SOURCE AND ACCURACY STATEMENT FOR THE 1989 PUBLIC USE FILES FROM THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

SOURCE OF DATA

The data were collected in the 1989 panel of the Survey of Income and Program Participation (SIPP). The SIPP universe is the noninstitutionalized resident population living in the United States. The population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

The 1989 panel of the SIPP sample is located in 230 Primary Sampling Units (PSUs) each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of two 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 containing clusters of four LQs was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panel.

In jurisdictions that don't issue building permits or have incomplete addresses, small land areas were sampled and expected clusters of four LQs within were listed by field personnel and then subsampled. In addition, sample LQs were selected from a supplemental frame that included LQs identified as missed in the 1980 census.

Approximately 17,500 living quarters were originally designated for the 1989 panel. For Wave 1 of the panel, interviews were obtained from occupants of about 11,900 of the 17,500 designated living quarters. Most of the remaining 5,600 living quarters in the panel were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. However, approximately 1,000 of the 5,600 living quarters in the 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 92 percent of all eligible living quarters participated in the first interview of the panel.

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

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 1 year beginning in February 1989[1]. 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.

A unique feature of the SIPP design is overlapping panels. The overlapping design allows panels to be combined and essentially doubles the sample sizes. Selected interviews for the 1989 panel can be combined with interviews from the 1988 panel. Information necessary to do this is included later in this statement.

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 1989 and 1988 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 1989 and 1988 panels respectively. For example, Wave 1 rotation group 2 of the 1989 panel was interviewed in February 1989 and data for the reference months October 1988 through January 1989 were collected.

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 factor was applied to the weight of every occupant of interviewed households to account for persons in noninterviewed occupied households which were eligible for the sample. (Individual nonresponse within partially interviewed households was treated with imputation. No special adjustment was made for noninterviews in group quarters.)

---

[1] Panels are usually about 2½ years long, but the 1989 panel was shortened due to budget cuts.

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.

The Bureau has used complex techniques to adjust the weights for nonresponse. For a further explanation of the techniques used, see the Nonresponse Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census, November 1988, Working paper 8823, by R. Singh and R. Petroni. The success of these techniques in avoiding bias is unknown. An example of successfully avoiding bias can be found in "Current Nonresponse Research for the Survey of Income and Program Participation" (paper by Petroni, presented at the Second International Workshop on Household Survey Nonresponse, October 1991).

An additional stage of adjustment to persons' weights was performed to reduce the mean square errors of the survey estimates. This was accomplished by ratio adjusting the sample estimates to agree with monthly Current Population Survey (CPS) type estimates of the civilian (and some military) noninstitutional population of the United States by demographic characteristics including age, race, and sex as of the specified date. The CPS estimates by age, race, and sex were themselves brought into agreement with estimates from the 1980 decennial census which have been adjusted to reflect births, deaths, immigration, emigration, and changes in the Armed Forces since 1980. In addition, SIPP estimates were controlled to independent Hispanic controls and an adjustment was made so that husbands and wives within the same household were assigned equal weights. All of the above adjustments are implemented for each reference month and the interview month.

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 1988. 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, December 1988 data is only available from rotations 2, 3, and 4 for Wave 1 of the 1989 panel (See table 3), 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 with four rotations worth of data are constructed from a wave file, factors greater than 1 must be applied. However, when core data from consecutive waves are used together, data from all four rotations may be available, in which case the factors are equal to 1.

These tapes contain no weight for characteristics that involve a persons's or household's status over two or more months (e.g., number of households with a 50 percent increase in income between November and December 1989).

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 5. (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 5 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 5 is 1.0. In all other states, the cases identified as not in the metropolitan subsample (METRO=2) are a mixture of non-metropolitan and metropolitan households. Only an indirect method of estimation is available: first compute an estimate for the total population, then subtract the estimates for the metropolitan population. The results of these tabulations will be slightly biased.

Combined Panel Estimates. Both the 1989 and 1988 panels provide data for October 1988-December 1989. Thus, estimates for these time periods may be obtained by combining the corresponding panels. However, since the Wave 1 questionnaire differs from the subsequent waves' questionnaire, we recommend that estimates not be obtained by combining Wave 1 data of the 1989 panel with data from another panel. In this case, use the estimate obtained from either panel. Additionally, even for other waves, care should be taken when combining data from two panels since questionnaires for the two panels differ somewhat and since the length of time in sample for interviews from the two panels differ.

Combined panel estimates may be obtained either (1) by combining estimates derived separately for the two panels or (2) by first combining data from the two files and then producing an estimate.

