Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Variance estimation for a ratio in the presence of imputed data

by David Haziza

December, 2007

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2007
Vol. 33, No. 2, pp. 159-166
Statistics Canada, Catalogue no. 12-001-X

Variance estimation for a ratio in the presence of imputed data

David Haziza [1]

Abstract

In this paper, we study the problem of variance estimation for a ratio of two totals when marginal random hot deck imputation has been used to fill in missing data. We consider two approaches to inference. In the first approach, the validity of an imputation model is required. In the second approach, the validity of an imputation model is not required but response probabilities need to be estimated, in which case the validity of a nonresponse model is required. We derive variance estimators under two distinct frameworks: the customary two-phase framework and the reverse framework.

Key Words: Imputation model; Nonresponse model; Marginal random hot deck imputation; Reverse framework; Two-phase framework; Variance estimation.

1. Introduction

Variance estimation in the presence of imputed data for simple univariate parameter such as population totals and population means has been widely treated in recent years; see for example, Särndal (1992), Deville and Särndal (1994), Rao and Shao (1992), Rao (1996) and Shao and Steel (1999). In practice, it is often of interest to estimate the ratio of two population totals, $R = Y / X$, where $(Y, X) = \sum_{i \in U} (y_i, x_i)$, y and x denote two variables of interest potentially missing and U denotes the finite population (of size N) under study. Although variance estimation for a ratio in the presence of imputed data is a problem frequently encountered in practice (especially in business surveys), it has not been, to our knowledge, fully studied in the literature. In this paper, we consider the case Marginal Random Hot Deck (MRHD) imputation performed within the same set of imputation classes for both variables y and x. In other words, to compensate for nonresponse, random hot deck imputation is performed separately for both variables within the same set of imputation classes. This situation occurs frequently in practice. For simplicity, we consider the case of a single imputation class. Extensions to multiple imputation classes are relatively straightforward for most derivations presented in this paper.

In this paper, we derive variance estimators that take sampling, nonresponse and imputation into account. Two distinct frameworks for variance estimation have been studied in the literature: (i) the customary two-phase framework (*e.g.*, Särndal (1992)) and (ii) the reverse framework (*e.g.*, Shao and Steel (1999)). In the two-phase framework, nonresponse is viewed as a second phase of selection. That is, a random sample is selected from the population according to a given sampling design. Then, given the selected sample, the set of respondents is generated according to the nonresponse mechanism. In the reverse framework, the order of sampling and response is reversed. That is, the population is first randomly divided into a population of respondents and a population of nonrespondents according to the nonresponse mechanism. Then, a random sample is selected from the population (containing respondents and nonrespondents) according to the sampling design. As we will see in section 4, the reverse framework facilitates the derivation of variance estimators but unlike the two-phase framework, it requires the additional assumption that the nonresponse mechanism does not depend on which sample is selected. This assumption is satisfied in many situations encountered in practice. For each framework, inference can be based either on an Imputation Model (IM) or a Nonresponse Model (NM). The IM approach requires the validity of an imputation model, whereas the NM approach requires the validity of a nonresponse model.

In section 2, we introduce notation, assumptions and the imputed estimator of a ratio under weighted MRHD imputation. The IM and NM approaches are then presented in sections 2.1 and 2.2. In section 2.3, the bias of the imputed estimator is discussed. In section 3, variance estimators are derived under the two-phase framework and the IM approach using the method proposed by Särndal (1992). We show that, under MRHD imputation, the naïve variance estimator (that treats the imputed values as observed values) generally overestimates the sampling variance when y and x are positively correlated. In section 4, we derive variance estimators under the reverse framework and both the IM and the NM approaches using the method proposed by Shao and Steel (1999). Finally, we conclude in section 5.

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, (Québec), H3C 3J7, Canada. E-mail: David.Haziza@umontreal.ca.

2. Notation and assumptions

Our goal is to estimate R. We select a random sample, s, of size n, according to a given sampling design $p(s)$. A complete-data estimator is given by

$$\hat{R} = \frac{\hat{Y}_{\text{HT}}}{\hat{X}_{\text{HT}}}, \tag{2.1}$$

where $(\hat{Y}_{\text{HT}}, \hat{X}_{\text{HT}}) = \sum_{i\in s} w_i(y_i, x_i)$ denote the Horvitz-Thompson estimators for Y and X, respectively and $w_i = 1/\pi_i$ denotes the sampling weight of unit i, where π_i is its probability of inclusion in the sample. The estimator $\hat{R}$ in (2.1) is asymptotically p-unbiased for R, *i.e.*, $E_p(\hat{R}) \approx R$, where the subscript p denotes the expectation and variance with respect to the sampling design $p(s)$. Since $\hat{R}$ is a nonlinear function of estimated totals, its exact design variance, $V_p(\hat{R})$, cannot be easily obtained. To overcome this problem, Taylor linearization is often applied in order to approximate the exact variance. An asymptotically p-unbiased estimator of the approximate variance of $\hat{R}$ is given by

