Pre-Sampling Model Based Inference IV

Stephen Woodruff

Specified Designs, 800 West View Terrace, Alexandria, VA 22301

Abstract

This paper suggests that sampling theory may usefully be expanded from random sampling of population units to both random sampling of population units and random construction of these population units. It continues previous work under this title, Woodruff (2011, 2010, 2009). In survey sampling, a sample unit`s study variables are expanded to population totals by probability design based or model based expansions that implicitly treat a unit`s study variables as totals over entities called atoms contained in each unit. When a unit`s atoms are random samples from a population of atoms, a model is imposed on sample data that leads to Pre-sampling Model Based (PSMB) Inference. PSMB inference provides estimates that retain the best properties of both model based and design based inference and that eliminate the main shortcomings of each. The result can be orders-of-magnitude error reduction. Techniques for deriving sampling error under repeated sampling from stratified cluster designs of PSMB estimators are derived. Applications to some common sampling problems are presented.

Key Words: Model Based Inference, Design Based Inference, Pre-sampling Inference

1. Introduction

Sampling inference should be based on randomization (both for access to the mathematics of probability and for impartiality), it should be multivariate since most surveys collect data on several study variables, it should be robust against problems encountered in sampling applications that require post sampling adjustments (frame changes, non-response, response error, outlier adjustments etc). Model conjecture based on sample data or historical data is another potential source of error and it should be minimized or eliminated. These are the goals of Pre-Sampling Model Based Inference (PSMB), a methodology that was developed piecemeal in earlier papers, Woodruff (2007, 2008, 2009, 2010,2011).

This paper provides an expanded introduction to some fundamentals that underlie Pre-Sampling Model Based (PSMB) inference. It also describes two applications in stratified cluster sampling.

The foundation of PSMB inference is the randomized construction or synthesis of population units. This random synthesis is called Pre-sampling. The Pre-sampling model imposed by this random synthesis provides the hypothesis for theorems on Best Linear Unbiased Estimation. This random process gives Pre-sampling models a degree of credibility similar to probabilities of selection in Design Based inference where the probabilities of selection are also determined by random process, thus Pre-

sampling also substantially eliminates concerns about model failure by avoiding data based model conjecture.

PSMB combines the best features of both Design Based inference and Model Based inference while avoiding the main shortcomings of each. Estimation error is defined and analyzed with respect to repeated sampling of population units under stratified cluster sampling designs. Pre-sampling is stochastically independent of sample selection and occurs naturally in many study populations. Repeated sampling variance expressions (for unit sampling) for several PSMB estimators are derived within the expanded context of both randomized unit selection and randomized unit synthesis. When sample design control is difficult and the resulting sample designs are inefficient for application of Design Based inference, PSMB inference is an alternative that substantially reduces sampling error compared to design based alternatives. PSMB inference also provides insights to the interaction between the sample design and the stochastic properties of the study variables being measured by the sample survey. *In all that follows, "sampling" refers to the sampling of units and "Pre-sampling" refers to randomized unit construction or synthesis which must proceed sampling of units (hence Pre-sampling).*

In Woodruff (2011) PSMB estimators were compared to common design based estimators like the Combined Ratio Estimator, Horwitz-Thompson estimator, and combinations of both. These comparisons explained the dramatic reduction in repeated sampling error of PSMB estimates compared to Design Based estimators observed in simulation studies documented in Woodruff (2007, 2008, 2009, 2010,2011).

In mail surveys the sampling frame of population units consists of mail containers. The USPS samples these containers to estimate total weight, postage, pieces etc. by mail flow stratum (e.g. all mail of a given class coming from France to New York by air in February). The atoms are the mail pieces within each container and the unit (container) study variables are the number of pieces it contains and its totals over these pieces of weight and postage. These sample unit totals are expanded to the population to provide estimates of population totals for the study variables. Woodruff (2010, 2009), derived PSMB Inference from this atom structure which assumes the randomized assignment of atoms to units, the random process from which the population model is deduced. This atom structure also helps identify sample design problems that can be avoided or at least reduced by looking at sample design through the lens PSMB inference provides. It also highlights properties of study variables that may magnify the sampling error of Design Based estimators.

This paper and the papers referenced above suggest that sampling theory may be usefully expanded from randomized sample selection to both randomized sample selection and randomized unit synthesis.

2. APM and PSMB Models and Consequent Estimators

2.1 Sample Design and Atom/Unit Structure

Stratified cluster designs were used to sample population units and compare different estimators (including the Combined Ratio Estimator) in earlier papers on this subject. In this paper, all estimators are sums of stochastically independent stratum estimators

(the Combined Ratio Estimator is not considered here) so it suffices to consider only a single stratum and hence stratification can be omitted from the sample design (all estimates and their variances are sums of independent stratum estimates and their variances).

2.1.1 Unit Sample Design

The population is a set of units denoted, U, of size N units partitioned into K clusters (each unit is in one and only one cluster). Let U_i be the set of units in cluster i so that $U = \bigcup_{i=1}^{K} U_i$ and let N_i denote the number of units in U_i.

The first stage of selection consists of an SRSWOR (simple random sample without replacement) of size k from the K clusters and this sample of clusters is denoted S. From each member, U_i, of S an SRSWOR of n_i units is selected from the N_i units in U_i and is denoted, S_i. The sample is the union of the sample units in the sampled clusters, $\bigcup_{i\epsilon S} S_i$.

2.1.2 Atom Population Model (APM)

Each unit in the population consists of a random sample (without replacement) of entities called atoms from a large population of atoms. **This randomized unit synthesis from a population of atoms is called Pre-sampling.** A column vector of study variables (same variables for all atoms) is attached to each atom contained within a unit. The vector of study variables attached to a unit is the sum of its atom study variable vectors. The number of atoms comprising a unit may be large but they are of relatively few types. Each atom type is described by a type specific Atom Population Model (APM) that is made explicit in the next paragraph.

