Accuracy of the Data (2006)

INTRODUCTION

The data contained in these data products are based on the American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) sample interviewed from January 1, 2006 through December 31, 2006. [Unless otherwise specified, the term "ACS" in this document will refer to both the ACS and PRCS.] The ACS sample is selected from all counties and county-equivalents in the United States, and all municipios in Puerto Rico (PR). In 2006, the ACS began collection of data from sampled persons in group quarters (GQs) – for example, military barracks, college dormitories, nursing homes, and correctional facilities. Persons in group quarters are included with persons in housing units (HUs) in all 2006 ACS estimates based on the total population. All ACS population estimates from previous years include only persons in housing units. The ACS, 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 ACS sample design, estimation methodology, and accuracy of the ACS data. The ACS is sponsored by the U.S. Census Bureau, and is part of the 2010 Decennial Census Program.

Additional information on the operational aspects of the ACS, including data collection and processing, can be found in the ACS Operations Plan (http://www.census.gov/acs/www/Downloads/OpsPlanfinal.pdf).

DATA COLLECTION

Housing Units

The ACS and PRCS employ three modes of data collection:

- Mailout/Mailback
- Computer Assisted Telephone Interview (CATI)
- Computer Assisted Personal Interview (CAPI)

With the exception of addresses in Remote Alaska, the general timing of data collection is:

Month 1: Addresses in sample that are determined to be mailable are sent a questionnaire via the U.S. Postal Service.
Month 2: All mail non-responding addresses with an available phone number are sent to CATI.
Month 3: A sample of mail non-responses without a phone number, CATI non-responses, and unmailable addresses are selected and sent to CAPI.

Note that mail responses are accepted during all three months of data collection.

All Remote Alaska addresses are assigned to one of two data collection periods, January-April, or September-December and are sampled for CAPI at a rate of 2-in-3. Data for these addresses are collected using CAPI only and up to four months are given to complete the interviews in Remote Alaska for each data collection period.

Group Quarters

Field representatives have several options available to them for data collection. These include completing the questionnaire while speaking to the resident in person or over the telephone, conducting an personal interview with proxy, such as a relative or guardian, or leaving paper questionnaires for residents to complete for themselves and then pick them up later. This last option is used for data collection in Federal prisons.

Group Quarters data collection spans six weeks, except in Remote Alaska and for Federal prisons, where the data collection time period is four months. As is done for HUs, Group Quarters in Remote Alaska are assigned to one of two data collection periods, January-April, or September-December and up to four months is allowed to complete the interviews. Similarly, all Federal prisons are assigned to September with a four month data collection window.

SAMPLING FRAME

Housing Units

The universe for the ACS consists of all valid, residential housing unit addresses in all county and county equivalents in the 50 states, including the District of Columbia, and Puerto Rico. The Master Address File (MAF) is a database maintained by the Census Bureau containing a listing of residential and commercial addresses in the U.S. and Puerto Rico. The MAF is updated twice each year with the Delivery Sequence Files provided by the U.S. Postal Service which cover only the U.S. These files identify mail drop points and provide the best available source of changes and updates to the housing unit inventory. The MAF is also updated with the results from various Census Bureau field operations, including the ACS.

Group Quarters

The group quarters (GQ) sampling frame is created from the Special Place (SP)/GQ facility files, obtained from decennial census operations, merged with the MAF. This frame includes GQs added from operations such as the GQ Incomplete Information Operation (IIO) at the Census Bureau's National Processing Center in Jeffersonville, Indiana, the Census Questionnaire Resolution (CQR) Program and GQs closed on Census day. The GQ frame underwent an unduplication process. GQs that were closed on Census day were not included in the SP/GQ inventory file received from Decennial Systems and Contract Management Office (DSCMO). These were added from a preliminary inventory file obtained from DSCMO since it was possible that while these GQs were closed on Census day, they could be open when the ACS contacts them. Headquarters Staff researched state prisons on the Internet to obtain the current operating

status and the current population counts for state prisons. After the frame was put together from these different sources, it was then sorted geographically.

SAMPLE DESIGN

Housing Units

The ACS employs a two-stage, two-phase sample design. The ACS first-stage sample consists of two separate samples, Main and Supplemental, each chosen at different points in time. Together, these constitute the first-stage sample. Both the Main and the Supplemental samples are chosen in two phases referred to as first- and second-phase sampling. Subsequent to second-phase sampling, sample addresses are randomly assigned to one of the twelve months of the sample year. The second-stage of sampling occurs when the CAPI sample is selected (see Section 2 below).

