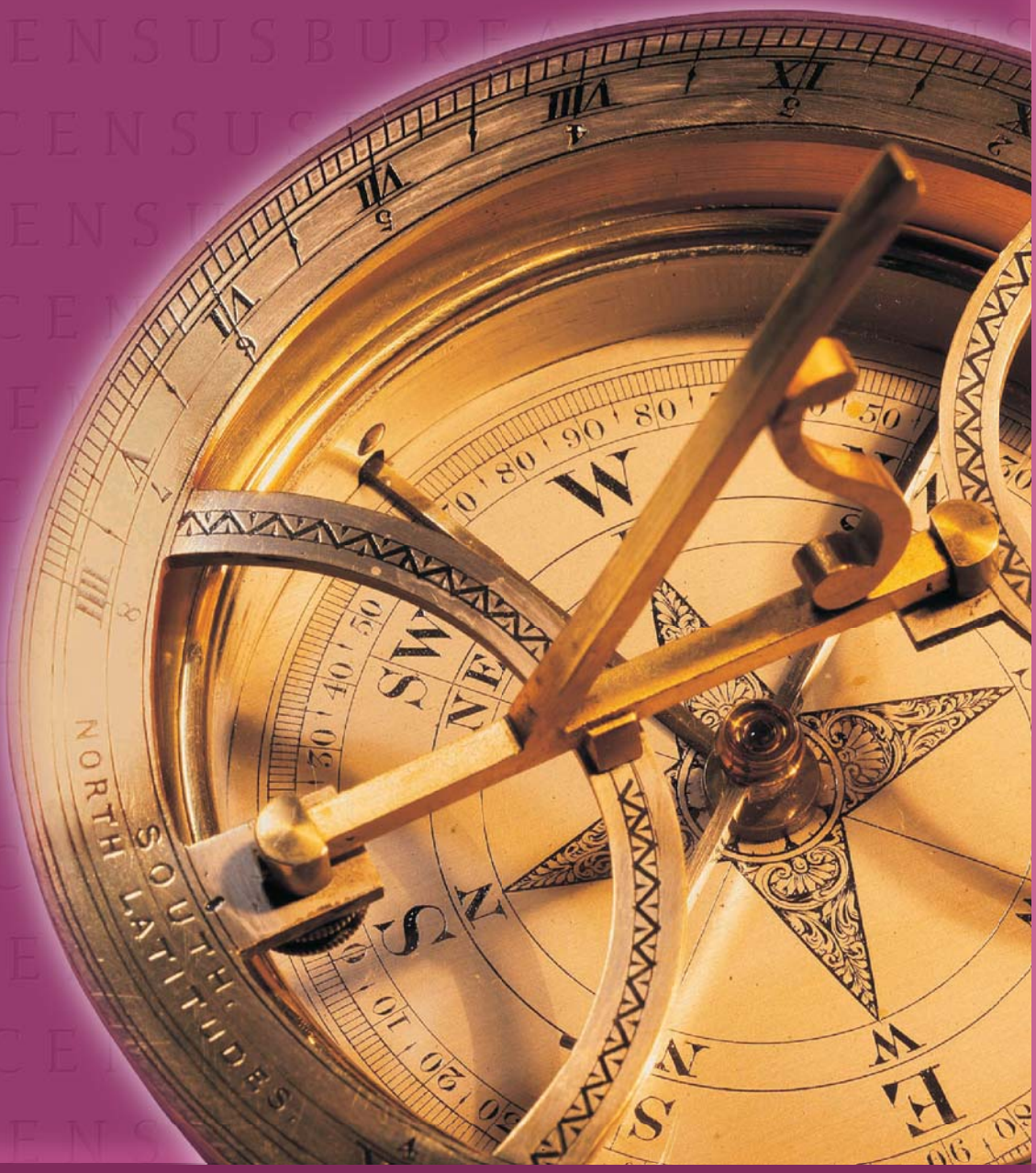
A Compass for Understanding and Using American Community Survey Data

What PUMS Data Users Need to Know

Issued
February 2009



USCENSUSBUREAU

Helping You Make Informed Decisions

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

United States Census 2010

Acknowledgments

Leonard P. Gaines, Consultant, drafted this handbook for the U.S. Census Bureau's American Community Survey Office. **Kennon R. Copeland** and **John H. Thompson** of National Opinion Research Center at the University of Chicago drafted the technical appendixes. **Edward J. Spar**, Executive Director, Council of Professional Associations on Federal Statistics, **Frederick J. Cavanaugh**, Executive Business Director, Sabre Systems, Inc., **Susan P. Love**, Consultant, **Linda A. Jacobsen**, Vice President, Domestic Programs, Population Reference Bureau, and **Mark Mather**, Associate Vice President, Domestic Programs, Population Reference Bureau, provided initial review of this handbook.

