SUCCESSIVE SAMPLING OF TWO OVERLAPPING FRAMES

Chapman P. Gleason and Robert D. Tortora
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture

1. Introduction

The theory of successive sampling (sampling on two or more occasions, double sampling or two-phase sampling) has been studied by many authors [1, 2, 5, 6, 7, 15]. The purpose has been to realize gains in precision of estimators or reduction in sample size for periodic surveys [14].

Multiple frame sampling theory has been studied by several authors [3, 4] and various estimators suggested. Hartley [9, 10], Lund [12] and Fuller and Burmeister [8] all give estimates of totals for samples selected from two overlapping frames.

In this paper the theory of successive sampling is applied to the estimation of the domain means in the multiple frame situation. The notion and nomenclature for integration of these two theories is given below.

2. Definitions and Notion

Consider two sampling frames A and B and assume that a simple random sample of size $_1n_A$ has been drawn from frame A and a simple random sample size $_1n_B$ has been drawn from frame B on the first occasion. Frame A has N_A units in the population and frame B has N_B units. Now assume that every unit in the population belongs to at least one of the frames and it is possible to record for each sample unit whether or not it belongs to the other frame (i.e. the duplication or overlap determination is made). Also, assume that frames A and B do not change between occasions. Hence, we can divide every unit in the population into the following domains:

domain (a) the unit belongs to Frame A only,

domain (b) the unit belongs to Frame B only,

domain (ab) the unit belongs to both frames.

The frame size N_A, N_B are known, but the domain sizes N_a, N_b, N_{ab} may be known or unknown.

We have,

$$N_A = N_a + N_{ab}, \quad N_B = N_b + N_{ab},$$

$$N = N_a + N_b + N_{ab} = N_a + N_B = N_b + N_A$$

On the first occasion let $_1n_a$, $_1n_{ab}$ be respectively the number of units in domain a and the number of units in the domain ab from frame A. For frame B, $_1n_b$ and $_1n_{ba}$ are similarly defined. On the second occasion, select a simple random sample of size m_a from $_1n_a$ units in domain a and a sample of size m_{ab} from the $_1n_{ab}$ units in domain ab. Select a simple random sample of size m_b and m_{ba} from the domains in frame B. Also on the second occasion, let a sample of size $_2u_A$ (u for unmatched) units be selected from the $N_A - {}_1n_A$ units not selected in the first sample from frame A. A similar selection of size $_2u_B$ is made from the $N_B - {}_2n_B$ units in frame B. Let $_2u_a$, $_2u_{ab}$, $_2u_b$ and $_2u_{ba}$ be the unmatched units on the 2nd occasion in domains a and ab respectively and frames A and B respectively. Similarly let, $_1u_a$, $_1u_{ab}$, $_1u_b$, and $_1u_{ba}$ be the units in the first sample unmatched on the second occasion. See Figure 1.

So,

m_A = Frame A total number of units matched between occasions = $m_a + m_{ab}$

m_B = Frame B total number of units matched between occasions = $m_b + m_{ba}$

$_2u_A$ = Frame A total number of unmatched units on 2nd occasion = $_2u_a + {}_2u_{ab}$

$_2u_B$ = Frame B total number of unmatched units on 2nd occasion = $_2u_a + {}_2u_{ba}$

Of course $_2u_a$, $_2u_{ab}$, $_2u_b$, $_2u_{ba}$ are random variables whereas m_a, m_{ab}, m_b, m_{ba} are fixed.

3. Multiple Frame Estimation

The multiple frame estimator of the population total proposed by Hartley [9] is

$$(1)\quad \hat{Y}_H = N_a \bar{y}_a + N_{ab}\,(p\,\bar{y}_{ab} + q\,\bar{y}_{ba}) + N_b \bar{y}_b$$

where $\bar{y}_a$ ($\bar{y}_b$) is an estimate of the domain a(b) mean (provided $n_a > 0$, $n_b > 0$) based on those units that were sampled in frame A(B) and also were determined to be in frame B(A), and p and q are optimally determined weights attached to the overlap domain ab from frame A and frame B, respectively.

