

Resampling methods for variance estimation in the presence of missing survey data

David Haziza

Abstract Resampling methods (e.g., jackknife and bootstrap) are often used in surveys. Unlike Taylor linearization procedures, these methods neither require separate derivation for each particular estimator nor require the second-order inclusion probabilities that may be difficult to obtain for complex designs. In the presence of imputed data, resampling methods have been adapted in order to take the nonresponse variance into account (e.g., Rao and Shao, 1992 and Shao and Sitter, 1996). In this paper, we review these methods and we describe some variance estimators of the total variance that are asymptotically unbiased and consistent when the sampling fraction is not negligible.

Key words: Bootstrap, jackknife, imputation, resampling methods, reverse framework, variance estimation

1 Introduction

The problem of variance estimation in the presence of imputed data has been widely treated in the literature. It is well known that treating the imputed values as if they were real observations will typically result in estimated variances (or coefficients of variation) that are usually too small because they fail to account for the variance due to nonresponse and imputation. As a result, inferences are generally invalid. For example, the coverage probability of 95% confidence intervals obtained by using naive variance estimation methods may be considerably smaller than the nominal level, especially if the nonresponse rate is high. This led researchers to develop variance estimation methods that account for the variance due to nonresponse and

David Haziza
Département de mathématiques et de statistique, Université de Montréal, e-mail:
david.haziza@umontreal.ca

imputation.

Traditionally, the total variance of the imputed estimator has been expressed as the sum of the sampling variance and the nonresponse variance. This decomposition of the total variance results from viewing nonresponse as a second-phase of selection. For this reason, this framework is often called the two-phase framework; see Rao (1990), Särndal (1992) and Deville and Särndal (1994), among others. In this paper, we consider an alternative framework, which we call the *reverse framework*; see Fay (1991) and Shao and Steel (1999). It consists of viewing the situation prevailing in the presence of nonresponse as follows: first, applying the nonresponse mechanism, the finite population U is randomly divided into a population of respondents U_r and a population of nonrespondents U_m . Then, given (U_r, U_m) , a sample s , containing both respondents and nonrespondents, is selected from U according to the chosen sampling design. Unlike the two-phase framework, it helps to clarify and to justify the properties of variance estimators obtained using replication methods in the presence of imputed data. The reader is referred to Rao and Shao (1992), Rao and Sitter (1995), Shao and Sitter (1996), Shao (2002) and Davison and Sardy (2007) for a description of replication methods in the presence of imputed data. In practice, replication methods are often used because, unlike Taylor linearization procedures, they do not require separate derivation for each particular estimator nor do they require second-order inclusion probabilities that may be difficult to obtain for complex designs.

2 Set-up

Let $U = \{1, 2, \dots, N\}$ be a population of N elements. We are interested in estimating a population mean $\bar{Y} = N^{-1} \sum_{i \in U} y_i$, where y_i denotes the i -th value of the variable of interest y , $i = 1, \dots, N$. We select a sample, s , of size n , according to a given sampling design $p(s)$. Let $d_i = 1/\pi_i$ denote the design weight attached to unit i , where π_i denote its first-order inclusion probability. In the absence of nonresponse, a design-unbiased estimator of $\bar{Y}$ is the expansion estimator given by $\hat{Y}_\pi = N^{-1} \sum_{i \in s} d_i y_i$. That

is, $E_p(\hat{Y}_\pi) = \bar{Y}$, where the subscript p denotes the sampling design $p(s)$. In the presence of nonresponse to item y , only a subset of s , of size n_r , respond to item y , in which case we define an imputed estimator $\hat{Y}_I$ by

$$\hat{Y}_I = \frac{1}{N} \left[\sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^* \right], \quad (1)$$

where y_i^* denotes the imputed value used to replace the missing value y_i and $r_i = 1$ if unit i responds to item y and $r_i = 0$, otherwise. For simplicity, we consider the case of mean imputation. The extension to the more general case of regression imputation is relatively straightforward. Under mean imputation, the imputed values y_i^* are given

by

$$y_i^* = \frac{\sum_{i \in s} d_i r_i y_i}{\sum_{i \in s} d_i r_i} \equiv \hat{Y}_r \quad (2)$$

and the imputed estimator (1) reduces to

$$\hat{Y}_I = \hat{Y}_r \equiv \psi \left(\sum_{i \in s} d_i r_i y_i, \sum_{i \in s} d_i r_i \right).$$

In this paper, we assume that the sampled units respond to item y independently of one another with probability p . This nonresponse mechanism is often called a uniform nonresponse mechanism.

