UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

February 6, 2018

MEMORANDUM FOR Jason Fields
Survey Director, Associate Director for Demographic Programs

From: Jennifer Guarino Tancreto
Acting Chief, Demographic Statistical Methods Division

Prepared by: Mahdi Sundukchi and Julia Yang
Survey of Income and Program Participation,
Demographic Statistical Methods Division

Subject: Social Security Administration (SSA) Supplement Source and Accuracy Statement for Public Use Files (S&A- 21)[1]

Attached is the Source and Accuracy Statement for 2014 Panel Public Use Files of the Social Security Administration (SSA) Supplement.

cc:

Matthew Marlay	(ADDP)	Nathaniel McKee	(DSD)
Daniel Doyle		Connie Bauer	
Holly Fee		James Farber	(DSMD)
Shelley Irving		Tracy Mattingly	
Cindy Easton		Mahdi Sundukchi	
David Waddington	(SEHSD)	Stephen Mack	
Denise Lewis		Ralph Culver III	
Trudi Renwick		Faith Nwaoha-Brown	
Alfred Gottschalck		Julia Yang	
Stephanie Galvin			
Edward Welniak			

[1]This source and accuracy statement can also be accessed through the U.S. Census Bureau website at *http://www.census.gov/programs-surveys/sipp/tech-documentation/source-accuracy-statements.html*

SOURCE AND ACCURACY STATEMENT FOR THE 2014 PANEL SOCIAL SECURITY ADMINISTRATION (SSA) SUPPLEMENT PUBLIC USE FILES[2]

DATA COLLECTION AND ESTIMATION

Source of Data

The data were collected in the 2014 Social Security Administration(SSA) Supplement. To reduce burden on the SIPP respondents, topical modules were removed from the 2014 SIPP. Instead, these questions were asked separately on the 2014 SSA Supplement from September 2014 through November 2014 via telephone. Although this is a separate survey, only the households that responded in the 2014 Survey of Income and Program Participation (SIPP) Wave 1 were eligible to be interviewed for the SSA Supplement. Therefore, the population represented in the 2014 SSA (the population universe) is the civilian noninstitutionalized population living in the United States, same as SIPP. The institutionalized population, which is excluded from the universe, is composed primarily of the population in correctional institutions and nursing homes (94 percent of the 4 million institutionalized people in Census 2010).

Since 2014 SSA sample is a subset of the SIPP 2014 Panel, they have the same survey design. Thus, the remaining section discusses the survey design of SIPP 2014.

The SIPP 2014 Panel sample is located in 820 Primary Sampling Units (PSUs), each consisting of a county or a group of contiguous counties. Of these 820 PSUs, 340 are self-representing (SR) and 480 are non-self-representing (NSR). SR PSUs have a probability of selection of one. NSR PSUs have a probability of selection less than one. Within PSUs, housing units (HUs) were systematically selected from the Master Address File (MAF), which is the Census Bureau's official inventory of known housing units. The frame was created using the decennial censuses, as well as the U.S. Postal Service's Delivery Sequence File (DSF). The Census Bureau continues to update the MAF using the DSF and various automated, clerical, and field operations.

Households were classified into two strata, such that one stratum had a higher concentration of low income households than the other. We oversampled the low income stratum by 24 percent to increase the accuracy of estimates for statistics of low income households and program participation. Analysts are strongly encouraged to use the SIPP weights when creating estimates since households are not selected with equal probability.

In Wave 1, the SIPP 2014 Panel began with a sample of about 53,070 HUs. About 10,722 of these HUs were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. Field Representatives (FRs) were able to obtain interviews for about 29,700 of the eligible HUs. FRs were unable to interview approximately 12,660 eligible HUs in

---

[2]For questions or further assistance with the information provided in this document contact: Tracy Mattingly of the Demographic Statistical Methods Division on 301-763-6445 or via email at Tracy.L.Mattingly@census.gov.

the panel because the occupants: (1) refused to be interviewed; (2) could not be found at home; (3) were temporarily absent; or (4) were otherwise unavailable. Thus, occupants of about 70 percent of all eligible HUs participated in the first interview of the panel.

Sample Loss

The weighted sample loss at SIPP 2014 Wave 1 and SSA were calculated and tabulated as shown in Table A below.

Table A. Sample Loss and Response Rate for SIPP 2014 Wave 1 and SSA

Wave	Eligible HUs	Interviewed HUs	Type As		Type Ds		Growth Factor	Weighted Sample Loss
			Total	Weighted Rate	Total	Weighted Rate		
1	42,348	29,685	12,663	31.2%				31.2%
SSA	29,502[3]	15,497	14,005	47.1%				47.1%

In addition, the percentage of each Type A outcome in the 2014 SSA are shown in Table B below.

