Probability Sample Designs that Impose Models on Survey Data

Stephen Woodruff

Abstract

For a substantial class of sampling problems, the sample design and the characteristics of the population being sampled, not only provide the probabilities of selection but also impose a regression model on the sample data. For many of these sampling problems this duality is an academic curiosity but there are some designs where inference based on the design and the probabilities of selection can be unacceptably inefficient. For such designs, the model provides a backup, a Best Linear Unbiased Estimator (BLUE), and avoids the inefficiencies of the Horvitz-Thompson (HT) based estimation methodologies.

Mail populations and their characteristics provide an example of this situation because an adequate frame is not available for sample design. The frame is only available after the sample has been selected and the data collected. In this situation, lack of design control and use of a combined ratio estimator result in estimates with sampling errors several times larger than those of the BLUE. The BLUE is a separate ratio estimator with a variance that is about 1/5th to 1/15th that of the Combined Ratio HT Estimator. This paper derives theory supporting the replacement of a combined ratio estimator with a Best Linear Unbiased Estimator for sampling problems where design control is prohibitively difficult and a model for the sample data is implicit.

Key Words: Combined and Separate Ratio Estimation, Best Linear Unbiased Estimation, Horvitz-Thompson Estimation.

1. Introduction

This paper describes a general sample design and three estimators. Two of these estimators are the Combined Ratio HT Estimator and the Separate Ratio HT Estimators, Cochran (1973). The third estimator is a Best Linear Unbiased Estimator (BLUE), Rao (1973) derived from the sample design and characteristics of the sampled population. This third estimator avoids weaknesses in the HT estimators imposed by lack of design control and exploits properties of mail populations which impose a model on the sample data. The variance of the BLUE is substantially smaller than the variance of the first two estimators.

In mail sampling, efficient sample design is difficult due to lack of a sampling frame at the design stage. The frame parameters are collected with sample data during data collection and the sampling clusters are formed haphazardly due to unpredictability induced by mail processing and transportation. The clusters within first stage strata are largely unrelated to actual characteristics of the study variables. These facts not only hinder efficient sample design that would normally control the variance of the Horwitz-Thompson estimator, but also make differential sample expansions unreflective of any real population differences for which sample expansions often adjust. See Woodruff, Lan (2004) for a more complete description of mail sampling and its particular difficulties.

The design is similar to a stratified cluster sample design but the second stage clusters that are randomly selected are not sub-sampled in the usual way. Instead, only a cluster's total over the individual cluster members is recorded. These cluster members are called atoms and the cluster is the ultimate sample unit since the study variables attached to each atom are not recorded. This structure imposes the model when the second stage clusters are appropriately designed.

In mail sampling, the clusters (ultimate sample units) are containers of mail (a bag, a tray, or a tub) and the atoms are the individual mail pieces within the containers. Only a container's piece totals or piece totals by mail class (total weight, total pieces, total postage) are recorded for each sampled container. This method of data collection imposes a model on the sample data when the strata are designed so that the content of each cluster is stochastically indistinguishable from a simple random sample of atoms from the stratum. Mail processing imposes this structure on containers of mail. The comparison of the three estimators is carried out with respect to sampling distributions (repeated sampling) under stratified cluster sampling designs. The populations being sampled are very large ($N=10^6$ atoms and larger) and the sample sizes small and therefore any distinction between finite population sampling and superpopulation sampling is inconsequential. Finite population corrections are ignored. The mathematical details of this model derivation are in Section 2.

The rest of the paper establishes the variance hierarchy of the three estimators. The Combined Ratio HT Estimator ($\hat{T}_{DC}$) has greatest variance, the Separate Ratio HT Estimator ($\hat{T}_{DS}$) has next greatest, and the BLUE ($\hat{\hat{T}}_{DS}$) derived from the model has smallest. The last section of the paper describes a simulation study that quantifies the size of the steps in this hierarchy from largest variance to smallest variance. The simulations in Section 5 show that the BLUE has roughly 1/5th to 1/15th the variance of The Combined Ratio HT Estimator – the estimator that is being replaced with the BLUE by the USPS.