$$\hat{V}_{\text{SAM}} = \sum_{i\in s}\sum_{j\in s} \Delta_{ij}\, e_i\, e_j, \tag{2.2}$$

where $e_i = 1/\hat{X}_{\text{HT}}(y_i - \hat{R}x_i)$, $\Delta_{ij} = (\pi_{ij} - \pi_i\,\pi_j)/\pi_{ij}\,\pi_i\,\pi_j$ and π_{ij} is the joint selection probability of units i and j. Note that $\pi_{ii} = \pi_i$. In the case of simple random sampling without replacement, the variance estimator (2.2) reduces to

$$\hat{V}_{\text{SAM}} = \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}^2}\left[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}\right], \tag{2.3}$$

where

$$s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2,\ s_x^2 = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2$$

and

$$s_{xy} = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})$$

with

$$(\bar{y},\ \bar{x}) = \frac{1}{n}\sum_{i\in s}(y_i,\ x_i).$$

We now turn to the case for which both variables x and y may be missing. Let a_i be the response indicator of unit i such that $a_i = 1$ if unit i responds to variable y and $a_i = 0$, otherwise. Similarly, let b_i be the response indicator of unit i such that $b_i = 1$ if unit i responds to variable x and $b_i = 0$, otherwise. Let $s_r^{(y)}$ be the set of respondents to variable y of size r_y and $s_r^{(x)}$ be the set of respondents to variable x of size r_x. Also, let r_{xy} be the number of respondents to both variables y an x. Finally, let y_i^* and x_i^* denote the imputed values to replace the missing values y_i and x_i, respectively. An imputed estimator of R is given by

$$\hat{R}_I = \frac{\sum_{i\in s} w_i\,\tilde{y}_i}{\sum_{i\in s} w_i\,\tilde{x}_i}, \tag{2.4}$$

where $\tilde{y}_i = a_i\,y_i + (1 - a_i)\,y_i^*$ and $\tilde{x}_i = b_i\,x_i + (1 - b_i)\,x_i^*$. Under weighted MRHD imputation, to compensate for the missing value y_i^*, a donor j is selected at random with replacement from $s_r^{(y)}$ so that

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{l\in s} w_l\,a_l}.$$

Similarly, to compensate for the missing value x_i^*, a donor k is selected at random with replacement from $s_r^{(x)}$ so that

$$P(x_i^* = x_k) = \frac{w_k}{\sum_{l\in s} w_l\,b_l}.$$

Note that, when both y_i and x_i are missing, j is generally not equal to k under weighted MRHD imputation.

Random hot-deck imputation within classes is widely used in practice because (i) it preserves the variability of the original data; and (ii) it leads to plausible values. The latter is particularly important in the case of categorical variables of interest. However, random hot-deck imputation within classes suffers from an additional component of variance due to the use of a random imputation mechanism. The main reason weighted MRHD imputation is used is that it leads to asymptotically unbiased estimator under the nonresponse model approach (see section 2.1) unlike unweighted MRHD imputation.

Let $E_I(.\mid s,\ s_r^{(y)},\ s_r^{(x)})$, $V_I(.\mid s,\ s_r^{(y)},\ s_r^{(x)})$ and $\text{Cov}_I(.,.\mid s,\ s_r^{(y)},\ s_r^{(x)})$ denote the conditional expectation, the conditional variance and the conditional covariance operators with respect to the random imputation mechanism (here, weighted MRHD imputation). Using a first-order Taylor expansion, it can be shown that

$$E_I(\hat{R}_I \mid s,\ s_r^{(y)},\ s_r^{(x)}) \approx \frac{\bar{y}_r}{\bar{x}_r} \equiv \hat{R}_r, \tag{2.5}$$

where

$$\bar{y}_r = \frac{\sum_{i\in s} w_i\,a_i\,y_i}{\sum_{i\in s} w_i\,a_i}$$

and

$$\bar{x}_r = \frac{\sum_{i\in s} w_i\,b_i\,y_i}{\sum_{i\in s} w_i\,b_i}$$

denote the weighted means of the respondents to variables y and x, respectively. The approximation in (2.5) will be valid if the sample size within classes is sufficiently large, which we assume to be the case. Now, let

$$s_{yr}^2 = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2$$

and

$$s_{xr}^2 = \frac{1}{\sum_{i \in s} w_i b_i} \sum_{i \in s} w_i b_i (x_i - \bar{x}_r)^2$$

denote the variability of the y-values and the x-values in the set of respondents $s_r^{(y)}$ and $s_r^{(x)}$, respectively. Noting that, under weighted MRHD imputation,

$$V_I(y_i^*) = s_{yr}^2,\ V_I(x_i^*) = s_{xr}^2$$

and

$$\mathrm{Cov}_I(y_i^*, x_i^*) = 0,$$

we can approximate $V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ by

$$V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{1}{\bar{x}_r^2}\left[\sum_{i \in s} w_i^2 (1 - a_i) s_{yr}^2 + \hat{R}_r^2 \sum_{i \in s} w_i^2 (1 - b_i) s_{xr}^2\right]. \quad (2.6)$$

Expressions (2.5) and (2.6) will be useful in subsequent sections when discussing the bias and the variance of the imputed estimator $\hat{R}_I$. As we will see in sections 3 and 4, the conditional variance (2.6) is a measure of the variability due to the imputation mechanism.

