Brazilian Journal of Probability and Statistics
2011, Vol. 25, No. 2, 171–182
DOI: 10.1214/10-BJPS115

Estimating the population mean when some responses are missing

Jingsong Lu[a], **Edward J. Stanek III**[b] **and Elaine Puleo**[b]

[a]*Adheris, Inc.*
[b]*University of Massachusetts*

Abstract. We develop a design-based prediction approach to estimate the finite population mean in a simple setting where some responses are missing. The approach is based on indicator sampling random variables that operate on labeled units (subjects). We define missing data mechanisms that may depend on a subject, or on a selection (such as when the study design assigns groups of selected subjects to different interviewers). Using an approach usually reserved for model-based inference, we develop a predictor that equals the sample total divided by the expected sample size. The methods are based on best linear unbiased prediction in finite population mixed models. When the probability of missing is estimated from the sample, the empirical estimator simplifies to the mean of the realized nonmissing responses. The different missing data mechanisms are revealed by the notation that accounts for the labels and sample selections. The mean squared error (MSE) of the empirical estimator, counterintuitively, is smaller than the MSE if the probability of missing is known.

1 Introduction

Statistical analysis in the presence of incomplete or missing data is a pervasive problem in sample surveys. A simple example illustrates the problem. Suppose that a voter opinion poll is conducted via a simple random telephone sample selected from a list of registered voters. Although a sample of size n is selected, response will most likely be obtained on $n_1 < n$ selected subjects. Some of the registered voters will have answering machines and screen calls, resulting in nonresponse. In addition, poor interviewing skills by some interviewers may result in refusals for other contacted subjects. The first type of nonresponse depends on the subject, while the second type of nonresponse depends on the interviewer.

In the simplest setting, the probability of nonresponse will be unrelated to the actual voter preference of the subject. If this is true, the missing responses are called missing completely at random (MCAR) (Little, 1987). For example, if the proportion of registered voters who screen calls among those who would vote for a candidate is the same for all candidates, then the missing responses are MCAR.

Key words and phrases. Simple random sampling, missing data, MCAR, finite population, best linear unbiased estimator (BLUE), prediction.

Received July 2009; accepted January 2010.

Also, if the proportion of refusals that result from poor interviewer skills is the same for voters of all candidates, the missing responses are MCAR. MCAR is the simplest kind of nonresponse assumption and it is often assumed as a starting point in an analysis, as we do here.

How should one estimate the voter preference for a candidate when response for some of the selected sample subjects is missing? An intuitive estimator is the simple proportion (i.e., mean) of the n_1 responding subjects who would vote for the candidate. This is the estimator described by Cochran (1977). Although intuition is a good guide in selecting this estimator, its statistical properties are difficult to derive, since the denominator is a random variable. A way around this complication is to condition on the observed sample size, n_1. Oh and Scheuren (1983) and Rao (1985) have used this approach to show that the estimator is unbiased. The conditional approach, however, draws into question the role of the underlying sampling scheme in statistical inference. We examine this simple problem and show how explicit specification of sampling indicator random variables will result in a probability model familiar to other problems. Straightforward application of prediction methods gives rise to a predictor that depends on the probability of missing response, which, when replaced by the sample estimate, reduces to the mean of the observed responses.

We define a finite population as a collection of a known number, N, of identifiable subjects labeled $s = 1, 2, \ldots, N$. Associated with subject s is response y_s, which we assume is potentially observable without error. In the voting preference survey, y_s corresponds to an indicator that assumes a value of one if subject s will vote for the incumbent, and zero otherwise, and the assumption of no response error corresponds to each subject having no uncertainty as to their vote. We summarize the set of population values in the vector $\mathbf{y} = (y_1, \ldots, y_N)'$ and assume that there is interest in a $p \times 1$ vector of parameters of the form $\boldsymbol{\beta} = \mathbf{Gy}$ where $\mathbf{G}$ is a matrix of known constants. Attention is limited to a single parameter, the population mean, given by $\beta = \mu = \frac{1}{N}\sum_{s=1}^{N} y_s$, that is, by taking $\mathbf{G} = \mathbf{g}' = \frac{1}{N}\mathbf{1}_N'$ and define the population variance as $\frac{N-1}{N}\sigma^2 = \frac{1}{N}\sum_{s=1}^{N}(y_s - \mu)^2$.

2 Sampling, missing data, and prediction

Suppose that a simple random sample without replacement is to be selected from the population represented by the first n elements in a permutation of the population, where each permutation is equally likely. This representation has been discussed by Cassel, Särndal and Wretman (1977) and explored in the context of superpopulation models by Rao and Bellhouse (1978). Our discussion is closely related, but follows the definition and notation used by Stanek, Singer and Lençina (2004).

