United States Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

April 10, 2024

National Survey of Children’s Health

Guide to Multi-Year Estimates

NSCH data files from 2016 from the same data series can be combined to increase the analytic sample size. Beginning in 2023, data files from 2016 to 2021 were re-released with improvements to weighting. These enhanced files (identified by naming convention, nsch_YYYYe_topical or nsch_YYYYe_screener) can be combined, but they should not be combined with redundant data files from 2016 to 2021 (nsch_YYYY_topical and nsch_YYYY_screener).

By leveraging a larger sample, data users can analyze smaller population groups and rare outcomes that are not sufficiently represented in a single year sample and produce national and state-level estimates with smaller standard errors. However, there are several important considerations and caveats that should be noted when analyzing multi-year survey data. This document provides discussion on the following topics, along with example software code in both SAS/SUDAAN and STATA to produce multi-year estimates.

1) Adjusting Survey Weights
2) Data Consistency Across Years
3) Presenting Multi-Year Estimates
4) Statistical Significance Testing
5) Trend Analysis

Adjusting Survey Weights

When analyzing combined years of data, the individual year survey weight will produce correct *prevalence estimates* reflecting a multi-year period. However, individual year survey weights need to be adjusted to produce the correct *weighted population sizes* that reflect an average annual or midpoint population rather than a cumulated or duplicated period population size. Since each survey year is individually weighted to represent the population of children residing in households for that year, the weight can simply be divided by the number of years being combined to derive multi-year estimates with an average annual or midpoint population size. For example, to calculate the combined 2016-2018 weight, each individual survey weight would be divided by 3 (i.e., number of survey years being combined for 2016-2018). The use of a combined weight is necessary for all analyses in which a weighted population size will be reported for the combined period.

It is also necessary to correctly define sampling strata to estimate variance and standard errors for statistical testing when analyzing multiple years of data. Whereas the two state-level sampling strata in 2016 were STRATUM=1 and STRATUM=2, sampling in 2017 and subsequent years split Stratum 2 into Strata 2a and 2b, with no households selected from Stratum 2b. When analyzing individual years, the strata can be used as defined on the individual data file. When analyzing combined years of data with 2016, it is recommended that STRATUM=2a records be recoded as STRATUM=2 to ensure that the combined file is correctly treated as having two mutually exclusive sampling strata rather than three. Guidance for variance estimation using the NSCH can be found in the NSCH Methodology Report under ‘Estimation and Hypothesis Testing’.

Example code to produce two and five-year estimates in SAS, SUDAAN, and STATA is provided below. These examples estimate the prevalence of children with special health care needs (CSHCN) from combined 2019-2020 and 2016-2020 datasets. Please note that a file containing multiple implicates for family poverty ratio [FPL] was released separately from the main topical file in 2016 (in subsequent years multiple implicates for FPL are included in the main topical file). Therefore, this file must be merged with the main 2016 topical file prior to appending data from 2017 and beyond. Additional detail and code to analyze implicate data can be found in the NSCH Guide for Analysis of Multiply Imputed Data.

Producing Two-Year Estimates in SAS and SAS-callable SUDAAN

```
/* All files saved in the same location */
libname file "<<Replace with file directory>>";

data NSCH19_20; * Create combined file by appending datasets;
set file.nsch_2019_topical file.nsch_2020_topical; * Append datasets;
fwc19_20=fwc/2; * Create average annual weight, dividing by # years;
if stratum='2A' then stratum='2'; * Recode to convert to numeric;
hhidnum = input(hhid,8.);    * Convert design variables to numeric for SUDAAN;
fipsstnum = input(fipsst,8.);
stratumnum = input(stratum,8.);
run;

proc surveyfreq data=NSCH19_20; * Example SAS surveyfreq;
strata stratum fipsst;
cluster hhid;
weight fwc19_20;
table sc_cshcn / row cl;
run;

proc crosstab data=NSCH19_20 design=wr notsorted; * Example SUDAAN crosstab;
nest fipsstnum stratumnum hhidnum / psulev=3;
weight fwc19_20;
class sc_cshcn;
table sc_cshcn;
print nsum wsum rowper serow lowrow uprow /style=nchs nsumfmt=f10.0
wsumfmt=f10.0;
run;
```

Producing Two-Year Estimates in Stata

```
/* All files saved in the same location */
cd "<<Replace with file directory>>"

use “nsch_2019_topical”, clear   /* open 2019 file */
append using “nsch_2020_topical”   /* append data set */

egen statacross=group(fipsst stratum)      /* create single cluster variable for svy */
gen fwc19_20=fwc/2   /* create average annual weight, dividing by # years  */

svyset hhid [pweight=fwc19_20], strata(statacross)   /* declare survey data */

svy: proportion sc_cshcn    /* request proportion */
```