We assume the domain sizes N_a, N_b and N_{ab} are known. When the same frames, or updated versions of the frames, are used for many surveys for a period of time, as is the case in agriculture, this assumption may be realistic. Further, since the survey has been conducted once on the first occasion we will have estimates of N_a, N_b, N_{ab} from this sample. We can then consider N_a, N_b, N_{ab} known on the second occasion and use the theory of known domain sizes. Then the variance of $\hat{Y}_H$ is, ignoring finite population correction factors,

$$\mathrm{Var}(Y_H) = \frac{N_A^2}{2n_A}\{(1-\alpha)\,\sigma_a^2 + \alpha\rho^2\,\sigma_{ab}^2\} + \frac{N_B^2}{2n_B}\{(1-\beta)\,\sigma_b^2 + \beta q^2\,\sigma_{ab}^2\},$$

where $= N_{ab}/N_A$, $\beta = N_{ab}/N_B$, and σ_a^2, σ_b^2, and σ_{ab}^2 are the domain population variances.

Now we turn to the successive sampling situation. The regression estimate of the i^{th} domain, i = a, b, ab(ba) based on the matched portion of the sample is

$$(3)\quad {}_2\bar{y}'_{im} = {}_2\bar{y}_{im} + b_i\,({}_1\bar{y}_i - {}_1\bar{y}_{im}),$$

i = a, b, ab, ba

where the pre-subscript (1 or 2) denotes the occasion, m denotes matched and b_i is the sample regression coefficient between the survey variable on occasions. Note we have taken a small liberty with the notation since i = ab and i = ba denote two different estimates of the overlap domain ab depending on whether the sample was drawn in frame A or frame B, respectively.

The best combined estimator of each of the respective domain means is found by weighting the two independent estimates together from each of the frames in the following manner:

$$(4)\quad {}_2\bar{y}'_i = \ell_i\,{}_2\bar{y}'_{iu} + (1-\ell_i)\,{}_2\bar{y}'_{im}$$

i = a, b, ab, ba

where ${}_2\bar{y}'_{iu} = {}_2\bar{y}_{iu}$, the mean of the units unmatched the 2nd occasion, for the i-th domain frame combination. The constant ℓ_i is determined so that the variance of ${}_2\bar{y}'_i$ is minimum [14].

The variance of ${}_2\bar{y}'_i$ is

$$(5)\quad \mathrm{Var}({}_2\bar{y}'_i) = \frac{\sigma_i^2\,({}_1n_i - \rho_i^2\,{}_1u_i)}{{}_1n_i\,{}_2n_i - \rho_i^2\,{}_1u_i\,{}_2u_i}$$

i = a, b, ab, ba

where, ${}_1u_i$ (${}_2u_i$) = sample size for the i-th domain of those questioned on the first (second) occasion only, and ρ_i is the domain correlation coefficient between variables on the 1st and 2nd occasion (note $\rho_{ab} = \rho_{ba}$), and σ_i^2 is the within domain variance. Note that the only random variable in (5) is ${}_2u_i$.

The optimum percent to match in each domain /frame combination can be found by minimization of (5) with respect to variation in $\lambda_i = m_i/{}_2n_i$. In this case one can show the optimum percent to match is

$$(6)\quad \lambda_{\mathrm{opt}\ i} = \frac{{}_1n_i\sqrt{1-\rho_i^2}}{{}_2n_i\,(1+\sqrt{1-\rho_i^2})}$$

i = a, b, ab, ba

Now we comine successive and multiple frame sampling to obtain an estimate of a population total. Assume first that the domain sizes are known. We use the form of Hartley's estimator given by equation (1) but substitute regression estimates obtained on the second equation for the domain means. We have

