Accuracy of the Data (2000)

INTRODUCTION

The data contained in these Profiles and Summary Tables are based on the sample interviewed in 2000 from the Census 2000 Supplementary Survey (C2SS) and the 2000 comparison site tests. The C2SS is designed to provide accurate estimates for the housing units and population for the 50 states and the District of Columbia. The C2SS, like any other statistical activity, is subject to error. The purpose of this documentation is to provide data users with a basic understanding of the sample design, estimation methodology, and accuracy of the 2000 data.

The "*Operational Overview of the Census 2000 Supplementary Survey*" provides information on the data collection and Master Address File.

SAMPLE DESIGN

The sample for the C2SS and the 2000 comparison site tests uses a two-stage stratified sample of approximately 890,000 housing units designed to measure socioeconomic and demographic characteristics of housing units and their occupants. The C2SS samples housing units from the Master Address File (MAF). The first stage of sampling involves dividing the United States into primary sampling units (PSUs) —most of which comprise a metropolitan area, a large county, or a group of smaller counties. Every PSU falls within the boundary of a state. The PSUs are then grouped into strata on the basis of independent information, that is, information obtained from the decennial census or other sources. The strata are constructed so that they are as homogeneous as possible with respect to social and economic characteristics that are considered important by C2SS data users. A pair of PSUs were selected from each stratum. The probability of selection for each PSU in the stratum is proportional to its estimated 1996 population. In the second stage of sampling, a sample of housing units within the sample PSUs is drawn. Ultimate sampling units (USUs) are housing units. The USUs sampled in the second stage consist of housing units which are systematically drawn from sorted lists of addresses of housing units from the MAF.

PSU Definitions

For the most part, the C2SS PSU definitions are the same as the 1990 PSU definitions for the Current Population Survey (CPS). In forming the C2SS PSUs, changes were made to the CPS PSU definitions for the following reasons:

- Revised Metropolitan Statistical Area (MSA) definitions from Office of Management and Budget (OMB)
- C2SS used county-based instead of minor civil division (MCD)-based PSUs in New England and Hawaii
- Changes in county geography since the 1990 census.

Many PSUs are groups of contiguous counties rather than single counties.

The following are the rules used in defining the CPS PSUs:

- PSUs are contained within state boundaries.
- Metropolitan areas are defined as separate PSUs using projected 1990 Metropolitan Statistical Area (MSA) definitions. (An MSA is defined to be at least one county.) If an MSA straddles state boundaries, each state-MSA intersection is a separate PSU.
- For most states, PSUs are either one county or two or more contiguous counties. For the New England states and part of Hawaii, minor civil divisions (towns or townships) define the PSUs. In some states, county equivalents are used: cities, independent of any county organization, in Maryland, Missouri, Nevada, and Virginia; parishes in Louisiana; and boroughs and census divisions in Alaska.
- The area of the PSU should not exceed 3,000 square miles except in cases where a single county exceeds the maximum area.
- The population of the PSU is at least 7,500 except where this would require exceeding the maximum area specified in number 4.
- In addition to meeting the limitation on total area, PSUs are formed to limit extreme length in any direction and to avoid natural barriers within the PSU.

The C2SS design had 1,925 PSUs.

PSU Stratification

Initially all PSUs with an estimated 1996 population of at least 250,000 persons were designated to be self-representing (SR); that is, each of the SR PSUs is treated as a separate stratum and is included in the sample. In addition, any PSU which contained a 1999 American Community Survey (ACS) county was made SR. All other PSUs were designated as nonself-representing (NSR). Note that some initially designated NSR PSUs became SR during the stratification process. The following states are entirely SR: Connecticut, Delaware, Maine, Massachusetts, New Hampshire, New Jersey, Rhode Island, Vermont, and the District of Columbia.

For stratification, estimates of the total population for each county in 1996 were used to compute the measure of size for each PSU. For states, projected populations for the year 2000 were used to compute projected sample sizes at that level. Using the state population projection for the year 2000 and the number of persons per housing unit in each state (computed from 1996 data), a projected number of housing units for the year 2000 was derived for each state.