2. Mail Sampling and a Model for Estimation

To emphasize the ephemeral nature of mail populations, these populations will be referred to as mail flows or just flows. Sampling mail is like sampling a river by collecting some of its water at specific times from a fixed location. Little is known about its content until after it is sampled and then the sampled portion of the river is gone. This is a major difference between sampling mail and sampling other relatively static populations like households, businesses, or people. This is also the reason the design cannot be fully known in advance or fully described here.

A mail flow is a small subset of the totality of mail pieces (letters, cards, magazines, and small packages) that move hourly through transportation and processing facilities on their way to their destination. The totality of mail is stratified into thousands of strata by processing facility, mail class, transportation mode, container type, reference period (month, quarter, or year), and country of origin or destination for international mail. Estimates of totals for study variables are needed for domains (D) that are aggregates of many mail strata. These domain estimates are sums of their strata estimates and the same is true for their variances. The ultimate sample units are containers holding mail pieces. A container is a bag, tray, or tub of mail pieces and is light enough for a single person to lift and carry; they average about 10 kilograms and only exist for a few hours.

Within strata, the first stage clusters are days, a random sample of days each month are selected, and the mail containers in the selected days are sub-sampled in the second stage of selection. There is usually ad-hoc sampling within sample days that spreads a fixed sample size over whatever arrives during the day. An exact description is peripheral to the statistical content of this paper (also impossible to known in advance). These necessarily omitted details aggravate the design problems described here but they don't effect the efficacy of the proposed solution.

The short lifetime of mail population units and the lack of prior knowledge about them imply that the sampling frame used for sample design must be dependent on predictions of frame parameters: the stratum sizes and cluster sizes. Questionable predictions of these parameters and administrative restrictions on sample size both militate against efficient design (roughly equal probabilities of selection within 1st stage flow-month strata). After the mail has arrived and been sampled, the actual frame parameters are recorded and used to derive probabilities of selection. The selection probabilities for the sample days are fairly uniform; the erratic flow volumes within days cause the large fluctuations in final selection probabilities. These design effects, Kish (1995) are unrelated to the rate per kilogram of the study variables and the highly heterogeneous sample expansion that these ad hoc designs impose on the sample data do not reflect any sub-population differences for which differential weighting usually adjusts. Consequently, domain estimates based on HT estimation can be quite inefficient.

An alternate road to inference is available through a model that is imposed by the sample design and the process which generates mail flows. This process forces the rates per kilogram of the study variables to remain relatively stable from day-to-day within strata in spite of large daily variations in piece totals, weight totals, and postage totals. Containers are filled with the sole goal of moving the mail through to its destination as fast as possible. The mail populations that are sampled in a processing facility during a day are generated by thousands of independent decisions by people all over the country (and world). Containers are filled as this mail arrives and may each be considered a random sample of all pieces (or atoms) in the strata provided the strata are carefully defined and relatively short in time duration (usually a month in length). This provides a deductive foundation for the model and is preferable to inference based on inductive modeling through data mining techniques. The derivation of this model within a stratum follows next.

Let y_{fji} denote value of the study variable attached to the i^{th} atom in the j^{th} sample unit for study variable y in stratum f. Let $y_{fj} = \sum_{i=1}^{n_{fj}} y_{fji}$ where n_{fj} is the number of atoms (mail pieces) in sample unit j (container j) of stratum f. Let μ_{fy} and σ^2_{fy} be the stratum mean and variance of study variable y for the population of atoms in stratum f when y_{fji} is selected from stratum f with uniform probability for each atom. Then $E\left(y_{fji}\right) = \mu_{fy}$ and $V\left(y_{fji}\right) = \sigma^2_{fy}$. Within this stratum the sample units are formed so that their mean and variance are given as: $y_{fji} \propto (\mu_{fy}, \sigma^2_{fy}) \ \forall \ j \ and \ i$ in each sample unit (container of mail) and these $\left\{y_{fji}\right\}$ are pair-wise uncorrelated. μ_{fy} is the stratum f mean of study variable y and σ^2_{fy} is the variance of the population of stratum atoms for y in stratum f.

The populations considered here are very large in terms of both clusters (containers of mail) and atoms. Thus finite population corrections are essentially unity and omitted and other finite population notation like S^2_{fy} is replaced with σ^2_{fy}.

