Variance Estimation for Establishment Surveys: An Introductory Overview

Phillip S. Kott

This overview begins by discussing the uses of models in establishment surveys. These range from providing some statistical structure when using estimation strategies with cutoff or convenience samples to improving the accuracy of more robust strategies based on probability sampling. In both situations, a statistician usually constructs an estimator unbiased (or nearly so) under a simple, yet plausible, linear model. A model-based measure of the estimator's variance *given a particular sample* can be estimated. Alternatively, one can measure the estimator's variance under the combination of the model and the random-sampling mechanism. Of particular interest are ratio estimators that are both unbiased under certain assumed models and nearly unbiased under the random-sampling mechanism. Robust model-based variance estimators for a ratio often possess better coverage properties than the conventional probability-based alternative. Results for the ratio extend to calibration estimators. Models can also be helpful in variance estimation when there is unequal-probability sampling within strata, additional phases of sample selection, reweighting for unit nonresponse, and imputation for item nonresponse.

KEY WORDS: Anticipated variance, calibration, finite population correction, nonresponse, ratio, robust.

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030. This paper was prepared for the third International Conference on Establishment Surveys, June 2007, Montreal, Quebec

1. Introduction: Why Do We Need Models?

It was once a common practice to base establishment surveys on nonrandom samples such as cut-off and convenience samples. Estimates were derived using purely *ad-hoc* methods, leading to inferences that were often dubious and never testable.

A *cut-off sample* of a population of N establishments consists of those $n < N$ establishments believed to be the largest in some sense. A *convenience sample* is based on expediency, containing n establishments that are easy to contact or eager to participate in the survey.

Scientifically defensible inferences are possible with a nonrandom sample when they are based on a plausible model of survey-variable behavior. By clearly stating his (her) model assumptions, the statistician often possesses tools for assessing the quality of the resulting inferences.

Nowadays, most establishment surveys, especially those conducted by governments, are based on randomly selected samples. Estimates are derived from such samples using probability-sampling theory.

A limitation of this theory is its asymptotic nature. With small samples, even random ones, models are needed for reasonable estimation. The skeptical reader is directed to Basu's infamous elephant example (Basu, 1971), where a random sampling scheme is designed to select what is believed to be the most representative of 50 elephants with a 99% probability. The resulting estimate for the total weight of the 50 elephants, although unbiased according to probability-sampling theory, has a 100% probability of being disastrously bad.

An *estimation strategy* is a sampling plan, usually random, coupled with an estimator. Even when a survey statistician intends to draw a random sample of respectable size and to compute an estimator with good probability-sampling-theory properties, invoking a plausible model can be of great assistance in choosing among alternative estimation strategies.

There are other ways models can be useful. Estimation strategies incorporating unequal probability sampling within strata often feature estimators

with hard-to-estimate measures of accuracy (variances and mean squared errors). Assuming a model can simplify this task considerably.

Coverage intervals derived from probability-sampling theory need to be based on large, sometimes very large, samples to justify the normal approximation. A model can speed up the asymptotics with middling-sized samples, improving the accuracy of calculated coverage intervals.

Finally, models are particularly helpful when adjusting for nonresponse and for errors in the frame (duplication and undercoverage). Models provide both methods for treating nonresponse and frame errors and means for assessing the accuracy of those methods.

2. Model-based Estimators (for totals)

To keep things simple, we focus this discussion on estimating a total for a finite population, U, in a situation where the population and sampling frame coincide. We further assume that there is no nonresponse (for now) and that all survey respondents provide correct information upon enumeration.

Our goal is to estimate the population total $T_y = \sum_{k \in U} y_k = \sum_U y_k$ with a sample, S, of n elements (establishments). Suppose there is an auxiliary variable, $x_k \in S$, about which $T_x = \sum_U x_k$ is known. The auxiliary-variable total may come from administrative data or a census. A commonly used auxiliary variable, although not often the best choice, is x_k =1 for all elements in the population.

We will use the following motivating example throughout this discussion. Consider a population U of farms, where
x_k is the total land (believed to be) on farm *k*, and
y_k is the planted corn acres on the farm.

Our goal is to estimate the total acres of planted corn in the population based on a sample. The total-land value is assumed to be known before

sampling and enumeration. *It cannot be changed by survey information* even if found to be in error for a particular farm k.

A characteristic of establishment surveys rarely shared by demographic surveys is the frequent existence of measures of size for members of the population whether in the sample of not. The total-land variable in our motivating example is one such measure of size. A plausible model can often be assumed relating a survey variable of interest, planted corn in the motivating example, to one or more of these measures of size. We will focus here on estimation strategies incorporating a single auxiliary variable/measure of size both because they are commonly used in practice and because they are mathematically simpler than strategies employing multiple auxiliary variables.

If we assume that each $y_k \in U$ satisfies the simple linear *prediction* model:

$$y_k = \beta x_k + \varepsilon_k, \tag{1}$$

where $E(\varepsilon_k \mid x_k) = 0$, then the ratio estimator,

$$t_y^r = T_x \frac{\sum_S y_k}{\sum_S x_k} = T_x \frac{\bar{y}_S}{\bar{x}_S}, \tag{2}$$

is a (model) unbiased estimator (predictor) for T_y in the sense that $E_\varepsilon(t_y^r - T_y) = 0$ *as long as the sampling mechanism is ignorable* . The sampling mechanism is ignorable when the distribution of $\varepsilon_k \mid x_k$ is that same whether or not $k \in S$. This means (among other things) that $E(\varepsilon_k \mid x_k) = E(\varepsilon_k \mid x_k;\, k \in S) = 0$

The ignorability assumption is the reason why the x-value cannot be changed by the survey: $\varepsilon_k = y_k - \beta x_k$ must be the same whether or not establishment k is in the sample. The assumption is unnecessary for simple random samples and cutoff samples based on the elements with the largest x-values (as long as population values aren't changed by survey information), but is needed for convenience samples.