The Main sample is selected during the summer preceding the sample year. Approximately 99 percent of the sample is selected at this time. Each address in sample is randomly assigned to one of the 12 months of the sample year. Supplemental sampling occurs in January/February of the sample year and accounts for approximately 1 percent of the overall first-stage sample. The Supplemental sample is allocated to the last nine months of the sample year. The sample of addresses from both Main and Supplemental is referred to as the initially selected sample. A sub-sample of non-responding addresses and of any addresses deemed unmailable is selected for the CAPI data collection mode.

Several of the steps used to select the first-stage sample are common to both Main and Supplemental sampling. The descriptions of the steps included in the first-stage sample selection below indicate which are common to both and which are unique to either Main or Supplemental sampling.

1. First-Stage Sample Selection

   - First-phase sampling (*performed during both Main and Supplemental sampling*) – First stage sampling defines the universe for the second stage of sampling through two steps. First, all addresses that were in a first-stage sample within the past four years are excluded from eligibility. This ensures that no address is in sample more than once in any five-year period. The second step is to select a 20 percent systematic sample of "new" units, i.e. those units that have never appeared on a previous MAF extract. Each new address is systematically assigned to either the current year or to one of four back-samples. This procedure maintains five equal partitions of the universe.

   - Assignment of blocks to a second-phase sampling stratum (*performed during Main sampling only*) – Second-phase sampling uses seven distinct sampling rates in the U.S. and five in PR. These rates are applied at a block level to addresses in the U.S. and PR by calculating a measure of size for each of the following entities:

- Counties
- Places (active, functioning governmental units)
- School Districts (elementary, secondary, and unified)
- American Indian Areas
- Alaska Native Village Statistical Areas
- Hawaiian Homelands
- Minor Civil Divisions (MCDs) – in Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin (these are the states where MCDs are active, functioning governmental units)
- Census Designated Places – in Hawaii only

The measure of size for all areas except American Indian and Alaska Native Village Statistical Areas is an estimate of the number of occupied HUs in the area. This is calculated by multiplying the number of ACS addresses by the occupancy rate from Census 2000 at the block level. A measure of size for each Census Tract is also calculated in the same manner.

For American Indian and Alaska Native Village Statistical Areas, the measure of size is the estimated number of occupied HUs multiplied by the proportion of people reporting American Indian or Alaska Native (alone or in combination) in Census 2000.

Each block is then assigned the smallest measure of size from the set of all entities of which it is a part. The second-phase sampling strata are shown in Table 1 below.

- Calculation of the second-phase sampling rates (*performed during Main sampling only*) – The sampling rates given in Table 1 are calculated using the distribution of ACS valid addresses by second-phase sampling stratum in such a way as to yield an overall target sample size for the year of approximately 3,000,000 in the U.S. and 36,000 in PR. These rates also account for expected growth of the HU inventory between Main and Supplemental of roughly 1 percent.

- Second-phase sample selection (*performed in Main and Supplemental*) – After each block is assigned to a second-phase sampling stratum, a systematic sample of addresses is selected from the second-phase universe (first-phase sample) within each county, county equivalent, and municipio.

- Sample Month Assignment (*performed in Main and Supplemental*) – After the second phase of sampling, all sample addresses are randomly assigned to a sample month. Addresses selected during Main sampling are allocated to each of the 12 months. Addresses selected during Supplemental sampling are assigned to the months of April-December.

Table 1. **First-Stage Sampling Rate Categories for the United States and Puerto Rico**

Sampling Rate Category	Sampling Rates	
	United States	**Puerto Rico**
Blocks in smallest governmental units (MOS[1] < 200)	10.0%	10.0%
Blocks in smaller governmental units (200 ≤ MOS < 800)	6.8%	8.1%
Blocks in small governmental units (800 ≤ MOS ≤ 1200)	3.4%	4.0%
Blocks in large tracts (MOS >1200, TRACTMOS[2] ≥ 2000) where Mailable addresses[3] ≥ 75% and predicted levels of completed mail and CATI interviews prior to second-stage sampling > 60%	1.6%	2.0%
Other Blocks in large tracts (MOS >1200, TRACTMOS ≥ 2000)	1.7%	
All other blocks (MOS >1200, TRACTMOS < 2000) where Mailable addresses ≥ 75% and predicted levels of completed mail and CATI interviews prior to second-stage sampling > 60%	2.1%	2.7%
All other blocks (MOS >1200, TRACTMOS < 2000)	2.3%	

[1]MOS = Measure of size.
[2]TRACTMOS = Census Tract measure of size.
[3]Mailable addresses: Addresses that have sufficient information to be delivered by the U.S. Postal Service (as determined by ACS).

