RELATIVE EFFICIENCY OF SOME TWO-FRAME ESTIMATORS

H. Huang, Minnesota State Department of Education

1. Introduction

In sample surveys, a complete frame is often unavailable or too expensive to construct. When these situations arise, a survey practitioner may use multiple frames. One of the first applications of the multiple frame procedure appeared in the "Sample Survey of Retail Stores" conducted by the United States Bureau of the Census in 1949, reported by Bershad [1]. Hartley [5]gave a complete description of multiple frame concepts. Cochran [2,3], Lund [7], and others have also considered the problem.

Fuller and Burmeister [4] proposed some alternative estimators. In this study, agricultural data is used to illustrate their multiple regression estimators for population totals. The relative efficiencies of these estimators to Hartley's estimator are presented.

2. Notation and Estimators for Population Totals

We assume that two frames, A and B, containing N_A and N_B elements respectively, are available. We denote by N_{ab} the number of elements included in both frame A and frame B, by N_a the number of elements occurring only in frame A, and by N_b the number of elements occurring only in frame B. Thus

$$N_A = N_a + N_{ab} ,$$

$$N_B = N_b + N_{ab}$$

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

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

We refer to the elements contained only in Frame A as domain a, the elements only in frame B as domain b and those elements in both frames A and B as domain ab. Domain ab is sometimes called the overlap domain.

Given that simple random samples of size n_A and n_B are selected from frame A and frame B, respectively, Hartley [5] proposed the following estimator of the population total for the characteristic, Y:

$$\hat{Y}_H = \hat{Y}_a + \hat{Y}_B + P(\hat{Y}'_{ab} - \hat{Y}''_{ab}) \qquad (2.1)$$

where

$$\hat{Y}_B = \hat{Y}_b + \hat{Y}''_{ab}$$

$\hat{Y}_a$ is the estimator of the total of Y for domain a obtained from the sample from frame A,

$\hat{Y}'_{ab}$ is the estimator of the total of Y for domain ab obtained from the sample from frame A,

$\hat{Y}_b$ is the estimator of the total of Y for domain b obtained from the sample from frame B,

$\hat{Y}''_{ab}$ is the estimator of the total of Y for domain ab obtained from the sample from frame B, and

P is the number chosen to minimize the variance of the estimator.

Fuller and Burmeister [4] suggested the estimator:

$$\tilde{Y}_r = \hat{Y}_a + \hat{Y}_B + b_1 (\hat{N}'_{ab} - \hat{N}''_{ab}) + b_2 (\hat{Y}'_{ab} - \hat{Y}''_{ab}) , \qquad (2.2)$$

where

$\hat{N}'_{ab}$ is an estimator of the number of elements in domain ab estimated from the sample from frame A,

$\hat{N}''_{ab}$ is an estimator of the number of elements in domain ab estimated from the sample from frame B,

b_1 and b_2 are numbers chosen to minimize the variance of the estimator.

The estimators $\hat{N}'_{ab} - \hat{N}''_{ab}$ and $\hat{Y}'_{ab} - \hat{Y}''_{ab}$ are unbiased estimators of zero. Both $\tilde{Y}_r$ and $\hat{Y}_H$ are recognizable as multiple regression estimators. Therefore, Hartley's estimator, $\hat{Y}_H$, is inefficient relative to the Fuller-Burmeister estimator $\tilde{Y}_r$, if the partial correlation between $\hat{Y}_a + \hat{Y}_B$ and $\hat{N}'_{ab} - \hat{N}''_{ab}$, after adjusting for $\hat{Y}'_{ab} - \hat{Y}''_{ab}$, is not zero.

In our application of the theory frame A is a stratified list frame and frame B is a complete area frame. The sample elements selected from the area frame can be identified as belonging or not belonging to the list frame A. The Hartley estimator remains the same for a stratified list, but the Fuller-Burmeister estimators can be extended to include additional unbiased estimators of zero. We define

$$\tilde{Y}_{mR} = \hat{Y}_B + \sum_{i=1}^{L} b_{1i} (\hat{Y}'_{iab} - \hat{Y}''_{iab}) + \sum_{j=1}^{m} b_{2j} (\hat{N}'_{jab} - \hat{N}''_{jab}) , \qquad (2.3)$$

where

$\hat{N}'_{jab}$ is an estimator of the number of elements in domain ab of the jth subgroup obtained from the sample of frame A,

$\hat{N}''_{jab}$ is an estimator of the number of elements in domain ab of the j^{th} subgroup obtained from the sample of frame B,

$\hat{Y}'_{iab}$ is an estimator of the total of Y for domain ab of the i^{th} stratum obtained from the sample of frame A,

$\hat{Y}''_{iab}$ is an estimator of the total of Y for domain ab of the i^{th} stratum obtained from the sample of frame B,

L is the total number of strata,

and

m is the number of subgroups on which the estimator of the number of elements in domain ab are obtained and included in the estimator.