Deborah H. Griffin, Special Assistant to the Chief of the American Community Survey Office, provided the concept and directed the development and release of a series of handbooks entitled *A Compass for Understanding and Using American Community Survey Data.* **Cheryl V. Chambers**, **Colleen D. Flannery**, **Cynthia Davis Hollingsworth**, **Susan L. Hostetter**, **Pamela M. Klein**, **Anna M. Owens**, **Clive R. Richmond**, **Enid Santana**, and **Nancy K. Torrieri** contributed to the planning and review of this handbook series.

The American Community Survey program is under the direction of **Arnold A. Jackson**, Associate Director for Decennial Census, **Daniel H. Weinberg**, Assistant Director for the American Community Survey and Decennial Census, and **Susan Schechter**, Chief, American Community Survey Office.

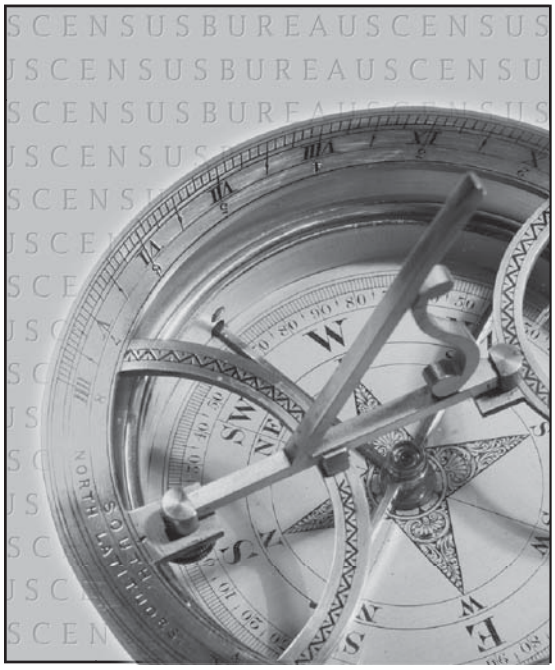
Other individuals who contributed to the review and release of these handbooks include **Dee Alexander**, **Herman Alvarado**, **Mark Asiala**, **Frank Ambrose**, **Maryam Asi**, **Arthur Bakis**, **Genora Barber**, **Michael Beaghen**, **Judy Belton**, **Lisa Blumerman**, **Scott Boggess**, **Ellen Jean Bradley**, **Stephen Buckner**, **Whittona Burrell**, **Edward Castro**, **Gary Chappell**, **Michael Cook**, **Russ Davis**, **Carrie Dennis**, **Jason Devine**, **Joanne Dickinson**, **Barbara Downs**, **Maurice Eleby**, **Sirius Fuller**, **Dale Garrett**, **Yvonne Gist**, **Marjorie Hanson**, **Greg Harper**, **William Hazard**, **Steve Hefter**, **Douglas Hillmer**, **Frank Hobbs**, **Todd Hughes**, **Trina Jenkins**, **Nicholas Jones**, **Anika Juhn**, **Donald Keathley**, **Wayne Kei**, **Karen King**, **Debra Klein**, **Vince Kountz**, **Ashley Landreth**, **Steve Laue**, **Van Lawrence**, **Michelle Lowe**, **Maria Malagon**, **Hector Maldonado**, **Ken Meyer**, **Louisa Miller**, **Stanley Moore**, **Alfredo Navarro**, **Timothy Olson**, **Dorothy Paugh**, **Marie Pees,** **Marc Perry**, **Greg Pewett**, **Roberto Ramirez**, **Dameka Reese**, **Katherine Reeves**, **Lil Paul Reyes**, **Patrick Rottas**, **Merarys Rios**, **J. Gregory Robinson**, **Anne Ross**, **Marilyn Sanders**, **Nicole Scanniello**, **David Sheppard**, **Joanna Stancil**, **Michael Starsinic**, **Lynette Swopes**, **Anthony Tersine**, **Carrie Werner**, **Edward Welniak**, **Andre Williams**, **Steven Wilson**, **Kai Wu**, and **Matthew Zimolzak**.

Linda Chen and **Amanda Perry** of the Administrative and Customer Services Division, **Francis Grailand Hall**, Chief, provided publications management, graphics design and composition, and editorial review for the print and electronic media. **Claudette E. Bennett**, Assistant Division Chief, and **Wanda Cevis**, Chief, Publications Services Branch, provided general direction and production management.