Then $E\left(y_{fj}\right) = \sum_{i=1}^{n_{fj}} E(y_{fji}) = n_{fj}\mu_{fy}$ and similarly $V\left(y_{fj}\right) = \sum_{i=1}^{n_{fj}} V(y_{fji}) = n_{fj}\sigma^2_{fy}$. Note that the $\left\{n_{fji}\right\}$ where $n_{fji} = 1 \ \forall \ j \ and \ i$ in stratum f are also atomic variables and $n_{fj} = \sum_{i=1}^{n_{fj}} n_{fji}$.

This structure for y_{fj} and n_{fj} can also be rewritten:

$$y_{fj} = n_{fj}\mu_{fy} + \varepsilon^y_{fj} \quad \text{where } \varepsilon^y_{fj} \propto (0, n_{fj}\sigma^2_{fy}) \tag{2.1}$$

This implies that all study variables are roughly proportional to an auxiliary variable (a study variable for which the stratum total is known). This is a consequence of the transitivity of proportionality as follows:

Let k_{fj} be a study variable that is also an auxiliary variable, then from (2.1):

$$k_{fj} = n_{fj}\mu_{fk} + \varepsilon^k_{fj} \tag{2.2}$$

By rearranging (2.2):

$$n_{fj} = \left(k_{fj} \Big/ \mu_{fk} \right) - \delta^k_{fj} \quad \text{where } \delta^k_{fj} = \varepsilon^k_{fj} \Big/ \mu_{fk} \quad \text{and}$$

$$\delta^k_{fj} \propto (0, n_{fj}\left(\sigma^2_{fk} \Big/ \mu^2_{fk} \right)) \tag{2.3}$$

From (2.1) , for all study variables, y :

$$\begin{aligned} y_{fj} &= \left(k_{fj} \Big/ \mu_{fk} - \delta^k_{fj} \right)\mu_{fy} + \varepsilon^y_{fj} \\ &= \left(\mu_{fy} \Big/ \mu_{fk} \right) k_{fj} + (\varepsilon^y_{fj} - \delta^k_{fj}\mu_{fy}) \\ &= \beta_f k_{fj} + \lambda^k_{fj} \quad \text{where } \beta_f = \mu_{fy} \Big/ \mu_{fk} \quad \text{and} \end{aligned}$$

$$\lambda^k_{fj} = (\varepsilon^y_{fj} - \delta^k_{fj}\mu_{fy})$$

Thus $\lambda^k_{fj} \propto (0, n_{fj}G_f)$ where

$$G_f = \left(\sigma^2_{fy} + \mu^2_{fy}\frac{\sigma^2_{fk}}{\mu^2_{fk}} - 2\frac{\mu_{fy}}{\mu_{fk}}\sigma^2_{fy} \right).$$

By (2.2) n_{fj} is approximately proportional to k_{fj} , $\lambda^k_{fj} \propto (0, k_{fj}G_f^{'})$, where $G_f^{'} = \dfrac{G_f}{\mu_{fk}}$.

Summarizing these results, the model for y_{fj} follows:

$$y_{fj} = \beta_f k_{fj} + \lambda^k_{fj} \quad \text{where } \lambda^k_{fj} \propto (0, k_{fj}G_f^{'}) \tag{2.4}$$

and under this model, the BLUE estimator for β_f in the stratum is:

$$\hat{\hat{\beta}}_f = \frac{\sum_{j\varepsilon s_f} y_{fj}}{\sum_{j\varepsilon s_f} k_{fj}} \tag{2.5}$$

where s_f is the stratum f sample, Rao (1973, pg 230). The BLUE estimate for the stratum total is:

$$\hat{\hat{\beta}}_f K_f = \frac{\sum_{j\varepsilon s_f} y_{fj}}{\sum_{j\varepsilon\, s_f} k_{fj}} K_f \quad \text{where } K_f \text{ is the stratum f}$$

total for the auxiliary variable, k.

