Predicting Random Effects from a Finite Population of Unequal Size Clusters based on Two-Stage Sampling

Edward J. Stanek III
Department of Public Health
401 Arnold House
University of Massachusetts
715 North Pleasant Street
Amherst, MA 01003-9304 USA
stanek@schoolph.umass.edu

Julio M. Singer
Departamento de Estatistica
Universidade de São Paulo
São Paulo, Brazil
jmsinger@ime.usp.br
ABSTRACT

Prediction of the latent value (the mean) of a realized cluster selected via two-stage sampling is an important problem, and one that has multiple solutions. The different solutions arise from different stochastic models posited for the random variables representing the sample. Models that require specific distributional assumptions, or require the notion of a superpopulation have been postulated, but are not based on the finite population sampling design usually employed in many practical situations. We develop predictors of the latent value of a realized cluster from a population with unequal size clusters using a design based two-stage random permutation model. We avoid introducing random subscripts for primary sampling units (PSUs) by expanding the number of random variables used to represent the problem. This expanded set is adequate to identify clusters and primary sampling units, but is complicated by linear dependencies. We propose three alternative linear combinations of the expanded random variables that simplify the linear dependencies while retaining the identity of PSUs. The collapsed random variables correspond to second stage sampling unit totals, weighted totals, or means for the sample and remaining PSU random variables. We develop predictors for each that are linear functions of the sample, unconditionally unbiased, and have minimum expected mean squared error. The unbiased constraint is satisfied for the first two sets of random variables only when second stage sampling is made with probability proportional to size. The target of the prediction differs for the third set of random variables, but the unbiased constraint can be satisfied for these variables under any simple random second stage sampling design. The developments illustrate the additional complexity that occurs with an unequal size clustered population, and reveal multiple avenues for future development.
Contact Address:
Edward J. Stanek III
Department of Public Health
401 Arnold House
University of Massachusetts at Amherst
Amherst, Ma.  01002

Phone: 413-545-3812
Fax:  413-545-1645
Email: stanek@schoolph.umass.edu

KEYWORDS: superpopulation, best linear unbiased predictor, random permutation, optimal estimation, design-based inference, mixed models.

ACKNOWLEDGEMENT. This work was developed with the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil and the National Institutes of Health (NIH-PHS-R01-HD36848, R01-HL071828-02), USA.
1. INTRODUCTION

We develop design-based predictors of the mean (or total) of a realized cluster selected via a two-stage cluster random sample from a finite population of unequal size clusters. Prediction of random effects is commonly based on mixed model theory (Goldberger 1962; Henderson 1984; McLean, Sanders and Stroup 1991; Robinson 1991; McCulloch and Searle 2001) leading to best linear unbiased predictors (BLUP). The mixed model assumes an infinite population of clusters and units, and hence does not account for differences in cluster sizes. In the survey sampling literature, predictors of the mean of realized clusters have been developed using superpopulation models (Scott and Smith 1969; Bolfarine and Zacks 1992; Valliant, Dorfman and Royall 2000) that account for different size clusters. Closely related are the empirical Bayes estimators of cluster means from stratified samples (Ghosh and Lahiri 1987). Design-based predictors obtained under a two-stage random permutation model possibly including response error have been developed by Stanek and Singer (2004), but are limited to equal cluster size settings.

We extend the random permutation framework to clustered populations where the number of units in clusters is not equal and there is no response error, and use it to predict target random variables that correspond to the mean (or total) response of units in a cluster. The predictors account for the different size clusters in the finite population since their derivation is based on the two-stage sampling design.

Mixed model, superpopulation model, and empirical Bayes modeling approaches have been directly applied to unequal size clustered populations. Part of their appeal is the ease with which these approaches may be applied to more complex settings. In contrast, design-based
methods become increasingly complex with more complex sampling designs (Sarndal, Swensson, and Wretman 1992; Cochran 1977). The additional complexity is due in part to an accounting for the population structure. In particular, the methods we develop and apply to unequal size clustered populations are more complex than similar design-based methods applied to equal size clustered populations (Stanek and Singer 2004). In order to appreciate the need for the complexity, it is helpful to review the cost of the apparent simplicity of alternative model-based approaches.

The simplicity of the model-based approaches can be attributed to conceptualizing the problem in terms of a set of random variables referred to as a superpopulation. These random variables may be structured as when they are divided into mutually exclusive groups, referred to as strata, or clusters. A probability model is then specified for the random variables. This model is not necessarily related to the sampling design, and hence can be simple. An average of the random variables in a group is optimally predicted conditional on the realized random variables in the sample, with optimality assessed relative to the superpopulation model. The simplicity of this recipe has popularized the model-based approach.

The model-based approach, however, has limitations when applied to a defined finite population. First, since the superpopulation is only conceptual, the number of superpopulation random variables included in the probability model is not defined by the problem. For instance, only random variables corresponding to sample clusters and units are included in the superpopulation probability model used for the usual mixed model. Also, while the superpopulation model defined by Ghosh and Lahiri (1987) includes random variables corresponding to all units in a sample of clusters, the superpopulation defined by Scott and Smith (1969) includes random variables corresponding to the entire finite population. The lack of a
link between the probability model for the random variables and the finite population allows for multiple solutions to the same problem, each of which satisfies some model-based optimality criterion. Arguments over the ‘best’ solution provide the practitioner little guidance, since the different assumptions relate to conceptual superpopulations, and not to the problem at hand.

A second limitation is the apparent lack of connection between the sampling design and the superpopulation probability model. The superpopulation probability model is commonly motivated by a population structure, often defined in terms of hierarchical units or spatial structures (Rao 2003). Motivating the superpopulation model by a population structure can simplify the model choice, since the same model can be used for different sampling designs. However, the superpopulation model does not capture the sampling distribution of predictors, nor their properties. Such a distribution requires specification of both the population structure and the sampling plan. As a result, the model-based optimality criteria are of questionable relevance since they do not account for resampling variability.

The limitations of the model-based approach have been addressed in the survey sampling literature. One approach is to compare the design-based bias and mean squared error of predictors developed under a superpopulation model (Sarndal et al. 1992). When covariables are present, generalized regression estimators (GREG) may be used to combine design-based Horvitz-Thompson estimators with model based predictors (Cassel, Sarndal, and Wretman 1976), or to calibrate predictors to covariables in the population to improve estimators (Deville and Sarndal 1992).

In light of the limitations of the model-based approaches, we develop design-based predictors of realized cluster means (or totals), assuming a two-stage probability sample was selected from a population formed by unequal size clusters. The approach accounts for the
population structure and sampling design. Two-stage sampling designs with probability proportional to size (PPS), or unequal probability sampling are shown to lead to conceptually different problems, and different predictors. In all cases, the results provide clear guidance for practitioners.

As an example, we apply our results to a cluster randomized controlled trial designed to evaluate the relative effectiveness of a training program for physician-delivered nutrition counseling, alone and in combination with an office support program, on dietary fat intake and blood LDL-cholesterol levels in patients with hyperlipidemia (the Worcester-Area Trial for Counseling in Hyperlipidemia (WATCH) Trial) (Ockene et al. 1999). Forty-five primary care internists were selected from a total of 60 internists at the Fallon Clinic, a central Massachusetts health maintenance organization (HMO), with each internist randomly assigned to one of three conditions: (I) Usual Care; (II) Physician nutrition counseling training; and (III) Physician nutrition counseling training plus an office support program. Twelve hundred and seventy-eight patients with blood LDL-cholesterol levels in the highest 25\textsuperscript{th} percentile, having previously-scheduled physician visits, were recruited into the study. The number of sampled patients per internist (cluster) ranged from 1 to 43.

The WATCH study had a cluster randomized design, with patients in physician practices forming the basic clusters. We focus our attention on a single dependent variable, change in percent of Kcal from saturated fat, defined as the baseline minus one year measure for each patient. There was considerable variability in the magnitude of the average patient’s saturated fat change between physician’s practices within a given intervention. We discuss estimating the change for eligible patients (units) in a physician’s practice (cluster) that accounts for the practice size.
1.1 Basic Notation

We define notation and terminology for a clustered finite population, and then use it to provide an overview of prediction based on mixed models (Searle, Casella, and McCulloch 1992; Verbeke and Molenberghs 2000; McCulloch and Searle 2001; Diggle, Heagerty, Liang, and Zeger 2002) and superpopulation sampling models (Bolfarine and Zacks 1992; Valliant et al. 2000). Let a finite population be defined by a listing of $M_s$ units labeled $t = 1, \ldots, M_s$ in each of $N$ clusters, labeled $s = 1, \ldots, N$, where the non-stochastic response for unit $t$ in cluster $s$ is given by $y_{st}$. We assume that a two-stage without replacement cluster sample of $n$ clusters from this population is to be selected, with a different (but known) number of units, $m_s$, selected from each of the selected clusters.

We represent the sample of clusters by defining a random permutation of the clusters in the original frame. A new label, $i = 1, \ldots, N$, is assigned to the positions in a permutation of clusters. Without loss of generality, we assume that the first $n$ positions contain the sample clusters. In a similar manner, we distinguish the listed units in cluster $s$ from a random permutation of units in cluster $s$, and label their positions by $j = 1, \ldots, M_s$. We again assume that the first $m_s$ positions contain the units in the sample if cluster $s$ is selected. For ease of exposition, we refer to the cluster that will occupy position $i$ in the permutation as primary sampling unit (PSU) $i$, and to the unit that will occupy position $j$ in the permutation as secondary sampling unit (SSU) $j$. Using these ideas and notation, we represent a two-stage permutation of the population as an ordered list of $N = \sum_{s=1}^{N} M_s$ random variables, where the
random variable representing response for SSU $j$ in PSU $i$ is $Y_{ij}$. PSUs and SSUs are indexed by positions ($i$ and $j$), whereas clusters and units are indexed by labels ($s$ and $t$) in the finite population.

