SOURCE AND ACCURACY STATEMENT

for the Survey of Income and Program Participation[1] from 1996 Public Use Files

SOURCE OF DATA

The data was collected in the 1996 panel of the Survey of Income and Program Participation (SIPP). The SIPP universe is the noninstitutionalized resident population living in the United States. The population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

The 1996 panel of the SIPP sample is located in 322 Primary Sampling Units (PSUs), each consisting of a county or a group of contiguous counties. Within these PSUs, living quarters (LQs) were systematically selected from lists of addresses prepared for the 1990 decennial census to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1990 census, a sample containing clusters of four LQs was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panel.

In jurisdictions that don't issue building permits or have incomplete addresses, we systematically sampled expected clusters of four LQs which were listed by field personnel and then subsampled in the field. In addition, we selected sample LQs from a supplemental frame that included LQs identified as missed in the 1990 census.

For the first interview of the panel, Wave 1, we obtained interviews from occupants of about 36,700 of the 49,200 designated living quarters. We found most of the remaining 12,500 living quarters in the panel to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. However, we did not interview approximately 3,400 of the 12,500 living quarters in the panel because the occupants, (1) refused to be interviewed, (2) could not be found at home, (3) were temporarily absent, or (4) were otherwise unavailable. Thus, occupants of about 92 percent of all eligible living quarters participated in the first interview of the panel.

For subsequent interviews, only original sample persons (those in Wave 1 sample households and interviewed in Wave 1) and persons living with them were eligible to be interviewed. We followed original sample persons if they moved to a new address, unless the new address was more than 100 miles from a SIPP sample area. Then, we attempted telephone interviews.

---

[1]For questions or further assistance with the information provided in this document, contact the Survey of Income and Program Participation Branch of the Demographic Statistical Methods Division on (301) 457-4192 or via the internet using Karen.C.King@ccmail.census.gov

Sample households within a given panel are divided into four random subsamples of nearly equal size. These subsamples are called rotation groups and one rotation group is interviewed each month. Each household in the sample was scheduled to be interviewed at 4 month intervals over a period of roughly 4 years beginning in April 1996. The reference period for the questions is the 4-month period preceding the interview month. In general, one cycle of four interviews covering the entire sample, using the same questionnaire, is called a wave.

The public use files include core and supplemental (**topical module**) data. Core questions are repeated at each interview over the life of the panel. Topical modules include questions which are asked only in certain waves. The 1996 panel topical modules are given in Table 1.

Table 2 indicates the reference months and interview months for the collection of data from each rotation group for the 1996 panel. For example, Wave 1 rotation group 1 of the 1996 panel was interviewed in April 1996 and data for the reference months December 1995 through March 1996 were collected.

Estimation. We used several stages of weight adjustments in the estimation procedure to derive the SIPP cross-sectional person weights. We gave each person a base weight **(BW)** equal to the inverse of probability of selection of a person's household. We applied two noninterview adjustment factors. One adjusted the weights of interviewed persons in interviewed households to account for households which were eligible for the sample but which field representatives could not interview at the first interview (F_{N1}). The second compensated for person noninterviews occurring in subsequent interviews (F_{N2}). We used a Duplication Control Factor **(DCF)** which adjusts for subsampling done in the field when the number of sample units is much larger than expected. We applied a Mover's Weight **(MW)**, which adjusts for persons in the SIPP universe who move into sample households after wave 1. The last weight applied is the Second Stage Adjustment Factor (F_{2s}). This weight adjusts estimates to population controls and causes husbands' and wives' weights to be equal.

The final cross-sectional weight is $\mathbf{Fw_c = BW \times DCF \times F_{n1} \times F_{2S}}$ for wave 1 and is $\mathbf{Fw_c = IW \times F_{n2} \times F_{2S}}$ for waves 2+, where **IW** is either $\mathbf{BW \times DCF \times F_{n1}}$ or **MW**. James (1995) and Siegel (1995a) describe SIPP cross-sectional weighting in greater detail.

Researchers both inside and outside the Census Bureau conducted evaluations of SIPP weighting methodology and researched alternative methodologies. We are making several improvements to SIPP weighting methods beginning with this panel. They are described below.

- We dropped the first stage factor (F_{1s}) from cross-sectional weighting. This factor adjusted for differences between the Census count of population and an estimate of that count based on Census data for sample PSUs. James (1994) found that it did not reduce variance as was previously believed. Jabine, et al (1990) describe the first stage factor used in earlier panels.

- We are using additional variables in nonresponse adjustment. We added high/low poverty stratum code to the Wave 1 nonresponse adjustment, and we added household income, geographic division, and number of imputations for selected income and asset items to the nonresponse adjustment for waves 2+. Research by Rizzo, et al (1994) and by Folsom and Witt (1994) pointed out the potential of the latter three variables in reducing nonresponse bias.

- We redefined nonresponse adjustment cells for waves 2+ weighting. We formed the nonresponse cells by successively partitioning data from five panels by whichever variable most reduced the bias of the household income to poverty threshold ratio. We used data from a sixth panel to evaluate the results. We calculated the nonresponse bias of six variables at waves two and seven for both the new cells and the original cells using initial weights and data from the most recent interview in the calculations. The new cells had lower bias for five of the six variables (Siegel, 1995b).

Research was conducted on a number of promising weighting improvements. Allen and Petroni (1994) reported on an adjustment for mover attrition. Folsom and Witt (1994) and Rizzo, et al (1994) studied alternative nonresponse adjustments using response propensity models. Each study computed weights using an alternative methodology. The researchers then compared estimates of various items to benchmarks. The benchmarks came from administrative records and survey data with less nonresponse than the SIPP. The comparisons did not provide strong evidence of lower bias using the alternative weighting methods.

