SAMPLING FROM TWO OVERLAPPING FRAMES WHEN THE DOMAIN SIZES ARE KNOWN

Leon F. Burmeister, University of Iowa

It may be occasionally necessary or economically convenient to utilize more than one frame in drawing a sample from a population. This paper assumes two frames cover the population and the observational units for the two frames are identical. If the sizes of both frames and the population size are known, the number of elements common to the two frames is also known, as well as the number of elements included only on each individual frame.

Estimators will be developed that have smaller variance than those previously suggested. However, since the number of duplicated elements included in the overall sample is utilized in the proposed estimators, the cost is greater than for previous estimators. Consequently, allocation procedures and situations likely to result in appreciably higher cost will be considered.

The adaptation of a previously suggested method to construct unbiased estimators of the population total with empty domain sample sizes is illustrated.

1.1 Examples

The opportunity to utilize two or more frames of known sizes does occur in practice. Comstock, et. al. (3) describe a study designed to evaluate immunization histories obtained from a sample of the population of Washington County, Maryland. The evaluation was completed by the comparison of historical information from interviews with the results of serologic determinations.

The study was done during the summer of 1968. A 1% systematic sample of the county population was drawn from a list of households obtained in a non-official census conducted in 1963, supplemented by a similar sample of dwelling units added since that time. It seems reasonable to assume that the number of households on both frames were known as well as the total number of households in the county. From this knowledge it is easily determined if the two frames cover the population and if they overlap. If there is an overlap, consideration should be given to the determination of the duplicated households. Since it is not obvious that the costs of doing so can be justified, the improvements in the resulting estimators need to be carefully evaluated.

Other examples of studies that have been conducted based on multiple and possibly overlapping frames include those described by Serfling, Cornell and Sherman (8), Bershad (1) and Cochran (2).

1.2 Notation

It is assumed two frames, A and B, containing N_A and N_B elements respectively are available. The notation of Hartly (5) is adopted and N_{ab} denotes the number of elements included on both frame A and Frame B. N_a is the number of elements occurring only on frame A and N_b is the number of elements occurring on Frame B. Thus:

$$N_A = N_a + N_{ab}, \tag{1.1}$$

$$N_B = N_b + N_{ab} \tag{1.2}$$

and the total number of elements in the population, N, is given by

$$N = N_a + N_b + N_{ab} = N_a + N_B = N_b + N_A. \tag{1.3}$$

The elements contained only on frame A are called domain a, the elements only on frame B domain b and those elements on both frames A and B domain ab. It is assumed that a simple random sample of size n_A is selected from Fr. A & and simple random sample of size n_B is selected from FrB. The number of elements sampled from frame A and contained in domain a is denoted by n_a. The number of elements sampled from frame A and contained in domain ab is denoted by n'_{ab}. The number of sampled elements in domains ab and b drawn from frame B are denoted by n''_{ab} and n_b respectively. Thus

$$n_A = n_a + n'_{ab} \tag{1.4}$$

and

$$n_B = n''_{ab} + n_b \tag{1.5}$$

This completes the description of the problem. It is only one of several frame problems that the sample designer may face. Kish (6) gives several interesting and informative discussions of these additional problems.

2. ESTIMATORS OF THE POPULATION TOTAL

Assume a sample of size n_A is drawn without replacement from frame A, and a sample of size n_B is drawn without replacement from frame B. Assuming simple random sampling in both frames, the probability of being included in the sample, Π_i, can be calculated for elements in each domain. These probabilities will be utilized in the construction of alternative estimators.

$$\text{Prob}\left[\, i^{th} \text{ element in domain a is included in the sample} \,\right] = \frac{n_A}{N_A}, \tag{2.1}$$

$$\text{Prob}\left[\, i^{th} \text{ element in domain b is included in the sample} \,\right] = \frac{n_B}{N_B}, \tag{2.2}$$

and

$$\text{Prob}\left[\, i^{th} \text{ element in domain ab is included in the sample at least once} \,\right] = \frac{n_A N_B + n_B N_A - n_A n_B}{N_A N_B}. \tag{2.3}$$

Lund (7) proposed the estimator

$$\hat{Y}_L = N_a \bar{y}_a + N_{ab} \bar{y}^*_{ab} + N_b \bar{y}_b \tag{2.4}$$

for the case of known domain sizes. The sample total for domain ab,

$$y^*_{ab} = \sum_{i=1}^{n'_{ab}} y_i + \sum_{i=1}^{n''_{ab}} y_i, \tag{2.5}$$

in $\hat{Y}_L$ is based on the $n'_{ab} + n''_{ab}$ elements sampled from frames A and B. If the duplicated elements in domain ab are excluded, the sample mean $\bar{y}^*_{ab}$ becomes

$$\bar{y}_{ab} = \frac{\sum_{i=1}^{n'_{ab}} y_i + \sum_{i=1}^{n''_{ab}} y_i - \sum_{i=1}^{n_d} y_i}{n'_{ab} + n''_{ab} - n_d}, \tag{2.6}$$

where n_d is the number of duplicated items.

