Estimating the Population Mean From a Simple Random Sample When Some Responses are Missing

Jingsong Lu , Edward J. Stanek III, and Elaine Puleo

Department of Biostatistics and Epidemiology

401 Arnold House

University of Massachusetts

Amherst, Massachusetts 01103

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). Missing data mechanisms are defined 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 direct extensions of best linear unbiased prediction (BLUP) 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 non-missing 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, counter-intuitively, is smaller than the MSE if the probability of missing is known.

KEYWORDS: Simple random sampling, Missing data, MCAR, finite population, Best linear unbiased estimator (BLUE), prediction.

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 non-response. In addition, poor interviewing skills by some interviewers may result in refusals for other contacted subjects. The first type of non-response depends on the subject, while the second type of non-response depends on the interviewer.

In the simplest setting, the probability of non-response 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 and Rubin 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. 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 non-response assumption. 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, the estimator is not a simple linear function of the sample data, 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 simple random sampling 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. Straight forward 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.

2. The Population

We define a finite population as a collection of a known number, N, of identifiable subjects labeled $s = 1, 2, \text{L}\ , N$. Associated with subject s is a 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. When there are more than two choices, response corresponds to a set of indicators for the candidates, with only one having a value of one. In this context, we will limit our interest to votes for a single candidate, and thus consider a single response variable. 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 $\mathbf{y} = \left(y_1, \text{L}\ , y_N \right)'$ and assume that there is interest in a $p \times 1$ vector of parameters of the form $\boldsymbol{\beta} = \mathbf{G}\mathbf{y}$ where $\mathbf{G}$ is a matrix of known constants. We limit our attention to a single parameter, the population mean given by

$\mathbf{g}'\mathbf{y} = \beta = \mu = \frac{1}{\mathrm{N}}\sum_{s=1}^{N} y_s$, defined by setting $\mathbf{g}' = \frac{\mathbf{1}_N'}{N}$, and define the population variance as

$\frac{N-1}{N}\sigma^2 = \frac{1}{\mathrm{N}}\sum_{s=1}^{N}\left(y_s - \mu\right)^2$.

3. Sampling, Missing Data, and Prediction

Suppose that a simple random sample without replacement is to be selected from the population. We define the possible samples as the set of all possible permutations of the population. In simple random sampling, each permutation is equally likely. This representation has been discussed by Cassel, Särndal and Wretman (1977) and explored in the context of super-population models by Rao and Bellhouse (1978). Our discussion is closely related to these presentations, but follows the definition and notation used by Stanek, Singer and Lençina (2004) for random variables corresponding to positions in a randomly selected permutation. We define the elements occupying the first n positions in the permutation to be the sample.

Let $i = 1, 2, \mathrm{L}\ , N$ 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 $\mathbf{Y} = \left(Y_1, \mathrm{L}\ , Y_N\right)'$ is a random permutation of the population (as in Cassel, Särndal and Wretman 1977). We can relate $\mathbf{Y}$ to $\mathbf{y}$ such that $\mathbf{Y} = \mathbf{U}\mathbf{y}$, where

$$\mathbf{U} = \begin{pmatrix} U_{11} & \mathrm{L} & U_{1N} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ U_{N1} & \mathrm{L} & U_{NN} \end{pmatrix}.$$

Note that $\mathbf{y}$ is a vector of constants indexed by the labeled subjects, while $\mathbf{Y}$ is a vector of random variables 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 value corresponding to the realized subject. 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, i.e. 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 of a position 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 $i = 1,...,N$, 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,

$\mathbf{Y}_I = (Y_1, \text{L}\ , Y_n)'$, indexed by $i = 1, 2, \text{L}\ , n$, and the remainder, $\mathbf{Y}_{II} = (Y_{n+1}, \text{L}\ , Y_N)'$, indexed by $i = n+1, \text{L}\ , N$, such that $\mathbf{Y} = \left(\mathbf{Y}_I' \mid \mathbf{Y}_{II}'\right)'$. The prediction approach to inference makes use of the fact that the realized sample is the realized values of $\mathbf{Y}_I$. 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. 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 based on a simple random sample 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 represent the population as $\mathbf{y} = (y_1 \quad y_2 \quad y_3 \quad y_4)'$ and a random permutation of the population as $\mathbf{Y} = (Y_1 \quad Y_2 \quad Y_3 \quad Y_4)'$. The first two random variables in the permutation make up the sample. A total of $N! = 24$ possible permutations can occur, with each of them equally likely. The results of three possible permutations are given in Figure 1.

