Inference in Sampling Problems Using Regression Models Imposed by Randomization in the Sample Design - Called Pre-Sampling

Steve Woodruff

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

Abstract

The variance of a probability expansion estimator is sensitive to sample design and can be large when the design is subject to administrative and physical constraints. Models and model based estimates provide a more efficient alternative but are dependent on models of questionable validity and sacrifice the impartiality of randomization. There is a third estimation technique that retains the comforting impartiality of randomization and uses this randomization to impose a model on the sample data under which there is a Best Linear Unbiased Estimator (BLUE). Since the model is imposed by the statistician through designed randomization, model failure tends toward a non-issue. Examples from actual surveys are provided where the sampling variance of the Combined Ratio Estimator is tens to hundreds of times greater than that of the BLUE.

Key Words: Probability Sampling, Regression Models, BLUE, Combined Ratio Estimator

1. Introduction

Models can be imposed on sample data by randomized construction of the sample units (and universe units). This randomized construction is called pre-sampling. In what follows, “pre-sampling” is used to describe the randomized construction of sample units and “sampling” is used to describe the selection of these randomly constructed sample units from the universe of all units. The models imposed by pre-sampling provide Best Linear Unbiased Estimators (BLUE) for all study variables, Rao (1973); their foundation is the pre-sampling design, not the potentially fickle or dated nature of historical sample data generally used to hypothesize models. Therefore the estimators based on pre-sampling imposed models retain the impartiality of randomization and the inferential power of model based BLUEs under models assured by the pre-sampling design. The methodologies described here were developed to provide an alternative to inefficient design based estimators when design control is difficult due to physical and administrative constraints.

An application of this methodology is found in two other papers, Woodruff (2006,2007). In these papers, the sample and population units consist of sets whose elements are called atoms. These sets can reasonably 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 subunits (atoms), have data for all population study variables for which population totals are to be estimated. This paper expands upon those two papers and applies the basic structure described in them to a multivariate framework.

First consider the structure of universe and sample units as samples of atoms. If y_{ik} denotes the i^{th} study variable attached to the k^{th} sample unit then y_{ik} can be written as the following sum over the atoms that comprise the k^{th} sample unit.

$y_{ik} = \sum_{j=1}^{n_k} \omega_{ikj}$ where ω_{ikj} is the study variable for the j^{th} atom of the k^{th} sample unit’s i^{th} study variable and n_k is the pre-sample size of atoms in the SRSWOR from the population of atoms in the stratum being sampled (the stratum containing the k^{th} unit).

These atom totals that comprise each sample unit, y_{ik}, can be recorded without measuring or recording these study variables for each atom. For example, if weight is a study variable attached to each atom, weighing the whole unit provides this sum. Other measures like be pressure, volume, or radioactivity may share this property. An example is found in Woodruff (2006, 2007) where containers holding mail pieces are sampled and used to measure mail characteristics. A mail container is a sample unit and its mail pieces are its atoms. The study variables are weight and postage to be found on each mail piece and their container sum is the unit’s study variable for weight and postage. Within tightly defined categories of mail, it is entirely appropriate to think of the pieces within a container as an SRSWOR from all the mail pieces within the mail category. Another example is measurement of stream pollution or other water born particulate. The quantity of particulate (atoms) in a bucket (sample unit) sampled from the stream is proportional to the weight of water in the bucket. Weight or volume is the auxiliary variable that is recorded for the entire volume of water passing the sampling point during a time interval defining a stratum.

In Section 2, a multivariate structure is described for deriving a population model under pre-sampling. This structure allows for several auxiliary variables (study variables for which population/strata totals are known) and several types of atoms within each sample unit. This structure provides greater flexibility and more appropriate sample expansions for different study variables, escaping the restriction of applying the same probability based expansion to all study variables (as with Horwitz-Thompson estimation). Probability based expansions may be appropriate for some variable estimates but can be very inefficient for others, especially in large multi-purpose surveys.

Section 3 describes a simulation study that compares the combined ratio estimator, Cochran (1977) with the model based BLUE under an inefficient stratified cluster sample design and SRSWOR pre-sampling. Section 4 concludes that this study is just the tip of an iceberg and that sampling theory may profit from the extension of this work to more general pre-sampling designs, their models, and their BLUEs.

