WEIGHTING SAMPLE DATA WHEN MULTIPLE SAMPLE FRAMES ARE USED

Barbara Lepidus Carlson, John W. Hall
John W. Hall, Mathematica Policy Research, Inc. P.O. Box 2393, Princeton, NJ 08543

Key Words: Weighting, Multiplicity, Nonresponse Adjustment

Abstract

In some surveys, multiple frames are used to more efficiently oversample population subgroups. In other cases, observations from independently selected samples are pooled for analysis. In either instance, care must be taken in weighting to account for multiple chances of selection. The use of multiple frames also complicates adjustments for nonresponse. In this paper we examine a survey of health insurance coverage, where list, random digit dialing, and area probability frames were used, and describe the weighting issues encountered.

Introduction

The target population of the survey described below fits the description in Hartley's (1962) paper well: "a frame known to cover approximately all units in the population ... is costly." Other frames are cheaper to use but cover an unknown portion of the population. The theoretical framework for multiple frame designs and comparisons of various estimation techniques are covered in Lessler and Kalsbeek (1992), Cochran (1967), Land (1968), and Fuller and Burmeister (1972).

Weighting sample data is complicated when multiple sample frames are used. If the frames are not mutually exclusive, the weighting must account for multiple chances of selection. The use of several frames can also complicate adjustments for nonresponse.

Below we discuss the construction of sample weights for ten state-level household surveys, conducted by Mathematica Policy Research, Inc. for the Robert Wood Johnson Foundation during 1993 and 1994. Households with uninsured members or those covered by Medicaid were oversampled. Each survey had a telephone and an in-person component, and each component employed one or more sample frames.

Overview of the Survey Design

The study population included household residents in each state. Households were sampled using area probability, list, or random digit dialing frames. Three units were defined for operational and weighting purposes: the household, the family, and the individual. The *household*, all persons residing at a housing unit, was the unit screened for survey eligibility. Households were subsampled based on screening reports of the health insurance coverage of their members. The *family* was the interviewing unit, and is a subunit within the household; we attempted to interview all families within a sampled household. What we call a family may be better thought of as a health insurance unit, or those persons who would be covered together under a typical health insurance policy with family coverage. An unmarried adult with no children would be classified as a one-person family within a household. In the interview, person-level information was obtained for all adults within a family and, if the family included children, for one randomly-selected child.

The survey relied heavily on telephone interviewing, but included an in-person component to cover households without telephones. The in-person component used area probability sampling to identify non-phone households, but for cost reasons included only those areas within each state with less than 95% telephone coverage. Each of the in-person frames had between two and four strata. In three of the states, listing areas were combined with lists of households identified by the State Medicaid offices.[1]

For the telephone component, households were contacted via one of three sampling frames: households with published telephone numbers, households with unpublished telephone numbers, and households identified by participating states as having at least one Medicaid recipient. Telephone numbers for the Medicaid frame were obtained from state records[2] and directory assistance searches. The published and unpublished telephone frames each had up to five sampling strata, based on estimated prevalence levels of the uninsured or persons covered by Medicaid.

Weighting Procedures

Each weight is the product of three or four factors: the first adjusts for differing probabilities of selection; the second for differences in response rates; the third adjusts for non-coverage (used only for in-person interviews); and the fourth is a ratio adjustment to external estimates of population totals. In general terms:

$$W_i = \min \{T95_d, \max \{TO5_d, (1/P_i)(1/RR_c)(NCADJ)(POPRATIO_g)\}\}$$

where:

W_i is the weight for the ith unit (household, family, and individual)

P_i is the ith unit's probability of selection

RR_c is the response rate for the cth response adjustment cell

NCADJ is the noncoverage adjustment

$TO5_d$ and $T95_d$ are the 5th and 95th percentiles of the weight distribution for insurance domain d

$POPRATIO_g$ is the ratio adjustment to external population totals for the gth group

These factors are discussed in more detail below. The computation of the probabilities of selection were the most complex.

Adjustment for Probabilities of Selection

Although there are three units of observation--household, family and individual--our discussion focuses on calculating households' probabilities. Since the design selects all families and all adults in a sampled household, the probability of selection of a household is the same as for its component families and adult members. Within each family, one child was selected at random, so a child's probability of selection equals the probability of selection for the child's household divided by the total number of children in his/her *family.*

In-Person Component Probabilities. The in-person sample designs for all states used area-probability designs, with probability proportional-to-size (PPS) selection. In most respects, the designs varied only slightly from state to state. However, there were two important differences: in the first three states, we selected PPS all the way to the interviewing area (IA) level, and within IA, used state lists to identify Medicaid households and oversampled them; for the final seven states, the method of selecting IAs differed, and we did not use Medicaid lists. The probability of selection for a household in first three of the ten states reflected four stages of selection:

$$\mathbf{P(HH) = P(PSU) * P(SSU) * P(IA) * P(HH\,|IA)}$$

where:

P(x) is the probability of selection at the x level,
PSU is the primary sampling unit, normally a county or group of counties,
SSU is the secondary sampling unit,
IA is the interviewing area (one or more Block Groups), and
P(HH |IA), the within-IA probability of selection equals the number of households contacted divided by the number of households listed.

