VARIANCE ESTIMATION WHEN A FIRST STAGE AREA SAMPLE IS RESTRATIFIED

Phillip S. Kott, National Agricultural Statistics Service
U.S. Department of Agriculture, Washington, DC 20250

1 Introduction

Many countries conduct large scale annual surveys of agricultural activity based on samples of geographically defined segments drawn from stratified area frames. Cotter and Nealon (1987) discuss the sampling design employed by the U.S. Department of Agriculture (USDA) for its major June survey.

It is also a common practice to conduct smaller surveys of agricultural activity throughout the year based (at least in part) on subsamples of the farm tracts enumerated in the annual survey (a farm tract is that part of a farm operation within a single area segment). Tracts within sampled area segments are first restratified based on their responses during the annual survey; subsamples of farm tracts are then drawn within each new stratum. To avoid confusion, this paper adopts the USDA practice of referring to the original area strata as <u>districts</u> and the new annual response-based strata as <u>strata</u>.

We will assume for simplicity that stratified simple random sampling (srs) without replacement is performed at both stages of the sampling design. This is not quite the case at the USDA (Kott, 1988; Kott and Johnston, 1988), but it is close enough to the mark for our present purpose. We will also ignore the fact that in practice area subsamples are often combined with samples drawn from list frames.

The <u>two-phase</u> sampling design described above gives statisticians the ability to use the information collected during the annual survey in a cost effective manner. On the down side, the variances of their estimates are themselves difficult to estimate.

This paper derives an unbiased estimation formula for the two-phase estimation strategy under discussion. The formula is a generalization of a suggestion by Cochran and Huddleston (1969, 1970), which assumed that the original area sample was the result of unstratified srs (something that was very close to true for USDA June samples at the time that paper was written).

2 Preliminaries

Suppose we started with an original (annual) area survey consisting of n_D (out of N_D) area segments from each of L districts. For a current survey, the farm tracts within the originally sampled area segments have been restratified into H strata. Within stratum h, v_h (out of T_h) tracts are in the subsample. Both phases of the sampling design are the result of stratified srs without replacement.

Let us concentrate on one farm value of interest. Furthermore, let

S^1 denote the set of all farm tracts within originally sampled area segments whether enumerated for the current survey or not,

S_j denote the set of all current (i.e., currently enumerated or subsampled) tracts in segment j,

S_D denote the set of all current tracts in district D,

R_h denote the set of all current tracts in stratum h,

x_i denote the farm value of interest for tract i,

$y_i = (N_D/n_D)x_i$ denote the first phase expanded farm value of tract i, $i \epsilon S_D$,

$z_i = (T_h/v_h)y_i$ denote the fully expanded farm value of tract i, $i \epsilon R_h$,

$y_{jh} = \sum_{i \in S_j \cap R_h} y_i$ denote the total first phase expanded farm value of all current tracts in stratum h and segment j,

$y_{Dh} = \sum_{i \in S_D \cap R_h} y_i$ denote the total first phase expanded farm value of all current tracts in stratum h and district D,

$y_{.h} = \sum_{i \in R_h} y_i$ denote the total first phase expanded farm value of all current tracts in stratum h,

$z_{j.} = \sum_{i \in S_j} z_i$ denote the total fully expanded farm value of all current tracts in segment j,

$z_{D.} = \sum_{i \in S_D} z_i$ denote the total fully expanded farm value of all current tracts in district D, and

$Y = \sum_{i \in S^1} y_i$ denote the total first phase expanded farm value of all tracts within the originally sampled area segments.

An unbiased estimator for X, the sum of x_i values across all tracts in the population is

$$\hat{X} = \sum_{h=1}^{H} \sum_{i \in R_h} z_i = \sum_{h=1}^{H} \sum_{i \in R_h} (T_h/v_h) y_{i.} \quad (1)$$

To see this, observe that
$Y = \sum_{i \in S^1} y_i = \sum_{i \in S^1} (N_D/n_D) x_i$ is an unbiased

estimator of X with respect to the first phase of sampling, while $\hat{X}$ is an unbiased estimator of Y with respect to the second sampling phase. Mathematically, $E_1(Y)=X$ and $E_2(\hat{X})=Y$, which implies $E(\hat{X})=E_1E_2(\hat{X})=X$.

3 The variance of $\hat{X}$

From any of a number of textbooks on sampling theory (e.g., Cochran, 2, p. 276), we know that the variance of a two-phase estimator like $\hat{X}$ is

$$\mathrm{var}(\hat{X}) = \mathrm{var}_1[E_2(\hat{X})] + E_1[\mathrm{var}_2(\hat{X})], \quad (2)$$

where E_k and var_k denote, respectively, expectation and variance with respect to the k^{th} phase of sampling.

