SOURCE AND ACCURACY STATEMENT FOR THE FIRST SURVEY OF PROGRAM DYNAMICS LONGITUDINAL FILE

DATA COLLECTION AND ESTIMATION

Source of Data

The Survey of Program Dynamics (SPD) universe is the noninstitutionalized resident population living in the United States. This population includes people (including children) 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 people, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. In addition, 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, people who were at least 15 years of age at the time of the interview were eligible to be asked income and job experience.

The calendar year data for 1996 were collected during April, May, and June of 1997 as part of the SPD Bridge Survey. Likewise, the calendar year data for 1997 were collected during May, June, and July of 1998 as part of the SPD 1998 Survey. The SPD Bridge calendar and SPD 1998 calendar year files consist principally of the calendar year data for 1996 and 1997, respectively. The first SPD longitudinal file (also known as the SPD 1998 longitudinal file) longitudinally combines the data from the SIPP Panels 1992 and 1993, and the SPD Bridge file and SPD 1998 calendar year files.

The goal of SPD program is to provide policy makers a survey to assess the effects of the recent welfare reforms and how these reforms interact with each other, and with employment, income and family circumstances. The SPD program eventually spans from the pre-reform through the post-reform period, 1992-2002. In order to obtain information about past economic history, employment, income, and program participation, two retired SIPP panels 1992 and 1993 were chosen as the SPD sample. A full potential of the SPD data is generally achieved when using the first SPD longitudinal file until the release of other subsequent SPD longitudinal files.

The SPD Bridge Survey data was collected in 1997 and intended to be a connection run between the SIPP and SPD data. Data that was merged from the previous SIPP surveys, the SPD Bridge survey, and the subsequent SPD survey (for example, SPD 1998) should give us the necessary pre-reform and post-reform information for sampled households.

Background of SIPP 1992 and 1993 Panels and SPD Bridge Survey

The 1992 and 1993 SIPP panel samples were located in 284 Primary Sampling Units (PSUs), each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of two or four 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 sub-sampled. In addition, sample LQs were selected from supplemental frames that included LQs identified as missed in the 1980 census and group quarters (GQs).

At the time of the initial visit of the SIPP panels, the occupants of about 19,600 living quarters were interviewed for the 1992 panel and 19,900 were interviewed for the 1993 panel. This accounts for approximately 72% (1992) and 73% (1993) of the LQs originally designated for the SIPP samples. Approximately 21% (1992) and 20% (1993) of the designated LQs were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. The remainder, approximately 2000 LQs, were not interviewed because the occupants refused to be interviewed, could not be found at home, were temporarily absent, or otherwise unavailable. Thus, occupants of about 91% of all eligible LQs participated in the first interview of the 1992 and 1993 SIPP panels.

For the remaining nine interviews, only original sample people (those in Wave 1 sample households and interviewed in Wave 1) and people living with them were eligible to be interviewed. With certain restrictions, original sample people were to be followed even if they moved to a new address. When original sample people moved without leaving a forwarding address or moved to extremely remote parts of the country and no telephone number was available, additional non-interviews resulted.

The 1992 10-Wave Longitudinal File consists of data collected from February 1992 to April 1995. Data for up to 39 reference months are available for people on this file. The 1993 Nine-Wave Longitudinal File consists of data collected from February 1993 to January 1996. Data for up to 36 reference months are available for people on this file.

Tables 1a-1c indicate the interview months for the collection of data from the 1992 Ten-Wave Longitudinal File, 1993 Nine-Wave Longitudinal File, and the 1998 SPD File. For the SIPP, a person was classified as interviewed or non-interviewed based on the following definitions. (Note: A person may be classified differently for calculating different weights). Interviewed sample people (including children) were defined to be: *those for whom self, proxy, or imputed responses were obtained for each month of the appropriated longitudinal period.*

The months for which people were deceased or residing in an ineligible address were identified on the file. Non-interviewed people were defined to be those for whom neither self nor proxy responses were obtained for one or more months of the appropriate longitudinal period (excluding imputed people and people who died or moved to an ineligible address).

It is estimated that roughly 56,300 (1992) and 57,200 (1993) people were initially designated in the sample for the SIPP. Approximately 51,100 (1992) and 51,900 (1993) people were interviewed in Wave 1; while the balance, residing in the 4,000 (1992 and 1993 combined) living quarters not interviewed at Wave 1 remained anonymous and became the initial source of the person non-response in the weighting procedures. For panel weighting, the eligible sample is considered to be all people initially classified as interviewed with a person non-response rate of 25 percent (1992) and 24 percent (1993). The longitudinal file contains approximately 59,700 (1992) and 62,700 (1993) people in all. This includes the Wave 1 interviewed people and about 8,600 (1992) and 10,600 (1993) people who entered survey households during the panel through births, marriages, and other reasons. Some respondents did not respond to some of the questions; therefore, item non-response rates, especially for sensitive income and money related items, are higher than the person non-response rates given above.

We define the SPD Bridge sample cohort as people in the 1992 and 1993 SIPP panels that were in an interviewed household in the last wave. However, only people considered interviewed (self or proxy or imputed response) longitudinally in SIPP *and* considered interviewed in SPD Bridge were eligible to go on further for the SPD 1998 Survey, since the SPD 1998 Survey was carried out only as a part of a long-term *longitudinal* data collection effort for the SPD. In addition, due to budget constraints, the SPD 1998 Survey was also subject to a sample cut based on the sub-sampling procedure described in the section below.