2. Second-Stage Sample Selection – Subsampling the Unmailable and Non-Responding Addresses

   All addresses determined to be unmailable are subsampled for the CAPI phase of data collection at a rate of 2-in-3. Unmailable addresses, which include Remote Alaska addresses, do not go to the CATI phase of data collection. Subsequent to CATI, all addresses for which no response has been obtained prior to CAPI are subsampled based on the expected rate of completed interviews at the tract level using the rates shown in Table 2.

Table 2. Second-Stage (CAPI) Subsampling Rates for the United States and Puerto Rico

Address and Tract Characteristics	CAPI Subsampling Rate
United States	
Unmailable addresses and addresses in Remote Alaska	2-in-3
Mailable addresses in tracts with predicted levels of completed mail and CATI interviews prior to CAPI subsampling between 0% and less than 36%	1-in-2
Mailable addresses in tracts with predicted levels of completed mail and CATI interviews prior to CAPI subsampling greater than 35% and less than 51%	2-in-5
Mailable addresses in other tracts	1-in-3
Puerto Rico	
Unmailable addresses	2-in-3
Mailable addresses	1-in-2

Group Quarters

The GQ sampling frame is divided into three strata: one for small GQs (having 15 or fewer people according to Census 2000 or updated information), one for GQs that were closed on Census Day 2000, and one for large GQs (having more than 15 people according to Census 2000 or updated information). GQs in the first two strata are sampled using the same procedure, and GQs in the large stratum are sampled using different a method. The small GQ stratum and the stratum for GQs closed on Census Day are combined into one sampling stratum and sorted geographically [1].

1. First Stage Sample Selection for Small Stratum

- First phase sampling - Small GQs are only eligible to be selected for the ACS once every five years. To accomplish this, the first phase sampling procedure systematically assigned all small GQs to one of five partitions of the universe. Each partition was assigned to a particular year (2006-2010) and the one assigned to 2006 became the first phase sample. In future years, each new GQ will be systematically assigned to one of the five samples. These samples are rotated over five year periods and become the universe for selecting the second phase sample.

- Second phase sampling - A simple 1-in-8 systematic sample of the GQs in the first phase sample is selected. Regardless of their actual size, all GQs in the small stratum have the

[1]Note that all references to the small GQ stratum include both small GQs and GQs closed on Census day.

same probability of selection. Since the first phase sample is 20% of the universe, this yields the targeted sampling rate of 2.5%.

2. First Stage Sample Selection for the Large Stratum

   - First Phase Sampling - Unlike housing unit address sampling and the small GQ sample selection, the large GQ sampling procedure has no first-phase in which sampling units are randomly assigned to one of five years. All large GQs are eligible for sampling each year.

   - Second Phase Sampling - In the large GQ stratum, GQ hits are selected using a one-phase systematic PPS (probability proportional to size) sample, with a target sampling rate of 2.5%. A hit refers to a grouping of 10 expected interviews. GQs are selected with probability proportional to its most current count of persons or capacity. For stratification, and for sampling the large GQs, a GQ measure of size (GQMOS) is computed, where GQMOS is the expected population of the GQ divided by 10. This reflects that the GQ data is collected in groups of 10 GQ persons. People are selected in hits of 10 in a systematic sample of 1-in-40 hits. All GQs in this stratum are eligible for sampling every year, regardless of their sample status in previous years. For large GQs, hits can be selected multiple times in the sample year. For most GQ types, the hits are randomly assigned throughout the year. Some GQs may have multiple hits with the same sample date if more than 12 hits are selected from the GQ. In these cases, the person sample within that month is unduplicated.

3. Sample Month Assignment

   In order to assign a panel month to each hit, all of the GQ samples from a state are combined and sorted by small/large stratum and second-phase order of selection. Consecutive samples are assigned to the twelve panel months in a predetermined order, starting with a randomly determined month, except for Federal prisons and remote Alaska. Remote Alaska GQs are assigned to January and September based on where the GQ is located. Correctional facilities have their sample clustered. All Federal prisons hits are assigned to the September panel. In non-Federal correctional facilities, all hits for a given GQ are assigned to the same panel month. However, unlike Federal prisons, the hits in state and local correctional facilities are assigned to randomly selected panels spread throughout the year.

