Pre-Sampling Model Based Inference III

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

Abstract

In survey sampling, a sample unit`s study variables are expanded to population totals by probability design based (DB) or model based (MB) expansions that implicitly treat a unit`s study variables as totals over entities called atoms contained in a unit. Expansions would make little sense if applied to unit statistics other than atom totals, for example, industry surveys where units are businesses, atoms are employees, and unit study variables are establishment totals for hours, wages, and number of workers. Other examples are household, mail, and ecological surveys. Woodruff (2010, 2009), derived Pre-sampling Model Based Inference from this atom structure, a structure that depends on probability sampling of population units and of atoms that comprise each unit. It provides estimates that retain the best properties of both MB and DB inference and that eliminate the main shortcomings of each. The result can be order-of-magnitude error reduction. Sampling error under repeated sampling from stratified cluster designs is the basis for comparison of the Combined Ratio Estimator and the Pre-sampling Model Based Estimator. Formulae for sampling errors are derived and analyzed.

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

1. Introduction

Sampling inference should be based on randomization, should be multivariate since most surveys collect data on more than one study variable, should be robust against common difficulties in sampling applications and seek to minimize or avoid post sampling adjustments like outlier and non-response corrections. Model conjecture based on sample data or historical data is another potential source of error that should be eliminated or at least minimized. These are the goals of Pre-Sampling Model Based Inference (PSMB), a methodology that was partially developed in several papers, Woodruff (2007, 2008, 2009, 2010).

PSMB models are not conjectured from sample data or historical data but rather derived deductively from stochastic structure found in many populations. This structure imposes a model on study variables that is a direct consequence of randomization and this model in turn provides access to the powerful theorems on Best Linear Unbiased Estimation. The mathematics supporting this methodology is initiated in this paper.

PSMB procedures expand sampling theory from randomized sample selection to both randomized sample selection and randomized unit synthesis. By doing so, it provides insights to the interaction between the sample design and the stochastic properties of the study variables being measured by the sample survey. It also provides estimates of population totals that are often far closer to the mark than commonly used Design Based estimates.

PSMB combines the best features of both Design Based inference and Model Based inference while avoiding their main shortcomings. In this paper, estimation error is defined and analyzed with respect to repeated sampling under stratified cluster sampling designs. The repeated sampling variance of the BLUE derived under Pre-sampling models can be orders of magnitude smaller than the repeated sampling variance of common design based alternatives. Repeated sampling variance expressions for several PSMB estimators and the standard Design Based estimators: Horwitz-Thompson, Combined Ratio Estimator, and Separate Ratio Estimator are derived within the expanded context of both randomized unit selection and randomized unit synthesis.

In survey sampling, a sample unit`s study variables are expanded to population totals. These expansions implicitly treat a unit`s study variables as totals over entities contained in each population unit. These entities are called "atoms" in PSMB theory. Neither Model Based nor Design Based expansions make much sense if applied to unit study variables other than these atom totals.

For example, 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 (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 unit totals are expanded to the population to provide estimates of population totals for these items. Woodruff (2010, 2009), derived PSMB Inference from this atom structure within population units, a structure that applies the randomized assignment of atoms to units in the population or a close approximation to randomized assignment.

The atom structure within the population units also helps identify sample design problems that can be avoided or at least reduced by looking at sample design through the lens of the theorems below. They highlight designs to avoid in Design Based inference and characteristics of population study variables that particularly aggravate these design problems. They make explicit the characteristics of population study variables that are implicitly assumed (and unstated) in most sampling texts. The theorems below deal with univariate data structures and are an initial effort to explain mathematically the simulation results contained in previous papers (referenced above) which provide mostly multivariate results.

This paper makes a modest beginning in expanding sampling theory from randomized sample selection to both randomized sample selection and randomized unit synthesis. These techniques express sampling error in terms of both the sample design and the stochastic structure of the study variables. This leads to improved inference by exposing situations where a particular sample design can be inefficient and suggests ways to improve inference.

2. Structures, Techniques, and Theorems

2.1 The Atom Structure for Population Units

A two component Atom Population Model (APM) is used to describe atom study variables and through them, the unit study variables. Let Y_{ijl} denote the value of the study variable for atom l of unit j in cluster i. Since there are two quite different random

processes considered here, a subscript A will distinguish expectation and variance with respect to the Atom Population Model (APM) from expectation and variance with respect to the sample design given in Section 2.2 below.

Definition 2.1.1 The notation $Y_{ijl} \sim (\mu, \sigma^2)$ means that the expected value of Y_{ijl} with respect to the APM, denoted $E_A(Y_{ijl})$, is μ and its variance, denoted $Var_A(Y_{ijl})$, is σ^2.

The atom population model (APM) is given by (2.1.1) and (2.1.2):

The $\{Y_{ijl}\}$ are iid, $Y_{ijl} \sim (\mu, \sigma^2)$, and their support is positive. (2.1.1)

The number of atoms in a unit is also a random variable and this number is denoted A_{ij} for unit j in cluster i.

The $\{A_{ij}\}$ are iid, $A_{ij} \sim (\delta, \alpha^2)$, and the support of the $\{A_{ij}\}$ is contained in the positive integers. (2.1.2)

The unit Y-variable for population unit j in cluster i is: $Y_{ij} = \sum_{l=1}^{A_{ij}} Y_{ijl}$. (2.1.3)

From this APM, the mean and variance of $Y_{ij} = \sum_{l=1}^{A_{ij}} Y_{ijl}$ are:

$E_A(Y_{ij}) = \mu\delta$ and $Var_A(Y_{ij}) = \delta\sigma^2 + \mu^2\alpha^2$ or $Y_{ij} \sim (\mu\delta\ ,\ \delta\sigma^2 + \mu^2\alpha^2)$, with the $\{Y_{ij}\}$ independent for all i and j.

Conditional on the number of atoms in a unit, the following unit population model is a direct consequence of (2.1.1):

$Y_{ij} = \mu A_{ij} + \epsilon_{ij}$ where $\epsilon_{ij} \sim (0, A_{ij}\sigma^2)$ and the $\{\epsilon_{ij}\}$ are independent. (2.1.4)

The APM (2.1.1) presumes little about the set of atoms $\{Y_{ijl}\}$ - stochastic independence and the existence of their mean, $\bar{Y}$ and variance, S_Y^2. This structure is found in many populations and although minimal, this APM allows great unit-to-unit variation in unit study variables according to the number of atoms in a unit. This model is generalized in Woodruff (2009, 2010) to several atom types within each unit which permits yet greater unit heterogeneity according the distribution of the different atoms types within a unit. The APM imposes a population model at the unit level that can be complex enough to substantially capture unit behaviour via these minimal APM assumptions. Model conjecture and potential model failure are lesser concerns.