A Compass for Understanding and Using American Community Survey Data

Issued February 2009

What PUMS Data Users Need to Know



U.S. Department of Commerce
Vacant,
Secretary

Vacant,
Deputy Secretary

Economics and Statistics Administration
Kim White,
Acting Under Secretary for Economic Affairs

U.S. CENSUS BUREAU
Thomas L. Mesenbourg,
Acting Director

Suggested Citation

U.S. Census Bureau,
A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know
U.S. Government Printing Office, Washington, DC, 2009.

Economics and Statistics Administration

Kim White,
Acting Under Secretary for Economic Affairs

U.S. CENSUS BUREAU

Thomas L. Mesenbourg,
Acting Director

Thomas L. Mesenbourg,
Deputy Director and
Chief Operating Officer

Arnold A. Jackson
Associate Director for Decennial Census

Daniel H. Weinberg
Assistant Director for ACS and Decennial Census

Susan Schechter
Chief, American Community Survey Office

Contents

Foreword

The American Community Survey (ACS) is a nationwide survey designed to provide communities with reliable and timely demographic, social, economic, and housing data every year. The U.S. Census Bureau will release data from the ACS in the form of both single-year and multiyear estimates. These estimates represent concepts that are fundamentally different from those associated with sample data from the decennial census long form. In recognition of the need to provide guidance on these new concepts and the challenges they bring to users of ACS data, the Census Bureau has developed a set of educational handbooks as part of *The ACS Compass Products*.

We recognize that users of ACS data have varied backgrounds, educations, and experiences. They need different kinds of explanations and guidance to understand ACS data products. To address this diversity, the Census Bureau worked closely with a group of experts to develop a series of handbooks, each of which is designed to instruct and provide guidance to a particular audience. The audiences that we chose are not expected to cover every type of data user, but they cover major stakeholder groups familiar to the Census Bureau.

General data users	Congress
High school teachers	Puerto Rico Community Survey data users (in Spanish)
Business community	Public Use Microdata Sample (PUMS) data users
Researchers	Users of data for rural areas
Federal agencies	State and local governments
Media	Users of data for American Indians and Alaska Natives

The handbooks differ intentionally from each other in language and style. Some information, including a set of technical appendixes, is common to all of them. However, there are notable differences from one handbook to the next in the style of the presentation, as well as in some of the topics that are included. We hope that these differences allow each handbook to speak more directly to its target audience. The Census Bureau developed additional *ACS Compass Products* materials to complement these handbooks. These materials, like the handbooks, are posted on the Census Bureau's ACS Web site: <www.census.gov/acs/www>.

These handbooks are not expected to cover all aspects of the ACS or to provide direction on every issue. They do represent a starting point for an educational process in which we hope you will participate. We encourage you to review these handbooks and to suggest ways that they can be improved. The Census Bureau is committed to updating these handbooks to address emerging user interests as well as concerns and questions that will arise.

A compass can be an important tool for finding one's way. We hope *The ACS Compass Products* give direction and guidance to you in using ACS data and that you, in turn, will serve as a scout or pathfinder in leading others to share what you have learned.

Background

The American Community Survey (ACS) is the new source for the information previously collected through the decennial census long form. This information includes topics such as income, employment status, housing costs, and housing conditions. Unlike the decennial census, ACS data are collected on a continuous basis. This presents a number of challenges and benefits for data users.

This handbook is primarily intended for users of the ACS who are looking for more information than is available in the profiles and tables produced by the Census Bureau. In this handbook, you will learn how the Public Use Microdata Sample (PUMS) files differ from the pretabulated products, how to access the data, and some ways to produce your own tables. Data users already familiar with the PUMS files available from the decennial censuses can learn how those files differ from the ACS PUMS.

While a common use of PUMS data is to develop statistical models describing the relationship between variables, that use of the data is beyond the scope of this handbook. Researchers developing these kinds of models should nonetheless find the information in this handbook helpful if they are not familiar with the ACS PUMS files.

A glossary and a series of technical appendixes are included at the back of this handbook for those interested in more advanced ACS issues.

What Is the ACS and Why Is It Important?

As in the past, the 2010 Census of Population and Housing will collect data about the number of people residing in the United States and their relationship within a household, age, race, Hispanic origin (ethnicity), and sex. It will also collect information about the number, occupancy status, and tenure (ownership status) of the nation's housing units. However, unlike previous censuses, information about topics such as income, education, employment status, disability status, housing value, housing costs, and number of bedrooms will not be asked as part of the 2010 Census. Instead, these data on these topics will come from the ACS. In this way, the ACS can be considered the replacement for the decennial census long form.

