

# **Sampling and Nonsampling Errors**



*This section discusses methods for computing sampling errors and highlights major sources of nonsampling error in SIPP.*

- *Computing Sampling Error*
  - Direct Variance Estimation*
  - Approximate Variance Estimation*
- *Sources of Nonsampling Error*
  - Differential Undercoverage*
  - Nonresponse*
  - Measurement Errors*
- *Effects of Nonsampling Error on Estimates*

## Computing Sampling Error

Analysts often mistakenly ignore a survey's complex design and treat the sample as a simple random sample (SRS) of the population. If analysts apply SRS formulas for variances to SIPP estimates, they will typically underestimate the true variances.

The following approaches are useful in obtaining variances for SIPP estimates.

### Direct Variance Estimation

The SIPP data files contain primary sampling unit (PSU) and stratum variables that were created for the purpose of variance estimation. When analysts use these variables with software designed for complex surveys, they can calculate appropriate variances of survey estimates.

**1990–1993 Panels.** In the public use data files, analysts should look for the following variable names for the variance stratum and variance unit codes associated with each sample member:

- HHSC and HSTRAT in the core wave files
- HALFSAMP and VARSTRAT in the full panel files

These codes can be used in any of the software packages for variance estimation with complex sample designs.

**1996 Panel.** For the 1996 Panel, analysts should use Fay's method for estimating variances. This modified balanced repeated replication method allows the use of both halves of the sample. Thus, no subset of the sample units in a particular classification will be totally excluded.

The variance formula for Fay's method is presented and discussed in Chapter 7 of the *SIPP Users' Guide*.



### **Approximate Variance Estimation**

The Census Bureau provides two forms for approximate variance estimation:

- Generalized variance functions (GVFs), which are updated annually
- Tables of standard errors for different estimated numbers and percentages

The use of GVFs and tables of standard errors is described in the source and accuracy statement included with each data file. Examples of their use appear in Chapter 7 of the *SIPP Users' Guide*.

### **Sources of Nonsampling Error**

A full discussion of nonsampling errors in SIPP is presented in the third edition of the *SIPP Quality Profile* (available at the SIPP Web site). In this tutorial, we briefly describe three broad sources of nonsampling error.

#### **Differential Undercoverage**

One source of error in SIPP is differential undercoverage of demographic subgroups, particularly young adult black males. Undercoverage in SIPP is due mainly to omissions within households rather than to omissions of entire households.

To compensate for undercoverage, the Census Bureau uses known population controls to adjust SIPP weights.

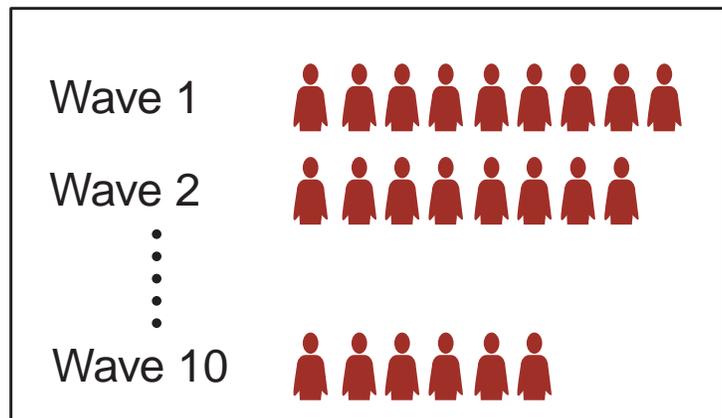
#### **Nonresponse**

Nonresponse is a major concern in SIPP because of the need to follow the same people over time. In SIPP, nonresponse can occur at several levels:

- Household nonresponse at the first wave and thereafter
- Person nonresponse in interviewed households
- Item nonresponse, including complete nonresponse to topical modules

Nonresponse reduces the effective sample size, thereby increasing sampling error, and may bias the survey estimates.

The Census Bureau uses weighting and imputation methods to reduce the potential biasing effects of nonresponse (see Chapters 4, 5, and 8 of the *SIPP Users' Guide*).



### **Measurement Errors**

Measurement errors occur during data collection and processing. They may vary across SIPP panels because of changes in data collection procedures. For example, SIPP switched from total face-to-face interviews in the early panels to a mix of telephone and face-to-face interviews since February 1992.

Response errors in SIPP include:

- Errors of recall
- Errors in proxy respondents' reports
- Errors associated with respondents' misinterpretation of questions
- Errors associated with the panel nature of SIPP

To reduce memory error, SIPP uses a relatively short recall period of 4 months for most questions. Also, interviewers encourage respondents to use financial records and event calendars to facilitate recall.

