Introduction to Pre-Sampling Inference II

Stephen M Woodruff[1]

[1]Specified Designs, 800 West View Terrace, Alexandria VA 22301-2750

1. Introduction

This paper describes a solution to survey sampling estimation problems that result from common sample design practices and from inadequate sample design control. These problems tend to magnify the variance of probability expansion estimators based on those designs – in particular, the Combined Ratio Estimator. Examples occur in flow sampling where a sampling frame and its design parameters are usually not available until after the universe has been sampled. Even when sample design parameters are available in timely fashion, generally accepted sample design practices can also cause unnecessarily large sampling error in the design based Combined Ratio Estimator. For these reasons, efficient design based strategies are a serious concern.

For some populations, the population members (or units) are composed of smaller entities called atoms and for many of these populations, Pre-sampling inference flows from the randomized assignment of atoms to the sample/population units. This randomized selection of a unit`s atoms is called Pre-sampling and regression population models follow from this randomized selection. The models imposed by Pre-sampling provide Best Linear Unbiased Estimators (BLUE) for the study variable population totals. Following common survey sampling terminology, `sampling`, is the random selection of sample units from the population of units (each constructed by Pre-sampling).

The foundation for the model and the BLUE is a deductive consequence of the Pre-sampling design rather than an inductive consequence of potentially fickle, anomalous or dated historical sample data that is generally used to hypothesize models and derive BLUEs. Estimators based on models imposed on sample units by Pre-sampling, combine the comforting impartiality of randomization with the inferential power of model based BLUEs that possess attributes detailed in the Gauss-Markov Theorem as summarized in Graybill (1961) – Consistent, Efficient, Unbiased, Sufficient, Complete, and Minimum Variance Unbiased under Normality.

The Pre-sampling BLUE was developed to provide an alternative to the Combined Ratio Estimator (CRE) when design control is difficult due to physical, financial, and administrative constraints. Pre-sampling is a useful addition to probability sampling theory. An addition that combines the advantages of design based inference and those of model based inference while eliminating their major shortcomings.

An application of this methodology is found in two other papers, Woodruff (2007, 2008). In these papers, the sample and population units consist of random samples of atoms where the atoms within a unit can be described as simple random samples without replacement (SRSWOR) from all the atoms in the population or from a stratum of that population. Each of these atoms have data for all population study variables for which population totals are to be estimated and a unit’s values for these study variables is the sum over the unit’s atoms of these same study variables. The 2009 paper extends the theory in those two papers from univariate data to vector valued data but is restricted in application to situations where the number of atom types is equal to the number of auxiliary variables. This paper extends the 2009 paper to situations where the number of atom types differs from the number of auxiliary variables. This extension permits general applicability to most survey sampling inference problems where population units can be described as random samples of atoms which make up the population.

Numerous examples of such populations are encountered in practice: business surveys where business establishments are the units and an establishment`s employees are the atoms, mail surveys where mail containers are the units and the mail pieces within a container are its atoms, pollution studies where containers of water drawn from rivers and streams are the units and the particulate within a container are its atoms. Other examples are: households of people, agricultural fields of plants, factory production of widgets, and so on.

The study variables attached to each population and sample unit are the sums over each unit`s atoms of the same study variables attached to each atom. Let the atoms in the population be of m distinct types labeled with i=1,2, ……m. Let y_{kij} denote the vector of study variables attached to the j^{th} atom of the i^{th} atom type in the k^{th} unit (population or sample) then the vector of study variables attached to unit k is $Y_k = \sum_{i=1}^{m} \sum_{j=1}^{n_{ki}} y_{kij}$ where n$_{ki}$ is the number of type i atoms in unit k.

An example for m=1 is found in Woodruff (2006, 2007,2008) where containers (of mail pieces) are sampled and used to measure mail volumes (total kilograms and pieces). A mail container is a sample unit and the mail pieces it contains are its atoms. The container (unit) study variables are its piece count and the total of its piece weights. Within tightly defined categories of mail, it is entirely appropriate to describe the pieces within a container as a simple random sample without replacement (SRSWOR) from all the mail pieces within the mail category being sampled.

In Section 2, a population model is deduced from Pre-sampling random selection. This structure models several auxiliary variables (study variables for which population/strata totals are known), several target variables (study variables for which estimates of their population/strata totals are needed), and several atom types. Section 3 describes simulation studies that compare the Combined Ratio Estimator (CRE), Cochran (1977) with the Pre-sampling BLUE under repeated sampling from stratified cluster sampling designs.

When the number of atom types is equal to the number of auxiliary variables, Pre-sampling usually imposes a unique model on sample data and does so by deliberate designed randomization eliminating concerns over model fit or failure. The BLUE derived from this model avoids sample design inefficiencies that can be a consequence of common operational and administrative constraints. This BLUE combines advantages of design based inference (randomization and impartiality) and those of model based inference (Gauss-Markov properties) while eliminating their main shortcomings (inefficient sample design and potential model failure).

Woodruff (2009) required that the number of atoms types be equal to the number of auxiliary variables. This paper finds that little efficiency is lost when they are unequal and a generalized inverse is required. Use of a generalized inverse forces a loss of complete model specificity and a bias that is estimable and that can be substantially reduced. The result is an inference strategy that depends on both the sampling distribution and the Pre-sampling distribution. This strategy produces Pre-sampling BLUEs adjusted for incomplete model bias that possess MSEs which are much smaller than the MSE of design based alternatives, in particular the CRE. These MSEs for estimators in this paper are all evaluated with respect to repeated sampling under a stratified cluster sample design.