Let $i = 1, 2, \ldots, N$ index the positions in a permutation. We represent the value in position i of a randomly selected permutation by the random variable $Y_i =$

$\sum_{s=1}^{N} U_{is} y_s$ where $U_{is} = 1$ if unit s is in position i and $U_{is} = 0$ otherwise. When all permutations are equally likely, the random vector $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ is a random permutation of the population (as in Cassel, Särndal and Wretman (1977)). We can relate $\mathbf{Y}$ to $\mathbf{y}$ by considering $\mathbf{Y} = \mathbf{U}\mathbf{y}$ where

$$\mathbf{U} = \begin{pmatrix} U_{11} & \cdots & U_{1N} \\ \vdots & \ddots & \vdots \\ U_{N1} & \cdots & U_{NN} \end{pmatrix}.$$

Note that $\mathbf{y}$ is a vector of constants indexed by the subject labels, while $\mathbf{Y}$ is a vector of random variable indexed by the positions. Realizing a value of Y_i will not reveal which subject is occupying position i in the permutation, although it will reveal the corresponding value. To know which subject occupies position i in a permutation, we need to know the realized value of the random variable $S_i = \sum_{s=1}^{N} U_{is} s$.

These subtle distinctions can be illustrated with the voting preference example. Suppose that the realized response for the first selected subject ($i = 1$) is a vote for the incumbent. Simply knowing the realized value of Y_1 does not tell us which subject voted for the incumbent, it only tells us that one of the subjects voted this way. In order to know which subject cast this vote, we need to know which subject occupied the first position in the permutation, that is, the realization of S_1. This could be recorded along with the realized value of Y_1, resulting in a bivariate response. Typically, the additional variate representing the labeled unit is dropped from the analysis. Although not relevant for the present discussion, the subtle difference between the realized value corresponding to a position in the sample and the realized value of a subject is what makes interpretation of realized random effects in mixed models so challenging (see Stanek, Singer and Lençina (2004) for additional discussion).

Since each subject has an equal chance of being assigned to a given position in a permutation, $E_\xi(Y_i) = \mu$ for $i = 1, \ldots, N$, where ξ denotes expectation over permutations. We can summarize this expected value structure in a linear model given by $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$ where $\mathbf{X} = \mathbf{1}_N$ and $\beta = \mu$.

We partition the vector of random variables into a subset which we call the sample, $\mathbf{Y}_I = (Y_1, \ldots, Y_n)'$, indexed by $i = 1, 2, \ldots, n$, and the remainder, $\mathbf{Y}_{II} = (Y_{n+1}, \ldots, Y_N)'$, indexed by $i = n + 1, \ldots, N$, such that $\mathbf{Y} = (\mathbf{Y}_I' | \mathbf{Y}_{II}')'$. In order to estimate a parameter that is a linear function of $\mathbf{Y}$, the basic problem is prediction of a linear function of $\mathbf{Y}_{II}$ that is not observed, since $\mathbf{Y}_I$ is realized. The linear function is determined by the parameter of interest. For example, since the population mean can be represented by $\mu = \frac{n}{N}\bar{y}_I + \frac{N-n}{N}\bar{Y}_{II}$ (where $\bar{y}_I = \frac{1}{n}\sum_{i=1}^{n} y_i$ with y_i representing the realized value of Y_i and $\bar{Y}_{II} = \frac{1}{N-n}\sum_{i=n+1}^{N} Y_i$), an estimator of μ requires prediction of $\bar{Y}_{II}$.

We illustrate this process with a simple example. Suppose we have a population with size $N = 4$ and select a sample without replacement of size $n = 2$. We

Table 1 *Example of realized sample for three possible permutations where $N = 4$ and $n = 2$*

uy	Permutation $\mathbf{Y}$	Realized sample $\mathbf{y}_I$
$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$	$\left(\begin{array}{c} y_2 \\ y_1 \\ \hline y_4 \\ y_3 \end{array}\right)$	$\begin{pmatrix} y_2 \\ y_1 \end{pmatrix}$
$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$	$\left(\begin{array}{c} y_3 \\ y_4 \\ \hline y_1 \\ y_2 \end{array}\right)$	$\begin{pmatrix} y_3 \\ y_4 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$
$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$	$\left(\begin{array}{c} y_1 \\ y_2 \\ \hline y_3 \\ y_4 \end{array}\right)$	$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

represent the population as $\mathbf{y} = (y_1, y_2, y_3, y_4)'$ and a random permutation of the population as $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)'$. The first two random variables in the permutation make up the sample. A total of $N! = 24$ possible equally likely permutations can occur. As an example, the results of three possible permutations are given in Table 1. The random variables Y_3 and Y_4 will not be observed. Predicting their sum in the expression $\bar{Y}_{II}$ is the basic problem.

