CALIBRATION ESTIMATORS BASED ON SEVERAL SEPARATE STRATIFICATIONS

Phillip S. Kott, USDA/NASS
3251 Old Lee Highway, Fairfax, VA 22030

Key Words: Design, Model, Permanent random number, Probabity proportional to size

Many of the list frame surveys conducted by the National Agricultural Statistics Service (NASS) are integrated in the sense that survey variables cover a range of heterogenous items such as planted crop acres and grain stock inventories. Bankier (1986), Skinner (1991), and Skinner et al. (1994) have shown how an old method of combining independently drawn stratified simple random samples - where each sample comes from a (list) frame with a different stratification scheme - can be made reasonably efficient (i.e., the variance of the estimation strategy would not be too large).

Even more appealing for many applications would be a sampling design that tends to select the same units from every frame. This paper explores several such designs. Three make use of permanent random numbers. The fourth, and best, uses a variation of systematic probability proportional to size sampling.

The paper shows how a calibration (i.e., reweighted) estimator can provide relative efficiency by capturing what we know about the original stratum sizes in the estimation. It also addresses variance estimation for the contemplated estimation strategies.

A final section points out that the use of a least-squares-based calibration technique can do more than simply reflect original stratum sizes. In fact, one may want to do away with stratification entirely in certain applications.

1. An Unbiased Multiplicity Estimator

Suppose we have F independent frames; for example, a sorghum frame, an oats frame, and a general grain stocks frame. Each frame is stratified independently, and without replacement simple random samples are drawn from each stratum of every frame. Frame f (say, the oats frame) contains H_f strata; stratum h (large oats operations) in frame f has N_{fh} population units, out of which n_{fh} units are selected. The union of the F frames must cover the entire (list) population, but no single frame need be complete.

One unbiased estimator for a population total $T = \sum_{i\in P} y_i$ is the simple multiplicity estimator:

$$t_M = \sum_{i\in P} y_i n_{(i)}/E[n_{(i)}], \qquad (1)$$

where P denotes the entire population, and $n_{(i)}$ is the number of times unit i is selected for the sample from any frame.

Observe that $n_{(i)} = 0$ for the population units not in the sample. In the great majority of applications, $n_{(i)}$ will be one for most sampled units, but $n_{(i)} > 1$ is a possibility with this design.

The expected number of times unit i will be selected for the sample is $E[n_{(i)}] = \sum^F p_{if}$, where p_{if} is the probability of selecting unit i in the stratified simple random sample from frame F; that is, $p_{if} = n_{fh}/N_{fh}$, where unit i is in stratum h of frame f.

Skinner et al. provide a variance estimator for t_M under this design. There is also a Horvitz-Thompson estimator for T under the design, namely $t_{HT} = \sum_{i\in S} y_i/\pi_i$, where S denotes the sample and $\pi_i = 1 - (1 - p_{i1})(1 - p_{i2}) \cdots (1 - p_{iF})$. Estimating the design variance of t_{HT}, however, is difficult. See Bankier (1986) for further discussion of this approach.

2. Sampling Strategies Using Permanent Random Numbers

The sampling design discussed above is independent across frames. For many of NASS's purposes, however, it would be convenient if the design were not independent across frames. In fact, it is often a desirable for a design to have a tendency to select the same operators in every frame.

To this end, suppose each unit began with a target p_{if} in each frame that was constant for all units in stratum h of frame f. One can then assign unit in the population a *permanent random number* (prn) drawn from the uniform distribution on the interval [0, 1). Unit i is selected for the frame f sample when its prn is less than p_{if}.

The result is a Poisson sample in which the probability of selecting unit i for the sample is $\pi_i = \max_f\{p_{if}\}$. An unbiased Horvitz-Thompson estimator for T under this design is

$$t_P = \sum_{i\in S} y_i/\max_f\{p_{if}\}.$$

A *collocated* variant of this design assigns each population unit a unique prn from among the members of the set {0, 1/N, 2/N, 3/N, ..., (N-1)/N}. In practice, one can first draw provisional prn's for each unit and then assign 0 to the unit with the smallest provisional prn, 1/N to the units with the second smallest provisional prn, and so on until (N-1)/N is assigned to the unit with the largest provisional prn. The estimator t_P remains unbiased under collocated sampling.

