Journal of Official Statistics
Vol. 10. No. 3, 1994, pp. 231–243

Experiments with Variance Estimation from Survey Data with Imputed Values

Hyunshik Lee[1], *Eric Rancourt*[1], *and Carl E. Särndal*[2]

Abstract: Missing data occur in almost all surveys and frequently some form of imputation is used to obtain a completed data set. It is well known that the ordinary variance formula applied to the data with imputed values generally underestimates the variance. There have been some proposals to remedy this problem. One is the well known multiple imputation. This method, however, requires generating two or more completed data sets, which may be seen as a disadvantage in some applications. Recently, Särndal proposed a variance estimation method for single imputation using a model-assisted approach to the problem. Rao also proposed a method based on the two-phase sampling approach for single imputation. In this article, these methods are studied by the Monte Carlo technique together with other methods including multiple imputation methods, under 12 artificially generated populations representing a variety of forms and three response mechanisms of which two are confounded, that is, they depend on the values of the variable of interest.

Key words: Model-assisted approach; ratio imputation; multiple imputation; Monte Carlo study.

1. Introduction

Let $\bar{y}_U = (1/N)\Sigma_U y_k$ be the mean of the finite population $U = \{1, \ldots, k, \ldots, N\}$. A simple random sample without replacement (SRSWOR), s, of size n is drawn from U to estimate $\bar{y}_U$. Denote by r the set of responding units; let m be the size of r. The nonresponse set is $s - r$; its size is $n - m$. For every unit $k \in r$, the value y_k is observed. However, for the units $k \in s - r$, the y_k-values are missing, and imputed values are derived with a specified imputation method. The six imputation methods studied in this paper are defined in the following.

If a single value imputation is used for each missing observation, the imputation leads to a completed data set, called the *data after imputation*. This data set is denoted as $\{y_{.k} : k \in s\}$, where $y_{.k}$ equals the observed value y_k if k is a responding unit, that is, if $k \in r$, and $y_{.k}$ equals the imputed value if k is a nonresponding unit, that is, if $k \in s - r$. The population mean is then estimated by $\bar{y}_{.s} = (1/n)\Sigma_s y_{.k}$.

[1] Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa K1A OT6, Canada.
[2] University of Montreal, Department of Mathematics and Statistics, CP6128, Succursale A, Montreal, Quebec H3C 3J7, Canada.
Acknowledgement: We are grateful to Y. Leblond for his participation in the early phase of the study. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. We also thank the associate editor and referees for their helpful comments. An earlier version of the article was presented at the annual meetings of the American Statistical Association in Atlanta, Georgia, August 1991.

We assume that imputation is carried out with the aid of an auxiliary variable, x, such that x_k, the value of x for unit k, is known and positive for every $k \in s$. That is, the data $x_s = \{x_k : k \in s\}$ are known.

Given s, the response set r is realized by an unknown probability distribution called the response mechanism. In general, when the sample s has been drawn, the probability that the response set r is realized is given by $q(r|s) = q(r|s, z_s, y_s)$, where $y_s = \{y_k : k \in s\}$. The response mechanism is called *unconfounded* in the particular case where $q(r|s) = q(r|z_s)$, and the response probabilities satisfy $\Pr(k \in r|s) > 0$ for all $k \in s$. Otherwise, the response mechanism is *confounded*. Our distinction between confounded and unconfounded mechanisms corresponds closely to the one made by Rubin (1983, 1992). An unconfounded mechanism may depend on the sample x-data, x_s, but not on the sample y-data, y_s, so we can still have response probabilities $\Pr(k \in r|s)$ that vary with k; for example, $\Pr(k \in r|s)$ may be a function of x_k but not of y_k. An unconfounded response mechanism such that the units respond independently with constant response probabilities $\Pr(k \in r|s)$ for all $k \in s$ is called *uniform*. (A distinction is often made between ignorable and nonignorable response mechanisms, see, for example Rubin (1983, 1987). This distinction is based on a property of the posterior distribution of the unobserved y-values. Because we are interested in randomization inference rather than Bayesian inference, the distinction between confounded and unconfounded mechanisms is more suited for this article.)

The imputation methods considered in the Monte Carlo study are: (i) ratio imputation; (ii) nearest neighbor imputation. The main objective in this study is to compare different estimators of the variance of the estimator $\bar{y}_{.s}$. Both single imputation and multiple imputation variance estimators are considered.

2. Description of the Variance Estimators Studied

The point estimator of $\bar{y}_U$ for the single imputation methods is given by the mean of the data after imputation, $\bar{y}_{.s} = (1/n)\Sigma_s y_{.k}$. The estimator formula is thus the same as the one that would be used in the case of 100% response. Since this formula is calculated on data after imputation, there is an implicit assumption that a negligible bias is caused by replacing missing data by imputed values. This assumption is often violated, particularly when the nonresponse mechanism is confounded.

For the multiple imputation methods, we used $M = 2$ repeated imputations. This implies that the point estimator is calculated as the average of the means of two sets of data after imputation. In the multiple imputation methods, too, the bias of the point estimator can be considerable when the nonresponse is confounded.