2. Derivation of Estimators

Sample selection and pre-sample selection in each stratum is independent of sample selection and pre-sample selection in all other strata. The BLUE for the population target totals is the sum of the independent strata target BLUEs. To minimize notation, the following derivation is for the BLUE of target totals in a single stratum and a stratum subscript is unnecessary. Population estimates are the sum of the strata estimates.

2.1 Models Imposed by Simple Random Pre-Sampling and the BLUE Derivation

The atoms in a sample unit are a random selection from all the atoms in the population or stratum thereof and are sampled without regard to atom type. n_{ki} (a random variable) is the number of type i atoms in sample unit k. The population size in atoms is large enough to be modeled as infinite and thus finite population considerations have negligible impact and are ignored in what follows.

y_{kij} is the column vector of study variables (auxiliary and target variables) attached to the j^{th} atom of type i in unit k for k=1,2,……r where r is the sample size in units for the SRSWOR of units in the stratum under consideration. Let $E\left(y_{kij}\right) = \mu_i$ for all k and j and let the covariance matrix of the components of y_{kij} be C_i for all k and j or $y_{kij} \sim (\mu_i, C_i)$ for i=1,2,⋯,m. [Define: $y_{kij} \sim (\mu_i, C_i)$ to mean $E\left(y_{kij}\right) = \mu_i$ and its covariance matrix is C_i.] The unit k study variable vector is $Y_k = \sum_{i=1}^{m} \sum_{j=1}^{n_{ki}} y_{kij}$ and for all ordered triples such that (k,i,j) $\neq$ $\left(k', i', j'\right)$, $y_{kij} \perp y_{k'i'j'}$. 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 an SRSWOR.

If U denotes the universe of population units, the structure described above may be easier to understand and follow 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 m 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.

By the definitions above, $E(Y_k) = \sum_{i=1}^{m} n_{ki}\mu_i$ and the covariance matrix of Y_k is Σ_k = $\sum_{i=1}^{m} n_{ki}\, C_i$.

Now let the vector of study variables for unit k be partitioned into two sub-vectors, the first is a vector of v auxiliary variables and the second is a vector of l target variables so that $y_{kij} = \begin{pmatrix} y^a_{kij} \\ y^t_{kij} \end{pmatrix}$ where y^a_{kij} is the v-vector of auxiliary variables and $y^t_{kij}$is the l-vector of target variables. Then

$$Y_k = \sum_{i=1}^m \sum_{j=1}^{n_{ki}} y_{kij} = \sum_{i=1}^m \sum_{j=1}^{n_{ki}} \begin{pmatrix} y^a_{kij} \\ y^t_{kij} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m \sum_{j=1}^{n_{ki}} y^a_{kij} \\ \sum_{i=1}^m \sum_{j=1}^{n_{ki}} y^t_{kij} \end{pmatrix} \overset{defn}{=} \begin{pmatrix} A_k \\ T_k \end{pmatrix} \text{ and}$$

$E(Y_k) = \sum_{i=1}^m n_{ki} \begin{pmatrix} \mu^a_i \\ \mu^t_i \end{pmatrix}$ where $\mu^a_i = E\left(y^a_{kij}\right) \forall$ k and j and similarly for μ^t_i. Letting $C_i = \begin{pmatrix} C_{ia} & C_{iat} \\ C_{ita} & C_{it} \end{pmatrix}$, where C_{ia} is the $v \times v$ covariance matrix of each y^a_{kij} for all k and j . C_{iat} is the $v \times l$ matrix of covariances between the components of y^a_{kij} and y^t_{kij} etc. and an upper prime denotes transpose, $C_{iat} = C'_{ita}$. Then

$$COV(Y_k) = \Sigma_k = \sum_{i=1}^m n_{ki} \begin{pmatrix} C_{ia} & C_{iat} \\ C_{ita} & C_{it} \end{pmatrix} \overset{defn}{=} \begin{pmatrix} \Sigma_{kA} & \Sigma_{kAT} \\ \Sigma_{kTA} & \Sigma_{kT} \end{pmatrix} \qquad (2.1.1)$$

Let $M_A = (\mu^a_1, \ \cdots \ , \mu^a_m)$, the matrix whose columns are the $\{\mu^a_i : i = 1,2, \dots, m\}$ and similarly $M_T = (\mu^t_1, \ \cdots \ , \mu^t_m)$ then

$E(Y_k) = \begin{pmatrix} M_A \\ M_T \end{pmatrix} \begin{pmatrix} n_{k1} \\ \vdots \\ n_{km} \end{pmatrix}$ or $Y_k = \begin{pmatrix} M_A \\ M_T \end{pmatrix} \begin{pmatrix} n_{k1} \\ \vdots \\ n_{km} \end{pmatrix} + \varepsilon_k$ where $\varepsilon_k \sim (0, \ \Sigma_k)$. Letting $N_k = \begin{pmatrix} n_{k1} \\ \vdots \\ n_{km} \end{pmatrix}$, ε_{kA} be the first v components of ε_k , and ε_{kT} be the last l components of ε_k then:

$$Y_k = \begin{pmatrix} M_A \\ M_T \end{pmatrix} N_k + \varepsilon_k \text{ where } \varepsilon_k = \begin{pmatrix} \varepsilon_{kA} \\ \varepsilon_{kT} \end{pmatrix} \sim (0, \ \Sigma_k) \text{ for k=1,2, …… r} \qquad (2.1.3)$$

M_A is $v \times m$ and M_T is $l \times m$.

