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

Chris Moriarity, Van Parsons
National Center for Health Statistics

2018 Summer Conference
Preview/Review
2018 Joint Statistical Meetings
July 16, 2018

Presentation outline

National Health Interview Survey (NHIS) overview

Major changes for the 2016 sample redesign

Discussion of the flexibility feature in the new sample design

NHIS features

A major source of official U.S. statistical information about the health of U.S. residents

Personal visit interview survey, operating continuously since 1957

Current annual sample size if no sample cuts/augmentations: ~87,500 persons in ~35,000 completed household interviews

NHIS sample design periods

Each sample design period is ~10 years long, based on information from previous decennial census

Most recent completed period: 2006-2015, based on Census 2000

Current: 2016-2025(?), based on 2010 Census

Some historic NHIS sample design features

Emphasis on producing precise national estimates - sample allocation by state approximately proportional to state population size

Most NHIS sample designs have sampled in all U.S. States and D.C. (an exception: the 1985-1994 design)

Recent NHIS sample redesigns: relatively minor changes

1995-2005 design: began using screening as part of the mechanism to oversample black and Hispanic persons

2006-2015 design: expanded oversampling to include Asian persons

2016 NHIS sample redesign: several major changes

- **Build in more flexibility to increase/decrease overall sample and/or shift sample allocations by State from year to year, if desired (lead time required to implement)**

- **New source of sample addresses**

How did we implement the increased flexibility?

Selected a large initial sample, called the "super sample", of groups of addresses

Assigned "entry orders" to govern which pieces come in/go out if there is a change in an annual sample size, and/or changes in the distribution of the sample

Super sample selection preparation

Independent sampling in each U.S. state and D.C.

Geographic areas (one or more contiguous counties) defined to delineate interviewer travel boundaries (personal visit survey)

Geographic areas assigned to one or two groups in each state

Super sample selection

In each state group, groups of addresses defined within the geographic areas

Selection of a systematic sample of groups of addresses

Where the selected address groups were located determined which geographic areas were in the super sample

Different sampling mechanism than previous NHIS sample designs

Previously, the geographic areas were primary sampling units (PSU)

First, a sample of PSUs would be selected, then a sample of address groups would be selected within the sampled PSUs

Previous sampling mechanism inhibits flexibility

We discovered in the previous NHIS sample design period that having a fixed sample of PSUs was not optimal if funding was provided for large-scale sample augmentation

More efficient sample augmentation possible with current sampling mechanism

Super sample to annual sample

We have done a thorough investigation that has determined the super sample is a "good" sample

Needed to redo part of the step from super sample to annual sample in 18 states in July 2017; changes took effect at the beginning of 2018

Two issues with the original sample in 18 states

Communication gap led to subsampling in 3 states using traditional "self-representing" methodology

Programming error (simple random subsample instead of systematic subsample) led to a less efficient annual subsample in 15 states

Principles to guide the revised subsampling

We wanted to retain the super sample, and as much of the existing annual sample as possible

We wanted to resequence the entire super sample, not just the annual sample, even in states where the annual sample was OK, to be prepared for contingencies of future sample cuts/augmentation

Preparing to resequence the super sample

The super sample pieces were all associated with geographic areas, so we could just work with the geographic areas (for brevity, referred to henceforth as PSUs)

Needed to identify an algorithm for the resequencing

First resequencing algorithm: Hill's Method

Used to apportion members of the U.S. House of Representatives to the U.S. states after each decennial census

We found that this algorithm favored the PSUs with larger measures of size (2010 Census housing unit counts) in the early stages

Second resequencing algorithm

Within a given state group of PSUs, we knew the population (2010 Census) proportions within the PSUs

Step 1: pick the PSU with largest population proportion

Later steps: consider all possible choices, pick the "best" one

Second resequencing algorithm example

Two PSUs: A, with 80% of the group population, B, with 20%

Step 1: pick A

Step 2: if pick A, sample is 100/0; compute abs(100-80)+abs*(0-20)=40. If pick B, sample is 50/50; compute abs(50-80)+abs(50-20)=60. As 40 is less than 60, at step 2, pick A.

Example, continued

Step 3: if pick A, sample is 100/0; compute abs(100-80)+abs*(0-20)=40. If pick B, sample is 67/33; compute abs(67-80)+abs(33-20)=26. At step 3, pick B.

Step 4: if pick A, sample is 75/25; compute abs(75-80)+abs*(25-20)=10. If pick B, sample is 50/50; compute abs(50-80)+abs(50-20)=60. At step 4, pick A.

Second algorithm results

We found that the second algorithm gave robust performance at all stages of resequencing

In a few state groups, needed to skip a few of the algorithm's choices near the end to match the existing super sample

Final step: re-index algorithm results to reduce transformation work from old to new

- Where possible, re-index to reduce the amount of work processing databases to change from existing entry orders to revised ones

Re-index example

A, with 80% of the group population, B, with 20%, 7 address groups in annual sample

As of July 2017: A,A,B,A,B,B,B

Resequenced: A,A,B,A,A,A,A

Can re-index up to the first four.

Example, continued

Re-indexing was done separately in the annual/non-annual pieces to assure nothing was moved across the boundary

Original B with entry order 3 was pushed into the non-annual group

Re-indexing was successful for entry orders 1,2,4.

Keeping track of all of the pieces: great complexity

Keeping track of all of the pieces of the super sample and annual sample has been very challenging

We're still working on resolving inconsistencies, etc., more than 2.5 years after the beginning of the sample design period

Summary

The NHIS undergoes periodic sample redesigns every ~10 years

Several major changes for the 2016 NHIS sample design

The flexibility requirement of the new sample design has been implemented; very complex