JACKKNIFE VARIANCE ESTIMATION WITH IMPUTED SURVEY DATA

J.N.K. Rao, Carleton University
Department of Mathematics & Statistics, Carleton University, Ottawa, Canada

KEY WORDS: Adusted imputed values, item nonresponse, stratified multistage sampling

Item nonresponse is usually handled by some form of imputation; in particular, deterministic or hot deck imputation is often used to assign values for missing item responses. We provide an account of our recent joint work on jackknife variance estimation based on adjusted imputed values, using only a single imputation and, hence, a single completed data set. We also present linearization versions of the proposed jackknife variance estimators which are asymptotically consistent under missing at random set-up. To implement the proposed variance estimators, the completed data set must carry identification flags to locate imputed and observed values. We study both simple random sampling and stratified multistage sampling.

1. INTRODUCTION

Item nonresponse is usually handled by some form of imputation. Two types of imputation are often used: (a) deterministic imputation which covers mean imputation, ratio and regression imputation and nearest neighbour imputation. The imputed values are deterministic, given the sample of respondents and any auxiliary information on nonrespondents. (b) Hot deck imputation which employs a sample drawn from the respondent values. Various versions of hot deck imputation have been proposed (Kalton, 1981, p.91; Sedransk, 1985). In the simplest form of hot deck imputation, a simple random sample is selected with replacement from the sample respondents to an item y, and the associated item values are used as donors. In practice, the accuracy of imputation is improved by first forming two or more imputation classes using auxiliary variables observed on all sample units, and then performing hot deck imputation separately within each imputation class for each item with missing values. Hot deck imputation has the following advantages: (i) it preserves the distribution of item values unlike mean imputation; (ii) results obtained from different analyses are consistent with one another, unlike the results of analyses from an incomplete data set; (c) it permits the use of same survey weight for all items, unlike the weighting adjustment method which is more appropriate for unit nonresponse.

It is a common practice to treat the imputed values as if they are true values, and then compute the variance estimates using standard formulae. This procedure, however, can lead to serious underestimation of the true variance of the estimates, when the proportion of missing values for the item of interest is appreciable. Rubin (1978) proposed multiple imputation to account for the inflation in the variance due to imputation. Multiple imputation leads to valid variance estimators when the imputation is "proper" in the sense that the imputed values are drawn from the posterior distribution of nonobserved y-values given the respondent values. However, it may not lead to consistent variance estimators for stratified multistage surveys in the common situation of imputation cutting across sample clusters (Fay, 1991). (Note that such imputations are not proper.) Moreover, several

statistical agencies seem to prefer single imputation, mainly due to operational difficulties in maintaining multiple complete data sets, especially in large-scale surveys.

Burns (1990) proposed jackknife variance estimation for multistage surveys, using pseudo-replicate hot deck imputation. This method uses independent imputations from the full sample of respondent values and from the samples of respondent values obtained by deleting each sampled cluster in turn. Unfortunately, Burns jackknife variance estimator can lead to serious overestimation (Rao and Shao, 1992).

In this paper, we provide an account of our recent joint work on jackknife variance estimation based on adjusted imputed values using only a single imputation and, hence, a single completed data set. To calculate the adjusted imputed values, the completed data set must carry identification flags to locate imputed and observed values. We also present linearized versions of the proposed jackknife variance estimators which are asymptotically consistent under missing at random set-up. These variance estimators are obtained by Taylor approximations to our jackknife variance estimators. An advantage of this approach is that the resulting linearization variance estimators retain certain important properties of the jackknife, unlike some other linearization variance estimators. The linearized versions can be implemented using software packages that employ the linearization (Taylor) method instead of the jackknife method for calculation of standard errors, such as SUDAAN and PC CARP.

We study both simple random sampling and stratified multistage sampling. Establishment surveys, based on list frames, often employ simple random sampling (within strata) while large-scale socio-economic surveys often use stratified multistage sampling.

2. DETERMINISTIC IMPUTATION

We focus on ratio and regression imputation. Nearest neighbour imputation can also be handled by the jackknife method, but limited simulation results (Zanutto, 1993) suggest that the jackknife variance estimate can lead to serious overestimation as the nonresponse rate increases, possibly due to nonsmoothness of imputed estimator.