Figure 1. Example of Realized Sample for Three Possible Permutations where $N = 4$ and $n = 2$

uy	Permutation **Y**	Realized Sample $\mathbf{y}_I$
$\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}$	$\begin{pmatrix}y_2\\y_1\\y_4\\y_3\end{pmatrix}$	$\begin{pmatrix}y_2\\y_1\end{pmatrix}$
$\begin{pmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}$	$\begin{pmatrix}y_3\\y_4\\y_1\\y_2\end{pmatrix}$	$\begin{pmatrix}y_3\\y_4\end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$
$\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}$	$\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}$	$\begin{pmatrix}y_1\\y_2\end{pmatrix}$

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 of inference.

Predictors of this function can be developed using the approach of Royall (1988) which has been recently summarized by Valliant, Dorfman, and Royall (2000) in the context of super-population models. It is not necessary to introduce a super-population to apply the approach to simple random sampling. We assume that the predictors are a linear function of the sample, are unbiased, and will result in minimum expected MSE. The resulting predictor, $\hat{\bar{Y}}_{II}$, is called the best linear unbiased predictor (BLUP). When combined as a weighted linear function with the sample mean, the estimator of μ is the best linear unbiased estimator (BLUE). 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).

3.1. 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 the probability of a missing response is constant, and equal to π. The first model represents the missing data mechanism by random variables indexed by the position of a subject in the sample, M_i, $i = 1,...,N$, 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 for example 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, $s = 1,...,N$, 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 mechanism emphasize the distinction between subject labels and sample positions.

3.1.1. 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 i^{th} random variable is given by $(1-M_i)Y_i$. If $i \le n$, 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, i.e. 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 i^{th} random variable is given by M_iY_i. If $i \le n$, 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_iY_i will correspond to the realized value of m_iY_i, i.e. 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_iY_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 the remainder, $\mathbf{Y}_{II}^{(o)}$, such that $(\mathbf{I}_N - \mathbf{M}^*)\mathbf{Y} = \left(\mathbf{Y}_I^{(o)} \mid \mathbf{Y}_{II}^{(o)}\right)'$, where the superscript is a

reminder that these random variables are potentially observed. The second N random variables corresponding to missing responses are given by the product $\mathbf{Y}^{(m)} = \mathbf{M}^{*}\mathbf{Y}$. We represent the vector of $2N$ random variables by $\mathbf{Z}_1 = \left(\mathbf{Y}_I^{(o)} \mid \mathbf{Y}_{II}^{(o)} \mid \mathbf{Y}^{(m)}\right)'$. Elements of this vector are given by $Z_{1i}^{(o)} = \left(1 - M_i\right)\sum_{s=1}^{N} U_{is} y_s$ and $Z_{1i}^{(m)} = M_i \sum_{s=1}^{N} U_{is} y_s$.

3.1.2. A Model for Response when Missing Data Depends on Labeled Subjects

When the probability of missing depends on the subject, we represent the 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 $\mathbf{U}\left(\mathbf{I}_N - \mathbf{H}^{*}\right)\mathbf{y}$, 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 an $n \times 1$ vector representing the sample, $\Upsilon_I^{(o)}$, and the remainder, $\Upsilon_{II}^{(o)}$, using the same notation, but where $\mathbf{U}\left(\mathbf{I}_N - \mathbf{H}^{*}\right)\mathbf{y} = \left(\Upsilon_I^{(o)} \mid \Upsilon_{II}^{(o)}\right)'$. Elements of $\Upsilon_I^{(o)}$ are now of the form $Z_{2i}^{(o)} = \sum_{s=1}^{N} U_{is}\left(1 - H_s\right) y_s$ for $i = 1,...,n$. 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 missing responses are given by the product $\Upsilon^{(m)} = \mathbf{U}\mathbf{H}^{*}\mathbf{y}'$ with elements $Z_{2i}^{(m)} = \sum_{s=1}^{N} U_{is} H_s y_s$.