Table B. Percent of Type As by Nonresponse Status for 2014 SSA

Wave	Household Refused	Other
SSA	41.71%	58.29%

Estimation

The SSA program produces person-level cross-sectional weights corresponding to September 2014. The reference period for all of the questions, except those pertaining to Personal Retirement Accounts, is as of the time of the interview.

The SSA estimation procedure involves several stages of weight adjustments to derive the person level weights. First, each person is given an initial weight (IW), which is the noninterview weight from SIPP 2014 Panel Wave 1. Next, IW of noninterviewed people are distributed to interviewed people through household noninterview adjustment factor (HHNIF). Finally, second stage adjustment is made to the interviewed people's new weight (IW*HHNIF) with second stage adjustment factor (SSF). This adjusts estimates to population controls from September 2014 and equalizes married spouses' weights.

The final weight is $FW_c = IW * HHNIF * SSF$.

3 Due to non-Type-A noninterviews, the number of eligible households (interviewed + Type A households) in SSA does not exactly match the number of interviewed households in SIPP 2014 Wave 1.

Population Controls
The 2014 SSA estimation procedure adjusts weighted sample results to agree with independently derived population estimates of the civilian noninstitutional population in September 2014. The national and state level population controls are obtained directly from the Population Division and are prepared each month to agree with the most current set of population estimates released by the U.S. Census Bureau's population estimates and projections program.

The national level controls are distributed by demographic characteristics as follows:

- Age, Sex, and Race (White Alone, Black Alone, and all other groups combined)
- Age, Sex, and Hispanic Origin

The state level controls are distributed by demographic characteristics as follows:

- State by Age and Sex
- State by Hispanic origin
- State by Race (Black Alone, all other groups combined)

The estimates begin with the latest decennial census as the base and incorporate the latest available information on births and deaths along with the latest estimates of net international migration.

The net international migration component in the population estimates includes a combination of:

- Legal migration to the U.S.,
- Emigration of foreign born and native people from the U.S.,
- Net movement between the U.S. and Puerto Rico,
- Estimates of temporary migration, and
- Estimates of net residual foreign-born population, which include unauthorized migration.

Because the latest available information on these components lags the survey date, to develop the estimate for the survey date, it is necessary to make short-term projections of these components.

Use of Weights
For the SSA 2014 Panel, weights will be produced at the person level. The person weight is for analyzing data at the person level. Each interviewed person on the SSA Supplement will be assigned a weight corresponding to the September 2014. This weight is called the SSA final weight. The benchmark population estimates for the second stage raking for the weight is for the corresponding month. Adults and children in the SSA Supplement universe who have completed self interviews or proxy interviews for the SSA Supplement and who were in scope for SIPP in December 2013 will receive a positive person weight.

For the SSA 2014 Panel, the household structure of an interviewed unit is only set for at the time of 2014 SIPP Wave 1 interview. Therefore, it is possible to tell which interviewed persons lived together and their relationships to each other during 2014 SIPP Wave 1, but the files do not

specify a household ID or reference person for September 2014. The same is true for families. If a data user would like to do analysis at the household or family level, the person weights can be used to specify a single household or family weight. To do this, one can take the average of the person month weights for all persons in the household or family.

ACCURACY OF ESTIMATES

SSA 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: sampling and nonsampling. For a given estimator, the difference between an estimate based on a sample and the estimate that would result if the sample were to include the entire population is known as sampling error. For a given estimator, the difference between the estimate that would result if the sample were to include the entire population and the true population value being estimated is known as nonsampling error. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error.

Nonsampling Error

Nonsampling errors can be attributed to many sources:

- 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
- 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, 1998 SIPP Working Paper Number 230*, issued May 1999.

Undercoverage in SIPP results from missed HUs and missed persons within sample HUs. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group.

A common measure of survey coverage is the coverage ratio, the estimated population before ratio adjustment divided by the independent population control. Table C below shows SSA coverage ratios for age-sex-race groups for September 2014.

Comparability with Other Estimates
Caution should be exercised when comparing this data with data from other SIPP products 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 discussions.

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

Standard Error Uses

Confidence Intervals
The sample estimate and its standard error enable one to construct a confidence interval. A confidence interval is a range about a given estimate that has a known probability of including the result of a complete enumeration. 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:

Table C. SSA Average Coverage Ratios for September 2014 for Age by Race and Sex						
Age	**White Only**		**Black Only**		**Residual**	
	Male	Female	Male	Female	Male	Female
<15	0.74	0.73	0.57	0.54	0.73	0.74
15	0.73	0.83	0.66	0.64	0.64	0.74
16-17	0.69	0.74	0.63	0.61	0.63	0.70
18-19	0.66	0.57	0.61	0.63	0.71	0.69
20-21	0.62	0.69	0.52	0.41	0.66	0.72
22-24	0.62	0.57	0.52	0.41	0.63	0.71
25-29	0.66	0.62	0.41	0.52	0.76	0.65
30-34	0.71	0.79	0.52	0.57	0.75	0.67
35-39	0.77	0.78	0.60	0.60	0.73	0.77
40-44	0.73	0.76	0.59	0.61	0.74	0.77
45-49	0.69	0.77	0.71	0.74	0.76	0.78
50-54	0.86	0.87	0.68	0.88	0.74	0.81
55-59	0.96	1.03	0.90	1.02	0.87	0.89
60-61	1.04	1.07	0.93	0.98	0.89	0.93
62-64	1.15	1.14	0.91	0.99	0.85	0.85
65-69	1.11	1.12	1.27	1.17	0.92	0.97
70-74	1.14	1.13	1.26	1.21	0.93	0.99
75-79	1.17	1.25	1.28	1.24	0.94	1.00
80-84	1.28	1.24	1.27	1.18	0.91	0.95
85+	1.06	0.96	1.36	1.21	0.97	0.98

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 1.645 standard errors below the estimate to 1.645 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 $(-1.645 \times S_{DIFF})$ 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 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. A Bonferroni correction can be done to account for this potential problem that consists of dividing your stated level of significance by the number of tests you are performing. This correction results in a conservative test of significance.

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 250,000. Also, nonsampling error in one or more of the small number of cases providing the estimation can cause large relative error in that particular estimate. 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 Computations

Calculating Standard Errors for SIPP Estimates
There are three main ways we calculate the Standard Errors (SEs) for SIPP Estimates. They are as follows:

- Direct estimates using replicate weight methods;
- Generalized variance function parameters (denoted as a and b); and
- Simplified tables of SEs based on the a and b parameters.

While the replicate weight methods provide the most accurate variance estimates, this approach requires more computing resources and more expertise on the part of the user. The Generalized Variance Function (GVF) parameters provide a method of balancing accuracy with resource

usage as well as smoothing effect on SE estimates across time. SIPP uses the Replicate Weighting Method to produce GVF parameters (see K. Wolter, *Introduction to Variance Estimation*, for more information). The GVF parameters are used to create the simplified tables of SEs.

Standard Error Parameters and Tables and Their Use

Most SIPP and SSA estimates have greater standard errors than those obtained through a simple random sample because of its two-stage cluster sample design. 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 as $\boldsymbol{a}$ and $\boldsymbol{b}$) 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 $\boldsymbol{a}$ and $\boldsymbol{b}$ parameters vary by characteristic and by demographic subgroup to which the estimate applies. Table 1 provides $\boldsymbol{a}$ and $\boldsymbol{b}$ parameters for the core domains to be used for the 2014 SSA estimates.

The $\boldsymbol{a}$ and $\boldsymbol{b}$ parameters given in Table 1 can be used directly for cross-sectional estimates.

For those users who wish further simplification, we have also provided base standard errors for estimates of totals and percentages in Tables 2 through 5. Note that these base standard errors must be adjusted by an f factor provided in Table 1. The standard errors resulting from this simplified approach are less accurate. Methods for using these parameters and tables for computation of standard errors for different estimates are given in the following sections. Later, we will describe how to use software packages to directly compute standard errors using replicate weights.

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. Note that neither method should be applied to dollar values.

The standard error may be obtained by the use of Formula (1):

$$s_x = f \times s, \tag{1}$$

where f is the appropriate f factor from Table 1, and s is the base standard error on the estimate obtained by interpolation from Tables 2 or 3.

Alternatively, s_x may be approximated by Formula (2):

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

This formula was used to calculate the base standard errors in Tables 2 and 3. Here x is the size of the estimate and a and b are the parameters from Table 1 which are associated with the

characteristic being estimated. Use of Formula (2) will generally provide more accurate results than the use of Formula (1).

Illustration 1.
Suppose 2014 SSA estimates show that there were 5,000,000 females aged 44 and up with monthly pension greater than $1,000. The appropriate parameters and factor from Table 1 and the appropriate general standard error from Table 3 are:

$$a = -0.00010160 \qquad b = 13{,}053 \qquad f = 1.027 \qquad s = 246{,}895$$

Using Formula (1), the approximate standard error is:

$$s_x = 1.027 \times 246{,}895 = 253{,}561$$

Using Formula (2), the approximate standard error is:

$$s_x = \sqrt{(-0.00010160 \times 5{,}000{,}000^2) + (13{,}053 * 5{,}000{,}000)} = 250{,}450\ females.$$

Using the standard error based on Formula (2), the approximate 90-percent confidence interval as shown by the data is from 4,588,010 to 5,411,990 females ($i.e., 5{,}000{,}000 \pm 1.645 \times 250{,}450$). Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90% of all samples.