4. Second Stage Sample: Selection of Persons in Small and Large GQs

   Small GQs in the second phase sampling are “take all,” i.e., every person in the selected GQ is eligible to receive a questionnaire. If the actual number of persons in the GQ exceeds 15, a field subsampling operation is performed to reduce the total number of sample persons interviewed at the GQ to 10. If the actual number of persons in the GQ is 10 or fewer, then the group size will be less than 10.

For each hit in the large GQs, the automated instrument uses the population count at the time of the visit and selects a subsample of 10 people from the roster. The people in this subsample receive the questionnaire.

ESTIMATION PROCEDURE

The estimates that appear in this product were obtained from a raking 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. 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. 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.

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 a value 40, all characteristics of that person or housing unit is tabulated with the weight of 40.

The weighting is conducted in two main operations: group quarters person weighting and a combined housing unit and household person weighting. The group quarters person weighting is conducted first with the household person weighting second. The household person weighting is dependent on the group quarters person weighting so that estimates for total population which include both group quarters and household population will be controlled to the Census Bureau's official 2006 total resident population estimates.

Group Quarters Person Weighting

Each GQ person is first assigned to an Estimates Area Major GQ Type Group (the type groups used by the Population Estimates Program). The major type groups used are:

Table 3: Estimates Area Major GQ Type Groups

Major GQ Type Group	Definition	Institutional / Non-Institutional
1	Correctional Institutions	Institutional
2	Juvenile Detention Facilities	Institutional
3	Nursing Homes	Institutional
4	Other Long-Term Care Facilities	Institutional
5	College Dormitories	Non-Institutional
6	Military Facilities	Non-Institutional
7	Other Non-Institutional Facilities	Non-Institutional

The procedure used to assign the weights to the GQ persons is performed independently within state. The steps are as follows:

- Base Weight—The initial base weight after the first stage of sampling is the inverse of its first-stage sampling rate, which is equal to 40 for all sample cases in 2006. This initial base weight is then adjusted for the second-stage sampling that occurs at the time of interview.

- Non-Interview Factor—This factor adjusted the weight of all responding GQ persons to account for both the responding and non-responding GQ persons including those non-responding persons contained in whole GQ non-respondents. The non-interview factor was computed and assigned using the following groups:

  State × Major GQ Type Group × County

- GQ Person Post-stratification Factor—This factor adjusted the GQ person weights so that the weighted sample counts matched independent population estimates by Estimates Area Major Type Group at the state level in both the U.S. and Puerto Rico. Because of collapsing of groups in applying this factor, only total GQ population is assured of agreeing with the Census Bureau's official 2006 population estimates at the state level. The GQ person post-stratification factor was computed and assigned using the following groups:

  State × Major GQ Type Group

- Rounding—The final GQ person weight 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:

  Major GQ Type Group
  Major GQ Type Group × County
  Major GQ Type Group × County × Race
  Major GQ Type Group × County × Race × Hispanic Origin
  Major GQ Type Group × County × Race × Hispanic Origin × Sex
  Major GQ Type Group × County × Race × Hispanic Origin × Sex × Age

Housing Unit and Household Person Weighting

The housing unit and household person weighting used weighting areas built from collections of whole counties. Census 2000 data were used to group counties of similar demographic and social characteristics. The characteristics considered in the formation included:

- Percent in poverty
- Percent renting
- Percent in rural areas
- Race, ethnicity, age, and sex distribution
- Distance between the centroids of the counties
- Core-based Statistical Area status

Each weighting area was also required to meet a threshold of 400 expected person interviews in the 2006 ACS. The stratification process then attempted to minimize the differences on the characteristics listed above between the counties within a weighting area. The process also tried to preserve as many counties that met the threshold to form their own weighting areas. In total, there were 2,006 weighting areas formed from the 3,219 counties and county equivalents including Puerto Rico.

The estimation procedure used to assign the weights was then performed independently within each of the ACS weighting areas.

1. Initial Housing Unit Weighting Factors—This process produced the following factors:

   - Base Weight (BW)—This initial weight was assigned to every housing unit as the inverse of its block'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 = 2.0, 2.5, or 3.0 according to Table 2
     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, CATI, and CAPI records to be tabulated in a month equal to the total base weight of all cases originally mailed for that month. For all cases, VMS was computed and assigned based on the following groups:

     Weighting Area × 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 stages. The first factor, NIF1, is a ratio adjustment that was computed and assigned to occupied housings units based on the following groups:

     Weighting Area × Building Type × Tract

     A second factor, NIF2, is a ratio adjustment that was computed and assigned to occupied housing units based on the following groups:

     Weighting Area × Building Type × Month

NIF was then computed by applying NIF1 and NIF2 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, the Mode Bias Factor.