Producing Five-Year Estimates in SAS and SAS-callable SUDAAN

```
/* All files saved in the same location */
libname file "<<Replace with file directory>>";

data NSCH2016; * Create 2016 dataset with fpl implicates;
merge file.nsch_2016_topical  file.nsch_2016_implicate;
by hhid;
run;

data NSCH16_20; * Create combined file by appending datasets;
length stratum $2;   * Averts an error message for differing variable length;
set NSCH2016 file.nsch_2017_topical file.nsch_2018_topical
file.nsch_2019_topical file.nsch_2020_topical; * Append datasets;
if stratum='2A' then stratum='2'; * Recode for 2 rather than 3 strata;
fwc16_20=fwc/5; * Create average annual weight, dividing by # years;
hhidnum = input(hhid,8.);   * Convert design variables to numeric for SUDAAN;
fipsstnum = input(fipsst,8.);
stratumnum = input(stratum,8.);
run;

proc surveyfreq data=NSCH16_20; * Example SAS surveyfreq;
strata stratum fipsst;
cluster hhid;
weight fwc16_20;
table sc_cshcn / row cl;
run;

proc crosstab data=NSCH16_20 design=wr notsorted; * Example SUDAAN crosstab;
nest fipsstnum stratumnum hhidnum / psulev=3;
weight fwc16_20;
class sc_cshcn;
table sc_cshcn;
print nsum wsum rowper serow lowrow uprow /style=nchs nsumfmt=f10.0
wsumfmt=f10.0;
run;
```

Producing Five-Year Estimates in Stata

```
/* All files saved in the same location */
cd "<<Replace with file directory>>"

use "nsch_2016_topical", clear
merge 1:1 hhid using "nsch_2016_implicate" /* merge 2016 file with fpl implicates */
tostring stratum, replace  /* convert to character for compatibility with subsequent files */

append using “nsch_2017_topical”            /* append data sets */
append using "nsch_2018_topical"
append using "nsch_2019_topical"
append using "nsch_2020_topical"

replace stratum="2" if stratum=="2A" /* recode for 2 rather than 3 strata */
egen statacross=group(fipsst stratum)     /* create single cluster variable for svy */
gen fwc16_20=fwc/5 /* create average annual weight, dividing by # years */
```

```
svyset hhid [pweight=fwc16_20], strata(statacross) /* declare survey data */ svy:

proportion sc_cshcn /* request proportion */
```

Data Consistency Across Years

The NSCH prioritizes consistency across years, but changes to question wording, response options, and data processing have occurred. One resource for assessing changes in questionnaire items across cycles of the NSCH is the NSCH Codebook. The codebook lists variables that are included in the redesigned NSCH public use files. Question wording, response options, reported ranges, and other details are organized by variable name. If there was a change in those details from one year to the next, the information is listed in separate panels under the variable name.

In the case of major modifications or item additions and deletions, variables may not be combined across years. In other cases, response options can be collapsed in one year to mimic the range of responses available in another. Data users must decide if a change to question wording represents a substantial change to the data series and should note any related limitations with their reported results.

Presenting Multi-Year Estimates

With the adjustment to survey weights, as described above, estimates of *population size* reflect an average across multiple years. However, *prevalence estimates* from multiple years (e.g., the percent of CSHCN) are not an exact average of single year estimates since weighted population sizes change from year to year. Thus, each annual prevalence estimate is not equally weighted in a multi-year average. To avoid misinterpretation, prevalence estimates should refer to a multi-year period rather than an average, such as the percent of CSHCN in 2016-2017.

With regard to weighted survey response and interview completion rates, several options exist for multi-year periods. Data users may choose to report these details from each year included in the multi-year estimates, the range, or a simple average from the years included.

Statistical Significance Testing

Significance testing of multi-year estimates is recommended only with non-overlapping (i.e., exclusive) samples to support accurate independent statistical tests using standard two-sample methods. For example, these methods could be used to determine whether there is a statistically significant difference in the percent of CSHCN in 2018-2019 versus 2016-2017 (i.e., exclusive, non-overlapping samples). Changes over time should be sustained over multiple assessments to confirm a change versus a single fluctuation or random error.

Prevalence comparisons of two estimates in a multi-year appended dataset can be obtained directly in standard statistical software with specific years selected in a domain variable for SAS, a subpopx statement for SUDAAN, or a subpop request in STATA. If two multi-year estimates are being compared to improve reliability at the state level (e.g., 2017-2018 and 2019-2020), a new year variable can be created to combine and distinguish the estimates (e.g, 2017 and 2018 observations assigned to 2017.5 and 2019 and 2020 observations assigned to 2019.5). In this case, a single multi-year weight would need to be created to produce annualized population estimates (e.g., fwc2=fwc/2).

A statistical significance testing tool is available from the Census Bureau to facilitate comparisons between years using estimates and standard errors that are pre-calculated or were derived separately. The Health Resources and Services Administration's Federally Available Data (FAD) Resource Document includes estimates and standard errors for select indicators at both national and state levels as well as statistical significance tests across time.

Trend Analysis

Ideally, when three or more data points are available, formal tests for trend should be used to incorporate all data points and examine the shape of a trend. Record level analysis, rather than analysis of aggregated estimates, is recommended to properly account for year-to-year correlation in variance estimation using complete survey design information. Linear trends can be assessed with tabular tests (e.g., Cochrane-Mantel-Haenszel), which are particularly helpful for multi-category outcomes, or regression-based approaches. Non-linearity can be assessed with polynomial terms in regression models, JoinPoint regression, and cubic spline models. When single year estimates are unstable, particularly at the state-level, multi-year estimates may be preferable to analyze and visualize underlying trends but may obscure the precise timing of trend changes. Either logistic or linear probability models may be used to assess trends for binary outcomes. More information on trend analysis recommendations can be located within the National Center for Health Statistics Guidelines for Analysis of Trends.