VARIANCE ESTIMATION FOR FINITE POPULATIONS WITH IMPUTED DATA

Philip Steel, Robert E. Fay, U.S. Bureau of the Census[1]
Philip Steel, U.S. Bureau of the Census, Washington, DC 20233

KEY WORDS: *Monte Carlo, Rao-Shao, Imputation, Jackknife*

Abstract One way of handling survey nonresponse is to impute data for each nonrespondent. When estimating sampling variances, however, treating the imputed data as a complete set frequently leads to underestimates of the true sampling variance. Techniques have been recently developed to yield valid variance estimates in the presence of imputed data for some estimators and sample designs.

Economic surveys frequently deal with highly skewed populations and employ high sampling rates, including selection with certainty for large units. It is also common for economic surveys to have administrative or historical data available for use in imputation. This paper describes a Monte Carlo study of variance estimation on skewed populations with high sampling fractions in some strata. We examine a variety of imputation techniques and patterns of nonresponse. We extend the Rao-Shao technique to finite populations and nearest neighbor imputation, and compare the resulting estimators to the true variance.

1. INTRODUCTION

Economic surveys and censuses conducted by the U.S. Census Bureau frequently employ imputation to supply missing values. Recently, King and Kornbau (1994) surveyed statistical practices in the Economic Area of the Census Bureau. They found imputation in widespread use for treatment of missing data. Hot deck, mean, and ratio imputation are among the frequently employed methods, as well as procedures that produce an estimate from an external source or adjust a previous value. With the exception of a few special studies, variance estimates have not attempted to reflect the effect of imputation on the reliability of the estimates.

This paper attempts to examine variance estimation for imputed data sets in an economic context. In general, economic surveys and censuses confront populations that are quite skewed, with relatively few members of the population contributing a substantial fraction of the overall total. The sample designs are typically stratified by estimated size and other characteristics, with typically the largest units included with certainty. Recent advances in variance estimation for imputed data, such as the Rao-Shao (1992) results for hot-deck imputation, are not immediately applicable to such survey and census designs.

In an economic setting, the statistical agency frequently has past data or administrative data helpful for making imputations for missing data. Often, several variations on an imputation method may be applied in the same survey to exploit whatever external information may be available.

Lee, Rancourt, and Särndal (1994) recently compared variance estimators for ratio and nearest neighbor imputation in a finite population context. Each of the estimators was an analytic expression. We propose to examine properties of replication-based variance estimators in this paper. Section 2 describes the creation of an artificial population modeled to have many of the characteristics of an economic population. By randomly generating the observations, we have created a population that we may share with other researchers without risk of disclosure of confidential data. In section 3 we describe an extension of the Rao-Shao (1992) variance estimator for hot deck imputation and ratio imputation (Rao and Sitter 1995) to ratio imputation for finite populations, including when some observations are sampled with certainty but possibly subject to nonresponse. Section 4 describes a Monte Carlo study of the performance of the variance estimator on the population under a variety of nonresponse conditions.

2. AN ARTIFICIAL ECONOMIC POPULATION

Our goal was to have a randomly generated population that mimicked the behavior of economic data. The majority of economic surveys begin with a frame that includes information that is well correlated with at least some of the variables of interest. In absence of response or historical information, the frame information can be used to impute response values. Payroll and receipts are typical: payroll is known administratively and receipts is the variable of interest. Using the payroll variable, a ratio estimate may be made, a nearest neighbor's value may be used, or a hot deck class created to obtain an imputation for receipts. We chose as our starting point economic census data containing payroll and receipts and randomly generated a similar population.

We focused first on generating an artificial payroll population. A brief survey of 1992 census industries was conducted; the distribution of payroll appeared to be lognormal or a mixture of lognormals in almost all cases. The best of these, whose departure from the normal under the log transformation was not significant ($p > 0.15$ for the Kolmogorov D statistic), was the basis of our x population. We fit the transformed data to obtain parameters for a random normal population ($\mu=5.41$ $\sigma=1.93$), then transformed the artificial population back.