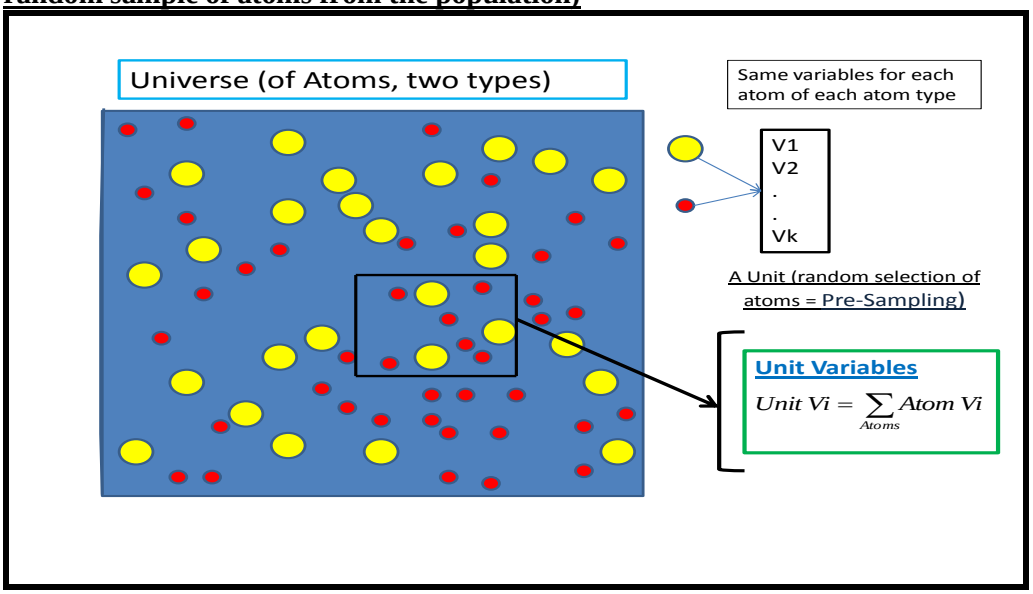
Population (and sample) units consist of V types of atoms (see Figure 1 below, an example where V=2 with red and yellow atom types). For $1 \le v \le V$, let a_{ijv} be the number of atoms of type v in the j^{th} unit of the i^{th} cluster, for $1 \le j \le N_i$ (refer to this unit as "unit (i,j)"). The column vector of study variables attached to the l^{th} atom for $1 \le l \le a_{ijv}$ of type v in unit (i,j) is denoted Y_{ijvl} and if Y_{ij} is the vector of study variables attached to unit (i,j), then $Y_{ij} = \sum_{v=1}^{V} \sum_{l=1}^{a_{ijv}} Y_{ijvl}$. The quantity to be estimated is Y= $\sum_{i=1}^{K} \sum_{j=1}^{N_i} Y_{ij}$, the population total for the unit study variables. Because the number of atoms in the population is very large, Pre-sampling that defines the study variables for one unit is very nearly stochastically independent of Pre-sampling of atoms defining any other unit and their independence is a reasonable approximating assumption: $Y_{ij} \perp Y_{\tilde{i}\tilde{j}}$ for $(i,j) \neq (\tilde{\imath},\tilde{\jmath})$.

Let $a_{ij} = \sum_{v=1}^{V} a_{ijv}$, the total number of atoms of all types in unit (i,j). In Figure 1 below the contents of the black box represent a unit (i,j), type 1 atoms are the red circles, and type 2 atoms are the yellow circles, then $a_{ij1} = 6$ and $a_{ij2} = 3$.

In case of stratification, a single Atom Population Model (APM) for several atoms types applies to all atoms in the stratum. The APM in one stratum may differ from the APM in other strata. This may influence the structure/definition of the strata.

Y_{ijvl} is the column vector of auxiliary and target variables attached to the l^{th} atom of type v in sample unit (i,j). Auxiliary variables are study variables whose population totals are known. Target variables are study variables whose population totals are to be estimated.

Figure 1. <u>Population of Atoms and a Unit (the contents of the black box, a random sample of atoms from the population)</u>



The expectation of Y_{ijvl} with respect to the APM denoted, $\varepsilon_M(.)$, is $\varepsilon_M(Y_{ijvl}) = \mu_v$ for all atoms, l, of type v in unit (i,j) $(1 \le v \le V)$. The covariance matrix of the components of Y_{ijvl} with respect to the APM is denoted $C_v = \varepsilon_M(Y_{ijvl}Y_{ijvl}^{'}) - \mu_v\mu_v^{'}$ and is the same for all atoms of type v in the population (the prime denotes transpose). The APM is the set of these V atom type models.

In what follows, the notation $W \sim [B,H]$ means the expectation and covariance matrix of the vector valued random variable W, are B and H respectively. The APM expressed this way is the set of V models:

$Y_{ijvl} \sim (\mu_v, C_v)$ for all (i,j) and type v atoms where $1 \le v \le V$.

The unit (i,j) study variable vector is $Y_{ij} = \sum_{v=1}^{V}\sum_{l=1}^{a_{ijv}} Y_{ijvl}$ and for all ordered quadruples such that $(i,j,v,l) \neq (\tilde{\imath},\tilde{\jmath},\tilde{v},\tilde{l})$, $Y_{i,j,v,l} \perp Y_{\tilde{\imath},\tilde{\jmath},\tilde{v},\tilde{l}}$. This is a consequence of the assumption that the population size in atoms is large enough to be approximated as infinite and the sample of atoms in a unit is a random selection from this population of atoms. These APMs for the V atom types, assume only the existence of the first two moments for each type. Each population unit is comprised of a random selection from all atoms without regard to type. Different units will probably have differing mixes of atom types. This helps capture the stochastic structure of widely diverse population units with relatively small values of V (different atom types).

The structure described above may be easier to understand if you think of a sample unit as a container of water drawn from a stream and its atoms as particulate or

bacteria of which there are V distinct types, each type with the same study variables which for different types may be distributed differently. Given that these atoms enter the stream some distance up-flow, they will be well mixed and it is appropriate to think of the atoms within a sampled container as an SRSWOR from all the atoms in the stream. Generally, the atom content in the stream will vary over time so time may be a stratification variable. For example, a population stratum of atoms may be all atoms in the stream flowing past a point during a specific time interval.

2.1.3 Pre-sampling Unit Model

By definition of the APM, $\varepsilon_M(Y_{ij}) = \sum_{v=1}^{V} a_{ijv}\mu_v$ and the APM covariance matrix of Y_{ij} is $\Sigma_{ij} = \sum_{v=1}^{V} a_{ijv}\, C_v$ (the atom vectors of study variables are independent of one another so the variance (covariance) of the sum is the sum of the variances (covariances).

Now let the vector of study variables for unit (i,j) and atom l of type v be partitioned into two sub-vectors, the first is an A-vector (A components) of the A auxiliary variables (variables for which the population total is known) and the second is a T-vector of the T target variables so that $Y_{ijvl} = \begin{pmatrix} Y^a_{ijvl} \\ Y^t_{ijvl} \end{pmatrix}$ where Y^a_{ijvl} is the A-vector (column) of auxiliary variables and Y^t_{ijvl} is the T-vector (column) of target variables. Then

$Y_{ij} = \sum_{v=1}^{V}\sum_{l=1}^{a_{ijv}} Y_{ijvl} = \sum_{v=1}^{V}\sum_{l=1}^{a_{ijv}} \begin{pmatrix} Y^a_{ijvl} \\ Y^t_{ijvl} \end{pmatrix} = \begin{pmatrix} \sum_{v=1}^{V}\sum_{l=1}^{a_{ijv}} Y^a_{ijvl} \\ \sum_{v=1}^{V}\sum_{l=1}^{a_{ijv}} Y^t_{ijvl} \end{pmatrix} \underset{=}{defn} \begin{pmatrix} Y^a_{ij} \\ Y^t_{ij} \end{pmatrix}$ and

$\varepsilon_M(Y_{ij}) = \sum_{v=1}^{V} a_{ijv} \begin{pmatrix} \mu^a_v \\ \mu^t_v \end{pmatrix}$ where $\mu^a_v = \varepsilon_M(Y^a_{ijvl})\ \forall$ (i,j), l, and v. Similarly $\mu^t_v = \varepsilon_M(Y^t_{ijvl})\ \forall$ (i,j), l, and v. Letting $C_v = \begin{pmatrix} C_{va} & C_{vat} \\ C_{vta} & C_{vt} \end{pmatrix}$, where C_{va} is the $A \times A$ covariance matrix of each Y^a_{ijvl} for all (i,j), l, and v in the population. C_{vat} is the $A \times T$ matrix of covariances between the components of Y^a_{ijvl} and Y^t_{ijvl} and so on for the other two covariance matrices in C_v. A prime denotes transpose, $C_{vat} = C'_{vta}$. Σ_{ij} is the covariance matrix of Y_{ij} under the APM so:

$\Sigma_{ij} = \sum_{v=1}^{V} a_{ijv} \begin{pmatrix} C_{va} & C_{vat} \\ C_{vta} & C_{vt} \end{pmatrix} \underset{=}{defn} \begin{pmatrix} \Sigma_{ija} & \Sigma_{ijat} \\ \Sigma_{ijta} & \Sigma_{ijt} \end{pmatrix}$.

