Survey Methodology, June 1990
Vol. 16, No. 1, pp. 99–103
Statistics Canada

Variance Estimation when a First Phase Area Sample is Restratified

PHILLIP S. KOTT[1]

ABSTRACT

This paper proposes an unbiased variance estimation formula for a two-phase sampling design used in many agricultural surveys. In this design, geographically defined primary sampling units (PSUs) are first selected via stratified simple random sampling; then secondary sampling units within sampled PSUs are restratified based on their characteristics and subsampled in a second phase of stratified simple random sampling.

KEY WORDS: Two-phase sample; Primary sampling unit; Secondary sampling unit; Unbiased.

1. INTRODUCTION

Suppose we have a sample of geographically defined primary sampling units (PSUs) drawn from a stratified area frame. Each sampled PSU contains a number of secondary sampling units (SSUs) which are restratified based on their characteristics. Subsamples of the SSUs are then drawn within each new stratum. To avoid confusion, only the original area strata will hereafter be referred to as strata; the new strata based on SSU characteristics will be referred to as *domains*. Stratified simple random sampling (srs) without replacement is performed at both phases of the sampling design.

This article derives an unbiased variance formula for the estimation strategy described above which is used in many agricultural surveys (for example, see Kott and Johnston 1988) but is not restricted to such surveys. The formula is a generalization of a suggestion by Cochran and Huddleston (1969, 1970), who assumed unstratified srs in the first sampling phase. It is also a special case of a variance formula in Särndal and Swensson (1987). The Särndal and Swensson formula (their equation (4.4)) depends on the calculation of a joint inclusion probability for each pair of subsampled SSUs. This proves cumbersome for the particular application under study because there are six distinct situations which need to be considered (depending on whether or not the two SSUs come from the same PSU, stratum, and/or domain). The derivation presented here follows a different line of reasoning entirely.

2. PRELIMINARIES

Suppose we start with an area survey consisting of n_h (out of N_h) PSUs from each of H strata. The SSUs within sampled PSUs are then restratified into D domains. Within domain d, m_d (out of M_d) SSUs are subsampled. Both phases of the sampling design are stratified srs without replacement.

Let us concentrate on estimating the total for a particular item of interest. To this end, let

[1] Phillip S. Kott, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, DC 20250, USA.

S^1 = denote the set of all SSUs within a PSU selected in the first phase of sampling whether these SSUs are in the subsample or not,

S_{hj} = denote the set of subsampled SSUs in PSU j of stratum h,

$S_{h\cdot}$ = denote the set of all subsampled SSUs in stratum h,

R_d = denote the set of all subsampled SSUs in domain d,

x_i = denote the value of interest for SSU i,

e_i = $(N_h/n_h)(M_d/m_d)x_i$ (assuming $i \in S_{h\cdot} \cap R_d$) be the "fully expanded" value of interest for SSU i,

$$e_{dhj} = \sum_{i \in S_{hj} \cap R_d} e_i,$$

$$e_{dh\cdot} = \sum_{i \in S_{h\cdot} \cap R_d} e_i,$$

$$e_{d\cdot\cdot} = \sum_{i \in R_d} e_i,$$

$$e_{\cdot hj} = \sum_{i \in S_{hj}} e_i, \text{ and}$$

$$e_{\cdot h\cdot} = \sum_{i \in S_{h\cdot}} e_i.$$

Note that when S_{hj} is empty, e_{dhj} and $e_{\cdot hj}$ are zero. Likewise when $S_{h\cdot}$ is empty, $e_{dh\cdot}$ and $e_{\cdot h\cdot}$ are zero, and when R_d is empty e_{dhj}, $e_{dh\cdot}$, and $e_{d\cdot\cdot}$ are zero.

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

$$\hat{X} = \sum_{d=1}^{D} \sum_{i \in R_d} e_i. \tag{1}$$

To see this, observe that $\tilde{X} = \sum_{i \in S^1} (N_D/n_D)x_i$ is an unbiased estimator of X with respect to the first phase of sampling, while $\hat{X}$ is an unbiased estimator of $\tilde{X}$ with respect to the second sampling phase. Mathematically, $E_1(\tilde{X}) = X$ and $E_2(\hat{X}) = \tilde{X}$, which implies $E(\hat{X}) = E_1E_2(\hat{X}) = X$.

3. VARIANCE OF $\hat{X}$

From any of a number of textbooks on sampling theory (*e.g.*, Cochran 1977, p. 276), we know that the variance of a two-phase estimator like $\hat{X}$ is

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

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 often called the first phase variance because it equals the variance that would be obtained if every SSU within a sampled PSU were part of the subsample. The second term in (2) is often called 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 value of interest for a PSU in the first phase sample can only be estimated using the subsample. As is well known, putting an estimated PSU total in place of a real total in the usual one-phase variance formula biases the resulting estimator.

3.1 Second Phase Variance Estimation

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

Now given our particular S^1,

$$\hat{\text{var}}_2 = \sum_{d=1}^{D} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i\in R_d} e_i^2\right\} - e_{d..}^2/m_d\right] \tag{3}$$

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 PSU j within stratum h. The value $e_{.hj}$ is an unbiased estimator of (N_h/n_h) times the total value among all SSUs in PSU j whether in the current subsample or not. Consequently, $E_2(e_{.hj})$ is exactly equal to (N_h/n_h) times the total value among all SSUs in PSU j. With this in mind, the following would be an unbiased estimator of the first phase variance of $\hat{X}$:

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

$$\sum_{h=1}^{H} (1 - n_h/N_h)[n_h/(n_h - 1)]\left[\sum_{j=1}^{n_h} \{E_2(e_{.hj})\}^2 - \{E_2(e_{.h.})\}^2/n_h\right]. \tag{4}$$

Taken as is, equation (4) is of little use since it supposes we know what the $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ 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 $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ to plug into (4).

