NONRESPONDENT SUBSAMPLING SCHEMES FOR LONGITUDINAL SURVEYS

Judith T. Lessler and Robert E. Mason, Research Triangle Institute
Richard A. Rosthal, Killalea Associates, Inc.
George E. Pugh, Decision-Science Applications, Inc.
J. George Caldwell, Vista Research, Inc.

1. INTRODUCTION

In the prototypic nonrespondent followup study the survey is conducted in two phases. An initial attempt is made to obtain measurements or responses from sample members by means of a relatively inexpensive survey procedure. Commonly, one does not obtain responses from every member of the sample as a result of this initial attempt. Because nonresponding individuals may differ in many important characteristics from those who respond, estimates based only on the respondent sample may not yield a correct picture of the population as a whole; i.e., there may be nonresponse bias. In order to eliminate the bias due to nonresponse to the initial survey procedure, a subsample of the nonrespondents is drawn and measurements are obtained for the units in the subsample using a more expensive survey procedure.

Procedures for choosing optimum sizes for the initial sample and the nonrespondent followup subsample have been developed and depend upon the relative cost of the two survey procedures and the differences between the initial respondents and nonrespondents (Hansen and Hurwitz, reference [1]). This paper deals with an extension of these procedures to the case of longitudinal surveys.

2. SPECIAL CONSIDERATIONS FOR SUBSAMPLING NONRESPONDENTS IN A LONGITUDINAL SURVEY

When attempting to extend the traditional approach to longitudinal surveys, special considerations arise. A longitudinal survey has quite different objectives than most typical surveys. Whereas in many surveys the objective is to measure certain characteristics of the target population at a particular point in time, the objective in a longitudinal study is to provide data that can be used to study cause and effect relationships within the target population over a period of time. To make such an analysis possible, it is essential to have, over time, comparative information on the same units. That is, for a longitudinal survey conducted at t=1,2,...,T times, domains of interest are defined by the relation,

$$D = \bigcap_{t=1}^{T} D_t \ .$$

If survey results are missing for one or more of the component domains, then membership in the domain of interest is indeterminate. This requirement is the essential feature of a longitudinal survey and distinguishes it from a one-time survey and from a series of cross-sectional surveys.

Because of this unique characteristic of the longitudinal survey, the standard method of subsampling nonrespondents is not directly applicable to a longitudinal survey. If nonresponding units are simply subsampled at each time, the necessary continuity is not maintained, since units do not remain in fixed respondent-nonrespondent strata.

The problem becomes one of defining a procedure which:

(1) maintains the necessary continuity of information,

(2) provides in the context of a longitudinal survey at least some of the cost-variance benefits of subsampling,

(3) insures that valid unbiased parameter estimates are provided.

3. MODIFICATIONS REQUIRED IN THE TRADITIONAL APPROACH WHICH PERMIT LONGITUDINAL ESTIMATES

The model that was developed for nonresponse in a longitudinal survey which will allow subsampling of nonrespondents is an adaption of double sampling for stratification. In this model a series of "post-strata" are formulated over time with the post stratification variable being the response/nonresponse history of the individuals in the sample.

Post strata consisting of units which are nonrespondents for the first time are defined at each t=1,2,...,T and carried forward. A subsample of the nonresponding units is selected and measurements are obtained using the costly survey procedure. Complete response under the costly procedure is assumed.

Two alternatives exist at this point.

(1) Subsampled nonresponding units are again subsampled and measurements obtained under the costly procedure.

(2) *All* subsampled units which do not respond at later times are followed up with the costly procedure.

If the sampling variance is similar for units contained in the initially defined post strata and the subset of subsequently nonresponding units from the same strata, Method (1) above collapses to Method (2). Also, unless per unit costs exhibit highly inflationary trends over the life of the survey, Method (1) would seem to be suboptimal to Method (2), although this suggestion has not been documented at this writing. Nonetheless, the material presented in the following sections relates to Method (2).