Stratification variables were chosen based on their relationship to variables considered important by C2SS data users. Variables used to stratify the PSUs included:

- Percent change in total PSU population between 1990 and 1996
- Number of vacant housing units (HUs) in 1990

- Percent change in number of HUs in PSU between 1980 and 1990
- Number of renter occupied HUs in 1990
- Rural farm population in 1990
- Number of related children under 18 below the poverty level in 1993 *(from the Census Bureau's Small Area Income and Poverty Estimates program)*
- Number of persons 16-19 in 1990 who are not enrolled in school and are not high-school graduates
- Total Hispanic population in 1990 (in states where Hispanics made up more than 10% of the projected total population for 2000): AZ, CA, CO, FL, IL, NV, NJ*, NM, NY, TX
- Total Black or African American population in 1990 (in states where blacks made up more than 10% of projected total population for 2000): AL, AR, CT*, DE*, DC*, FL, GA, IL, LA, MD, MI, MS, MO, NJ*, NY, NC, OH, PA, SC, TN, TX, VA

Note that the states marked with '*' are entirely self-representing (SR). Other information used in the stratification included target workloads and sample sizes in each state.

The sampling rate was based on a targeted annual national sample size of 890,000 housing units in both the C2SS and 2000 comparison site tests. For some small states this sampling interval yielded a sample size that was below the minimum annual state sample size of 7,000 persons. For these states, the sampling interval that yielded the minimum annual state sample size was used. Because of reductions that were made to some state sampling intervals during the stratification process (resulting in larger samples in those states), the final sampling interval for most states was determined to be 186.

For the estimation procedure, collapsed estimation strata were formed from the original PSU strata. There were three requirements placed on the collapsed strata:

1. Any ACS site was its own collapsed estimation stratum.
2. Any county with a Census 2000 household population of 250,000 or more which was self-representing was its own collapsed estimation stratum.
3. All other collapsed strata were formed by collapsing one or more PSU strata together in order to have a minimum of 400 sample interviews from C2SS.

In the third requirement, collapsed strata were formed of demographically similar and/or geographically contiguous PSU strata where possible. Generally, geography was used as the first criteria for grouping PSUs. The first two requirements are present so that the total housing unit and population estimates for published counties will agree with the Census 2000 estimates used for the controls.

The total number of collapsed estimation strata and total sample size by state is given in Table 1.

Table 1. Number and Sample Sizes of Strata by State

State	Number of Strata	Sample Size	Number of Interviews
Total	607	890,698	587,519
Alabama	13	10,975	6,873
Alaska	6	6,734	3,607
Arizona	4	25,825	16,780
Arkansas	9	8,059	5,132
California	30	83,323	52,597
Colorado	9	9,747	6,904
Connecticut	7	7,900	5,502
Delaware	3	6,542	4,227
District of Columbia	1	6,677	3,796
Florida	26	61,414	39,508
Georgia	14	19,243	11,812
Hawaii	2	6,587	4,119
Idaho	10	5,802	3,778
Illinois	17	33,931	22,892
Indiana	11	15,291	10,502
Iowa	12	14,162	10,756
Kansas	8	9,606	6,842
Kentucky	11	15,504	10,377
Louisiana	16	15,220	9,464
Maine	10	5,942	3,808
Maryland	13	14,933	10,228
Massachusetts	11	22,107	15,304
Michigan	19	23,472	16,286
Minnesota	11	11,062	8,549
Mississippi	11	15,187	9,353
Missouri	16	14,996	10,367
Montana	11	8,732	5,954
Nebraska	7	15,154	11,164
Nevada	4	6,283	3,810
New Hampshire	7	5,985	4,073
New Jersey	18	18,687	12,334
New Mexico	8	7,563	4,593
New York	25	60,779	36,315
North Carolina	14	19,008	12,149
North Dakota	9	6,271	4,450
Ohio	24	37,412	26,704
Oklahoma	8	8,417	5,438
Oregon	7	19,718	14,135
Pennsylvania	24	33,761	23,626
Rhode Island	3	6,608	4,373
South Carolina	10	9,996	6,047
South Dakota	8	8,342	6,194
Tennessee	14	14,922	9,909
Texas	35	52,444	32,369
Utah	5	5,806	3,975
Vermont	9	6,141	3,856
Virginia	18	16,744	11,686
Washington	9	17,197	11,622
West Virginia	10	12,627	8,208
Wisconsin	11	15,739	11,163
Wyoming	9	6,121	4,009

