Appendix C. Source and Reliability of Estimates

SOURCE OF DATA

The data were collected during the fourth wave of the 1984 panel of the Survey of Income and Program Participation (SIPP). The SIPP universe is the noninstitutionalized resident population of persons living in the United States.[1]

The 1984 panel 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.

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 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 forwarding addresses or moved to extremely remote parts of the country, additional noninterviews resulted.

[1]The noninstitutionalized resident 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, U.S. citizens residing abroad 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 interviewed.

Noninterviews. Tabulations in this report were drawn from interviews conducted from September through December 1984. Table C-1 summarizes information on nonresponse for the interview months in which the data used to produce this report were collected.

Some respondents do not respond to some of the questions. Therefore, the overall nonresponse rate for some items such as income and other money-related items is higher than the nonresponse rates in table C-1. The Bureau has used complex techniques to handle nonresponse, but the success of these techniques in avoiding bias is unknown.

Estimation. 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 following movers.

A noninterview adjustment factor was applied to the weight of every occupant of interviewed households to account for persons in noninterviewed occupied households which were eligible for the sample. (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.

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

Month	Household units eligible			
	Total	Interviewed	Not interviewed	Nonresponse rate
September 1984	5,600	4,800	800	*14
October 1984	5,600	4,800	800	15
November 1984	5,600	4,700	900	15
December 1984	5,600	4,700	900	17

*Due to rounding of all numbers at 100, there are some inconsistencies. The percentage was calculated using unrounded numbers.

An additional stage of adjustment to person 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 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).[2] 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. The magnitude of SIPP sampling error can be estimated, but this is not true of nonsampling error. Found below are descriptions of sources of SIPP nonsampling error, followed by a discussion of sampling error, its estimation, and its use in data analyses.

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 universe (undercoverage). Quality control and edit procedures were used to reduce errors made by respondents, coders, 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.

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 sources such as the seasonal patterns for many characteristics, definitional differences, 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 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.

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 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 parameters are different when, in fact, they are identical.

All statements of comparison in the report have passed a hypothesis test at the 0.10 level of significance or better. This means that, for differences cited in

[2]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.

Table C-2. Distribution of Monthly Earnings Among Wage and Salary Workers 25 Years and Over

Monthly earnings	Number (thous.)	Percent with at least as much as lower bound of interval
Total	52,727	(X)
Under $500	3,601	100.0
$500 to $999	18,531	93.2
$1,000 to $1,499	12,486	77.0
$1,500 to $1,999	10,171	53.3
$2,000 to $2,499	7,791	34.0
$2,500 to $2,999	4,242	19.2
$3,000 to $3,499	2,450	11.2
$3,500 to $3,999	1,191	6.6
$4,000 and over	2,264	4.3

X Not applicable.

the report, the estimated absolute difference between parameters is greater than 1.6 times the standard error of the difference.

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, in case of a borderline difference, even a small amount of nonsampling error can lead to a wrong decision about the hypotheses, thus distorting a seemingly valid hypothesis test.

Standard error parameters and tables 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 vary by subgroup. Table C-5 provides "a" and "b" parameters for characteristics of interest in this report. The "a" and "b" parameters may be used to directly calculate the standard error for estimated numbers and percentages. Because the actual variance behavior was not identical for all statistics within a group, the standard errors computed from parameters provide an indication of the order of magnitude of the standard error for any specific statistic.

For those users who wish further simplification, we have also provided general standard errors in tables C-3 and C-4. Note that these standard errors must be adjusted by a factor from table C-5. 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.