Standard Error of a Mean
A mean is defined here to be the average quantity of some item (other than persons, families, or households) per person, family or household. For example, it could be the average monthly household income of females aged 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:

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

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 i^{th} unit. (A 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 lower and upper boundaries of interval j are Z_{j-1} and Z_j, respectively. Each unit, x_i, is placed into one of c intervals such that $Z_{j-1} < x_i \leq Z_j$. The estimated population mean, $\bar{x}$, and variance, s^2, are given by the formulas:

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

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

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

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

In the second method, the estimated population mean, $\bar{x}$, and variance, s^2 are given by:

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

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

where there are n units with the item of interest and w_i is the final weight for i^{th} unit. (Note that $\sum w_i = y$.)

<u>Illustration 2.</u>
<u>Method 1</u>

Suppose that based on the 2014 SSA data, the distribution of monthly pension income for persons aged 34 to 54 as of September 2014 is given in Table 6. Using these data, the mean monthly pension income for persons aged 34 to 54 is $1,940. Applying Formula (4), the approximate population variance, s^2, is:

$$s^2 = \left(\frac{137}{3{,}984}\right)(150)^2 + \left(\frac{165}{3{,}984}\right)(450)^2 + \cdots + \left(\frac{149}{3{,}984}\right)(3{,}800)^2 - (1{,}940)^2 = 772{,}612\,.$$

Using Formula (3) and a base b parameter of 13,053, the estimated standard error of a mean $\bar{x}$ is:

$$s_{\bar{x}} = \sqrt{\frac{13{,}053}{3{,}984{,}000} \times 772{,}612} = \$50.31$$

Thus, the approximate 90-percent confidence interval as shown by the data ranges from $1857.24 to $2022.76.

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 Formula (5) and b be the parameter associated with the particular type of item. The standard error of an aggregate is:

$$s_x = \sqrt{b \times y \times s^2}. \tag{6}$$

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 people sharing a particular characteristic such as the percent of people owning their own home. 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 percent of total wealth held by people with high income and the percent of total income received by people on welfare.

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

$$s_{(x,p)} = f \times s, \tag{7}$$

where f is the appropriate f factor from Table 1 and s is the base standard error of the estimate from Tables 4 or 5.

Alternatively, it may be approximated by the formula:

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

from which the standard errors in Tables 4 and 5 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 Formula (8)

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 3.
Suppose that in September 2014, 6.7 percent of the 16,812,000 persons in nonfarm households with monthly household pension income of $2,000 to $2,999 were black. Using Formula (8), with a b parameter of 12,992, the approximate standard error is:

$$s_{(x,p)} = \sqrt{\frac{12{,}992}{16{,}812{,}000} \times 6.7 \times (100 - 6.7)} = 0.70\%$$

Consequently, the 90 percent confidence interval as shown by these data is from 5.55% to 7.85%.

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:

$$p_I = 100\left(\frac{x_A}{x_N}\right),$$

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

$$p_I = 100\left(\hat{p}_A\left(\frac{\bar{x}_A}{\bar{x}_N}\right)\right),$$

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

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

where s_p is the standard error of $\hat{p}_A, s_A$ is the standard error of $\bar{x}_A$ and s_N 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.

Illustration 4.
Suppose that as of September 2014, 9.8% of the people invested their lump sum in real properties, the mean value of real property investment is $10,121, the mean value of total lump sum is $37,734, and the corresponding standard errors are 0.18%, $1,468, and $1,703, respectively. In total, there are 38,899,630 people who received lump sum. Then the percent of all investment in real properties among lump sum recipients is:

$$100\left(0.098 \times \frac{10{,}121}{37{,}734}\right) = 2.63\%$$

Using Formula (9), the appropriate standard error is:

$$s_I = \sqrt{(2.63)^2\left[\left(\frac{0.0018}{0.098}\right)^2 + \left(\frac{1{,}468}{10{,}121}\right)^2 + \left(\frac{1{,}703}{37{,}734}\right)^2\right]} = 0.40\%.$$

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

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

where s_x and s_y are the standard errors of the estimates *x* and *y*.

The estimates can be numbers, percent, ratios, etc. The correlation between x and y is represented by *r*. The above formula assumes that the correlation coefficient between the characteristics estimated by *x* and *y* is non-zero. If no correlations have been provided for a given set of x and y estimates, assume *r* = 0. However, if the correlation is really positive (negative), then this assumption will tend to cause overestimates (underestimates) of the true standard error.

Illustration 5.
Suppose that for September 2014, SSA estimates show the number of persons age 35-44 years with monthly pension income of $1,000 to $1,999 was 103,010 and the number of persons age 25-34 years with monthly pension income of $1,000 to $1,999 in the same time period was 64,003. Then, using the parameters $a = -0.00005255$ and $b = 13{,}053$ from Table 1 and Formula (2), the standard errors of these numbers are approximately 36,661 and 28,900 respectively. The difference in sample estimates is 39,007 and using Formula (10), the approximate standard error of the difference is:

$$\sqrt{36{,}661^{\,2} + 28{,}900^2} = \$46{,}682\,.$$

Suppose that it is desired to test at the 10 percent significance level whether the number of persons with monthly cash income of $1,000 to $1,999 was different for people age 35-44 years than for people age 25-34 years. To perform the test, compare the difference of 39,007 to the product $1.645 \times 46{,}682 = 76{,}792$. Since the difference is not greater than 1.645 times the

standard error of the difference, the data show that the two age groups are not significantly different at the 10 percent significance level.