In Summary, pre-sampling imposes a model on sample data that is substantially immune from model failure and inefficiency in the sample design. This methodology combines the strengths of design based inference (randomization/impartiality) with the strengths model based inference (existence of a BLUE) while eliminating their respective shortcomings (inefficient sample design and potential model failure).

2. Models Induced by Simple Random Pre-Sampling

For the population considered here, the atoms contained within a sample unit are an SRSWOR from all atoms in the population or stratum of the population. The number of universe units and their constituent atoms in the study population are both sufficiently large to be appropriately described as coming from a continuous density function (infinite population). Finite population considerations are omitted in this development.

The study variables attached to each population unit, k, consist of auxiliary variables and target variables and can be described as a vector, Y_k, where:

$Y_k = \begin{pmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{mk} \end{pmatrix} = \begin{pmatrix} A_k \\ T_k \end{pmatrix} = \begin{pmatrix} \vec{\mu}_A \\ \vec{\mu}_T \end{pmatrix} n_k + \begin{pmatrix} \vec{\epsilon}_{Ak} \\ \vec{\epsilon}_{Tk} \end{pmatrix}$ and n_k is the number of atoms in the k^{th} universe (or sample) unit, A_k is the m_A-vector (column) of the m_A auxiliary variables, the variables whose population means or totals are known, $\vec{\mu}_A = E(A_k)$, $\vec{\epsilon}_{Ak}$ is a vector valued random variable with $E(\vec{\epsilon}_{Ak})$ = zero vector, and with $m_A \times m_A$ covariance matrix, Σ_A . T_k is the m_T-vector of target variables, the variables whose population means or totals are to be estimated, $m = m_A + m_T$, and $\vec{\mu}_T$, $\vec{\epsilon}_{Tk}$, & Σ_T defined analogously to $\vec{\mu}_A$, $\vec{\epsilon}_{Ak}$, & Σ_A above.

Let $A_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{m_A k} \end{pmatrix}$ and $T_k = \begin{pmatrix} t_{1k} \\ t_{2k} \\ \vdots \\ t_{m_T k} \end{pmatrix}$. Then

$$E\begin{pmatrix} \vec{\epsilon}_{Ak} \\ \vec{\epsilon}_{Tk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and the covariance matrix of } \begin{pmatrix} \vec{\epsilon}_{Ak} \\ \vec{\epsilon}_{Tk} \end{pmatrix} \text{ is } \begin{pmatrix} \Sigma_A & \Sigma_{AT} \\ \Sigma_{TA} & \Sigma_T \end{pmatrix} n_k. \qquad (2.0)$$

A prime is used to denote matrix transpose, ie $\Sigma_{AT}^{'} = \Sigma_{TA}$.

Expectation and variance are matrix-proportional to the number of atoms (n_k) and by the transitivity of matrix-proportionality (given non-singularity of a matrix given below), both expectation and variance of the target variables are matrix-proportional to the auxiliary variables yielding a regression model under which a Best Linear Unbiased Estimator (BLUE) is available for $\vec{\mu}_T$.

Let U denote the universe of population units.

The number of atoms in unit k, n_k, is a random variable but its value is not required to derive a model, only its existence. However, the numbers of atoms per unit are required for the estimation of the variance-covariance matrices in (2.0). The atoms that comprise a population unit are a random sample of random size from the set of all atoms and for any unit k, the $\{\omega_{ikj}\}_{j=1}^{n_k}$ may be considered independent and identically distributed for any fixed pair (i, k) and all $1 \le j \le n_k$. Let their common mean and variance be $\mu_{\mathrm{i}} = E(\omega_{ikj})$ and $\sigma_i^2 = Var(\omega_{ikj})$ for all i , k , and j. Let $\mu_{\mathrm{n}} = E(n_k)$ and $\sigma_n^2 = Var(n_k)$ for all k, and let the $\{n_k\}$ and $\{\omega_{ikj}\}_{j=1}^{n_k}$ be uncorrelated for any fixed pair (i, k) and all $1 \le j \le n_k$.