CONFIDENTIALITY OF THE DATA

Confidentiality Edit -- The sample itself provides adequate protection for most areas for which sample data are published since the resulting data are estimates of the actual characteristics. The non-ACS counties had a confidentiality edit implemented by identifying a subset of individual housing units from the sample data files as having a unique combination of specified person and household characteristics within a county. Because of the larger sample in the ACS data that is included in the C2SS data, the confidentiality edit was applied at the tract level. The confidentiality edit is controlled so that the basic structure of the data is preserved.

ESTIMATION PROCEDURE

The estimates that appear in this product were obtained from a ratio estimation procedure that resulted in the assignment of two sets of weights: a weight to each sample person record and a weight to each sample housing unit record. For any given tabulation area, a characteristic total was estimated by summing the weights assigned to the persons, households, families or housing units possessing the characteristic in the tabulation area. Estimates of person characteristics were based on the person weight. Estimates of family, household, and housing unit characteristics were based on the housing unit weight.

Each sample person or housing unit record was assigned exactly one weight to be used to produce estimates of all characteristics. For example, if the weight given to a sample person or housing unit had the value 160, all characteristics of that person or housing unit would be tabulated with the weight of 160. The estimation procedure, however, did assign weights varying from person to person or housing unit to housing unit.

The estimation procedure used to assign the weights was performed independently within each of the C2SS collapsed estimation strata.

- Initial Housing Unit Weighting Factors - This process produced the following factors:

    - Base Weight (BW) - This factor was assigned to every housing unit based on its counties' stratum times the inverse of the housing unit's sampling rate.

    - CAPI Subsampling Factor (SSF) - The weights of the CAPI cases were adjusted to reflect the results of CAPI subsampling. This factor was assigned to each record as follows:

        Selected in CAPI subsampling: SSF = 3.0
        Not selected in CAPI subsampling: SSF = 0.0
        Not a CAPI case: SSF = 1.0

    Some sample addresses were unmailable. A two-thirds sample of these were sent directly to CAPI and for these cases SSF = 1.5.

- Variation in Monthly Response by Mode (VMS) - This factor made the total weight of the Mail, Delivery, CATI, and CAPI records to be tabulated in a month equal to the total base weight of all cases originally mailed for that month. The value of VMS for Mail and Delivery cases was 1.0. For CATI and CAPI cases, VMS was computed and assigned based on the following groups.

  Strata x Month

- Noninterview Factor (NIF) - This factor adjusted the weight of all responding occupied housing units to account for both responding and nonresponding housing units. The factor was computed in two states. For the ACS sites only, a ratio adjustment NIF1 was computed and assigned to occupied housings units based on the the following groups.

  County x Building Type x Tract

  For both the C2SS national counties and the ACS sites, a second factor, assigned by a ratio adjustment NIF2, was computed and assigned to occupied housing units based on the following groups.

  Strata x Building Type x Month

  NIF was then computed by applying NIF1 and NIF2 for the ACS sites and just NIF2 for the C2SS national counties for each occupied housing unit. Vacant housing units were assigned a value of NIF = 1.0. Nonresponding housing units were now assigned a weight of 0.0.

- Noninterview Factor - Mode (NIFM) - This factor adjusted the weight of just the responding CAPI occupied housing units to account for both CAPI respondents and all nonrespondents. This factor was computed as if NIF had not already been assigned to every occupied housing unit record. This factor was not used directly but rather as part of computing the next factor: MBF. NIFM was computed and assigned to occupied CAPI housing units based on the following groups.

  Strata x Building Type x Month

  Mail and CATI cases received a value of NIFM = 1.0. Vacancies received a value of NIFM = 1.0.

- Mode Bias Factor (MBF) - This factor made the total weight of the housing units in the groups below the same as if NIFM had been used instead of NIF. MBF was computed and assigned to occupied housing units based on the following groups.