Consideration is thus given to

$$\hat{Y}_d = N_a\bar{y}_a + N_{ab}\bar{y}_{ab} + N_b\bar{y}_b \cdot \qquad (2.7)$$

The notation $\hat{Y}_d$ in (2.7) indicates that this estimator of the total is based on the distinct elements included in the sample.

If it can be assumed that n_a, n_{ab} and n_b are each greater than zero, $\hat{Y}_d$ is unbiased. The conditional expectation becomes

$$E(\hat{Y}_d \,|\, n_a>0,\ n_{ab}>0,\ n_b>0)$$

$$= N_aE(\bar{y}_a|n_a>0) + N_{ab}E(\bar{y}_{ab}|n_{ab}>0) + N_bE(\bar{y}_b|n_b>0)$$

$$= N_a\bar{Y}_a + N_{ab}\bar{Y}_{ab} + N_b\bar{Y}_b = Y. \qquad (2.8)$$

It should be noted that Fuller (4) has devised a method of constructing unbiased post-stratified estimators that does not require each domain size to be positive. This method will be considered below.

A comparison of the variances of $\hat{Y}_L$ and $\hat{Y}_d$ reduces to a comparison of the variances of $\bar{y}_{ab}$ and $\bar{y}^*_{ab}$ since the estimators differ only in the estimated mean of the overlap domain. To facilitate this comparison, let

$$\bar{y}^*_{ab} = \frac{y_{ab} + n_d\bar{y}_d}{n_{ab} + n_d} \qquad (2.9)$$

and

$$\bar{y}_{ab} = \frac{y_{ab}}{n_{ab}} \qquad (2.10)$$

where $n_{ab} = n'_{ab} + n''_{ab} - n_d$, y_{ab} is the total of the n_{ab} distinct elements sampled from domain ab and $\bar{y}_d$ is the mean of the n_d duplicated elements. It is assumed that n_{ab} is greater than zero. Conditional on n_{ab} and n_d,

$$\mathrm{Var}(\bar{y}^*_{ab}) = \frac{\mathrm{Var}(y_{ab})+n_d^2\mathrm{Var}(\bar{y}_d)+2n_d\mathrm{Cov}(y_{ab},\bar{y}_d)}{(n_{ab}+n_d)^2} \qquad (2.11)$$

and

$$\mathrm{Var}(\bar{y}_{ab}) = \frac{\mathrm{Var}(y_{ab})}{n_{ab}^2} \cdot \qquad (2.12)$$

To evaluate (2.11) and (2.12), assume the variance-covariance structure of the y_i's to be

$$\begin{bmatrix} 1-1/N_{ab} & -1/N_{ab} & -1/N_{ab} & \cdots & -1/N_{ab} \\ -1/N_{ab} & 1-1/N_{ab} & -1/N_{ab} & & -1/N_{ab} \\ -1/N_{ab} & -1/N_{ab} & 1-1/N_{ab} & & -1/N_{ab} \\ \vdots & & & & \vdots \\ -1/N_{ab} & -1/N_{ab} & -1/N_{ab} & \cdots & 1-1/N_{ab} \end{bmatrix} S^2_{ab} \qquad (2.13)$$

Then the variance of $\bar{y}^*_{ab}$ is greater than or equal to that of $\bar{y}_{ab}$ if

$$n_{ab}^2\,\mathrm{Var}(y_d) + 2n_{ab}^2\,\mathrm{Cov}(y_{ab}, y_d)$$

$$\geq 2n_dn_{ab}\,\mathrm{Var}(y_{ab}) + n_d^2\mathrm{Var}(y_{ab}). \qquad (2.14)$$