1998 SPD Sub-sampling

Due to budget constraints, the SPD 1998 Survey did not visit all 35,000 Bridge households. The budget only allowed for SPD to visit 21,000 households. Roughly 19,100 cases were sampled in this operation, since we needed to account for an expected 12.5 percent non-response and a growth of 10 percent of the total sample size due to household spawning.

In the sub-sampling (sample cut), the SPD Bridge sample households were demographically divided into six strata as shown in the table at the end of this section. The stratification was performed using the household information collected from the SPD Bridge. In each stratum, the households were sampled independently with the sampling rate as provided in the table below. As indicated among sampling rates in this table, the low income sample households were generally not subjected to the sample cut at all.

As a result of the sample cut, the actual number of the households selected for interview was 19,288. Among the 19,288 households selected for interview and their spawned households, 16,395 households were interviewed.

Strata	Description	Sampling Rate	Designated Number	Projected Interviews
1	Households where the primary family or the primary individual has a total family income below 150% of the poverty threshold	1-in-1	6,182	5,950
2	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are children under 18	1-in-1	1,075	1,035
3	Households where the primary family or the primary individual has a total family income above 200% of the poverty threshold and there are children under 18	1-in-1.11	6,623	6,375
4	Households where the primary family or the primary individual has a total family income between 150% and 200% of the poverty threshold and there are no children under 18	1-in-1.22	1,461	1,406
5	Households in the balance	1-in-3.70	3,707	3,568
6	Households entirely institutionalized (Outcome code = 228)	1-in-3.70	81	DK
Total			19,129	18,334

ESTIMATION

In the estimation procedure described below, all the sample people classified as longitudinally interviewed for the entire longitudinal period spanning the SIPP, SPD Bridge, and SPD 1998 were assigned positive final longitudinal weights in the first SPD longitudinal while all those classified otherwise were assigned zero final longitudinal weights, except for children aged six or less if spawned in the SIPP Panel 1992 and aged five or less if spawned in the SIPP Panel 1993. If the child's designated parent (biological or adopted or guardian) is an original sample person then assign the child's weight to be the same as the designated parent's weight, otherwise assign the child's weight as zero. In the first SPD longitudinal file, the weights of these children were already assigned accordingly. A description of the weighting procedure and corresponding terminologies for calculating the final longitudinal weights of the sample people in the first SPD longitudinal file were provided earlier in the subsection "Weighting" (of the section "File Information").

Estimation of Person Characteristics

For the estimation of the person characteristics in the SPD universe, the final longitudinal weights of the sample people in the first SPD longitudinal file can be used. Hereinafter, the term "*the final longitudinal* *weights of the sample people in the first SPD longitudinal file*" will be simply referred to as "*the longitudinal person weights*." Some basic types of longitudinal estimates (using the first SPD longitudinal file) can be constructed using the longitudinal person weights are described below in terms of estimated numbers.

1. The number of people who have ever experienced a characteristic or situation during a given period of time (for example, the number of people who experience unemployment during 1997). To construct such an estimate, sum the weights over all people who possessed the characteristic of interest at some point during the time period of interest.

2. The amount of a characteristic accumulated by people during a given time period (for example, the amount of unemployment compensation received by unemployed people during 1997). To construct such an estimate, compute the product of the weight times the amount of the characteristic and sum this product over all appropriate people.

3. The average number of consecutive months or years of possession of a characteristic (i.e., the spell length for a characteristic.) For example, one could estimate the average spell of unemployment that elapsed before a person found a new job. (*Note that the first SPD longitudinal file provides the employment data only in terms of week numbers with and without employment in a given year. Thus, for calculation the average unemployment spell length in a time period of interest, the data user needs to match the sample person's record back to the one on the SIPP longitudinal file to determine the number of spells in the time period and/or needs to make some justifiable approximation on the number of unemployment spells within the time period of interest.*) To construct such an estimate, first identify the sample persons possessing the characteristic at some point during the time period of interest. Then, create two sums of these (longitudinal person) weights: Sum 1 is sum of the products of the weights times the number of months (or years) the spell lasted, and Sum 2 is the sum of the weights only. The average spell length in months (or years) is given by Sum 1 divided by Sum 2. A person who experienced two spells during the time period of interest would be treated as two persons and appear twice in Sum 1 and Sum 2. An alternate method of calculating the average can be found in the section "Standard Error of a Mean or an Aggregate."

   Note that spells extending before or after the time period of interest are cut off (censored) at the boundaries of the time period. If they are used in estimating average spell length, a downward bias will result.

4. The number of year-to-year changes in the status of a characteristic (i.e., number of transitions) summed over every set of two consecutive years during the time of interest. To construct such estimate, sum the longitudinal person weights each time a change is reported between two consecutive years during the time period of interest. For example, to estimate the number of

persons who changed from receiving any public assistance in 1996 to not receiving in 1997 add together the longitudinal person weights of each person who had such a change.

5. Yearly estimates of a characteristic average over a number of consecutive years. For example, we could estimate the yearly average number of food stamp recipients over 1996 and 1997. To construct such an estimate, first form an estimate for each year in the time period of interest by summing up the longitudinal person weights of those possessed the characteristic of interest. Then sum the yearly estimates and divide by the number of years in the time period of interest.

ACCURACY OF ESTIMATES

SPD 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: non-sampling and sampling. We are able to provide estimates of the magnitude of SPD sampling error, but this is not true of non-sampling error. The next sections describe sources of SPD non-sampling error, followed by a discussion of sampling error, its estimation, and its use in data analysis.

