

INFERENCE FOR DOMAIN MEANS UNDER IMPUTATION FOR MISSING DATA

David Haziza¹ and J.N.K. Rao²

ABSTRACT

A population mean can be estimated unbiasedly under mean imputation for missing data and uniform response or ignorable response, but the imputed estimator of a domain mean is generally biased. Following the approach of Skinner and Rao (1999), we obtain a bias-adjusted estimator of a domain mean under simple random sampling. We derive consistent variance estimators using a method introduced by Fay (1991) in which the usual sample-response path is reversed.

KEYWORDS: Domain; Mean imputation; Model assisted framework; Design-based framework; Ignorable response; Uniform response mechanism.

RÉSUMÉ

Il est possible d'obtenir un estimateur sans biais pour une moyenne sous un mécanisme de réponse uniforme ou ignorable lorsque l'imputation par la moyenne a été utilisée. Cependant, quand il s'agit d'une moyenne d'un domaine, l'estimation obtenue est généralement biaisée. Grâce à une correction, nous obtenons un estimateur sans biais pour la moyenne d'un domaine en considérant l'approche développée par Skinner et Rao (1999) dans le cas d'un échantillon aléatoire simple sans remise. De plus, en utilisant une méthode proposée par Fay (1991), qui suggère de renverser l'ordre habituel de plan de sondage-réponse, nous développons des estimateurs de variance convergents.

MOTS-CLÉS: Domaine; Imputation par la moyenne; Approche assistée d'un modèle; Approche basée sur le plan de sondage; Réponse ignorable; Mécanisme de réponse uniforme.

1. INTRODUCTION

Most surveys use imputation to handle item nonresponse but one should be warned about some of the dangers when imputation is used to compensate for nonresponse: (a) It is common practice to treat the imputed values as if they are true values, and then compute the variance estimates using standard formulas. This can lead to serious underestimation of the true variance of the estimates when the proportion of missing values is not small. (b) The relationships between variables may be distorted.

Marginal item imputation often gives unbiased estimators of marginal population means under uniform or ignorable response. In this article, we are interested in estimating a domain mean when imputation has been used to compensate for nonresponse. Because a domain mean involves a

product term of the form $\sum x_i y_i$, where x_i is the domain indicator variable and y_i is the response variable, marginal imputation may not lead to unbiased estimators under uniform response. Following the idea of Skinner and Rao (1999), we propose a bias-adjusted estimator of the domain mean and derive consistent estimators for its variance.

Traditionally, researchers have used the following sample-response path (sometimes called the two-phase approach) for variance estimation: Population \rightarrow complete sample \rightarrow sample with nonrespondents, but it can lead to cumbersome evaluations. A simpler approach by Fay (1991) proposed a reversing of the order of sampling and response (we will call it the reverse approach) that can be depicted as: Population \rightarrow census with nonrespondents \rightarrow sample with nonrespondents. In this case, (see Shao and Steel, 1999)

¹ David Haziza, Carleton University, Ottawa, Canada, K1S 5B6, dhaziza@math.carleton.ca

² J.N.K. Rao, Carleton University, Ottawa, Canada, K1S 5B6, jrao@math.carleton.ca

$$V(\hat{\theta} - \theta) = E_r(V_s(\hat{\theta} - \theta)) + V_r(E_s(\hat{\theta} - \theta)) \quad (1)$$

where θ denotes an arbitrary parameter and $\hat{\theta}$ denotes its estimator based on the observed and imputed data, E_s and V_s denotes respectively the expectation and the variance operators with respect to the sampling design and E_r and V_r denotes respectively the expectation and the variance operators with respect the response mechanism.

One can show that the order of $\frac{V_r E_s(\hat{\theta} - \theta)}{V_s(\hat{\theta})}$ is $O(\frac{n}{N})$.

Hence, when the sampling fraction is negligible the second component in (1) is negligible in which case we can omit the derivation of its estimator which is generally quite tedious.