Additional Methodology

Use of Weights. Each household and each person within each household on each wave tape has four weights. These four weights are reference month specific and therefore can be used only to form reference month estimates. Reference month estimates can be averaged to form estimates of monthly averages over some period of time.

> **Example**, using the proper weights, one can estimate the monthly average number of households in a specified income range over November and December 1996. To estimate monthly averages of a given measure (e.g., total, mean) over a number of consecutive months, sum the monthly estimates and divide by the number of months.

To form an estimate for a particular month, use the <u>reference month</u> weight for the month of interest, summing over all persons or households with the characteristic of interest whose reference period includes the month of interest. Multiply the sum by a factor to account for the number of rotations contributing data for the month. This factor equals four divided by the number of rotations contributing data for the month. For example, December 1995 data is only available from rotations 2, 3, and 4 for Wave 1 of the 1996 panel (See Table 2), so a factor of 4/3 must be applied.

When estimates for months with less than four rotations worth of data are constructed from a wave file, factors greater than 1 must be applied. However, when core data from consecutive waves are used together, data from all four rotations may be available, in which case the factors are equal to 1.

These tapes contain no weight for characteristics that involve a persons's or household's status over two or more months (e.g., number of households with a 50 percent increase in income between November and December 1995).

Producing Estimates for Census Regions and States. The total estimate for a region is the sum of the state estimates in that region. Using this sample, estimates for individual states are subject to very high variance and are not recommended. The state codes on the file are primarily of use in linking respondent characteristics with appropriate contextual variables (e.g., state-specific welfare criteria) and for tabulating data by user-defined groupings of states.

Producing Estimates for the Metropolitan Population. For Washington, DC and 14 other states, metropolitan or non-metropolitan residence is identified (variable H*-METRO). In 28 additional states, where the non-metropolitan population in the sample was small enough to present a disclosure risk, a fraction of the metropolitan sample was recoded to be indistinguishable from non-metropolitan cases (H*-METRO = 2). In these states, therefore, the cases coded as metropolitan (H*-METRO = 1) represent only a subsample of that population.

In producing state estimates for a metropolitan characteristic, multiply the individual, family, or household weights by the metropolitan inflation factor for that state, presented in Table 3. (This inflation factor compensates for the subsampling of the metropolitan population and is 1.0 for the states with complete identification of the metropolitan population.)

The same procedure applies when creating estimates for particular identified MSA's or CMSA's--apply the factor appropriate to the state. For multi-state MSA's, use the factor appropriate to each state part. For example, to tabulate data for the Maine, ME-VT, apply the Vermont factor of 1.57953 to weights for residents of the Vermont part of the MSA; Maine residents require no modification to the weight (i.e., their factors equal 1.57953).

In producing regional or national estimates of the metropolitan population, it is also necessary to compensate for the fact that no metropolitan subsample is identified within two states (Mississippi and West Virginia). Thus, factors in the right-hand column of Table 3 should be used for regional and national estimates. The results of regional and national tabulations of the metropolitan population will be biased slightly. However, less than one-half of one percent of the metropolitan population is not represented.

Producing Estimates for the Non-Metropolitan Population. State, regional, and national estimates of the non-metropolitan population cannot be computed directly, except for Washington, DC and the 13 states where the factor for state tabulations in Table 3 is 1.0. In all other states, the cases identified as not in the metropolitan subsample (METRO = 2) are a mixture of non-metropolitan and metropolitan households. Only an indirect method of estimation is available: first compute an estimate for the total population, then subtract the estimates for the metropolitan population. The results of these tabulations will be slightly biased.

ACCURACY OF ESTIMATES

SIPP estimates are based on a sample; they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaire, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: nonsampling and sampling. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error. Found in the next sections are descriptions of sources of SIPP nonsampling error, followed by a discussion of sampling error, its estimation, and its effect in data analyses.

Nonsampling Error. Nonsampling errors can be attributed to many sources:

- inability to obtain information about all cases in the sample
- definitional difficulties
- differences in the interpretation of questions
- inability or unwillingness on the part of the respondents to provide correct information
- inability to recall information, errors made in the following: collection such as in recording or coding the data, processing the data, estimating values for missing data
- biases resulting from the differing recall periods caused by the interviewing pattern used
- and undercoverage.

Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers. More detailed discussions of the existence and control of nonsampling errors in the SIPP can be found in the <u>SIPP Quality Profile</u> by Thomas B. Jabine, Karen E. King and Rita J. Petroni, issued May 1990.

Undercoverage in SIPP results from missed living quarters and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for nonBlacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group. Further, the independent population controls used have been adjusted for undercoverage in the Census.

A common measure of survey coverage is the coverage ratio, the estimated population before ratio adjustment divided by the independent population control. The Table below shows SIPP coverage ratios for age-sex-race groups for one month-April 1996 prior to the weighting adjustment. The SIPP coverage ratios exhibit some variability from month to month, but these are a typical set of coverage ratios. Other Census Bureau household surveys [like the Current Population Survey] experience similar coverage.