3.1 Estimators and Variances of Estimators

The derivation of estimation and variance formulas is made difficult by the necessity to identify the various subsamples in terms of their response-nonresponse history, and the notation becomes cumbersome.

An initial sample of size n is selected and canvassed using the low cost procedure at time t=0. At any time, t, let N(t) and $N(\bar{t})$ denote the population sizes of respondents and nonrespondents, respectively. Corresponding sample values are n(t) and $n(\bar{t})$. The notation is extended to indicate the response history of units up to any time, t, during the survey.

For example, denoting the responding set by R_t and the nonresponding set by R_t^c, the cardinalities of interest for units

$$u_i \in \{R_1 \cap R_2 \cap R_3^c\}$$

are $N(1,2,\bar{3})$ and $n(1,2,\bar{3})$. Sampling variances and subsample sizes are similarly identified, for example, by $S^2(1,2,\bar{3})$ and

$m(1,2,\tilde{3})$, respectively. Finally, let $X(t)_i$ stand for the measurement obtained for the i-th unit at time, t.

Consider, first, time t=1 for which the usual Hansen-Hurwitz [1] formulas apply. In the above notation, a subsampling fraction 1/K(1) is chosen for the nonresponse subsample, and

$$K(1) = \frac{n(\tilde{1})}{m(\tilde{1})} .$$

The unbiased estimate of the population mean, $\bar{X}(1)$, is given by

$$\bar{x}(1) = \frac{1}{n}\left[\sum_{u_i \in R_1} X(1)_i + \frac{n(\tilde{1})}{m(\tilde{1})} \sum_{u_i \in R_1^c} X(1)_i\right].$$

The variance of $\bar{x}(1)$ is given by

$$\mathrm{Var}\left\{\bar{x}(1)\right\} = \left[\frac{1}{n}\ \frac{N-n}{N} S^2 + (K(1)-1)\frac{N(\tilde{1})}{N} S^2(\tilde{1})\right],$$

where S^2 is the population variance.

Now consider time t=2. Let,

$$K(2) = \frac{n(1,\tilde{2})}{m(1,\tilde{2})} .$$

Under Method (2), the subsampling fraction of t=1 nonrespondents is unity. Hence,

$$\bar{x}(2) = \frac{1}{n}\left[\sum_{u_i \in \cap_{t=1}^{2} R_t} X(2)_i + \frac{n(1,\tilde{2})}{m(1,\tilde{2})} \sum_{u_i \in R_1 \cap R_2^c} X(2)_i + \frac{n(\tilde{1})}{m(\tilde{1})} \sum_{u \in R_1^c} X(2)_i\right],$$

and

$$\mathrm{Var}\left\{\bar{x}(2)\right\} = \frac{1}{n}\left[\frac{N-n}{N} S^2 + (K(1)-1)\frac{N(\tilde{1})}{N} S^2(\tilde{1}) + (K(2)-1)\frac{N(1,\tilde{2})}{N} S^2(1,\tilde{2})\right].$$

In general, define

$$x(T) = \sum_{u_i \in \cap_{t=1}^{T} R_t} X(t)_i ,$$

and

$$x(t,s) = \sum_{u_i \in \cap_{t=1}^{T-1} R_t \cap R_T^c} X(t)_i .$$

Then

$$\bar{x} = \frac{1}{n}\left[x(T) + \sum_{t=1}^{T} K(t)\, x(t,s)\right],$$

and

$$\mathrm{Var}\left\{\bar{x}\right\} = \frac{1}{n}\left[\frac{N-n}{N} S^2 + \sum_{t=1}^{T} (K(t)-1)\frac{N(1,2,\ldots,\tilde{t})}{N} S^2(1,2,\ldots,\tilde{t})\right].$$

