we consider the case of stratified random sampling. In this case, the population P is partitioned into H strata P_h with N_h sampling units in the h^{th} stratum; $P = \bigcup_{h=1}^H P_h$, $N = \sum_{h=1}^H N_h$. We then independently select a simple random sample without replacement s_h of size n_h from each stratum; $s = \bigcup_{h=1}^H s_h$ and $n = \sum_{h=1}^H n_h$. Two situations may occur in practice: (1) Imputation is done independently in each stratum (*i.e.*, the imputation classes coincide with the strata). In this case, under unweighted ratio imputation, the imputed estimator is approximately unbiased under uniform response within strata. (2) The imputation is done across strata. In this case, we note from (11) that the imputed estimator is approximately unbiased if and only if $n_h = n(N_h / N)$ (proportional allocation).

A bias-adjusted estimator of $\bar{Y}$ under unweighted ratio imputation is given by

$$\bar{y}_{\mathrm{IR}}^a = \hat{p}^{-1}\ \bar{y}_{\mathrm{IR}} + (1 - \hat{p}^{-1})\ \frac{\bar{z}}{\bar{z}^{un}}\ \bar{y}_{\mathrm{IR}}^{\mathrm{un}}, \tag{16}$$

where $\hat{p} = (\sum_{s_r} w_i / \sum_s w_i)$ is a consistent estimator of the response probability p, $\bar{z}^{\mathrm{un}} = 1/n \sum_s z_i$ and $\bar{y}_{\mathrm{IR}}^{\mathrm{un}}$ is the unweighted mean of the observed values y_i and the imputed values $y_i^* = \hat{R}_r^{\mathrm{un}} z_i$. This estimator may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\bar{y}_{\mathrm{IR}}) = p\bar{Y} + (1-p)\ \frac{\bar{Y}_\pi}{\bar{Z}_\pi}\ \bar{Z}$$

for $\bar{Y}$ and replacing $E(\bar{y}_{\mathrm{IR}})$ by its estimator $\bar{y}_{\mathrm{IR}}$, $(\bar{Y}_\pi / \bar{Z}_\pi)\, \bar{Z}$ by its estimator

$$\hat{R}_r^{\text{un}}\ \bar{z} = \left(\frac{\bar{z}}{\bar{z}^{\text{un}}}\right)\bar{y}_{\text{IR}}^{\text{un}}, \tag{17}$$

and p^{-1} by its estimator $\hat{p}^{-1}$. Note that the estimator $\bar{z}$ of $\bar{Z}$ makes use of the full sample $z-$values, unlike $\bar{z}_r$. If $\bar{z}_r$ is used to estimate $\bar{Z}$, then the bias-adjusted estimator requires response identifiers, unlike (16).

We now show that the bias-adjusted estimator (16) is approximately unbiased under both the design-based and the model-based approaches. Hence, unlike the unadjusted estimator (10), the adjusted estimator is robust in the sense of validity under both approaches. First, noting that $\bar{y}_{\text{IR}}$ may be expressed as $\hat{p}\,\bar{y}_r + \hat{R}_r^{\text{un}}(\bar{z} - \hat{p}\bar{z}_r)$ and using (17), the bias-adjusted estimator (16) reduces to

$$\bar{y}_{\text{IR}}^a = \bar{y}_r + \hat{R}_r^{\text{un}}(\bar{z} - \bar{z}_r). \tag{18}$$

Comparing (9) and (18), we see that $\bar{y}_{\text{IR}}$ under weighted ratio imputation is not equal to the bias-adjusted estimator $\bar{y}_{\text{IR}}^a$ under unweighted ratio imputation, unless $z_i = 1$ for all i. In the latter case, both estimators reduce to $\bar{y}_r$. However, the form (16) for $\bar{y}_{\text{IR}}^a$ does not require response identifiers, provided $\hat{p}$ is available.

Since $E_m(\bar{y}_{\text{IR}}^a) = \beta\bar{z}$ and $E_m(\bar{Y}) = \beta\bar{Z}$ under the ratio model (7), we have $E_p\,E_m(\bar{y}_{\text{IR}}^a - \bar{Y}) \approx 0$; that is, the adjusted estimator is approximately unbiased under the model-based approach. On the other hand, since $E_p\,E_r(\bar{y}_r) \approx \bar{Y}$ and $E_r(\bar{z} - \bar{z}_r) \approx 0$ under uniform response, it follows that $E_p E_r(\bar{y}_{\text{IR}}^a) \approx \bar{Y}$ so that the adjusted estimator is approximately design-unbiased under uniform response.

We note several points here: (1) The survey analyst can easily implement the adjusted estimator $\bar{y}_{\text{IR}}^a$, given by (16), from the imputed data file without response identifiers, *i.e.*, $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$. Note that the response identifiers are not needed on the data file, but the response rate $\hat{p}$ should be available to the analyst, which we assume to be the case here. In the case of multiple imputation classes, response rates within classes and imputation class identifiers need to be provided with the file. (2) The bias-adjusted estimator coincides with the unadusted estimator $\bar{y}_{\text{IR}}$, given by (10), under a self-weighting design $w_i = w$. (3) The adjusted estimator $\bar{y}_{\text{IR}}^a$ in (18) has the form of a regression estimator in two-phase sampling. (4) Under mean imputation, (18) reduces to the weighted mean of respondents $\bar{y}_r$, so the correction made to the unadjusted estimator eliminates the effect of using unweighted mean imputation.