We think of sampling as selecting a permutation. Prior to sample selection, the cluster that will correspond to PSU $i$ is random, and hence the cluster mean in position $i$ is a random variable. Once the sample has been selected, for positions $i \leq n$, we will observe which cluster corresponds to a particular PSU in the sample. We refer to that cluster as the realized PSU, and to the average (or total) response for units in that cluster as the latent value of the realized PSU. In this context, mixed model and superpopulation modeling strategies may be advocated for predicting the latent value of a realized PSU, or linear combinations of such latent values. We review these approaches, setting the stage for development of the main results in the subsequent section.

### 1.2 Mixed Models

In the mixed model literature, the model for the response of SSU $j$ in PSU $i$ is given by

$$Y_{ij} = \mu + B_i + E_{ij} \quad (1)$$

with $B_i \sim iid \ N(0, \sigma_i^2)$ independent of $E_{ij} \sim iid \ N(0, \sigma^2)$. In this model, $\mu$ corresponds to the overall population mean, $B_i$ is a random effect that corresponds to the deviation of the latent value of PSU $i$ from the population mean, and $E_{ij}$ is a random error corresponding to the deviation of response for SSU $j$ from the mean of PSU $i$. The corresponding density, when
maximized jointly with respect to the fixed and random effects, leads to Henderson’s mixed model equations (Henderson, Kempthorne, Searle and von Krosigk 1959).

In model (1), assuming known variances, the estimator of the fixed effect is the weighted least squares estimator, \[ \hat{\mu} = \frac{1}{m} \sum_{i=1}^{n} w_i Y_i, \]
where
\[ w_i = \frac{1}{\sqrt{\sum_{j=1}^{m} \frac{1}{v_{ij}}}}, \quad \bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij}, \]
and \[ v_i = \sigma^2 + \frac{\sigma_i^2}{m_i}. \]

Solution of the mixed model equations results in the BLUP of \( B_i \). Combining the estimator of \( \mu \) and the predictor of \( B_i \), the predictor of \( \mu + B_i \) is
\[ \hat{P}_{\text{MM}} = \hat{\mu} + k_i (\bar{Y}_i - \hat{\mu}) \quad (2) \]
where \[ k_i = \frac{\sigma^2}{v_i}. \]

In practice, the variance parameters are replaced by the maximum likelihood or restricted maximum likelihood estimates. McCulloch and Searle (2001) note that there are many other derivations of the BLUP (discussed extensively by Robinson (1991)). A derivation of particular note is the Bayesian derivation, where under normality, the BLUP is the posterior mode. The normality assumptions are not necessary, and can be relaxed by assuming only the existence of the first and second moments. None of these derivations or discussions account for the impact that different cluster size has on the predictors.

1.3 Superpopulation Sampling Models

Predictors of a linear combination of fixed and random effects (corresponding to the average response of SSUs in a PSU) in two-stage sampling settings were developed by Scott and Smith (1969) using a model-based survey sampling approach according to which a finite
population is viewed as the realization of a set of random variables in a superpopulation. The parameter corresponding to the latent value of a realized PSU is represented by the average response of a PSU in the superpopulation. Since only the sample random variables are observed, the essential statistical problem is how best to predict the remaining random variables in the realized PSU. Predictors are constructed using the joint distribution assumed for the superpopulation model (Royall 1976; Bolfarine and Zacks 1992; Valliant et al. 2000).

The random variables $Y_{ij}$, $i = 1, \ldots, N$; $j = 1, \ldots, M_i$ in the superpopulation are assumed to satisfy $E(Y_{ij}) = \mu$ and

$$\text{cov}(Y_{ik}, Y_{kl}) = \delta^2 + \sigma^2_i \text{ when } i = k; j = l$$

$$= \delta^2 \text{ when } i = k; j \neq l$$

$$= 0 \text{ otherwise.}$$

Scott and Smith (1969) use this superpopulation model to derive predictors of the average (over SSUs) response of a PSU, requiring the predictor to be a linear function of the sample. A Bayesian derivation (assuming the superpopulation is normally distributed), and a distribution free derivation (based on minimizing the expected mean squared error (MSE) of a linear predictor assuming that PSUs and SSUs within PSUs are exchangeable) are given, and shown to result in the same predictor, namely

$$\hat{P}_{SSU} = f_i \bar{Y} + (1 - f_i) \left[ \hat{\mu}^* + k^* \left( \bar{Y} - \hat{\mu}^* \right) \right] \quad (3)$$

if PSU $i$ is partially realized in the sample, or $\hat{\mu}^*$ if PSU $i$ is not in the sample, where

$$\hat{\mu}^* = \sum_{i=1}^{n} w_i^* \bar{Y}_i \quad f_i = \frac{m_i}{M_i} \quad w_i^* = \frac{1/v_i^*}{\sum_{i=1}^{n} 1/v_i^*} \quad v_i^* = \delta^2 + \frac{\sigma_i^2}{m_i} \quad \text{and} \quad k_i^* = \frac{\delta^2}{v_i^*}$$

This predictor has an appealing interpretation as the weighted sum of two terms: the sample mean for the realized PSU, and the predictor of the average of the remaining SSUs for the realized PSU. The weighting
factors are the proportions of SSUs that are and are not observed. For PSUs that are not in the sample, the predictor simplifies to the weighted sample mean.

There is an obvious similarity between (2) and (3). When $\delta^2 = \sigma^2$ and $f_i$ is small enough so that the first term in equation (3) can be ignored, $1 - f_i \cong 1$, and the two expressions are identical. The best linear unbiased predictor derived under the mixed model (2) predicts unobserved SSUs in a realized PSU. If the number of SSUs in the realized PSU is so large that the observed SSUs are a negligible fraction of the total for the PSU, then the realized PSU mean is estimated by the predicted values of the unobserved SSUs for the PSU. When a non-trivial portion of the SSUs are observed in the sample for a realized PSU, then the estimate of the realized PSU mean is a weighted average of the sampled and predicted values of the unobserved SSUs. This provides a strong intuitive appeal to the prediction-based approach as advocated by Valliant et al. (2000).

As discussed by Scott and Smith (1969), the assumption of normality is not needed, and can be replaced by assuming that the PSUs are exchangeable, and that the SSUs are exchangeable within PSUs. Coefficients of the predictor are derived that minimize the expected MSE over the superpopulation model. The MSE is required to be bounded, leading to a constraint in the minimization process.

A similar result was obtained by Ghosh and Lahiri (1987) using empirical Bayes estimation. They define a nested superpopulation model for the prior distribution nearly identical to Scott and Smith (1969), but setting $i = 1, \ldots, n$. Then, assuming the expected value of the posterior mean for PSU $i$ is a linear function of the sample (posterior linearity), they show that the empirical Bayes estimator is given by (3).
2. A TWO-STAGE SAMPLING MODEL FOR A FINITE UNEQUAL SIZE CLUSTERED POPULATION

We use design-based methods to predict the mean (or total) of a realized cluster based on a two-stage cluster random sample from a finite population formed by unequal size clusters. We first define target parameters explicitly, along with the random variables that represent a two-stage random permutation of the population. In Section 3, we use these results to develop optimal predictors. The development extends the results of Stanek and Singer (2004) for design-based predictors of realized random effects in a balanced clustered sampling design.

2.1 Finite Population Parameters and Reparameterizations

Recall that the finite population is defined by a listing of \( M_s \) units labeled \( t = 1, \ldots, M_s \) in each of \( N \) clusters, labeled \( s = 1, \ldots, N \) where response for unit \( t \) in cluster \( s \) is given by \( y_{st} \). The finite population parameters correspond to the mean and variance of cluster \( s \), \( s = 1, \ldots, N \), defined by

\[
\mu_s = \frac{1}{M_s} \sum_{t=1}^{M_s} y_{st} \quad \text{and} \quad \sigma^2_s = \frac{1}{M_s} \sum_{t=1}^{M_s} \left( y_{st} - \mu_s \right)^2
\]

(where we use the survey sampling definition of the parameter \( \sigma^2_s \)). Similarly, the population mean, and the variance between cluster means are defined as

\[
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s \quad \text{and} \quad \sigma^2 = \frac{1}{N} \sum_{s=1}^{N} \left( \mu_s - \mu \right)^2.
\]
Using these parameters, we represent the potentially observable response for unit $t$ in cluster $s$ as

$$y_{st} = \mu + \beta_s + \varepsilon_{st} \quad (4)$$

where $\beta_s = (\mu_s - \mu)$ is the deviation of the mean for cluster $s$ from the overall mean, and $\varepsilon_{st} = (y_{st} - \mu_s)$ is the deviation of unit $t$’s response (in cluster $s$) from the mean for cluster $s$.

Model (4) is called a derived model (Hinkelmann and Kempthorne 1994). Defining

$$y = \left( y_1', y_2', \ldots, y_N' \right)'$$

where $y_s = \left( y_{s1}, y_{s2}, \ldots, y_{sM_s} \right)'$, model (4) can be summarized as

$$y = X\mu + Z\beta + \varepsilon \quad (5)$$

where $X = \mathbf{I}_N$, $Z = \bigoplus_{s=1}^N \mathbf{I}_{M_s}$, $N = \sum_{s=1}^N M_s$, and $\beta' = (\beta_1, \beta_2, \ldots, \beta_N)$. Here, $\mathbf{I}_a$ is an $a \times 1$ column vector of ones, $\bigoplus_{s=1}^N \mathbf{A}_s$ denotes a block diagonal matrix with blocks given by $\mathbf{A}_s$ (Graybill 1983), and $\varepsilon$ is defined similarly to $y$. None of the terms in model (5) are random variables.

2.2 Random Variables and The Two Stage Random Permutation Model

We explicitly define a vector of random variables that represents a two stage random permutation of the population, similarly to $y$ in (5). Assuming that each realization of the permutation is equally likely, the random variables formally represent two-stage sampling (Cochran 1977). We assume that the sample clusters are in the first $n$ positions in a permutation.
of clusters. Similarly, we assume that the sample units in cluster \( s \) correspond to the units in the first \( m_s \) positions in a permutation of the cluster’s units.