4. OPTIMUM ALLOCATION FOR FIXED EXPECTED COST

Let

$$V(j) = \frac{N(1,2,\ldots,j\tilde{-}1)}{N} S^2(1,2,\ldots,j\tilde{-}1) ,$$

for values of j = 2,3,...,T+1, and,

$$V(1) = S^2 - \sum_{j=2}^{T+1} V(j) .$$

Then

$$\mathrm{Var}\left\{\bar{x}\right\} = \sum_{j=1}^{T+1} \frac{V(j)}{\ell(j)} - \frac{S^2}{N} ,$$

where

$$\ell(j) = nh(j) ,$$

with

$$h(i) = 1 ,$$

$$h(j) = \frac{1}{K(j-1)} , \quad j = 2,3,\ldots,T+1 .$$

The cost model corresponding to the above variance is given by

$$C = \sum_{j=1}^{T+1} C(j)\,\ell(j) + C_o .$$

In this expression, C_o is the component of the total cost, C, which is not affected by changes in sample sizes. The remaining cost compounds, C(j), are difficult to express algebraically although they are not difficult to compute.

A recursion relation is used to express the cost components, with the j-subscript taking the values

$$j = 1,2,\ldots,g+1 ,$$

for each value of

$$g = 1,2,\ldots,T .$$

The notation C_1 and C_2 is used to distinguish between the per unit cost of the low cost and high cost procedures, respectively. The notation $C(j)_g$ implies the j-th component of cost at the g-th computational step.

At g=1, define

$$C(1)_1 = C_1 ,$$

$$C(2)_1 = C_2 \frac{N(\tilde{1})}{N} .$$

At each subsequent step, i.e., g=2,3,...,T, a new cost component is defined for the value j=g+1 by

$$C(j)_g = C_2 \frac{N(1,2,...,\tilde{g})}{N} .$$

Otherwise, for g = 2,3,...,T,

$$C(1)_g = C(1)_{g-1} + C_1 \frac{N(1,2,...,g-1)}{N} ,$$

and for j = 1,2,...,g,

$$C(j)_g = C(j)_{g-1} + C_1 \frac{N(1,2,...,\tilde{j-1})}{N} + C_2 \sum_{\overset{\circ}{t}=1}^{s} \frac{N(1,2,...,\tilde{j-1},\overset{\circ}{t}_j,\overset{\circ}{t}_{j+1},...,\overset{\circ}{t}_{g-1},\tilde{g})}{N} ,$$

where the summation over $\overset{\circ}{t}$ implies all combinations of t and $\tilde{t}$ in the interval [j,g-1]. There are

$$s = \sum_{r=0}^{g-j} \binom{g-j}{r}$$

terms in this sum which can be arrayed as follows:

$\overset{\circ}{t}$	=	1	2	...	s
$\overset{\circ}{t}_j$	=	j	$\tilde{j}$	...	$\tilde{j}$
$\overset{\circ}{t}_{j+1}$	=	j+1	j+1	...	$\tilde{j+1}$
.		.	.		.
.		.	.		.
.		.	.		.
t_{g-1}	=	g-1	g-1		$\tilde{g-1}$

Given the cost and variance components as expressed above, the optimal solutions for fixed expected cost are

$$\ell(j) = \frac{C - C_o}{\sum_{j=1}^{T+1} \left[V(j)\, C(j)\right]^{1/2}} \left[\frac{V(j)}{C(j)}\right]^{1/2} ,$$

for all j = 1,2,...,T+1, provided that

$$\left[\frac{C(1)}{C(j)}\right] \left[\frac{V(j)}{V(1)}\right] \leq 1 ,$$

which arises from the necessity that the sampling fractions not exceed unity.

5. RESULTS FOR SIMULATED CASES

In planning a longitudinal survey, it is not unusual that knowledge of the magnitudes of population variances and nonresponse biases involved is lacking. Information is likely available concerning:

TABLE 1. SUBSAMPLING FRACTIONS FOR SELECTED DESIGN PARAMETER VALUES, T=4 PERIODS.

