A STUDY OF DUAL FRAME ESTIMATORS FOR THE NATIONAL HEALTH INTERVIEW SURVEY

Robert J. Casady, Cecelia B. Snowden, Monroe G. Sirken
National Center for Health Statistics

INTRODUCTION

In recent years the interview component of cost for personal interview surveys has increased rapidly while total survey dollars have tended to remain fixed or even decline. Hence, if estimator precision is to be maintained, alternative data collection modes and/or survey designs must be developed and implemented.

As telephone interviews tend to be relatively inexpensive, one very attractive alternative is the random digit dialing (RDD) sample design developed by Waksberg (1978). Unfortunately, telephone frame surveys are vulnerable to coverage bias. In a study by Thornberry and Massey (1978) it was reported that in 1977 over 94% of the U.S. population resided in a household with access to a telephone. However, they also found that for certain demographic, socio-economic and geographic subdomains telephone coverage was very poor. For example, only 81 percent of blacks, 70 percent of persons with family income less than $3000 and 86 percent of persons in the south were covered by the telephone frame. Estimates for any of these subdomains are subject to a potentially large bias.

Another alternative is to utilize Hartley's (1962, 1974) dual frame approach to construct a survey design which combines a telephone frame and an area/list household frame. The basic logic for this approach is to

(1) eliminate the potential for coverage bias by using the area household list, as one of the frames, and
(2) lower average interview cost by conducting telephone interviews using the telephone list as the other frame.

The dual frame design based on household and telephone frames has been studied by Lund (1968) and Casady and Sirken (1980). The principal design consideration is the allocation of the sample to the two frames to attain a minimum variance for a fixed (expected) cost. The papers by Lund and by Casady and Sirken analyzed the problem under the assumption of simple random sampling. In the next section of this paper the variance of the dual frame estimator is derived under the assumption of cluster sampling from both frames. In the third section, these results are applied to a possible redesign of the National Health Interview Survey (NHIS).

DUAL FRAME ESTIMATION

Assume we have a self weighting sample of m_1 clusters (or segments) of households from an area/list household frame and an independent SRS of m_2 clusters of telephone households from a telephone number frame. Suppose the sample from the household frame yields n_{11} persons from telephone households and n_{12} persons from non-telphone households and the sample from the telephone frame yields n_2 persons (obviously all from telephone households). The dual frame estimator of the mean level of a characteristic, say characteristic Y, for the population is

$$\bar{y} = \hat{p}_{11}\,\bar{y}_{11} + (1-\hat{p}_{11})\,[a\,\bar{y}_{12} + (1-a)\,\bar{y}_2]$$

where $\hat{p}_{11} = n_{11}/(n_{11} + n_{12})$ = estimator of p_1, the proportion of non-telephone persons in the population.

$\bar{y}_{11}$ = estimator of mean level of y for non-telephone population (based on household frame sample)

$\bar{y}_{12}$ = estimator of mean level of y for telephone population (based on household frame sample)

$\bar{y}_2$ = estimator of mean level of y for telephone population (based on telephone frame sample)

and a = arbitrary real constant $(0 \le a \le 1)$.

Following Lund (1968) the constant a is chosen to minimize

$\text{Var}\,(\bar{y} \mid n_{11}, n_{12}, n_2)$ so that

$$a = (\frac{\delta_2}{n_2})\,(\frac{n_{12}\,n_2}{n_{12}\,\delta_2 + n_2\,\delta_{12}})$$

$$X \left[1-\rho\,\frac{p_1\,\sigma_1}{(1-p_1)\,\sigma_2}\,(\frac{\delta_{11}\,\delta_{12}\,n_2^2}{n_{11}\,n_{12}\,\delta_2^2})^{1/2} \right]$$

where δ_{11} = deff for $\bar{y}_{11}$

δ_{12} = deff for $\bar{y}_{12}$

δ_2 = deff for $\bar{y}_2$

σ_1^2 = population variance of y for non-telephone persons

σ_2^2 = population variance of y for telephone persons

ρ = correlation of $\bar{y}_{11}$ and $\bar{y}_{12}$.