3. The Combined Horvitz-Thompson Ratio Estimator and Design Effect in Mail Surveys

The correlation between domain D's kilogram total of mail and its total for the study variable is exploited to strengthen estimation by using a combined ratio estimator. This estimator has been used by the USPS to estimate flow totals of mail characteristics. It was probably chosen because it has been the standard across many US Federal Government statistical sampling programs for sampling households, businesses, and people. The frame problems described above may make it an unfortunate choice for measurement of mail characteristics.

A container's probability of selection is computed from the number of containers sampled and total number of containers available for sampling as recorded along with the study variables during data collection.

Resource constraints limit the sample size for a selected day to less than about 10 containers, without regard to the amount of mail that arrives during the day. The sample design is at the mercy of erratic daily flow population sizes generated by transportation and processing. A stratum (mail flow) population can change from a dozen or fewer containers on one day to several hundred or more the next. This fact creates widely different probabilities of selection, that can vary by a factor of a hundred or more between a container selected on a light volume day and one selected on a heavy volume day from the same stratum.

As mentioned in Section 2 there may also be second stage strata within sample days for containers that arrive during different times of the day. These ad hoc sample designs within selected days are done by the data collector, they adjust for the actual mail that day, and can be quite complex. This part of sample design adjusts for office workload with the goal of spreading a small sample over whatever arrives during the day. It also injects additional variability into the container probabilities of selection within strata and is a major factor in inefficient HT estimation.

The Combined Ratio HT Estimator for the domain D total of the study variable y is:

$$\hat{T}_{DC} = K_D \frac{\sum_{f=1}^{F}\sum_{j=1}^{n_f} \frac{y_{fj}}{\pi_{fj}}}{\sum_{f=1}^{F}\sum_{j=1}^{n_f} \frac{k_{fj}}{\pi_{fj}}} \qquad (3.1)$$

where the pair (f,j) represents the j^{th} sample container in stratum f for $f = 1,2,3,.....F$, where F is the number of strata that comprise domain D. n_f is the number of sample containers in stratum f. y_{fj}, π_{fj}, and k_{fj} denote respectively the study variable y for the j^{th} sample container in stratum f , its probability of selection, and its kilogram weight. K_D is the known total domain D kilograms of mail.

Let $r_{fj} = {}^{y_{fj}}\!/_{k_{fj}}$, then: $\hat{T}_{DC} = K_D \frac{\sum_{f=1}^{F}\sum_{j=1}^{n_f} r_{fj}\frac{k_{fj}}{\pi_{fj}}}{\sum_{f=1}^{F}\sum_{j=1}^{n_f} \frac{k_{fj}}{\pi_{fj}}}$

This can also be written as:

$$\hat{T}_{DC} = K_D \sum_{f=1}^{F}\sum_{j=1}^{n_f} W_{fj}^{*} r_{fj} = K_D \hat{\beta}_D \qquad (3.2)$$

where $W_{fj}^{*} = \frac{\frac{k_{fj}}{\pi_{fj}}}{\sum_{f=1}^{F}\sum_{j=1}^{n_f}\frac{k_{fj}}{\pi_{fj}}}$, and

$$\hat{\beta}_D = \left(\sum_{f=1}^{F}\sum_{j=1}^{n_f} W_{fj}^{*} r_{fj} \right) . \qquad (3.3)$$

$\sum_{f=1}^{F}\sum_{j=1}^{n_f} W_{fj}^{*} = 1$, and $W_{fj}^{*} \geq 0$ for all sample containers in domain D.

(3.2) can be written:

$$\hat{T}_{DC} = K_D \sum_{f=1}^{F}\sum_{j=1}^{n_f} W_{fj}^{*} r_{fj} = K_D \sum_{f=1}^{F} \hat{W}_f \hat{\beta}_f \qquad (3.4)$$

where: $\hat{W}_f = \sum_{j=1}^{n_f} W_{fj}^{*}$ and $\hat{\beta}_f = \sum_{j=1}^{n_f} \frac{W_{fj}^{*}}{\hat{W}_f} r_{fj}$ (3.5).

Note: $\sum_{f=1}^{F} \hat{W}_f = 1$ and $E\left(\hat{W}_f\right) \doteq \frac{K_f}{K_D}$

(3.5)

$\hat{W}_f$ is an estimate of the proportion of total domain D kilograms in stratum f and $\hat{\beta}_f$ is an estimate of the rate per kilogram of the study variable y in stratum f.

