MULTIPLE FRAME SAMPLE SURVEYS

Robert S. Cochran, University of Wyoming

I. INTRODUCTION

In the theory and practice of sampling finite populations two concepts are carefully distinguished in the literature. They are

(1) the population of units, and

(2) the frame for sampling the units.

Concept (1) does not involve much more than a clear definition of the units deemed to belong to the population. However, concept (2) goes beyond a mere definition of an aggregate of units.

In many situations it is not possible to designate a unique reference frame for some reason. It then becomes necessary to supplement the original frame with an additional frame or frames in order to obtain full coverage of the population, and the investigator must design a survey based upon a multiplicity of sampling frames. In other situations it is possible to designate one frame that will give complete coverage but it may be possible to use another frame to cover a subset of the original frame. Here, again, it may be advantageous to the investigator to consider his problem as being one of multiple frame sampling.

Historically, most cases of multiple frames have been concerned with a "master" frame with 100 coverage and a "cheap" frame not possessing 100 per cent coverage. Usually a sample design had been designated for both frames, but whenever a unit sampled from the "master" frame was encountered which was also in the "cheap" frame it was discarded.

At the 1962 meetings of the American Statistical Association, H. O.Hartley presented a paper entitled "Multiple Frame Surveys." (1962) In his paper he suggested a weighting system whereby it would not be necessary to discard any information obtained in order to arrive at an estimate of the population total or the population mean. In this paper the basis for Hartley's work will be reviewed and a comparison will be presented between using the multiple frame estimator and the screening estimator outlined above.

II. MULTIPLE FRAME TERMINOLOGY AND NOTATION IN THE TWO FRAME CASE

Consider the general situation of two interlocking frames, A and B, where their union gives 100 per cent coverage to the populationof interest. In this situation there are three subsets consisting of those units that are only in A, (domain a), those units that are only in B (domain b), and those units in both frames (domain ab).

Also assume a simple random sample is to be drawn from each frame and that the number of population units in the three domains are known. The notation introduced by Hartley will be used for this discussion. It is presented in Table I.

III. ESTIMATION OF THE POPULATION TOTAL

Hartley's method of weight coefficients actually creates non-overlapping strata out of the overlapping frames. Letting y_i be the value of the Y characteristic attached to the i^{th} sampling unit define

$$u_{Ai} = y_i, \quad \text{if the } i^{th} \text{ unit is in } \underline{a}$$
$$= py_i, \quad \text{if the } i^{th} \text{ unit is in } \underline{ab}$$

when sampling from the A frame. Also define

$$u_{Bi} = qy_i, \quad \text{if the } i^{th} \text{ unit is in } \underline{ab}$$
$$= y_i, \quad \text{if the } i^{th} \text{ unit is in } \underline{b}$$

when sampling from the B frame. Thus, the number of units in the intersection, ab, have been artificially doubled, creating two mutually exclusive strata with U characteristic attached to the units of each strata. Clearly the population total of the original Y characteristic is equal to the population total of the newly constructed U characteristic whenever $p + q = 1$.

The estimate of the population total of the Y characteristic is

$$\hat{Y} = N_a\bar{y}_a + N_{ab}p\bar{y}'_{ab} + N_{ab}q\bar{y}''_{ab} + N_b\bar{y}_b \qquad (1)$$

This estimator is in the form of a post-stratified sampling estimator for the U characteristic. Because of this the variance of $\hat{Y}$ when the usual finite population correction can be ignored and the sample is sufficiently large from each frame is approximately

$$V(\hat{Y}) = \frac{N_A}{n_A}\left\{\sigma_a^2 N_a + \sigma_{ab}^2 N_{ab} p^2\right\} + \frac{N_B}{n_B}\left\{\sigma_b^2 N_b + \sigma_{ab}^2 N_{ab}\, q^2\right\} \qquad (2)$$

where σ_a^2, σ_b^2, and σ_{ab}^2 are the wihin post stratum variances.

After some algebra this becomes

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\left\{\sigma_a^2(1-\alpha) + \alpha\, p^2\, \sigma_{ab}^2\right\} \frac{N_B^2}{n_B}\left\{\sigma_b^2(1-\beta) + \beta\, q^2\, \sigma_{ab}^2\right\}$$

where $\alpha = N_{ab}/N_A$, and $\beta = N_{ab}/N_B$. (3)

