# 2017 National Survey of Children's Health

## Guide to Analysis of Multi-Year NSCH Data

U.S. Census Bureau

10/25/2018

United States® Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

@uscensusbureau

Data across multiple years of the redesigned NSCH (2016 and later) can be combined to increase the analytic sample size. 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) Discussing Multi-Year Estimates
4) Statistical Significance Testing

**Adjusting 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 2016-2017 weight, each individual survey weight would be divided by 2 (i.e., number of survey years being combined for 2016-2017). The use of a combined weight is necessary for all analyses in which a weighted population size will be reported.

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 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 data file. When analyzing combined years of data, it is recommended that 2017 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 can be found in the 2017 NSCH Methodology Report under 'Estimation and Hypothesis Testing'.

Example code to produce multi-year estimates in SAS, SUDAAN, and STATA is provided below. This example estimates the prevalence of children with special health care needs (CSHCN) from a combined 2016-2017 dataset. 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 2017 NSCH Guide for the Analysis of Multiply Imputed Data.

*Producing Multi-Year Estimates in SAS and SAS-callable SUDAAN*

```
/* 2016 Topical file, 2016 Implicate file and 2017 Topical file should all be
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_17; * Create combined file by appending datasets;
length stratum $2;  * Averts an error message for differing variable length;
set NSCH2016 file.nsch_2017_topical; * Append datasets;
if stratum='2A' then stratum='2'; * Recode for 2 rather than 3 strata;
fwc16_17=fwc/2; * Create average annual weight;
hhidnum = input(hhid,8.);  * Convert design variables to numeric for SUDAAN;
fipsstnum = input(fipsst,8.);
stratumnum = input(stratum,8.);
run;

proc surveyfreq data=NSCH16_17; * Example SAS surveyfreq;
strata stratum fipsst;
cluster hhid;
weight fwc16_17;
table sc_cshcn / row cl;
run;

proc crosstab data=NSCH16_17 design=wr notsorted; * Example SUDAAN crosstab;
nest fipsstnum stratumnum hhidnum / psulev=3;
weight fwc16_17;
class sc_cshcn;
table sc_cshcn;
print nsum wsum rowper serow lowrow uprow /style=nchs nsumfmt=f10.0
wsumfmt=f10.0;
run;
```

*Producing Multi-Year Estimates in Stata*

```
/* 2016 Topical file, 2016 Implicate file and 2017 Topical file should all be
saved in the same location */
local file = "<<Replace with file directory>>"

use "`file'\nsch_2016_topical", clear
merge 1:1 hhid using "`file'\nsch_2016_implicate" /* create 2016 file with fpl implicates */
save  "`file'\nsch_2016_topical", replace

use "`file'\nsch_2017_topical", clear  /* open 2017 file */
replace stratum="2" if stratum=="2A" /* recode for 2 rather than 3 strata */
destring stratum, replace /* convert to numeric for compatibility with 2016 data */

append using "`file'\nsch_2016_topical"    /* append data sets */
egen statacross=group(fipsst stratum)    /* create single cluster variable for svy */
gen fwc16_17=fwc/2 /* create average annual weight */
```

```
svyset hhid  [pweight=fwc16_17], strata(statacross)  /* declare survey data */

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

**Data Consistency (2016 to 2017)**

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 Crosswalk. The crosswalk lists variables (alphabetically) that are included in the redesigned NSCH public use files. Changes in question wording, response options, and reported ranges are highlighted in this resource, as are the deletion and addition of survey items.

When combining data from the 2016 and 2017 NSCH, data users should pay particular attention to the response categories included in the derived variables of INSTPYE and FAMILY. For example, the unknown category of insurance type (INSTYPE=4) was not included in 2017; a combined 2016-2017 coding would eliminate that category in 2016. In other cases, response options for 2017 can be collapsed in one year to mimic the range of responses available in 2016. 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.

**Discussing 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 with non-overlapping (i.e. exclusive) samples is done 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). The use of non-overlapping samples to produce estimates is preferred. If overlapping samples are used, change will be muted due to shared data and the statistical adjustment to remove the contribution of the overlapping sample can introduce some error.

Furthermore, the interpretation of statistical significance when comparing overlapping samples is driven by the non-overlapping years. Thus, from a technical standpoint, it is best to examine change over time with non-overlapping estimates.

Nested multi-year estimates, a special case of estimates from overlapping samples in which one sample is wholly represented in the second sample, may be compared statistically using a t-test for overlapping groups. For example, 2016 estimates may be compared with 2016-2017 estimates because 2016 is a subset of 2016-2017. The formula involves an adjustment for the proportion of variance that is overlapping or shared.

Whereas a Z-test for independent samples divides the difference between two means ($\bar{X}$) by the square root of the sum of the squared standard errors ($SE$), the nested Z-test also removes the share of variance that is redundant with two times the proportion of observations that are shared times the squared standard error of the shared observations. The formula for this is as follows:

$$Z = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{SE_i^2 + SE_j^2 - 2P * SE_j^2}}$$

Where j is a subset of i (e.g., 2016 within 2016-2017) and P is the proportion of the weighted denominator for a given indicator that is specific to j (e.g., 2016 weighted denominator divided by 2 times the 2016-2017 weighted denominator). For 2016 and 2016-2017, P is ~0.5 so the denominator of the test statistic is approximately the 2016-2017 SE and essentially reduces to a comparison between 2016 and 2017. The difference and interpretation of the test is driven by the non-overlapping years, 2016 and 2017. For example, a nested Z-test comparing a 2016-2017 estimate to a 2016 estimate should be interpreted as comparing 2017 to 2016.