$\hat{Y}_{PS}$ [from (2.3.2)] is the Best Linear Unbiased Estimator (BLUE) under model (2.1.4). Since (2.1.4) holds for all study variables (each with its own unique constants μ and σ^2) it follows immediately that if X_{ij} is another study variable that is also an auxiliary variable (population total of the $\{X_{ij}\}$ known), then there exist constants θ and γ^2 such that:

$Y_{ij} = \theta X_{ij} + \varphi_{ij}$ where $\varphi_{ij} \sim (0, A_{ij}\gamma^2)$ and the $\{\varphi_{ij}\}$ are independent. (2.1.5)

In fact, $\theta = \frac{E_A(Y_{ij})}{E_A(X_{ij})}$ and $\gamma^2 = \sigma_Y^2 + \theta^2\sigma_X^2 - 2\theta\sigma_{YX}$ where σ_Y^2 is the σ^2 in (2.1.1) for the Y-variate, σ_X^2 is the σ^2 for the X-variate, and σ_{YX} is their atom level covariance (See Section 3). Since all study variables are related to the $\{A_{ij}\}$ by (2.1.4) with their own particular constants, (μ, σ^2), all study variables are related to one another by (2.1.5) with their own unique constants, θ and γ^2, for each different pair. Thus, in particular, all study variables are related to the auxiliary variable, X, by (2.1.5) and a PSMB BLUE is an immediate consequence.

Given (2.1.5), this PSMB BLUE for the population total is $\hat{Y}_p$ (3.1.1), the estimator available when the $\{A_{ij}\}$ are not auxiliary variables (their population total is unknown) but the $\{X_{ij}\}$ are auxiliary variables. This estimator was used by the US Postal Service to estimate mail volumes (units were mail containers and atoms were their contents, mail pieces) and by the Bureau of Labor Statistics to estimate employment totals (units were business establishments and atoms were their employees). Both these agencies derived this estimator from the model (2.1.5) which was conjectured by data mining. The APM is an alternative to these model conjectures but with substantially less conjecture from potentially misleading historical or sample data.

2.2 Sample Design and Sample Design Parameters

The population consists of 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 and let U be the set of all the N units in the population so that $U = \bigcup_{i=1}^{K} U_i$.

Let S denote an SRSWOR (simple random sample without replacement) of size k from the K clusters.

Let N_i denote the number of units in U_i, and S_i be an SRSWOR of size n_i from the N_i units in U_i.

2.3 Estimators and their Sampling Errors

The Horwitz-Thompson estimator for the population total of the $\{Y_{ij}\}$ under the design in Section 2.2 is denoted $\hat{Y}_{HT}$ and is:

$$\hat{Y}_{HT} = \frac{K}{k}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{j\epsilon s_i} Y_{ij} \qquad (2.3.1)$$

Let $A=\sum_{i=1}^{K}\sum_{j=1}^{N_i} A_{ij}$, the population`s total number of atoms and $\check{A}=\sum_{i\epsilon s}\sum_{j\epsilon s_i} A_{ij}$, the total number of atoms in the sample.

The Pre-Sampling Model Based estimator (PSMB) for the finite population total of the study variables is $A\hat{\mu}$ where $\hat{\mu}$ is the BLUE for μ under (2.1.1) and (2.1.2). Denote this Pre-Sampling Model Based estimator $\hat{Y}_{PS}$, then

$$\hat{Y}_{PS} = \frac{A}{\check{A}}\sum_{i\epsilon s}\sum_{j\epsilon s_i} Y_{ij} = A\frac{\check{Y}}{\check{A}} \quad \text{where } \check{Y}=\sum_{i\epsilon s}\sum_{j\epsilon s_i} Y_{ij}. \qquad (2.3.2)$$

$\hat{Y}_{PS}$ weights all sample atom data, $\{Y_{ijl}\}$ equally with the weight, $\frac{1}{\check{A}}$ .The expectation of $\check{A}$ under repeated sampling from the cluster sample design is:
$E(\check{A}) = \frac{k}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}\sum_{j=1}^{N_i}A_{ij}$.

Definition 2.3.1 Expectation, $E(.)$, without subscript and variance, $Var(.)$, without a subscript denote expectation and variance with respect to repeated sampling under the sample design in Section 2.2.

Mean squares that occur in the expressions for sampling variance derived under the design in Section 2.2 are approximated with their model expectations under (2.1.1) and (2.1.2). These expectations seem appropriate approximations for large cluster sizes since the mean squares converge to their expected values as the cluster sizes increase [Law of Large Numbers (LLN)].

By the Taylor Series linear approximation of the ratio, $\frac{\check{Y}}{\check{A}}$, around the expected values of numerator and denominator, the variance of this ratio can be approximated as:

$$Var\left(\frac{\check{Y}}{\check{A}}\right) \doteq \frac{1}{E^2(\check{A})}Var\left(\check{Y} - \frac{E(\check{Y})}{E(\check{A})}\check{A}\right) = \frac{1}{E^2(\check{A})}Var\left(\sum_{i\epsilon s}\sum_{j\epsilon s_i}\left(Y_{ij} - \frac{E(\check{Y})}{E(\check{A})}A_{ij}\right)\right).$$

Let $Z_{ij} = Y_{ij} - \frac{E(\check{Y})}{E(\check{A})}A_{ij}$, (the approximation $\frac{E(\check{Y})}{E(\check{A})} \doteq \mu$ is used below in the estimation of mean squares). Then $Var\left(\sum_{i\epsilon s}\sum_{j\epsilon s_i}\left(Y_{ij} - \frac{E(\check{Y})}{E(\check{A})}A_{ij}\right)\right) = Var\left(\sum_{i\epsilon s}\sum_{j\epsilon s_i}Z_{ij}\right)$ and writing this as the sum of the expected value of the conditional variance given the sample outcome S and the variance of the expected value likewise given S,

$$Var\left(\sum_{i\epsilon s}\sum_{j\epsilon s_i}Z_{ij}\right) = Var\left(\sum_{i\epsilon s}\frac{n_i}{N_i}\sum_{j=1}^{N_i}Z_{ij}\right) +$$
$$E\left(\sum_{i\epsilon s} n_i\left(1-\frac{n_i}{N_i}\right)\frac{1}{N_i-1}\sum_{j=1}^{N_i}\left(Z_{ij}-\bar{Z}_i\right)^2\right)$$
$$= k\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{i=1}^{K}\left(\frac{n_i}{N_i}\sum_{j=1}^{N_i}Z_{ij} - \frac{1}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}\sum_{j=1}^{N_i}Z_{ij}\right)^2 + \frac{k}{K}\sum_{i=1}^{K}n_i\left(1-\frac{n_i}{N_i}\right)S_{Zi}^2$$

where $\bar{Z}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}Z_{ij}$, and $S_{Zi}^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}\left(Z_{ij}-\bar{Z}_i\right)^2$.

