INTEGRATED MULTIPLE FRAME SAMPLE SURVEYS

Robert Vogel and Donna Brogan, Emory University

Introduction

The topic of overlapping or multiple frames in sample surveys has been investigated by Hartley (1962, 1974), R. Cochran (1967), Fuller and Burmeister (1972), Lund (1968) and Ali (1967). In these investigations the objective was to make inference to the union of two or more overlapping frames under the assumption that one frame by itself does not cover the entire inference population of interest. Different estimators were proposed under various assumptions whether domain and frame sizes are known or unknown.

This research utilizes the previous work to address a somewhat different problem, that of making inference to each of two different populations when the two sampling frames overlap. One sampling plan is to conduct two separate sample surveys, one for each sampling frame, ignoring the fact that the intersection of the two frames is nonempty. However, by recognizing the overlap new sampling plans and corresponding estimators can be developed such that the same or better precision is obtained for the two point estimates of interest at a cost not exceeding that of conducting two separate sample surveys. This procedure is called integrated multiple frame sampling, since the two separate sample surveys are integrated into one survey.

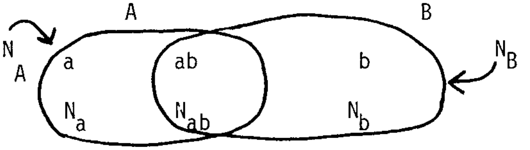
Example

The federal government mandates periodic quality control checks to determine what percentage of AFDC recipients and what percentage of Food Stamp (FS) recipients have an error (overpayment) in their monthly benefit payment. Originally each agency conducted its own sample survey using its own frame of benefit recipients. When a recipient was selected into either sample, an extensive investigation was undertaken of the recipient's income sources, with the information collected by AFDC similar to that collected by FNS (Food Nutrition Service). Preliminary data from Georgia indicated that it took 6.2 hours to collect data for a FS recipient, 7.3 hours for an AFDC recipient, and 9.4 hours to collect both AFDC and FS information for a recipient receiving both benefits. Thus, if a recipient benefits from both programs, a combined AFDC/FNS data collection effort could be done in less time than two separate investigations on the same person. This reduces respondent burden as well as reducing cost of the surveys and/or increasing precision. In Georgia in 1981 there were about 230,000 FS recipients and about 90,000 AFDC recipients, with the intersection of these two frames containing about 69,000 recipients. Thus, 77% of the AFDC recipients also received Food Stamps and 30% of the FS recipients also received AFDC. With such a substantial overlap, it is reasonable to expect that an integrated survey would yield greater precision and/or reduced cost.

Assumptions

Let A and B denote two sampling frames with a nonempty intersection as shown in Figure 1.

Figure 1



N_A and N_B are the sizes of frames A and B, respectively. Frames A and B form three nonempty mutually exclusive and exhaustive domains a, ab, and b of sizes N_a, N_{ab} and N_b, respectively. Thus, a contains units that can be measured only on attribute A, b can be measured only on attribute B and ab can be measured on both attributes A and B.

This paper assumes that all frame and domain sizes are known. Further, it is assumed that sampling can be done from each of the three domains a, b, and ab. (Work in process considers other situations where the three domains cannot be stratified and/or where domain sizes are not known.)

Two Separate Surveys

Under the assumptions above, and assuming that the reciprocals of the three stratum sizes are negligible, the best sampling plan for frame A alone is stratified random sampling (with strata a and ab) using Neyman allocation. A similar statement holds for frame B with strata b and ab. Thus, the integrated methods developed in this paper will be compared to two independent stratified random samples based on Neyman allocation with fixed cost functions given by

$$C_A = c_A n_A \text{ and } C_B = c_B n_B \tag{1}$$

where C_A and C_B are the total survey budgets for survey A and B; c_A and c_B are the costs of collecting information on units sampled from frame A and frame B; and n_A and n_B are the total sample sizes for frames A and B.