Another approach to getting a bias-adjusted estimator, $\bar{y}_{\text{IR}}^a$, is to subtract an estimator, $b(\bar{y}_{\text{IR}})$, of the bias of $\bar{y}_{\text{IR}}$, from $\bar{y}_{\text{IR}}$, *i.e.*,

$$\bar{y}_{\text{IR}}^a = \bar{y}_{\text{IR}} - b(\bar{y}_{\text{IR}}). \tag{19}$$

It follows from (11) that an estimator of the bias of $\bar{y}_{\text{IR}}$ is given by

$$b^{(1)}(\bar{y}_{\text{IR}}) = (1-\hat{p})\quad(\hat{R}_r^{\text{un}}\,\bar{z} - \bar{y}_r). \tag{20}$$

But the resulting bias-adjusted estimator is not identical to (16), and it depends on response identifiers, unlike (16). On the other hand, if one uses

$$b^{(2)}(\bar{y}_{\text{IR}}) = (1-\hat{p})(\hat{R}_r^{\text{un}}\,\bar{z}_r - \bar{y}_r), \tag{21}$$

it is easy to verify that the resulting bias-adjusted estimator is identical to (16).

3.2 Variance Estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

3.2.1 Variance Estimation Under Weighted Ratio Imputation

In this subsection, we obtain a linearization variance estimator of the imputed estimator (9) based on weighted ratio imputation, using the reverse approach of Fay (1991). First, express (9) as

$$\bar{y}_{\text{IR}} = \frac{\sum_s w_i\,a_i\,y_i}{\sum_s w_i\,a_i\,z_i}\bar{z},$$

where a_i is a response indicator to item y such that $a_i = 1$ if $i \in s_r$ and $a_i = 0$, otherwise. It follows from (4) that the variance $V(\bar{y}_{\text{IR}})$ of $\bar{y}_{\text{IR}}$ can be estimated by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\bar{y}_{\text{IR}} - \bar{Y})$ conditional on the a_i's, and v_2 is an estimator of $V_r\,E_p(\bar{y}_{\text{IR}} - \bar{Y})$. Denote the estimator of the variance of the estimated total $\hat{Y} = \sum_s w_i\,y_i$ based on the full sample as $v(y_i)$. Then, using the delta method, a linearization variance estimator, v_1, in the operator notation $v(.)$, is given by

$$v_1 = v(\hat{\xi}), \tag{22}$$

where the value of $\hat{\xi}$ for $i \in s$ is given by

$$\hat{\xi}_i = \frac{1}{\sum_s w_i}[\hat{\xi}_{1i} - \bar{y}_{\text{IR}}],$$

with

$$\hat{\xi}_{1i} = a_i\,y_i + (1-a_i)\hat{R}_r\,z_i + \hat{c}a_i(y_i - \hat{R}_r z_i),$$

where

$$\hat{c} = \frac{\sum_s w_i(1-a_i)z_i}{\sum_s w_i\,a_i\,z_i}.$$

Note that v_1 is valid regardless of the response mechanism and the imputation model. The derivation of (22) is given in Appendex A. Shao and Steel (1999) derived a linearization variance estimator of the imputed estimator $\hat{Y} = \sum_s w_i\,a_i\,y_i + \sum_s w_i(1-a_i)\hat{R}_r\,z_i$ of the total Y. They first expressed $\hat{Y}$ as

$$\hat{Y} = \sum_s w_i[a_i\,y_i + (1-a_i)R_a\,z_i] + \hat{c}\sum_s w_i\,a_i\,(y_i - R_a z_i),$$

where $R_a = Y_a / Z_a$ with $(Y_a, Z_a) = \sum_p a_i (y_i, z_i)$ and then replaced $\hat{c}$ by $\tilde{c} = \sum_p (1-a_i) z_i / \sum_p a_i z_i$ to get linear approximation for $\hat{Y} \approx \sum_s w_i \eta_i$, where

$$\eta_i = a_i y_i + (1-a_i) R_a z_i + \tilde{c} a_i (y_i - R_a z_i).$$

Now replacing R_a by $\hat{R}_r$ and $\tilde{c}$ by $\hat{c}$ in the above expression for η_i we get $\hat{\eta}_i = a_i y_i + (1-a_i)\hat{R}_r z_i + \hat{c} a_i (y_i - \hat{R}_r z_i)$ which leads to the linearization variance estimator $v_1 = v(\hat{\eta})$. The delta method in Appendix A may be used to obtain this result in straightforward manner.

Next, using the delta method,

$$V_r E_p (\bar{y}_{\mathrm{IR}} - \bar{Y}) \approx p(1-p) \left(\frac{Z}{E_r(Z_a)} \right)^2 \frac{S_e^2}{N}, \qquad (23)$$

Under assumption DB where $Z = \sum_p z_i$, and $S_e^2 = 1/N \sum_p (y_i - E_r(R_a) z_i)^2$. The component v_2 is then obtained by substituting estimators for the unknown quantities in (23).

We obtain

$$v_2 = \hat{p}(1-\hat{p}) \left(\frac{\hat{Z}}{\hat{Z}_a} \right)^2 \frac{s_{er}^2}{\hat{N}}, \qquad (24)$$

where $\hat{Z} = \sum_s w_i z_i$, $\hat{Z}_a = \sum_s w_i a_i z_i$, $\hat{N} = \sum_s w_i$ and

$$s_{er}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i (y_i - \hat{R}_r z_i)^2.$$

The sum of (22) and (24) gives v_t, the estimator of the overall variance of $\bar{y}_{\mathrm{IR}}$.