Next, we describe two approaches to inference that will be used to obtain variance estimators in sections 3 and 4: the Nonresponse Model (NM) approach and the Imputation Model (IM) approach.

2.1 The nonresponse model approach

In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the nonresponse model. The nonresponse model is a set of assumptions about the unknown distribution of the response indicators $\mathbf{R}_s = \{(a_i, b_i);\ i \in s\}$. This unknown distribution is often called the nonresponse mechanism. Let $p_{yi} = P(a_i = 1 \mid s, \mathbf{Z}_s)$ be the response probability of unit i to variable y, where $\mathbf{Z}_s = \{\mathbf{z}_i;\ i \in s\}$ and $\mathbf{z}_i$ is a vector of auxiliary variables available for all sample units used to form the imputation classes. Similarly, let $p_{xi} = P(b_i = 1 \mid s, \mathbf{Z}_s)$ be the response probability of unit i to variable x. We assume that units respond independently; *i.e.*, $p_{yij} = P(a_i = 1, a_j = 1 \mid s, \mathbf{Z}_s) = p_{yi}\, p_{yj}$ for $i \neq j$ and $p_{xij} = P(b_i = 1, b_j = 1 \mid s, \mathbf{Z}_s) = p_{xi}\, p_{xj}$ for $i \neq j$. However, we do not assume that, for a given unit i, response to variable y is independent of response to variable x. In other words, if we let $p_{xyi} = P(a_i = 1, b_i = 1 \mid s, \mathbf{Z}_s)$, then we have $p_{xyi} \neq p_{xi}\, p_{yi}$, in general. Within an imputation class, we assume a uniform response mechanism such that $p_{yi} = p_y$, $p_{xi} = p_x$ and $p_{xyi} = p_{xy}$.

We also assume that, after conditioning on s and $\mathbf{Z}_s$, the nonresponse mechanism is independent of all other variables involved in the imputed estimator (2.4) as well as the joint selection probabilities. In other words, the distribution of $\mathbf{R}_s$ does not depend on $\mathbf{Y}_s = \{y_i;\ i \in s\}$, $\mathbf{W}_s = \{w_i;\ i \in s\}$ and $\mathbf{\Pi}_s = \{\pi_{ij};\ i \in s,\ j \in s\}$, after conditioning on s and $\mathbf{Z}_s$. As a result, except for the response indicators a_i and b_i, we assume that all the variables involved in the imputed estimator (2.4) as well as the joint selection probabilities are treated as fixed when taking expectations and variances with respect to the nonresponse model. From this point on, we use the subscript q to denote the expectation and variance with respect to the nonresponse mechanism.

2.2 The imputation model approach

In the IM approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design and the nonresponse model. The imputation model is a set of assumptions about the unknown distribution of $(\mathbf{Y}_U, \mathbf{X}_U) = \{(y_i, x_i);\ i \in U\}$. Within an imputation class, the imputation model, m, in the case of MRHD imputation, is given by

$$m: \begin{cases} y_i = \mu_y + \varepsilon_i \\ x_i = \mu_x + \eta_i \end{cases} \quad (2.7)$$

where ε_i is a random error term such that $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$, $V_m(\varepsilon_i) = \sigma_\varepsilon^2$ and η_i is a random error term such that $E_m(\eta_i) = 0$, $E_m(\eta_i \eta_j) = 0$, for $i \neq j$, $V_m(\varepsilon_i) = \sigma_\eta^2$. Furthermore, we assume that $E_m(\varepsilon_i \eta_i) = \sigma_{\varepsilon\eta}$. Here, $E_m(.)$, $V_m(.)$ and $\mathrm{Cov}_m(.)$ denote respectively the expectation, the variance and the covariance operators with respect to model m. It is implicit in the notation that expectations or variances with respect to model m are conditional on $\mathbf{Z}_U = \{\mathbf{z}_i;\ i \in U\}$. In this approach, we assume that the distribution of the model errors $(\mathbf{\varepsilon}_U, \mathbf{\eta}_U) = \{(\varepsilon_i, \eta_i);\ i \in U\}$ does not depend on s, $s_r^{(y)}$, $s_r^{(x)}$, $\mathbf{W}_U = \{w_i;\ i \in U\}$ and $\mathbf{\Pi}_U = \{\pi_{ij};\ i \in U,\ j \in U\}$, after conditioning on $\mathbf{Z}_U$. As a result, except for the variables of interest y and x, all variables involved in the imputed estimator (2.4) are treated as fixed when taking expectations and variances with respect to the imputation model.

