INCORPORATING THE FINITE POPULATION CORRECTION INTO VARIANCE ESTIMATES FROM IMPUTED DATA

Hyunshik Lee and Jill M. Montaquila, Westat
Jill M. Montaquila, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Single imputation, jackknife, bootstrap, without replacement sampling

1. Introduction

The finite population correction (fpc) is used to adjust a variance estimator when the sample is selected without replacement from a finite population. Incorporation of the fpc into the resampling variance estimators such as the jackknife and the bootstrap becomes more complex in the presence of imputation. When the fpc is non-negligible, these variance estimators without the fpc-correction can appreciably overestimate the variance. On the other hand, a naïve application of the fpc can cause an appreciable underestimation. This paper addresses this problem.

The main variance components of estimators based on imputed data are the sampling variance and imputation variance, which is due to imputation. The standard variance estimators traditionally used underestimate the total variance in the presence of imputation by treating imputed values as observed. Many remedies to the problem have been proposed. An early starter is multiple imputation proposed by Rubin (1987). For single imputation, traditional variance estimation techniques have been modified to address the problem (see Lee, Rancourt, and Särndal, 2000, and Shao, 2000 for review).

Methods based on resampling techniques for imputed data do not provide the two variance components separately and this makes the application of the fpc difficult since it has to be applied only to the sampling variance component.

In the case of the jackknife, there are some solutions to this problem. For example, for simple random sampling with uniform response mechanism, Lee, Rancourt, and Särndal (1995) considered a simple solution, where the sampling variance is separately estimated using the respondents only so that the fpc is properly applied.

Rao and Sitter (1995) considered linearized jackknife estimator assuming the set of respondents is a second phase sample from the full sample. Once the variance is linearized, the fpc can be easily incorporated.

Steel and Fay (1995) studied another solution based on the adjusted jackknife variance estimator of Rao and Shao (1992).

Shao and Sitter (1996) proposed a bootstrap variance estimator that can be applied to imputed data. Unlike the adjusted jackknife, which overestimates the variance, the Shao-Sitter bootstrap estimator underestimates the variance. As far as we know, no solution has been proposed for this problem.

The all-cases imputation (ACI) method (Montaquila and Jernigan 1997) involves imputing for respondents as well as nonrespondents, then using the differences between imputed and actual values for respondents to estimate the imputation error variance component. Since the ACI variance estimator contains separate terms for the sampling and imputation error variance components, the fpc may be easily incorporated by applying it to only the sampling error variance component.

In this article, we propose some solutions to address the fpc problem. We will discuss our proposed new methods in detail in section 2. We evaluated the new methods using a simulation study, of which the results are presented in section 3. In section 4, we give some concluding remarks

2. New Approaches to the Problem

To introduce the idea, let's assume a simple random sample design, where a sample, *s*, of size *n* is selected by simple random sampling without replacement (SRSWOR) from a universe of size *N*. The variable of interest is denoted by *y* and indexed as y_k to denote the value of the variable for the *k*-th unit. The set of *m* respondents is denoted by *r*. An auxiliary variable *x* is observed for all units in *s*. The parameter of interest is the population mean and it is estimated by the sample mean. Let $D^* = \{y_k^* \mid k \in s)$ be imputed data set where $y_k^* = y_k$ for $k \in r$ and $y_k^* = \hat{y}_k$ is the imputed value for $k \in s - r$. If ratio imputation is used, then $\hat{y}_k = \hat{\beta} x_k$, $k \in s - r$ with $\hat{\beta} = \sum_r y_k / \sum_r x_k$. A random ratio imputation is achieved by adding a randomly selected residual $\left(\hat{e}_k^*\right)$ to $\hat{\beta} x_k$ where (observed) residuals are defined by $\hat{e}_k = y_k - \hat{\beta} x_k$. Setting $x_k = 1$ for all $k \in s$, we obtain mean imputation or hot deck imputation.

After imputation, the population mean is then estimated by

$$\bar{y}_s^* = \frac{1}{n} \sum_{k \in s} y_k^* .$$

Ignoring the fpc, the naïve jackknife variance estimator applied to D^* is given by

$$v_J = \frac{n-1}{n} \sum_{j=1}^{n} (\bar{y}_s^*(-j) - \bar{y}_s^*)^2$$

where $\bar{y}_s^*(-j)$ is calculated using the reduced data set with j-th unit deleted. This variance estimator underestimates the total variance by missing the imputation variance component completely.