  NIFM was computed and assigned to occupied CAPI housing units based on the following groups:

  Weighting Area × Building Type × Month

  Vacant housing units or non-CAPI (mail and CATI) housing units 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:

  Weighting Area × Tenure (Owner or renter) × Month × Marital Status of the Householder (married/widowed or single)

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

- Housing unit Post-stratification Factor (HPF)—This factor made the total weight of all housing units agree with the 2006 independent housing unit estimates at the weighting area level.

  These independent housing unit estimates exist only for the U.S. and not Puerto Rico. Thus, all housing units in Puerto Rico received a value of HPF = 1.0.

2. 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 × … × HPF). At this point everyone in the household has the same weight. Beginning in 2006, the person weighting is done in a series of three steps which are repeated until a stopping criterion is met. These person weights were individually adjusted based for each person as described below.

   The three steps are as follows:

   - Spouse Equalization Raking Factor (SPEQRF)—This factor was applied to individuals based on their status of being in a married-couple or unmarried-partner household. All persons were assigned to one of three groups:

Householder in a married-couple or unmarried-partner household
Spouse or unmarried partner in a married-couple or unmarried-partner household
All others

The first two groups are adjusted so that the sum of their person weights is equal to the total estimate of married-couple or unmarried-partner households using the housing unit weight (BW × … × HPF). The goal of this step is to produce more consistent estimates of spouses or unmarried partners and married-couple and unmarried-partner households.

- Householder Equalization Raking Factor (HHEQRF)— This factor was applied to individuals based on their householder/non-householder status. All persons were assigned to one of two groups:

  Householders
  Non-householders

  The first group is adjusted so that the sum of their person weights is equal to the total estimate of occupied housing units using the housing unit weight (BW × … × HPF). The goal of this step is to produce more consistent estimates of householders, occupied housing units, and households.

- Demographic Raking Factor (DEMORF)—This factor was applied to individuals based on their age, race, sex and Hispanic origin in the U.S. and based on their age and sex in Puerto Rico. It adjusted the person weights so that the weighted sample counts matched independent population estimates by age, race, sex, and Hispanic origin at the weighting area level in the U.S. and matched the independent population estimates by age and sex in Puerto Rico at the weighting area level. Because of collapsing of groups in applying this factor, only total population is assured of agreeing with the official 2006 intercensal population estimates at the weighting area level.

  For U.S., this used the following groups:

  Weighting Area × Race / Ethnicity (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)) × Sex × Age Groups.

  In Puerto Rico, this used only the Sex × Age Groups.

  These three steps were repeated several times until the estimates at the national level achieved their optimal consistency with regard to the spouse and householder equalization. The effect Person Post-Stratification Factor (PPSF) is

then equal to the product (SPEQRF × HHEQRF × DEMORF) from all of iterations of these three adjustments. The unrounded person weight is then the equal to the product of PPSF times the housing unit weight (BW × … × HPF × PPSF).

- Rounding—The final product of all person weights (BW × … × HPF × 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 × Race
  County × Race × Hispanic Origin
  County × Race × Hispanic Origin × Sex
  County × Race × Hispanic Origin × Sex × Age
  County × Race × Hispanic Origin × Sex × Age × Tract
  County × Race × Hispanic Origin × Sex × Age × Tract × 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.

3. Final Housing Unit Weighting Factors—This process produced the following factors:

   - Householder Factor (HHF)—This factor adjusted for differential response depending on the race, Hispanic origin, sex, and age of the householder. The value of HHF for an occupied housing unit was the PPSF of the householder. Since there is no householder for vacant units, the value of HHF = 1.0 for all vacant units.

   - Rounding—The final product of all housing unit weights (BW × … × HHF) was rounded to an integer. For occupied units, the rounded housing unit weight is the same as the rounded person weight of the householder. This ensures that both the rounded and unrounded householder weight is equal to the occupied housing unit weight. The rounding for vacant housing units was then 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 × Tract
     County × Tract × Block

CONFIDENTIALITY OF THE DATA

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 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.