3.2.2 Variance Estimation Under Unweighted Ratio Imputation

We now give a linearization estimator of variance of the imputed estimator (10) based on unweighted ratio imputation. Using the delta method, see Appendix A, we obtain

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

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i} [\hat{\xi}_{1i} - \bar{y}_{\mathrm{IR}}],$$

with

$$\hat{\xi}_{1i} = a_i y_i + (1-a_i) \hat{R}_r^{\mathrm{un}} z_i + \hat{d} \frac{a_i}{w_i} (y_i - \hat{R}_r^{\mathrm{un}} z_i)$$

and $\hat{d} = \sum_s w_i (1-a_i) z_i / \sum_s a_i z_i$. The component v_2 is given by (B.2) in Appendix B.

3.2.3 Variance Estimation for the Bias-Adjusted Estimator

In this subsection, we give a linearization variance estimator of the bias-adjusted estimator (18). Using the delta method, we obtain

$$v_1 = v(\hat{\xi}), \qquad (26)$$

where

$$\hat{\xi}_i = \frac{a_i}{\sum_s w_i a_i} [(y_i - \bar{y}_r) + \hat{R}_r^{\mathrm{un}} (z_i - \bar{z}_r)] + \frac{\hat{R}_r^{\mathrm{un}}}{\hat{N}} (z_i - \bar{z})$$

$$+ (\bar{z} - \bar{z}_r) \frac{1}{\sum_s a_i z_i} \frac{a_i}{w_i} (y_i - \hat{R}_r^{\mathrm{un}} z_i);$$

see Appendix A. The component v_2 is given by (C.2) in Appendix C.

4. Random Hot-Deck Imputation

In this section, we study the properties of the imputed estimator (6) under weighted and unweighted random hot-deck imputation. We also derive linearization variance estimators under uniform response.

4.1 Estimation of a Mean

In section 4.1 we study point estimators under weighted and unweighted random hot-deck imputation.

4.1.1 Weighted Random Hot-Deck Imputation

Under weighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with selection probabilities $w_j / \sum_{s_r} w_i$ and use $y_i^* = y_j$, $i \in s_m$. The imputed estimator, $\bar{y}_{\mathrm{IH}}$, is given by (6) with the above imputed values. It is approximately unbiased for the population mean $\bar{Y}$ under both the design-based and the model-based approaches. The latter uses the mean model (8).

4.1.2 Unweighted Random Hot-Deck Imputation

Under unweighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with equal probabilities $1/r$ and use $y_i^* = y_j$, $i \in s_m$. The imputed estimator, $\bar{y}_{\mathrm{IH}}$, is given by (6) with the above imputed values. It is approximately unbiased for $\bar{Y}$ under the mean model (8), but biased under uniform reponse. The bias of $\bar{y}_{\mathrm{IH}}$ is given by

$$B(\bar{y}_{\mathrm{IH}}) = (1-p)(\bar{Y}_\pi - \hat{Y}). \qquad (27)$$

A biased-adjusted estimator of $\bar{Y}$ under unweighted random hot-deck imputation is given by

$$\bar{y}_{\mathrm{IH}}^a = \hat{p}^{-1} \bar{y}_{\mathrm{IH}} + (1 - \hat{p}^{-1}) \bar{y}_{\mathrm{IH}}^{\mathrm{un}}, \qquad (28)$$

where $\hat{p} = (\sum_{s_r} w_i / \sum_s w_i)$ is a consistent estimator of the response probability p and $\bar{y}_{\mathrm{IH}}^{\mathrm{un}}$ is the unweighted mean of the observed values y_i and the imputed values y_i^*. The estimator (28) may be derived from the method of moments following Skinner and Rao (2002), by solving

$$E(\bar{y}_{\mathrm{IH}}) = p \bar{Y} + (1-p) \bar{Y}_\pi$$

for $\bar{Y}$ replacing by $E(\bar{y}_{\mathrm{IH}})$ its estimator $\bar{y}_{\mathrm{IH}}$, $\bar{Y}_\pi$ by its estimator $\bar{y}_{\mathrm{IH}}^{\mathrm{un}}$ and p^{-1} by its estimator $\hat{p}^{-1}$. The adjusted estimator is approximately unbiased for $\bar{Y}$ under both the design-based and the model-based approaches. As in section

3.1.2, note that the survey analyst can easily implement the adjusted estimator $\bar{y}_{\mathrm{IR}}^{a}$ from the imputed data file without response identifiers, *i.e.*, $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$, provided the response rate, $\hat{p}$, is available.

Note that the method of subtracting an estimator of the bias of $\bar{y}_{\mathrm{I}}$ from $\bar{y}_{\mathrm{I}}$, using (27), will lead to a bias-adjusted estimator that depends on response identifiers, unlike (28). It is not possible to obtain the bias-adjusted estimator (28) by this approach, unlike in the case of deterministic ratio imputation studied in subsection 3.1.2.