SIPP Coverage Ratios - Age by Nonblack/Black Status and Sex

	NonBlack		Black	
Age	**M**	**F**	**M**	**F**
15	0.9175	1.1235	0.7044	0.7749
16-17	0.8640	0.9289	0.8826	0.9433
18-19	0.8620	0.8647	0.8274	0.8339
20-21	0.8848	0.8041	0.6255	0.9596
22-24	0.7859	0.8692	0.5857	0.6705
25-29	0.8022	0.8254	0.8504	0.8386
30-34	0.8721	0.9063	0.8792	0.7991
35-39	0.9212	0.9855	0.7119	0.8982
40-44	0.9058	0.9321	0.8059	0.9653
45-49	0.9009	0.9761	0.6856	0.7758
50-54	0.9667	0.9181	0.8993	1.2103
60-61	0.8405	0.8961	1.0210	0.9877
62-64	0.9866	1.0698	0.9914	0.9618
65-69	0.9304	0.9423	1.0646	0.7759
70-74	0.8836	0.9362	0.7896	1.3338
75-79	0.8952	1.0046	--------	0.9104
80-84	0.8974	0.9651	--------	--------
85 +	0.9558	0.9669	--------	--------

These coverage ratios are for April 1996.

Comparability with Other Estimates. Caution should be exercised when comparing data from this with data from other SIPP products or with data from other surveys. The comparability problems are caused by such sources as the seasonal patterns for many characteristics, different nonsampling errors, and different concepts and procedures. Refer to the SIPP Quality Profile for known differences with data from other sources and further discussions.

Sampling Error. Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample rather than the entire population was surveyed.

USES AND COMPUTATION OF STANDARD ERRORS

Confidence Intervals. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average result of all possible samples.

2. Approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

3. Approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Hypothesis Testing. Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population characteristics using sample estimates. The most common types of hypotheses tested are 1) the population characteristics are identical versus 2) they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

To perform the most common test, compute the difference $X_A - X_B$, where X_A and X_B are sample estimates of the characteristics of interest. A later section explains how to derive an estimate of the standard error of the difference $X_A - X_B$. Let that standard error be S_{DIFF}. If $X_A - X_B$ is between -1.6 times S_{DIFF} and +1.6 times S_{DIFF}, no conclusion about the characteristics is justified at the 10 percent significance level. If, on the other hand, $X_A - X_B$ is smaller than -1.6 times S_{DIFF} or larger than +1.6 times S_{DIFF}, the observed difference is significant at the 10 percent level. In this event, it is commonly accepted practice to say that the characteristics are different. Of course, sometimes this conclusion will be wrong. When the characteristics are the same, there is a 10 percent chance of concluding that they are different.

Note that as more tests are performed, more erroneous significant differences will occur. For example, at the 10 percent significance level, if 100 independent hypothesis tests are performed in which there are no real differences, it is likely that about 10 erroneous differences will occur. Therefore, the significance of any single test should be interpreted cautiously.

Note Concerning Small Estimates and Small Differences. Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000. Care must be taken in the interpretation of small differences since even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Calculating Standard Errors for SIPP Estimates. There are three main ways we calculate the Standard Errors for SIPP Estimates. They are as follows:

- Replicate Weighting Methods,
- Generalized Variance parameters (denoted as "a" and "b"),
- Simplified tables using the "a" and "b" parameters.
  The most reliable method is the Replicate Weighting Method. SIPP uses the Replicate Weighting Method to produce Generalized Variance parameters. Using the Generalized Variance parameters, we create simplified tables.

Standard Error Parameters and Tables and Their Use. Most SIPP estimates have greater standard errors than those obtained through a simple random sample because PSUs are sampled and clusters of living quarters are sampled for the SIPP in the area and new construction frames. To derive standard errors that would be applicable to a wide variety of estimates and could be prepared at a moderate cost, a number of approximations were required. Estimates with similar standard error behavior were grouped together and two parameters (denoted "a" and "b") were developed to approximate the standard error behavior of each group of estimates. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. These "a" and "b" parameters vary by characteristic and by demographic subgroup to which the estimate applies. Table 4 provides base "a" and "b" parameters to be used for the 1996 panel estimates. Table 10 provides parameters for calculating 1996 topical module variances.

The factors provided in Table 5 when multiplied by the base parameters of Table 4 for a given subgroup and type of estimate give the "a" and "b" parameters for that subgroup and estimate type for the specified reference period. For example, the base "a" and "b" parameters for total number of households are -0.00002480 and 2,474, respectively. For Wave 1 the factor for March 1996 is 1 since 4 rotation months of data is available. So, the "a" and "b" parameters for total household income in March 1996 based on Wave 1 are -0.00002480 and 2,474, respectively. Also for Wave 1, the factor for the first quarter of 1996 is 1.2222 since 9 rotation months of data are available (rotations 1 and 2 provide 3 rotations months each, while rotations 3 and 4 provide 1 and 2 rotation months, respectively). So the "a" and "b" parameters for total number of households in the first quarter of 1992 are -0.00003031 and 3,024, respectively for Wave 1.

The "a" and "b" parameters may be used to calculate the standard error for estimated numbers and percentages. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. Methods for using these parameter for computation of approximate standard errors are given in the following sections.

For those users who wish further simplification, we have also provided general standard errors in Tables 6 through 9. Note that these standard errors only apply when data from all four rotations are used and must be adjusted by a factor from Table 4. The standard errors resulting from this simplified approach are less accurate. Methods for using these parameters and tables for computation of standard errors are given in the following sections.

The procedures described below apply only to reference month estimates or averages of reference month estimates. Refer to the section "Use of Weights" for a more detailed discussion of the construction of estimates.

Variance stratum codes and half sample codes are included on the tapes to enable the user to compute the variances directly and more accurately by methods such as balanced repeated replications (BRR). William G. Cochran provides a list of references discussing the application of this technique. (See Sampling Techniques, 3rd Ed., New York: John Wiley and Sons, 1977, p. 321.)

