CHAPTER 5

ACCURACY OF THE MICRODATA SAMPLE ESTIMATES

INTRODUCTION

The tabulations prepared from the Public Use Microdata Sample (PUMS) files are based on a 10-percent sample of the 2010 Census. The data summarized from these files are estimates of the actual figures that were obtained from the census tabulation and are subject to sampling error. Sampling error in data arises from the selection of people and housing units to be included in the sample. Because the PUMS is a sample of the census records, other types of errors that occurred during the data collection and data processing phases of the census, nonsampling errors, are inherent in the PUMS data. This chapter provides information about both sampling and nonsampling error and a description of how to estimate the sampling error for PUMS estimates.

MEASURING SAMPLING ERROR

Since the estimates derived from the PUMS files are based on a sample, they will differ somewhat from counts obtained from the census. The sample estimate also may differ from other samples of housing units, people within those housing units, and people living in group quarters.

The *standard error* of a sample estimate is a measure of the variation among the estimates from all possible samples. Thus, it measures the precision with which an estimate from a particular sample approximates the average result of all possible samples, or the census value in this case. Sampling error and some types of nonsampling error, such as item nonresponse, are estimated, in part, by the standard error.

Estimating the Sampling Error. There is more than one way to estimate the sampling error. In the following sections, we present two methods for estimating the standard error of PUMS estimates: (1) a generalized variance method and (2) the delete-a-group jack-knife variance method. The generalized approach uses design factors to adjust a standard error calculated assuming simple random sampling. The delete-a-group jack-knife technique directly estimates the standard error from the PUMS data, requiring additional data processing.

The generalized standard error approach produces an acceptable measure of reliability, particularly for estimates of totals and percentages. The delete-a-group jack-knife method will generally produce a more accurate estimate of the standard error and is more appropriate for a wider variety of statistics, such as means and ratios, and for detailed data tabulated over more than one characteristic. The trade-off is an increase in precision for more data processing. It is important to keep in mind that there will be differences between the standard error estimates computed by the two methods.

Calculating the Confidence Interval from the Standard Error. A confidence interval is a range of values that describes the uncertainty surrounding an estimate. A confidence interval is

indicated by its endpoints, (*Lower bound*, *Upper bound*). A confidence interval is also itself an estimate, a function of the sample estimate and its estimated standard error.

$$Lower\ bound\ =\ Estimate\ -(z_{\alpha/2}\ \times\ Standard\ Error) \quad (1)$$

$$Upper\ bound\ =\ Estimate\ +\ (z_{\alpha/2}\ \times\ Standard\ Error) \quad (2)$$

*w*here
α = level of significance (the complement of the confidence level), and
$z_{\alpha/2}$ = the value from the standard normal distribution for level of significance, α.

The selected confidence level represents a level of certainty about our estimate, for example, a 90-percent confidence level. This means that if we were to repeatedly create new estimates using the same procedure (by drawing a new sample, and calculating new estimates and confidence intervals), the confidence intervals would contain the census value 90 percent of the time. The Census Bureau routinely uses 90-percent confidence levels for which $z_{\alpha/2} = 1.645$.

When constructing confidence intervals, be aware of any "natural" limits on the bounds. For example, if a characteristic estimate for the population is near 0, the calculated value of the lower confidence bound may be negative. However, a negative number of people does not make sense, so the lower confidence bound should be reported as 0 instead. For other estimates such as income, negative values do make sense. Take into consideration the context and meaning of the estimate when creating these bounds. Another of these natural limits would be 100 percent for the upper bound of a percent estimate.

Limitations. Users should keep in mind a couple of points when computing and interpreting standard errors and confidence intervals for the PUMS data.

- The estimated standard errors do not include all portions of the variability due to nonsampling error that may be present in the data. For example, the standard errors do not reflect the effect of systematic errors introduced by interviewers or data processing. Consider the standard errors to be a lower bound of the total error (sampling error plus nonsampling error) and be conservative when making inferences. This caution is particularly relevant for small estimates close to 0 and very large estimates close to the total population for which sampling error may be a relatively smaller proportion of total error.
- Percentage estimates of 0 and estimated totals of 0 are subject to both sampling and nonsampling error even though the methods presented here will yield standard error estimates of 0.