Let $\hat{\hat{T}}_{DS} = \sum_{f=1}^{F} K_f \hat{\hat{\beta}}_f \quad = \sum_{f=1}^{F} K_D E\left(\hat{W}_f\right)\hat{\hat{\beta}}_f$ (3.6)

where $\hat{\hat{\beta}}_f$ is given by (2.5) .

Let:

$$\hat{T}_{DS} = \sum_{f=1}^{F} K_f \hat{\beta}_f = \sum_{f=1}^{F} K_D E\left(\hat{W}_f\right)\hat{\beta}_f . \qquad (3.7)$$

Estimation of the study variable totals for a domain reduces to estimating their rates per kilogram for the domain, β_D. By writing $\hat{T}_{DC}$ as proportional to the weighted average of the $\left\{\hat{\beta}_f\right\}$ in (3.4), it may be clearer how inefficient design works against $\hat{T}_{DC}$.

Minimizing the variance of the HT estimator requires that a unit's selection probability be proportional to its study variable. Thus minimizing variance requires nearly uniform selection probabilities within strata since weights of mail containers tend to be similar (lie within a relatively small range of about 3 to 20 kilograms) and the study variables are proportional to these container kilograms. This optimal design is a nearly self-weighting design. This optimality condition is violated under sample designs imposed by mail processing and transportation. If an optimal sample design could be executed, then the Separate Ratio HT Estimator would be the BLUE derived in Section 2.

Large day-to-day fluctuations in mail volumes available for sampling and fixed daily sample sizes result in HT based estimates with extremely variable weights, $\left\{\dfrac{W_{fj}^{*}}{\hat{W}_f}\right\}$, in (3.5). Since these weights add to one and some will be many times greater than others, sample containers with the small weights are effectively pushed from the sample by those with the large weights. This reduces the effective sample size quite substantially.

4. Comparison of the Three Estimators, $\hat{T}_{DC}$, $\hat{T}_{DS}$, and $\hat{\hat{T}}_{DS}$

All three of these estimators are unbiased. In this section the following inequality is established:

$$V\left(\hat{T}_{DC}\right) \;\geq\; V\left(\hat{T}_{DS}\right) \;\geq\; V\left(\hat{\hat{T}}_{DS}\right). \qquad (4.1)$$

From (3.4):

$$\hat{T}_{DC} = K_D \sum_{f=1}^{F}\sum_{j=1}^{n_f} W_{fj}^{*} r_{fj} = K_D \sum_{f=1}^{F} \hat{W}_f \hat{\beta}_f$$

where: $\hat{W}_f = \sum_{j=1}^{n_f} W_{fj}^{*}$ and $\hat{\beta}_f = \sum_{j=1}^{n_f} \dfrac{W_{fj}^{*}}{\hat{W}_f} r_{fj}$

By independence of sampling between different strata:

$$V\left(\hat{T}_{DC}\right) = K_D^2 \sum_{f=1}^{F} V\left(\hat{W}_f \hat{\beta}_f\right) \qquad (4.2)$$

Apply the Taylor-Series expansion about expectations to each variance term in this sum to get:

$$V\left(\hat{T}_{DC}\right) \doteq$$
$$K_D^2 \sum_{f=1}^{F}\left[\left(E(\hat{W}_f)\right)^2 V\left(\hat{\beta}_f\right) + \left(E(\hat{\beta}_f)\right)^2 V\left(\hat{W}_f\right)\right]$$

(4.3)

$E(\hat{\beta}_f) = \beta_f$, the stratum's population rate per kilogram of the study variable, and $E(\hat{W}_f) = W_f$. (4.3) becomes:

$$V\left(\hat{T}_{DC}\right) \doteq K_D^2 \sum_{f=1}^{F}\left[W_f^2 V\left(\hat{\beta}_f\right) + \beta_f^2 V\left(\hat{W}_f\right)\right]$$

(4.4)

The variance of $\hat{T}_{DS}$ from (3.7) and independence of sampling between different strata is

$$V\left(\hat{T}_{DS}\right) = \sum_{f=1}^{F} K_f^2 V\left(\hat{\beta}_f\right) = K_D^2 \sum_{f=1}^{F} W_f^2 V\left(\hat{\beta}_f\right).$$

(4.5)

Comparing (4.4) and (4.5) , $V\left(\hat{T}_{DC}\right)$ differs from $V\left(\hat{T}_{DS}\right)$ by $K_D^2 \sum_{f=1}^{F}\left[\beta_f^2 V\left(\hat{W}_f\right)\right]$, which is always greater than or equal to zero.