Suppose each sample unit, k, is composed of m_A types of atoms: type 1, type 2, …….. , type m_A. Let n_{kr} be the number of type r atoms for each r=1,2,…… m_A. Then $n_k = \sum_{r=1}^{m_A} n_{kr}$. Assume that for each r, the n_{kr} type r atoms are a random sample from all type r atoms in the population/stratum. Assume the study variables attached to each type r atom have a common $(m_A + m_T) \times 1$, mean vector, $\vec{\mu}_r = \begin{pmatrix} \vec{\mu}_{Ar} \\ \vec{\mu}_{Tr} \end{pmatrix}$. The stochastic structure given above can describe units as containers of water drawn from a stream, and atoms as stream particulate. For this case, the above assumptions seem appropriate. We then have: $y_{ik} = \sum_{r=1}^{m_A} \sum_{j=1}^{n_{kr}} \omega_{ikrj}$ where ω_{ikrj} denotes the value of the j^{th} atom of type r for the i^{th} study variable of the k^{th} sample unit. The population totals for the m_A types of atoms are known for all the m_A auxiliary variables. Given the randomization in the selection of atoms and the very large population size of atoms, it is appropriate to assume the $\{\omega_{ikrj}\}_{r=1,2,\ldots\ldots m_A \& j=1,2\ldots.,n_{kr}}$ for sample unit k are uncorrelated with one another for each k in the sample.

The preceding implies, that conditional on the sample unit's realized values for the $\{n_{kr}\}_{r=1}^{m_A}$:

$E(y_{ik} | \{n_{kr}\}_{r=1}^{m_A}) = \sum_{r=1}^{m_A} \mu_{ir} n_{kr}$, where $\mu_{ir} = E(\omega_{ikrj})$ for all j and k in group r.

Let: $Cov(y_{ik}, y_{lk} | \{n_{kr}\}_{r=1}^{m_A}) = \sum_{r=1}^{m_A} n_{kr} C_{ilr}$ (2.1)

where C_{ilr} is the covariance between the i^{th} and l^{th} study variables attached to an atom of type r. Assume the covariance between the i^{th} study variable of type r and the l^{th} study variable of type r' are zero for all i, l, and $r' \ne r$. All co-variances between study variables of different atoms are zero and all between unit co-variances are also zero. From (2.1) the vector of auxiliary variables, A_k, is:

$$A_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{m_A k} \end{pmatrix} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1m_A} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2m_A} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{m_A 1} & \mu_{m_A 2} & \cdots & \mu_{m_A m_A} \end{pmatrix} \begin{pmatrix} n_{k1} \\ n_{k2} \\ \vdots \\ n_{km_A} \end{pmatrix} + \begin{pmatrix} \epsilon_{1k} \\ \epsilon_{2k} \\ \vdots \\ \epsilon_{m_A k} \end{pmatrix} \tag{2.2}$$

And $\vec{\epsilon}_{Ak} = \begin{pmatrix} \epsilon_{1k} \\ \epsilon_{2k} \\ \vdots \\ \epsilon_{m_A k} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} , \Sigma_A \right)$ where $\Sigma_A = \sum_{r=1}^{m_A} n_{kr} \Sigma_{Ar}$ and the matrix of the (C_{ilr}), $1 \le i \le m_A$ and $1 \le j \le m_A$ is Σ_{Ar} , the covariance matrix for the group r auxiliary variables.

The $\{\Sigma_{Ar}\}_{r=1}^{m_A}$ involve none of the $\{n_{kr}\}_{r=1}^{m_A}$. Let $N_k = \begin{pmatrix} n_{k1} \\ n_{k2} \\ \vdots \\ n_{km_A} \end{pmatrix}$ and

$M_A = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1m_A} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2m_A} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{m_A 1} & \mu_{m_A 2} & \cdots & \mu_{m_A m_A} \end{pmatrix}$, then (2.2) can be expressed as:

$$A_k = M_A N_k + \vec{\epsilon}_{Ak} .$$

A similar equation holds for the m_T-vector of target variables: $T_k = M_T N_k + \vec{\epsilon}_{Tk}$

Where M_T is the m_TXm_A matrix of means analogous to M_A and $\vec{\epsilon}_{Tk}$ is the m_T-vector of error terms for the target variables.

