VARIANCE ESTIMATION IN THE PRESENCE OF IMPUTED DATA

Jill M. Montaquila, Westat, Inc. and Robert W. Jernigan, American University
Jill M. Montaquila, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

KEY WORDS: Hot-deck; Model-assisted

1. Introduction

Imputation, a technique that fills in hypothetical values for missing data, is widely used to account for item nonresponse in surveys. Traditionally, the imputed values have been treated as if they had actually been observed or reported, and estimates and variance estimates have been computed using standard complete data methods. While the conditions under which estimators based on imputed survey data are unbiased or consistent are well-known, the problem of obtaining valid variance estimates for imputed data has long troubled survey practitioners. Treating imputed values as if they had actually been observed or reported leads to underestimation of the variance of the estimator. Even for items with relatively low nonresponse rates, this downward bias in the variance estimates may be substantial.

Over the past two decades, several methods have been proposed to account for the additional variance due to imputation error. Rubin (1977, 1978) developed the method of multiple imputation. Rao and Shao (1992) proposed an adjusted jackknife variance estimation procedure. Rancourt *et al.* (1994) developed a model-assisted approach. Fay (1996b) presented a variation on the Rao-Shao jackknife called fractionally weighted imputation. Shao and Sitter (1996) proposed a bootstrap procedure to account for the imputation error variance.

It was our goal to develop a procedure for variance estimation for imputed data that would yield valid variance estimates under different imputation schemes. Due to the proliferation of auxiliary data generally available, survey practitioners frequently use variations of nearest neighbor, ratio, or regression imputation methods. Occasionally, when little auxiliary information is available or the auxiliary data are not highly correlated with the characteristic of interest or with the response propensity, the random hot-deck procedure is used.

We sought to develop a procedure that could be used regardless of the imputation method, and could be extended "naturally" to complex sample designs (By "naturally," we mean using the same types of derivations used to extend the variance of the mean under simple random sampling to the variance of the mean for stratified random sampling, for instance.). We have developed such an approach, which we will refer to as "all-cases imputation," or ACI. This approach imputes for all cases (including respondents), and then uses information about the relationship of imputed values to actual values for respondents to reflect uncertainty introduced by missing data and the imputation process.

2. The All-Cases Imputation Variance Estimator: A New Approach to Variance Estimation for Imputed Data

In this section, we present a new approach to variance estimation for imputed data. Section 2.1 defines and describes the variance estimator under simple random sampling, and Section 2.2 extends the variance estimator to stratified random sampling. In Section 2.3, we describe the general approach for extending this variance estimator to other sample designs or estimators other than means.

2.1 Simple Random Sampling

For the sake of discussion, we will focus on estimates of population means. Let y be the characteristic of interest. We will partition the sample (S) into the set of respondents to y (R) and the set of nonrespondents to y (NR). Throughout the discussion of simple random sampling, we will assume the sampling fraction (n/N) is negligible, so that we can ignore the finite population correction, $fpc = 1 - \frac{n}{N}$. Later, during the discussion of stratified sampling, we will incorporate *fpc*, since sampling fractions within some strata may not be small.

With imputed data, the standard estimate of the population mean is:

$$\bar{y}_I = \frac{1}{n}\left\{ \sum_{i \in R} y_i + \sum_{i \in NR} y_i^* \right\} \qquad (1)$$

where y_i is the actual (observed or reported) value for respondent *i*, and y_i^* is the imputed value for nonrespondent *i*.

Under the assumption of ignorable item nonresponse, the variance of the estimator $\bar{y}_I$ can be decomposed as follows:

$$v(\bar{y}_I) = v(\bar{y}_S) + \frac{1}{n^2}\left[\sum_{i \in NR} v\left(\tau_i^{(k)}\right) + 2 \sum_{i \in NR} \sum_{j \in NR} \operatorname{cov}\left(\tau_i^{(k)}, \tau_j^{(l)}\right)\right] \qquad (2)$$

where

$$\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_i \ , \ \tau_i^{(k)} = y_i^* - y_i = y_k - y_i$$

is the imputation error incurred when respondent k is used as the donor for nonrespondent i, and $\operatorname{cov}(x,y)$ denotes the covariance between x and y.

We will refer to the first term in expression (2), $v(\bar{y}_S)$, as the sampling error variance component. We will refer to the second term, $\frac{1}{n^2}\sum_{i \in NR} v\left(\tau_i^{(k)}\right)$, as the imputation error variance component, and will refer to the third term, $\frac{2}{n^2}\sum_{i \in NR}\sum_{\substack{j \in NR \\ j>i}} \operatorname{cov}\left(\tau_i^{(k)}, \tau_j^{(l)}\right)$, as the imputation error covariance component.

