SURVEY OF INCOME AND PROGRAM PARTICIPATION, (SIPP)
1984 PANEL ANNUAL WEIGHTS FILE

This file documentation consists of the following materials:

Attachment 1	Abstract
Attachment 2	Source and Reliability Statement
Attachment 3	Data Dictionary

NOTE

Questions about the accompanying documentation should be directed to Data User Services Division, Data Access and Use Staff, Bureau of the Census, Washington, D.C. 20233. Phone: (301) 763-2074.

Questions about the tape should be directed to Data User Services Division, Customer Services (Tapes), Bureau of the Census, Washington, D.C. 20233. Phone: (301) 763-4100.

For additional general information about SIPP, contact Daniel Kasprzyk (763-5764) or David McMillen (763-7958) in Population Division, Bureau of the Census, Washington, D.C. 20233.

ABSTRACT

Survey of Income and Program Participation (SIPP), 1984 Panel Annual Weights File [machine-readable data file] / conducted by the U.S. Bureau of the Census.--Washington: The Bureau [producer and distributor], 1987.

Type of File:

Microdata; unit of observation is persons.

Universe Description:

Universe is the resident population of the United States, excluding persons living in institutions and military barracks.

Subject-Matter Description:

The file provides the appropriate identification match fields for all respondents in Waves 2-5 of the 1984 panel as well as two weights. One weight is controlled to the December 1983 population estimates; the second weight is controlled to the March 1985 population estimates.

Geographic Coverage:

There are no geographic identification codes on this file.

Technical Description:

File Structure: Rectangular

File Size: 34,788 records; record length is 40 characters.

File Sort Sequence: Sequential by person identification fields (positions 1-14).

Reference Materials:

Survey of Income and Program Participation (SIPP), 1984 Panel Annual Weights Technical Documentation. The documentation includes this abstract, a record layout and an extensive source and reliability statement, including caveats on using the weights.

Related Machine-Readable Data Files:

SIPP files are available from Data User Services Division, Customer Services, Washington, D.C. 20233. An order form is on the following page for your convenience. Files related to this file are listed below:

1984 Panel- Rectangular Files	Number of Reels 1600 bpi	6250 bpi	Cost 1600 bpi	6250 bpi
Wave 2 Core	6	2	$1,050	$350
Wave 3 Core	8	2	$1,400	$350
Wave 3 Core & Topical Module	9	3	$1,575	$525
Wave 4 Core	8	2	$1,400	$350
Wave 4 Core & Topical Module	10	3	$1,750	$525
Wave 5 Core	7	2	$1,225	$350

1984 Panel- Relational Files	Number of Reels 1600 bpi	6250 bpi	Cost 1600 bpi	6250 bpi
Wave 2 Core	9	3	$1,575	$525
Wave 3 Core	12	3	$2,100	$525
Wave 4 Core	12	3	$2,100	$525
Wave 5 Core	12	3	$2,100	$525

A machine-readable data dictionary is available at the end of the last reel at either density or may be purchased separately on 1 reel for $175.

File Availability:

The file is available without charge to users who purchased Wave 2, 3, 4, or 5 from the Census Bureau. Others may order the file from the Data User Services Division using the order form on the following page. It is available for $175 in 1 reel at either 1600 or 6250 bpi.

Customer Services (Microfiche and Data Files)

Mail To: Data User Services Division, Customer Services, Bureau of the Census, Washington, D.C. 20233 Phone: 301/763-4100.