The sample data for unit k, given the $\{n_{kr}\}_{r=1}^{m_A}$ is summarized by:

$$\begin{pmatrix} A_k \\ T_k \end{pmatrix} = \begin{pmatrix} M_A \\ M_T \end{pmatrix} N_k + \begin{pmatrix} \vec{\epsilon}_{Ak} \\ \vec{\epsilon}_{Tk} \end{pmatrix} \text{ where } \begin{pmatrix} \vec{\epsilon}_{Ak} \\ \vec{\epsilon}_{Tk} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \sum_{r=1}^{m_A} n_{kr} \begin{pmatrix} \Sigma_{Ar} & \Sigma_{ATr} \\ \Sigma_{TAr} & \Sigma_{Tr} \end{pmatrix} \right) \tag{2.3}$$

Where the $\{\Sigma_{Ar}, \Sigma_{Tr}, \Sigma_{ATr}\}_{r=1}^{m_A}$ are not functions of the $\{n_{kr}\}_{r=1}^{m_A}$. Note that in (2.2) above $\Sigma_A = \sum_{r=1}^{m_A} n_{kr}\Sigma_{Ar}$ and similarly for Σ_T and Σ_{AT} .

When there are m_A auxiliary variables, m_A types of atoms, and M_A is nonsingular, then the model given by (2.3) above can be transformed into one in which the target variables are matrix-proportional to the auxiliary variables as follows.

A_k=$M_A N_k$+ $\vec{\epsilon}_{Ak}$ can be rewritten as: $N_k = M_A^{-1}(A_k - \vec{\epsilon}_{Ak})$ = $M_A^{-1}A_k$ - $M_A^{-1}\vec{\epsilon}_{Ak}$

or $$N_k = M_A^{-1}A_k + \vec{\epsilon}^{\,\circ}_{Ak} \quad \text{where } \vec{\epsilon}^{\,\circ}_{Ak} = -\, M_A^{-1}\vec{\epsilon}_{Ak}, \;\; \vec{\epsilon}^{\,\circ}_{Ak} \sim (0, \quad M_A^{-1}\Sigma_A(M_A^{-1})') \tag{2.4}$$

Thus $T_k = M_T(M_A^{-1}A_k + \vec{\epsilon}^{\,\circ}_{Ak}) + \vec{\epsilon}_{Tk}$ = $M_T M_A^{-1}A_k + (M_T\vec{\epsilon}^{\,\circ}_{Ak} + \vec{\epsilon}_{Tk})$. Let B= $M_T M_A^{-1}$, then

$$T_k = BA_k + \delta_k \qquad \text{for k=1,2,……..,n.} \tag{2.5}$$

Where $\delta_k = (M_T\vec{\epsilon}^{\,\circ}_{Ak} + \vec{\epsilon}_{Tk})$, $E(\delta_k)$=0, and the covariance matrix of δ_k is

$\Sigma_\delta = B\Sigma_A B' - B\Sigma_{AT} - \Sigma_{TA}B' + \Sigma_T$ for all k. Let $B = (b_{ij})$, an $m_T \times m_A$ matrix and let the transpose of its i^{th} row be $B_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{im_A} \end{pmatrix} for\ i = 1,2, \ldots\ldots, m_T$. Then (2.5) can be written as:

$$T_k = \begin{pmatrix} B_1' \\ \vdots \\ B_{m_T}' \end{pmatrix} \begin{pmatrix} a_{1k} \\ \vdots \\ a_{m_A k} \end{pmatrix} + \delta_k \quad \text{for k=1,2,……..,n} \tag{2.6}$$

$$= (I \otimes A_k') \begin{pmatrix} B_1 \\ \vdots \\ B_{m_T} \end{pmatrix} + \delta_k \tag{2.7}$$

for k=1,2,……..,n where I is the $m_T \times m_T$ identity matrix.

