SOURCE AND ACCURACY STATEMENT
SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)
1986 AND 1987 PANELS

SOURCE OF DATA

The data were collected in the 1986 and 1987 panels 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 exception noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

Each of the 1986 and 1987 panels of the SIPP sample are located in 230 Primary Sampling Units (PSUs) each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of 2 living quarters (LQs) were systematically selected from lists of addresses prepared for the 1980 decennial census to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1980 census, a sample was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panel. In jurisdictions that do not issue building permits, small land areas were sampled and the LQs within were listed by field personnel and then clusters of 4 LQs were subsampled. In addition, sample LQs were selected from supplemental frames that included LQs identified as missed in the 1980 census and persons residing in group quarters at the time of the Census.

Approximately 16,300 living quarters were originally designated for the 1986 panel and approximately 16,700 for the 1987 panel. For Wave 1 of the 1986 panel, interviews were obtained from the occupants of about 11,500 of the 16,300 designated living quarters. For Wave 1 of the 1987 Panel about 11,700 interviews were obtained from the 16,700 designated living quarters. Most of the remaining 4800 living quarters in the 1986 panel and 5000 living quarters in the 1987 panel were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. However, approximately 900 of the 4800 living quarters in the 1986 panel and 800 of the 5000 living quarters in the 1987 panel 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 93 percent of all eligible living quarters participated in Wave 1 of the Survey for both the 1986 and 1987 panels.

For Waves 2-7, 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. With certain restrictions, original sample persons were to be followed if they moved to a new address. When original sample persons moved without leaving a forwarding address or moved to extremely remote parts of the country and no telephone number was available, additional noninterviews resulted.

Sample households within a given panel are divided into four subsamples of nearly equal size. These subsamples are called rotation groups 1, 2, 3, or 4 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 2½ years beginning in February 1986 for the 1986 panel and February 1987 for the 1987 panel. 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 exception is Wave 3 for the 1986 panel which covers three interviews.

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 1986 and 1987 panel topical modules are given in tables 1 and 2, respectively.

Tables 3 and 4 indicate the reference months and interview months for the collection of data from each rotation group for the 1986 and 1987 panels. For example, Wave 1 rotation group 2 of the 1986 panel was interviewed in February 1986 and data for the reference months October 1985 through January 1986 were collected.

Table 1 1986 Panel Topical Modules

Wave	Topical Module
1	None
2	Welfare History Recipiency History Employment History Work Disability History Education and Training History Family Background Marital History Migration History Fertility History Household Relationships
3	Child Care Arrangements Child Support Agreements Support of Non-household Members Health Status and Utilization of Health Care Services Long-term Care Disability Status of Children Job Offers
4	Assets and Liabilities Retirement Expenditures and Pension Plan Coverage Real Estate Property and Vehicles
5	Taxes Annual Income and Retirement Accounts Educational Financing and Enrollment
6	Child Care Arrangements Child Support Agreements Support for Non-household Members Work Related Expenses Shelter Costs/Energy Usage
7	Assets and Liabilities Pension Plan Coverage Real Estate Property and Vehicles

Table 2 1987 Panel Topical Modules

Wave	Topical Module
1	None
2	Welfare History Recipiency History Employment History Work Disability Education and Training History Family Background Marital History Migration History Fertility History Household Relationships
3	Child Care Arrangements Child Support Agreements Support for Non-household Members Work Related Expenses Shelter Costs
4	Assets and Liabilities Real Estate Property and Vehicles
5	Taxes Annual Income Educational Financing and Enrollment
6	Child Care Arrangements Child Support Agreements Support for Non-household Members Health Status and Utilization of Health Care Services Long-term Care Disability Status of Children Job Offers
7	Selected Financial Assets Medical Expenses Work Disability Real Estate, Shelter Costs, Dependent Care and Vehicles