We represent the vector of $2N$ random variables by $\mathbf{Z}_2 = \left(\Upsilon_I^{(o)} \mid \Upsilon_{II}^{(o)} \mid \Upsilon^{(m)}\right)'$. The

random variables in the $n \times 1$ vector $\mathbf{Y}_I^{(o)}$ are observed as a result of sampling. The elements of $\Upsilon_{II}^{(o)}$ and $\Upsilon^{(m)}$ 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'.

3.2. First and Second Moments.

We develop the expected value and variance of the $2N \times 1$ vector of random variables representing the population next. 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 of the form $Z_{1i}^{(o)} = (1 - M_i) \sum_{s=1}^{N} U_{is} y_s$ and $Z_{1i}^{(m)} = M_i \sum_{s=1}^{N} U_{is} y_s$. Using conditional expectation,

$E_{\xi_1 \xi_2}\left(Z_{1i}^{(o)}\right) = E_{\xi_1 \xi_2}\left[E_{\xi_2|\xi_1}\left(Z_{1i}^{(o)}\right)\right]$, and since $E_{\xi_2|\xi_1}\left(Z_{1i}^{(o)}\right) = (1 - M_i)\mu$, $E_{\xi_1 \xi_2}\left(Z_{1i}^{(o)}\right) = (1 - \pi)\mu$.

Similarly, $E_{\xi_1 \xi_2}\left(Z_{1i}^{(m)}\right) = \pi\mu$. Combining these expressions, $E_{\xi_1 \xi_2}(\mathbf{Z}_1) = \mu\left[\begin{pmatrix} 1-\pi \\ \pi \end{pmatrix} \otimes \mathbf{1}_N\right]$.

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, $Z_{1i}^{(o)}$, $i = 1, ..., n$ is 80% of μ. The

expected value doesn't imply that there is bias, but simply that the expected value will be closer to zero than the population mean. The identical results are obtained taking the expected value of the random variables $\mathbf{Z}_2$, such that $E_{\xi_1\xi_2}(\mathbf{Z}_2) = \mu\left[\begin{pmatrix}1-\pi\\ \pi\end{pmatrix}\otimes\mathbf{1}_N\right]$.

The variance can be developed in a similar manner. Using a conditional expansion, $\text{var}_{\xi_1\xi_2}(\mathbf{Z}_1) = \text{var}_{\xi_1}\left[E_{\xi_2|\xi_1}(\mathbf{Z}_1)\right] + E_{\xi_1}\left[\text{var}_{\xi_2|\xi_1}(\mathbf{Z}_1)\right]$. To evaluate this expression, we define $\mathbf{M} = \left(\mathbf{I}_N - \mathbf{M}'^* \mid \mathbf{M}'^*\right)'$, such that $\mathbf{Z}_1 = \mathbf{MUY}$. Then $E_{\xi_2|\xi_1}(\mathbf{Z}_1) = \mathbf{M1}_N\,\mu$, where $\mathbf{M1}_N = \begin{pmatrix}\mathbf{1}_N\\ \mathbf{0}_N\end{pmatrix} + \begin{pmatrix}-\mathbf{m}\\ \mathbf{m}\end{pmatrix}$, and $\mathbf{m} = \begin{pmatrix}M_1 & M_2 & \cdots & M_N\end{pmatrix}$. Note that $\text{var}_{\xi_1}(\mathbf{M1}_N) = \begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\otimes \text{var}_{\xi_1}(\mathbf{m})$. Since we assume the missing data random variables are independent, $\text{var}_{\xi_1}(\mathbf{m}) = \pi(1-\pi)\mathbf{I}_N$, and hence $\text{var}_{\xi_1}\left[E_{\xi_2|\xi_1}(\mathbf{Z}_1)\right] = \pi(1-\pi)\mu^2\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\otimes\mathbf{I}_N$.