$$(7)\quad \hat{Y}'_H = N_a\,{}_2\bar{y}'_a + N_{ab}\,(p\,{}_2\bar{y}'_{ab} + q\,{}_2\bar{y}'_{ba}) + N_b\,{}_2\bar{y}'_b$$

when the domain sizes are assumed known. Lund [12], Fuller and Burmeister, [8], Bosecker and Ford [3], and Huang [11] have improved Hartley's estimator. The latter two sets of authors extend the concepts introduced by Fuller and Burmeister of adding unbiased estimates of zeroes to a stratified design using regression estimators. To compute the variance of $\hat{Y}'_H$ we condition on the set of random variables $S = \{{}_2u_a,\ {}_2u_{ab},\ {}_2u_b,\ {}_2u_{ba}\}$ and use the formula

$$(8)\quad \mathrm{Var}(\hat{Y}'_H) = E\,[\mathrm{Var}(\hat{Y}'_H\,|S)] + \mathrm{Var}\,[E(\hat{Y}'_H\,|S)]$$

Now for any value of p, $E(\hat{Y}'_H\,|S)$ is an unbiased estimate of the total Y. Hence, the last term in (8) is 0. Therefore we need only to calculate $E[\mathrm{Var}(\hat{Y}'_H\,|S)]$.

We have

$$(9)\quad \mathrm{Var}(\hat{Y}'_H\,|S) = N_a^2\,\sigma^2_{2\bar{y}'_a} + p^2\,N_{ab}^2\,\sigma^2_{2\bar{y}'_{ab}} + q^2\,N_{ab}^2\,\sigma^2_{2\bar{y}'_{ba}} + N_b^2\,\sigma^2_{2\bar{y}'_b}.$$

Taking expected values of both sides of (9) we have four terms to evaluate. Consider the first term, viz., $E(N_a^2\,\sigma^2_{2\bar{y}'_a})$. Upon substitution of equation (5) we get

$$(10)\quad E(N_a^2\,\sigma^2_{2\bar{y}'_a}) = N_a^2\;E\left\{\frac{\sigma_a^2\,({}_1n_a - \rho_a^2\,{}_1u_a)}{{}_1n_a\,{}_2n_a - \rho_a^2\,{}_1u_a\,{}_2u_a}\right\}$$

Unfortunately equation (10) involves the inverse of the hypergeometric random variable ${}_2u_a$. We avoid computing the exact expection for the inverse of any random variable Z by using the well-known approximation

$$(11)\quad E(\tfrac{1}{Z}) \doteq \frac{1}{E(Z)}\{1 + [CV(Z)]^2\} \doteq \frac{1}{E(Z)}$$ provided the CV(Z) is small.

Letting $Z = {}_1n_a\,{}_2n_a - \rho_a^2\,{}_1u_a\,{}_2u_a$

we get,

$$(12)\quad E(N_a^2\ {}_2\bar{y}_a^2) = \frac{N_a^2\ {}_2\sigma_a^2\ ({}_1n_a - \rho_a^2\ {}_1u_a)}{{}_1n_a\ m_a + ({}_1n_a - \rho^2\ {}_1u_a)\ {}_2u_A\ \phi_A}$$

where,

$$\phi_A = \frac{N_a - {}_1n_a}{N_A - {}_1n_A}\ .$$

In a similar manner all the remaining terms of $E[Var(\hat{Y}_H\ |S)]$ can be evaluated thus the $Var(\hat{Y}_H')$ can be approximately written as

$$(13)\quad Var(\hat{Y}_H') \doteq$$

$$\frac{N_a^2\ \sigma_a^2\ ({}_1n_a - \rho_a^2\ {}_1u_a)}{{}_1n_a\ m_a + ({}_1n_a - \rho_a^2\ {}_1u_a)\ \phi_A\ {}_2u_A} +$$