- 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 Avoidance: Disclosure avoidance is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual that has provided information under a pledge of confidentiality. For data tabulations, the Census Bureau uses disclosure avoidance 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, 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.

- Data Swapping: Data swapping is a method of disclosure avoidance 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 were used again in Census 2000.

- Synthetic Data: The goals of using synthetic data are the same as the goals of data swapping, namely to protect the confidentiality in tables of frequency data. Persons are identified as being at risk for disclosure based on certain characteristics. The synthetic

data technique then models the values for another collection of characteristics to protect the confidentiality of that individual.

ERRORS IN THE DATA

- Sampling Error — The data in the ACS 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 ensure 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 data entry from questionnaires and editing 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 ACS 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.

MEASURES OF SAMPLING ERROR

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 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 – in the form of margins of error – are provided with all published ACS 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.

Confidence Intervals and Margins of Error

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 ACS 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 1.645 times the estimated standard error below the estimate to 1.645 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.

Margin of Error – Instead of providing the upper and lower confidence bounds in published ACS tables, the margin of error is provided instead. The margin of error is the difference between an estimate and its upper or lower confidence bound. Both the confidence bounds and the standard error can easily be computed from the margin of error. All ACS published margins of error are based on a 90 percent confidence level.

Standard Error = Margin of Error / 1.645

Lower Confidence Bound = Estimate - Margin of Error

Upper Confidence Bound = Estimate + Margin of Error

Note that for 2006 and earlier estimates, ACS margins of error and confidence bounds were calculated using a 90 percent confidence level multiplier of 1.65. Beginning with the 2006 data release, we are now employing a more accurate multiplier of 1.645. Margins of error and confidence bounds from previously-published products will not be updated with the new multiplier. When calculating standard errors from margins of error or confidence bounds using published data for 2006 and earlier, use the 1.65 multiplier.

When constructing confidence bounds from the margin of error, the user should be aware of any "natural" limits on the bounds. For example, if a population estimate is near zero, 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 zero instead. However, for other estimates such as income, negative values do make sense. The context and meaning of the estimate must be kept in mind when creating these bounds. Another of these natural limits would be 100 percent for the upper bound of a percent estimate.

If the margin of error is displayed as '*****' (five asterisks), the estimate has been controlled to be equal to a fixed value and so it has no sampling error. When using any of the formulas in the following section, use a standard error of zero for these controlled estimates.

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

- The estimated standard errors (and thus margins of error) included in these data products do not include 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. Nor do they reflect the error from imputed values due to missing responses. 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.

- Zero or small estimates; very large estimates — The value of almost all ACS 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.

CALCULATION OF STANDARD ERRORS

Direct estimates of the standard errors were calculated for all estimates reported in this product. The standard errors, in most cases, are calculated using a replicate-based methodology that takes into account the sample design and estimation procedures. Exceptions include:

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

2. There are either no sample observations available to compute an estimate or standard error of a median, an aggregate, a proportion, or some other ratio, or there are too few sample observations to compute a stable estimate of the standard error.. The estimate is represented in the tables by "-" and the margin of error by "**" (two asterisks).

3. The estimate of a median falls in the lower open-ended interval or upper open-ended interval of a 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 margin of error is represented in the tables by "***" (three asterisks).

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 — The statistic of interest may be the ratio of two estimates. First is the case where the numerator *is not* a subset of the denominator. The standard error of this ratio between two sample estimates is approximated as:

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

Proportions/percents – For a proportion (or percent), a ratio where the numerator *is* a subset of the denominator, a slightly different estimator is used. Note the difference between the formulas for the standard error for proportions (below) and ratios (above) - the plus sign in the previous formula has been replaced with a minus sign. If the value under the square root sign is negative, use the ratio standard error formula above, instead. If $\hat{P} = \hat{X} / \hat{Y}$, then

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

If $\hat{Q} = 100\% \times \hat{P}$ (P is the proportion and Q is its corresponding percent), then $SE(\hat{Q}) = 100\% \times SE(\hat{P})$.