1. <u>Combining Separate Estimates</u>

   Corresponding estimates from two consecutive year panels can be combined to create joint estimates by using the formula

$$\hat{\vartheta} = W\hat{\vartheta}_1 + (1-W)\hat{\vartheta}_2 \qquad \text{(A)}$$

$\hat{\vartheta}$ = joint estimate (total, mean, proportion, etc);

$\hat{\vartheta}_1$ = estimate from the earlier panel;

$\hat{\vartheta}_2$ = estimate from the later panel;

W = weighting factor of the earlier panel.

To combine the 1988 and 1989 panels use a W value of 0.509 unless one of the panels contributes no information to the estimate. In that case, the panel contributing information receives a factor of 1. The other receives a factor of zero.

2. Combining Data from Separate Files

Start by first creating a file containing the data from the two panel files. Apply the weighting factor, W, to the weight of each person from the earlier panel and apply (1-W) to the weight of each person from the later panel. Estimates can then be produced using the same methodology as used to obtain estimates from a single panel.

Illustration for computing combined panel estimate.

Suppose SIPP estimates for Wave 5, 1988 panel show there were 441,000 households with monthly May income above $6,000. Also, suppose SIPP estimates for Wave 2, 1989 panel show there were 435,000 households with monthly May income above $6,000. Using formula (A), the joint level estimate is

$$\hat{\vartheta} = (0.509)(441{,}000) + (0.491)(435{,}000) = 438{,}000$$

ACCURACY OF ESTIMATES

SIPP estimates are based on a sample; they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaire, instructions,

and enumerators. There are two types of errors possible in an estimate based on a sample survey: nonsampling and sampling. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error. Found in the next sections are descriptions of sources of SIPP nonsampling error, followed by a discussion of sampling error, its estimation, and its 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 the following: collection such as in recording or coding the data, processing the data, estimating values for missing data; biases resulting from the differing recall periods caused by the interviewing pattern used; and undercoverage. Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers. More detailed discussions of the existence and control of nonsampling errors in the SIPP can be found in the SIPP Quality Profile.

Undercoverage in SIPP results from missed living quarters and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for nonBlacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group. Further, the independent population controls used have not been adjusted for undercoverage in the Census.

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

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.

USES AND COMPUTATION OF STANDARD ERRORS

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

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

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

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

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

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

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

fact, the same, there is a 10 percent chance of concluding that they are different.

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

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

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 for the SIPP. To derive standard errors that would be applicable to a wide variety of estimates and could be prepared at a moderate cost, a number of approximations were required. Estimates with similar standard error behavior were grouped together and two parameters (denoted "a" and "b") were developed to approximate the standard error behavior of each group of estimates. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. These "a" and "b" parameters vary by characteristic and by demographic subgroup to which the estimate applies. Table 6 provides base "a" and "b" parameters to be used for the 1989 panel estimates.

The factors provided in table 7 when multiplied by the base parameters of table 6 for a given subgroup and type of estimate give the "a" and "b" parameters for that subgroup and estimate type for the specified reference period. For example, the base "a" and "b" parameters for total number of households are -0.0001144 and 10,623, respectively. For Wave 1 the factor for October 1988 is 4 since only 1 rotation month of data is available. So, the "a" and "b" parameters for total household income in October 1988 based on Wave 1 are -0.0004576 and 42,492, respectively. Also for Wave 1, the factor for the first quarter of 1989 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 number of households in

the first quarter of 1989 are -0.0001398 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 parameter for computation of approximate standard errors are given in the following sections.

For those users who wish further simplification, we have also provided general standard errors in tables 8 through 11. Note that these standard errors only apply when data from all four rotations are used and must be adjusted by a factor from table 6. 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.

For the 1988, 1989 combined panel parameters, multiply the parameters in table 6 by a factor of 0.5232. The factors provided in table 12 adjust parameters for the number of rotation months available for a given estimate. These factors, when multiplied by the combined panel parameters derived from table 6 for a given subgroup and type of estimate, give the "a" and "b" parameters for that subgroup and estimate type for the specified combined reference period.

Table 13 provides base "a" and "b" parameters for calculating 1989 topical module variances. Table 14 provides base "a" and "b" parameters for computing the 1988, 1989 combined panel topical module variances.

Procedures for calculating standard errors for the types of estimates most commonly used are described below. Note specifically that these procedures apply only to reference month estimates or averages of reference month estimates. Refer to the section "Use of Weights" for a more detailed discussion of the construction of estimates. Stratum codes and half sample codes are included on the tapes to enable the user to compute the variances directly by methods such as balanced repeated replications (BRR). William G. Cochran provides a list of references discussing the application of this technique. (See Sampling Techniques, 3rd Ed., New York: John Wiley and Sons, 1977, p. 321.)