We then modeled the relationship in the Census data between x (payroll) and y (receipts). The data is heteroscedastic, which in Knaub (1993) is modeled as:

$$y = \alpha + \beta x + x^{\gamma} \epsilon$$

Unfortunately it also exhibits more or less constant error in the initial segment (by payroll) of the data. This is not captured in the heteroscedastic model. To account for this and because of its amenable character we used an exponential form for the x-dependent component of the error term:

$$y = \alpha + \beta x + \gamma_0 e^{\gamma_1 x^g} \epsilon$$

Roughly, γ_0 describes the variability when x is small, γ_1 and g control the onset and magnitude of x's influence on the variance. Application of our model to the census data under a reweighting scheme converged nicely to $\alpha=861.04$, $\beta=15.5842$, $\gamma_0=0.091$, $\gamma_1=0.008329$ with $g=0.05$. The studentized residuals show a gamma-like distribution and we have incorporated that into our model. To obtain the artificial y population, we applied the model to the xs already generated, with random gamma errors.

A population of 1500 was then stratified on x by the Lavallée and Hidiroglou method (Lavallée and Hidiroglou 1988), using a Census Bureau stratification package (Sweet and Sigman 1995). This stratification consisted of four noncertainty and one certainty strata. The target CV was 0.05 and the resulting sampling fractions for the noncertainty were 0.0632, 0.1689, 0.2732, and 0.6364, with a fifth certainty stratum. The sample sizes were 51, 64, 53, 63, and 21, in the five strata, respectively.

3. EXTENDING THE RAO-SHAO VARIANCE ESTIMATOR TO FINITE POPULATIONS

We follow the notation of Lee, Rancourt, and Särndal (1994) and Rancourt, Särndal, and Lee (1994) closely. Let $\bar{y}_U = (1/N)\sum_U y_k$ be the mean of the finite population $U = \{1,\ldots,k,\ldots,N\}$. Suppose that a simple random sample, s, of size n is drawn without replacement from U to estimate $\bar{y}_U$. Let r denote the set of m respondents and $s - r$ the set of $n - m$ nonrespondents in the sample. An auxiliary variable, x, is observed for all sample cases. For each nonrespondent, $k \in s - r$, an imputation, $y_{.k}$, is made. The imputations are then included in the estimation of the overall mean or total as if they had been observed.

We assume that response is unconfounded with the missing information. In other words, we consider different alternative relationships between the probability of nonresponse and the auxiliary variable, x, but do not consider the probability of nonresponse to depend on y given x.

Ratio imputation exploits a linear relationship between y_k and x_k through:

$$\begin{aligned} y_{.k} &= y_k, \quad \text{if } k \in r \\ &= \hat{\beta} x_k, \quad \text{if } k \in s - r \end{aligned} \tag{1}$$

where $\hat{\beta} = (\sum_r y_k)/(\sum_r x_k)$ and resulting estimate

$$\bar{y}_{.sRAT} = (1/n)\sum_{k \in s} y_{.k} = \left(\frac{\bar{y}_r}{\bar{x}_r}\right)\bar{x}_s. \tag{2}$$

where $\bar{x}_s = (1/n)\sum_s x_k$, $\bar{y}_r = (1/m)\sum_r y_k$, and $\bar{x}_r = (1/m)\sum_r x_k$.