2.1 *Simple Random Sampling*

Suppose in a simple random sample, s, of size n, m units respond to item y and $n-m$ do not. Let $\bar{y}_m$ be the mean for the respondents s_r and y_i^* be the imputed value for unit $i \in s - s_r$, the set of nonrespondents. The imputed estimator of the population mean, $\bar{Y}$, is then given by

$$\bar{y}_I = \frac{1}{n}\Big(\sum_{i\in s_r} y_i + \sum_{i\in s-s_r} y_i^*\Big). \qquad (2.1)$$

The jackknife variance estimator, under deterministic imputation, is calculated in the usual way except that when a respondent $j \in s_r$ is to be deleted, each of the imputed values y_i^* is adjusted by an amount $y_i^*(j) - y_i^*$, where $y_i^*(j)$ is the value one would impute for the i-th nonrespondent if j-th respondent is deleted from the sample. Thus, the adjusted imputed value equals the "correct" value $y_i^*(j)$ if $j \in s_r$ and remains unchanged, i.e. equals y_i^*, if a nonrespondent j is deleted. Throughout the paper, we assume that the completed data set carries identification flags to locate imputed and observed values.

Denote the imputed estimator based on the adjusted imputed values as $\bar{y}_I^a(j)$ when j-th sample unit is deleted. The jackknife variance estimator is then given by the standard formula applied to the adjusted estimators $\bar{y}_I^a(j)$:

$$v_J(\bar{y}_I) = \frac{n-1}{n}\sum_{j=1}^{n}\Big[\bar{y}_I^a(j) - \bar{y}_I\Big]^2. \qquad (2.2)$$

In the case of a nonlinear statistic of the form $\hat{\theta}_I = g(\bar{y}_I)$, the jackknife variance estimator is simply obtained from (2.2) by changing $\bar{y}_I^a(j)$ to $\hat{\theta}_I^a(j) = g[\bar{y}_I^a(j)]$ and $\bar{y}_I$ to $\hat{\theta}_I$. For simplicity, we have ignored the finite population correction in (2.2).

Ratio Imputation

Suppose an auxiliary variable, x, closely related to y is observed on all sample units, s. Ratio imputation uses $y_i^* = (\bar{y}_m/\bar{x}_m)x_i$ for the missing values, where $\bar{y}_m$ and $\bar{x}_m$ are the means of y- and x-values for the respondents s_r. In this case, the imputed estimator (2.1) reduces to

$$\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}, \tag{2.3}$$

where $\bar{x}$ is the x-mean for the full sample s. Under a uniform response mechanism, that is independent response across sample units and equal response probabilities p, the estimator (2.3) has the same properties as the standard two-phase sampling ratio estimator. This follows by noting that, conditionally given m, s_r is a simple random sample of fixed size m drawn from s. The estimator (2.3) is, therefore, approximately design (or p-) unbiased under uniform response.

Under ratio imputation, we have $y_i^*(j) = [\bar{y}_m(j)/\bar{x}_m(j)]x_i$, where $\bar{y}_m(j) = (m\bar{y}_m - y_j)/(m-1)$ and $\bar{x}_m(j) = (m\bar{x}_m - x_j)/(m-1)$. Using these values, the jackknife variance estimator is obtained from (2.2). Linearized version of the jackknife variance estimator is given by (Rao and Sitter, 1992):

$$v_L(\bar{y}_I) = \Big(\frac{\bar{x}}{\bar{x}_m}\Big)^2 \frac{A}{m} + 2\Big(\frac{\bar{x}}{\bar{x}_m}\Big)\frac{B}{n} + \frac{C}{n}, \tag{2.4}$$

where

$$A = \frac{1}{m-1}\sum_{i\in s_r}\Big(y_i - \frac{\bar{y}_m}{\bar{x}_m}x_i\Big)^2,$$

$$B = \Big(\frac{\bar{y}_m}{\bar{x}_m}\Big)\frac{1}{m-1}\sum_{i\in s_r}\Big(y_i - \frac{\bar{y}_m}{\bar{x}_m}x_i\Big)^2$$

and

$$C = \Big(\frac{\bar{y}_m}{\bar{x}_m}\Big)^2 \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2.$$

Under two-phase sampling, a standard p-consistent variance estimator is given by (Sukhatme and Sukhatme, 1970, p. 176):

$$v_0 = \Big(\frac{1}{m} - \frac{1}{n}\Big)A + \frac{1}{n}\Big[\frac{1}{m-1}\sum_{i\in s_r}\Big(y_i - \bar{y}_m\Big)^2\Big]. \tag{2.5}$$