Name _____ Telephone _____

Organization _____

Address _____

City, State, ZIP _____ Date _____

CHECK ONE

- ☐ Enclosed is check or money order, payable to "Commerce-Census"
- ☐ Charge to Census Bureau Deposit Account No. 9 _____
- ☐ Credit card (below)

Name of Data File, Documentation* or Microfiche	No. of Tapes	No. of Microfiche	No. of Document copies	Cost	
*One copy of the Technical Documentation is sent at no extra charge with tape orders.			TOTAL $		

Characteristics of Tape (Check One):

- ☐ 9 track, 1600 bpi, EBCDIC
- ☐ 9 track, 6250 bpi, EBCDIC
- ☐ 9 track, 1600 bpi, ASCII
- ☐ 9 track, 6250 bpi, ASCII

Labeling (Check One): ☐ Standard ☐ Unlabeled

Maximum block size (If less than 32K bytes)? _____

Credit Card Orders

Total charges $_____ Fill in the boxes below.

Credit Card No. ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Expiration Date Month/Year ☐☐☐☐

MasterCard and VISA accepted.

SOURCE AND RELIABILITY STATEMENT FOR ESTIMATES DERIVED FROM THE SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP) LONGITUDINAL WEIGHTS

SOURCE OF DATA

The weights were calculated upon the request of several federal agencies for the purpose of comparing longitudinal 1984 Calendar Year estimates from SIPP with those obtained from the 1985 CPS March supplement. Two final longitudinal weights were calculated for each person. The first was based on householder /nonhouseholder type and age as of December 1983 (the cohort appropriate for SIPP longitudinal analysis). The second weight was based on householder/nonhouseholder type and age as of March 1985 (the cohort associated with CPS March type estimates).

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. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

The 1984 panel SIPP sample is located in 174 areas comprising 450 counties (including one partial county) and independent cities. Within these areas, clusters of 2 to 4 living quarters (LQs) were systematically selected from lists of addresses prepared for the 1970 decennial census to form the bulk of the sample. In jurisdictions requiring a building permit for new private residential construction and where Census address information was considered inadequate, an area sample (small land areas were sampled and the LQs within were listed by field personnel and then subsampled) for all units except those constructed since the census was performed. To account for LQs built within each of the sample areas after the 1970 census, a sample was drawn of permits issued for construction of residential LQs through March 1983. In jurisdictions that do not issue building permits, an area sample to represent all existing units was performed. In addition, sample LQs were selected from supplemental frames that included mobile home parks and new construction for which permits were issued prior to January 1, 1970, but for which construction was not completed until after April 1, 1970.

Sample households within a given panel are divided into four subsamples of nearly equal size. These subsamples are called rotation groups and one rotation group is interviewed each month. Each household in the sample was scheduled to be interviewed at 4 month intervals over a period of 2 1/2 years beginning in October 1983. The reference period for the questions is the 4-month period preceding the interview. In general, one cycle of four interviews covering the entire sample, using the same questionnaire, is called a wave.

The longitudinal weights were created from data collected prior to August 1985. Persons who have positive weights generally must have at least 20 reference months of data. This period was used to ensure that December 1983 age and relationship to householder characteristics and March 1985 age and relationship to householder characteristics were available for weighting purposes.

Table 1 indicates the reference months and interview month for the collection of data from the first six interviews for each rotation group. For example, rotation group 2 was interviewed in November 1983 and data for the reference months July 1983 through October 1983 were collected. Rotation group 2 was interviewed five more times during the next twenty months, in March, July, and November 1984 and March and July 1985.

Approximately 26,000 living quarters were originally designated for the sample. For Wave 1, interviews were obtained from the occupants of about 19,900 of the 26,000 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