We now define the variance estimators used in the study.

2.1. *Ratio imputation*

Consider first the single value imputation. If unit k requires imputation, the value $\hat{B}x_k$ is imputed, where $\hat{B} = (\Sigma_r y_k)/\Sigma_r x_k)$. The data after imputation are therefore

$$y_{.k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}x_k, & \text{if } k \in s - r. \end{cases} \tag{2.1}$$

For this imputation method, the point estimator $\bar{y}_{.s} = (1/n)\Sigma_s y_{.k}$ becomes

$$\bar{y}_{.s\text{RAT}} = \frac{\bar{y}_r}{\bar{x}_r}\bar{x}_s \tag{2.2}$$

where $\bar{x}_s = (1/n)\Sigma_s x_k$, $\bar{y}_r = (1/m)\Sigma_r y_k$ and $\bar{x}_r = (1/m)\Sigma_r x_k$. Three different variance

estimators are considered with this imputation method.

RAT-O. This method uses the ordinary variance estimator formula, computed using the data after imputation, that is

$$\hat{V} = (1/n - 1/N)S^2_{y\cdot s} \tag{2.3}$$

where $S^2_{y\cdot s} = \Sigma_s(y_{\cdot k} - \bar{y}_{\cdot s})^2/(n-1)$. The method is known to underestimate the real variance and is included in the study only to assess the underestimation caused by acting as if imputed data were as good as actual data.

RAT-R. Using a two-phase sampling argument, Rao (1990) suggested the variance estimator

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{yr} + \left(\frac{1}{m} - \frac{1}{n}\right)S^2_{er}$$

where $S^2_{yr} = \Sigma_r(y_k - \bar{y}_r)^2/(m-1)$ and $S^2_{er} = \Sigma_r e^2_k/(m-1)$, with $e_k = y_k - \hat{B}x_k$.

An alternative variance estimator, also suggested by Rao (1990), is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{B}^2 S^2_{xs} + 2\left(\frac{1}{n} - \frac{1}{N}\right)\hat{B}S_{xer} + \left(\frac{1}{m} - \frac{1}{N}\right)S^2_{er} \tag{2.4}$$

where $S_{xer} = \Sigma_r e_k x_k/(m-1)$.

Both of these variance estimators are unbiased in large samples under the assumption that the response set r is generated by a uniform response mechanism, given s. This holds regardless of the regression relationship between x and y. However, both variance estimators are biased if the response mechanism is nonuniform, for example, if the response probabilities depend on x_k or y_k (or on both). In our empirical study, we examined both of these variance estimators but results are reported only for the latter formula since it performed consistently better than the former.

RAT-S. This model-assisted variance estimator, derived in Särndal (1990) and also in Deville and Särndal (1991) for more general cases, is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)\{S^2_{y\cdot s} + C_0\hat{\sigma}^2\} + \left(\frac{1}{m} - \frac{1}{n}\right)C_1\hat{\sigma}^2 \tag{2.5}$$

where

$$C_0 = \frac{1}{n-1} \times \left(\sum_{s-r} x_k - \frac{\sum_{s-r} x_k^2}{\sum_r x_k} + \frac{1}{n}\frac{\sum_{s-r} x_k \sum_s x_k}{\sum_r x_k}\right)$$

$$C_1 = \frac{\bar{x}_s \bar{x}_{s-r}}{\bar{x}_r},$$

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2/(m-1)}{\bar{x}_r\{1 - (cv_{xr})^2/m\}},$$

with $\bar{x}_{s-r} = \Sigma_{s-r} x_k/(n-m)$, $e_k = y_k - \hat{B}x_k$ and $\text{cv}_{xr} = S_{xr}/\bar{x}_r$, which is the coefficient of variation of x in the response set r. The term $(1/n - 1/N)\{S^2_{y\cdot s} + C_0\hat{\sigma}^2\}$ estimates the sampling variance component, and $(1/m - 1/n)C_1\hat{\sigma}^2$ estimates the imputation variance component. This variance estimator is based on the regression model ξ stating that $y_k = \beta x_k + \epsilon_k$, for $k = 1, \ldots, N$, where $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2 x_k$ and the model errors ϵ_k are independent. As shown in Särndal (1990), the RAT-S variance estimator is ξpq-unbiased if (i) the model ξ holds and (ii) the response mechanism q is unconfounded. The ξpq-unbiasedness implies that $E_\xi E_p E_q\ (\hat{V} - V) = 0$, where $V = E_\xi E_p E_q\ (\bar{y}_{\cdot s} - \bar{y}_U)^2$ is the ξpq-variance of the point estimator $\bar{y}_{\cdot s} = \bar{x}_s \bar{y}_r/\bar{x}_r$. Here, the operators E_ξ, E_p and E_q denote expectation over the model ξ, over SRS sampling, and over

the unconfounded mechanism q, respectively. The RAT-S variance estimator is therefore expected to perform particularly well when the finite population scatter (y_k, x_k) agrees closely with the model ξ and when the response probability of the unit k depends on x_k only and not on y_k. When m is moderate to large, $C_0 \doteq (1 - m/n)\bar{x}_{s-r}$ and $\hat{\sigma}^2 \doteq \Sigma_r e_k^2 / \Sigma_r x_k$ are good approximations of the more cumbersome exact expressions.