The linear relationship summarizing all the sample data for k=1, 2, …,n is:

$\begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{n-1} \\ T_n \end{pmatrix} = \begin{pmatrix} I\otimes A_1' \\ I\otimes A_2' \\ \vdots \\ I\otimes A_{n-1}' \\ I\otimes A_n' \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{m_T-1} \\ B_{m_T} \end{pmatrix} + \Delta$ where I is the $m_T \times m_T$ identity matrix, $\otimes$ is the Kronecker product, and Δ is the nm_T random column vector $\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}$ with expectation of 0 and covariance matrix $\mathcal{I}\otimes\Sigma_\delta$ where $\mathcal{I}$ is the $n \times n$

identity matrix. The Kronecker product of two matrices is defined as the matrix result of multiplying each component of the first matrix by the second matrix. The BLUE for $\beta = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{m_T-1} \\ B_{m_T} \end{pmatrix}$ is:

$$\hat{\beta} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \vdots \\ \hat{B}_{m_T-1} \\ \hat{B}_{m_T} \end{pmatrix} = \left(\sum_{k=1}^{n}(I\otimes A_k)\Sigma_\delta^{-1}\left(I\otimes A_k'\right)\right)^{-1}\sum_{k=1}^{n}(I\otimes A_k)\Sigma_\delta^{-1}T_k \ , \quad \text{Rao (1973).}$$

Substituting estimates for M_T , M_A, Σ_A , Σ_T and Σ_{AT} , into Σ_δ, $\hat{B}$ can be approximated directly from the sample data. Then the BLUE for the vector of target variable population totals and its co-variance matrix are:

$$\hat{T}_{TOT} = (I\otimes A)\hat{\beta} \quad \text{and} \quad \text{Var}(\hat{T}_{TOT})=(I\otimes A)\left(\sum_{k=1}^{n}(I\otimes A_k)\Sigma_\delta^{-1}\left(I\otimes A_k'\right)\right)^{-1}(I\otimes A)' \tag{2.8}$$

Where A $= \left(\sum_{k=1}^{N} a_{1k} \quad , \dots\dots\dots\dots\dots\dots , \sum_{k=1}^{N} a_{m_A k}\right)$, N is the number of universe units, and I is the $m_T \times m_T$ identity matrix.

The sampling and estimation methodology described here is applicable to flow sampling where a flow consists of a population in constant one-way movement past a point where it is sampled, much like sampling a population whose strata are rivers and streams over a time period (i.e. a month). This occurs for actual rivers and streams but also for other types of populations like parts moving through an assembly process, autos moving along roads to or from a city or country, or mail moving through processing centers. Auxiliary data may occur naturally or be a part of the design that is added to the flows upstream in known quantities to provide auxiliary data. For example, known quantities of a chemical marker may be added to a river or stream far upstream from the sampling point (to insure complete mixing). The setup presented in this section with m_A types of atoms and m_A auxiliary variables can model a flow consisting of m_A inputs from tributaries to the stream where total flow volume or weight is measured for each tributary to provide an auxiliary variable for each tributary.

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

The following study compares the combined ratio estimator (with probability based expansions) to the BLUE based on pre-sampling with a single auxiliary variable - the simplest case (with m_A=1) of the setup described in Section 2. A single auxiliary variable may be the case most experienced in sampling practice and simulates a government survey. Analysis of the estimators is done with respect to repeated sampling under a complex design (although the BLUE is based on a model). The results described here may seem extreme, but they are readily explained in the repeated sampling context. Design based procedures are inherently problematic for the population being measured due to several factors that amplify each other to produce large levels of sampling error in the Combined Ratio Estimator while leaving the BLUE with orders of magnitude less sampling error. Apparently survey design can be a more difficult endeavor than suggested by the orthodoxy of good survey practice. This is particularly true under administrative and operational constraints encountered in practice. The government agency providing this example is purposely left opaque.

In this example, the population consists of F strata (F≅150) 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 consists of a simple random pre-sample of atoms selected from all atoms making up the units in U_f. For example, a bucket of water taken from a stream (stratum) could be a sample unit and the particulate matter in the bucket's water, the atoms. It would be quite reasonable to assume that this particulate was an SRSWOR from all the particulate in the stream flowing past a fixed point over a brief time period. For sampling flows, time period and sampling location are the primary components of strata definition. In this example, estimates of the total number of atoms in a population consisting of the streams flowing into an inlet or lake during a month are required. The total weight in kilograms of water containing the universe units in U_f is known for all the strata and recorded for each sample unit. These kilogram measures are the auxiliary variables.

