Category: Dissemination

Revision Date: 9/24/02
Previously issued: 3/23/01

Census Bureau Standard: Source and Accuracy Statements for Census and Survey Data Tabulations and Model-Based Estimates

Introduction

The Census Bureau Methodology and Standards Council sets statistical and survey standards for Census Bureau surveys and censuses. This responsibility encompasses setting relevant standards for Census Bureau data products. This standard establishes a three tiered approach for informing data users of the statistical reliability of data as part of the release of tabular survey and census data, and of model-based estimates, through all electronic and paper formats

Scope

These standards apply to all tabulated data and model-based estimates disseminated by the Census Bureau, however collected, derived, or assembled, and irrespective of the medium of dissemination or the source of funding.

Standard

The approach consists of three standard statements that provide increasingly more detail about the statistical quality of the data and its appropriate uses. Divisions may modify the statements to meet the needs of individual programs but the standard language should be used to the extent possible, because it has been tested to ensure that users understand it. The components of the approach are described below.

1. **Headnote for all Tables.** Each table contains a source note (labeled SOURCE). This note is positioned at the top of multi-page tables; at the top or bottom of single-page tables. The note does not typically exceed two printed lines. The first sentence of the note provides a description of the data in the table — sometimes referred to as the source-type. The second sentence provides the location of relevant information on confidentiality protection, sampling error, nonsampling error, model error, and definitions. For electronic tables, the location source has a hyperlink. Examples of the Headnote are provided in Attachment 1.

2. **Location Source Hyperlink (electronic) or Intermediate Page (paper) Statement**. The reader is referred to a statement that has the following six components. Specific examples are provided in Attachment 2. Bracketed information in the italicized statements below indicate alternate wording for different programs.

    - The source-type for the data used in the tabulations or to develop the model-based estimates (sample, census, administrative records, etc.)

- Statement on disclosure procedures where appropriate for the census or survey tabulations:

  To maintain confidentiality, the Census Bureau [applies statistical procedures that introduce some uncertainty into data for geographic areas with small population groups/ suppresses data to protect the identity of any business or individual.]

  Statement on disclosure procedures where appropriate for the model-based estimates:

  To maintain confidentiality, the Census Bureau uses procedures to assure that the estimates and related information that are released cannot be used to disclose individual data or violate other confidentiality restrictions applicable to the source data.

- Statement on sampling and nonsampling error for tabulations:

  The survey [census] results in this table contain sampling error and nonsampling error [or contain nonsampling error but do not contain sampling error].

  Statement on model error, sampling error, and nonsampling error for model-based estimates:

  The estimates contain error stemming from model error, sampling error, and nonsampling error. Standard errors are provided to indicate the quality of the estimates and may be used to produce confidence intervals. [or, Confidence intervals are provided to indicate the quality of the estimates.] Subject to the validity of the underlying model assumptions, these reflect uncertainty due to the effects of model error and sampling error, but do not account for the effects of nonsampling error.

- Statement on uses of data extrapolated from tables:

  Data users who create their own estimates using data from [American FactFinder/ Ferret/Census Bureau] tables should cite the Census Bureau as the source of the original data only.

- Statement and location of other definitions of potential interest:

  See also [definitions of subject characteristics and geographic definitions/explanation of terms and geographic definitions].

- Location of full technical documentation for tabulations:

  For the full technical documentation [for the data set, which is the source of data in this table], see [www location/ page in documentation].

Location of full technical documentation for model-based estimates:

For the full technical documentation of the estimates see [www location/ page in documentation]. For technical documentation of the [survey, census, or other, e.g., administrative records] source data, see [www location/ page in documentation].

3. **Technical Documentation**. This consists of the full technical documentation for the program. It should specifically include the statements on nonsampling error, sampling error, confidentiality, Title 13 United States Code, disclosure limitation, and the statements on suppression, data swapping, and protection of microdata files as relevant that are provided in Attachment 3. An attempt was made to develop statements that addressed the methodology of all program areas. If the language in the statement does not address the specific needs of an individual program, modifications may be made. But the standard language should be used to the extent possible because it has been cognitively tested to ensure that users understand it.

Responsibilities

Program areas will be responsible for the following:

- the preparation of the appropriate statistical reliability components based on these standards;
- the clearance of the statistical reliability components.