Since y is not observed for nonrespondents, the imputation error variance and covariance terms cannot be estimated directly based on the set of nonrespondents (NR) alone. Our proposed imputation procedure, the all-cases imputation (ACI) method, involves imputing y for respondents as well as nonrespondents and using the imputation error for respondents to estimate the imputation error variance and covariance for nonrespondents.

The all-cases imputation variance estimator for the variance of the population mean is

$$\hat{v}_{ACI}(\bar{y}_I) = \frac{1}{n}\left\{\frac{1}{n-1}\left[\sum_{i \in R}(y_i - \bar{y}_I)^2 + \sum_{i \in NR}\left(y_i^* - \bar{y}_I\right)^2\right]\right\} + \frac{m}{n^2(r-1)}\sum_{i \in R}(\tau_i - \bar{\tau})^2 + \frac{2}{n^2}\left(\frac{m(m-1)}{r(r-1)}\right) \sum_{k \in R}\sum_{\substack{i,j \in R \\ j>i}} I_{k_{ij}}(\tau_i - \bar{\tau})(\tau_j - \bar{\tau}) \qquad (3)$$

where $I_{k_{ij}}$ is equal to 1 if respondent k is used as a donor for both i and j, and is equal to 0 otherwise. In Montaquila (1997), we have shown that under certain general conditions, $\hat{v}_{ACI}(\bar{y}_I)$ is an unbiased estimator of $v(\bar{y}_I)$.

2.2 Stratified Random Sampling

Here, we consider the sample design where the N units in the population are stratified into L strata indexed by h. There are N_h units in stratum h, and we will select n_h units, such that each unit in stratum h has an equal probability of selection (but the probabilities of selection may differ from stratum to stratum). That is, within each stratum, a simple random sample is drawn. The sampling is assumed to be independent from stratum to stratum.

For a stratified random sample design, the complete-data estimator of the population mean per unit is

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad (4)$$

where $W_h = \frac{N_h}{N}$ and $\bar{y}_h = \frac{1}{n_h} = \sum_{i=l}^{n_h} y_{hi.}$.

In the presence of missing data, the imputed estimator of the population mean per unit is:

$$\bar{y}_{st,I} = \sum_{h=1}^{L} W_h \left\{\frac{1}{n_h}\left[\sum_{i \in R_h} y_{hi} + \sum_{i \in NR_h} y_{hi}^*\right]\right\} \qquad (5)$$

where R_h denotes the set of respondents to item y in stratum h, NR_h is the set of nonrespondents to item y in stratum h, and y_{hi}^* is the imputed value for case i in stratum h. The imputation error is defined as:

$$\tau_{hi} = y_{hi}^* - y_{hi} = y_{hk} - y_{hi},$$

where unit k in stratum h is used as a donor for unit i in stratum h.

For stratified random sampling, we propose the same general approach as for simple random sampling, an all-cases imputation approach, to estimating the variance of the imputed estimator. Imputations will be obtained for all cases, both respondents and nonrespondents. Here, the imputations will be selected from among respondents in the same stratum; that is, stratum boundaries cannot be crossed when identifying donors. The imputation error for respondents will be used to estimate the variance components due to imputation error.

The all-cases imputation variance estimator for the mean per unit in a stratified random sample design is

$$\hat{v}_{ACI}(\bar{y}_{st,I}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 s_{h,I}^2 \qquad (6)$$

where

$$s_{h,I}^2 = \frac{N_h - n_h}{N_h} \frac{1}{n_h} \left\{ \frac{1}{n_h - 1} \left[\sum_{i \in R_h} \left(y_{hi} - \bar{y}_{h,I} \right)^2 \right] \right\}$$

$$+ \frac{m_h}{n_h^2 (r_h - 1)} \sum_{i \in R_h} \left(\tau_{hi} - \bar{\tau}_h \right)^2$$

$$+ \frac{2}{n_h^2} \frac{m_h (m_h - 1)}{r_h (r_h - 1)} \sum_{k \in R_h} \sum_{i \in R_h} \sum_{\substack{j \in R_h \\ j > i}} I_{k_{ij}} (\tau_{hi} - \bar{\tau}_h)(\tau_{hj} - \bar{\tau}_h)$$

Within each stratum, the sample is a simple random sample. Furthermore, note that $s_{h,I}^2$ is simply the ACI variance estimator for simple random sampling (with the finite population correction factor), indexed by h. Thus, the result that $E\left(s_{h,I}^2\right) = v\left(\bar{y}_{h,I}\right) \ \forall \ h = 1,2,\ldots,L$ follows directly from the result that the ACI variance estimator for simple random sampling is unbiased. Therefore, this form of the ACI variance estimator for stratified random sampling is also unbiased (under the same general conditions).