Therefore: $V\left(\hat{T}_{DC}\right) \;\geq\; V\left(\hat{T}_{DS}\right)$.

Both $\hat{\beta}_f$ and $\hat{\hat{\beta}}_f$ are linear and unbiased in the sample data; therefore $V\left(\hat{\beta}_f\right) \geq V\left(\hat{\hat{\beta}}_f\right)$ since $\hat{\hat{\beta}}_f$ is minimum variance among the linear unbiased estimators. This implies:

$$V\left(\hat{T}_{DS}\right) = \sum_{f=1}^{F} K_f^2 V\left(\hat{\beta}_f\right) \quad \geq \sum_{f=1}^{F} K_f^2 V\left(\hat{\hat{\beta}}_f\right) =$$

$V\left(\hat{\hat{T}}_{DS}\right)$ and (4.1) is established.

$$V\left(\hat{T}_{DC}\right) \geq V\left(\hat{T}_{DS}\right) \geq V\left(\hat{\hat{T}}_{DS}\right).$$

Next note that the bias in $\hat{\hat{T}}_{DS}$ is proportional to a linear combination of the: $\left\{Cov\left(\sum_{j\varepsilon s_f} k_{fj}, \hat{\hat{\beta}}_f\right), f\varepsilon D\right\}$. Under (2.4), these covariances are all zero independent of sample size. An analogous statement is true for the $\left\{\hat{\beta}_f\right\}$. This implies that both $\hat{T}_{DS}$ are $\hat{\hat{T}}_{DS}$ are unbiased. These theoretical statements about bias are supported by the simulation results in Section 5 where the size of the steps between $V\left(\hat{T}_{DC}\right)$ and $V\left(\hat{\hat{T}}_{DS}\right)$ are also measured. For typical mail flow designs $V\left(\hat{T}_{DC}\right)$ will be many times larger than $V\left(\hat{\hat{T}}_{DS}\right)$.

5. Quantifying the Difference Between the Sampling Errors of the Three Estimators

The size of the steps in the inequality, $V\left(\hat{T}_{DC}\right) \geq V\left(\hat{T}_{DS}\right) \geq V\left(\hat{\hat{T}}_{DS}\right)$, were estimated using simulations where three mail stream domains were repeatedly sampled in order to estimate both bias and variance. Five hundred independent sample replicates were selected following an approximation of sampling plans used by the USPS. For each replicate, the three estimators were computed and these 500 replicate estimates were used to estimate the variance and bias of each estimator. This was done for three population domains, each with five degrees of day-to-day volume volatility. Summarizing these 15 simulation studies, the variance of $\hat{\hat{T}}_{DS}$ ran between 1/5th and 1/15th the variance of $\hat{T}_{DC}$ and the bias of all three estimators were statistical zeros.

The three study domains were constructed using sample data for airmail entering the US for nine months and 3 countries together with their total population kilograms by stratum. These domains were constructed to mimic the characteristics of the totality of mail containers for these flows. There are three domains: Great Britain with 156 strata (a stratum for a country domain is a mail group defined by processing office, container type, and month), France with 62 strata, and Belgium with 27 strata. These populations are clean and the sample design does not fully capture the complex and unpredictable elements of sub-sampling that is a necessary part of mail data collection. For this reason, the design difficulties that afflict the Combined Ratio HT Estimator are probably understated in this study. These design difficulties have little effect on the BLUE.

The comparisons in the tables below are relative (ratios) and attempt to illuminate the effect on the estimators of the design difficulties inherent in mail sampling.

