Expanding the Number of Primary Sampling Units for the National Health Interview Survey

Chris Moriarity, Van Parsons
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782 USA

Abstract
The National Health Interview Survey (NHIS) is a continuous survey that has collected health data using personal interviews since 1957. The NHIS sample design is complex, with multiple stages of sampling, beginning with primary sampling units (PSU) and then additional stages of sampling within each PSU to obtain a sample of addresses. The current NHIS sample design period (2006-2015) has a state-level stratification, but only a limited number of states normally have adequate sample to support reliable estimation. Recently, additional funding became available for sample expansion in select states. The NHIS has a time-static PSU selection procedure that keeps the first-stage sample areas fixed over the length of a survey sample design period. To increase sampling efficiency, we decided to put resources into expanding the number of sampled PSUs rather than only expanding the within-PSU samples. The preferred method was to extend the probability proportionate to size (PPS) procedure that occurred at the beginning of the current design so that the expansion probabilities also remained PPS. In this paper we discuss issues that arise in such an expansion.

Key Words: Sample Survey

1. Introduction

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the U.S. It is a continuous survey that has been in operation since 1957. The current NHIS sample design was implemented in 2006. NCHS anticipates obtaining completed interviews at approximately 35,000 living quarters (households and noninstitutional group quarters such as college dormitories) each calendar year if there are no sample reductions or augmentations. All persons at a sampled address are included in the NHIS interview, yielding a sample of approximately 87,500 persons each year if there are no sample reductions or augmentations. Sample sizes can increase or decrease appreciably, according to the availability of funding. Each interview is conducted via a personal visit to the living quarters by an employee of the U.S. Bureau of the Census, which is the data collection agent for the NHIS.

The NHIS sample consists of clusters of living quarters (addresses) chosen within a first-stage sample of U.S. counties. The NHIS sample design is complex, with

multiple stages of sampling. The first stage of sampling is selection of primary sampling units (PSU), which are one or more contiguous U.S. counties. The PSUs are assigned to sampling strata, and then one or two PSUs are selected from each sampling stratum. Once the sample PSUs are selected for a given sample design period, they delineate the geographic areas where the NHIS sample addresses will be selected for the entire sample design period. There are additional stages of sampling within each PSU, yielding a sample of addresses after the final stage of sampling. This sampling method is used to control the costs related to personal visit interviewing. The cost of conducting personal visit interviews in a simple random sample of U.S. living quarters would be prohibitive, due to the amount of travel that would be required.

The current sample design of the NHIS is based on Census 2000 information. The current sample design is very similar to the previous sample design, which was in effect from 1995 to 2005 and was based on 1990 Census information. The NHIS has undergone changes in the sample design at intervals of approximately ten year duration, using information from the previous decennial census.

Additional information about the NHIS is available online at the NHIS home page, http://www.cdc.gov/nchs/nhis.htm. The reference section includes publications that describe the NHIS sample designs all the way back to when the survey began in 1957.

NCHS started receiving supplemental funding in 2010 to expand the NHIS sample in select states, beginning with the 2011 NHIS. The current NHIS sample design period (2006-2015) has a state-level stratification, but only a limited number of states normally have adequate sample to support reliable estimation. The 2011 and 2012 NHIS samples were augmented by expanding the number of sample addresses in existing primary sampling units (PSU). The two initial augmentation sources were addresses eliminated from the NHIS sample in previous years due to budget shortfalls, and addresses assigned for interview beyond 2015 ("reserve sample"). Beginning in 2012, addresses initially subsampled out during a within-PSU stage of sampling also were used as an augmentation source. By the end of 2012, both of the two initial augmentation sources had been used up.

For the 2013 NHIS, we wanted to consider the additional option of augmenting the NHIS using sample addresses in new sample PSUs. We first needed to determine if such an option was feasible. We discuss below our PSU expansion research and results.

2. PSUs Selected For the 2006-present NHIS

The current NHIS design had a fixed number of PSUs selected. Contingency planning for sample expansion using new PSUs was not part of the survey design. NHIS PSUs are either self-representing (selected with certainty) or non-self-

representing. Durbin's method (1967) was used to select PSUs in the sampling strata that contained non-self-representing PSUs. Two PSUs were selected from most non-self-representing strata; in a few cases, one PSU was selected. The NHIS sample includes one or more PSUs in all 50 states and the District of Columbia, although the number of PSUs usually is small in the least populous states.

3. Motivation For Selecting New PSUs

The sample augmentation that began with the 2011 NHIS was in 32 states and the District of Columbia. No sample augmentation occurred in the 18 most populous states because their typical annual sample sizes were considered to be adequate for precise state-level estimates. Some of the states where the NHIS sample was augmented contained few sample PSUs. Our motivation for considering the possibility of selecting new sample PSUs was to provide a mechanism for additional sample expansion in states where it is possible to select new sample PSUs. (There are several states where the entire state already is in sample.) This was particularly desirable in states where the augmentation methods that began in the 2011 NHIS within existing PSUs gave smaller state-level samples than desired. Additionally, we realized that some of the address sources that were being used for NHIS sample augmentation were going to be exhausted after a few years of use.

4. Research For Selecting New PSUs

We conducted research to determine the feasibility of extending the Durbin PSU selection method that was used to select the current NHIS sample PSUs in order to select additional sample PSUs. Extending the Durbin method would allow us to update selection probabilities for the existing sample PSUs in a straightforward way, while using a probability proportional to size method for selecting new sample PSUs.