  Strata x Tenure (Owner or renter) x Month x Marital Status (married/widowed or other)

Vacant housing units received a value of MBF = 1.0. MBF is applied to the weights computed through NIF.

- Housing control Factor (HPF1) - This factor made the total weight of all housing units agree with the Census 2000 number of housing units at the collapsed strata level.

- Person Weighting Factors - Initially the person weight of each person in an occupied housing unit was the product of the weighting factors of their associated housing unit (BW x . . . xHPF1). At this point everyone in the household would have the same weight. These person weights were then individually adjusted based on each person's age, race, sex, and Hispanic origin as described below.

    - Person Post-Stratification Factor (PPSF) - This factor was applied to individuals based on their age, race, sex and Hispanic origin. It adjusted the person weights so that the weighted sample counts matched Census 2000 population counts by collapsed strata, age, race, sex, and Hispanic origin.

      This used the following groups:

      Strata x Race (non-Hispanic White, non-Hispanic Black, non-Hispanic American Indian or Alaskan Native, non-Hispanic Asian, non-Hispanic Native Hawaiian or Pacific Islander, and Hispanic(any race)) x Sex x Age Groups.

    - Rounding - The final product of all person weights (BW x . . . x HPF1 x PPSF) was rounded to an integer. Rounding was performed so that the sum of the rounded weights was within one person of the sum of the unrounded weights for any of the groups listed below:

      County
      County x Race
      County x Race x Hispanic Origin
      County x Race x Hispanic Origin x Sex
      County x Race x Hispanic Origin x Sex x Age
      County x Race x Hispanic Origin x Sex x Age x Tract
      County x Race x Hispanic Origin x Sex x Age x Tract x Block

      For example, the number of White, Hispanic, Males, Age 30 estimated for a county using the rounded weights was within one of the number produced using the unrounded weights.

- Final Housing Unit Weighting Factors - This process produced the following factors:

    - Principal Person Factor (PPF) - This factor adjusted for differential response depending on the race, Hispanic origin, sex, and age of the principal person in the household. The principal person was defined as the female spouse of the responding householder. If there was no such person, then the responding

householder was the principal person. The value of PPF for a housing unit was the PPSF of the principal person.

- Final Housing Unit Controls (HPF2) - The final product of the principal person weights (BW x . . . x HPF1 x PPF) was then assigned to the housing unit. The total number of weighted housing unit counts are then made to agree to the Census 2000 housing unit counts at the collapsed strata level.

- Rounding - The final product of all housing unit weights (BW x . . . x PPF x HPF2) was rounded to an integer. Rounding was performed so that total rounded weight was within one housing unit of the total unrounded weight for any of the groups listed below:

  County
  County x Tract
  County x Tract x Block

SPECIAL LIMITATIONS OF THE ESTIMATES

The Census 2000 Supplementary Survey estimates have some advantages which they will not have in future years. The numbers use the Census 2000 person and housing unit counts as controls in the weighting. In the future, these numbers will only be estimates from the population estimates program for the person weighting and Master Address File estimates for the housing unit controls.

ERRORS IN THE DATA

- Sampling Error -- The data in the C2SS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology. The estimates from the chosen sample also differ from other samples of housing units and persons within those housing units. Sampling error in data arises due to the use of probability sampling, which is necessary to insure the integrity and representativeness of sample survey results. The implementation of statistical sampling procedures provides the basis for the statistical analysis of sample data.

- Nonsampling Error -- In addition to sampling error, data users should realize that other types of errors may be introduced during any of the various complex operations used to collect and process survey data. For example, operations such as editing, reviewing, or keying data from questionnaires may introduce error into the estimates. These and other sources of error contribute to the nonsampling error component of the total error of survey estimates. Nonsampling errors may affect the data in two ways. Errors that are introduced randomly increase the variability of the data. Systematic errors which are consistent in one direction introduce bias into the results of a sample survey. The Census Bureau protects against the effect of systematic errors on survey estimates by conducting

extensive research and evaluation programs on sampling techniques, questionnaire design, and data collection and processing procedures. In addition, an important goal of the C2SS is to minimize the amount of nonsampling error introduced through nonresponse for sample housing units. One way of accomplishing this is by following up on mail nonrespondents during the CATI and CAPI phases.