When all clusters are of equal size such that \( M_s = M \) for all \( s = 1, \ldots, N \), Stanek and Singer (2004) define indicator random variables to explicitly represent a random vector of dimension \( NM \times 1 \) corresponding to a two-stage permutation of the population. We follow a similar strategy when clusters sizes differ. However, in order to clearly identify the PSUs and SSUs in the vector representing a random permutation, it is necessary to expand the number of random variables from \( N \) to \( NN \). We define the random variables and illustrate the need for the expansion next.

Sample indicator random variables relate \( y_{st} \) (response for unit \( t \) in cluster \( s \)) to \( Y_{ij} \) (response of the unit in position \( j \) for the cluster in position \( i \)). The indicator random variable \( U^{(s)}_{it} \) takes on a value of one when SSU \( j \) in cluster \( s \) is unit \( t \), and zero otherwise. As a consequence, the random variable corresponding to SSU \( j \) in cluster \( s \) is given by

\[
\tilde{Y}_{sj} = \sum_{t=1}^{M_s} U^{(s)}_{it} y_{st}.
\]

Defining \( U^{(s)} = \left( U^{(s)}_1, U^{(s)}_2, \ldots, U^{(s)}_{M_s} \right) \) as an \( M_s \times M_s \) matrix with columns

\[
U^{(s)}_i = \left( U^{(s)}_{1i}, U^{(s)}_{2i}, \ldots, U^{(s)}_{Mi} \right),
\]

a permutation of units in cluster \( s \) is represented by the \( M_s \times 1 \) vector \( \tilde{Y}_s = U^{(s)} y_s = \left( \tilde{y}_j \right) \).

A permutation of clusters is defined by using the indicator random variables \( U_{is} \), \( i = 1, \ldots, N \) and \( s = 1, \ldots, N \) that take on a value of one when PSU \( i \) is cluster \( s \), and a value of zero otherwise. We summarize these random variables in the \( N \times N \) matrix \( U = \left( U_{is} \right) \), and represent the random variable corresponding to PSU \( i \) and SSU \( j \) by
\[ Y_j = \sum_{x=1}^{N} U_{ix} \tilde{Y}_{sj} \quad (6). \]

In a permutation of clusters, all units in a cluster are permuted together. It is tempting to represent the permuted units in PSU \( i \) by \( \sum_{x=1}^{N} U_{ix} \tilde{Y}_s \). Notice, however, that this representation is not well defined since when clusters differ in size, the dimensions of \( \tilde{Y}_s \), \( s = 1, ..., N \), will differ.

To avoid the problem of random dimensions, we expand the number of variables used to represent a permutation of clusters by defining the \( N \times 1 \) vector

\[ Y_i = \left( \begin{array}{c}
U_{i1} \tilde{Y}_1' \\
U_{i2} \tilde{Y}_2' \\
\vdots \\
U_{iN} \tilde{Y}_N'
\end{array} \right) \].

Then, a two stage random permutation of the population is represented by the \( N \times 1 \) vector, \( Y = \left( \begin{array}{c} Y_1' \\
Y_2' \\
\vdots \\
Y_N'
\end{array} \right) \).

The following simple example illustrates the notation. Suppose the population consists of three clusters (\( N = 3 \)), where the first two clusters have two units (\( M_1 = M_2 = 2 \)), and the third cluster has three units (\( M_3 = 3 \)), such that \( N = 7 \). We represent permutations of units within clusters by \( \tilde{Y}_s \), \( s = 1, ..., 3 \). Suppose the first permutation of clusters results in clusters \( s = 1, \ s = 2 \) in positions \( i = 1, ..., 3 \) respectively, while a second permutation results in clusters \( s = 3, \ s = 2 \), and \( s = 1 \) in positions \( i = 1, ..., 3 \) respectively. The representation of the random variables realized by the first permutation of PSUs is the random vector \( \left( \tilde{Y}_1', \tilde{Y}_2', \tilde{Y}_3' \right)' \), and by the second permutation of PSUs is the random vector \( \left( \tilde{Y}_1', \tilde{Y}_2', \tilde{Y}_3' \right)' \). Although both vectors are of dimension \( N \times 1 \), the third SSU in the first permutation is in PSU \( i = 2 \), while the third SSU in the second permutation is in PSU \( i = 1 \). The position of a SSU in the permuted population is not sufficient to retain the identity of the PSU for the SSU. In contrast, using the
expanded random variable representation, the random variables realized by the first permutation of PSUs are represented by \((\tilde{Y}_1' \ 0_2' \ 0_3' \ | \ 0_2' \ \tilde{Y}_2' \ 0_3' \ | \ 0_2' \ \tilde{Y}_3'\)'\), while those realized by the second permutation are represented by \((0_2' \ 0_2' \ \tilde{Y}_3' \ | \ 0_2' \ \tilde{Y}_2' \ 0_3' \ | \ \tilde{Y}_1' \ 0_2' \ 0_3'\)'\). This notation preserves the identity of the PSU for each SSU.

2.3 Fixed and Mixed Effect Models for the Expanded Random Variables

We use the subscript \(\xi_1\) to indicate expectation with respect to permutations of clusters and the subscript \(\xi_2\) to indicate expectation with respect to permutations of units in a cluster, resulting in \(E_{\xi_1\xi_2}(Y_i) = \frac{1}{N}1_N\mu + \frac{1}{N}Z\beta\) for PSU \(i\) so that

\[
E_{\xi_1\xi_2}(Y) = \frac{1}{N}1_{NN}\mu + \frac{1}{N}(1_N \otimes Z)\beta
\]

Similarly, as indicated in the Appendix, we may show that

\[
\text{var}_{\xi_1\xi_2}(Y) = \frac{1}{N-1}P_a \otimes \left[ \bigoplus_{s=1}^{N}1_M, \mu_s \right] \bigoplus_{s=1}^{N}1_M, \sigma_s^2 \right]
\]

where \(P_a = I_a - \frac{1}{a}J_a\) and \(J_a\) denotes an \(a \times a\) matrix with all elements equal to 1. For all PSUs, the fixed effect model is given by the following identity

\[
Y = \frac{1}{N}1_{NN}\mu + \frac{1}{N}(1_N \otimes Z)\beta + \left[ Y - E_{\xi_1\xi_2}(Y) \right]. \tag{7}
\]

The term \(Y - E_{\xi_1\xi_2}(Y)\) represents a vector of residuals which may be written as the sum of two terms to express the mixed model. For PSU \(i\), let \(Y_i - E_{\xi_1\xi_2}(Y_i) = \left[ E_{\xi_2}(Y_i) - E_{\xi_1\xi_2}(Y_i) \right] + \varepsilon_i\),
where \( E_i = Y_i - E_{\xi_2}(Y_i) \). The first term, \( E_{\xi_2}(Y_i) - E_{\xi_2}(Y_i) \), represents random effects that depend on the PSU \( i \); the second term, \( E \), represents deviations of SSUs within the PSU.

Explicitly, the random effect for PSU \( i \) is defined by

\[
E_{\xi_2}(Y_i) - E_{\xi_2}(Y_i) = \left( \bigoplus_{s=1}^{N} 1_{M_s} \right) M_i + \left( \bigoplus_{s=1}^{N} 1_{M_s} \right) B_i
\]

where \( M_i = \left( (M_is) \right) \) with \( M_is = (U_is - E_{\xi_2}(U_is)) \mu \) and \( B_i = \left( (B_is) \right) \) with \( B_is = (U_is - E_{\xi_2}(U_is)) \beta_s \). The term \( M_is \) corresponds to a random effect multiplied by the population mean, while the second term, \( B_is \), corresponds to a similar random effect multiplied by the deviation of the cluster mean from the population mean. In subsequent sections, these effects can be related to different aspects of the sample design.

We summarize the random effects in the vectors \( M = (M'_1 \ M'_2 \ \cdots \ M'_N)' \) and \( B = (B'_1 \ B'_2 \ \cdots \ B'_N)' \). As a result, the mixed model may be expressed as

\[
Y = \frac{1}{N} 1_{N \times 1} \mu + \frac{1}{N} (1_N \otimes Z) \beta + \left[ \bigoplus_{s=1}^{N} 1_{M_s} \right] (M + B) + E \quad (8)
\]

where \( E = (E'_1 \ E'_2 \ \cdots \ E'_N)' \). The variance of the random effects is

\[
\text{var}_{\xi_2} \left[ \left( \bigoplus_{s=1}^{N} 1_{M_s} \right) (M + B) \right] = \frac{1}{N-1} P_N \otimes \left[ \left( \bigoplus_{s=1}^{N} 1_{M_s} \mu_s \right) P_N \left( \bigoplus_{s=1}^{N} 1'_{M_s} \mu_s \right) \right], \text{ while}
\]

\[
\text{var}_{\xi_2} (E) = \bigoplus_{s=1}^{N} \left( \frac{1}{N} P_{M_s} \sigma^2_s \right).
\]

2.3. Defining Random Variables of Interest
We assume that there is interest in a linear combination of PSUs totals (or means),

\[ P = g'Y, \]

where \( g' = b' \otimes \left( I'_{M_1} \quad I'_{M_2} \quad \ldots \quad I'_{M_N} \right) \) for totals, \( g' = b' \otimes \left( \frac{I'_{M_1}}{M_1} \quad \frac{I'_{M_2}}{M_2} \quad \ldots \quad \frac{I'_{M_N}}{M_N} \right) \) for means, and where \( b = \left( \left( b_i \right) \right) \) is an \( N \times 1 \) vector of constants. Of principal interest is the linear combination that represents the total, \( P^*_i \), or the mean, \( P_i \), of PSU \( i \), defined by setting \( b = e_i \) where \( e_i \) is an \( N \times 1 \) vector with all elements equal to zero, except for element \( i \) which has the value of one. We develop predictors of such random variables next.

