2013-2015 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 2013-2015 data release. The summary document is referred to as "summary methodology document" below and is entitled "2013-2015 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods".

This document pertains to the 2013-2015 data from the NSFG. This is second of two data releases from an overall period of planned fieldwork spanning 2011-2019. Data from the first release covers 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".

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 2013-2015, 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 are collapsed together to form new strata for variance estimation purposes.

The National Survey of Family Growth (NSFG) 2011-2019 has a continuous 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. In the prior NSFG, fielded in 2006-2010, to account for the time dimension, two key aims of the coding scheme were to (a) be consistent across two data releases (2.5-year 2006-2008 and 4-year 2006-2010) and (b) for the 4 year data release to allow users to make comparisons between estimates at different points in time. 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 2011-2019 shares the continuous design employed for 2006-2010 NSFG. The NSFG 2011-2019 design creates datasets from each of four 2-year intervals in the data collection period – 2011-2013, 2013-2015, and so forth. During the sample design stage, two types of strata were identified for the full 8-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 2-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 2-year or 4-year interval is then calculated and used to develop weights for each 2- or 4-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, 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 8 years of data collection. These are non-overlapping in the sense that they are formed within each two-year dataset. NSFG 2013-2015 has sampling error strata and clusters that are unique to that sample. This should simplify the task of combining and splitting 2-year datasets as the same SEST and SECU codes will work for single 2-year intervals or combined datasets (e.g. a 4-year dataset including 2011-2013 and 2013-2015). The same number of strata and clusters are formed in each 2-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 release involves the release of 2-year files, which can then be combined with prior 2-year releases to yield the 4-year, 6-year, and 8-year files. Appropriate weights will be provided for these combined files. The NSFG 2013-2015 public use file has 18 pseudo-strata and 72 pseudo-clusters. Combining the 2011-2013 and 2013-2015 datasets will produce a file with 36 psedu-strata and 154 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	18	72
Four Years	36	154
Six Years	54	216
Eight Years	72	288

Data users are reminded that standard statistical procedures are based on the assumption that data are generated via simple random sampling (SRS) generally will 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).

Analysts are strongly encouraged to use appropriate software to reflect the complex sample design in their analyses. Several software packages are available for analyzing complex samples. The key design variables for analysis of 2013-2015 NSFG data are:

- Stratum variable: SEST
- Cluster: SECU

- Final weight: WGT2013_2015

Examples of analyses using the survey procedures in SAS, Stata, and SPSS can be found here. Also see the 2013-2015 User's Guide for additional analysis information.

In addition, a weight variable has been created that will enable analyses of combined datasets consisting of data from 2011-2013 and 2013-2015. These two datasets can be combined by "stacking" them. These weights are adjusted to account for the sampling rates that accrue over a four-year interval. They are also poststratified to estimated control totals for July 1, 2013. As such, these weights are appropriate for analyses involving four-year datasets. The SEST and SECU variables are numbered so that they are non-overlapping and as such, they will work for combined datasets.

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 "2013-2015 National Survey of Family Growth (NSFG): Design and Data Collection Methods".