Predicting Realized Cluster Parameters from Two Stage Samples of Unequal Size Clustered Population
(formerly c04ed21.doc, c04ed11.doc, c02ed43v1.doc)

Edward J. Stanek III
Department of Biostatistics and Epidemiology
404 Arnold House
University of Massachusetts
715 North Pleasant Street
Amherst, MA 01003-9304 USA
stanek@schoolph.umass.edu

Julio Singer
Departamento de Estatística
Universidade de São Paulo
São Paulo, Brazil
jmsinger@ime.usp.br
[Comment for discussion or abstract: Basically, the problem is that the dimension of the random variables in a PSU depends on the dimension of the cluster that is realized. This problem is a result of limitations in the traditional representation used for random variables in the context of two stage permutations. The ambiguity is a result of inadequate notation; it gives rise to ambiguity in analysis and interpretation.]
Contact Address:
Edward J. Stanek III
Department of Biostatistics and Epidemiology
404 Arnold House
University of Massachusetts at Amherst
Amherst, Ma. 01002

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

KEYWORDS: super-population, best linear unbiased predictor, two-stage sampling, random permutation, optimal estimation, design-based inference

ACKNOWLEDGEMENT. Many people contributed to the development of ideas that lead to this manuscript, including Dalton Andrade, Heleno Bolfarine, John Buonaccorsi, Viviana Lencina, and Oscar Loureiro. 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), FINEP (PRONEX), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil and the National Institutes of Health (NIH-PHS-R01-HD36848), USA.
1. INTRODUCTION

We develop design-based methods for predicting the mean or total of a realized cluster based on a two-stage cluster random sample from a finite population formed from 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 super-population models (Scott and Smith 1969; Bolfarine and Zacks 1992; Valliant, Dorfman and Royall 2000) that account for different size clusters of super-population random variables. Design-based predictors based on a two-stage random permutation model have been developed, but are limited to equal cluster size settings (Stanek and Singer 2004), possibly including response error.

We extend the random permutation framework to clustered populations where the number of units in clusters are not equal, and use it to predict target random variables that correspond to the average (or total) response for units in a cluster (which we call the latent value). The prediction is based solely on the two stage sampling design; no hypothetical super-population is required, and accounting is made for different size clusters in the finite population.

Mixed model and super-population model approaches can be directly applied to estimate fixed effects in unequal size clustered populations. These approaches can be contrasted with design-based approaches in survey sampling (Sarndal, 1992 #1024) (Cochran 1977) where estimators depend on sampling fractions, and become more complex along with the sampling design. The additional complexity is due in part to maintaining an account of the population complexity.

As an example, we consider 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, Hebert, Ockene, Saperia, Nicolosi, Merriam and Hurley 1996). Forty-five primary care internists were selected from 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 25th percentile, having previously-scheduled physician visits, were recruited into the study. The number of sampled patients per 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, percent of Kcal from saturated fat, which we consider as a difference constructed by taking 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 basic 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 super-population sampling models (Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). Let a finite population be defined by a listing of \( t = 1, \ldots, M_s \) units in each of \( s = 1, \ldots, N \) clusters, where response for unit \( t \) in cluster \( s \) is given by \( y_{st} \). We assume that a two-stage cluster sample of \( n \) clusters from this population (without replacement) 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 \( i = 1, \ldots, n \) 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 \( s \), whose positions are labeled by \( j = 1, \ldots, M_s \). We again assume that the first \( j = 1, \ldots, m_s \) 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 the \( i^{th} \) primary sampling unit (PSU), and to the unit that will occupy position \( j \) in the permutation as the \( j^{th} \) secondary sampling unit (SSU). Using these ideas and notation, we represent a two-stage permutation of the population as an ordered list of

\[
N = \sum_{s=1}^{N} M_s \text{ random variables, where the random variable representing response for SSU } j \text{ in PSU } i \text{ is } Y_{ij}.
\]

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 expected value of response for the cluster 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 (noting that such values are often defined as an overall mean plus a deviations relative to the overall mean). In this context, mixed model and super-population modeling strategies may be advocated for predicting the latent value of a realized PSU, or linear combinations of such latent values. We first review these approaches. We then define the context for finite population sampling prediction 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

\[
Y_{ij} = \mu + B_i + E_{ij} \quad (1)
\]
where \( \mu \) corresponds to the overall population mean, \( B_i \) is a random effect that corresponds to the deviation from the population mean of the latent value of PSU \( i \), and \( E_{ij} \) is the random error corresponding to the deviation of response for SSU \( j \) from the mean of PSU \( i \).

Typically, additional assumptions are made. The random effects, \( B_i \), are assumed independent and identically normally distributed, \( B_i \sim iid \ N(0, \sigma^2_B) \), and independent of the random errors, also assumed independent and identically distributed for all \( j \) in \( i \), \( E_{ij} \sim iid \ N(0, \sigma^2) \). The model and assumptions enable the joint distribution of fixed and random effects to be specified. This 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, \( \hat{\mu} = \sum_{i=1}^{n} w_i \bar{Y}_i \) where

\[
w_i = \frac{1}{\sum_{i=1}^{a} 1/\nu_i}, \quad \bar{Y}_i = \frac{\sum_{j=1}^{m_i} Y_{ij}}{m_i}, \quad \text{and} \quad \nu_i = \sigma^2 + \frac{\sigma^2_B}{m_i}.
\]

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 given by

\[
\hat{\mu} + k_i \left( \bar{Y}_i - \hat{\mu} \right) \quad (2),
\]

where \( k_i = \frac{\sigma^2}{\nu_i} \).

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 (2) is the expected value of the posterior distribution, assuming normality. The normality assumptions are not necessary, and can be replaced by knowledge 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 Super-population Sampling Models

Predictors of a linear combination of fixed and random effects (corresponding to the average response for a PSU) in two-stage sampling settings were developed by Scott and Smith (1969) using a model based survey sampling approach. In their approach, a finite population is viewed as the realization of a set of random variables. We refer to the set of random variables as the super-population, and specify a model for it. The parameters of interest are linear combinations of values in the finite population, such as the latent value for a realized PSU. Such parameters can be defined for all clusters in the finite population. Since only a part of the finite population is observed in a sample, the essential statistical problem is how best to predict the values of the remaining random variables. Predictors are constructed using the joint distribution assumed for the super-population (Royall 1976; Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). This approach to statistical
inference in survey sampling is called model-based, since inference is based on the probability model assumed for the super-population.