To correct the underestimation, Rao and Shao (1992) proposed to use an adjustment to the imputed data for hot deck imputation. The adjustment is applied only to the imputed values depending on whether the deleted unit for the jackknife estimation is a respondent or not. The basic principle is that when the deleted unit is a respondent, the imputed value should be modified to reflect the effect of the reduced data set without the respondent to the imputation procedure. Reimputation was tried to incorporate this effect (Burns, 1990) for hot deck imputation but Rao and Shao (1992) proved that the resulting jackknife variance estimator overestimates the variance and proposed the adjustment approach.

The adjustment proposed by Rao and Shao (1992) is defined as follows:

$$y_k^{*(a)}(j) = \begin{cases} y_k^* & \text{if } j \in s - r \\ y_k^* + E'(y_k^*) - E(y_k^*) & \text{if } j \in r \end{cases}$$

where E is the expectation under the imputation procedure with the full data set and E' is the same expectation but with the j-th unit deleted. The adjusted jackknife variance estimator, denoted by $v_J^{(a)}$, is obtained by applying the jackknife variance estimator to the adjusted values. This variance estimator is valid if the sampling fraction is negligible. However, if the sampling fraction is appreciable, the adjusted jackknife variance estimator overestimates the variance of the imputed estimator, $\bar{y}^*$. On the other hand, if the fpc is naively applied (i.e., $(1-f)v_J^{(a)}$), then the variance is underestimated because the fpc is applied not only to the sampling variance but also to the imputation variance, which does not need the correction.

Our approach to handle this problem is to first estimate the sampling variance by the ordinary jackknife variance estimator, v_J, applied to unadjusted data set D^*. The imputation variance is then estimated by $v_J^{(a)} - v_J$, and is set to be zero if $v_J^{(a)} - v_J$ is negative. Applying the fpc to the sampling variance component and adding the imputation variance component to it, we obtain the following new fpc-corrected adjusted jackknife variance estimator for imputed data:

$$v_J^{(ca)} = \begin{cases} v_J^{(a)} - f v_J & \text{if } v_J^{(a)} - v_J \ge 0 \\ (1-f)v_J & \text{otherwise} \end{cases}$$

Turning to the bootstrap method, the basic principle of the modification Shao and Sitter (1996) proposed for imputed data is to follow the spirit of the bootstrap method, where the sampling and response behaviors are mimicked through simulation. To incorporate the imputation variance the modified bootstrap uses reimputation to replace the nonrespondents' original imputed values in the bootstrap sample using the respondents in the bootstrap sample. The bootstrap sampling for survey data that provides the basic framework for the Shao-Sitter proposal is done to mimic the finite population sampling and so it is supposed to incorporate the fpc in the bootstrap sampling procedure. When this finite population bootstrap sampling is modified for imputed data, the finite population sampling feature is carried over and thus, the fpc is automatically incorporated if present. The variance estimate obtained from the modified bootstrap method, however, include the imputation variance and it gets corrected unnecessarily, which results in underestimation of the variance. This was also noticed by Lee, Rancourt, and Särndal (2000).

In our solution to this problem, we assume that the ordinary bootstrap variance estimator (without the Shao-Sitter modification) estimates the sampling variance correctly. This is usually the case for stochastic imputation methods such as hot-deck and random ratio imputation.

Let v_{OB} be the ordinary bootstrap variance estimator applied to imputed data and let v_{SSB} be the Shao-Sitter modified bootstrap variance estimator, which contains the imputation variance (let this be denoted by $v_{\text{B-IM}}$ although it is not estimated in the procedure). The above discussion can be summarized by the following equation:

$$v_{\text{SSB}} = v_{\text{OB}} + (1-f)v_{\text{B-IM}} \cdot$$

The unnecessary multiplication of $(1-f)$ to $v_{\text{B-IM}}$ causes the underestimation. Then the correct variance is given as $v_{\text{OB}} + v_{\text{B-IM}}$. However, the imputation variance is not directly estimated. Our solution is to obtain an estimate of the imputation variance using v_{OB} and v_{SSB} by algebraic manipulation from the above equation, namely, $v_{\text{B-IM}} = (1-f)^{-1}(v_{\text{SSB}} - v_{\text{OB}})$. Inserting this in the correct formula, we get

$$\begin{aligned} v_{\mathrm{SSB}}^{(c)} &= v_{\mathrm{OB}} + (1-f)^{-1}(v_{\mathrm{SSB}} - v_{\mathrm{OB}}) \\ &= (1-f)^{-1}(v_{\mathrm{SSB}} - f v_{\mathrm{OB}}). \end{aligned}$$