We note that $\hat{N}'_{jab} - \hat{N}''_{jab}$ may be an estimator of zero obtained from a particular stratum or from a combination of several strata. We also define n_{Ai}, $i = 1, L$, as the size of sample selected from the i^{th} stratum of frame A.

When frame B is a complete area frame, the variance of $\hat{Y}_H$ and $\tilde{Y}_r$ are given as follows:

$$V(\hat{Y}_H) = V(\hat{Y}_B) - \frac{[\mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab})]^2}{V(\hat{Y}'_{ab}) + V(\hat{Y}''_{ab})} , \qquad (2.4)$$

$$V(\tilde{Y}_r) = V(\hat{Y}_B) - b_1 \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{ab}) - b_2 \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab}) \qquad (2.5)$$

where

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} V(\hat{N}''_{ab}) & \mathrm{Cov}(\hat{N}''_{ab}, \hat{Y}''_{ab}) \\ \mathrm{Cov}(\hat{N}''_{ab}, \hat{Y}''_{ab}) & V(\hat{Y}'_{ab}) + V(\hat{Y}''_{ab}) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{ab}) \\ \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab}) \end{bmatrix} .$$

To obtain the variance of $\tilde{Y}_{mR}$, we write (2.3) as

$$\tilde{Y}_{mR} = \hat{Y}_B + \hat{X}\hat{\beta} \qquad (2.6)$$

where

$$\hat{\beta}' = (b_{11}, b_{12}, \ldots, b_{1L}, b_{21}, b_{22}, \ldots, b_{2m})$$

$$\hat{X} = \hat{X}_1 - \hat{X}_2 = (\hat{Y}'_{1ab} - \hat{Y}''_{1ab}, \hat{Y}'_{2ab} - \hat{Y}''_{2ab}, \ldots, \hat{Y}'_{Lab} - \hat{Y}''_{Lab}, \hat{N}'_{1ab} - \hat{N}''_{1ab}, \ldots, \hat{N}'_{mab} - \hat{N}''_{mab}) .$$

Then

$$V(\tilde{Y}_{mR}) = V(\hat{Y}_B) - \hat{\beta}' \mathrm{Cov}(\hat{Y}_B, \hat{X}_2) \qquad (2.7)$$

where

$$\hat{\beta} = V^{-1} \mathrm{Cov}(\hat{Y}_B, \hat{X}_2),$$

$$\mathrm{Cov}(\hat{Y}_B, \hat{X}_2) = (\mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{1ab}), \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{2ab}), \ldots, \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{Lab}), \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{1ab}), \ldots, \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{mab}))' ,$$

and V is the covariance matrix of $\hat{X}$.

3. Application of Two-Frame Estimators to California Fruit Data

3.1. Description of the frames

Some data on fruit collected by USDA in California in 1972 are used to illustrate the relative efficiency of the Fuller-Burmeister estimator to Hartley's estimator. These data represent a complete listing of acreages of certain fruits organized on an area basis. The basic unit is an area segment. The area segments are grouped into clusters to form an area frame of 187 area clusters. Some of the clusters contain no acreage in fruit.

A "list frame" of area segments was constructed using the list of segments. This list was constructed to simulate the type of list that might be constructed using producer lists. Such lists traditionally contain a larger fraction of the large operators. Therefore the list frame contained 95% of the segments with area over 500 acres devoted to fruits, 60% of the segments having fruit acreage greater than or equal to 100 acres but less than 500 acres, and 28% of the segments having some fruit acreage but less than 100 acres. The list frame created in this manner contained a total of 310 segments, representing 50% of the non-zero area segments.

Two characteristics, the number of acres under fruit and the number of fruit trees (in hundreds), are studied.

