VARIANCE ESTIMATION FOR SAMPLES WITH RANDOM IMPUTATION

Margot Tollefson and Wayne A. Fuller
Margot Tollefson, Department of Statistics, Iowa State University, Ames, IA 50011

Key words: Nonresponse, survey samples, missing at random, imputation classes

1. Introduction

In survey sampling, practitioners are often faced with the problem of not being able to take measurements on some of the units, or on some items for some units, that are in the chosen sample. This is called the problem of nonresponse or missing data. When nonresponse is present, there can be serious problems with bias in estimators of population quantities if nothing is done to adjust the responding parts of the sample for the nonresponse. Random imputation is one form of adjustment for nonresponse. Our work on random imputation is an extension of work by Hansen, Hurwitz, and Madow (1953), Kalton (1983), and Little and Rubin (1987). These authors studied the variance of the usual estimator of the population mean constructed with imputed values in the place of missing values. They derived the variance of the usual estimator for the population mean for simple random sampling without replacement. Different authors employed different models. We extend the results to general designs, under the nonresponse model of Little and Rubin (1987).

Let there be a finite population of N units, where N is known. Let the population be divided into H exhaustive and mutually exclusive imputation classes of size N_h, h=1,...,H . Let there be a characteristic associated with each unit, $\{Y_{hi}\}_{h=1,..,H,\ i=1,..,N_h}$, and let

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi} \qquad (1)$$

be the overall finite population total for the characteristic.

Assume a probability sample of size n is selected from the population, where the only conditions on the sample are that the probabilities of inclusion are greater than zero for all units in the population and that the joint probabilities of inclusion are greater than zero for all pairs of units in the population.

Let n_h denote the number of units in the sample that fall in imputation class h . Suppose that, within imputation class h , we are able to measure the characteristic for r_h of the units and unable to measure the characteristic for m_h of the units. Let $m_h = k_h r_h + t_h$, where k_h and t_h are nonnegative integers and $t_h < r_h$. Let

$$R = \begin{bmatrix} r_1 & m_1 \\ \vdots & \vdots \\ r_H & m_H \end{bmatrix} \qquad (2)$$

be the matrix of the number of units that responded and the number of units that are missing for all of the H imputation classes.

Let the portion of the sample that is in imputation class h be written $\{Y_{hi}\}_{i=1}^{n_h}$, where $Y_{h1},...,Y_{hr_h}$ are the values for the units that responded and $Y_{h,r_h+1},...,Y_{hn_h}$ are the values for the units for which there was no response. Let the augmented sample in imputation class h be $\{Y^*_{hi}\}_{i=1}^{n_h}$, where $Y^*_{hi} = Y_{hi}$ for $i=1,...,r_h$ and Y^*_{hi} is an imputed value for $i=r_h+1,...,n_h$. Let π_{hi}, h=1,...,H, $i=1,...,n_h$ be the probability that sample unit hi is chosen in the random sample. Let $\pi_{(hi)(gj)}$, h=1,...,H, $i=1,...,n_h$, g = 1, ..., H, $j=1,...,n_g$, $hi \neq gj$ be the probability that both sample units hi and gj are in the sample.

Assume that, within a given class, the missing values are missing at random.

Definition: Missing values are missing at random within an imputation class if the mechanism by which they are missing is equivalent to the selection of an equal probability sample from the intended sample within the imputation class.

Let the imputation method be as follows. Each respondent in imputation class h is used at least k_h times for imputation and t_h of the respondents in imputation class h are used $k_h + 1$ times. The t_h of the respondents that are used $k_h + 1$ times are chosen by simple random sampling without replacement from the respondents in imputation class h . The donor respondents are assigned to the missing units randomly.