The last term in the equation differed by whether a household was on the Medicaid sample frame. For the remaining seven states, the probability of selection reflected five stages of selection:

$$\mathbf{P(HH) = P(PSU) * P(SSU) * P(replicate) * P(chunking) * P(HH\,|IA)}$$

where:

P(replicate) is the proportion of replicates released for listing[3] and

P(chunking) is the proportion of chunks released for listing in areas that were chunked[4]. Other terms are defined above.

Calculating Telephone Component Probabilities. In the telephone component of the survey, the probability of selection for a household depends on the frame (Medicaid, published phone number, unpublished) of its telephone number(s), stratum, household insurance status, any alternate frame and stratum in which the household had a probability of selection, and number of telephone numbers reaching the household.

In our discussion below, we first discuss the problem of dual probabilities for households with members on Medicaid. We next turn to the simple case of a household with one telephone number and then present our method for accounting for households with multiple telephone listings.

Dual Probabilities in the Telephone Component. Because the in-person surveys included only households without telephones, the telephone frames and in-person frames do not overlap. Within the general population telephone sample, we created separate frames of published and unpublished numbers to eliminate overlap in coverage. However, there may be overlap between those frames and the Medicaid frames, so the telephone sample weights must account for multiple chances of selection.

All Medicaid frame households also had a chance of selection from the general population telephone frames. However, only those general population frame households who had a member receiving Medicaid could have had a chance of selection from the Medicaid frame. (We assumed that households correctly reported the Medicaid status of their members.) For households having a chance of selection from two frames, the probability depended on the alternate frame and stratum. For households sampled under the Medicaid frame, we determined whether the phone was published or not and assigned the alternate sampling frame (published or unpublished) accordingly; an alternate stratum was assigned based on phone prefix. For Medicaid households identified from the general population frame, in the seven states which provided a Medicaid frame with at least some phone numbers, the state's Medicaid[5] office matched the household against its Medicaid phone list to determine whether there was a dual probability of selection.

Because the state-provided lists of telephone numbers for the Medicaid frame were incomplete, they were supplemented through directory assistance searches. Thus, a Medicaid household identified from the general population frame might have also had a chance of selection on the Medicaid frame, even if the state had no telephone number for it. Because we could not determine whether a household fell into this category, we used the proportion of Medicaid frame households whose phone numbers were obtained by directory search as an estimate of the proportion of general population frame Medicaid households with a dual probability of selection. For general population frame households with at least one Medicaid person and a published phone number (unpublished numbers could not have been found by directory searching), but for which the state found no match on their Medicaid frame phone list, LPROP/PPROP of these households were randomly assigned dual probability of selection on the Medicaid frame where:

LPROP = the proportion of phone numbers in Medicaid frame found through directory searches out of those sent for searching.

PPROP = the proportion of Medicaid frame households with a published phone number.

One state provided phone numbers only for Medicaid frame cases in its five largest counties or for those not receiving AFDC benefits. Therefore, an automated check for matches of Medicaid households to the Medicaid frame could be carried out by the state only for non-AFDC households. Because phone numbers were provided for households in the five largest counties through a manual look-up, the state could not verify whether a general population frame Medicaid household was one for which they provided a phone number. Therefore, a different proportion of non-matched Medicaid households was assigned a dual probability depending on whether the household had a published or unpublished phone number, was an AFDC household or not, and was in one of the largest five counties or not.

Once dual probabilities and alternate frames and strata were assigned, the overall probability of selection could be calculated for each household. The probability of selection for a household sampled from the Medicaid frame with insurance domain *d* and alternate frame/stratum *hi* is:

$$P(HH) = p_m * pSUB_{dm} + (1 - p_m) * p_{hi} * pSUB_{dhi},$$

where:

p_m = probability of being screened from the Medicaid frame
= $(n'_m / N_m)\ (n''_m / (n'_m - nBAD_m))$,

$pSUB_{dm}$ = probability of being selected from the Medicaid frame if domain *d*
= $nSEL_{dm} / (nSEL_{dm} + nNOSEL_{dm})$,

p_{hi} = probability of being screened from frame *i* , stratum *h*
= $(n'_{hi} / N_{hi})\ (n''_{hi} / (n'_{hi} - nBAD_{hi}))$,