Standard errors of estimated numbers. The approximate standard error, s_x, of an estimated number of persons, households, families, unrelated individuals and so forth, can be obtained in two ways. Both apply when data from all four rotations are used to make the estimate. However, only the second method should be used when less than four rotations of data are available for the estimate. Note that neither method should be applied to dollar values.

The standard error may be obtained by the use of the formula

$$s_x = fs \qquad (1)$$

where f is the appropriate "f" factor from Table 4, and s is the standard error on the estimate obtained by interpolation from Table 6 or 7. Alternatively, s_x may be approximated by the formula

$$s_x = \sqrt{ax^2 + bx} \qquad (2)$$

from which the standard errors in Tables 8 and 9 were calculated. Here x is the size of the estimate and "a" and "b" are the parameters associated with the particular type of characteristic being estimated. Use of formula 2 will provide more accurate results than the use of formula 1.

Illustration.

Suppose SIPP estimates for Wave 1 of the 1996 panel show that there were 1,700,000 black households with monthly household income above $4,000. The appropriate parameters and factor from Table 4 and the appropriate general standard error from Table 6 are

a = -0.00018540 b = 2,160 f = 0.61 s = 117,000

Using formula 1, the approximate standard error is

$s_x = 71,370$

Using formula 2, the approximate standard error is

$$\sqrt{(-0.00018540)(1,700,000)^2 + (2,160)(1,700,000)} = 56,002$$

Using the standard error based on formula 2, the approximate 90-percent confidence interval as shown by the data is from 1,610,397 to 1,789,603. Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90% of all samples.

Standard Error of a Mean. A mean is defined here to be the average quantity of some item (other than persons, families, or households) per person, family or household. For example, it could be the average monthly household income of females age 25 to 34. The standard error of a mean can be approximated by formula 3 below. Because of the approximations used in developing formula 3, an estimate of the standard error of the mean obtained from this formula will generally underestimate the true standard error. The formula used to estimate the standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\left(\frac{b}{y}\right) s^2} \qquad (3)$$

where y is the size of the base, s^2 is the estimated population variance of the item and b is the parameter associated with the particular type of item.

The population variance s^2 may be estimated by one of two methods. In both methods we assume x_i is the value of the item for unit I. (Unit may be person, family, or household). To use the first method, the range of values for the item is divided into c intervals. The upper and lower boundaries of interval j are Z_{j-1} and Z_j, respectively. Each unit is placed into one of c groups such that $Z_{j-1} < x_i \leq Z_j$.

The estimated population variance, s^2, is given by the formula:

$$s^2 = \sum_{j=1}^{c} p_j m_j^2 - \bar{x}^2, \qquad (4)$$

where p_j is the estimated proportion of units in group j, and $m_j = (Z_{j-1} + Z_j/2)$. The most representative value of the item in group j is assumed to be m_j. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate value for m_c is

$$m_c = \frac{3}{2} Z_{c-1}.$$

The mean, $\bar{x}$ can be obtained using the following formula:

$$\bar{x} = \sum_{j=1}^{c} p_j m_j.$$

In the second method, the estimated population variance is given by

$$s^2 = \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \bar{x}^2 , \qquad (5)$$

where there are n units with the item of interest and w_i is the final weight for unit I. The mean, $\bar{x}$, can be obtained from the formula

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

<u>Illustration.</u>

Suppose that based on Wave 1 data, the distribution of monthly cash income for persons age 25 to 34 during the month of January 1996 is given in Table 11.

Using formula 4 and the mean monthly cash income of \$2,530 the approximate population variance, s^2, is

$$s^2 = \left(\frac{1,371}{39,851}\right)(150)^2 + \left(\frac{1,651}{39,851}\right)(450)^2 + \ldots\ldots + \left(\frac{1,493}{39,851}\right)(9,000)^2 - (2,530)^2 = 3,159,887.$$

Using formula 3 and the appropriate base "b" parameter from Table 4, the estimated standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\left(\frac{3,476}{39,851,000}\right)(3,159,887)} = \$16.60$$

Standard error of an aggregate. An aggregate is defined to be the total quantity of an item summed over all the units in a group. The standard error of an aggregate can be approximated using formula 6.

As with the estimate of the standard error of a mean, the estimate of the standard error of an aggregate will generally underestimate the true standard error. Let y be the size of the base, s^2 be the estimated population variance of the item obtained using formula (4) or (5) and b be the parameter associated with the particular type of item. The standard error of an aggregate is:

$$s_x = \sqrt{(b)\,(y)s^2} \qquad (6)$$

Standard Errors of Estimated Percentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which the percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more, e.g., the percent of people employed is more reliable than the estimated number of people employed. When the numerator and denominator of the percentage have different parameters, use the parameter (and appropriate factor) of the numerator. If proportions are presented instead of percentages, note that the standard error of a proportion is equal to the standard error of the corresponding percentage divided by 100.

There are two types of percentages commonly estimated. The first is the percentage of persons, families or households sharing a particular characteristic such as the percent of persons owning their own home. The second type is the percentage of money or some similar concept held by a particular group of persons or held in a particular form. Examples are the percent of total wealth held by persons with high income and the percent of total income received by persons on welfare.

For the percentage of persons, families, or households, the approximate standard error, $s_{(x,p)}$, of the estimated percentage p can be obtained by the formula

$$s_{(x,p)} = fs \qquad (7)$$

when data from all four rotations are used to estimate p.

In this formula, f is the appropriate "f" factor from Table 6 and s is the standard error of the estimate from Table 10 or 11.