3. Model Groups

Some populations naturally separate into G mutually exclusive groups, each denoted by U_g, where the behavior of the y-variable is consistent within each group but (may) vary across groups. For example, a population of farms may naturally be separated into agricultural districts, each with distinct soil and weather characteristics.

When the group x-totals, $T_{xg} = \sum_{U_g} x_k$, are known, it is possible to compute the *separate ratio estimator*:

$$t_y^{sr} = \sum_{g=1}^{G} T_{xg} \frac{\sum_{S_g} y_k}{\sum_{S_g} x_k} = \sum_{g=1}^{G} T_{xg} \frac{\sum_{S_g} y_k / n_g}{\sum_{S_g} x_k / n_g} = \sum_{g=1}^{G} T_{xg} \frac{\bar{y}_{gS}}{\bar{x}_{gS}} , \tag{3}$$

where S_g is that part of the sample from group G. This estimator is not only unbiased for T_y under the simple linear model in equation (1), but also the *group-ratio model*:

$$y_k = \beta_g x_k + \varepsilon_k \quad \text{for} \quad k \in U_g , \tag{4}$$

where $E(\varepsilon_k \mid x_k) = E(\varepsilon_k \mid x_k ; k \in S) = 0$. The simple ratio estimator t_y^r in equation (2) is generally *not* unbiased for T_y under this model.

When all the x_k =1, the group-ratio model is called the "group-mean model." The *separate ratio estimator* has the familiar form:

$$t_y^{PS} = \sum_{g=1}^{G} N_g \frac{\sum_{S_g} y_k}{n_g} = \sum_{g=1}^{G} N_g \bar{y}_{gS} .$$

Although the G groups look like H design strata, the difference is that samples need not be selected independently across groups. A popular name for t_y^{PS} is the "post-stratified estimator" (hence the superscript "PS") so named because groups are often formed after the survey design stage where the strata are created.

4. Prediction Form

Returning to the simple linear model in equation (1), an estimator for T_y having the form:

$$t_y^{pred} = \sum_S y_k + \sum_{U-S} x_k \hat{\beta}$$

is model unbiased when $\hat{\beta}$ is an unbiased estimator for β. This estimator uses the model to *predict* the y-values for the nonsampled elements. (Purist would say that since T_y is itself a random variable under the model, t_y^{pred} is more properly called a "predictor" rather than an "estimator" for T_y.)

If we assume the ε_k are uncorrelated and each has variance σ_k^2, then the best linear unbiased (BLU) estimator for T_y is

$$t_y^{BLU} = \sum_S y_k + \sum_{U-S} x_k \frac{\sum_S \left(x_j y_j / \sigma_j^2 \right)}{\sum_S \left(x_k^2 / \sigma_j^2 \right)}.$$

With some work, one can show that this collapses into the ratio estimator if and only if $\sigma_k^2 \propto x_k$ (so that $\hat{\beta} = b = \sum_S y_j / \sum_S x_j$).

5. Model-assisted Randomization-based Estimators

Unfortunately, models can fail. In particular, the ratio $r_k = \frac{y_k}{x_k}$ may tend to increase or decrease along with x_k. In our motivating example, the corn-to-land ratio on a farm may tend to increase along with the size of the farm (smaller farms tend to concentrate on more specialized crops than corn). Nevertheless, from a probability-sampling viewpoint, t_y^r under simple random sampling is nearly unbiased. In fact, one can show its relative mean squared error under probability-sampling theory will tend to zero as the sample size grows arbitrarily large under mild conditions. This property is called "randomization (or design) consistency." Henceforth, we will use the modifiers "randomization" and "randomization-based" when referring to properties derived under probability-sampling theory. Although the asymptotic properties of strategies under probability-sampling (randomization) theory are central to many of the methods discussed here, a more thorough discussion of them is beyond the scope of this paper.

Let $B = T_y / T_x$ be the corn-to-land acres ratio in the population and $\bar{y}_U = T_y / N$ be the average size of a farm in land acres. When the element-by-element values of $e_k = y_k - Bx_k$ tend to be smaller than those of $y_k - \bar{y}_U$, the ratio estimator will likely have less randomization mean squared error than the expansion estimator, $t_y^e = \frac{N}{n} \sum_S y_k = N\,\bar{y}_S$. In other words, $y_k = \beta x_k + \varepsilon_k$, however imperfect, is a better implicit prediction model than $y_k = \mu + \varepsilon_k$, and the ratio estimator a better estimator for the population total under simple random sampling than the expansion estimator.

There are more than two ways to estimate a total in a nearly randomization unbiased fashion based on a sample of a specified size. Without a model of survey-variable behavior, it is often impossible to choose among them. As a result, it is often helpful to:

1. Restrict attention (if possible) to nearly randomization-unbiased or randomization-consistent estimation strategies, and

2. Look at the model expectation of the randomization mean squared error (*mse*) of these strategies.

We will call the model expectation of the randomization mse of an estimator its "combined mse" (the term "anticipated variance" or "anticipated *mse*" is also commonly used).