Consider now multiple ratio imputation. The multiple imputation is carried out as outlined in Rubin (1987, pp. 166–168), under the assumption that $y_k \sim N(\beta x_k, \sigma^2 x_k)$ and that response mechanism is ignorable in the sense defined by Rubin (1983). First, β and σ^2 are estimated, respectively, by $\hat{B}$ and

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_r \frac{(y_k - \hat{B}x_k)^2}{x_k}.$$

Then, for each $i = 1, \ldots, M$ (where M is the number of imputations), perform the following steps:

Step 1. Draw a χ^2 random variate with $(m-1)$ degrees of freedom, say g, and let $\sigma_i^2 = \hat{\sigma}^2(m-1)/g$.

Step 2. Draw a $N(0,1)$ random variate, say z, and let $\beta_i = \hat{B} + \sigma_i z \, (\Sigma_r x_k)^{-1/2}$.

Step 3. For each $k \in s - r$, draw a $N(0,1)$ variate independently, say u, and let $e_{ik}^* = u\sqrt{x_k}\sigma_i$.

In our study, $M = 2$. Thus we obtain two data sets after imputation defined by

$$y_{\cdot 1k} = \begin{cases} y_k, & \text{if } k \in r \\ \beta_1 x_k + e_{1k}^*, & \text{if } k \in s - r, \end{cases}$$

$$y_{\cdot 2k} = \begin{cases} y_k, & \text{if } k \in r \\ \beta_2 x_k + e_{2k}^*, & \text{if } k \in s - r \end{cases}$$

A modification to the above procedure is to replace Step 3 by the following:

Step 3′. For each $k \in s - r$, draw a number, say w_k, with replacement from the set of standardized residuals $\{(y_l - \hat{B}x_l)/\sqrt{(1 - 1/m)x_l\hat{\sigma}^2};\ l \in r\}$ and let $e_{ik}^* = w_k\sqrt{x_k}\sigma_i$.

We tried both methods but report results only for the latter method here because it appeared to be the better of the two.

The point estimator of the population mean is

$$\bar{y}_{\cdot s\text{RAT-M}} = (\bar{y}_{\cdot 1s} + \bar{y}_{\cdot 2s})/2 \tag{2.6}$$

where $\bar{y}_{\cdot 1s}$ and $\bar{y}_{\cdot 2s}$ are the means of the data sets $\{y_{\cdot 1k} : k \in s\}$ and $\{y_{\cdot 2k} : k \in s\}$, respectively. The corresponding variance estimator is defined as follows:

RAT-M. The variance estimator suggested by Rubin (1983, 1986) is given by

$$\hat{V} = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{n} - \frac{1}{N}\right)S_{y\cdot js}^2 + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\times \sum_{j=1}^{M}(\bar{y}_{\cdot js} - \bar{y}_{\cdot s\text{RAT-M}})^2 \tag{2.7}$$

with $M = 2$ in our study where $S_{y\cdot js}^2 = \Sigma_s(y_{\cdot jk} - \bar{y}_{\cdot js})^2/(n-1)$ is the variance calculated from the jth data set after imputation, $\{y_{\cdot jk} : k \in s\}$.

2.2. *Nearest neighbor imputation*

We examined two variance estimators for single imputation by nearest neighbor, NN-O and NN-S, and one variance estimator for the multiple imputation analog, NN-M. We now describe these estimators.

Consider first single value nearest neighbor imputation. If the unit k requires imputation, the imputed value, $y_{\text{NN}k}$, equals the y-value of a donor unit that is as close

as possible to k, as measured by the x-variable. More specifically, the donor unit is the one for which the distance $|x_k - x_l|$ is minimum among all potential donors l such that $l \in r$, $l \neq k$. The data after imputation are

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{\mathrm{NN}k}, & \text{if } k \in s - r. \end{cases}$$

The point estimator is the mean of the n values $y_{\cdot k}$, that is,

$$\bar{y}_{\cdot s\mathrm{NN}} = \frac{1}{n}\left(\sum_r y_k + \sum_{s-r} y_{\mathrm{NN}k}\right). \qquad (2.8)$$

Two variance estimators were used with this imputation method:

NN-O. This consists of the ordinary variance estimator, $\hat{V} = (1/n - 1/N)S^2_{y\cdot s}$, computed on the data after imputation. We can expect underestimation of the true variance with this method.

NN-S. This variance estimator is given by the formula $\hat{V}$ defined as in the RAT-S method by (2.5) with the only difference that $S^2_{y\cdot s}$ is computed using the data with nearest neighbor imputation values $y_{\mathrm{NN}k}$ (instead of $\hat{B}x_k$) for $k \in s - r$, whereas $\hat{\sigma}^2$ is computed with the aid of the residuals from the ratio imputation method, that is, $e_k = y_k - \hat{B}x_k$, for $k \in r$. (The residuals $y_k - y_{\mathrm{NN}k}$ are not used because they would lead to overestimation of σ^2.)