Alternatively, it may be approximated by the formula

$$s_{(x,p)} = \sqrt{\frac{b}{x}\,(p)\,(100-p)} \qquad (8)$$

from which the standard errors in Tables 10 and 11 were calculated. Here x is the size of the subclass of social units which is the base of the percentage, p is the percentage ($0 < p < 100$), and b is the parameter associated with the characteristic in the numerator. Use of this formula will give more accurate results than use of formula 7 above and should be used when data from less than four rotations are used to estimate p.

Illustration.

Suppose that, in the month of January 1996, 6.7 percent of the 16,812,000 persons in nonfarm households with a mean monthly household cash income of $4,000 to $4,999, were black. Using formula 8 and the "b" parameter of 5,053 from Table 4 and a factor of 1 for the month of January 1996 from Table 7, the approximate standard error is

$$\sqrt{\frac{4{,}611}{(16{,}812{,}000)}\,(6.7)\,(100-6.7)} = 0.41\ \textit{percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 6.3 to 7.1 percent.

For percentages of money, a more complicated formula is required. A percentage of money will usually be estimated in one of two ways. It may be the ratio of two aggregates:

$$p_I = 100\,(X_A / X_N)$$

or it may be the ratio of two means with an adjustment for different bases:

$$p_I = 100\,(\hat{p}_A\,\bar{X}_A / \bar{X}_N)$$

where x_A and x_N are aggregate money figures, $\bar{x}_A$ and $\bar{x}_N$ are mean money figures, and $\hat{p}_A$ is the estimated number in group A divided by the estimated number in group N. In either case, we estimate the standard error as

$$s_I = \sqrt{\left(\frac{\hat{p}_A \bar{x}_A}{\bar{x}_N}\right)^2 \left[\left(\frac{s_p}{\hat{p}_A}\right)^2 + \left(\frac{s_A}{\bar{x}_A}\right)^2 + \left(\frac{s_B}{\bar{x}_N}\right)^2\right]} \quad , \tag{9}$$

where s_p is the standard error of $\hat{p}_A$, s_A is the standard error of $\bar{x}_A$ and s_B is the standard error of $\bar{x}_N$. To calculate s_p, use formula 8. The standard errors of $\bar{x}_N$ and $\bar{x}_A$ may be calculated using formula 3.

It should be noted that there is frequently some correlation between $\hat{p}_A$, $\bar{x}_N$, and $\bar{x}_A$. Depending on the magnitude and sign of the correlations, the standard error will be over or underestimated.

Illustration.

Suppose that in January 1996, 9.8% of the households own rental property, the mean value of rental property is \$72,121, the mean value of assets is \$78,734, and the corresponding standard errors are 0.31%, \$5799, and \$2867. In total there are 86,790,000 households. Then, the percent of all household assets held in rental property is

$$= 100\left((0.098)\frac{72121}{78734}\right) = 9.0\%$$

Using formula (9), the appropriate standard error is

$$s_I = \sqrt{\left(\frac{(0.098)(72121)}{78734}\right)^2 \left[\left(\frac{0.0031}{0.098}\right)^2 + \left(\frac{5799}{72121}\right)^2 + \left(\frac{2867}{78734}\right)^2\right]}$$

$= 0.008$

$= 0.8\%$

Standard Error of a Difference. The standard error of a difference between two sample estimates is approximately equal to

$$s_{(x-y)} = \sqrt{s_x^2 + s_y^2} \tag{10}$$

where s_x and s_y are the standard errors of the estimates x and y. The estimates can be numbers, percents, ratios, etc. The above formula assumes that the correlation coefficient between the characteristics estimated by x and y is zero. If the correlation is really positive (negative), then this assumption will tend to cause overestimates (underestimates) of the true standard error.

Illustration.

Suppose that SIPP estimates show the number of persons age 35-44 years with monthly cash income of $4,000 to $4,999 was 3,186,000 in the month of January 1996 and the number of persons age 25-34 years with monthly cash income of $4,000 to $4,999 in the same time period was 2,619,000. Then, using parameters from Table 4 and formula 2, the standard errors of these numbers are approximately 104,414 and 94,801, respectively. The difference in sample estimates is 9,439 and using formula 10, the approximate standard error of the difference is

$$\sqrt{(104,414)^2 + (94,801)^2} = 95,371$$

Suppose that it is desired to test at the 10 percent significance level whether the number of persons with monthly cash income of $4,000 to $4,999 was different for persons age 35-44 years than for persons age 25-34 years. To perform the test, compare the difference of 9,439 to the product 1.6 x 95,371 = 152,594. Since the difference is less than 1.6 times the standard error of the difference, the data show that the two age groups are not significantly different at the 10 percent significance level.

Standard Error of a Median. The median quantity of some item such as income for a given group of persons, families, or households is that quantity such that at least half the group have as much or more and at least half the group have as much or less. The sampling variability of an estimated median depends upon the form of the distribution of the item as well as the size of the group. To calculate standard errors on medians, the procedure described below may be used.

An approximate method for measuring the reliability of an estimated median is to determine a confidence interval about it. (See the section on sampling variability for a general discussion of confidence intervals.) The following procedure may be used to estimate the 68-percent confidence limits and hence the standard error of a median based on sample data.

1. Determine, using either formula 7 or formula 8, the standard error of an estimate of 50 percent of the group.

2. Add to and subtract from 50 percent the standard error determined in step 1.

3. Using the distribution of the item within the group, calculate the quantity of the item such that the percent of the group with more of the item is equal to the smaller percentage found in step 2. This quantity will be the upper limit for the 68-percent confidence interval. In a similar fashion, calculate the quantity of the item such that the percent of the group with more of the item is equal to the larger percentage found in step 2. This quantity will be the lower limit for the 68-percent confidence interval.