- Standard Errors -- The standard error is a measure of the deviation of a sample estimate from the average of all possible samples. Sampling errors and some types of nonsampling errors are estimated by the standard error. The sample estimate and its estimated standard error permit the construction of interval estimates with a prescribed confidence that the interval includes the average result of all possible samples.

CONTROL OF NONSAMPLING ERROR

As mentioned earlier, sample data are subject to nonsampling error. This component of error could introduce serious bias into the data, and the total error could increase dramatically over that which would result purely from sampling. While it is impossible to completely eliminate nonsampling error from a survey operation, 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 for control of this error. The success of these programs, however, is contingent upon how well the instructions actually were carried out during the survey.

- Undercoverage -- It is possible for some sample housing units or persons to be missed entirely by the survey. The undercoverage of persons and housing units can introduce biases into the data. A major way to avoid undercoverage in a survey is to ensure that its sampling frame, for C2SS an address list in each state, is as complete and accurate as possible.

  The source of addresses was the Master Address File (MAF). The MAF is created by combining the 1990 Census Address Control File, the Delivery Sequence File of the United States Postal Service, and addresses listed for Census 2000. An attempt is made to assign all appropriate geographic codes to each MAF address via an automated procedure using the Census Bureau TIGER files. A manual coding operation based in the appropriate regional offices is attempted for addresses which could not be automatically coded. The MAF was used as the source of addresses for selecting sample housing units and mailing questionnaires. TIGER produced the location maps for personal visit CAPI assignments.

  In the CATI and CAPI nonresponse follow-up phases, efforts were made to minimize the chances that housing units that were not part of the sample were interviewed in place of units in sample by mistake. If a CATI interviewer called a mail nonresponse case and was not able to reach the exact address, no interview was conducted and the case was eligible for CAPI. During CAPI follow-up, the interviewer had to locate the exact address for each sample housing unit. In some multi-unit structures the interviewer could

not locate the exact sample unit or found a different number of units than expected. In these cases the interviewers were instructed to list the units in the building and follow a specific procedure to select a replacement sample unit.

- Respondent and Interviewer Error -- The person answering the questionnaire or responding to the questions posed by an interviewer could serve as a source of error, although the questions were phrased as clearly as possible based on testing, and detailed instructions for completing the questionnaire were provided to each household. In addition, respondents' answers were edited for completeness, and problems were followed up as necessary.

    - Interviewer monitoring -- The interviewer may misinterpret or otherwise incorrectly enter information given by a respondent; may fail to collect some of the information for a person or household; or may collect data for households that were not designated as part of the sample. To control these problems, the work of interviewers was monitored carefully. Field staff were prepared for their tasks by using specially developed training packages that included hands-on experience in using survey materials. A sample of the households interviewed by CAPI interviewers was reinterviewed to control for the possibility that interviewers may have fabricated data.

- Item Nonresponse -- Nonresponse to particular questions on the survey questionnaire and instrument allows for the introduction of bias into the data, since the characteristics of the nonrespondents have not been observed and may differ from those reported by respondents. As a result, any imputation procedure using respondent data may not completely reflect this difference either at the elemental level (individual person or housing unit) or on average.

  Some protection against the introduction of large biases is afforded by minimizing nonresponse. In the C2SS, nonresponse for the CATI and CAPI operations was reduced substantially by the requirement that the automated instrument receive a response to each question before the next one could be asked. For mail responses, the automated clerical review and follow-up operations were aimed at obtaining a response for every question on selected questionnaires. Values for any items that remain unanswered were imputed by computer using reported data for a person or housing unit with similar characteristics.

- Automated Clerical Review -- Questionnaires returned by mail were edited for completeness and acceptability. They were reviewed by computer for content omissions and population coverage. If necessary, a telephone follow-up was made to obtain missing information. Potential coverage errors were included in this follow-up, as well as questionnaires with too many omissions to be accepted as returned.

- Processing Error -- The many phases involved in processing the survey data represent potential sources for the introduction of nonsampling error. The processing of the survey questionnaires includes the keying of data from completed questionnaires, automated clerical review, and follow-up by telephone; the manual coding of write-in responses;

and the electronic data processing. The various field, coding and computer operations undergo a number of quality control checks to insure their accurate application.