Note that estimates from this sample for individual states are subject to very high sampling errors 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.

Non-sampling Errors

Non-sampling errors can be attributed to many sources, for examples, inability to obtain information about all cases in the sample, difficulties in precisely stating some definitions, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, and the following errors made. These errors generally include 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 rotation pattern used, and under coverage. Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers.

Under-coverage in SPD results from missed living quarters and missed people within sample households. It is known that under coverage varies with age, race, and gender. Generally, under-coverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-gender population controls (benchmark estimates) partially corrects for the bias due to survey under-coverage. However, biases exist in the estimates to the extent that people in missed households or missed people in interviewed households have characteristics different from those of interviewed people in the same age-race-gender group. In addition, the independent population controls used have not been adjusted for under-coverage in the decennial census. The Census Bureau has used complex techniques to adjust the weights for non-response. For an explanation of the techniques used, see the “Non-response Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census,” November 1988, Working Paper 8823, by R. Singh and R. Petroni. An example of successfully avoiding bias can be found in "Current Non-response Research for the Survey of Income and Program Participation" (paper by Petroni, presented at the Second International Workshop on Household Survey Non-response, October 1991). The procedure for calculating the longitudinal person weights on the first SPD longitudinal file was derived based on such complex techniques.

Unlike SIPP data that can be analyzed from a cross-sectional or longitudinal view point, the SPD data are solely longitudinal and must be used as such. Thus, the income and poverty estimates in a given single year may not be comparable with those from other surveys such as the Current Population Survey (CPS) and the SIPP. This is principally attributable to the fact that the sample per se and the longitudinal person weights on the first SPD longitudinal file essentially represents just the cohort of people around March 1993. As the SPD sample aged more, it will become less adequate to represent the more current population (say, the 1998 population). In addition, the high non-response rate (roughly 50 percent) in the SPD may reduce the degree of the effectiveness of the non-interview adjustment process to fully compensate for differential attrition. Note that the non-response rate has three components: 27 percent sample loss inherited from the SIPP, 14 percent occurred from the SPD Bridge interview, and an additional 9 percent occurred at the SPD 1998 interview.

Comparability with Other Estimates

Caution should be exercised when comparing data from this file with data from SIPP publications or with data from other surveys, such as Current Population Survey (CPS). The comparability problems are caused by such sources as the seasonal patterns for many characteristics, different non-sampling 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 non-sampling 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 90 percent of the intervals from 1.645 standard errors below the estimate to 1.645 standard errors above the estimate would include the average result of all possible samples.

2. Approximately 95 percent of the intervals from 1.960 standard errors below the estimate to 1.960 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 the population characteristics are identical versus 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 -1.645 times s_{DIFF} and +1.645 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 -1.645 times s_{DIFF} or larger than +1.645 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. We recommend that users report only those differences that are significant at the 10 percent level or better. 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.

Caution 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. Also, non-sampling error in one or more of the small number of cases providing the estimate can cause large relative error in that particular estimate. Therefore, care must be taken in the interpretation of small differences since even a small amount of non-sampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Standard Error Parameters

Most SPD estimates have greater standard errors than those obtained through a simple random sample because clusters of living quarters are sampled for the SIPP, SPD Bridge, and SPD 1998. 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. The *a* and *b* parameters are also known as “generalized variance parameters.” For the first SPD longitudinal file, the *a* and *b* parameters for various groups of the populations are provided in Table 3. Hereinafter, *the a and b parameters in Table 3* will be referred to as *the base a and b parameters*.

Computation of Standard Error Parameters

In this section we discuss the adjustment of base *a* and *b* parameters (Table 3) to provide *a* and *b* parameters appropriate for each type of longitudinal described in the section "Estimation of Person Characteristics." Later sections will discuss the use of the adjusted parameters in various formulas to compute standard errors of estimated numbers, percents, averages, etc. Table 3 provides the base *a* and *b* parameters needed to compute the approximate standard errors for estimates.

The creation of appropriate *a* and *b* parameters for the types of estimates discussed in the section “Estimation of Person Characteristics” is described below. It is assumed that the full sample is used for the estimation.

1. The number of people who have ever experienced a characteristic during a given time period. The appropriate *a* and *b* parameters are taken directly from Table 3 (the base *a* and *b* parameters). The choice of parameter depends on the characteristic of interest and the demographic subgroup of interest.

2. Amount of a characteristic accumulated by people during a given time period. The appropriate *a* and *b* parameters are also taken directly from Table 3.

3. The average number of consecutive months or years of possession of a characteristic per spell (i.e., the average spell length for a characteristic) during a given time period. Start with the appropriate base *a* and *b* parameters from Table 3. The parameters are then inflated by an additional factor, *g* to account for persons who experience multiple spells during the time period of interest. The *g* factor is computed by Formula 1 below.

$$g = \frac{\sum_{i=1}^{n} m_i^2}{\sum_{i=1}^{n} m_i} \qquad (1)$$

   where there are *n* persons with at least one spell and m_i is the number of spells experienced by person *i* during the time period of interest.

4. The number of years-to-year changes in the status of a characteristic (i.e., number of transitions) summed over every set of two consecutive years during the time period of interest. Obtain a set of adjusted *a* and *b* parameters exactly as just described in 3, then multiply these parameters by an additional factor of 2.0. The factor of 2.0 is based on the assumption that each spell produces two transitions within the time period of interest.