While the ACS takes the place of the long form as the source for similar information, it is not the same thing. Instead of collecting data from about 1 in every 6 households once every 10 years, like the decennial census long form, the ACS samples about 1 in every 40 addresses every year, or 250,000 addresses every month. This allows the Census Bureau to produce data every year rather than every decade. For areas with large populations (65,000 or more), survey estimates are based on 12 months (1 year) of ACS data. For all areas with populations of 20,000 or more, the survey estimates are based on 36 months (3 years) of ACS data. The Census Bureau will produce estimates for all areas, down to the census tract and block group levels, based on 60 months (5 years) of ACS data. How these estimates are produced is detailed in the *ACS Design and Methodology* report (Technical Paper 67) on the Census Bureau's Web site at <http://www.census.gov/acs/www/Downloads/tp67.pdf>. Information on the basic set of ACS data products and guidance on how to interpret ACS data are provided in other handbooks in the *ACS Compass Products.* The appendixes at the back of this report also provide important information about the use and interpretation of multiyear estimates.

What Are the Public Use Microdata Sample (PUMS) Files?

The Census Bureau produces a large number of data profiles, tables, and maps showing a massive amount of pretabulated data from the ACS. However, these products cannot meet the needs of every data user. The Census Bureau produces the Public Use Microdata Sample (PUMS) files so that data users can create custom tables that are not available through pretabulated ACS products.

The PUMS files are a set of untabulated records about individual people or housing units. They differ from the ACS summary products, which show data that have already been tabulated for specific geographic areas. The difference between these kinds of products can be seen in Table 1. Summary products display summary statistics such as estimates of the number of males and females; the median age of the population; and estimates of the number of occupied housing units by tenure. These estimates are specific to a geographic area. (In Table 1, for example, State 1 and County A.) PUMS files, in contrast, include population and housing unit records with individual response information such as relationship, sex, educational attainment, and employment status.

The Census Bureau plans to produce 1-year, 3-year, and 5-year ACS PUMS files. The 3-year and 5-year PUMS files are multiyear combinations of the 1-year PUMS files with appropriate adjustments to the weights and inflation adjustment factors described later in this handbook.

Table 1. **Conceptual Comparison of ACS Summary Products and Public Use Microdata Samples**

Example Summary Product						
Geography	**Males**	**Females**	**Median age**	**Occupied housing units**	**Owner-occupied units**	**Renter-occupied units**
State 1	7,345,968	7,952,709	35.9	5,689,354	3,005,973	2,683,381
County A	45,678	49,852	33.5	40,678	15,961	24,717

Example Public Use Microdata Sample Population Records						
Household ID	**Person ID**	**PUMA**	**Relationship**	**Sex**	**Educational attainment**	**Employment status**
105	1	00100	Householder	Female	Bachelors	Working
105	2	00100	Spouse	Male	Masters	Working
105	3	00100	Child	Male	Some high school	N/A

Example Public Use Microdata Sample Housing Unit Household Records						
Household ID	**PUMA**	**Tenure**	**Rooms**	**Bedrooms**	**Value**	**Contract rent**
105	00100	Owned	8	3	236,500	N/A
106	00100	Rented	3	1	N/A	1,250

Note: In the actual PUMS files, many variables, such as tenure and relationship, are represented by numeric codes rather than descriptive text.
Source: U.S. Census Bureau, artificial data.

Confidentiality of the ACS PUMS Data

As required by federal law, the confidentiality of the ACS respondents is protected through a variety of means, ensuring that it is impossible to identify individuals who provide any response. The first means of protecting confidentiality is the removal of all personal identification, such as name and address, from the record. Next, a small number of records are switched with similar records from a neighboring area, reducing the ability to identify individuals from their responses. Then, the answers to open-ended questions, such as age, income, or housing unit value—where an extreme value might identify an individual—are top-coded. Top coding is the process of taking any response exceeding a particular value and replacing it with a predetermined value. These predetermined values vary by state. For example, if someone in New York reports their age as 103, it will be recorded in the ACS PUMS file as 94 (the maximum value shown for New York).

In addition to modifying the individual records, respondent confidentiality is protected in the PUMS because only a sample of ACS responses is included in the PUMS. The 1-year tabulated ACS products found on the American FactFinder are based on all of the ACS data collected for that year, about 2.5 percent of the population. The 1-year ACS PUMS files contain a sample of the ACS housing unit and group quarters population records representing about 1 percent of the population. So, in New York State's 2006 ACS PUMS file, there are 187,143 population records or 0.969 percent of the estimated 19,306,183 people residing in the state. There are also a total of 85,108 records on the PUMS housing unit file for the state of New York (79,075 housing unit records and 6,033 group quarters person placeholders).