2. FRAMEWORK AND ASSUMPTIONS

Let P be a finite population of known size N and let P_d be a domain (subpopulation) of P of size N_d generally unknown. The objective is to estimate the domain mean \bar{Y}_d when imputation has been used to compensate for nonresponse, where $\bar{Y}_d = (\sum_{i=1}^N x_i y_i) / (\sum_{i=1}^N x_i)$ with $x_i = 1$ if unit i belongs to P_d and $x_i = 0$, otherwise. We assume that x_i is known for all the units in the sample.

Suppose a simple random sample, s , of size n is selected without replacement from P . Let s_r be the sample of respondents of size r and let s_m be the sample of nonrespondents of size m ; $r + m = n$. Also, let n_d be the size of $s \cap P_d$ and r_d be the size of $s_r \cap P_d$. To reduce or eliminate biases of imputed estimators, imputation is often done by first dividing P into J nonoverlapping imputation cells and then imputing nonrespondents in an imputation cell using respondents as donors within the same imputation cell, independently across the J imputation cells. For simplicity of notation, we assume that $J = 1$; the extension to $J > 1$ imputation cells is straightforward. We consider two distinct frameworks: (i) design-based, (ii) model-assisted. Under the design-based framework, we assume a uniform response mechanism within cells so that the following assumption holds:

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

Under the model-assisted framework, the following assumption holds:

Assumption MA: Within an imputation cell 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 used for imputation. Imputation is performed according to an imputation model. For simplicity, we consider only mean imputation in which case the imputation model is given by:

$$E_m(y_i) = \beta, \quad V_m(y_i) = \sigma^2 \\ Cov_m(y_i, y_j) = 0 \quad \text{if } i \neq j$$

where β and σ^2 are unknown parameters, E_m , V_m , and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the imputation model. In the model assisted framework, we replace E_r and V_r in (1) by E_m and V_m respectively.

3. POINT ESTIMATION

In this section, we consider two imputed estimators of the domain mean \bar{Y}_d : the unadjusted estimator and the adjusted estimator.

3.1 The unadjusted estimator

An imputed estimator of \bar{Y}_d , denoted by \bar{y}_{dl} , is given by

$$\bar{y}_{dl} = \frac{1}{n_d} \left(\sum_{s_r} x_i y_i + \sum_{s_m} x_i y_i^* \right) \quad (2)$$

where $y_i^* = \bar{y}_r$ and $\bar{y}_r = \frac{1}{r} \sum_{s_r} y_i$ under mean imputation. It is easy to show that under the design-based framework, conditionally given r , the bias of the unadjusted estimator is given by

$$Bias(\bar{y}_{dl}) = (1 - \hat{p})(\bar{Y} - \bar{Y}_d)$$

where $\hat{p} = \frac{r}{n}$ (response rate), and $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$ (overall population mean). The bias will be zero in the full response case (i.e., $\hat{p} = 1$) or if the domain mean \bar{Y}_d and the overall mean \bar{Y} are equal. On the other hand, under the model-assisted framework, the unadjusted estimator (2) is always unbiased for the domain mean \bar{Y}_d , provided the imputation model is true.

3.2 The adjusted estimator

Following Skinner and Rao (1999), a simple bias-adjusted estimator for the domain mean \bar{Y}_d , is given by

$$\bar{y}_{dl}^a = \hat{p}^{-1} \bar{y}_{dl} + (1 - \hat{p}^{-1}) \bar{y}_i \quad (3)$$

where $\bar{y}_l = \frac{1}{n} \left(\sum_{s_r} y_i + \sum_{s_m} y_i^* \right) = \bar{y}_r$ is an unbiased imputed estimator of the overall mean \bar{Y} under both the design-based and the model-assisted frameworks. It is easy to show that the adjusted estimator is unbiased for the domain mean \bar{Y}_d under both the design-based and the model-assisted frameworks. Hence, the adjusted estimator (3) is robust in the sense of validity under both frameworks.