We define the nested super-population by a model, and first and second moments. The random variables $Y_{ij}, i = 1, \ldots, N; j = 1, \ldots, M_i$ in the super-population are assumed to satisfy $E(Y_{ij}) = \mu$ and

$$
\text{cov}(Y_{ij}, Y_{kl}) = \delta^2 + \sigma_i^2 \text{ when } i = k; j = l \\
= \delta^2 \text{ when } i = k; j \neq l \\
= 0 \text{ otherwise.}
$$

Scott and Smith (1969) used this super-population model to derive predictors of linear combinations of elements in the super-population, and in particular, the linear combination that corresponds to the latent value of a PSU. Scott and Smith frame the statistical problem as prediction of the remaining unobserved SSUs in a PSU. A Bayesian derivation (assuming the super-population 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 given by

$$
f_i \bar{Y}_i + (1 - f_i) \left[ \hat{\mu}^* + k_i \left( \bar{Y}_i - \hat{\mu}^* \right) \right] \quad (3)
$$

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

where $\hat{\mu}^* = \sum_{i=1}^{n} w_i \bar{Y}_i$, $f_i = \frac{m_i}{M_i}$, $w_i = \frac{1}{\sum_{i=1}^{n} 1/v_i}$, $v_i = \delta^2 + \frac{\sigma_i^2}{m_i}$, and $k_i = \frac{\delta^2 \sigma_i}{v_i}$. This predictor has an appealing interpretation as the weighted sum of two terms: the sample mean for the realized PSU, and predictors of the remaining SSUs for the realized PSU. The weighting factors are the proportion of SSUs that are observed and the proportion that 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 equation (2) and equation (3). When $f_i$ is small enough that the first term in equation (3) can be ignored, and $1 - f_i \geq 1$, the two equations appear to be identical, with the exception of different representations for the variance components. The best linear unbiased predictor (in equation (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, Dorfman, and Royall (2000).

As discussed by Scott and Smith, the assumption of normality is not needed. Instead, the model assumptions correspond to assuming PSUs are exchangeable, and SSUs are exchangeable within PSUs. The predictor is assumed to be a linear function of the sample responses. Coefficients of the predictor are derived so to minimizes the expected MSE over the super-population. The MSE is required to be bounded, leading to a constraint in the minimization.
2. A TWO-STAGE SAMPLING MODEL FOR A FINITE UNEQUAL SIZE CLUSTERED POPULATION

Mixed model and super-population model predictors are both model based. This means that the problem definition begins with an assumption about a probability model, including parameters for the model. Underlying the model is a concept of a population, but there is no explicit definition the population, or of parameters for population units. An advantage of such approaches is that the methods appear to apply to problems in many settings, since the approach does not account for differences in the setting details. The model based approaches to inference can be contrasted with a design based approach, where the problem definition is explicit, and not tied to assumptions about a stochastic model. For example, consider the problem of determining the average percent Kcal intake from saturated fat per practice among patients with elevated LDL cholesterol levels in HMO physician practices. The problem could conceivably be answered by enumeration, without sampling or statistics. Alternatively, an optimal strategy (sampling plus estimation) for estimating such levels can be selected. Notice that the strategy is clearly separated from the problem definition. Although not all aspects of the problem definition need to be accounted for in the strategy, better estimator may result from strategies that account for more details. We use this design based approach to develop predictors of realized random effects in an unequally size clustered population.

We derive predictors of realized random effects from a two-stage sample of a finite unequal size clustered population using a random permutation model similar to {Stanek, 2004 #1697}. A principal distinction is the use of a probability model based directly on the two-stage sampling and thus is based on the sample design. The results, while similar in form to those of Scott and Smith 1969 and Valliant, Dorfman et al. 2000, do not require positing a super-population. We first define the population and target parameters. Next we define the random variables that represent a two-stage random permutation of the population, and variations in the representation that preserve representation of the cluster’s identity.

2.1 Finite Population Parameters and Re-parameterizations

We begin by defining finite population parameters and representing response for a subject in a simple deterministic model. The mean and variance of cluster $s$ are defined as

$$
\mu_s = \frac{1}{M_s} \sum_{i=1}^{M_s} y_{si}
$$

and

$$
\sigma^2_s = \frac{1}{M_s} \sum_{i=1}^{M_s} (y_{si} - \mu_s)^2
$$

for $s = 1, \ldots, N$ (where we use the survey sampling definition of the parameter $\sigma^2_s$). Similarly, the mean of clusters in the population, and the variance between cluster means are defined as

$$
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s
$$

and

$$
\sigma^2 = \frac{1}{N} \sum_{s=1}^{N} (\mu_s - \mu)^2.
$$

Using these parameters, we represent the potentially observable response for subject $t$ in cluster $s$ as

$$
y_{st} = \mu + \beta_s + \epsilon_{st}
$$

where $\beta_s = (\mu_s - \mu)$ is the deviation of the mean for cluster $s$ from the overall mean, and $\epsilon_{st} = (y_{st} - \mu_s)$ is the deviation of subject $t$'s response (in cluster $s$) from the mean for cluster $s$. Model (4) is called a derived model (Hinkelmann and Kempthorne 1994).
Defining \( y = \begin{pmatrix} y'_1 & y'_2 & \cdots & y'_N \end{pmatrix}' \) where \( y'_s = \begin{pmatrix} y_{s1} & y_{s2} & \cdots & y_{sM_s} \end{pmatrix}' \), model (4) can be summarized as
\[
y = X\mu + Z\beta + \epsilon \quad (5)
\]
where \( X = 1_{nN} \), \( Z_{s} = \bigoplus_{s=1}^{N} 1_{M_s} \), \( N = \sum_{s=1}^{N} M_s \), and \( \beta' = (\beta_1, \beta_2, \cdots, \beta_N) \). Here, \( 1_{a} \) is an \( a \times 1 \) column vector of ones, \( \bigoplus A_s \) denotes a block diagonal matrix with blocks given by \( A_s \) (Graybill 1983), and \( \epsilon \) 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 define a two stage random permutation model that parallels model (5) and can be used to represent two-stage cluster sampling. We assume that the sample clusters are the clusters in the first \( n \) positions in a permutation of clusters, \( y'_s \). 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. We define random variables whose realization corresponds to a two-stage permutation of the population. Assuming each realization of the permutation is equally likely, the random variables formally represent two-stage sampling (Cochran 1977).