Assuming a simple cost function

$$C = c_A n_A + c_B n_B \qquad (4)$$

TABLE I
HARTLEY'S NOTATION FOR TWO FRAME DESIGNS AND ESTIMATES

Category	Frame		Domain			
	A	B	a	b	ab	
Population number	N_A	N_B	N_a	N_b	N_{ab}	
Sample number*	n_A	n_B	n_a	n_b	n'_{ab},	n''_{ab}
Population total	Y_A	Y_B	Y_a	Y_b	Y_{ab}	
Population mean	$\bar{Y}_A$	$\bar{Y}_B$	$\bar{Y}_a$	$\bar{Y}_b$	$\bar{Y}_{ab}$	
Sample total*	y_A	y_B	y_a	y_b	y'_{ab}	y''_{ab}
Sample mean*	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_a$	$\bar{y}_b$	$\bar{y}'_{ab}$	$\bar{y}''_{ab}$
Cost of sampling unit*	c_A	c_B				

*Applies to case of drawing random samples from both frames

where C is the total cost of sampling, c_A is the cost of obtaining an observation from the A frame and c_B is the cost of obtaining an observation from the B frame, it is possible to contemplate finding the values of n_A, n_B, and p that will give a minimum value for the variance whenever the cost is fixed or vice versa.

After some labor, the optimum value of p is found to be one of the solutions of

$$p^2\rho\ [\phi_B(1-\beta) + \beta q^2] = q^2\ [\phi_A(1-\alpha) + \alpha p^2] \tag{5}$$

where $\rho = c_A/c_B$, $\phi_B = \sigma_b^2/\sigma_{ab}^2$, and $\phi_A = \sigma_a^2/\sigma_{ab}^2$. Once the value of p has been determined from the above the values of n_A and n_B can be found from

$$\frac{n_A}{N_A} = \theta \left\{(\sigma_a^2 (1-\alpha) + \alpha p^2 \sigma_{ab}^2) / c_A\right\}^{1/2}$$
$$\frac{n_B}{N_B} = \theta \left\{\frac{(\sigma_b^2 (1-\beta) + \beta q^2 \sigma_{ab}^2)}{c_B}\right\}^{1/2} \tag{6}$$

where θ would be determined by the budget available.

IV. ONE FRAME IS 100 PER CENT

An important special case of the above exists whenever one of the frames, say A, gives 100 per cent coverage to the population of interest and the other frame covers only a subset of the population. An example of such a situation is a survey of households in a city which would be sampled by the block sampling plan while it may be possible to sample most of the population by using the telephone directory. However, there are households that do not have a telephone and can only be sampled by more expensive personal interviews.

When using simple random sampling from both frames we have a special case $\beta = 1$ and $\sigma_B^2 = \sigma_{ab}^2$. With this simplification the estimator becomes

$$\hat{Y} = N_a \bar{y}_a + p\, N_{ab}\, \bar{y}'_{ab} + q\, N_{ab}\, \bar{y}''_{ab} \tag{7}$$

and the variance becomes

$$V(\bar{Y}) = \frac{N_A^2}{n_A} [\sigma_a^2(1-\alpha) + \alpha p^2 \sigma_{ab}^2] + \frac{N_B^2}{n_B} \sigma_{ab}^2 q^2. \tag{8}$$

Using the same cost equation as above the optimum p now becomes

$$p = \sqrt{\frac{\phi_A (1-\alpha)}{-\alpha}} \tag{9}$$

An alternate sampling plan for such a situation is the one mentioned in the introduction as the plan that has been usually used when such situations have developed historically. This alternate is actually a special case of the above with p = 0 and q = 1. Therefore, the estimator is

$$\hat{Y}_0 = \frac{N_a}{n_a}\ y_a + \frac{N_B}{n_B}\ y_B \tag{10}$$

and the approximate variance is

$$V(\hat{Y}_0) = \frac{N_A^2}{n_A} (1-\alpha)\ \sigma_a^2 + \frac{N_B^2}{n_B} \sigma_{ab}^2 = \frac{N_A^2}{n_A}\ \sigma_{ab}^2\ [(1-\alpha)\ \phi_A + \frac{n_A}{n_B} \alpha^2] \tag{11}$$

For this procedure the cost equation would be

$$C = n_a\, c_a + n'_{ab}\ c'_A + n_B\, c_B \tag{12}$$

because n_a and n'_{ab} are random we have

$$E(C) = (1-\alpha)\, c_A + \alpha c'_A\ n_A + c_B\, n_B = c_A^*\, n_A + c_B\, n_B$$

and

c_A = the cost of interviewing sampled units from the 100 per cent frame
c_B = the cost of interviewing sampled units from the list
c'_A = the cost of determining that units sampled from the 100 per cent frame are also on the list (screening cost)