4. Divide the difference between the two quantities determined in step 3 by two to obtain the standard error of the median.

To perform step 3, it will be necessary to interpolate. Different methods of interpolation may be used. The most common are simple linear interpolation and Pareto interpolation. The appropriateness of the method depends on the form of the distribution around the median. If density is declining in the area, then we recommend Pareto interpolation. If density is fairly constant in the area, then we recommend linear interpolation. Note, however, that Pareto interpolation can never be used if the interval contains zero or negative measures of the item of interest. Interpolation is used as follows. The quantity of the item such that "p" percent have more of the item is

$$X_{pN} \;'\; \exp\left[\left(Ln\left(\frac{pN}{N_1}\right) \;/\; Ln\left(\frac{N_2}{N_1}\right)\right) \quad Ln\left(\frac{A_2}{A_1}\right)\right] A_1 \tag{11}$$

if Pareto Interpolation is indicated and

$$X_{pN} \;'\; \left[\frac{PN\&N_1}{N_2\&N_1} \quad (A_2\&A_1) \;\%\; A_1\right] \tag{12}$$

if linear interpolation is indicated, where

N	is the size of the group,
A_1 and A_2	are the lower and upper bounds, respectively, of the interval in which X_{pN} falls,
N_1 and N_2	are the estimated number of group members owning more than A_1 and A_2, respectively,
exp	refers to the exponential function and
Ln	refers to the natural logarithm function.

Illustration.

To illustrate the calculations for the sampling error on a median, we return to Table 14. The median monthly income for this group is \$2,158. The size of the group is 39,851,000.

1. Using formula 8, the standard error of 50 percent on a base of 39,851,000 is about 0.6 percentage points.

2. Following step 2, the two percentages of interest are 49.4 and 50.6.

3. By examining Table 14, we see that the percentage 49.4 falls in the income interval from 2000 to 2499. (Since 55.5% receive more than \$2,000 per month, the dollar value corresponding to 49.4 must be between \$2,000 and \$2,500). Thus, A_1 = \$2,000, A_2 = \$2,500, N_1 = 22,106,000, and N_2 = 16,307,000.

In this case, we decided to use Pareto interpolation. Therefore, the upper bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(Ln\left(\frac{(.494)(39{,}851{,}000)}{22{,}106{,}000}\right) / Ln\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right) Ln\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2177$$

Also by examining Table 11, we see that 50.6 falls in the same income interval. Thus, A_1, A_2, N_1 and N_2 are the same. We also use Pareto interpolation for this case. So the lower bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(Ln\left(\frac{(.506)(39{,}851{,}000)}{22{,}106{,}000}\right) / Ln\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right) Ln\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2139$$

Thus, the 68-percent confidence interval on the estimated median is from \$2139 to \$2177. An approximate standard error is

$$\frac{\$2177 - \$2139}{2} = \$19$$

Standard Errors of Ratios of Means and Medians. The standard error for a ratio of means or medians is approximated by:

$$s_{\frac{x}{y}} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2\right]} \qquad (13)$$

where x and y are the means or medians, and s_x and s_y are their associated standard errors. Formula 13 assumes that the means are not correlated. If the correlation between the population means estimated by x and y are actually positive (negative), then this procedure will tend to produce overestimates (underestimates) of the true standard error for the ratio of means.

Table 1. 1996 Panel Topical Modules

Wave	Topical Module
1	Recipiency History and Employment History
2	Work Disability; Education & Training; Marital; Migration; and Fertility Histories; and Household Relationships
3	Eligibility and Assets & Liabilities
4	Annual Income & Retirement Accounts; Taxes; Work Schedule; and Child Care
5	School Enrollment & Financing; Child Support; Support for Non-Household Members; Disability; and variable modules to be determined
6	Eligibility and Well-Being
7	Annual Income & Retirement Accounts; Taxes; and Retirement & Pension Plan Coverage
8	Variable modules to be determined
9	Eligibility and Assets & Liabilities
10	Annual Income & Retirement Accounts; Taxes; Work Schedule; and Child Care
11	Child Support; Support for Non-Household Members; Disability; and variable modules to be determined
12	Eligibility; and variable modules to be determined

Table 2. Reference Months for Each Interview Month - 1996 Panel

Reference Period

Month of Interview	Wave/ Rotation	1st Quarter (1996)			2nd Quarter (1996)			3rd Quarter (1996)			4th Quarter (1996)			…	3rd Quarter (1999)			4th Quarter (1999)		
		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec		Jul	Aug	Sep	Oct	Nov	Dec
Apr 96	1/1	X	X	X																
May	1/2	X	X	X	X															
June	1/3		X	X	X	X														
July	1/4			X	X	X	X													
Aug	2/1				X	X	X	X												
Sept	2/2					X	X	X	X											
Oct	2/3						X	X	X	X										
Nov	2/4							X	X	X	X									
Dec	3/1								X	X	X	X								
Jan 97	3/2									X	X	X	X							
Feb	3/3										X	X	X							
.																				
.														…						
.														…	.					
Aug 99	11/1																			
Sept	11/2																			
Oct	11/3														X	X	X			
Nov	11/4														X	X	X	X		
Dec	12/1															X	X	X	X	
Jan	12/2													. . .			X	X	X	X
Feb	12/3																	X	X	X
Mar 2000	12/4																		X	X

Table 3. Metropolitan Subsample Factors to be Applied to Compute National and Subnational Estimates