The model given by (2.1.3) above can be transformed into one in which the target variables are matrix-proportional to the auxiliary variables as follows.

By definition, A_k=$M_A N_k$+ ε_{kA} and can be rewritten as: $N_k = M_A^-(A_k - \varepsilon_{kA})$ = $M_A^- A_k$ - $M_A^- \varepsilon_{kA}$ where M_A^- is a generalized inverse of M_A. When M_A is nonsingular (v=m), then M_A^-=M_A^{-1}. There are many different choices of M_A^- that satisfy $N_k = M_A^-(A_k - \varepsilon_{kA})$. In Section 2.2 below, a particular version (the s-inverse) is defined and used throughout the simulation studies and applications that follow.

N_k = $M_A^- A_k$ + ε^o_{kA} where ε^o_{kA} = - $M_A^- \varepsilon_{kA}$, and

$$\varepsilon^o_{kA} \sim (0, \ M_A^- \Sigma_{kA} (M_A^-)') \qquad (2.1.4)$$

Thus from (2.1.3) and substituting for N_k, $T_k = M_T(M_A^- A_k + \ \varepsilon^o_{kA}) + \varepsilon_{kT}$ =

$M_T M_A^- A_k + (M_T \varepsilon^o_{kA} + \ \varepsilon_{kT})$.

Let B= $M_T M_A^-$ (2.1.5)

then $T_k = BA_k \; + \delta_k$ for k=1,2,……..,r. (2.1.6)

Where $\delta_k \; = (M_T\varepsilon_{kA}^{o} + \; \varepsilon_{kT})$, $E(\delta_k)$=0, and from 2.1.1, 2.1.4, 2.1.5, and 2.1.6 the covariance matrix of δ_k is

$$\Sigma_{k\delta} \; = \; B\Sigma_{kA}B^{'} - B\Sigma_{kAT} - \Sigma_{kTA}B^{'} + \Sigma_{kT} \qquad (2.1.7)$$

Let B $= \left(b_{ij}\right)$, an $l \times v$ matrix and let the transpose of its i^{th} row be $B_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{im} \end{pmatrix} for\ i = 1,2, \dots.., l.$ Then (2.1.6) can be written as:

$$T_k = \begin{pmatrix} B_1^{'} \\ \vdots \\ B_l^{'} \end{pmatrix} A_k + \; \delta_k \text{ for k=1,2,……..,r} \qquad (2.1.8)$$

$$= \left(I \otimes A_k^{'}\right) \begin{pmatrix} B_1 \\ \vdots \\ B_l \end{pmatrix} + \; \delta_k = \; \left(I \otimes A_k^{'}\right)\beta + \; \delta_k \text{ where } \beta = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{l-1} \\ B_l \end{pmatrix} \qquad (2.1.9)$$

for k=1,2,……..,r where I is the $l \times l$ 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.

Stacking the $\{T_k\}_{k=1}^{r}$ from (2.1.9) into a column vector, the linear relationship summarizing all the sample data for k=1, 2, …,r is:

$$\begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{r-1} \\ T_r \end{pmatrix} = \begin{pmatrix} I \otimes A_1^{'} \\ I \otimes A_2^{'} \\ \vdots \\ I \otimes A_{r-1}^{'} \\ I \otimes A_r^{'} \end{pmatrix} \beta \; + \; \Delta$$

where I is the $l \times l$ identity matrix, and Δ is the rl random column vector $\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_r \end{pmatrix}$ with expectation of 0 and its covariance matrix is the block diagonal matrix of the $\{\Sigma_{k\delta}\}_{k=1}^{r}$, all off diagonal blocks are zero matrices. The BLUE (Rao 1973) for β is:

$$\hat{\beta} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \vdots \\ \hat{B}_{l-1} \\ \hat{B}_l \end{pmatrix} = \; \left(\sum_{k=1}^{r}(I \otimes A_k)\Sigma_{k\delta}^{-1}\left(I \otimes A_k^{'}\right)\right)^{-1} \sum_{k=1}^{r}(I \otimes A_k)\Sigma_{k\delta}^{-1}T_k \, , \qquad (2.1.10)$$

Substituting estimates for M_T , M_A, Σ_{kA} , Σ_{kT} and Σ_{kAT} , into $\Sigma_{k\delta}$, $\hat{\beta}$ can be approximated directly from the atom level sample data. Then the BLUE for the vector of target variable population totals and its model covariance matrix are:

$\hat{T}^{M}_{TOT}$ = $(I\otimes A'_{TOT})\hat{\beta}$ (2.1.11)

and Var($\hat{T}_{TOT}$)=$(I\otimes A'_{TOT})\left(\sum_{k=1}^{r}(I\otimes A_k)\Sigma^{-1}_{k\delta}\left(I\otimes A^{'}_{k}\right)\right)^{-1}(I\otimes A'_{TOT})'$

where $A_{TOT} = \sum_{k=1}^{N} A_k$, is the known vector of population (or stratum) auxiliary variable totals, N is the number of population (or stratum) units, and I is the $l \times l$ identity matrix. Note that $VAR(\hat{T}_{TOT})$ 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 $VAR(\hat{T}_{TOT})$, Woodruff (2009).

2.2 Bias Correction to the Pre-Sampling BLUE under an Incomplete Model

The sampling distribution can be used to estimate the bias due to incomplete model specificity ($v < m$ and singular M_A), This alleviates some of the affects of incomplete model fit due to fewer auxiliary variables than data types. The following theorem, the proof of which is immediate, makes it easier to analyze bias in $\hat{T}^{M}_{TOT}$ due to lessened model specificity in the general case (singular M_A).

