Properties of the Combined Ratio Estimator and a Best Linear Unbiased Estimator When Design Control is Problematic

StephenWoodruff
Retired,800WestViewTerrace,Alexandria,Virginia,22301-275[0]

Abstract

This is a study of a Best Linear Unbiased Estimator (BLUE) and the Combined Ratio Horwitz-Thompson Estimator (CRHT) for sampling problems where self-weighting designs are impractical. Sampling of mail is an example of this. A measure of "self-weighting" is derived and their variances as functions of this measure are compared under repeated sampling from stratified cluster designs. For self-weighting designs, the variances of CRHT and BLUE are similar but that of CRHT increases rapidly with this measure. The BLUE is insensitive to this measure. When proper design control is impossible, the variance of CRHT can be many times (2 to 100s) that of the BLUE. The BLUE is derived from population characteristics of mail and the sample design, making model failure a virtual non-issue. Simulation results, supporting mathematics, their implications, and applications are presented.

KEY WORDS: Combined Ratio Estimation, Best Linear Unbiased Estimation, Stratified Cluster Designs

1. Introduction

This paper describes a sample design and three estimators of population totals. Two of these estimators are the Combined Ratio HT Estimator, $\hat{T}_C$ (used currently by USPS) and the Separate Ratio HT Estimators, $\hat{T}_S$ Cochran (1973). The third estimator is a Best Linear Unbiased Estimator, $\hat{\hat{T}}_S$ (BLUE), Rao (1973) derived from the sample design and characteristics of the sampled population, Woodruff (2006). $\hat{\hat{T}}_S$ avoids problems stemming from inadequate design control affecting $\hat{T}_S$ and $\hat{T}_C$. $\hat{\hat{T}}_S$ is derived from a model that captures the haphazard manner in which mail containers are filled during mail processing. The model is not used in the comparison of these estimators however, instead, $\hat{T}_C$, $\hat{T}_S$ and $\hat{\hat{T}}_S$ are all evaluated with respect to repeated sampling under a stratified two stage cluster sample design. This paper should serve to discourage the general use of the Combined Ratio HT Estimator in mail sampling and estimation.

Mail populations are referred to as mail flows or just flows. Sampling mail is like sampling a river by collecting some of its water in buckets at specific times from a fixed location. Little is known about its content (think sampling frame - stratum and cluster sizes in both numbers of units and auxiliary variable totals) until after it is sampled and then the sampled portion is gone. This is a major difference between sampling mail and sampling other relatively static populations like households, businesses, or people. This feature of flow sampling inhibits sample design since the frame is not available until after the sample is selected and observed.

A mail flow is a small subset of the totality of mail pieces (letters, cards, magazines, and small packages) that move through transportation and processing facilities from their origin to their destination. There are thousands of mail flows for which volume estimates are needed (mail volumes are totals for pieces, postage, and kilograms); each flow is defined by mail class, transport mode, origin, destination, and possibly other characteristics. These flows are stratified into dozens to hundreds of strata by processing facility, mail class, transportation mode, container type, month, and country of origin or destination for international mail. The sample unit is a container (a bag, tray, or tub of mail pieces), is light enough for a single person to lift and carry, and *only exists for a few hours*.

Within a stratum's month, the first stage clusters are days, a random sample of days is selected each month for sampling a mail flow at a processing facility and within each selected day, the second stage selection is a sample of mail containers from the flow-stratum. Piece, weight, and postage totals are recorded for each sampled container (possibly by mail class or shape within the container). There is often ad-hoc sample design within sample days which is conducted by the data collector for the purpose of spreading a fixed sample size (around 5 mail containers) over however many mail containers arrive during the sample day for the mail flow being sampled. An exact description of this ad hoc sampling is peripheral to the statistical content of this paper and, in

any case, impossible to know in advance. These final design stages further aggravate design problems for $\hat{T}_C$ and $\hat{T}_S$ but they don't affect the viability of the BLUE, $\hat{\hat{T}}_S$. See Woodruff, Lan (2004) for a more complete description of mail sampling and its particular difficulties. See Woodruff (2006) for a derivation of the BLUE, an unweighted separate ratio estimator.

This paper focuses on a special property of mail, "volume volatility", and its effect on these three estimators. Volume volatility, is a USPS term used to describe the day-to-day variability in the quantity of mail arriving at a processing facility for a given mail flow. A measure for day-to-day volume volatility is defined and its effect on the variance of the three estimators is derived. The variance of $\hat{T}_C$ and $\hat{T}_S$ increases with increasing volume volatility which forces the between day components of variance to increase. The variance of $\hat{\hat{T}}_S$ is unaffected by volume volatility. For most mail flows, day-to-day volume volatility is substantial and the day's total flow kilograms or numbers of containers cannot be accurately predicted.