3.2. Simple Random Sampling From List Frame

In the first study, we assume selection of simple random samples of segments from the list frame (frame A) and of clusters from the area frame (frame B). Variances of the estimated totals of the two characteristics for various sample sizes were computed both with and without the finite population correction (fpc) for both frames. The variances were computed using the optimal values of p for Hartley's estimator and optimal values of b_1 and b_2 for the Fuller-Burmeister estimator.

The percentage gain in efficiency of the Fuller-Burmeister estimator, $\tilde{Y}_r$, relative to the Hartley estimator, $\hat{Y}_H$, is defined by $100[V(\hat{Y}_H) - V(\tilde{Y}_r)]/V(\tilde{Y}_r)$. The results for selected sample sizes with fpc, are given in Table 1. Substantial gains are evident for most sample combinations. The gain increases as the fraction of the sample selected from the area frame increases.

The procedure used in the 1949 'Sample Survey of Retail Stores' consisted of observing only that portion of the area frame that fell in the nonoverlap domain. If a screening process is applied and the data on that portion of the area frame sample elements belonging to the overlap domain not collected, then the Hartley estimator reduces to

$$\hat{Y}_c = \hat{Y}'_{ab} + \hat{Y}_b \quad . \tag{3.1}$$

The Fuller-Burmeister estimator for this particular situation is

$$\hat{Y}_{cr} = \hat{Y}'_{ab} + \hat{Y}_b + \hat{\beta}_c(N_{ab} - \hat{N}''_{ab}) \quad . \tag{3.2}$$

The gains in efficiency from using $\hat{Y}_{cr}$, rather than $\hat{Y}_c$, for the set of sample sizes given in Table 1 were computed. The largest gain was 26% associated with a list sample size of 60 and area sample size of 10. For a fixed sample size selected from the list frame, the gain decreases as the size of the sample selected from the area frame increases. This is also apparent from the efficiency gain formula,

$$\frac{V(\hat{Y}_c) - V(\hat{Y}_{cr})}{V(\hat{Y}_{cr})} = \frac{[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2}{V(\hat{N}''_{ab})} \cdot \left[V(\hat{Y}'_{ab}) + V(\hat{Y}_b) - \frac{[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2}{V(\hat{N}''_{ab})} \right]^{-1} \tag{3.3}$$

Since $[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2 \; [V(\hat{N}''_{ab})]^{-1}$ and $V(\hat{Y}_b) - [\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2 \; [V(\hat{N}''_{ab})]^{-1}$ are multiples of n_B^{-1}, the ratio must decrease as n_B increases.

3.3. Stratified Sampling From the List Frame

To investigate efficiencies for stratified sampling of the list frame, we divided the list frame into three strata on the basis of our original construction of the frame. The three strata were sampled in the ratio 4:2:1.

Three forms of Fuller-Burmeister estimators, $\tilde{Y}_{mR}$, were considered. They are

$$\tilde{Y}_{1R} = \hat{Y}_B + b_{11}(\hat{N}'_{ab} - \hat{N}''_{ab}) + b_{12}(\hat{Y}'_{ab} - \hat{Y}''_{ab}) \tag{3.4}$$

$$\tilde{Y}_{2R} = \hat{Y}_B + b_{21}(\hat{N}'_{ab} - \hat{N}''_{ab}) + b_{22}(\hat{Y}'_{1ab} - \hat{Y}''_{1ab}) + b_{23}(\hat{Y}'_{2ab} - \hat{Y}''_{2ab}) + b_{24}(\hat{Y}'_{3ab} - \hat{Y}''_{3ab}) \tag{3.5}$$

$$\tilde{Y}_{3R} = \hat{Y}_B + b_{31}(\hat{N}'_{1ab} - \hat{N}''_{1ab}) + b_{32}(\hat{N}'_{2ab} - \hat{N}''_{2ab}) + b_{33}(\hat{N}'_{3ab} - \hat{N}''_{3ab}) + b_{34}(\hat{Y}'_{1ab} - \hat{Y}''_{1ab}) + b_{35}(\hat{Y}'_{2ab} - \hat{Y}''_{2ab}) + b_{36}(\hat{Y}'_{3ab} - \hat{Y}''_{3ab}) \tag{3.6}$$

where $\hat{Y}_B$, $\hat{Y}'_{iab}$, $\hat{Y}''_{iab}$, $\hat{N}'_{ab}$, and $\hat{N}''_{ab}$ are previously defined, while $\hat{N}'_{iab}$ and $\hat{N}''_{iab}$ are the estimators of the number of elements in domain ab of the ith stratum obtained from the sample of frame A and frame B respectively.