Theorem 2.2: If M is a nonsingular matrix and is partitioned vertically as follows, $M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}$ where M_1 is has n rows and $M^{-1} = (F_1, \quad F_2)$ where and F_1 has n columns then F_1 is a generalized inverse of M_1.

Call this F_1 an s-inverse of M_1. The s-inverse is dependent on M_2. The mean matrix M_A can easily be expanded to include non-auxiliary target variables until it is non-singular, its inverse available, and Theorem 2.2 applied to find an s-inverse for the mean matrix of actual auxiliary variables.

In case of a singular M_A there will generally be some bias in $\hat{T}^{M}_{TOT}$. Let this bias of $\hat{T}^{M}_{TOT}$ for estimating the target variable totals be H. Then H = $E\left(\hat{T}^{M}_{TOT} - T_{TOT}\right)$= $E\left((I\otimes A'_{TOT})\hat{\beta} - T_{TOT}\right)$ from (2.1.11).

Substituting the Horwitz-Thompson estimates for A_{TOT} and T_{TOT} into this expression, define $\hat{H}$ as: $\hat{H} = \left(I\otimes \sum_{k=1}^{r}\frac{A'_k}{\pi_k}\right)\hat{\beta} - \sum_{k=1}^{r}\frac{T_k}{\pi_k}$ = $\sum_{k=1}^{r}\frac{\left((I\otimes A'_k)\hat{\beta} - T_k\right)}{\pi_k}$.

Then $E(\hat{H})$ = H where this expectation is with respect to both sampling and Pre-sampling distributions.

Let $\hat{T}^{Mc}_{TOT}$ = $\hat{T}^{M}_{TOT}$ - $\hat{H}$, (2.2.1)

then $E\left(\hat{T}^{Mc}_{TOT}\right) = T_{TOT}$.

$\hat{T}^{Mc}_{TOT}$ is the bias corrected Pre-sampling BLUE (under the incomplete model with s-inverse).

$\hat{T}^{Mc}_{TOT}$ is the estimator to be studied in the simulation studies in the next section that relies on both sampling and Pre-sampling distributions. It avoids pitfalls in stratified cluster designs that enlarge sampling error. It relies on only the local sample design within each stratum (a simple random cluster sample that avoids variance magnifying design complexities encountered when stratum estimates are combined across strata as done in the CRE).

The analysis in Section 3 compares the generalized Pre-sampling BLUE from (2.1.11) under a singular M_A (and requiring a generalized inverse of M_A) to the Pre-sampling BLUE with additional information (an additional auxiliary variable, and row in M_A that makes it nonsingular and the model unique and totally specified given this additional information).

Theorem 2.2 facilitates comparison of these two Pre-sampling estimators. In particular, it expresses the bias of the Pre-sampling BLUE under a non-singular M_A as the expected value of an expression that can be approximated and its expected value estimated. The s-inverse from this theorem is used in all that follows in this paper. This s-inverse gives Pre-sampling BLUEs with expectations and variances that are very similar to the BLUEs derived with the Moore-Penrose generalized inverse.

The covariance matrices needed in (2.2.1) are linear combinations of the atom covariance matrices, $\{C_{ia}, C_{iat}, C_{it}\}_{i=1}^{m}$. These can be estimated from the many atoms in relatively few sample units using the standard variance estimate, the MLE under Normal Theory, based on hundreds to thousands of atoms. This is described in the next section for a stratified cluster sample design and 5 study variables (two auxiliary variables and three target variables).

Recall that in case of more than one stratum the derivations in Sections: 2.1 and 2.2 for $\widehat{\mathrm{T}}_{\mathrm{TOT}}^{\mathrm{Mc}}$ are for a single stratum. Summing these $\widehat{\mathrm{T}}_{\mathrm{TOT}}^{\mathrm{Mc}}$ over the strata is the estimator for the population total of the target variables in a population with multiple strata.

2.3 Special Cases

In case the stratum totals for the components of N_k in (2.1.3) are known for all k, the derivation for a BLUE is immediately available following the above procedure except that M_A^- from (2.1.2) is no longer needed, and the transformation given by (2.1.7) is unnecessary. The equation used in place of (2.1.8) is: $T_k = M_T N_k + \varepsilon_{kT}$ where the BLUE for M_T is used to derive the BLUE for all target variables totals paralleling the procedure used above to find the BLUE for B and the BLUE for the target variable totals. This BLUE uses sample and population unit counts only, paralleling the Horvitz-Thompson Estimator and will be evaluated in a later paper.

3. Pre-sampling BLUE Compared to the Combined Ratio Estimator - A Simulation Study

Section 2 derives the versions of the Pre-sampling model based BLUEs for a single stratum. The population total estimates from these three stratum estimators is the sum over the strata of the individual stratum total estimates. The notation for this sum in this section is the same as the stratum level notation. Hopefully this notational simplification will cause little confusion.

The following study compares the Combined Ratio Estimator (CRE) for a stratified population to the three Pre-sampling model based estimators:

1) The bias adjusted BLUE under incomplete model, $\hat{T}_{TOT}^{Mc}$ from (2.2.1) summed over the strata,

2) The BLUE under incomplete model, $\hat{T}_{TOT}^{M}$ from (2.1.11), without bias adjustment, summed over the strata, and using a single auxiliary variable with two atom types,

3) The BLUE under complete model, $\hat{T}_{TOT}^{Mn}$ from (2.1.11), summed over the strata and using two auxiliary variables and two atom types where M_A is nonsingular.