Let $M_a = (\mu^a_1, \cdots, \mu^a_V)$, the matrix whose columns are the $\{\mu^a_v\}$ and similarly $M_t = (\mu^t_1, \cdots, \mu^t_V)$ then

$\varepsilon_M(Y_{ij}) = \begin{pmatrix} M_a \\ M_t \end{pmatrix} \begin{pmatrix} a_{ij1} \\ \vdots \\ a_{ijV} \end{pmatrix}$ and $Y_{ij} = \begin{pmatrix} Y^a_{ij} \\ Y^t_{ij} \end{pmatrix} = \begin{pmatrix} M_a \\ M_t \end{pmatrix} \begin{pmatrix} a_{ij1} \\ \vdots \\ a_{ijV} \end{pmatrix} + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim (0, \Sigma_{ij})$, M_a is the $A \times V$ matrix of atom type means by auxiliary variable, its $(q,v)^{th}$ element is the mean of the v^{th} atom type for the q^{th} auxiliary variable. M_t is the $T \times V$ matrix of atom type means by target variable, its $(q,v)^{th}$ element is the mean of the v^{th} atom type for the q^{th} target variable. Let $D_{ij} = \begin{pmatrix} a_{ij1} \\ \vdots \\ a_{ijV} \end{pmatrix}$, ε^a_{ij} be the first A components of ε_{ij}, and ε^t_{ij} be the last T components of ε_{ij} then:

$Y_{ij} = \begin{pmatrix} M_a \\ M_t \end{pmatrix} D_{ij} + \varepsilon_{ij}$ where $\varepsilon_{ij} = \begin{pmatrix} \varepsilon_{ij}^a \\ \varepsilon_{ij}^t \end{pmatrix} \sim (0, \quad \Sigma_{ij})$ for $i \epsilon S$ and $j \epsilon S_i$. (2.2.1)

When V=A and M_a is non-singular then the model given by (2.2.1) above can be transformed into one in which the target variables are matrix-proportional to the auxiliary variables as follows.

By definition, $Y_{ij}^a = M_a D_{ij} + \varepsilon_{ij}^a$ and can be rewritten as:

$D_{ij} = M_a^{-1}\left(Y_{ij}^a - \varepsilon_{ij}^a\right) = M_a^{-1} Y_{ij}^a - M_a^{-1}\varepsilon_{ij}^a$ or

$D_{ij} = M_a^{-1} Y_{ij}^a + \tilde{\varepsilon}_{ij}^a$ where $\tilde{\varepsilon}_{ij}^a = - M_a^{-1}\varepsilon_{ij}^a$,

$\tilde{\varepsilon}_{ij}^a \sim \left(0, \quad M_a^{-1}\Sigma_{ija}(M_a^{-1})'\right)$ (2.2.2)

Substituting $D_{ij} = M_a^{-1} Y_{ij}^a + \tilde{\varepsilon}_{ij}^a$ into $Y_{ij}^t = M_t D_{ij} + \varepsilon_{ij}^t$,

$Y_{ij}^t = M_t\left(M_a^{-1} Y_{ij}^a + \quad \tilde{\varepsilon}_{ij}^a\right) + \varepsilon_{ij}^t = M_t M_a^{-1} Y_{ij}^a + \left(M_t \tilde{\varepsilon}_{ij}^a + \quad \varepsilon_{ij}^t\right)$.

Let B= $M_t M_a^{-1}$, then

$Y_{ij}^t = B Y_{ij}^a + \delta_{ij}$ for $i \epsilon S$ and $j \epsilon S_i$. (2.2.5)

Where $\delta_{ij} = \left(M_t \tilde{\varepsilon}_{ij}^a + \quad \varepsilon_{ij}^t\right)$, $\varepsilon_M(\delta_{ij})$=0, and the covariance matrix of δ_{ij} is

$\Sigma_{ij\delta} = B\Sigma_{ija}B' - B\Sigma_{ijat} - \Sigma_{ijta}B' + \Sigma_{ijt}$ (2.2.6)

Let the transpose of the α^{th} row of B be $B_\alpha = \begin{pmatrix} b_{\alpha 1} \\ b_{\alpha 2} \\ \vdots \\ b_{\alpha V} \end{pmatrix} for\ \alpha = 1,2, \ldots\ldots, T.$ Then (2.2.5) can be written as:

$Y_{ij}^t = \begin{pmatrix} B_1' \\ \vdots \\ B_T' \end{pmatrix} Y_{ij}^a + \delta_{ij}$ for $i \epsilon S$ and $j \epsilon S_i$. (2.2.7)

$= \left(I \otimes Y_{ij}^{a\prime}\right) \begin{pmatrix} B_1 \\ \vdots \\ B_T \end{pmatrix} + \delta_{ij}$ (2.2.8)

for $i \epsilon S$ and $j \epsilon S_i$ where I is the $T \times T$ identity matrix and $\otimes$ denotes Kronecker product. The Kronecker product of two matrices is defined as the matrix result of multiplying each component of the first matrix by the second matrix.

(2.2.7) or (2.2.8) is the Pre-sampling unit model derived from randomized unit synthesis (Pre-sampling). This model is imposed by the random process of Pre-sampling and is not conjectured from sample or population data.

2.2 PSMB Best Linear Unbiased Estimator

The linear relationship summarizing all the sample data for $i \epsilon S$ and $j \epsilon S_i$ is:

$\begin{pmatrix} \begin{pmatrix} Y_{11}^t \\ \vdots \\ Y_{1n_1}^t \end{pmatrix} \\ \vdots \\ \begin{pmatrix} Y_{k1}^t \\ \vdots \\ Y_{kn_k}^t \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} I \otimes Y_{11}^{a\prime} \\ \vdots \\ I \otimes Y_{1n_1}^{a\prime} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} I \otimes Y_{k1}^{a\prime} \\ \vdots \\ I \otimes Y_{kn_k}^{a\prime} \end{pmatrix} \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{T-1} \\ B_T \end{pmatrix} + \Delta$ where I is the $T \times T$ identity matrix, k is

the number of clusters sampled and n_k is the sample size (number of units sampled) from sample cluster k, and Δ is the $m_S \times T$ random column vector ($where\ m_S = \sum_{k\epsilon S} n_k$) , $\Delta = \begin{pmatrix} \begin{pmatrix} \delta_{11} \\ \vdots \\ \delta_{1n_1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \delta_{k1} \\ \vdots \\ \delta_{kn_k} \end{pmatrix} \end{pmatrix}$ with expectation of 0 and its covariance matrix is the block diagonal matrix of the $\{\Sigma_{ij\delta}\}$ for $i\epsilon S$ and $j\epsilon S_i$, all off diagonal blocks are zero matrices. The BLUE for $\beta = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{T-1} \\ B_T \end{pmatrix}$ is:

$$\hat{\beta} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \vdots \\ \hat{B}_{T-1} \\ \hat{B}_T \end{pmatrix} = \left(\sum_{i\epsilon S}\sum_{j\epsilon S_i}\left(I\otimes Y_{ij}^a\right)\Sigma_{ij\delta}^{-1}\left(I\otimes Y_{ij}^{a\prime}\right)\right)^{-1}\sum_{i\epsilon S}\sum_{j\epsilon S_i}\left(I\otimes Y_{ij}^a\right)\Sigma_{ij\delta}^{-1}Y_{ij}^t ,$$

Rao (1973).