5. Yearly estimates of characteristic averaged over a number of consecutive years. Appropriate base *a* and *b* parameters are taken directly from Table 3.

Standard Errors of Estimated Numbers

The approximate standard error s_x of an estimated number *x* of people, families and so forth, can be obtained by using Formula 2 provided below.

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

Here *a* and *b* are the standard error parameters associated with the particular type of characteristic for the appropriate longitudinal time period. For the analysis using the SPD data on either the 1998 longitudinal file or the 1998 calendar year file, the *a* and *b* parameters are provided in Table 3.

An illustration would be to suppose that using 1998 SPD data, the estimate of the number of people ever receiving Social Security since 1993 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* = -0.0000812, *b* = 13,858. Using Formula (2), the approximate standard error s_x is

$$s_x = \sqrt{(-0.0000812)(34{,}122{,}000)^2 + (13{,}858)(34{,}122{,}000)} = 687{,}650 \quad \textit{people}$$

The 90-percent confidence interval as shown by the data is from 32,990,816 to 35,253,184. 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. Similarly, the 95-percent confidence interval as shown by the data is from 32,774,206 to 35,469,794 and we could conclude that the average estimate derived from all possible samples lies within this interval.

Standard Error of a Mean or an Aggregate

A mean $\overline{x}$ is defined here to be the average quantity of some characteristic (other than the number of people, families, or households) per person, family, or household. An aggregate *k* is defined to be the total quantity of some characteristic summed over all units in a sub-population. For example, a mean could be the average annual income of females age 25 to 34. The standard error $s_{\overline{x}}$ of a mean can be approximated by Formula 3 and the standard error s_k of an aggregate can be approximated by Formula 4. Because of the approximations used in developing Formulas 3 and 4, an estimate of the standard error of the mean or aggregate obtained from these formulas will generally underestimate the true standard error. The formula used to estimate the standard error $s_{\overline{x}}$ of a mean $\overline{x}$ is

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

where *y* is the base s^2 is the estimated population variance of the characteristic and *b* is the standard error parameter associated with the type of the characteristic. The standard error s_k of an aggregate *k* is

$$s_k = \sqrt{by\,s^2} \tag{4}$$

The population variance s^2 may be estimated by one of two methods: the first method uses data that has been grouped into intervals, the second method uses ungrouped data. The second method is recommended because it is more precise. However, the first method will be easier to implement if grouped data are already being used as part of the analysis. In both methods, let x_i denote the value of the characteristic for the i^{th} person.

To use the first method, the range of values for the characteristic is divided into c intervals, where the lower and upper boundaries of interval j are Z_{j-1} and Z_j, respectively. Each person is placed into one of the c groups such that the value of the characteristic, x_i is between Z_{j-1} and Z_j. The estimated population variance, s^2 is then given by Formula 5 below.

$$s^2 = \sum_{j=1}^{c} p_j m_j^2 - \bar{x}^2 \qquad (5)$$

where p_j is the estimated proportion of people in group j (based on weighted data), and m_j is given by the equation below.

$$m_j = \frac{Z_{j-1} + Z_j}{2}, \quad for \quad j = 1, \ 2, \ \ldots, \ c$$

The most representative value of the characteristic in group j is assumed to be m_j. If group c is open-ended, that is, no upper interval boundary exists, then an approximate value for m_c is by the equation below.

$$m_c = \left(\frac{3}{2}\right) Z_{c-1}$$

The mean $\bar{x}$ can be obtained using Formula 6 below.

$$\bar{x} = \sum_{j=1}^{c} p_j m_j \qquad (6)$$

In the second method, the estimated population variance s^2 is given Formula 7 below.

$$s^2 = \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \bar{x}^2 \qquad (7)$$

where there are n sample people with the characteristic of interest and w_i is the final weight for person i. The mean $\bar{x}$ can be obtained from Formula 8 below.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (8)$$

Note that, by definition, y (the size of the base) in Formulas 3 and 4 can be obtained from the equation below.

$$y = \sum_{i=1}^{n} w_i$$

An illustration of Method 1 would be to suppose that the 1997 distribution of annual incomes is given in Table 2 for people aged 25 to 34 who were employed for all 12 months of 1997. The mean annual cash income from Formula 8 is

$$\bar{x} = \frac{1{,}371}{39{,}851}(2{,}500) + \frac{1{,}651}{39{,}851}(6{,}250) + \quad \ldots \quad + \frac{1{,}493}{39{,}851}(105{,}000) = \$26{,}717$$

Using Formula 7 and the mean annual cash income of $26,717 the estimated population variance, s^2 is

$$s^2 = \frac{1{,}371}{39{,}851}(2{,}500)^2 + \frac{1{,}651}{39{,}851}(6{,}250)^2 + \quad \ldots \quad + \frac{1{,}493}{39{,}851}(105{,}000)^2 = 468{,}331{,}633$$

The appropriate *b* parameter from Table 3 is 7,566. Now, using Formula 3, the estimated standard error of the mean is

$$s_{\bar{x}} = \sqrt{\frac{7{,}566}{39{,}851{,}000}(468{,}331{,}633)} \quad = \quad \$298$$