2.3 Bias of the imputed estimator

To study the bias of the imputed estimator (2.4), we use the standard decomposition of the total error of $\hat{R}_I$:

$$\hat{R}_I - R = \left[\hat{R} - R\right] + \left[E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}\right] + \left[\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})\right]. \quad (2.8)$$

The first term $\hat{R} - R$ on the right-hand side of (2.8) is called the sampling error, the second term $E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}$ is called the nonresponse error, whereas the third term $\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ is called the imputation error.

Using a first-order Taylor expansion, it can easily be shown that, under the NM approach, the imputed estimator (2.4) is asymptotically pqI-unbiased; that is, $E_{pqI}(\hat{R}_I - R) \approx 0$. Also, under the IM approach and model (2.7), it can be shown that the imputed estimator (2.4) is asymptotically $mpqI$-unbiased under the IM approach; that is, $E_{mpqI}(\hat{R}_I - R) \approx 0$. Thus, the imputed estimator is robust in the sense that it is valid under either the NM approach or the IM approach. Note that for the asymptotic bias to be equal to 0 under both approaches, we require that the sample size within each imputation class is sufficiently large. From this point on, we thus assume that the bias of $\hat{R}_I$ is negligible.

3. Variance estimation: The two-phase framework

In this section, we derive variance estimators under the two-phase framework and the IM approach according to the method proposed by Särndal (1992) and Deville and Särndal (1994). Using the decomposition (2.8), the total variance of $\hat{R}_I$ can be approximated by

$$V_{mpqI}(\hat{R}_I - R) \approx E_{mpqI}(\hat{R}_I - R)^2 = V_{\text{SAM}} + V_{\text{NR}} + \tilde{V}_I + 2V_{\text{MIX}}, \quad (3.1)$$

where $V_{\text{SAM}} = E_m V_p(\hat{R}) = E_m(V_{\text{SAM}})$ is the sampling variance of the complete-data estimator $\hat{R}$, $V_{\text{NR}} = E_{pq} V_m(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ is the nonresponse variance of the imputed estimator $\hat{R}_I$, $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \mid s, s_r^{(y)}, s_r^{(x)})$ is the imputation variance of the imputed estimator $\hat{R}_I$, and $V_{\text{MIX}} = E_{pqm}[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}]$ is a mixed component. Note that the expression (3.1) contains only one cross product term, $2V_{\text{MIX}}$, because the other cross product terms are all asymptotically equal to 0.

3.1 Estimation of the sampling variance V_{SAM}

Let $\hat{V}_{\text{ORD}}$ be the naive variance estimator of $\hat{R}_I$; *i.e.*, the variance estimator obtained by treating the imputed values as observed values. The variance estimator $\hat{V}_{\text{ORD}}$ is thus obtained by replacing e_i by $\tilde{e}_i = 1/\hat{X}_I(\tilde{y}_i - \hat{R}_I \tilde{x}_i)$ in (2.2) which leads to

$$\hat{V}_{\text{ORD}} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \tilde{e}_i \tilde{e}_j. \quad (3.2)$$

As we show now in the case of simple random sampling without replacement, $\hat{V}_{\text{ORD}}$ overestimates V_{SAM} under MRHD imputation whenever $\sigma_{\varepsilon\eta} > 0$ (as it is usually the case in practice). After some algebra, we obtain

$$E_{mI}(\hat{V}_{\text{ORD}} - \hat{V}_{\text{SAM}} \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{2}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\right)\left(1 - \frac{n}{N}\right)\left(1 - \frac{r_{xy}}{n}\right)\frac{\sigma_{\varepsilon\eta}}{n}. \quad (3.3)$$