For each permutation, we assign a new label, \( i = 1, \ldots, N \), to the clusters according to their position in the permuted list. In a similar manner, we assign a new label, \( j = 1, \ldots, M_s \), to the units according to their position in the permuted list for cluster \( s \). We refer to the cluster that will occupy position \( i \) in the permutation of clusters as primary sampling unit (PSU) \( i \), and to the unit that will occupy position \( j \) in a permutation of units in a PSU as secondary sampling unit (SSU) \( j \). In the finite population, we represent the response for unit \( t \) in cluster \( s \) by \( y''_{st} \); in the permuted population, we represent the random variable for response of the unit in position \( j \) for the cluster in position \( i \) by \( Y''_{ij} \). 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.

Sample indicator random variables relate \( y''_{st} \) to \( Y''_{ij} \). The indicator random variable \( U''_{is} \) takes on a value of one when PSU \( i \) is cluster \( s \), and a value of zero otherwise; the indicator random variable \( U''_{jt}^{(s)} \) 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 PSU \( i \) and SSU \( j \) in a permutation is given by
\[
Y''_{ij} = \sum_{s=1}^{N} U''_{is} \sum_{t=1}^{M_s} U''_{jt}^{(s)} y''_{st} \quad (6).
\]
Although the random variable \( Y''_{ij} \) is clearly defined, the notation is not adequate to define a vector of random variables representing a permutation of units for a randomly selected cluster. The problem is that the dimension of the vector of units for a cluster is not equal when clusters differ in size. Since the cluster in position \( i \) is random, the dimension of the vector representing a permutation of cluster units in position \( i \) will also be random.
We can avoid vectors of random dimension by expanding the random variable representation of the two stage random permutation. We expand $Y_{ij}$ to a vector of random variables, where elements in the vector are given by $Y_{ij} = U_{is} \sum_{t=1}^{M_s} U_{jt}^{(i)} y_{st}$ for $s = 1, \ldots, N$. We organize these elements by $j = 1, \ldots, M_s$ in each $s = 1, \ldots, N$. First, we represent a permutation of units in cluster $s$ by the $M_s \times 1$ random vector $Y^*_s = \left( (Y_{ij}) \right) = U^{(i)} y_s$, where $Y_{ij} = \sum_{t=1}^{M_s} U_{jt}^{(i)} y_{st}$ and $U^{(i)}$ is a matrix of indicator random variables,

$$U^{(i)} = \left( U^{(i)}_1 \quad U^{(i)}_2 \quad \cdots \quad U^{(i)}_{M_s} \right)$$

Next, we expand random variables represented by $Y_{ij}$ (for all possible $j$) to the $N \times 1$ vector $Y_i = \left( U_{i1} Y^{*}_1 \quad U_{i2} Y^{*}_2 \quad \cdots \quad U_{in} Y^{*}_N \right)$. Finally, we represent a two stage random permutation of the population by the $NN \times 1$ vector, $Y = \left( Y'_1 \quad Y'_2 \quad \cdots \quad Y'_N \right)$.

A simple example may help illustrate the notation and organization. Suppose the population consists of three clusters ($N = 3$), with the first two having two units ($M_1 = M_2 = 2$), and the third having three units ($M_3 = 3$), such that $N = 7$. We represent permutations of units within clusters by $Y^*_s$, $s = 1, \ldots, 3$. Suppose the first permutation of clusters results in clusters $s = 1$, $s = 2$, and $s = 3$ in positions $i = 1, \ldots, 3$ respectively, while a second permutation results in clusters $s = 3$, $s = 2$, and $s = 1$ in positions $i = 1, \ldots, 3$ respectively. A simple representation of the random variables resulting from the first permutation is the random vector $\left( Y'_1 \quad Y'_2 \quad Y'_3 \right)$, and from the second permutation is the random vector $\left( Y'_3 \quad Y'_2 \quad Y'_1 \right)$. 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 the SSUs 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 first permutation is represented by

$$\left( Y'_1 \quad 0'_1 \quad 0'_1 \quad 0'_2 \quad Y'_2 \quad 0'_2 \quad 0'_2 \quad Y'_3 \right),$$

while the second permutation is represented by

$$\left( 0'_2 \quad 0'_2 \quad Y'_3 \quad 0'_2 \quad Y'_2 \quad 0'_2 \quad Y'_1 \quad 0'_2 \quad 0'_2 \right).$$

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 denote expectation with respect to permutations of the clusters and the subscript $\xi_2$ to denote expectation with respect to permutations of units in a cluster. Equating $Y_i$ to its expected value, $E_{\xi_1 \xi_2} (Y_i) = \frac{1}{N} 1_N \mu + \frac{1}{N} Z \beta$, plus a residual,
\( Y_i - E_{\xi_2} (Y_i) \), is a fixed effect model for the random variables representing responses for the SSUs in PSU \( i \). For all PSUs, the fixed effect model is given by

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

where \( \text{var}_{\xi_2} (Y) = \frac{1}{N-1} P_N \otimes \left[ \sum_{i=1}^{N} 1_{M_i} \mu_i \right] P_N \left( \sum_{i=1}^{N} 1_{M_i} \mu_i \right) + \left[ \sum_{i=1}^{N} \frac{1}{N} P_{M_i} \sigma^2_x \right] \) (see Appendix A).

We can divide the residuals into two terms to express this model as a mixed model. Defining \( E_i = Y_i - E_{\xi_2} (Y_i) \), we represent the residuals as

\[
Y_i - E_{\xi_2} (Y_i) = \left[ E_{\xi_2} (Y_i) - E_{\xi_2} (Y_i) \right] + E_i
\]

where the random effects for PSU \( i \) are defined by

\[
E_{\xi_2} (Y_i) = \left( \sum_{i=1}^{N} 1_{M_i} \right) M_i + \left( \sum_{i=1}^{N} 1_{M_i} \right) B_i
\]

where \( M_i = \left( M_{i1} \right) \), \( B_i = \left( B_{i1} \right) \), \( M_{is} = \left( U_{is} - E_{\xi_2} (U_{is}) \right) \mu \) and \( B_{is} = \left( U_{is} - E_{\xi_2} \left( U_{is} \right) \right) \beta_s \).

