Appendix C. Source and Reliability of Estimates

SOURCE OF DATA

The data were obtained in the second and third interview waves of the 1984 panel of the Survey of Income and Program Participation (SIPP). The SIPP universe is the noninstitutionalized resident population living in the United States. This 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. Similarly, 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 these qualifications persons who were at least 15 years of age at the time of interview were eligible to be in the survey.

The 1984 SIPP sample is located in 174 areas comprising 450 counties (including one partial county) and independent cities. Within these areas, the bulk of the sample consisted of clusters of two to four living quarters (LQ's), systematically selected from lists of addresses prepared for the 1970 decennial census. The sample was updated to reflect new construction through March 1983.

Approximately 26,000 living quarters were designated for the sample. For wave 1, interviews were obtained from the occupants of about 19,900 of the designated living quarters. Most of the remaining 6,100 living quarters were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. However, approximately 1,000 of the 6,100 living quarters were not interviewed because the occupants refused to be interviewed, could not be found at home, were temporarily absent, or were otherwise unavailable. Thus, occupants of about 95 percent of all eligible living quarters participated in wave 1 of the survey.

For the subsequent waves, only original sample persons (those interviewed in the first wave) and persons living with them were eligible to be interviewed. With certain restrictions, original sample persons were to be followed even if they moved to a new address. All noninterviewed households from wave 1 were automatically designated as noninterviews for all subsequent waves. When original sample persons moved without leaving a forwarding address or moved to extremely remote parts of the country, additional noninterviews resulted.

Tabulations in this report were drawn from interviews conducted from February to July 1984. February to May interviews come from the second time in sample. June and July interviews result from the third time in sample for the respondents. Table C-1 summarizes information on nonresponse for the first three times in sample.

Table C-1. Sample Size, by Sample Wave and Interview Status

Sample wave	Household units eligible			
			Not interviewed (cumulative)	
	Total	Inter-viewed	Number	Non-response rate
Wave 1.........	20,900	19,900	1,000	4.8
Wave 2.........	21,500	19,400	2,100	9.8
Wave 3.........	22,000	19,000	2,900	13.2

The estimation procedure used to derive SIPP person weights involved several stages of weight adjustments. In the first wave, each person received a base weight equal to the inverse of his/her probability of selection. For each subsequent interview, each person received a base weight that accounted for differences in the probability of selection caused by the following of movers.

A noninterview adjustment factor was applied to the weight of each interviewed person to account for persons in occupied living quarters who were eligible for the sample but were not interviewed. (Individual nonresponse within partially interviewed households was treated with imputation. No special adjustment was made for noninterviews in group quarters.) A factor was applied to each interviewed person's weight to account for the SIPP sample areas not having the same population distribution as the strata from which they were selected.

An additional stage of adjustment to persons' weights was performed to bring the sample estimates into agreement with independent monthly estimates of the civilian (and some military) noninstitutional population of the United States by age, race, and sex. These independent estimates were based on statistics from the 1980 Decennial Census of Population; statistics on births, deaths, immigration, and emigration; and statistics on the strength of the Armed Forces. To increase accuracy, weights were further adjusted in such a manner that SIPP sample estimates would closely agree with special Current Population Survey (CPS) estimates by type of householder (married, single with relatives or single without relatives by sex and race) and relationship to householder (spouse or other).[1] The estimation procedure for the data in the report also involved an adjustment so that the husband and wife of a household received the same weight.

RELIABILITY OF ESTIMATES

SIPP estimates in this report 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. Descriptions of sources of SIPP nonsampling error, along with a discussion of sampling error, its estimation, and its use in data analyses follow.

Nonsampling variability. Nonsampling errors can be attributed to many sources, e.g., 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 collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, biases resulting from the differing recall periods caused by the rotation pattern, and failure to represent all units within the sample (undercoverage). Quality control and edit procedures were used to minimize errors made by respondents and interviewers.