Let π_k denote the selection probability of establishment *k*. Suppose we want to estimate T_y with an arbitrary random sample and a ratio of two randomization-unbiased expansion estimators. This is known as the *Hajek estimator*:

$$t_y^{Hajek} = T_x \frac{\sum_S \frac{y_k}{\pi_k}}{\sum_S \frac{x_k}{\pi_k}} = T_x \frac{t_y^e}{t_x^e}. \tag{5}$$

Each π_k can take on any value in the interval (0, 1] (this interval excludes 0 but includes 1). Under mild conditions, the Hajek estimator is nearly randomization unbiased.

Suppose we want to find the set of selection probabilities that produces the t_y^{Hajek} with the smallest randomization *mse* given an (expected) sample size n. Formally, this sample-size restriction means that $\sum_U \Pr(k \in S) = \sum_U \pi_k = n$. This task would be impossible unless we assume a structure for survey-value behavior. By assuming the linear model in equation (1) and further that the ε_k are uncorrelated, each with variance σ_k^2, one can show that the combined *mse* of the Hajek estimator is approximately

$$E_{\varepsilon}\left\{E_{rand}\left[\left(t_y^{Hajek}-T_y\right)^2\right]\right\} \approx \sum_U \left(\tfrac{1}{\pi_k}-1\right)\sigma_k^2. \tag{6}$$

This approximation works especially well when the sample size is on the large side (say n > 20).

The large-sample combined *mse* on the right hand size of equation (6) is minimized under *Brewer selection*:

$$\pi_k = n\frac{\sigma_k}{\sum_U \sigma_j} \quad .$$

This allocation requires $\sigma_k \le \sum_U \sigma_j / n$ for all elements in the population. Any k for which this inequality fails should be selected with certainty and Brewer selection applied to the remaining elements of the population.

Although the σ_k need only be known up to a constant to use Brewer selection, even that is unusual in practice. Often, it is speculated in establishment surveys that $\sigma_k \propto x_k^{\gamma}$ for some γ between 1/2 and 1 (equivalently, $\sigma_k^2 \propto x_k^{2\gamma}$). It is important to recognize that the *frame value* x_k must be known for all elements in U to use this speculation.

In our motivating example, γ = 1 means that the corn-to-land ratio, r_k, acts like a random variable with a common mean, β, and variance, *c*, across independent farms. Consequently, $E(y_k) = E(x_k r_k) = x_k\beta = \beta x_k$, and $\sigma_k^2 = V(\varepsilon_k) = V(y_k) = V(x_k r_k) = x_k^2 c = c x_k^2$. At the other extreme, γ = 1/2 means that r_k acts like a random variable with a common mean and variance across independent *acres of land*, so that $E(y_k) = \beta x_k$ again, but now $\sigma_k^2 = V(y_k) = \sum_{i=1}^{x_k} V(r_k) = x_k c = c x_k$.

Any assumption about the σ_k is speculative, even under the most ideal conditions (such as the frame x-values never being in error). Fortunately, a faulty assumption about the σ_k does not diminish the randomization consistency or model unbiasedness of the Hajek estimator.

On the other hand, insisting that the estimation strategy be randomization consistent rules out systematic probability-proportional-to-σ_k sampling from an *purposefully-ordered* list since such a designed effectively produces a cluster sample of size one. It also rules out using a separate ratio estimator (equation (3)) under stratified simple random sampling (*stsrs*) when the samples sizes within strata (the groups of equation (3)) are too small.

A very popular estimation strategy for establishment surveys is the combined ratio estimator under *stsrs*:

$$t_y^{cr} = T_x \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{S_h} y_k}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{S_h} x_k}. \tag{7}$$

It is special case of the Hajek estimator in equation (5) with $\pi_k = n_h / N_h$ when k is a sampled establishment in stratum h (denoted S_h). The strategy is randomization consistent and model unbiased under the simple linear model in equation (1) but not the group-ratio model in (4).

6. Estimating the Model Variance of t_y^r

Observe that under the simple linear model:

$$\begin{aligned} t_y^r - T_y &= \sum_U x_k \frac{\sum_S y_k}{\sum_S x_k} - \sum_U y_k \\ &= \sum_U x_k \frac{\sum_S (\beta x_k + \varepsilon_k)}{\sum_S x_k} - \sum_U (\beta x_k + \varepsilon_k) \\ &= \frac{N\bar{x}_U}{n\bar{x}_S} \sum_S \varepsilon_k - \sum_U \varepsilon_k, \end{aligned}$$

which explains why t_y^r is model unbiased for T_y. If the ε_k are uncorrelated and each has variance σ_k^2, then the model variance of t_y^r (as an estimator for T_y) is

$$E_\varepsilon\left[\left(t_y^r - T_y\right)^2\right] = \left(\frac{N\,\bar{x}_U}{n\bar{x}_S}\right)^2 \sum_S \sigma_k^2 - 2\frac{N\,\bar{x}_U}{n\bar{x}_S}\sum_S \sigma_k^2 + \sum_U \sigma_k^2, \tag{8}$$

which can be estimated in a unbiased fashion if we knew an unbiased estimator for the σ_k^2.