2.3 Extensions to Other Sample Designs or Other Estimators

The ACI variance estimator can easily be extended to other sample designs and other estimators. Regardless of the sample design or the form of the estimator, the approach will be to impute for all cases. That is, any data item that requires imputation will be imputed for all cases, such that the imputations preserve features of the sample design (e.g., not crossing stratum boundaries to find donors). The standard complete-data variance estimate will be computed using the actual data whenever possible and the imputed data for cases with missing data. Imputation error variance and covariance components will be estimated using the imputation errors for respondents. For nonlinear estimators, one possible approach would be to first linearize the estimator, and then derive the imputation error variance components. Since the imputation errors would be linearized in the process of linearizing the estimator, these extensions are straightforward.

3. Simulation Study

To evaluate the ACI variance estimator and compare this procedure to the Rao-Shao jackknife and the Shao-Sitter bootstrap, we conducted a simulation study using simple random sampling. We generated samples of size n=300 from a prespecified distribution F. In each sample, we designated $100*(m/n)$ percent of the cases as missing. We imputed for the m missing cases using the r respondents as the donor set, using the random hot-deck procedure. We also imputed values for each of the r respondents, using the other respondents as the donor set. We then computed $\bar{y}_I$ and $\hat{v}_{ACI}(\bar{y}_I)$ for each sample.

The process described above was repeated for 500 iterations. For each iteration, the normal 95% confidence intervals were computed. The coverage rates across the 500 iterations were then computed. For comparison purposes, variance estimates and coverage rates were also computed using the Rao-Shao jackknife and the Shao-Sitter bootstrap procedure.

The variance in the 500 Monte Carlo estimates, $v_M(\bar{y}_I) = \frac{1}{M} \sum_{i=1}^{M} \left(\bar{y} - \bar{\bar{y}}_I\right)^2$, was computed. Although this measure is subject to a small amount of sampling error, it was assumed to be the "truth." That is, $v_M(\bar{y}_I)$ was assumed to be equal to $v(\bar{y}_I)$.

As we stated previously, treating imputed values as if they had actually been observed or reported leads to underestimation of the variance of the estimator. This may result in confidence interval coverage rates which are far below the nominal levels. In Table 1, the mean variance estimates and confidence interval coverage rates based on $\hat{v}_{ACI}(\bar{y}_I)$ are compared to those based on $\hat{v}_{SAMP}$, the sampling error variance component of the ACI variance estimator. Note that $\hat{v}_{SAMP}$ is the naive variance estimate which treats the imputed values as if they had actually been observed or reported. Table 1 shows that even with small proportions of missing data, the naive variance estimator leads to serious underestimation of the variance and poor confidence interval coverage.

The distributions of variance estimates generated using each method were compared to the Monte Carlo variance $v_M(\bar{y}_I)$. Table 1 also presents a comparison of the Monte Carlo variance to the mean variance estimate obtained using each of the three approaches. The three variance estimates tend to be very similar, and there is no distinct pattern among the three. As the proportion of missing data increases, all three variance estimates increase, as does the Monte Carlo variance.

Rubin (1996) emphasizes the fact that rarely is the variance of an estimator itself an estimand. That is, rarely is the sole purpose to estimate the variance of an estimator. Rather, the goal is to obtain valid inferences. The variance estimator is merely a vehicle used *en route* to obtaining valid inferences. Thus, in comparing methods, it is important to assess the validity of inferences obtained using the methods.

Table 1 also compares the coverage rates for the nominal 95% confidence intervals obtained using the variance estimate from each approach. The Shao-Sitter bootstrap procedure tends to yield confidence intervals with higher coverage rates than those obtained using the other two procedures. However, the coverage rates obtained using all three approaches are very good; that is, very close to the nominal 95% coverage rate. Although there is a slight drop in coverage rates for data sampled from skewed distributions, even with 70% missing data, the drop is not substantial.