Two special sources of response error arise from the panel nature of SIPP:

- **The Time-in-Sample Effect (or Panel Conditioning).** This effect refers to the tendency of sample members to “learn the survey” over time. The concern is that sample members will alter their responses in an effort to conceal sensitive information or to shorten the length of the interview.

- **The Seam Phenomenon.** Research has consistently shown that SIPP respondents tend to report the same status (e.g., program participation) and the same amounts (e.g., Social Security income) for all 4 months within a wave. Thus, most changes in status are reported to occur between the last month of one wave and the first month of the next wave, which is the seam between the two waves.

The seam phenomenon results in an overstatement of changes at the on-seam months and an understatement of changes at the off-seam months. 

## ***Effects of Nonsampling Error on Survey Estimates***

Despite extensive research on nonsampling error in SIPP, it is difficult to quantify the combined effects of nonsampling error on SIPP estimates. A full discussion of this issue appears in the *SIPP Quality Profile*.

Some of the research findings that users should keep in mind when conducting their analyses and examining the results include the following:

- Demographic subgroups underrepresented in SIPP include:
  - Young black males
  - Metropolitan residents
  - Renters
  - People who changed addresses during a panel
  - People who were divorced, separated, or widowed

Census Bureau adjustments to correct the underrepresentation may not fully address potential biases.

- Differences exist between SIPP and CPS estimates of the working population, people without any health insurance coverage, and, for pre-1996 panels, people in poverty.

## **SIPP** *tip*

*Because of the rotation group design used in SIPP, the seam phenomenon has relatively small effects on cross-sectional estimates based on all four rotation groups. Its effects on longitudinal estimates are not well known.*

- SIPP estimates of interest and dividend income are prone to error and tend to be underreports. SIPP estimates of assets, liabilities, and wealth are low relative to estimates from the Federal Reserve Board.
- Compared with estimates based on administrative records, SIPP estimates of income from Social Security, Railroad Retirement, and Supplemental Security programs are similar, but SIPP estimates of unemployment income, worker's compensation income, veteran's income, and public assistance income are low.
- SIPP and CPS estimates of number of births are comparable, but are low relative to records from the National Center for Health Statistics.

## ***Sampling Weights***

*This section briefly describes why weights are important in SIPP analyses and how to use them.*

- *Purpose of Using Weights*
- *Weights Available in SIPP Files*
- *Choosing Weights*
- *Using Weights in SIPP Analyses*
  - Core Wave Files*
  - Topical Module Files*
  - Full Panel Files*
  - Estimation with Full Panel Files*



## Purpose of Using Weights

SIPP data analysts need to understand the importance of using weights to minimize bias in survey estimates. Biased estimates will likely occur if the responding units in a survey do not reflect the target population and the units are not adjusted with weights.

In general, weighting is necessary when:

- Population units are sampled with different selection probabilities
- Coverage rates and response rates vary across subpopulations

In the 1990 and 1996 SIPP Panels, subpopulations were sampled at different rates. In addition, there have been minor variations in sampling rates in all SIPP panels as well as appreciable variations in response and coverage rates across subpopulations.

To compensate for the differential representation in SIPP, the Census Bureau constructs weights for all responding units. The weight for each unit is an estimate of the number of units in the target population that the responding unit represents.

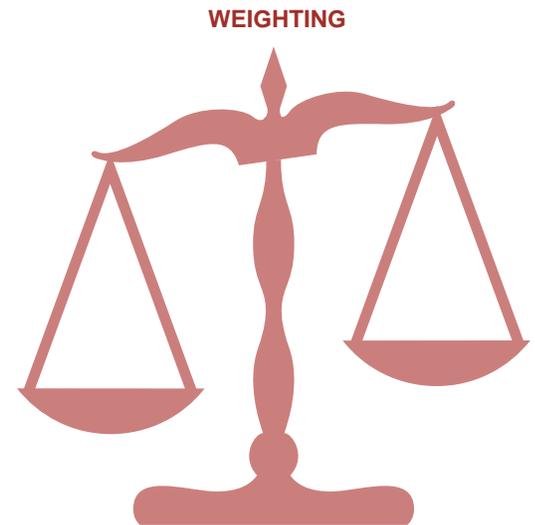
If analysts do not use these weights in their analyses, or if they use them incorrectly, their survey estimates will likely be biased.

Analysts also need to use weights so that they can benchmark their estimates to those of other sources.

## Weights Available in SIPP Files

Each SIPP file contains a number of sets of weights for use in data analysis. The different sets of weights are needed to address the different possible units of analysis and time periods for which survey estimates may be required.