Alternatively, a bootstrap sampling method that is used for a with-replacement sample may be used to obtain an fpc corrected modified bootstrap variance estimator for imputed data. We will use primed symbols to denote the bootstrap sampling quantities (for the with-replacement sample) that correspond to the bootstrap sampling quantities (v_{OB} and v_{SSB} for the without-replacement sample). Then an fpc-corrected variance estimator is obtained by

$$v_{\mathrm{SSB}}'^{(c)} = v_{\mathrm{SSB}}' - f v_{\mathrm{OB}}'.$$

However, since the imputation variance is estimated by the difference of two variance estimates, it can be negative. In this case, it would be more sensible to insert zero for $v_{\text{B-IM}}$ in the above derivations. Consequently, we recommend using the following variance estimator:

$$\begin{aligned} v_{\mathrm{SSB}}^{(c)} &= \frac{v_{\mathrm{SSB}} - f v_{\mathrm{OB}}}{1-f} \quad \text{if } v_{\mathrm{SSB}} - v_{\mathrm{OB}} \geq 0 \\ &= v_{\mathrm{OB}} \quad \text{otherwise.} \end{aligned}$$

or

$$\begin{aligned} v_{\mathrm{SSB}}'^{(c)} &= v_{\mathrm{SSB}}' - f v_{\mathrm{OB}}' \quad \text{if } v_{\mathrm{SSB}}' - v_{\mathrm{OB}}' \geq 0 \\ &= (1-f) v_{\mathrm{OB}}' \quad \text{otherwise.} \end{aligned}$$

Note that the latter formula is in the same form as for the corrected jackknife formula. We used this formula in the simulation.

The above discussion is readily applicable for a stratified simple random sampling design if imputation is carried out stratum by stratum. Note that resampling variance estimation methods such as the jackknife and the bootstrap are not directly applicable for the case of census and in the discussion in this section, we tacitly assume that the sampling fraction is less than 1.

3. Simulation Study

In order to evaluate the approach described in section 2, a Monte Carlo simulation study was undertaken. For the simulation study, two finite populations of size $N = 300$ were generated. The first population was generated by sampling from a normal distribution, and the second population was generated by sampling from a lognormal distribution. In the first population, the variable y has a population mean of 99.2 and population standard deviation of 19.1. In the second population, the population mean and standard deviation of y are 113.2 and 182.4, respectively.

Next, for each iteration of the simulation, a simple random sample was selected. Different combinations of the sampling fractions for SRSWOR and item response rates with the uniform response mechanism. The sampling fractions considered in this simulation were 1/3, 2/3, and 1, and the item nonresponse rates were ¼, ½, and ¾. For item nonrespondents, y was imputed using a single-cell random hot-deck. The completed dataset (i.e., the full dataset containing the actual values of y for item respondents and the imputed values of y for item nonrespondents) was used to obtain an estimate of the overall mean of y, and the approach described in section 2 was used to obtain estimates of the variance of the mean using the fpc-corrected ACI, jackknife $\left(v_J^{(ca)}\right)$, and bootstrap $\left(v_{\mathrm{SSB}}'^{(c)}\right)$ variance estimators.

For each combination of sampling fraction and item nonresponse rate, this process was repeated for 1,000 iterations. For each of the fpc-corrected variance estimators, the mean variance estimate was computed and compared to the Monte Carlo variance in the estimates across the 1,000 iterations. Additionally, the confidence interval coverage rates were computed for nominal 95 percent confidence intervals.

Table 1 gives a comparison of the mean uncorrected variance estimates to the Monte Carlo variance. The results are very similar across the three methods. The ratios in this table demonstrate that using variance estimates for imputed data without corrections for without-replacement finite population sampling may result in substantial overestimation of the variance.