**3. PREDICTING A PSU TOTAL OR MEAN BASED ON A TWO-STAGE SAMPLE**

We develop a predictor of a PSU total (or mean) based on the two-stage without replacement sampling model (8). We assume that sampling will result in observing a single non-stochastic value on each of the sample SSUs selected from each of the \( n \) sample PSUs. We require the predictor to be a linear function of the sample, to be unbiased, and to minimize the expected value (over the design) of the mean squared error (MSE).

The basic strategy for developing a predictor under a model-based approach is given in many places (Scott and Smith 1969; Royall 1976; Bolfarine and Zacks 1992; Valliant et al. 2000), and applied in a design-based framework to balanced two stage cluster sampling by Stanek and Singer (2004). First, the elements of \( Y \) are partitioned into a sample and remaining portion. We assume that the elements in the sample portion will be observed, and express \( P_i \) as the sum of two parts, one which is a function of the sample, and the other which is a function of the remaining random variables. Then, requiring the predictor to be a linear function of the sample random variables and to be unbiased, coefficients are evaluated that minimize
\[ \text{var}_{\hat{\xi} \hat{\xi}} \left( \hat{P} - P \right) \], the expected value of the MSE. The general development is given in Theorem 2.1 by Royall (1976).

Conceptually, Royall’s theorem can be applied directly to derive predictors, but practically, its application to model (8) is complicated by singularities in \( \text{var} \left( Y \right) \). The singularities are the result of the expansion of random variables needed to explicitly track clusters. We avoid these complications by developing solutions to a simpler problem based on a ‘collapsed’ set of random variables. The traditional representation collapses the expanded random variables to the terms \( Y_{ij} = \sum_{s=1}^{N} \sum_{r=1}^{M} U_{is} U_{jt} Y_{st} \), leading to random dimensions for the vectors of PSU random variables. We define alternative collapsed random variables that avoid this problem.

We consider three types of collapsed random variables. The first two enable the development of unbiased predictors for a PSU total, or mean when the samples of SSUs are selected with PPS sampling. The third type enables predictors of a PSU mean to be developed under any second stage simple random sampling plan, but imposes a weaker unbiased requirement. In the first case the random variables are collapsed to PSU totals for the sample and the remainder. In the second case the random variables are collapsed to scaled PSU totals for the sample and the remainder using the cluster size as the scaling factor. In the third strategy, the random variables are collapsed to PSU means for the sample and the remaining SSUs.

In order to simplify the derivation of the predictors, we introduce some additional notation. We represent the sampling fraction for cluster \( s \) as \( f_s = \frac{m_s}{M_s} \), and the average sampling fraction as \( \bar{f} = \frac{\overline{m}}{\overline{M}} \), where \( \overline{m} = \frac{1}{N} \sum_{s=1}^{N} m_s \) and \( \overline{M} = \frac{1}{N} \sum_{s=1}^{N} M_s \). Ratios of size for the
sample, remainder, and cluster are given by $r_{sl} = \frac{m_s}{\bar{m}}$, $r_{sll} = \frac{M_s - m_s}{M - \bar{m}}$, and $r_s = \frac{M_s}{M}$, and are used to define the weighted cluster means $\mu_i = \frac{1}{N} \sum_{s=1}^{N} \mu_{si}$, $\mu_{II} = \frac{1}{N} \sum_{s=1}^{N} \mu_{sll}$, $\mu_R = \frac{1}{N} \sum_{s=1}^{N} \mu_{sR}$, and

$$\mu_i = \frac{1}{N} \sum_{s=1}^{N} \mu_{si}^{*} \quad \text{and} \quad \mu_{II} = \frac{1}{N} \sum_{s=1}^{N} \mu_{sll}^{*} \quad \text{with} \quad \mu_{si}^{*} = r_{si} \mu_s, \quad \mu_{sll}^{*} = r_{sll} \mu_s, \quad \mu_{sll}^{*} = \frac{r_{sll}}{r_s} \mu_s, \quad \mu_{s}^{*} = \frac{r_{s}}{r_s} \mu_s, \quad \mu_{sll}^{*} = \frac{r_{sll}}{r_s} \mu_s, \quad \text{and} \quad \mu_{R} = r_s \mu_s.$$

Similarly, we define average deviations from these means by $\beta_i = \frac{1}{N} \sum_{s=1}^{N} \beta_{si}$, $\beta_{II} = \frac{1}{N} \sum_{s=1}^{N} \beta_{sll}$, $\beta_R = \frac{1}{N} \sum_{s=1}^{N} \beta_{sR}$ with $\beta_{si} = \mu_{si} - \mu_i$, $\beta_{sll} = \mu_{sll} - \mu_{II}$, $\beta_{sll} = \mu_{sll} - \mu_{II}$, and $\beta_{sR} = \mu_{sR} - \mu_R$. Vectors of length $N \times 1$ containing these means are denoted by $\mathbf{\mu}_i = (\mu_i, \ldots, \mu_i)^T$, $\mathbf{\mu}_{II} = (\mu_{II}, \ldots, \mu_{II})^T$, $\mathbf{\mu}_R = (\mu_R, \ldots, \mu_R)^T$, $\mathbf{\beta}_i = (\beta_i, \ldots, \beta_i)^T$, $\mathbf{\beta}_{II} = (\beta_{II}, \ldots, \beta_{II})^T$, $\mathbf{\beta}_R = (\beta_R, \ldots, \beta_R)^T$.

Variance components are defined by $\sigma_{i}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2}$, $\sigma_{I}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{sll}^{2}$, $\sigma_{II}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{sll}^{2}$, and $\sigma_{R}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} (\mu_{sR} - \mu_R)^2$. Also, $\sigma_{i}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2}$, $\sigma_{II}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{sll}^{2}$, and $\sigma_{R}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{sR}^{2}$. Additional weighted average variance components for within cluster variances are given by $\sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{f (1-f)} r_s \sigma_{s}^{2}$, $\sigma_{eI}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{r_s (1-f)} \sigma_{s}^{2}$, $\sigma_{eII}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{r_{sll} (1-f)} \sigma_{sll}^{2}$, $\sigma_{el}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{r_{s}} \sigma_{s}^{2}$, $\sigma_{eII}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{r_{sll}} \sigma_{sll}^{2}$, and $\sigma_{el}^{2} = \frac{1}{N} \sum_{s=1}^{N} \frac{f_s (1-f_s)}{r_{sll}} \sigma_{sll}^{2}$. We also
define \( k^* = \frac{\bar{m} \sigma^2_R}{\sigma^2_e + \bar{m} \sigma^2_R}, \ k^* = \frac{\bar{m} \sigma^2_R}{\sigma^2_e + \bar{m} \sigma^2_R}, \ k^* = \frac{\bar{m} \sigma^2_R}{\sigma^2_e + \bar{m} \sigma^2_R}, \) where \( \sigma^2_R = \sigma^2_R - \frac{\sigma^2_e}{M}, \)

\( \sigma^2 = \sigma^2 - \frac{\sigma^2_e}{M} \) and \( \sigma^2 = \sigma^2 - \frac{\sigma^2_e}{M}. \) The details of the development of predictors are given next.

3.1. Predictors of PSU Totals Based on Collapsing to Cluster Totals

We consider a vector of PSU totals for the sample and remainder by letting \( Y^* = C^* Y \)

where \( C^* = (I_N \otimes (C^*_i 1_N)) \cdot I \otimes (C^*_i 1_N)^\prime, \ C^*_i = \sum_{s=1}^{N} \left( 1'_{m_s} \ 0 \right) \) and

\( C^*_i = \sum_{s=1}^{N} \left( 1'_{m_s} \ 0 \right). \) The vector \( Y^* \) has dimension \( 2N \times 1 \) with the first \( N \) random variables consisting of totals of the sample SSUs for each PSU, namely \( Y^*_{i} = \sum_{s=1}^{N} U_{is} m_s \bar{Y}^*_{is}, \) where

\( \bar{Y}^*_{is} = \frac{1}{m_s} \sum_{j=1}^{m_s} \bar{Y}_{ij}, \ i = 1,...,N, \) and the second set of \( N \) random variables consisting of totals of the remaining SSUs for each PSU, namely \( Y^*_{i} = \sum_{s=1}^{N} U_{is} (M_{s} - m_s) \bar{Y}^*_{is}, \ i = 1,...,N \) with

\( \bar{Y}^*_{is} = \frac{1}{M_{s} - m_s} \sum_{j=m_{s}+1}^{M_{s}} \bar{Y}_{ij}. \) The model for \( Y^* \) is given by

\[
Y^* = X^* \begin{pmatrix} \mu_i^* \\ \mu_{ii}^* \end{pmatrix} + Z^* \begin{pmatrix} B^*_i \\ B^*_{ii} \end{pmatrix} + E^*
\]

where \( X^* = \bar{M} \begin{pmatrix} \bar{f} & 0 \\ 0 & 1 - \bar{f} \end{pmatrix} \otimes 1_N, \ Z^* = \bar{M} \begin{pmatrix} \bar{f} & 0 \\ 0 & 1 - \bar{f} \end{pmatrix} \otimes I_N, \ B^*_i = \left( (B_{ii}^*_i) \right) = U \beta_{ji} \) and

\( B^*_{ii} = \left( (B_{ii}^*_i) \right) = U \beta_{ii}. \) The fixed effects correspond to the expected PSU total for the potential
sample SSUs, \( \bar{m}_I \), and the expected PSU total for the remaining SSUs, \( (M - \bar{m}) \mu_II \). Notice that these parameters are defined as averages over all PSUs in the population. The random effects correspond to deviations of PSU totals (for the sample and the remainder) about their respective expected values, i.e. \( B_{ii}^s = \sum_{i=1}^N U_{ii} \beta_{ii}^s \) and \( B_{ill} = \sum_{i=1}^N U_{ii} \beta_{ill} \).