Substituting estimates for M_t , M_a, Σ_{ija} , Σ_{ijt} and Σ_{ijat} , into $\Sigma_{ij\delta}$, $\hat{\beta}$ can be approximated directly from the atom sample data. For example, the approximated mean for the ω^{th} study variable of type v atoms is the average of the ω^{th} study variable for all the type v atoms in the sample units. Likewise, the covariance matrices, $\{C_v\}_{v=1}^{V}$, are the estimates of these quantities given that the type v atoms in the sample are an SRSWOR from all type v atoms in the population (or stratum). *Unit sample design (2.1.1) plays no role in these model estimates and of course, atom data must be collected from the sample units in addition to the unit study variables.*

Then the BLUE for the vector of target variable population totals and its model covariance matrix are:

$\hat{T}_{TOT} = (I\otimes A)\hat{\beta}$ and $\quad \mathrm{Var}(\hat{T}_{TOT}) = (I\otimes A)\left(\sum_{i\epsilon S}\sum_{j\epsilon S_i}\left(I\otimes Y_{ij}^a\right)\Sigma_{ij\delta}^{-1}\left(I\otimes Y_{ij}^{a\prime}\right)\right)^{-1}(I\otimes A)'$ (2.2.9)

Where A = $\sum_{k=1}^{N} Y_k^{a\prime}$, is the known vector of population (or stratum) auxiliary variable totals, N is the number of population (or stratum) units, and I is the $T \times T$ identity matrix. Note that $VAR\left(\hat{T}_{TOT}\right)$ is **not** the repeated sampling variance which is the measure for evaluating sampling and estimation strategies in this paper. The repeated sampling variance is generally a relatively small component of $VAR\left(\hat{T}_{TOT}\right)$.

2.3 Applications

The following lemma provides the repeated sampling variance (and covariances) under the design described in Section 2.1.1. of un-weighted sample sums of study variables.

Lemma 2.3.1: Let a_{ij} *and* b_{ij} be two study variables attached to the j^{th} unit of the i^{th} cluster. Under repeated sampling using the cluster sampling design described in Section 2.1.1 the repeated sampling covariance between the sample sums, $\sum_{i\epsilon S}\sum_{j\epsilon S_i} a_{ij}$ and $\sum_{i\epsilon S}\sum_{j\epsilon S_i} b_{ij}$ is given by:

$$Cov\left(\sum_{i\epsilon S}\sum_{j\epsilon S_i} a_{ij}, \sum_{i\epsilon S}\sum_{j\epsilon S_i} b_{ij}\right) = k\left(1-\frac{k}{K}\right)S_{n_i\bar{a}_i, n_i\bar{b}_i} + \frac{k}{K}\sum_{i\epsilon U} n_i\left(1-\frac{n_i}{N_i}\right)S_{a_ib_i} \quad (2.3.9)$$

Where

$$S_{n_i\bar{a}_i, n_i\bar{b}_i} = \frac{1}{K-1}\sum_{i\epsilon U}\left(\frac{n_i}{N_i}\sum_{j\epsilon U_i} a_{ij} - \frac{1}{K}\sum_{l\epsilon U}\frac{n_l}{N_l}\sum_{j\epsilon U_l} a_{lj}\right)\left(\frac{n_i}{N_i}\sum_{j\epsilon U_i} b_{ij} - \frac{1}{K}\sum_{l\epsilon U}\frac{n_l}{N_l}\sum_{j\epsilon U_l} b_{lj}\right)$$

$$= \frac{1}{K-1}\sum_{i=1}^{K}\left(n_i\bar{a}_i - \frac{1}{K}\sum_{l=1}^{K} n_l\bar{a}_l\right)\left(n_i\bar{b}_i - \frac{1}{K}\sum_{l=1}^{K} n_l\bar{b}_l\right) \text{ where } \bar{a}_i = \frac{1}{N_i}\sum_{j\epsilon U_i} a_{ij} \text{ and } \bar{b}_i = \frac{1}{N_i}\sum_{j\epsilon U_i} b_{ij}.$$

$$S_{a_ib_i} = \frac{1}{N_i-1}\sum_{j\epsilon U_i}(a_{ij}-\bar{a}_i)(b_{ij}-\bar{b}_i).$$

When $a_{ij} = b_{ij} = c_{ij}$, the repeated sampling variance of, $\sum_{i\epsilon S}\sum_{j\epsilon S_i} c_{ij}$ is given by (2.3.9).

2.3.1 An Example, V=A=2 and T=1

This example is the case of two auxiliary variables, 2 atom types, and one target variable. When there are two auxiliary variables and multiple target variables the results below may be applied by using this solution for T=1 multiple times to obtain estimates (and their repeated sampling variances under design in section 2.1.1) for each of the targets variables separately. Some efficiency is lost by this approach which fails to capture the additional strength borrowed from related data on the other target variables within a unit.

Let $Y_{ij} = \begin{pmatrix} Y_{ij}^a \\ Y_{ij}^t \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} Y_{ij1}^a \\ Y_{ij2}^a \end{pmatrix} \\ Y_{ij1}^t \end{pmatrix}$.

In this case the covariance matrix given in (2.2.6), $\Sigma_{ij\delta}$, reduces to a scalar, $\Sigma_{ij\delta} = \sigma_{ij\delta}^2$.

Then, $\hat{\beta} = \left(\sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij}^a Y_{ij}^{a\prime}\right)^{-1}\sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij}^t Y_{ij}^a = \begin{pmatrix} g_5 & g_3 \\ g_3 & g_1 \end{pmatrix}^{-1}\begin{pmatrix} g_2 \\ g_4 \end{pmatrix}$

where $g_1 = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij2}^a Y_{ij2}^a = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 g_{1ij}$ where $g_{1ij} = Y_{ij2}^a Y_{ij2}^a$.

$g_2 = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij1}^a Y_{ij1}^t = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 g_{2ij}$ where $g_{2ij} = Y_{ij1}^a Y_{ij1}^t$

$g_3 = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij2}^a Y_{ij1}^a = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 g_{3ij}$ where $g_{3ij} = Y_{ij2}^a Y_{ij1}^a$

$g_4 = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij2}^a Y_{ij1}^t = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 g_{4ij}$ where $g_{4ij} = Y_{ij2}^a Y_{ij1}^t$

$g_5 = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 Y_{ij1}^a Y_{ij1}^a = \sum_{i\epsilon S}\sum_{j\epsilon S_i}\sigma_{ij\delta}^2 g_{5ij}$ where $g_{5ij} = Y_{ij1}^a Y_{ij1}^a$

Thus: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \frac{1}{g_5 g_1 - g_3^2}\begin{pmatrix} g_1 & -g_3 \\ -g_3 & g_5 \end{pmatrix}\begin{pmatrix} g_2 \\ g_4 \end{pmatrix} = \begin{pmatrix} \frac{g_1 g_2 - g_3 g_4}{g_1 g_5 - g_3^2} \\ \frac{g_4 g_5 - g_2 g_3}{g_1 g_5 - g_3^2} \end{pmatrix}$.