In table 2, the mean fpc-corrected variance estimates are compared to the Monte Carlo variance. Again, the ratios are close to 1 for most of the cases and very similar across the three methods studied. This table demonstrates substantial reduction in the bias of the imputation variance estimates when the fpc correction described in Section 2 is applied.

Tables 3 and 4 give the confidence interval coverage rates for nominal 95 percent confidence intervals constructed using the uncorrected and fpc-corrected variance estimates, respectively. As depicted in table 3, confidence intervals based on the uncorrected variance estimates tend to have coverage rates that are higher than the nominal rate, due to the overestimation of the variance. For the case of sampling from the normal population, Table 4 demonstrates that when the fpc-corrected variance estimators are used, the confidence intervals tend to have coverage rates that are much closer to the nominal rate. This pattern in the confidence interval coverage rates is not clear for the simulations involving the population generated by sampling from the lognormal distribution. However, we believe that this is not indicative of a problem with the fpc-corrected variance estimators, but rather due to the failure of asymptotic results with small sample sizes from such a skewed distribution.

4. Summary and Conclusions

In the case of finite population sampling, variance estimators that have been developed for imputed data must be adapted to account for the fact that the sample was drawn without-replacement from a finite population if that is the case. When the variance estimator contains separate terms for the sampling error and imputation error variance components, the adaptation may involve simply applying the fpc to the estimate of the sampling error component. However, when these components are not explicitly estimated, this approach is not feasible and other alternatives must be explored.

We have described and evaluated three fpc-corrected variance estimators for imputed data in the context of simple random sampling with single-cell hot-deck imputation (with direct extensions to stratified random sampling). The first, based on the all-cases imputation approach, involves adapting a variance estimator that contains separate terms for the sampling and imputation error variance components, as described above. The other two approaches involve using different jackknife and bootstrap variance estimators to estimate variances that reflect the imputation error component and variances that do not reflect the imputation error component; estimating the sampling error component by subtraction; and applying the fpc to only the sampling error component. Simulation results demonstrate substantial improvement in the variance estimates (compared with variance estimates that do not reflect the fpc) and in confidence interval coverage rates.

5. References

Burns, R.M. (1990). Multiple and replicate item imputation in a complex sample survey. *Proceedings of the Sixth Annual Research Conference*, pp. 655-665. Washington, D.C.: U.S. Bureau of the Census.

Lee, H., Rancourt, E., Särndal, E.R. (1995). Jackknife Variance Estimation for Data with Imputed Values. *Proceedings of Survey Methods Section*, Statistical Society of Canada, pp. 11-116.

Lee, H., Rancourt, E., Särndal, E.R. (2000). Chapter 21. Variance Estimation from Survey Data under Single Imputation, in *Survey Nonresponse*, New York: Wiley (to be published).

Montaquila, J.M. and Jernigan, R.W. (1997). Variance Estimation in the Presence of Imputed Data. *Proceedings of Survey Research Methods Section*, American Statistical Association, pp. 273-278.

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

Rao, J.N.K., Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, vol. 82, pp. 453-460.

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

Shao, J. (2000). Chapter 20: Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In *Survey Nonresponse*, New York: Wiley (to be published).

Shao, J. and Sitter, R. (1996). Bootstrap for imputed survey data. *Journal of American Statistical Association*, 91, 1278-1288.

Steel, P., and Fay, R. (1995). Variance Estimation for Finite Populations with Imputed Data. *Proceedings of Survey Research Methods Section*, American Statistical Association, pp. 111-116.

Table 1. Comparison of uncorrected variance estimates from simulation study to Monte Carlo variance

Population distribution	Sampling fraction (percent)	Item nonresponse rate (percent)	Ratio of mean uncorrected variance estimate to Monte Carlo variance		
			ACI	Rao-Shao jackknife	Shao-Sitter bootstrap
Normal	33	25	1.29	1.30	1.29
	33	50	1.14	1.17	1.16
	33	75	1.08	1.19	1.16
	67	25	1.70	1.70	1.69
	67	50	1.32	1.33	1.32
	67	75	1.12	1.16	1.14
	100	25	2.84	2.84	2.83
	100	50	1.85	1.86	1.86
	100	75	1.28	1.30	1.29
Lognormal	33	25	1.29	1.29	1.29
	33	50	1.03	1.05	1.03
	33	75	0.92	1.03	1.00
	67	25	1.76	1.76	1.76
	67	50	1.43	1.42	1.40
	67	75*	1.12	1.23	1.22
	100	25	2.81	2.83	2.82
	100	50	1.71	1.72	1.71
	100	75	1.28	1.30	1.28

*Results for this combination of parameters are based on 700 iterations (rather than 1,000).