$$\frac{p^2\ N_{ab}^2\ \sigma_{ab}^2\ ({}_1n_{ab} - \rho_{ab}^2\ {}_1u_{ab})}{{}_1n_{ab}\ m_{ab} + ({}_1n_{ab} - \rho_{ab}^2\ {}_1u_{ab})(1 - \phi_A)\ {}_2u_A} +$$

$$\frac{N_{ab}^2\ (1 - p)^2\ \sigma_{ab}^2\ ({}_1n_{ba} - \rho_{ab}^2\ {}_1u_{ba})}{{}_1n_{ba}\ m_{ba} + ({}_1n_{ba} - \rho_{ab}^2\ {}_1u_{ba})(1 - \phi_B)\ {}_2u_B} +$$

$$\frac{N_b^2\ \sigma_b^2\ ({}_1n_b - \rho_b^2\ {}_1u_b)}{{}_1n_b\ m_b + ({}_1n_b - \rho_b^2\ {}_1u_b)\phi_B\ {}_2u_B}$$

The estimator $\hat{Y}_H'$ is an improvement over $\hat{Y}_H$ since it has smaller variance. We have for all i, $Var({}_2\bar{y}_i'|S) \leq Var({}_2\bar{y}_i'|S)$, so $E[(Var({}_2\bar{y}_i'|S)] \leq E[(Var({}_2\bar{y}_i'|S)]$. Summing over the 3 domains of interest (but four terms) we see that $Var(\hat{Y}_H') \leq Var(\hat{Y}_H)$. Hence the estimator $\hat{Y}_H'$ has greater precision than $\hat{Y}_H$.

By differentiating equation (13) with respect to p, setting the result equal to zero and solving for p we obtain the correct weights for the two estimates, ${}_2\bar{y}_{ab}'$ and ${}_2\bar{y}_{ba}'$, of the domain (ab) total. The optimum for $\hat{Y}_H'$ is

$$(14)\quad p_{opt} = \left(\frac{{}_1n_{ab}\ m_{ab}}{{}_1n_{ab} - \rho_{ab}^2\ {}_1u_{ab}} + (1 - \phi_A)\ {}_2u_A\right) \div$$

$$\left(\frac{{}_1n_{ab}\ m_{ab}}{{}_1n_{ab} - \rho_{ab}^2\ {}_1u_{ab}} + (1 - \phi_A)\ {}_2u_A + \frac{{}_1n_{ba}\ m_{ba}}{{}_1n_{ba} - \rho_{ab}^2\ {}_1u_{ba}} + (1 - \phi_B)\ {}_2u_B\right)$$

The variance of $\hat{Y}_H'$ with this p becomes

$$(15)\quad Var(\hat{Y}_{H,opt}') =$$

$$\frac{N_a^2\ \sigma_a^2}{\frac{{}_1n_a\ m_a}{{}_1n_a - \rho_a^2\ {}_1u_a} + {}_2u_A\ \phi_A} + \frac{N_b^2\ \sigma_b^2}{\frac{{}_1n_b\ m_b}{{}_1n_b - \rho_b^2\ {}_1u_b} + {}_2u_B\ \phi_B}$$

$$+ N_{ab}\ \sigma_{ab}^2 \div \left(\frac{{}_1n_{ab}\ m_{ab}}{{}_1n_{ab} - \rho_{ab}^2\ {}_1u_{ab}} + (1 - \phi_A)\ {}_2u_A + \frac{{}_1n_{ba}\ m_{ba}}{{}_1n_{ba} - \rho_{ab}^2\ {}_1u_{ba}} + (1 - \phi_B)\ {}_2u_B\right)$$