We summarize these random effects in the vectors \( M = \left( M'_1 \quad M'_2 \quad \cdots \quad M'_{N} \right)' \) and \( B = \left( B'_1 \quad B'_2 \quad \cdots \quad B'_{N} \right)' \). As a result, the mixed model is given by

\[
Y = \left[ \frac{1}{N} 1_{NN} \mu + \frac{1}{N} (1_N \otimes Z) \beta \right] + \left( \sum_{i=1}^{N} \left( \sum_{i=1}^{N} 1_{M_i} \right) \mu_i \right) (M + B) + E
\]

where \( E = \left( E'_1 \quad E'_2 \quad \cdots \quad E'_{N} \right)' \). The variance of the random effects is given by

\[
\text{var}_{\xi_2} \left[ \left( \sum_{i=1}^{N} \sum_{j=1}^{N} 1_{M_i} \right) (M + B) \right] = \frac{1}{N-1} P_N \otimes \left[ \sum_{i=1}^{N} 1_{M_i} \mu_i \right] P_N \left( \sum_{i=1}^{N} 1_{M_i} \mu_i \right) , \quad \text{while}
\]

\[
\text{var}_{\xi_2} (E) = \frac{1}{N} \left( \sum_{i=1}^{N} P_{M_i} \sigma^2_x \right).
\]

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 \( g' = b' \otimes \left( 1'_{M_1} \quad 1'_{M_2} \quad \cdots \quad 1'_{M_N} \right) \) for totals, and

\[
g' = b' \otimes \left( \frac{1'_{M_1}}{M_1} \quad \frac{1'_{M_2}}{M_2} \quad \cdots \quad \frac{1'_{M_N}}{M_N} \right) \quad \text{for means, where} \quad b = \left( b_1 \right) \quad \text{is an} \quad N \times 1 \quad \text{vector of constants.}
\]

Of principle interest are linear combinations, \( P_i \), defined by setting \( b = e_i \) where \( e_i \) as 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 these 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 simple random sampling probability model. We assume the realized sample results in 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 of the mean squared error (MSE). The basic strategy is given in many places (Scott and Smith 1969; Royall 1976; Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). The elements of $Y$ are partitioned into a sampled and remaining portion. We assume that the elements in the sampled 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 sampled random variables and to be unbiased, coefficients are evaluated that minimize the expected value of the mean squared error, $\text{var}_{\hat{y}_1} (\hat{P}_i - P_i)$. This general development was given in Theorem 2.1 by (Royall 1976).

Conceptually, Royall’s theorem can be applied directly to form predictors, but practically is complicated by singularities in $\text{var}(Y)$. 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. A key step in the developments is the definition of the ‘collapsed’ set of random variables. We consider three different definitions. The first definition, $Y^* = C''Y$, collapses random variables to PSU totals for the sample and remainder. The second definition, $Y^* = C''y$, collapses random variables to scaled PSU totals for the sample and
remainder where the scaling factor is the cluster size. A third strategy, \( Y^* = C^* Y \), is to collapse random variables to PSU means for the sample and remaining units. We proceed to outline details of each of these developments.

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

We develop the best linear unbiased predictor of the total for PSU \( i \) based on the vector of PSU totals for the sample and remainder given by \( Y^* = \left( I_N \otimes C^* \right) R \) where

\[
C^* = \left( I_N' \otimes I_2 \right) \left[ \bigoplus_{s=1}^{N} \left( I_m' \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}_{(M_s - m_s)} \right) \right].
\]

For PSU \( i \), the corresponding elements of \( Y^* \) are given by \( Y_i^* = \left( Y_{il}^* \right)' \), where \( Y_{il}^* = \sum_{s=1}^{N} U_{is} y_{is} \).

The mixed model for PSU \( i \) reduces to

\[
Y_i^* = X^* \alpha + Z_i D + E_i^* \quad (7)
\]

where \( X^* = \begin{pmatrix} \bar{m} & 0 & 1 \\ \bar{M} - \bar{m} & 1 & -1 \end{pmatrix} \), \( Z_i^* = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \otimes e_i' \), \( \alpha = \begin{pmatrix} \mu \\ \bar{\delta} \\ \bar{\delta}_f \end{pmatrix}' \), \( \delta_s = M_s \beta_s \),

\[
\bar{\delta} = \frac{1}{N} \sum_{s=1}^{N} \delta_s, \quad f_s = \frac{m_s}{M_s}, \quad \delta_{fs} = f_s \delta_s, \quad \bar{\delta}_f = \frac{1}{N} \sum_{s=1}^{N} \delta_{fs}, \quad D = \begin{pmatrix} m_i' & M_i' & D_{ji}' & D_{j}' \end{pmatrix}',
\]

\[
m_i = \sum_{s=1}^{N} U_{is} m_s, \quad M_i = \sum_{s=1}^{N} U_{is} M_s, \quad \bar{m} = \frac{1}{N} \sum_{s=1}^{N} m_s, \quad \bar{M} = \frac{1}{N} \sum_{s=1}^{N} M_s, \quad D_i = \sum_{s=1}^{N} U_{is} \delta_s,
\]
\[ D_{fi} = \sum_{s=1}^{N} U_s \delta_{fi} \quad \text{and} \quad \begin{pmatrix} E_{ii}^\ast \\
\n \end{pmatrix} = \begin{pmatrix} \sum_{s=1}^{N} U_{is} Y_{is} - \sum_{s=1}^{N} U_{is} m_s \mu_s \\
\n \sum_{s=1}^{N} U_{is} Y_{is} - \sum_{s=1}^{N} U_{is} (M_s - m_s) \mu_s \\
\n \end{pmatrix}. \] 

[See c04ed18.doc and c04ed20.doc] The first term corresponds to fixed effects, with \( \mu \) corresponding to the average of the cluster means, \( \delta \) the weighted (by cluster size) average deviation of the cluster means about \( \mu \), and \( \delta_j \) the weighted (by cluster sample size) average deviation of the cluster means about \( \mu \). The fixed effects in the model can be expressed equivalently in terms of average of the cluster totals, \( \bar{\tau} = \frac{1}{N} \sum_{s=1}^{N} \tau_s \), where \( \tau_s = M_s \mu_s \) such that

\[
\begin{pmatrix} \bar{f} & -\bar{f} \\
1 - \bar{f} & \bar{f} \\
\end{pmatrix} \alpha^* \text{ where } \bar{f} = \frac{1}{N} \sum_{s=1}^{N} m_s \text{ and } \alpha^* = \left( \bar{\tau} \quad \delta \quad \delta_j \right)^\prime. \] 

[See c04ed14.doc p29 and c04ed20.doc].