If we can assume $\sigma_k^2 \propto \upsilon_k$ for a known set of υ_k (say x_k^γ for some γ), then

$$\hat{\sigma}_k^2 = \frac{1}{n-1}\sum_{i\in S}\left(\frac{y_i}{\sqrt{\upsilon_i}} - \frac{x_i}{\sqrt{\upsilon_i}}\frac{\sum_{j\in S}\left(x_j y_j / \upsilon_j\right)}{\sum_{j\in S}\left(x_j^2 / \upsilon_j\right)}\right)^2 \upsilon_k$$

is an unbiased estimator for σ_k^2

Since assumptions about the σ_k^2 are very often dubious, a more robust approach to modeling variance estimation is advisable. For $k \in S$, let

$$s_k^2 = \left[\frac{\bar{x}_S}{\bar{x}_S - \frac{1}{n}x_k}\right]\left(y_k - \frac{\bar{y}_S}{\bar{x}_S}x_k\right)^2 = \left[\frac{n}{n - \left(x_k / \bar{x}_S\right)}\right]\left(y_k - \frac{\bar{y}_S}{\bar{x}_S}x_k\right)^2.$$

This is an unbiased estimator for σ_k^2 when $\sigma_k^2 \propto x_k$, and nearly (asymptotically) unbiased otherwise when n is large because

$$s_k^2 \approx \left(y_k - \frac{\bar{y}_S}{\bar{x}_S}x_k\right)^2 = r_k^2 \approx \left(y_k - \beta x_k\right)^2 = \varepsilon_k^2 \;.$$

Observe that $\sigma_k^2 \propto x_k$ implies $\frac{N\,\bar{x}_U}{n\bar{x}_S}\sum_S \sigma_k^2 = \sum_U \sigma_k^2$. Consequently,

$$v_{\varepsilon}^{robust} = \left(\frac{N\,\bar{x}_U}{n\bar{x}_S}\right)^2 \sum_S s_k^2 - \frac{N\,\bar{x}_U}{n\bar{x}_S} \sum_S s_k^2. \tag{9}$$

is an exactly unbiased as an estimator for the model variance of t_y^r when $\sigma_k^2 \propto x_k$. Moreover, v_{ε}^{robust} is almost unbiased as long as both n and N/n are large, in which case s_k^2 can be approximated by r_k^2 (or $\frac{n}{n-1} r_k^2$), and v_{ε}^{robust} by

$$v_{\varepsilon}^{approx} = \left(\frac{N\,\bar{x}_U}{n\bar{x}_S}\right)^2 \sum_S r_k^2 .$$

N/n large means that the sampling fraction, n/N, is small. This is often, but not always, the case with establishment surveys.

7. Estimating the Randomization *MSE* of the Hajek Estimator

A conventional randomization *mse* estimator for the Hajek under an arbitrary sampling design is

$$\begin{aligned} v_1 &= \sum_{k,j\in S} \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} \frac{r_k}{\pi_k} \frac{r_j}{\pi_j} \\ &= \sum_{k\in S} (1-\pi_k) \left(\frac{r_k}{\pi_k}\right)^2 + \sum_{\substack{k,j\in S \\ k\neq j}} \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} \frac{r_k}{\pi_k} \frac{r_j}{\pi_j}, \end{aligned} \tag{10}$$

where $r_k = y_k - \left(t_y^e / t_x^e\right) x_k$, and π_{kj} is the probability that both elements k and j are in the sample (by definition, $\pi_{kk} = \pi_k$). This estimator is nearly randomization unbiased under mild conditions so long as π_{kj} is positive for all pairs of elements in the population. That condition rules out systematic (equal or unequal)

probability sampling from a purposefully ordered list because, except under rare circumstances, it is impossible to jointly select adjacent elements on the list.

The *Horvitz-Thompson mse* estimator in equation (10) can be re-expressed for the combined ratio estimator under *stsrs*, t_y^{cr}, as

$$v_1^{cr} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{n_h}{n_h - 1} \sum_{k \in S_h} \left(r_k - \bar{r}_{hS} \right)^2. \tag{11}$$

For the special case of t_y^r under *srs*, H = 1, and $\bar{r}_{1S} = 0$.

The expression v_1^{cr} also estimates the (large-sample) combined *mse* of t_y^{cr}:

$$E_\varepsilon \left\{ E_{rand} \left[\left(t_y^{cr} - T \right)^2 \right] \right\} \approx \sum\nolimits_U \left(\frac{1}{\pi_k} - 1 \right) \sigma_k^2$$
$$= \sum_{h=1}^{H} \left(\frac{N_h}{n_h} - 1 \right) \sum_{k \in S_h} \sigma_k^2$$

under the simple linear model when the ε_k are uncorrelated. This is a special case of equation (6).

With many sampling designs, we cannot always avoid incorporating π_{kj} terms for distinct k and j when estimating the randomization *mse* of a Hajek estimator. By assuming the ε_k are uncorrelated, however, the *combined mse* of the general Hajek can be estimated with

$$v_1^{Hajek} = \sum\nolimits_S \left(1 - \pi_k \right) \left(\frac{r_k}{\pi_k} \right)^2 = \sum\nolimits_S \left[\left(\frac{1}{\pi_k} \right)^2 - \frac{1}{\pi_k} \right] r_k^2 . \tag{12}$$

Here $1-\pi_k$ serves as an element-specific model-based finite population correction (*fpc*) term. Its average values is $1-\sum_S \pi_j^2 / n$ (a pooled, model-based *fpc* term), not the *ad-hoc* formulation, 1 – n/N .

Unlike v_1, the *mse* estimator in equation (12) *can* be used with systematic probability sampling from a purposefully-ordered list. Remember, however, it is the combined (model and randomization) *mse* that is being estimated. A combined *mse* requires the model to be correct or at least a reasonably good approximation to be an effective accuracy measure.