A third version of this design begins with target n_{fh} values. The units in stratum h of frame f with the n_{fh} smallest prn's are selected for the sample. A Horvitz-Thompson estimator under this *sample size prn* design requires one to compute the selection probabilities of the sampled units - a difficult task which may have to be approximated by simulation.

3. A Systematic Probability Proportional to Size Design

Another sampling design with the same selection probabilities as the Poisson (and collocated) sampling scheme described in the last section consists of the following steps:

0) When necessary, create an additional "stratum" for each frame consisting of those units not in any design stratum.

1) Divide up the population into mutually exclusive cells such that every unit in a particular cell is in the same stratum of each frame (e.g., the large oats stratum, the medium grain stocks stratum, and the no sorghum stratum).

2) Randomly order the population units in each cell and then sort the cells themselves in any order. This results in a list of all population units.

3) Draw a systematic probability proportional to "size" (pps) sample from this list using the π_i described in the discussion of Poisson sampling as the measures of size (the word "size" is in quotes because the π_i are not really size measures in a conventional sense).

The systematic pps sampling design introduced above will always result in a sample of size close to $\sum_{i\in P} \pi_i$. In fact, if $\sum_{i\in P} \pi_i$ is an integer, then the sample size will exactly equal that sum. Otherwise, the sample size will be one of the two integers closest to $\sum_{i\in P} \pi_i$. Similarly, the expected number of sampled units in a cell, C, will be $\sum_{i\in C} \pi_i$, while the actual sample size will either be $\sum_{i\in C} \pi_i$ or one of the two integers closest to it.

Consider now a particular stratum h in a particular frame f with target sample size n_{fh}. For a unit i in this stratum, $\pi_i \geq n_{fh} / N_{fh}$ by design. Let P(fh) denote the set of population units stratum fh. The expected number of sampled units in fh is $\sum_{i\in P(fh)} \pi_i \geq n_{fh}$. There is no ***guarantee*** that the realized sample size in the stratum will be greater or equal to n_{fh}. Nevertheless, given the above inequality and the lower bounds on the sample sizes of the cells within fh, the sample size in stratum fh will never be far below n_{fh}.

The advantages of this design over Poisson and collocated sampling is obvious. The sample it produces has a more stable size and a greater likelihood of meeting frame/stratum requirements. Sample size prn, by contrast, will always meet frame/stratum requirements, but it does so at a cost: the design has a less stable overall sample size, and selection probabilities can be very difficult to determine. See Amrhein et al. (1996) for more on these four sampling designs.

4. Calibration

The problem with both t_M and t_P (or t_{HT}) is that they are not usually very good estimators for T in term of precision (variance). One of the properties of single-frame, stratified simple random sampling is that the conventional expansion estimator estimates the stratum population size perfectly (i.e., with zero variance). In our multiple frame set up, however, neither t_M nor t_P will estimate the N_{fh} perfectly in most applications.

Let us define $w_i^0 = n_{(i)}/E[n_{(i)}]$ as the *original sampling weight* of unit i in t_M. Similarly, $w_i^0 = 1/\max_f\{p_{if}\}$ in t_P and $1/\pi_i$ more generally for a Horvitz-Thompson estimator. The Skinner articles (following Bankier's original suggestion) propose raking to create a set of adjusted weights $\{w_i^{\,C}\}$ such that

$$\sum_{i\in S_{fh}} w_i^C = N_{fh} \qquad (2)$$

for each stratum h in every frame f, where S_{fh} is that part of the sample that is in stratum h of frame f regardless of the frame(s) from which the units were selected.

The Skinner articles do not say so, but for equation (2) to be logically coherent, units not assigned to any stratum of a particular frame for sampling purposes must be assigned to such a stratum now. One way to do that is to add an *estimation stratum* to every incomplete frame containing all population units not in the frame. This may already have been done to draw a systematic pps sample.