The Pre-sampling BLUE in 3) is included for comparison with $\hat{T}_{TOT}^{Mc}$ & $\hat{T}_{TOT}^{M}$ and helps quantify the effect of incomplete auxiliary data (one auxiliary variable) compared to complete auxiliary data (same number of auxiliary variables as atom types).

There are five study variables consisting of two auxiliary variables and three target variables. Analysis of the estimators is done with respect to repeated sampling under a stratified cluster sample design using 1000 replications of sampling and estimation to produce 1000 independent estimates for the three target variable population totals. These 1000 replicate estimates are used to estimate mean, variance, and mean squared error (MSE) for $\hat{T}_{TOT}^{Mn}$, $\hat{T}_{TOT}^{Mc}$, $\hat{T}_{TOT}^{M}$, and CRE. The results are similar to those in Woodruff (2007, 2008, 2009) , demonstrate the effects inefficient sample design on the probability based Combined Ratio Estimator, and indicate that the general application (fewer auxiliary variables than atom types) of Pre-sampling inference suffers relatively little additional sampling error in face of this data deficiency (M_A singular and the attendant lack of total model specificity).

The populations studied below have F strata (F≅50) where U_f denotes the set of universe units in stratum f and N_f for f=1, 2, 3, …….F denotes the number of universe units in U_f. Each unit in U_f is itself a simple random pre-sample of atoms from all the atoms making up units in U_f. The units in each stratum are partitioned into first stage clusters. Let M_f be the set of clusters in U_f and $G_f \cong 40$ be the number of clusters in M_f. Let s_f be an SRSWOR of size n_f from M_f. Let U_{fd} be the set of second stage universe units in cluster d of stratum f for d=1,2,3,…., G_f. Let N_{fd} be the number of universe units in U_{fd} $\left(\sum_{d=1}^{G_f} N_{fd} = N_f\right)$. Let s_{fd} be an SRSWOR of size n_{fd} selected from the universe units in U_{fd}. Both n_f and n_{fd} are roughly 4 for all f and d in the simulations below.

Let $Y_{fdk} = \begin{pmatrix} A_{fdk} \\ T_{fdk} \end{pmatrix}$ be the vector of study variables attached to the k^{th} unit in U_{fd} ***(Note that Y with 3 subscripts in this section is a cluster-unit breakout of the single subscript, k, used in Section 2 without the stratum f notation)***. Let π_{fdk} be the probability of selection of the k^{th} unit in U_{fd}. Then $\pi_{fdk} = \frac{n_f}{G_f}\frac{n_{fd}}{N_{fd}}$ for a k in U_{fd}. Let Y_{fdk}^{s} be the vector of study variables for the k^{th} **sample** unit from s_{fd}. The Horwitz-Thompson Estimator (probability expansion) for the stratum total of the vectors $\left\{Y_f = \sum_{d\varepsilon U_f}\sum_{k\varepsilon U_{fd}} Y_{fdk}\right\}$ in U_f is $\hat{Y}_f = \begin{pmatrix} \hat{A}_f \\ \hat{T}_f \end{pmatrix} = \sum_{d\epsilon s_f}\sum_{k\varepsilon s_{fd}} \frac{1}{\pi_{fdk}} Y_{fdk}^{s}$. The auxiliary variable totals in stratum f are known and denoted, $A_f = \sum_{d=1}^{G_f}\sum_{k=1}^{N_{fd}} A_{fdk}$, then $A_f = E(\hat{A}_f)$ (expectation under repeated sampling). Let $A = \sum_{f=1}^{F} A_f$ and similarly for Y and T. The first auxiliary variable in A_{fdk} is used for ratio adjustment in the CRE, $\beta_f = \frac{Y_f}{A_{f1}}$ where A_{f1} is the first component of A_f and $\hat{A}_{f1}$ is the first component of $\hat{A}_f$. β_f is an $l \times 1$ vector of ratios of target variable totals to the first auxiliary variable total. Four estimators for each of the three target variable totals are compared in the tables and graphs below. $\hat{T}_C$ is the CRE for the vector of target variable totals, $A_{(1)}$ is the first component of A, and

$$\hat{T}_C = A_{(1)} \frac{1}{\sum_{f=1}^{F} \hat{A}_{f1}} \sum_{f=1}^{F} \hat{T}_f \qquad (3.1)$$

the Combined Ratio Estimator (CRE) for the population total of the vector of target variables.

For many populations, the variance of the CRE is substantially governed by (an increasing function of) population parameters, Q and $\{\Delta_i\}_{i=1}^{3}$ (i denotes target variable here, rather than atom type) defined next.

Q = $\frac{1}{F}\sum_{f=1}^{F} \frac{1}{(G_f-1)} \sum_{d=1}^{G_f} \left(N_{fd} - \bar{N}_f\right)^2$ and $\bar{N}_f = \frac{1}{G_f}\sum_{d=1}^{G_f} N_{fd}$.

The $\{\Delta_i\}$ are defined as $\Delta_i = \frac{1}{F}\sum_{f=1}^{F}\left(\beta_{fi} - \bar{\beta}_i\right)^2$ for each target variable i where β_{fi} is the i^{th}component of β_f and $\bar{\beta}_i = \frac{1}{F}\sum_{f=1}^{F} \beta_{fi}$.

As was shown in Woodruff (2007, 2008,2010), the sampling error of a combined ratio Horwitz-Thompson estimator (CRE) for the i^{th} target total increases with Q and Δ_i for populations where study variable totals are roughly proportional to unit counts.