Table 2. Comparison of fpc-corrected variance estimates from simulation study to Monte Carlo variance

Population distribution	Sampling fraction (percent)	Item nonresponse rate (percent)	Ratio of mean fpc-corrected variance estimate to Monte Carlo variance		
			fpc-corrected ACI	fpc-corrected Rao-Shao jackknife	fpc-corrected Shao-Sitter bootstrap
Normal	33	25	1.02	1.03	1.02
	33	50	0.98	1.02	1.00
	33	75	1.00	1.11	1.08
	67	25	0.98	0.99	0.98
	67	50	0.97	0.98	0.97
	67	75	0.96	1.00	0.98
	100	25	1.05	1.05	1.04
	100	50	1.12	1.12	1.12
	100	75	1.02	1.03	1.02
Lognormal	33	25	1.02	1.02	1.02
	33	50	0.89	0.91	0.89
	33	75	0.85	0.96	0.93
	67	25	1.02	1.02	1.02
	67	50	1.05	1.05	1.02
	67	75*	0.95	1.06	1.05
	100	25	1.03	1.05	1.05
	100	50	1.03	1.04	1.04
	100	75	1.01	1.03	1.02

*Results for this combination of parameters are based on 700 iterations (rather than 1,000).

Table 3. Confidence interval coverage rates (nominal 95% confidence intervals) from simulation study based on uncorrected variance estimates

Population distribution	Sampling fraction (percent)	Item nonresponse rate (percent)	Confidence interval coverage rate (nominal 95% confidence interval)		
			ACI	Rao-Shao jackknife	Shao-Sitter bootstrap
Normal	33	25	97.7	97.5	98.3
	33	50	95.7	96.9	98.4
	33	75	92.7	96.3	97.3
	67	25	98.9	98.9	99.7
	67	50	97.1	97.3	98.6
	67	75	94.5	96.4	97.5
	100	25	100.0	100.0	100.0
	100	50	98.9	99.2	100.0
	100	75	95.5	97.5	99.1
Lognormal	33	25	92.7	93.4	93.6
	33	50	88.9	89.1	90.3
	33	75	80.9	84.6	86.0
	67	25	97.4	97.4	98.9
	67	50	95.0	95.1	96.9
	67	75*	86.7	90.9	91.1
	100	25	99.6	99.6	99.7
	100	50	97.7	97.2	98.5
	100	75	91.8	92.8	93.9

*Results for this combination of parameters are based on 700 iterations (rather than 1,000).

Table 4. Confidence interval coverage rates (nominal 95% confidence intervals) from simulation study based on fpc-corrected variance estimates

Population distribution	Sampling fraction (percent)	Item nonresponse rate (percent)	Confidence interval coverage rate (nominal 95% confidence interval)		
			fpc-corrected ACI	fpc-corrected Rao-Shao jackknife	fpc-corrected Shao-Sitter bootstrap
Normal	33	25	95.7	95.7	97.1
	33	50	93.5	95.2	97.9
	33	75	91.3	96.0	97.1
	67	25	95.2	95.4	97.6
	67	50	93.6	94.2	96.9
	67	75	92.8	95.2	96.7
	100	25	95.7	95.8	98.6
	100	50	96.1	97.0	98.3
	100	75	92.3	95.4	96.9
Lognormal	33	25	90.4	91.2	91.7
	33	50	86.4	87.6	88.9
	33	75	78.4	84.0	85.6
	67	25	92.7	93.0	96.6
	67	50	90.6	92.8	94.0
	67	75*	82.7	88.7	89.3
	100	25	93.7	93.6	96.7
	100	50	91.9	94.3	96.2
	100	75	86.1	91.0	92.0

*Results for this combination of parameters are based on 700 iterations (rather than 1,000).