4. VARIANCE ESTIMATION

To estimate the variance of the estimators proposed in section 3, we use the reverse approach of Fay (1991).

4.1 The unadjusted estimator

To apply the reverse approach, we first write the unadjusted estimator \bar{y}_{dl} as

$$\bar{y}_{dl} = \frac{1}{\sum_s w_i x_i} \left[\sum_s w_i a_i x_i y_i + \left(\frac{\sum_s w_i a_i y_i}{\sum_s w_i a_i} \right) \times \left(\sum_s w_i x_i - \sum_s w_i a_i x_i \right) \right] = \varphi(\hat{\mathbf{T}})$$

where

$$\hat{\mathbf{T}} = \left(\sum_s w_i t_{1i}, \sum_s w_i t_{2i}, \sum_s w_i t_{3i}, \sum_s w_i t_{4i}, \sum_s w_i t_{5i} \right)^T,$$

with $t_{1i} = x_i, t_{2i} = a_i, t_{3i} = a_i y_i, t_{4i} = a_i x_i y_i,$

$t_{5i} = a_i x_i, \varphi$ is a smooth function of totals,

$w_i = n/N$ under simple random sampling, $a_i = 1$ if unit i belongs to s and $a_i = 0$, otherwise. It follows

from (1) that $V(\hat{\bar{Y}}_{dl})$ can be estimated by $v_i = v_1 + v_2$, where v_1 is an estimator of $V_s(\bar{y}_{dl})$, conditional on the a_i 's and v_2 is an estimator of $V_r[E_s(\bar{y}_{dl} - \bar{Y}_d)]$. Using Taylor linearization, we get

$$v_1 = \nabla(\varphi(\hat{\mathbf{T}}))^T \mathbf{V} \nabla(\varphi(\hat{\mathbf{T}}))$$

where ∇ is the vector of partial derivatives and \mathbf{V} is a matrix whose (k,l) element is a standard design-based estimator of $\text{cov}_s(\sum_s w_i t_{ki}, \sum_s w_i t_{li})$. Denote the variance estimator of $\hat{Y} = \sum_s w_i y_i$ as $v(y_i)$. Then we can show that v_1 reduces to:

$$v_1 = v(\hat{\xi}_i) \quad (4)$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i x_i} (\hat{\xi}_{li} - \hat{R}_{\xi_1} x_i),$$

$$\hat{\xi}_{li} = a_i x_i y_i + (1 - a_i) x_i \bar{y}_r + [(n_d - r_d)/r] a_i (y_i - \bar{y}_r)$$

$$\text{and } \hat{R}_{\xi_1} = \sum_s \hat{\xi}_{li} / n_d.$$

Note that v_1 , given by (4) does not depend on the response mechanism or the model. Hence, it is valid under both design-based and model-assisted approaches. To determine v_2 , we need to distinguish between the design-based and the model-assisted approaches since it depends on the assumptions on the response mechanism and/or model. Using Taylor linearization, we have

$$V[E_s(\bar{y}_{dl}^a - \bar{Y}_d)] \approx \nabla(\phi E_r(\tilde{\mathbf{T}}))^T \mathbf{C} \nabla(\phi E_r(\tilde{\mathbf{T}})) \quad (5)$$

where $\phi(\tilde{\mathbf{T}}) = E_s(\bar{y}_{dl}^a - \bar{Y}_d) = \phi(E_s \hat{\mathbf{T}}) - \bar{Y}_d$, ϕ is a smooth function of totals, and

$$\tilde{\mathbf{T}} = (\bar{Y}_d, \sum_p x_i, \sum_p a_i, \sum_p a_i y_i, \sum_p a_i x_i y_i, \sum_p a_i x_i)^T.$$