The Census Bureau also protects confidentiality by limiting the geographic area codes available on the PUMS files. The only geographic codes available in the PUMS records are those for regions, divisions, states, and Public Use Microdata Areas or "PUMAs."[1] PUMAs, which are described in more detail in a later section, were defined for Census 2000 to represent geographic areas with populations of at least 100,000.

[1] Regions and divisions are collections of states.

Who Should Use the PUMS and Why?

The PUMS files should be used by people who are looking for data tables that are not presented by the Census Bureau in the pretabulated products available through American FactFinder.[2] These files can be used to extract custom data for particular population groups (e.g., veterans, college students) or when it is not possible to get particular data categories from the standard tables (e.g., families with income between 90 and 99 percent of the official poverty threshold). While it is possible to request that the Census Bureau produce custom tabulations for a fee, the PUMS files provide a much less expensive—and often faster—way to get the data.

One common group of users of the PUMS files are academic researchers interested in modeling relationships between the variables collected as part of the ACS. Another common group of users are researchers working in government and business looking at either characteristics that are not usually cross-tabulated against each other or are categorized in different ways than is done in the standard tables.

While the standard ACS products answer the majority of questions data users are interested in, some questions cannot be answered by these products. For example, the standard products do not provide a table showing the poverty status of foreign-born residents by education. This can be produced using the PUMS files.

The PUMS files also can be used when the standard tables do not provide the categories that a data user is interested in seeing. For example, many of the standard tables use age breaks like 55 to 64 years and 65 years and older. But in New York State, many of the programs administered by the Office for the Aging are designed for the population aged 60 and older. So if the Office for the Aging wants to study the impact of changing any of their programs, they would need to look to the PUMS files as a primary source of information.

PUMS Geography

As noted earlier, to ensure the confidentiality of ACS respondents, PUMS files present data for a much more limited set of geographic areas than the pretabulated ACS products. PUMS files cannot be used to summarize data for individual counties, cities, or other small areas. It is possible to summarize data for the nation, each of the states, the District of Columbia, Puerto Rico, and areas known as Public Use Microdata Areas (PUMAs).

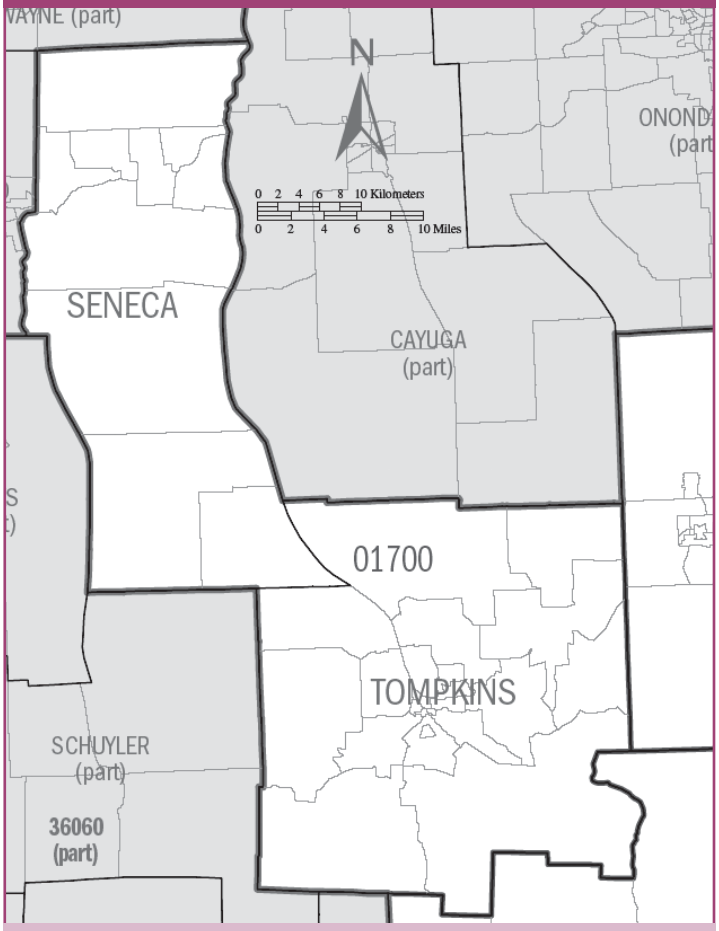
As part of Census 2000, PUMAs were defined as areas with 100,000 residents or more based on the populations reported in Census 2000. The ACS uses these same PUMAs. In most states, PUMA boundaries were defined by the State Data Center. If the State Data Center chose not to define these areas, the Census Bureau's regional office geographic staff defined them.