The second term in (2.5) is simply obtained by using the sample variance $s_{y_m}^2 = (m-1)^{-1}\sum_{i\in s_r}(y_i - \bar{y}_m)^2$ to estimate the population variance S_y^2. An alternative p-consistent variance estimator that uses x-information to estimate S_y^2 is given by (Rao and Sitter, 1992):

$$v_1 = \frac{A}{m} + 2\frac{B}{n} + \frac{C}{n}. \tag{2.6}$$

It follows from (2.4) and (2.6) that $v_L(\bar{y}_I)$ is asymptotically equivalent to v_1 under uniform response, since $\bar{x}/\bar{x}_m \doteq 1$. Thus, $v_L(\bar{y}_I)$ and $v_J(\bar{y}_I)$ are p-consistent under uniform response.

Särndal (1992) assumed the following "imputation" model ξ:

$$E_\xi(y_i) = \beta x_i,\ V_\xi(y_i) = \sigma^2 x_i,\ \mathrm{cov}_\xi(y_i, y_j) = 0,\quad i \neq j \in s. \tag{2.7}$$

The imputed estimator (2.3) is design-model (or $p\xi$-) unbiased for $\bar{Y}$ irrespective of the response mechanism, provided the model also holds for the respondons s_r, i.e., if selection bias is absent. Särndal (1992) proposed the following approximately $p\xi$-unbiased linearization variance estimator, irrespective of the response mechanism:

$$v_S(\bar{y}_I) = \Big(\frac{\bar{x}}{\bar{x}_m}\Big)^2 \frac{A}{m} + 2\Big(\frac{m}{n}\Big)\frac{B}{n} + \frac{C}{n}. \tag{2.8}$$

Comparing (2.6) with (2.8), it follows, however, that $v_S(\bar{y}_I)$ is design-inconsistent under uniform response.

Comparing (2.4) with (2.8) and noting that $E_\xi B = 0$, it follows that the jackknife variance estimator and its linearized version are approximately $p\xi$-unbiased, irrespective of the response mechanism. Thus, the jackknife variance estimator (and its linearized version) remain p-consistent, unlike (2.8), under uniform response irrespective of any underlying model, as well as approximately $p\xi$-unbiased under model (2.7), irrespective of the response mechanism. This robustness property is an important feature of the jackknife method.

Our jackknife variance estimator (2.2) can also be used in standard two-phase sampling when "mass" imputation is used, i.e., when the y-values of the units not sampled at the second phase are imputed, using the first phase x-information. Whitridge and Kovar (1990) discuss the practical advantages of mass imputation and give applications in business surveys.

For nearest neighbour imputation, the proposed jackknife variance estimator, with adjusted values under ratio imputation, may be used as an approximation. Simulation results by Kovar and Chen (1992) indicate good performance of this variance estimator, provided the correlation between y and x is high.

Rao and Sitter (1992) considered jackknife variance estimation when the auxiliary variable, x, is not observed on all the sample units.

Regression Imputation

We again assume that x is observed on all sample units, s. Simple linear regression imputation uses $y_i^* = \bar{y}_m + \hat{\beta}_m(x_i - \bar{x}_m)$ for the missing values, where $\hat{\beta}_m$ is the usual least squares regression coefficient based on the respondents, s_r. In this case, the imputed estimator reduces to the standard double sampling regression estimator

$$\bar{y}_I = \bar{y}_m + \hat{\beta}_m(\bar{x} - \bar{x}_m). \qquad (2.9)$$

The estimator (2.9) is approximately p-unbiased under uniform response. It is also $p\xi$-unbiased under the "imputation" model

$$E_\xi(y_i) = \alpha + \beta x_i,\ V_\xi(y_i) = \sigma^2,$$
$$\mathrm{cov}_\xi(y_i, y_j) = 0, \quad i \neq j \in s \qquad (2.10)$$

irrespective of the response mechanism, provided the model also holds for the respondents, s_r.

Under regression imputation, we have

$$y_i^*(j) = \bar{y}_m(j) + \hat{\beta}_m(j)(x_i - \bar{x}_m(j)),$$

where $\hat{\beta}_m(j)$ is the least squares regression coefficient when the j-th sample unit is deleted.

Rao and Sitter (1992) have shown that the jackknife variance estimator remains p-consistent under uniform response irrespective of any underlying model, as well as approximately $p\xi$-unbiased under model (2.10), irrespective of the response mechanism. They have also given a linearized version of the jackknife variance estimator.