Utilization of the variance-covariance matrix, (2.13), allows us to express each term of (2.14) as follows:

$$n_{ab}^2\mathrm{Var}(y_d) = n_{ab}^2n_d\left[\frac{1-(n_d+1)}{N_{ab}}\right]S^2_{ab} \qquad (2.15)$$

$$2n_{ab}^2\mathrm{Cov}(y_{ab},y_d) = 2n_{ab}^2n_d\left[\frac{1-n_{ab}}{N_{ab}}\right]S^2_{ab} \qquad (2.16)$$

$$2n_{ab}n_d\mathrm{Var}(y_{ab}) = 2n_{ab}^2n_d\left[\frac{1-(n_{ab}+1)}{N_{ab}}\right]S^2_{ab} \qquad (2.17)$$

and

$$n_d^2\mathrm{Var}(y_{ab}) = n_{ab}n_d^2\left[\frac{1-(n_{ab}+1)}{N_{ab}}\right]S^2_{ab} \cdot \qquad (2.18)$$

If n_d equals zero, equality holds in expression (2.14) since the estimators $\bar{y}_{ab}$ and $\bar{y}^*_{ab}$ are identical. If n_{ab} and n_d are both greater than zero, (2.14) may be written as

$$(n_{ab}-n_d) > -\frac{(n_{ab}+n_d)}{N_{ab}}. \qquad (2.19)$$

Thus $\bar{y}_{ab}$ has smaller variance than does y^*_{ab} for all positive values of n_{ab} and n_d.

One of the reasons for calculating the probability of selection, Π_i, for each element in the population now is evident. The Π_i's clearly demonstrate that every element in domain ab has the same probability of being included in the sample. This is required if the simple mean of the distinct elements sampled from domain ab is to be used as an unbiased estimator of the domain mean.

3. UNBIASED ESTIMATOR OF Y WITH EMPTY DOMAINS

It was noted above that Fuller (4) has devised a scheme to construct unbiased post-stratified estimators. This scheme does not require the usual assumption of non-empty strata. Fuller's general construction will be reviewed and then his approach will be applied to this problem.

Assume a random sample of size n has been drawn from a population. After the sample has been taken, the sampled elements are classified as members of two strata. Assume that the population is such that the population proportion of elements contained in stratum one, P_1, and the proportion contained in stratum two, $P_2 = 1 - P_1$, are known. Fuller then considers the general estimator

$$A_i\bar{y}_1 + (1-A_i)\bar{y}_2 \,, \qquad (3.1)$$

where

$\bar{y}_1$ = the sample mean of the characteristic y for stratum one,

$\bar{y}_2$ = the sample mean of the characteristic y for stratum two

and

A_i = the weight applied to the mean of stratum one for samples with i (i = 0, 1, . . . , n) sample elements in stratum one.

It is further assumed that $0 \leq A_i \leq 1$ for all i, $A_o = 0$ and $A_n = 1$.

Minimization of the conditional mean square error of (3.1) yields

$$A_i = \frac{if_{2i}S_2^2 + i(n-i)P_1(\bar{Y}_1-\bar{Y}_2)^2}{(n-i)f_{1i}S_1^2 + if_{2i}S_2^2 + i(n-i)(\bar{Y}_1-\bar{Y}_2)^2} \quad (3.2)$$

where $f_{1i} = \frac{N_1 - i}{N_1}$

and $f_{2i} = \frac{N_2 - n + i}{N_2}$.

The estimator (3.1) employing the weight (3.2) is in general biased. However, Fuller shows it is possible to derive weights A_i such that the estimator (3.1) is unbiased.

Fuller also extends the above development for two post-strata to the general case of more than two strata. First the strata must be arranged in a natural order. Then the strata are repeatedly divided into groups of two. Beginning at the finest subdivision, an unbiased estimator is constructed for each pair of strata.

In this problem, the situation is as shown in Table 1.

Table 1 A DESCRIPTION OF THE POST-STRATA RESULTING FROM TWO OVERLAPPING SAMPLING FRAMES

Stratum	Proportion of pop. in stratum	Stratum ID	Sample number	Sample mean
a	N_a/N	1 1	n_a	$\bar{y}_a$
ab	N_{ab}/N	2 1	n_{ab}	$\bar{y}_{ab}$
b	N_b/N	2 2	n_b	$\bar{y}_b$

Table 1 indicates that the first division of the strata is into stratum a and strata ab and b. Thus an unbiased estimator will be constructed first for strata ab and b. It is to be remembered that n_a, n_{ab} or n_b may be zero in the development of this unbiased estimator.

In the development of the scheme, it is convenient to let $n = n_a + n_{ab} + n_b$, $n_1 = n_a$, $n_2 = n_{ab} + n_b$, $P_{21} = N_{ab}/N_B$, $P_{22} = N_b/N_B$ and $P_1 = N_a/N$.