In addition to having a minimum population of 100,000 residents, the PUMAs had to be combinations of contiguous counties or census tracts. While attempts were made to create PUMAs that represented entire communities on their own, this was not always possible. For example, the city of Albany, New York, had a population of 95,658—not quite large enough to be its own PUMA. So, in order to get the population over 100,000 and create a PUMA that essentially represents the city, Albany was combined with one census tract from the adjacent town with similar characteristics.

As much as possible, PUMAs were designed to contain areas with similar characteristics. However, this was not always possible, so users need to consider the potential impacts of these tract combinations on the overall PUMA populations. Figure 1 shows one PUMA (01700), that comprises two counties in New York State: Seneca County and Tompkins County. In one regard these are very different counties. As shown in Table 2, Tompkins County's college and graduate school population accounts for about 31 percent of the population, while in Seneca County, this group represents about 5 percent of the total population. Yet outside the urban center of Tompkins County, these two counties are very similar. In a situation like this, the data user needs to consider the impact of such a large difference in the composition of the PUMA.

[2] You can determine if the data you are interested in are found in a standard ACS table by going to American FactFinder and searching the tables by subject or keyword. A detailed list of American FactFinder table shells for the 2007 ACS is also available at <http://www.census.gov/acs/www/Products/users_guide/2007/index.htm>.

Figure 1. **Sample PUMA Map**



Source: U.S. Census Bureau, accessed at <http://ftp2.census.gov/geo/maps/puma/puma2k/ny_puma5.pdf>.

Table 2. **Selected Characteristics of New York State PUMA 01700 and Component Counties**

	Seneca County	Tompkins County	PUMA 01700
Population	34,279	100,590	134,869
College or graduate school enrollees	1,822	31,326	33,148
Percent of population enrolled in college or graduate school	5.3	31.1	24.6

Source: U.S. Census Bureau, 2005–2007 ACS 3-year Estimates, Social Data Profile.

Identifying PUMAs

PUMAs are identified by a 5-digit number that is unique within state. Generally, the 5-digit PUMA codes are not useful in identifying where in a state a PUMA is located. To show where a particular PUMA is located, the Census Bureau has provided both maps and geographic equivalency files. The geographic equivalency files show which counties, places, and census tracts are included in each PUMA.

Finding the PUMA maps on the Census Bureau's Web site is fairly easy. As shown in Figure 2, from the Census Bureau's home page <www.census.gov>, click on "Maps" under the "Geography" section.

Figure 2. **Census Bureau Home Page**

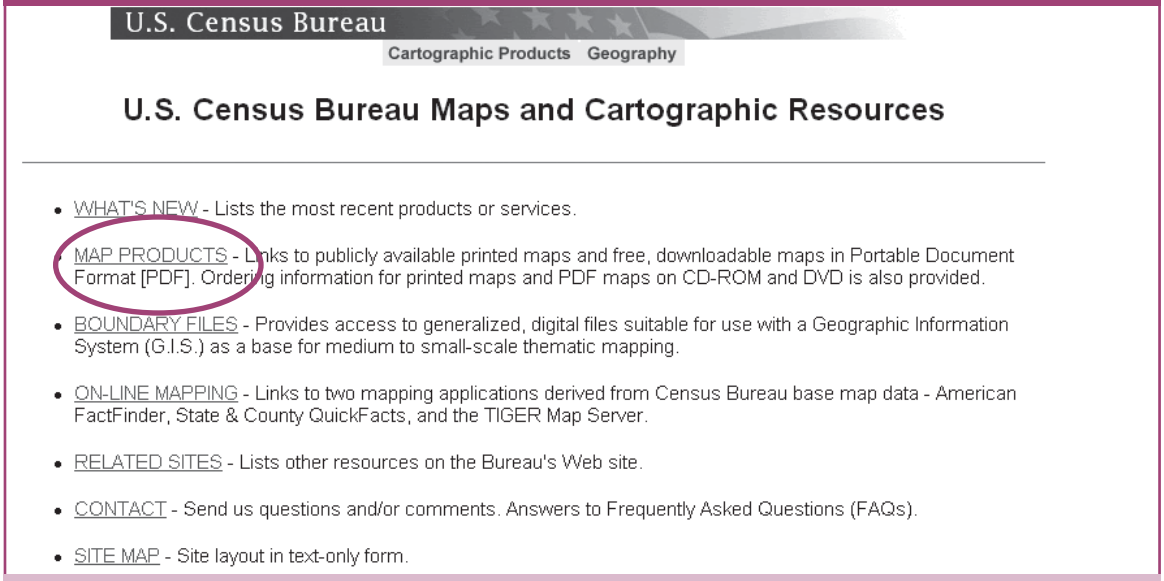
Source: U.S. Census Bureau, accessed at <www.census.gov>.

