2017-2019 National Survey of Family Growth (NSFG): Sample Error Estimation Design

This document describes the 2017-2019 NSFG sample error estimation design and is a detailed supplement to the briefer overview of all aspects of the methodology and survey operations for the 2017-2019 NSFG, posted on the NSFG webpage as "2017-2019 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods." The NSFG has been conducted by the National Center for Health Statistics (NCHS) since 1973 and is a principal source of national estimates on a variety of fertility, health and family formation topics. The target population for the NSFG consists of all non-institutionalized women and men aged 15-49 years (15-44 before 2015) as of first contact for the survey, living in households, and whose usual place of residence is the 50 United States and the District of Columbia. As with NSFG surveys in 2002, 2006-2010, 2011-2013, 2013-2015, and 2015-2017, sample design and fieldwork for the 2017-2019 NSFG were conducted by the University of Michigan's Institute for Social Research under a contract with NCHS.

This document pertains to the 2017-2019 public-use data from the NSFG. This is the fourth of four 2-year public-use data releases from an overall period of fieldwork spanning 2011-2019. The first three public-use file releases are listed below, along with links to their sample error estimation design documents:

- Data from the first release covered September 2011 through September 2013, and a report analogous to this one can be found in "2011-2013 National Survey of Family Growth (NSFG): Sample Error Estimation Design."
- Data from the second release covered September 2013 through September 2015, and a similar report can be found in "2013-2015 National Survey of Family Growth (NSFG): Sample Error Estimation Design."
- Data from the second release covered September 2015 through September 2017, and a similar report can be found in "2015-2017 National Survey of Family Growth (NSFG): Sample Error Estimation Design."

NSFG Sampling Error Estimation Codes

Multi-stage area probability samples require coding schemes for the strata and clusters in order for users to estimate sampling variance appropriately. These coding schemes are used for at least three reasons. First, the sample design often includes clustering at multiple levels (e.g., Primary Sampling Units - PSUs and Secondary Sampling Units - SSUs). These clusters are often collapsed to the highest level under an "ultimate cluster" model for estimating sampling variance. Second, coding schemes are used to limit disclosure risk for the geographic areas

included in the sample. The combination of collapsing and combining PSUs can allow survey designers to disguise the identity of specific PSUs included in the sample. Third, combining first-stage selections increases the number of observations within a stratum, leading to greater reliability in estimates of variance. This is especially important in datasets where small subgroups are to be analyzed. The coding scheme should be designed such that in expectation every stratum has at least two clusters that have observations from key population subgroups. On the other hand, the collapsing should not be based on observed values as this would tend to bias variance estimates in a downward direction.

In general, these coding schemes involve creating pseudo-strata and pseudo-clusters. At a minimum, each pseudo-stratum will contain two or more pseudo-clusters. More clusters per stratum can increase the reliability of variance estimates and will increase the degrees of freedom for confidence intervals and inference. However, reducing the number of strata will tend to result in over-estimates of sampling variance. These losses in precision tend to be small as most of the gains from stratification occur with few strata (see Cochran, 1977, p. 132; Kish, 1965, p. 102). In the case of the 2017-2019 NSFG, strata from the original design were collapsed, but this was done in a way that would minimize losses. Strata that were expected to be most similar to each other using the design information available from the sampling frame were collapsed together. Specifically, strata with similar geography or urbanicity were collapsed together to form new strata for variance estimation purposes.

The 2011-2019 NSFG survey period used a continuous fieldwork design, where new samples of PSUs were released annually, while new samples of SSUs and housing units were released each quarter. In a continuous design, the sampling error estimation coding schemes are complicated by the additional dimension of time. The key aim of the coding scheme was to allow users to make comparisons between estimates at different points in time for the four-year data release. For the variance estimation, this meant that each pseudo-stratum had to include at least two clusters that were measured at each point in time. As a result, there were four pseudo-clusters included in each pseudo-stratum.

While the NSFG during the 2011-2019 data collection period shared the continuous design employed for the 2006-2010 NSFG, the 2011-2019 period had a somewhat different challenge with regard to defining pseudo-strata. This is because (unlike in 2006-2010) the 2011-2019 design produced separate, non-overlapping datasets from each of four 2-year intervals in the data collection period – 2011-2013, 2013-2015, 2015-2017, and 2017-2019. During the sample design stage, two types of strata were identified for the full eight-year data collection: those strata that were “self-representing,” and those that were not. In general, the self-representing strata were organized into three groups. Self-representing PSUs were included for different amounts of time to allow the appropriate sampling rates for the different sizes of these PSUs.