Products – For a product of two estimates - for example if you want to estimate a proportion's numerator by multiplying the proportion by its denominator - the standard error can be approximated as

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

Significant differences – Users may conduct a statistical test to see if the difference between an ACS estimate and any other chosen estimates is statistically significant at a given confidence level. "Statistically significant" means that the difference is not likely due to random chance alone. With the two estimates (Est_1 and Est_2) and their respective standard errors (SE_1 and SE_2), calculate

$$Z = \frac{Est_1 - Est_2}{\sqrt{(SE_1)^2 + (SE_2)^2}}$$

If Z > 1.645 or Z < -1.645, then the difference can be said to be statistically significant at the 90 percent confidence level. [Note that we are now recommending that +/-1.645 be used to determine significance. Previous ACS Accuracy of the Data documents suggested using +/-1.65.] Any estimate can be compared to an ACS estimate using this method, including other ACS estimates from the current year, the ACS estimate for the same characteristic and geographic area but from a previous year, Census 2000 100 percent counts and long form estimates, estimates from other Census Bureau surveys, and estimates from other sources. Not all estimates have sampling error — Census 2000 100 percent counts do not, for example, although Census 2000 long form estimates do — but they should be used if they exist to give the most accurate result of the test.

Users are also cautioned to *not* rely on looking at whether confidence intervals for two estimates overlap to determine statistical significance, because there are circumstances where that method will not give the correct test result. The Z calculation above is recommended in all cases.

All statistical testing in ACS data products is based on the 90 percent confidence level. Users should understand that all testing was done using *unrounded* estimates and standard errors, and it may not be possible to replicate test results using the rounded estimates and margins of error as published.

EXAMPLES OF 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 39,401,560 from summary table B12001 for the United States for 2006. The margin of error is 99,234.

Standard Error = Margin of Error / 1.645

Calculating the standard error using the margin of error, we have:

SE(39,401,560) = 99,234 / 1.645 = 60,325.

Example 2 - Calculating the Standard Error of a Sum

We are interested in the number of people who have never been married. From Example 1, we know the number of males, never married is 39,401,560. From summary table B12001 we have the number of females, never married is 33,385,649 with a margin of error of 77,920. So, the estimated number of people who have never been married is 39,401,560 + 33,385,649 = 72,787,209. 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 60,325. The standard error for the number of females never married is calculated using the margin of error:

SE(33,385,649) = 77,920 / 1.645 = 47,368.

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

$$SE(72,787,209) = \sqrt{60,325^2 + 47,368^2} = 76,699$$

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 72,787,209 using the standard error, simply multiply 76,699 by 1.645, then add and subtract the product from 72,787,209. Thus the 90 percent confidence interval for this estimate is [72,787,209 - 1.645(76,699)] to [72,787,209 + 1.645(76,699)] or 72,661,039 to 72,913,379.

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 33,385,649 and the number of people who have never been married is 72,787,209. 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 47,368 and the standard error for the number of people never married calculated from example 2 as 76,699.

The estimate is (33,385,649 / 72,787,209) * 100% = 45.9%

So, using the formula for the standard error of a proportion or percent, we have:

$$\text{SE}(45.9\%) = 100\% * \left(\frac{1}{72{,}787{,}209} \sqrt{47{,}368^2 - 0.459^2 \times 76{,}699^2} \right) = 0.04\%$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 45.9 using the standard error, simply multiply 0.05 by 1.645, then add and subtract the product from 45.9. Thus the 90 percent confidence interval for this estimate is [45.9 - 1.645(0.04)] to [45.9 + 1.645(0.04)], or 45.8% to 46.0%.

Example 4 - Calculating the Standard Error of a Ratio

Now, let us calculate the estimate of the ratio of the number of unmarried males to the number of unmarried females and its standard error. From the above examples, the estimate for the number of unmarried men is 39,401,560 with a standard error of 60,325, and the estimates for the number of unmarried women is 33,385,649 with a standard error of 47,368.

The estimate of the ratio is 39,401,560 / 33,385,649 = 1.180.

The standard error of this ratio is

$$\text{SE}(1.18) = \left(\frac{1}{33{,}385{,}649} \sqrt{60{,}325^2 + 1.180^2 \times 47{,}368^2} \right) = 0.00246$$

The 90 percent margin of error for this estimate would be 0.00246 multiplied by 1.645, or about 0.004. The 90 percent lower and upper 90 percent confidence bounds would then be [1.180 – 0.004] to [1.180 + 0.004], or 1.176 and 1.184.