Deville and Särndal (1992) call (2) a *calibration equation*. They point out that there are a number of ways to compute the *calibration weights*, the w_i^C, so that equation (2) is satisfied and w_i^C/w_i^0 is in some sense close to 1 for all i. One method is raking as suggested in the Skinner articles. Another method uses least squares. Either way, the resulting estimator

$$t_C = \sum_{i\in S} w_i^C y_i,$$

where S denotes the entire sample, will be nearly design unbiased because w_i^C/w_i^0 is close to 1 for all i.

The estimator t_C is also unbiased under the model:

$$y_i = \beta_0 + \sum_{f=1}^{F} \sum_{h=2}^{H_f} d_{ifh}\beta_{fh} + \epsilon_i, \tag{3}$$

where the dummy variable, d_{ifh}, is 1 when unit i is in stratum h of frame f (sampled or not) and zero otherwise, while ϵ_i is a random variable with a mean of zero. The β_0 and the β_{fh} are unknown constants (β_0 represents the mean y-value for a unit in the first stratum of every frame; that is why the second sum excludes h=1). The same d_{ifh} values apply to every survey item (y) of interest, while the β values change with the survey item. For many survey items, β_{fh} values will be zero when frame f (say, grain stocks) is irrelevant to the item (say, planted oat acres).

5. Simple Post-Stratification

It is the satisfaction of the calibration equations in equation (2) that assures the unbiasedness of t_C under the model in equation (3). One question that needs to be asked before proceeding in practice is whether the effort needed for satisfying the calibration equations in (2) is worthwhile.

For example, we may not care about the other frame stratifications when estimating planted oat acres in a state. Let f=1 be the oats frame. A useful set of calibrations weights for oats estimates would be

$$w_i^C = (N_{1h} / \sum_{j \in S_{1h}} w_j^0) w_i^0, \tag{4}$$

when i is in stratum h of the oats frame. (At noted above, if the oats frame is incomplete, we may need to add an additional oats "stratum" to cover sample units with oats from the other frame samples that are not in the oats frame; this will only be possible, of course, if such units exist.)

The simple weight adjustment in equation (4) is identical to post-stratification. Skinner calls the resulting t_C a "ratio estimator" ("separate ratio estimator" would be more accurate). It is not hard to see that this estimator is unbiased under a model in which the units within the same oats stratum have a common mean.

The logic behind equation (4) allows there to be a different set of weights for each frame. It is usually convenient, however, to have as few weights as possible. As a result, we will concentrate on developing a single set of weights for all estimators from now on.

6. General Calibration Using Least Squares

In this section we discuss a general method of satisfying calibration equations. Let $\mathbf{x}_i = (x_{i1}, ..., x_{iG})$ be a row vector containing known values for every unit in the population P. The general set of calibration equations we want to satisfy are

$$\sum_{i \in S} w_i^C x_{ig} = \sum_{i \in P} x_{ig} \quad \text{for } g = 1, 2, ..., G. \tag{5}$$

Expressed in vector notation, this is $\sum_S w_i^C \mathbf{x}_i = \sum_P \mathbf{x}_i$.

Suppose x_{ig} in equation (5) is 1 when $i \in P(fh)$ and is 0 otherwise. Then the general calibration equation (5) collapses to the specific calibration equation (2).

One good method for calculating general calibration weights, a variant of least squares, is discussed below. Let

$$w_i^C = \sum_{j \in P} \mathbf{x}_j \, (\sum_{j \in S} w_j^0 \mathbf{x}_j' \mathbf{x}_j)^{-1} w_i^0 \mathbf{x}_i' \tag{6}$$

$$= w_i^0 + (\sum_{j \in P} \mathbf{x}_j - \sum_{j \in S} w_j^0 \mathbf{x}_j)(\sum_{j \in S} w_j^0 \mathbf{x}_j' \mathbf{x}_j)^{-1} w_i^0 \mathbf{x}_i'$$

for all sampled units i. The last equality assume the existence of a row vector $\mathbf{p}$ such that $\mathbf{p}\mathbf{x}_i' = 1$ for all i.