Each of the populations has different Q and Δ_i for i=1,2,3, and each population has several hundred thousand units spread over 50 strata. Each unit is generated with two types of atoms until a randomly determined size threshold is reached (as measured by the unit's total for the first auxiliary variable, a_1). All study variables are greater than or equal to zero. This process yields units that are roughly similar in size as measured by its a_1 total (the calibration variable used by the CRE). This process models the way mail containers are filled or water quality tested where the size of a unit is determined by the limits (weight) of what can easily be carried or handled by an individual. The distribution of atom types per unit is also random, modeling the occurrence of particulate type that would be contained in a water sample taken from a bucket dipped into a stream. For each stratum, the same model for constructing the units is applied with different models in different strata. Cluster sizes (in numbers of units) are randomly determined for each stratum and the clusters within a stratum can be described as SRSWORs from all the stratum`s units.

The covariance matrix used in the BLUE is estimated from sample data collected at the atom level from 5 sample units in each stratum. Although there are a total of about 16 sample units per stratum, only a subset of 5 is used for parameter estimation. This follows proposed applications where only atom totals for each sample unit are necessary for most sample members, avoiding the expense of enumerating atom level data from all sample units.

The covariance matrix for the vector of study variables, Y_{fdk}^{s}, (auxiliary variables and target variables) for the k^{th} sample unit in s_{fd} is $\Sigma_{fk} = \sum_{i=1}^{m} n_{fki} \begin{pmatrix} C_{fia} & C_{fiat} \\ C_{fita} & c_{fit} \end{pmatrix}$ = $\sum_{r=1}^{m_A} n_{fki}\, C_{fi}$. Note that this is similar to notation in part 2 except that a stratum subscript is necessarily added. Let the set of type i atoms in these 5 sample units in stratum f be $\varLambda_{fi}$. $\varSigma_{fk}$ is estimated with:

$\hat{\varSigma}_{fk} = \sum_{i=1}^{m} n_{fki}\, \hat{C}_{fi}$ where $\hat{C}_{fi}$ = MLE under normality from the atom vectors $\{y_{fiq}\}_{q\varepsilon\varLambda_{fi}}$ = $\left\{\begin{pmatrix} y_{fiq}^{a} \\ y_{fiq}^{t} \end{pmatrix}\right\}_{q\varepsilon\varLambda_{if}}$. Let D_{fi} be the number of atoms in $\varLambda_{fi}$. $\hat{C}_{fi}$ = $\frac{1}{D_{fi}-1}\left(S_{fi} - \left(\frac{1}{D_{fi}}\right) V_{fi}V_{fi}^{'}\right)$ where S_{fi} =$\sum_{q=1}^{D_{fi}} y_{fiq}\, y_{fiq}^{'}$ and V_{fi} = $\sum_{q=1}^{D_{fi}} y_{fiq}$ and both these sums are over elements of $\varLambda_{fi}$.

The matrices M_{fT} and M_{fA} are stratum f versions of M_T and M_A, consisting of study variable means by atom type and are estimated from the atoms in $\bigcup_{i=1}^{m} \varLambda_{fi}$. If μ_{fji} is the $(j, i)^{th}$ element of M_{fT} (mean of j^{th} target variable for type i atoms in stratum f) then μ_{fji} is estimated with $\hat{\mu}_{fji}$ $= \frac{1}{D_{fi}} \sum_{q=1}^{D_{fi}} t_{fjiq}$ where t_{fjiq} is the value of the j^{th} target variable of type i for the q^{th} atom in $\varLambda_{fi}$. The M_{fA} are estimated similarly. This provides the estimates, $\widehat{M}_{fT}$ and $\widehat{M}_{fA}$, for M_{fT} and M_{fA} in (2.1.5). $\hat{B}_f = \widehat{M}_{fT}\widehat{M}_{fA}^{-}$ where $\widehat{M}_{fA}^{-}$ is the s-inverse defined in Theorem 2.2. This $\hat{B}_f$ is substituted for B in (2.1.7) to estimate $\varSigma_{fk\delta}$. This estimate is denoted $\hat{\varSigma}_{fk\delta}$ and used in (2.1.10) to compute the BLUE when there are two atom types and one auxiliary variable.

There are 4 simulation studies summarized in the tables below for a variety of values for root Q and regression coefficient variability measures, $(\Delta_1, \Delta_2, \Delta_3)$. These 4 were selected from about 60 other simulated populations and generally illustrate the relative magnitudes of the MSEs of the four estimators for the 60 populations. The four estimators studied in the tables are:

BLUE, Complete Model, two auxiliary variables & two atom types (2.1.11)	BLUE, Generalized Inverse with HT bias correction, one auxiliary variable & two atom types (2.2.1)	BLUE, Generalized Inverse w/o HT bias correction, one auxiliary variable & two atom types (2.1.11)	Combined Ratio Estimator (3.1)
$\hat{T}_{TOT}^{Mn}$	$\hat{T}_{TOT}^{Mc}$	$\hat{T}_{TOT}^{M}$	$\hat{T}_C$

The Squared Bias, Variance, and Mean Squared Error of these estimators are tabulated below for each of the three target variable estimates. Below each table title (Simulation Results 1 through 4) are found the population parameters (Q, Δ_1, Δ_2, Δ_3) for the study population. As these four parameters increase in size, the differences between the MSE of the CRE and the MSEs of the three Pre-sampling BLUEs increase.