Notice that we can express \( \mu_{ii} = \mu + \bar{\beta}_I \) and \( \mu_{II} = \mu + \bar{\beta}_II \), which in matrix notation may be written as, 

\[
\left( \begin{array}{c} \mu_{I} \\ \mu_{II} \end{array} \right) = \left( \begin{array}{cc} 1 & 0 \\ 1 & 1 \end{array} \right) \left( \begin{array}{c} \mu \\ \bar{\beta}_I \\ \bar{\beta}_II \end{array} \right).
\]

Neither \( \bar{\beta}_I \) nor \( \bar{\beta}_II \) are necessarily equal to zero. The two additional parameters correspond to weighted averages of sample cluster effects, and weighted averages of remainder cluster effects. The variance of the collapsed random variables is

\[
\text{var}_{\gamma_{II}} (Y^c) = \bar{M}^2 \left( \begin{array}{cc} \bar{f}^2 \sigma_{I}^2 & \bar{f} (1 - \bar{f}) \sigma_{I,II} \\ \bar{f} (1 - \bar{f}) \sigma_{I,II} & (1 - \bar{f}) \sigma_{II}^2 \end{array} \right) \otimes P_N + \bar{M} \bar{f} (1 - \bar{f}) \sigma_{I}^2 \left( \begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array} \right) \otimes I_N.
\]

Next, we develop a predictor of the total for PSU \( i \) expressed as \( P_i^* = g_i^* Y^c \) where \( g_i = 1_2 \otimes e_i \). We require the predictor to be a linear function of the random variables in the sample, to be unconditionally unbiased, and to minimize the expected value of the mean squared error. First, we partition the random variables into a vector for the sample and a vector for the remainder by letting \( \left( Y^c_i \right)^T = (K'_I \ K'_II)^T \ Y^c \) where \( K'_I = \left( I_n \mid 0 \right)_{n \times (2N-n)} \) and

\[
K'_{II} = \left( \begin{array}{c} 0 \\ \left( K'_{II} \right)_{2 \times (N-n)} \end{array} \right).\]

Consequently, \( E_{\gamma_{II}} (Y_i^c) = \bar{m} I_n \mu_{I}^* \) and

\[
E_{\gamma_{II}} (Y_{II}^c) = \left( \begin{array}{cc} \bar{m} I_{N-n} & 0 \\ 0 & (M - \bar{m}) I_N \end{array} \right) \left( \begin{array}{c} \mu_{I}^* \\ \mu_{II}^* \end{array} \right).\]

The total for PSU \( i \) is given by \( P_i^* = g_{II}^* Y_i^c + g_{ill}^* Y_{II}^c \).
where \( g'_{ii} = e'_{ii}, \ g'_{III} = (e'_{III} \ | \ e'_{ii}) \), and \( e'_{i} = (e'_{ii} \ | \ e'_{III}) \) with the dimensions of \( e_{ii} \) and \( e_{III} \) being respectively, \( n \times 1 \) and \( (N - n) \times 1 \). Collection of the study data will result in realizing the values of \( Y'_i \).

We require the predictor of \( Y'_i \) to be an unbiased linear function of the sample data,

\[
\hat{P}'_i = (g' + a')Y'_i, \quad \text{such that} \quad E_{\hat{z}z} \left( \hat{P}'_i - P'_i \right) = 0. \quad \text{This unbiased constraint implies that}
\]

\[
a' E_{\hat{z}z} \left( Y'_i \right) - g' E_{\hat{z}z} \left( Y'^*_i \right) = 0, \quad \text{a condition that can be expressed as}
\]

\[
\frac{1}{N} \sum_{s=1}^{N} (c - \frac{1}{f_s}) m_s u_s = 0 \quad \text{for some constant} \ c. \quad \text{To satisfy the unbiased constraint, we require} \ c = \frac{1}{f_s} \quad \text{for all} \ s = 1, ..., N. \quad \text{Thus, an unbiased predictor of a PSU total exists only when second stage sampling fractions are equal for all clusters, i.e. second stage sampling is PPS.}
\]

Thus, we assume PPS second stage sampling to develop an unbiased predictor of the PSU total that minimizes the expected value of the MSE. The common second stage sampling fraction is represented by \( f \). With PPS sampling, \( f = \overline{f} \) and \( r_s = r_{ii} = r_{III} \) so that \( \mu_R = \mu'_i = \mu'_{II} \) and \( \sigma_R^2 = \sigma_i^2 = \sigma_{II}^2 = \sigma_{i,II}^2 \), and hence

\[
\var_{\hat{z}z} \left( Y'_i \right) = \overline{M}^2 \sigma_R^2 \left( \frac{f}{1-f} \right) ^2 \otimes P_N + \overline{M} f (1-f) \sigma_e^2 \left( \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right) \otimes I_N.
\]

The model for the sample is \( Y'_i = X'_i \mu_R + Z'_i \beta_R + E'_i \) where \( X'_i = \overline{M} f I_n \),

\[
Z'_i = \overline{M} \left( A_{n \times (N-n)} \right) \text{ and } B_R = U \beta_R, \quad \text{while the model for the remainder is}
\]
\(Y^*_{II} = X^*_{II}\mu^*_{R} + Z^*_{II}B_{R} + E^*_{II}\) where \(X^*_{II} = \bar{M}\left(\frac{f1_{N-n}}{(1-f)1_N}\right)\) and \(Z^*_{II} = \bar{M}\left(\frac{0}{(1-f)1_N}\right)\).

Partitioning the variance accordingly, we get \(\text{var}_{\hat{\theta}}\left(Y^*_{II}, Y^*_{II}\right) = \left(\begin{array}{c} V^*_i \\ V^*_{I,II} \\ V^*_i \\ V^*_{II} \end{array}\right)\) (see Appendix).

The best linear unbiased predictor of the total for PSU \(i\) is given by

\[\hat{P}_i = \mathbf{g}'_i Y_i + \mathbf{g}'_{II} \left[ X^*_i \hat{\mu}_R + V^*_i V^{-1}_i \left( Y_i - X^*_i \hat{\mu}_R \right) \right],\]

where \(\hat{\mu}_R = \left( X^*_i V^{-1}_i X^*_i \right)^{-1} X^*_i V^{-1}_i Y_i \). We express the total for PSU \(i\) as \(\mathbf{g}'_i Y_i\) where \(\mathbf{g}'_i = (\mathbf{g}'_d, \mathbf{g}'_{II})\), \(\mathbf{g}'_{II} = \mathbf{e}'_{II}\) and \(\mathbf{g}'_{II} = (\mathbf{e}'_{II}, \mathbf{e}'_d)\). Simplifying terms it follows that \(\hat{\mu}_R = \bar{Y}_{II}\) where \(\bar{Y}_{II} = \frac{1}{n} \sum_{i=1}^{n} Y_{II}\) and

\[\bar{Y}_{II} = \sum_{s=1}^{N} U_{is} r_s \bar{Y}_{II}\]. Then, using \(\bar{Y}_{II} = \left(\bar{Y}_{II}\right)\), \(X_i = \bar{M} f 1_n, X^*_i = \bar{M}\left(\frac{f1_{N-n}}{(1-f)1_N}\right)\), \(\mathbf{g}'_i = \mathbf{e}'_i\) and \(\mathbf{g}'_{II} = (\mathbf{e}'_{II}, \mathbf{e}'_d)\), when \(i \leq n\), we obtain

\[\hat{P}_i = \bar{M}\left(\mathbf{f}\bar{Y}_{II} + (1-f) \left[ \bar{Y}_{II} + k^* \left( \bar{Y}_{II} - \bar{Y}_{II} \right) \right] \right).\]

When \(i > n\), it follows that \(\hat{P}_i = \bar{M}\bar{Y}_{II}\).

Because of the unbiased condition, the MSE of the predictor is \(\text{var}_{\hat{\theta}}\left(\hat{P}_i - P_i^*\right)\). Using the expressions for the partitioned variance and simplifying, it follows that when \(i \leq n\),

\[\text{var}_{\hat{\theta}}\left(\hat{P}_i - P_i^*\right) = \bar{M}^2 (1-f) \left[\frac{\sigma^2_x}{nm} + \frac{n-1}{n} \left(1-k^*\right) \frac{\sigma^2_R}{n} \right]\]

while when \(i > n\),

\[\text{var}_{\hat{\theta}}\left(\hat{P}_i - P_i^*\right) = \bar{M}^2 (1-f) \left[\frac{\sigma^2_x}{nm} + \frac{n+1}{n} \frac{\sigma^2_R}{1-f} \right].\]

3.2. Predictors of PSU Means based on Weighted PSU Totals
Predictors of the PSU mean, \( P_i \), can be developed from a collapsed set of random variables corresponding to weighted PSU totals by letting \( Y^* = C'Y \) with

\[
C' = \left( I_N \otimes \left( C_i^1 I_N \right) \right) \left( I_N \otimes \left( C_i^N I_N \right) \right)', \quad \text{where} \quad C_i^1 = \bigoplus_{s=1}^{N} \left( \frac{1}{M_s} I'_{m_s} 0 \right)' \quad \text{and}
\]

\[
C_i^N = \bigoplus_{s=1}^{N} \left( 0 \frac{1}{M_s} I'_{M_s-M_s} \right)'.
\]

The vector of random variables \( Y^* \) has dimension \( 2N \times 1 \), with the first \( N \) elements consisting of weighted totals of the sample SSUs for each PSU,

\[
\bar{Y}_{il} = \sum_{s=1}^{N} U_{is} f_s \bar{Y}_{isl}, \quad i = 1, \ldots, N,
\]

and the second set of \( N \) elements consisting of weighted totals of the remaining SSUs for each PSU, \( \bar{Y}_{ill} = \sum_{s=1}^{N} U_{is} (1-f_s) \bar{Y}_{isl}, \quad i = 1, \ldots, N \). The model for \( Y^* \) is given by

\[
Y^* = X^* \left( \begin{array}{c} \mu_i^* \\ \mu_{II}^* \end{array} \right) + Z^* \left( \begin{array}{c} B_i^* \\ B_{II}^* \end{array} \right) + E^*.
\]

where \( X^* = \frac{1}{M} X^* \), \( Z^* = \frac{1}{M} Z^* \), \( B_i^* = \left( (B_{ii}^*) \right) = U \beta^*_i \) and \( B_{II}^* = \left( (B_{II}^*) \right) = U \beta_{II}^* \). The fixed effects can be reparameterized by writing the models for the PSU sample and remainder totals in terms of the population mean and weighted average deviations, such that

\[
\left( \begin{array}{c} \mu_i^* \\ \mu_{II}^* \end{array} \right) = \frac{f}{f} \begin{pmatrix} 1 & 0 \\ 1-f \end{pmatrix} \begin{pmatrix} 1-f \end{pmatrix} \begin{pmatrix} 1 \\ 1-f \end{pmatrix} \begin{pmatrix} \mu \\ \beta_i \\ \beta_{II} \end{pmatrix}.
\]