If a is chosen as shown, then

$$\mathrm{Var}\ (\bar{y}) \doteq \frac{p_1\ \sigma_1^2\ \delta_{11}}{m_1\ \bar{N}_1}$$

$$+ \frac{(1-p_1)^2\ \delta_{12}\ \delta_2\ \sigma_2^2\ E\ (\theta)}{m_1\ (1-p_1)\ \bar{N}_1\ \delta_2 + m_2\ \bar{N}_2\ \delta_{12}}$$

$$+ \frac{p_1\ (1-p_1)\ \gamma\ (E(\bar{y}_{11}) - E(\bar{y}_{12}))^2}{m_1\ \bar{N}_1}$$

where $\bar{N}_1$ = average number of persons per cluster from the household frame

$\bar{N}_2$ = average number of persons per cluster from telephone frame

γ = deff for $\hat{p}_{11}$

and $$E(\theta) = 1 + 2\ \rho\ (\frac{p_1}{1-p_1})^{1/2}\ \frac{\sigma_1}{\sigma_2}\ \delta_{12}\ \delta_2$$

$$- \rho^2\ \frac{\sigma_1^2\ \delta_{11}\ m_2\ \bar{N}_2}{(1-p_1)\ \sigma_2^2\ \delta_2\ m_1\ \bar{N}_1}$$

It should be noted that if ρ is negligible (as it is for the household survey considered in the next section) then E $(\theta) \doteq 1$ and the expression for Var $(\bar{y})$ simplifies considerably.

Provided that a realistic cost function depending on m_1 and m_2 can be obtained, the expression for Var $(\bar{y})$ can be utilized in the usual manner to optimally allocate survey resources to data collection from the two frames.

A DUAL FRAME DESIGN FOR NHIS

The sample design assumed for the household frame is essentially the same as for the current NHIS design. This design calls for the selection of 376 PSU's (counties or groups of contiguous counties) and approximately 11,400 s.s.u's which are geographic compact clusters of approximately four households. For the telephone frame a RDD design was assumed with the telephone household itself being the cluster unit. Actually, the use of Waksberg's telephone sampling procedure would probably be more efficient. However, the data from NHIS do not permit the estimator of the intra-bank correlation coefficient. Thus, the RDD design was assumed. Based on 1976 NHIS data the average cluster for the household frame, $\bar{N}_1$, is 9.9 persons and the average cluster size for the telephone frame, $\bar{N}_2$, is 2.9 persons. Using the 1976 NHIS data base, the various population and design parameters specified in the preceding section were estimated for a selected set of eight health characteristics. The parameter ρ was found to be between 0 and .05 for all of the variables and hence was ignored for the purpose of sample allocation.

The $4,000,000 budget for NHIS can be broken into three major categories:

(a) Central office administration and survey maintenance costs
(b) Regional office administration, survey maintenance and supervision costs, and,
(c) Direct and indirect interviewing costs.

For the 1976 NHIS, it was estimated that all of the costs in (a) and approximately half of the costs in (b) could be considered as fixed costs. This amounted to approximately $800,000. The costs included in (c) and the remaining costs in (b) were considered to be variable interviewing costs. Thus the variable interviewing cost for NHIS was estimated to be about $280 per cluster of households or about $80 per household.

To determine the cost model for the dual frame survey, it was estimated that added administrative and survey maintenance costs would increase fixed costs by about 50 percent to $1,200,000, leaving $2,800,000 for interviewing. The cost of a telephone interview was assumed to be only one half as expensive as a personal household interview or about $40 per household. Hence, the cost equation for purposes of sample allocation to the two frames is

$$\$2{,}800{,}000 = \$280\ m_1 + \$40\ m_2.$$

Next, the sample was allocated to the two frames for each of the eight health characteristics so as to minimize the variance of the estimator of the mean level of the health characteristic subject to the above cost constraint. Depending on the health characteristic considered the optimal allocation ranged from

$$\left\{ \begin{array}{l} m_1 = 3{,}300 \text{ clusters (or 11,550 personal interview households)} \\ m_2 = 46{,}600 \text{ telephone interview households} \end{array} \right.$$

to

$$\left\{ \begin{array}{l} m_1 = 5{,}200 \text{ clusters (or 18,200 personal interview households)} \\ m_2 = 33{,}400 \text{ telephone interview households} \end{array} \right.$$

Four design options were selected for study:

Option 1 - m_1 = 3,300 and m_2 = 46,600; one of the two extreme allocations above

Option 2 - $m_1 = 4{,}500$ and $m_2 = 38{,}000$; a compromise between the two extremes above.