$$\text{Var}(\hat{Y}_{PS}) = \text{Var}\left(A\frac{\check{Y}}{\check{A}}\right) = A^2 Var\left(\frac{\check{Y}}{\check{A}}\right) =$$
$$\frac{A^2}{\left(\frac{k}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}\sum_{j=1}^{N_i}A_{ij}\right)^2}\left(k\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{i=1}^{K}\left(\frac{n_i}{N_i}\sum_{j=1}^{N_i}Z_{ij} - \frac{1}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}\sum_{j=1}^{N_i}Z_{ij}\right)^2 + \right.$$
$$\left.\frac{k}{K}\sum_{i=1}^{K}n_i\left(1-\frac{n_i}{N_i}\right)S_{Zi}^2\right).$$

Similarly (found in most sampling texts) the variance of the Horwitz-Thompson estimator under the clustered design in Section 2.2 is:

Var($\hat{Y}_{HT}$)=

$$\frac{K}{k}\sum_{i=1}^{K} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{1}{n_i}S_{Yi}^2 + \frac{K^2}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{i=1}^{K}\left(\sum_{j=1}^{N_i} Y_{ij} - \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{N_i} Y_{ij}\right)^2$$

where $S_{Yi}^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}\left(Y_{ij}-\bar{Y}_i\right)^2$ and $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$.

Let $Q_i = \frac{n_i}{N_i}\sum_{j=1}^{N_i} Z_{ij}$, $\bar{Q}=\frac{1}{K}\sum_{i=1}^{K} Q_i$, $Q_i^{'} = \sum_{j=1}^{N_i} Y_{ij}$, $\bar{Q}^{'} = \frac{1}{K}\sum_{i=1}^{K} Q_i^{'}$,

$S_Q^2 = \frac{1}{K-1}\sum_{i=1}^{K}(Q_i-\bar{Q})^2$, and $S_{Q^{'}}^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left(Q_i^{'}-\bar{Q}^{'}\right)^2$

then writing Var($\hat{Y}_{PS}$) and Var($\hat{Y}_{HT}$) in terms of these $\{Q_i\}$ and $\{Q_i^{'}\}$,

$$\text{Var}(\hat{Y}_{PS}) = \frac{A^2}{\left(\frac{k}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}\sum_{j=1}^{N_i} A_{ij}\right)^2}\left(k\left(1-\frac{k}{K}\right)S_Q^2 + \frac{k}{K}\sum_{i=1}^{K} n_i\left(1-\frac{n_i}{N_i}\right)S_{Zi}^2\right) \qquad (2.3.3)$$

$$\text{Var}(\hat{Y}_{HT}) = \frac{K^2}{k}\left(1-\frac{k}{K}\right)S_{Q^{'}}^2 + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{1}{n_i}S_{Yi}^2 \qquad (2.3.4)$$

Var($\hat{Y}_{HT}$) and Var($\hat{Y}_{PS}$) each consist of two parts, one part is sample design components (number of clusters, cluster sizes, number of sample clusters, and cluster sample sizes) and the other part is means squares (the $S^2-Terms$). The Law of Large Numbers (LLN) implies that the mean square terms converge to their expectations under the APM in section 2.1 as cluster sizes increase. This suggests these mean square terms can be approximated with their expectations under the APM (2.1.1) and (2.1.2). These APM expectations and approximations are as follows:

$S_{Zi}^2 \doteq \sigma^2\delta$, from: $E_A(S_{Zi}^2) = \sigma^2\delta$ where $E_A(.)$ is defined in Section 2.1.

$S_{Yi}^2 \doteq \sigma^2\delta + \mu^2\alpha^2$, from $E_A(S_{Yi}^2) = \sigma^2\delta + \mu^2\alpha^2$,

$S_Q^2 \doteq \frac{\sigma^2\delta}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}$, from $E_A\left(S_Q^2\right) \doteq \frac{\sigma^2\delta}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}$,

$S_{Q^{'}}^2 \doteq \mu^2\delta^2\frac{1}{(K-1)}\sum_{i=1}^{K}(N_i-\bar{N})^2 + (\mu^2\alpha^2+\sigma^2\delta)\bar{N}$ from

$E_A\left(S_{Q^{'}}^2\right) = \mu^2\delta^2\frac{1}{(K-1)}\sum_{i=1}^{K}(N_i-\bar{N})^2 + (\mu^2\alpha^2+\sigma^2\delta)\bar{N}$.

Then Var($\hat{Y}_{HT}$) and Var($\hat{Y}_{PS}$) with respect to repeated sampling can be approximated:

$$Var\left(\hat{Y}_{PS}\right) \doteq N^2\frac{\delta\sigma^2}{k\bar{n}^2K}\left(K\bar{n} - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right) = \frac{KN^2}{kn^2}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right) \text{ where}$$

$\bar{n} = \frac{1}{K}\sum_{i=1}^{K} n_i$ and $n = K\bar{n}$ (2.3.5)

Note: $\bar{n}$ is the average of the cluster sample sizes assigned to each and every cluster (whether selected or not).

$$\text{Var}(\hat{Y}_{HT}) \doteq \left(\frac{K^2}{k}-K\right)\mu^2\delta^2S_N^2 + (\sigma^2\delta+\mu^2\alpha^2)\left(\frac{K}{k}M-N\right) \qquad (2.3.6)$$

where $N = \sum_{i=1}^{K} N_i$, $\bar{N} = \frac{1}{K}\sum_{i=1}^{K} N_i$, $S_N^2 = \frac{1}{K-1}\sum_{i=1}^{K}(N_i - \bar{N})^2$, and M= $\sum_{i=1}^{K} \frac{N_i^2}{n_i}$.