2.2. *Stratified Multistage Sampling*

Suppose we have L strata with n_h primary sampling units (psu) sampled from stratum h. Let n_{hi} be the number of ultimate units (elements) sampled from i-th psu in h-th stratum, and $n = \Sigma\Sigma n_{hi}$ be the total sample size. In the absence of nonresponse on item y, an estimator of population total Y is of the form

$$\hat{Y} = \sum_{(hik)\in s} w_{hik}\, y_{hik}, \qquad (2.10)$$

where s is the total sample, w_{hik} and y_{hik} respectively denote the survey weight and y- value attached to the (hik)-th element. Typically, the basic weights w_{hik} are subject to post-stratification adjustment. Here we confine ourselves to the basic weights and the resulting imputed estimators. Extensions to post-stratification and "calibration" regression estimators are being investigated by a Ph.D. student, Wesley Yung.

A customary estimator of variance of $\hat{Y}$ is given by

$$v(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2 = v(r_{hi}), \quad (2.11)$$

where

$$r_{hi} = \sum_k (n_h w_{hik}) y_{hik}, \quad \bar{r}_h = n_h^{-1} \sum_i r_{hi}$$

and the operator notation $v(r_{hi})$ denotes that $v(\hat{Y})$ depends only on the psu totals r_{hi}. This variance estimator is p-unbiased if the psu's are sampled with replacement, but generally it tends to overestimate the variance.

In the presence of nonresponse on item y, let y^*_{hik} be the imputed values for the nonrespondents under a specified imputation procedure. The imputed estimator of Y is given by

$$\hat{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s-s_r} w_{hik} y^*_{hik}, \quad (2.12)$$

where s_r is the sample of respondents.

A jackknife variance estimator is obtained by first calculating the adjusted imputed estimator $\hat{Y}^a_I(gj)$ when each sample psu (gj) is deleted in turn and then using the usual formula:

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g}{n_g - 1} \sum_{j=1}^{n_g} [\hat{Y}^a_I(gj) - \hat{Y}_I]^2. \quad (2.13)$$

Here $\hat{Y}^a_I(gj)$ is simply obtained from (2.12) by retaining the original weights w_{hik} for $h \neq g$, changing w_{gik} to $[n_g/(n_g-1)]w_{gik}$ for $i \neq k$, setting $w_{gjk} = 0$, and finally adjusting each of the imputed values, y^*_{hik}, by an amount $y^*_{hik}(gj) - y^*_{hik}$ where $y^*_{hik}(gj)$ is the value one would impute for the (hik)-th nonrespondent if (gj)-th sample psu is deleted from the sample. Thus, the adjusted imputed value equals the "correct" value $y^*_{hik}(gj)$ if (gj)-th sample psu is deleted.

Mean Imputation

Mean imputation is equivalent to customary weight adjustment. It uses $y^*_{hik} = \hat{S}/\hat{T}$, where

$$\hat{S} = \sum_{s_r} w_{hik} y_{hik}, \quad \hat{T} = \sum_{s_r} w_{hik}.$$

In this case, $\hat{Y}_I$ reduces to

$$\hat{Y}_I = (\hat{S}/\hat{T})\hat{U}$$

where $\hat{U}/\hat{T}$ is the nonrespondent adjustment factor and

$$\hat{U} = \sum_s w_{hik}.$$

Moreover, $\hat{Y}^a_I(gj)$ reduces to $[\hat{S}(gj)/\hat{T}(gj)]\hat{U}(gj)$, where $\hat{S}(gj)$ is obtained from $\hat{S}$ by changing the weights as described above, and $\hat{T}(gj)$ and $\hat{U}(gj)$ are similarly obtained from $\hat{T}$ and $\hat{U}$. The jackknife variance estimator is readily obtained from (2.13) using these adjusted imputed estimators.

By a Taylor expansion of $\hat{Y}^a_I(gj) - \hat{Y}_I$ around $(\hat{S}, \hat{T}, \hat{U})$, we get a linearized version of the jackknife variance estimator. It is given by

$$v_L(\hat{Y}_I) = v(\hat{r}_{hi}) \quad (2.14)$$

with

$$\hat{r}_{hi} = \frac{\hat{U}}{\hat{T}}(s_{hi} - \hat{R}t_{hi}) + \hat{R}u_{hi},$$

where

$$s_{hi} = \sum_{k \in s_r} (n_h w_{hik}) y_{hik}, \quad t_{hi} = \sum_{k \in s_r} (n_h w_{hik})$$

$$u_{hi} = \sum_{k \in s} (n_h w_{hik}), \ \hat{R} = \hat{S}/\hat{T}.$$

Thus, the linearized version is simply obtained from the standard variance formula (2.11) with r_{hi} changed $\hat{r}_{hi}$ which is a linear combination of cluster totals s_{hi}, t_{hi} and u_{hi}.