- Automated Editing -- After data collection was completed, any remaining incomplete or inconsistent information was imputed during the final automated edit of the collected data. Imputations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, were needed most often when an entry for a given item was lacking or when the information reported for a person or housing unit on that item was inconsistent with other information for that same person or housing unit. As in other surveys and previous censuses, the general procedure for changing unacceptable entries was to assign an entry for a person or housing unit that was consistent with entries for persons or housing units with similar characteristics. Assigning acceptable values in place of blanks or unacceptable entries enhances the usefulness of the data.

CALCULATION OF STANDARD ERRORS

Direct Standard Errors

Methodology Used -- Direct estimates of the standard errors were calculated for all estimates reported in this product. They are provided in the summary tables and profiles as 90 percent confidence intervals. The standard errors, in most cases, are calculated using standard variance estimation software using a methodology that takes into account the sample design and estimation procedures.

Exceptions -- There are seven cases for which the direct standard error estimates are not appropriate.

1. The estimate of the number or proportion of people, households, housing units or families in a geographic area with a specific characteristic is zero. A special procedure was used to estimate the standard error.

2. There are no sample observations available to compute an estimate of a proportion or other ratio or an estimate of its standard error. The estimate is represented in the tables by "-" and the lower and upper bounds of the 90 percent confidence interval by "**".

3. There are no sample observations available to compute an estimate of a median or an estimate of its standard error. The estimate is represented in the tables by "-" and the lower and upper bounds of the 90 percent confidence interval by "**".

4. Only a small number of identical values are reported and used to calculate an aggregate, mean, or per capita amount. In this case, there are too few sample observations to compute a stable estimate of the standard error. The lower and upper bounds of the 90 percent confidence interval are represented in the tables by "*".

5. The estimate of a median falls in the lowest interval or upper interval of an open-ended distribution. If the median occurs in the lowest interval, then a "-" follows the estimate, and if the median occurs in the upper interval, then a "+" follows the estimate. In both cases the lower and upper bounds of the 90 percent confidence interval are represented in the tables by "***".

6. The estimate of the number of people having a specified characteristic is controlled to be equal to an independently derived population estimate. These estimates are those from Census 2000. For these cases the standard error is zero. The lower and upper bounds of the 90 percent confidence interval are represented in the tables by "*****". (See "ESTIMATION PROCEDURE" for a further explanation.)

7. The estimate of the number of housing units is controlled to be equal to an independently derived housing unit estimate. These estimates are those from Census 2000. For these cases the standard error is zero. The lower and upper bounds of the 90 percent confidence interval are represented in the tables by "*****". (See "ESTIMATION PROCEDURE" for a further explanation.)

Calculating Standard Errors from the 90 Percent Confidence Interval -- In most cases you can calculate the standard error using the estimate and the upper bound. If the upper bound has been set to its largest admissible value (See Limitation 2. below) then the lower bound should be used instead of the upper bound.

Standard Error = (upper bound - estimate) / 1.65

or

Standard Error = (estimate - lower bound) / 1.65

Sums and Differences of Direct Standard Errors -- The standard errors estimated from these tables are for individual estimates. Additional calculations are required to estimate the standard errors for sums of and differences between two sample estimates. The estimate of the standard error of a sum or difference is approximately the square root of the sum of the two individual standard errors squared; that is, for standard errors $SE(\hat{X})$ and $SE(\hat{Y})$ of estimates $\hat{X}$ and $\hat{Y}$:

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

This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated.

Ratios -- Frequently, the statistic of interest is the ratio of two variables, where the numerator *is not* a subset of the denominator. The standard error of the ratio between two sample estimates is approximated as follows:

$$SE\left(\frac{\hat{X}}{\hat{Y}}\right) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 + \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$$

Proportions/percents - The statistic of interest may be a proportion or percent, where the numerator *is* a subset of the denominator. Note the difference between the formulas for the standard error for proportions and ratios.

$$SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$$

Confidence Intervals

Confidence Intervals -- A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that will contain the average value of the estimated characteristic that results over all possible samples, with a known probability.