$k\bar{n} = E(\sum_{i \in S} n_i)$, the expected sample size in units where expectation is with respect to repeated sampling under the cluster design in Section 2.2. When the cluster sample sizes in units are proportional to the cluster sizes in units, $\{n_i = CN_i \; for \; all \; i \; where \; C \; is \; a \; constant\}$, then:

$$\frac{Var(\hat{Y}_{HT})}{Var(\hat{Y}_{PS})} = 1 + \frac{\mu^2\alpha^2}{\delta\sigma^2} + \frac{\mu^2}{\sigma^2}(\delta k\bar{n})\frac{S_N^2}{N^2}\frac{\left(1-\frac{k}{K}\right)}{k\left(1-\frac{k\bar{n}}{N}\right)}.$$

In particular, the condition $\{n_i = CN_i \; for \; all \; i\}$ implies the design is self-weighting and under a self weighting design:

$$\frac{Var(\hat{Y}_{HT})}{Var(\hat{Y}_{PS})} > \frac{\mu^2}{\sigma^2}(\delta k\bar{n})\frac{S_N^2}{N^2}\frac{\left(1-\frac{k}{K}\right)}{k\left(1-\frac{k\bar{n}}{N}\right)} \qquad > \frac{\mu^2}{\sigma^2}(\delta k\bar{n})\frac{S_N^2}{N^2}\frac{\left(1-\frac{k}{K}\right)}{k} \qquad (2.3.7)$$

$\delta k\bar{n}$ is the expected sample size in atoms, a number that can be in the hundreds, thousands, or greater. Thus if S_N^2 or μ are not zero or near zero, this ratio can be quite large. Simulation results in Woodruff (2009), show the ratio, $\frac{Var(\hat{Y}_{HT})}{Var(\hat{Y}_{PS})}$, can easily be in the hundreds or greater.

When all the $\{N_i\}$ are identical, all the $\{n_i\}$ are identical, and all the $\{A_{ij}\}$ are identical, then $\hat{Y}_{HT} = \hat{Y}_{PS}$ and when these conditions are inserted into (2.3.5) and (2.3.6), $Var(\hat{Y}_{HT}) = Var(\hat{Y}_{PS})$.

Substituting the model expectations under (2.1.1) and (2.1.2) of S_{Zi}^2, S_{Yi}^2, S_Q^2 , and S_Q^2 in (2.3.5) and (2.3.6) expresses repeated sampling error in terms of both sample design parameters and stochastic properties of the study variable(s) under the APM and helps determine when a sample design may be problematic for a particular study variable.

Inequalities relating $Var(\hat{Y}_{HT})$ and $Var(\hat{Y}_{PS})$ need the following lemmas. Let n_s be a proposed total sample size in units determined by administrative considerations. The expected total sample size under the design 2.2 is $k\bar{n}$ and thus the reasonable constraint is that $k\bar{n} = n_s$ and this is used in the lemmas below.

Lemma 2.1 For a fixed set of positive integers $\{N_i\}_{i=1}^{K}$, the set of positive integers $\{n_i\}_{i=1}^{K}$ that minimizes $\sum_{i=1}^{K}\frac{n_i^2}{N_i}$ subject to the constraint that $k\bar{n} = n_s$, is $n_i = \frac{Kn_s}{kN}N_i$ $for \; all \; 1 \le i \le K$ and the value of $\sum_{i=1}^{K}\frac{n_i^2}{N_i}$ at these $\{n_i\}$ is $\frac{K^2\bar{n}^2}{N}$.

Proof: Use the method of Lagrange multipliers to solve the system of K+1 linear equations for the $\{n_l\}$:
$0 = \frac{\partial L}{\partial n_l} = \frac{2n_l}{N_l} - \lambda\frac{k}{K}$ for $1 \le l \le K$ and $0 = \frac{\partial L}{\partial \lambda} = -(k\bar{n} - n_s)$

where $L = \sum_{i=1}^{K} \frac{n_i^2}{N_i} - \lambda(k\bar{n} - n_s)$.

The solution is as stated in the Lemma 2.1. Q.E.D.

Lemma 2.2 For a fixed set of positive integers $\{N_i\}_{i=1}^{K}$, the set of positive integers $\{n_i\}_{i=1}^{K}$ that minimizes $\sum_{i=1}^{K} \frac{N_i^2}{n_i}$ subject to the constraint $k\bar{n} = n_s$ is $n_i = \frac{Kn_s}{kN} N_i$ $for\ all\ 1 \le i \le K$ and the value of $\sum_{i=1}^{K} \frac{N_i^2}{n_i}$ at these $\{n_i\}$ is $\frac{kN^2}{Kn_s}$.

Proof: Similar to Lemma 2.1 proof. Q.E.D.

Theorem 2.3 $Var(\hat{Y}_{HT}) \ge Var(\hat{Y}_{PS}) + \left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2$ under the cluster design described in Section 2.2 and where $k\bar{n} = n_s$.

Proof: From (2.3.6)

$$Var(\hat{Y}_{HT}) \doteq \left(\frac{K}{k}M - N\right)\sigma^2\delta + \left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2$$

and by Lemma 2.2, $M \ge \frac{K}{k}\frac{N^2}{n_s}$ so that ,

$$Var(\hat{Y}_{HT}) \ge \left(\frac{K}{k}\frac{kN^2}{Kn_s} - N\right)\sigma^2\delta + \left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2$$
$$= \left(\frac{N^2}{k\bar{n}} - N\right)\sigma^2\delta + \left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2 \quad (2.3.8)$$

$$Var(\hat{Y}_{PS}) \doteq \frac{K}{k}\frac{N^2}{n^2}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\delta\sigma^2$$

and by Lemma 2.1, $\sum_{i=1}^{K} \frac{n_i^2}{N_i} \ge \frac{K^2\bar{n}^2}{N}$ so that

$$Var(\hat{Y}_{PS}) \le \frac{K}{k}\frac{N^2}{n^2}\left(n - \frac{k}{K}\frac{K^2\bar{n}^2}{N}\right)\delta\sigma^2 = \left(\frac{N^2}{k\bar{n}} - N\right)\sigma^2\delta \quad (2.3.9)$$

Combining these two inequalities, (2.3.8) and (2.3.9):

$$Var(\hat{Y}_{HT}) \ge Var(\hat{Y}_{PS}) + \left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2 \quad (2.3.10)$$

Q.E.D.