It is shown that the viability of $\hat{T}_C$ and $\hat{T}_S$ requires day-to-day volume stability (roughly constant daily totals for kilograms of mail or numbers of mail containers). Volume stability is an unrealistic expectation in mail flows because of the complex nature of mail processing and transportation. On the other hand, the viability of $\hat{\hat{T}}_S$ is based on day-to-day stability of these first stage cluster **averages** (average number of pieces per container, average weight per container, and average total postage per container) and these daily averages tend to be quite stable regardless of how much the daily totals fluctuate within a flow-stratum.

Each estimator's variance is expressed as the sum of several variance components which isolate the sources of sampling error and quantify the role of volume volatility in the total sampling variance of each estimator. Sections 2 gives the variance formulae for each estimator under the stratified cluster sample design. Section 3 describes the simulation study that quantifies the size of the variances of these three estimators and the relative contributions to total variance of their variance components. These variances are graphed as functions of the volume volatility measure, Q, defined in Section 3.

2. Sample Design, Notation, and Variances

Each mail flow is partitioned into F strata (F may be several dozen to several hundred). Let U_f denote the set of mail containers in stratum f and let N_f for f=1, 2, 3, …….F be the number of mail containers in U_f. A stratum is typically the set of all mail containers passing through a processing office, moving by a transportation mode (air or surface, or SAL), in a particular container type (tray, tub or bag), within a particular mail class, and during a month. Estimates of total mail pieces, total kilograms, and total postage are required for mail flows each of which may consist of several hundred strata. For example, the flow of all letter post arriving by airmail from Great Britain during 2006 is a typical flow and its strata are defined by month (12), container type (3), and processing office (10) to give 360 strata.

Let D_f for f=1, 2, 3, …….F be the number of days in stratum f's month and let M_f be the set of days in stratum f's month. Let s_f be a simple random sample without replacement of days from M_f and let n_f be the number of days in s_f. Let U_{fd} be the set of mail containers for the mail flow in day d of stratum f and let N_{fd} be the number of mail containers in U_{fd} ($\sum_{d=1}^{D_f} N_{fd} = N_f$). Let s_{fd} be a simple random sample without replacement selected from U_{fd} and let n_{fd} be the number of containers in s_{fd}.

Let K_{fdj} be the weight in kilograms of the j^{th} container in U_{fd} and let π_{fdj} be the probability of selection of the j^{th} container in U_{fd}. Then $\pi_{fdj} = \frac{n_f}{D_f}\frac{n_{fd}}{N_{fd}}$. 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 for total kilograms of mail in U_f is $\hat{k}_f = \sum_{d \varepsilon s_f} \sum_{j \varepsilon s_{fd}} \frac{k_{fdj}}{\pi_{fdj}}$ 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, $K_f = \sum_{d=1}^{D_f}\sum_{j=1}^{N_{fd}} K_{fdj}$, then $K_f = E(\hat{k}_f)$. Define $\hat{y}_f$ and $Y_f = \sum_{d=1}^{D_f}\sum_{j=1}^{N_{fd}} Y_{fdj}$ similarly.

Let: $S^2_{fd}(y,k) = \frac{1}{N_{fd}-1}\sum_{j=1}^{N_{fd}}(Y_{fdj} - \bar{Y}_{fd})(K_{fdj} - \bar{K}_{fd})$ (2.1)

where $\bar{Y}_{fd} = \frac{1}{N_{fd}}\sum_{j=1}^{N_{fd}} Y_{fdj}$. Similarly for $\bar{K}_{fd}$.

Let: $S^2_f(y,k) = \frac{1}{D_f - 1}\sum_{d=1}^{D_f}(Y_{fd} - \bar{Y}_f)(K_{fd} - \bar{K}_f)$ (2.2)

where $K_{fd} = N_{fd}\bar{K}_{fd}$, $Y_{fd} = N_{fd}\bar{Y}_{fd}$, and $\bar{K}_f = \left(\frac{1}{D_f}\right)\sum_{d=1}^{D_f}\sum_{j=1}^{N_{dj}} K_{fdj}$ with $\bar{Y}_f$ defined similarly.

Under the design described above:

$Cov(\hat{y}_f, \hat{k}_f) =$

$$\frac{D_f}{n_f}\sum_{d=1}^{D_f}\frac{N^2_{fd}}{n_{fd}}\left(1 - \frac{n_{fd}}{N_{fd}}\right)S^2_{fd}(y,k) + \frac{D^2_f}{n_f}\left(1 - \frac{n_f}{D_f}\right)S^2_f(y,k) \quad (2.3)$$

The variance of $\hat{y}_f$, $V(\hat{y}_f)$, is similar with $S^2_{fd}(y,y)$ and $S^2_f(y,y)$ in place of $S^2_{fd}(y,k)$ and $S^2_f(y,k)$. Similarly for $V(\hat{k}_f)$. By independence of sampling in different strata, $Cov(\hat{y}_f, \hat{k}_l) = 0$ for all $f \neq l$.