$pSUB_{dhi}$ = probability of being selected from frame *i* , stratum *h*, if domain *d*
= $nSEL_{dhi} / (nSEL_{dhi} + nNOSEL_{dhi})$,

and,

N_m , N_{hi} = the total number of households in the Medicaid frame/frame *i*, stratum *h*

n'_m, n'_{hi} = the number of households initially selected in the Medicaid frame/ frame *i*, stratum *h*

n''_m, n''_{hi} = the number of households subselected in the Medicaid frame/frame *i*, stratum *h*

$nBAD_m$, $nBAD_{hi}$ = the number of nonworking, nonresidential, or duplicate phone numbers in the Medicaid frame/frame *i*, stratum *h*

$nSELd_m$, $nSELd_{hi}$ = the number of selected households in domain *d* in the Medicaid frame/frame *i*, stratum *h*

$nNOSELd_m$, $nNOSELd_{hi}$ = the number of households not selected in domain *d* in the Medicaid frame/frame *i*, stratum *h*

For three of the ten states, the Medicaid frame and Medicaid population total were given by the state in terms of individuals, rather than households. In these cases, the Medicaid frame screening probability formula changes slightly:

$\mathbf{p_m}$ = **probability of being screened from the Medicaid frame**

$$= (n'_m / N_m)\ (n''_m / (n'_m - nBAD_m)),$$

where n'_m is the number of *individuals* initially selected in the Medicaid frame, and N_m is the total number of individuals in the Medicaid frame.

The probability of selection for a household sampled from frame *i* (non-Medicaid frame), stratum *h*, with domain *d*, is:

$$\mathbf{P(HH) = p_m * pSUB_{dm} + (1 - p_m) * p_{hi} * pSUB_{dhi}}$$

This is the same as for households sampled from the Medicaid frame, but here p_m is equal to zero if there was no alternate probability of selection from the Medicaid frame.

Adjustment for Multiple Telephones. The probability of selection was more complicated for households with more than one residential telephone number, since the additional numbers afforded these households additional possibilities of selection. The probabilities associated with these additional numbers depend on whether they were published or unpublished, and whether the household was sampled from the Medicaid frame. We assumed that the additional numbers could be selected only through the general population frames.

The probability of selection was determined for each additional number using: (1) the household's insurance domain, (2) published or unpublished status as reported by respondents or imputed from the sampled phone number, (3) stratum of the telephone number used to reach the household (where Medicaid was the original frame, we used the alternate frame). For households with more than one telephone number, the probability of selection was calculated separately for each phone number and combined as follows. For a household with T additional telephone numbers:

$$P_{mult}(HH) \quad = \quad 1 - \prod_{t=0}^{T} [1 - P_t(HH)]$$

where $P_t(HH)$ is the probability of selecting telephone t, and $P_{mult}(HH)$ is the combined probability.

Calculating Response Rates

For each of the two survey components, a response rate adjustment[6] was carried out at the household and family levels. Since there was generally one respondent for the entire family, there was virtually no person-level nonresponse for the interview. The screener response rate was calculated at the household level by frame and stratum. To calculate interview response rates we grouped households and families into response rate cells based on frame, stratum, and domain.

Each response rate cell with fewer than 30 selected households was collapsed with another cell or cells within the same domain for purposes of the response rate calculations. For non-Medicaid frame households, we collapsed cells (or more often a group of cells) from adjacent strata, within the same frame, resulting in a combined cell size of 30 or greater. For Medicaid frame cells with fewer than 30 households in a domain, the households were combined with households of the same domain, but in their alternate frame and stratum (see above)[7]. The overall response rate at the household and family levels was the product of the screener and respective interview response rates.

The household screener response rate was calculated as:

$$\frac{\text{number of eligible households that completed the screener}}{\text{no. of eligible households + (no. of households with undetermined eligibility * ELIGRATE),}}$$

where:

ELIGRATE = number of eligible households/ (number of eligible + ineligible households).

The household interview response rate is calculated as:

$$\frac{\text{number of households with at least one family responding to the interview}}{\text{number of selected households.}}$$

The *overall household response rate* is:

household screener response rate * household interview response rate.

The family interview response rate is calculated as:

$$\frac{\text{number of families responding to the interview}}{\text{number of families in selected households}^{8}\text{.}}$$

The *overall family response rate* is:

household screener response rate * family interview response rate.

Preliminary Weights

Using the probabilities of selection and the response rates, we derived preliminary weights (W_iPRE) as the product of the inverse of the probability of selection, the inverse of the response rate and a non-coverage adjustment. Using the notation introduced above:

$$W_iPRE = (1/P_i)(1/RR_c)(NCADJ)$$

where:

NCADJ = 1.0 for the telephone components

NCADJ = (1/proportion of non-phone households covered by the survey) for the in-person components

and the probabilities (P_i) and response rates (RR_c) are derived as discussed in the preceding sections.