Theorem 2.1 states that $\hat{Y}_{PS}$, a purely model based estimator, has considerably smaller repeated sampling variance than the Horwitz-Thompson estimator. This proof also provides a lower bound for the difference between $Var(\hat{Y}_{HT})$ and $Var(\hat{Y}_{PS})$, a difference that can apparently be quite large as observed in numerous simulations and by inspection of:

$$\left(\frac{K}{k}M - N\right)\mu^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2 .$$

The proof of Theorem 2.3 implies that (2.3.7), $\frac{Var(\hat{Y}_{HT})}{Var(\hat{Y}_{PS})} > \frac{\mu^2}{\sigma^2}(\delta k\bar{n})\frac{S_N^2}{N^2}\frac{\left(1-\frac{k}{K}\right)}{k}$ which was derived under a self-weighting design, is true in general for all designs given by Section 2.2.

(2.3.10) clarifies the roll of cluster size variation in the variance of the Horwitz-Thompson estimator ($\hat{Y}_{HT}$). When S_N^2 increases, the variance of $\hat{Y}_{HT}$ increases while $Var(\hat{Y}_{PS})$ is unaffected. This situation occurs in mail sampling where cluster sizes are unknown in advance, highly variable, and cluster sample sizes are based on available resources which are roughly constant.

In Theorem 2.3, unit atom counts are used as auxiliary variables. These will seldom be available for units not in the sample and therefore seldom available as auxiliary variables. This still makes a useful introduction to the sampling properties of the $\hat{Y}_{PS}$ compared to design based alternatives and possibly provides a more appropriate definition of design effect.

$\hat{Y}_{HT}$ will next be compared to the ratio of two Horwitz-Thomson estimators using the A-variate as auxiliary variable. Denote and define this estimator as:

$$\hat{Y}_{R(\frac{Y}{A})} = A\frac{\hat{Y}_{HT}}{\hat{A}_{HT}}, \qquad (2.3.11)$$

where $\hat{Y}_{HT}=\frac{K}{k}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{j\epsilon s_i}Y_{ij}$, $\hat{A}_{HT}=\frac{K}{k}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{j\epsilon s_i}A_{ij}$, and $A=\sum_{i=1}^{K}\sum_{j=1}^{N_i}A_{ij}$.

Approximating $\hat{Y}_{R(\frac{Y}{A})}$ with a Taylor series about the expected values (under repeated sampling) of numerator and denominator, the variance of $\hat{Y}_{R(\frac{Y}{A})}$ is approximately:

$$Var\left(\hat{Y}_{R(\frac{Y}{A})}\right) = \frac{K}{k}\sum_{i=1}^{K}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S_{T_i}^2 + \frac{K^2}{k}\left(1-\frac{k}{K}\right)S_{Q''}^2$$

where $T_{ij} = Y_{ij} - RA_{ij}$, $R = E(\hat{Y}_{HT})\Big/E(\hat{A}_{HT})$, $S_{Ti}^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(T_{ij}-\bar{T}_i)^2$

$\bar{T}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}T_{ij}$, and $S_{Q''}^2 = \frac{1}{K-1}\sum_{i=1}^{K}(Q_i''-\bar{Q}'')^2$ where $Q_i'' = \sum_{j=1}^{N_i}T_{ij}$ and $\bar{Q}''=\frac{1}{K}\sum_{i=1}^{K}Q_i''$.

As above, these quantities can be approximated with their expectations under (2.1), and (2.2) : R $\doteq \mu$, from $E_A(R)=\mu$. $S_{Q''}^2 \doteq \delta\sigma^2\bar{N}$, from $E_A\left(S_{Q''}^2\right)=\delta\sigma^2\bar{N}$, and $S_{T_i}^2 \doteq \delta\sigma^2$, from $E_A\left(S_{T_i}^2\right)=\delta\sigma^2$.

Then:

$$Var\left(\hat{Y}_{R(\frac{Y}{A})}\right) \doteq \delta\sigma^2\left(\frac{K}{k}M-N\right) \text{ and} \qquad (2.3.12)$$

$$Var(\hat{Y}_{HT}) = \delta\sigma^2\left(\frac{K}{k}M-N\right)+\left(\frac{K^2}{k}-K\right)\mu^2\delta^2S_N^2+\mu^2\alpha^2\left(\frac{K}{k}M-N\right)$$

$$= Var\left(\hat{Y}_{R\left(\frac{Y}{A}\right)}\right) + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2 + \mu^2\alpha^2\left(\frac{K}{k}M - N\right) \quad (2.3.13)$$

Since the last two terms in (2.3.13) are positive, the variance of $\hat{Y}_{HT}$ is greater than the variance of $\hat{Y}_{R\left(\frac{Y}{A}\right)}$ and (2.3.13) proves the following theorem.

Theorem 2.4 $Var\left(\hat{Y}_{HT}\right) = Var\left(\hat{Y}_{R\left(\frac{Y}{A}\right)}\right) + \left(\frac{K^2}{k} - K\right)\mu^2\delta^2 S_N^2 + \mu^2\alpha^2\left(\frac{K}{k}M - N\right)$ under the cluster design in Section 2.2 and the APM in Section 2.1.

$\frac{Var(\hat{Y}_{HT})}{Var\left(\hat{Y}_{R\left(\frac{Y}{A}\right)}\right)} > \frac{\mu^2}{\sigma^2}\left(\frac{1}{k} - \frac{1}{K}\right)k\bar{n}\delta\frac{S_N^2}{N^2}$ under a self-weighting design & in general. (2.3.14)

The $\{Q_i'\}$ are cluster totals of study variables and to the degree that these totals very from cluster to cluster, the variance of $\hat{Y}_{HT}$ increases. The analogous term for the $\hat{Y}_{PS}$ is the variance of the cluster means of the study variable which tend to be relatively stable (compared to cluster totals). This explains the increase in repeated sampling variance of $\hat{Y}_{HT}$ compared to $\hat{Y}_{PS}$ as cluster size variability increases; more on this is found in Woodruff (2010, 2009, and 2008).

3. More Than One Study Variable

3.1 Horwitz-Thompson Ratio Estimator

Consider the ratio estimator based on a more realistic auxiliary variable, a study variable that is not a unit`s atom count, but it`s atom sum of another study variable, X. Let X_{ijl} be the value of this study variable for atom l of unit j in cluster i, just as Y_{ijl}, was defined.

Independence of X_{ijl} and Y_{ijl} is no longer an appropriate assumption and a multivariate APM is more appropriate.