ESTIMATING A STANDARD ERROR BY THE GENERALIZED VARIANCE METHOD WITH DESIGN FACTORS

To produce generalized standard error estimates, one obtains (1) the standard error for the characteristic that would result from a simple random sample (SRS) design (of people, families, or housing units) and estimation methodology; and (2) a design factor for the geography and the particular characteristic estimated. In general, this method provides conservative estimates of the standard error.

The design factors provided in Tables A.1 through A.53 can be used to approximate the standard errors of most sample estimates of totals and percentages. Design factors are given by characteristic for the United States, each of the 50 states, the District of Columbia, and Puerto Rico. The design factors reflect the effects of the sample design and estimation procedure used for the 2010 Census PUMS.

Totals and Percentages. To approximate the standard error of an estimated total or percentage, follow the steps below.

Step 1. Compute the SRS standard error for estimated totals or percentages.

For estimated totals, the general formula for the SRS standard error is

$$SE(Y) = \sqrt{N^2(1-f)\frac{\frac{Y}{N}\left(1-\frac{Y}{N}\right)}{n}} \qquad (3)$$

where
Y = estimate (weighted) of the characteristic,
N = population size of the geography of interest,
f = sampling rate (or probability of selection), and
n = size of the sample.

The population size, N, is the estimated total number of people, housing units, households, or families in the geography for which the user is interested.

For an estimated percentage, the general formula for the estimated standard error assuming SRS is

$$SE(p_d) = \sqrt{(1-f)\frac{p_d(100-p_d)}{n_d}} \qquad (4)$$

where
p_d = estimated percentage,

f is defined above, and
n_d = size of the subpopulation of interest in the sample.

A percentage is defined here as the ratio of a numerator to a denominator multiplied by 100, where the numerator is a subset of the denominator, $p_d = \frac{Y}{N_d} \times 100$, and N_d is the estimated number of people, housing units, households, or families in the subpopulation for which the user is interested. If the base of the percentage is the estimated total number of people, housing units, households, or families in the geography of interest, then $N_d = N$ and $n_d = n$.

Step 2. Use the appropriate table to obtain the appropriate design factor, based on the geography and the characteristic. Use the table for the United States (Table A.1) when estimating characteristics for the United States or geographic areas that cover more than one state. Use the table for the specific state, the District of Columbia, or Puerto Rico (Tables A.2 through A.53) when estimating characteristics for that state or geographic areas that are contained entirely within that state. If the estimate is a cross-tabulation of more than one characteristic, use the largest design factor.

Step 3. Multiply the SRS standard error from Step 1 by the design factor found in Step 2.

For estimated percentages that are less than 2 or greater than 98, use p_d equal to 2 or 98 percent in formula (4) to protect against severely understating the error present in very small or very large estimates.

Sums and Differences. To estimate the standard error of a sum or difference of two sample estimates, we use an approximation that assumes the estimates are uncorrelated:

$$SE(X + Y) = SE(X - Y) = \sqrt{[SE(X)]^2 + [SE(Y)]^2} \qquad (5)$$

However, it is likely that the two estimates of interest are correlated. If the two quantities X and Y are positively correlated, this method underestimates the standard error of the sum of X and Y and overestimates the standard error of the difference between the two estimates. If the two estimates are negatively correlated, this method overestimates the standard error of the sum and underestimates the standard error of the difference.

Ratios. Frequently, the statistic of interest is the ratio of two variables, where the numerator is not a subset of the denominator. An example is the ratio of males to females. (Note that this method cannot be used to compute a standard error for a sample mean.) The standard error of the ratio between two sample estimates is approximated by using the formula,

$$SE\left(\frac{X}{Y}\right) = \left(\frac{X}{Y}\right)\sqrt{\frac{[SE(X)]^2}{X^2} + \frac{[SE(Y)]^2}{Y^2}} \qquad (6)$$

Similar to sums and differences above, the estimates of X and Y are assumed to be uncorrelated. For reasonably large samples, ratio estimates are approximately normally distributed, particularly for the census population. Therefore, if we can calculate the standard error of a ratio estimate, then we can form a confidence interval around the ratio.