4.2 Variance Estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

4.2.1 Variance Estimation Under Weighted Random Hot-Deck Imputation

We now obtain a linearization variance estimator of the imputed estimator $\bar{y}_{\mathrm{IH}}$ under weighted random hot-deck imputation. First, note that under weighted random hot-deck imputation, $E_*(\bar{y}_{\mathrm{IH}}) = \bar{y}_r$. This is a particular case of (9) with $z_i = 1$ for all i. Hence, using (22), v_1 is given by

$$v_1 = v(\hat{\xi}), \tag{29}$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i}[\hat{\xi}_{1i} - \bar{y}_r],$$

$$\hat{\xi}_i = a_i\, y_i + (1 - a_i)\,\bar{y}_r + \hat{c}\, a_i (y_i - \bar{y}_r),$$

with $\hat{c} = \sum_s w_i (1 - a_i) / \sum_s w_i a_i$. Straightforward algebra shows that $\hat{\xi}_i$ simplifies to $\hat{\xi}_i = a_i (y_i - \bar{y}_r) / \sum_s w_i a_i$. Now, noting that $V_*(y_i^*) = (1/\sum_s w_i a_i) \sum_s w_i a_i (y_i - \bar{y}_r)^2 = s_{yr}^2$, we have

$$v_* = V_*(\bar{y}_{\mathrm{IH}} - \bar{Y}) = \frac{\sum_s w_i^2 (1 - a_i)}{\left(\sum_s w_i\right)^2} s_{yr}^2. \tag{30}$$

As noted in section 1, v_2 is the same as for the deterministic case. Hence, under weighted random hot-deck imputation, v_2 is given by (24) with $z_i = 1$ for all i, which lead to

$$v_2 = \hat{p}(1 - \hat{p}) \left(\frac{\hat{N}}{\sum_s w_i a_i} \right)^2 \frac{s_{yr}^2}{\hat{N}}. \tag{31}$$

The sum of (29), (30) and (31) gives v_t, the estimator of overall variance.

4.2.2 Variance Estimation Under Unweighted Random Hot-Deck Imputation

We now obtain a linearization estimator of variance of the imputed estimator (6) under unweighted random hot-deck imputation. First, note that $E_*(\bar{y}_{\mathrm{IH}})$ reduces to (10) with $z_i = 1$ for all i. Hence, v_1 is given by

$$v_1 = v(\hat{\xi}), \tag{32}$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i}[\hat{\xi}_{1i} - E_*(\bar{y}_{\mathrm{IH}})],$$

$$\hat{\xi}_{1i} = a_i\, y_i + (1 - a_i)\,\bar{y}_r^{\mathrm{un}} + \hat{d}\, \frac{a_i}{w_i} (y_i - \bar{y}_r^{\mathrm{un}}),$$

with $\hat{d} = \sum_s w_i (1 - a_i) / \sum_s a_i$. Now, nothing that $V_*(y_i^*) = (1/\sum_s a_i) \sum_s a_i (y_i - \bar{y}_r^{\mathrm{un}})^2 s_{yr}^{2\mathrm{un}}$, we have

$$v_* = \frac{\sum_s w_i^2 (1 - a_i)}{\left(\sum_s w_i\right)^2} s_{yr}^{2\mathrm{un}}. \tag{33}$$

As noted in section 1, v_2 is the same as for the deterministic case. Hence, under unweighted random hot-deck imputation, v_2 is given by (B.2) with $z_i = 1$ for all i. The sum of (32), (33) and (B.2) gives v_t.

4.2.3 Variance Estimation for the Bias-Adjusted Estimator

We now obtain a linearization variance estimator of the bias-adjusted estimator given by (28). First, note that, $E_*(\bar{y}_{\mathrm{IH}}^{a})$ reduces to $\bar{y}_r$, the mean of the $y-$values respondent. Hence, v_1 is given by (29) and v_2 is given by (31). Now, noting that $V_*(y_i^*) = (1/\sum_s a_i) \sum_s a_i (y_i - \bar{y}_r^{\mathrm{un}})^2 = s_{yr}^{2\mathrm{un}}$ and $\mathrm{Cov}_*(y_i^*, y_j^*) = 0$ for $i \neq j$, one can show that $V_*(\bar{y}_{\mathrm{IH}}^{a} - \bar{Y})$ is given by

$$v_* = \left[\frac{\hat{p}^{-2}}{\left(\sum_s w_i\right)^2} \sum_s w_i^2 (1 - a_i) - (1 - \hat{p}^{-1})^2 \left(\frac{r + n}{n^2} \right) \right] s_{yr}^{2\mathrm{un}}. \tag{34}$$

The sum of (29), (31) and (34) gives v_t. Note that even though v_* given by (34) is expressed as the difference between two terms, it is always nonnegative, as shown in Appendix D.

5. Simulation Study

As a complement to the theory, we present some results from a limited simulation study. We generated a population of $N = 800$ values (y_i, z_i) according to the ratio model $y = \beta z + \varepsilon$, where z and ε were generated from a normal distribution such that the correlation, ρ_{yz}, between y and z equaled 0.05, 0.30, 0.70 and 0.90. The objective is to estimate the population total $Y = \sum_p y_i$. We drew $R =$ 10,000 PPS samples, each of size $n = 75$, according to Sampford's pps sampling method, using item z as the measure of size. Nonresponse to item y was then generated from each PPS sample according to a uniform response

mechanism with a response rate of 0.7; item z was observed for all units in the sample. We used weighted and unweighted random hot-deck imputation to compensate for nonresponse to item y.