Model (7) is a mixed model where \( E_{\xi_i} \left( Y_i^\ast \right) = X_i^\ast \alpha \). There are two types of random effects. The first type of random effects corresponds to weighed products of the deviation in PSU sample size (or deviation in the number of non-sampled SSUs) from the average cluster sample size (or average number of non-sampled SSU), multiplied by the population mean. These terms are random since the number of sample SSUs for PSU \( i \) is a random variable, and depends upon which cluster is assigned to the PSU. The second type of random effect corresponds to a deviation in the PSU sample total from the cluster average sample total for PSU \( i \), and a similar deviation for the PSU non-sample totals.

The variance simplifies to
We develop a predictor of the total for PSU $i$ next, given by $P_i^* = g_i Y^*$ where $g_i = e_i \otimes I_2$. We require the predictor to be a linear function of the sampled random variables, to be unbiased, and to minimize the expected value of the mean squared error. First, we re-arrange terms into a sampled and remaining vector,

$$\begin{pmatrix} Y_i^* & Y_{ii}^* \end{pmatrix} = (K_i' \quad K_{II}') Y^*$$

where $K_I = \left( I_n \mid 0 \mid _{n(N-n)} \right) \otimes (1 \mid 0)$ and $K_{II} = \left[ \begin{array}{c} \left( I_n \mid 0 \mid _{n(N-n)} \right) \otimes (0 \mid 1) \\ 0 \end{array} \right]$, $(N-n) \times 2N$.

In terms of the partitioned random variables, the total for PSU $i$ is given by

$$P_i^* = g_{ii} Y_i^* + g_{II} Y_{II}^*$$

where $g_{ii} = e_{ii}'$ and $g_{II} = (e_{II}' \quad e_{III}' \quad e_{IIV}')$, and $e_i' = (e_{II}' \mid e_{IIV}')$ where $e_{ii}$ is an $n \times 1$ vector, and $e_{II}$ has dimension $(N-n) \times 1$. Collection of the study data will result in realizing the values of $Y_i^*$. We require the predictor of $P_i^*$ to be an unbiased linear function of the sample data, $\hat{P}_i^* = (g_i' + a') Y_i^*$, such that $E_{\xi_{ii}^{*}} \left( \hat{P}_i^* - P_i^* \right) = 0$. The unbiased constraint implies that $(a' X_i^* - g_{II} X_{II}^*) \alpha = 0$, a condition that will be satisfied when $1/N \sum_{i=1}^{N} \left[ m_i c - M_1 \right] \mu_s = 0$, where $c = (a' + e_{II}') I_n$. This then implies that the predictor will be unbiased only if $f_i^{-1} = f^{-1}$ is constant. Thus, an unbiased predictor of a PSU total exists only when second stage sampling fractions are equal for all clusters, ie. that SSUs are selected with probability proportional to size (pps).

We assume pps second stage sampling to develop an unbiased predictor based on minimizing the expected value of the mean squared error. With this assumption, the model
for the sample SSUs in the \( n \) sample PSUs is given by \( Y^* = X_i^* \bar{\tau} + Z^*_i T_i + E_i^* \) while the model for the remainder is given by \( Y^* = X_i^* \bar{\tau} + Z^*_i T_i + E_i^* \)

where \( X_i^* = f I_n, \ Z_i^* = f I_n, \ X_i^* = (1 - f) \left( \begin{array}{c} I_n \\ 0 \end{array} \right), \ Z_i^* = (1 - f) \left( \begin{array}{c} I_n \\ 0 \end{array} \right), \)

\[
f = \left( \frac{f}{1 - f} \right)^{1'}, \ T = (T'_1 \ T'_2 \ \cdots \ T'_n)' \]

\( T_i = \sum_{s=1}^N U_{is} \tau_s - \bar{\tau} \). The variance reduces to

\[
\text{var}_i \left( Y^* \right) = (1 - f)^2 \left( \sigma^2_e \right)^2 \otimes \left[ \left( f(1 - f)^{-2} \right)^2 0 \right] - \sigma^2_e \mathbf{P}_N \otimes \mathbf{f}'
\]

where \( \sigma^2_e = \frac{\sum_{s=1}^N (\delta_s - \bar{\delta})^2}{N - 1} \) and \( \sigma^2_e = \frac{\sum_{s=1}^N \mathbf{M}_s \sigma^2_s}{N} \) (see c04ed20.doc), which we partition such that \( \text{var}_i \left( Y^* \right) = \left\{ \begin{array}{c} V_i^* \\ V_{I,II}^* \\ V_{II}^* \end{array} \right\} \). The best linear unbiased predictor of the \( i^{th} \) selected cluster total is given by \( \hat{P}^* = g_i^* Y_i^* + \mathbf{g}_{II}^* \left[ X_i^* \hat{\tau} + V_{I,II}^* \left( Y_i^* - X_i^* \hat{\tau} \right) \right] \) where \( \hat{\tau} = \frac{1}{f n} Y_i^* \),

\[
V_i^* = \text{var}_i \left( Y_i^* \right) = \mathbf{f}' \mathbf{I}_n - \frac{f^2 \sigma^2_e}{N} \mathbf{J}_n \), \( v^* = (1 - f) \sigma^2_e + \mathbf{f}' \sigma^2_e \), and

\[
V_{I,II}^* = \text{var}_{i,i'} \left( Y_i^*, Y_{i'}^* \right) = f (1 - f) \left( \sigma^2_e - \sigma^2_e \right) \left( \mathbf{I}_n \bigg| \frac{0}{m=2(N-n)} \right) - \frac{\sigma^2_e}{N} \left( \mathbf{J}_n \bigg| \frac{\mathbf{f}' \otimes \mathbf{J}_{m=2(N-n)} }{n} \right)
\]

(see c02ed12.doc and c02ed42.doc for details, and c04ed19.doc (page 15+16 – note different order for \( J \)) and c04ed20.doc). In addition,

\[
V_{II}^* = \text{var}_{i,i'} \left( Y_{II}^* \right) = f (1 - f) \left( \sigma^2_e \right)^2 \left( \mathbf{I}_n \bigg| \frac{0}{m=2(N-n)} \right) - \frac{\sigma^2_e}{N} \left( \mathbf{f}' \otimes \mathbf{J}_{m=2(N-n)} \right)
\]

(see c04ed20.doc). Using these expressions,

\[
\hat{P}^* = \mathbf{e}' \mathbf{Y}^* + \left( \frac{1 - f}{f} \right) \mathbf{e}_{II} \left( \frac{\mathbf{J}_n + k^* \mathbf{P}_n}{n} \right) \mathbf{Y}^*_i + \frac{1}{f} \mathbf{e}'_{II} \left( \frac{\mathbf{J}_{(N-n)xn}}{n} \right) \mathbf{Y}^*_i
\]

where \( k^* = \frac{f \left( \sigma^2_e - \sigma^2_e \right)}{f \left( \sigma^2_e - \sigma^2_e \right) + \sigma^2_e} \).