The point estimate for the mean of attribute A in frame A is

$$\bar{y}_{Ast} = W_a \bar{y}_a + W_{Aab} \bar{y}_{Aab} \tag{2}$$

with variance

$$V(\bar{y}_{Ast}) = W_a^2 \, S_a(1-f_a)/n_a + W_{Aab}^2 \, S_{Aab}^2 \times (1-f_{Aab})/n_{Aab} \tag{3}$$

where

$$W_a = N_a/N_A;\ W_{Aab} = N_{ab}/N_A;\ f_a = n_a/N_a;\ f_{Aab} = n_{Aab}/N_{ab}\ ;\ \bar{y}_a = \sum_{i=1}^{n_a} y_i/n_a \text{ for } y_i \varepsilon a; \text{ and}$$

$$\bar{y}_{Aab} = \sum_{i=1}^{n_{Aab}} y_i/n_{Aab} \text{ for } y_i \ \varepsilon ab \tag{4}$$

Further, S_a^2 is the stratum variance for domain a and S_{Aab}^2 is the stratum variance with respect to attribute A in domain ab. The point estimate $\bar{y}_{Bst}$ and its variance $V(\bar{y}_{Bst})$ are defined analogously.

Integrated Survey

As an alternative to selecting two independent stratified samples based on Neyman allocation, consider the following algorithm for an integrated survey.

Step 1: Determine n_a and n_b as if two independent stratified samples with Neyman allocation are to be taken using the following Neyman formula:

$$n_a = C_A N_a S_a / c_A(N_a S_a + N_{ab} S_{Aab}) \qquad (5)$$

$$n_b = C_B N_b S_b / c_B(N_b S_b + N_{ab} S_{Bab}) \qquad (6)$$

Step 2: Define

$$m_{Aab} = (C_A - c_A n_a)/c_{Aab} \qquad (7)$$

and

$$m_{Bab} = (C_B - c_B n_b)/c_{Bab} \qquad (8)$$

where $c_{Aab} = c_A(1+k_A),\ k_A \geq 0$ (9)

and $c_{Bab} = c_B(1-k_b),\ k_B \geq 0$ (10)

In equations (9) and (10), c_{Aab} is the cost of selecting a unit from stratum ab when sampling with respect to frame A as if two independent samples are to be drawn and then measuring the variable of interest from both frame A and frame B. k_A is the proportionate increase over c_A in collecting information about units from both frames. c_{Bab} and k_B are defined in an analogous manner. Most likely c_{Aab} will equal c_{Bab} even though k_A and k_B are unequal; however, equality of c_{Aab} and c_{Bab} is not assumed here.

Step 3: Select a sample of size n_a from stratum a, a sample of size n_b from stratum b, and a sample of size $(m_{Aab}+m_{Bab})$ from stratum ab, remembering to measure the variables of interest from both frame A and frame B.

Note that this scheme uses the total budget of the two separate surveys in the definition of the sample sizes for the overlap domain ab in (7) and (8). Using this algorithm to determine sample sizes, the point estimate for the population mean for the attribute of interest from frame A is given by:

$$\bar{y}_{IAst} = W_a \bar{y}_a + W_{Aab} \bar{y}_{IAab} \qquad (11)$$

with variance

$$V(\bar{y}_{IAst}) = W_a^2 S_a^2(1-fa)/n_a + W_{Aab}^2 S_{Aab}^2 \times (1-f_{IAab})/(m_{Aab}+m_{Bab}) \qquad (12)$$

The point estimate $\bar{y}_{IBst}$ and its variance $V(\bar{y}_{IBst})$ are defined analogously. $\bar{y}_{IAab}$ and $\bar{y}_{IBab}$ are the sample means from stratum ab for the attribute of interest from frame A and frame B, respectively, each based on $(m_{Aab} + m_{Bab})$ units.

Theorem 1: a.) $V(\bar{y}_{IAst}) \leq V(\bar{y}_{Ast})$ iff

$$(m_{Aab} + m_{Bab}) \geq n_{Aab} \qquad (13)$$

b.) $V(\bar{y}_{IBst}) \leq V(\bar{y}_{Bst})$ iff

$$(m_{Aab} + m_{Bab}) \geq n_{Bab} \qquad (14)$$

Proof: Can be shown by comparing the variance formulae to each other.