An illustration of Method 2 would be to suppose that we are interested in estimating the average length of spell of receiving public assistance during 1992-1994 (just prior to the Welfare Reform) for a given sub-population. Also, suppose there are only 10 sample persons in the sub-population who were public assistance recipients. (This example is for illustrative purpose only; in reality, 10 sample units or cases would be too few for a reliable estimate.) The number of consecutive years of receiving public assistance during 1992-1994 are given for each sample persons in the table below. *(Caveat - In reality, only the total number of months of receiving public assistance in a given year is available in the first SPD longitudinal file. Thus, the actual number of spells in a time period of interest is not known or equivalently the actual spell length is not known. Consequently, to use the such data in the first SPD longitudinal file for assessing average spell length in a time period of interest, it is the responsibility of the data user to match back the sample person record to the one in the SIPP longitudinal file to determine the number of spells in the time period of interest and/or make his/her own justifiable assumption on what should be the number of spells in the time period of interest given the total number of months in a year that a sample person possessed a spell characteristic, e.g., receiving public assistance.)*

Sample Person Number	Number of Spells During 1992-1994	Spell Lengths in Months	Final Longitudinal Weight
1	2	12, 6	5300
2	1	2	7100
3	1	5	4900
4	2	3, 6	6500
5	1	13	4700
6	1	14	5500
7	2	3, 6	4100
8	1	24	4200
9	1	6	4500
10	1	4	6100

Using Formula 8, the average spell $\overline{x}$ of receiving public assistance is estimated to be

$$\overline{x} = \frac{5300 \times 12 + 5300 \times 6 + 7100 \times 2 + \quad \ldots \quad + 6100 \times 4}{5300 + 5300 + 7100 + \quad \ldots \quad + 6100} = \frac{472800}{68800} = 6.872 \quad \textit{months}$$

The standard error $s_{\overline{x}}$ will be computed by Formula 3. First, estimate the population variance s^2 by Formula 7

$$s^2 = \frac{5300 \times 12^2 + 5300 \times 4^2 + 7100 \times 2^2 + \ldots + 6100 \times 4^2}{5300 + 5300 + 7100 + \ldots + 6100} - 6.872^2 = 41.92 \quad \textit{months}^2$$

Next, the base *b* parameter from Table 3 is 14601. To account for the multiple number of spells during 1992-1994 of three sample persons (two spells for Sample Persons 1, 4, and 7), multiply the base *b* parameter by a factor *g* computed from Formula 1 as shown below.

$$g = \frac{2^2 + 1 + 1 + 2^2 + 1 + 1 + 2^2 + 1 + 1 + 1}{2 + 1 + 1 + 2 + 1 + 1 + 2 + 1 + 1 + 1} = 1.462$$

Therefore, the adjusted b parameter is 14601×1.462 = 21347 and the standard error $s_{\overline{x}}$ of the mean is

$$s_{\overline{x}} = \sqrt{\frac{21347}{68800} \times 41.92} = 3.606 \quad \textit{months}$$

Standard Errors of Estimated Percentages

This section refers to the percentages of a group of people, families, or households possessing a particular attribute and to percentages of money or related concepts. 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 more than 50 percent. For example, the percent estimate of employed people is more reliable than the estimated number of employed people. 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 standard error of the corresponding percentage divided by 100.

There are two types of percentages commonly estimated. The first type is the percentage of people sharing a particular characteristic such as the percentage of people owning their own home or the percentage of 1996 food stamp recipients who were also receiving food stamps in 1997. The second type is the percentage of money or some similar concept held by a particular group of people or held in a particular form. Examples are the percentage of wealth held by people with high income and the percentage of annual income received by females.

For the percentage of people, the approximate standard error, $s_{x,p}$, of the estimated percentage, p, can be obtained by Formula 9 below.

$$s_{x,p} = \sqrt{\frac{b}{x} p(100 - p)} \qquad (9)$$

Here, x is the base of the percentage p is the percentage (0<p<100), and b is parameter for the numerator of the percentage calculation. For the analysis using the SPD data on either the 1998 longitudinal file or the 1998 calendar year file, the b parameters are provided in Table 3.
An illustration would be to suppose that, in 1997, an estimate of number of male aged 22 to 55 was 46,023,000. Among all the males in this age group, an estimate of 2.4 percent was unemployed. The b parameter associated with the numerator (the number of unemployed male) is 7,566 (from Table 3). Using Formula 9, the approximate standard error $s_{x,p}$ is

$$s_{x,p} = \sqrt{\frac{7{,}566}{46{,}023{,}000}(2.4)(1 - 2.4)} \quad = \quad 0.20\%$$

Consequently, the 90-percent confidence interval for the unemployment estimate is 2.1% to 2.7%.

To calculate the percentages of money, the formula is more complicated. A percentage of money will usually be estimated in one of two ways. It may be the ratio, p_M of two aggregates as defined in Formula 10 below.

$$p_M = 100\left(\frac{X_A}{X_N}\right) \qquad (10)$$

or it may be the ratio, p_M of two means with an adjustment, $\hat{p}_A$ for different bases as defined in Formula11 below.

$$p_M = 100\left(\frac{\overline{X}_A}{\overline{X}_N}\right)\hat{p}_A \tag{11}$$

where X_A and X_N in Formula 10 are aggregate money figures, $\overline{X}_A$ and $\overline{X}_N$ in Formula 11 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 way of estimating p_M (Formula 10 or 11), we estimate the standard error s_{p_M} of p_M using Formula 12 provided below.

$$s_{p_M} = \sqrt{\left(\frac{\hat{p}_A\overline{X}_A}{\overline{X}_N}\right)^2\left[\left(\frac{s_{\hat{p}_A}}{\hat{p}_A}\right)^2 + \left(\frac{s_{\overline{X}_A}}{\overline{X}_A}\right)^2 + \left(\frac{s_{\overline{X}_N}}{\overline{X}_N}\right)^2\right]} \tag{12}$$

where $s_{\hat{p}_A}$ is the standard error of $\hat{p}_A$, $s_{\overline{X}_A}$ is the standard error of $\overline{X}_A$ and $s_{\overline{X}_N}$ is the standard error of $\overline{X}_N$. To calculate $s_{\hat{p}_A}$, use Formula 9. The standard errors $s_{\overline{X}_A}$ and $s_{\overline{X}_N}$ are calculated using Formula 3.