The Methodology and Standards Council will be responsible for the following:

- the initiation of periodic evaluations and reviews of statistical reliability statements, and updates to the standards, as necessary;
- guidance to program areas in the development and usage of statistical reliability statements.

Implementation

This standard is ready for implementation immediately in the American FactFinder and in the next round of estimates for the Small Area Income and Poverty Estimates (SAIPE) program. Other electronic and paper census and survey data dissemination programs should incorporate this standard into their tabular products by the end of the calendar year. New programs for data tabulation or model-based estimates should incorporate this standard into the release of their first set of estimates. Specific situations where this three-tiered approach is not viable for the statistical product or the implementation cannot occur according to the above schedule should be brought to the attention of the Methodology and Standards Council so that a reasonable alternative can be developed.

Inquiries

Inquiries relating to the interpretation of this standard should be addressed to the Census Bureau Methodology and Standards Council.

Recommended by Census Bureau Methodology and Standards Council:
(signed 09/24/02)

_____
Cynthia Z.F. Clark,
Associate Director for Methodology and Standards

_____
Robert Tinari, CASRO

_____
David Whitford, (Acting) DSSD

_____
Al Tupek, DSMD

_____
Howard Hogan, ESMPD

_____
Ruth Ann Killion, PRED

_____
Tommy Wright, SRD

Concurrence: (_____)

_____
Preston Jay Waite
Associate Director for Decennial Programs

_____
Nancy M. Gordon
Associate Director for Demographic Programs

_____
Frederick T. Knickerbocker
Associate Director for Economic Programs

_____
Marvin D. Raines
Associate Director for Field Operations

_____
Cynthia Z.F. Clark
Associate Director for Methodology and Standards

Examples of Headnotes for Tables

The following examples are designed for internet dissemination. Analogous notes for paper dissemination should reference the appropriate technical documentation in the second sentence. That technical documentation should contain the respective Intermediate Page Statement (see Attachment 2 for examples), and the relevant technical documentation statements (see Attachment 3).

1. For 100 percent Census 2000 data.

SOURCE: Census 2000. For information on confidentiality protection, nonsampling error, and definitions, see http://www.census.gov/explanation.html.

2. For 100 percent Census 2000 data with corrections for coverage errors from A.C.E.

SOURCE: Census 2000. Data corrected for coverage errors based on the Accuracy and Coverage Evaluation. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www.census.gov/explanation.html

3. For Census 2000 sample data.

SOURCE: Census 2000 long form sample. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www.census.gov/explanation.html

4. For American Community Survey data.

SOURCE: American Community Survey, data based on twelve monthly samples during 2000. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http;://www.acs/explanation.html

5. For Demographic Survey data (example - Survey of Income and Program Participation)

SOURCE: Survey of Income and Program Participation . For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www.sipp.census.gov/sipp/methmain.html

6. For Economic Census data.

SOURCE: 2002 Economic Census. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www.econcensus/explanation.html

7. For Economic Survey data (example - Service Annual Survey)

SOURCE. 1998 Service Annual Survey. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www.census.gov/svsd/www/cv.html.

8. For Small Area Income and Poverty Estimates (SAIPE).

SOURCE: Estimates for income year X obtained from models applied to CPS data. Model inputs also include Census 2000 and administrative records data. For information on the data sources, measures of uncertainty, confidentiality protection, sources of error, and comparisons with decennial census income and poverty estimates, see http://www.census.gov/hhes/www/saipe/documentation.html.

Examples of Location Source Hyperlink (electronic) or Intermediate Page (paper) Statement

Hyperlink statements for confidentiality, sampling error, nonsampling error, and model error (for model-based estimates) for the examples below are provided in Attachment 3.