Table 1. REFERENCE MONTHS FOR EACH INTERVIEW MONTH - LONGITUDINAL RESEARCH FILE

Reference Period

Month of Inter-view	Rota-tion	2nd Quarter (1983)			3rd Quarter (1983)			4th Quarter (1983)			1st Quarter (1984)			2nd Quarter (1984)			3rd Quarter (1984)			4th Quarter (1984)			1st Quarter (1985)			2nd Quarter (1985)		
		Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Oct. 83	1		X		X	X	X																					
Nov.	2				X	X	X	X																				
Dec.	3					X	X	X	X																			
Jan. 84	4						X	X	X	X																		
Feb.	1							X	X	X	X																	
March	2								X	X	X	X																
Apr.	3									X	X	X	X															
May	4										X	X	X	X														
June	1											X	X	X	X													
July	2												X	X	X	X												
Aug.	3													X	X	X	X											
Sept.	4														X	X	X	X										
Oct.	1															X	X	X	X									
Nov.	2																X	X	X	X								
Dec.	3																	X	X	X	X							
Jan. 85	4																		X	X	X	X						
Feb.	1																			X	X	X	X					
March	2																				X	X	X	X				
Apr.	3																					X	X	X	X			
May	4																						X	X	X	X		
June	1																							X	X	X	X	
July	2																								X	X	X	X

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 a forwarding address or moved to extremely remote parts of the country, additional noninterviews resulted. Approximately 52,800 persons were counted as initially interviewed, but this count excludes about 1,300 people who were members of households containing "type Z" noninterviews (i.e., those occurring in households in which only some member(s) declined to be interviewed). Approximately 34,100 persons subsequently retained positive longitudinal weights. Of the remaining 18,700, approximately 9,400 were dropped from sample in March 1985 because of budgetary reasons and 9,300 remaining in sample were noninterviews some time after their initial interview.

The following sample persons were treated as "interviewed" persons in the weighting procedure: 1) those who responded to each of the interviews conducted before August 1985 and who during the first interview lived in a household in which all eligible members responded to the interview (call this a wave 1 interviewed household) and 2) those who resided in a wave 1 interviewed household but after the wave 1 interview and prior to August 1985 are known to have died or moved to an ineligible address.

The following sample persons were treated as "noninterviewed"

persons in the weighting procedure: 1)those who at the time of the first interview lived in a household in which at least one household member failed to respond to the first interview (call this a wave 1 noninterviewed household), 2) those who resided in a wave 1 interviewed household but did not respond to at least one of the interviews conducted before August 1985 because of household or person nonresponse, and 3) those who resided in a wave 1 interviewed household but who moved in with members of another wave 1 interviewed household after the first interview. These persons are treated as noninterviews because an imputation system for handling missing interviews is not yet available and because the processing system is unable to handle households defined in 1) and 3) above.

ESTIMATION OF WEIGHTS

The requested weights are constructed from three components. The first component is the longitudinal person weight calculated for use with the first three interview longitudinal research file. The second is a factor which adjusts these longitudinal person weights for the March 1985 sample cut. The third is a raking procedure which is expected to reduce the mean square error of important annual estimates derived from the weight. A brief description of procedures used to create the three interview longitudinal person weights and the adjustment used to create the requested longitudinal person weights are given below.

The estimation procedure used to derive the SIPP three interview longitudinal person weights involved several stages of weighting

adjustments. First, each person interviewed in wave 1 was assigned an initial weight comprised of 1)a base weight which was equal to the inverse of his/her probability of selection; 2) a noninterview adjustment factor which was applied to the weights of interviewed persons in interviewed households to account for persons in noninterviewed households who were eligible for the sample but were not interviewed, and 3) a factor which 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.

Next, a second noninterview adjustment factor was applied to the weights of interviewed persons to account for persons who were not interviewed in either the second or third interviews.

The last stage of adjustment to the SIPP three interview longitudinal person weights was performed to reduce the mean square error of the survey estimates. This was accomplished by bringing the sample estimates into agreement with independent monthly estimates of the civilian (and some military) noninstitutional population of the United States by age, race, Hispanic ethnicity, and sex as of December 1, 1983 and with special Current Population Survey (CPS) estimates by type of householder and relationship to householder using a raking procedure.[1]

---

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

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.