Each stratum, U_f, is partitioned into first stage clusters consisting of the stream flow past a fixed point during an hour of each day. Let M_f be the set of clusters in U_f and N_f =24 be the number clusters in M_f. Let s_f be an SRSWOR from M_f and let n_f be the number of clusters in s_f. Let U_{fd} be the set of second stage universe units in cluster d of stratum f for d=1,2,3,…., N_f. Let N_{fd} be the number of universe units in U_{fd} $\left(\sum_{d=1}^{N_f} N_{fd} = N_f\right)$. Let s_{fd} be an SRSWOR selected from the universe units in U_{fd} and let n_{fd} be the number of units in s_{fd}. For the survey being studied, workload constraints restrict both n_f and n_{fd} to being roughly 4, (between 3 and 5) for all f and d.

Let K_{fdj} be the weight in kilograms of the j^{th} unit in U_{fd} and let π_{fdj} be the probability of selection of the j^{th} unit in U_{fd}. Then $\pi_{fdj} = \frac{n_f}{N_f}\frac{n_{fd}}{N_{fd}}$. When referring to population units, upper case is used and for sample units, lower case is used - y_{fdj} is the value of the study variable for the j^{th} sample unit from s_{fd} and Y_{fdj} denotes the value of the study variable for the j^{th} population unit in U_{fd}. The Horwitz-Thompson Estimator (probability expansion) for total kilograms of the units in U_f is $\hat{k}_f = \sum_{d \epsilon s_f} \sum_{j \varepsilon s_{fd}} \frac{k_{fdj}}{\pi_{fdj}}$ where k_{fdj} and K_{fdj} are defined analogously to y_{fdj} and Y_{fdj}. Let the total kilograms in stratum f be known and denoted, $K_f = \sum_{d=1}^{N_f} \sum_{j=1}^{N_{fd}} K_{fdj}$, then $K_f = E(\hat{k}_f)$. Define $\hat{y}_f$ and Y_f similarly. Let $K = \sum_{f=1}^{F} K_f$ and similarly for Y. Let $\beta_f = \frac{Y_f}{K_f}$. The four estimators to be studied are:

$\hat{T}_{HT} = \hat{y}_f$ is the Horwitz-Thompson estimator for Y.

$\hat{T}_C = K \frac{\sum_{f=1}^{F} \hat{y}_f}{\sum_{f=1}^{F} \hat{k}_f}$ is the combined ratio estimator for Y.

$\hat{T}_S = K \sum_{f=1}^{F} W_f \hat{\beta}_f$ is the separate ratio estimator for Y where $W_f = \frac{K_f}{K}$, and $\hat{\beta}_f = \frac{\hat{y}_f}{\hat{k}_f}$.

$\hat{T}_B = K \sum_{f=1}^{F} W_f \hat{\hat{\beta}}_f$ is the BLUE for Y where $\hat{\hat{\beta}}_f = \frac{\bar{y}_f}{\bar{k}_f}$, $\bar{y}_f = \frac{\sum_{d \varepsilon s_f} \sum_{j \varepsilon s_{fd}} y_{fdj}}{\sum_{d \varepsilon s_f} n_{fd}}$, and $\bar{k}_f$ is defined similarly to $\bar{y}_f$.

These four estimators are studied under the sample design just described where the size of the first stage cluster sizes vary from nearly uniform – each cluster about the same size in units and weight – to quite diverse in size – clusters have widely different numbers of units and total weights. This analysis studies the variance of the four estimators as functions of a measure of first stage cluster size variability given by