This will bring you to the geography page, shown in Figure 3. Clicking on "MAP PRODUCTS" will take you to the Census Bureau's Map Products page, Figure 4, which includes links to a variety of map products, including the Census 2000 PUMA maps that show the current set of PUMAs being used for the ACS PUMS files. The Census 2000 PUMAs will be used until the new PUMAs are delineated using 2010 Census counts. You will have to scroll down the page to see these files. Clicking on the 2000 link under the 5 percent sample will take you to a list of the states.[3] Simply choose the state you are interested in seeing and you will get a PDF file with a map showing the Super-PUMAs.[4] Individual maps for each Super-PUMA show the individual PUMAs.

If you would prefer to look at a list of the components of PUMAs, you need to look at the geographic equivalency files online. A separate geographic equivalency file exists for each state. They are found in the main PUMS link from the Census Bureau Web site: <http://www.census.gov/main/www/pums.html>. For instance, using the New York equivalency file and sorting the data by summary level code, one can see which census tracts are grouped together in a PUMA or which PUMAs compose New York County [Manhattan], (03801-03810). Geographic Equivalency Files can be accessed at the following FTP site: <http://www2.census.gov/census_2000/datasets/PUMS/FivePercent/>. The Missouri State Data Center created a tool that allows PUMA users to enter the geography that they are interested in to identify PUMA codes and equivalent geographies. For more information, go to <http://mcdc2.missouri.edu/websas/geocorr2k.html>.

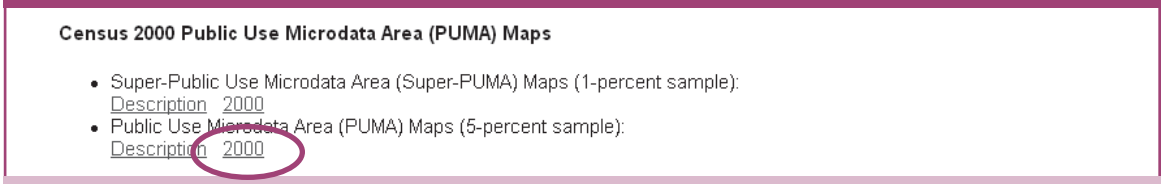
[3] The Census Bureau produced both 1percent and 5 percent PUMS files for Census 2000. The ACS PUMS uses the same geography as the Census 2000, 5 percent PUMS files even though it contains a 1 percent sample.
[4] A Super-PUMA is a collection of PUMAs with a minimum combined population of 400,000. These were used as the geographic areas for the Census 2000, 1 percent PUMS files. The ACS does not use Super-PUMAs.

Figure 3. **Geography Page**



Source: U.S. Census Bureau, accessed at <www.census.gov/geo/www/maps/>.

Figure 4. **Census 2000 PUMA Map Products Page**



Source: U.S. Census Bureau, accessed at <www.census.gov/geo/www/maps/CP_MapProducts.htm>.

Creating PUMS Tabulations

Creating tables from the PUMS files is easy if you have the right software and some basic knowledge to use the files correctly. As a rule, these files are too large to work with in a spreadsheet, and databases are not well suited to cross-tabulating data. Fortunately, a number of other applications do this very well. Many of these are general statistical programs, such as SPSS, SAS, S-Plus, and R Statistical Software. The Census Bureau also provides the ability to produce cross-tabulations using the DataFerrett system. Each of these programs has its own advantages and disadvantages, and users can choose the one that best meets their needs.

This section describes how to access the ACS PUMS files for use in any of the general statistical programs. It also shows how to use the DataFerrett program. For some PUMS applications it is necessary to apply inflation adjustments. For the 1-year PUMS the inflation adjustment variable (ADJUST) should be used to produce income characteristics. The 3-year and 5-year PUMS will carry two inflation adjustment variables—ADJINC for income applications and ADJHOUS for housing cost applications. For additional information about dollar-denominated data in the ACS, refer to Appendix 5.

Accessing PUMS Files

The Census Bureau has made it fairly easy to access the PUMS data files. These files are available in two basic formats, as ASCII text files with comma-separated values (CSV) and in two versions of SAS data sets (PC-SAS files and UNIX files). Most statistical programs can read files in at least one of these formats.