Consider first variance estimation for sampling from an infinite population, or one with negligible sampling fraction. A naive variance estimator for the variance of (2), based on the jackknife, is given by:

$$v_{J(1)} = \frac{n-1}{n}\sum_{k \in s} (\bar{y}(-k) - \bar{y})^2 \tag{3}$$

where

$$\bar{y}(-k) = \frac{1}{(n-1)}(n\bar{y} - y_{.k}) \tag{4}$$

represents the mean of y computed by omitting observation k. Thus, this estimator treats imputed values as if they were observed, and may appropriately be called "naive" for doing so. Rao and Shao (1992) modified (3) and (4) in the context of hot deck imputation to correct the resulting understatement of the variance. The modification of (3) and (4) appropriate for ratio imputation (Rao and Sitter 1995) is:

$$v_J = \frac{n-1}{n}\sum_{k \in s} (\bar{y}^a(-k) - \bar{y})^2 \tag{5}$$

where

$$\begin{aligned} \bar{y}^a(-k) &= \left(\frac{1}{n-1}\right)\left[m\bar{y}_r - y_k + \sum_{j \in s-r} x_j \hat{\beta}(-k)\right] \\ &\qquad \text{if } k \in r \\ &= \left(\frac{1}{n-1}\right)\left[n\bar{y} - y_{.k}\right] \quad \text{if } k \in s - r \end{aligned} \tag{6}$$

and where $\hat{\beta}(-k) = \bar{y}_r(-k)/\bar{x}_r(-k) = (m\bar{y}_r - y_k)/(m\bar{x}_r - x_k)$. In other words, if $k \in s - r$, then (6) is computed in the same way as (4), by omitting the imputed value for k. If $k \in r$, then y_k is omitted and the imputed values are adjusted to reflect y_k*'s* influence on the imputed values.

Sampling without replacement in the absence of nonresponse requires that the variance estimator (3) be multiplied by the factor *(N-n)/N*. The extension of the Rao-Shao variance estimator to be discussed here begins by incorporating this factor into (5). With this adjustment, the variance estimator omits components of variance; the extended variance estimator incorporates two additional sets of replicate values to account for these components.

First, under a model with $E_\xi(y_k \mid x_k) = \beta x_k$, $Cov_\xi(y_j, y_k \mid x_j, x_k) = 0$, for $j \neq k$, the incorporation of *(N-n)/N* into (5) results in an effective underestimation of $Var_\xi(\hat{\beta})$, which in turn underestimates its contribution to the variance of the prediction of $E_\xi(y_k \mid x_k)$ by $\hat{\beta} x_k$.

Secondly, the extension must account for additional error in predicting the true y_k, $k \in s - r$ by $E_\xi(y_k \mid x_k)$. One approach is to identify two nearest neighbors, *nn1(k)* and *nn2(k)*, on the basis of their x values, with reported values for y.

One proposed extension is:

$$v_{J1} = \frac{n-1}{n} \frac{N-n}{N} \sum_{k \in s} \left[\bar{y}^{a}(-k) - \bar{y}\right]^2$$

$$+ \frac{n-1}{n} \frac{1}{Nn} \sum_{k \in r} \sum_{j \in s-r} \left[x_j(\hat{\beta}(-k) - \hat{\beta})\right]^2 \qquad (7)$$

$$+ \frac{1}{2} \frac{1}{Nn} \sum_{k \in s-r} \left[\frac{y_{nn1(k)}}{x_{nn1(k)}} x_k - \frac{y_{nn2(k)}}{x_{nn2(k)}} x_k\right]^2$$

Note that the last term in (7) reflects an adjustment for differences between $x_{nn1(k)}$ and $x_{nn2(k)}$. Because both $x_{nn1(k)}$ and $x_{nn2(k)}$ should consequently be close to x_k, the difference inside the braces could be estimated by $y_{nn1(k)} - y_{nn2(k)}$ instead.

We note that the added terms to the variance estimate make the following assumptions: 1) the variance in the estimated ratio coefficient assumes that the observed x and y in the finite population are sampled from a superpopulation; 2) the variance estimate based on nearest neighbors assumes that the ratio model is conditionally correct in expectation. In other words, the second assumption does not allow for conditional lack of fit of the regression line through the origin.