A key model assumption for estimating the combined *mse* estimation under purposefully-ordered systematic sampling is that the ε_k are uncorrelated no matter where they are on the list. When the ε_k for adjacent elements on the list tend to be more positively correlated than other pairs, v_1^{Hajek} can be shown to overestimate the combined *mse*. That is to say, there are potential efficiency gains (*mse* reduction) to be had from purposefully ordering the list not reflected within v_1^{Hajek}.

As for the combined ratio estimator under *stsrs*, some work reveals

$$\begin{aligned} v_1^{cr} &= v_1^{Hajek} + \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right)^2 \frac{1}{n_h - 1} \left[\sum_{k \in S_h} r_k^2 - \left(\sum_{k \in S_h} r_k \right)^2 \right], \\ &= v_1^{Hajek} + A \end{aligned}$$

where A has a model expectation of (nearly) zero, but will, in fact, be negative if the ε_k within a stratum tend to be correlated. When there is only one stratum (i.e., under *srs*), the model expectation for v_1^{cr} will be slightly higher than that for v_1^{Hajek} because $\sum_S r_k = 0$.

8. Randomization-assisted Model-based Variance Estimation

The *weighted residual variance estimator* for the Hajek estimator is

$$v_{wrve} = \sum_{k,j \in S} \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} \left(\frac{T_x}{t_x^e} \right)^2 \frac{r_k}{\pi_k} \frac{r_j}{\pi_j} = \left(\frac{T_x}{t_x^e} \right)^2 v_1 , \tag{13}$$

where $t_x^e = \sum_S x_k / \pi_k$. Like v_1 in equation (10), v_{wrve} is a good estimator for the randomization *mse* of t_y^{Hajek}.

Using reasoning similar to that producing equation (8), the model variance of the Hajek estimator can be shown to be

$$V_\varepsilon^{robust} = \left(\frac{T_x}{t_x^e} \right)^2 \sum_S \frac{\sigma_k^2}{\pi_k^2} - 2 \frac{T_x}{t_x^e} \sum_S \frac{\sigma_k^2}{\pi_k} + \sum_U \sigma_k^2 .$$

When the π_k are all small (<< 1), this expression is dominated by

$$V_\varepsilon^{approx} = \left(\frac{T_x}{t_x^e} \right)^2 \sum_S \frac{\sigma_k^2}{\pi_k^2} .$$

Consequently, v_{wrve} is a nearly unbiased estimator for the model variance of the Hajek (since $r_k \approx \varepsilon_k$, $E(r_k r_j) \approx E(\varepsilon_k \varepsilon_j) = 0$ when k and j are distinct).

Estimates of model variance are only as good as the models on which they are based. When the model is a reasonably good representation of reality, however, a coverage interval based on a model-variance estimate and the normal approximation will often be superior to an analogous one derived solely using randomization theory. This is because model variance is conditioned on the sample actually drawn while a randomization *mse* is averaged over all possible samples.

The following expressions are even better than v_{wrve} at estimating the model variance of the combined ratio estimator and the general Hajek estimator:

$$v_2^{cr} = \sum_{h=1}^{H} \left[\left(\frac{N_h}{n_h} \frac{T_x}{t_x^e} \right)^2 - \frac{N_h}{n_h} \frac{T_x}{t_x^e} \right] \frac{n_h}{n_h - 1} \sum_{k \in S_h} \left(r_k - \bar{r}_{hS} \right)^2,$$

and

$$v_2^{Hajek} = \sum\nolimits_S \left[\left(\frac{T_x}{t_x^e} \frac{1}{\pi_k} \right)^2 - \frac{T_x}{t_x^e} \frac{1}{\pi_k} \right] r_k^2 \,. \tag{14}$$

The estimator v_2^{cr} retains the large-sample randomization-based properties of v_1^{cr}, and is nearly model unbiased when $\sigma_k^2 \propto x_k$ even when some of the π_k are not small.

The estimator v_2^{Hajek} is also nearly model unbiased when $\sigma_k^2 \propto x_k$, but tends to have a small downward bias since

$$\begin{aligned} E(r_k^2) &= \sigma_k^2 - 2 \frac{x_k / \pi_k}{\sum_S x_j / \pi_j} \sigma_k^2 + x_k^2 \frac{\sum_S \sigma_j^2 / \pi_j^2}{\left(\sum_S x_j / \pi_j \right)^2} \\ &= \left(1 - \frac{x_k / \pi_k}{\sum_S x_j / \pi_j} \right) \sigma_k^2 - \frac{x_k \pi_k}{\sum_S x_j / \pi_j} \left(\frac{\sigma_k^2}{\pi_k^2} - \frac{x_k / \pi_k}{\sum_S x_j / \pi_j} \sum_S \frac{\sigma_j^2}{\pi_j^2} \right) \\ &= \left(\frac{t_x^e - \frac{1}{\pi_k} x_k}{t_x^e} \right) \sigma_k^2 - \frac{x_k \pi_k}{\sum_S x_j / \pi_j} \left(\frac{\sigma_k^2}{\pi_k^2} - \frac{x_k / \pi_k}{\sum_S x_j / \pi_j} \sum_S \frac{\sigma_j^2}{\pi_j^2} \right) \end{aligned}$$

is less than σ_k^2 for most k. A slightly improved version of v_2^{Hajek} replaces the r_k^2 with $\frac{n}{n-1} r_k^2$ or, better yet,

$$s_k^2 = \left(\frac{t_x^e}{t_x^e - \frac{1}{\pi_k} x_k} \right) r_k^2 = \left(\frac{n}{n - \left(\frac{n x_k}{\pi_k t_x^e} \right)} \right) r_k^2 \,.$$

Like v_1^{Hajek}, the quality of v_2^{Hajek} as a measure for the accuracy of t_y^{Hajek} depends on the quality of the model. By contrast v_1^{cr} and v_2^{cr} are reasonable measures for pure randomization mean squared error of t_y^{cr}.