Consider now multiple imputation by nearest neighbor. For each nonrespondent, the two nearest neighbors are identified based on the distance $|x_k - x_l|$. Of the two y-values thus obtained, one is randomly picked and assigned to the first data set, $y_{\mathrm{NN}1k}$, and the remaining y-value is assigned to the second data set, $y_{\mathrm{NN}2k}$. Two data sets after imputation are thus obtained, namely

$$y_{\cdot 1k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{\mathrm{NN}1k}, & \text{if } k \in s - r, \end{cases}$$

$$y_{\cdot 2k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{\mathrm{NN}2k}, & \text{if } k \in s - r, \end{cases}$$

The point estimator of the population mean is

$$\bar{y}_{\cdot s\mathrm{NN-M}} = (\bar{y}_{\cdot 1s} + \bar{y}_{\cdot 2s})/2 \qquad (2.9)$$

where $\bar{y}_{\cdot 1s}$ and $\bar{y}_{\cdot 2s}$ are the means of the two data sets $\{y_{\cdot 1k} : k \in s\}$ and $\{y_{\cdot 2k} : k \in s\}$, respectively. The variance estimator used for this method is:

NN-M. The variance estimator is calculated in the same way as RAT-M, that is, from (2.7). The only difference is that nearest neighbor imputation values are used to calculate the quantities $\bar{y}_{\cdot js}$ and $S^2_{y\cdot js}$, $j = 1, 2$. This method is not "proper" (see Rubin 1987, pp. 118–128, for the definition of a proper multiple imputation) but was mentioned in Rubin (1986, 1987).

Remark 1: The single value imputation variance estimators require that the imputed values be flagged in the data file, whereas flags are not required for the multiple imputation variance estimators.

Remark 2: Ties can occur in reality when donors are identified. In that case, a donor is selected randomly. However, in our simulation, the occurrence of ties was almost nonexistent because we used untruncated random numbers.

Remark 3: If one tries to modify RAT-R for nearest neighbor imputation using $e_k = y_k - \hat{B}x_k$ in the same way as for the NN-S method, the resulting estimator remains identical to RAT-R. This is the reason why we did not try the modification of RAT-R.

3. Monte Carlo Simulation Experiments

The performance of the different variance estimators was studied with the aid of the customary Monte Carlo summary measures: mean, bias and variance of the variance estimators and also coverage rate

Table 1. Characteristics of the 12 populations used in the simulation study

Pop.	Type	a	b	c	d	g	ρ	Mean
1	RATIO	0	1.50	0.00	13.78	.25	.773	70.44
2	RATIO	0	1.50	0.00	5.13	.50	.775	73.47
3	RATIO	0	1.50	0.00	1.84	.75	.755	72.93
4	CONCAVE	0	3.00	−.01	15.04	.25	.760	112.93
5	CONCAVE	0	3.00	−.01	5.60	.50	.765	117.10
6	CONCAVE	0	3.00	−.01	2.01	.75	.746	110.31
7	CONVEX	0	.25	.01	13.20	.25	.761	44.77
8	CONVEX	0	.25	.01	4.91	.50	.759	40.94
9	CONVEX	0	.25	.01	0.75	.75	.755	34.93
10	SIM-REG	20	1.50	0.00	13.79	.25	.746	91.97
11	SIM-REG	20	1.50	0.00	5.13	.50	.763	91.29
12	SIM-REG	20	1.50	0.00	1.84	.75	.767	91.22

of the confidence interval. The performance of the different imputation methods was also investigated in terms of mean and bias of the point estimators of the population mean.

The Monte Carlo simulations were carried out using 12 different artificially generated populations of values (y_k, x_k). These populations were generated as follows: a set of $N = 100$ x-values was generated according to a Γ-distribution with mean 48 and variance 768. Then, for each fixed value of x, we generated the corresponding value of y according to a Γ-distribution with mean $\mu(x) = a + bx + cx^2$ and variance $\sigma^2(x) = d^2x^{2g}$ with appropriately chosen constants a, b, c, d, and g. If the density of the Γ-distribution is written as

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) \text{ for } x > 0$$

then the mean and the variance are, respectively, $\alpha\beta$ and $\alpha\beta^2$. We thus have the equations $\mu(x) = a + bx + cx^2 = \alpha\beta$ and $\sigma^2(x) = d^2x^{2g} = \alpha\beta^2$, which imply that the constants α and β used to generate the y-value associated with a given x-value are determined by

$$\alpha = \frac{\{\mu(x)\}^2}{\sigma^2(x)} = \frac{(a + bx + cx^2)^2}{d^2x^{2g}}$$

$$\beta = \frac{\sigma^2(x)}{\mu(x)} = \frac{d^2x^{2g}}{a + bx + cx^2}.$$

The coefficient of correlation between x and y is also a function of the five constants a, b, c, d, and g. We first specified the values for a, b, c, and g, and then determined the remaining constant, d, as a consequence of the desired theoretical correlation, which we fixed at 0.75 for all populations. The values of a, b, c, d, and g are given for the 12 populations in Table 1, as well as the correlation coefficient ρ and the mean of y calculated from the $N = 100$ pairs of (x_k, y_k), $k = 1, \cdots, 100$, that were generated by the procedure.