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 non-Blacks. 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 different characteristics than interviewed persons in the same age-race-sex group. Further, the independent population controls used have not been adjusted for undercoverage in the decennial census.

[1]These special CPS estimates are slightly different from the published monthly CPS estimates. The differences arise from forcing counts of husbands to agree with counts of wives.

As noted in table C-1, there was a 5-percent noninterview rate in wave 1 and a cumulative noninterview rate increasing with each additional time in sample. In addition, it should be noted that nonresponse for income and money-related items is often greater than that for other items. The Bureau has used complex techniques to adjust the weights for nonresponse, but the success of these techniques in avoiding bias is unknown.

Comparability with other statistics. Caution should be exercised when comparing data from this report with data from earlier SIPP publications or with data from other surveys. The comparability problems are caused by seasonal factors and different nonsampling errors.

Sampling variability. 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.

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.

Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common types of hypotheses tested are 1) the population parameters are identical or 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 parameters are different when, in fact, they are identical.

To perform the most common test, let x and y be sample estimates of two parameters of interest. A subsequent section explains how to derive a standard error $S_{(x-y)}$ on the difference x-y. Compute $R = \frac{x-y}{S_{(x-y)}}$ the ratio of the difference to the standard error of the difference. If this ratio is between -2 and +2, no conclusion about the parameters is justified at the 5-percent significance level. If, on the other hand, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 5-percent level. In this event, it is commonly accepted practice to say that the parameters are different. Of course, sometimes this conclusion will be wrong. When the parameters are, in fact, the same, there is a 5 percent chance of concluding that they are different.

All statements of comparison in the report have passed a hypothesis test at the 0.10 level of significance or better, and most have passed a hypothesis test at the 0.05 level of significance or better. This means that, for most differences cited in the report, the estimated absolute difference between parameters is greater than twice the standard error of the difference. For the other differences mentioned, the estimated absolute difference between parameters is between 1.6 and 2.0 times the standard error of the difference. When this is the case, the statement of comparison will be qualified in some way (e.g., by use of the phrase "some evidence").

Note when using small estimates. Summary measures (such as means, medians, and percent distributions) are shown in the report only when the base is 200,000 or greater. Because of the large standard errors involved, there is little chance that summary measures would reveal useful information when computed on a smaller base. Estimated numbers are shown, however, even though the relative standard errors of these numbers are larger than those for the corresponding percentages. These smaller estimates are provided primarily to permit such combinations of the categories as serve each user's needs. Also, care must be taken in the interpretation of small differences. For instance, even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Standard error parameters and their use. To derive standard errors that would be applicable to a wide variety of statistics and could be prepared at a moderate cost, a number of approximations were required. Most of the SIPP statistics have greater variance than those obtained through a simple random sample of the same size because clusters of living quarters are sampled for SIPP.

Two parameters (denoted "a" and "b") were developed to calculate variances for each type of characteristic. The "a" and "b" parameters were computed directly using the 1983 SIPP 3rd quarter data adjusted for several differences in the sample size. These "a" and "b" parameters are used in estimating standard errors of survey estimates.

The "a" and "b" parameters vary by type of estimate and by subgroup to which the estimate applies. Table C-2 provides the first quarter "a" and "b" parameters for various subgroups and types of estimates.

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

Standard errors of estimated numbers. The approximate standard error of an estimated number can be obtained by using formula (1).

$$S_x = \sqrt{ax^2 + bx} \qquad (1)$$

Table C-2. "a" and "b" Parameters for Direct Computation of Standard Errors of Estimated Numbers and Percentages of Households and Persons: First Quarter 1984

Characteristic	Parameters	
	a	b
HOUSEHOLDS		
All races or White........	-0.0000764	6,766
Black	-0.0004661	4,675
ALL PERSONS		
All Races or White		
Both sexes..........	-0.0000864	19,911
Male......................	-0.0001786	19,911
Female....................	-0.0001672	19,911
Black		
Both sexes..........	-0.0002670	7,366
Male......................	-0.0005737	7,366
Female....................	-0.0004993	7,366
PERSONS 16 YEARS AND OVER		
Income and Labor Force		
Both sexes..........	-0.0000321	5,475
Male......................	-0.0000677	5,475
Female....................	-0.0000612	5,475
Program Participation and Benefits		
Both sexes..........	-0.0000943	16,059
Male......................	-0.0001984	16,059
Female....................	-0.0001796	16,059

Here x is the size of the estimate and "a" and "b" are the parameters associated with the particular type of characteristic being estimated.