Table 3. Reference Months for Each Interview Month - 1986 Panel

Reference Period

Month of Interview	Wave/ Rotation	4th Quarter (1985) Oct	Nov	Dec	1st Quarter (1986) Jan	Feb	Mar	2nd Quarter (1986) Apr	May	Jun	3rd Quarter (1986) Jul	Aug	Sep	4th Quarter (1986) Oct	Nov	Dec	. . .	4th Quarter (1987) Oct	Nov	Dec	1st Quarter (1988) Jan	Feb	Mar
Feb 86	1/2	X	X	X	X																		
March	1/3		X	X	X	X																	
April	1/4			X	X	X	X																
May	1/1				X	X	X	X															
June	2/2					X	X	X	X														
July	2/3						X	X	X	X													
Aug	2/4							X	X	X	X												
Sept	2/1								X	X	X	X											
Oct	3/2									X	X	X	X										
Nov	3/3										X	X	X	X									
Dec	3/4											X	X	X	X								
.													.	.	.								
.															.		. . .						
.																	. . .		. .				
April 88	7/4																			X	X	X	X

Table 4. Reference Months for Each Interview Month - 1987 Panel

Reference Period

Month of Interview	Wave/ Rotation	4th Quarter (1986)			1st Quarter (1987)			2nd Quarter (1987)			3rd Quarter (1987)			4th Quarter (1987)			. . .	1st Quarter (1989)			2nd Quarter (1989)		
		Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec		Jan	Feb	Mar	Apr	May	Jun
Feb 87	1/2	X	X	X	X																		
March	1/3		X	X	X	X																	
April	1/4			X	X	X	X																
May	1/1				X	X	X	X															
June	2/2					X	X	X	X														
July	2/3						X	X	X	X													
Aug	2/4							X	X	X	X												
Sept	2/1								X	X	X	X											
Oct	3/2									X	X	X	X										
Nov	3/3										X	X	X	X									
Dec	3/4											X	X	X	X								
.													.	.	.								
.															.		. . .						
.																	. . .						
May 89	7/1																	X	X	X	X		

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 households which 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 reduce the mean square error of the survey estimates by ratio adjusting SIPP sample estimates to monthly Current Population Survey (CPS) estimates[1] of the civilian (and some military) noninstitutional population of the United States by age, race, Spanish origin, sex, type of householder (married, single with relatives, single without relatives), and relationship to householder (spouse or other). The CPS estimates were themselves brought into agreement with estimates from the 1980 decennial census which were adjusted to reflect births, deaths, immigration, emigration, and changes in the Armed Forces since 1980. Also, an adjustment was made so that a husband and wife within the same household were assigned equal weights.

Use of Weights.

Each household and each person within each household on each wave tape has five weights. Four of these 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. For example, using the proper weights, one can estimate the monthly average number of households in a specified income range over November and December 1986. 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.

The remaining weight is interview month specific. This weight can be used to form estimates that specifically refer to the interview month (e.g., total persons currently looking for work), as well as estimates referring to the time period including the interview month and all previous months (e.g., total persons who have ever served in the military).

To form an estimate for a particular month, use the reference month 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, February 1986 data is only available from rotations 1, 3, and 4 for Wave 1 of the 1986 panel, so a factor of 4/3 must be applied. To form an estimate for an interview month, use the procedure discussed above using the interview month weight provided on the file.

When estimates for months without 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 person'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 1986).

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.

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 for 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 11 states, metropolitan or non-metropolitan residence is identified (variable H*-METRO). In 34 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 8. (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 Washington, DC-MD-VA MSA, apply the Virginia factor of 1.0521 to weights for residents of the Virginia part of the MSA; Maryland and DC residents require no modification to the weights (i.e., their factors equal 1.0).

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) and one state-group (North Dakota - South Dakota - Iowa). Thus, factors in the right-hand column of table 8 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 11 states where the factor for state tabulations in table 8 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 estimate for the metropolitan population. The results of these tabulations will be slightly biased.

ACCURACY OF THE ESTIMATES

SIPP estimates obtained from public use files 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 analysis.