Next, we examine the components of the ACI variance estimator. Recall that we refer to the first term in the ACI variance estimator, $v(\bar{y}_S)$, as the sampling error variance component; we refer to the second term, $\frac{1}{n^2}\sum_{i\in NR} v\left(\tau_i^{(k)}\right)$, as the imputation error variance component; and we refer to the third term, $\frac{2}{n^2}\sum_{i\in NR}\sum_{\substack{j\in NR\\ j>i}} \operatorname{cov}\left(\tau_i^{(k)},\tau_j^{(l)}\right)$, as the imputation error covariance component. The mean estimates of each variance component of $\hat{v}_{ACI}(\bar{y}_I)$ are given in Table 2. The sampling error variance component, $\hat{v}_{SAMP}$, remains essentially fixed as the proportion of missing data increases, while the imputation error variance component, $\hat{v}_{IMP}$, and the imputation error covariance component, $\hat{c}_{IMP}$, increase as the proportion of missing data increases.

4. Summary and Conclusions

We have presented a new approach to variance estimation for imputed data, the all-cases imputation variance estimator. In this paper, we have developed this approach for the special case where the sampled cases are chosen using simple random sampling and the imputation method is the random hot-deck, and have described the extensions of the variance estimator to stratified, multistage designs. The extensions from simple random sampling to more complex sample designs are straightforward, given the extensions of the complete-data variance estimator from simple random sampling to complex designs.

Our approach, all-cases imputation, is a model-assisted method in which the variance of the estimator is decomposed into components which reflect the sampling error variance, the imputation error variance, and the imputation error covariance. Since imputation errors--the differences between the actual (but unobserved) value and the imputed value--are unknown for item nonrespondents, we use the imputation errors for the item respondents to estimate the variance components involving imputation errors.

We have empirically compared our method to two other proposed methods--the Rao-Shao jackknife and the Shao-Sitter bootstrap--as well as the naive approach. We have demonstrated, both analytically and empirically, that the naive approach to variance estimation--treating the imputed values as if they had actually been observed or reported--underestimates the variance of the estimator. The downward bias is substantial, even with small proportions of missing data. The ACI variance estimator has been shown to yield unbiased variance estimates and randomization-valid confidence intervals for the problem of estimating the population mean.

An advantage of the ACI variance estimator over the Rao-Shao variance estimator is that the ACI variance estimator may be directly extended to situations where imputation methods other than the random hot-deck are used. For example, the nearest neighbor and regression imputation methods are commonly used in practice, since sampling frames tend to have an abundancy of auxiliary data that may be correlated with the characteristic of interest or the response propensity. The ACI variance estimator, as presented here, can be directly applied to situations where nearest neighbor or regression imputation are used; the Rao-Shao jackknife cannot.

The ACI variance estimator is not as computationally intensive as the bootstrap approach. With the ACI procedure, one imputation is generated for each case, and one variance calculation--with three components--is required. The bootstrap involves drawing numerous bootstrap samples, imputing independently within each bootstrap sample, and computing an estimate for each bootstrap sample. To properly implement this procedure and ensure its validity for a specific problem, the validity checks used in full-sample imputation must be used for each bootstrap sample. This can be quite time-consuming and labor-intensive. Complicated skip patterns which are sometimes present in survey instruments will further muddle this procedure.

We believe the ACI variance estimator has more intuitive appeal than the model-assisted approach developed by Rancourt *et al.* Each component of the ACI variance estimator has a very clear interpretation. Furthermore, the extensions of the ACI variance estimator to complex sample designs, to other estimators, and to other imputation methods are straightforward.

5. Directions for Future Research

There are many extensions of the ACI variance estimator that we wish to pursue in the future. These include:

- The extension to stratified, multistage sample designs. We have presented our proposed approach for stratified, multistage designs; an empirical study will enable us to evaluate the performance of the ACI variance estimator for such designs.
- An empirical evaluation of the ACI variance estimator when imputation methods other than the random hot-deck are used. We expect that the ACI variance estimator will be directly applicable, without any changes in its form, to such situations.
- The extension of the ACI variance estimator to imputation of multivariate data.
- Extensions of the ACI variance estimator to the problem of estimating quantities other than population means and totals; for example, extensions to ratio estimators or estimators of population quantiles.

References

Cochran, W.G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.

Fay, R.E. (1992), "When are Inferences from Multiple Imputation Valid?," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 227-232.

Fay, R.E. (1993), "Valid Inferences from Imputed Survey Data," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 41-48.

Fay, R.E. (1994), "Analyzing Imputed Survey Data Sets with Model-Assisted Estimators," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 900-905.

Fay, R.E. (1996a), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91, 490-498.

Fay, R.E. (1996b), "Replication-Based Variance Estimators for Imputed Survey Data from Finite Populations," draft manuscript.

Kalton, G., and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1-16.