Fay (1995a) shows a close connection between (7) and an estimator studied by Rao and Sitter (1995)

$$v_2 = \left(\frac{1}{n} - \frac{1}{N}\right) \hat{\beta}^2 S_{xs}^2 + 2\left(\frac{\bar{x}_s}{\bar{x}_r}\right)\left(\frac{1}{n} - \frac{1}{N}\right) \hat{\beta} S_{xer}$$

$$+ \left(\frac{\bar{x}_s}{\bar{x}_r}\right)^2 \left(\frac{1}{m} - \frac{1}{N}\right) S_{er}^2 \qquad (8)$$

where $S_{xs}^2 = \sum_s (x_k - \bar{x}_s)^2 / (n-1)$, $S_{xer}^2 = \sum_r e_k x_k / (m-1)$ and $S_{er}^2 = \sum_r e_k^2 / (m-1)$ with $e_k = y_k - \hat{\beta} x_k$.

Estimator (8) suggests a second replication method, different in the last term from (7):

$$v_{J2} = \frac{n-1}{n} \frac{N-n}{N} \sum_{k \in s} \left[\bar{y}^{a}(-k) - \bar{y}\right]^2$$

$$+ \frac{n-1}{n} \frac{1}{Nn} \sum_{k \in r} \sum_{j \in s-r} \left[x_j(\hat{\beta}(-k) - \hat{\beta})\right]^2 \qquad (9)$$

$$+ \frac{1}{Nn} \sum_{k \in s-r} \left[\frac{y_{nn1(k)}}{x_{nn1(k)}} x_k - \hat{\beta} x_k\right]^2$$

The preceding formulas describe the situation for a single stratum, but the methods are readily extended to multiple strata by including weights and summing the variance contribution from each stratum.

4. MONTE CARLO RESULTS

Monte Carlo samples of 5000 draws were performed for each assumption about response. Each Monte Carlo draw consisted of an independent stratified simple random sample drawn without replacement with the sampling rates specified in Section 2. File preparations and most estimators were done in SAS, and the extended Rao-Shao estimator was calculated using the Census Bureau's variance package VPLX (Fay 1995b). The platform for all processing was a DEC 3000 Model 300LX workstation.

Five different assumptions about response were employed:

1) Uniform response rate of 80%.
2) Uniform response rate of 70%.
3) 90% response in the certainty, 68.2% in noncertainty (70% overall rate)
4) Probability of response is approximately $1 - x^{-0.25}$ -(70% overall rate)
5) Probability of response uniform within stratum at 60.0%, 78.4%, 83.8%, 87.9% and 91.6% for strata 1 through 5 respectively. (This pattern takes the stratum level response rates generated by pattern 4) and applies them uniformly to their respective strata).

Alternative 3) was selected on the basis that it is common practice at the Census Bureau to direct followup resources to the largest units in the sample. Thus, response rates typically are highest in certainty strata as a result of effort expended. In addition, the smallest units in the population often are the most prone to nonresponse, hence, alternatives 4) and 5) examine whether a further dependence of response on x affects the performance of the variance estimators.

Table 1 reports the results for strata 3-5 and the total. The naive variance estimator consistently understates the actual uncertainty. By comparison, the results obtained for estimators (7), (8) and (9) are highly encouraging. The results for strata 1-2 (not shown) and 3 are consistently satisfactory. We note that in these strata most of the variance estimate derives from the original component of the Rao-Shao variance estimator, that is, the first terms on the right hand side of (7) and (9).

In strata 4 and 5, the variance estimators depend more on the extensions to finite population sampling. In stratum 5 the first terms of (7) and (9) vanish, as do the first two terms of (8). Estimator (7) appears to underestimate the variance in stratum 4 and overestimate in 5. Our current interpretation is that the ratio model is not met adequately in stratum 5, leading (7) to underestimate. Estimator (9) does somewhat better in stratum 4 but overestimates in 5. In these two strata (8) is generally the best of the three variance estimators.