Standard Error of a Median

The median quantity of some items such as income for a given group of people is the 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.

The median, like the mean, can be estimated using either data which have been grouped into intervals or ungrouped data. If grouped data are used, the median is estimated using Formulas (11) or (12) with $p = 0.5$. If ungrouped data are used, the data records are ordered based on the value of the characteristic, then the estimated median is the value of the characteristic such that the weighted estimate of 50 percent of the subpopulation falls at or below that value and 50 percent is at or above that value. Note that the method of standard error computation which is presented here requires the use of grouped data. Therefore, it should be easier to compute the median by grouping the data and using Formulas (11) or (12).

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:

$$X_{pN} = A_1 \times \exp\left[\left(\frac{\ln\left(\frac{pN}{N_1}\right)}{\ln\left(\frac{N_2}{N_1}\right)}\right)\ln\left(\frac{A_2}{A_1}\right)\right] \tag{11}$$

if Pareto Interpolation is indicated and:

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

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 6</u>.
To illustrate the calculations for the sampling error on a median, we return to Table 6. The median monthly income for this group is $1,912. The size of the group is 3,984,000.

1. Using Formula (8), the standard error of 50 percent on a base of 3,984,000 is about 2.9 percentage points.

2. Following step 2, the two percentages of interest are 47.1 and 52.9.

3. By examining Table 6, we see that the percentage 47.1 falls in the income interval from $1,800to $2,099. (Since 55.5% receive more than $1,800 per month, the dollar value corresponding to 47.1 must be between $1,800 and $2,100.) Thus, $A_1 = \$1{,}800, A_2 = \$2{,}100, N_1 = 2{,}210{,}000$ *and* $N_2 = 1{,}630{,}000$.

In this case, we decided to use Pareto interpolation. Therefore, using Formula (11), the upper bound of a 68% confidence interval for the median is

$$\$1{,}800 \times \exp\left[\left(\frac{\ln\left(\frac{0.471 \times 3{,}984{,}000}{2{,}210{,}000}\right)}{\ln\left(\frac{1{,}630{,}000}{2{,}210{,}000}\right)}\right) \times \ln\left(\frac{2{,}100}{1{,}800}\right)\right] = \$1{,}955.$$

Also by examining Table 6, we see that the percentage 52.9 falls in the interval \$1,800 to \$2,099, same as before. 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

$$\$1{,}800 \times \exp\left[\left(\frac{\ln\left(\frac{0.529 \times 3{,}984{,}000}{2{,}210{,}000}\right)}{\ln\left(\frac{1{,}630{,}000}{2{,}210{,}000}\right)}\right) \times \ln\left(\frac{2{,}100}{1{,}800}\right)\right] = \$1{,}844$$

Thus, the 68-percent confidence interval on the estimated median is from \$1,844 to \$1,955.

4. Then the approximate standard error of the median is

$$\frac{\$1{,}955 - \$1{,}844}{2} = \$56$$

Standard Errors of Ratios of Means and Medians

The standard error for a ratio of means or medians is approximated by:

$$s_{\frac{x}{y}} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2\right]}, \tag{13}$$

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.

Standard Errors Using Software Packages

Standard errors and their associated variance, calculated by statistical software packages such as SAS or Stata, do not accurately reflect the SIPP's complex sample design. Erroneous conclusions will result if these standard errors are used directly. We provide adjustment factors by characteristics that should be used to correctly compensate for likely under-estimates. The factors called design effects (DEFF), available in Table 1, must be applied to SAS or Stata generated variances. The square root of DEFF can be directly applied to similarly generated standard errors. These factors approximate design effects which adjust statistical measures for sample designs more complex than simple random sample.

Replicate weights for SIPP are also provided and can be used to estimate more accurate standard errors and variances. While replicate weighting methods require more computing resources, many statistical software packages, including SAS, have procedures that simplify the use of replicate weights for users. To calculate variances using replicate weights use the formula:

$$Var(\theta_0) = \frac{1}{G(0.5)^2} \times \sum_{i=1}^{G} (\theta_i - \theta_0)^2 \tag{14}$$

where G is the number of replicates, θ_0 is the estimate using full sample weights, and θ_i is the estimate using the replicate weights. For the 2014 panel, G=240 for the number of replicate weights provided in the public use files. Replicate weights are created using Fay's method, with a Fay coefficient of 0.5.