Under the design-based approach the matrix \mathbf{C} is a matrix whose (k,l) element is $\text{cov}_r(\sum_p w_i t_{ki}, \sum_p w_i t_{li})$ and under the model assisted approach \mathbf{C} is a matrix whose (k,l) element is $\text{cov}_m(\sum_p w_i t_{ki}, \sum_p w_i t_{li})$. The second variance

component, denoted by v_{2DB} or v_{2MA} (DB and MA stand for design-based and model-assisted, respectively), is then obtained by substituting estimators for the unknown quantities in (5). We can show that under simple random sampling and uniform response, v_{2DB} is given by

$$v_{2DB} = \left(\frac{n}{N} - \frac{r}{N} \right) \times \left[\left(\frac{r_d}{n_d} - 1 \right)^2 \frac{s_r^2}{r} + \left(\frac{r + 2(n_d - r_d)}{n_d} \right) \frac{s_d^2}{n_d} \right] \quad (6)$$

where $(r-1)s_r^2 = \sum_s a_i (y_i - \bar{y}_r)^2$ and

$$(r-1)s_d^2 = \sum_s a_i x_i (y_i - \bar{y}_r)^2.$$

We can also show that under simple random sampling and ignorable response, v_{2MA} is given by

$$v_{2MA} = \frac{s_r^2}{N} \left[\left(\frac{n}{n_d} \right) \left(1 - \frac{r_d}{n_d} \right) + \left(\frac{n}{r} \right) \left(1 - \frac{r_d}{n_d} \right)^2 \right] \quad (7)$$

4.2 The adjusted estimator

We have also derived the two variance components, v_1 and v_2 for the adjusted estimator. The derivation

involves tedious but fairly straightforward algebra. We obtained

$$v_1 = v(\tilde{\xi}_i) \quad (8)$$

where

$$\tilde{\xi}_i = \frac{1}{N} \left[\left(\frac{n}{r} \right) \left(\frac{n}{n_d} \right) (\tilde{\xi}_{1i} - \tilde{R}_{\xi_1} x_i) + \left(1 - \frac{n}{r} \right) (\tilde{\xi}_{2i} - \tilde{R}_{\xi_2}) \right]$$

$$\tilde{\xi}_{1i} = a_i x_i y_i + (1 - a_i) x_i \bar{y}_r + [(n_d - r_d)/r] a_i (y_i - \bar{y}_r)$$

$$\tilde{\xi}_{2i} = a_i y_i + (1 - a_i) \bar{y}_r + \left(\frac{n}{r} - 1 \right) a_i (y_i - \bar{y}_r),$$

$$\tilde{R}_{\xi_1} = \sum_s \tilde{\xi}_{1i} / n_d, \text{ and } \tilde{R}_{\xi_2} = \sum_s \tilde{\xi}_{2i} / n.$$

Note that if $x_i = 1$ for all i in the population, (i.e., if $\bar{Y}_d = \bar{Y}$), then (4) and (8) reduce to (15) obtained by Shao and Steel (1999) in the case of mean imputation.

We also obtained v_{2DB} and v_{2MA} as

$$v_{2DB} = \left(\frac{n}{N} - \frac{r}{N} \right) \left(1 - \left(\frac{n}{r} \right) \left(\frac{r_d}{n_d} \right) \right)^2 \frac{s_r^2}{r} + \left(1 - 2 \frac{r_d}{r} \right) \left(\frac{n}{n_d} \right)^2 \frac{s_d^2}{r} - \left(\frac{n}{r} \right)^2 \left(\frac{r_d}{n_d} \right)^2 \frac{1}{r} (\bar{y}_{rd} - \bar{y}_r)^2 \quad (9)$$

where $\bar{y}_{rd} = \sum_s a_i x_i y_i / r_d$, and

$$v_{2MA} = \frac{s_r^2}{N} \left[\left(\frac{n}{r} \right) + \left(\frac{n}{n_d} \right) - 2 \left(\frac{n}{r} \right) \left(\frac{r_d}{n_d} \right) + \left(\frac{n}{r} \right)^2 \left(\frac{r_d}{n_d} \right)^2 - 2 \left(\frac{n}{n_d} \right)^2 \left(\frac{r_d}{n_d} \right) + \left(\frac{n}{n_d} \right) \left(\frac{r_d}{n_d} \right) \left(\frac{n}{r} \right)^2 - \left(\frac{r_d}{n_d} \right)^2 \left(\frac{n}{r} \right)^3 \right] \quad (10)$$

Note that if $x_i = 1$ for all i in the population, then (9) and (10) reduce to (18) and (26) obtained by Shao and Steel (1999) in the case of mean imputation.