In the weighting process for the requested longitudinal weights, weights for all persons with positive three interview longitudinal weights and who were classified as interviewed were adjusted further. All other persons were given weights of zero and were not further adjusted. The first adjustment was application of a constant factor to compensate for the sample cut. The second and final adjustment was intended to reduce the mean square error of survey estimates derived from the resulting weights. The raking procedure used in this adjustment is identical to the one used for the longitudinal weights described above. Two different ratio adjustments were done on the longitudinal weights adjusted for the sample cut. One adjustment controlled a December 1983 cohort to CPS March type population estimates for December 1983 by age, race, sex, householder/not householder, relationship to householder and with population estimates of persons with Spanish origin for December 1983. The other adjustment controlled a December 1983 cohort to CPS March type population estimates for March 1985 by the characteristics listed above for March 1985. Two adjustments yielding two distinct sets of weights were carried out because there was concern over the effect on estimates of controlling a December 1983 cohort to March 1985 independent estimates. Note, in both this raking procedure and the one described above, the weights of husbands and their wives were not forced to be equal; thus, theoretically a difference in the estimates of husbands

and wives should be expected.

USE OF WEIGHTS

The following paragraphs give information on the proper use of the requested weights and associated caveats.

Each person has two weights; a weight based on December 1983 population estimates and a weight based on March 1985 population estimates. The March 1985 based weights were created only for comparing CPS 1985 March Supplement estimates with 1984 annual estimates from SIPP.

To form 1984 annual estimates for persons, sum the weights of all persons possessing the characteristic of interest. To form an estimate for a particular month, sum weights over all persons with the characteristic of interest whose reference period includes the month of interest. To estimate monthly averages of a given measure over a number of consecutive months, sum the estimates from each month and divide by the number of months.

Note the following caveats:

* The cohort of the requested weights were based on the set of persons interviewed in wave one rather than the set of persons interviewed at the start of the calendar year 1984.

* No special noninterview adjustment was done to account for the additional sample loss which occurred at the fourth, fifth and

sixth waves due to nonresponse. However, the final raking adjustment procedure would account for some of the additional sample loss.

* The number of males that were married with spouses present for March 1985 (based on March 1985 characteristics) was estimated at 51,944,000, while the number of females married with husbands present was roughly 50,814,000, a difference of 1,130,000. For December 1983 (based on December 1983 characteristics), the difference between estimated numbers of males, married spouse present and females, married spouse present was 799,000.

* Sample persons that were known to have died or left the universe by March 1985 were treated as interviewed persons for weighting purposes and were given positive weights through March 1985. These persons were given positive weights and included on the file so that estimates of the numbers of such persons and the economic effects on the affected households could be made. Note that there is no field on the weight file that indicates that these persons left the universe. This information can be obtained from the Cross-Sectional Public Use Files.

* These weights were based on 20 to 24 reference months of data for each person. Only the first 12 months of these data were longitudinally edited. As a result, the weights created with March 1985 population controls were based on householder relationship that had been only cross-sectionally edited while weights created with December 1983 controls were based on longi-

tudinally edited data. (Edited age data was used to create both sets of weights.)

* The weights were designed to be used soley for producing person estimates. They were not developed for producing family or household estimates. Hence, use of these weights may provide inappropriate family or household estimates. Estimates at the individual state level are subject to very high variance and are not recommended.

RELIABILITY OF THE ESTIMATES

Estimates created using the weights on this file 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 the sampling error, but this is not true of nonsampling error. Found below are descriptions of sources of 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 sample (undercoverage). Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers.

Undercoverage results from missed living quarters and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for blacks than for nonblacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group. Further, the independent population controls used have not been adjusted for undercoverage in the decennial census. The Bureau has used complex techniques to adjust for nonresponse, but the success of these techniques in avoiding bias is unknown.

In addition, some respondents do not respond to some of the questions. Therefore, the overall nonresponse rate for some items, such as income and money-related items is higher than the nonresponse rates for other items.

Comparability with other statistics. Caution should be exercised when comparing estimates created using the weights on this file with estimates from other SIPP products or with estimates from other surveys. The comparability problems are caused by the seasonal patterns for many characteristics, by different nonsampling errors, and by different concepts and procedures in other surveys.