Using this cost equation the optimum values of n_A and n_B yield

$$\frac{n_A}{n_B} = \frac{1}{\alpha} \sqrt{\frac{\phi_A (1-\alpha)}{\rho^*}} \tag{13}$$

where $\rho^* = \frac{c_A^*}{c_B}$

Using this expression for n_A/n_B in $V(\hat{Y}_0)$ yields

$$V(\hat{Y}_0) = \frac{N_A^2}{n_A}\sigma_{ab}^2\left[(1-\alpha)\ \phi_A \quad \alpha\sqrt{\frac{\phi_A\ (1-\alpha)}{\rho^*}}\right] \tag{14}$$

Likewise in C it yields

$$C = n_A c_A^*\left(1 + \frac{\alpha}{\sqrt{\rho^*}\sqrt{\phi_A\ (1-\alpha)}}\right) \tag{15}$$

Thus the variance of the screening estimator is

$$V_0 = \frac{N_A^2\ \sigma_{ab}^2}{C}\quad c_A^*\left(1 + \frac{\alpha}{\sqrt{\rho^*}\sqrt{\phi_A\ (1-\alpha)}}\right)^2\phi_A\ (1-\alpha) \tag{16}$$

On page 4 equation (3), the corresponding multiple frame estimator is given to be

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\left[(1-\alpha)\ \sigma_a^2 + \alpha\ p^2\ \sigma_{ab}^2\right] + \frac{N_B^2}{n_B}\ q^2\ \sigma_{ab}^2. \tag{17}$$

The optimum values for p and q given values of n_A and n_B lead to

$$\frac{p}{q} = \frac{n_A}{n_B}\quad\frac{N_B}{N_A} = \frac{n_A}{n_B}\ \alpha. \tag{18}$$

Substitution of the above into $V(\hat{Y})$ leads to

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\ \sigma_{ab}^2\quad\left[(1-\alpha)\ \phi_A + \alpha\ p\right] \tag{19}$$

Also substituting p/q above into the cost equation on page 4, equation (4), leads to

$$C = c_A n_A\ \left(1 + \frac{q}{\rho}\frac{\alpha}{p}\right)^2 \tag{20}$$

Therefore, the variance of the multiple frame estimator is

$$V(\hat{Y}) = \frac{N_A^2\ \sigma_{ab}^2}{C}\quad c_A\ p^2\rho\quad\left(1+\frac{q}{p}\frac{\alpha}{\rho}\right)^2 \tag{21}$$

When the total budget C is the same for the two types of investigations the ratio of $V(\hat{Y})$ to $V(\hat{Y}_0)$ will give an indication of the relative efficiency of the screening estimator as compared to the multiple frame estimator.

After some algebra this ratio becomes

$$\frac{V(\hat{Y})}{V(\hat{Y}_0)} = \frac{V_P}{V_0} = \frac{c_A}{c_A^*}\ \frac{\rho}{\rho-\alpha}\ \frac{\left(1 + \frac{q}{p}\frac{\alpha}{\rho}\right)^2}{\left(1 + \frac{\alpha}{p\sqrt{\rho^*}\sqrt{\rho-\alpha}}\right)^2}$$

Letting

$$w = \frac{c_A}{c_A^*}$$

ρ^* becomes

$$\rho^* = \frac{\rho}{w}$$

and the variance becomes

$$\frac{V_p}{V_0} = w\ \frac{\rho}{\rho-\alpha}\ \frac{\left(1 + \frac{q}{p}\frac{\alpha}{\rho}\right)^2}{\left(1 + \frac{\alpha\sqrt{w}}{p\sqrt{\rho}\sqrt{\rho-\alpha}}\right)^2} \tag{24}$$

The screening estimator will have the lower variance whenever this ratio is greater than 1. In order to determine parametric conditions that will result in such a situation let

$$A = \frac{\rho}{\rho-\alpha}$$
$$B = \left(1 + \frac{q}{p}\frac{\alpha}{\rho}\right)^2 \tag{25}$$
$$C^{-1} = \frac{\alpha}{p\sqrt{\rho}\sqrt{\rho-\alpha}}$$

With these definitions $V/V_0 > 1$ becomes

$$\frac{w\ A\ B}{\left(1 + \sqrt{w}\right)^2} > 1 \tag{26}$$

From this it can be shown that

$$\sqrt{w} > \left(\sqrt{AB} - C^{-1}\right)^{-1} \tag{27}$$

and because ρ is usually greater than α it can be shown that

$$\sqrt{AB} > C^{-1}. \tag{28}$$

Therefore,

$$w > \left(\sqrt{AB} - C^{-2}\right) \tag{29}$$

which yields

$$w > \left[\left(1 + \frac{q}{p}\frac{\alpha}{\rho}\right)\sqrt{\frac{\rho}{\rho-\alpha}} - \frac{\alpha}{p\sqrt{\rho}\sqrt{\rho-\alpha}}\right]^{-2} = \frac{\rho}{\rho-\alpha} > 1 \tag{30}$$