The first term in equation (2) is called the first phase variance because it equals the variance that would be obtained if every tract within an originally sampled segment were part of current subsample.

The second term in (2) is called the second phase variance, but that is not strictly speaking true. The second phase variance (really $\mathrm{var}_2(\hat{X})$) can only be defined with respect to a particular original sample. What the second term in (2) is is the average of second phase variances taken over all possible original samples (and weighted by the probability of drawing each sample).

Despite this slight confusion about the second phase variance, it is easier to estimate than the first phase variance and we will attack it first. The problem with first phase variance estimation is that total current values for the segments in the original sample can only be estimated using the current subsample. As is well known, putting estimated segment totals in place of real totals in the usual one-phase variance formula biases the resulting estimator.

3.1 Second Phase Variance Estimation

First note that a formula for an unbiased estimator of $\mathrm{var}_2(\hat{X})$ given any original sample is automatically an unbiased estimator of $E_1[\mathrm{var}_2(\hat{X})]$. To see this, suppose that v_2 is an unbiased estimator of $\mathrm{var}_2(\hat{X})$ given any sample. Since $E_2[v_2-\mathrm{var}_2(\hat{X})]=0$ for all S^1, the first phase expectation of $E_2[v_2-\mathrm{var}_2(\hat{X})]$ must also be zero. Consequently,
$E(v_2)=E_1E_2(v_2)=E_1[\mathrm{var}_2(\hat{X})]$.

Now given our particular S^1,

$$\hat{\mathrm{var}}_2 = \sum_{h=1}^{H} (T_h^2/v_h - T_h)*[1/(v_h-1)] \quad (3)$$
$$* [(\sum_{i \in R_h} y_i^2) - y_{.h}^2/v_h]$$

is the conventional unbiased estimator for $\text{var}_2(\hat{X})$. Moreover, equation (3) would hold whatever first phase sample obtained. As a result, $\hat{\text{var}}_2$ is also an unbiased estimator for $E_1[\text{var}_2(\hat{X})]$.

3.2 First Phase Variance Estimation

Consider a segment j within district D. The value $z_{j.}$ is an unbiased estimator of (N_D/n_D) times the total farm value among all tracts in segment j whether in the current subsample or not. Consequently, $E_2(z_{j.})$ is exactly (N_D/n_D) times the total farm value among all tracts in segment j. With this in mind, the following would be an unbiased estimator of the first phase variance of $\hat{X}$:

$$\hat{\text{var}}_1[E_2(\hat{X})]) = \tag{4}$$

$$\sum_{D=1}^{L} (1 - n_D/N_D)[n_D/(n_D-1)] * \left[\sum_{j=1}^{n_D}\{E_2(z_{j.})\}^2 - \{E_2(z_{D.})\}^2/n_D\right].$$

Taken as is, equation (4) is useless since it supposes we know what the $\{E_2(z_{j.})\}^2$ and $\{E_2(z_{D.})\}^2$ are. Nevertheless, it does suggest that $\text{var}_1[E_2(\hat{X})]$ would be estimated in an unbiased manner if one could find unbiased estimators for the $\{E_2(z_{j.})\}^2$ and $\{E_2(z_{D.})\}^2$ to plug into (4).

Observe first that $z_{j.}^2$ and $z_{D.}^2$ are not unbiased estimators of $\{E_2(z_{j.})\}^2$ and $\{E_2(z_{D.})\}^2$. In fact,

$$E_2(z_{j.}^2)=\{E_2(z_{j.})\}^2+\text{var}_2(z_{j.}), \text{ while} \tag{5}$$

$$E_2(z_{D.}^2)=\{E_2(z_{D.})\}^2+\text{var}_2(z_{D.}).$$

These equations hint towards alternative estimators for $\{E_2(z_{j.})\}^2$ and $\{E_2(z_{D.})\}^2$. If v_{2j} and v_{2D}, say, were unbiased estimators of $\text{var}_2(z_{j.})$ and $\text{var}_2(z_{D.})$, respectively, then $z_{j.}^2 - v_{2j}$ would be an unbiased estimator of $E_2(z_{j.}^2)$, while $z_{D.}^2 - v_{2D}$ would be an unbiased estimator of $E_2(z_{D.}^2)$.