Note that there is frequently some correlation among the characteristics estimated by $\hat{p}_A$, $\overline{X}_A$, and $\overline{X}_N$. These correlations, if present, will cause a tendency toward overestimates or underestimates, depending on the relative sizes of the correlations and whether they are positive or negative.

An illustration would be to suppose that, in 1998, an estimated 8.8% of males aged 16 and over was Black, the mean annual earning of these Black males was $15,456, the mean annual earning of all males aged 16 and over was $22,932, and the corresponding standard errors are 0.37 percent, $432, and $324, respectively. Then, the percent (p_M) of male earnings made by Blacks in 1998 per Formula 11 is

$$p_M = 100\left(\frac{15{,}456}{22{,}932}\right)(0.088) = 5.9\%$$

Using Formula 12, the approximate standard error, s_{p_M} is

$$s_{p_M} = \sqrt{\left(\frac{(0.088)(15{,}456)}{22{,}932}\right)^2\left[\left(\frac{0.0037}{0.088}\right)^2 + \left(\frac{432}{15{,}456}\right)^2 + \left(\frac{324}{22{,}932}\right)^2\right]} = 0.31\%$$

Standard Error of a Difference

The standard error s_{x-y} of a difference between two sample estimates *x* and *y* is equal to

$$s_{x-y} = \sqrt{{s_x}^2 + {s_y}^2 - 2rs_xs_y} \qquad (13)$$

where s_x and s_y are the standard errors of the estimates *x* and *y*. The estimates can be numbers, averages, percents, ratios, etc. The correlation between *x* and *y* is represented by *r* (*0 # r # 1*). If *r* is assumed to be zero and the true correlation is really positive (negative), then this assumption will result in a tendency toward overestimates (underestimates) of the true standard error.

An illustration would be to suppose that we are interested in the difference in the average annual number of adult males (aged 16 and above) versus adult females with annual cash income above $9,000 in 1998. An estimate of the number of adult people in this income bracket has been obtained for both males and females. For females, the estimate is 1,619,000. A similar estimate for males is 2,198,000. The difference in estimates is 579,000.

The standard error of the adult female estimate is computed next. The *a* and *b* parameters from Table 3 for females are -0.0000845 and 7,566, respectively. Based on Formula 2, the standard error, s_x of the female estimate is

$$s_x = \sqrt{(-0.0000845)(1{,}619{,}000)^2 + (7{,}566)(1{,}619{,}000)} = 109{,}672$$

Similarly, the a and b parameters from Table 3 for males are -0.0000936 and 7,566, respectively. Based on Formula 2, the standard error, s_y of the male estimate is

$$s_y = \sqrt{(-0.0000936)(2{,}198{,}000)^2 + (7{,}566)(2{,}198{,}000)} = 127{,}192$$

Now, the standard error of the difference is computed using the above two standard errors. The correlation *r* for this example is assumed to be zero. The standard error, s_{x-y} of the difference is computed by Formula 13 as shown below.

$$s_{x-y} = \sqrt{(109{,}672)^2 + (127{,}192)^2} = 167{,}946$$

Suppose that it is desired to test at the 10 percent significance level whether the number of adult males and females with monthly cash income above $9,000 were different in 1998, one can compare the difference of 579,000 to the product 1.645 x 167,946 = 276,271. Since the difference is larger than 1.645 times the standard error (s_{x-y}) of the difference, the data allow us to conclude that, in 1998, the number of adult males with annual cash income above $90,000 is significantly higher than the number of the adult females at the 10 percent confidence level.

Standard Error of a Median

The median quantity, X_{med} of some item (characteristic), *X* such as income for a given group of people, families, or households is that quantity such that at least half the group has as much or more and at least half the group has as much or less. The sampling variability of an estimated median $\hat{X}_{med}$ depends upon the form of the distribution of the item as well as the size of the group. To estimate the median (X_{med}) and the standard error of the median $s_{X_{med}}$ the procedure described below may be used.

The median (X_{med}) like the mean, can be estimated using either data which has been grouped into intervals (e.g., income intervals) or ungrouped data. If grouped data are used, the median (X_{med}) is estimated using either Formula 15 or 16 with $p = 0.5$. If ungrouped data are used, the data records are ordered based on the value of the item (e.g., income level), then the estimated median is the value of the item such that the weighted estimate of 50 percent of the sub-population falls at or below that value and 50 percent is at or above that value. The method of standard error computation presented here requires the use of grouped data, because it is deemed easier to compute the median by grouping the data and then using Formula 15 or 16.

An approximate method for measuring the reliability of an estimated median ($\hat{X}_{med}$) is to determine a confidence interval about it. (See the section on "Confidence Intervals.") The following procedure (four steps) may be used to estimate the 68-percent confidence limits (i.e., approximately ± one standard error from the median) and hence the standard error (of a median based on sample data.

Step 1 - Determine, using Formula 9, the standard error ($s_{x,p\,=\,50}$) of an estimate of 50 percent of the group (sub-population).