The constants, a, b and c are shape parameters. The populations are classified into four population types as shown in Table 1. Populations 1 to 3 represent a linear regression through the origin. Populations 4 to 9 have a second degree polynomial regression through the origin with slight curvature. Populations 10 to 12 are based on simple linear regression with a nonzero intercept.

Figures 1–4 show plots of populations 2, 5, 8, and 11.

For each of the 12 populations, three different response mechanisms were used. These have the following features, where

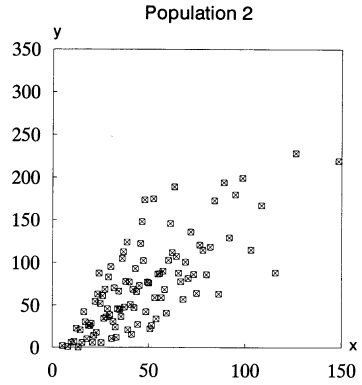

Fig. 1. Scatter plot of population 2

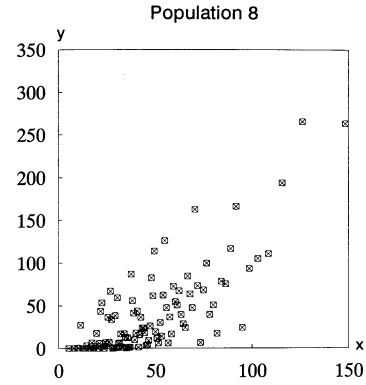

Fig. 3. Scatter plot of population 8

θ_k denotes the probability of nonresponse for the unit k:

i. The probability θ_k decreases as y_k increases where $\theta_k = \exp(-c_1 y_k)$ and the constant c_1 is chosen so that the average nonresponse probability over the whole finite population is 0.3. (A numerical method was used to achieve this goal.) This confounded response mechanism, which is such that small y-values are underrepresented among the respondents, is denoted $\downarrow$.

ii. The probability θ_k increases as y_k increases where $\theta_k = 1 - \exp(-c_2 y_k)$ and the constant c_2 is chosen so that the average nonresponse probability over the whole finite population is 0.3. This mechanism, which is such that small y-values are overrepresented among the respondents, is denoted $\uparrow$.

iii. The probability θ_k is constant at 0.3 for all $k \in U$. Both large and small y-values are evenly represented among the respondents. This mechanism is denoted $\rightarrow$.

In cases (i) and (ii) the nonresponse probability depends on the value y_k of the variable of interest; these nonresponse mechanisms are confounded. In case (iii), the probability of nonresponse is constant through the population; the response mechanism is unconfounded and uniform.

For each population, we drew 1,000 samples, each of size $n = 30$. For each of these samples, 50 realizations for each

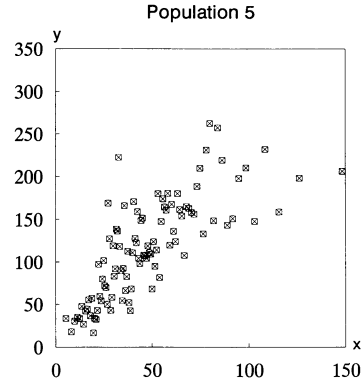

Fig. 2. Scatter plot of population 5

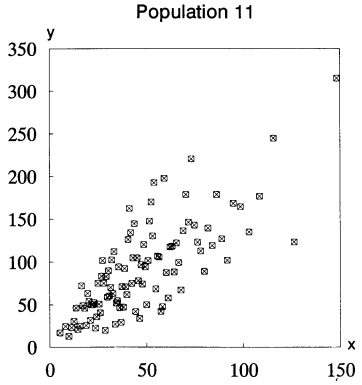

Fig. 4. Scatter plot of population 11

of the three response mechanisms were obtained by performing independent Bernoulli trials, one for each of the 30 sample units. For each of the $12 \times 3 = 36$ different combinations of population and response mechanism, we thus obtained 50,000 realized nonresponse sets. With this number of repetitions, the Monte Carlo error should not exceed two-thirds of one percent. The size of the nonresponse set, $(n - m)$, is random with the expected value $30 \times 0.3 = 9$ for each of the three mechanisms.

In summary, the following factors were held fixed in our Monte Carlo study:

i. the sampling fraction $(n/N = 30\%)$,
ii. the expected nonresponse rate $(E_q(m)/n = 30\%)$,
iii. the x-value distribution (generated as a Γ-distribution with mean 48 and variance 768),
iv. the correlation between x and y (around 0.75).