Ratio Imputation

Suppose covariate values x_{hik} closely related to y_{hik} are observed on all the elements in the sample s. In this case, ratio imputation uses $y^*_{hik} = (\tilde{S}/\tilde{T})x_{hik}$, where

$$\tilde{S} = \sum_{s_r} w_{hik} y_{hik}, \quad \tilde{T} = \sum_{s_r} w_{hik} x_{hik}.$$

The imputed estimator $\hat{Y}_I$ reduces to

$$\hat{Y}_I = (\tilde{S}/\tilde{T})\tilde{U},$$

where

$$\tilde{U} = \sum_{s} w_{hik} x_{hik}.$$

Further, $\hat{Y}^a_I(gj)$ reduces to $[\tilde{S}(gj)/\tilde{T}(gj)]$ $\tilde{U}(gj)$, where $\tilde{S}(gj)$, $\tilde{T}(gj)$ and $\tilde{U}(gj)$ are obtained from $\tilde{S}, \tilde{T}$ and $\tilde{U}$ by changing the weights as described above. The jackknife variance estimator is readily obtained from (2.13) using these adjusted imputed estimators.

A linearized version of the jackknife variance estimator is simply given by

$$v_L(\hat{Y}_I) = v(\tilde{r}_{hi}) \qquad (2.15)$$

with

$$\tilde{r}_{hi} = \frac{\tilde{U}}{\tilde{T}}(\tilde{s}_{hi} - \tilde{R}\tilde{t}_{hi}) + \tilde{R}\tilde{u}_{hi},$$

where

$$\tilde{s}_{hi} = \sum_{k \in s_r} (n_h w_{hik}) y_{hik},$$

$$\tilde{t}_{hi} = \sum_{k \in s_r} (n_h w_{hik}) x_{hik})$$

$$\tilde{u}_{hi} = \sum_{k \in s} (n_h w_{hik}) x_{hik},$$

$$\tilde{R} = \tilde{S}/\tilde{T}.$$

3. HOT DECK IMPUTATION

We consider ratio and regression hot deck imputation as well as imputing all missing item values from a common donor. The latter method preserves multivariate relationships (Kalton and Kasprzyk, 1986, p.11). It was used in the 1975 Canadian Census of Construction, as noted by Ford (1983). Under this method, we consider general parameters which cover population means, variances and covariances, correlation and regression coefficients, domain means and cell proportions in two-way tables.

3.1. *Simple Random Sampling*

The imputed estimator of $\bar{Y}$ and the jackknife variance estimator are again given by (2.1) and (2.2) except that y^*_i is adjusted by an amount $E^{-j}_* y^*_i - E_* y^*_i$ when a respondent j is deleted, where E_* denotes the same expectation with the donor set modified by excluding unit j. The imputed value y^*_i remains unchanged, as before, if a nonrespondent j is deleted.

Ratio Hot Deck

As in ratio imputation, we assume that an auxiliary variable, x, is observed on all sample units, s. In the ratio hot deck method, donors i are selected by simple random sampling with replacement from the respondents s_r, and the associated ratio residuals $e^*_i = y_i - (\bar{y}_m/\bar{x}_m)x_i$ are then added to the deterministic imputed values $(\bar{y}_m/\bar{x}_m)x_i$,

$i \in s - s_r$ to get the hot deck imputed values:

$$y_i^* = (\bar{y}_m/\bar{x}_m)x_i + e_i^*, \quad i \in s - s_r \quad (3.1)$$

Noting that $E_*(e_i^*) = 0$, the imputed estimator $\bar{y}_I$ is approximately p-unbiased under uniform response.

The adjusted imputed values used in the jackknife variance estimator (2.2) are given by $y_i^* + [\bar{y}_m(j)/\bar{x}_m(j)]x_i - (\bar{y}_m + \bar{x}_m)x_i$ if j-th respondent is deleted and remains y_i^* if j-th nonrespondent is deleted.