Step 2 - Subtract from and add to 50 percent the standard error determined in Step 1 to obtain the percentages associated with the lower and upper limits of the 68-percent confidence interval of the item. Namely, the smaller percentage is $50 - s_{x,p\,=\,50}$ percent, and the larger percentage is 50 + $s_{x,p\,=\,50}$ percent.

Step 3 - Using the distribution of the item within the group, calculate the quantity, X_{UCL} of the item such that the percent of the group owning more of the item is equal to the smaller percentage $(50 - s_{x,p\,=\,50})$ found in Step 2. This quantity (X_{UCL}) will be the upper limit for the 68-percent confidence interval (assuming that the interval with higher item value is ranked at lower percentile as illustrated in Table 2.) In a similar fashion, calculate the quantity, X_{LCL} of the item such that the percent of the group owning more of the item is equal to the larger percentage $(50 + s_{x,p\,=\,50})$ found in Step 2. This quantity (X_{LCL}) will be the lower limit for the 68-percent confidence interval. (Note that a median computed from ungrouped data may or may not fall in this confidence interval).

Step 4 - Divide the difference between the two quantities (X_{UCL} and X_{LCL}) determined in Step 3 by two to obtain the standard error estimate ($s_{\hat{X}_{med}}$) of the median estimate ($\hat{X}_{med}$). Namely,

$$\hat{s}_{\hat{X}_{med}} = \frac{X_{UCL} - X_{LCL}}{2} \tag{14}$$

To perform Step 3, it will be necessary to interpolate, which may be done using different methods. The most common is simple linear interpolation (Formula 15) and Pareto interpolation (Formula 16). 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, X_{pN} such that p percent own more of the item is

$$X_{pN} = A_1 \exp\left[\frac{\ln\left(\frac{pN}{N_1}\right)}{\ln\left(\frac{N_2}{N_1}\right)}\ln\left(\frac{A_2}{A_1}\right)\right] \tag{15}$$

if Pareto Interpolation is indicated and

$$X_{pN} = \left[\frac{pN - N_1}{N_2 - N_1}\left(A_2 - A_1\right) + A_1\right] \tag{16}$$

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 numbers 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. One should note that a mathematically equivalent result is obtained by using common logarithms (base 10) and antilogarithms.

An illustration would be in order to calculate the standard error of a median, we return to the first example used to illustrate the standard error of a mean. As indicated in Table 2, the size (*N*) of the group is 39,851,000 and the median annual income estimate ($\hat{X}_{med}$) for the group falls in between $17,500 and $19,999. With *p = 0.5*, A_1 *= $17,500*, A_2 *= $19,999*; N_1 *= 5,799,000 + 4,730,000 + ... +1,493,000 = 22,106,000,* and N_2 = *4,730,000 + 3,723,000 + ... + 1,493,000 = 16,307, 000*; the median annual income estimate, $\hat{X}_{med}$ for this group is computed using Formula 6.C-14 to be $18,317. The standard error estimate ($s_{\hat{X}_{med}}$) of the median annual income estimate is calculated using the above four step procedure as follows.

Step 1 - Using Formula 9 and the appropriate *b* parameter of 7,566, the standard error estimate of 50 percent on a base of 39,851,000 is about 0.7 percentage points, (i.e., $s_{x,p\,=\,50}$ = 0.7%).

Step 2 - Obtain the two percentages associated with the lower and upper limits of the 68-percent confidence: the smaller percentage = 50 - $s_{x,p\,=\,50}$ = 49.3 and the larger percentage = 50 + $s_{x,p\,=\,50}$ = 50.7.

Step 3 - By examining Table 2, we see that the percentage 49.3 falls in the income interval from $17,500 to $19,999. Thus as determined previously, A_1 *= $17,500*, A_2 *= $19,999*, N_1 *= 22,106,000*, N_2 *= 16,307,000*, and *N = 39,851,000* and *p = 49.3*. Based on Formula 15, the upper bound (X_{UCL}) of a 68-percent confidence interval for the median estimate ($\hat{X}_{med}$) is

$$X_{UCL} = 17{,}500 \exp\left[\frac{\ln\left(\dfrac{0.493\times 39{,}851{,}000}{22{,}106{,}000}\right)}{\ln\left(\dfrac{16{,}307{,}000}{22{,}106{,}000}\right)}\ln\left(\frac{19{,}999}{17{,}500}\right)\right] = \$18{,}429$$

Also by examining Table 2, the 50.7 percent fall in the same income interval. Thus, A_1, A_2, N_1, and N_2 are the same as above, but *p = 0.507*. The lower bound (X_{LCL}) of a 68-percent confidence interval for the median ($\hat{X}_{med}$) is

$$X_{LCL} = 17{,}500 \exp\left[\frac{\ln\left(\dfrac{0.507 \times 39{,}851{,}000}{22{,}106{,}000}\right)}{\ln\left(\dfrac{16{,}307{,}000}{22{,}106{,}000}\right)} \ln\left(\frac{19{,}999}{17{,}500}\right) \right] = \$18{,}204$$

Step 4 - Based on Formula 14, the standard error estimate ($s_{\hat{X}_{med}}$) of the median annual income estimate ($\hat{X}_{med}$) is

$$s_{\hat{X}_{med}} = \frac{\$18{,}429 - \$18{,}204}{2} = \$113$$

If the linear interpolation is used, the median is then estimated using Formula 16 to be $18,440 and the 68-percent confidence interval of the estimated median is from $18,319 to $18,560. The standard error estimate is $120.