The optimal p's of the Hartley estimator and the optimal b's of Fuller-Burmeister estimators for various sample sizes and the associated variances of the estimators, $V(\hat{Y}_H)$, $V(\hat{Y}_{1R})$, $V(\hat{Y}_{2R})$, and $V(\hat{Y}_{3R})$ were computed retaining the finite population correction. The gains in efficiency from using $\tilde{Y}_{1R}$, $\tilde{Y}_{2R}$, and $\tilde{Y}_{3R}$ relative to Hartley's estimator, $\hat{Y}_H$, are shown in Tables 2-4.

The gains from including additional estimators of zero in the estimator for the total are substantial. As before the gain increases as the area sample size increases.

A summary of the efficiency of $\tilde{Y}_r$ in simple random sampling, and $\tilde{Y}_{3R}$ in stratified sampling, relative to the Hartley estimator is presented in Table 5.

3.4. Optimum Allocation

For any given cost structure, we can obtain the gain in efficiency under optimum allocation among the two frames for each estimator. We now assume the cost for each unit in the area sample is six times as great as that for a unit in the list sample. We study optimal allocation only for the data of acreage in fruit. In simple random sampling, ignoring the finite population correction terms, the optimum allocation for the Hartley estimator is specified by the ratio $n_A/n_B = 4.34$. For the Fuller-Burmeister estimator the optimal ratio is $n_A/n_B = 3.12$. The gain in efficiency of the Fuller-Burmeister procedure relative to the Hartley procedure given optimum allocation for each procedure is 13.64%.

We now investigate the behavior of these estimators under the optimum allocation among the strata. We assume the cost of a unit in one stratum is the same as that of a unit in other strata. Using the iteration procedure, we found that, for $\hat{Y}_H$, the optimum stratum allocation is 49:45:6 and the optimum frame sample ratio is $n_A/n_B = 2.18$, while, for $\tilde{Y}_{3R}$, the optimum stratum allocation is 62:37:1 and the optimum frame sample ratio is $n_A/n_B = 0.79$. Under these best conditions for each estimator, the gain in efficiency from using $\tilde{Y}_{3R}$ relative to $\hat{Y}_H$ is 19.26%.

By comparing the gains in efficiency under the best conditions for each estimator with the data in Table 4, we can see that the relative efficiency of the Hartley estimator is slightly better under optimum sample allocation than under nonoptimum allocation. That is, as we improve the efficiency with which we select the sample, the potential for reduction in variance associated with the inclusion of estimators of zero is reduced.

4. Summary

The variances of alternative multiple-frame estimators are compared using data collected in a census of fruit trees in California in 1972.

In one comparison, we assumed the selection of a simple random sample of individual segments from the list frame and of clusters of segments from the area frame. The gain in efficiency of the Fuller-Burmeister estimator relative to the Hartley estimator was a function of the relative rates at which the two frames were sampled. The gain in efficiency increases as the sampling rate in the area frame increases. In a second comparison the optimum sampling procedure for a fixed budget was used for each estimator under reasonable cost assumptions, the gain of the Fuller-Burmeister estimator relative to the Hartley estimator is about fourteen percent.

The efficiency of the Fuller-Burmeister estimators were also investigated for stratified sampling. When stratified sampling is used, there are a number of estimators of zero that can be used in the regression estimator. The regression estimators displayed considerable gains in efficiency when several estimators of zero were used. As in simple random sampling, the gain in efficiency from using the Fuller-Burmeister estimators is largest for samples where the ratio of the size of the list sample to the size of the area sample size is small. When the optimum sample allocation is used for each estimator, the gain is about nineteen percent.

REFERENCES

[1] Bershad, M. A., "The Sample of Retail Stores," in Hansen, Hurwitz, and Madow, Sample Survey Methods and Theory, Vol. I. Wiley (1953), 516-558.

[2] Cochran, R. S., "Multiple Frame Sample Surveys." Proceedings of the Social Statistics Section of the American Statistical Association (1964), 16-19.

[3] ________, "The Estimation of Domain Sizes When Sampling Frames are Interlocking." Proceedings of the Social Statistics Section of the American Statistical Association (1967), 332-335.