It appears that incomplete auxiliary variable data (fewer auxiliary variables than atom types) is not a serious obstacle to Pre-sampling inference when the HT bias adjustment is subtracted from $\hat{T}_{TOT}^{M}$ to yield $\hat{T}_{TOT}^{Mc}$. $\hat{T}_{TOT}^{Mc}$ achieves nearly as much MSE reduction compared to the Combined Ratio Estimator as achieved with complete auxiliary data through $\hat{T}_{TOT}^{Mn}$, included as a benchmark to measure the penalty for incomplete auxiliary data model.

Across stratum calibration, a defining feature of the Combined Ratio Estimator, magnifies its variance when the stratum ratios of target to auxiliary variable vary widely (large values of the (Δ_1, Δ_2, Δ_3)). If the separate ratio estimator were used in these simulations, the differences between the MSEs of this estimator and the Pre-sampling estimators would be greatly reduced. The Combined Ratio Estimator is the basis for comparison because it seems to be the standard across many sample survey applications.

$\hat{T}_{TOT}^{M}$ tends to be biased due to incomplete model specificity. The bias adjustment, $\hat{H}$, through Horwitz-Thompson estimation of the residual is included in $\hat{T}_{TOT}^{Mc}$ and largely corrects for this bias. $\hat{T}_{TOT}^{Mc}$ achieves an MSE that is nearly as small as the complete model BLUE, $\hat{T}_{TOT}^{Mn}$. All three of these estimators have much smaller MSE than the Combined Ratio Estimator, $\hat{T}_{C}$.

The bias corrected incomplete data BLUE, $\hat{T}_{TOT}^{Mc}$, is the sum of the incomplete data BLUE, $\hat{T}_{TOT}^{M}$, and - $\hat{H}$. Since these two terms are nearly uncorrelated, this implies that the variance of $\hat{T}_{TOT}^{Mc}$ is roughly the sum of the variances of these two terms and therefore generally greater than the variance of $\hat{T}_{TOT}^{M}$. This is exhibited in all the tables above, Simulation Results 1 through Simulation Results 4. The squared bias reduction resulting from the HT bias correction term is greater than this increase in variance with a net result that $\hat{T}_{TOT}^{Mc}$ has an MSE smaller than the MSE of $\hat{T}_{TOT}^{M}$. Thus a singular M_A and lack of complete model specificity results in relatively little additional MSE compared to a complete model BLUE.

Simulation Results 1 (Data in Trilions for Universe 'apr11')

$\sqrt{Q}$ = 194 Δ_1= .36 Δ_2 =.57 Δ_3 = .47

Estimator	Bias squared	Variance	MSE
---------------------------------- variable=t1 useable=988 --------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	25	421	446
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	5	933	938
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	391	767	1,158
Combined Ratio Est CRE	1	15,039	15,040
---------------------------------- variable=t2 useable=988 --------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	34	564	598
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	3	1,100	1,103
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	487	907	1,394
Combined Ratio Est CRE	19	20,709	20,728
---------------------------------- variable=t3 useable=988 --------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	48	611	659
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	3	1,278	1,281
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	610	1,052	1,662
Combined Ratio Est CRE	23	20,425	20,448

Simulation Results 2 (Data in Trilions for Universe 'apr4')

$$\sqrt{Q} = 202 \quad \Delta_1 = 3.9 \quad \Delta_2 = 3.46 \quad \Delta_3 = 8.87$$

Estimator	Bias squared	Variance	MSE
------------------------------ variable=t1 useable=998 ----------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	810	11,170	11,980
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	10	15,885	15,895
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	2,595	16,771	19,366
Combined Ratio Est CRE	57	1,004,821	1,004,878
----------------------------- variable=t2 useable=998 -----------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	1,405	16,278	17,683
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	7	24,990	24,997
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	3,848	41,115	44,963
Combined Ratio Est CRE	129	1,365,217	1,365,346
----------------------------- variable=t3 useable=998 -----------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	1,642	28,771	30,413
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	8	36,249	36,257
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	4,774	35,938	40,712
Combined Ratio Est CRE	291	1,636,229	1,636,520

Simulation Results 3 (Data in Billions for Universe 'mar23')

$$\sqrt{Q} = 69 \quad \Delta_1 = .11 \quad \Delta_2 = .15 \quad \Delta_3 = .20$$

Estimator	Bias squared	Variance	MSE
------------------------------ variable=t1 useable=995 ----------------------- -			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	27	382	409
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	2	910	912
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	300	844	1,144
Combined Ratio Est CRE	17	6,137	6,154
----------------------------- variable=t2 useable=995 -----------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	128	809	936
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	0	1,496	1,496
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	623	1,454	2,077
Combined Ratio Est CRE	25	7,900	7,917
----------------------------- variable=t3 useable=995 -----------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	1,011	2,026	3,037
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	0	2,165	2,165
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	1,179	2,107	3,286
Combined Ratio Est CRE	36	10,666	10,702

Simulation Results 6 (Data in Billions for Universe 'mar3'**)**

$\sqrt{Q} = 71 \quad \Delta_1 = .23 \quad \Delta_2 = .40 \quad \Delta_3 = .47$

Estimator	Bias squared	Variance	MSE
---------------------------- variable=t1 useable=996 --------------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	73	3,210	3,283
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	1	7,286	7,287
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	608	6,816	7,425
Combined Ratio Est CRE	189	183,106	183,112
--------------------------- variable=t2 useable=996 --------------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	217	7,308	7,526
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	14	12,811	12,825
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	1,092	12,001	13,093
Combined Ratio Est CRE	246	375,222	375,091
-------------------------- variable=t3 useable=996 -------------------------			
BLUE Complete Model $\hat{T}_{TOT}^{Mn}$	406	9,944	10,350
BLUE Gen Inverse with HT $\hat{T}_{TOT}^{Mc}$	3	15,292	15,295
BLUE Gen Inverse w/o HT $\hat{T}_{TOT}^{M}$	1,441	14,622	16,063
Combined Ratio Est CRE	10	465,033	464,576