		Factors for use in State or CMSA (MSA) Tabulations	Factors for use in Regional or National Tabulations
Northeast:	Connecticut	1.00000	1.00000
	Maine	1.57953	0.65171
	Massachusetts	1.03252	1.03252
	New Hampshire	1.24580	1.24580
	New Jersey	1.00000	1.00000
	New York	1.00000	1.00000
	Pennsylvania	1.00000	1.00000
	Rhode Island	1.00000	1.00000
	Vermont	1.57953	0.65171
Midwest:	Illinois	1.00735	1.00735
	Indiana	1.00000	1.00000
	Iowa	1.30446	1.30446
	Kansas	1.16632	1.16632
	Michigan	1.02281	1.02281
	Minnesota	1.06701	1.06701
	Missouri	1.00000	1.00000
	Nebraska	1.30873	1.30873
	North Dakota	---	---
	Ohio	1.00000	1.00000
	South Dakota	---	---
	Wisconsin	1.00908	1.0098
South:	Alabama	1.07631	1.07631
	Arkansas	1.28386	1.28386
	Delaware	1.49701	1.49701
	D.C.	1.00000	1.00000
	Florida	1.01184	1.01184
	Georgia	1.01513	1.01513
	Kentucky	1.07446	1.07446
	Louisiana	1.06406	1.06406
	Maryland	1.00000	1.00000
	Mississippi	---	---
	North Carolina	1.00000	1.00000
	Oklahoma	1.07759	1.07759
	South Carolina	1.08096	1.08096
	Tennessee	1.00980	1.00980
	Texas	1.01112	1.01112
	Virginia	1.01554	1.01554
	West Virginia	---	---

- indicates no metropolitan subsample is identified for the state

Table 3.cont'd. Metropolitan Subsample Factors to be Applied to Compute National and Subnational Estimates

		Factors for use in State or CMSA (MSA) Tabulations	Factors for use in Regional or National Tabulations
West:	Alaska	---	---
	Arizona	1.02596	1.02596
	California	1.00000	1.00000
	Colorado	1.13327	1.13327
	Hawaii	1.00000	1.00000
	Idaho	---	---
	Montana	---	---
	Nevada	1.00000	1.00000
	New Mexico	1.66611	1.66611
	Oregon	1.03327	1.03327
	Utah	1.00000	1.00000
	Washington	1.03799	1.03799
	Wyoming	---	---

- indicates no metropolitan subsample is identified for the state

Table 4: SIPP Indirect Generalized Variance Parameters for the 1996 Panel

Characteristics		**Parameters**			
PERSONS		**a**	**b**	**DEFF**	**f**
Poverty and Program Participation		-0.00002071	4,241	1.80	0.72
	Male	-0.00004305	4,241	1.80	0.72
	Female	-0.00003999	4,241	1.80	0.72
Income and Labor Force		-0.00001697	3,476	1.47	0.65
	Male	-0.00003528	3,476	1.47	0.65
	Female	-0.00003278	3,476	1.47	0.65
Other (Person) Items		-0.00002073	5,479	2.32	0.82
	Male	-0.00004245	5,479	2.32	0.82
	Female	-0.00004053	5,479	2.32	0.82
Black (Person) Items		-0.00013740	4,611	1.95	0.75
	Male	-0.00029645	4,611	1.95	0.75
	Female	-0.00025609	4,611	1.95	0.75
Hispanic (Person) Items		-0.00026708	5,746	2.43	0.84
	Male	-0.00052410	5,746	2.43	0.84
	Female	-0.00054462	5,746	2.43	0.84
Metro/NonMetro (Person) Items		-0.00003100	8,191	3.47	1.00
	Male	-0.00006347	8,191	3.47	1.00
	Female	-0.00006059	8,191	3.47	1.00
Poverty and Program Participation Demographic Person Items (age/race/sex/marital status)		-0.00001361	2,788	1.18	0.58
	Male	-0.00002830	2,788	1.18	0.58
	Female	-0.00002629	2,788	1.18	0.58
HOUSEHOLDS					
Total or White		-0.00002480	2,474	1.05	0.65
Black		-0.00018540	2,160	0.92	0.61
Hispanic		-0.00041675	2,968	1.26	0.72
Metro/NonMetro		-0.00005798	5,783	2.45	1.00

Note 1: For Wave 4 and beyond, to account for sample attrition, multiply the a and b parameters by 1.06 for estimates which include data.

Use the "Other (Person) Items" parameters for tabulations of persons 15+ in the labor force, retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of person tabulations not specifically covered by another characteristic in this Table.

Table 5. Factors to be Applied to Table 6 Base Parameters to Obtain Parameters for Various Reference Periods

# of available rotation months[1]	factor
Monthly estimate	
1	4.0000
2	2.0000
3	1.3333
4	1.0000
1st Quarter 1996 to 4th Quarter 2000	1.000

Note 1: The number of available rotation months for a given estimate is the sum of the number of rotations available for each month of the estimate.

Table 6. Standard Errors of Estimated Numbers of Households, Families, or Unrelated Persons (Numbers in Thousands)

Size of Estimate	Standard Error*	Size of Estimate	Standard Error
200	34	25,000	329
300	42	30,000	348
500	54	40,000	372
750	66	50,000	380
1,000	76	60,000	372
2,000	106	70,000	347
3,000	130	75,000	328
5,000	166	80,000	303
7,500	200	90,000	225
10,000	228	95,000	162
15,000	271	99,500	37

*** To account for sample attrition, multiply the standard error of the estimate by 1.06 for estimates which include data from Wave 4 and beyond.**

Table 7. Standard Errors of Estimated Numbers of Persons (Numbers in Thousands)