Using the calibration weights in equation (6) renders $t_C = (\sum_{i \in P} \mathbf{x}_i)\mathbf{b}$, where

$$\mathbf{b} = (\sum_{j \in S} w_j^0 \mathbf{x}_j' \mathbf{x}_j)^{-1} \sum_{i \in S} w_i^0 \mathbf{x}_i' y_i.$$

Observe that $\mathbf{b}$ is the weighted least squares estimator for $\boldsymbol{\beta}$ in the model:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{7}$$

where $E(\epsilon_i) = 0$. It is now easy to see that t_C is unbiased under this model (because

$$E_\epsilon(t_C) = (\sum_{i \in P} \mathbf{x}_i)\boldsymbol{\beta} = E_\epsilon(\sum_{i \in P} y_i) = E_\epsilon(T)).$$

In the simple case of post-stratification (equation (4)) with f being the relevant frame, $x_{ig} = 1$ when i is in stratum g of frame f and zero otherwise. For the model in equation (3),

$$\mathbf{x}_i = (1, d_{i12}, ..., d_{i1H(1)}, d_{i22}, ..., d_{i2H(2)}, ..., d_{iF2}, ..., d_{iFH(F)}),$$

where H(f) denotes the last stratum in frame f. Note that will this $\mathbf{x}_i$, $\mathbf{p} = (1, 0, 0, ..., 0)$.

7. A Generalization

Recently, Brewer (1994) has proposed the following alternative to equation (6):

$$w_i^C = w_i^0 + (\sum_{j \in P} \mathbf{x}_j - \sum_{j \in S} w_j^0 \mathbf{x}_i)(\sum_{j \in S} \{[w_j^0 - 1]/z_j\} \mathbf{x}_j' \mathbf{x}_j)^{-1} \{[w_i^0 - 1]/z_i\} \mathbf{x}_i, \tag{6'}$$

where the z_j are to-be-determined constants. Brewer's problem with equation (6) is its tendency to produce calibration weights that are less than unity. Equation (6'), while it too can produce very small (and even negative) w_i^C usually will behave better than (6) - especially when the z_j are well chosen (exactly how depends on the context). For example, consider the version of (6') that is analogous to equation (4):

$$w_i^C = w_i^0 + \frac{N_{1h} - \sum_{j \in S_{1h}} w_j^0}{\sum_{j \in S_{1h}} (w_j^0 - 1)} (w_i^0 - 1) . \qquad (4')$$

When w_i^0 is close to unity, the w_i^C produced by equation (4) can dip below 1, while the calibration weights produced by (4') cannot.

Equations (6) and (6') can both be viewed as special cases of the general form:

$$w_i^C = w_i^0 + \left(\sum_{j \in P} \mathbf{x}_j - \sum_{j \in S} w_j^0 \mathbf{x}_j \right) \left(\sum_{j \in S} d_j w_j^0 \mathbf{x}_j{}' \mathbf{x}_j \right)^{-1} d_i w_i^0 \mathbf{x}_i . \qquad (8)$$

In equation (6), $d_j = 1$, while in equation (8), $d_j = (w_j^0 - 1)/(w_j^0 d_j)$. The scalar d_i can be called a "tuning" constant. Its function, when it has one, is to keep the values of the w_j^C within desired bounds.

8. The Variance Under Non-Replacement Sampling Designs

Suppose we have a calibration estimator with weights satisfying equation (8). Let $\mathbf{B} = (\sum_P d_j \mathbf{x}_j{}' \mathbf{x}_j)^{-1} \sum_P d_i \mathbf{x}_i{}' y_i$, and $e_i = y_i - \mathbf{x}_i \mathbf{B}$. Now the design mean squared error of $t_C = \sum_S w_i^C y_i$ is identical to that of $\sum_S w_i^C e_i$ because

$$\sum_S w_i^C y_i - \sum_P y_i = \sum_S w_i^C e_i - \sum_P e_i .$$

Given the sampling designs under consideration, it is not hard to show that w_i^C / w_i^0 is $1 + O_p(1/\surd n)$ for all i, and $\sum_p x_i / \sum_S w_i^0 x_i$ is also $1 + O_p(1/\surd n)$. As a result,