Expression (3.3) shows that $\hat{V}_{\text{ORD}}$ is $mpqI$-biased for $\hat{V}_{\text{SAM}}$ unless $\sigma_{\varepsilon\eta} = 0$, $r_{xy} = n$ (which is the case of complete data) or $n = N$ (which is the census case). The fact that $\hat{V}_{\text{ORD}}$ is not a valid estimator of V_{SAM} can be easily explained by noting that although MRHD imputation preserves the variability, s_x^2 and s_y^2, corresponding to variables x and y, it does not preserve the covariance, s_{xy}, in (2.3). Indeed, imputation tends to underestimate relationships between variables that are positively correlated. As a result, $\hat{V}_{\text{ORD}}$ overestimates V_{SAM} because of the presence of the minus sign in front of s_{xy} in (2.3). To overcome this difficulty, Särndal (1992) proposed to estimate $V_{\text{DIF}} = E_{mI}(\hat{V}_{\text{SAM}} - \hat{V}_{\text{ORD}} \mid s, s_r^{(y)}, s_r^{(x)})$ by a mI-unbiased estimator $\hat{V}_{\text{DIF}}$; *i.e.*, $E_{mI}(\hat{V}_{\text{DIF}} \mid s, s_r^{(y)}, s_r^{(x)}) = V_{\text{DIF}}$. However, the derivation of this component for an arbitrary design involves very tedious algebra in the case of a ratio. Therefore, we propose an alternative that does not require any derivation but involves the construction of a new set of imputed values. It can be described as follows: whenever $a_i = 0$ and/or $b_i = 0$, select a donor j at random with replacement from the set of respondents to both variables y and x (*i.e.*, the set of sampled units for which $a_i = 1$ and $b_i = 1$) with probability $w_j / \sum_{l \in s} w_l a_l b_l$ and impute the vector (x_j, y_j). In other words, whenever one variable is missing, the observed value is discarded and set to missing; the missing values are then replaced by the values of a donor selected at random among the set of respondents to both variables x and y (often called the set of common donors). Similarly, when both variables are missing, the vector (x_j, y_j) of a donor j is imputed. Then, use the standard variance estimator (2.2) valid in the complete response case using these imputed values. Let

$\hat{V}^*_{\text{ORD}}$ denote the resulting variance estimator. Note that this new set of imputed values is used only to obtain a valid estimator of the sampling variance and is not used to estimate the parameter of interest R. It can be shown that $\hat{V}^*_{\text{ORD}}$ is an asymptotically $mpqI$-unbiased estimator of V_{SAM}. In practice, one could, for example, create a variance estimation file containing the new set of imputed values and use standard variance estimation systems (used in the complete data case) to obtained an estimate of the sampling variance.

3.2 Estimation of the nonresponse variance V_{NR}

An estimator $\hat{V}_{\text{NR}}$ of $V_{\text{NR}} = E_{pq} V_m (E_I (\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ can simply be obtained by estimating $V_m (E_I (\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$. Using a first-order Taylor expansion, we obtain

$$
\begin{aligned}
V_{\text{NR}} \approx\ & \frac{1}{\mu_x^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a}\right]\sigma_\varepsilon^2\right. \\
& + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\left(\frac{\mu_y}{\mu_x}\right)^2\sigma_\eta^2 \\
& \left. -2\left[\frac{\sum_{i\in s} w_i^2 a_i b_i}{\hat{N}_a\hat{N}_b} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\left(\frac{\mu_y}{\mu_x}\right)\sigma_{\varepsilon\eta}\right\}, \quad (3.4)
\end{aligned}
$$

where $(\hat{N}, \hat{N}_a, \hat{N}_b) = \sum_{i\in s} w_i (1, a_i, b_i)$. Now, let $s_{Ix}^2 = 1/\hat{N}\sum_{i\in s} w_i (\tilde{x}_i - \bar{x}_I)^2$ and $s_{Iy}^2 = 1/\hat{N}\sum_{i\in s} w_i (\tilde{y}_i - \bar{y}_I)^2$ with $(\bar{x}_I, \bar{y}_I) = 1/\hat{N}\sum_{i\in s} w_i (\tilde{x}_i, \tilde{y}_i)$. Note that s_{Ix}^2 and s_{Iy}^2 denote respectively the sample variability of the x-values and the y-values after imputation. It can be shown that s_{Ix}^2 and s_{Iy}^2 are respectively asymptotically mI-unbiased for the model variances σ_η^2 and σ_ε^2. Also, let $s_{xyr} = 1/\hat{N}_{ab}$ $\sum_{i\in s} w_i a_i b_i (x_i - \bar{x}_{rr})(y_i - \bar{y}_{rr})$, where $\hat{N}_{ab} = \sum_{i\in s} w_i a_i b_i$ and $(\bar{x}_{rr}, \bar{y}_{rr}) = \hat{N}_{ab}^{-1}\sum_{i\in s} w_i a_i b_i (x_i, y_i)$. Note that s_{xyr} is m-unbiased for the model covariance $\sigma_{\varepsilon\eta}$. It follows that $\hat{V}_{\text{NR}}$ is obtained by estimating the unknown quantities in (3.4), which leads to

$$
\begin{aligned}
\hat{V}_{\text{NR}} =\ & \frac{1}{\bar{x}_I^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a}\right]s_{Iy}^2\right. \\
& + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\hat{R}_I^2 s_{Ix}^2 \\
& \left. -2\left[\frac{\sum_{i\in s} w_i^2 a_i b_i}{\hat{N}_a\hat{N}_b} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\hat{R}_I s_{xyr}\right\}. \quad (3.5)
\end{aligned}
$$