Evaluating the Weighting Process

MPR undertook three steps to evaluate the need for sample weights and for making adjustments to the basic weights. These steps included a pre-weighting sensitivity analysis, checks for outliers among the weights, and comparisons of the sample distribution to alternative estimates of the population distribution (usually taken from the 1990 Census), which we assumed were more accurate.

Weighting survey data normally increases sample variance. Therefore, prior to computing the weights, we conducted a sensitivity analysis at the family and person levels for each survey component to determine the need for weighting. Using various demographic and health care utilization variables, we evaluated differences between weighted and unweighted estimates. These comparisons indicated that the weights had a significant effect on sample estimates.

After computing preliminary weights, we examined weights with extreme values to determine the need to truncate them; also, demographic distributions were evaluated to determine the desirability of poststratification. Our examination revealed several outlier weights that were almost always associated with children, whose probability of selection is multiplied by the number of children in the family. Based on these analyses, we decided to truncate the weights.

To evaluate the need for poststratification, telephone and in-person components were combined, and population and subpopulation totals, as well as distributions on age, race, Hispanic ethnicity, and sex, were compared with Census figures. Then we computed confidence intervals around the survey's weighted population total to evaluate whether estimated Census totals fell within sampling error of our survey estimates. We decided to poststratify each state's weighted sample totals to match Census estimates of population totals.

Truncation and Poststratification

Truncated weights were created by capping weights at the 5% and 95% quantiles by person domain (defined as Medicaid, uninsured, insured and other). The next step was to poststratify both the untruncated and truncated weights to Census figures. For all states, the in-person (nontelephone) component of the survey yielded significantly lower estimates of the number of nontelephone households than we estimated from Census figures, even after adjusting for undercoverage of nontelephone households. Therefore, the first step in the poststratification process was to inflate the weights in the in-person component so that the sum of the household level weights reflected our best estimate of the number of nontelephone households in each state.

Next, weighted distributions of race[9], whether or not Hispanic, sex, and median age, were examined and compared with 1990 Census distributions. This was carried out separately for both untruncated and truncated weights. Race and Hispanic status were first imputed for children according to the status of the oldest parent or guardian in their family, since these questions were not asked of children. The final step in the poststratificationprocess was to inflate all weights so that the population total reflected our estimate[10] of the civilian non-institutionalized populationof each state in 1993.

ENDNOTES

1. The Medicaid frame was used for the in-person componentin the first three of the ten states to aid in oversampling Medicaid households. This strategy proved inefficient and was dropped for the other seven states.

2. Only seven of the ten states provided Medicaid lists with at least some phone numbers. Two of the states provided lists but no phone numbers, and one state provided no Medicaid list.

3. The block groups (BGs) in an SSU were divided into 10 replicates of roughly equal size. We selected four to ten replicates which were released for listing, based on estimated cost and the desire to reduce variabilityin probabilities of selection.

4. Some very large block groups were assigned to more than one replicate. If some, but not all, of the replicates including the large BG were selected, we divided the BG into chunks and subsampled chunks to maintain the appropriate probability of selection.

5. For three states, other state programs were included on the Medicaid frame, and were accounted for in determining dual probabilities of selection.

6. This adjustment was carried out unweighted. Weighting cells generally contained households with similar probabilities of selection. Using preliminary weights would thus have had little impact on the response rate.

7.This involved imputing an alternate frame and stratum for a handful of either non-complete or no-alternate-determined households.

8. Note that the number of families was unknown for most of the selected households where no families responded to the interview. For these households, the number of families was imputed according the distribution found for households with known number of families, by frame, stratum, and domain.

9. Race categories were collapsed differently for each state, with race categories containing less than five percent of the population being grouped together. The Census "other race" category was, for purposes of poststratification, matched with missing responses to the survey question on race, because "other" was not an option in the Family Survey.

10. Using CPS figures of each state's civilian population in 1991 and 1992, a civilian population estimate was extrapolated for 1993 (assuming constant levels of population growth). Then, using 1990 Census figures for the state's proportion institutionalized, we estimated the 1993 civilian non-institutionalized population.

REFERENCES

Cochran, Robert S. (1967). The Estimation of Domain Sizes When Frames are Interlocking. American Statistical Association. Proceedings of the Social Statistics Section, pp. 332-335.

Fuller, Wayne A. and Leon F. Burmeister (1972). Estimators for Samples Selected From Two Overlapping Frames. American Statistical Association. Proceedings of the Social Statistics Section, pp. 245-249.

Hartley, H.O. (1962). Multiple Frame Surveys. American Statistical Association. Proceedings of the Social Statistics Section, pp. 203-205.

Lessler, Judith T. and William Kalsbeek (1992). **Nonsampling Error in Surveys**. John Wiley & Sons, pp. 83-88.

Lund, Richard E. (1968). Estimators in Multiple Frame Surveys. American Statistical Association. Proceedings of the Social Statistics Section, pp. 282-288.