Standard errors of estimated numbers. The approximate standard error, S_x, of an estimated number of persons shown in this report can be obtained in two ways. Note that neither method should be applied to dollar values. It may be obtained by use of the formula

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

where f is the appropriate factor from table C-5, and s is the standard error on the estimate obtained by interpolation from table C-3. Alternatively, S_x may be approximated by the formula,

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

from which the standard errors in table C-3 were calculated. Use of this formula will provide more accurate results than the use of formula (1) above. 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. SIPP estimates given in text table C show that there were 12,486,000 workers covered by a pension plan with monthly earnings in the range of $1,000 to $1,499. The appropriate parameters and factor from table C-5 and the appropriate general standard error from table C-3 are

$a = -.0000588$, $b = 10{,}027$, $f - .71$, $s = 484{,}000$

Table C-3. Standard Errors of Estimated Numbers of Persons

(Numbers in thousands)

Size of estimate	Standard error[1]	Size of estimate	Standard error[1]
200	63	30,000	721
300	77	50,000	883
600	109	80,000	1,020
1,000	141	100,000	1,062
2,000	199	130,000	1,062
5,000	312	135,000	1,055
8,000	392	150,000	1,021
11,000	457	160,000	987
13,000	494	180,000	886
15,000	528	200,000	725
17,000	560	210,000	609
22,000	629	220,000	446
26,000	678		

[1]These values must be multiplied by the appropriate factor in table C-5 to obtain the correct standard error.

Using formula (1), the approximate standard error is

$$S_x = .71 \times 484{,}000 = 344{,}000$$

Using formula (2), the approximate standard error is

$$\sqrt{(-.0000588)\,(12{,}486{,}000)^2 + (10{,}027)\,(12{,}486{,}000)} = 341{,}000$$

Using the standard error based on formula (2), the approximate 90-percent confidence interval as shown by the data is from 11,940,000 to 13,032,000.

Standard error of a mean. A mean is defined here to be the average quantity of some item per person. For example, we may discuss the mean monthly earnings level of wage and salary workers. Standard errors are usually provided in the detailed tables for all displayed means. However, if the reader desires to calculate standard errors on means for collapsed groups, formula (3) may be used. Because of the approximations used in developing formula (3), an estimate of the standard error of a mean obtained from this formula 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 and b be the parameter associated with the particular type of item.

The standard error of a mean is:

$$S_{\bar{x}} = \sqrt{(b/y)\,S^2} \qquad (3)$$

The estimated population variance, S^2, is given by

$$S^2 = \sum_{i=1}^{c} p_i\, x_i^2 - \bar{x}^2 \qquad (4)$$

Where

$$\bar{x} = \sum_{i=1}^{c} p_i\, x_i \qquad (5)$$

each sample unit falls in one of c groups; p_1 is the estimated proportion of group i; $x_1 = (Z_{1-1} + Z_1/2$ where Z_{1-1} and Z_1 are the lower and upper interval boundaries, respectively, for group i. x_1 is assumed to be the most representative value for the characteristics of interest in group i. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate average value of x_c is

$$x_c = \frac{3}{2} Z_{c-1} \qquad (6)$$

Illustration. The distribution of monthly earnings levels of wage and salary workers is given in text table C. Using formulas (4), (5), (6), and the mean monthly earnings amount of $1,584, the approximate population variance for all workers, S^2, is

$$S^2 = \left(\frac{9{,}535}{78{,}619}\right)(250)^2 + \left(\frac{16{,}405}{78{,}619}\right)(750)^2 + \ldots + \left(\frac{2{,}766}{78{,}619}\right)(6{,}000)^2 - (1{,}584)^2 = 1{,}477{,}014$$

Using formula (3) the estimated standard error of a mean x is

$$S_{\bar{x}} = \sqrt{\left(\frac{5{,}475}{78{,}619{,}000}\right)(1{,}477{,}014)} = \$103$$