In general, the slight model bias of a nearly model-unbiased variance estimator v for t can be removed under the assumption $\sigma_k^2 \propto \upsilon_k$. One way to do this is with the ratio adjustment:

$$\tilde{v} = \frac{E_\varepsilon\left[(t-T)^2\right]}{E_\varepsilon(\tilde{v})} v.$$

Usually, $\tilde{v}$ will retain the large-sample randomization-based properties of v, if it has any.

9. Calibration and the General Regression Estimator

For our present purposes, a calibration estimator for T_y is one that can be put in the form $t_y^{cal} = \sum_S w_k y_k$, where the vector (of) calibration equation(s)

$$\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_\mathbf{x} \tag{15}$$

holds for a row vector $\mathbf{x}_k$ of auxiliary variables, and the vector of auxiliary-variable totals, $T_\mathbf{x}$, is known.

This t_y^{cal} is unbiased under the multivariate model:

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \tag{16}$$

where $E(\varepsilon_k \mid \mathbf{x}_k) = E(\varepsilon_k \mid \mathbf{x}_k;\, k \in S) = 0$. A calibration estimator is nearly randomization unbiased when w_k converges to $d_k = 1/\pi_k$ for each element in the sample as *n* grows arbitrarily large.

In surveys with multiple survey variables of interest (like planted corn acres, planted wheat acres, planted soy bean acres, etc.), there may be one auxiliary variable associated with each survey variable (a corn-frame value for each farm in the population, a wheat-frame value, etc.). A single set of calibration weights satisfying equation (15) can produce estimators that are nearly randomization unbiased for the estimated total of each survey variable (total planted corn acres, total planted wheat acres, etc.) yet are also unbiased under a simple linear model in equation (1) applicable for each survey variable separately (a linear model relating planted corn to frame corn, one relating planted wheat to frame wheat, etc.).

A possible calibration estimator sets each

$$w_k = d_k + (T_{\mathbf{x}} - \sum_S d_j \mathbf{x}_j)\left(\sum_S d_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} d_k \mathbf{x}_k'.$$

When $\mathbf{x}_k$ is not a scalar, this formulation produces the most popular of the general regression (GREG) estimators. The model parameter $\boldsymbol{\beta}$ in equation (16) is effectively estimated by $\mathbf{b}_{GREG} = (\sum_S d_j \mathbf{x}_j' \mathbf{x}_j)^{-1} \sum_S d_j \mathbf{x}_j' y_j$.

When $\mathbf{x}_k$ is a scalar, usually denoted by x_k, a different version of the GREG is more common. This formulation sets $w_k = \left(T_x / t_x^e\right) d_k$ and results in the Hajek estimator.

Variance estimators for t_y^{cal} mimic what we have seen before. Paralleling equations (13) and (14) we have.

$$v_{wrve} = \sum_{k,j \in S} \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} w_k r_k w_j r_j \,, \tag{17}$$

and

$$v_2^{"Hajek"} = \sum_S \left(w_k^2 - w_k\right) r_k^2.$$

where $r_k = y_k - \mathbf{x}_k \mathbf{b}_{GREG}$. Several methods are available to reduce or remove the slight downward model bias in $v_2^{"Hajek"}$.

10. The Group-ratio Model

Suppose the population is divided into *G* mutually exclusive groups. Let $\mathbf{x}_k$ be a row vector containing *G* components all but one of which are zero. If element k is in group g, the gth component of $\mathbf{x}_k$ is the scalar x_k. The estimator

$$t_y^{grm} = \sum_{g=1}^{G} T_{xg} \frac{\sum_{S_g} d_k y_k}{\sum_{S_g} d_k x_k}$$

can be put in calibration form with $w_k = \left(T_{xg} \Big/ \sum_{S_g} d_j x_j \right) d_k$. It is a straightforward generalization of the separate ratio estimator in equation (3) and is likewise unbiased under the *group-ratio model*:

$$y_k = \beta_g x_k + \varepsilon_k \quad \text{for} \quad k \in U_g. \tag{4}$$

Let $r_k = y_k - b_g x_k$ for $k \in S_g$ and $b_g = \sum_{S_g} d_j y_j \Big/ \sum_{S_g} d_j x_j$. A very good estimator for model variance of t_y^{grm}, especially when $\sigma_k^2 \propto x_k$, is

$$v_\varepsilon^{grm} = \sum_{g=1}^{G} \sum_{k \in S_g} \left[\left(\frac{T_{xg}}{t_{xg}} d_k \right)^2 - \frac{T_{xg}}{t_{xg}} d_k \right] \left(\frac{\sum_{j \in S_g} d_j x_j}{\sum_{j \in S_g} d_j x_j - d_k x_k} \right) r_k^2.$$

Compare and contrast this with the weighted residual variance estimator in equation (13) under *stsrs* ($g = h$):

$$v_{wrve} = \sum_{g=1}^{G} \sum_{S_g} \left(\frac{T_{xg}}{t_{xg}} \right)^2 \left(d_k^2 - d_k \right) \frac{n_g}{n_g - 1} r_k^2$$
$$= \sum_{g=1}^{G} \left(\frac{T_{xg}}{t_{xg}} \right)^2 \left[\left(\frac{N_g}{n_g} \right)^2 - \frac{N_g}{n_g} \right] \frac{n_g}{n_g - 1} \sum_{S_g} r_k^2 .$$

When the scalar x_k is a constant (say 1) across all elements, t_y^{grm} becomes a generalized version of the post-stratified estimator with $w_k = \left(N_g \Big/ \sum_{S_g} d_j \right) d_k$, and $r_k = y_k - \sum_{S_g} d_j y_j \Big/ \sum_{S_g} d_j$.