Instead of direct computation, various SAS procedures include options to use replicate weights when estimating standard errors or variances. To use replicate weights in SAS include the VARMETHOD=BRR(FAY=0.5) option in the PROC statement and specify the replicate weights with a REPWEIGHTS. Other computer packages have similar methods.

Formula (14) produces variance estimates close to zero for the median when multiple obervations have value equal to the median. In this case, two methods can be used to estimate the variance of the median. The first technique incorporates replicate weights in Woodruff's method for estimating variability (Woodruff, 1952). Gossett et al (2002) documents the procedure for combining Woodruff's method with Jackknife replication and provides sample codes adapted by Mack and Tekansik (2011) for Fay's BRR. The second method uses VARMETHOD=TAYLOR option, a direct application of Woodruff's method, along with the cluster and strata statements instead of replicate weights to account for SIPP's complex design.

<u>Illustration 7</u>
In SAS, the SURVEYMEANS procedure is used to estimate statistics such as means, totals, proportions, quantiles, and ratios for a survey sample. An example syntax for estimating the mean of monthly pension income (TPENSAMT using SSA replicate weights is:

```
proc surveymeans data=Pu2014ssa varmethod=brr(Fay=0.5) mean;
   var TPENSAMT;
   weight SSA_PFINWGT;
   repweights REPWGT1-REPWGT240;
run;
```

Similarly, replicate weights can be used to estimate standard errors in the SURVEYFREQ (for frequency tables and cross-tabulations), SURVEYREG (for regression analysis), SURVEYLOGISTIC (for logistic regression analysis), and SURVEYPHREG (for proportional hazards regression analysis) SAS procedures by using the same VARMETHOD = BRR(FAY=0.5) option and REPWEIGHTS statement.

In Stata, the SVY command is used to fit a statistical model to a complex survey dataset. SVYSET is used to determine the survey design and provide information about the variance estimation. The following Stata syntax is equivalent to using SURVEYMEANS by SAS:

```
use pu2014ssa.dta
svyset [pweight=ssa_pfinwgt], brrweight(repwgt1-repwgt240) fay(.5) vce(brr)
mse
svy: mean tpensamt
```

REFERENCES

Girard, Claude (2005). *An Extension of Woodruff's Method for Calculating Sampling Variability for the Median*. ASA Section on Nonparametric Statistics, pp1644-1648.

Gossett, J.M. et al (2002). "How Complex Can Complex Survey Analysis Be with SAS?," Statistics and Data Analysis Conference.

Mack, Stephen and Tekansik, Sarah. "*Calculating Standard Errors and Confidence Intervals of Medians*. Internal Memorandum for Documentation. U.S. Census Bureau, May 6, 2011.

U.S. Census Bureau (1999). *SIPP Quality Profile, 1998,* SIPP Working Paper No. 230. Washington, DC: U.S. Census Bureau, May 1999.

U.S. Census Bureau (2016). "Chapter 7: Nonsampling Error, Sampling Error and Weighting," *Survey of Income and Program Participation: 2014 Panel Users' Guide,* 1st Ed. Washington, DC: U.S. Census Bureau.

Wolter, Kirk M. (2007). "Chapter 7: Generalized Variance Functions," *Introduction to Variance Estimation,* 2nd Ed. New York: Springer, pp. 272-297.

Woodruff, Ralph S. "*Confidence Intervals for Medians and Other Position Measures,*" Journal of the American Statistical Association, Vol. 47, No. 260 (Dec. 1952), pp 635-646.

Table 1. SSA Generalized Variance Parameters

Domain	Parameters		DEFF[4]	*f*
	a	***b***		
Poverty and Program Participation, Persons 15+				
Total	-0.00005125	12,729	4.93	1.013
Male	-0.00010616	12,729		
Female	-0.00009908	12,729		
Income and Labor Force Participation, Persons 15+				
Total	-0.00005255	13,053	5.06	1.027
Male	-0.00010886	13,053		
Female	-0.00010160	13,053		
Other, Persons 0+				
Total (or White)	-0.00003951	12,389	4.80	1.000
Male	-0.00008121	12,389		
Female	-0.00007768	12,389		
Black, Persons 0+	-0.00032020	12,992	5.04	1.025
Male	-0.00068486	12,992		
Female	-0.00060137	12,992		
Hispanic, Persons 0+	-0.00026602	14,452	5.60	1.080
Male	-0.00052784	14,452		
Female	-0.00053632	14,452		
Households				
Total (or White)	-0.00007972	9,909	3.84	0.894
Black	-0.00060515	9,909		
Hispanic	-0.00062151	9,909		

Notes on Domain Usage for Table 1:

Poverty and Program Participation	Use these parameters for estimates concerning poverty rates, welfare program participation (e.g., SNAP, SSI, TANF), and other programs for adults with low incomes.
Income and Labor Force	These parameters are for estimates concerning income, sources of income, labor force participation, economic well being other than poverty, employment related estimates (e.g., occupation, hours worked a week), and other income, job, or employment related estimates.
Other Persons	Use the "Other Persons" parameters for estimates of total (or white) persons aged 0+ in the labor force, and all other characteristics not specified in this table, for the total or white population.
Black/Hispanic Persons	Use these parameters for estimates of Black and Hispanic persons 0+.
Households	Use these parameters for all household level estimates.