5. SIMULATION STUDY

We conducted a small simulation study using a population of size $N = 2000$ (Lohr 1999, Appendix C, p. 441) containing two domains (males and females) of sizes $N_d = 1000$, $d = 1, 2$. The variable of interest, y , is the variable *Height* in cm. The objective is to estimate the mean of this variable for females in the population. We drew $R = 1000$ repeated simple random samples, s , of size varying between 50 and 120. Nonresponse was generated using Bernoulli trials, which corresponds to a uniform response mechanism. The response rate was set at 0.7.

To measure the relative bias of an estimator $\hat{\theta}$, we used $B_{rel}(\hat{\theta}) = Bias(\hat{\theta}) / s.e(\hat{\theta})$. Table 1 reports the simulation values of relative bias of the unadjusted and adjusted estimators, the MSE of the two estimators and the bias ratios of the design-based estimators v_1 and $v_i = v_1 + v_2$ of the unadjusted and adjusted estimators, where

$$\text{bias ratio } B_r(\text{var. est.}) = \frac{E(\text{var. est.}) - MSE(\text{est.})}{MSE(\text{est.})}$$

It is clear from Table 1 that the relative bias of the unadjusted estimator is substantial while that of the adjusted estimator is negligible, as expected. As a result the MSE of the adjusted estimator is considerably smaller than the MSE of the unadjusted estimator, especially when n increases. The variance estimator $v_i(\text{adj.})$ tracks the MSE of the adjusted estimator well, whereas the variance estimator $v_i(\text{unadj.})$ is a serious underestimate of the MSE of the unadjusted estimator, as judged by the values of bias ratio reported in Table 1. Also, the sampling variance estimator v_1 performs well in tracking MSE even when the sampling fraction is appreciable; note from Table 1 that the bias ratio of v_1 is close to the bias ratio of v_i .

Table 1: Relative biases and MSE of estimators and bias ratios of variance estimators

	$n = 50$	$n = 80$	$n = 120$
$B_{rel}(\text{unadj.})$	1.128	1.400	1.412
$B_{rel}(\text{adj.})$	0.0346	0.00186	0.00381
$B_r(v_1)(\text{unadj.})$	-0.563	-0.654	-0.757
$B_r(v_i)(\text{unadj.})$	-0.556	-0.650	-0.753
$B_r(v_1)(\text{adj.})$	0.053	0.057	0.047
$B_r(v_i)(\text{adj.})$	0.069	0.075	0.087
$MSE(\text{unadj.})$	7.4532	5.6237	5.2551
$MSE(\text{adj.})$	4.0385	2.0156	1.5777

6. CONCLUSION

This article focussed on estimating domain means under imputation for missing data. We have proposed a simple bias-adjusted estimator which, in our view, is attractive since its use can be justified under both the design-based and the model-assisted frameworks. We also derived consistent variance estimators using an approach proposed by Fay (1991) and developed by Shao and Steel (1999). For simplicity, we have presented the special case of simple random sampling and mean imputation but we have extended the results to the case of ratio imputation, stochastic imputation and stratified multistage sampling under uniform response. We are presently investigating the extension to general functions $z(x, y)$ such as regression and correlation coefficients and cell proportions in two-way tables.

REFERENCES

- [1] Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the census*, pp. 429-440.
- [2] Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- [3] 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.
- [4] Skinner, C.J. and Rao, J.N.K. (1999). Jackknife Variance for Multivariate Statistics under Hot-deck Imputation from Common Donors. *Journal of Statistical Planning and Inference*, in press.