These factors were fixed since changing them is not expected to have any significant effect on the ranking of the procedures studied. A nonresponse rate of 30% or more is often encountered in actual surveys and represents a rate high enough so that uncritical use of standard formulae might be grossly misleading. The gamma shape is fairly typical for survey populations, in particular for many populations encountered in business surveys.

The factors that were varied in the Monte Carlo study were:

v. the populations (12 cases, as described above, representing four different types of regression between x and y, crossed with three different patterns for the variance around the regression curve),
vi. the response mechanisms (three cases, as described above).

The population types and the response mechanisms have considerable effect on the procedures we are comparing, so it is essential to let these factors vary.

4. Summary of the Simulation Results

The principal object of the simulation study was to examine the performance of the variance estimators. However, we first comment on some observed features of the four point estimators $\bar{y}_{\cdot sRAT}$, $\bar{y}_{\cdot sRAT-M}$, $\bar{y}_{\cdot sNN}$ and $\bar{y}_{\cdot NN-M}$ defined, respectively, by (2.2), (2.6), (2.8) and (2.9).

4.1. *Bias and variance of the point estimators*

For the mechanism $\rightarrow$, we found that the Monte Carlo means of all four point estimators, (2.2), (2.6), (2.8) and (2.9), agree well with the respective population means. This holds for all 12 populations. That is, all point estimators are essentially unbiased when the response mechanism is uniform, and this is true regardless of population type. Also as expected, all four point estimators are noticeably biased for the mechanism $\downarrow$ (where the bias is ordinarily positive) and for the mechanism $\uparrow$ (where the bias is ordinarily negative). The bias can be substantial and is particularly pronounced for the convex regression populations 7, 8 and 9, where the absolute relative bias ranges between 12% and 22% with the $\downarrow$ mechanism and between 30% and 37% with the $\uparrow$ mechanism. For the other nine populations, the absolute relative bias is less than 13%. We conclude that confounded nonresponse is a severe problem in that it generates considerable bias in all point estimators studied; consequently, the squared bias can be a large component of the mean squared error (MSE). Our study indicates that the convex regression populations in combination with

the ↑ mechanism are particularly likely to generate a large bias.

In particular, the ratio imputation estimator $\bar{y}_{.sRAT} = \bar{x}_s\bar{y}_r/\bar{x}_r$ is considerably biased for the confounded mechanisms ↑ and ↓. This is true even for population 2, despite the fact that this population is suited for the ratio estimation in the sense that it conforms to a linear regression through the origin with variance increasing proportionally to x. The bias is explained in this case by the fact that $\bar{y}_r/\bar{x}_r$ is a biased estimate of the slope when nonresponse is confounded.

We also noted that under the confounded mechanisms the nearest neighbor estimators $\bar{y}_{.sNN}$ and $\bar{y}_{.sNN-M}$ are more biased than $\bar{y}_{.sRAT}$ and $\bar{y}_{.sRAT-M}$ for most populations. Exceptions to this are found among the convex regression populations.

As for the variances of the four point estimators, we noted in a majority of the 36 cases that $\bar{y}_{.sRAT}$ and $\bar{y}_{.sRAT-M}$ have distinctively lower variances than $\bar{y}_{.sNN}$ and $\bar{y}_{.sNN-M}$. The averaging over repeated imputations leads to some variance reduction; we noted that $\bar{y}_{.sNN-M}$ has slightly lower variance than $\bar{y}_{.sNN}$.

We now turn to the main objective of the simulation study, that is, comparison of the variance estimators. The argument can be made that variance estimation should not be attempted if the relative bias of the point estimator is too high because then the confidence intervals are far from valid. This is why we report results separately for 27 out of 36 cases where the bias is not considered too large.

4.2. *Simulation results for the variance estimators for ratio imputation*

Table 2 summarizes the Monte Carlo simulation results for the four variance estimators for ratio imputation defined by (2.3), (2.4), (2.5) and (2.7). Results are reported for three performance criteria: absolute relative bias (ARB), coverage rate (COVR), and root mean squared error (RMSE).

To define these criteria, let the sampling variance of a point estimator $\hat{\bar{y}}_U$ of $\bar{y}_U$ be $V = E_pE_q(\hat{\bar{y}}_U - \bar{y}_U)^2$. Then the ARB and RMSE of a variance estimator $\hat{V}$ of V are defined, respectively, as

$$\mathrm{ARB}(\hat{V}) = 100 \times |E_pE_q(\hat{V}) - V|/V,$$

$$\mathrm{RMSE}(\hat{V}) = \sqrt{E_pE_q(\hat{V} - V)^2}.$$

The expectations in these formulae were evaluated by Monte Carlo simulation using 50,000 realized response sets. The COVR is defined as the actual coverage probability of the 95% confidence interval constructed by $\hat{\bar{y}}_U \pm 1.96\sqrt{\hat{V}}$. The Monte Carlo COVR was calculated as 100 times the proportion of the 50,000 response sets such that the interval included the true mean $\bar{y}_U$.