To derive the variances of the three estimators, the following results are needed. Let $\hat{W}_f = \frac{\hat{k}_f}{\sum_{\alpha=1}^{F}\hat{k}_\alpha}$, the estimated kilogram proportion of the total mail flow kilograms in stratum f, $\frac{K_f}{K}$, where $K = \sum_{f=1}^{F} K_f$.

Expanding $\hat{W}_f$ in a Taylor Series about the expected values of the $\{\hat{k}_f\}$, the variance of $\hat{W}_f$ is approximately:

$V(\hat{W}_f) \doteq$

$$\left(\frac{K - K_f}{K}\right)^2\left(\frac{1}{K^2}\right)V(\hat{k}_f) + \left(\frac{K^2_f}{K^4}\right)\sum_{\substack{\alpha=1\\ \alpha\neq f}}^{F} V(\hat{k}_\alpha) \quad (2.4).$$

Let $\hat{\beta}_f = \frac{\hat{y}_f}{\hat{k}_f}$, the estimated rate per kilogram of the study variable y in stratum f, and let the actual rate be $\beta_f = \frac{Y_f}{K_f}$. Expanding $\hat{\beta}_f$ in a Taylor Series about the expected values of $\hat{y}_f$ and $\hat{k}_f$, ignoring terms of order 2 and greater, the variance of $\hat{\beta}_f$ is approximately

$$V(\hat{\beta}_f) = \frac{1}{K^2_f}V(\hat{y}_f) + \frac{Y^2_f}{K^4_f}V(\hat{k}_f) - 2\frac{Y_f}{K^3_f}Cov(\hat{y}_f, \hat{k}_f) \quad (2.5)$$

By a similar linearization,

$$Cov(\hat{W}_f, \hat{\beta}_f) = \frac{1}{K}\left(\frac{1}{K_f} - \frac{1}{K}\right)\left[Cov(\hat{k}_f, \hat{y}_f) - \beta_f V(\hat{k}_f)\right] \quad (2.6)$$

and

$$Cov(\hat{W}_i\hat{\beta}_i, \hat{W}_j\hat{\beta}_j) = \frac{1}{K^2}\left[\frac{Y_iY_j}{K^2}\sum_{f=1}^{F}V(\hat{k}_f) - \frac{Y_j}{K}Cov(\hat{y}_i, \hat{k}_i) - \frac{Y_i}{K}Cov(\hat{y}_j, \hat{k}_j)\right] \quad (i \neq j) \quad (2.7)$$

These are all the terms needed for $V(\hat{T}_C)$ and $V(\hat{T}_S)$.

$\hat{T}_C = K\frac{\sum_{f=1}^{F}\hat{y}_f}{\sum_{f=1}^{F}\hat{k}_f}$ is the combined ratio estimator for the mail flow total, $Y = \sum_{f=1}^{F} Y_f$ where $K = \sum_{f=1}^{F} K_f$ is known.

$$\hat{T}_C = K\sum_{f=1}^{F}\hat{W}_f\hat{\beta}_f \quad (2.8)$$

and

$$V(\hat{T}_C) = K^2\left[\sum_{f=1}^{F}V(\hat{W}_f\hat{\beta}_f) + \sum_{i=1}^{F}\sum_{\substack{j=1\\ j\neq i}}^{F} Cov(\hat{W}_i\hat{\beta}_i, \hat{W}_j\hat{\beta}_j)\right]$$

by Taylor Series Approximation:

$V(\hat{W}_f\hat{\beta}_f)$

$\doteq V\left[W_f\beta_f + \beta_f(\hat{W}_f - W_f) + W_f(\hat{\beta}_f - \beta_f)\right]$ and

$$V\left(\hat{T}_C\right) \doteq K^2\left[\sum_{f=1}^{F}\left(W_f^2 V\left(\hat{\beta}_f\right)+\beta_f^2 V\left(\hat{W}_f\right)+2W_f\beta_f Cov\left(\hat{\beta}_f,\hat{W}_f\right)\right)+\sum_{i=1}^{F}\sum_{\substack{j=1\\ j\neq i}}^{F} Cov\left(\hat{W}_i\hat{\beta}_i,\hat{W}_j\hat{\beta}_j\right)\right] \quad (2.9)$$

$\hat{T}_S = K\sum_{f=1}^{F} W_f\hat{\beta}_f$, the separate ratio estimator where $W_f = \frac{K_f}{K}$ and $V\left(\hat{T}_S\right) = K^2\sum_{f=1}^{F} W_f^2 V\left(\hat{\beta}_f\right)$ (2.10)