The estimator (3.5) is asymptotically $mpqI$-unbiased for V_{NR}. In the special case of simple random sampling without replacement, expression (3.5) reduces to

$$
\begin{aligned}
\hat{V}_{\text{NR}} =\ & \frac{1}{\bar{x}_I^2}\left[\left(1 - \frac{r_y}{n}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(1 - \frac{r_x}{n}\right)\frac{s_{Ix}^2}{r_x}\right. \\
& \left. - 2\hat{R}_I\left(1 - \frac{r_x r_y}{n r_{xy}}\right)\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right].
\end{aligned}
$$

3.3 Estimation of the imputation variance $\tilde{V}_I$

An estimator $\hat{V}_I$ of $\tilde{V}_I = E_{mpq} V_I (\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ can simply be obtained by estimating $V_I (\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ given by (2.6). An asymptotically I-unbiased of $V_I (\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ is then given by

$$
\hat{V}_I = \frac{1}{\bar{x}_I^2}\left[\sum_{i\in s} w_i^2 (1 - a_i) s_{Iy}^2 + \hat{R}_I^2 \sum_{i\in s} w_i^2 (1 - b_i) s_{Ix}^2\right]. \quad (3.6)
$$

It follows that $\hat{V}_I$ in (3.6) is asymptotically $mpqI$-unbiased for $\tilde{V}_I$. In the special case of simple random sampling without replacement, expression (3.6) reduces to

$$
\hat{V}_I = \frac{N^2}{\bar{x}_I^2}\left[\left(1 - \frac{r_y}{n}\right)\frac{s_{Iy}^2}{n} + \hat{R}_I^2\left(1 - \frac{r_x}{n}\right)\frac{s_{Ix}^2}{n}\right].
$$

3.4 Estimation of the mixed component V_{MIX}

Finally, we obtain an estimator $\hat{V}_{\text{MIX}}$ of V_{MIX} by estimating

$$E_m[(E_I(\hat{R}_I \mid s,\ s_r^{(y)},\ s_r^{(x)}) - \hat{R})\ (\hat{R} - R) \mid s,\ s_r^{(y)},\ s_r^{(x)}].$$

Using a first-order Taylor expansion, we obtain

$$E_m[(E_I(\hat{R}_I \mid s,\ s_r^{(y)},\ s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s,\ s_r^{(y)},\ s_r^{(x)}]$$

$$\approx \frac{1}{\mu_x^2}\left\{\left[\frac{\sum_{i\in s} w_i^2\, a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\sigma_\varepsilon^2\right.$$

$$+\left[\frac{\sum_{i\in s} w_i^2\, b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\left(\frac{\mu_y}{\mu_x}\right)^2\sigma_\eta^2$$

$$\left.-2\left[\frac{\sum_{i\in s} w_i^2\, a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i\in s} w_i^2\, b_i}{\hat{N}\hat{N}_b} - 2\frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\left(\frac{\mu_y}{\mu_x}\right)\sigma_{\varepsilon\eta}\right\}. \quad (3.7)$$

An estimator of (3.7) is thus given by

$$\hat{V}_{\text{MIX}} =$$

$$\frac{1}{\bar{x}_I^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]s_{Iy}^2 + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\hat{R}_r^2 s_{Ix}^2\right.$$

$$\left.-2\left[\frac{\sum_{i\in s} w_i^2\, a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i\in s} w_i^2\, b_i}{\hat{N}\hat{N}_b} - 2\frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\hat{R}_I\, s_{xyr}\right\}. \quad (3.8)$$

The estimator (3.8) is asymptotically *mpqI*-unbiased for V_{MIX}. In the case of simple random sampling without replacement, the component $\hat{V}_{\text{MIX}}$ is equal to zero. More generally, the component $\hat{V}_{\text{MIX}}$ is equal to zero for any unistage self-weighting design (*i.e.*, a sampling design for which the sampling weight are all equal). For unequal probability designs, it is important to include the component $\hat{V}_{\text{MIX}}$ because its contribution (positive or negative) to the overall variance could be substantial (Brick, Kalton and Kim (2004)).

Finally, an asymptotically *mpqI*-unbiased estimator of the total variance $V_{\text{TOT}} = V_{mpqI}(\hat{R}_I - R)$ is thus given by

$$\hat{V}_{\text{TOT}}^{(\text{TP})} = \hat{V}_{\text{ORD}}^{*} + \hat{V}_{\text{NR}} + \hat{V}_I + 2\hat{V}_{\text{MIX}}.$$

4. Variance estimation: The reverse framework

In this section, we derive variance estimators under the reverse framework and both the NM and the IM approaches according to the method proposed by Shao and Steel (1999). Recall that, under this framework, we require the additional assumption that the response probabilities do not depend on the sample s. Under the NM approach, the total variance of $\hat{R}_I$ can be approximated by

$$V(\hat{R}_I - R) \approx E_q V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$$

$$+ E_{pq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}), \quad (4.1)$$

where $\mathbf{a} = (a_1, \ldots, a_N)'$ and $\mathbf{b} = (b_1, \ldots, b_N)'$ denote the vectors of response indicators to variables y and x, respectively.