3 Variance estimation under the reverse framework

Under the reverse framework, the total variance of $\hat{Y}_I$, denoted by V_T , can be expressed as

$$V_T = E_q V_p(\hat{Y}_I | \mathbf{r}) + V_q E_p(\hat{Y}_I | \mathbf{r}) \equiv V_1 + V_2, \quad (3)$$

where the subscript q denotes the nonresponse mechanism and $\mathbf{r} = (r_1, \dots, r_N)'$ is the vector of response indicators. To estimate the term V_1 , it suffices to estimate $V_p(\hat{Y}_I | \mathbf{r})$, which represents the variance due to sampling conditional on the vector of response indicators $\mathbf{r}$. Because the response indicator r_i can now be viewed as a characteristic of unit i , the problem of estimating V_1 reduces to that of estimating the sampling variance of a smooth function (namely, $\psi(\cdot)$) of estimated totals, which is a standard problem in design-based variance estimation. An estimator of V_1 is readily obtained by using any standard variance estimation technique available in the complete data case. For example, Taylor linearization procedures or replication methods such as the jackknife or the bootstrap can be used. Estimating the term V_2 requires the specification of a nonresponse model and can be obtained using a Taylor linearization procedure.

Under mild regularity conditions, the term V_1 is of order $O(n^{-1})$, whereas the term V_2 is of order $O(N^{-1})$. Therefore, the contribution of V_2 to the total variance, $\frac{V_2}{V_1 + V_2}$ is of order $O\left(\frac{n}{N}\right)$ and it follows that the contribution of V_2 is negligible when the sampling fraction, $f \equiv n/N$, is negligible. In this case, the term V_2 may be omitted from the calculations and an estimator of the total variance is given by $\hat{V}_T = \hat{V}_1$. Under mild regularity conditions, the estimator $\hat{V}_1$ is asymptotically unbiased and consistent for V_1 even if the underlying assumptions about the nonresponse mechanism are not verified. For example, suppose that the sample s was selected according to simple random sampling without replacement, in which case $d_i = N/n$ for all i . A first-order Taylor expansion leads to

$$\hat{V}_1 = (1 - f) \frac{s_r^2}{n_r}, \quad (4)$$

where $s_r^2 = (n_r - 1)^{-1} \sum_{i \in s} r_i (y_i - \hat{Y}_r)^2$. When the sampling fraction is not negligible, we must take the term V_2 into account. Noting that $E_p(\hat{Y}_I | \mathbf{r}) \approx \sum_{i \in U} r_i y_i / \sum_{i \in U} r_i \equiv \bar{Y}_r$, we approximate V_2 under uniform nonresponse (in which case $E_q(r_i) = p$ and $V_q(r_i) = p(1 - p)$) by

$$V_2 \approx \left(\frac{N}{\sum_{i \in U} r_i} \right)^2 p(1 - p) \sum_{i \in U} r_i (y_i - \bar{Y}_r)^2. \quad (5)$$

Under simple random sampling without replacement, an estimator of V_2 , denoted by $\hat{V}_2$, is obtained by estimating the unknown quantities in (5), which leads to

$$\hat{V}_2 = \frac{n_m}{N} \frac{s_r^2}{n_r}, \quad (6)$$

where $n_m = n - n_r$ denotes the number of nonrespondents. The estimator $\hat{V}_2$ is asymptotically unbiased for V_2 under uniform nonresponse; that is $E_q E_p(\hat{V}_2 | \mathbf{r}) \approx V_2$. Finally, under simple random sampling without replacement, an estimator of V_T is given by

$$\hat{V}_T = \hat{V}_1 + \hat{V}_2 = \left(1 - \frac{n_r}{N}\right) \frac{s_r^2}{n_r}, \quad (7)$$

where $\hat{V}_1$ and $\hat{V}_2$ are given by (4) and (6), respectively. The estimator $\hat{V}_T$ is the so-called linearization variance estimator of V_T .

4 Resampling methods under complete response

In this section, we briefly describe two resampling methods in the complete data case: the jackknife and the Rao-Wu bootstrap. We assume that the sample s is selected according to simple random sampling without replacement, in which case $\hat{Y}_\pi = n^{-1} \sum_{i \in s} y_i \equiv \bar{y}$.

