# Nested Subsamples: a Method for Achieving Flexibility in Annual Sample Sizes for a Continuous Multiyear Survey

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

**Abstract**
A new requirement for the most recent National Health Interview Survey (NHIS) sample redesign, implemented in 2016, was annual sample size flexibility. The new sample design accomplished this requirement by selecting a large initial sample, and then assigning sequential identifiers within the large sample. A given annual sample is obtained by including sample parts associated with the sequential identifiers, in order, until the desired annual sample size is obtained. We describe the research undertaken to implement this flexibility, including what was done after the discovery that one part of the sequential identifier assignment process needed to be revised.

**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 2016 and is anticipated to be in place through the end of 2025. Within the 2016 sample design, the National Center for Health Statistics (NCHS) obtains 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 eligible persons residing at a sampled address are covered by 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. Census Bureau, which is the data collection agent for the NHIS.

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

http://www.cdc.gov/nchs/nhis/methods.htm

Disclaimer:
The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Two major changes were implemented at the beginning of the current NHIS sample design. The first was a major change in the main source of sample addresses, and the second was increased flexibility to accommodate changes in annual sample sizes. This paper describes the increased flexibility feature, along with work carried out to revise some of the original assignments related to this feature.

## 2. Continuing Features of the 2016 NHIS Sample Design

The NHIS has always been a personal interview survey, and will continue to be so in the foreseeable future. The sample distribution for a personal interview survey usually has some level of geographic clustering, to reduce interviewer travel. This has been, and will continue to be, a feature of the NHIS sample design.

The precision of national-level annual estimates has been and remains a high priority for the NHIS sample design. To meet other geographical estimation objectives, while still maximizing precision for national-level estimates, the NHIS "base" sample (i.e., the sample with no reduction/augmentation) usually has been allocated proportional to population size within each state. The current base sample allocation is close to being proportional to state population size. A small amount of undersampling was done in the most populous states to increase the sample sizes in the least populous states to a projected 250 completed household interviews annually, which will support state-level estimates based on aggregations of three annual samples. (For convenience, any future reference in this paper to "states" is a reference to the fifty states and the District of Columbia.)

## 3. New Features of the 2016 NHIS Sample Design

There are two major new features of the 2016 NHIS sample design: a change in the sample address source, and increased flexibility. Each is briefly described in turn, and more details of the second appear later.

Sample Frame Change: The NHIS sample that is selected for interviewing is a sample of addresses. The NHIS sample frame of addresses is a list of addresses within the geographic areas selected into sample. NHIS sample designs from 1985 to 2015 used field listing to develop most of the sample frame of addresses. This was feasible because NCHS shared the cost of field listing with other federal agencies that sponsor demographic surveys (e.g., The Current Population Survey) conducted by the Census Bureau. This was not feasible for the 2016 NHIS sample design because the other federal agencies that sponsor demographic surveys conducted by the Census Bureau decided to discontinue using field listing nationwide. In 2016 and beyond, field listing is only a small part of the NHIS sample frame development process. Instead, in most geographic areas, NCHS is using addresses purchased from a commercial vendor, Marketing Systems Group (MSG).

Flexibility: A main goal of the 2016 NHIS sample design planning process was to have flexibility to increase or decrease annual sample sizes, and/or flexibility to shift annual sample sizes on a state-by-state level. In order to maintain year-to-year stability, the "base" NHIS sample, which as mentioned above is the anticipated annual sample size if there are no sample reductions or augmentations, consists of two parts. One part (~70% of the total) will remain the same from year to year, with state-level sample allocation proportional to state population size. The other part (~30% of the total) will be allowed to change from year to year during the sample design period, as directed by NCHS leadership. The 70/30 split provides a large degree of stability, a feature of all previous NHIS sample designs, while allowing for the possibility of change.

## 4. NHIS's Conceptual Sampling Structure

The conceptual structure of the NHIS survey sample is delineated in advance for the entire planned ten year sample design period, plus a contingency of two additional years. The conceptual structure is then filled with actual sample addresses a few months prior to interviewing in a given month. The conceptual structure allows the Census Bureau to do advance planning for logistics such as hiring and assignment of survey interviewers.

The conceptual sample structure planning process considered the time length (10 plus 2 years), interviewer workload (~100 sample addresses annually), and a doubling factor as a contingency for a future reinstatement of oversampling race and ethnicity groups such as black persons, Hispanic persons, and/or Asian persons. The combination of these factors results in the conceptual sampling process selecting groups of approximately 2500 geographically clustered addresses ("address groups") into sample.

Within each state, the counties (or county equivalents) were partitioned into geographic areas consisting of one or more counties. Almost without exception, the counties in the multi-county geographic areas are contiguous. In some states, the geographic areas were divided into two groups, roughly along urban/rural lines. For states with two groups, the geographic areas in the groups were designated as "Type A"/"Type B". For the remaining states, the geographic areas in each were all designated as either Type A or Type B.

Each geographic area was assigned a measure of size: the count of 2010 Census housing units. Using this measure of size, an integer number of address groups were defined within each geographic area.

## 5. Implementing the Flexibility Feature

The flexibility feature was implemented by the selection of a very large initial conceptual sample, called the "super sample", in each of the states. The size of the super sample was large enough to accommodate any possible sample augmentation/expansion in any state, relative to the historic NHIS sample allocations since the inception of the survey.

The super sample was selected independently within each state. For states with two groups of geographic areas, the sampling was done independently within each group.

The sampling departed from the historic method of designating the geographic areas as primary sampling units (PSU), selecting a sample of PSUs, then selecting address groups within each of the sampled PSUs. Instead, within each state, the conceptual address groups were sorted geographically across the collection of geographic areas, and a systematic sample was selected. In states with Type A and Type B areas, the process occurred separately within each area. The locations of the sampled address groups determined which geographic areas were part of the super sample.