The corresponding variance is given by

\[
\text{var}_{\hat{\sigma}^2_{i,II}}(Y^*) = \frac{\hat{f}^2 \hat{\sigma}^2_{i} \sigma^2_{i,II}}{\hat{f}(1-\hat{f}) \sigma^2_{i,II}} \otimes P_N + \frac{\hat{f}(1-\hat{f}) \sigma^2_{e}}{M},
\]

\[
\otimes P_N + \frac{\hat{f}(1-\hat{f}) \sigma^2_{e}}{M} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes I_N.
\]
We develop a predictor of the mean for PSU $i$ defined by $P_i = g_i'Y^*$ where $g_i = I_2 \otimes e_i$. The development is similar to that in Section 3.1. We require the predictor to be a linear function of the random variables in the sample, to be unconditionally unbiased, and to minimize the expected value of the mean squared error. First, we partition the random variables into a vector for the sample and the remainder by setting $(Y^*_{i1} \ Y^*_{i2})' = (K_i' \ K''_i)' Y^*$, and express the mean for PSU $i$ as $P_i = g'_iY^*_{i1} + g''_iY^*_{i2}$. We require the predictor of $P_i$ to be an unbiased linear function of the sample data, $\hat{P}_i^* = (g'_i + a')Y^*_{i1}$, such that $E_{\xi\xi_{i2}}(\hat{P}_i^* - P_i) = 0$. The unbiased constraint will be satisfied only when $f_x$ is constant for all $s = 1,...,N$, implying that an unbiased predictor of the mean for PSU $i$ exists only when the second stage sampling is PPS.

Assuming PPS sampling, we have $\mu_{i1}^* = \mu_{i2}^* = \mu_s$ and hence $\mu_i^* = \mu_{i2}^* = (\mu_s) = \mu$ and $\mu_i^* = \mu_{i2}^* = \mu$. Defining $B = U\beta$ and $\beta = ((\beta_s)) = \mu - \mu 1_N$, the model for the sample simplifies to $Y_s^* = X_i^*\mu + Z_i^*B + E_i^*$ where $X_i^* = fI_n$ and $Z_i^* = \left(f I_n 0 \atop n(1-n)\right)$, while the model for the remainder is given by $Y_{i2}^* = X_{i2}^*\mu + Z_{i2}^*B + E_{i2}^*$ where $X_{i2}^* = \left(f I_{N-n} (1-f) I_N\right)'$ and $Z_{i2}^* = \left(0 n(1-n) \atop n(1-n)\right)(1-f)I_N$. The corresponding variance simplifies to $\text{var}_{\xi\xi_{i2}}(Y^*) = \sigma^2 \left( f \begin{pmatrix} 1 \ 1-f \end{pmatrix} \otimes P_N + f (1-f) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes I_N \right)$ since $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_{i1,ii}^2 = \sigma^2$. We partition the variance in a similar manner as in Section 3.1 to form the predictor.

The best linear unbiased predictor of a PSU mean when $i \leq n$ is
\[
\hat{P}_i^* = f\bar{Y}_i + (1-f)\left[\bar{Y} + k^* (\bar{Y}_i - \bar{Y})\right]
\]

and when \( i > n \)

\[
\hat{P}_i^* = \bar{Y}
\]

where \( \bar{Y}_i = \left(\left(\bar{Y}_i\right)\right) \), \( \bar{Y}_i = \sum_{s=1}^{N} U_{is} \bar{Y}_{sl} \) and \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i \). The MSE of the predictor when \( i \leq n \) is

\[
\text{var}_{\hat{P}_i} (\hat{P}_i^* - P_i) = (1-f)\left[\frac{\sigma^2}{mn} + \left(\frac{n-1}{n}\right)(1-k^*)\sigma^2\right].
\]

while when \( i > n \), \( \text{var}_{\hat{P}_i} (\hat{P}_i^* - P_i) = (1-f)\left[\frac{\sigma^2}{mn} + \left(\frac{n+1}{n}\right)\sigma^2\right].\]

3.3. Predictors of a PSU Mean based on Sampled and Remaining SSU Averages

Predictors of a PSU mean can be developed from a collapsed set of random variables corresponding to sampled and remaining SSU averages by setting \( \mathbf{C}^* \mathbf{Y} = \mathbf{Y}^* \) where

\[
\mathbf{C}^* = \left(\mathbf{I}_N \otimes \left(\mathbf{C}_i^* 1_N\right)\right) + \left(\mathbf{I}_N \otimes \left(\mathbf{C}_f^* 1_N\right)\right), \quad \text{with} \quad \mathbf{C}_i^* = \bigoplus_{s=1}^{N} \left(\frac{1}{m_s} \mathbf{1}_{m_s} - \mathbf{0}_{(M_s - m_s)}\right) \quad \text{and} \quad \mathbf{C}_f^* = \bigoplus_{s=1}^{N} \left(\mathbf{0}_{m_s} - \frac{1}{M_s - m_s} \mathbf{1}_{(M_s - m_s)}\right).
\]

This form of collapsing the relevant random variables enables best linear unbiased predictors to be developed for second stage sampling plans when sampling fractions differ between clusters. The \( 2N \times 1 \) vector of random variables \( \mathbf{Y}^* \) is such that its first \( N \) elements consist of PSU sample means, \( \bar{Y}_i \), \( i = 1, \ldots, N \), and the second set of \( N \) elements consist of means for the remaining SSUs in each PSU, \( \bar{Y}_{sl} = \sum_{s=1}^{N} U_{is} \bar{Y}_{sl} \). Using this collapsed set of random variables, the model simplifies to
\[ Y^* = X^* \mu + Z^* B + E^* \]

where \( X^* = \mathbf{1}_{2N} \) and \( Z^* = \mathbf{1}_2 \otimes \mathbf{1}_N \), and \( E^* = \mathbf{C}^r \mathbf{E} \). The variance is

\[
\operatorname{var}_{\mathbf{Y}^*} \left( Y^* \right) = \begin{pmatrix} \frac{\sigma^2}{m} & 0 \\ 0 & \frac{\sigma^2}{s} \end{pmatrix} \otimes \mathbf{I}_N + \sigma^2 \mathbf{J}_2 \otimes \mathbf{I}_N - \frac{\sigma^2}{N} \mathbf{J}_2 \otimes \mathbf{J}_N.
\]

In order to develop a predictor of the PSU mean, we define a linear combination of the random variables as

\[ P_i^* = g_i^* \mathbf{Y}^* = \sum_{s=1}^{N} U_{is} \left( c_{i} \tilde{Y}_{i} + \left(1 - c_{i}\right) \bar{Y}_{i} \right) \]

where \( g_i^* = (c_{i} \ 1 - c_{i}) \otimes \mathbf{e}^r_i \) for some constant \( c_{i} \). Notice that \( c_{i} \) depends on the position of the PSU, but not on the realized cluster. This random variable, although not necessarily equal to the PSU mean, has an expected value (over SSUs) that equals the mean for PSU \( i \), i.e., \( E_{\tilde{Y}} \left( P_i^* \right) = P_i \).

We develop a predictor of \( P_i^* \), not \( P_i \).

Again, we partition \( \mathbf{Y}^* \) into a sample and remaining vector by letting

\[ \begin{pmatrix} \mathbf{K}_I \\ \mathbf{K}_{II} \end{pmatrix} \mathbf{Y}^* = \begin{pmatrix} \mathbf{Y}_I^* \\ \mathbf{Y}_{II}^* \end{pmatrix} \]

and similarly, we partition \( g_i^* \) by setting \( g_i^* \left( \mathbf{K}_I' \quad \mathbf{K}_{II}' \right) = \left( g_{ii}^* \quad g_{iil}^* \right) \) such that \( g_{ii}^* = c_{i} \mathbf{e}_{ii}' \) and \( g_{iil}^* = (c_{i} \mathbf{e}_{iil}' \ 1 - c_{i}) \mathbf{e}^r_i \). Collection of the study data will result in realizing the values of \( \tilde{Y}_I \).