Corollary 1: a.) $V(\bar{y}_{IAst}) \leq V(\bar{y}_{Ast})$ iff

$$n_{Bab}/(1+k_B) \geq k_A n_{Aab}/(1+k_A) \qquad (15)$$

b.) $V(\bar{y}_{IBst}) \leq V(\bar{y}_{Bst})$ iff

$$n_{Aab}/(1+k_A) \geq k_B n_{Nab}/(1+k_B) \qquad (16)$$

Proof: Substitute (7) and (8) and the calculated values of n_{Aab} and n_{Bab} from Neyman allocation into (13) and (14).

Theorem 2: Let $c_{Aab} = k\, c_{Bab}$ where $k > 0$. If $k_A k_B \leq 1$ then $V(\bar{y}_{IAst}) \leq V(\bar{y}_{Ast})$ and $V(\bar{y}_{IBst}) \leq V(\bar{y}_{Bst})$ iff

$$\frac{k_A c_A}{k c_B} \leq \frac{n_{Bab}}{n_{Aab}} \leq \frac{c_A}{k_B c_B k} \qquad (17)$$

Proof: Use corollary 1.

Thus, theorem 2 can be used to determine if an integrated survey will give greater precision than two separate surveys at the same cost. For example, let A and B be the FS and AFDC frames, respectively. Earlier we noted that N_A=230,000; N_B=90,000; N_a=161,000; N_b=21,000; N_{ab}=69,000; c_A=6.2, c_B=7.3, and $c_{Aab}=c_{Bab}$=9.4 (or k = 1). Each agency's survey cost is fixed by a federal mandate that $n_A = n_B$ = 1200, resulting in C_A=7440 hours and C_B=8760 hours. Past experience indicates that $S_a^2 \doteq 0.166$ and $S_{Aab}^2 \doteq .214$ for frame A and $S_b^2 \doteq .096$ and $S_{Bab}^2 \doteq .104$ for frame B.

Using two independent surveys yields $V(\bar{y}_{Ast})$= .000150 and $V(\bar{y}_{Bst})$ = .000085. Using (9) and (10), k_A = .52 and k_B = .29. Since $k_A k_B \leq 1$, and since 0.44 = $\frac{k_A c_A}{c_B} \leq \frac{n_{Bab}}{n_{Aab}} =$

$$\frac{928}{392} \leq 2.367 \leq \frac{c_A}{k_B c_B} = 2.93,$$

a gain in precision is guaranteed for the integrated survey, yielding

$V(\bar{y}_{IAst}) = .000121$ and $V(\bar{y}_{IBst}) = .000082$. The percentage of relative reduction in variance at equal cost is 19.3% for frame A and 3.5% for frame B.

Modified Integrated Survey

Consider the example above with k_A and k_B now assuming the hypothetical values of 0.74 and 0.48. With k = 1, $c_A/k_Bc_B = 1.76$ and $k_Ac_A/c_B = 0.65$. Hence with $n_{Bab}/n_{Aab} = 2.367$, inequality (17) fails to hold. With these new values of k_A and k_B, m_{Aab} and m_{Bab} can be calculated and $m_{Aab} + m_{Bab} = 853$. Thus, with $n_{Aab} = 392$ and $n_{Bab} = 928$, it is easy to see that when sampling with respect to frame A, the integrated plan yields a sample size substantially larger for stratum ab than that of the stratified sample using Neyman allocation. On the other hand, when sampling with respect to frame B, the overlap stratum is "undersampled" when using the integrated plan as opposed to that of the stratified sample using Neyman allocation. In this event, a modification to the integrated plan may be made in which not all of the $(m_{Aab} + m_{Bab}) = 853$ units from stratum ab are used to measure both variables. Since only 392 units are needed for the variable of interest from frame A, the savings in measuring both variables on 392 units instead of 853 units may allow the sampler enough resources to sample the additional needed units from stratum ab for the variable of interest with respect to frame B.

Thus, consider the following algorithm:

Step 1: Compute all values of n_a, n_b, n_{Aab}, and n_{Bab} as if two independent stratified samples based on Neyman allocation are to be taken.