Using the result from Stanek, Singer and Lençina (2004) that $\text{var}_{\xi_1}[\mathbf{UY}] = \sigma^2\left(\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N\right)$, and hence

$$E_{\xi_1}\left[\text{var}_{\xi_2|\xi_1}(\mathbf{Z}_1)\right] = \sigma^2 E_{\xi_1}\left[\mathbf{M}\left(\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N\right)\mathbf{M}'\right] = \sigma^2\pi(1-\pi)\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\otimes\mathbf{I}_N + \sigma^2\begin{pmatrix}1-\pi\\ \pi\end{pmatrix}\begin{pmatrix}1-\pi\\ \pi\end{pmatrix}'\otimes\mathbf{J}_N.$$

Combining these expressions,

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

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

We can summarize the model for the population that includes missing data. The model is

given by

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

where $\mathbf{X} = \mathbf{I}_2 \otimes \mathbf{1}_N$, and $\boldsymbol{\gamma} = \begin{pmatrix} (1-\pi)\mu \\ \pi\mu \end{pmatrix}$. 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.

3.3. Developing Predictors of the Mean

We use the prediction approach to estimate the population mean. First, note that we can express the population mean as a simple linear combination of the random variables, $\mu = \mathbf{g}'\mathbf{Z}$, where $\mathbf{g} = \frac{1}{N}\mathbf{1}_{2N}$. Also, we can partition $\mathbf{Z}$ into a set of random variables corresponding to the sample, $\mathbf{Z}_I$ (corresponding to $\mathbf{Y}_I^{(o)}$ or $\Upsilon_I^{(o)}$), and the remaining random variables, $\mathbf{Z}_{II}$. We partition $\mathbf{g} = \begin{pmatrix} \mathbf{g}_I' & \mathbf{g}_{II}' \end{pmatrix}'$ in a similar manner, where $\mathbf{g}_I = \frac{n}{N}\left(\frac{\mathbf{1}_n}{n}\right)$. The values of the sample random variables will be observed, and correspond to the response for a non-missing selected unit, or the value zero for a selected unit where response is missing. 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 prediction of $\mathbf{g}_{II}'\mathbf{Z}_{II}$.

We require the predictor to be linear function of the sample data, $\mathbf{p}'\mathbf{Z}_I$, to be unbiased, such that $E_{\xi_1\xi_2}\left(\mathbf{p}'\mathbf{Z}_I - \mathbf{g}_{II}'\mathbf{Z}_{II}\right)$, resulting in the constraint that $\mathbf{p}'\mathbf{1}_n(1-\pi) = 1 - \frac{n}{N}(1-\pi)$, and to

minimize the variance, $\text{var}_{\xi_1\xi_2}\left(\mathbf{p}'\mathbf{Z}_I - \mathbf{g}_{II}'\mathbf{Z}_{II}\right)$. Minimizing the variance subject to this constraint using Lagrange multipliers and simplifying leads to the best linear unbiased estimator (Lu, 2004) given by

$$\hat{\mu} = \frac{\mathbf{1}_n'\mathbf{Y}_1}{n(1-\pi)}. \qquad (3.1)$$

The denominator, $n(1-\pi)$, corresponds to the expected number of responding sample subjects. We refer to $\hat{\mu}$ as the average of the expected respondents. The numerator is simply the total of the realized sample, $\sum_{i=1}^{n} Y_i$, 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 un-observed random variables. We express it as the weighted sum of three terms: the sample mean, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$; 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

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

There is a simple intuition that corresponds to the choice of predictors. 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 $\hat{P}_1 = \bar{Y}$. 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,

$$\hat{\mu} = \frac{1}{N}\left[n\overline{Y} + \left(N - n\right)\overline{Y} + N\pi\hat{\mu} \right],$$

an expression which readily can be seen to be equal to (3.1). 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 un-observed random variables. This provides an intuitive guide to the statistical inference that links directly to the actual statistical methods.