We require the predictor of \( P_i^* \) to be a linear function of the sample data, \( \hat{P}_i^* = (g_{ii}^* + a') \tilde{Y}_I \) and be unbiased so that \( E_{\tilde{Y}_I} \left( \hat{P}_i^* - P_i^* \right) = 0 \). The constraint will be satisfied if \( a' = 1 - c_{i} \) for \( i \leq n \), or if \( a' = 1 \) for \( i > n \). The best linear unbiased predictor is

\[ \hat{P}^* = g_{ii}^* \tilde{Y}_I + g_{iil}^* \left[ X^* \tilde{Y} + V^* i_{iI, II} V_i^{-1} (\tilde{Y}_I - X^* \tilde{Y}) \right] \]
which simplifies to
\[
\hat{P}_i^* = c_iY_i + (1-c_i)\left[\bar{Y} + k^*\left(\bar{Y} - \bar{Y}\right)\right]
\]
when \(i \leq n\) and to \(\hat{P}_i^* = \bar{Y}\) when \(i > n\). When \(i \leq n\), the expected MSE simplifies to
\[
\text{var}_{\xi_{Y_i}}(\hat{P}_i^* - P_i^*) = (1-c_i)\left[\frac{\sigma_{\varepsilon_i}^2}{\bar{m}n} + \left(\frac{n-1}{n}\right)(1-k^*)\sigma^2\right] + (1-c_i)\left[\left(\frac{1-c_i}{1-f}\right)\frac{\sigma_{\varepsilon_{\ell \ell}}^2}{M} - \left(\frac{\sigma_{\varepsilon_i}^2 + c_i\bar{M}\sigma_{\varepsilon_{\ell \ell}}^2}{\sigma_{\varepsilon_{\ell \ell}}^2 + f\bar{M}\sigma_{\varepsilon_{\ell \ell}}^2}\right)\frac{\sigma_{\varepsilon_{\ell \ell}}^2}{\bar{M}}\right]
\]
and when \(i > n\) it reduces to
\[
\text{var}_{\xi_{Y_i}}(\hat{P}_i^* - P_i^*) = \frac{\sigma_{\varepsilon_{\ell \ell}}^2}{nm} + \sigma^2\left(\frac{n+1}{n}\right) + \left[\frac{c_i}{f}\right]\left(\frac{\sigma_{\varepsilon_{\ell \ell}}^2}{M}\right) + (1-c_i)\left(\frac{1-c_i}{1-f}\right)\left(\frac{\sigma_{\varepsilon_{\ell \ell}}^2}{M} - \frac{\sigma_{\varepsilon_i}^2}{\bar{M}}\right).
\]

4. APPLICATION

We illustrate predictors for a selected physician’s practice in the randomized controlled trial described in the Introduction. Patients enrolled in the study had duplicate venous cholesterol measures made at baseline and at 1 year follow-up. In addition, a 24 hour dietary recall was reported at baseline and at the 1 year follow-up. Data from these measures, in addition to the baseline questionnaire data and intervention data were used for study evaluation.

We limit discussion to a single treatment arm (III) to which fourteen physician practices were randomly assigned from \(N = 60\). A summary of the average change for different sample physician practices assigned to this intervention is given in Table 1.
Table 1. List of Average Change in Saturated Fat as Pct of Kcal by Physician for n=14 physicians in the Counseling and Support Intervention

<table>
<thead>
<tr>
<th>ID #</th>
<th>76</th>
<th>92</th>
<th>84</th>
<th>81</th>
<th>71</th>
<th>91</th>
<th>75</th>
<th>93</th>
<th>95</th>
<th>83</th>
<th>73</th>
<th>82</th>
<th>74</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>1</td>
<td>30</td>
<td>11</td>
<td>42</td>
<td>25</td>
<td>20</td>
<td>25</td>
<td>15</td>
<td>10</td>
<td>29</td>
<td>15</td>
<td>38</td>
<td>16</td>
<td>33</td>
</tr>
<tr>
<td>Mean</td>
<td>-3.25</td>
<td>-2.31</td>
<td>-2.16</td>
<td>-2.09</td>
<td>-1.47</td>
<td>-1.38</td>
<td>-0.92</td>
<td>-0.82</td>
<td>-0.69</td>
<td>-0.46</td>
<td>0.36</td>
<td>0.37</td>
<td>0.75</td>
<td>0.91</td>
</tr>
<tr>
<td>Var</td>
<td>25.36*</td>
<td>35.48</td>
<td>33.44</td>
<td>34.18</td>
<td><strong>19.88</strong></td>
<td>21.82</td>
<td>23.29</td>
<td>14.59</td>
<td>22.39</td>
<td>18.70</td>
<td>28.73</td>
<td>27.70</td>
<td>17.53</td>
<td>20.97</td>
</tr>
</tbody>
</table>

* Weighted average of variances based on estimated variances from other realized PSUs

We focus on predicting the average change in the percent of Kcal from saturated fat for a particular realized physician practice (ID=71) for which the observed average change was -1.47% Kcal for the 25 sample patients. This estimate of the latent value for practice #71 is derived from a model that represents physician practices as fixed effects. This estimator is the best linear unbiased estimator of the parameter for cluster #71. Many practitioners still use such an estimator, ignoring the first stage sampling of clusters.

Using a model that accounts for the two stage design, the parameter for a cluster corresponds to the latent value of a realized PSU. Predictors of the realized latent value have smaller expected MSE than the sample cluster mean. We can represent the predictors as the sum of two terms, one for the sample and the other for the remainder. For comparison, we have
represented the sample mean as a similar sum by including an arbitrary non-zero constant, \( c_i^* \) (see Table 2).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Target</th>
<th>Sample SSUs</th>
<th>Remaining SSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{Y}_i )</td>
<td>Simple Mean</td>
<td>( \mu,</td>
<td>U_{is}=1 )</td>
</tr>
<tr>
<td>( \hat{P}_{imm} )</td>
<td>Mixed Model</td>
<td>( P_i )</td>
<td>( \hat{\mu} + k_i (\bar{Y}_i - \hat{\mu}) )</td>
</tr>
<tr>
<td>( \hat{P}_{iss} )</td>
<td>Scott&amp;Smith Model</td>
<td>( P_i )</td>
<td>( f_i \bar{Y}_i ) + ( (1-f_i) \left[ \hat{\mu}^* + k_i^* (\bar{Y}_i - \hat{\mu}^*) \right] )</td>
</tr>
<tr>
<td>( \hat{P}_i^* )</td>
<td>RP (PPS)</td>
<td>( P_i )</td>
<td>( f_i \bar{Y}_i ) + ( (1-f) \left[ \bar{Y} + k^* (\bar{Y} - \bar{Y}) \right] )</td>
</tr>
<tr>
<td>( \hat{P}_i^* )</td>
<td>RP (General)</td>
<td>( P_i^* )</td>
<td>( c_i \bar{Y}_i ) + ( (1-c_i) \left[ \bar{Y} + k^* (\bar{Y} - \bar{Y}) \right] )</td>
</tr>
</tbody>
</table>

The model underlying the mixed model predictor is posited for a conceptually very large clustered population; it does not account for possibly different size clusters. The superpopulation model and the random permutation model incorporate cluster and sample sizes, and require these characteristics to evaluate the predictors. When second stage samples are not selected with PPS
sampling, the target random variable, $P_i^*$, is defined in terms of a non-zero constant $c_i$. This constant can be arbitrary and still $E_{\xi_i}(P_i^*) = P_i$.

In the Watch II study, physician practice sizes differed, but the practice size was not recorded. We make two assumptions about population sizes to illustrate differences in the predictors, first assuming half the eligible patients in a practice are included in the sample, and second assuming all physician practices have $M = 100$ eligible patients. The predictors of the latent value for a realized PSU require knowledge (or estimates) of the population variance components. They are based on different models. To evaluate a predictor, we replace cluster variance components by their unbiased estimates. We use the mixed model restricted maximum likelihood estimate of the variance between clusters, $\hat{\sigma}^2 = 0.431$, and the results are summarized in Table 3.
Table 3. Predictors of the Latent Value for Physician Practice #71 using Empirical Variances assuming \( M = 100 \) Eligible Patients in each Cluster.

**PPS Sampling (\( f = 0.5 \))**

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor=</th>
<th>Sample SSUs</th>
<th>Remaining SSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mean</td>
<td>(-1.47 = ) ( c_i^* (-1.47) )</td>
<td>((1 - c_i^*)[-1.47])</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>(-0.96 = )</td>
<td>(-0.69 + 0.35[-1.47 - (-0.69)])</td>
<td></td>
</tr>
<tr>
<td>Scott and Smith</td>
<td>(-1.22 = (0.5)(-1.47) ) + ((1-0.5)(-0.69 + 0.35[-1.47 - (-0.69)]))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rand. Perm.*</td>
<td>(-1.19 = (0.5)(-1.47) ) + ((1-0.5)(-0.94 - 0.04[-1.47 - (-0.94)]))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Unequal Sampling of SSUs (\( M = 100 \) for all PSUs, varying \( f_i \))**

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor=</th>
<th>Sample SSUs</th>
<th>Remaining SSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mean</td>
<td>(-1.47 = )</td>
<td>(-1.47)</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>(-0.96 = )</td>
<td>(-0.69 + 0.35[-1.47 - (-0.69)])</td>
<td></td>
</tr>
<tr>
<td>Scott and Smith</td>
<td>(-1.09 = (0.25)(-1.47) ) + ((1-0.25)(-0.69 + 0.35[-1.47 - (-0.69)]))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rand. Perm.*</td>
<td>(-1.08 = (0.22)(-1.47) ) + ((1-0.22)(-0.94 + 0.06[-1.47 - (-0.94)]))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* \( c_i \) is set equal to \( \bar{m}/\bar{M} \)
The differences in the predictors in each panel in Table 3 highlight the different model assumptions. Each of the predictors has lower expected mean squared error relative to the simple mean estimator when evaluated over the same probability framework. For example, with PPS sampling, the expected MSE for the sample mean and random permutation model are 1.48 and 0.42, respectively.

5. DISCUSSION

Predicting random effects based on a two-stage random permutation model for cluster sampling is complicated by unequal size clusters, necessitating careful tracking of target random variables through the development of predictors. Tracing through this development, while more complex, retains a clear connection between the target random variable and the latent value of a the PSU. Avoiding these complications with other methods requires additional assumptions and introduces ambiguity in the interpretation of predictors. Perhaps more importantly, the usual mixed model methods ignores the structure of the actual population, and thus prevents tailoring of methods that account for such structure.

We have illustrated that it is possible to directly account for the physical randomization in a two stage simple random sample of an unequal size clustered population. This provides a design-based framework for prediction similar to that used in equal size cluster populations by Stanek and Singer (2004). In so doing, expanded sets of random variables are introduced to avoid random dimensions of PSUs, and then collapsed to PUS random variables prior to prediction. This strategy has the potential to be extended to include additional cluster level
covariates, similar to work of Li (2003). We anticipate that extensions to include unit level covariates will require further expansion of the random variables for SSUs.