Predictors of this function can be developed using the approach of Royall (1988), as summarized by Valliant, Dorfman and Royall (2000) in the context of superpopulation models, but it is not necessary to introduce a superpopulation to apply the approach to simple random sampling. We require that the predictors be a linear function of the sample, be unbiased, and have minimum expected MSE, that is, $\hat{\bar{Y}}_{II}$, is the best linear unbiased predictor (BLUP). A weighted linear function of this predictor and the sample mean leads to the best linear unbiased estimator (BLUE) of μ. Under simple random sampling, the BLUP of $\bar{Y}_{II}$ is $\bar{y}_I$, so that the BLUE of μ is $\bar{y}_I$, the simple sample mean (Stanek, Singer and Lençina, 2004).

3 Two methods of specifying missing data

Under the assumption of MCAR, we specify two models that account for missing data. In each model, we assume that the probability of a missing response is constant, and equal to π. The first model represents the missing data mechanisms by random variables indexed by the position of a subject in the sample, $M_i, i = 1, \ldots, N$, where M_i takes on a value of one if response is missing for position i, and zero otherwise. Such random variables may represent a missing data mechanism for factors determined by the study design, as when different interviewers are assigned groups of sample subjects to interview. The second model represents a missing data mechanism by random variables indexed by subjects, $H_s = 1, \ldots, N$, where H_s takes on a value of one if response is missing for subject s, and zero otherwise. Such random variables may represent a missing data mechanism where a factor, such as answering machine screening, depends on individual subjects. The two missing data mechanisms emphasize the distinction between subject labels and sample positions.

3.1 A model for response when missing data depends on sample subject positions

We first consider the setting where the missing data mechanism is indexed by the position of subjects in the sample, as might occur if interviewers are assigned to consecutive selected subjects. We incorporate the missing data mechanism into the random permutation model by augmenting the N random variables to a vector of $2N$ random variables.

The first N random variables in the vector correspond to potentially observed responses. The ith random variable is given by $(1 - M_i)Y_i$. If $i \leq n$, the random variable will be realized in the sample. When the realized value of M_i is $m_i = 0$, response for the subject selected in position i is given by the realized value of Y_i, that is, y_i. When the realized value of M_i is $m_i = 1$, response for the subject selected in position i is missing, and the value of the realization, $(1 - m_i)Y_i$, is zero. Thus, the first N random variables are the potentially observable responses for random variables representing a permutation.

The second N random variables in the vector correspond to missing responses. The ith random variable is given by $M_i Y_i$. If $i \leq n$, the random variable will be realized (but the value of the random variable will not be observed) in the sample. For example, when the realized value of M_i is $m_i = 1$, response for the subject selected in position i is missing, but the realized value of $M_i Y_i$ will correspond to the realized value of $m_i Y_i$, that is, y_i. Although this value will not be observed by the investigator, it will be contained in the second set of random variables. When the realized value of M_i is $m_i = 0$, response for the subject selected in position i is not missing, but the value of the realization, $m_i Y_i$, is zero. Thus, the second N random variables are the potentially observable responses for realized random variables representing a permutation where response is missing.

When the probability of missing depends on position, we represent the first N random variables by the product $(\mathbf{I}_N - \mathbf{M}^*)\mathbf{Y}$ where $\mathbf{M}^* = \bigoplus_{i=1}^{N} M_i$ is a diagonal matrix with diagonal elements given by M_i. We partition this vector into an $n \times 1$ vector representing the sample, $\mathbf{Y}_I^{(o)}$, and an $(N - n) \times 1$ vector representing the remainder, $\mathbf{Y}_{II}^{(o)}$, such that $(\mathbf{I}_N - \mathbf{M}^*)\mathbf{Y} = (\mathbf{Y}_I^{(o)\prime} | \mathbf{Y}_{II}^{(o)\prime})'$, where the superscript is a reminder that these random variables are potentially observed. The second N random variables correspond to missing responses and are given by the product $\mathbf{Y}^{(m)} = \mathbf{M}^{*}\mathbf{Y}$. We represent the vector of $2N$ random variables by $\mathbf{Z}_1 = (\mathbf{Y}_I^{(o)\prime}|\mathbf{Y}_{II}^{(o)\prime}|\mathbf{Y}^{(m)\prime})^{\prime}$. Elements of this vector are given by $Z_{1i}^{(o)} = (1 - M_i)\sum_{s=1}^{N} U_{is} y_s$ and $Z_{1i}^{(m)} = M_i \sum_{s=1}^{N} U_{is} y_s$.