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

The following tables summarize information on household nonresponse for the interview months for Wave 1 of the 1986 and 1987 panels, respectively.

Table 5. 1986 Panel: Sample Size, by Month and Interview Status

Household Units Eligible

Month	Total	Interviewed	Noninterviewed	Nonresponse Rate (%)
Feb. 1986	3200	3000	300	8
Mar. 1986	3100	2900	200	9
Apr. 1986	3100	2800	200	7
May 1986	3000	2800	200	7
	12,400	11,500	900	

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

Table 6. 1987 Panel: Sample Size, by Month and Interview Status

Household Units Eligible

Month	Total	Interviewed	Noninterviewed	Nonresponse Rate (%)
Feb. 1987	3100	2900	200	7
Mar. 1987	3200	2900	200	7
Apr. 1987	3000	2900	200	6
May 1987	3200	3000	200	8
	12,500	11,700	800	

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

Sample loss at Wave 1 of the 1986 and 1987 Panels was about 7% and increased to roughly 19% at the end of Wave 5 of the 1986 Panel and to roughly 18% at the end of Wave 5 for the 1987 Panel. Further noninterviews increased the sample loss about 1% for each of the remaining waves.

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 the above tables.

The Bureau uses complex techniques to adjust the weights for nonresponse, but the success of these techniques in avoiding bias is unknown.

Unique to the 1986 Panel, maximum telephone interviewing was tested in Waves 2,3, and 4. Specifically, half of the sample in rotations 4 and 1 of Wave 2, rotations 2 and 3 of Wave 3 and rotations 2,3, and 4 of Wave 4 were designated for telephone interviews. Analysis has not yet been completed so the affect on data quality is not yet known. Hence, caution should be used when interpreting analytical results, especially for Waves 2 through 4 of the 1986 panel. Again, this test was conducted in the 1986 panel only and will have no bearing on the 1987 Panel data.

Comparability With Other Statistics.

Caution should be exercised when comparing data from these files with data from other SIPP products 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 variability. 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.

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 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.

To perform the most common hypothesis test, compute the difference $X_A - X_B$, where X_A and X_B are sample estimates of the parameters 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 parameters 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 parameters are different. Of course, sometimes this conclusion will be wrong. When the parameters are, in fact, the same, there is a 10 percent chance of concluding that they are different.

Note when using small estimates.

Because of the large standard errors involved, there is little chance that summary measures would reveal useful information when computed on a smaller base than 200,000. 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.

Most SIPP estimates have greater standard errors than those obtained through a simple random sample because clusters of living quarters are sampled. 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. These "a" and "b" parameters are used in estimating standard errors and vary by type of estimate and by subgroup to which the estimate applies. Table 9 provides base "a" and "b" parameters to be used for estimates in this file.

The factors provided in table 10 when multiplied by the base parameters 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 income of households are -0.0001168 and 10,623, respectively.

For Wave 1 the factor for October 1985 is 4 since only 1 rotation of data is available. So, the "a" and "b" parameters for total household income in October 1985 based on Wave 1 are -0.0004672 and 42,492, respectively. Also for Wave 1, the factor for the first quarter of 1986 is 1.2222 since 9 rotation months of data are available (rotations 1 and 4 provide 3 rotations months each, while rotations 2 and 3 provide 1 and 2 rotation months, respectively). So, the "a" and "b" parameters for total household income in the first quarter of 1986 are -0.0001428 and 12,983, 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 parameters 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 11 through 14 for making estimates with the use of data from all four rotations. Note that these standard errors must be adjusted by a factor from table 9. 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, 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.