Finally, the variance of the BLUE will be derived under the design described above and evaluated with respect to repeated sampling under this design. The BLUE is denoted, $\hat{\hat{T}}_S$, and is defined as $\hat{\hat{T}}_S = K\sum_{f=1}^{F} W_f\hat{\hat{\beta}}_f$ 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 similarly for $\bar{k}_f$. The daily sample size of mail containers is dictated by workload constraints and is constant (or nearly so). Thus, $n_{fd} = n_{fd'}\ \forall\ d\ and\ d'$. Then $\bar{y}_f = \frac{1}{n_f n_{fd}}\sum_{d\varepsilon s_f}\sum_{j\varepsilon s_{fd}} y_{fdj}$ and similarly for $\bar{k}_f$. Under the sample design just described, $E\left(\bar{k}_f\right) = \frac{1}{D_f}\sum_{d=1}^{D_f}\frac{1}{N_{fd}}\sum_{j=1}^{N_{fd}} K_{fdj}$ and similarly for $E\left(\bar{y}_f\right)$. Then linearizing $\hat{\hat{\beta}}_f$,

$$V\left(\hat{\hat{\beta}}_f\right) \doteq V\left(\frac{\bar{y}_f}{E\left(\bar{k}_f\right)} - \frac{E\left(\bar{y}_f\right)}{E^2\left(\bar{k}_f\right)}\bar{k}_f\right) = \frac{1}{E^2\left(\bar{k}_f\right)}V\left(\bar{y}_f\right)+\frac{E^2\left(\bar{y}_f\right)}{E^4\left(\bar{k}_f\right)}V\left(\bar{k}_f\right)-2\frac{E\left(\bar{y}_f\right)}{E^3\left(\bar{k}_f\right)}Cov\left(\bar{y}_f,\bar{k}_f\right) \quad (2.11)$$

Where:

$$Cov(\bar{y}_f,\bar{k}_f) = \frac{1}{n_f D_f}\sum_{d=1}^{D_f}\frac{1}{n_{fd}}\left(1-\frac{n_{fd}}{N_{fd}}\right)S_{fd}^2(y,k) + \frac{1}{n_f}\left(1-\frac{n_f}{D_f}\right)\frac{1}{D_f-1}\sum_{d=1}^{D_f}\left(\bar{Y}_{fd}-\frac{1}{D_f}\sum_{i=1}^{D_f}\bar{Y}_{fd}\right)\left(\bar{K}_{fd}-\frac{1}{D_f}\sum_{i=1}^{D_f}\bar{K}_{fd}\right) \quad (2.12)$$

and similarly for $V\left(\bar{y}_f\right)$ and $V\left(\bar{k}_f\right)$. Then $V\left(\hat{\hat{T}}_S\right) = K^2\sum_{f=1}^{F} W_f^2 V\left(\hat{\hat{\beta}}_f\right)$.

The expressions for $V(\hat{T}_C)$ and $V(\hat{T}_S)$ are functions of the $\left\{V\left(\hat{y}_f\right)\right\}_{f=1}^{F}$, $\left\{V\left(\hat{k}_f\right)\right\}_{f=1}^{F}$, and $\left\{Cov(\hat{y}_f,\hat{k}_f\right\}_{f=1}^{F}$. The second term (between day variance) for each of these variances(covariances) that define $V(\hat{T}_C)$ and $V(\hat{T}_S)$ is a variance(covariance) of daily totals for the k and y variates over the month. Thus this second term is an increasing measure of day-to-day volume volatility and will drive up the variance of $V(\hat{T}_C)$ and $V(\hat{T}_S)$ as volume volatility increases.

The expression for $V(\hat{\hat{T}}_S)$ is a function of the $\left\{V\left(\bar{y}_f\right)\right\}_{f=1}^{F}$, $\left\{V\left(\bar{k}_f\right)\right\}_{f=1}^{F}$, and $\left\{Cov(\bar{y}_f,\bar{k}_f\right\}_{f=1}^{F}$. The second term of (2.12) (between day variance) for each of these variances(covariances) that define $V(\hat{\hat{T}}_S)$ is a variance(covariance) of daily means each month for the k and y variates. These daily means are quite stable and therefore this between day component of variance/covariance is not only small but also unaffected by volume volatility. Day-to-day volume changes occur through fluctuations in numbers of containers, not through the amount of mail each contains (most are nearly filled with mail).

These last two paragraphs explain much about the simulations and empirical computations in the next section where both $V(\hat{T}_C)$ and $V(\hat{T}_S)$ increase as volume volatility increases but $V(\hat{\hat{T}}_S)$ remains unaffected by volume volatility.