Illustration of the computation of the standard error of an estimated number. Table 1 shows that there were 15,029,000 persons in nonfarm households with a mean monthly household cash income during the first quarter of 1984 of $4,000 to $4,999. The appropriate "a" and "b" parameters to use in calculating a standard error for the estimate are obtained from table C-2 and are a = -.0000864 and b = 19,911.

Using formula (1), the approximate standard error is

$$\sqrt{(-.0000864)\ (15{,}029{,}000)^2\ +\ (19{,}911)\ (15{,}029{,}000)} \doteq 529{,}000$$

The 68-percent confidence interval as shown by the data is from 14,500,000 to 15,588,000. 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 68 percent of all possible samples.

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. When the numerator and denominator of the percentage have different parameters, use the larger of the two parameters. The approximate standard error, $S_{(x,p)}$, of the estimated percentage can be obtained by the formula

$$S_{(x,p)} = \sqrt{\frac{b}{x} \cdot p(100-p)} \qquad (2)$$

Here x is the size of the subclass of households or persons in households which is the base of the percentage, p is the percentage ($0 < p < 100$), and b is the larger of the "b" parameters of the numerator and denominator.

Illustration of the computation of the standard error of an estimated percentage. Continuing the example from above, of the 15,029,000 persons in nonfarm households where the mean monthly household cash income of $4,000 to $4,999, 91.1 percent were White. Using formula (2) and the "b" parameter from table C-2, the approximate standard error is

$$S_{(x,p)} = \sqrt{\frac{(19{,}911)}{(15{,}029{,}000)}\ (91.1)(100-91.1)} \doteq 1.0$$

Consequently, the 68-percent confidence interval as shown by these data is from 90.1 to 92.1 percent, and the 95-percent confidence interval is from 89.1 to 93.1 percent.

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

$$S_{(x-y)} = \sqrt{S_x^2 + S_y^2 - 2\varrho S_x S_y} \qquad (3)$$

where S_x and S_y are the standard errors of the estimates x and y and ϱ denotes the correlation between the two estimates. The estimates can be numbers, percents, ratios, etc. The user should assume ϱ equals zero. If ϱ is really positive (negative), then this assumption will lead to overestimates (underestimates) of the true standard error.

Illustration of the computation of the standard error of a difference within a quarter. Table 1 shows that the number of persons aged 35 to 44 years in nonfarm households with mean monthly household cash income of $4,000 to $4,999 during the first quarter of 1984 was 2,565,000 and the number of persons age 25 to 34 years in nonfarm households with mean monthly household cash income of $4,000 to $4,999 was 2,364,000. The standard errors of these numbers are 118,000 and 113,000, respectively. Assuming that these two estimates are not correlated, the standard error of the estimated difference of 201,000 is

$$S_{(x-y)} = \sqrt{(118{,}000)^2\ +\ (113{,}000)^2} \doteq 163{,}000$$

Suppose that it is desired to test at the 5-percent significance level whether the number of persons with mean monthly household cash income of $4,000 to $4,999 during the first quarter of 1984 was different for persons age 35 to 44 years in nonfarm households than for persons age 25 to 34 years in nonfarm households. The difference divided by the standard error of the difference is 1.23. Since this is less than 2, the data does not provide any evidence of a significant difference between the two age groups at the 5-percent significance level.