Option 3 - $m_1 = 5{,}200$ and $m_2 = 33{,}400$; the other extreme allocation above

Option 4 - $m_1 = 11{,}300$ and $m_2 = 0$; only the household frame is used (i.e., the current NHIS design)

To determine the overall performance for each of the design options the variances of the estimators for each of the eight health characteristics were calculated for the total population and are given in the table below.

The most important fact to be noted in the above table is that variances for the health characteristic estimators using Option 4 are uniformly greater than or equal to the variances for each of the dual frame designs. In fact, except for health variable 4, the variances are strictly larger for the single frame survey. Although these results must be reviewed as tentative, they do seem to indicate that it may be possible to construct a dual frame design that will greatly improve the efficiency of NHIS with respect to most variables of interest.

SUMMARY AND CONCLUSIONS

As a design feature of population surveys, the sampling frame has several options including an area/list household frame and a frame of telephone numbers. The data collection mode of the former is face-to-face interviewing and of the latter it is telephone interviewing. Numerous studies have investigated the design and cost effects of both sampling frame options and by now it is well established that the survey costs are greater for the area/list frame and coverage errors are greater for the telephone frame. Since maximum benefits can not necessarily be obtained by opting for either one or the other type of frame, we are proposing the dual frame which is a combination of both frames, as a third option.

The hallmark of dual frame designs is the allocation of resources among the frames. The statistical theory for optimizing resource allocations was generalized in this paper to accommodate cluster sampling which is somewhat more realistic than the simple random sampling assumptions of earlier work.

The dual frame design was illustrated by application to the National Health Interview Survey (NHIS) which is currently based solely on an area/list household frame. In the decennial redesign of NHIS, serious consideration is being given to adding the telephone frame by shifting some resources away from the area/list frame. The preliminary findings that were presented in this paper are encouraging. They indicate that a redesigned NHIS based on combined frames would be subject to smaller sampling errors than a redesigned NHIS based solely on an area/list frame. Much work remains to be completed, however, before this design issue is resolved. The cost and coverage error effects assumed in these calculations need to be verified and substantiated. And, other nonsampling error effects of sampling frame options, such as bias due to nonresponse, need to be investigated.

Health Characteristic	Allocation Option			
	Option 1	Option 2	Option 3	Option 4
1	$4.9\text{X}10^{-4}$	$5.0\text{X}10^{-4}$	$5.2\text{X}10^{-4}$	$7.4\text{X}10^{-4}$
2	$8.8\text{X}10^{-4}$	$8.5\text{X}10^{-4}$	$8.6\text{X}10^{-4}$	$10.1\text{X}10^{-4}$
3	$.18\text{X}10^{-4}$	$.18\text{X}10^{-4}$	$.18\text{X}10^{-4}$	$.23\text{X}10^{-4}$
4	$.03\text{X}10^{-4}$	$.03\text{X}10^{-4}$	$.03\text{X}10^{-4}$	$.03\text{X}10^{-4}$
5	$8.6\text{X}10^{-4}$	$8.3\text{X}10^{-4}$	$8.3\text{X}10^{-4}$	$9.8\text{X}10^{-4}$
6	$15.2\text{X}10^{-4}$	$13.4\text{X}10^{-4}$	$13.1\text{X}10^{-4}$	$20.1\text{X}10^{-4}$
7	$14.3\text{X}10^{-4}$	$13.4\text{X}10^{-4}$	$13.4\text{X}10^{-4}$	$15.3\text{X}10^{-4}$
8	$.29\text{X}10^{-4}$	$.28\text{X}20^{-4}$	$.27\text{X}10^{-4}$	$.31\text{X}10^{-4}$

REFERENCES

Casady, R.J. and M.G. Sirken, "A Multiplicity Estimator for Multiple Frame Sampling", Proceedings of the American Statistical Association, Social Statistics Section (1980) pp. 601-605.

Hartley, H.O., 'Multiple Frame Surveys", Proceedings of the American Statistical Association, Social Statistics Section (1962) pp. 203-206.

Hartley, H.O., "Multiple Frame Methodology and Selected Application", Sankhya (1974) pp. 99-118.

Lund, R.E., "Estimators in Multiple Frame Surveys", Proceedings of the American Statistical Association, Social Science Section (1968).

Thornberry, O.T. and J.T. Massey, "Correcting the Undercoverage Bias in Random Digit Dialed National Health Surveys", Proceedings of the American Statistical Association, Survey Research Section (1978) pp. 224-229.

Waksberg, J., "Sampling Methods for Random Digit Dialing", Journal of the American Statistical Association, Vol. 73, March, 1978.