Defining \( \bar{Y}_{ij} = \sum_{s=1}^m Y_{ij} \), \( \bar{Y}_{i} = (\bar{Y}_1 \ \bar{Y}_2 \ \cdots \ \bar{Y}_n)' \), and \( \bar{Y} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i \), when in addition all cluster sizes are equal, the predictor simplifies to
\[ \hat{P}_i = m e_{iI} \bar{Y}_i + (M - m) e_i \left( 1_n \bar{Y} + k P_n \bar{Y}_i \right) + M e_{iI} \cdot 1_{N-n} \bar{Y} \]

where \( k = \frac{m \sigma^2}{m \sigma^2 + \sigma^2_e} \), and \( \sigma^2 = \sigma^2 - \frac{\sigma^2_e}{M} \). (see c02ed39.doc and c02ed42.doc, and c04ed20.doc). This result matches the predictor developed by Stanek and Singer (2004).

The variance when \( i \leq n \) simplifies to

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

and when \( i > n \) reduces to

\[
\text{var}_{n_i} \left( \hat{P}_i - P^* \right) = \left[ 2f (1 - f) + \frac{1}{n} \left( \frac{1-f}{f} \right) \sigma^2_e \right]\sigma^2_s + \left[ 1 - 2f (1 - f) + \frac{1}{n} + \frac{1}{N} \left[ 1 - \left( \frac{1}{1-f} \right)^2 \right] \right] \sigma^2_s.
\]

(see c04ed23.doc -which differs from the results in c02ed49.doc).

3.2. Predictors of PSU Means based on Weighted PSU Totals

Predictors of the PSU mean can be developed from a collapsed set of random variables corresponding to a weighted PSU totals by substituting \( C_s^{**} = \frac{C_s^*}{M_s} \) in place of \( C_s^* \)

to define \( C^* R = Y^* \) where \( C^* = (C_1^* \quad C_2^* \quad \cdots \quad C_N^*) \). The elements in \( Y^* \) corresponding to PSU \( i \) are given by

\[
Y_i^* = \left( \sum_{s=1}^{N} U_{is} f_s \bar{Y}_{sl} \right) \left( \sum_{s=1}^{N} U_{is} (1-f_s) \bar{Y}_{sll} \right)
\]

where \( \bar{Y}_{sl} = \frac{1}{m_s} \sum_{j=1}^{m_s} Y_{sj} \) and

\[
\bar{Y}_{sll} = \frac{1}{M_s - m_s} \sum_{j=m_s+1}^{M_s} Y_{sj}, \text{ and result in a mixed model given by}
\]

\[
Y_i^* = X^* \alpha^* + Z_i^* D^* + E_i^*
\]
where \( \mathbf{X}^* = \left( \frac{\bar{f}^*}{(1 - \bar{f}^*)} \right)^{-1} \), \( \mathbf{Z}_i^* = \left( \begin{array}{ccc} 1 & 1 & 0 \\ -1 & -1 & 1 \end{array} \right) \) \( \otimes \mathbf{e}' \), \( \mathbf{\alpha}^* = (\mu \ \bar{\delta}_i)' \), \( \bar{f}^* = \frac{1}{N} \sum_{s=1}^{N} f_s \),

\[
f_s = \frac{m_s}{M_s}, \quad \bar{\delta}_i = \frac{1}{N} \sum_{s=1}^{N} f_s \beta_s, \quad \mathbf{D}' = \left( \mathbf{f}'_i \ \mathbf{D}'_{ii} \ \mathbf{B}' \right)', \quad \mathbf{f}_i = \left( \left( (f_i - \bar{f}^*) \mu \right) \right), \quad \mathbf{D}'_{ii} = \left( \left( D_{ii}^* - \bar{\delta}_i \right) \right),
\]

\[
\mathbf{B} = \left( (B_i) \right), \quad f_i = \sum_{s=1}^{N} U_{is} f_s, \quad D_{fi} = \sum_{s=1}^{N} U_{is} \delta_{fi} = \sum_{s=1}^{N} U_{is} f_s \beta_s, \quad B_i = \sum_{s=1}^{N} U_{is} \beta_s \text{ and }
\]

\[
\mathbf{E}_i = \left( \sum_{s=1}^{N} U_{is} (Y_{si}^* - f_s \mu_s) \right) \left( \sum_{s=1}^{N} U_{is} \left[ Y_{si}^* - (1 - f_s) \mu_s \right] \right)'.
\]

We develop a predictor of the mean for PSU \( i \) given by \( P_i^* = \mathbf{g}' \mathbf{Y}^* \) where \( \mathbf{g} = \mathbf{e} \otimes \mathbf{1}_2 \) and \( \mathbf{Y}^* = \left( \mathbf{Y}_i^* \ \mathbf{Y}_2^* \ \cdots \ \mathbf{Y}_N^* \right)' \). We require the predictor to be a linear function of the sampled random variables, to be unbiased, and to minimize the expected value of the mean squared error. Proceeding in a similar manner as in Section 3.1, we first re-arrange \( \mathbf{Y}^* \) into a sampled and remaining vector, and express the target random variable in terms of the partitioned vector, \( P_i^* = \mathbf{g}_i' \mathbf{Y}_i^* + \mathbf{g}_R' \mathbf{Y}_R^* \) (with terms defined similar to those in Section 3.1). [see page 6, c04ed24.doc] Collection of the study data will result in realizing the values of \( \mathbf{Y}_i^* \). We require the predictor of \( P_i^* \) to be an unbiased linear function of the sample data, \( \hat{P}_i^* = (\mathbf{g}_i' + \mathbf{a}' \mathbf{1}) \mathbf{Y}_i^* \), such that \( E_{i=1}^N \left( \hat{P}_i^* - P_i^* \right) = 0 \). For sample PSUs, the constraint implies that \( \left( \frac{1}{N} \sum_{s=1}^{N} \left[ \mathbf{a}' \mathbf{1}_n f_s - (1 - f_s) \right] \right) \mu + \frac{1}{N} \sum_{s=1}^{N} f_s \beta_s = 0 \), which will be true if \( \mathbf{a}' \mathbf{1}_n = \left( \frac{1 - f_i}{f_s} \right) \), and \( f_s = f \) for all \( s = 1, \ldots, N \). Thus, unbiased prediction of a sample PSU mean requires pps sampling and \( \mathbf{a}' \mathbf{1}_n = (1 - f)/f \). When \( i > n \), the unbiased constraint requires only that \( \mathbf{a}' \mathbf{1}_n = \frac{1}{f} \), and pps sampling is not required. (see c04ed24.doc for more details)