[4] Fuller, W. A. and Burmeister, L. F., "Estimators for Samples Selected from Two Overlapping Frames." Research Report for the Bureau of the Census, Iowa State University, Ames, Iowa (1973).

[5] Hartley, H. O., "Multiple Frame Surveys." Proceedings of the Social Statistics Section of the American Statistical Associaticn (1962), 203-206.

[6] Huang, H. T., "The Relative Efficiency of Some Two-Frame Estimators." A report for the Statistical Reporting Service, USDA, Iowa State University, Ames, Iowa (1974).

[7] Lund, R. E., "Estimators in Multiple Frame Surveys." Proceedings of the Social Statistics Section of the American Statistical Association (1968), 282-288.

Table 1. Percentage Gain in Efficiency of the Fuller-Burmeister Estimator ($\tilde{Y}_r$) relative to the Hartley Estimator ($\hat{Y}_H$) for Various Sample Sizes for California Fruit Data.

List frame sample size (n_A)	Area frame sample size (n_B)				
	10	15	20	25	30
Acres in Fruit					
20	25.30	38.67	51.87	64.77	77.33
30	16.18	25.14	34.36	43.71	53.12
40	11.63	18.11	24.96	32.08	39.40
50	8.95	13.87	19.18	24.78	30.63
60	7.21	11.07	15.29	19.81	24.59
No. of trees					
20	7.35	12.69	17.90	22.90	27.69
30	3.75	7.29	10.97	14.69	18.39
40	2.05	4.50	7.22	10.06	12.98
50	1.12	2.87	4.92	7.15	9.48
60	0.60	1.84	3.41	5.17	7.07

Table 2. Percentage Gain in Efficiency of $\tilde{Y}_{1R}$ Relative to the $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	6.07	9.41	13.10	17.08	21.32
16	8	4	4.51	6.82	9.41	12.25	15.33
20	10	5	3.62	5.33	7.27	9.41	11.76
32	16	8	2.39	3.25	4.24	5.35	6.58
40	20	10	2.01	2.62	3.32	4.10	4.97
No. of trees							
12	6	3	2.86	5.86	9.06	12.35	15.67
16	8	4	1.50	3.55	5.88	8.38	10.96
20	10	5	0.79	2.22	3.98	5.91	7.98
32	16	8	0.07	0.54	1.31	2.29	3.41
40	20	10	0.00	0.17	0.60	1.22	1.98

Table 3. Percentage Gain in Efficiency of $\tilde{Y}_{2R}$ Relative to $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	12.07	20.19	28.80	37.65	46.60
16	8	4	8.30	14.05	20.42	27.19	34.23
20	10	5	6.18	10.45	15.34	20.68	26.35
32	16	8	3.41	5.45	7.98	10.90	14.16
40	20	10	2.66	4.01	5.74	7.81	10.17
No. of trees							
12	6	3	24.87	35.16	43.84	51.20	57.50
16	8	4	19.51	28.48	36.52	43.66	50.02
20	10	5	16.00	23.85	31.21	38.01	44.24
32	16	8	10.45	15.93	21.58	27.23	32.78
40	20	10	8.61	13.06	17.87	22.86	27.93

Table 4. Percentage Gain in Efficiency of $\tilde{Y}_{3R}$ Relative to the Hartley Estimator $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	15.07	26.50	39.08	52.48	66.57
16	8	4	9.89	17.74	26.68	36.43	46.86
20	10	5	7.07	12.75	19.44	26.88	34.96
32	16	8	3.56	6.11	9.38	13.23	17.58
40	20	10	2.70	4.30	6.48	9.14	12.22
No. of trees							
12	6	3	32.33	41.26	48.81	55.46	61.50
16	8	4	28.04	36.14	43.11	49.27	54.87
20	10	5	25.19	32.68	39.27	45.15	50.51
32	16	8	20.49	26.78	32.71	38.25	43.44
40	20	10	18.85	24.60	30.28	35.76	41.02

Table 5. Efficiency of Fuller-Burmeister Estimator Relative to the Hartley Estimator.

	Acres in fruit		No. of trees	
n_A/n_B	Simple random	Stratified	Simple random	Stratified
6.0	107	103	101	120
5.0	109	104	101	121
4.0	111	106	102	124
3.0	116	109	104	128
2.0	125	115	107	132
1.3	139	129	113	141
1.0	152	139	118	149
0.8	165	152	123	155
0.7	177	167	128	162