This is the PSMB BLUE for β.

The $\{\sigma_{ij\delta}^2 g_{uij}\}$ for $u = 1,2,3,4,5$, $i\epsilon S$ *and* $j\epsilon S_i$ are called "derived" study variables and are functions of the study variables whose data are collected from sample units

during data collection. They behave with respect to repeated sampling just like the study variables whose data is directly recorded during data collection. Sampling variances of nonlinear functions of the study variables can usually be well approximated as the sampling variances of sample sums of derived study variables. An example of this follows.

The PSMB BLUE for the population total of the target variable (first component of Y^t) is: $\hat{Y}_{TOT} = \hat{\beta}_1 Y^a_{TOT1} + \hat{\beta}_2 Y^a_{TOT2}$ where Y^a_{TOT1} is the population total for the first auxiliary variable and Y^a_{TOT2} is the population total for the second auxiliary variable.

$\hat{\beta}_1 = \frac{g_1 g_2 - g_3 g_4}{g_1 g_5 - g_3^2}$, a function of five random variables, $g = (g_1, g_2, g_3, g_4, g_5)$ can be approximated (Taylor Series) with a plane tangent to $\hat{\beta}_1$ at the expected values (with respect to repeated sampling) of $(g_1, g_2, g_3, g_4, g_5)$ which are denoted, $G = (G_1, G_2, G_3, G_4, G_5)$ where $E(g_u) = G_u$ for u=1,2,3,4,5. This approximation is given by:

$\hat{\beta}_1 \doteq \hat{\beta}_1\big|_{g=G} + \sum_{u=1}^{5} \frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} (g_u - G_u)$ and

$\hat{\beta}_2 \doteq \hat{\beta}_2\big|_{g=G} + \sum_{u=1}^{5} \frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G} (g_u - G_u)$, where $\frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G}$ is the partial derivative of $\hat{\beta}_1$ with respect to g_u evaluated at g=G and similarly for $\frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G}$.

Thus $\hat{Y}_{TOT}$ can be approximated with

$$\hat{Y}_{TOT} \doteq Y^a_{TOT1}\left[\hat{\beta}_1\big|_{g=G} + \sum_{u=1}^{5} \frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} (g_u - G_u)\right] + Y^a_{TOT2}\left[\hat{\beta}_2\big|_{g=G} + \sum_{u=1}^{5} \frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G} (g_u - G_u)\right]$$

$$= \left[Y^a_{TOT1}\hat{\beta}_1\big|_{g=G} + Y^a_{TOT2}\hat{\beta}_2\big|_{g=G}\right] + \sum_{u=1}^{5}\left(Y^a_{TOT1}\frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} + Y^a_{TOT2}\frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G}\right)(g_u - G_u)$$

Discarding constant terms, the repeated sampling variance of $\hat{Y}_{TOT}$ is thus approximated as:

$Var(\hat{Y}_{TOT}) \doteq Var\left(\sum_{u=1}^{5}\left(Y^a_{TOT1}\frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} + Y^a_{TOT2}\frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G}\right)g_u\right)$, a linear function of g.

Each g_u is the sample sum of derived study variables and thus letting $\omega_u = Y^a_{TOT1}\frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} + Y^a_{TOT2}\frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G}$,

$\sum_{u=1}^{5}\left(Y^a_{TOT1}\frac{\partial \hat{\beta}_1}{\partial g_u}\Big|_{g=G} + Y^a_{TOT2}\frac{\partial \hat{\beta}_2}{\partial g_u}\Big|_{g=G}\right)g_u = \sum_{u=1}^{5}\omega_u g_u$ and, by interchanging summations, is the sample sum of a derived study variable, a linear function of the $\{\sigma^2_{ij\delta} Y^a_{ij2} Y^a_{ij2}\}$ etc. The sampling variance of this sample sum, which is $Var(\hat{Y}_{TOT})$, is provided by Lemma 2.3.1.

$\sum_{u=1}^{5} \omega_u g_u = \sum_{u=1}^{5} \omega_u \sum_{i\epsilon S} \sum_{j\epsilon S_i} \sigma^2_{ij\delta} g_{uij} = \sum_{i\epsilon S} \sum_{j\epsilon S_i} \sum_{u=1}^{5} \omega_u \sigma^2_{ij\delta} g_{uij}$ and thus

$\Omega_{ij} = \sum_{u=1}^{5} \omega_u \sigma^2_{ij\delta} g_{uij}$ is a derived variable for unit (i,j) and the repeated sampling variance, $Var\left(\hat{Y}_{TOT}\right) = Var\left(\sum_{i\epsilon S} \sum_{j\epsilon S_i} \Omega_{ij}\right)$ and is given by Lemma 2.3.1.

The sampling variance of the PSMB BLUE for the population total is the sampling variance of the sample sum of the $\{\Omega_{ij}\}$. This sampling variance will tend to increase with cluster size variability, same as with more mundane unit study variables. This same methodology can be applied to finding the repeated sampling variance of more complex BLUEs, consisting of iterated products and inverses of matrices. This example is also easily adapted to the case A=1 and V=2 using generalized inverses, see Woodruff 2010.

2.3.2 An Application to Gross and Net Weight - two types of atoms: items within the container (several atoms) and the container (a single atom).

An estimator of total revenue (also called postage) for mail is derived using PSMB inference. The mail travels in containers of various types and its total gross weight (weight of mail pieces and containers) in the population is known. A sample of mail containers is selected and total weight and postage recorded for each container in the sample (the sample units are containers of mail). Each unit consists of two atom types, the first type ($v = 1$) is container (only a single atom per unit) and the second type of atom ($v = 2$) is mail piece (usually many per container). With this definition of atom, the preceding theory from Section 2.1.3 and 2.2 is readily applicable and is described in detail below.

There are two study variables, an auxiliary variable, weight, and a target variable, postage. As described in Section 2, Y_{ijvl} is the vector of study variables attached to the l^{th} atom of type v in unit (i,j). The APM is motivated by populations of containers (and their contents) where the weight of the container and its contents is the auxiliary variable. $Y^a_{\psi 11}$ is the container tare weight (weight of the empty container) and $Y^t_{ij11} = 0$, since the container has no postage. Y^a_{ij2l} is the weight of the l^{th} mail piece in unit (i,j) and Y^t_{ij2l} is the postage of the l^{th} mail piece in unit (i,j).

The vector of study variables attached to unit (i,j) is $Y_{ij} = \begin{pmatrix} Y^a_{ij} \\ Y^t_{ij} \end{pmatrix}$ where

$Y^a_{ij} = \sum_{v=1}^{2} \sum_{l=1}^{a_{ijv}} Y^a_{ijvl}$ is a scalar, the gross weight of unit (i,j). Since a_{ij1}=1, $Y^a_{ij} = Y^a_{ij11} + \sum_{l=1}^{a_{ij2}} Y^a_{ij2l}$, the container tare weight is Y^a_{ij11}, and the weight of its contents is $\sum_{l=1}^{a_{ij2}} Y^a_{ij2l}$.