From Cochran (1977, p. 143, eq. (5A.68)), one can see that

$$\hat{\text{var}}_{2j}= \sum_{h=1}^{H} (T_h^2/v_h-T_h)[1/(v_h-1)] * [\{\sum_{i\epsilon S_j\cap R_h} y_i^2\} - y_{jh}^2/v_h]$$

and (6)

$$\hat{\text{var}}_{2D}= \sum_{h=1}^{H} (T_h^2/v_h-T_h)[1/(v_h-1)] * [\{\sum_{i\epsilon S_D\cap R_h} y_i^2\} - y_{Dh}^2/v_h]$$

are, respectively, unbiased estimators of $\text{var}_2(z_{j.})$ and $\text{var}_2(z_{D.})$. Armed with equations (3) through (6), we are now in position to provide an unbiased estimator for the variance of $\hat{X}$.

3.3 Putting It All Together

Plugging $z_{j.}^2 - \hat{\text{var}}_{2j}$ and $z_{D.}^2 - \hat{\text{var}}_{2D}$ respectively into $\{E_2(z_{j.})\}^2$ and $\{E_2(z_{D.})\}^2$ of equation (4), we have an estimator for the first phase variance of $\hat{X}$. This can then be added to (3) to yield (after some manipulation) the following estimator for the variance of $\hat{X}$ in (1):

$$\hat{\text{var}} = A + B + C, \tag{7}$$

where

$$A= \sum_{D=1}^{L} [n_D/(n_D-1)][\{\sum_{j=1}^{n_D} z_{j.}^2\} - z_{D.}^2/n_D],$$

$$B= \sum_{h=1}^{H} \{(T_h^2/v_h - T_h)\ [1/(v_h\{v_h-1\})]\} * \sum_{D=1}^{L}[n_D/(n_D-1)][\{\sum_{j=1}^{n_D} y_{jh}^2\} - y_{Dh}^2/n_D] - y_{.h}^2)\},$$

$$C= -\sum_{D=1}^{L} f_D n_D/(n_D-1)[\{\sum_{j=1}^{n_D} z_{j.}^2 - \hat{\text{var}}_{2j}\} - \{z_{D.}^2 - \hat{\text{var}}_{2D}\}/n_D],$$

$f_D = n_D/N_D$ is the first phase sampling fraction in district D, and $\hat{\text{var}}_{2j}$ and $\hat{\text{var}}_{2D}$ are defined by equation (6).

Observe that if all the first phase sampling fractions are very small, then the contribution of C to (7) can be ignored. In any event dropping would at worst give $\hat{\text{var}}$ a upward bias, since $E(C)\leq 0$.

Observe further that var would collapse to A if - in addition C being ignorably small - the sampling design had been conventional two-stage sampling; that is, if each strata had been contained within one of the originally sampled area segments, so that $y_{.h}=y_{jh}=y_{Dh}$ and B=0. This should not be surprising, since A is the standard variance estimator in two stage sampling when the first stage is srs with replacement (Cochran, 1977, p. 307). Ignorable first stage sampling fractions blur the distinction between srs with and without replacement.

The right hand side of (7) can, in principle, be negative. This is because B is usually negative (since $y_{.h} \geq y_{Dh} \geq y_{jh}$), while A can theoretically be as small as zero. I doubt, however, that $\hat{var}$ will very often be negative in practice. This contention is supported in empirical work conducted by Kott and Johnston (1988). Applying a formula similar to (6) to data from a USDA survey, they did not (in 41 cases) find a B that was as large in absolute value as even 7% of A.

One final note. Since $B \leq 0$ and $E(C) \leq 0$, using A alone provides a conservative, unambiguously non-negative, estimate for $\hat{var}(\hat{X})$.

References

Cochran, R. and Huddleston, H. (1969). Unbiased Estimates for Stratified Subsample Designs. U.S. Department of Agriculture, Statistical Reporting Service.

Cochran, R. and Huddleston, H. (1970). Unbiased estimates for stratified subsample design. American Statistical Association Proceedings of the Section on Social Statistics, 265-267.

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

Cotter, Jim and Nealon, Jack (1987). Area Frame Design for Agricultural Surveys, U.S. Department of Agriculture, National Agricultural Statistics Service.

Kott, P. S. (1988). Estimating Variances for the June Enumerative Survey. RAD Staff Report No. SRB-NERS-8806, U.S. Department of Agriculture, National Agricultural Statistics Service.

Kott, P. S. and Johnston, R. (1988). Estimating the Non-Overlap Variance Component for Multiple Frame Agricultural Surveys. RAD Staff Report No. SRB-NERS-8805, U.S. Department of Agriculture, National Agricultural Statistics Service.