Means. A mean is defined here as the average quantity of some characteristic (other than the number of people, housing units, households, or families) per person, housing unit, household, or family. For example, a mean could be the average age of females living in an urban residence. The standard error of a mean can be approximated by the formula below. Because of the approximation used in developing this formula, the estimated standard error will generally underestimate the true standard error.

$$SE(\bar{x}) = \sqrt{(1-f) \times \frac{s^2}{n_d}} \times Design\ Factor \qquad (7)$$

where
$\bar{x}$ = estimated sample mean,
s^2 = estimated population variance of the characteristic, and
n_d = size of the subpopulation of interest in the sample.

1. When the characteristic of interest is available as a continuous or quantitative variable, the estimated population variance, s^2, can be computed as follows:

$$s^2 = \frac{1}{n_d} \times \sum_{i=1}^{n_d} (x_i - \bar{x})^2 \qquad (8)$$

   where
   n_d is defined above,
   x_i = value of the characteristic for the i^{th} sample record, and
   $\bar{x}$ is the estimated mean.

   Because all persons, families, and housing units in the PUMS have a weight of 10, the mean can be calculated as

$$\bar{x} = \frac{1}{n_d} \times \sum_{i=1}^{n_d} x_i \qquad (9)$$

2. When continuous or quantitative values for a characteristic are categorized into ranges, the population variance, s^2, can be estimated from the grouped data. Suppose there are *c* intervals where the lower and upper boundaries of interval *j* are L_j and U_j, respectively. Each person is placed into one of the *c* intervals such that the value of the characteristic is between L_j and U_j. The estimated population variance, s^2, is then given by

$$s^2 = \sum_{j=1}^{c} p_j m_j{}^2 - (\overline{x})^2 \quad (10)$$

where
p_j = weighted proportion of people, housing units, households or families in interval *j* and
m_j = midpoint of the j^{th} interval, calculated as

$$m_j = \frac{L_j + U_j}{2} \quad (11)$$

If the c^{th} interval is open-ended (i.e., no upper interval boundary exists), then approximate m_c by

$$m_c = \left(\frac{3}{2}\right) L_c \quad (12)$$

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

$$\overline{x} = \sum_{j=1}^{c} p_j m_j \quad (13)$$

EXAMPLES OF GENERALIZED STANDARD ERROR CALCULATIONS AND CONFIDENCE INTERVALS

Note: The following examples do not contain actual estimates or standard errors derived from this data product. The numbers are used for illustration purposes only.

For each of the following examples, the sampling rate, *f*, is 0.1.

Example 1: Computing the Standard Error and Confidence Interval for a Total. Suppose the estimate for the total number of persons who are age 16 years and over and live in urban residences in county A in state B is 59,950, denoted by *Y* in formula (3). From the 10-percent PUMS for state B, suppose that for county A, the number of persons in sample is 15,432, denoted by *n* in formula (3), and the sum of the PUMS weights for all persons is 154,320, denoted by N.

Using formula (3), the estimated standard error under SRS is

$$SE(59{,}950) = \sqrt{154{,}320^2(1-0.1)\frac{\frac{59{,}950}{154{,}320}\left(1-\frac{59{,}950}{154{,}320}\right)}{15{,}432}}$$

$\approx$ 574 people.

For state B, suppose that the design factor for "Type of residence (urban/rural)" is 1.20 and is larger than that for "Age". Therefore, the appropriate design factor is that for "Type of residence (urban/rural)" for State B. Then, the estimated standard error is

$$SE(59{,}950) = 574 \times 1.20 = 689 \text{ people.}$$

We can obtain a 90-percent confidence interval for the total number of persons who are age 16 years and over and live in urban residences in county A in state B by using formulas (1) and (2). Thus, a 90-percent confidence interval for this estimated total is

[59,950 – (1.645 × 689)] to [59,950 + (1.645 × 689)]
or
(58,817, 61,083).