$V(\hat{T}_C)$ contains the term $K^2\sum_{f=1}^{F} V(\beta_f\hat{W}_f) =$

$$K^2 V\left(\sum_{f=1}^{F}\beta_f\hat{W}_f\right) - K^2\sum_{f=1}^{F}\sum_{\substack{j=1\\ j\neq f}}^{F} Cov\left(\beta_f\hat{W}_f,\beta_j\hat{W}_j\right),$$

substituting this into (2.9) and noting that by linearizing $\hat{\beta}_f$, we have that approximately

$$\sum_{f=1}^{F}\sum_{\substack{j=1\\ j\neq f}}^{F} Cov\left(\beta_f\hat{W}_f,\beta_j\hat{W}_j\right) \doteq \sum_{f=1}^{F}\sum_{\substack{j=1\\ j\neq f}}^{F} Cov\left(\hat{\beta}_f\hat{W}_f,\hat{\beta}_j\hat{W}_j\right)$$

. Then (2.9) can be written:

$V\left(\hat{T}_C\right)$

$$\doteq K^2 \sum_{f=1}^{F} \left(W_f^2 V\left(\hat{\beta}_f\right) + 2W_f \beta_f Cov\left(\hat{\beta}_f, \hat{W}_f\right) \right) + K^2 V\left(\sum_{f=1}^{F} \beta_f \hat{W}_f \right) \quad (2.13).$$

The covariance term, $2W_f \beta_f Cov\left(\hat{\beta}_f, \hat{W}_f\right)$, is approximately zero. This follows from (2.6) and the linear model that relates the $\{y_{fdj}\}$ and the $\{k_{fdj}\}$, $y_{fdj} = \beta_f k_{fdj} + \varepsilon_{fdj}$ where the $\{\varepsilon_{fdj}\}$ are pairwise uncorrelated and $\varepsilon_{fdj} \propto \left(0, k_{fdj}\sigma_f^2\right)$, see Woodruff (2006) for details. Under this linear relationship the approximation,

$$Cov\left(\hat{W}_f, \hat{\beta}_f\right) = \frac{1}{K}\left(\frac{1}{K_f} - \frac{1}{K}\right)\left[Cov(\hat{k}_f, \hat{y}_f) - \beta_f V\left(\hat{k}_f\right)\right]$$

$$\doteq \frac{1}{K}\left(\frac{1}{K_f} - \frac{1}{K}\right)\left[Cov(\hat{k}_f, \beta_f \hat{k}_f + \hat{\varepsilon}_f) - \beta_f V\left(\hat{k}_f\right)\right] = 0$$

holds

where $\hat{\varepsilon}_f$ is the Horwitz-Thompson estimator of zero applied to the sample $\{\varepsilon_{fdj}\}$. Using this result, both (2.9) and (2.13) can be further simplified:

$$V\left(\hat{T}_C\right) \doteq K^2 V\left(\sum_{f=1}^{F} W_f \hat{\beta}_f \right) + K^2 V\left(\sum_{f=1}^{F} \beta_f \hat{W}_f \right) \quad (2.14)$$

(2.14) explains the behaviour of $\hat{T}_C$ that is observed in Table 1 of Section 3. As volume volatility increases, $V\left(\hat{T}_C\right)$ increases with it but most of this increase comes from the second term in (2.14). This second term accounts for roughly 90% of $V\left(\hat{T}_C\right)$ for the levels of volume volatility typically encountered in mail flows. The first term will tend to be inherently small due to the correlation between numerator and denominator of each $\{\hat{\beta}_f\}$. No such correlation will limit the size of the second term.

$\sum_{f=1}^{F} \beta_f \hat{W}_f$ is the randomly weighted average (weights are random variables, each positive, and they add to one) of the $\{\beta_f\}_{f=1}^{F}$. $\sum_{f=1}^{F} \beta_f \hat{W}_f$ is always in the interval $[\beta_{\min}, \beta_{\max}]$ where $\beta_{\min} = \min_{1 \le f \le F} \{\beta_f\}$ and $\beta_{\max}$ is the maximum of the same set. This constrains the variance of $\sum_{f=1}^{F} \beta_f \hat{W}_f$ to be small when $[\beta_{\min}, \beta_{\max}]$ is short but opens the possibility for it to be large when $[\beta_{\min}, \beta_{\max}]$ is long. For many mail flows this interval is relatively long. For the mail flow considered in section 3, the $\{\beta_f\}_{f=1}^{F}$ vary from near one to over 50. This characteristic of a mail flows causes $V(\hat{T}_C)$ to grow rapidly with volume volatility through the increase in $K^2 V\left(\sum_{f=1}^{F} \beta_f \hat{W}_f \right)$. $K^2 \sum_{f=1}^{F} V\left(W_f \hat{\beta}_f\right)$ is less affected by volume volatility, due surely to the relative stability of the ratios of correlated variables, the $\{\hat{\beta}_f\}$. It is shown in section 3 that $K^2 \sum_{f=1}^{F} 2W_f \beta_f Cov\left(\hat{\beta}_f, \hat{W}_f\right)$ is inconsequentially small as also indicated analytically above.