The estimator of the first component in the variance formula (4) was computed using the well know Sen-Yates-Grundy estimator. Let $v(\xi)$ denote the variance estimator of $\sum_s w_i \xi_i$. The Sen-Yates-Grundy estimator of variance is then given by

$$v(\xi) = \frac{1}{2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\xi_i}{\pi_i} - \frac{\xi_j}{\pi_j} \right)^2, \qquad (35)$$

where $\pi_{ij} = P(i \in s \text{ and } j \in s)$ is the joint probability of inclusion of units i and j in the sample. Sampford's method ensures $\pi_i \pi_j - \pi_{ij} \geq 0$ for all i, j so that the variance estimator in (35) us always nonnegative.

As a measure of the bias of an imputed estimator $\hat{Y}_I$ of Y, we used the bias ratio $B_r(\hat{Y}_I) = \text{Bias}(\hat{Y}_I)/\text{s.e.}(\hat{Y}_I)$, where $\text{s.e.}(\hat{Y}_I)$ denotes the standard error of $\hat{Y}_I$. To compare the efficiencies, we used the coefficient of variation of $\hat{Y}_I$, denoted $\text{CV}(\hat{Y}_I)$ and given by $\text{CV}(\hat{Y}_I) = (\sqrt{\text{MSE}}/Y)$. The variance estimators were compared in terms of their relative bias and CV. The relative bias of a variance estimator, v_t, is measured by $B_{\text{rel}}(v_t) = (E(v) - \text{MSE}(\hat{Y}_I))/\text{MSE}(\hat{Y}_I)$ and its CV is given by $\text{CV}(v_t) = \sqrt{\text{MSE}}(v_t)/\text{MSE}(\hat{Y}_I)$. Values of the above measures were calculated from the simulated PPS samples.

Table 1 reports the simulation results on the bias ratio (B_r) of the three imputed estimators of Y, denoted B_r (weighted), B_r (unweighted) and B_r (adjusted) and the CVs of the estimators, denoted CV (weighted), CV (unweighted) and CV (adjusted). It is clear from Table 1 that the bias ratio of the estimator under unweighted imputation is large $(\geq 30\%)$ if $\rho_{xy} \geq 0.5$, while the bias ratios of the estimator under weighted imputation and the adjusted estimator are small $(\leq 4\%)$ for all values of ρ_{xy}. Due to large bias, the CV of the unweightede estimator is large that the CV of the weighted estimator if $\rho_{xy} \geq 0.5$ and also larger than the CV of the adjusted estimator if $\rho_{xy} \geq 0.7$, but the increase in CV is not large. Also, CV (weighted) is slightly smaller than CV (adjusted) for all values of ρ_{xy}.

Table 2 reports the relative bias (B_{rel}) and the CV ratios of the variance estimators. As expected, the variance estimator v_t (unweighted) leads to serious underestimation of MSE of the estimator for large $\rho_{xy}\ (\geq 0.7)$, while the absolute relative bias of the variance estimators v_t (weighted) and v_t (adjusted) is small $(\leq 6\%)$ for all values of ρ_{xy}. Turning to the CV ratios of the variance estimators, Table 2 shows that v_t (unweighted) has the smallest CV followed by v_t (weighted) and v_t (adjusted) for $\rho_{xy} \geq 0.3$.

Table 1
Bias Ratio (%) and CV (%) of the Imputed Estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
B_r (weighted)	–0.78	1.99	–0.79	0.40	3.27
B_r (unweighted)	1.82	18.60	30.50	49.20	64.20
B_r (adjusted)	–1.12	1.47	0.01	0.61	2.94
CV(weighted)	18.80	15.30	11.60	5.87	4.69
CV(unweighted)	18.00	15.20	12.50	6.83	5.93
CV(adjusted)	20.90	16.80	13.50	6.10	4.78

Table 2
Relative Bias (%) of the Variance Estimators and
Comparisons of the CV Ratios of the Variance Estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
$B_{\text{rel}}(v_t)$ (weighted)	–2.43	–4.78	–4.28	3.96	–1.95
$B_{\text{rel}}(v_t)$ (unweighted)	–1.03	–3.47	–11.80	–18.50	–29.30
$B_{\text{rel}}(v_t)$ (adjusted)	–5.42	–1.06	–4.21	1.61	0.07
$\frac{\text{CV}(v_t) \text{ (unweighted)}}{\text{CV}(v_t) \text{ (weighted)}}$	1.016	0.984	0.931	0.875	0.781
$\frac{\text{CV}(v_t) \text{ (unweighted)}}{\text{CV}(v_t) \text{ (adjusted)}}$	1.032	0.829	0.701	0.819	0.692
$\frac{\text{CV}(v_t) \text{ (weighted)}}{\text{CV}(v_t) \text{ (adjusted)}}$	1.016	0.843	0.751	0.935	0.886

6. Concluding Remarks

Unweighted imputation methods are often used in practice to compensate for item nonresponse when the survey weights are not available at the imputation stage. Also, unweighted imputation is appealing to users even when the weights are available at the imputation stage. But it leads to biased estimators under uniform response within imputation classes. We have proposed bias-adjusted estimators under ratio imputation and random hot-deck imputation. These estimators can be implemented from the imputed data file, even if the imputation flags within classes are not given, provided estimates of response rates within classes are reported. We have shown that the bias-adjusted estimator performs better than the unadjusted estimator under unweighted imputation, and is robust in the sense of validity under both the frequentist and model-based approaches.