In what follows, the notation $W \sim [A, B]$ means the expectation and covariance matrix with respect to the APM of the vector valued random variable W, are A and B respectively. Then let:

$\begin{pmatrix} Y_{ijl} \\ X_{ijl} \end{pmatrix} \sim \left[\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{pmatrix}\right]$ and the $\left\{\begin{pmatrix} Y_{ijl} \\ X_{ijl} \end{pmatrix}\right\}$ are iid for all i, j, and l.

Defining the $\{Y_{ij}\}$ and $\{X_{ij}\}$ as the $\{Y_{ij}\}$ are given by (2.1.3) it follows that

$\begin{pmatrix} Y_{ij} \\ X_{ij} \end{pmatrix} \sim \left[\begin{pmatrix} \delta\mu_Y \\ \delta\mu_X \end{pmatrix}, \begin{pmatrix} \delta\sigma_Y^2 + \alpha^2\mu_Y^2 & \delta\sigma_{YX} + \alpha^2\mu_X\mu_Y \\ \delta\sigma_{YX} + \alpha^2\mu_X\mu_Y & \delta\sigma_X^2 + \alpha^2\mu_X^2 \end{pmatrix}\right]$ & these are iid for all i & j.

Letting $\Sigma_{YX} = \begin{pmatrix} \delta\sigma_Y^2 + \alpha^2\mu_Y^2 & \delta\sigma_{YX} + \alpha^2\mu_X\mu_Y \\ \delta\sigma_{YX} + \alpha^2\mu_X\mu_Y & \delta\sigma_X^2 + \alpha^2\mu_X^2 \end{pmatrix}$,

$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \Sigma_{j=1}^{N_i}\begin{pmatrix} Y_{ij} \\ X_{ij} \end{pmatrix} \sim \left[\begin{pmatrix} N_i\delta\mu_Y \\ N_i\delta\mu_X \end{pmatrix}, N_i\Sigma_{YX}\right]$, for all i and j.

$$\begin{pmatrix} Y \\ X \end{pmatrix} = \sum_{i=1}^{K} \begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim \left[\begin{pmatrix} N\delta\mu_Y \\ N\delta\mu_X \end{pmatrix}, N\Sigma_{YX} \right] \text{ \& } \begin{pmatrix} \bar{Y} \\ \bar{X} \end{pmatrix} = \frac{1}{K}\sum_{i=1}^{K} \begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim \left[\bar{N}\delta \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \frac{\bar{N}}{K}\Sigma_{YX} \right]$$

.

Thus it follows immediately that,

$E_A(\bar{Y}\bar{X}) = \frac{\bar{N}}{K}(\delta\sigma_{YX} + \alpha^2\mu_Y\mu_X) + \bar{N}^2\delta^2\mu_Y\mu_X$ and this implies,

$E_A(S_{YX}) = (\delta\sigma_{YX} + \alpha^2\mu_Y\mu_X)\bar{N} + \delta^2\mu_Y\mu_X S_N^2$

where $S_{YX} = \frac{1}{K-1}\sum_{i=1}^{K}(X_i - \bar{X})(Y_i - \bar{Y})$

From the above,

$\begin{pmatrix} \bar{Y}_i \\ \bar{X}_i \end{pmatrix} = \frac{1}{N_i}\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim \left[\begin{pmatrix} \delta\mu_Y \\ \delta\mu_X \end{pmatrix}, \frac{1}{N_i}\Sigma_{YX} \right]$ which implies that

$E_A(\bar{Y}_i\bar{X}_i) = \frac{1}{N_i}(\delta\sigma_{YX} + \alpha^2\mu_Y\mu_X) + \delta^2\mu_Y\mu_X$ and

$E_A(S_{Y_iX_i}) = \delta\sigma_{YX} + \alpha^2\mu_Y\mu_X$ where $S_{Y_iX_i} = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)$

The Horwitz-Thompson Estimator for the population total of the X-variate, its sampling variance, and its covariance with the Horwitz-Thompson Estimator for the Y-variate are:

$\hat{X}_{HT} = \frac{K}{k}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{j\epsilon s_i} X_{ij}$

$\text{Var}(\hat{X}_{HT}) \doteq \left(\frac{K^2}{k} - K\right)\mu_X^2\delta^2 S_N^2 + \left(\frac{K}{k}M - N\right)(\mu_X^2\alpha^2 + \sigma_X^2\delta).$

$Cov(\hat{Y}_{HT}, \hat{X}_{HT}) \doteq \left(\frac{K^2}{k} - K\right)\delta^2\mu_X\mu_Y S_N^2 + \left(\frac{K}{k}M - N\right)(\delta\sigma_{YX} + \alpha^2\mu_Y\mu_X).$

S_{Xi}^2, $S_{Q_X}^{2'}$, S_{YX}, and $S_{Y_iX_i}$ are defined analogously to S_{Yi}^2, $S_Q^{2'}$, S_{YA}, and $S_{Y_iA_i}$ respectively. The APM expectations that provide the APM approximations to S_{Xi}^2, $S_{Q_X}^{2'}$, S_{YX}, and $S_{Y_iX_i}$ are:

$E_A(S_{Xi}^2) = (\sigma_X^2\delta + \mu_X^2\alpha^2)$
$E_A\left(S_{Q_X}^{2'}\right) = (\mu_X^2\delta^2 S_N^2 + (\mu_X^2\alpha^2 + \sigma_X^2\delta)\bar{N})$
$E_A(S_{YX}) = \left((\delta\sigma_{XY} + \alpha^2\mu_X\mu_Y)\bar{N} + \delta^2\mu_X\mu_Y S_N^2\right)$
$E_A(S_{Y_iX_i}) = (\delta\sigma_{XY} + \alpha^2\mu_X\mu_Y)$

Similarly for the ratio estimator $\hat{Y}_{R\left(\frac{Y}{X}\right)} = X\frac{\hat{Y}_{HT}}{\hat{X}_{HT}}$,

$Var\left(\hat{Y}_{R\left(\frac{Y}{X}\right)}\right) = \left(\frac{K}{k}M - N\right)\delta\left(\sigma_Y^2 + \frac{\mu_Y^2}{\mu_X^2}\sigma_X^2 - 2\frac{\mu_Y}{\mu_X}\sigma_{YX}\right)$ where $X = \sum_{i=1}^{K}\sum_{j=1}^{N_i} X_{ij}$ the population total for the auxiliary variable X.