1. Census 2000 data

The data in [this table/these tables] are from Census 2000 [and have been adjusted for estimated net census coverage error based on the results of the Accuracy and Coverage Evaluation]. To maintain confidentiality, the Census Bureau applies statistical procedures that introduce some uncertainty into data for geographic areas with small population groups. The results in [this table/these tables] contain sampling error and nonsampling error. Data users who create their own estimates using data from American FactFinder tables should cite the Census Bureau as the source of the original data only. See also definitions of subject characteristics and geographic definitions. For the full technical documentation for the Census 2000 Summary File 1 (SF1), which is the source of data in this table, see http://www.census.gov/techdoc/sf1.pdf

2. American Community Survey data

The data in [this table/these tables] are based on the American Community Survey conducted in 2000. To maintain confidentiality, the Census Bureau applies statistical procedures that introduce some uncertainty into data for geographic areas with small population groups. The data in [this table/these tables] contain sampling error and nonsampling error. Data users who create their own estimates using data from American FactFinder tables should cite the Census Bureau as the source of the original data only. See also definitions of subject characteristics and geographic definitions. For the full technical documentation, see http://www.census.gov/techdoc/acs.pdf

3. Demographic Survey Data (example - Survey of Income and Program Participation)

The data in [this table/these tables] are based on the Survey of Income and Program Participation. The data in [this table/these tables] contain sampling error and nonsampling error. Data users who create their own estimates using data from Census Bureau tables should cite the Census Bureau as the source of the original data only. See also definitions of subject characteristics and geographic definitions. For the full technical documentation, see http://www.sipp.census.gov/sipp/methmain.htm.

4. Economic Census data

The data in [this table/these tables] are based on the 2002 Economic Census. To maintain confidentiality, the Census Bureau suppresses data to protect the identity of any business or individual. The census results in [this table/these tables] contain nonsampling error. Data users who create their own estimates using data from American FactFinder tables should cite the Census Bureau as the source of the original data only. See also explanation of terms and geographic definitions. For the full technical documentation, see http://www.census.gov/techdoc/econcensus.pdf.

5. Economic Survey data (example - Service Annual Survey)

The data in [this table/these tables] are based on the 1998 Service Annual Survey. To maintain confidentiality, the Census Bureau suppresses data to protect the identity of any business or individual. The data in [this table/these tables] contain sampling error and nonsampling error. Data users who create their own estimates using data from American FactFinder tables should cite the Census Bureau as the source of the original data only. See also explanation of terms and geographic definitions. For the full technical documentation for the 1998 Service Annual Survey, see http://www.census.gov/techdoc/sas.pdf

6. Model Based Estimates (example — Small Area Income and Poverty Estimates (SAIPE))

The data in [this table/these tables] are indirect estimates produced by statistical model-based methods using sample survey, decennial census, and administrative data sources. To maintain confidentiality, the Census Bureau uses procedures to assure that the estimates and related information that are released cannot be used to disclose individual data or violate other confidentiality restrictions applicable to the source data. The estimates contain error stemming from model error, sampling error, and nonsampling error. Confidence intervals are provided to indicate the quality of the estimates. Subject to the validity of the underlying model assumptions, these reflect uncertainty due to the effects of model error and sampling error, but do not account for the effects of nonsampling error. In addition, one should exercise caution when making comparisons between these estimates for different years or different places, or between these model-based estimates and other Census Bureau estimates. The SAIPE web site contains technical documentation for each set of estimates produced (see http://www.census.gov/hhes/www/saipe/documentation.html). Additional information is available in a final report prepared by the National Academy of Sciences that evaluates the methods used to produce the estimates.

Technical Documentation Statements

Nonsampling error may occur during the development or execution of a survey. There are several sources of nonsampling error. These errors can occur because of circumstances created by the interviewer, the respondent, the survey instrument, or the way the data are collected and processed. For example, errors could occur because:

- the interviewer records the wrong answer, the respondent provides incorrect information, the respondent estimates the requested information, or an unclear survey question is misunderstood by the respondent (measurement error);
- some individuals or businesses which should have been included in the census or survey were omitted (coverage error);
- responses are not collected from all those in the sample (nonresponse error);
- forms may be lost, data may be incorrectly keyed, coded or recoded, etc. (processing error).

Information about nonsampling error for a specific program is provided or referenced with the data. The Census Bureau recommends that data users incorporate this information into their analyses, as nonsampling error could impact the conclusions drawn from the results.

Sampling error is the difference between an estimate based on a sample and the corresponding value that would be obtained if the estimate were based on the entire population (as from a census). Note that sample-based estimates will vary depending on the particular sample selected from the population. Measures of the magnitude of sampling error in direct survey tabular estimates (variances, standard deviations, or coefficients of variation) reflect the variation in the estimates over all possible samples that could have been selected from the population using the same sampling methodology.