Table C-4. Standard Errors of Estimated Percentages of Persons

Base of estimated percentage (thousands)	Estimated percentage[1]					
	1 or 99	2 or 98	5 or 95	10 or 90	25 or 75	50
200	3.1	4.4	6.9	9.5	13.7	15.8
300	2.6	3.6	5.6	7.7	11.2	12.9
600	1.8	2.6	4.0	5.5	7.9	9.1
1,000	1.4	2.0	3.1	4.2	6.1	7.1
2,000	1.0	1.4	2.2	3.0	4.3	5.0
5,000	0.6	0.9	1.4	1.9	2.7	3.2
8,000	0.5	0.7	1.1	1.5	2.2	2.5
11,000	0.4	0.6	0.9	1.3	1.8	2.1
13,000	0.4	0.5	0.8	1.2	1.7	2.0
17,000	0.34	0.5	0.7	1.0	1.5	1.7
22,000	0.29	0.4	0.7	0.9	1.3	1.5
26,000	0.28	0.4	0.6	0.8	1.2	1.4
30,000	0.26	0.4	0.6	0.8	1.1	1.3
50,000	0.20	0.3	0.4	0.6	0.9	1.0
80,000	0.16	0.2	0.3	0.5	0.7	0.8
100,000	0.14	0.2	0.3	0.4	0.6	0.7
130,000	0.12	0.17	0.3	0.4	0.5	0.6
220,000	0.10	0.13	0.2	0.3	0.4	0.5

[1]These values must be multiplied by the appropriate factor in table C-5 to obtain the correct standard error.

Table C-5. SIPP Generalized Variance Parameters

Characteristic	Parameters a	Parameters b	f factor
ALL RACES OR WHITE			
16 Years and Over			
Program participation and benefits (3)			
Both sexes	-0.0000943	16,059	0.90
Male	-0.0001984	16,059	0.90
Female	-0.0001796	16,059	0.90
Income and labor force (5)			
Both sexes	-0.0000321	5,475	0.52
Male	-0.0000677	5,475	0.52
Female	-0.0000612	5,475	0.52
Pension coverage[1] (4)			
Both sexes	-0.0000588	10,027	0.71
Male	-0.0001240	10,027	0.71
Female	-0.0001121	10,027	0.71
All Others[1] (6)			
Both sexes	-0.0000864	19,911	1.00
Male	-0.0001786	19,911	1.00
Female	-0.0001672	19,911	1.00
BLACK			
Poverty (1)			
Both sexes	-0.0004930	13,698	0.83
Male	-0.0010522	13,698	0.83
Female	-0.0009274	13,698	0.83
All Others (2)			
Both sexes	-0.0002670	7,366	0.61
Male	-0.0005737	7,366	0.61
Female	-0.0004933	7,366	0.61

[1]Use the "16 years and over" "Pension Plan" parameters for pension plan tabulations of persons 16 years and over in the labor force. Use the "All Others" parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.

Note: For cross-tabulations, use the parameters of the characteristics with the smaller number within the parentheses.

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 parameter (and appropriate factor) of the numerator.

The type of percentage presented in this report is the percentage of persons sharing a particular characteristic such as the percent of workers covered by a pension plan.

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

$$S_{(x,p)} = f_s \qquad (7)$$

In this formula, f is the appropriate "f" factor from table C-5 and s is the standard error on the estimate from table C-4. Alternatively, it may be approximated by the formula

$$S_{(x\text{-}p)} = \sqrt{(b/x)\ (p)\ (100\text{-}p)} \qquad (8)$$

from which the standard errors in table C-4 were calculated. Use of this formula will give more accurate results than use of formula (7) above. Here x is the size of the subclass of persons 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.

Illustration. Text table C shows that 69.4 percent of workers covered by a pension plan had a monthly earnings level of $1,000 to $1,499. Using formula (7) with the factor from table C-5 and the appropriate standard error from table C-4, the approximate standard error is

$$S_{(x,p)} = 0.71 \times 1.5\% = 1.1\%$$

Using formula (8) with the "b" parameter from table C-5, the approximate standard error is

$$S_{(x\text{-}p)} = \sqrt{\frac{10{,}027}{18{,}002{,}000} 69.4\% \; (100\% - 69.4\%)} = 1.1\%$$

Consequently, the approximate 90-percent confidence interval as shown by these data is from 67.6 to 71.2 percent.