3.2 A model for response when missing data depends on labeled subjects

When the probability of a missing value depends on the subject, we represent potentially observable random variables by a $2N \times 1$ vector in a similar manner. We form the first vector of N random variables that are potentially observed by the product $\mathbf{U}(\mathbf{I}_N - \mathbf{H}^{*})\mathbf{y}$, where $\mathbf{H}^{*} = \bigoplus_{s=1}^{N} H_s$ is a diagonal matrix with diagonal elements given by H_s. We partition this vector into the sample, $\boldsymbol{\gamma}_I^{(o)}$, and the remainder, $\boldsymbol{\gamma}_{II}^{(o)}$, using the same notation, but where $\mathbf{U}(\mathbf{I}_N - \mathbf{H}^{*})\mathbf{y} = (\boldsymbol{\gamma}_I^{(o)^{\prime}}|\boldsymbol{\gamma}_{II}^{(o)^{\prime}})^{\prime}$. Elements of $\boldsymbol{\gamma}_I^{(o)}$ are now of the form $Z_{2i}^{(o)} = \sum_{s=1}^{N} U_{is}(1 - H_s) y_s$ for $i = 1, \ldots, n$. When $h_s = 0$, the realized value for the subject s is not missing and may be observed; when $h_s = 1$, the realized value of the random variable $Z_{2i}^{(o)}$ is zero. The N random variables in the vector corresponding to the missing responses are given by the product $\boldsymbol{\gamma}^{(m)} = \mathbf{U}\mathbf{H}^{*}\mathbf{y}$ with elements $Z_{2i}^{(m)} = \sum_{s=1}^{N} U_{is} H_s y_s$.

We represent the vector of $2N$ random variables by $\mathbf{Z}_2 = (\boldsymbol{\gamma}_I^{(o)^{\prime}}|\boldsymbol{\gamma}_{II}^{(o)^{\prime}}|\boldsymbol{\gamma}^{(m)^{\prime}})^{\prime}$. The random variables in the $n \times 1$ vector $\boldsymbol{\gamma}_I^{(o)}$ are observed as a result of sampling. The elements of $\boldsymbol{\gamma}_{II}^{(o)}$ and $\boldsymbol{\gamma}^{(m)}$ are not observed. Notice that unobserved random variables correspond to both the missing data, and to the portion of the population that is not included as part of the sample. Although these random variables are represented distinctly, they share the common status of ‘missing data.’

4 Predicting the mean response

We develop the expected value and variance of the $2N \times 1$ vector of random variables representing the population next, and use these expressions to predict mean response. Expectation is taken with respect to random variables representing the missing data mechanism, ξ_1, and with respect to random permutations of the population ξ_2. For example, the elements of $\mathbf{Z}_1$ are either of the form $Z_{1i}^{(o)} = (1 - M_i)\sum_{s=1}^{N} U_{is} y_s$ or $Z_{1i}^{(m)} = M_i \sum_{s=1}^{N} U_{is} y_s$. Using conditional expectation, $E_{\xi_1\xi_2}(Z_{1i}^{(o)}) = E_{\xi_1}[E_{\xi_2|\xi_1}(Z_{1i}^{(o)})]$, and since $E_{\xi_2|\xi_1}(Z_{1i}^{(o)}) = (1 - M_i)\mu$, $E_{\xi_1\xi_2}(Z_{1i}^{(o)}) = (1 - \pi)\mu$. Similarly, $E_{\xi_1\xi_2}(Z_{1i}^{(m)}) = \pi\mu$. Combining these expressions, we get $E_{\xi_1\xi_2}(\mathbf{Z}_1) = \mu\left[\begin{pmatrix} 1-\pi \\ \pi \end{pmatrix} \otimes \mathbf{1}_N\right]$.

The results illustrate that the expected value of random variables in the sample, $Z_{1i}^{(o)}$ are not equal to the population mean. This result is intuitive if we recall that when a response is missing, the observed response is zero (as a result of introducing the missing data random variables in the model). For example, if the probability of a missing response is $\pi = 0.20$, the expected value of a potentially observable random variable, $Z_{1i}^{(o)}$, $i = 1, \dots, n$ is 0.8μ. The expected value does not imply that there is bias, but simply that the expected value will be closer to zero than the population mean. Identical results are obtained taking the expected value of the random variables $\mathbf{Z}_2$, that is, $E_{\xi_1\xi_2}(\mathbf{Z}_2) = \mu\Big[\binom{1-\pi}{\pi} \otimes \mathbf{1}_N\Big]$.