[4] DEFF=b/sample interval, where sample interval=2,580

Table 2. Base Standard Errors of Estimated Numbers of Households or Families			
Size of Estimate	**Standard Error**	**Size of Estimate**	**Standard Error**
200,000	44,482	30,000,000	474,892
300,000	54,457	40,000,000	518,467
500,000	70,246	50,000,000	544,197
750,000	85,947	60,000,000	554,570
1,000,000	99,143	70,000,000	550,456
2,000,000	139,639	80,000,000	531,519
3,000,000	170,322	90,000,000	496,062
5,000,000	218,064	95,000,000	471,044
7,500,000	264,260	99,500,000	443,506
10,000,000	301,858	105,000,000	401,910
15,000,000	361,522	110,000,000	354,088
25,000,000	444,860	117,610,000	250,403

Note: These estimates are calculations using the Household Total (or White) a and b parameters from Table 1 and Formula (2).

Table 3. Base Standard Errors of Estimated Numbers of Persons			
Size of Estimate	**Standard Error**	**Size of Estimate**	**Standard Error**
200,000	49,762	110,000,000	940,595
300,000	60,936	120,000,000	957,985
500,000	78,642	130,000,000	971,005
750,000	96,278	140,000,000	979,829
1,000,000	111,128	150,000,000	984,568
2,000,000	156,907	160,000,000	985,284
3,000,000	191,863	170,000,000	981,983
5,000,000	246,895	180,000,000	974,626
7,500,000	301,156	190,000,000	963,119
10,000,000	346,322	200,000,000	947,312
15,000,000	420,649	210,000,000	926,984
25,000,000	533,883	220,000,000	901,829
30,000,000	579,751	230,000,000	871,430
40,000,000	657,529	240,000,000	835,215
50,000,000	721,578	250,000,000	792,386
60,000,000	775,309	260,000,000	741,798
70,000,000	820,750	270,000,000	681,726
80,000,000	859,218	275,000,000	647,326
90,000,000	891,616	280,000,000	609,373
100,000,000	918,586	299,340,000	410,185

Notes: (1) These estimates are calculations using the Other Persons 0+ a and b parameters from Table 1 and Formula (2).

(2) To calculate the standard for another domain multiply the standard error from this table by the appropriate f factor from Table 1.

Table 4. Base Standard Errors for Percentages of Households or Families						
Base of Estimated Percentages	**Estimated Percentages**					
	≤ 1 or ≥ 99	**2 or 98**	**5 or 95**	**10 or 90**	**25 or 75**	**50**
200,000	2.21%	3.12%	4.85%	6.68%	9.64%	11.13%
300,000	1.81%	2.54%	3.96%	5.45%	7.87%	9.09%
500,000	1.40%	1.97%	3.07%	4.22%	6.10%	7.04%
750,000	1.14%	1.61%	2.51%	3.45%	4.98%	5.75%
1,000,000	0.99%	1.39%	2.17%	2.99%	4.31%	4.98%
2,000,000	0.70%	0.99%	1.53%	2.11%	3.05%	3.52%
3,000,000	0.57%	0.80%	1.25%	1.72%	2.49%	2.87%
5,000,000	0.44%	0.62%	0.97%	1.34%	1.93%	2.23%
7,500,000	0.36%	0.51%	0.79%	1.09%	1.57%	1.82%
10,000,000	0.31%	0.44%	0.69%	0.94%	1.36%	1.57%
15,000,000	0.26%	0.36%	0.56%	0.77%	1.11%	1.29%
25,000,000	0.20%	0.28%	0.43%	0.60%	0.86%	1.00%
30,000,000	0.18%	0.25%	0.40%	0.55%	0.79%	0.91%
40,000,000	0.16%	0.22%	0.34%	0.47%	0.68%	0.79%
50,000,000	0.14%	0.20%	0.31%	0.42%	0.61%	0.70%
60,000,000	0.13%	0.18%	0.28%	0.39%	0.56%	0.64%
70,000,000	0.12%	0.17%	0.26%	0.36%	0.52%	0.59%
80,000,000	0.11%	0.16%	0.24%	0.33%	0.48%	0.56%
90,000,000	0.10%	0.15%	0.23%	0.31%	0.45%	0.52%
105,000,000	0.10%	0.14%	0.21%	0.29%	0.42%	0.49%
110,000,000	0.09%	0.13%	0.21%	0.28%	0.41%	0.47%
117,610,000	0.09%	0.13%	0.20%	0.28%	0.40%	0.46%

Note: These estimates are calculations using the Households Total (or White) *b* parameter from Table 1 and Formula (8).