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

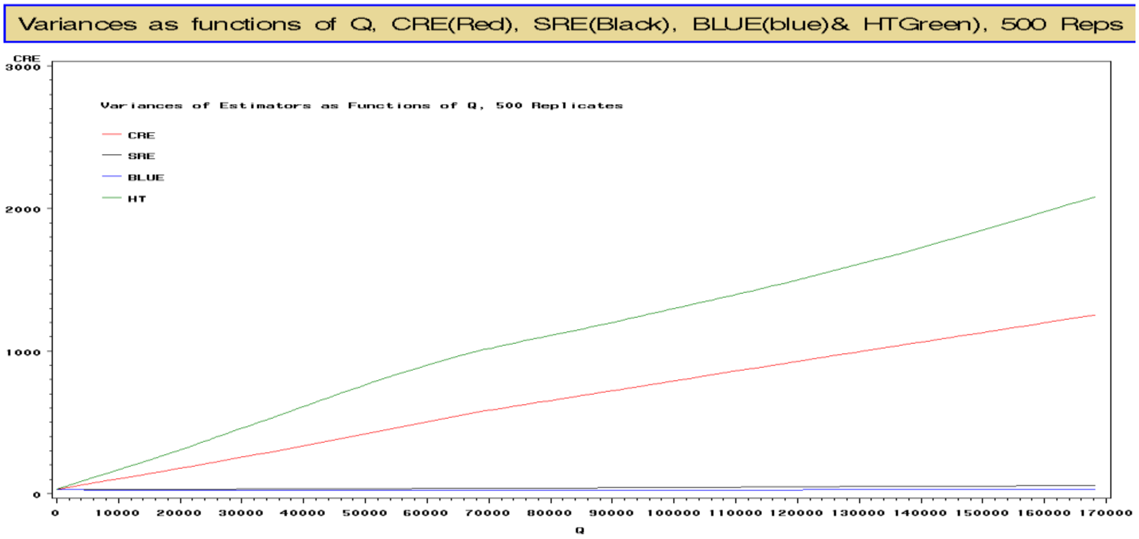
Q is an increasing function of the variance between first stage cluster sizes, that is, the average variability of the $\left\{N_{fd}\right\}_{d=1}^{N_f}$ for f=1,2,……,F over the F strata.

The population studied here consists of about 44 million atoms, about 700,000 units, and 150 strata. Within each stratum f, the ${y_{fdj}}\Big/{k_{fdj}}$ are relatively homogeneous and vary around β_f. The 150 different values of the $\left\{\beta_f\right\}$ vary from rough unity to more than 50. The study that follows summarizes variance of the four estimators for about 50 populations that are similar except for first stage cluster size variability within strata. That is, once the units have been randomly assigned atoms, the units within each stratum are randomly assigned to clusters. These assignments vary from nearly uniform (small Q) to widely variable (large Q).

The variance of each estimator can be either derived directly from the sample design and population parameters (cluster means, totals, and variances) or by brut force simulations – selecting 500 samples according to the above design and estimating means and variances of all four estimators from these 500 independent replicates. The direct derivation and the derivation based on 500 replications yield similar results, summarized below. For a detailed description of these two studies see Woodruff (2007). The direct derivation provides the variance components that explain the behavior of the four estimators as functions of Q and these formulae are also given in Woodruff (2007).

In Graph 1 below, CRE is $\hat{T}_C$, SRE is $\hat{T}_S$, BLUE is $\hat{T}_B$, and HT is $\hat{T}_{HT}$. Q is represented along the horizontal axis and variance of the four estimators (in units of 10^{10}) on the vertical axis. These four estimators are analyzed as functions of Q over a wide range because, in practice, Q varies a great deal over time in an unpredictable manner making a study based on a range of Q necessary.

Graph 1



Let $V(.)$ denote variance with respect to repeated sampling, then given the pre-sampling as SRSWOR the variance of the Combined Ratio Estimator can be expressed as:

$$V(\hat{T}_C) \cong K^2 V\left(\sum_{f=1}^{F} W_f \hat{\beta}_f\right) + K^2 V\left(\sum_{f=1}^{F} \beta_f \hat{W}_f\right), \text{ where } \hat{W}_f = {}^{\hat{k}_f}\Big/_{\sum_{m=1}^{F} \hat{k}_m} \tag{3.1}$$

The first term of $V(\hat{T}_C)$ is the variance of $\hat{T}_S$ but the $\{\hat{\beta}_f\}$ are ratios of correlated random variables and so this first term should be relatively small compared to the second term of $V(\hat{T}_C)$. The $\{\beta_f\}_{f=1}^{150}$ vary from about 1 to over 50 and this second term is the variance of a randomly weighted average of these $\{\beta_f\}_{f=1}^{150}$ (the weights, $\{\hat{W}_f\}$, are random variables) and each $\hat{W}_f$ is the ratio of nearly uncorrelated random variables and thus the variance of this second term can be substantial. Graph 1 demonstrates the dominance of the second term in $V(\hat{T}_C)$ – the second term of $V(\hat{T}_C)$ is the difference between the red and black lines and accounts for over 90% of $V(\hat{T}_C)$.