11. Two Phase Sampling

It is increasingly popular to use two-phase sampling in establishment surveys. The first phase is often used to remove deadwood (establishments no longer in operation) and to collect enough information to divide the remaining sample into mutually exclusive and exhaustive groups. The second phase of sampling, which is often *srs* within groups, is used to collect more detailed information.

Suppose we have a first-phase sample as described above. The sample is segregated into groups, one of which is reserved for the deadwood. Since we know $y_k = 0$ for all establishments in the deadwood group, we can pretend the group is completely enumerated in the second-phase sample.

Under the group-ratio model, a good estimator for T is

$$t_y^{grm} = \sum_{g=1}^{G} \hat{T}_{xg} \frac{\sum_{s_g} d_k y_k}{\sum_{s_g} d_k x_k}, \tag{18}$$

where $\hat{T}_{xg}$ is an estimate for T_{xg} based on the first phase sample.

For near randomization unbiasedness, d_k is the inverse of the product and the first and second-phase selection probabilities, and s_g is the second-phase sample within group g. For future reference, let $s = \bigcup_{g=1}^{G} s_g$ be the total second-phase sampling.

When the second-phase sampling design is *srs* within groups, the d_k in equation (18) can be replaced by the first-phase selection probabilities without changing its value.

The auxiliary variable x_k (which may be 1 for all elements) is collected during the first phase of sampling. When all x_k = 1, $\hat{T}_{xg} = \hat{N}_g$. This value is an estimate because the population size of each group is unknown before the first-phase sample is enumerated.

Now the combined randomization/model variance of t_y^{grm} can be estimated by adding together two term. The first is

$$v_{\varepsilon}^{grm} = \sum_{g=1}^{G} \sum_{k \in s_g} \left[\left(\frac{\hat{T}_{xg}}{t_{xg}} d_k \right)^2 - \frac{\hat{T}_{xg}}{t_{xg}} d_k \right] \left(\frac{\sum_{s_g} d_j x_j}{\sum_{j \in s_g} d_j x_j - d_k x_k} \right) r_k^2. \qquad (19)$$

The second is an estimator of the randomization variance from the first-phase of sampling using $\hat{y}_k = \dfrac{\sum_{s_g} d_j y_j}{\sum_{s_g} d_j x_j} x_k$ in place of y_k for elements in group g. Call it v_{rand}^{mb2}.

The expression v_{ε}^{grm} in equation (19) estimates the model variance from *both* phases of sampling (in a nearly unbiased fashion when n_g and N_g are large

enough or $\sigma_k^2 \propto x_k$), while v_{rand}^{mb2} estimates the randomization-based expectation of the square of the model bias resulting from $\hat{T}_{xg}$ being used in place of T_{xg}.

Of course, if we knew T_{xg}, then we could compute

$$t_y^{grm} = \sum_{g=1}^{G} T_{xg} \frac{\sum_{s_g} d_k y_k}{\sum_{s_g} d_k x_k},$$

and estimate its model variance (and combined *mse*) with v_ε^{grm} in equation (19) (each $\hat{T}_{xg}$ equaling T_{xg} in this case).

12. Adjusting for Nonresponse

We treat unit (whole-element) response/nonresponse in a single-phase survey as if it were a second-phase of sampling (the respondent sample, R, being the second-phase sample s). Under the group ratio model, t_y^{grm} in equation (18) – with T_{xg} either known or estimated from the original sample before nonresponse, and d_k equal to the original-sample weight of k – serves as a model- or combined-unbiased estimator for T_y, and v_ε^{grm} in equation (19) (or $v_\varepsilon^{grm} + v_{rand}^{mb2}$) as an estimator for its model (or combined) variance.

The model, again, is $y_k = \beta_g x_k + \varepsilon_k$ for $k \in U_g$ (equation (4)) with $E(\varepsilon_k \mid \mathbf{x}_k) = E(\varepsilon_k \mid \mathbf{x}_k;\, k \in R) = 0$. The key assumption in prediction modeling for unit nonresponse is the ignorability of the response mechanism. It is *not* necessary for all the elements in a group to have the same probability of response. That is the assumption of an alternative and complementary paradigm: quasi-random response modeling.

Now t_y^{grm} can be put in reweighted form as $t_y^{grm} = \sum_R w_k y_k$, where $w_k = \dfrac{\hat{T}_{xg}}{\sum_{R_g} d_j x_j} d_k$ for $k \in R_g$. We can likewise, rewrite v_ε^{grm} as

$$v_\varepsilon^{grm} = \sum_{g=1}^{G} \sum_{k \in S} \left[\left(w_k \right)^2 - w_k \right] \left(\frac{\sum_{R_g} w_j x_j}{\sum_{j \in R_g} w_j x_j - w_k x_k} \right) r_k^2$$

with $r_k = y_k - \dfrac{\sum_{R_g} w_j y_j}{\sum_{R_g} w_j x_j} x_k$ for $k \in R_g$.