The volatility in daily mail volumes is modeled by random assignment of the containers in each strata to days in that stratum's month. The amount of day-to-day volatility is measured by the *-weight ratios from (3.3) that are in the tables below adjacent to "Weight Ratio". This is the ratio between the largest and smallest *-weights for the simulation and it varies from minimal (simulations A, F, and K where daily volumes are kept nearly equal within each stratum's month) to fairly large (far right in each table). For each stratum's month, a random sample of 4 days is selected and within each selected day, a random sample of 4 mail containers is selected. The study variable to be estimated is the total number of mail pieces in the domain and thus the study variable attached to each container is its piece count and its rate per kilogram (r) is its number of pieces per kilogram. $\hat{T}_{DC}$, $\hat{T}_{DS}$, and $\hat{\hat{T}}_{DS}$ are computed for each of 500 replicates of the sampling process. These 500 estimates are subtracted from the known population total and used to estimate mean square error, bias and variance, $\hat{V}(.)$, for each of the three estimators and 15 simulations tabled below. All three estimators have simulation bias estimates that are statistical zeros, so Mean Square Error and variance are statistically indistinguishable.

The tables below contain estimated variance ratios for the three estimators.

Ratio 1 = $\hat{V}\left(\hat{T}_{DS}\right) \big/ \hat{V}\left(\hat{T}_{DC}\right)$ and

Ratio 2 = $\hat{V}\left(\hat{\hat{T}}_{DS}\right) \Big/ \hat{V}\left(\hat{T}_{DC}\right)$.

Ratio 1 compares the Combined HT Ratio Estimator to the Separate HT Ratio Estimator and shows that the variance change from $\hat{T}_{DC}$ to $\hat{T}_{DS}$ is the largest step from $V\left(\hat{T}_{DC}\right)$ to $V\left(\hat{\hat{T}}_{DS}\right)$. The Combined Ratio HT Estimator is extremely sensitive to day-to-day volume volatility while the two separate ratio estimators are much less affected by this feature of mail sampling.

A comparison of Ratio 1 and Ratio 2 shows that $\hat{\hat{T}}_{DS}$ provides an additional 30% to 60% reduction in variance compared to $\hat{T}_{DS}$. For each country, simulations A, F, and K model variance ratios given no day-to-day variability in mail volumes within each stratum, thus $\hat{\beta}_f$ and $\hat{\hat{\beta}}_f$ are the same. For these three simulations, this column shows a relatively small variance reduction. All columns to the right of the first column (A, F, K) show that even a little day-to-day volatility in mail volumes interact badly in the Combined Ratio HT Estimator, $\hat{T}_{DC}$, and substantially increase its variance compared to the other two estimators. Apparently, if mail volumes didn't vary from day-to-day within strata, the Combined Ratio HT Estimator would be acceptable if not optimal (see simulations A, F, and K). For all the other simulations, increasing amounts of day-to-day volatility are modeled as you go from left to right in the tables. This increase is measured by the Weight Ratio. A modest degree of day-to-day volatility results in a large drop off in efficiency in $\hat{T}_{DC}$ compared to $\hat{\hat{T}}_{DS}$ and $\hat{T}_{DS}$. This may reflect the fact that the separate ratio estimators confine the reduction of effective sample size caused by highly differential weighting to individual sampling strata. The Combined Ratio HT Estimator extends this effect across the entire estimation domain and reduces the effective sample to the small handful of containers sampled on the busiest days at the largest processing center in the estimation domain.

Table 1. Great Britain

Weight Ratio	Simul A 2.04	Simul B 11.9	Simul C 20.8	Simul D 35.1	Simul E 40.6
Ratio 1	.684	.139	.124	.143	.112
Ratio 2	.684	.101	.082	.094	.071

Table 2. Belgium

Weight Ratio	Simul F 1.88	Simul G 4.7	Simul H 5.79	Simul I 7.0	Simul J 8.85
Ratio 1	.781	.196	.21	.201	.132
Ratio 2	.781	.165	.174	.169	.102

Table 3. France

Weight Ratio	Simul K 1.66	Simul L 7.8	Simul M 12.67	Simul N 19.68	Simul O 23.05
Ratio 1	.87	.128	.117	.114	.127
Ratio 2	.87	.083	.069	.063	.071

These simulated designs reflect the general characteristics of mail flow sampling and estimation. Actual mail designs are far more complex and therefore the actual inefficiency of the Combined Ratio HT Estimator compared to the BLUE is probably understated in these simulations.