Thus w will be greater than $\frac{\rho}{\rho-\alpha}$ whenever V/V_0 is greater than 1. It can also be shown that w greater than $\frac{\rho}{\rho-\alpha}$ implies that V/V_0 will be greater than 1. (31)

In terms of the cost conditions in its definition

$$w = \frac{c_A}{c_A^*} = \frac{c_A}{(1-\alpha)\ c_A + c_A'\ \alpha} = \frac{1}{(1-\alpha) + w'\alpha}$$

where

$$w' = \frac{c'_A}{c_A} .$$

Since the advantage is to the screening estimator whenever $w > \frac{\rho}{\rho - \alpha}$ using the definition of w in (31) we find

$$\frac{1}{(1-\alpha) + w'\alpha} > \frac{\rho}{\rho - \alpha} \qquad (32)$$

or

$$\frac{\rho - \alpha}{\rho} > 1 - \alpha + \alpha w'$$

and

$$1 - \frac{1}{\rho} > w'. \qquad (33)$$

Using the definitions of ρ and w' this can be written as

$$1 - \frac{c_B}{c_A} > \frac{c'_A}{c_A} \quad \text{or} \quad c_B < c_A - c'_A \qquad (34)$$

This indicates that on the average the screening estimator will have the lower variance whenever the cost of sampling from the supplementary frame is less than the difference between sampling from the 100 per cent frame and screening members of the 100 per cent frame in the supplementary frame. For example, if it does cost less to ask informative questions of a person on the telephone than it does to ask them face to face, the screening estimator will have the lower variance.

Some illustrations of the relationship between the variance ratio V/V_0 and the screening cost ratio w' under various parametric conditions are given in the following tables and graphs. These are set up for a low (.25) and a high (.75) values of $\phi_A = \sigma_a^2/\sigma_{ab}^2$, for a low (.20), a medium (.50), and a high (.90) values of $\alpha = N_{ab}/N_A = N_B/N_A$, and for a low (1), a medium (2), and high (10) values of $\rho = c_A/c_B$. The range presented for $w' = c'_A/c_A$ is from 0.0 to 1.0 in increments of 1.

TABLE II

$\phi_A = .75$

	$\alpha = .50$			$\alpha = .20$			$\alpha = .90$		
ρ	1	2	10	1	2	10	1	2	10
w' \ p	.87	.50	.20	.87	.58	.25	.87	.26	.09
.0	1.00	1.26	1.63	1.00	1.10	1.20	1.00	1.35	2.12
.1	.95	1.20	1.52	.99	1.08	1.18	.99	1.24	1.82
.2	.92	1.14	1.43	.97	1.06	1.15	.89	1.16	1.58
.3	.88	1.09	1.34	.95	1.04	1.12	.86	1.10	1.44
.4	.84	1.04	1.27	.93	1.02	1.11	.83	1.05	1.32
.5	.82	1.00	1.20	.92	1.00	1.08	.80	1.00	1.23
.6	.79	.96	1.14	.90	.98	1.06	.78	.96	1.15
.7	.76	.92	1.09	.88	.96	1.04	.76	.93	1.09
.8	.74	.90	1.04	.87	.94	1.02	.74	.91	1.03
.9	.72	.87	1.00	.86	.93	1.00	.72	.88	.98
1.0	.70	.84	.91	.84	.91	.98	.71	.85	.93

TABLE III

$\phi = .25$

	$\alpha = .20$			$\alpha = .50$			$\alpha = .90$		
ρ	1	2	10	1	2	10	1	2	10
w' \ p	.5	.33	.14	.50	.29	.11	.50	.15	.05
.0	1.00	1.09	1.19	1.00	1.20	1.52	1.00	1.20	1.67
.1	.99	1.07	1.17	.96	1.14	1.44	.96	1.14	1.50
.2	.97	1.05	1.14	.94	1.11	1.36	.93	1.09	1.38
.3	.96	1.03	1.12	.91	1.07	1.29	.91	1.06	1.30
.4	.94	1.02	1.10	.88	1.03	1.23	.89	1.02	1.23
.5	.93	1.00	1.07	.86	1.00	1.18	.87	1.00	1.17
.6	.92	.99	1.06	.83	.96	1.13	.85	.97	1.12
.7	.90	.96	1.04	.81	.94	1.08	.84	.95	1.08
.8	.89	.95	1.02	.81	.92	1.05	.83	.94	1.04
.9	.88	.94	1.00	.79	.91	1.00	.81	.91	1.00
1.0	.87	.92	.98	.77	.86	.97	.80	.90	.97