Step 2: Set $m = \min \left\{ n_{Aab}, n_{Bab} \right\}$ (18)

and define m' as

$m' = \max \left\{ n_{Aab}, m_{Bab} \right\} - m$ (19)

Step 3: Select $(m + m')$ units from stratum ab in which for m of these units both variables are measured and for m' units only measure the variable of interest from the "undersampled" frame. To ensure all units in the ab stratum are selected with equal probability, select $(m + m')$ units with a single random sample and then select a subset at random of size m from these $(m + m')$ units.

With the above modified algorithm, the point estimate for the population mean for the variable of interest from frame A is given by:

$$\bar{y}_{MAst} = \begin{cases} W_a\bar{y}_a + W_{Aab}\,\bar{y}'_{MAab} & \text{if } m = n_{Aab} \\ W_a\bar{y}_a + W_{Aab}\,\bar{y}''_{MA_{ab}} & \text{if } m = n_{Bab} \end{cases} \quad (20)$$

with variance:

$$V(\bar{y}_{MAst}) = \begin{cases} W_a^2S_a^2(1-f_a)/n_a + W_{Aab}^2\,S_{Aab}^2\,(1-f'_{MAab})/m \\ \qquad \text{if } m = n_{Aab} \\ W_a^2S_a^2(1-f_a)/n_a + W_{Aab}^2\,S_{Aab}^2\,(1-f''_{MAab})/ \\ \qquad (m + m')\text{ if } m = n_{Bab} \end{cases} \quad (21)$$

where $\bar{y}'_{MAab}$ is the sample mean for stratum ab computed from the m units in which the variable of interest from frame A was measured and $\bar{y}''_{MAab}$ is the sample mean for stratum ab computed from the $(m + m')$ units in which the variable of interest from frame A was measured. f'_{MAab} is the sampling fraction based on m units and f''_{MBst} is the sampling fraction based on $(m + m')$ units from stratum ab. The estimate $\bar{y}_{MBst}$ and its variance $V(\bar{y}_{MBst})$ are defined analogously.

Theorem 3: The precision of the estimated population means under the modified integrated sampling plan will always be greater than or equal to the precision of two independent stratified samples based on Neyman allocation iff

$$m' \geq n_{Aab} - n_{Bab} \quad (22)$$

Proof: can be shown by comparing variance formulae to each other.

As an example, consider the hypothetical values of .74 and .48 for k_A and k_B with k = 1. Using the previous values for this FS/AFDC example and the modified integrated survey yields m = 392 and $|n_{Aab} - n_{Bab}| = |392 - 928| = 536$. Thus, using m = 392 and $m' = 536$ there is a total reduction of 1608 case worker hours over two independent stratified samples based on Neyman allocation.

Conclusion

Under certain conditions, it is shown that integrating two simultaneous independent sample surveys into a single integrated sample survey can result in greater precision at the same cost or equivalent precision at reduced cost. The sample sizes for the particular integrated schemes are defined in terms of cost and variance parameters of the two separate surveys.

List of References

Ali, M.S. (1967), Multiple Frame Sample Surveys Involving Different Units and Nonresponse, unpublished dissertation, Texas A&M University Library.

Cochran, R. (1967),"The estimation of domain sizes when sampling frames are interlocking," Proceedings of the Social Statistics Section of the American Statistical Association, 332-335.

Fuller, W.A. and Burmeister, L.F. (1972), "Estimators for samples selected from two overlapping frames," Proceedings of the Social Statistics Section of the American Statistical Association, 245-249.

Hartley, H.O. (1962), "Multiple frame surveys," Proceedings of the Social Statistics Section of the American Statistical Association, 203-206.

Hartley, H.O. (1974), "Multiple frame methodology and selected applications," Sankhyā: The Indian Journal of Statistics, Series C, 36:99-118.

Lund, R.E. (1968), "Estimators in multiple frame surveys," Proceedings of the Social Statistics Section of the American Statistical Association, 282-288.

Acknowledgement

This research was partially supported by NHLBI Contract #N01-HV-92912 awarded to Georgia Department of Human Resources and Emory University School of Medicine and by NHLBI grant #HL 24004 awarded to Fort Valley State College and Emory University School of Medicine.