It can be shown that the linearization version of (2.2) is given by (Rao, 1993)

$$v_L(\bar{y}_I) = \quad (2.4) \qquad (3.2)$$
$$+ \left(1 - \frac{m}{n}\right)\left(2\frac{D_e^*}{n} + \frac{s_e^{*2}}{n} + \frac{m}{n^2}\bar{e}_{n-m}^{*2}\right),$$

where

$$D_e^* = \left(\frac{\bar{y}_m}{\bar{x}_m}\right)\frac{1}{n-m-1}\sum_{s-s_r} e_i^*(x_i - \bar{x})$$

$$s_e^{*2} = \frac{1}{n-m-1}\sum_{s-s_r}(e_i^* - \bar{e}_{n-m}^*)^2$$

and $\bar{e}_{n-m}^*$ is the mean of e_i^*s. It follows from (3.2) that the linearization variance estimator is obtained by adding a term due to hot deck to the formula under ratio imputation.

The $p\xi$-properties of $v_L(\bar{y}_L)$ under "imputation" model (2.7) are being investigated.

Regression Hot Deck

As in the ratio hot deck, donors i are selected by simple random sampling with replacement from the respondents s_r. The associated regression residuals $e_i^* = (y_i - \bar{y}_m) - \hat{\beta}_m(x_i - \bar{x}_m)$ are then added to the deterministic imputed values $\bar{y}_m + \hat{\beta}_m(x_i - \bar{x}_m)$, $i \in s - s_r$ to get the hot deck imputed values:

$$y_i^* = \bar{y}_m + \hat{\beta}_m(x_i - \bar{x}_m) + e_i^*, \quad i \in s - s_r. \qquad (3.3)$$

Noting that $E_*(e_i^*) = 0$, the imputed estimator $\bar{y}_I$ is approximately p-unbiased under uniform response.

The adjusted imputed values used in the jackknife variance estimator (2.2) are given by $y_i^* + \{\bar{y}_m(j) + \hat{\beta}_m(j)(x_i - \bar{x}_m(j))\} - \{\bar{y}_m + \hat{\beta}_m(x_i - \bar{x})\}$ if j-th respondent is deleted and remains y_i^* if j-th nonrespondent is deleted.

A linearized version of the jackknife variance estimator is given by Rao (1993).

Common Donor Hot Deck

Skinner and Rao (1993) consider general parameters either of the form

$$\bar{Z} = N^{-1}\sum_U z(y_{1i}, y_{2i}) \qquad (3.4)$$

for some function $z(\cdot,\cdot)$, where (y_{1i}, y_{2i}) is a pair of values associated with unit i in a finite population U of size N, or of the form $g(\bar{Z})$ for some function $g(\cdot)$. These parameters cover means, variances and covariances, correlation and regression coefficients, domain means and cell proportions in a two-way table. For example, a domain mean is of the form $N^{-1}\sum_U y_{1i}y_{2i}/\bar{Y}_1$ if y_{1i} is an indicator variable taking the value 1 when i belongs to the domain of interest and 0 otherwise.

The subsample responding to both y_{1i} and y_{2i} is denoted by s_{rr} with $s_{r\bar{r}}$, $s_{\bar{r}r}$ and $s_{\bar{r}\bar{r}}$ defined similarly and with n_{rr}, $n_{r\bar{r}}$, $n_{\bar{r}r}$ and $n_{\bar{r}\bar{r}}$ denoting their respective sizes. A common donor hot deck uses with imputed values drawn by simple random sampling with replacement from s_{rr}, the donor set. All missing item values of a recipient are imputed from the responses of the donor.

The imputed estimator of $\bar{Z}$ is given by

$$\bar{z}_I = n^{-1}\sum_s z(\tilde{y}_{1i}, \tilde{y}_{2i}), \qquad (3.5)$$

where $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ denote either the actual response or the imputed value if the former is missing. Assuming uniform response, Skinner and Rao (1993) have shown that the conditional expectation of $\bar{z}_I$, given n_{rr}, $n_{\bar{r}r}, n_{r\bar{r}}$, is $a\ \bar{Z} + (1-a)\overline{\overline{Z}}$, where $a = (n_{rr} + n_{\bar{r}\bar{r}})/n$ and

$$\overline{\overline{Z}} = [N(N-1)]^{-1} \sum_{i,j \in U: i \neq j} z(y_{1i}, y_{2i}). \tag{3.6}$$

In general, $\bar{Z} \neq \overline{\overline{Z}}$ so that $\bar{z}_I$ is p-biased. Note that if both y_{1i} and y_{2i} are either observed or missing, i.e. if $n_{\bar{r}r} = n_{r\bar{r}} = 0$, then $a = 1$ and $\bar{z}_I$ will be p-unbiased under uniform response.