When the domain sizes N_a, N_b, N_{ab} are unknown we must estimate these quantities. The quantities $N_A \frac{{}_2n_a}{{}_2n_A}$ and $N_B \frac{{}_2n_b}{{}_2n_B}$ are unbiased estimates of N_a and N_b. Further, $N_A \frac{{}_2n_{ab}}{{}_2n_A}$ and $N_B \frac{{}_2n_{ba}}{{}_2n_B}$ are both unbiased estimators of N_{ab}. So substituting these estimators into $\hat{Y}_H'$ (equation 7) an an unbiased estimator of the population total becomes

$$(16)\quad \hat{Y}_H' = N_A \frac{{}_2n_a}{{}_2n_A}\ {}_2\bar{y}_a' + N_A \frac{{}_2n_{ab}}{{}_2n_A}\ p\ {}_2\bar{y}_{ab}' +$$

$$N_B \frac{{}_2n_b}{{}_2n_B}\ {}_2\bar{y}_b + N_B \frac{{}_2n_{ba}}{{}_2n_B}\ q\ {}_2\bar{y}_{ba}'$$

To compute the variance of (16) we use equation (8) again. Since the domain sizes are unknown the $Var(E(\hat{Y}_H'|S))$ in (8) is not 0. The terms that must be added to compute the variance of $\hat{Y}_H'$ when N_a, N_b, N_{ab} unknown are:

$$(17)\quad Var(E(\hat{Y}_H'|S)) =$$

$$\frac{N_a\ N_{ab}\ (N_A - {}_2n_A)}{{}_2n_A\ (N_A - 1)}\ ({}_2\bar{y}_a - p\ {}_2\bar{y}_{ab})^2 +$$

$$\frac{N_b\ N_{ab}\ (N_B - {}_2n_B)}{{}_2n_B\ (N_B - 1)}\ ({}_2\bar{y}_b - q\ {}_2\bar{y}_{ab})^2$$

These two terms can make a substantial contribution to increasing the variance unless the size of the overlap domain is nearly complete or relatively small.

Minimization of (17) plus (13) with respect to p gives the optimum p when N_a, N_b, N_{ab} unknown.

$$p_{opt} = \frac{\dfrac{N_{ab}^2 \sigma_{ab}^2 \rho'_{ba}}{1^{n_{ba}} m_{ba} + \rho'_{ba}(1-\phi_B) 2^{u_B}} + \dfrac{\bar{Y}_{ab}^2 N_b g_B 2^{n_A} + \bar{Y}_a \bar{Y}_{ab} N_a g_A 2^{n_B} - \bar{Y}_b \bar{Y}_{ab} N_b g_B 2^{n_A}}{2^{n_A} 2^{n_B}}}{N_{ab} \sigma_{ab}^2 \dfrac{\rho'_{ab}}{1^{n_a} m_a + \rho'_{ab}(1-\phi) 2^{u_A}} + \dfrac{\rho'_{ba}}{1^{n_{ba}} m_{ba} + \rho'_{ba}(1-\phi) 2^{u_B}} + \dfrac{\bar{Y}_{ab}^2 (N_a g_A 2^{n_B} + N_b g_B 2^{n_A})}{2^{n_A} 2^{n_B}}}$$

where $\rho'_a = 1^{n_a} - \rho_a^2 1^{u_a}$, $g_A = \dfrac{N_A - 2^{n_A}}{N_A - 1}$ and the other quantities ρ'_b, ρ'_{ab}, ρ'_{ba}, g_B are similarly defined.

4. Application

As an example of applying this theory to an ongoing survey, ESCS conducts a biannual survey to estimate cattle inventories using multiple frame sampling. This survey uses a stratified list of cattle operators and an area frame sample of segments of land. The area frame sample is conceptually a complete sampling frame of farms. A "screening" estimator of the form $\hat{Y}_H = N_a \bar{y}_a + N_{ab} \bar{y}_{ba}$ is used to estimate the total inventory for a state.