Standard Error of Ratio of Means or Medians

The standard error for a ratio of means or medians is approximated by Formula 17 provided below.

$$s_{\frac{X}{Y}} = \sqrt{\left(\frac{X}{Y}\right)^2 \left[\left(\frac{s_X}{X}\right)^2 + \left(\frac{s_Y}{Y}\right)^2 \right]} \qquad (17)$$

where X and Y are the means or medians, and s_X and s_Y are their associated standard errors. Formula 17 assumes that the means or medians are not correlated. If the correlation between the population means or medians 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 or medians.

Table 1a - Reference months for each interview month of the SIPP 1992 Panel, SIPP 1993 Panel, SPD Bridge (1997), and SPD 1998 Surveys.

Survey	Months of Interview	Reference Months
SIPP Panel 1992	February 1992 - April 1995	October 1991 - March 1995
SIPP Panel 1993	February 1993 - January 1996	October 1992 - December 1995
SPD Bridge (1997)	April 1997 - June 1997	January 1996 - December 1996 (also January 1995 - December 1995 for SIPP Panel 1992 for only selected questions)
SPD 1998	May 1998 - July 1998	January 1997 - December 1997

Table 1b - Reference months for the SIPP Panel 1992, SIPP Panel 1993, SPD Bridge (1997), and SPD 1998 Surveys.

```
October                                                            March
1991                                                                1995
|<! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! !  SIPP Panel 1992 Survey  ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! >|


                       October                                                       December
                       1992                                                              1995
                       |<! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! SIPP Panel 1993 Survey ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! >|


                                                                                                 January     December
                                                                                                 1996             1996
                                                                                                 |<! ! SPD Bridge ! ! >|
                                                                                                       Survey


                                                                                                                     January     December
                                                                                                                     1997             1997
                                                                                                                     |<! !  SPD 1998 ! ! >|
                                                                                                                           Survey
```

Table 1c - Interview months for the SIPP Panel 1992, SIPP Panel 1993, SPD Bridge (1997), and SPD 1998 Surveys.

```
February                                                    April
1992                                                        1995
|<! ! ! ! ! ! ! ! ! ! ! ! ! SIPP Panel 1992 Survey  ! ! ! ! ! ! ! ! ! ! ! ! !>|

                    February                                            January
                    1993                                                  1996
                    |<! ! ! ! ! ! ! ! ! ! ! SIPP Panel 1993 Survey ! ! ! ! ! ! ! ! ! ! ! !>|

                                                                                  April              June
                                                                                  1997               1997
                                                                                  |<! ! SPD Bridge ! ! >|
                                                                                         Survey

                                                                                                          May              July
                                                                                                          1998             1998
                                                                                                          |<! !  SPD 1998 ! ! >|
                                                                                                                 Survey
```

Table 2 - Distribution of annual income among people 25 to 34 years old.

	Total Number of People	Number of People in Annual Income Interval												
		Under $5000	$5000 to $7499	7500 to $9999	$10000 to $12499	$12500 to $14999	$15000 to $17499	$17500 to $19999	$20000 to $29999	$30000 to $39999	$40000 to $49999	$50000 to $59999	$60000 to $69999	$70000 and Over
Number of People (in Thousands)	39851	1371	1651	2259	2734	3452	6278	5799	4730	3723	2591	2619	1223	1493
Percent with at Least as Much as Lower Bound of Interval	N/A	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

Note: This table contains a fictitious distribution of annual income and is used only to illustrate standard error calculation.

Table 3 - SPD Generalize variance parameters for estimates using the final longitudinal weights on the first SPD longitudinal file.

Characteristic	Parameters	
	a	b
TOTAL OR WHITE PEOPLE		
16+ Program Participation and Benefits, Poverty (3)*		
Both Sexes	-0.0000858	14,601
Male	-0.0001805	14,601
Female	-0.0001633	14,601
16+ Income and Labor Force (5)*		
Both Sexes	-0.0000443	7,566
Male	-0.0000936	7,566
Female	-0.0000845	7,566
16+ Pension Plan** (4)*		
Both Sexes	-0.0000812	13,858
Male	-0.0001714	13,858
Female	-0.0001549	13,858
All Others*** (6)*		
Children Aged Less Than 18		
Both Sexes	-0.0000798	18,398
Male	-0.0001649	18,398
Female	-0.0001546	18,398

Characteristic	Parameters	
	a	b
Adults Aged 18 and Over		
Both Sexes	-0.0001193	27,519
Male	-0.0002466	27,519
Female	-0.0002313	27,519
BLACK PEOPLE		
Poverty (1)*		
Both Sexes	-0.0004513	12,453
Male	-0.0009700	12,453
Female	-0.0008443	12,453
All Others*** (2)*		
Children Aged Less Than 18		
Both Sexes	-0.0002469	6,806
Male	-0.0005301	6,806
Female	-0.0004613	6,806
Adults Aged 18 and Over		
Both Sexes	-0.0003693	10,180
Male	-0.0007929	10,180
Female	-0.0006901	10,180
HOUSEHOLDS		
Total or Whites	-0.0001054	9,352

Characteristic	Parameters	
	a	b
Black	-0.0006441	6,461

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

** Use the "16+ Pension Plan" parameters for pension plan tabulations of people aged 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.

*** Use the "All Others" parameters for any type of tabulation not specifically covered by another characteristic in this table.