Example 5 - Calculating the Standard Error of a Product

We are interested in the number of 1-unit detached owner-occupied housing units. The number of owner-occupied housing units is 75,086,485 with a margin of error of 218,471 from subject table S2504 for 2006, and the percent of 1-unit detached owner-occupied housing units is 81.4% (0.814) with a margin of error of 0.1 (0.001). So the number of 1-unit detached owner-occupied housing units is 75,086,485 * 0.814 = 61,120,399. Calculating the standard error for the estimates using the margin of error we have:

SE(75,086,485) = 218,471 / 1.645 = 132,809

and

SE(0.814) = 0.001 / 1.645 = 0.0006079

The standard error for number of 1-unit detached owner-occupied housing units is calculated using the formula for products as:

$$SE(61,120,399) = \sqrt{75,086,485^2 \times 0.0006079^2 + 0.814^2 \times 132,809^2} = 117,348$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 61,118,243 using the standard error, simply multiply 117,348 by 1.645, then add and subtract the product from 61,120,399. Thus the 90 percent confidence interval for this estimate is [61,120,399 - 1.645(117,348)] to [61,120,399 + 1.645(117,348)] or 60,927,362 to 61,313,436.

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 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 ACS an address list in each state, is as complete and accurate as possible.

The source of addresses was the MAF. The MAF is created by combining the Delivery Sequence File of the United States Postal Service and the address list 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 (Topologically Integrated Geographic Encoding and Referencing) 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 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. If the interviewer could not locate the exact sample unit in a multi-unit structure, or found a different number of units than expected, the interviewers were instructed to list the units in the building and follow a specific procedure to select a replacement sample unit.

- Overcoverage — It is possible for some sample housing units or persons to be sampled multiple times in the survey. The overcoverage of persons and housing units can introduce biases into the data, and increase respondent burden and survey costs. As with undercoverage, a major way to avoid housing unit overcoverage in a survey is to ensure that its sampling frame, for ACS an address list in each state, is as complete and accurate as possible.

  The source of addresses was the MAF. The MAF is created by combining the Delivery Sequence File of the United States Postal Service and the address list for Census 2000. Sometimes the MAF has an address that is the duplicate of another address already on the MAF. This could occur when there is a slight difference in the address such as 123 Main Street versus 123 Maine Street.

  Person overcoverage can occur when an individual is included as a member of a housing unit but does not meet ACS residency rules.

- Nonresponse Error — Survey nonresponse is a well-known source of nonsampling error. There are two types of nonresponse error – unit nonresponse and item nonresponse. Nonresponse errors affect survey estimates to varying levels depending on amount of nonresponse and the extent to which nonrespondents differ from respondents on the characteristics measured by the survey. The exact amount of nonresponse error or bias on an estimate is almost never known. Therefore, survey researchers generally rely on proxy measures, such as the nonresponse rate, to indicate the potential for nonresponse error.

- Unit Nonresponse — Unit nonresponse is the failure to obtain data from housing units in the sample. Unit nonresponse may occur because households are unwilling or unable to participate, or because an interviewer is unable to make contact with a housing unit. Unit nonresponse is problematic when there are systematic or variable differences between interviewed and noninterviewed housing units on the characteristics measured by the survey. Nonresponse bias is introduced into an estimate when differences are systematic, while nonresponse error for an estimate evolves from variable differences between interviewed and noninterviewed households.

  The ACS made every effort to minimize unit nonresponse, and thus, the potential for nonresponse error. First, the ACS used a combination of mail, CATI, and CAPI data collection modes to maximize response. The mail phase included a series of three to four mailings to encourage housing units to return the questionnaire. Subsequently, mail nonrespondents (for which phone numbers are available) were contacted by CATI for an interview. Finally, a subsample of the mail and telephone nonrespondents was contacted for by personal visit to attempt an interview. Combined, these three efforts resulted in a very high overall response rate for the ACS.

- Item Nonresponse — Nonresponse to particular questions on the survey questionnaire and instrument allows for the introduction of error or 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 errors or biases is afforded by minimizing nonresponse. In the ACS, item nonresponse for the CATI and CAPI operations was minimized by the requirement that the automated instrument receive a response to each question before the next one could be asked. 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.

- Measurement and Processing Error — The person completing the questionnaire or responding to the questions posed by an interviewer could serve as a source of error, although the questions were cognitively tested for phrasing, and detailed instructions for completing the questionnaire were provided to each household.

  - 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.

- 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, follow-up by telephone, manual coding of write-in responses, and automated data processing. The various field, coding and computer operations undergo a number of quality control checks to insure their accurate application.

- Content Editing — After data collection was completed, any remaining incomplete or inconsistent information was imputed during the final content 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 missing 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 allocate an entry for a person or housing unit that was consistent with entries for persons or housing units with similar characteristics. Imputing acceptable values in place of blanks or unacceptable entries enhances the usefulness of the data.