The APM for the type 1 atoms (container weight & postage) in this population is, $Y^a_{ij11} \sim (\mu_{a1}, \sigma^2_{a1})$, $Y^t_{ij11} \sim (0,0)$. The APM for the type 2 atoms (mail piece weight & postage) is:

$\begin{pmatrix} Y^a_{ij2l} \\ Y^t_{ij2l} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{a2} \\ \mu_{t2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{a2} & \sigma_{at2} \\ \sigma_{ta2} & \sigma^2_{t2} \end{pmatrix} \right)$, for all $l \; and \; (i,j)$. The $\left\{ \begin{pmatrix} Y^a_{ij2l} \\ Y^t_{ij2l} \end{pmatrix} \right\}$ are independent for all $l \; and \; (i,j)$. Summing across atoms (item types and items) within each unit the unit study variable is:

$Y_{ij} = \begin{pmatrix} Y_{ij}^a \\ Y_{ij}^t \end{pmatrix} = \begin{pmatrix} Y_{ij11}^a \\ 0 \end{pmatrix} + \sum_{l=1}^{a_{ij2}} \begin{pmatrix} Y_{ij2l}^a \\ Y_{ij2l}^t \end{pmatrix}$ and from the APM above:

$$Y_{ij} = \begin{pmatrix} Y_{ij}^a \\ Y_{ij}^t \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{a1} + a_{ij2}\mu_{a2} \\ a_{ij2}\mu_{t2} \end{pmatrix}, \begin{pmatrix} \sigma_{a1}^2 + a_{ij2}\sigma_{a2}^2 & a_{ij2}\sigma_{at2} \\ a_{ij2}\sigma_{ta2} & a_{ij2}\sigma_{t2}^2 \end{pmatrix} \right)$$

This can be rewritten as:

$$Y_{ij} = \begin{pmatrix} \mu_{a1} & \mu_{a2} \\ 0 & \mu_{t2} \end{pmatrix} \begin{pmatrix} 1 \\ a_{ij2} \end{pmatrix} + \begin{pmatrix} \epsilon_{ij}^a \\ \epsilon_{ij}^t \end{pmatrix} \text{ where } \begin{pmatrix} \epsilon_{ij}^a \\ \epsilon_{ij}^t \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{a1}^2 + a_{ij2}\sigma_{a2}^2 & a_{ij2}\sigma_{at2} \\ a_{ij2}\sigma_{ta2} & a_{ij2}\sigma_{t2}^2 \end{pmatrix} \right)$$

$$= \begin{pmatrix} M_a \\ M_t \end{pmatrix} \begin{pmatrix} 1 \\ a_{ij2} \end{pmatrix} + \begin{pmatrix} \epsilon_{ij}^a \\ \epsilon_{ij}^t \end{pmatrix}, \text{ where } M_a = (\mu_{a1} \quad \mu_{a2}) \text{ and } M_t = (0 \quad \mu_{t2}).$$

This implies: $Y_{ij}^a = (\mu_{a1} \quad \mu_{a2}) \begin{pmatrix} 1 \\ a_{ij2} \end{pmatrix} + \epsilon_{ij}^a$

or $Y_{ij}^a - \epsilon_{ij}^a = \mu_{a1} + \mu_{a2} a_{ij2}$ and solving for a_{ij2}, $a_{ij2} = \frac{Y_{ij}^a - \epsilon_{ij}^a - \mu_{a1}}{\mu_{a2}}$.

Substituting $a_{ij2} = \frac{Y_{ij}^a - \epsilon_{ij}^a - \mu_{a1}}{\mu_{a2}}$ into, $Y_{ij}^t = M_t \begin{pmatrix} 1 \\ a_{ij2} \end{pmatrix} + \epsilon_{ij}^t$,

$Y_{ij}^t = \frac{\mu_{t2}}{\mu_{a2}} Y_{ij}^a - \frac{\mu_{t2}\mu_{a1}}{\mu_{a2}} + \left(\epsilon_{ij}^t - \frac{\mu_{t2}}{\mu_{a2}} \epsilon_{ij}^a \right)$. Let $\beta = \frac{\mu_{t2}}{\mu_{a2}}$, $\alpha = -\frac{\mu_{t2}\mu_{a1}}{\mu_{a2}}$, and

$\gamma_{ij} = \epsilon_{ij}^t - \frac{\mu_{t2}}{\mu_{a2}} \epsilon_{ij}^a$, then

$\gamma_{ij} \sim \left(0, \beta^2 \sigma_{a1}^2 + a_{ij2}(\sigma_{t2}^2 + \beta^2 \sigma_{a2}^2 - 2\beta\sigma_{at2}) \right)$

and $Y_{ij}^t = \beta Y_{ij}^a + \alpha + \gamma_{ij}$ for all (i, j). (2.3.2)

Let $\sigma_{ij}^2 = \beta^2 \sigma_{a1}^2 + a_{ij2}(\sigma_{t2}^2 + \beta^2 \sigma_{a2}^2 - 2\beta\sigma_{at2})$.

This implies that all $\{Y_{ij}^t\}$ in the sample stacked in a column vector are described by the linear equation:

$$\begin{pmatrix} Y_{11}^t \\ Y_{12}^t \\ \vdots \\ Y_{k(n_k-1)}^t \\ Y_{kn_k}^t \end{pmatrix} = \begin{pmatrix} Y_{11}^a & 1 \\ Y_{12}^a & 1 \\ \vdots & \vdots \\ Y_{kn_k}^a & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \begin{pmatrix} \gamma_{11} \\ \gamma_{12} \\ \vdots \\ \gamma_{k(n_k-1)} \\ \gamma_{kn_k} \end{pmatrix} \text{ where}$$

$$\begin{pmatrix} \gamma_{11} \\ \gamma_{12} \\ \vdots \\ \gamma_{k(n_k-1)} \\ \gamma_{kn_k} \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_{12}^2 & 0 & \ldots & \vdots \\ \vdots & 0 & . & 0 & \vdots \\ \vdots & \vdots & \vdots & . & 0 \\ 0 & 0 & \ldots & 0 & \sigma_{kn_k}^2 \end{pmatrix} \right]$$

Under (2.3.2) then the Best Linear Unbiased Estimator (BLUE) for β and α is:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = \frac{1}{f_1 f_2 - f_3^2} \begin{pmatrix} f_4 f_2 - f_3 f_5 \\ f_1 f_5 - f_3 f_4 \end{pmatrix} \text{ where } f_1 = \sum_{i \epsilon S} \sum_{j \epsilon S_i} \frac{\left(Y_{ij}^a\right)^2}{\sigma_{ij}^2}, \; f_2 = \sum_{i \epsilon S} \sum_{j \epsilon S_i} \frac{1}{\sigma_{ij}^2},$$

$$f_3 = \sum_{i \epsilon S} \sum_{j \epsilon S_i} \frac{Y_{ij}^a}{\sigma_{ij}^2}, \; f_4 = \sum_{i \epsilon S} \sum_{j \epsilon S_i} \frac{Y_{ij}^t Y_{ij}^a}{\sigma_{ij}^2}, \text{ and } \; f_5 = \sum_{i \epsilon S} \sum_{j \epsilon S_i} \frac{Y_{ij}^t}{\sigma_{ij}^2}.$$

Let $f_{1ij} = \frac{\left(Y_{ij}^a\right)^2}{\sigma_{ij}^2}$, $f_{2ij} = \frac{1}{\sigma_{ij}^2}$, $f_{3ij} = \frac{Y_{ij}^a}{\sigma_{ij}^2}$, $f_{4ij} = \frac{Y_{ij}^t Y_{ij}^a}{\sigma_{ij}^2}$, and $f_{5ij} = \frac{Y_{ij}^t}{\sigma_{ij}^2}$.