It may be obtained by the use of the formula

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

where f is the appropriate "f" factor from table 9, and s is the standard error on the estimate obtained by interpolation from table 11 or 12. Alternatively, s_x may be approximated by the formula

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

from which the standard errors in tables 11 and 12 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 1986 panel show that there were 472,000 households with monthly household income above $6,000. The appropriate parameters and factor from table 9 and the appropriate general standard error from table 11 are

$$a = -0.0001168 \qquad b = 10{,}623 \qquad f = 1.0 \qquad s = 71{,}000$$

Using formula 1, the approximate standard error is

$$s_x = 71{,}000$$

Using formula 2, the approximate standard error is

$$\sqrt{(-0.0001168)\ (472{,}000)^2 + (10{,}623)\ (472{,}000)} \approx 70{,}600$$

Using the standard error based on formula 2, the approximate 90-percent confidence interval as shown by the data is from 359,000 to 585,000. 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 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 person i. 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 person 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 persons 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 persons with the item of interest and w_i is the final weight for person i. The mean, $\bar{x}$, can be obtained from the formula

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

Illustration

Suppose that based on Wave 1 data, the distribution of monthly cash income for persons age 25 to 34 during the month of January 1986 is given in table 7.

Table 7 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	39,851	1371	1651	2259	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

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, the appropriate base "b" parameter and factor from table 9, the estimated standard error of a mean x is

$$s_{\bar{x}} = \sqrt{\left(\frac{8,596}{39,851,000}\right) (3,159,887)} = \$26$$

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:

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

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 9 and s is the standard error of the estimate from table 13 or 14. 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 13 and 14 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.

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

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

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 $\bar{x}_A$. If these correlations are positive, then formula 9 will tend to overestimate the true standard error. If they are negative, underestimates will tend to result.

Illustration.

Suppose that, in the month of January 1986, 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 11,565 and a factor of 1 for the month of January 1986 from table 9, the approximate standard error is

$$\sqrt{\frac{11,565}{(16,812,000)} \quad (6.7) \quad (100-6.7)} \approx 0.66 \text{ percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 5.6 to 7.8 percent.

Standard Error of a Difference.

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

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

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, r, between the characteristics estimated by x and y is zero. If r 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 1986 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 and factors from table 9 and formula 2, the standard errors of these numbers are approximately 164,000 and 149,000, respectively. The difference in sample estimates is 567,000 and, using formula 10, the approximate standard error of the difference is

$$\sqrt{(164,000)^2 + (149,000)^2} \approx 222,000$$

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 567,000 to the product 1.6 x 222,000 = 355,200. Since the difference is greater than 1.6 times the standard error of the difference, the data show that the two age groups are 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 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 appropriatenessof 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

$$X_{pN} = \exp\left[\left(\mathrm{Ln}\left(\frac{pN}{N_1}\right) \Big/ \mathrm{Ln}\left(\frac{N_2}{N_1}\right)\right) \mathrm{Ln}\left(\frac{A_2}{A_1}\right)\right] A_1 \qquad (11)$$

if Pareto Interpolation is indicated and

$$X_{pN} = \left[\frac{pN-N_1}{N_2-N_1}\,(A_2-A_1) + A_1\right] \qquad (12)$$

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 the same table 7. The median monthly income for this group is $2,158. The size of the group is 39,851,000.

1. Using the formula 8, the standard error of 50 percent on a base of 39,851,000 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 7, we see that the percentage 49.3 falls in the income interval from 2000 to 2499. (Since 55.5% receive more than $2,000 per month, the dollar value corresponding to 49.3 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 bond of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(\mathrm{Ln}\left(\frac{(.493)\ (39{,}851{,}000)}{22{,}106{,}000}\right)\Big/\mathrm{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right)\ \mathrm{Ln}\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2181$$

Also by examining table 7, we see that 50.7 falls in the same income interval. Thus, A_1, A_2, N_1, and N_2 are the same. We also decided to use Pareto interpolation for this case. So the lower bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(\mathrm{Ln}\left(\frac{(.507)\ (39{,}851{,}000)}{22{,}106{,}000}\right)\Big/\mathrm{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right)\ \mathrm{Ln}\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2136$$

Thus, the 68-percent confidence interval on the estimated median is from $2136 to $2181. An approximate standard error is

$$\frac{\$2181 - \$2136}{2} = \$23$$

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, 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 rror for the ratio of means.