We have obtained linearization variance estimators for the bias-adjusted estimators. For variance estimation, imputation flags should be provided in the variance estimation file.

If the imputation flags are available in the data file and imputation is deterministic, the imputed values can be replaced by those under weighted imputation. For example, in the case of unweighted ratio imputation, $y_i^* = \bar{y}_r^{\text{un}} / \bar{z}_r^{\text{un}} z_i$, one could either multiply each imputed value by $\bar{z}_r^{\text{un}} / \bar{y}_r^{\text{un}} \times \bar{y}_r / \bar{z}_r$ to reproduce the values $\bar{y}_r / \bar{z}_r z_i$ under weighted ratio imputation, provided edits are not applied after imputation. Alternatively, one could reimpute values using the sampling weights w_i. In both cases, the adjusted estimator does not present advantages over the imputed estimator based on weighted imputation other than assuring that the imputed values in the data file are not changed.

In the case of random hot-deck imputation, however, the only way to implement weighted random hot-deck imputation is to reimpute using a weighted hot-deck scheme. We believe that analysts do not like to change the imputed values on the data file produced by the edit and imputation system.

The imputed estimator (10) can use poststratification (or calibration) weights, $\tilde{w}_i(s)$, based on known population auxiliary information, instead of design weights w_i. Note that the calibration weights, $\tilde{w}_i(s)$, depend on the whole sample s unlike the design weights w_i. If the calibration weights are used for ratio imputation, then we simply replace w_i by $\tilde{w}_i(s)$ in section 3.1.1 and the resulting linearization variance estimator, v_1, uses $\hat{\xi}$ in (22) with w_i changed to $\tilde{w}_i(s)$. However, $v(.)$ in (22) now refers to the linearization variance estimator of the full sample post-stratified estimator $\sum_s \tilde{w}_i(s)\, y_i$.

Under unweighted imputation, linearization variance estimation becomes more complex because the bias-adjusted estimator based on the calibration weights will involve both design weights and calibration weights. If the design weights, w_i, are available at the imputation stage but not the calibration weights, $\tilde{w}_i(s)$, the design weights can be used for imputation and the calibration weights for estimation. The resulting imputed estimator (6) based on calibration weights remains asymptotically unbiased under uniform response (within classes), but linearization variance estimation becomes more complex because both sets of weights are involved in the imputed estimator. We propose to study poststratification and some other extensions in a separate paper, and derive corresponding linearization variance estimators.

Acknowledgement

The authors would like to thank the referee, the Associate Editor and Jae Kim of Hankuk University, Korea, for useful comments and constructive suggestion, David Haziza would also like to thank Jean-Francois Beaumont and Eric Rancourt of Statistics Canada for their encouragement and valuable discussions.

Appendix

A. Derivation of v_1

Suppose that an estimator $\hat{\theta}$ is expresses as

$$\hat{\theta} = \frac{1}{\hat{Y}_1}\left[\hat{Y}_2 + \frac{\hat{Y}_3}{\hat{Y}_4}(\hat{Y}_5 - \hat{Y}_6)\right] =: g(\hat{\mathbf{Y}}), \qquad (A.1)$$

where $\hat{Y}_j = \sum_s w_i y_{ji}$, $j = 1, ..., 6$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, ..., \hat{Y}_6)'$. Letting $\theta = g(\mathbf{Y})$, $R_{34} = Y_3 / Y_4$, $\hat{Y}_j = Y_j(1 + \delta\hat{Y}_j)$ with $\delta\hat{Y}_j = (\hat{Y}_j - Y_j)/Y_j$ and $Y_j = E_p(\hat{Y}_j)$, we have

$$\hat{\theta} - \theta = \frac{1}{Y_1(1+\delta\hat{Y}_1)}\left\{\begin{array}{l} Y_2(1+\delta\hat{Y}_2) \\ + R_{34}\dfrac{(1+\delta\hat{Y}_3)}{(1+\delta\hat{Y}_4)}[Y_5(1+\delta\hat{Y}_5)] \\ - Y_6(1+\delta\hat{Y}_6)]\end{array}\right\} - \theta$$

$$\approx \frac{1}{Y_1}\left\{\begin{array}{l}(\delta\hat{Y}_2 - \delta\hat{Y}_1)Y_2 \\ \quad + R_{34}\, Y_5(\delta\hat{Y}_3 - \delta\hat{Y}_4 + \delta\hat{Y}_5 - \delta\hat{Y}_1) \\ \quad - R_{34}\, Y_6(\delta\hat{Y}_3 - \delta\hat{Y}_4 + \delta\hat{Y}_6 - \delta\hat{Y}_1)\end{array}\right\}, \qquad (A.2)$$

neglecting higher order terms in $\delta\hat{Y}_j$'s. The expression (A.2) reduces to

$$\hat{\theta} - \theta \approx \frac{1}{Y_1}\left\{\begin{array}{l}\hat{Y}_2 + R_{34}(\hat{Y}_5 - \hat{Y}_6) \\ \quad + \dfrac{Y_5 - Y_6}{Y_4}(\hat{Y}_3 - R_{34}\hat{Y}_4) - \theta\hat{Y}_1\end{array}\right\}$$

$$= \sum_s w_i \xi_i,$$

where

$$\xi_i = \frac{1}{Y_1}(\xi_{1i} - \theta) \qquad (A.3)$$

with

$$\xi_{1i} = y_{2i} + R_{34}(y_{5i} - y_{6i}) + \frac{Y_5 - Y_6}{Y_4}(y_{3i} - R_{34}y_{4i}).$$