2. The estimator and its expected value in a finite population

In this section we consider the random imputation procedure using imputation classes. Let

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi}^{-1} Y_{hi} \qquad (3)$$

be the Horvitz—Thompson estimator of the population total based on the original complete sample. Let

$$\hat{Y}^* = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi}^{-1} Y_{hi}^* \qquad (4)$$

be the Horvitz—Thompson estimator of the population total constructed with the augmented sample. We now find the conditional expected value of $\hat{Y}^*$ for the finite population under the assumption that the missing values are missing at random within imputation classes.

Theorem 1: Assume that a sample is taken from a finite population, as described in Section 1. Let $\hat{Y}$ and $\hat{Y}^*$ be as defined in (3) and (4). Assume that the finite population is made up of H mutually exclusive and exhaustive classes and that, within a class, the nonrespondents are missing at random from the portion of the sample that falls in that class. Let the imputation be done as described in Section 1. Then

$$E(\hat{Y}^* \,|\, FP, S, R) = \hat{Y}$$

$$+ \sum_{h=1}^{H} \frac{m_h}{n_h - 1} \sum_{i=1}^{n_h} \pi_{hi}^{-1}(\bar{y}_h - Y_{hi}) , \qquad (5)$$

where FP denotes the finite population, S stands for the intended sample, **R** is the number responding and missing as defined in (2), and $\bar{y}_h$ is the mean of the intended sample in imputation class h .

Proof: If the set of respondents and the set of nonrespondents is held constant, the expected value of the estimator over all possible imputation patterns is

$$E(\hat{Y}^* \,|\, FP, S, rp)$$

$$= \sum_{h=1}^{H} \left[\sum_{i=1}^{r_h} \pi_{hi}^{-1} Y_{hi} + \sum_{i=r_h+1}^{n_h} \pi_{hi}^{-1} \bar{y}_{rh} \right] , \qquad (6)$$

where rp stands for the response pattern and $\bar{y}_{rh}$ is the mean of the respondents in imputation class h . Because the probability of a response is equal for all elements within a cell

$$E\left\{ \sum_{i=1}^{r_h} \pi_{hi}^{-1} Y_{hi} \,\middle|\, FP, S, R \right\}$$

$$= n_h^{-1} r_h \sum_{i=1}^{n_h} \pi_{hi}^{-1} Y_{hi} \qquad (7)$$

and

$$E\left\{ \sum_{i=r_h+1}^{n_h} \pi_{hi}^{-1} \bar{y}_{rh} \,\middle|\, FP, S, R \right\}$$

$$= E \left\{ \sum_{i=r_h+1}^{n_h} \pi_{hi}^{-1} r_h^{-1} \sum_{j=1}^{r_h} Y_{hj} \,\middle|\, FP, S, R \right\}$$

$$= r_h^{-1} r_h m_h n_h^{-1} (n_h - 1)^{-1}$$

$$\times \sum_{i=1}^{n_h} \pi_{hi}^{-1} \sum_{\substack{j=1\\ i\neq j}}^{n_h} Y_{hj}$$

$$= n_h^{-1} m_h \sum_{i=1}^{n_h} \pi_{hi}^{-1}(n_h - 1)^{-1}(n_h \bar{y}_h - Y_{hi}) . \tag{8}$$

Therefore

$$E\left\{\sum_{i=1}^{r_h} \pi_{hi}^{-1} Y_{hi} + \sum_{i=r_h+1}^{n_h} \pi_{hi}^{-1} \bar{y}_{rh} \,|\, \text{FP, S, R}\right\}$$

$$= \sum_{i=1}^{n_h} \pi_{hi}^{-1} Y_{hi} + \frac{m_h}{n_h - 1} \sum_{i=1}^{n_h} \pi_{hi}^{-1}(\bar{y}_h - Y_{hi}) . \tag{9}$$

□

By Theorem 1, we see that the imputation procedure is unbiased for the finite population total for simple random sampling. Since $E[\Sigma_{h=1}^{H}(n_h-1)^{-1} m_h \Sigma_{i=1}^{n_h} \pi_{hi}^{-1}(\bar{y}_h - Y_{hi}) \,|\, \text{FP}]$ is not generally equal to zero, $\hat{Y}^*$ is not generally an unbiased estimator of the population total in a finite sample. To derive unbiasedness for $\hat{Y}^*$, we add the assumption of an underlying superpopulation to our set of assumptions.