Table 8. 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	1.0387	1.0387
	Maine	1.2219	1.2219
	Massachusetts	1.0000	1.0000
	New Hampshire	1.2234	1.2234
	New Jersey	1.0000	1.0000
	New York	1.0000	1.0000
	Pennsylvania	1.0096	1.0096
	Rhode Island	1.2506	1.2506
	Vermont	1.2219	1.2219
Midwest:	Illinois	1.0000	1.0110
	Indiana	1.0336	1.0450
	Iowa	–	–
	Kansas	1.2994	1.3137
	Michigan	1.0328	1.0442
	Minnesota	1.0366	1.0480
	Missouri	1.0756	1.0874
	Nebraska	1.6173	1.6351
	North Dakota	–	–
	Ohio	1.0233	1.0346
	South Dakota	–	–
	Wisconsin	1.0188	1.0300
South:	Alabama	1.1574	1.1595
	Arkansas	1.6150	1.6179
	Delaware	1.5593	1.5621
	D.C.	1.0000	1.0018
	Florida	1.0140	1.0158
	Georgia	1.0142	1.0160
	Kentucky	1.2120	1.2142
	Louisiana	1.0734	1.0753
	Maryland	1.0000	1.0018
	Mississippi	–	–
	North Carolina	1.0000	1.0018
	Oklahoma	1.0793	1.0812
	South Carolina	1.0185	1.0203
	Tennessee	1.0517	1.0536
	Texas	1.0113	1.0131
	Virginia	1.0521	1.0540
	West Virginia	–	–
West:	Alaska	1.4339	1.4339
	Arizona	1.0117	1.0117
	California	1.0000	1.0000
	Colorado	1.1306	1.1306
	Hawaii	1.0000	1.0000
	Idaho	1.4339	1.4339
	Montana	1.4339	1.4339
	Nevada	1.0000	1.0000
	New Mexico	1.0000	1.0000
	Oregon	1.1317	1.1317
	Utah	1.0000	1.0000
	Washington	1.0456	1.0456
	Wyoming	1.4339	1.4339

– indicates no metropolitan subsample is identified for the state

Table 9. SIPP Indirect Generalized Variance Parameters for the 1986+Panels

CHARACTERISTICS[1]	a	b	f
PERSONS			
Total or White			
16+ Program Participation and Benefits, Poverty (3)			
Both Sexes	-0.0001481	25,213	.90
Male	-0.0003115	25,213	
Female	-0.0002820	25,213	
16+ Income and Labor Force (5)			
Both Sexes	-0.0000504	8,596	.52
Male	-0.0001063	8,596	
Female	-0.0000961	8,596	
16+ Pension Plan[2] (4)			
Both Sexes	-0.0000923	15,742	.71
Male	-0.0001947	15,742	
Female	-0.0001760	15,742	
All Others[2] (6)			
Both Sexes	-0.0001356	31,260	1.00
Male	-0.0002804	31,260	
Female	-0.0002625	31,260	
Black			
Poverty (1)			
Both Sexes	-0.0007740	21,506	.83
Male	-0.0016520	21,506	
Female	-0.0014560	21,506	
All Others (2)			
Both Sexes	-0.0004192	11,565	.61
Male	-0.0009007	11,565	
Female	-0.0007839	11,565	
HOUSEHOLDS			
Total or White	-0.0001168	10,6231	.00
Black	-0.0007318	7,340	.83

1. To account for sample attrition, multiply the a and b parameters by 1.09 for estimates which include data from Wave 5 and beyond.

   For cross-tabulations, use the parameters of the characteristic with the smaller number within the parentheses.

2. Use the "16+ Pension Plan" parameters for pension plan tabulations of persons 16+ 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.