4.1 The jackknife

In the absence of nonresponse, a jackknife variance estimator is obtained as follows:

- (i) remove the unit $j = 1$ from the sample;

- (ii) adjust the design weights $d_i = N/n$ to obtain the so-called jackknife weights $\tilde{d}_{i(j)}$, where $\tilde{d}_{i(j)}$ is given by

$$\tilde{d}_{i(j)} = \begin{cases} \frac{N}{n} \frac{n}{n-1} = \frac{N}{n-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

- (iii) compute the estimator $\bar{y}_{(j)}$ which is calculated the same way as $\bar{y}$ but using the adjusted weights $\tilde{d}_{i(j)}$ instead of the design weights d_i ;
 (iv) insert back unit $j = 1$ deleted in step (i);
 (v) repeat the steps (i)-(iv) for $j = 2, \dots, n$.

A jackknife variance estimator of $\bar{y}$ is then given by

$$\hat{V}_J = \left(\frac{n-1}{n} \right) \sum_{j=1}^n (\bar{y}_{(j)} - \bar{y})^2. \quad (8)$$

It is well known that $\hat{V}_J$ in (8) reduces to $\hat{V}_J = s_y^2/n$, where $s_y^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y})^2$. That is, $\hat{V}_J$ corresponds to the textbook variance estimator under simple random sampling with replacement. A trivial modification of $\hat{V}_J$ consists of incorporating the finite population correction, $1-f$, to obtain $\hat{V}_J^* = (1-f)\hat{V}_J$ which corresponds to the textbook variance estimator under simple random sampling without replacement.

4.2 The Rao-Wu bootstrap

Rao and Wu (1988) proposed a rescaling bootstrap method in order to estimate the variance of $\hat{Y}_\pi$. Their method draws bootstrap samples of size n' with replacement from the rescaled sample s . Note that n' may be different from n . The rescaling factor, denoted by C , is chosen so that the variance under resampling matches the usual variance estimator of the population mean. The Rao-Wu bootstrap may be described as follows:

- (i) Let n' be the bootstrap sample size. Let $z_i = \bar{y} + \sqrt{C}(y_i - \bar{y})$, $i = 1, \dots, n$, where $C = n'(1-f)/(n-1)$.
 (ii) Draw a simple random sample $\{z_i^*\}_{i=1}^{n'}$ with replacement from $\{z_i\}_{i=1}^n$. Let $\bar{z}^* = \sum_{i=1}^{n'} z_i^* / n'$.
 (iii) Repeat (ii) a large number of times, B , to get $\bar{z}_{(1)}^*, \dots, \bar{z}_{(B)}^*$.

The Rao-Wu bootstrap variance estimator is given by

$$\hat{V}_{RW} = \frac{1}{B-1} \sum_{b=1}^B (\bar{z}_{(b)}^* - \bar{z}_{(\cdot)}^*)^2,$$

where $\bar{z}_{(\cdot)}^* = \sum_{b=1}^B \bar{z}_{(b)}^* / B$. The estimator $\hat{V}_{RW}$ is a consistent estimator of $(1-f)s_y^2/n$, which corresponds to the textbook variance estimator under simple random sampling without replacement.

5 Resampling methods in the presence of imputed data

Resampling methods have been adapted to take the nonresponse into account. In the presence of imputed data, Rao and Shao (1992) proposed a jackknife variance estimator and Shao and Sitter (1996) proposed a bootstrap variance estimator. As we argue in sections 5.1 and 5.2, both variance estimators can be justified by the reverse framework as they provided an estimator of V_1 (see section 3).

5.1 The Rao-Shao jackknife

In the presence of nonresponse to item y , the use of (8) may lead to serious underestimation of the variance of the estimator, especially if the nonresponse rate is important. Rao and Shao (1992) proposed an adjusted jackknife method that is calculated in a similar fashion as (8) except that, whenever a responding unit is deleted, the imputed values are adjusted. The imputed values are unchanged if a nonresponding unit is deleted. Let $y_{i(j)}^{a*}$ denote the adjusted imputed value for unit i when unit j was deleted. For mean imputation, we have $y_{i(j)}^{a*} = \bar{y}_{r(j)}$ if $r_j = 1$ and $y_{i(j)}^{a*} = \bar{y}_r$ if $r_j = 0$, where $\bar{y}_{r(j)}$ denotes the mean of the respondents excluding unit j . The Rao-Shao jackknife variance estimator is then given by

$$\hat{V}_{JRS} = \left(\frac{n-1}{n} \right) \sum_{j=1}^n \left(\hat{Y}_{I(j)}^a - \hat{Y}_I \right)^2 \approx \frac{s_r^2}{n_r}, \quad (9)$$