Surveys are usually designed to estimate totals for a number of variables. Often reweighting for unit nonresponse is done by setting all the x_k = 1, so that the same set of adjusted weights, $\{w_k\}$, can be used for every survey variable.

When there is item nonresponse for some y-variable closely related to an auxiliary x-variable known for all sampled elements, a missing y-value is often imputed with:

$$\hat{y}_k = \frac{\sum_{j \in R_g} a_{kj} y_j}{\sum_{j \in R_g} a_{kj} x_j} x_k ,$$

where the a_{kj} can have any number of forms. With *weighted-ratio imputation*, $a_{kj} = d_j$, while $a_{kj} = 1/x_j$ with *unweighted-mean-of-ratio imp*utation. Another popular technique is *random-ratio imputation*, where *j* is randomly selected from among R_g, and y_k is imputed with $\hat{y}_k = x_k \left(y_j / x_j \right)$. Here a_{kj} = 1 when *j* is the element randomly selected to impute for the missing y_k and 0 otherwise.

In each of these variations we can rewrite the estimated total in the form of a reweighted estimator and compute v_ε^{grm} as in equation (19). Note that there is a possibly different reweighted estimator for each survey variable, so that this reweighting is done for variance-estimation purposes only.

There are other approaches.

13. Concluding Remarks

With the sole exception of Basu (1971), I have avoided putting references in the text. The interested reader is directed instead to the reference section for further readings on this topic. Important early papers demonstrating the usefulness of models in choosing among estimation strategies with good probability-sampling properties are Brewer (1963) and Isaki and Fuller (1982). Brewer selection (although not named such) is found in the former, while the latter sets up an asymptotic structure for analyzing randomization theory that has been used ever since.

Brewer (2002) is a text book on combining models and probability sampling that uses Basu's elephant example as its starting point. By contrast Valliant *et al*. (2000) is written from a purely model-based perspective.

The weighted residual variance estimator is introduced in Särndal *et al.* (1989) and calibration weighting in Deville and Särndal (1992). Kott (2004) discusses in depth the approach to variance estimation often taken here: simultaneously estimating the model variance and randomization mean squared error of a calibration estimator (the simple, combined, and separate ratio estimators being special cases). Two co-authored articles of mine in the references contain other ideas finding expression here.

Let us turn briefly to extensions of the methods described in the text. When the frame exhibits coverage errors (duplications or missing population elements), and true group population totals are known for some auxiliary variable x, then we can treat the process of moving from the population to the frame as the first phase of a two-phase sampling scheme, where the actual sample is the second phase of sampling. One can show that t_y^{grm} in equation (18) with $\hat{T}_{xg} = T_{xg}$ is a model unbiased estimator for T_y under the group-ratio model, assuming the coverage mechanism is ignorable. The expression v_ε^{grm} in equation (19) serves as a model (and combined) variance estimator.

We can extend the treatment of two-phase estimation to incorporate additional phases (or stages) of sampling. We can also include a more complex set of benchmark variables than the estimated group totals for a scalar x.

We can incorporate nonlinear prediction models for the y_k. Some care must be taken, however, because $T_y = \sum_U y_k$ is itself a linear function of random variables under the prediction model. See, for example, Breidt *et al.* (2005).

Finally, when all the sampling fractions (the n_h/N_h are small enough to be ignored,

$$v_{2-approx}^{cr} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \frac{T_x}{t_x^e} \right)^2 \frac{n_h}{n_h - 1} \sum_{k \in S_h} \left(r_k - \overline{r}_{hS} \right)^2,$$

is a good estimator for the randomization mean squared error of t_y^{cr} estimator under *stsrs* and an even better estimator for its model variance.

It can be shown that $v_{2-approx}^{cr}$ is asymptotically closer to the stratified jackknife variance estimator than to the conventional estimator of randomization mean squared error (ignoring the sampling fractions):

$$v_{1-approx}^{cr} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right)^2 \frac{n_h}{n_h - 1} \sum_{k \in S_h} \left(r_k - \overline{r}_{hS} \right)^2.$$

This suggests the stratified jackknife should be viewed as a robust (to misspecification of the σ_k^2) model-based variance estimator rather than a randomization-based variance estimator.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1 (with discussion). In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, Eds., Toronto: Holt, Rinehart and Winston, 203-242.

F.J. Breidt, G. Claeskens and J.D. Opsomer (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika,* **92**, 831-846.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93-105.

Brewer, K. R. W. (2002). *Combined survey and sampling inference: Weighing Basu's elephants.* London: Arnold.

Deville, J- C. and C- E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376.

Isaki, C.T. and W. A. Fuller (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association* **77**, 89-96.

Kott, P. S. (2004). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference* **48**, 263-277.

Kott, P. S. and Bailey, J. T. (2000). The theory and practice of maximal Brewer selection. *Proceedings of the Second International Conference on Establishment Surveys, invited papers,* 269-278.

Kott, P. S. and K. R. W. Brewer (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington DC.

Särndal, C- E, B. Swensson, and J. Wretman (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika* **76**, 527-537.

Valliant, R, Dorfman, A.H., and Royall, R.M. (2000). *Finite population sampling and inference: a prediction approach*. New York: Wiley.