The variance can be developed in a similar manner by taking $\mathrm{Var}_{\xi_1\xi_2}(\mathbf{Z}_1) = \mathrm{Var}_{\xi_1}[E_{\xi_2|\xi_1}(\mathbf{Z}_1)] + E_{\xi_1}[\mathrm{Var}_{\xi_2|\xi_1}(\mathbf{Z}_1)]$. To evaluate this expression, we let $\mathbf{M} = (\mathbf{I}_N - \mathbf{M}^*|\mathbf{M}^*)'$, so that $\mathbf{Z}_1 = \mathbf{MUy}$. Then $E_{\xi_2|\xi_1}(\mathbf{Z}_1) = \mathbf{M1}_N\mu$, where $\mathbf{M1}_N = \binom{\mathbf{1}_N}{\mathbf{0}_N} + \binom{-\mathbf{m}}{\mathbf{m}}$, and $\mathbf{m} = (\,M_1 \quad M_2 \quad \cdots \quad M_N\,)'$. Note that $\mathrm{Var}_{\xi_1}(\mathbf{M1}_N) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathrm{Var}_{\xi_1}(\mathbf{m})$. Since we assume that the missing data random variables are independent, it follows that $\mathrm{Var}_{\xi_1}(\mathbf{m}) = \pi(1-\pi)\mathbf{I}_N$, and hence $\mathrm{Var}_{\xi_1}[E_{\xi_2|\xi_1}(\mathbf{Z}_1)] = \pi(1-\pi)\mu^2\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_N$. From Stanek, Singer and Lençina (2004) we have $\mathrm{Var}_{\xi_1}[\mathbf{UY}] = \sigma^2(\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N)$, and hence

$$E_{\xi_1}[\mathrm{Var}_{\xi_2|\xi_1}(\mathbf{Z}_1)] = \sigma^2 E_{\xi_1}\Big[\mathbf{M}\Big(\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N\Big)\mathbf{M}'\Big],$$

where $E_{\xi_1}(\mathbf{MJ}_N\mathbf{M}') = \sigma^2\pi(1-\pi)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_N + \sigma^2\binom{1-\pi}{\pi}\binom{1-\pi}{\pi}' \otimes \mathbf{J}_N$. Combining these expressions, it follows that

$$\mathrm{Var}_{\xi_1\xi_2}(\mathbf{Z}_1) = \left(\sigma^2\begin{bmatrix} (1-\pi)^2 & \pi(1-\pi) \\ \pi(1-\pi) & \pi^2 \end{bmatrix} \otimes \left(\mathbf{I}_N - \frac{\mathbf{J}_N}{N}\right)\right) \\ + \left[\pi(1-\pi)\Big(\frac{N-1}{N}\sigma^2 + \mu^2\Big)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_N\right].$$

Identical results are obtained evaluating the variance of the random variables $\mathbf{Z}_2$.

The model for the population that includes missing data is given by

$$\mathbf{Z}_1 = \mathbf{X}\boldsymbol{\alpha} + \mathbf{E}_1$$

where $\mathbf{X} = \mathbf{I}_2 \otimes \mathbf{1}_N$ and $\boldsymbol{\alpha} = \binom{(1-\pi)\mu}{\pi\mu}$. Notice that in this model, the sum of the parameters is equal to the population mean, μ. A similar model can be expressed for $\mathbf{Z}_2$. We drop the subscripts for $\mathbf{Z}$ in the subsequent development since the two models have the same expected value and variance.

We use the prediction approach to estimate the population mean, $\mu = \mathbf{g}'\mathbf{Z}$, where $\mathbf{g} = \frac{1}{N}\mathbf{1}_{2N}$. To do so, we partition $\mathbf{g} = (\,\mathbf{g}_I' \quad \mathbf{g}_{II}'\,)'$ with $\mathbf{g}_I = \frac{n}{N}\frac{\mathbf{1}_n}{n}$ and partition $\mathbf{Z}$ into the sample set of random variables, $\mathbf{Z}_I$, with elements $Z_i^{(o)}$ (corresponding to $\mathbf{Y}_I^{(o)}$ or $\boldsymbol{\gamma}_I^{(o)}$), and the remaining random variables, $\mathbf{Z}_{II}$. The sample random variables will be observed, and correspond either to the responses for the nonmissing selected units, or to zero for the selected units with missing responses. As a result, once the sample is realized, $\mu = \mathbf{g}_I'\mathbf{z}_I + \mathbf{g}_{II}'\mathbf{Z}_{II}$, and the basic problem is the prediction of $\mathbf{g}_{II}'\mathbf{Z}_{II}$.