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 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.
2. 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 test, let x and y be sample estimates of two parameters of interest. A subsequent section explains how to derive a standard error on the difference x-y. If the absolute difference between x and y is greater than 1.6 times the standard error of the difference, 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% chance of concluding that they are different. If the absolute difference between x and y is less than 1.6 times the standard error

of the difference, no conclusion about the parameters is justified at the 10% significance level.

Note when using small estimates. Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000 if they are obtained using a small number of cases from SIPP. Nonsampling error can occasionally occur in one of the small number of cases providing the estimate, causing large relative error in that particular estimate. Also, care must be taken in the interpretation of small differences. 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 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. All statistics do not have the same variance behavior; statistics with similar variance behavior were grouped together. Most of the SIPP statistics have greater variance than those obtained through a simple random sample because clusters of living quarters are sampled for SIPP. Two parameters (denoted "a" and "b") were developed to quantify these increases in variance. 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 group to which the estimate applies. Table 3 provides "a" and "b" parameters for various groups and types of estimates.

The "a" and "b" parameters may be used directly to 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 these parameters provide an indication of the order of magnitude of the standard error for any specific statistic. Methods for using these parameters for direct computation of standard errors are given in the following sections.

The user can create far more types of estimates than standard errors are provided for here. Procedures for calculating standard errors for the types of estimates most commonly used are described below. Note specifically that these procedures apply only to annual estimates or averages of annual estimates. Refer to the section "Use of Weights" for a discussion of construction of estimates.

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

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

Here x is the size of the estimate and "a" and "b" are the parameters associated with the particular type of characteristic for the appropriate reference period.

Illustration. Suppose that the 1984 calendar year estimate of the number of persons ever receiving Social Security is 34,122,000. The appropriate "a" and "b" parameters to use in calculating a standard error for the estimate are obtained from table 3.

They are a = -0.0001054, b = 19,020.

Using formula (1), the approximate standard error is

$$\sqrt{(-0.0001054)(34{,}122{,}000)^2 + (19{,}020)(34{,}122{,}000)} \approx 725{,}000$$

The 90-percent confidence interval as shown by the data is from 32,962,000 to 35,282,000. 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 percent of all samples.

Standard error of mean or of an aggregate. A mean is defined here to be the average quantity of some item per unit or person. An aggregate is defined to be the total quantity of the item summed up for all units in a group. For example, a mean could be the average monthly household income of females age 25 to 34; an aggregate, the total income for that group. The standard error of a mean and aggregate can be approximated by formulas (2) and (3) respectively. Because of the approximations used in developing formulas (2) and (3), an estimate of the standard error of the mean or of an aggregate obtained from these formulas will generally underestimate the true standard error.

The formula used to estimate the standard error of a mean $\bar{x}$ is

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

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 standard error of an aggregate k is estimated by:

$$S_k = \sqrt{b\ y\ s^2} \qquad (3)$$

and, the estimated population variance, s^2, is given by formula (4):

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

where it is assumed that each person was placed in one of c groups; p_i is the estimated proportion of 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 value for x_c is

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

and $\bar{x}$ can be obtained using the following formula:

$$\bar{x} = \sum_{i=1}^{c} p_i x_i$$

Illustration. Suppose that the mean monthly household incomes of persons age 25 to 34 for the twelve months, of calendar year 1984 are given in the table 2. Note that this is a person level characteristic, not a household characteristic.