3. The expected value and variance of $\hat{Y}^*$ under the superpopulation model

Assume that there is a superpopulation made up of H subpopulations. Assume that within subpopulation h , h=1,...,H, the elements of the subpopulation are identically and independently distributed with mean μ_h and variance σ_h^2. Assume that between subpopulations, the elements are independent. Assume that equal probability samples from the subpopulations of sizes N_h, h=1,...,H , form the H imputation classes of the finite population.

Let

$$\bar{\mu} = \sum_{h=1}^{H} N^{-1} N_h \mu_h \tag{10}$$

be the mean of the finite population under the superpopulation structure. The estimator (4) is unbiased under the superpopulation model.

Theorem 2: Let the assumptions of Theorem 1 hold. Assume that the superpopulation structure of this section holds. Assume $n_h \geq 2$ for all h and all samples. Then

$$E(\hat{Y}^*) = N\bar{\mu} , \tag{11}$$

$$E(\hat{Y}^* - Y) = 0 , \tag{12}$$

$$V(\hat{Y}^* - N\bar{\mu}) = V(\hat{Y} - N\bar{\mu}) + A_h , \tag{13}$$

and,

$$V(\hat{Y}^* - Y) = V(\hat{Y} - Y) + A_h , \tag{13a}$$

where

$$A_h = \sum_{h=1}^{H} \sigma_h^2 E\left\{B_h\left[\sum_{\substack{i=1\\ i\neq j}}^{n_h} \sum_{j=1}^{n_h} \pi_{hi}^{-1} \pi_{hj}^{-1}\right]\right\} , \tag{14}$$

and

$$B_h = \frac{k_h n_h + (k_h + 2)t_h}{n_h(n_h - 1)} .$$

Proof: Under the model, every element in the h–th class of the finite population (and, thus, the sample) has the same expected value when the expectation is over draws from the superpopulation. It follows that expressions (11) and (12) hold.

To find the variance of $\hat{Y}^*$, we use three levels of conditioning. At the lowest level (level 3), we hold the choice of the sample (cs) , the response pattern (rp) , the choices for imputation (ic) , and the choices of which donor goes with which missing value (dm) constant, and take the expectations with respect to draws from the superpopulation. At level 2 we hold the choice of the sample (cs) and the n_h's and the r_h's (R) constant and average over the possible response patterns (given R) and over the possible imputation choices and over the possible choices of which donors go with which missing values. At level 1 we average over everything left random in the expression. We never actually evaluate the variance at level 1.

It is straightforward to show that

$$V(\hat{Y} - N\bar{\mu}) = V_1\left\{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}^{-1}\mu_h\right\} + E_1\left\{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}^{-2}\sigma_h^2\right\} . \qquad (15)$$

We now find the variance of $\hat{Y}^* - N\bar{\mu}$ in terms of $V(\hat{Y} - N\bar{\mu})$. We have

$$V(\hat{Y}^* - N\bar{\mu}) = V_1\{E_2[E_3(\hat{Y}^*)]\} + E_1\{V_2[E_3(\hat{Y}^*)]\} + E_1\{E_2[V_3(\hat{Y}^*)]\} . \qquad (16)$$

Also,

$$E_3\{\hat{Y}^* \mid cs, rp, ic, dm\} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}^{-1}\mu_h , \qquad (17)$$

since the Y_{hi}^*'s in the augmented sample are identically distributed in imputation class h. The Y_{hi}^*'s are identically distributed in imputation class h since the missing values are missing at random and the imputed values are chosen and assigned randomly within the imputation classes and since the real values in subpopulation h are identically distributed. It follows that

$$V(\hat{Y}^*) = V_1\left\{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}^{-1}\mu_h\right\} + E_1\{E_2[V_3(\hat{Y}^* \mid cs, rp, ic, dm) \mid cs, R]\} . \qquad (18)$$