assuming that $n/n-1 = 1$ and $n_r/n_r-1 = 1$, where $\hat{Y}_{I(j)}^a$ is computed the same way as $\hat{Y}_I$ but with the adjusted imputed values $y_{i(j)}^{a*}$ instead of the original imputed values y_i^* . That is, $\hat{V}_{JRS}$ is asymptotically equivalent to $\hat{V}_1$ in (4) if we neglect the finite population correction, $(1-f)$. In fact $\hat{V}_{JRS}$ is an estimator of $V_p(\hat{Y}_I|\mathbf{r})$ that we would have obtained had the sampling been performed using simple random sampling with replacement. Hence, $\hat{V}_{JRS}$ is asymptotically unbiased and consistent for V_1 under simple random sampling with replacement, or equivalently, if the sampling fraction f is negligible. Note that the second term, V_2 , is not accounted for. Intuitively, this can be explained by the fact that once an unit is deleted, nonresponse is not generated in each jackknife replicate before the imputation process is performed. In other words, the Rao-Shao adjusted jackknife simulates the effect of (with replacement) sampling conditional only on the vector of response indicators

r. Now, since the contribution of the term V_2 to the total variance is negligible when the sampling fraction f is negligible, $\hat{V}_{JRS}$ is asymptotically unbiased and consistent for the total variance of $\hat{Y}_I$. This is not true anymore if f is appreciable. In this case, it is important to take the term V_2 into account. An estimator of the total variance is given by

$$\hat{V}_J^* = (1 - f)\hat{V}_{JRS} + \hat{V}_2,$$

where $\hat{V}_2$ is given by (6); see Haziza and Picard (2009). Since $\hat{V}_{JRS}$ is an estimator of $V_p(\hat{Y}_I|\mathbf{r})$, and $\hat{Y}_I = \psi(\sum_{i \in S} d_i r_i y_i, \sum_{i \in S} d_i r_i)$, where $\psi(\cdot)$ is a smooth function of estimated totals (see section 2), $\hat{V}_{JRS}$ can alternatively be obtained by using a standard (complete-data) jackknife procedure designed for estimating the variance of smooth functions of totals. Therefore, under deterministic imputation, no adjustment of the imputed values is necessary if an appropriate standard jackknife procedure is applied. As a result, the Rao-Shao jackknife variance estimator can be obtained using software designed for complete data jackknife variance estimation. This is an attractive property from a practical point of view.

5.2 The Shao-Sitter bootstrap

Applying the Rao-Wu bootstrap in the presence of missing responses and treating the missing values as true values, may lead to serious underestimation of the variance of the estimator. In the presence of imputed data, Shao and Sitter (1996) proposed a bootstrap procedure for imputed survey data that may be described as follows:

- (i) Let $z_i = \bar{y}^I + \sqrt{C}(y_i^I - \bar{y}^I)$, for $i = 1, \dots, n$, where $C = n'(1 - f)/(n - 1)$.
- (ii) Draw a simple random sample of size n' , $\{(z_i^*, r_i^*)\}_{i=1}^{n'}$, with replacement from $\{(z_i, r_i)\}_{i=1}^n$.
- (iii) Let $z_i^{*I} = z_i^*$ if $r_i^* = 1$ and $z_i^{*I} = \bar{z}_r^*$ if $r_i^* = 0$, where $\bar{z}_r^* = \sum_{i=1}^{n'} r_i^* z_i^* / \sum_{i=1}^{n'} r_i^*$. Also, let $\bar{z}^{*I} = \sum_{i=1}^{n'} z_i^{*I} / n'$.
- (iv) Repeat (ii)-(iii) a large number of times, B , to get $\bar{z}_{(1)}^{*I}, \dots, \bar{z}_{(B)}^{*I}$.

The Shao-Sitter bootstrap variance estimator is then given by

$$\hat{V}_{BSS} = \frac{1}{(B-1)} \sum_{b=1}^B (\bar{z}_{(b)}^{*I} - \bar{z}_{(\cdot)}^{*I})^2,$$

where $\bar{z}_{(\cdot)}^{*I} = \sum_{b=1}^B \bar{z}_{(b)}^{*I} / B$. The estimator $\hat{V}_{BSS}$ estimates $(1 - f)s_r^2/r$ consistently, which is identical to $\hat{V}_1$ in (4). Thus, as for $\hat{V}_{JRS}$, $\hat{V}_{BSS}$ is an estimator of $V_p(\hat{Y}_I|\mathbf{r})$ under simple random sampling without replacement. As for the Rao-Shao jackknife variance estimator, the term V_2 is not accounted for. Intuitively, this can be explained by the fact nonresponse is not generated in each bootstrap sample before the imputation process is performed. In other words, a respondent in the original