The estimator for this specific problem is

$$\bar{y}_{\bar{y}} = N\left[A_1\bar{y}_a + A_2(A_{21}\bar{y}_{ab} + A_{22}\bar{y}_b)\right] , \quad (3.3)$$

where the general expressions for the weights in (3.3) are

$$A_1 = (1-A_2) = \frac{n_a + n_a(n_{ab}+n_b)\left[\frac{N_aM_1}{N} + \lambda_{M_1}\right]}{n+n_a(n_{ab}+n_b)M_1} \quad (3.4)$$

and

$$A_{21} = (1-A_{22}) = \frac{n_{ab}+n_{ab}n_b\left[\frac{N_{ab}}{N_B}M_2 + \lambda_{M_2}\right]}{(n_b+n_{ab}) + n_{ab}n_bM_2} , \quad (3.5)$$

where

$$\lambda_{M_1} = \frac{(\bar{Y}_a-\bar{Y}_b)^2}{S_w^2} ,$$

$$\lambda_{M_2} = \frac{(\bar{Y}_{ab}-\bar{Y}_b)^2}{S_w^2} ,$$

$$S_1^2 = S_2^2 = S_w^2 > 0,$$

$$\lambda_{M_1} = \frac{F_1 - \sum_{i=1}^{n-1} \frac{i+i(n-i)P_1M_1}{n+i(n-i)M_1}}{\sum_{i=1}^{n-1} \frac{P_i\, i(n-i)}{n+i(n-i)M_1}}$$

and

$$\lambda_{M_2} = \frac{F_{21} - \sum_{i=1}^{n_2-1} \frac{i+i(n_2-i)P_{21}M_2}{n_2+i(n_2-i)M_2}}{\sum_{i=1}^{n_2-1} \frac{P_i\, i(n_2-i)}{n_2+i(n_2-i)M_2}} . \quad (3.6)$$

F_{21} in (3.6) is P_{21}-Prob $[n_{ab} = (n_{ab}+n_b)$ given strata ab and b contain $n_{ab}+n_b$ sample elements$]$. That is, F_{21} is N_{ab}/N_B minus the probability that all n_B sampled elements fall in domain ab. P_i in expression (3.6) is the probability domain ab contains i units given that n_2 units have been selected from domains ab and b.

4. COST CONSIDERATIONS

Cost considerations are now introduced into the efficiency comparisons. The employment of the proposed estimator, $\hat{\bar{Y}}_d$, necessitates the identification of duplicated elements in domain ab. Lund's estimator (2.4) is a special case of Hartley's(5) procedure which utilizes a weighted average of $\bar{y}'_{ab}$ and $\bar{y}''_{ab}$ and does not require this identification. Hartley gave expressions for the sampling fractions n_A/N_A and n_B/N_B that minimize the variance of his estimator subject to the cost restraint

$$C = c_An_A + c_Bn_B . \quad (4.1)$$

Lund's estimator results when these optimum sampling fractions are used to solve the bi-quadratic equation given by Hartley for the value of the weight p. Another possibility is the retention of elements from domain ab from one frame only. This procedure was employed by the Bureau of the Census in a 1949 study (1). This procedure is a special case of Hartley's procedure with the weight p = 1.

A third procedure is to merge the two frames before sampling and remove the duplicated elements Once the merging has been completed, any number of sampling schemes could be utilized. For example, one could employ stratified random sampling where the strata are the three domains a, b and ab.

The variance of $\hat{\bar{Y}}_d$ (2.7) can be minimized subject to the cost constraint (4.1). A system of three equations in three unknowns, f_A, f_B and λ results. Utilization of the ratio of two of these equations reduces the system to the following system of two equations in two unknowns, f_A and f_B:

(4.2)

$$\frac{c_A N_A}{c_B N_B} = \frac{f_B^2\left[N_a\sigma_a^2(f_A+f_B-f_Af_B)^2 + N_{ab}\sigma_{ab}^2 f_A^2(1-f_B)\right]}{f_A^2\left[N_b\sigma_b^2(f_A+f_B-f_Af_B)^2 + N_{ab}\sigma_{ab}^2 f_B^2(1-f_A)\right]}$$

and

$$C = c_A f_A N_A + c_B f_B n_B \quad . \qquad (4.3)$$

Solution of these two equations requires the solution of a sixth degree equation.