$$\sum_{i \in S} w_i^C e_i = \sum_{i \in S} w_i^0 e_i +$$

$$\left(\sum_{j \in P} \mathbf{x}_j - \sum_{j \in S} w_j^0 \mathbf{x}_j \right) \left(\sum_{j \in S} d_j w_j^0 \mathbf{x}_j{}' \mathbf{x}_j \right)^{-1} \sum_{i \in S} d_i w_i^0 \mathbf{x}_i{}' e_i$$

$$\approx \sum_{i \in S} w_i^0 e_i .$$

In this section, we will confine our attention to non-replacement sampling schemes, like the three prn designs and the systematic pps design discussed previously. Let π_{ik} denote the joint selection probability of units i and k. Then

$$MSE_D(t_C) \approx Var_D(\textstyle\sum_S w_i^0 e_i)$$

$$= \sum_{i \in P} (w_i^0 - 1) e_i^2 + \sum_{i \neq k} [(\pi_{ik} / \pi_i \pi_k) - 1] e_i e_k .$$

Under the Poisson version of prn sampling, the selection of one unit is independent of whether another unit has been selected. This means that $\pi_{ik} = \pi_i \pi_k$. The design mean squared error of t_C thus collapses to $\sum_P (1/\pi_i - 1) e_i^2$.

For the other two versions of prn sampling, π_{ik} will be slightly less than $\pi_i \pi_k$, and no simple expression for the design mean squared error of t_C appears to exist. Nevertheless, given the nature of the e_i's (they are sometimes positive and sometimes negative, with an average value around zero), it may not be unreasonable to assume that

$$\sum_{i \neq k} [(\pi_{ik} / \pi_i \pi_k) - 1] e_i e_k \approx 0.$$

For the systematic pps design, π_{ik} may be considerable less than $\pi_i \pi_k$ when i and k are in the same cell. Moreover, when i and k are in the same cell, $e_i e_k$ will - if anything - have a propensity to be positive reflecting an "interaction" effect not captured by the model in equation (7). *Consequently,* $\sum_P (1/\pi_i - 1) e_i^2$ *may be an overstatement of true design mean squared error.*

Isaki and Fuller (1982) call the model expectation of the design mean squared error of t_C the "anticipated mean squared error" of the estimator. This value is of most use at the planning stage of a sample survey.

If the model in equation (7), $y_i = x_i \beta + \epsilon_i$, holds and the ϵ_i are uncorrelated, then the anticipated mean squared of t_C is

$$E_\epsilon[MSE_D(t_C)] = E_\epsilon\{ E_D[(\textstyle\sum_S w_i^C \epsilon_i - \sum_P \epsilon_i)^2]\}$$

$$\approx \sum_P (1/\pi_i - 1) E_\epsilon(\epsilon_i^2) +$$

$$\sum_{i \neq k} [(\pi_{ik} / \pi_i \pi_k) - 1] E_\epsilon(\epsilon_i \epsilon_k)$$

$$\approx \sum_P (1/\pi_i - 1) E_\epsilon(\epsilon_i^2). \qquad (9)$$

It is of some interest to note that using Poisson, collocated, and systematic pps sampling result in estimators with identical anticipated mean squared errors.

Suppose we had used stratified simple random sampling and selected unit i with probability $p_{if} \leq \pi_i$, where f is the frame relevant to y. It is not hard to show that theanticipated variance of the simple expansion estimator would have been $\sum_P (1/p_{if} - 1) E_\epsilon(\epsilon_i^2)$, which is at least as large as the right hand side of equation (9). Thus, there are gains from "integrating" the samples from various frames as we have effectively done.

9. Variance/Mean Squared Error Estimation

If we accept that the design mean squared error of t_C is approximately $V_D = \sum_P (1/\pi_i - 1)e_i^2$, then a nearly unbiased estimator of this design mean squared error is

$$v_D = \sum_S (1/\pi_i^2 - 1/\pi_i)r_i^2,$$

where $r_i = y_i - \mathbf{x}_i\mathbf{b} \approx e_i$. Since $w_i^C \approx (1/\pi_i)(1 + O_p(\surd n))$, the alternative

$$v = \sum_S ([w_i^C]^2 - w_i^C)r_i^2 \tag{10}$$

is also nearly design unbiased.