The final step in each sampled area was to form a nested sequence of subsamples of the super sample by the assignment of entry orders. Once it was determined how much sample was to be allocated in a given area for a given year, entry orders 1, 2, … were taken into the final annual sample until the total number of addresses in the subsample met the target allocation.

The departure from the traditional method of selecting PSUs/selecting sample within PSUs provides the desired flexibility, allowing changes in sample sizes that can be implemented in a manner that retains stability in the sampling weights.

NCHS began the sampling process by selecting the super sample in states with Type B areas, and assigning entry orders. NCHS and the Census Bureau worked together to develop the specifications for selecting the super sample in states with Type A areas. The Census Bureau then selected the Type A super sample, and later assigned entry orders within that part of the super sample.

## 6. Revisions of Entry Order Assignments in Some Areas

The combination of the major changes and resource constraints meant that much of the sample redesign work was done very quickly, and then the focus moved on to the next set of priority tasks, leaving little or no time to assess the accuracy of work recently done. This increased the risk that errors would not be detected promptly.

NCHS and the Census Bureau came to the realization in June 2017 that a review of the annual sample was needed in some states. In order to implement any desired changes to the annual sample that would take effect at the beginning of 2018, a deadline in early July 2017 had to be met, which meant there was very little time for investigation and research.

The investigation began with assessing the super samples. The quick conclusion was that the super samples were satisfactory, no revisions were required.

Similarly, an investigation of the Type B annual samples, and the Type A annual samples in states containing only a single Type A geographic area, led to a quick conclusion that no revisions were required.

The focus was then on the annual sample in states with more than one Type A geographic area (multiple Type A units), where it was clear that some revisions were needed.

The next step was to develop a methodology for evaluating the annual samples in these states, to identify the specific revisions that were needed.

The first methodology tried was the method that is used to apportion the 435 seats in the U.S. House of Representatives to the states after each decennial census, which is known as "Hill's Method", or the "Method of Equal Proportions". (Note that the initial step that is done in House apportionment to first assign one seat to each state was not done.) This method did not give satisfactory results. In the early stages of assigning entry orders, this method gave preferential treatment to Type A units with larger populations. This is not ideal; the ideal situation is that at all points in the entry order assignment process, the distribution of the given subset of the super sample is similar to the population distribution of the super sample. Research continued to try to find a method that gave good results at all stages of entry order assignment.

A review of the output for a subsequent methodology being considered led to the development of the methodology that was used to assign revised entry orders. The central concept of the algorithm is to consider, at each step, all possibilities for assigning the next entry order, and then choosing the one that minimizes a distance function. For a given group of "n" geographic areas, each has a measure of size, the count of 2010 Census housing units. The measures of size can then be summed to compute the group's population proportion within each geographic area. The distance function to be minimized at each step is the sum of the absolute values of the differences between the population proportions and the sample proportions.

The output from application of this algorithm was examined carefully and found to be satisfactory at all stages in the entry order assignment process.

The algorithm was then applied to all states with multiple Type A units. About 70% of the states required some revision of the annual sample. The remaining states with multiple Type A units did not require any revisions to the annual sample. To be prepared for any potential future changes in the national sample allocation, the algorithm was applied to the entire super sample in all states with multiple Type A areas.

The 2016 public use file release had occurred a short time after the initial discovery that revision was needed. Research carried out after the entry order revision work was completed led to the decision to re-release the 2016 public use files with revised weights.

## 7. Subsequent Enhancements of the Entry Order Revision Algorithm

The original version of the algorithm was designed for use within a state, for a group of geographic areas.

The first generalization was to reformulate the algorithm to be applied across the 50 U.S. states, and examine the result of running the algorithm 435 times using 2010 Census population counts at the state level. The result was identical to the current U.S. House of Representatives apportionment to the 50 states. This indicates that the algorithm could be a suitable replacement for Hill's Method that does not lead to the paradoxes of historic apportionment methods (see, e.g., documentation available at the Mathematical Association of America website, https://www.maa.org/press/periodicals/convergence/apportioning-representatives-in-the-united-states-congress-paradoxes-of-apportionment).

The next generalization was to reformulate the algorithm to be applied across the entire United States (50 states and the District of Columbia), and take account of different sample correspondences in different states. First, there is some state-to-state variability in the average size of the individual annual address groups in sample. Second, in states with both Type A and Type B units, NCHS and the Census Bureau has created a correspondence between the samples in the two groups. If there is a change in the sample allocation to these states, there is increase/decrease in sample in both Type A and Type B areas, using the correspondence code, to keep the state-level base weights constant as part of the change.

This generalization was used to determine the stable part of the NHIS base sample (see Moriarity and Parsons (2015)), accounting for the variability described in the preceding paragraph. The algorithm output was modified slightly to decrease the variability of the base weights associated with a hypothetical base sample of this size.

## 8. Conclusion

The implementation of the flexibility feature has been successful. As modified in 2017, it provides the capability of implementing sample size changes (with some lead time required) in a way that weight stability is maintained.

The implementation of the flexibility feature has a cost, however; increased complexity. A number of new parameters are required to keep track of the various pieces of the super sample, the part of the super sample that is the annual sample, etc. NCHS and the Census Bureau are still facing challenges with maintaining the integrity of the complex system.

## 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.

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

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

Moriarity C, Parsons V. 2016 Sample Redesign of the National Health Interview Survey. 2015 Proceedings of the Joint Statistical Meetings, 2444-2451.

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.

Parsons V, Moriarity C, Jonas K, Moore T, Davis K, and Tompkins L. Design and Estimation for the National Health Interview Survey, 2006-2015. Vital Health Stat 2(165). 2014.