sample is automatically a respondent in the bootstrap samples and the same is true for a nonrespondent. Therefore, the rescaling Shao-Sitter bootstrap variance estimator simulates the effect of (without replacement) sampling conditional only on the vector of response indicators $\mathbf{r}$. As for the Rao-Shao jackknife variance estimator, $\hat{V}_{BSS}$ can alternatively be obtained by using a standard (complete-data) Rao-Wu bootstrap procedure to a smooth function of means. Therefore, under deterministic imputation, there is no need to re-impute within each bootstrap sample and the Shao-Sitter bootstrap variance estimator can be obtained using software designed for complete data bootstrap variance estimation.

Now, since the contribution of the term V_2 to the total variance is negligible when the sampling fraction f is negligible, $\hat{V}_{BSS}$ is asymptotically unbiased and consistent for the total variance of $\hat{Y}_I$. Once again, this is not true anymore if f is appreciable. In this case, an estimator of the total variance is given by

$$\hat{V}_B^* = \hat{V}_{BSS} + \hat{V}_2,$$

Alternatively, it is possible to use a bootstrap procedure, which tracks the total variance $V_T = V_1 + V_2$, in the case of nonnegligible f . It is presented in the next section.

5.3 Bootstrap procedure for estimating the total variance

Haziza, Léger and Mashreghi (2010) proposed a bootstrap procedure, which leads to an asymptotically unbiased and consistent estimator of the total variance, $V_T = V_1 + V_2$, even if the sampling fraction f is not negligible. It can be described as follows:

- (i) Let $z_i = \bar{y}^I + \sqrt{C}(y_i^I - \bar{y}^I)$, for $i = 1, \dots, n$, where $C = n'[1 - (n_r/N)]/(n_r - 1)$.
- (ii) Draw a simple random sample of size n' , $\{z_i^*\}_{i=1}^{n'}$, with replacement from $\{z_i\}_{i=1}^n$ and then independently, draw a bootstrap sample of response indicator of size n' , $\{r_i^*\}_{i=1}^{n'}$, with replacement from the original response indicators $\{r_i\}_{i=1}^n$.
- (iii) Let $\bar{z}_i^{*I} = z_i^*$ if $r_i^* = 1$ and $\bar{z}_i^{*I} = \bar{z}_r^*$ if $r_i^* = 0$, where $\bar{z}_r^* = \sum_{i=1}^{n'} r_i^* z_i^* / \sum_{i=1}^{n'} r_i^*$. Also, let $\bar{z}^{*I} = \sum_{i=1}^{n'} \bar{z}_i^{*I} / n'$.
- (v) Repeat Steps (ii)-(iii) a large number of times, B , to get $\bar{z}_{(1)}^{*I}, \dots, \bar{z}_{(B)}^{*I}$.

The resulting bootstrap variance estimator is then given by

$$\hat{V}_B = \frac{1}{(B-1)} \sum_{b=1}^B (\bar{z}_{(b)}^{*I} - \bar{z}_{(\cdot)}^{*I})^2.$$

The estimator $\hat{V}_B$ estimates $(1 - n_r/N)s_r^2/n_r$ consistently, which coincides with the linearization variance estimator given by (7).

References

1. Davison, A. C. and Sardy, S. (2007). Resampling Variance Estimation in Surveys with Missing Data. *Journal of Official Statistics*, 23, 371-386.
2. Deville, J. C. and Särndal, C. E. (1994), Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 23, 33-40.
3. 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.
4. Haziza, D. and Picard, F. (2009). On doubly robust point and variance estimation in the presence of imputed data . Submitted for publication.
5. Haziza, D., Léger, C. and Mashreghi, Z. (2010). Bootstrap variance estimation in the presence of imputed data . Unpublished manuscript.
6. Rao, J. N. K. (1990), Variance estimation under imputation for missing data. *Technical report*, Statistics Canada, Ottawa.
7. Rao, J. N. K. and Shao, J. (1992). On Variance Estimation under Imputation for Missing Data. *Biometrika*, 79, 811-822.
8. Rao, J. N. K. and Sitter, R. R. (1995), Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
9. Rao, J. N. K. and Wu, C.F.J. (1988), Resampling inference for complex survey data. *Journal of American Statistical Association*, 83, 231-241.
10. Särndal, C. E. (1992), Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
11. Shao, J. (2002), Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In *Survey Nonresponse edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little*, New York: John Wiley and Sons, 303-314.
12. Shao, J. and Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association*, 93, 819-831.
13. 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.