The sample design is not one that would be chosen if proper sample control could be exercised. In particular, it is not self-weighting. The ratio, H= ${}^{V(\hat{T}_C)}\Big/_{V(\hat{T}_B)}$, for values of Q experienced historically were in the interval (20,40). Although it is impractical to use a self-weighting design in the real survey, a self-weighting design was tried in the simulation by increasing the second stage sample sizes to achieve a self weighting design. With the self-weighting design, H went from the interval (20,40) to (80, 160). Thus H can be over 100 under what is considered the gold standard in survey design! Apparently, the roughly five fold increase in sample size to achieve a self weighting design results in a roughly five fold decrease in sampling variance in the BLUE, $V(\hat{T}_B)$, but only a 10%

to a 20% decrease in $V(\hat{T}_C)$. Sample allocation to achieve a self-weighting design places much sample in strata where it does little to reduce $V(\hat{T}_C)$ and starves additional sample from strata where it would most reduce $V(\hat{T}_C)$.

The model based estimation strategy derived from pre-sampling retains randomization as an essential feature but avoids the inefficiencies that plague the sample design and the Combined Ratio Estimator. The result is a model based estimation methodology under a model that is imposed by statistical design and is largely immune from the usual weakness of model based inference – the question of model suitability.

4. Conclusions

Pre-sampling uses designed randomization to force a model on sample data under which there is a BLUE. Pre-sampling is not always possible but when it can be applied or when it occurs naturally, the BLUE derived from it can manifestly improve the precision of survey estimates. This is especially true when operational and administrative constraints force an inefficient sample design. Pre-sampling avoids questions about model fit by imposing the model deductively. Thus Pre-sampling methodologies retain the main advantages of both model based and design based inference while avoiding their most notable shortcomings.

Pre-sampling methodologies provide a more appropriate set of sample expansions that better capture the stochastic structure of sample data when there are many different population totals to be estimated. These methodologies replace the "one size fits all" expansions based on probabilities of selection. The pre-sampling BLUE is based on a multi-parameter model that provides considerable flexibility for a better fit of estimator to the parameter being estimated. Since the model is deduced from designed randomization it is not dependent upon sample data or historical data that could be anomalous or otherwise poorly reflect the process that generates the sample data.

Section 3 examines a real survey where accepted sample design based inference performs poorly and pre-sampling model based inference provides a gratifying improvement. This example provides a comparison of the pre-sampling model based BLUE to the Combined Ratio Estimator under a stratified cluster sample design in which first stage cluster size variability and the large difference in strata ratios of study variables to the auxiliary variable cause disturbingly large variance in the Combined Ratio Estimator. The variance of the pre-sampling BLUE remains small and unaffected by the cluster totals and strata ratios. This is a univariate application of Section two with a pre-sampling BLUE that achieves a 20 to 40 fold variance reduction compared to the Combined Ratio Estimator. It may be noted that the sample design in Section three is far from self-weighting for moderate to large values of Q. When the sample was increased to achieve a self-weighting design, the pre-sampling BLUE experienced an 80 to 160 fold variance reduction compared to the Combined Ratio Estimator!

Simple random pre-sampling as presented here can readily be extended to more complex pre-sampling designs. A two stage cluster pre-sampling design yields sample units that follow a regression model similar to that developed under simple random pre-sampling in section two.

The application in Section 3 of pre-sampling avoids the more challenging issues encountered in multivariate applications, estimation of the parameter matrices, Σ_A, Σ_T , Σ_{AT} , M_A, and M_T. The estimation of these parameters was omitted due to space limitations. Although their estimation is basic, it requires some calculation that may be informative and therefore deserves consideration. Further study using the full multivariate structure described in Section 2 is underway and will be the subject of future work.

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