Results are reported as averages for different subsets of 27 cases formed by crossing nine populations with three response mechanisms (RM's). The nine populations are those where the bias of the point estimator was found to be fairly limited, that is, populations 1, 2 and 3 (which have linear regression through the origin and are called Ratio populations), and populations 4, 5, 6, 10, 11, and 12 (which have a regression other than linear through the origin and are called Nonratio populations). These 27 cases are broken down into four cells formed by contrasting on the one hand Ratio populations with Nonratio populations and on the other Uniform RM with Nonuniform RM. Average performance measures are shown for each cell, for the marginals, and for all 27 cases.

In addition, the far right column in Table 2, headed All populations, shows

Table 2. Performance of the four variance estimators under ratio imputation

Response mechanism	Estimator	Populations with smaller point estimation bias (Pop 1, 2, 3, 4, 5, 6, 10, 11, 12)									All populations (Including Pop 7, 8, 9)		
		Ratio (Pop 1, 2, 3)			Non-Ratio (Others)			All					
		ARB	COVR	RMSE	ARB	COVR	RMSE	ARB	COVR	RMSE	ARB	COVR	RMSE
Uniform	RAT-O	32.8	87.3	39.1	23.0	89.8	32.4	26.3	89.0	34.7	31.5	86.7	39.9
	RAT-R	3.1	92.6	33.0	1.8	93.3	30.5	2.2	93.1	31.3	3.7	91.9	36.2
	RAT-S	3.6	92.9	32.6	13.2	94.9	37.1	10.0	94.2	35.6	11.0	92.6	38.4
	RAT-M	4.5	91.6	48.9	6.9	93.8	55.1	6.1	93.1	53.0	7.4	91.5	54.7
Non-uniform	RAT-O	30.6	80.4	38.4	21.2	88.9	31.6	24.3	86.1	33.9	29.8	78.0	39.8
	RAT-R	18.1	87.1	37.9	11.7	92.5	32.9	13.8	90.7	34.6	19.6	84.0	41.9
	RAT-S	4.2	87.8	32.6	14.3	94.1	37.0	11.0	92.0	35.5	12.2	85.1	38.9
	RAT-M	5.7	85.4	48.2	8.1	92.8	53.9	7.3	90.4	52.0	8.3	82.9	55.0
All	RAT-O	31.3	82.7	38.7	21.8	89.2	31.9	25.0	87.1	34.1	30.4	80.9	39.8
	RAT-R	13.1	88.9	36.2	8.4	92.8	32.1	9.9	91.5	33.5	14.3	86.6	40.0
	RAT-S	4.0	89.5	32.6	14.0	94.3	37.0	10.6	92.7	35.6	11.8	87.6	38.8
	RAT-M	5.3	87.5	48.4	7.7	93.1	54.3	6.9	91.3	52.3	8.0	85.8	54.9

Note: ARB: Absolute relative bias; COVR: Coverage rate; RMSE: Root mean squared error.

Table 3. Performance of the three variance estimators under nearest neighbor imputation

Response mechanism	Estimator	Populations with smaller point estimation bias (Pop 1, 2, 3, 4, 5, 6, 10, 11, 12)									All populations (Including Pop 7, 8, 9)		
		Ratio (Pop 1, 2, 3)			Non-Ratio (Others)			All					
		ARB	COVR	RMSE	ARB	COVR	RMSE	ARB	COVR	RMSE	ARB	COVR	RMSE
Uniform	NN-O	37.2	85.5	44.0	34.2	86.6	40.5	35.2	86.2	41.7	36.0	85.1	44.0
	NN-S	10.1	91.6	34.1	4.1	93.1	31.1	6.1	92.6	32.1	6.2	91.5	36.5
	NN-M	12.4	89.6	49.3	10.3	90.5	44.9	11.0	90.2	46.4	12.4	89.0	48.8
Non-uniform	NN-O	39.9	73.2	46.2	37.1	79.6	42.9	38.0	77.5	44.0	40.7	72.5	47.6
	NN-S	14.2	81.9	34.0	6.6	88.2	30.6	9.1	86.1	31.8	13.0	81.1	36.3
	NN-M	17.6	77.6	48.5	15.0	83.8	45.0	15.9	81.8	46.2	19.2	76.6	50.0
All	NN-O	39.0	77.3	45.5	36.1	82.0	42.1	37.1	80.4	43.2	39.1	76.7	46.4
	NN-S	12.8	85.1	34.1	5.7	89.9	30.8	8.1	88.3	31.9	10.7	84.6	36.4
	NN-M	15.9	81.6	48.8	13.4	86.1	45.0	14.3	84.6	46.2	16.9	80.7	49.6

Note: ARB: Absolute relative bias; COVR: Coverage rate; RMSE: Root mean squared error.

average performance for all 36 cases (12 populations by 3 RM's). That is, the three convex regression populations (those with considerable point estimator bias) are also included. Of these 36 cases, 12 correspond to Uniform RM and 24 to Nonuniform RM.