Hence, the variance estimator of $\hat{\theta}$ from the delta method may be expressed as $v(\xi)$. Now, replacing unknown quantities in (A.3) by their estimators, we get

$$\text{estvar}(\hat{\theta}) = v(\hat{\xi}),$$

where

$$\hat{\xi}_i = \frac{1}{\hat{Y}_1}(\hat{\xi}_{1i} - \hat{\theta})$$

with

$$\hat{\xi}_{1i} = y_{2i} + \hat{R}_{34}(y_{5i} - y_{6i}) + \frac{\hat{Y}_5 - \hat{Y}_6}{\hat{Y}_4}(y_{3i} - \hat{R}_{34}y_{4i}).$$

Note that the delta method avoids evaluation of partial derivatives of $g(\hat{\mathbf{Y}})$ with respect to its components Y_j, unlike the usual Taylor linearization method.

Letting $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_s w_i a_i y_i$, $\hat{Y}_4 = \hat{Y}_6 = \sum_s w_i a_i z_i$ and $\hat{Y}_5 = \sum_s w_i z_i$ in (A.1), we get the variance estimator (22) of $\bar{y}_{\text{IR}}$ based on weighted ratio imputation. Also, letting $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \sum_s w_i a_i y_i$, $\hat{Y}_3 = \sum_s w_i a_i (y_i / w_i)$, $\hat{Y}_4 = \sum_s w_i a_i (z_i / w_i)$, $\hat{Y}_5 = \sum_s w_i z_i$ and $\hat{Y}_6 = \sum_s w_i a_i z_i$ in (A.1), we get the variance estimator (25) of $\bar{y}_{\text{IR}}$ based on unweighted imputation. Finally, we note that the bias-adjusted estimator (16) written in the form (18) can be expressed as the sum of three components: $\bar{y}_r$, $\hat{R}_r^{\text{un}}\bar{z}$ and $-\hat{R}_r^{\text{un}}\bar{z}_r$. Each of these components is a special case of (A.1). Indeed, the component $\bar{y}_r$ is a special case of (A.1) with $\hat{Y}_1 = \sum_s w_i a_i$, $\hat{Y}_2 = \sum_s w_i a_i y_i$ with $\hat{Y}_5 = \hat{Y}_6$. The component $\hat{R}_r^{\text{un}}\bar{z}$ is a special case of (A.1) with $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_s w_i a_i (y_i / w_i)$, $\hat{Y}_4 = \hat{Y}_6 = \sum_s w_i a_i (z_i / w_i)$ and $\hat{Y}_5 = \sum_s w_i a_i z_i$. We apply the delta method to each component separately to obtain $v_1 = v(\xi)$ given by (26).

B. Derivation of v_2 for the Estimator $\bar{y}_{\text{IR}}$ Under Unweighted Imputation

Using the delta method, it can be shown that $V_r E_p(\bar{y}_{\text{IR}}^{(1)} - \bar{Y})$ under unweighted ratio imputation is given by

$$V_r\, E_p(\bar{y}_{\text{IR}} - \bar{Y}) \approx p(1-p)\frac{1}{N} \times \left[S_{e(1)}^2 + \left(\frac{Z - E_r(Z_a)}{E_r(Z_{\pi a})}\right)^2 S_{e(2)}^2 + 2\left(\frac{Z - E_r(Z_a)}{E_r(Z_{\pi a})}\right) S_{e(3)}^2 \right], \qquad (B.1)$$

where

$$S_{e(1)}^2 = \frac{1}{N}\sum_p (y_i - E_r(R_{\pi a})z_i)^2,$$

$$S_{e(2)}^2 = \frac{1}{N}\sum_p \pi_i^2 (y_i - E_r(R_{\pi a})z_i)^2,$$

$$S_{e(3)}^2 = \frac{1}{N}\sum_p \pi_i (y_i - E_r(R_{\pi a})z_i)^2,$$

with $R_{\pi a} = Y_{\pi a} / Z_{\pi a}$ and $(Y_{\pi a}, Z_{\pi a}) = \sum_P \pi_i a_i (y_i, z_i)$. The component v_2 is obtained by estimating unknown quantities in (B.1). It is given by

$$v_2 \approx \hat{p}(1-\hat{p}) \times \frac{1}{\hat{N}}\left[s_{er(1)}^2 + \left(\frac{\hat{Z} - \hat{Z}_a}{\hat{Z}_{\pi a}}\right)^2 s_{er(2)}^2 + 2\left(\frac{\hat{Z} - \hat{Z}_a}{\hat{Z}_{\pi a}}\right) s_{er(3)}^2 \right], \qquad (B.2)$$

where

$$s_{er(1)}^2 = \frac{1}{\sum_s w_i a_i}\sum_s w_i a_i (y_i - \hat{R}_r^{\text{un}} z_i)^2,$$

$$s_{er(2)}^2 = \frac{1}{\sum_s w_i a_i}\sum_s w_i^{-1} a_i (y_i - \hat{R}_r^{\text{un}} z_i)^2,$$

$$s_{er(3)}^2 = \frac{1}{\sum_s w_i a_i}\sum_s a_i (y_i - \hat{R}_r^{\text{un}} z_i)^2$$

and $\hat{Z}_{\pi a} = \sum_s a_i z_i$.