Durbin's 1967 paper "Design of Multi-stage Surveys for the Estimation of Sampling Errors" describes an algorithm for the selection of PSUs, where each PSU has a measure of size assigned to it. The first PSU is selected with probability proportional to size. The paper describes an algorithm for assigning probabilities for selecting a second PSU, given the selection of a first one, an algorithm for assigning probabilities for selecting a third sample PSU, given the selection of two others, etc. The algorithm always is correct for selecting a single PSU because the algorithm specifies that the first PSU is selected with probability proportional to size. Eventually, the algorithm fails because it yields invalid probabilities (negative, or greater than 1). For example, if there is a PSU with selection probability greater than 0.5, and it is selected first, the algorithm produces an inclusion probability greater than 1 if a second PSU is drawn.

The software that was used to do the original PSU selection using Durbin's method in 2002 no longer was available, so we could not actually begin with the exact conditions that yielded the existing sample PSUs. We created software that implemented Durbin's method. The software then went through iterations that tried various random starts in a given stratum until one was found that selected all existing sample PSUs in the correct order in that stratum; then, given the selection of the existing sample PSUs, the software could then randomly select one or two additional sample PSUs in the stratum.

We systematically reviewed the non-self-representing strata in the states where we had an interest in selecting new PSUs to explore the possibilities for selecting one or two new sample PSUs to give a total of up to 3 sample PSUs per stratum. As part of our research, we created software that enumerated the sample space in each stratum to check for problems with the algorithm generating negative probabilities and/or inclusion probabilities greater than 1.

We imposed constraints to prevent the selection of new sample PSUs from making major alterations to existing sample PSU definitions. For example, if a stratum contained a total of 3 PSUs, and one already had been selected into sample, we would only allow the selection of one additional sample PSU. The reason for this is that selecting both of the remaining PSUs into sample would transform the existing sample PSU from being non-self-representing to self-representing. If we noticed the occurrence of the algorithm generating negative probabilities and/or inclusion probabilities greater than 1 in the sample space enumeration, we would not allow the number of new PSUs to increase to the point that one or both of these conditions could occur.

5. Results

In many of the strata where we wanted to expand the total number of sample PSUs to three, we were successful. The exceptions usually were due to the occurrence of one of the following four conditions:

1. A stratum contained a total of 3 PSUs, with 1 initially selected. We selected only one additional sample PSU so the existing sample PSU definition of non-self-representing would be retained.

2. A stratum contained a total of 2 PSUs, with 1 initially selected. We did not select the remaining PSU so the existing sample PSU definition of non-self-representing would be retained.

3. A stratum had 1 PSU initially selected. One PSU in the stratum had selection probability greater than 0.5, which leads to the algorithm generating an inclusion probability greater than 1 and/or negative selection probabilities if more than one sample PSU is selected.

4. A stratum had 2 PSUs initially selected. One PSU in the stratum had selection probability greater than 1/3, which leads to the algorithm generating an inclusion probability greater than 1 and/or negative selection probabilities if a third sample PSU is selected.

In all of the states that we wished to select new sample PSUs, we were able to select at least one new one. Our constraints and/or the limitations of Durbin's algorithm sometimes prevented us from selecting as many new sample PSUs in a given state as we had hoped to do.

6. Conclusion

We found that we were able to recreate a sampling mechanism that selected existing sample PSUs in the order they had been selected initially. This enabled us to select one or two additional sample PSUs in several sampling strata, given the selection of the existing sample PSUs.

We imposed constraints to keep existing sample PSU definitions unchanged, and to avoid the possibility of invalid probabilities being generated by the algorithm, i.e., an inclusion probability greater than 1, or a negative selection probability.

Our selection process for the new sample PSUs enables us to decouple the sample augmentation from the regular NHIS sample, and continue to treat the regular NHIS sample as a probability sample of the U.S. civilian noninstitutionalized population. This gives us the flexibility to re-use some of the augmentation addresses for research purposes.

This experience has influenced us to use different methodology for selecting the non-self-representing sample PSUs for the next NHIS sample design, which is scheduled to be implemented in 2016. We also will plan for the contingency of having new sample PSUs available if we receive sufficient funding to augment the NHIS sample.

References

Botman S, Moore T, Moriarity C, and Parsons V. Design and Estimation for the National Health Interview Survey, 1995-2004. Vital Health Stat 2(130). 2000.

Durbin, J. (1967). Design of Multi-Stage Surveys for the Estimation of Sampling Errors. Applied Statistics, 16:152-167.

Kovar MG, Poe GS. The National Health Interview Survey design, 1973–1984, and procedures, 1975–83. Vital Health Stat 1(18). 1985.

Massey JT, Moore TF, Parsons VL, Tadros W. Design and estimation for the National Health Interview Survey, 1985–94. Vital Health Stat 2(110). 1989.

National Center for Health Statistics. The statistical design of the Health Household-Interview Survey. Health Statistics. PHS Pub. No. 584-A2. Public Health Service. Washington: U.S. Government Printing Office. 1958.

National Center for Health Statistics. Health Interview Survey procedure, 1957–1974. National Center for Health Statistics. Vital Health Stat 1(11). 1975.