Table 2. Distribution of Monthly Household 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	2619	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

The mean monthly household cash income is

$$\bar{x} = [(1371)(150) + (1651)(450) + \ldots + (1493)(9000)]/39,851$$

$$= 2,527$$

Using formula (4) and the mean monthly household cash income of $2,527 the approximate population variance, s^2, is

$$s^2 = \left(\frac{1,371}{39,851}\right)(150)^2 + \left(\frac{1,651}{39,851}\right)(450)^2 + \ldots$$

$$+ \left(\frac{1,493}{39,851}\right)(9,000)^2 - (2,527)^2 = 3,144,716.$$

Using formula (2), and the appropriate "b" parameter from table 3, the estimated standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\frac{6,485}{39,851,000}(3,144,716)} = \$23$$

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 of the numerator. If proportions are presented instead of percentages, note that the standard error of a proportion is equal to the quotient of the standard error of the corresponding percentage and 100.

There are two types of percentages commonly estimated. The first is the percentage of persons 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 wealth held by persons with high income and the percent of income for persons on welfare.

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

$$S_{x,p} = \sqrt{\frac{b}{x}\, p(100-p)} \qquad (5)$$

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 "b" parameter for the numerator.

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:

$$\hat{P}_1 = \frac{X_A}{X_N},$$

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

$$\hat{P}_1 = \hat{P}_A \bar{X}_A / \bar{X}_N$$

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 percent of group N that is in group A. In either case, we estimate the standard error as

$$S_1 = \sqrt{\left(\frac{\hat{P}_A \bar{X}_A}{\bar{X}_N}\right)^2 \left[\left(\frac{S_P}{\hat{P}_A}\right)^2 + \left(\frac{S_A}{\bar{X}_A}\right)^2 + \left(\frac{S_B}{\bar{X}_N}\right)^2\right]} \qquad (6)$$

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 (5). The standard errors of $\bar{X}_N$ and $\bar{X}_A$ may be calculated using formula (2).

It should be noted that there is frequently some correlation between $\hat{P}_A$ and $\overline{X_N}$ and $\overline{X_A}$ and $\overline{X_N}$. If these correlations are positive, then this formula will overestimate the standard error. If they are negative, underestimates will result.

Illustration. Suppose that the 1984 calendar year estimate for the total number of Black persons is 27,202,000 and that 47.5 percent of Black persons received non-cash benefits during this period. Using formula (5) and the "b" parameter from table 3, the approximate standard error is

$$\sqrt{\frac{(8,724)}{(27,202,000)}} \quad (47.5) \quad (100-47.5) \approx 0.9 \text{ percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 46.1 to 48.9 percent.

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

$$S_{(x-y)} = \sqrt{S_x^2 + S_y^2} \qquad (7)$$

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. Suppose that 1984 calendar year estimates show the number of persons age 35-44 years in non-farm households with mean monthly household cash income of $4,000 to $4,999 was 3,186,000 and the number of persons age 25-34 years in non-farm households with mean monthly household cash income of $4,000 to $4,999 in the same time period was 2,619,000. Then the standard errors of these numbers are approximately 143,000 and 129,000, respectively. Assuming that these two estimates are not correlated, the standard error of the estimated difference of 567,000 is

$$\sqrt{(143{,}000)^2 + (129{,}000)^2} = 193{,}000.$$

Suppose that it is desired to test at the 10 percent significance level whether the number of persons with mean monthly household cash income of $4,000 to $4,999 was different for persons age 35-44 years in non-farm households than for persons age 25-34 years in non-farm households. The absolute difference between the two sample estimates, x and y, is greater than 1.6 times the standard error of the difference. Therefore, we conclude that the parameters are 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 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.

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 formula (5), 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. We recommend Pareto interpolation in most instances. Interpolation is used as follows. The quantity of the item such that "p" percent own more is

$$X_{pN} = A_1 \exp\left[\operatorname{Ln}\left(\frac{pN}{N_1}\right) \operatorname{Ln}\left(\frac{A_2}{A_1}\right) \Big/ \operatorname{Ln}\left(\frac{N_2}{N_1}\right) \right] \qquad (8)$$

if Pareto interpolation is indicated and

$$X_{pN} = \frac{N_1 - pN}{N_1 - N_2} (A_2 - A_1) + A_1 \qquad (9)$$

if linear interpolation is indicated,

where

N	is 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.