Adjusted imputed values, $z_i^a(j)$, are defined as follows:

$$\begin{aligned} z_i^a(j) &= \tilde{z}_i + E_*^{-j}\tilde{z}_i - E_*\tilde{z}_i \text{ if } j \in s_{rr} \\ &= \tilde{z}_i \text{ if } j \notin s_r, \text{ if } j \notin s_{rr}, \end{aligned} \tag{3.7}$$

where E_*^{-j} and E_* are defined as before and $\tilde{z}_i = z(\tilde{y}_{1i}, \tilde{y}_{2i})$. A p-consistent, jackknife estimator of variance of $\bar{z}_I$ is then given by

$$v_J(\bar{z}_I) = \frac{n-1}{n} \sum_s [\bar{z}_I^a(j) - \bar{z}_I]^2, \tag{3.8}$$

where $\bar{z}_i^a(j)$ is the imputed estimator based on the adjusted imputed values when j-th sample unit is deleted:

$$\bar{z}_I^a(j) = (n-1)^{-1} \sum_{i \in s: i \neq j} z_i^a(j). \tag{3.9}$$

Skinner and Rao (1993) also consider an adjusted estimator of $\bar{Z}$ which is asymptotically unbiased.

A jackknife variance estimator of $g(\bar{z}_I)$ is readily obtained from (3.8) by changing $\bar{z}_I^a(j)$ and $\bar{z}_I$ to $g[\bar{z}_I^a(j)]$ and $g(\bar{z}_I)$ respectively, where $\bar{z}_I$ could be a vector of imputed estimators.

Linearized versions of the jackknife variance estimators are currently being investigated.

3.2. *Stratified Multistage Sampling*

For simplicity, we consider only the case of a single imputation class with uniform response mechanism. The results, however, may be extended to multiple imputation classes, along the lines of Rao and Shao (1992), by allowing for separate imputation within different imputation classes.

The imputed estimator of Y and the jackknife variance estimator are again given by (2.12) and (2.13) except that y_{hik}^* is adjusted by an amount $E_*^{-gj} y_{hik}^* - E_* y_{hik}^*$ when (gj)-th psu is deleted, where E_* denotes the expectation with respect to hot deck imputation and E_*^{-gj} denotes the same expectation with the donor set modified by excluding (gj)-th psu.

Simple Hot Deck

The estimator $\hat{Y}_I$ will be biased if simple random sampling is used to select the donors from s_r, the set of respondents to item y, unless y_{hik}^* is chosen as $y_{gj\ell}(w_{gj\ell}/w_{hik})$, where $(gj\ell) \in s_r$ is the selected donor (Platek and Gray, 1983). The latter choice, however, may not be practically appealing when y takes only integer values, for example, when $y = 0$ or 1. A simple alternative, proposed by Rao and Shao (1992), is to select the donors $(gj\ell) \in s_r$ with replacement with probabilities $w_{gj\ell}/\sum_{s_r} w_{hik}$ and use $y_{hik}^* = y_{gj\ell}$. Under this hot deck scheme and uniform response, the imputed estimator $\hat{Y}_I$ is asymptotically unbiased for Y as $n = \sum n_h \to \infty$.

The adjusted imputed values are obtained by adjusting y_{hik}^* by an amount $\hat{S}(gj)/\hat{T}(gj) - \hat{S}/\hat{T}$. Rao and Shao (1992) have shown that the resulting jackknife variance estimator (2.13) is asymptotically consistent as $n \to \infty$.

A linearized version of the jackknife variance estimator is obtained using a Taylor expansion of $\hat{Y}_I^a(gj) - \hat{Y}_I$ around $(\hat{S}, \hat{T}, \hat{U})$. It is simply given by (Rao, 1993)

$$v_L(\hat{Y}_I) = v(r_{hi}^*), \qquad (3.10)$$

where

$$r_{hi}^* = \hat{r}_{hi} + a_{hi}^*$$

and

$$a_{hi}^* = \sum_{k \in s - s_r} (n_h w_{hik})(y_{hik}^* - \hat{S}/\hat{T}).$$

That is, we simply add a_{hi}^* to the deterministic component $\hat{r}_{hi}$, defined below (2.14), to get r_{hi}^* and then use the standard formula (2.11) with r_{hi} changed to r_{hi}^*.