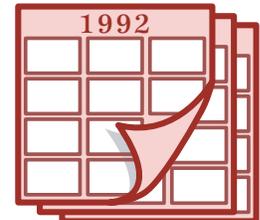
[Link to a table that lists the weight variables in SIPP files for the 1996 and 1990–1993 Panels.](#)



## Choosing Weights

Users must first determine the population of interest in a particular analysis, then select the corresponding set of weights. The weights in the SIPP files are constructed for sample cohorts defined by:

- Month (e.g., the reference month weights in the core wave files and the interview month weights in the pre-1996 topical module files)
- Year (e.g., the calendar year weights in the full panel file)
- Panel (e.g., the full panel weight in the full panel file)



Users can choose to base their analyses on:

- A cross-sectional sample at a given month
- A longitudinal sample that provides continuous monthly data over a year
- A longitudinal sample that provides monthly data over the life of a panel
- A subset of the sample and/or the period in any of the above

Monthly (cross-sectional) weights allow the use of all available data for a given month. For this type of analysis, users can choose among the following units of analysis:

- Person
- Household
- Family
- Related subfamily

Analysts can use SIPP longitudinal samples to follow the same people over time and thus study the dynamics of program participation, lengths of poverty spells, and changes in other circumstances, such as household composition.

The longitudinal weights allow the inclusion of all people for whom data were collected for every month of the period involved (calendar year or full panel). The weights include those who left the target population through death or by moving to ineligible addresses (institutions, foreign living quarters, or military barracks), as well as those for whom data were imputed for missing months.

The Census Bureau makes two types of adjustments to the longitudinal weights:

- Nonresponse adjustments to compensate for panel attrition
- Poststratification adjustments to make the weighted sample totals conform to known population totals for key variables

## ***Using Weights in SIPP Analyses***

Users should consult Chapter 8 and Appendix C of the *SIPP Users' Guide* for a full discussion of how SIPP weights are constructed and used in the core wave, topical module, and full panel files. In this section of the tutorial we highlight only a few issues.

### ***Core Wave Files***

Each core wave file contains reference month weights for persons, households, families, and subfamilies.

For all pre-1996 panels, each core wave file also contains interview month weights for persons and households.

(Interview month weights are not computed for families and related subfamilies.) Beginning with the 1996 Panel, the core wave files no longer provide interview month weights.

In the 1989 and earlier panels, each person's record in a core wave file contained 18 weight variables. For the 1990 and later panels, the file structure was changed to a person-month format (see Chapter 10 of the *SIPP Users' Guide*) and each person-month record has only 6 weights.

### **Topical Module Files**

The topical module files contain one weight variable. Prior to 1996, this weight was the person interview month weight for people who provided data for a topical module. For the 1996 Panel, this weight is the person cross-sectional weight for the fourth reference month.

### **Full Panel Files**

The weight variables in the full panel file are the calendar year weights and the full panel weight.

**Calendar Year Weights.** These weights apply to sample persons who have interviews covering the control date of the corresponding calendar year and who have complete data (either reported or imputed) for every month of the year (excluding months of ineligibility).

People are assigned calendar year weights equal to zero when they do not have interviews covering the control date, have missing data for one or more months of the year, or both.

The number of calendar year weights on the file depends on the duration of the panel. Most panels before the 1996 Panel have two calendar year weights. The exceptions are the 1989 Panel, which has one calendar year weight, and the 1992 Panel, which has three calendar year weights. When the 1996 full panel file is complete, it will have four calendar year weights.

**Panel Weight.** This weight applies to sample persons who are in the sample in Wave 1 of the panel and who have complete data (either reported or imputed) for every month of a panel (excluding months of ineligibility).

People are assigned a panel weight equal to zero if they were not in-sample in Wave 1, have missing data for one or more months of the panel, or both.

Infants born after the beginning of the panel are assigned a panel weight equal to zero. Similarly, infants born after the control date are assigned a calendar year weight equal to zero for that year. 

## **SIPP** *tip*

*The weighting procedures for infants can have important implications for analysts studying young children when infants are a sizable fraction of the population. For example, infants constitute 20 percent of the WIC program population.*

### **Estimation with the Full Panel File**

Analysts can use the full panel files to construct calendar year estimates of quantities, such as total annual income, by extracting records with positive calendar year weights.

Annual estimates computed with the full panel files are based on monthly data from the same person collected at three or four times (depending on the rotation group of the respondent). **tip**

Analysts can also take full advantage of the longitudinal nature of SIPP to construct spell estimates that allow dynamic studies of household composition, labor force activity, health insurance coverage, and welfare reciprocity.

### **SIPP tip**

*The 4-month recall period used by SIPP is generally believed to provide estimates of annual measures with less nonsampling error than estimates derived from surveys that have a 12-month recall period.*