We focus on predictors of PSU means under pps sampling. With this assumption, \( \bar{f}^* = f \), and the mixed model for PSU means for the sample simplifies to

\[
\mathbf{Y}_i^* = \mathbf{X}_i^* \mu + \mathbf{Z}_i^* \mathbf{B}_i + \mathbf{E}_i^* \quad \text{while the model for the remainder is given by}
\]

\[
\mathbf{Y}_R^* = \mathbf{X}_R^* \mu + \mathbf{Z}_R^* \mathbf{B} + \mathbf{E}_R^* \quad \text{(see c04ed25.doc)}
\]

Expressions for the variance simplify to

\[
\mathbf{V}_i^* = f^2 \mathbf{I}_n - \frac{f^2 \sigma^2}{N} \mathbf{J}_n \quad \text{where} \quad v^* = \left( 1 - f \right) \sigma^2 + f \sigma^2,
\]
\[ V_{I,II} = f(1-f) \left[ (\sigma^2 - \sigma_e^* ) \left( I_n + \begin{bmatrix} 0 \\ \sigma^2 \\ \begin{bmatrix} f' \otimes \begin{bmatrix} J_n + \sigma^2 N \end{bmatrix} \end{bmatrix} \end{bmatrix} \right) \right] , \]

and

\[ V_{II} = f(1-f)\sigma^* I_{n+2(N-n)} + \sigma^2 \begin{bmatrix} (1-f)^2 I_n \\ 0 \\ f^2 I_{N-n} \\ 0 \\ 0 \\ (1-f)^2 I_{N-n} \end{bmatrix} - \frac{\sigma^2}{N} \begin{bmatrix} (1-f) I_n \\ f \otimes I_{N-n} \\ f \otimes I_{N-n} \end{bmatrix} \]

. (see c04ed25.doc) The best linear unbiased predictor of a PSU mean is given by

\[ \hat{P}_i = \mathbf{e}_{i} \mathbf{Y}_i + (1-f)\mathbf{e}_{i}' \left( \mathbf{1}_n \bar{Y} + k^* \mathbf{P}_n \bar{Y}_i \right) + \mathbf{e}_{ili} \mathbf{1}_{N-n} \bar{Y} \]

where \( k^* = \frac{f(\sigma^2 - \sigma_e^* )}{f(\sigma^2 - \sigma_e^* ) + \sigma_e^2} \). When in addition all cluster sizes are equal, the predictor simplifies to

\[ \hat{P}_i = \mathbf{fe}_i \mathbf{Y}_i + (1-f)\mathbf{e}_i' \left( \mathbf{1}_n \bar{Y} + k^* \mathbf{P}_n \bar{Y}_i \right) + \mathbf{e}_{ili} \mathbf{1}_{N-n} \bar{Y} , \]

the predictor presented by Stanek and Singer (2004). (see c04ed25.doc)

The variance when \( i \leq n \) simplifies to

\[ \text{var}_{n\geq i} \left( \hat{P}_i - \mathbf{P}_i \right) = (1-f)^2 \left[ f(1-f)k^2 + (1-fk^*) \right] \left( 1 - \frac{1}{n} \right) \sigma^2 + (1-f) \left( 1 - \frac{1}{n} + \frac{1}{nf} \right) \sigma_e^2 \]

and when \( i > n \) reduces to \( \text{var}_{n\geq i} \left( \hat{P}_i - \mathbf{P}_i \right) = \sigma^2 + \frac{1}{n} \left( \sigma^2 + \frac{1-f}{f} \sigma_e^2 \right) \). (see c02ed49.doc)

TO HERE
Appendix A. Evaluation of $\text{var}_{\xi \eta} (Y)$

We outline the steps leading to $\text{var}_{\xi \eta} (Y)$ using the conditional expansion

$$\text{var}_{\xi \eta} (Y) = \text{var}_{\xi} \left( E_{\eta | \xi} [Y] \right) + E_{\eta} \left( \text{var}_{\xi | \eta} [Y] \right).$$

The vector $Y = (Y_1' \ Y_2' \ \cdots \ Y_N')'$ is constructed from sub-vectors for each PSU given by $Y_i = \left( U_{i1} Y_1' \ U_{i2} Y_2' \ \cdots \ U_{in} Y_N' \right)'$

where $Y_s = \left( (Y_s) \right)' = U^{(s)} Y_s$, $Y_{sj} = \sum_{t=1}^{M_s} U_{jt}^{(s)} Y_{st}$ and $U^{(s)}$ is an $M_s \times M_s$ matrix of indicator random variables, $U^{(s)} = \left( U_1^{(s)} \ U_2^{(s)} \ \cdots \ U_{M_s}^{(s)} \right)$ with $U_{ij}^{(s)} = \left( U_{ij1}^{(s)} \ U_{ij2}^{(s)} \ \cdots \ U_{ijM_s}^{(s)} \right)'$. We begin by developing the expected value and variance of these sub-vectors.