The

$$E_2\left[V_3\left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}^{-1}Y_{hi}^* \middle| cs, rp, ic, dm\right] \middle| cs, R\right]$$

$$= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\pi_{hi}^{-1}\pi_{hj}^{-1}E_2\{H_3(Y_{hi}^*, Y_{hj}^*) \mid cs, R\} ,$$

where

$$H_3(Y_{hi}^*, Y_{hj}^*) = Cov_3(Y_{hi}^*, Y_{hj}^* \mid cs, rp, ic, dm) , \qquad (19)$$

since the Y_{hi}^*'s are independent between the imputation classes. Now the

$$H_3(Y_{hi}^*, Y_{hi}^*) = \sigma_h^2 \qquad (20)$$

for all hi , since the units in the augmented sample in imputation class h are identically distributed. Also, because, given the choice of the sample, the $H_3(Y_{hi}^*, Y_{hj}^*)$'s are identically distributed for $i = 1, \ldots, n_h - 1$, $j = i+1, \ldots, n_h$,

$$E_2[H_3(Y_{hi}^*, Y_{hj}^*) \mid cs, R]$$

$$= E_2\left[\sum_{\substack{\ell=1 \\ \ell \neq m}}^{n_h}\sum_{m=1}^{n_h}\frac{H_3(Y_{h\ell}^*, Y_{hm}^*)}{n_h(n_h - 1)} \middle| cs, R\right]$$

$$= E_2\,[B_h\,\sigma_h^2 \mid cs, R\,]$$

$$= B_h\,\sigma_h^2 \qquad (21)$$

for $i \neq j$. The last two expressions in (21) follow since each of the n_h elements in imputation class h in the augmented sample appears at least $(k_h + 1)$ times in the augmented sample and $(k_h + 2)t_h$ of the elements appear one more time and since we need to subtract out the n_h of the elements that are represented by expression (20). The sum in expression (21) is independent of the sample (given R), the response pattern (given R), the imputation choices, and the choices of which donors go with which missing values. From expressions (19), (20), and (21), the

$$E_2[V_3(\hat{Y}^* \mid cs, rp, ic, dm) \mid cs, R]$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi}^{-2} \sigma_h^2$$

$$+ \sum_{h=1}^{H} \sigma_h^2 B_h \left[\sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ i \neq j}}^{n_h} \pi_{hi}^{-1} \pi_{hj}^{-1} \right] . \quad (22)$$

Expression (13) follows from expressions (15), (18), and (22).

Expression (13a) follows from expression (13), since $\text{Cov}(\hat{Y}, Y)$ equals $\text{Cov}(\hat{Y}^*, Y)$ by the result in Theorem 1. □

4. Estimation of the variance of $\hat{Y}^* - Y$

We now look at the estimation of the variance of $\hat{Y}^* - Y$. We start by noting that

$$V(\hat{Y} - Y) = E[V(\hat{Y} - Y \mid FP)] , \quad (23)$$

since

$$E(\hat{Y} - Y \mid FP) = 0 . \quad (24)$$

Cochran (1977) gives

$$\hat{V}(\hat{Y} - Y \mid FP) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (1 - \pi_{hi}) \pi_{hi}^{-2} Y_{hi}^2$$

$$+ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{\substack{g=1 \\ hi \neq gj}}^{H} \sum_{j=1}^{n_g} \omega_{higj} Y_{hi} Y_{gj},$$

where

$$\omega_{higj} = \frac{\pi_{(hi)(gj)} - \pi_{hi}\, \pi_{gj}}{\pi_{hi}\, \pi_{gj}\, \pi_{(hi)(gj)}} , \quad (25)$$

as an unbiased estimator of $V(\hat{Y} - Y \mid FP)$.