Using a successive sampling regression estimator in just the nonoverlap domain and calling this $\hat{Y}'_H = N_a \bar{y}'_a + N_{ab} \bar{y}_{ba}$), we get an absolute gain in precision of 23% for the variance of the nonoverlap domain. The percent variance of total variance for the nonoverlap domain decreases from 81% to 77.4% using successive sampling in just this domain, as can be seen in the following tabulation.

Domain	Percent Estimate of the total	Percent of $\hat{Var}(Y_H)$	Percent of $\hat{Var}(Y'_H)$	Correlation Coefficient between surveys
Stratified List Frame (Domain ba)	81.2	18.4	22.6	--
Nonoverlap (Domain a)	18.8	81.6	77.4	.588

5. Future Research

It is planned to apply this theory to the estimation of livestock inventories using USDA/ESCS surveys and compare the estimators to a single time multiple frame estimator. Further theoretical research into "screening" estimators (i.e. p = 0, when the A frame is complete) and the optimum percent to match in each frame is also planned.

6. Summary

A theory for the estimator of domain means using successive sampling in the multiple frame situation has been presented. It was shown that the successive sampling estimator had greater precision than a single time multiple frame estimator.

The case of known and unknown domain sizes is considered. The sampling plan requires that the matched sample on the second occasion be drawn from the sampling units that fall into each of the domains on the first occasion. On the other hand, the unmatched sample (in both frames) is drawn from all units not sampled on the first occasion.

References

1. Avadhani, M.S., Sukhatme, B.V., "Estimation in Sampling on Two Successive Occasions", Statistica Neerlandica, Vol. 26, No. 2, 1972.

2. Avadhani, M.S., Sukhatme, B.V., "A Comparison of Two Sampling Procedures With an Application to Successive Sample", Journal Ryal Stat. Soc. Series C, Vol. 19, pp 251-259.

3. Bosecker, R.R., Ford, B.L., "Multiple Frame Estimation With Stratified Overlap Domain", Sample Survey Research Branch, Research Division, Statistical Reporting Service, USDA, Washington, D.C., 20250, 1976.

4. Cochran, R.S., "Theory and Application of Multiple Frame Surveys", Ph.D. Dissertation, Iowa State University Library, Ames, 1965.

5. DeGraft-Johnson, "Some Contributions to the Theory of Two-Phase Sampling", Ph.D. Dissertation, Iowa State University, Ames, Iowa, 1969.

6. DeGraft-Johnson, K.T., Sedransk, J., "Estimation of Domain Means Using Two-Phase Sampling", Biometrika, 60 pp 387-393, 1973.

7. Eckler, A.R., "Rotation Sampling", Annals of Math Statistics, Vol. 26, pp 664-685.

8. Fuller, W.A., Burmeister, L.F., "Estimation for Samples Selected From Two Overlapping Frames", Statistical Laboratory, Iowa State

University, Ames, Iowa, January 1973.

9. Hartley, H.O., "Multiple Frame Surveys" Proceedings of the Social Statistics Section, ASA, 1962.

10. Hartley, H.O., "Multiple Frame Methodology and Selected Applications", Sankyā, Vol. 36, Series C, Part 3, pp 99-118, 1974.

11. Huang, Her Tzai, "The Relative Efficiency of Some Two-Frame Estimators", Statistical Laboratory, Iowa State University, Ames, Iowa, March 1974.

12. Lund, Richard E., "Estimators in Multiple Frame Surveys", Proceedings of the Social Science Section, ASA, Pittsburgh, PA, 1968.

13. Raj, Des, "On Sampling Over Two Occasions With Probability Proportional to Size", Ann. Math. Statistics, 36, pp 327-330, 1965.

14. Sen, A.R., Sellers, S., Smith, G.E.J., "The Use of a Ratio Estimate in Successive Sampling", Biometrics Vol. 51, pp 673-684, 1975.

15. Singh, D., "Estimates in Successive Sampling Using a Multi-Stage Design", JASA, Vol. 63, pp 99-112, 1968.

Figure 1