From Table 2 it is clear that the RAT-O variance estimator is highly biased under all conditions. What Table 2 does not show is that this large bias is negative, as one would expect. We need not comment further on the RAT-O estimator which is clearly unsatisfactory. For the other three contenders, RAT-R, RAT-S and RAT-M, we conclude:

a. *Absolute Relative Bias (ARB)*: For the cell Ratio populations with Uniform RM (three cases), all three variance estimators perform well; the ARBs lie between 3% and 5%. The remaining three cells put the differences between RAT-R, RAT-S and RAT-M into very clear focus.

For the cell Ratio populations with Nonuniform RM (six cases), RAT-S is the best, as theory would lead us to expect, RAT-M follows closely, whereas RAT-R is considerably more biased. For the cell Nonratio populations with Uniform RM (six cases), RAT-R is the best, again as theory would suggest, followed closely by RAT-M; RAT-S is considerably more biased. Finally, the cell Nonratio populations with Nonuniform RM (six cases) can be said to test resistance to breakdown of assumptions of Ratio population and Uniform RM. In this cell, the conditions are a priori unfavourable both to RAT-S (which builds on the assumption of Ratio population) and to RAT-R (which builds on the assumption of Uniform RM). We see in this cell that the multiple imputation estimator RAT-M resists best (the ARB is 8.1), followed by RAT-R and RAT-S (the ARBs are 11.7 and 14.3, respectively).

The marginal averages confirm the finding that for Ratio populations, RAT-S is the best method, while for Uniform RM, RAT-R is the best.

Averaging over the 27 cases, RAT-M has a clear advantage (the ARB is 6.9) over RAT-R and RAT-S. Between the latter two, there is little to choose (9.9 and 10.6, respectively). The better performance of RAT-M is again explained by the fact that RAT-S and RAT-R build each on one important assumption, whereas RAT-M does not appeal explicitly to either of these conditions.

Averaging over all 36 cases again shows RAT-M ahead (the ARB is 8.0), followed by RAT-S (11.8) and RAT-R (14.3).

b. *Coverage Rate (COVR)*: On this criterion, RAT-S has an advantage over both RAT-R and RAT-M for all of the breakdowns in Table 2. That is, RAT-S comes closer to the nominal 95% confidence level. However, the differences are small; RAT-R and RAT-M follow closely. All three methods show satisfactory performance in all cells except the nonuniform/ratio cell, although the COVR is always somewhat on the low side.

c. *Root Mean Squared Error (RMSE)*: No great differences between RAT-S and RAT-R are recorded in any of the four cells in the left upper corner of Table 2. A striking contrast is the much higher RMSE of RAT-M. To decrease the RMSE of RAT-M would be possible by increasing the number of repeated imputations to more than two. This, on the other hand, implies a considerably increased data handling effort, and then RAT-M loses some of its attractiveness.

4.3. *Simulation results for the variance estimators for nearest neighbor imputation*

Table 3 summarizes the Monte Carlo simulation results for the three variance estimators for nearest neighbor imputation (NN-O, NN-S and NN-M) described in Section 2. The results are reported in the same manner as in Table 2, that is, with 27 cases broken down into four cells, and with all 36 cases covered in the far right column.

The NN-O variance estimator is again clearly unsatisfactory. We go on to compare the variance estimators NN-S and NN-M.

For the 27 cases, NN-S has smaller ARB than NN-M in all cells, as well as overall. This also holds for all 36 cases.

The coverage rate, COVR, is also more satisfactory for NN-S than for NN-M. The RMSE is as in the case of ratio imputation, considerably higher for the multiple imputation method.

5. Conclusion

The simulation shows very clearly that vast improvement on the "ordinary formula" (RAT-O and NN-O) can be achieved by properly constructed variance estimators either with single value imputation (RAT-R, RAT-S and NN-S) or with multiple imputation (RAT-M and NN-M).

The most important conclusion regarding RAT-R, RAT-S and RAT-M is that none of these is ideal from all points of view. Both RAT-R and RAT-S are sensitive to departure from the conditions on which they build: uniform RM for RAT-R and ratio population for RAT-S. On the other hand, RAT-M is considerably more volatile than the other two competitors, when only two repeated imputations are used.

6. References

Deville, J.C. and Särndal, C.E. (1991). Estimation de la variance en présence de données imputées. In Proceedings of Invited Papers for the 48th Session of the International Statistical Institute, Cairo, Egypt, September 9–17, 1991, Book 2, Topic 17, pp. 3/17.

Rao, J.N.K. (1990). Variance Estimation Under Imputation for Missing Data. Technical Paper, Statistics Canada, Ottawa.

Rubin, D.B. (1983). Conceptual Issues in the Presence of Nonresponse. In Incomplete Data in Sample Surveys, W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Vol. 2. New York: Academic Press, 123–142.

Rubin, D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, 12, 37–47.

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

Rubin, D.B. (1992). Discussion. In Proceedings, 1992 Annual Research Conference, Washington, Bureau of the Census, March 22–25, 540–544.

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. In Proceedings of Statistics Canada's Symposium '90: Measurement and Improvement of Data Quality, Ottawa, October 29–31, 337–347.

Received February 1992
Revised May 1994