Observe first that $e_{.hj}^2$ and $e_{.h.}^2$ are *not* unbiased estimators of $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$. In fact,

$$E_2(e_{.hj}^2) = \{E_2(e_{.hj})\}^2 + \text{var}_2(e_{.hj}),$$

while (5)

$$E_2(e_{.h.}^2) = \{E_2(e_{.h.})\}^2 + \text{var}_2(e_{.h.}).$$

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

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

$$\hat{\text{var}}_{2hj} = \sum_{d=1}^{D} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i \in S_{hj} \cap R_d} e_i^2\right\} - e_{dhj}^2/m_d\right] \tag{6}$$

and

$$\hat{\text{var}}_{2h} = \sum_{h=1}^{H} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i \in S_{h.} \cap R_d} e_i^2\right\} - e_{dh.}^2/m_d\right]$$

are, respectively, unbiased estimators of $\text{var}_2(e_{.hj})$ and $\text{var}_2(e_{.h.})$.

3.3 Putting It All Together

Observe that combining equations (3) and (6) can yield (after some manipulation) this estimator for the second phase variance of $\hat{X}$:

$$\hat{\text{var}}_2 = \sum_{h=1}^{H} [n_h/(n_h - 1)] \sum_{j=1}^{n_h} \hat{\text{var}}_{2hj} - \hat{\text{var}}_{2h}/(n_h - 1) +$$

$$\sum_{d=1}^{D} \left\{(1 - m_d/M_d)[1/(m_d - 1)] \cdot \right.$$

$$\left. \left(\sum_{h=1}^{H} [n_h/(n_h - 1)]\left[\left\{ \sum_{j=1}^{n_h} e_{dhj}^2 \right\} - e_{dh.}^2/n_h\right] - e_{d..}^2 \right) \right\}. \tag{7}$$

By plugging $e_{.hj}^2 - \hat{\text{var}}_{2hj}$ and $e_{.h.}^2 - \hat{\text{var}}_{2h}$ respectively into $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ of equation (4), we have the following estimator for the first phase variance of $\hat{X}$:

$$\hat{\text{var}}_1[E_2(\hat{X})] = \sum_{h=1}^{H} (1 - n_h/N_h)[n_h/(n_h - 1)] \cdot$$

$$\left[\left\{ \sum_{j=1}^{n_h} e_{.hj}^2 - \hat{\text{var}}_{2hj} \right\} - \{(e_{.h.}^2 - \hat{\text{var}}_{2h})\}/n_h\right].$$

This can then be added to (7) to yield the following estimator for the variance of $\hat{X}$ in (1):

$$\hat{\text{var}} = A + B + C, \tag{8}$$

where

$$A = \sum_{h=1}^{H} [n_h/(n_h - 1)]\left[\left\{\sum_{j=1}^{n_h} e_{\cdot hj}{}^2\right\} - e_{\cdot h\cdot}{}^2/n_h\right],$$

$$B = \sum_{d=1}^{D}\left\{(1 - m_d/M_d)\ [1/(m_d - 1)] \cdot \left(\sum_{h=1}^{H} [n_h/(n_h - 1)]\left[\left\{\sum_{j=1}^{n_h} e_{dhj}{}^2\right\} - e_{dh\cdot}{}^2/n_h\right] - e_{d\cdot\cdot}{}^2\right)\right\},$$

$$C = -\sum_{h=1}^{H} f_h n_h/(n_h - 1)\left[\sum_{j=1}^{n_h} \{e_{\cdot hj}{}^2 - \hat{\text{var}}_{2hj}\} - \{e_{\cdot h\cdot}{}^2 - \hat{\text{var}}_{2h}\}/n_h\right],$$

$f_h = n_h/N_h$ is the first phase sampling fraction in stratum h, and $\hat{\text{var}}_{2hj}$ and $\hat{\text{var}}_{2h}$ are defined by equation (6).

Observe that if all the first phase sampling fractions are very small, then the contribution of C to (8) can be ignored. In any event dropping C would at worst give vâr an upward bias, since $E(C) \leq 0$.

Observe further that vâr would collapse to A if – in addition to C being ignorably small – the sampling design had been conventional two-stage sampling; that is, if each domain had been contained within one of the originally sampled PSU's so that $y_{d\cdot\cdot} = y_{dhj} = y_{dh\cdot}$ 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 (8) can, in principle, be negative. This is because B is often negative (since $y_{d\cdot\cdot} \geq y_{dh\cdot} \geq y_{dhj}$), while A can theoretically be as small as zero. Kott and Johnston (1988) applied a formula similar to (6) to data from a US Department of Agriculture survey. In the 41 cases they examined the absolute value of B was always less than 7% of A.

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

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. *Proceedings of the Section on Social Statistics, American Statistical Association*, 265-267.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd. ed.). New York: Wiley.

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.

SÄRNDAL, C.E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.