Table 10. Factors to be Applied to 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
Quarterly estimate	
6	1.8519
8	1.4074
9	1.2222
10	1.0494
11	1.0370
12	1.0000

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

Size of Estimate	Standard Error[1]	Size of Estimate	Standard Error[1]
200	46	15,000	365
300	56	25,000	439
500	73	30,000	462
750	89	40,000	488
1,000	102	50,000	489
2,000	144	60,000	466
3,000	176	70,000	414
5,000	224	80,000	320
7,500	270	90,000	100
10,000	307		

1. To account for sample attrition, multiply the standard error of the estimate by 1.04 for estimates which include data from Wave 5 and beyond.

Table 12. Standard Errors of Estimated Numbers of Persons

Size of Estimate	Standard Error[1]	Size of Estimate	Standard Error[1]
200	79	50,000	1,106
300	97	80,000	1,278
600	137	100,000	1,330
1,000	176	130,000	1,331
2,000	249	135,000	1,322
5,000	391	150,000	1,280
8,000	491	160,000	1,237
11,000	572	180,000	1,111
13,000	619	200,000	910
15,000	662	210,000	765
17,000	702	220,000	560
22,000	789		
26,000	849		
30,000	903		

1. To account for sample attrition, multiply the standard error of the estimate by 1.04 for estimates which include data from Wave 5 and beyond.

Table 13 Standard Errors of Estimated Percentages of Households Families or Unrelated 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	2.3	3.2	5.0	6.9	10.0	11.5
300	1.9	2.6	4.1	5.6	8.1	9.4
500	1.5	2.0	3.2	4.4	6.3	7.3
750	1.2	1.7	2.6	3.6	5.2	6.0
1,000	1.0	1.4	2.2	3.1	4.5	5.2
2,000	0.7	1.0	1.6	2.2	3.2	3.6
3,000	0.6	0.8	1.3	1.8	2.6	3.0
5,000	0.5	0.6	1.0	1.4	2.0	2.3
7,500	0.4	0.5	0.8	1.1	1.6	1.9
10,000	0.3	0.46	0.7	1.0	1.4	1.6
15,000	0.26	0.37	0.6	0.8	1.2	1.3
25,000	0.21	0.29	0.4	0.6	0.9	1.0
30,000	0.19	0.26	0.41	0.56	0.8	0.9
40,000	0.16	0.23	0.36	0.49	0.7	0.8
50,000	0.15	0.20	0.32	0.44	0.6	0.7
60,000	0.13	0.19	0.29	0.40	0.58	0.66
80,000	0.11	0.16	0.25	0.35	0.50	0.58
90,000	0.11	0.15	0.24	0.33	0.47	0.54

1. To account for sample attrition, multiply the standard error of the estimate by 1.04 for estimates which include data from Wave 5 and beyond.

Table 14 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	3.9	5.5	8.6	11.9	17.1	19.8
300	3.2	4.5	7.0	9.7	14.0	16.1
600	2.3	3.2	5.0	6.8	10.0	11.4
1,000	1.8	2.5	3.9	5.3	7.7	8.8
2,000	1.2	1.8	2.7	3.8	5.4	6.3
5,000	0.8	1.1	1.7	2.4	3.4	4.0
8,000	0.6	0.9	1.4	1.9	2.7	3.1
11,000	0.53	0.75	1.2	1.6	2.3	2.7
13,000	0.49	0.69	1.1	1.5	2.1	2.5
17,000	0.43	0.60	0.9	1.3	1.9	2.1
22,000	0.38	0.53	0.8	1.1	1.6	1.9
26,000	0.35	0.49	0.76	1.0	1.5	1.7
30,000	0.32	0.45	0.70	0.97	1.4	1.6
50,000	0.25	0.35	0.54	0.75	1.1	1.3
80,000	0.20	0.28	0.43	0.60	0.9	1.0
100,000	0.18	0.25	0.39	0.53	0.8	0.9
130,000	0.15	0.22	0.34	0.47	0.67	0.77
220,000	0.12	0.17	0.26	0.36	0.52	0.60

1. To account for sample attrition, multiply the standard error of the estimate by 1.04 for estimates which include data from Wave 5 and beyond.