Size of Estimate	Standard Error*	Size of Estimate	Standard Error
200	40	90,000	697
300	50	100,000	714
500	64	110,000	725
750	78	120,000	732
1,000	90	130,000	735
2,000	128	140,000	734
3,000	156	150,000	729
5,000	200	160,000	719
7,500	244	170,000	705
10,000	281	180,000	686
15,000	340	190,000	661
25,000	431	200,000	631
30,000	467	210,000	594
40,000	527	220,000	549
50,000	576	230,000	494
60,000	616	240,000	425
70,000	649	250,000	332
75,000	663	260,000	185
80,000	676	264,000	43

*** To account for sample attrition, multiply the standard error of the estimate by 1.06 for estimates which include data from Wave 4 and beyond.**

Table 8. Standard Errors of Estimated Percentages of Households, Families, or Unrelated Persons

Base of Estimated Percentage (Thousands)	Estimated Percentages*					
	< =1 or > =9	2 or 98	5 or 95	10 or 90	25 or 75	50
200	1.69	2.38	3.71	5.10	7.36	8.50
300	1.38	1.94	3.03	4.17	6.01	6.94
500	1.07	1.51	2.34	3.23	4.66	5.38
750	0.87	1.23	1.91	2.63	3.80	4.39
1,000	0.76	1.06	1.66	2.28	3.29	3.80
2,000	0.54	0.75	1.17	1.61	2.33	2.69
3,000	0.44	0.61	0.96	1.32	1.90	2.20
5,000	0.34	0.48	0.74	1.02	1.47	1.70
7,500	0.28	0.39	0.61	0.83	1.20	1.39
10,000	0.24	0.34	0.52	0.72	1.04	1.20
15,000	0.20	0.27	0.43	0.59	0.85	0.98
25,000	0.15	0.21	0.33	0.46	0.66	0.76
30,000	0.14	0.19	0.30	0.42	0.60	0.69
40,000	0.12	0.17	0.26	0.36	0.52	0.60
50,000	0.11	0.15	0.23	0.32	0.47	0.54
60,000	0.10	0.14	0.21	0.29	0.43	0.49
70,000	0.09	0.13	0.20	0.27	0.39	0.45
75,000	0.09	0.12	0.19	0.26	0.38	0.44
80,000	0.08	0.12	0.19	0.26	0.37	0.43
90,000	0.08	0.11	0.17	0.24	0.35	0.40
95,000	0.08	0.11	0.17	0.23	0.34	0.39
99,500	0.08	0.11	0.17	0.23	0.33	0.38

*** To account for sample attrition, multiply the standard error of the estimate by 1.06 for estimates which include data from Wave 4 and beyond.**

Table 9. Standard Errors of Estimated Percentages of Persons

Base of Estimated Percentage (Thousands)	Estimated Percentages*					
	< =1 or > =9	2 or 98	5 or 95	10 or 90	25 or 75	50
200	2.01	2.83	4.41	6.07	8.76	10.12
300	1.64	2.31	3.60	4.96	7.15	8.26
600	1.16	1.64	2.55	3.51	5.06	5.84
1,000	0.90	1.27	1.97	2.72	3.92	4.53
2,000	0.64	0.90	1.39	1.92	2.77	3.20
5,000	0.40	0.57	0.88	1.21	1.75	2.02
7,500	0.33	0.46	0.72	0.99	1.43	1.65
10,000	0.28	0.40	0.62	0.86	1.24	1.43
15,000	0.23	0.33	0.51	0.70	1.01	1.17
20,000	0.20	0.28	0.44	0.61	0.88	1.01
25,000	0.18	0.25	0.39	0.54	0.78	0.91
30,000	0.16	0.23	0.36	0.50	0.72	0.83
50,000	0.13	0.18	0.28	0.38	0.55	0.64
75,000	0.10	0.15	0.23	0.31	0.45	0.52
100,000	0.09	0.13	0.20	0.27	0.39	0.45
125,000	0.08	0.11	0.18	0.24	0.35	0.40
150,000	0.07	0.10	0.16	0.22	0.32	0.37
200,000	0.06	0.09	0.14	0.19	0.28	0.32
225,000	0.06	0.08	0.13	0.18	0.26	0.30
250,000	0.06	0.08	0.12	0.17	0.25	0.29
260,000	0.06	0.08	0.12	0.17	0.24	0.28
264,000	0.06	0.08	0.12	0.17	0.24	0.28

*** To account for sample attrition, multiply the standard error of the estimate by 1.06 for estimates which include data from Wave 4 and beyond.**

Table 10. 1996 Wave 1 Topical Module Generalized Variance Parameters

	a	b
Employment History		
Both Sexes 18+	-0.00001632	3,476
Males 18+	-0.00003392	3,476
Females 18+	-0.00003152	3,476
Recipiency History		
Both Sexes 18+	-0.00001991	4,241
Males 18+	-0.00004139	4,241
Females 18+	-0.00003845	4,241

Use the "15+ Income and Labor Force" core parameter for tabulations of reasons for not working/reservation wage and work related income.

Table 11. Distribution of Monthly Cash Income Among Persons 25 to 34 Years Old

	Total	under $300	$300 to $599	$600 to $899	$900 to $1,199	$1,200 to $1,499	$1,500 to $1,999	$2,000 to $2,499	$2,500 to $2,999	$3,000 to $3,499	$3,500 to $3,999	$4,000 to $4,999	$5,000 to $5,999	$6,000 and over
Thousands in interval	39,85	1371	165	225	2734	3452	6278	5799	4730	3723	2519	2619	1223	1493
Percent with at least as much as lower bound of interval	--	100.0	96.6	92.4	86.7	79.9	71.2	55.5	40.9	29.1	19.7	13.4	6.8	3.7