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 (4) below. Because of the approximations used in developing formula (4), an estimate of the standard error of the mean obtained from that formula will generally underestimate the true standard error. The formula used to estimate the standard error of a mean $\overline{x}$ is

$$S_{\overline{x}} = \sqrt{\frac{b}{y}\ s^2} \qquad (4)$$

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 estimated population variance, s^2, is given by formula (5):

$$S^2 = \sum_{i=1}^{c} p_i x_i^2 - \overline{x}^2 \qquad (5)$$

where it is assumed that each person or other unit was placed in one of c groups based on the quantity of the item associated with it; p_i is the estimated proportion of the group of interest whose values for the characteristic (x-values) being considered fall in group i; $x_i = (Z_{i-1} + Z_i)/2$ where Z_{i-1} and Z_i are the lower and upper interval boundaries, respectively, for group i. x_i is assumed to be the most representative value for the characteristic of interest in group i. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate average value is

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

Note that the standard error of the mean given in the tables may not agree with those computed using this formula since those in the tables were computed using the raw data and not grouped data.

Standard error of a mean number of persons with characteristic per family or household. Mean values for persons in families or households may be calculated as the ratio of two numbers. The denominator, y, represents a count of families or households of a certain class, and the numerator, x, represents a count of persons with the characteristic under consideration who are members of these families or households. For example, the mean number of children per family with children is calculated as

$$\frac{x}{y} = \frac{\text{total number of children in families}}{\text{total number of families with children}}$$

For means of this kind, the standard error is approximated by the following formula:

$$S_{\left(\frac{x}{y}\right)} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{S_y}{y}\right)^2 + \left(\frac{S_x}{x}\right)^2 - 2\varrho\left(\frac{S_x}{x}\right)\left(\frac{S_y}{y}\right)\right]} \quad (6)$$

The standard error of the estimated number of families or households is S_y, and the standard error of the estimated number of persons with the characteristic is S_x. In formula (6), ϱ represents the correlation coefficient between the numerator and the denominator of the estimate. If at least one member of each family or household in the class possesses the characteristic, then use 0.7 as an estimate of ϱ. If, on the other hand, it is possible that no member of a family or household has the characteristic, then use $\varrho = 0$.

Standard error of a median. To compute a median, first group the units of interest into cells by the value of the statistic under consideration. Then form a cumulative density for the cells. Identify the first cell with cumulative density greater than 0.5. Use interpolation to find the value of the characteristic that corresponds to cumulative density 0.5. That value is the estimated median. 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. The best procedure is to define the cells (income intervals) to be so small that the method of interpolation does not matter.

The sampling variability of an estimated median depends upon the form of the distribution as well as the size of its base. Given that the data were grouped into intervals (e.g., income intervals), then the standard error of a median is given by

$$\frac{\sqrt{bN}\,(A_2 - A_1)}{2(N_2 - N_1)} = \frac{\sqrt{bN}\,W}{2F} \quad (7)$$

or

$$\frac{\sqrt{b}\,M \ln(A_2/A_1}{\sqrt{N}\,\ln\,[(N-N_1)/(N-N_2)]} \quad (8)$$

depending on whether the linear (7) or the Pareto (8) interpolation was used for estimating the median, where

M = the estimated median,
A_1 and A_2 = the lower and upper boundaries of the interval in which the median falls,
W = A_2-A_1, the width of the interval in which the median falls,
N_1 and N_2 = the number of units with the characteristic (e.g., income) less than A_1 and A_2, respectively,
F = N_2-N_1, the number of units in the interval in which the median lies,
N = the total number of units in the frequency distribution,
b = the appropriate value of the parameter "b".

Standard errors of ratios of means and medians. The standard error for a ratio of means or medians is approximated by formula (9):

$$S_{\left(\frac{x}{y}\right)} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{S_y}{y}\right)^2 + \left(\frac{S_x}{x}\right)^2\right]}$$

where x and y are the means or medians, and S_x and S_y are their associated standard errors. Formula (9) assumes that the means or medians are not correlated. If the correlation between the two means or medians is actually positive (negative), then this procedure will provide an overestimate (underestimate) of the standard error for the ratio of means and medians.