Response Rate	0.90	0.70	0.90	0.90	0.90	0.70	0.90	0.90	0.90	0.70	0.90	0.90	0.90	0.70	0.90	0.90
Cost Ratio	0.29	0.29	0.07	0.29	0.29	0.29	0.07	0.29	0.29	0.29	0.07	0.29	0.29	0.29	0.07	0.29
Proportion Bias	0.20	0.20	0.20	0.60	0.20	0.20	0.20	0.60	0.20	0.20	0.20	0.60	0.20	0.20	0.20	0.60
Relative Domain Size	**Subsampling Fractions Period 1**				**Subsampling Fractions Period 2**				**Subsampling Fractions Period 3**				**Subsampling Fractions Period 4**			
0.01	0.72	0.80	0.41	0.47	0.80	0.94	0.44	0.53	0.91	1.00	0.48	0.62	1.00	1.00	0.52	0.75
0.10	0.73	0.81	0.42	0.48	0.81	0.95	0.45	0.55	0.92	1.00	0.49	0.64	1.00	1.00	0.53	0.78
0.20	0.74	0.84	0.43	0.50	0.82	0.98	0.46	0.57	0.94	1.00	0.49	0.67	1.00	1.00	0.54	0.81
0.30	0.76	0.87	0.44	0.53	0.84	1.00	0.47	0.60	0.96	1.00	0.50	0.70	1.00	1.00	0.55	0.85
0.40	0.78	0.92	0.45	0.57	0.86	1.00	0.48	0.64	0.98	1.00	0.52	0.75	1.00	1.00	0.56	0.90
0.50	0.81	0.99	0.47	0.62	0.90	1.00	0.50	0.70	1.00	1.00	0.54	0.81	1.00	1.00	0.58	0.97
0.60	0.86	1.00	0.50	0.69	0.95	1.00	0.53	0.78	1.00	1.00	0.56	0.90	1.00	1.00	0.61	1.00
0.70	0.94	1.00	0.55	0.82	1.00	1.00	0.58	0.92	1.00	1.00	0.62	1.00	1.00	1.00	0.66	1.00
0.80	1.00	1.00	0.65	1.00	1.00	1.00	0.69	1.00	1.00	1.00	0.73	1.00	1.00	1.00	0.78	1.00
0.90	1.00	1.00	0.96	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.93	1.00	1.00	1.00	0.92	1.00
0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(1) the relative sizes of longitudinal domains of interest to the study,

(2) the response rate likely to be achieved under the low cost procedure, and its rate and direction of change, if any, over the total period of the study,

(3) the proportion of the maximum or minimum bias possible, given the response rate, that might be experienced.

Simulations were conducted over a range of values for these quantities. Results are presented in Table 1. The table is based on a longitudinal survey of T=4 periods, and shows the subsampling fractions to be used for the post strata defined at each time period. Cost ratios in the table are of the form C_1/C_2. The particular cost ratios chosen might represent an initial mail survey, followed up by a telephone survey in the case of the higher ratio, and a personal interview survey in the case of the lower ratio. Positive nonresponse biases are used in the table, however, values for negative biases can be obtained by subtracting the relative domain sizes in the table from unity. All data in the table are based on a single total cost constraint.

As illustrated in the table, subsampling fractions increase over the successive periods in the longitudinal survey. Solutions tend rapidly to one of "take all nonrespondents" as domain sizes increase, given positive biases, or as domain sizes decrease, given negative biases. Low response rates to the low cost procedure increases the subsampling fraction of nonrespondents, while decreasing cost ratios decrease the subsampling fractions, as would be expected. The decrease in subsampling fractions in response to increase in the proportion of the maximum bias reflects the binomial nature of the variance of the relative domain size in the nonresponding population.

6. REFERENCES CITED

[1] Hansen, M. H., and W. N. Hurwitz, 1946, The Problem of Nonresponse in Sample Surveys, *Journ. Am. Statistical Assoc.*, 41, p517-529.

The early part of this research was conducted under subcontract to Killalea Associates, Inc., and was sponsored by the National Center for Education Statistics. In its final stages, the research was supported by the Research Triangle Institute.