Next consider a simple un-weighted ratio estimator: $\hat{Y}_p = X\frac{\sum_{i\epsilon S}\sum_{j\epsilon S_i} Y_{ij}}{\sum_{i\epsilon S}\sum_{j\epsilon S_i} X_{ij}}$. (3.1.1)

This is the BLUE under (2.1.5).

$Var(\hat{Y}_p) = \frac{N^2}{G^2}\frac{k}{K}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\delta\left(\sigma_Y^2 + \frac{\mu_Y^2}{\mu_X^2}\sigma_X^2 - 2\frac{\mu_Y}{\mu_X}\sigma_{YX}\right)$ where $n = \sum_{i=1}^{K} n_i$ and the ratio:

$\frac{Var\left(\hat{Y}_{R\left(\frac{Y}{X}\right)}\right)}{Var(\hat{Y}_P)} = \frac{\frac{K}{k}M-N}{\frac{N^2k}{G^2K}\left(n-\frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)} \geq \frac{\frac{K}{k}\left(Min_{\{n_i\}}\sum_{i=1}^{K}\frac{N_i^2}{n_i}\right)-N}{\frac{N^2k}{G^2K}\left(n-\frac{k}{K}\left(Min_{\{n_i\}}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\right)} = \frac{\frac{K}{k}\frac{kN^2}{K\bar{k}\bar{n}}}{\frac{N^2k}{(k\bar{n})^2K}\left(n-\frac{k}{K}\left(\frac{K^2\bar{n}^2}{N}\right)\right)} = \frac{1}{1-\frac{k\bar{n}}{N}} \geq 1$ by Lemma 2.1 and Lemma 2.2.

This proves the following theorem.

Theorem 2.5__ Under the cluster sample design considered in this paper, $\mathrm{Var}\left(\hat{Y}_p\right) \leq \mathrm{Var}\left(\hat{Y}_{R\left(\frac{Y}{X}\right)}\right)$.

In case the cluster sample sizes are constant, say $n_i \doteq n_0 \; for \; all \; 1 \leq i \leq K$, then the ratio: $\frac{Var\left(\hat{Y}_{R\left(\frac{Y}{X}\right)}\right)}{Var(\hat{Y}_P)} \geq 1 + \frac{S_N^2}{N^2}$. Thus when the cluster sizes vary a great deal and cluster sample sizes are nearly constant, the un-weighted ratio estimator, $\hat{Y}_p$, should replace the HT-weighted ratio estimator (this situation was encountered with USPS mail surveys) where $\frac{S_N^2}{N^2}$ ranged from 1 to 5 and $\hat{Y}_p$ was used in place of $\hat{Y}_{R\left(\frac{Y}{X}\right)}$. In that case, it was justified by a model conjectured from historical data – a second route to the PSMB BLUE, $\hat{Y}_p$.

3.2 Stratification and the Combined Ratio Estimator

Finally consider the Combined Ratio Estimator, Cochran (1977), for the population total of the Y-variate where there are H strata. Let h be the stratum subscript $1 \leq h \leq H$ and within each stratum let there be a unique APM with APM parameters subscripted by h.

Let these APM parameters be denoted $(\mu_{Xh}, \sigma_{Xh}^2, \mu_{Yh}, \sigma_{Yh}^2, \delta_h, \alpha_h^2)$ analogous to the $(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2, \delta, \alpha^2)$ in Section 3.1. Let K_h, k_h, N_{hi}, n_{hi}, S_h, and S_{hi} be the stratum h sampling parameters defined exactly as are $K,\ k,\ N_i,\ n_i,\ S, and\ S_i$ in Section 2.2 for an un-stratified population.

The population total of the Y-variate is $Y = \sum_{h=1}^{H} Y_h$ where $Y_h = \sum_{i=1}^{K_h}\sum_{j=1}^{N_{hi}} Y_{hij}$ and similarly for X.

Let $\hat{Y}_{HTh}$ be the Horwitz-Thompson estimator for the stratum h total of the Y-variate, Y_{hij}. Then $\hat{Y}_{HTh} = \frac{K_h}{k_h}\sum_{i\epsilon S_h}\frac{N_{hi}}{n_{hi}}\sum_{j\epsilon S_{hi}} Y_{hij}$. The Combined Ratio Estimator is:

$$\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)} = X\frac{\sum_{h=1}^{H}\hat{Y}_{HTh}}{\sum_{h=1}^{H}\hat{X}_{HTh}}.$$

By approximating $\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)}$ with a plane passing through the expected values of numerator and denominator, the variance of $\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)}$ can be approximated as:

$Var\left(\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)}\right) \doteq \sum_{h=1}^{H} Var\left(\hat{G}_{HTh}\right)$ where $G_{hij} = Y_{hij} - BX_{hij}$ and $B = \frac{Y}{X}$.

Let $\mu_{Gh} = E\left(G_{hij}\right) \doteq \mu_{Yh}\delta_h - B\delta_h\mu_{Xh}$. $\mu_{Gh} = 0$ for all h if and only if all stratum ratios $\left\{\mu_{Yh}/\mu_{Xh}\right\}$ are identical. Let $M_h = \sum_{i=1}^{N_h}\frac{N_{hi}^2}{n_{hi}}$, $S_{Nh}^2 = \frac{1}{K_h - 1}\sum_{i=1}^{K_h}(N_{hi} - \bar{N}_h)^2$, and $\bar{N}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} N_{hi}$ then from (2.3.6).

$$Var\left(\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)}\right) = \sum_{h=1}^{H}\left[\left(\frac{K_h^2}{k_h} - K_h\right)\mu_{Gh}^2\delta_h^2 S_{Nh}^2 + \left(\sigma_{Gh}^2\delta_h + \mu_{Gh}^2\alpha_h^2\right)\left(\frac{K_h}{k_h}M_h - N_h\right)\right]$$

When the $\left\{\mu_{Yh}/\mu_{Xh}\right\}$ are not all identical then $\mu_{Gh} \neq 0$ for at least some h and the variance of $\hat{Y}^{C}_{R\left(\frac{Y}{X}\right)}$ does not reduce to sums of $\sigma_{Gh}^2\delta_h\left(\frac{K_h}{k_h}M_h - N_h\right)$ but rather to sums of $\left(\frac{K_h^2}{k_h} - K_h\right)\mu_{Gh}^2\delta_h^2 S_{Nh}^2 + \left(\sigma_{Gh}^2\delta_h + \mu_{Gh}^2\alpha_h^2\right)\left(\frac{K_h}{k_h}M_h - N_h\right)$. In particular, $\left(\frac{K_h^2}{k_h} - K_h\right)\mu_{Gh}^2\delta_h^2 S_{Nh}^2$ contains the square of the mean number of atoms per unit in stratum h so unless S_{Nh}^2 is vanishingly small the variance of the combined ratio estimator will be quite large. This result was found in Woodruff (2009) by a more tortuous route. It apparently implies that in many sampling problems where the APMs described in this paper apply, one should use a Combined Ratio Estimator only when all stratum ratios are nearly identical, otherwise the separate ratio estimator is better.