In Theorem 2, we presented the variance of $\hat{Y}^*$ in terms of the variance of the estimator of the population total based on the intended sample and a term for the increase in variance due to imputation. From Theorem 2 and expression (23), under the superpopulation model, a natural estimator for $V(\hat{Y}^* - Y)$ is

$$\hat{V}_1(\hat{Y}^* - Y) = \hat{V}^*(\hat{Y} - Y \mid FP)$$

$$+ \sum_{h=1}^{H} s_{rh}^2 B_h \left[\sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ i \neq j}}^{n_h} \pi_{hi}^{-1} \pi_{hj}^{-1} \right] , \quad (26)$$

where s_{rh}^2 is the sample variance of the Y's for the responding units in class h, $h=1,\dots,H$ and where $\hat{V}^*(\hat{Y} - Y \mid FP)$ is $\hat{V}(\hat{Y} - Y \mid FP)$ calculated using the augmented sample instead of the intended sample. If $W_{hi} = \pi_{hi}^{-1}$ is the design weight on unit hi, $h=1,\dots,H$, $i=1,\dots,n_h$, then

$$\hat{Y}^* = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^* \quad (27)$$

and

$$\hat{V}_1(\hat{Y}^* - Y) = \hat{V}^*(\hat{Y} - Y \mid FP)$$

$$+\sum_{h=1}^{H} s^2_{r_h} B_h \left[\left[\sum_{i=1}^{n_h} W_{hi}\right]^2 - \sum_{i=1}^{n_h} W^2_{hi}\right] . \tag{28}$$

Estimator (26) is biased because $\hat{V}^*(\hat{Y}-Y|FP)$ is a biased estimator of $V(\hat{Y}-Y|FP)$. The bias is given in Lemma 1.

Lemma 1: Let the assumptions of Theorem 2 hold. Then

$$E[\hat{V}^*(\hat{Y}-Y|FP)] = V(\hat{Y}-Y)$$

$$+\sum_{h=1}^{H} \sigma^2_h E\left\{B_h \sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ i \neq j}}^{n_h} \omega_{hihj}\right\} . \tag{29}$$

Proof: See Tollefson (1992) for the proof. □

From the results in Theorem 2 and Lemma 1, an unbiased estimator of the $V(\hat{Y}^*-Y)$ is

$$\hat{V}_2(\hat{Y}^*-Y) = \hat{V}^*(\hat{Y}-Y|FP)$$

$$+\sum_{h=1}^{H} s^2_{rh} B_h \left[\sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ i \neq j}}^{n_h} \pi^{-1}_{(hi)(hj)}\right] . \tag{30}$$

Expression (30) is obtained from expression (26) by replacing $(\pi_{hi}\pi_{hj})^{-1}$ with $(\pi_{(hi)(hj)})^{-1}$ to take into account the bias in $\hat{V}^*(\hat{Y}-Y|FP)$. In practice, it is preferable to use the inverses of the marginal probabilities of selection rather than the inverses of the joint probabilities of selection in calculations because marginal probabilities of selection are generally available in the data files, while joint probabilities generally are not available in the data files. For most designs, the bias in estimating $V(\hat{Y}-Y)$ using an imputed sample is negligible in comparison to the increase in variance due to imputation.

ACKNOWLEDGEMENTS

This research was partly supported by the U.S. Department of Agriculture's Soil Conservation Service under SCS Cooperative Agreement No. 68–3A75–2–64.

REFERENCES

Cochran, W. (1977). Sampling Techniques, 3rd ed., Ch. 13, Wiley, New York.

Hanson, M. H., Hurwitz, W. N., and Madow, W. G. (1953). Sample Survey Methods and Theory. Volume II Theory. 139–141. Wiley, New York.

Kalton, G. (1983). Compensating for Missing Survey Data. Survey Research Center, The University of Michigan, Ann Arbor.

Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. Ch. 4. Wiley, New York.

Tollefson, M. (1992). Variance Estimation Under Random Imputation. Unpublished dissertation, Iowa State University