3. Simulations and Computational Analysis

The studies here are based on sampling mail flows that differ only in degree of volume volatility as measured by

$$Q = \frac{1}{F}\sum_{f=1}^{F} \frac{1}{D_f - 1} \sum_{d=1}^{D_f} (N_{fd} - \overline{N}_f)^2$$ where

$\overline{N}_f = \frac{1}{D_f}\sum_{d=1}^{D_f} N_{fd}$ with the notation defined in Section 2.

Let MF(Q) denote the mail flow with a volume volatility measure of Q. Each mail flow population, MF(Q) is identical to the others except for the assignment of a stratum's mail containers to days in the month. The mail flow used to construct these populations is all letter post airmail from Great Britain to the US between January 2004 and September 2004. This mail flow consisted of about 43.7 million pieces and about 700,000 containers and was stratified into 150 strata by USPS facility where it entered the US, the container type in which it arrived, and the month in which it arrived. Similar simulations and computations were done with other mail flows and similar results were obtained. Great Britain airmail was documented here because it was the largest and most diverse mail flow (most arrival facilities and container types represented).

There are two studies described here. The first is a set of 20 simulation studies conducted on each of the 20 members of $\{MF(Q)\}$ where Q ranged from 0 to about 170,000. For each of these 20 simulations, sampling and

estimation of total number of mail pieces is replicated 500 times following the procedures described in Section 2 (y_{fdj} is total number of mail pieces in container j, on day d, in stratum f). For each MF(Q), these 500 replications produce 500 independent estimates of total pieces for each estimator: $\hat{T}_C$, $\hat{T}_S$, $\hat{\hat{T}}_S$, and $\hat{y}_{HT} = \sum_{f=1}^{F} \hat{y}_f$. These 500 replicate estimates are used to estimate bias and variance for the four estimators for each of the 20 values of Q. This provides a picture of each estimator's behaviour as Q (volume volatility) increases.

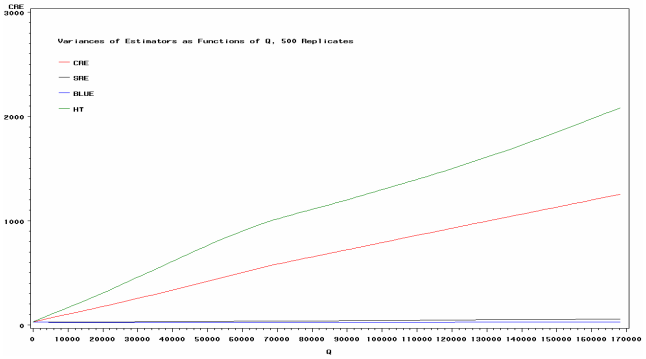
The second study uses the formulae derived in Section 2 to compute the variance of each estimator directly from the known population parameters for each MF(Q). This second approach permits the variance of each estimator to be broken down into a sum of major components so that the components contributing to variance can be isolated and quantified as functions of Q. These two approaches to computing variance should give similar results and therefore also serve as a check on one another.

In the first set of simulation studies with the 500 replicates, a degree of randomness was built into sample selection to capture variations in both numbers of days per month sampled and numbers of containers sampled per day that characterize actual conditions experienced in sample selection and data collection. The average total sample size in the 20 simulation studies on each of the $\{MF(Q)\}$ was about 2,200. This total of about 2,200 varied because these simulations randomly selected 3 to 5 days per month and 3 to 5 containers per day. For this reason, the simulation studies produce variance estimates that are about 10% to 15% larger than those derived directly from theory in Section 2 where exactly 4 sample days per month and 4 sample containers per sample day were assumed - 16 containers per stratum month. For this second study, the total sample size was about 2,400.

Graphs 1 and 2 show the variances (vertical axis) of the 4 estimators as functions of Q (horizontal axis). Graph 1 is from the 20 simulation studies of 500 sample replicates for each MF(Q) and Graph 2 is from the population parameters derived in Section 2. These graphs are smoothed representations of scatter plots and show the straight line relationship between Q and the variances of the four estimators, a relationship somewhat clouded by the unsmoothed scatter plots. The vertical scale units in Graph 1 and Graph 2 are in 10^{10} of $(Mail \quad Pieces)^2$. In Graphs 1 and 2, CRE=$\hat{T}_C$, SRE=$\hat{T}_S$, BLUE=$\hat{\hat{T}}_S$, and HT=Horwitz-Thompson Estimator= $\sum_{f=1}^{F} \hat{y}_f$. The variance components in Table 1 are in 10^{10} of $(Mail \quad Pieces)^2$