4. Conclusions

This continues the development of a probability based methodology called Pre-sampling that imposes a model on sample data and that avoids onerous design effects often encountered in design based inference. Pre-sampling is based on randomized construction of sample units (as opposed to their randomized selection in design based inference). This methodology provides a Best Linear Unbiased Estimator from a model deduced from the Pre-sampling design. Pre-sampling inference largely eliminates questions of model fit or failure. Comparisons in Section 3 between Pre-sampling estimates and the Combined Ratio Estimator under a stratified cluster design highlight the potential for extreme design effects that can occur with the Combined Ratio Estimator. These design effects result from inadequate sample control, common sample design practice, and can be quite large.

When the Auxiliary Mean Matrix (M_A in Section 2) is nonsingular, the model is complete and totally specified by the Pre-sampling design. This case was examined in detail in Woodruff (2009). This paper generalizes these 2009 results to the case where M_A is singular and the model is consequently incomplete (not uniquely specified). In this general case, stratum residual adjustments using probabilities of selection, alleviate the worst effects of an incomplete model and BLUEs based on the incomplete model. The residual adjustment, $\hat{H}$, included in $\hat{T}_{TOT}^{Mc}$, provides reductions in mean squared error (MSE) similar to those expected in case of a completely specified model. This can be observed in the tables in Section 3 where the first three rows of MSE estimates for each target variable total (one for the three versions of the BLUE under complete and incomplete data) are similar and all three are much smaller than the fourth row (the MSE of the Combined Ratio Estimator).

It appears that the generalized methodology developed here using generalized inverse matrices can be usefully applied to a wide variety of problems where sample control is problematic or where stratum and/or cluster parameters will generate unnecessarily large sampling error regardless of sample control. There seems to be relatively few applied sampling problems where some type of atom structure within sample and population units does not obtain. Some examples of these Sampling/Pre-sampling problems for estimating population totals with a unit/atom structure are listed in the following table. Within carefully designed strata, a unit's atom sample can be appropriately modeled as a simple random sampling without replacement from all the atoms in the stratum or population.

Some Examples of Populations with Atom Structure

Populations/Programs	**Units**	**Atoms and Atom Types (by)**
Household Surveys	Households	Household Occupants (by age)
Business Establishment Surveys	Business Establishments	Employees (by occupation)
Mail Surveys	Bags, Trays, or Tubs of mail pieces	Mail pieces (by shape or category)
Bioassay of a Species	Salmon (any other species)	Parasites (by type)
Agricultural Inspection	Containers of fruit	Pieces of fruit (by type)
Factory Production/Quality Control	Establishments	Widgets (by type) produced each day, week, or month.
Agricultural Production	Fields or farms	Plants (by Type)

Pre-sampling inference changes the focus of finite population sampling from randomized sample unit selection to randomized sample unit construction. In Pre-sampling inference, randomized unit selection plays a relatively minor role where it is applied as a refinement Pre-Sampling estimates where there is bias due to an incomplete Pre-Sampling model. The extraordinary magnitude of sampling error reduction in the Pre-sampling BLUE, $\hat{T}_{TOT}^{M}$, compared to the Combined Ratio Estimator that were observed in Woodruff (2009) still hold in the generalized methodology described in Section 3 where $\hat{T}_{TOT}^{Mc}$ is derived, the BLUE corrected for an incomplete model.

The variance estimator described in Woodruff (2009) can be readily applied to the generalized Pre-sampling BLUE developed in Section 2 above. The s-inverse of a matrix defined in Section 2 provided Pre-sampling BLUEs that were little different (virtually same mean and variance) from the Moore-Penrose Inverse. The s-inverse was used here for purely analytic reasons – The s-inverse makes a components of variance analysis somewhat easier, an analysis that may be published next year so that estimator comparisons that don't depend solely on simulation studies can be made. Simple random Pre-sampling was studied here but clearly other Pre-sampling designs would better describe Pre-sampling for some of the populations in the table of populations above.

In summary, this paper describes a technique for using both sampling and Pre-sampling distributions to produce estimates of population totals. This technique requires far fewer restrictions on sampled populations than needed in Woodruff (2009) where Pre-Sampling inference was introduced. The emphasis is on Pre-Sampling model based inference where design based inference is applied to make minor bias adjustments to the Pre-Sampling model based estimates in general applications where incomplete models describe the sample data. As in the applications presented in Woodruff (2009), the MSEs of Pre-sampling model based BLUEs are orders of magnitude smaller than those of the standard design based estimator, the Combined Ratio Estimator.

It would be informative to complete a purely analytic study that explains the simulation results in Section 3 with formulae that better clarify the reasons for large differences between the sampling errors of the Pre-sampling model based estimators and those of the Combined Ratio Estimator. This analysis seems to be contingent upon theorems that permit a representation of the inverse for the sum of several matrices as a linear combination (or some similar structure) of the inverses of each of these matrices – possibly an interesting problem or more likely, a problem with no general solution.

References

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

Graybill, F. A. (1961). An Introduction to Linear Statistical Models, Volume 1, McGraw Hill Inc., PP 114.

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

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” Proceedings of the American Statistical Association, Survey Research Methods