It should be noted that a mathematically equivalent result is obtained by using common logarithms (base 10) and antilogarithms.

Illustration. To illustrate the calculations for the sampling error on a median, we return to the same example used to illustrate the standard error of a mean. The median monthly income for this group is $2,158. The size of the group is 39,851,000.

1. Using formula (5), the standard error of 50 percent on a base of 39,851,000 is about 0.6 percentage points.

2. Following step (2), the two percentages of interest are 49.4 a 50.6.

3. By examining table 2, we see that the percentage 49.4 falls in the income interval from $2,000 to $2,499. Thus A_1 = $2,000, A_2 = $2,500, N_1 = 22,106,000, and N_2 = 16,307,000. Implementing Pareto interpolation, the upper bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\operatorname{Ln}\left(\frac{(.494)(39{,}851{,}000)}{22{,}106{,}000}\right) \operatorname{Ln}\left(\frac{2{,}500}{2{,}000}\right) \Big/ \operatorname{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right] = \$2{,}160$$

Also by examining table 2, we see that the percentage of 50.6 falls in the same income interval. Thus, A_1, A_2, N_1, and N_2 are the same as above. The lower bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\operatorname{Ln}\left(\frac{(.506)(39{,}851{,}000)}{22{,}106{,}000}\right) \operatorname{Ln}\left(\frac{2{,}500}{2{,}000}\right) \Big/ \operatorname{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right] = \$2{,}140$$

and the 68-percent confidence interval on the estimated median is from $2,140 to $2,160. An approximate standard error is

$$\frac{\$2,180 - \$2,140}{2} = \$20.$$

Using linear interpolation, the 68-percent confidence interval of the estimated median is from $2,167 to $2,209 and the approximate standard error is $20.

Standard errors of ratios of means and medians. The standard error for a ratio of means or medians is approximated by formula (10):

$$S_{x/y} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{S_y}{y}\right)^2 + \left(\frac{S_x}{x}\right)^2\right]} \qquad (10)$$

where x and y are the means or medians, and S_x and S_y are their associated standard errors. Formula (10) 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.

Table 3: Generalized Variance Parameters for Use With Weights

PERSONS[1]	a	b
Total or White		
15+ Program Participation and Benefits (3)		
Both Sexes	-0.0001054	19,020
Male	-0.0002203	19,020
Female	-0.0002022	19,020
15+ Income and Labor Force (4)		
Both Sexes	-0.0000360	6,485
Male	-0.0000751	6,485
Female	-0.0000689	6,485
All Others[2] (5)		
Both Sexes	-0.0001016	23,583
Male	-0.0002092	23,583
Female	-0.0001977	23,583
Black		
Poverty (1)		
Both Sexes	-0.0005839	16,224
Male	-0.0012462	16,224
Female	-0.0010984	16,224
All Others (2)		
Both Sexes	-0.0003140	8,724
Male	-0.0006701	8,724
Female	-0.0005907	8,724
HOUSEHOLDS/Families/Unrelated Individuals		
Total or White	-0.0000881	8,014
Black	-0.0005521	5,537

[1]For cross-tabulations, apply the parameters of the category showing the smaller number in parentheses.

[2]These parameters are to be used for all tabulations not specifically covered by any other category in this table, e.g., for retirement and pension tabulations, for 0+ benefits, 0+ income, and 0+ labor force tabulations.

Attachment 3

SIPP 1984 PANEL ANNUAL WEIGHTS FILE DATA DICTIONARY

MNEUMONIC	SIZE	BEGIN POSITION	ITEM
SU-ID	9	1	Scrambled identifier
PP-ENTRY	2	10	Person entry address identifier
PP-PNUM	3	12	Person number
DEC83WGT	10	15	Weight controlled to December 1983 population estimates
MAR85WGT	10	25	Weight controlled to March 1985 population estimates
FILLER	6	35	Zero filler