Example 2: Computing the Standard Error and Confidence Interval for a Percentage. Suppose the estimate for the percentage of persons who are age 16 years and over who live in urban residences in county B in state A, p_d, is 62.6. From the 10-percent PUMS for state A, suppose that for county B, there are 9,576 persons age 16 years and over in sample, denoted by n_d in formula (4). Therefore, using formula (4), the estimated standard error under SRS is found to be approximately 0.47 percent.

$$SE(62.6) = \sqrt{(1 - 0.1)\frac{62.6(100 - 62.6)}{9{,}576}} \approx 0.47$$

For state A, suppose that the design factor for "Type of residence (urban/rural)" is 1.25 and is larger than that for "Age". Therefore, the appropriate design factor is that for "Type of residence (urban/rural)" for State A. The estimated standard error for the estimated 62.6 percent of persons 16 years and over who live in urban residences is 0.47 × 1.25 = 0.59 percent.

The 90-percent confidence interval for this estimated percentage is

[62.6 – (1.645 × 0.59)] to [62.6 + (1.645 × 0.59)]
or
(61.6, 63.6).

Example 3: Computing the Standard Error and Confidence Interval for a Difference. Suppose the estimate for the percentage of males in county A in state C age 16 years and over who live in urban residences is 76.1, and the sample size of males 16 years and over is 4,627. Using formula (4), the estimated SRS standard error is approximately 0.59 percent. Assume the design factor to be 1.20 for "Type of residence (urban/rural)" for state C. Thus, the approximate standard error of the percentage (76.1 percent) is 0.59 × 1.20 = 0.71 percent.

Suppose the estimated percentage of females 16 years and over who live in urban residences is 48.2 percent with an approximate standard error of 0.82.
The difference in the two estimates is

$$76.1 - 48.2 = 27.9 \text{ percent.}$$

Using formula (5), the estimated standard error of the difference is

$$SE(27.9) = \sqrt{[SE(76.1)]^2 + [SE(48.2)]^2} = \sqrt{[0.71]^2 + [0.82]^2}$$

$$= 1.08 \text{ percent.}$$

The 90-percent confidence interval for the difference is

$$[27.9 - (1.645 \times 1.08)] \text{ to } [27.9 + (1.645 \times 1.08)]$$
$$\text{or}$$
$$(26.1, 29.7).$$

When, as in this example, the interval does not include 0, one can conclude, again with 90-percent confidence, that the difference observed between the two sexes for this characteristic is greater than can be attributed to sampling error.

Example 4: Computing the Standard Error and Confidence Interval for a Ratio. Suppose that one wished to obtain the standard error of the estimated ratio of males to females who were 16 years and over and who lived in urban residences. If the estimates for males and females are 35,200 and 23,850, respectively, then the ratio of the two estimates of interest is

$$\frac{35{,}200}{23{,}850} = 1.48.$$

After having applied the appropriate design factors to each SRS standard errors, suppose the estimated standard errors are 579 and 504, respectively. Using formula (6), the estimated standard error of the ratio is

$$SE(1.48) = \left(\frac{35{,}200}{23{,}850}\right)\sqrt{\frac{[579]^2}{[35{,}200]^2} + \frac{[504]^2}{[23{,}850]^2}} = 0.04.$$

Using the results above, the 90-percent confidence interval for this ratio is

$$[1.48 - (1.645 \times 0.04)] \text{ to } [1.48 + (1.645 \times 0.04)]$$
$$\text{or}$$
$$(1.41, 1.55).$$

Example 5: Computing the Standard Error for a Mean of Categorized Data. This example shows the steps for calculating the standard error for the average age of Asian householders in a hypothetical city, city C, in state D. The frequency distribution is given in Table 1.

Table 1. Frequency Distribution for Age of Asian Householder

Age of Asian Householder	Weighted Frequency
15 to 24 years	44,600
25 to 34 years	69,070
35 to 44 years	107,160
45 to 54 years	138,190
55 to 64 years	109,730
65 years and over	72,880

1. Cumulating the frequencies over the 6 categories yields an estimated population count of 541,630 Asian householders age 15 years and over. Suppose that we have 54,163 Asian households age 15 years and over in the 10-percent PUMS sample in city C, in state D, denoted n_d in formula (7).