Standard errors of estimated numbers. The approximate standard error, s_x, of an estimated number of persons, households, families, unrelated individuals and so forth, can be obtained in

two ways. Both apply when data from all four rotations are used to make the estimate. However, only the second method should be used when less than four rotations of data are available for the estimate. Note that neither method should be applied to dollar values.

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

$$s_x = fs \tag{1}$$

where f is the appropriate "f" factor from table 6, and s is the standard error on the estimate obtained by interpolation from table 8 or 9. Alternatively, s_x may be approximated by the formula

$$s_x = \sqrt{ax^2 + bx} \tag{2}$$

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

Illustration.

Suppose SIPP estimates for Wave 1 of the 1989 panel show that there were 472,000 households with monthly household income above $6,000. The appropriate parameters and factor from table 6 and the appropriate general standard error from table 8 are

a = -0.0001144 b = 10,623 f = 1.00 s = 71,000

Using formula 1, the approximate standard error is

$$s_x = 71{,}000$$

Using formula 2, the approximate standard error is

$$\sqrt{(-0.0001144)(472{,}000)^2 + (10{,}623)(472{,}000)} = 70{,}600$$

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.

Illustration for computing standard errors for combined panel estimates.

Suppose the combined SIPP estimate for total number of households for Wave 5, 1988 panel and Wave 2, 1989 panel was 92,398,000. The combined panel parameters for total households are obtained by multiplying the appropriate "a" and "b" values from table 6 by g = 0.5232 and the appropriate factor from table 12. The 1989 parameters and factors are a = -0.0001144, b = 10,623, g = 0.5232 and factor = 1.0000, respectively. Thus, the combined panel parameters are a = -0.0000599 and b = 5,558. Using formula 2, the approximate standard error is

$$S = \sqrt{(-0.0000599)(92,398,000)^2 + (5558)(92,398,000)} = 46,500$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

When forming combined estimates using formula (A) from the section on combined panel estimates, s^2, given by formula (4), should be calculated by forming a distribution for each panel. The range of values for the item will be divided into intervals. Combined estimates for each interval can be obtained using formula (A). Formula (4) can be applied to the combined distribution. To calculate $\bar{x}$ and s^2 given by formula (5), replace x_i by Wx_i for x_i from the earlier panel and $(1-W)x_i$ for x_i from the later panel.

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 1989 is given in table 15.

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 6, the estimated standard error of a mean $\bar{x}$ 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:

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

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

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

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

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

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

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

Alternatively, it may be approximated by the formula

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

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

Illustration.

Suppose that, in the month of January 1989, 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 from table 6 and a factor of 1 for the month of January 1989 from table 7, the approximate standard error is

$$\sqrt{\frac{11,565}{(16,812,000)}\ (6.7)\ (100-6.7)} = 0.66\ \textit{percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 5.7 to 7.7 percent.

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

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

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

$$p_I = 100\ (\hat{P}_A\ \overline{X}_A\ /\ \overline{X}_N)$$

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

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

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

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

<u>Illustration.</u>

Suppose that in January 1991, 9.8% of the households own rental property, the mean value of rental property is $72,121, the mean value of assets is $78,734, and the corresponding standard errors are 0.31%, $5799, and $2867. In total there are 86,790,000 households. Then, the percent of all household assets held in rental property is

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

Using formula (9) the appropriate standard error is

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

$$= 0.008$$

$$= 0.8\%$$

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

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

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 between the characteristics estimated by x and y is zero. If the correlation is really positive (negative), then this assumption will tend to cause overestimates (underestimates) of the true standard error.

Illustration.

Suppose that SIPP estimates show the number of persons age 35-44 years with monthly cash income of $4,000 to $4,999 was 3,186,000 in the month of January 1989 and the number of persons age 25-34 years with monthly cash income of $4,000 to $4,999 in the same time period was 2,619,000. Then, using parameters from table 6 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

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

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

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

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

of the item of interest. Interpolation is used as follows. The quantity of the item such that "p" percent have more of the item is

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

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.

<u>Illustration.</u>

To illustrate the calculations for the sampling error on a median, we return to table 15. The median monthly income for this group is $2,158. The size of the group is 39,851,000.

1. Using formula 8, the standard error of 50 percent on a base of 39,851,000 is about 0.7 percentage points.

2. Following step 2, the two percentages of interest are 49.3 and 50.7.

3. By examining table 15, 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 bound of a 68% confidence interval for the median is

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

Also by examining table 14, 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 use Pareto interpolation for this case. So the lower bound of a 68% confidence interval for the median is

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

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