Let us now look at the model in equation (7), and again assume that the ϵ_i are uncorrelated with $E(\epsilon_i^2) = \sigma_i^2$. The model variance of t_C as estimator for T is

$$\begin{aligned} E_\epsilon[(t_C - T)^2] &= E_\epsilon[(\sum_S w_i^C y_i - \sum_P y_i)^2] \\ &= E_\epsilon[(\sum_S w_i^C \epsilon_i - \sum_P \epsilon_i)^2 \\ &= \sum_S (w_i^C)^2\sigma_i^2 - 2\sum_S w_i^C\sigma_i^2 + \sum_P \sigma_i^2. \end{aligned}$$

If we add the additional assumption $\sigma_i^2 = x_i\gamma$ for some vector γ, then the model variance of t_C is

$$V_\epsilon = \sum_S ([w_i^C]^2 - w_i^C)\sigma_i^2.$$

Since $E(r_i^2) = \sigma_i^2 + O(1/n)$ under mild conditions, v in equation (10) is a nearly unbiased estimator of the model variance of t_C as well as the design mean squared error of t_C. It should also be noted that the assumption $\sigma_i^2 = x_i\gamma$ is not really necessary when all $w_i^C >> 1$ so that $V_\epsilon \approx \sum_S (w_i^C)^2\sigma_i^2$.

The appendix (available from the author upon request) explores an alternative method of variace/mean squared error estimation using a delete-a-group jackknife.

10. Is Stratification Necessary?

In Section 4, it was pointed out that if calibration weights were design to satisfy equation (2), the resulting estimator would be unbiased under the model in equation (3). In Sections 6 and 7, a general way of producing calibration weights was discussed that produced an estimator unbiased under the model in equation (7), of which the model in (3) was a special case. In many applications, there may be a more appropriate model on which to based calibration than the one in equation (3). For example, if there was a continuous control variable used to stratify a particular frame, it makes more sense to use that variable directly in the model rather than indirectly through frame/stratum identifiers. For a survey of chemical use on vegetables, say, it makes more sense to treat the farm acres of each vegetable of interest as the components of $\mathbf{x}_i$ in equation (7), rather than creating a separate dummy variable (a d_{ifh}) for all but one stratum of every vegetable frame.

In fact, for some applications, we may want to abandon stratification at the design stage as well. For example, in the vegetable survey discussed above, rather than stratifying for each vegetable, a tentative univariate pps selection probability, $p_{ic} = n_c v_{ic} / \sum_{j\in P} v_{jc}$, can be computed for farm i based on its acreage, v_{ic}, of vegetable c, where n_c is the target sample size for the vegetable. The farms' actual selection probability would then be its largest tentative selection probability across all vegetables, $\pi_i = \max_c\{p_{ic}\}$. Sample selection would be by systematic pps with farms sorted by, first, the presence or absence of the least common vegetable of interest to the survey, then the second least common, and so forth. As a result of this selection process, most should vegetables at least meet their sample targets. See Hicks et al. (1996).

References

AMRHEIN, J., HICKS, S. & KOTT, P. (1996). Methods to control selection when sampling from multiple list frames. *ASA Proc. Surv. Res. Meth. Sec.* (this volume).

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *J. Amer. Statist. Assoc.*, **81**, 1074-1079.

BREWER, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pak. J. Statist.* **10(1)A**, 213-33.

DEVILLE, J-C. & SÄRNDAL, C-E. (1992). Calibration estimator in survey sampling. *J. Amer. Statist. Assoc,*. **87**, 376-382.

HICKS, S., AMRHEIN, J. & KOTT, P. (1996). Methods to meet target sample sizes under a multivariate PPS sampling strategy. *ASA Proc. Surv. Res. Meth. Sec.* (this volume).

ISAKI, C.T. & FULLER, W.A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, **77**, 89-96.

SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *J. Amer. Statist. Assoc.*, **86**, 779-84.

SKINNER, C.J., HOLMES, D.J. &HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *Int. Statist. Rev.*, **62, 3**, 333-47.