We require the predictor to be a linear function of the sample data, $\mathbf{p}'\mathbf{Z}_I$, to be unbiased, such that $E_{\xi_1\xi_2}(\mathbf{p}'\mathbf{Z}_I - \mathbf{g}'_{II}\mathbf{Z}_{II}) = 0$, resulting in the constraint that $\mathbf{p}'\mathbf{1}_n(1-\pi) = 1 - \frac{n}{N}(1-\pi)$, and to minimize the variance $\text{Var}_{\xi_1\xi_2}(\mathbf{p}'\mathbf{Z}_I - \mathbf{g}'_{II}\mathbf{Z}_{II})$. Minimizing the variance subject to this constraint using Lagrange multipliers and simplifying leads to the best linear unbiased estimator (Lu, 2004) given by

$$\hat{\mu} = \frac{\mathbf{1}'_n\mathbf{Z}_I}{n(1-\pi)}. \tag{4.1}$$

The denominator, $n(1-\pi)$, corresponds to the expected number of responding sample subjects and $\hat{\mu}$ to the average of the expected respondents. The numerator is simply the total of the realized sample, $\sum_{i=1}^{n} Z_i^{(o)}$, using a response of zero for random variables where the response is missing. The variance of the estimator is given by

$$\text{Var}(\hat{\mu}) = \frac{1}{n(1-\pi)}\left[\pi\mu^2 + \frac{N - n(1-\pi) - \pi}{N}\sigma^2\right].$$

The estimator can be written in a manner that emphasizes the interpretation of predicting the unobserved random variables. We express it as the weighted sum of three terms: the sample mean, $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{(o)}$, the predictor of response for a subject not selected in the sample, $\hat{P}_1$ and the predictor of response for the $N\pi$ subjects where response is expected to be missing, $\hat{P}_2$. Using this notation, the estimator is given by

$$\hat{\mu} = \frac{1}{N}[n\bar{Z} + (N-n)\hat{P}_1 + N\pi\hat{P}_2].$$

To interpret this result, note that the predictor of response for a subject not selected in the sample who will respond is equal to the average response over the sample, and given by $\hat{P}_1 = \bar{Z}$. The predictor for subjects whose response will be missing corresponds to the average response of the expected respondents, $\hat{P}_2 = \hat{\mu}$. Combining these expressions, we get

$$\hat{\mu} = \frac{1}{N}[n\bar{Z} + (N-n)\bar{Z} + N\pi\hat{\mu}],$$

which readily can be seen to be equal to (4.1). A key feature of this decomposition is the ability to interpret terms in the estimator as a sum of realized sample values, and predictors of unobserved random variables. This provides an intuitive guide to the statistical inference that links directly to the actual statistical methods.

4.1 The empirical predictor

In practice, it is common to replace the unknown probability of missing responses with estimates that may come as additional data, or directly from the sample. We refer to the resulting predictor as an empirical predictor.

We can estimate π by the proportion of missing responses in the sample. Notice that if response consists solely of the realized values of $\mathbf{Z}_I$, then we will not be able to distinguish whether or not response for position i in the sample is missing, or simply represents a response with value zero for the selected subject. As a result, we cannot form an unbiased estimate of π without additional information which we assume consists of the realized values, m_i of M_i (or $\sum_{s=1}^{N} U_{is} H_s$) for $i = 1, \ldots, n$. This allows us to know whether or not response is missing for each position in the sample. Defining n_0 as the number of elements of $\mathbf{Z}_I$ where response is missing, we estimate π by $\hat{\pi} = \frac{n_0}{n}$. Representing the number of nonmissing sample responses as $n_1 = n - n_0$, the empirical predictor simplifies to

$$\hat{\mu}_0 = \frac{\mathbf{1}_n' \mathbf{Z}_I}{n(1-\hat{\pi})} = \frac{1}{n_1} \sum_{i=1}^{n} Z_i^{(o)},$$

that is, the simple mean of the nonmissing sample respondents. The empirical predictor simplifies to the intuitive estimator widely used, although rarely motivated in a formal fashion. Using the finite population random permutation model approach and the additional data on M_i or $\sum_{s=1}^{N} U_{is} H_s$ for $i = 1, \ldots, n$, this corresponds to the BLUP.