For example, if all possible samples that could result under the C2SS sample design were independently selected and surveyed under the same conditions, and if the estimate and its estimated standard error were calculated for each of these samples, then:

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

2. Approximately 90 percent of the intervals from 1.65 times the estimated standard error below the estimate to 1.65 times the estimated standard error above the estimate would contain the average result from all possible samples.

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

The intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively.

Lower and Upper Bounds -- The lower and upper bounds presented in the summary tables and profiles are the bounds based upon a 90 percent confidence interval.

Limitations -- The user should be careful when computing and interpreting confidence intervals.

1. The estimated standard errors included in this data product do not include all portions of the variability due to nonsampling error that may be present in the data. In particular, the standard errors do not reflect the effect of correlated errors introduced by interviewers, coders, or other field or processing personnel. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (i.e., 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this data product based on the estimated standard errors.

2. Zero or small estimates; very large estimates -- The value of almost all C2SS characteristics is greater than or equal to zero by definition. For zero or small estimates, use of the method given previously for calculating confidence intervals relies on large sample theory, and may result in negative values which for most characteristics are not admissible. In this case the lower limit of the confidence interval is set to zero by default. A similar caution holds for estimates of totals close to a control total or estimated proportions near one, where the upper limit of the confidence interval is set to its largest admissible value. In these situations the level of confidence of the adjusted range of values is less than the prescribed confidence level.

EXAMPLES- STANDARD ERROR CALCULATIONS

We will present some examples based on the real data to demonstrate the use of the formulas.

Example 1 - Calculating the Standard Error from the Confidence Interval

The estimated number of males, never married is 30,952,067 from summary table P031 in the US. The lower bound is 30,823,400 and the upper bound is 31,080,734.

Standard Error = (upper bound - estimate) / 1.65 = (estimate - lower bound) / 1.65

Calculating the standard error using the upper bound we have:

SE(30,952,067) = (31,080,734 - 30,952,067) / 1.65 = 77,980.

Example 2 - Calculating the Standard Error of a Sum

We are interested in the number of people who have never been married. From summary table P031 we have the number of males, never married is 30,952,067 with an upper bound of 31,080,734; and the number of females, never married is 26,977,973 with an upper bound of 27,106,155. So the estimated number of people who have never been married is 30,952,067 + 26,977,973 = 57,930,040. To calculate the standard error of this sum, we need the standard errors of the two estimates in the sum. We have the standard

error for the number of males never married from example 1 as 77,980. The standard error for the number of females never married is calculated using the upper bound:

SE(26,977,973) = (27,106,155 - 26,977,973) / 1.65 = 77,686.

So using the formula for the standard error of a sum or difference we have:

$$SE(57,930,040) = \sqrt{77,980^2 + 77,686^2} = 110,073.$$

Caution: This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated.

To calculate the lower and upper bounds of the 90 percent confidence interval around 57,930,040 using the standard error, simply multiply 110,073 by 1.65, then add and subtract the product from 57,930,040. Thus the 90 percent confidence interval for this estimate is [57,930,040 - 1.65(110,073)] to [57,930,040 + 1.65(110,073)] or 57,748,420 to 58,111,660.

Example 3 - Calculating the Standard Error of a Percent

We are interested in the percentage of females who have never been married to the number of people who have never been married. The number of females, never married is 26,977,973 and the number of people who have never been married is 57,930,040. To calculate the standard error of this sum, we need the standard errors of the two estimates in the sum. We have the standard error for the number of females never married from example 2 as 77,686 and the standard error for the number of people never married calculated from example 2 as 110,073.

The estimate is (26,977,973 / 57,930,040) * 100 = 46.6%

So using the formula for the standard error of a ratio we have:

$$SE(46.6) = \left(\frac{1}{57,930,040} \sqrt{77,686^2 - 0.466^2 \times 110,073^2} \right) * 100 = 0.1\%.$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 46.6 using the standard error, simply multiply 0.1 by 1.65, then add and subtract the product from 46.6. Thus the 90 percent confidence interval for this estimate is [46.6 - 1.65(0.1)] to [46.6 + 1.65(0.1)] or 46.4% to 46.8%.