Kott, P. (1995), "A Paradox of Multiple Imputation," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 380-383.

Lee, H., Rancourt, E., and Särndal, C.E. (1995), "Variance Estimation in the Presence of Imputed Data for the Generalized Estimation System," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 384-389.

Montaquila, J. (1997), "A New Approach to Variance Estimation in the Presence of Imputed Data," Ph.D. dissertation, American University.

Rancourt, E., Särndal, C.E., and Lee, H. (1994), "Estimation of the Variance in the Presence of Nearest Neighbour Imputation," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 888-893.

Rao, J.N.K. (1993), "Jackknife Variance Estimation with Imputed Survey Data," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 31-40.

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.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20-34.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D.B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473-489.

Shao, J., and Sitter, R.R. (1996), "Bootstrap for Imputed Survey Data," Technical Report 227, Carleton University, Laboratory for Research in Statistics and Probability.

Steel, P., and Fay, R.E. (1995), "Variance Estimation for Finite Populations with Imputed Data," in *Proceedings of the Survey Research Methods Section, American Statistical Association* , pp. 374-379.

Table 1. An empirical comparison of the ACI variance estimator to three alternative approaches

F	$^{m}/_{n}$	$100{*}v_M(\bar{y}_I)$	100*(Mean variance estimate)				Conf. interval coverage rate (%)			
			Naive	ACI	RSB[1]	SSB[2]	Naive	ACI	RSB[1]	SSB[2]
Normal(0,1)	0.1	0.39	0.33	0.40	0.40	0.40	91.8	93.2	93.4	94.8
	0.3	0.60	0.33	0.57	0.58	0.57	86.0	94.0	94.4	95.4
	0.5	0.83	0.33	0.83	0.83	0.82	79.8	95.0	95.2	97.0
	0.7	1.22	0.33	1.32	1.35	1.33	70.6	95.0	95.8	96.2
Poisson(λ=3)	0.1	1.12	1.00	1.21	1.21	1.20	92.8	95.2	95.0	96.2
	0.3	1.71	1.00	1.73	1.73	1.72	85.4	96.2	96.4	97.8
	0.5	2.41	1.00	2.48	2.49	2.46	79.0	94.4	94.8	95.0
	0.7	4.17	1.00	4.01	4.09	4.01	66.2	93.8	93.8	95.4
Gamma(α=1,β=1)	0.1	1.32	1.00	1.21	1.21	1.20	91.2	93.8	93.8	95.2
	0.3	1.63	1.01	1.74	1.74	1.71	87.6	94.6	95.2	96.2
	0.5	2.26	1.00	2.51	2.53	2.50	81.6	96.2	96.0	98.0
	0.7	4.31	0.99	3.99	4.04	3.95	63.2	92.6	94.0	94.4
Lognormal(0,1)	0.1	1.81	1.64	1.97	1.99	1.96	92.4	95.0	94.6	94.0
	0.3	2.73	1.52	2.61	2.64	2.58	82.4	90.4	90.8	92.8
	0.5	3.73	1.50	3.85	3.81	3.83	77.4	91.8	91.6	94.4
	0.7	6.05	1.52	5.95	6.36	6.24	60.8	89.8	92.0	93.6

[1] Rao-Shao jackknife

[2] Shao-Sitter bootstrap

Table 2. Mean estimates of variance components in the ACI variance estimator

F	$^{m}/_{n}$	Variance component‡ [100*(mean estimate)]		
		$\hat{v}_{SAMP}$	$\hat{v}_{IMP}$	$\hat{c}_{iMP}$
Normal(0,1)	0.1	0.33	0.07	0.00
	0.3	0.33	0.20	0.04
	0.5	0.33	0.34	0.16
	0.7	0.33	0.47	0.53
Poisson(λ=3)	0.1	1.00	0.20	0.01
	0.3	1.00	0.60	0.12
	0.5	0.99	1.00	0.49
	0.7	1.00	1.41	1.60
Gamma(α=1,β=1)	0.1	1.00	0.20	0.01
	0.3	1.01	0.60	0.12
	0.5	1.00	1.01	0.50
	0.7	0.99	1.41	1.59
Lognormal(0,1)	0.1	1.64	0.31	0.01
	0.3	1.52	0.91	0.18
	0.5	1.50	1.55	0.80
	0.7	1.52	2.21	2.22

‡ For notational convenience, $\hat{v}_{SAMP}$, $\hat{v}_{IMP}$, and $\hat{c}_{iMP}$ are used to denote the sampling error variance, imputation error variance, and imputation error covariance components of the ACI variance estimator.