To estimate the MSE we replace π by $\hat{\pi} = \frac{n_0}{n}$, $\frac{1}{N} \sum_{s=1}^{N} y_s^2$ by $T^2 = \frac{1}{n_1} \sum_{i=1}^{n} Z_i^{(o)2}$ and σ^2 by $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n} (1 - m_i)(Z_i^{(o)} - \hat{\mu}_0)^2$. It follows that $\hat{V}(\hat{\mu}_0) = \frac{n_0}{n n_1} T^2 + (\frac{N-n}{N}) \frac{S_1^2}{n}$. The first term in this expression inflates the variance to account for variability resulting from division by the expected number of nonmissing sample responses, as opposed to the actual number of nonmissing sample responses. Although for the empirical estimator, we use the actual number of nonmissing sample responses, the expression for the MSE still retains this term. The second term is similar to the variance of the sample mean under simple random without replacement sampling. The difference is that S_1^2 is an estimate of σ^2 that depends only on nonmissing sample respondents.

4.2 An example

We illustrate the empirical predictor with the voting example. Suppose that a telephone interview survey of $n = 400$ voters in Amherst, Massachusetts is conducted to estimate the proportion of voters who favor same sex marriages. We assume that the sample is selected based on simple random sampling of the town voter registration list containing $N = 8000$ registered voter names. We also assume that the probability of response being missing is independent of the actual subject's response for all voters.

As a result of the survey, suppose that $n_1 = 250$ responses are obtained, where 200 ($\hat{\mu}_0 = 0.80$) favor same sex marriages. The simple sample average is $\bar{z} = \frac{200}{400} = 0.5$, while the estimate of the probability of missing response is $\hat{\pi} = \frac{150}{400} = 0.375$. We construct the estimator of the proportion of voters favoring same sex marriages by adding the number of voters favoring same sex marriages in three groups, namely, the sampled voters who respond, $n\bar{z} = 400(\frac{200}{400}) = 200$, the predicted number of voters who would respond, but were not included in the sample, $(N-n)\bar{z} = 7600(\frac{200}{400}) = 3800$ and the predicted number of voters who would not respond, $[N\hat{\pi}]\hat{\mu}_0 = [8000(0.375)]0.8 = (3000)0.8 = 2400$. Adding the observed number of voters favoring same sex marriages who would respond and those who would not respond, we get

$$\hat{\mu}_0 = \frac{1}{8000}[200 + 3800 + 2400] = 0.8.$$

When the response is dichotomous, the expression for the MSE simplifies to

$$\hat{V}(\hat{\mu}_0) = \frac{1}{n_1}\left(\frac{N-n}{N-1}\right)\hat{\mu}_0(1-\bar{Z}) + \left(\frac{1}{N-1}\right)\frac{n_0}{n_1}\left(\frac{n-1}{n}\right)\hat{\mu}_0,$$

so that $\hat{V}(\hat{\mu}_0) = 0.0015202 + 0.000059857 = 0.00158$. We compare this to the variance corresponding to the finite population where the sample size is assumed to be equal to the number of nonmissing sample responses. This variance is given by $\hat{\sigma}^2 = \frac{1}{n_1}(\frac{N-n_1}{N})S_1^2$. When response is dichotomous, $S_1^2 = n_1(\frac{\hat{\mu}_0(1-\hat{\mu}_0)}{n_1-1})$, and in our example $\hat{\sigma}^2 = 0.0006225$. Simulation studies reveal that $\hat{\sigma}^2$ is a better approximation for the variance of $\hat{\mu}_0$ than $\hat{V}(\hat{\mu}_0)$. Using this expression and assuming asymptotic normality, we may construct an approximate 95% confidence interval for the expected response obtaining (0.751, 0.849) in the example.

5 Discussion

The simple example illustrates a design based method that frames statistical inference as a problem of predicting values not in the sample. When some responses are missing, predictors are needed both for the remaining units in the population, and for the sampled units for which response is missing. This is very similar to the approach advocated by Valliant, Dorfman and Royall (2000) in which optimal predictors are constructed for unobserved random variables based on a superpopulation model. Both approaches distinguish between the values in a finite population and the set of random variables whose realization corresponds to the population values. The difference in the two approaches stems from accounting for the unit labels. In the superpopulation approach, labels are ignored. The starting point is a set of exchangeable random variables that form a superpopulation. The finite population is considered to be a realization of a set of N superpopulation random variables and the predictors are developed from the superpopulation model, and not from the finite population sampling. Additional discussion of the superpopulation model approach is given by Bolfarine (1989), with other missing data perspectives given by Orchard and Woodbury (1972) and Bolfarine (1987–1988).