3.3 Summary

The sampling variances under the cluster sampling design and the APM described above of the seven estimators considered in this paper are:

1) $Var\left(\hat{Y}_{PS}\right) = N^2\frac{\delta\sigma_Y^2}{k\bar{n}^2K}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right) = \frac{K}{k}\frac{N^2}{n^2}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\delta\sigma_Y^2$

2) $Var\left(\hat{Y}_{R\left(\frac{Y}{A}\right)}\right) = \left(\frac{K}{k}M - N\right)\delta\sigma_Y^2$

3) $\text{Var}(\hat{Y}_{HT}) \doteq \left(\frac{K}{k}M - N\right)\left(\mu_Y^2\alpha^2 + \sigma_Y^2\delta\right) + \left(\frac{K^2}{k} - K\right)\mu_Y^2\delta^2 S_N^2$

$$= Var\left(\hat{Y}_{R\left(\frac{Y}{A}\right)}\right) + \left(\frac{K}{k}M - N\right)\mu_Y^2\alpha^2 + \left(\frac{K^2}{k} - K\right)\mu_Y^2\delta^2 S_N^2$$

4) $Var\left(\hat{Y}_{R\left(\frac{Y}{X}\right)}\right) = \left(\frac{K}{k}M - N\right)\delta\left(\sigma_Y^2 + \frac{\mu_Y^2}{\mu_X^2}\sigma_X^2 - 2\frac{\mu_Y}{\mu_X}\sigma_{YX}\right)$

5) $Var(\hat{Y}_p) = \frac{K}{k}\frac{N^2}{n^2}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\delta\left(\sigma_Y^2 + \frac{\mu_Y^2}{\mu_X^2}\sigma_X^2 - 2\frac{\mu_Y}{\mu_X}\sigma_{YX}\right)$

$$= Var(\hat{Y}_{PS}) + \frac{K}{k}\frac{N^2}{n^2}\left(n - \frac{k}{K}\sum_{i=1}^{K}\frac{n_i^2}{N_i}\right)\delta\left(\frac{\mu_Y^2}{\mu_X^2}\sigma_X^2 - 2\frac{\mu_Y}{\mu_X}\sigma_{YX}\right)$$

6)

$$Var\left(\hat{Y}^C_{R\left(\frac{Y}{X}\right)}\right) = \sum_{h=1}^{H}\left[\left(\frac{K_h^2}{k_h} - K_h\right)\mu_{Gh}^2\delta_h^2 S_{Nh}^2 + (\sigma_{Gh}^2\delta_h + \mu_{Gh}^2\alpha_h^2)\left(\frac{K_h}{k_h}M_h - N_h\right)\right]$$

7) $Var\left(\hat{Y}^S_{R\left(\frac{Y}{X}\right)}\right) = \sum_{h=1}^{H}\left(\frac{K_h}{k_h}M_h - N_h\right)\delta_h\sigma_z^2$ where $\sigma_z^2 = \sigma_{Yh}^2 + \frac{\mu_{Yh}^2}{\mu_{Xh}^2}\sigma_{Xh}^2 - 2\frac{\mu_{Yh}}{\mu_{Xh}}\sigma_{YXh}$

and $\hat{Y}^S_{R\left(\frac{Y}{X}\right)}$ is the separate ratio estimator, $\hat{Y}^S_{R\left(\frac{Y}{X}\right)} = \sum_{h=1}^{H}X_h\frac{\hat{Y}_{HTh}}{\hat{X}_{HTh}}$.

These formulae can be used to estimate variance via estimation of the APM parameters from the sample atoms – the number of sample atoms can be quite large even when sample size in units is modest. This is an alternative to the variance estimation methodology described in Woodruff (2009).

4. Conclusions

This paper provides some mathematical foundation for Pre-sampling Model Based Inference where repeated sampling error under stratified cluster designs is the criterion for comparison. Previous papers, Woodruff (2010,2009,2008,2007) relied on simulation studies to compare repeated sampling error. The mathematics derived in this paper provides a more general foundation for PSMB and provides explanation for the simulation results in those earlier papers.

The theorems presented above highlight situations (sample designs and study variables) where design based inference should be avoided. They do this by expressing sampling error in terms of both sample design parameters and APM parameters (the stochastic structure the study variables inherit from the APM).

The goals of PSMB inference were stated in the Introduction`s first paragraph. Sampling inference should be based on the impartiality of randomization, it should be multivariate (as are most sample surveys), and it should avoid (or at least minimize) well intentioned but opinioned tinkering (outlier adjustment and model conjecture). To accomplish these goals, it should provide a population model imposed by randomization, it should avoid a model conjectured from sample data or historical population data, it should be structured for application of theorems about Best Linear Unbiased Estimation under an imposed population model, and provide estimates that are robust against sample design inefficiencies, non-response, and outliers.

Pre-sampling Model Based Inference is motivated by these goals and developed to exploit data structures found in many populations. Expansion estimation, whether Model Based or Design Based, implicitly assumes the atom structure that is made explicit through the APMs in Section 2.1. 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. 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. This is particularly the case when design inefficiencies magnify the repeated sampling error of Design Based estimates and reliable models are not available.

The theorems in this paper require further generalization to cases of several auxiliary variables and several atom types as studied in Woodruff (2009, 2010). Although unit study variables cannot be considered iid, since each unit has a different mix of atom types, the individual atom study variables may well be approximated as such and the multivariate version of PSMB in Woodruff (2009, 2010) applied. This paper and those referenced above suggest that probability sampling theory can be usefully expanded from randomized unit selection to both randomized unit selection and randomized unit synthesis. Continuing along this path, probability sampling theory can be enhanced by paralleling and expanding upon subjects found in the standard sampling texts, Cochran (1977), using methodologies initiated here.

References

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

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

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