3.4. The Empirical Predictor

In practice, we need to know the probability of missing response in order to compute the predictor. A common practice when parameters are unknown is to replace the parameters by estimates of the parameters. The estimators may come as additional data, or directly from the sample. We refer to the resulting predictor as an empirical predictor.

In order to estimate the population mean, we need an estimate of π. We can estimate this parameter by the proportion of missing responses in the sample. Notice that if response consists solely of the realized values of $\mathbf{Y}_I^{(o)}$, then we will not be able distinguish whether or not response for position i in the sample is missing, or simply represents a response of zero for the selected subject. As a result, we can not form an unbiased estimate of π without more information. We assume that such additional information is available. The additional information consists of the realized values, x_i of M_i (or $\sum_{s=1}^{N} U_{is} H_s$) for $i = 1,...,n$, allowing us to

know for each position in the sample whether or not response is missing. Defining n_0 as the number of elements of $\mathbf{Y}_I^{(o)}$ (or $\Upsilon_I^{(o)}$) where response is missing, we estimate π by $\hat{\pi} = \frac{n_0}{n}$. Representing the number of non-missing sample responses as $n_1 = n - n_0$, the empirical predictor simplifies to

$$\hat{\mu}_0 = \frac{\mathbf{1}_n' \mathbf{Y}_1}{n(1-\hat{\pi})} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i ,$$

equal to the simple mean of the non-missing 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 $i = 1,...,n$, this predictor emerges as best.

We estimate the MSE by replacing π by $\hat{\pi} = \frac{n_0}{n}$, $\frac{1}{N}\sum_{s=1}^{N} y_s^2$ by $T^2 = \frac{1}{n_1}\sum_{i=1}^{n} Y_i^2$ and σ^2 by $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(1-x_i)(Y_i - \hat{\mu}_0)^2$. Using these estimators, $\hat{V}(\hat{\mu}_0) = \frac{n_0}{n n_1} T^2 + \left(\frac{N-n}{N}\right)\frac{S_1^2}{n}$. The first term in this expression inflates the variance to account for variability resulting from division by the expected number of non-missing sample responses, as opposed to the actual number of non-missing sample responses. Although for the empirical estimator, we use the actual number of non-missing sample responses, the expression for the MSE still retains this term. The second term is similar to the variance of 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 non-missing sample respondents.

3.5. 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 given by $\bar{y} = \frac{200}{400} = 0.5$, 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 the sum of the voters favoring same sex marriages in three groups, the sampled voters who respond, $n\bar{y} = 400\left(\frac{200}{400}\right) = 200$; the predicted number of voters who would respond, but were not included in the sample, $(N-n)\bar{y} = 7600\left(\frac{200}{400}\right) = 3800$; and the predicted number of voters who would not respond, $[N\hat{\pi}]\hat{\mu}_0 = \left[8000(0.375)\right]0.8 = (3000)0.8 = 2400$. Adding the observed number of voters favoring same sex marriages in the sample to the predicted number favoring same sex marriage who would respond and those who would not respond,

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

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

$\hat{V}(\hat{\mu}_0) = \frac{1}{n_1}\left(\frac{N-n}{N-1}\right)\hat{\mu}_0(1-\bar{Y}) + \left(\frac{1}{N-1}\right)\frac{n_0}{n_1}\left(\frac{n-1}{n}\right)\hat{\mu}_0$, which is given by

$\hat{\mathrm{v}}(\hat{\mu}_0) = 0.0015202 + 0.000059857 = 0.00158$.

We compare the expression for the variance with an expression for the variance corresponding to the finite population variance where the assumption is made that the sample size is equal to the number of non-missing sample responses. This variance is given by $\hat{\sigma}^2 = \frac{1}{n_1}\left(\frac{N-n_1}{N}\right)S_1^2$. When response is dichotomous, $S_1^2 = n_1\left(\frac{\hat{\mu}_0(1-\hat{\mu}_0)}{n_1-1}\right)$, and in our example $\hat{\sigma}^2 = 0.0006225$. Simulation studies (Lu 2004) 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 estimate a 95% confidence interval for response as (0.751, 0.849).