Then $f_1 = \sum_{i\epsilon S}\sum_{j\epsilon S_i} f_{1ij}$, $f_2 = \sum_{i\epsilon S}\sum_{j\epsilon S_i} f_{2ij}$,

$f_3 = \sum_{i\epsilon S}\sum_{j\epsilon S_i} f_{3ij}$, $f_4 = \sum_{i\epsilon S}\sum_{j\epsilon S_i} f_{4ij}$, and $f_5 = \sum_{i\epsilon S}\sum_{j\epsilon S_i} f_{5ij}$.

Let f be a 5×1 column vector such that its transpose, $f' = (f_1, \ \ldots \ , f_5)$. Let $F_i = E(f_i)$ for i=1,2,3,...,5 and where $E(.)$ is the expectation with respect to repeated sampling under the cluster design (2.1.1) and F such that $F' = (F_1, \ \ldots \ , F_5)$.

Then approximating $\hat{\beta}$ as a function of $(f_1, \ \ldots \ , f_5)$ with a hyperplane tangent to $\hat{\beta}$ at F, then the variance of $\hat{\beta}$ can be approximated as:

$Var(\hat{\beta}) \doteq Var\left(\sum_{u=1}^{5} f_u \frac{\partial \hat{\beta}}{\partial f_u}_{f=F}\right)$ and $Var(\hat{\alpha}) \doteq Var\left(\sum_{u=1}^{5} f_u \frac{\partial \hat{\alpha}}{\partial f_u}_{f=F}\right)$ where $\frac{\partial \hat{\beta}}{\partial f_u}_{f=F}$ is the partial derivative of $\hat{\beta}$ with respect to f_u evaluated at $f = F$ and similarly for $\frac{\partial \hat{\alpha}}{\partial f_u}_{f=F}$.

Let $A_u = \frac{\partial \hat{\beta}}{\partial f_u}_{f=F}$ and $B_u = \frac{\partial \hat{\alpha}}{\partial f_u}_{f=F}$ for u=1,2,....,5. The following table gives the values of the $\{A_u\}_{u=1}^{5}$ and the $\{B_u\}_{u=1}^{5}$.

	$\frac{\partial \hat{\beta}}{\partial f_u}_{f=F} = A_u$	$\frac{\partial \hat{\alpha}}{\partial f_u}_{f=F} = B_u$
u=1	$\frac{F_3F_5 - F_2F_4}{(F_1F_2 - F_3^2)^2}F_2$	$\frac{F_2F_3F_4 - F_5F_3^2}{(F_1F_2 - F_3^2)^2}$
u=2	$\frac{F_1F_5 - F_3F_4}{(F_1F_2 - F_3^2)^2}F_3$	$\frac{F_3F_4 - F_1F_5}{(F_1F_2 - F_3^2)^2}F_1$
u=3	$\frac{2F_3F_4F_2 - F_5F_1F_2 - F_5F_3^2}{(F_1F_2 - F_3^2)^2}$	$\frac{2F_1F_5F_3 - F_3^2F_4 - F_1F_2F_4}{(F_1F_2 - F_3^2)^2}$
u=4	$\frac{F_2}{F_1F_2 - F_3^2}$	$\frac{-F_3}{F_1F_2 - F_3^2}$
u=5	$\frac{-F_3}{F_1F_2 - F_3^2}$	$\frac{F_1}{F_1F_2 - F_3^2}$

Let A be the 5×1 column vector such that its transpose, A'=$(A_1, \ \ldots \ , A_5)$ and similarly B such that B'=$(B_1, \ \ldots \ , B_5)$.

Let $H_{ij} = \left(\frac{\left(Y_{ij}^a\right)^2}{\sigma_{ij}^2}, \frac{1}{\sigma_{ij}^2}, \frac{Y_{ij}^a}{\sigma_{ij}^2}, \frac{Y_{ij}^a Y_{ij}^t}{\sigma_{ij}^2}, \frac{Y_{ij}^t}{\sigma_{ij}^2}\right) = \left(f_{1ij}, f_{2ij}, f_{3ij}, f_{4ij}, f_{5ij}\right)$.

Then $Var(\hat{\beta}) \doteq Var\left(\sum_{u=1}^{5} A_u f_u\right) = Var\left(\sum_{i=1}^{k}\sum_{j=1}^{n_k} H_{ij}A\right)$

and $Var(\hat{\alpha}) \doteq Var\left(\sum_{u=1}^{5} B_u f_u\right) = Var\left(\sum_{i=1}^{k}\sum_{j=1}^{n_k} H_{ij}B\right)$.

Let $\Omega_{ij} = H_{ij}A$ (or $\Omega_{ij} = H_{ij}B$) the repeated sampling variance of $\hat{\beta}$ $(or\ \hat{\alpha})$ is given in Lemma 1 for appropriate choice of the $\{\Omega_{ij}\}$

Define $\Omega_{ij} = H_{ij}A$ (for estimating $Var(\hat{\beta})$) and $\Omega_{ij} = H_{ij}B$ (for estimating $Var(\hat{\alpha})$) as derived variables who`s sample sums have the same repeated sampling variance as the sampling variance of the BLUE for the PSMB model parameters, β and α.
Letting $a_{ij} = b_{ij} = \Omega_{ij}$ in Lemma 2.3.1, the repeated sampling variance of $\hat{\beta}$ is:

$$Var(\hat{\beta}) = Var\left(\sum_{i\epsilon S}\sum_{j\epsilon S_i}\Omega_{ij}\right) = k\left(1-\frac{k}{K}\right)S^2_{n_i\bar{\Omega}_i} + \frac{k}{K}\sum_{i\epsilon U} n_i\left(1-\frac{n_i}{N_i}\right)S^2_{\Omega_i} \quad (2.3.9)$$

where $S^2_{n_i\bar{\Omega}_i} = \frac{1}{K-1}\sum_{i\epsilon U}\left(\frac{n_i}{N_i}\sum_{j\epsilon U_i}\Omega_{ij} - \frac{1}{K}\sum_{l\epsilon U}\frac{n_l}{N_l}\sum_{j\epsilon U_l}\Omega_{lj}\right)^2$

and $S^2_{\Omega_i} = \frac{1}{N_i-1}\sum_{j\epsilon U_i}\left(\Omega_{ij} - \bar{\Omega}_i\right)^2$, where $\bar{\Omega}_i = \frac{1}{N_i}\sum_{j\epsilon U_i}^{N_i}\Omega_{ij}$.
$Var(\hat{\beta})$ is estimated with the sample data as:

$$\widehat{Var}(\hat{\beta}) = k\left(1-\frac{k}{K}\right)s^2_{n_i\bar{\Omega}_i} + \sum_{i\epsilon S} n_i\left(1-\frac{n_i}{N_i}\right)s^2_{\Omega_i}$$