Rather than solving a sixth degree equation, the following iterative procedure may be used. Let

$$r = \frac{f_A}{f_B} \quad .$$

An initial value for r is

$$r_1 = \sqrt{\frac{c_B}{c_A}} \quad .$$

Note that (4.2) may be expressed as (4.4)

$$\frac{n_A^2}{n_B^2} = \frac{c_B N_A\left[N_a\sigma_a^2\left[f_B+f_A(1-f_B)\right]^2 + N_{ab}\sigma_{ab}^2 f_A^2(1-f_B)\right]}{c_A N_B\left[N_b\sigma_b^2\left[f_B+f_A(1-f_B)\right]^2 + N_{ab}\sigma_{ab}^2 f_B^2(1-f_A)\right]}$$

Set $1-f_A$ and $1-f_B$ equal to 1 and divide each term of the right hand side of (4.4) by f_B^2 to obtain

$$r_{i+1}^2 = \frac{c_B N_B\left[N_a\sigma_a^2(1+r_i)^2 + N_{ab}\sigma_{ab}^2 r_i^2\right]}{c_A N_A\left[N_b\sigma_b^2(1+r_i)^2 + N_{ab}\sigma_{ab}^2\right]} \qquad (4.5)$$

Thus this procedure is repeated until r_i shows an arbitrarily small change from one iteration to the next.

The c_A term in (4.1) includes all the costs involved in taking a sample of size n_A from frame A. Therefore, define c_A as

$$c_A = c_{sA} + c_{cA} + c_{oA} \qquad (4.6)$$

That is, c_A includes the cost of selection, the cost of classification into the proper domain and the cost of observation. The scheme used in constructing $\hat{Y}_d$ includes the costs of selection and classification from both frames. It may be possible to select elements to be included in the sample and remove the duplicated elements before the actual observations are made. In this situation, the costs of observation are diminished since n_d fewer observations are made; however, the elements drawn from domain ab must now be checked for duplication. The new cost equation then becomes

$$C' = (c_{sA}+c_{cA}+c_{oA})n_A + (c_{sB}+c_{cB}+c_{oB})n_B - n_d c_o' + c_d n_{ab}' n_{ab}'' \quad ,$$

where

$$c_o' = \max(c_{oA}, c_{oB}) \qquad (4.7)$$

and c_d is the cost of determining duplications. In (4.7) it is assumed that each sampled element is classified into its proper domain upon selection. Thus c_d is multiplied by the product of n_{ab}' and n_{ab}''.

We conclude that if (4.1) is the correct cost equation, $\hat{Y}_d$ is preferred to $\hat{Y}_L$. The superiority of $\hat{Y}_d$ over $\hat{Y}_L$ is not so clear if (4.7) is the appropriate cost equation, but it should be noted that the increase in cost when employing $\hat{Y}_d$ may be small. It is quite likely the cost of observation, which would include travel costs and expenses of an enumerator, would be larger than the cost of checking elements for duplication in the office before the fieldwork is started. That is, if it is reasonable to assume that $n_d c_o'$ is of the same magnitude as $n_{ab}' n_{ab}'' c_d$, $\hat{Y}_d$ would be superior to $\hat{Y}_L$ or any of the other special forms of Hartley's original estimator. However, it may be impossible to determine the duplicated elements before the fieldwork is done. In this case, $\hat{Y}_d$ would require more expense than $\hat{Y}_L$.

REFERENCES

1. Bershad, M.A., "A Sample Survey of Retail Stores," Sample Survey Methods and Theory, Vol. I, New York: John Wiley and Sons, Inc., 1953, 516-58.

2. Cochran, R. S., "The Estimation of Domain Sizes when Sampling Frames are Interlocking," Proceedings of the Social Science Section of the American Statistical Association Meeting, Washington, D.C., 1967.

3. Comstock, G. W., et. al., "Validity of Interview Information in Estimating Community Immunization Levels," Health Services Reports 88 (October, 1973), 750-7.

4. Fuller, W. A., "Estimation Employing Post Strata," Journal of the American Statistical Association, 61 (December, 1966), 1172-83.

5. Hartley, H. O., "Multiple Frame Surveys," Proceedings of the Social Science Section of the American Statistical Association meeting, Minneapolis, Minnesota, 1962.

6. Kish, L., Survey Sampling, New York: John Wiley and Sons, Inc., 1965.

7. Lund, R. E., "Estimators in Multiple Frame Surveys," Proceedings of the Social Science Section of the American Statistical Association meeting, Pittsburg, Pennsylvania, 1968.

8. Serfling, R.E., Cornell, R.G. and Sherman, I. L., "The CDC Quota Sampling Technique with Results of 1959 Poliomyelitis Vaccination Surveys," American Journal of Public Health, 50 (December, 1960), 1847-57.