Simulation results by Kovar and Chen (1992) that the adjusted jackknife variance estimator has negligible relative bias for all values of response probability p, confirming its asymptotic consistency.

Zanutto (1993) extended the results of Rao and Shao (1992) on estimating Y or $g(Y)$ to the case of common donor hot deck.

Ratio Hot Deck

Let $e_{hik} = y_{hik} - \tilde{R}x_{hik}$ be the respondent residuals, $(hik) \in s_r$. We select donors as before and add the donor values e_{hik}^* to the deterministic imputed values $\tilde{R}x_{hik}$ to get the hot deck imputed values:

$$y_{hik}^* = \tilde{R}x_{hik} + e_{hik}^*, \quad (hik) \in s - s_r. \qquad (3.11)$$

In this case, the imputed estimator $\hat{Y}_I$ is approximately unbiased under uniform response, noting that $E_* e_{hik}^* = 0$. The adjusted values used in the jackknife variance estimator (2.13) are given by $y_{hik}^* + [\tilde{S}(gj)/\tilde{T}(gj) - \tilde{S}/\tilde{T}]x_{hik}$ when (gj)-th psu is deleted.

A linearized version of the jackknife variance estimator is simply given by

$$v_L(\hat{Y}_I) = v(\tilde{r}_{hi}^*), \qquad (3.12)$$

where

$$\tilde{r}_{hi}^* = \tilde{r}_{hi} + \tilde{a}_{hi}^*$$

and

$$\tilde{a}_{hi}^* = \sum_{k \in s - s_r} (n_h w_{hik}) e_{hik}^*.$$

That is, we simply add $\tilde{a}_{hi}^*$ to the deterministic component $\tilde{r}_{hi}$, defined below (2.15), to get $\tilde{r}_{hi}^*$ and then use the standard formula (2.11) with r_{hi} changed to $\tilde{r}_{hi}^*$.

Extensions to general parameters $\bar{Z}$ of the form (3.4) under common donor hot deck are being investigated.

REFERENCES

Burns, R.M. (1990). Multiple and replicate item imputation in a complex sample survey. In *Proc. Sixth Annual Res. Conf.*, pp. 655-65. Washington, D.C.: U.S. Bureau of the Census.

Fay, R.E. (1991). A design-based perspective on missing data variance. In *Proc. Seventh Annual Res. Conf.*, pp. 429-440. Washington, D.C.: U.S. Bureau of the Census.

Ford, B.L. (1983). An overview of hot-deck procedures. In Incomplete Data in Sample Surveys, Vol. 2 (W.G. Madow, I. Olkin and D.B. Rubin eds.), pp. 185-207. New York: Academic Press.

Kalton, G. (1981). *Compensating for Missing Data*, ISR research report series. Ann Arbor. Survey Research Center, University of Michigan.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, 1-16.

Kovar, J.G. and Chen, E.J. (1992). Variance estimation under imputation: an empirical investigation. Technical Report. Business Survey Methods Division, Statistics Canada, Ottawa.

Rao, J.N.K. (1993). Linearized variance estimators under imputation for missing data. Technical Report, Business Survey Methods Division, Statistics Canada, Ottawa.

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. (1992). Jackknife variance estimation under imputation for missing survey data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Rubin, D.B. (1978). Multiple imputation in sample surveys — a phenomenological Bayesian approach to nonresponse. In *Proc. Sect. Survey Res. Meth.*, pp. 20-34. Washington, D.C.: American Statistical Association.

Särndal, G.E. (1992). Methods of estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241-252.

Sedransk, J. (1985). The objective and practice of imputation. In *Proc. First Annual Res. Conf.*, pp. 445-452. Washington, D.C.: Bureau of the Census.

Skinner, C.J. and Rao, J.N.K. (1993). Jackknife variance estimation for multivariate statistics under hot deck imputation. Paper to be presented at the ISI meetings, Florence, Italy, Aug. 1993.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*, 2nd ed., London: Asia Publishing House.

Whitridge, P. and Kovar, J.G. (1990). Use of mass imputation to estimate for subsample variables. In *Proc. Business and Economic Statistics*, pp. 132-137, Washington, D.C.: American Statistical Association.

Zanutto, E. (1993). *Jackknife Variance Estimation Under Imputation for Missing Survey Data.* Masters Thesis, Carleton University, Ottawa.