In contrast, the probability model presented in Section 4 arises directly from the sampling and a missing data model. Units in the population are identifiable, and the labels can be traced through the process of describing the missing data mechanism. This enables a clear interpretation of the physical processes of sampling, and generation of the missing data. No superpopulation is needed. However, similar to the superpopulation model approach, the essential statistical problem is framed as a prediction problem, and uses the same tools in developing the best linear unbiased predictors as in the model-based approach.

The basic design-based prediction approach was presented in the context of simple random sampling by Stanek, Singer and Lençina (2004). Innovative aspects to the application of this approach to the missing data problem include a clear distinction between missing data mechanisms that depend on sample positions from those that depend on units, and the representation of the problem as a double set of random variables. The empirical estimates provide an additional interesting aspect of the development. In the context of best linear unbiased predictors in mixed models, empirical estimates are commonly constructed by replacing variance component parameters by sample estimators. Usually, such substitutions result in larger expected MSE due to additional variance introduced by substituting the estimators for parameters. In our application, the predictor involves a single unknown parameter, π. Replacing this parameter by the sample estimator does inflate the expected MSE. However, the expected MSE appears to dramatically overstate the variability when compared with the variance evaluated from simulation studies. In a sense, substituting n_1 for the sample size reduces the variance by accounting for the ignorable missing data, since the response is recorded as a value of zero when the sample respondent's response was missing.

Using finite population sampling models and a prediction approach connects estimation and prediction, since an estimator of the population mean can be interpreted as a predictor of the linear combination of random variables not included in the sample. The design-based prediction approach to finite populations can be extended to other situations. Predictors of realized random effects have been developed by Stanek and Singer (2004, 2008) in the context of two stage sampling with response error. Additional extensions have been made to settings where there are auxiliary variables associated with each unit in the context of simple random sampling by Li (2003). These extensions begin to develop design based methods that may be useful for modeling survey data. Other extensions, as for example to nonignorable missing response mechanisms, remain to be explored.

References

Bolfarine, H. and Rodrigues, J. (1987–1988). On the simple projection predictor in finite populations. *Estadistica* **39–40**, 55–59. MR1007418

Bolfarine, H. and Rodrigues, J. (1989). A missing value approach to the prediction problem in finite population. *Pub. Inst. Stat., Univ. Pari*, Vol. XXXIV, Fasc. 2, 59–66. MR1744759

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. New York, NY: John Wiley. MR0652527

Cochran, W. G. (1977). *Sampling Techniques*. New York, NY: John Wiley. MR0474575

Li, W. (2003). Use of random permutation model in rate estimation and standardization. Ph.D Thesis. Dept. Biostatistics and Epidemiology, Univ. Massachusetts, Amherst, MA. Available at http://www.umass.edu/cluster/ed/publication/Li-2003-Dissertation.pdf.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York, NY: John Wiley. MR0890519

Lu, J. (2004). Estimating parameters when considering the unobserved units as missing values in simple random sampling. Masters Thesis. Dept. Biostatistics and Epidemiology, Univ. Massachusetts, Amherst, MA. Available at http://www.umass.edu/cluster/ed/publication/jingsonglu-full-thesis-2004.pdf.

Oh, H. L. and Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography* (W. G. Madow, I. Olkin and D. B. Rubin, eds.) 143–184. New York: Academic Press.

Orchard, T. and Woodbury, M. (1972). A missing value information principle. *Proceedings of the 6th Berkeley Symposium on Math. Statist. and Prob.* **1**, 697–715. MR0400516

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11**, 15–31.

Rao, J. N. K. and Bellhouse, D. R. (1978). Estimation of finite population mean under generalized random permutation model. *Journal of Statistical Planning and Inference* **2**, 125–141. MR0507362

Royall, R. M. (1988). *The Prediction Approach to Sampling Theory* (P. R. Krishnaiah and C. R. Rao, eds.). *Handbook of Statistics* **6** 399–413. New York, NY: Elsevier Science Publishers. MR1020090

Stanek E. J. III and Singer, J. M. (2004). Predicting random effects from finite population clustered samples with response error. *Journal of the American Statistical Association* **99**, 1119–1130. MR2109500

Stanek, E. J. III, Singer, J. M. and Lençina, V. B. (2004). A unified approach to estimation and prediction under simple random sampling. *Journal of Statistical Planning and Inference* **121**, 325–338. MR2038825

Stanek, E. J. III and Singer, J. D. (2008). Predicting random effects with an expanded finite population mixed model. *Journal of Statistical Planning and Inference* **138**, 2991–3004. MR2526218

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York, NY: John Wiley. MR1784794

401 Arnold House
University of Massachusetts
Amherst, Massachusetts 01003
USA
E-mail: stanek@schoolph.umass.edu