C. Derivation of v_2 for the estimator $\bar{y}_{\text{IR}}^a$

Using the delta method, it can be shown that $V_r E_p(\bar{y}_{\text{IR}}^a - \bar{Y})$ for the bias-adjusted estimator is given by

$$V_r\, E_p(\bar{y}_{\text{IR}}^a - \bar{Y}) \approx p(1-p)\frac{N}{E_r\left[\left(\sum_p a_i\right)^2\right]} \times \left\{ \begin{array}{l} S_{ay}^2 + E_r(R_{\pi a}^2) S_{az}^2 - 2E_r(R_{\pi a}) S_{ayz} \\ \quad + E_r\left[h\sum_p a_i\right]^2 S_{e(2)}^2 \\ \quad + 2E_r\left[h\sum_p a_i\right](S_{\pi ey} - E_r(R_{\pi a}) S_{\pi ez}) \end{array} \right\}, \qquad (C.1)$$

where

$$S_{ay}^2 = \frac{1}{N}\sum_p (y_i - E_r(\bar{Y}_a))^2,$$

$$S_{az}^2 = \frac{1}{N}\sum_p (z_i - E_r(\bar{Z}_a))^2,$$

$$S_{ayz} = \frac{1}{N}\sum_p (y_i - E_r(\bar{Y}_a))(z_i - E_r(\bar{Z}_a)),$$

$$S_{e(2)}^2 = \frac{1}{N}\sum_p \pi_i^2 (y_i - E_r(R_{\pi a})z_i)^2,$$

$$S_{\pi ey} = \frac{1}{N}\sum_p \pi_i (y_i - E_r(\bar{Y}_a))(y_i - E_r(R_{\pi a})z_i),$$

$$S_{\pi ez} = \frac{1}{N}\sum_p \pi_i (z_i - E_r(\bar{Z}_a))(y_i - E_r(R_{\pi a})z_i),$$

$(\bar{Y}_a,\ \bar{Z}_a) = \sum_P a_i (y_i,\ z_i) / \sum_P a_i$ and $h = (\bar{Z} - \bar{Z}_a) / Z_{\pi a}$. The component v_2 is obtained by estimating unknown quantities in (C.1). It is given by

$$v_2 \approx \hat{p}(1-\hat{p}) \frac{\hat{N}}{\sum_s w_i\ a_i} \times \left\{ \begin{array}{l} s_{yr}^2 + (\hat{R}_r^{\text{un}})^2\, s_{zr}^2 - 2\hat{R}_r^{\text{un}} s_{yzr} \\ \quad + \left(\hat{h} \sum_s w_i\ a_i \right)^2 s_{er(2)}^2 \\ \quad + 2\left(\hat{h} \sum_s w_i\ a_i \right)(s_{eyr} - \hat{R}_r^{\text{un}}\ s_{ezr}) \end{array} \right\}, \quad \text{(C.2)}$$

where

$$s_{yr}^2 = \frac{1}{\sum_s w_i\ a_i} \sum_s w_i\ a_i\ (y_i - \bar{y}_r)^2,$$

$$s_{zr}^2 = \frac{1}{\sum_s w_i\ a_i} \sum_s w_i\ a_i\ (z_i - \bar{z}_r)^2,$$

$$s_{yzr} = \frac{1}{\sum_s w_i\ a_i} \sum_s w_i\ a_i\ (y_i - \bar{y}_r)(z_i - \bar{z}_r),$$

$$s_{er(2)}^2 = \frac{1}{\sum_s w_i\ a_i} \sum_s w_i^{-1}\ a_i\ (y_i - \hat{R}_r^{\text{un}} z_i)^2,$$

$$s_{eyr} = \frac{1}{\sum_s w_i\ a_i} \sum_s a_i\ (y_i - \bar{y}_r)(y_i - \hat{R}_r^{\text{un}} z_i),$$

$$s_{ezr} = \frac{1}{\sum_s w_i\ a_i} \sum_s a_i\ (z_i - \bar{z}_r)(y_i - \hat{R}_r^{\text{un}} z_i),$$

and $\hat{h} = (\bar{z} - \bar{z}_r) / \sum_s a_i\ z_i$.

D. Nonegativity of $V_*(\bar{y}_{\text{IH}}^a - \bar{Y})$

We show that the variance formula in (34) is always nonnegative. First, note that this expression can be expressed as

$$V_*(\bar{y}_{\text{IH}}^a - \bar{Y}) = \frac{n^2 \sum_{s_m} w_i^2 - (r+n)\left(\sum_{s_m} w_i\right)^2}{n^2 \left(\sum_{s_r} w_i\right)^2} \geq 0$$

$$\Leftrightarrow n^2 \sum_{s_m} w_i^2 - (r+n)\left(\sum_{s_m} w_i\right)^2 \geq 0$$

$$\Leftrightarrow n^2 \sum_{s_m} w_i^2 - m(r+n) \frac{\left(\sum_{s_m} w_i\right)^2}{m} \geq 0.$$

On one hand, $n^2 \geq m(r+n) \Leftrightarrow n \geq m$ which is always true. On the other hand, using Cauchy-Schwarz inequality, it is easily seen that $\sum_{s_m} w_i^2 \geq (\sum_{s_m} w_i)^2 / m$. The result follows.

References

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. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā*, Series A, 28, Part 1, 47-59.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

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.

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

Skinner, C.J., and Rao, J.N.K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*, 102, 149-167.