For estimating the totals across strata, which is the primary analytic interest, we note that the overall results are quite good for all three estimators. Figures 1 and 2 display some of the results for variance estimators, including results for strata 1 and 2 omitted from Table 1. Figures 3 and 4 display coverages of 95% confidence intervals based on a normal approximation. Intervals based on the naive variance estimator are consistently inadequate. Coverage of the individual strata is moderately good, but differences among estimators appear particularly for the certainty stratum. Coverage for the overall total is consistently good over the range of response assumptions studied.

Although the Monte Carlo studies identified some differences in performance among the estimators, each did quite well within the context of the simulated economic survey under the various assumptions of unconfounded response. Fay (1995a) reports further studies of these and other estimators applied to the populations studied by Lee, Rancourt, and Särndal, C.E. (1994).

These results suggest new topics for research, including how followup effort to reduce nonresponse might be optimally directed to improve accuracy and whether assumptions about nonresponse should be considered in the design phase, when forming sampling strata.

[1] This article reports results of research undertaken by staff members of the Census Bureau. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. Our thanks to Richard Sigman for his ideas on modeling and Mary Ann Cochran for editorial assistance.

REFERENCES

Fay, R.E. (1995a), "Replication-Based Variance Estimators for Imputed Survey Data from Finite Populations," unpublished Census Bureau report.

______ (1995b), "VPLX: Variance Estimation for Complex Samples, Program Documentation," unpublished Census Bureau report.

King, C. and Kornbau, M. (1994), "Inventory of Economic Area Statistical Practices, Phase 2: Editing, Imputation, Estimation, and Variance Estimation," unpublished Census Bureau report, ESMD Report Series, ESMD-9401.

Knaub Jr., James R. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," *Proceedings of the International Conference on Establishment Surveys, Invited and Contributed Papers*. American Statistical Association, Alexandria, VA, pp. 520-525.

Lavallé, P. and Hidiroglou, M. A. (1988), "On the Stratification of Skewed Populations," *Survey Methodology*, **14**, pp. 33-43.

Lee, H., Rancourt, E., and Särndal, C.E. (1994), "Experiments with Variance Estimation from Survey Data with Imputed Values," *Journal of Official Statistics*, **10**, 231-243.

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

Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, **79**, 811-822.

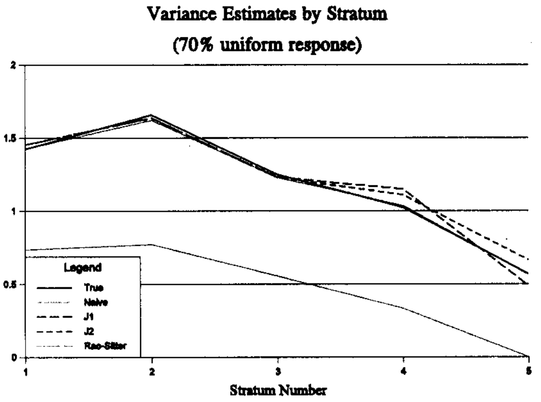
Rao, J.N.K. and Sitter, R.R. (1995), "Variance Estimation Under Two-Phase Sampling with Application to Imputation for Missing Data," *Biometrika,* 82, 453-460.

Sweet, E. M. and Sigman, R. S., (1995) "Evaluation of Model-assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data," *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, forthcoming.

Table 1. True and Estimated Variances ($\times 10^{12}$) for Strata 3-5 and Total.