2. Find the midpoint m_j for each of the 6 categories. Multiply each category's proportion p_j by the square of the midpoint and sum this product over all categories.

   For example, using formula (11), the midpoint of category 1 "15 to 24 years" is

   $$m_1 = \frac{15 + 24}{2} = 19.5,$$

   while the midpoint of the 6th category "65 years and over" is

   $$m_6 = \left(\frac{3}{2}\right) 65 = 97.5.$$

   The proportion of units in the first category, p_1, is

   $$p_1 = \frac{44{,}600}{541{,}630} = 0.08.$$

Information necessary to calculate the standard error is provided in Table 2.

Table 2. Calculations for Age of Asian Householder

Age of Asian Householder	p_j	m_j	$p_j m_j^2$	$p_j m_j$
15 to 24 years	0.08	19.5	30.42	1.56
25 to 34 years	0.13	29.5	113.13	3.84
35 to 44 years	0.20	39.5	312.05	7.90
45 to 54 years	0.26	49.5	637.07	12.87
55 to 64 years	0.20	59.5	708.05	11.90
65 years and over	0.13	97.5	1,235.81	12.68
		Total	3,036.53	50.75

3. To estimate the mean age of Asian householders, multiply each category's proportion by its midpoint and sum over all categories in the universe. Table 2 shows an estimated mean age of Asian householders, $\bar{x}$, of 50.75 years.

4. Calculate the estimated population variance using formula (10).

$$s^2 = 3{,}036.53 - (50.75)^2 = 460.97$$

5. Suppose the design factor for the population characteristic "Race of householder (race alone or in combination with one or more other races)" is larger than that for "Age of householder" and that the design factor for "Race of householder (race alone or in combination with one or more other races)" is 1.30. Using this information, formula (7), and the results from steps 1 through 4, the estimated standard error for the mean is

$$SE(50.75) = \sqrt{(1 - 0.1) \times \frac{460.97}{54{,}163}} \times 1.30$$

$$\approx 0.10 \text{ years.}$$

ESTIMATING A STANDARD ERROR BY THE DELETE-A-GROUP JACK-KNIFE VARIANCE METHOD

The delete-a-group jack-knife method is a replication technique that uses the PUMS sample directly to compute a standard error. This achieves a more accurate estimate of the standard error than using the generalized formulas. However, it increases processing time somewhat since it requires that the statistic of interest be computed separately for each of up to 100 replicate groups.

The general idea is to divide the full sample into replicate groups, calculate estimates for the full sample and the full sample without each specific replicate, and then use them to calculate a variance estimate. Using this method, it is also possible to compute standard errors for means, ratios, indexes, correlation coefficients, or other statistics for which the formulas presented earlier do not apply.

The delete-a-group jack-knife estimate of the variance is given by

$$v(\theta) = \frac{k-1}{k}\sum_{i=1}^{k}\left(\theta_{(i)} - \theta_{(\cdot)}\right)^2 \qquad (14)$$

where
θ = estimate of interest,
k = number of replicate groups,
$\theta_{(i)}$ = estimate excluding the i^{th} replicate group, and
$\theta_{(\cdot)}$ = full sample estimate.

Similar to how we use sample weights to create the full-sample estimates, we generate replicate weights to create the replicate estimates. In general, the replicate weight for each sample unit not in the i^{th} replicate group should be set to the sample weight adjusted by a factor of k / (k - 1). These replicate weights are used to create the replicate estimate excluding the i^{th} replicate group.

The standard error of the estimate is the square root of $v(\theta)$:

$$SE(\theta) = \sqrt{v(\theta)} \qquad (15)$$

An important aspect to consider with regard to the reliability of the delete-a-group jack-knife variance estimator is the number of groups, k. Often, a larger value of k produces a more reliable variance estimator. When using the 10-percent PUMS, k = 100 replicate groups is recommended. Use the subsample number assigned to each sample case to form the 100 groups. The subsample number has values from 00 to 99 as discussed in Chapter 4.