Table 5. Base Standard Errors for Percentages of Persons

Base of Estimated Percentages	Estimated Percentages					
	≤ 1 or ≥ 99	2 or 98	5 or 95	10 or 90	25 or 75	50
200,000	2.48%	3.48%	5.42%	7.47%	10.78%	12.44%
300,000	2.02%	2.85%	4.43%	6.10%	8.80%	10.16%
500,000	1.57%	2.20%	3.43%	4.72%	6.82%	7.87%
750,000	1.28%	1.80%	2.80%	3.86%	5.57%	6.43%
1,000,000	1.11%	1.56%	2.43%	3.34%	4.82%	5.57%
2,000,000	0.78%	1.10%	1.72%	2.36%	3.41%	3.94%
3,000,000	0.64%	0.90%	1.40%	1.93%	2.78%	3.21%
5,000,000	0.50%	0.70%	1.08%	1.49%	2.16%	2.49%
7,500,000	0.40%	0.57%	0.89%	1.22%	1.76%	2.03%
10,000,000	0.35%	0.49%	0.77%	1.06%	1.52%	1.76%
15,000,000	0.29%	0.40%	0.63%	0.86%	1.24%	1.44%
25,000,000	0.22%	0.31%	0.49%	0.67%	0.96%	1.11%
30,000,000	0.20%	0.28%	0.44%	0.61%	0.88%	1.02%
40,000,000	0.18%	0.25%	0.38%	0.53%	0.76%	0.88%
50,000,000	0.16%	0.22%	0.34%	0.47%	0.68%	0.79%
60,000,000	0.14%	0.20%	0.31%	0.43%	0.62%	0.72%
70,000,000	0.13%	0.19%	0.29%	0.40%	0.58%	0.67%
100,000,000	0.11%	0.16%	0.24%	0.33%	0.48%	0.56%
110,000,000	0.11%	0.15%	0.23%	0.32%	0.46%	0.53%
120,000,000	0.10%	0.14%	0.22%	0.30%	0.44%	0.51%
130,000,000	0.10%	0.14%	0.21%	0.29%	0.42%	0.49%
140,000,000	0.09%	0.13%	0.21%	0.28%	0.41%	0.47%
150,000,000	0.09%	0.13%	0.20%	0.27%	0.39%	0.45%
160,000,000	0.09%	0.12%	0.19%	0.26%	0.38%	0.44%
170,000,000	0.08%	0.12%	0.19%	0.26%	0.37%	0.43%
180,000,000	0.08%	0.12%	0.18%	0.25%	0.36%	0.41%
190,000,000	0.08%	0.11%	0.18%	0.24%	0.35%	0.40%
200,000,000	0.08%	0.11%	0.17%	0.24%	0.34%	0.39%
210,000,000	0.08%	0.11%	0.17%	0.23%	0.33%	0.38%
220,000,000	0.07%	0.11%	0.16%	0.23%	0.32%	0.38%
230,000,000	0.07%	0.10%	0.16%	0.22%	0.32%	0.37%
240,000,000	0.07%	0.10%	0.16%	0.22%	0.31%	0.36%
250,000,000	0.07%	0.10%	0.15%	0.21%	0.30%	0.35%
280,000,000	0.07%	0.09%	0.14%	0.20%	0.29%	0.33%
299,340,000	0.06%	0.09%	0.14%	0.19%	0.28%	0.32%

Notes: (1) These estimates are calculations using the Other Persons 0+ a and b parameter from Table 1 and Formula (8).

(2) To calculate the standard for another domain multiply the standard error from this table by the appropriate f factor from Table 1.

Table 6. Distribution of Monthly Pension Income Among People 34 to 54 Years Old
(Not Actual Data, Only Use for Calculation Illustrations)

	Interval of Monthly Pension Income												
	Under $300	$300 to $599	$600 to $899	$900 to $1,199	$1,200 to $1,499	$1,500 to $1,799	$1,800 to $2,099	$2,100 to $2,399	$2,400 to $2,699	$2,700 to $2,999	$3,000 to $3,299	$3,300 to $3,599	$3,600 -$4,000
Number of People in Each Interval (in thousands)	137	165	226	273	345	628	580	473	372	252	262	122	149
Cumulative Number of People with at Least as Much as Lower Bound of Each Interval (in thousands)	3,984 (Total People)	3,847	3,682	3,456	3,183	2,838	2,210	1,630	1,157	785	533	271	149
Percent of People with at Least as Much as Lower Bound of Each Interval	100	96.6	92.4	86.7	79.9	71.2	55.5	40.9	29.0	19.7	13.4	6.8	3.7