Actual daily mail flow volatility had to be artificially generated as did actual domain mail populations but these simulations do suggest that an alternative approach to estimating mail volumes needs consideration. The Combined Ratio HT Estimator is probably not a good choice for two reasons. First, it lacks robustness compared to either version of the separate ratio estimator, when there is substantial day-to-day volatility in mail volumes and when large volume differences between strata (processing facilities) exist. Both of these features characterize mail populations and while the separate ratio estimators are little affected by them, the Combined Ratio HT Estimator is remarkably sensitive to them. Its variance explodes with day-to-day volume volatility. Second, for sampling mail populations both separate ratio estimators are unbiased, without regard to sample size. This second fact eliminates the rationale often cited for using the Combined Ratio HT Estimator.

6. Conclusions

For mail surveys, the sample design together with characteristics of the population being sampled, impose a model on the sample data and under this model there is a best linear unbiased estimator (BLUE). This BLUE has a much smaller variance than the Combined Ratio HT Estimator. The order of magnitude of this variance reduction is measured by the simulation studies in Section 5. Since this model is based on the sample design and the general properties of the population being sampled, model failure is a peripheral issue – like design failure when traditional probability sample designs and the HT estimator are the basis for inference.

This BLUE provides an alternative to HT based methodologies when design control is problematic. This is the case with mail surveys where unpredictability forced by weather, transportation, and processing creates large and random fluctuations in strata and cluster sizes that frustrate efficient survey design. The second source of estimation error is the use of a combined ratio estimator to avoid the potential bias of the separate ratio estimator, Cochran (1977). For mail populations, the separate ratio estimator is also unbiased. The Combined Ratio HT Estimator, $\hat{T}_{DC}$, is particularly sensitive to the lack of design control. The simulation studies demonstrate that even small deviations from self-weighting designs result in a large loss of efficiency for $\hat{T}_{DC}$ compared to $\hat{T}_{DS}$ and $\hat{\hat{T}}_{DS}$. As a result, $\hat{\hat{T}}_{DS}$ has a much smaller variance than $\hat{T}_{DC}$, and still substantially smaller than $\hat{T}_{DS}$. Apparently, the Combined Ratio HT Estimator, which was used to address a nonexistent bias problem creates a real variance problem given the design challenges that are a part of mail sampling.

In Section 5, a simulation study quantifies the variance differences between the three estimators. These differences were shown to be substantial – the Combined Ratio HT Estimator has roughly 5 to 15 times more variance than the BLUE. This is equivalent to discarding 80% to 93% of the sample data, the dollars it took to collect that data, and these figures probably understate this loss in efficiency.

For mail flows where the $\{K_f\}$are not available, this paper strongly suggests that mail processing include weighing the mail to obtain the strata weights so that a separate ratio estimator can be used. This is already done for international mail.

The procedures described above are not unique to mail sampling. They have application to general flow sampling where each stratum is sufficiently mixed so that a contiguous set of its atoms selected from the flow can be reasonably modeled as a simple random sample from the stratum population of atoms. This holds for certain biological populations, for example, the sampling of rivers for their particulate of microbial content.

The stochastic structure (atoms) within the ultimate sample units (mail containers) may be useful to inference in similar situations where design control is difficult. For such situations, the population model (or superpopulation model) is dependent on the sample design and inference flows from both design and model.

This paper formalizes the theory supporting procedures implemented in some mail surveys to deal with an historical instability in the Combined Ratio Horvitz-Thompson Estimator. Prior to this paper these procedures were based on an intuitive understanding achieved through years of observational experience. It also suggests further directions for research into the unique problems of mail sampling.

References

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

Des Raj, (1968), Sampling Theory, McGraw-Hill, PP 33.

Kendall, M.G. & Stuart A., (1969), The Advanced Theory of Statistics, Volume I- Distribution Theory, PP 60.

Kish, L., (1995), Survey Sampling, New York: Wiley.

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

Woodruff, S. M., Lan F. (2004), "Measurement of Mail Volumes - An Application of Model Assisted Estimation", Proceedings of the American Statistical Association, Survey Research Methods