	Naive	True	J1	J2	Rao-Sitter
Stratum 3 $n/N = .27$					
70%	.557	1.252	1.234	1.232	1.228
80%	.633	1.056	1.045	1.044	1.043
90%/68.2%	.544	1.260	1.287	1.282	1.277
$x^{.25}$	.665	.955	.964	.962	.964
Strat. ave.	.662	.975	.990	.990	.990
Stratum 4 $n/N = .64$					
70%	.334	1.027	1.149	1.109	1.036
80%	.371	.804	.858	.827	.782
90%/68.2%	.327	1.089	1.226	1.188	1.085
$x^{.25}$	.401	.610	.643	.624	.612
Strat. ave.	.398	.648	.682	.657	.630
Stratum 5 $n/N = 1.00$					
70%	-	.574	.491	.668	.566
80%	-	.303	.253	.366	.314
90%/68.2%	-	.138	.105	.159	.138
$x^{.25}$	-	.099	.077	.111	.098
Stratum ave.	-	.113	.085	.129	.113
All strata					
70%	2.402	5.923	5.974	6.106	5.878
80%	2.697	4.864	4.823	4.902	4.781
90%/68.2%	2.359	5.730	5.829	5.839	5.657
$x^{.25}$	2.679	4.530	4.547	4.559	4.511
Strat. ave.	2.580	5.050	4.943	4.959	4.850

Note: All variance estimators except the naive estimator performed uniformly well in strata 1-2, similar to stratum 3 shown here.



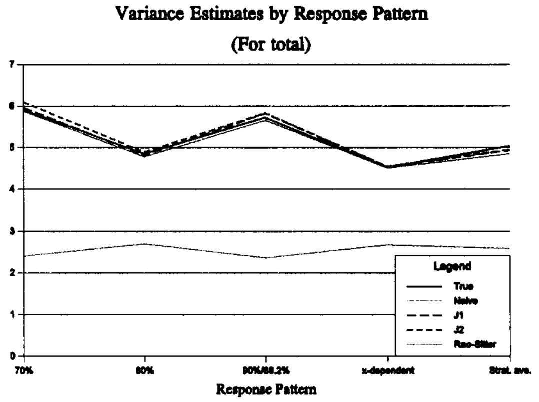
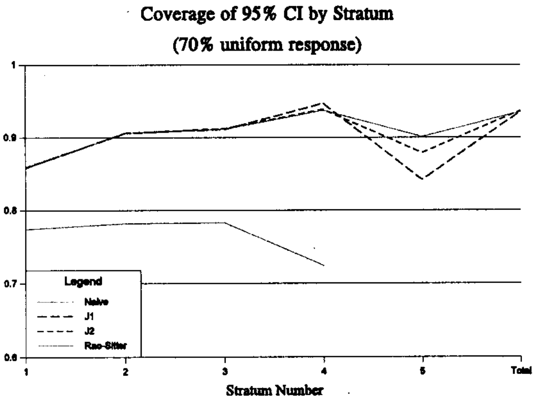


Figure 1. Comparison of true variance with four variance estimators by sampling stratum.

Figure 2. Comparison of true variance with four variance estimators for each of the response patterns.



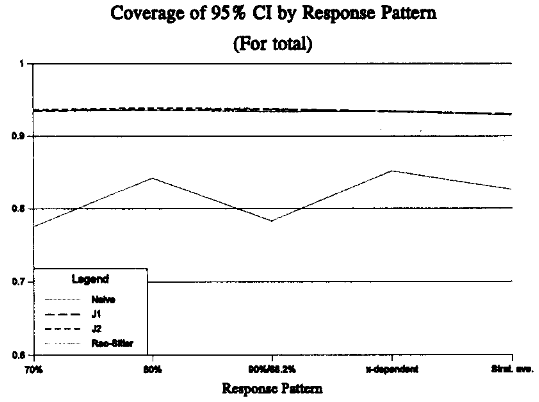


Figure 3. Coverages of 95% confidence intervals based on the normal approximation for four variance estimators.

Note: In stratum 5 the naive estimate of variance is 0, and confidence intervals were not computed. The coverage of the naive estimate for the total is .775.

Figure 4. Coverage of 95% confidence intervals based on the normal approximation for four variance estimators for the estimated total across strata.