If the user chooses to use fewer than 100 replicate groups, use appropriate combinations of the two-digit subsamples to define the replicate groups. For example, to construct 50 replicate groups assign all records in which the subsample number is 01 or 51 to the first replicate group; all records in which the subsample number is 02 or 52 to the second replicate group; etc.

NONSAMPLING ERROR

In any large-scale statistical operation, such as the 2010 Census, human- and computer-related errors occur. These errors are commonly referred to as nonsampling errors. Such errors include not enumerating every household or every person in the population, not obtaining all required information from the respondents, obtaining incorrect or inconsistent information, and recording information incorrectly. In addition, errors can occur during the field review of the enumerators' work, during clerical handling of the census questionnaires, or during the electronic processing of the questionnaires.

Nonsampling error may affect the data in two ways: (1) errors that are introduced randomly will increase the variability of the data and, therefore, should be reflected in the standard error and (2) errors that tend to be consistent in one direction will make data biased in that direction. For

example, if respondents consistently tend to underreport their age, then the resulting counts of households or families by age of householder will tend to be understated for the higher ages and overstated for the lower ages. Such systematic biases are not reflected in the standard error.

While it is impossible to completely eliminate nonsampling error from an operation as large and complex as the decennial census, the Census Bureau attempts to control the sources of such error during the collection and processing operations. Described below are the primary sources of nonsampling error and the programs instituted to control this error in the 2010 Census. The success of these programs, however, was contingent upon how well the instructions actually were carried out during the census.

Nonresponse. Nonresponse to particular questions on the census questionnaire or the failure to obtain any information for a housing unit allows for the introduction of bias into the data because the characteristics of the nonrespondents were not observed and may differ from those reported by respondents. Minimizing nonresponse provides some protection against the introduction of large biases. Characteristics for the nonresponses were imputed by using reported data for a person or housing unit with similar characteristics. In some cases, this imputation filled in all the information for a person, called *whole-person imputation*. In other situations, it filled in individual characteristics for a person, called *characteristic imputation*.

As a result of the editing and imputation, there are no blank fields or missing data in the PUMS files. Each field contains a data value or a "not applicable" indicator, except for the few items where imputation was not appropriate and a "not reported" indicator is included. For every characteristic item, it is possible for the user to differentiate between entries that were imputed by means of imputation flags, referred to as "allocation flags" in the microdata files. For all items, it is possible to compute the imputation rate and compute the distribution of actually observed values (with imputed data omitted) and compare it with the overall distribution including imputed values.

Respondent and Enumerator Error. The person answering the mail questionnaire for a household or responding to the questions posed by an enumerator could serve as a source of error, although the question wording was extensively tested in several studies prior to the census. The mail respondent may overlook or misunderstand a question, or answer a question in a way that cannot be interpreted correctly by the data capture system. The enumerator may also misinterpret or otherwise incorrectly record information given by a respondent, or may fail to collect some of the information for a person or household. To control problems such as these with the field enumeration, the work of enumerators was monitored carefully. Field staff were prepared for their tasks by using standardized training packages that included hands-on experience in using census materials. A sample of the households interviewed by each enumerator was reinterviewed to control for the possibility of fabricated data being submitted by enumerators.

Processing Error. The many phases involved in processing the census data represent potential sources for the introduction of nonsampling error. The processing of the census questionnaires completed by enumerators included field review by the crew leader, check-in, and transmittal of

completed questionnaires. No field reviews were done on the mail return questionnaires for this census. Error may also be introduced by the misinterpretation of data by the data capture system, or the failure to capture all the information that the respondents or enumerators provided on the forms. Write-in entries go through coding operations, which may also be a source of processing error in the data. Many of the various field, coding, and computer operations undergo a number of quality control checks to help ensure their accurate application.

Disclosure Avoidance Activities. As mentioned in Chapter 2, disclosure avoidance techniques were applied to protect confidentiality. Some of these techniques such as data swapping, synthetic data, top-coding of selected variables, and age perturbation for large households change information to disguise data.