Estimates of the magnitude of sampling errors for a specific estimates program are provided or referenced with the data. The Census Bureau recommends that data users incorporate this information into their analyses, as sampling error in survey estimates could impact the conclusions drawn from the results.

Error in model-based estimates arises from the effects of model error, sampling error, and nonsampling error. The relative contribution of these error components to the error in the model-based estimates depends on the model used and the properties of the data. Standard errors provided for the model-based estimates reflect, to the extent possible and subject to the model assumptions, the contributions of model error and sampling error, but do not reflect the contribution of nonsampling error.

Model error refers to error that would result in predictions from a statistical model even with no errors in the data (no sampling or nonsampling error). It can generally be broken down into three components: error that would occur in the predictions even if the true model were known; error resulting from estimation of model parameters; and error resulting from differences between the form of the assumed model and that of the true, unknown, model.

Confidentiality

For data tabulations:

> The Census Bureau has modified or suppressed some data on this site to protect confidentiality. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's or business' data can be identified.
>
> The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed. For more information on how the Census Bureau protects the confidentiality of data, see the disclosure limitation topics listed below as appropriate.

For model-based estimates:

> Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's or business' data can be identified. The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. For more information on how the Census Bureau protects the confidentiality of data used in producing model-based estimates, see the discussions of Disclosure Limitation and Release of Source Data.

Disclosure Limitation Procedures

- Suppression (for data tabulations)
- Data Swapping (for data tabulations)
- Protection of Microdata Files (for data tabulations)
- Release of Source Data (for model-based estimates)

Questions about confidentiality may be addressed to: [webmaster@census.gov /name of contact person]

Title 13, United States Code: Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to five years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information.

Disclosure Limitation: Disclosure limitation is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify either an individual or business that has provided information under a pledge of confidentiality. For data tabulations the Census Bureau uses disclosure limitation procedures to modify or remove the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual or business, the Census Bureau has taken steps to disguise or suppress the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of

data. For model-based estimates the Census Bureau uses other procedures to assure that the estimates and related information that are released cannot be used to disclose individual data.

Suppression: Suppression is a method of disclosure limitation used to protect individuals' confidentiality by not showing (suppressing) the cell values in tables of aggregate data for cases where only a few individuals or businesses are represented. The cells that are not shown are called *primary suppressions*. To make sure the primary suppressions cannot be closely estimated by subtracting the other cells in the table from the marginal totals, additional cells are also suppressed. These additional suppressed cells are called *complementary* or *secondary suppressions*. The process of suppression does not change the marginal totals, so the integrity of the data is not adversely affected. Before the Census Bureau releases data, computer programs check published tables for both primary and complementary disclosures. Suppression was used for the 1980 Census of Population and Housing and is now used for economic surveys and censuses.

Example -- With Disclosure

Value of Shipments
by County and Industry

	Industry				
County	W	X	Y	Z	Total
Alpha	15	1*	3*	1*	20
Beta	20	10	10	15	55
Gamma	3*	10	10	2*	25
Delta	12	14	7	2*	35
Total	50	35	30	20	135

NOTE: * Indicates cells in which data may be identifiable due to the low number in the cell.

Example -- Without Disclosure, Protected by Suppression

Value of Shipments
by County and Industry

	Industry				
County	W	X	Y	Z	Total
Alpha	15	D	D	D	20
Beta	20	10	10	15	55
Gamma	D	D	10	D	25
Delta	D	14	D	D	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.

Data Swapping: Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals. Data swapping procedures were first used in the 1990 Census, and will be used for Census 2000 and for the Census Bureau's new American Community Survey. For a description of the disclosure limitation procedures used in the economic census and surveys see the discussion on suppression.

Protection of Microdata Files: The Census Bureau sometimes releases microdata files which contain data from the censuses of the United States population and household surveys, which it conducts. These files contain individuals' responses that represent only samples of the population and have had all individual identifiers (such as name and address) removed from the records. In addition, to protect confidentiality, the Census Bureau may modify distinguishing characteristics (such as high levels of income), and restrict geographic identifiers (such as the name of a city) so that populations are composed of at least 100,000 people. This is done to protect the identity of individuals.

Release of Source Data: Source data used in the production of model-based estimates are released to the public only when such release would not disclose individual data or violate other confidentiality restrictions applicable to the source data. This includes both source data obtained from Census Bureau surveys and censuses and also data obtained from other sources such as other government agencies.