Under the IM approach, the total variance of $\hat{R}_I$ can be approximated by

$$V(\hat{R}_I - R) \approx E_{mq} V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$$

$$+ E_{mpq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$$

$$+ E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}). \quad (4.2)$$

Under both the NM and the IM approaches, an estimator of the first term on the right hand side of (4.1) and (4.2) can be obtained by finding an asymptotically *pI*-unbiased estimator of $V_p E_I(\hat{R}_I \mid \mathbf{a}, \mathbf{b})$. Also, the second term on the right hand side of (4.1) and (4.2) can be estimated by $\hat{V}_I$ given by (3.6). Under the NM approach, an estimator the last term on the right hand side of (4.1) can be obtained by estimating $V_q E_{pI}(\hat{R}_I \mid \mathbf{a}, \mathbf{b})$, whereas an estimator of the last term on the right hand side of (4.2) can be obtained by estimating $V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ under the IM approach. As a result, the estimators of the first two terms in (4.1) and (4.2) are identical and thus are valid regardless of the approach (NM or IM) used for inference. Only the third term on the right hand side of (4.1) and (4.2) will depend on the approach used. In the case of the IM approach, specification and validation of the imputation model is crucial to achieve asymptotic unbiasedness of the third component, whereas in the case of the NM approach, the asymptotic unbiasedness of the third component relies of the correct specification of the nonresponse model.

4.1 Estimation of $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$

Using a first-order Taylor expansion and expression (2.5), an estimator of $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$, denoted by $\hat{V}_1$, is given by

$$\hat{V}_1 = \sum_{i\in s}\sum_{j\in s} \Delta_{ij}\, \xi_i\, \xi_j, \quad (4.3)$$

where

$$\xi_i = \frac{1}{\bar{x}_r}\left[\frac{1}{\hat{N}_a} a_i (y_i - \bar{y}_r) - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)\frac{1}{\hat{N}_b} b_i\ (x_i - \bar{x}_r)\right].$$

In other words, the estimator $\hat{V}_1$ is obtained from the complete data variance estimator (2.2) by replacing e_i by ξ_i. In the case of simple random sampling without replacement, the estimator (4.3) reduces to

$$\hat{V}_1 = \left(1 - \frac{n}{N}\right)\frac{1}{\bar{x}_r^2}\left[\frac{s_{yr}^2}{r_y} + \hat{R}_r^2\frac{s_{xr}^2}{r_x} - 2\hat{R}_r\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \quad (4.4)$$

4.2 Estimation of $V_q\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ under the NM approach

First, note that

$$E_{pI}(\hat{R}_I) \approx \frac{Y_a N_b}{N_a X_b},$$

where $(Y_a, N_a) = \sum_{i\in U} a_i(y_i, 1)$ and $(X_b, N_b) = \sum_{i\in U} b_i(x_i, 1)$.

Using a first-order Taylor expansion, it can be shown that $V_q\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ can be approximated by

$$V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) \approx \frac{1}{N\bar{X}^2}\left[\left(\frac{1-p_y}{p_y}\right)S_y^2 + R^2\left(\frac{1-p_x}{p_x}\right)S_x^2 - 2R\left(\frac{p_{xy} - p_x p_y}{p_x p_y}\right)S_{xy}\right], \quad (4.5)$$

where

$$S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{i\in U}(x_i - \bar{X})^2$$

and

$$S_{xy} = \frac{1}{N-1}\sum_{i\in U}(x_i - \bar{X})(y_i - \bar{Y})$$

with

$$(\bar{Y},\ \bar{X}) = \frac{1}{N}\sum_{i\in U}(y_i,\ x_i).$$

An estimator of $V_q\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ is obtained by estimating unknown quantities in (4.5), which leads to

$$\hat{V}_2^{(\mathrm{NM})} = \frac{1}{\hat{N}\bar{x}_I^2}\left[\left(\frac{1-\hat{p}_y}{\hat{p}_y}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1-\hat{p}_x}{\hat{p}_x}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{p}_{xy} - \hat{p}_x\hat{p}_y}{\hat{p}_x\hat{p}_y}\right)s_{xyr}\right]$$

$$= \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b\hat{N}_b} - \frac{1}{N}\right)s_{xyr}\right], \quad (4.6)$$

where $\hat{p}_y = \frac{\hat{N}_a}{\hat{N}}$, $\hat{p}_x = \frac{\hat{N}_b}{\hat{N}}$ and $\hat{p}_{xy} = \frac{\hat{N}_{ab}}{\hat{N}}$.