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

$$S_{(x\text{-}y)} = \sqrt{S^2_x + S^2_y} \qquad (9)$$

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 sample correlation coefficient, r, between the two estimates is zero. If r is really positive (negative), then this assumption will lead to overestimates (underestimates) of the true standard error.

Illustration. Again, using text table C, 69.4 percent of workers with a monthly earnings level of $1,000 to $1,499 were covered by a pension plan and 76.2 percent of workers with a monthly earnings level of $1,500 to $1,999 were covered in the same manner. The standard errors for these percentages are computed using formula (8), to be 1.1 percent and 1.2 percent. Assuming that these two estimates are not correlated, the standard error of the estimated difference of 6.8 percentage points is

$$S_{(x\text{-}y)} = \sqrt{(1.1\%)^2 + (1.2\%)^2} = 1.6\%$$

The approximate 90-percent confidence interval is from 4.2 to 9.4 percentage points. Since this interval does not contain zero, we conclude that the difference is significant at the 10-percent level.

Standard error of a median. The median quantity of some item such as income for a given group of persons 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 owning more 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 owning more 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 own more is

$$\text{Pareto: } X_{pN} = \exp\left[\frac{\text{Ln }(pN/N_1)}{\text{Ln }(N_2/N_1)} \; \text{Ln }(A_2/A_1)\right] A_1 \qquad (10)$$

if Pareto interpolation is indicated and

$$\text{Linear: } X_{pN} = \frac{pN - N_1}{N_2 - N_1} \; (A_2 - A_1) + A_1 \qquad (11)$$

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. Again using text table C the median monthly earnings amount of workers covered by a pension plan was $1,586. The size of this group of workers was 52,727,000.

1. Using formula (8), the standard error of 50 percent on a base of 52,727,000 is about 0.7 percentage points.
2. Following step 2, the two percentages of interest are 49.3 and 50.7.
3. By examining table C-2, we see that the percentage 49.3 falls in the income interval from $1,500 to $1,999. (Since 53.3 percent receive more than $1,499 per month, but only 34.0 percent receive more than $1,999 per month, the quantity that exactly 49.3 percent receive more than must be between $1,500 and $1,999.) Thus A_1 = $1,500, A_2 = $1,999, N_1 = 28,109,000, and N_2 = 17,938,000. In this case, we decided to use Pareto interpolation.

Therefore, the upper bound of a 68-percent confidence interval for the median is

$$\exp\left[\left(\mathrm{Ln}\left(\frac{(.493)(52{,}727{,}000)}{28{,}109{,}000}\right)\Big/\mathrm{Ln}\left(\frac{17{,}938{,}000}{28{,}109{,}000}\right)\right)\mathrm{Ln}\left(\frac{1{,}999}{1{,}500}\right)\right](1{,}500) = \$1{,}577$$

Also by examining table C-2, we see that the percentage of 50.7 falls in the income interval from $1,500 to $1,999. Thus, A_1 = $1,500, A_2 = $1,999, N_1 = 28,109,000, and N_2 = 17,938,000. We also decided to use Pareto interpolation for this case. So the lower bound of a 68-percent confidence interval for the median is

$$\exp\left[\left(\mathrm{Ln}\left(\frac{(.507)(52{,}727{,}000)}{28{,}109{,}000}\right)\Big/\mathrm{Ln}\left(\frac{17{,}938{,}000}{28{,}109{,}000}\right)\right)\mathrm{Ln}\left(\frac{1999}{1500}\right)\right](1500) = \$1549$$

Thus, the 68-percent confidence interval on the estimated median is from $1,549 to $1,577. An approximate standard error is

$$\frac{\$1{,}577 - \$1{,}549}{2} = \$14$$

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

$$S_{x/y} = \sqrt{\left(\frac{x}{y}\right)^2\left[\left(\frac{S_y}{y}\right)^2 + \left(\frac{S_x}{x}\right)^2\right]} \qquad (12)$$

where x and y are the means or medians, and S_x and S_y are their associated standard errors. Formula (12) 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.