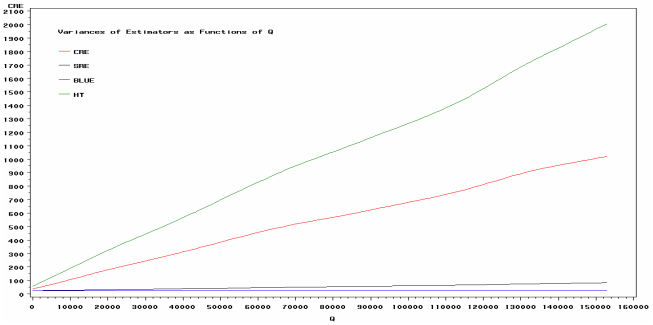
GRAPH 1

Variances as functions of Q, CRE(Red), SRE(Black), BLUE(blue)& HTGreen), 500 Reps



GRAPH 2

Variances as functions of Q, CRE (Red), SRE (Black), BLUE (blue)& HT Green)



The components of variance are summarized in Tables 1, isolate the main contributors to the variance of the combined ratio estimator, and confirm that the main contributor to $V(\hat{T}_C)$ as Q increases, is the substantial difference between strata of the rates per kilogram of the study variables (the sum the third, fourth, and fifth columns). The Covariance Between Strata (Table 1 , Column 5) which is always negative fails to compensate for the rapid increase of column 3 and the net result is the rapid increase in total variance of the combined ratio estimator with increasing Q (Table 1 , Column 6).

$$V(\hat{T}_C) = K^2[\sum_{f=1}^{F}(W_f^2 V(\hat{\beta}_f) + \beta_f^2 V(\hat{W}_f) + 2W_f\beta_f Cov(\hat{\beta}_f, \hat{W}_f)) + \sum_{i=1}^{F}\sum_{\substack{j=1 \\ j\neq i}}^{F} Cov(\hat{W}_i\hat{\beta}_i, \hat{W}_j\hat{\beta}_j)]$$

(Col 6, Table 1)

Separate Variance Component $= V\left(\hat{T}_S\right)$

$$= K^2 \sum_{f=1}^{F} W_f^2 V\left(\hat{\beta}_f\right)$$

Randomly Weighted Variance Component

$$= K^2 \sum_{f=1}^{F} \beta_f^2 V\left(\hat{W}_f\right)$$

Covariance $Cov(\hat{\beta}_f, \hat{W}_f)$ Component

$$= K^2 \sum_{f=1}^{F} 2W_f \beta_f Cov\left(\hat{\beta}_f, \hat{W}_f\right)$$

Covariance Between Strata Component

$$= K^2 \sum_{i=1}^{F} \sum_{\substack{j=1 \\ j \neq i}}^{F} Cov\left(\hat{W}_i \hat{\beta}_i, \hat{W}_j \hat{\beta}_j\right).$$

Table 1 Variance Components for the combined Ratio HT Estimator

Q (Volume Volatility)	Separate Variance Component	Randomly Weighted Variance Component	Covariance $Cov(\hat{\beta}_f, \hat{W}_f)$ Component	Covariance Between Strata Component	Total Variance of Combined Ratio HT
0.0	**23.1 ***	33.5	-0.4	-20.5	35.7
164.0	23.3	35.2	-0.4	-21.2	36.9
272.0	23.2	37.2	-0.3	-22.4	37.8
592.0	23.5	39.9	-0.6	-23.0	39.7
1002.0	23.7	45.8	-0.3	-26.5	42.6
2256.0	24.2	61.7	-0.4	-35.2	50.3
5518.0	25.3	97.9	-0.8	-48.6	73.7
7362.0	25.5	108.1	-0.7	-50.2	82.8
9984.0	27.2	155.8	-0.6	-81.3	101.1
15062.0	29.2	216.0	0.3	-118.7	126.7
19768.0	31.6	299.7	-0.3	-156.1	174.9
22846.0	32.9	345.5	-2.0	-133.0	243.4
39318.0	36.4	429.7	-0.9	-212.3	252.8
40424.0	40.0	551.5	1.0	-226.1	366.4
70772.0	52.0	978.5	0.8	-437.0	594.4
91188.0	55.3	1047.4	-3.1	-453.7	645.8
116588.0	67.3	1408.5	0.6	-701.1	775.3
120246.0	66.6	1307.9	-2.8	-591.0	780.8
127920.0	74.4	1580.9	-3.2	-741.5	910.6
129552.0	76.0	1686.3	5.7	-839.5	928.4
136302.0	75.2	1675.6	-3.7	-755.7	991.3
152836.0	83.8	1872.8	-3.8	-941.9	1010.9
184370.0	96.9	2340.4	-1.6	-1225.3	1210.4

*Also variance of the BLUE for all Volatilities

Column 2 of Table 1, Separate Variance Component, is also the total variance of $\hat{T}_S$ and the first entry of this column (Q=0) is the variance of the BLUE for all values of Q.