The first group was large enough to be in the sample every year. The second group was large enough to be in the sample two out of three years. The third group was large enough to be in the sample once every three years. This created a rotation of "self-representing" PSUs. Technically, in any two-year interval, some of these PSUs are not, in fact, "self-representing." Therefore, it is not a self-representing PSU in the first two-year dataset. Within each of these three groups, PSUs were organized into "super-strata," based on geography and urbanicity with most similar PSUs being grouped together. PSUs within each super-stratum were randomly sorted and then systematically assigned across the eight years. Additional details are available in "2017-2019 National Survey of Family Growth (NSFG): Sample Design Documentation." The probability that each "self-representing" PSU would be assigned to a two-year, four-year, six-year, or eight-year interval was then calculated and used to develop weights for each two-, four-, six-, or eight-year interval.

These pseudo-strata and pseudo-clusters have been coded as variables on the public-use dataset. The pseudo-strata are contained in the variable SEST, while the pseudo-clusters are contained in the variables SECU. The SECU values are nested within the SEST. That is, there are four pseudo-clusters in each pseudo-stratum and they are numbered 1, 2, 3, and 4. To uniquely identify each cluster, both the SEST and SECU must be specified. This coding scheme works with the major software packages available for the estimation of variance from complex sample surveys, including SAS and Stata.

The pseudo-strata and clusters have been coded for all eight years of data collection. These are non-overlapping in the sense that they are formed within each two-year dataset. The 2017-2019 NSFG has sampling error strata and clusters that are unique to that sample. This should simplify the task of combining two-year datasets as the same SEST and SECU codes will work for single two-year datasets or combined 4-, 6-, or 8-year datasets based on combining or "stacking" two, three, or all four of the public-use data sets. The same number of strata and clusters are formed in each two-year interval. Table 1 shows the number of pseudo-strata and pseudo-clusters for sequential cumulations of all of the planned two-year public-use releases of NSFG data. Each public-use data release involves the release of two-year files, which can then be combined with prior two-year releases to yield four-year, six-year, and eight-year files. The weights for these combined files are provided here. Based on the fact that each two-year release has 18 pseudo-strata and 72 pseudo-clusters, Table 1 shows the total numbers of pseudo-strata and pseudo-clusters associated with using two-, four-, six-, or eight-year data sets.

Table 1. Number of Strata and Clusters for Datasets Based on Two, Four, Six, or Eight Years of NSFG Data, 2011-2019

Cumulated Public Release Data Files	Number of Pseudo-Strata	Number of Pseudo-Clusters
Two Years[a]	18	72
Four Years[b]	36	154
Six Years[b]	54	216
Eight Years (2011-2019)	72	288

[a]Refers to any two-year file release.

[b]All possible combinations of any consecutive data releases are included.

Data users are reminded that standard statistical procedures are based on the assumption that data are generated via simple random sampling (SRS) and will generally produce incorrect estimates of variances and standard errors when used to analyze data from the NSFG. Analysts who apply SRS techniques to NSFG data will generally produce standard error estimates that are, on average, too small, and are likely to produce results that are subject to excessive Type I error. For further details on analysis of complex sample survey data, see Heeringa, West, and Berglund (2010).

Analysts are strongly encouraged to use appropriate software to account for the NSFG's complex sample design in their analyses. Several software packages are available for analyzing complex samples. The key design variables for analysis of 2017-2019 NSFG data are:

- SEST: Stratum variable
- SECU: Cluster
- WGT2017_2019: Final weight

Guidance and further details on using these survey design and weight variables can be found in the "Sample Weights and Variance Estimation" section of the NSFG User's Guide.
As noted above, along with the release of the 2017-2019 NSFG public-use data, a total of six combined-file weights for the 2011-2019 survey period have been provided here. On this NSFG page, users will find further technical guidance related to potential analyses to facilitate using 4-, 6-, or 8-year combined or "stacked" datasets over the 2011-2019 survey period.

References

Cochran, W. G. (1977). Sampling Techniques. New York, Wiley.

Heeringa, S., B. T. West and P. A. Berglund (2010). Applied Survey Data Analysis. Boca Raton, FL, Chapman & Hall/CRC.

Kish, L. (1965). Survey Sampling. New York, Wiley.

For a Glossary of terms used in this document and related documents, see Appendix I in "2017-2019 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods."