7. Sampling Error

This chapter discusses methods for obtaining the sampling error estimates derived from the Survey of Income and Program Participation (SIPP) panels. The sample selected for each SIPP panel is a stratified multistage probability sample. This complex sample design needs to be taken into account when estimating the variances of SIPP estimates. The SIPP data files contain variables, related to the sample design, that are created for the purpose of variance estimation. Several software packages are now available for computing variance estimates for a wide range of statistics based on complex sample designs. Using the variables that specify the design, these programs can calculate appropriate variances of survey estimates. The Census Bureau also provides generalized variance functions (GVFs) that can be used to obtain approximate estimates of sampling variance for SIPP estimates.

A common mistake in the estimation of sampling error for survey estimates is to ignore the complex survey design and treat the sample as a simple random sample (SRS) of the population. That mistake occurs because most standard software packages for data analyses assume simple random sampling for variance estimation. When applied to SIPP estimates, SRS formulas for variances typically underestimate the true variances. This chapter describes how appropriate variance estimates, which take into account the complex sample design, can be obtained for SIPP estimates.

The topics discussed in this chapter are:

- Direct variance estimation;
- Approximate variance estimates obtained from GVFs;

  and
- Variance estimation when some data are imputed.

Direct Variance Estimation

The primary sampling unit (PSU) plays a key role in variance estimation with a multistage sample design. SIPP PSUs are mostly counties, groups of counties, or independent cities (*SIPP Quality Profile*, 3rd Ed. [U.S. Census Bureau,1998a, Chapter 3]). The PSUs are sampled without replacement so that no PSU is selected more than once for the sample. Some PSUs are sampled with probability proportional to size within strata and usually called nonself-representing PSUs

(NSR PSUs). Other PSUs are so large that they are included in the sample with certainty and therefore are called self-representing PSUs (SR PSUs). Because no sampling is involved, SR PSUs are, in fact, not PSUs but strata. The actual PSUs for those certainty selections are the enumeration districts and other units selected within them. Although the SIPP PSUs are selected without replacement (as is the case with most multistage designs), for the purpose of variance estimation they are treated as if they were sampled with replacement. The with-replacement assumption greatly facilitates variance estimation since it means that variance estimates can be computed by taking into account only the PSUs and strata, without the need to consider the complexities of the subsequent stages of sample selection. This widely used simplifying assumption leads to an overestimation of variances, but the overestimation is not great.

Several software packages are available for computing variances of a wide range of survey estimates (e.g., means and proportions for the total sample and for subclasses, for differences in means and proportions between subclasses, and for regression and logistic regression coefficients) from complex sample designs. Many of these packages are listed on the Web: http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. Lepkowski and Bowles (1996) examined eight of the packages.

These packages use a variety of methods for variance estimation. Some use an approach based on a Taylor-series approximation, or linearization, method. Others use a replication method, such as jackknife repeated replications or balanced repeated replications. Although some methods have advantages in some situations, there is generally little to recommend one method over another. The variance estimates they produce are not identical, but the differences are usually small. See Wolter (1985) and Rust (1985) for discussions of these methods.

Variance Units and Variance Strata, 1990–2004 Panels

For the 1990-2004 SIPP Panels, the sample member record contains information concerning the PSU and stratum within which the member was sampled. This information is needed as input for all of the specialized software packages. The original PSU and strata codes are not included in the SIPP public use data files, however, to avoid potential identification of small geographic areas and sampled individuals. Instead, sets of PSUs are combined across strata to produce variance units and variance strata, with two variance units in each variance stratum. Variance units and variance strata may be treated as PSUs and strata for variance estimation purposes. Their use does not give rise to any bias in the variance estimates. The variance estimates are somewhat less precise, however, than those obtained from the use of the PSUs and strata that have not been combined.

Under the complex sample design, the number of degrees of freedom for variance estimation depends on the number of variance strata. The 1984 SIPP Panel consists of 142 variance units in 71 variance strata; the panels between 1985 and 1991 have 144 variance units and 72 variance strata; the 1992-1993 Panels have 198 variance units and 99 variance strata; the 1996-2001 Panels have 210 variance units and 105 variance strata; and the 2004 Panel has 228 variance units and 114 variance strata. As a rough approximation, the number of degrees of freedom for a variance estimate is the number of variance strata. Thus, for national estimates, the variance estimates have about 71 degrees of freedom for the 1984 Panel, 72 degrees of freedom for the 1985-1991 Panels, and 99 degrees of freedom for the 1992-1993 Panels, 105 degrees of freedom for the 1996-2001 Panels, and 114 degrees of freedom for the 2004 Panel. Regional estimates will have fewer degrees of freedom because such estimates include only some of the variance strata.

Table 7-1 displays the variable names for the variance stratum and variance unit code in the SIPP core wave files and the SIPP full panel files. These codes can be employed as stratum and PSU codes in any of the software packages for variance estimation with complex sample designs.

Table 7-1. Variance Stratum Code and Variance Unit Code in SIPP Files, 1990-2004

Variable for Variance Estimation	**SIPP Core Wave File**		**SIPP Longitudinal File**	
	1990-1993	**1996-2004**	**1990-1993**	**1996-2004**
Variance unit (or half-sample) code	HHSC	GHLFSAM	HALFSAMP	GHLFSAM
Variance stratum code	HSTRAT	GVARSTR	VARSTRAT	GVARSTR

Replication Weights for the SIPP Panels

Analysts should use Fay's method for estimating variances for the SIPP Panels. Fay's method is a modified balanced repeated replication (BRR) method of variance estimation. The difference between the basic BRR method and Fay's method is that the BRR method uses replicate factors of 0 and 2, whereas Fay's method uses one factor, *k*, which is in the range (0, 1), with the other factor equal to 2-*k*. In Fay's method, the introduction of the perturbation factor (1-*k*) allows the use of both halves of the sample. Thus, Fay's method has the advantage that no subset of the sample units in a particular classification will be totally excluded. The variance formula for Fay's method is

$$Var(\theta_0) = \{1/[G(1-k)^2]\} \sum_{i=1}^{G} (\theta_i - \theta_0)^2, \qquad (7\text{-}1)$$

where

G = number of replicates;
$1 - k$ = perturbation factor;
i = replicate, i = 1 to G;
θ_i= i^{th} estimate of the parameter θ based on the observations included in the i^{th} replicate;
θ_0 = survey estimate of the parameter θ based on the full sample.

The 1996 SIPP Panel uses 108 replicate weights, which are calculated on the basis of a perturbation factor of 0.5 ($k = 0.5$). Inserting those values into Equation (7-1) results in the 1996 SIPP Panel variance formula of

$$Var(\theta_0) = \{1/[180 * 0.5^2]\} \sum^{180} (\theta_i - \theta_0)^2.$$

The 2004 SIPP Panel uses 120 replicate weights, which are calculated on the basis of a perturbation factor of 0.5 ($k = 0.5$).

The Census Bureau used VPLX and SAS software to compute the replicate weights that are available through Data FERRET and the SIPP FTP Site.

Using GVFs to Approximate Variance Estimates

The Census Bureau provides two forms for approximate variance estimation: GVFs and tables of standard errors (the square root of the variance) for different estimated numbers and percentages. The generalized estimates provide indications of the magnitude of the sampling error in the survey estimates. They serve as convenient ways to summarize the sampling errors for a broad variety of estimates.

The GVFs for SIPP were derived by modeling the standard error behavior of groups of estimates with similar standard errors. The mathematical form of the function adopted is

$$s = (ax^2 + bx)^{1/2}, \tag{7-2}$$

where s represents the standard error and x the value of an estimate. The parameters a and b are derived on the basis of a selected group of estimates. They are updated annually and are included in the source and accuracy statement that accompanies each SIPP data file for a panel. It is essential to use the parameter estimates for a specific panel and to follow the instructions to apply necessary adjustments to obtain the correct estimates for subgroups. Besides GVFs, the Census Bureau provides summary tables of general standard errors. Those estimates are also available in the source and accuracy statements. The following examples show how to use GVFs to estimate the standard errors of estimated numbers and of sample means. The use of GVFs and tables of standard errors is described in the source and accuracy statements for each panel.

Before looking at the examples, the user should note that the generalized variance estimates for estimating the standard errors of other statistics may not be accurate for small subgroups. Using the 1984 SIPP Panel, Bye and Gallicchio (1989) developed variance functions for participants of Old Age, Survivors, and Disability Insurance (OASDI) and Supplemental Security Income (SSI) programs. They found that for estimates of less than 10 million, the generalized standard error estimates provided by the Census Bureau were 1.20 to 1.75 times larger than those obtained from the variance functions developed specifically for that subgroup.

Using GVFs for Standard Errors of Estimated Numbers

The approximate standard error, *s*, of an estimated number of persons (or households, and families) can be obtained by the formula

$$s = (ax^2 + bx)^{1/2},$$

where *a* and *b* are the parameters associated with the estimate for the particular reference period, and *x* is the weighted estimate. This equation is appropriate for the standard errors of estimated numbers and should not be applied to estimates of dollar values.

Suppose that the number of households with monthly household income above $20,000 is estimated from Wave 1 of the 2004 Panel to be 1,637,500. The approximate values of *a* and *b* from Table 3 of the source and accuracy statement of the 2004 Panel are *a* = -0.00002809, and *b* = 3,153. Then, the standard error, *s*, of this estimated number is given by

$$s = [(-0.00002809 * 1{,}637{,}500^2) + (3{,}153 * 1{,}637{,}500)]^{1/2}. \qquad (7\text{-}3)$$

The approximate 90 percent confidence interval for the estimated number can be computed as $x \pm 1.645 * s$, which ranges from 1,520,165 and 1,754,835. Therefore, a conclusion that the average estimate derived from all possible samples lies within an interval computed in this way would be correct for roughly 90 percent of all samples.

Using GVFs for the Standard Error of a Mean

A mean is defined here to be the average quantity of some characteristic (other than the number of persons or households) per person or household. For example, a mean could be the average monthly household income of females 25 to 54 years of age. The formula used to estimate the standard error of a mean, $\overline{x}$ is

$$s_{\overline{x}} = \sqrt{\frac{b}{y} s^2}, \qquad (7\text{-}4)$$

where *y* is the size on which the estimate is based, s^2 is the estimated population variance of the characteristic, and b is the parameter associated with the particular type of characteristic. Because of the approximations used in developing this formula, an estimate of the standard error of the mean obtained from this formula will generally underestimate the true standard error.

The estimated population mean, $\overline{x}$, and the population variance, s^2, are given by the formulas:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

and

$$s^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \overline{x})^2}{\sum_{i=1}^{n} w_i} \; or \; \frac{\sum_{i=1}^{n} w_i (x_i - \overline{x})^2}{\sum_{i=1}^{n} w_i - 1}$$

where there are n units with the item of interest, and w_i is the final weight for the i^{th} unit. (Note that $\sum^{n} w_i = y$). Suppose that, based on January of the 2004 data of the Wave 1, 2004 Panel, the mean monthly cash household income for females aged 25 to 54 is $5,826, the weighted number of females in this age range is y = 62,346,000, and the population variance is estimated to be 64,900,000. When the appropriate b parameter of 3,153 from Table 3 of the Source and accuracy statement for Panel 2004 is used, the estimated standard error of this mean is

$$s_{\overline{x}} = [(3{,}153 * 64{,}900{,}000)/62{,}346{,}000]^{1/2} = \$57.$$

Thus, the 90 percent confidence interval, computed

$$\overline{x} \pm 1.645 * s_{\overline{x}}.$$

ranges from $5,732 to $5,920. Therefore, a conclusion that the average estimate derived from all possible samples lies within an interval computed in this way would be correct for roughly 90 percent of all samples.

Variance Estimation with Imputed Data

Imputation methods are used to fill in several types of missing data in SIPP. They are used to complete some item nonresponse, person-level nonresponse within households (Type Z nonresponse), and some wave nonresponse (intermittent responses bounded by two responding waves). Imputation fills in gaps in the data set and makes data analyses easier. It also allows more people to be retained as panel members for longitudinal analyses. The concern, however, is that imputation fabricates data to some degree. Treating the imputed values as actual values in estimating the variance of survey estimates leads to an overstatement of the precision of the estimates (Brick and Kalton, 1996). It is important to recognize this fact when sizable proportions of values are imputed.