Recall from Section 2 that the row sums of columns 3, 4, and 5 in Table 1 are also equal to:

$$K^2 V\left(\sum_{f=1}^{F} \beta_f \hat{W}_f\right) \doteq K^2 \sum_{f=1}^{F} \beta_f^2 V\left(\hat{W}_f\right)$$

$$+ K^2 \sum_{f=1}^{F} 2W_f \beta_f Cov\left(\hat{\beta}_f, \hat{W}_f\right)$$

$$+ K^2 \sum_{i=1}^{F} \sum_{\substack{j=1 \\ j \neq i}}^{F} Cov\left(\hat{W}_i \hat{\beta}_i, \hat{W}_j \hat{\beta}_j\right).$$

The populations used for these simulations are clean and the sample design does not fully capture many of the complex and unpredictable elements of final stage (within day) 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.

Q is roughly proportional to two things, the square of the total number of containers in the mail flow stratum and the relative day-to-day volatility within strata defined next as QR_f. Let p_{fd} be the proportion of the stratum f's mail that is processed in day d and $\bar{p}_f = \frac{1}{D_f} \sum_{d=1}^{D_f} p_{fd}$ then $QR_f = \frac{1}{D_f} \sum_{d=1}^{D_f} (p_{fd} - \bar{p}_f)^2$.

The relationship between Q and the ratio, R = $V\left(\hat{T}_C\right) \Big/ V\left(\hat{\hat{T}}_S\right)$ as a function of mail flow stratum size is a complex one since QR_f probably decreases as mail flow size (as measured by number of containers) increases. Relative volatility and flow size should seldom be expected to cancel each other in Q. The same simulation study documented here was run on airmail from Belgium. This mail flow is about 2% the size of Great Britain's. For Belgium, R attained values of over 100 compared to Great Britain's largest R-values of 40 to 50 in the lower part of Table 1.

4. Conclusions

The USPS uses the Combined Ratio HT Estimator ($\hat{T}_C$) for many of its mail volume estimates but characteristics of mail flows and the volume volatility that mail processing and transportation impose on them make this estimator a particularly unfortunate choice. The manner in which mail containers are filled impose a model on the sample data and under this model there is a Best Linear Unbiased Estimator ($\hat{\hat{T}}_S$ or BLUE). This model describes the approximate proportionality of container

study variables to container kilogram weight within strata (stability across containers of average container pieces per kilogram and postage per kilogram) independent of volume volatility and this stability of rates per kilogram implies that the variance of the BLUE is not affected by volume volatility (Q). For moderate values of Q, the variance of the BLUE is a tiny fraction of that for the Combined Ratio HT Estimator.

This BLUE provides an alternative to HT based methodologies when design control is problematic. This lack of design control is forced by day-to-day volume volatility in mail sampling where unpredictability forced by weather, transportation, and processing creates large and unpredictable fluctuations in the day cluster sizes (number of containers in the mail flow each day). Equation (2.14) establishes that the Combined Ratio HT Estimator, $\hat{T}_C$, can be particularly sensitive to volume volatility when there are large differences between strata in the rates per kilogram of the study variables.

The empirical studies in Section 3, quantify the variances of four estimators (including the straight HT estimator) as functions of Q. These differences increase linearly with increasing day-to-day volume volatility but even with no volume volatility they are substantial - the Combined Ratio HT Estimator's variance is about 50% larger than that of the BLUE when Q=0 in Table 1. These differences increase with increasing volume volatility until they are truly extraordinary, over 4000% for the mail flow studied in Section 3. This is due to the concurrence of the following things: volume volatility, homogeneity of rates per kilogram within strata, and the heterogeneity of these rates between strata.

It is shown in Section 2 that the viability of the Combined Ratio HT Estimator depends upon day-to-day stability (Q=0) of mail volumes (daily total kilograms, total pieces, etc) within each of the mail flow strata. Similarly, the viability of the BLUE depends upon the stability of day-to-day averages (average container weight, average number of mail pieces per container, and average postage per container) within each mail flow stratum. Because of the way mail containers are filled, processed, and transported, any assumption of day-to-day volume stability is widely violated while day-to-day stability of these container averages is substantially assured.

As long as mail sample design involves sample clusters defined in terms of fixed time intervals (days for most mail surveys), there will be a variance component in a Horwitz-Thompson based estimator that is an increasing measure of the variability of mail volumes between these time intervals (day-to-day volume variability in mail surveys). This variance component can potentially dominate sampling error until such estimators approach white noise. Lengthening the time interval would reduce this variance but would involve an increase in processing time that would violate the fundamental goal of mail processing – minimization of time and cost. It is better to deal with this situation through an alternative estimator like the BLUE, $\hat{\hat{T}}_S$, under a model imposed by established features of mail populations.

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 (Woodruff , 2006) selected from the flow can be modeled as a simple random sample from the totality of atoms in the stratum. This holds for certain biological populations, for example, the sampling of rivers for their particulate or microbial content.

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

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