The development highlights other aspects of mixed model and super-population model predictors that differ from those developed here. Mixed model predictors are functions of the cluster sample size and cluster variances conditional on the realized cluster. As a result, the predictors can not be motivated as linear predictors of the random variables unconditionally in any of the collapsed settings. It may be possible to specify a different class of predictors in a random permutation model framework that come closer to the mixed model predictors. Additional work is being conducted in this direction.
APPENDIX: EVALUATION OF $\text{var}_{\xi_1 \xi_2}(Y)$, $\text{var}_{\xi_1 \xi_2}(Y')$

We outline the steps leading to the expression for $\text{var}_{\xi_1 \xi_2}(Y)$ using the conditional expansion $\text{var}_{\xi_1 \xi_2}(Y) = \text{var}_{\xi_1}(E_{\xi_2|\xi_1}[Y]) + E_{\xi_1}(\text{var}_{\xi_2|\xi_1}[Y])$. Recall that the vector $Y = (Y'_1 \ Y'_2 \ \cdots \ Y'_N)'$ is constructed from sub-vectors for each PSU given by

$$Y_i = \left( U_{i1} \tilde{Y}'_1 \ U_{i2} \tilde{Y}'_2 \ \cdots \ U_{iN} \tilde{Y}'_N \right)'$$

where $\tilde{Y}_i = \left( (\tilde{y}_{sj}) \right) = U^{(s)y}_i Y_s, \ \tilde{y}_{sj} = \sum_{t=1}^{M_i} U^{(s)}_{jt} y_{st}$ and $U^{(s)}$ is an $M_s \times M_s$ matrix of indicator random variables, $U^{(s)} = \left( U^{(s)}_1 \ U^{(s)}_2 \ \cdots \ U^{(s)}_{M_s} \right)$ with

$$U_i^{(s)} = \left( U_i^{(s)} \ U_s^{(s)} \ \cdots \ U_{M_s}^{(s)} \right)'$$

We begin by obtaining the expected value and variance of these sub-vectors.

First, since $E_{\xi_2|\xi_1}(U^{(s)}) = \frac{1}{M_s} J_{M_s}$, it follows that $E_{\xi_2|\xi_1}(Y_i) = \left( \bigoplus_{s=1}^{N} \mathbf{1}_{M_s}, \mu_s \right)$.

Defining

$$U = \left( U_1 \ U_2 \ \cdots \ U_N \right)$$

where $U_s = \left( U_{1s} \ U_{2s} \ \cdots \ U_{Ns} \right)'$ and combining terms,

$$E_{\xi_2|\xi_1}(Y) = \left[ I_N \otimes \left( \bigoplus_{s=1}^{N} \mathbf{1}_{M_s}, \mu_s \right) \right] \text{vec}(U)$$

Now, we develop an expression for

$$\text{var}_{\xi_1}(\text{vec}(U)) = E_{\xi_1} \begin{pmatrix} U_1 U_1' & U_1 U_2' & \cdots & U_1 U_N' \\ U_2 U_1' & U_2 U_2' & \cdots & U_2 U_N' \\ \vdots & \vdots & \ddots & \vdots \\ U_N U_1' & U_N U_2' & \cdots & U_N U_N' \end{pmatrix} - \frac{1}{N^2} J_N \otimes J_N$$

by noting that under simple random without replacement sampling, $E_{\xi_1}(U_s U_s') = \frac{1}{N} \mathbf{1}_N$, while when $s \neq s^*$.
\[
E_{s_i}(U_s U'_{s_i}) = \frac{1}{N(N-1)}(J_N - I_N).
\]
Substituting these terms and simplifying,

\[
\text{var}_{s_i}(\text{vec}(U)) = \frac{1}{N-1}P_N \otimes P_N,
\]
and hence

\[
\text{var}_{s_i}(E_{s_i}[Y]) = \frac{1}{N-1}P_N \otimes \left[ \left( \bigoplus_{s=1}^{N} I_{M_s}, \mu_s \right) P_N \left( \bigoplus_{s=1}^{N} I_{M_s}, \mu_s \right) \right].
\]

To evaluate \( E_{s_i}(\text{var}_{s_i}[Y]) \) we expand the terms in

\[
\text{var}_{s_i}[Y] = \begin{pmatrix}
\text{var}_{s_i}(Y_1) & \text{var}_{s_i}(Y_1, Y'_2) & \cdots & \text{var}_{s_i}(Y_1, Y'_N) \\
\text{var}_{s_i}(Y_2, Y'_1) & \text{var}_{s_i}(Y_2) & \cdots & \text{var}_{s_i}(Y_2, Y'_N) \\
\vdots & \vdots & \ddots & \vdots \\
\text{var}_{s_i}(Y_N, Y'_1) & \text{var}_{s_i}(Y_N, Y'_2) & \cdots & \text{var}_{s_i}(Y_N)
\end{pmatrix}
\]

Now

\[
\text{var}_{s_i}(Y, Y') = \begin{pmatrix}
U_{i1}U_{j1} \text{var}_{s_i}(\tilde{Y}_1) & U_{i1}U_{j2} \text{var}_{s_i}(\tilde{Y}_1, \tilde{Y}_2') & \cdots & U_{i1}U_{jN} \text{var}_{s_i}(\tilde{Y}_1, \tilde{Y}_N') \\
U_{i2}U_{j1} \text{var}_{s_i}(\tilde{Y}_2, \tilde{Y}_1') & U_{i2}U_{j2} \text{var}_{s_i}(\tilde{Y}_2) & \cdots & U_{i2}U_{jN} \text{var}_{s_i}(\tilde{Y}_2, \tilde{Y}_N') \\
\vdots & \vdots & \ddots & \vdots \\
U_{iN}U_{j1} \text{var}_{s_i}(\tilde{Y}_N, \tilde{Y}_1') & U_{iN}U_{j2} \text{var}_{s_i}(\tilde{Y}_N, \tilde{Y}_2') & \cdots & U_{iN}U_{jN} \text{var}_{s_i}(\tilde{Y}_N)
\end{pmatrix}
\]

Furthermore, since permutations of SSUs are independent for different PSUs,

\[
\text{var}_{s_i}(\tilde{Y}, \tilde{Y}') = \bigoplus_{s=1}^{N} \text{var}_{s_i}(\tilde{Y}_s).
\]

Since \( \tilde{Y}_i = \text{vec}(\tilde{Y}_i) = (I_{M_s} \otimes y'_s) \text{vec}(U^{(s)}) \), we find that

\[
\text{var}_{s_i}(\text{vec}(U^{(s)})) = \frac{1}{M_s-1}P_{M_s} \otimes P_{M_s},
\]
and hence, \( \text{var}_{s_i}(\tilde{Y}_s) = P_{M_s} \sigma_s^2 \) where \( \sigma_s^2 = \frac{1}{M_s-1}y'_s P_{M_s} y_s \).

Using these expressions, we get

\[
\text{var}_{s_i}(Y) = \begin{pmatrix}
\bigoplus_{s=1}^{N} U_{1s} U_{1s} P_{M_s} \sigma_s^2 & \bigoplus_{s=1}^{N} U_{1s} U_{2s} P_{M_s} \sigma_s^2 & \cdots & \bigoplus_{s=1}^{N} U_{1s} U_{Ns} P_{M_s} \sigma_s^2 \\
\bigoplus_{s=1}^{N} U_{2s} U_{1s} P_{M_s} \sigma_s^2 & \bigoplus_{s=1}^{N} U_{2s} U_{2s} P_{M_s} \sigma_s^2 & \cdots & \bigoplus_{s=1}^{N} U_{2s} U_{Ns} P_{M_s} \sigma_s^2 \\
\vdots & \vdots & \ddots & \vdots \\
\bigoplus_{s=1}^{N} U_{Ns} U_{1s} P_{M_s} \sigma_s^2 & \bigoplus_{s=1}^{N} U_{Ns} U_{2s} P_{M_s} \sigma_s^2 & \cdots & \bigoplus_{s=1}^{N} U_{Ns} U_{Ns} P_{M_s} \sigma_s^2
\end{pmatrix}
\]
Finally, evaluating the expected value over permutations of PSUs,

\[ E_{\pi} \left( \text{var}_{x_{i|N}} \left( \mathbf{Y} \right) \right) = \mathbf{I}_N \otimes \left( \sum_{s=1}^{N} \frac{1}{N} \mathbf{P}_s \sigma^2_s \right). \]

Under PPS sampling,

\[ \text{var}_{x_{i|N}} \left( \mathbf{Y} \right) = \bar{M}^2 \sigma_R^2 \left( \frac{f}{1-f} \mathbf{I}_{N-N} \right) \otimes \mathbf{P}_N + \bar{M}f \left( 1-f \right) \sigma^2 \left( \frac{1}{1-f} \right) \otimes \mathbf{I}_N \] and

\[ \text{var}_{x_{i|N}} \left( \mathbf{Y}^{\prime} \right) = \begin{pmatrix} \mathbf{V}_I & \mathbf{Y}_{I} \mathbf{V}_{II} \\ \mathbf{V}_{I}^* & \mathbf{V}_{II} \mathbf{V}_{II}^* \end{pmatrix}, \text{ such that} \]

\[ \text{var}_{x_{i|N}} \left( \mathbf{Y} \right) = \begin{pmatrix} f^2 \left( \mathbf{I}_{n-N} - \frac{1}{N} \mathbf{J}_n \right) & \left( \frac{f}{1-f} \right) \mathbf{I}_{n-N} - \frac{1}{N} \mathbf{J}_n \left( 1-f \right) \end{pmatrix} \]

\[ \bar{M}^2 \sigma_R^2 \begin{pmatrix} -f^2 \frac{1}{N} \mathbf{J} & \mathbf{I}_{n-N} - \frac{1}{N} \mathbf{J}_{n-N} \\ \mathbf{I}_{n-N} - \frac{1}{N} \mathbf{J}_{n-N} & -f \mathbf{I}_{n-N} - \frac{1}{N} \mathbf{J}_{n-N} \end{pmatrix} \]

\[ + \bar{M}f \left( 1-f \right) \sigma^2 \begin{pmatrix} \mathbf{I}_{n-N} & \mathbf{I}_{n-N} \\ \mathbf{I}_{n-N} & \mathbf{I}_{n-N} \end{pmatrix} \]
REFERENCES