Where $s^2_{n_i\bar{\Omega}_i} = \frac{1}{k-1}\sum_{i\epsilon S}\left(\sum_{j\epsilon S_i}\Omega_{ij} - \frac{1}{k}\sum_{l\epsilon S}\sum_{j\epsilon S_l}\Omega_{lj}\right)^2$, $s^2_{\Omega_i} = \frac{1}{n_i-1}\sum_{j\epsilon S_i}\left(\Omega_{ij} - \bar{\bar{\Omega}}_i\right)^2$,
and $\bar{\bar{\Omega}}_i = \frac{1}{n_i}\sum_{j\epsilon S_i}^{n_i}\Omega_{ij}$
The variance of $\hat{\alpha}$ and an estimator of its variance is obtained by using $\Omega_{ij} = H_{ij}B$ in the formulae directly above. The covariance between $\hat{\beta}$ and $\hat{\alpha}$, $Cov(\hat{\beta},\hat{\alpha})$ is given by Lemma 2.3.1 with $a_{ij} = H_{ij}A$ and $b_{ij} = H_{ij}B$ and its estimator obtained analogously to the formulae above.

These variances can be approximated by their expectations under the APM, but these expectations may be a bit too lengthy to be very informative.

From this the BLUE, the population total of the $\{Y^t_{ij}\}$ is:

$$\hat{Y}_{TOT} = \hat{\beta}\sum_{i\epsilon U}\sum_{j\epsilon U_i} Y^a_{ij} + \hat{\alpha}N. \quad (2.3.10)$$

And its repeated sampling variance is:

$$Var(\hat{Y}_{TOT}) = \left(\sum_{i\epsilon U}\sum_{j\epsilon U_i} Y^a_{ij}\right)^2 Var(\hat{\beta}) + N^2 Var(\hat{\alpha}) + 2\left(\sum_{i\epsilon U}\sum_{j\epsilon U_i} Y^a_{ij}\right)NCov(\hat{\beta},\hat{\alpha}).$$

An estimator for this Variance is:

$$\widehat{Var}(\hat{Y}_{TOT}) = \left(\sum_{i\epsilon U}\sum_{j\epsilon U_i} Y^a_{ij}\right)^2 \widehat{Var}(\hat{\beta}) + N^2\widehat{Var}(\hat{\alpha}) + 2\left(\sum_{i\epsilon U}\sum_{j\epsilon U_i} Y^a_{ij}\right)N\widehat{Cov}(\hat{\beta},\hat{\alpha}).$$

$\hat{Y}_{TOT}$ is the BLUE for total postage in the mail flow. In case $Y^t_{ij2l} = 1$ for all (i,j) and l, then $\hat{Y}_{TOT}$ (2.3.10) is the BLUE estimate of total mail pieces for the mail flow.

The USPS uses stratified cluster designs to estimate mail volumes. Their sample design parameters are not available until after the sample survey has been conducted (in fact they are collected with the sample data). Thus design effect can severely inhibit Design Based Inference. Within carefully defined strata, assignment of mail

pieces to mail containers is reasonably viewed as random. This structure imposes the model (2.3.2) directly from random process without appeal to data based conjecture. The BLUE derived from this model is a striking improvement over Designed Based procedures, Woodruff (2007,2008,2009,2010,2011).

3. Conclusions

This paper is an introduction to Pre-Sampling Model Based (PSMB) inference in survey sampling. PSMB models are deduced from randomized unit synthesis called Pre-sampling. This avoids data based model conjecture, the usual approach to model building. Model failure in PSMB inference has a status similar to design failure in Design Based inference – usually a topic of marginal concern. This gives PSMB inference a degree of credibility equivalent to inference from probabilities of selection (also derived from random process) in Design Based inference.

Pre-Sampling based models can capture considerable complexity by combining small numbers of simple data structures in many different combinations and thereby modelling substantial unit diversity. Applications of Pre-sampling Model Based Inference are examined where repeated sampling error under cluster designs is the criterion for evaluating estimators. Previous papers, Woodruff (2011, 2010, 2009, 2008, 2007) compared repeated sampling error under stratified cluster designs both theoretically and via simulation. The mathematics derived in Woodruff (2011) provided a foundation for PSMB inference and helped explain simulation results in earlier papers. PSMB estimators have smaller repeated sampling error than Design Based competitors and in most cases, much smaller errors. Design Based inference provides only unbiased estimation while PSMB inference controls both variance and bias, particularly in case of large design effects in Design Based Inference (design inefficiency).

Just as Design Based inference is founded on random sampling of population units, PSMB inference is based on randomized construction or synthesis of population units. Population units are comprised of entities called atoms and these atoms are a random selection from a very large (~ infinite) population of atoms. This randomization may occur naturally in many populations. For example, population units may be containers of mail filled with mail pieces, the atoms.

This paper looks in detail at two specific applications of PSMB theory. The first application to inference is a case of two auxiliary variables, two atom types, and one target variable. The second application is to sampling a population of containers to determine a total for some variables attached to each of the many items within a container (for example, to estimate total postage attached to all mail pieces in a mail flow where units are containers used to transport these mail pieces). This problem has one target variable, one auxiliary variable, and two atom types.

This atom structure appears to be a valid description of many types of population units – containers of mail, buckets of water draw from a stream, fields of crops on a farm, business establishments, tissue or blood samples. If the context in which the contents of these containers, buckets, or fields is such that their randomized synthesis is a reasonable description, then PSMB inference provides a viable alternative to Model Based and Design Based Inference.

In summary, probability sampling theory can be usefully expanded from randomized unit selection to both randomized unit selection and randomized unit synthesis through Pre-Sampling Model Based inference. Under this expansion, randomization remains the foundation of inference by providing impartiality and the foundation for the mathematics of probability theory. The population model is imposed by random process, not conjectured from data. Indeed, without this randomization it is hard to justify the application of probability theory to sample data. It is this randomization that distinguishes statistic from accounting. This gives the BLUE derived from the model a solid foundation where model failure is less likely than for models conjectured from either current sample data or dated historical data. Repeated sampling error can be expressed in terms of both the repeated sampling design and the stochastic properties the unit study variables inherit from the stochastic properties of the atom study variables expressed in the Atom Population Model.

References

Cochran, W.G., (1977), Sampling Techniques, 3rd ed., New York: Wiley, PP 167.

Rao, C.R. (1973), Linear Statistical Inference and its Applications, New York: Wiley, PP 230.

Woodruff, S. M. (2006), "Probability Sample Designs that Impose Models on Survey Data", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2007), "Properties of the Combined Ratio Estimator and a Best Linear Unbiased Estimator When Design Control is Problematic", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2008), "Inference in Sampling Problems Using Regression Models Imposed by Randomization in the Sample Design - Called Pre-Sampling", Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2009), "An Introduction to Pre-Sampling Inference I" Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2010), "An Introduction to Pre-Sampling Inference II" Proceedings of the American Statistical Association, Survey Research Methods

Woodruff, S. M. (2011), "An Introduction to Pre-Sampling Inference III" Proceedings of the American Statistical Association, Survey Research Methods
Pre-Sampling Model Based Inference IV