2015-2017 National Survey of Family Growth (NSFG): Sample Error Estimation Design

This document is a detailed supplement to another document that serves as a brief, summary description of all aspects of the methodology and operations for the 2015-2017 NSFG data release. The summary document is referred to as "summary methodology document" below and is entitled "2015-2017 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods."

This document pertains to the 2015-2017 data from the NSFG. This is the third of four planned two-year data releases from an overall period of planned fieldwork spanning 2011-2019. The first two are listed below:

- 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 fourth release will cover September 2017 through September 2019, and a similar report will be released sometime after this release in late 2020.

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 NSFG 2015-2017, 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 NSFG 2011-2019 survey period uses a continuous fieldwork design, where new samples of PSUs are released annually, while new samples of SSUs and housing units are released each quarter. In a continuous design, the sampling error estimation coding schemes are complicated by the additional dimension of time. To account for the time dimension in the prior NSFG survey period (2006-2010) where a continuous fieldwork design was also used, two key aims of the coding scheme were to: (a) be consistent across two data releases (2.5-year 2006-2008 and four-year 2006-2010) and (b) 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.

NSFG in 2011-2019 shares the continuous design employed for the 2006-2010 NSFG. The NSFG 2011-2019 design creates datasets from each of four two-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 are 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 the Sample Design Documentation. The probability that each "self-representing" PSU would be assigned to

a two-year, four-year, or six-year interval is then calculated and used to develop weights for each two-, four-, or six-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 2015-2017 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 intervals or combined datasets (e.g., a four-year dataset including 2011-2013 and 2013-2015 or a six-year dataset that adds 2015-2017 to the four-year dataset). 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 datafile release involves the release of two-year files, which can then be combined with prior two-year releases to yield the four-year, six-year, and eight-year files. Appropriate weights have been or will be provided for these combined files. The 2015-2017 NSFG public-use file has 18 pseudo-strata and 72 pseudo-clusters. Combining the 2013-2015 and 2015-2017 datasets will produce a file with 36 pseudo-strata and 154 pseudo-clusters. Combining 2011-2013, 2013-2015, and 2015-2017 datasets will produce a file with 54 pseudo-strata and 216 pseudo-clusters.

Table 1. Number of Strata and Clusters for each NSFG Public-Use Data Release

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	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 generally will 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). Also see the NSFG User's Guide, Main Text, page 5.

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 2015-2017 NSFG data are:

- Stratum variable: SEST
- Cluster: SECU
- Final weight: WGT2015_2017

Guidance and further details on using the weights and survey design variables can be found in the "Sample Weights and Variance Estimation" section of the NSFG User's Guide.

With the release of the 2015-2017 data for public use, two additional weight variables have been provided that will enable analyses of combined or "stacked" datasets from the 2011-2017 survey period.

- A four-year weight variable has been created that will enable analyses of combined datasets consisting of data from 2013-2015 and 2015-2017. (Note that a four-year weight variable for combining 2011-2013 and 2013-2015 data was previously released with the 2013-2015 public-use file release.) The four-year weights are adjusted to account for the sampling rates that accrue over the four-year interval. They are also post-stratified to estimated population control totals for the midpoint of this survey period - July 1, 2015. As such, these weights are appropriate for analyses involving 2013-2017 datasets.
- A six-year weight variable has been created that will enable analyses based on combining data for 2011-2013, 2013-2015, and 2015-2017. These six-year weights are adjusted to account for the sampling rates that accrue over this six-year interval. They are also post-stratified to estimated population control totals for the midpoint of this survey period - July 1, 2014. As such, these weights are appropriate for analyses involving the 2011-2017 datasets.

The SEST and SECU variables are numbered such that they will work for either the four-year or six-year combination of datasets. For further guidelines and SAS and Stata language examples for combining data across file releases, see the 2015-2017 NSFG User's Guide, Appendix 2.

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 "2015-2017 National Survey of Family Growth (NSFG): Design and Data Collection Methods".