The estimator (4.6) is asymptotically pqI-unbiased for the approximate variance (4.5), noting that s_{Iy}^2, s_{Ix}^2 and s_{xyr} are asymptotically pqI-unbiased for S_y^2, S_x^2 and S_{xy}, respectively. In the case of simple random sampling without replacement, the estimator (4.6) reduces to

$$\hat{V}_2^{(\mathrm{NM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{n}{N} - \frac{r_y}{N}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(\frac{n}{N} - \frac{r_x}{N}\right)\frac{s_{Ix}^2}{r_x} - 2\hat{R}_I\left(\frac{n}{N} - \frac{r_x r_y}{N r_{xy}}\right)\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \quad (4.7)$$

4.3 Estimation of $E_q V_m\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ under the IM approach

Using a first-order Taylor expansion, it can be shown that $E_q V_m\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ can be approximated by

$$E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) \approx \frac{1}{\mu_x^2}\left[\left(\frac{1}{E_q(N_a)} - \frac{1}{N}\right)\sigma_\varepsilon^2 + \left(\frac{\mu_y}{\mu_x}\right)^2\left(\frac{1}{E_q(N_b)} - \frac{1}{N}\right)\sigma_\eta^2 - 2\left(\frac{\mu_y}{\mu_x}\right)\left(E_q\left(\frac{N_{ab}}{N_b N_b}\right) - \frac{1}{N}\right)\sigma_{\varepsilon\eta}\right], \quad (4.8)$$

where $N_{ab} = \sum_{i\in U} a_i b_i$. An estimator of $E_q V_m\, E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ is obtained by estimating unknown quantities in (4.8), which leads to

$$\hat{V}_2^{(\mathrm{IM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b\hat{N}_b} - \frac{1}{N}\right)s_{xyr}\right]. \quad (4.9)$$

The estimator (4.9) is asymptotically $mpqI$-unbiased for the approximate variance (4.8). It is interesting to note that, under weighted MRHD imputation, the estimator $\hat{V}_2^{(\mathrm{NM})}$ in (4.6) obtained under the NM approach is identical to $\hat{V}_2^{(\mathrm{IM})}$ in (4.9) obtained under the IM approach. However, this may not be the case with a different imputation method. Also, the component $\hat{V}_2$ is negligible with respect to $\hat{V}_1$ when the sampling fraction n/N is negligible, where $\hat{V}_2$ stands for $\hat{V}_2^{(\mathrm{NM})}$ or $\hat{V}_2^{(\mathrm{IM})}$. In this case, the component $\hat{V}_2$ may be omitted from the calculations.

Finally, an estimator of the total variance under the reverse framework is given by

$$\hat{V}_{\mathrm{TOT}}^{(\mathrm{RE})} = \hat{V}_1 + \hat{V}_I + \hat{V}_2.$$

Under the reverse framework, both the NM approach and the IM approach lead to the same estimator of the total variance. Thus, the variance estimator $\hat{V}_{\text{TOT}}^{(\text{RE})}$ is robust in the sense that it is valid under either the NM approach or the IM approach.

5. Summary and conclusions

In this paper, we have derived variance estimators for the imputed estimator of a ratio under two different frameworks. The reverse framework facilitates the derivation of the variance expressions (in comparison with the customary two-phase framework), especially if the sampling fraction is small, in which case we can omit the component $\hat{V}_2$. However, unlike the two-phase framework, it requires an additional assumption that the response probabilities do not depend on the realized sample s. Also, the two-phase framework uses a natural decomposition of the total error that leads to a natural decomposition of the total variance. That is, the total variance can be expressed as the sum of the sampling variance, the nonresponse variance and the imputation variance, which allows the survey statistician to get an idea of the relative magnitude of each component. Under the reverse approach, there is no easy interpretation for the variance components (except the imputation variance).

We have considered the case of weighted MRHD imputation within classes. Another version of weighted random hot-deck imputation, which we call weighted joint random hot deck (JRHD) imputation, is identical to weighted MRHD imputation, except that when both variables are missing, a donor j is selected at random from the set of common donors (*i.e.*, the set of respondents to both variables y and x) with probability $w_j / \sum_{l \in s} w_l a_l b_l$ and the vector (x_j, y_j) is imputed. This version of the method helps preserving relationships between survey variables, contrary to imputing independently each variable. The results for JRHD imputation can be obtained using similar techniques presented in this paper. Finally, the results presented in this paper can be easily extended to the case of both deterministic and random regression imputation performed within imputation classes.

Acknowledgements

The author thanks an Associate Editor for constructive comments and suggestions that helped improving the quality of the paper. David Haziza's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Brick, M.J., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.

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

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

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

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.