4. 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 response is missing, predictors are needed both of the remaining units in the population, and of the sampled units where response is missing. This approach to inference is very similar to the approach advocated by Vallient, Dorfman, and Royall (2000) in which optimal predictors are constructed for un-observed random variables based on a model for a super-population. Both approaches distinguish between the values in a finite population and a set of random variables whose realization is the population values. The difference in the two approaches stems from accounting

for the unit labels. In the super-population approach, labels are ignored. The starting point is a set of random variables that form a super-population and satisfy certain statistical properties, such as exchangeability. The finite population is considered to be a realization of at set of N super-population random variables. The predictors are developed from the super-population model, and not from the finite population sampling. In the survey sampling literature, the predictors are referred to as model-based, since their derivation is based on the super-population model.

In contrast, the probability model presented in Section 3 arises directly from the sampling. 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 accounting and interpretation of the physical processes of sampling, and processes that may result in missing data. Unlike the model-based survey approach, the random variables and their properties are based solely on the sampling design and do not require additional assumptions. However, similar to the model-based approach, the essential statistical problem is framed as a prediction problem, and makes use of the same tools in developing the best linear unbiased predictors as are used 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). There are several innovative aspects to the application of this approach to the missing data problem. First, identifying the labeled units enables a clearer specification of the missing data mechanism. We have distinguished the missing data mechanisms that depend on sample positions (such as interviewers), from missing data mechanisms that depend on the labeled unit (such as having an answering machine). It is clearly possible to have more complex missing data models where the missing data mechanism

depends on both sample position, *i*, and unit, *s*. Although the development in Section 3 assumes that the probability of response being missing is independent and identically distributed, other assumptions are possible, and will likely lead to different predictors.

A second innovative feature of the development is the representation of the problem as a double set of random variables. The two sets of random variables correspond to one set where a response will be potentially observed, and a second set consisting of the response values when response is missing. In the first set, realizations of the random variables where response is missing have a value of zero; in the second set, realizations of the random variables where response is not missing have a value of zero. Summing these random variables gives rise to a set of random variables where there is no missing data. The idea of expanding the representation of random variables for missing data is similar in concept to the expansion of random variables considered by Stanek, Singer, and Lençina (2004) used to distinguish prediction of response for a unit based on a simple random sample.

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 components parameters by sample estimators. Usually, such substitutions result in an elevated expected MSE for the empirical predictor 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 non-

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. This is clearly illustrated in the simple random sampling/missing data setting. The example provides a setting for distinguishing terms commonly used in statistics. The term 'estimator' is reserved for a statistic that comes close (in terms of having small variance) to a parameter. The term 'predictor' is reserved for a statistic that comes close (in terms of having small mean squared error) to a random variable. Since we define a parameter as a linear combination of population values, but then define a random permutation of these values as a set of random variables, the parameter has an equivalent definition as a linear combination of random variables. This process implies that 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) in the context of two stage sampling with response error. Their development is limited to populations with equal size clusters, but the methods can be extended. Additional extensions have been made to settings where there are auxiliary variables associated with each unit in the context of simple random sampling (Li 2003; Li and Stanek, 2004). These extensions begin to develop design based methods that may be useful for modeling survey data. Other extensions, as for example to non-ignorable missing response mechanisms, remain to be explored.

References

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

Cochan, W.G. (1977), *Sampling Techniques*, New York, NY; John Wiley.

Li, W. (2003), *Use of Random Permutation Model in Rate Estimation and Standardization*, Ph.D Thesis, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA.

Li, W. and Stanek, E.J. III. (2004), "Covariance Adjusted Estimation Under a Design Based Random Permutation Model," *Journal of Statistical Planning and Inference*, under review.

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

Lu, J. (2004), *Estimating parameters when considering the unobserved units as missing values in simple random sampling*, Masters Thesis, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA.

Oh, H.L. and Scheuren, H.F. (1983), "Weighting Adjustment for Unit Nonresponse," *Incomplete data in Sample Survey*, 12, 142-184.

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.

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

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*, in press.

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.

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