First, since $E_{\xi | \eta} \left( U^{(s)} \right) = \frac{1}{M_s} J_{M_s}$, $E_{\xi | \eta} \left( Y_i \right) = \left( \bigoplus_{s=1}^{N} I_{M_s} \mu_s \right) U_i$ where

$$U_i = \left( U_{i1} \ U_{i2} \ \cdots \ U_{iN} \right)'.$$ Defining $U = \left( U_1 \ U_2 \ \cdots \ U_N \right)$ and combining terms,

$$E_{\xi | \eta} \left( Y \right) = \left[ I_N \bigotimes \left( \bigoplus_{s=1}^{N} I_{M_s} \mu_s \right) \right] \text{vec}(U).$$

We develop an expression for

$$\text{var}_{\xi} \left( \text{vec}(U) \right) = E_{\xi} \left[ \begin{pmatrix} U_1 U_1' & U_1 U_2' & \cdots & U_1 U_N' \\ U_2 U_1' & U_2 U_2' & \cdots & U_2 U_N' \\ \vdots & \vdots & \ddots & \vdots \\ U_N U_1' & U_N U_2' & \cdots & U_N U_N' \end{pmatrix} \right] - \frac{1}{N^2} J_N \bigotimes J_N$$

by noting that under simple random without replacement sampling, $E_{\xi} \left( U_i U_i' \right) = \frac{1}{N} I_N$, while when $i \neq i^*$,

$$E_{\xi} \left( U_i U_{i^*}' \right) = \frac{1}{N(N-1)} \left( J_N - I_N \right).$$

Substituting these terms and simplifying,

$$\text{var}_{\xi} \left( \text{vec}(U) \right) = \frac{1}{N-1} P_N \bigotimes P_N,$$

where $P_N = I_N - \frac{1}{N} J_N \bigotimes J_N$, and hence

$$\text{var}_{\xi} \left( E_{\xi | \eta} \left[ Y \right] \right) = \frac{1}{N-1} P_N \bigotimes \left[ \bigoplus_{s=1}^{N} I_{M_s} \mu_s \right] P_N \left( \bigoplus_{s=1}^{N} I_{M_s} \mu_s \right).$$
We evaluate $E_{\hat{\xi}_1}(\text{var}_{\hat{\xi}_1}[Y])$ by expanding terms in

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

Since permutations of SSUs are independent for different PSUs, $\text{var}_{\hat{\xi}_1}(Y_i, Y'_i) = \bigoplus_{s=1}^N \text{var}_{\hat{\xi}_1}(Y'_i)$. Now

$$\text{var}_{\hat{\xi}_1}(Y_i, Y'_i) = \begin{pmatrix} U_{i1}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_i) & U_{i1}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_1, Y'_2) & \cdots & U_{i1}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_1, Y'_N) \\ U_{i2}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_2, Y'_1) & U_{i2}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_2) & \cdots & U_{i2}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_2, Y'_N) \\ \vdots & \vdots & \ddots & \vdots \\ U_{iN}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_N, Y'_1) & U_{iN}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_N, Y'_2) & \cdots & U_{iN}U_{1^s} \text{var}_{\hat{\xi}_1}(Y'_N) \end{pmatrix}.$$ 

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

$$\text{var}_{\hat{\xi}_1}(\text{vec}(U^{(s)})) = \frac{1}{M_s - 1} P_{M_s} \otimes P_{M_s} \text{ where } P_{M_s} = I_{M_s} - \frac{J_{M_s}}{M_s}.$$ 

Hence, $\text{var}_{\hat{\xi}_1}(Y'_i) = P_{M_s} \sigma^2_{y_s}$. Using these expressions,

$$\text{var}_{\hat{\xi}_1}(Y) = \begin{pmatrix} \bigoplus_{s=1}^N U_{1^s}P_{M_s} \sigma^2_{y_s} & \bigoplus_{s=1}^N U_{1^s}U_{2^s}P_{M_s} \sigma^2_{y_s} & \cdots & \bigoplus_{s=1}^N U_{1^s}U_{N^s}P_{M_s} \sigma^2_{y_s} \\ \bigoplus_{s=1}^N U_{2^s}P_{M_s} \sigma^2_{y_s} & \bigoplus_{s=1}^N U_{2^s}U_{2^s}P_{M_s} \sigma^2_{y_s} & \cdots & \bigoplus_{s=1}^N U_{2^s}U_{N^s}P_{M_s} \sigma^2_{y_s} \\ \vdots & \vdots & \ddots & \vdots \\ \bigoplus_{s=1}^N U_{N^s}P_{M_s} \sigma^2_{y_s} & \bigoplus_{s=1}^N U_{N^s}U_{2^s}P_{M_s} \sigma^2_{y_s} & \cdots & \bigoplus_{s=1}^N U_{N^s}U_{N^s}P_{M_s} \sigma^2_{y_s} \end{pmatrix}.$$ 

Finally, evaluating the expected value over permutations of PSUs,

$$E_{\hat{\xi}_1}(\text{var}_{\hat{\xi}_1}[Y]) = I_N \otimes \left( \bigoplus_{s=1}^N \frac{1}{N} P_{M_s} \sigma^2_{y_s} \right).$$ 

We use these expressions to express $\text{var}_{\hat{\xi}_2}(Y) = \text{var}_{\hat{\xi}_1}(E_{\hat{\xi}_1}[Y]) + E_{\hat{\xi}_1}(\text{var}_{\hat{\xi}_1}[Y])$ as

$$\text{var}_{\hat{\xi}_2}(Y) = \frac{1}{N-1} P_N \otimes \left[ \bigoplus_{s=1}^N I_{M_s} \otimes P_{M_s} \left( \bigoplus_{s=1}^N I'_M \otimes \mu_s \right) \right] + I_N \otimes \left( \bigoplus_{s=1}^N \frac{1}{N} P_{M_s} \sigma^2_{y_s} \right).$$
Appendix B.

We summarize the model for the sample SSUs in the \( n \) sample PSUs as a mixed model for the vector \( \mathbf{Y}_i^m \) given by

\[
\mathbf{Y}_i^m = \left[ \mathbf{1}_n \otimes (\bar{m} \ 0 \ 1) \right] \alpha + \left[ (1 \ 0 \ 1 \ 0) \otimes \left( \mathbf{I}_n \mid \mathbf{0}_{n(n-N)} \right) \right] \mathbf{D} + \mathbf{E}_i,
\]

and

\[
\mathbf{Y}_{ii}^m = \left( \begin{array}{c} \bar{M} - \bar{m} \\ \bar{m} \\ \bar{M} - \bar{m} \end{array} \right) \otimes \mathbf{1}_{N-n} \alpha + \left( \begin{array}{c} -1 \ 1 \ -1 \\ 1 \ 0 \ 1 \\ -1 \ 1 \ -1 \end{array} \right) \otimes \left( \begin{array}{c} \mathbf{I}_n \mid \mathbf{0}_{n(N-n)} \\ \mathbf{0}_{N-n,n} \mid \mathbf{I}_{N-n} \end{array} \right) \mathbf{D} + \mathbf{E}_{ii}^m,
\]

(see page 8 in c04ed19.doc).

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 predictor. Avoiding these complications with other methods comes with a cost- more extensive assumptions and ambiguity in the interpretation of predictors.
References
The Worcester-area trial for counseling in hyperlipidemia (WATCH).