The easiest way to access the PUMS files is through the Census Bureau's American FactFinder system. One way to get to American FactFinder is to click on the "**American FactFinder**" button on the left side of the Census Bureau's home page <www.census.gov> as shown in Figure 5. This will take you to the American FactFinder home page, shown in Figure 6. To get to the ACS PUMS data, click on "**Data Sets**" and then "**American Community Survey**." This will bring you to the ACS data sets page shown in Figure 7.

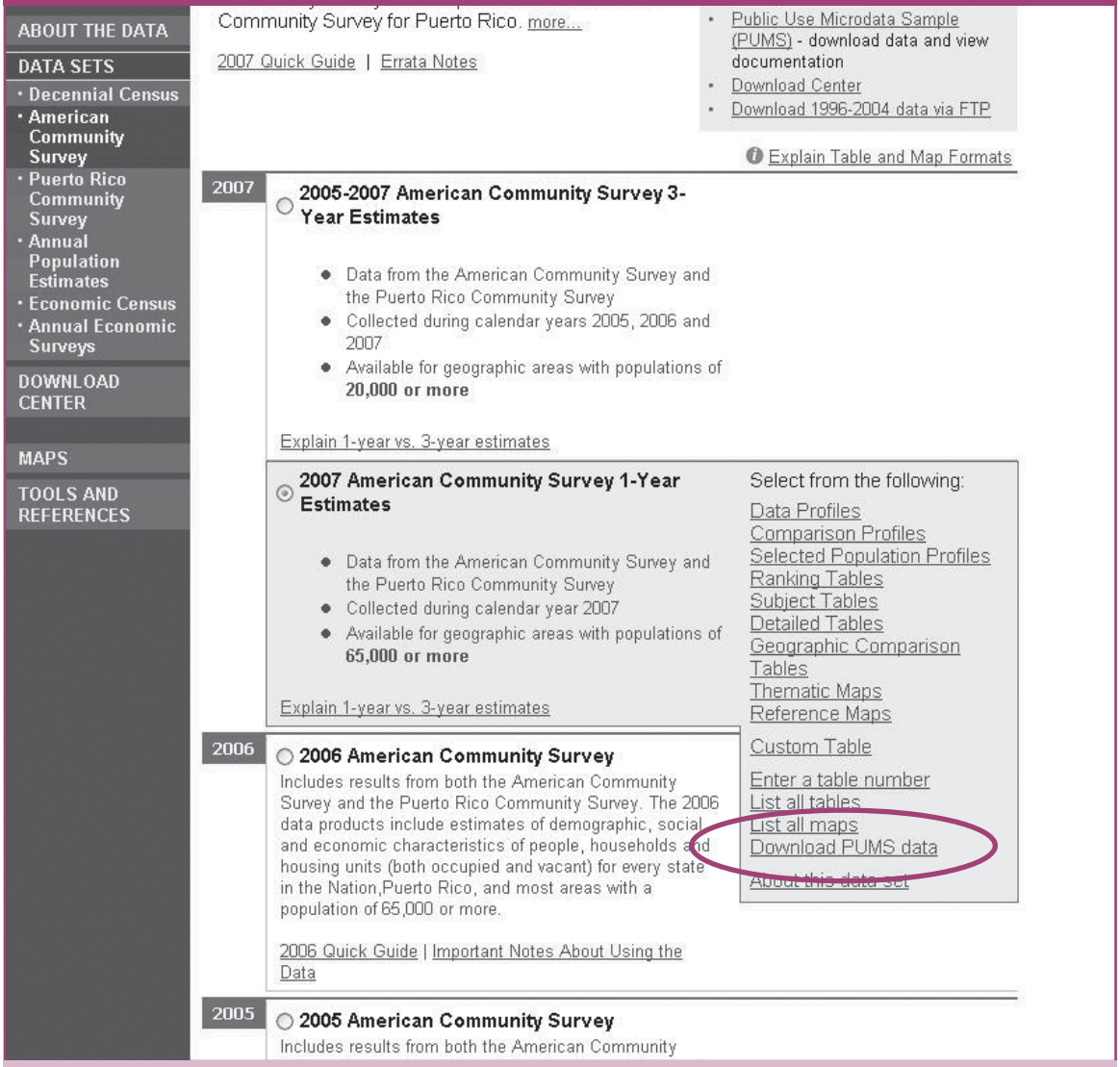
Figure 5. **Census Bureau Home Page Link to American FactFinder**

Source: U.S. Census Bureau, American FactFinder, accessed at <http://www.census.gov>.

Figure 6. **American FactFinder Home Page**

Source: U.S. Census Bureau, American FactFinder, accessed at <http://factfinder.census.gov>.

Figure 7. **American FactFinder ACS Data Sets Page**



Source: U.S. Census Bureau, American FactFinder, accessed at <http://factfinder.census.gov>.

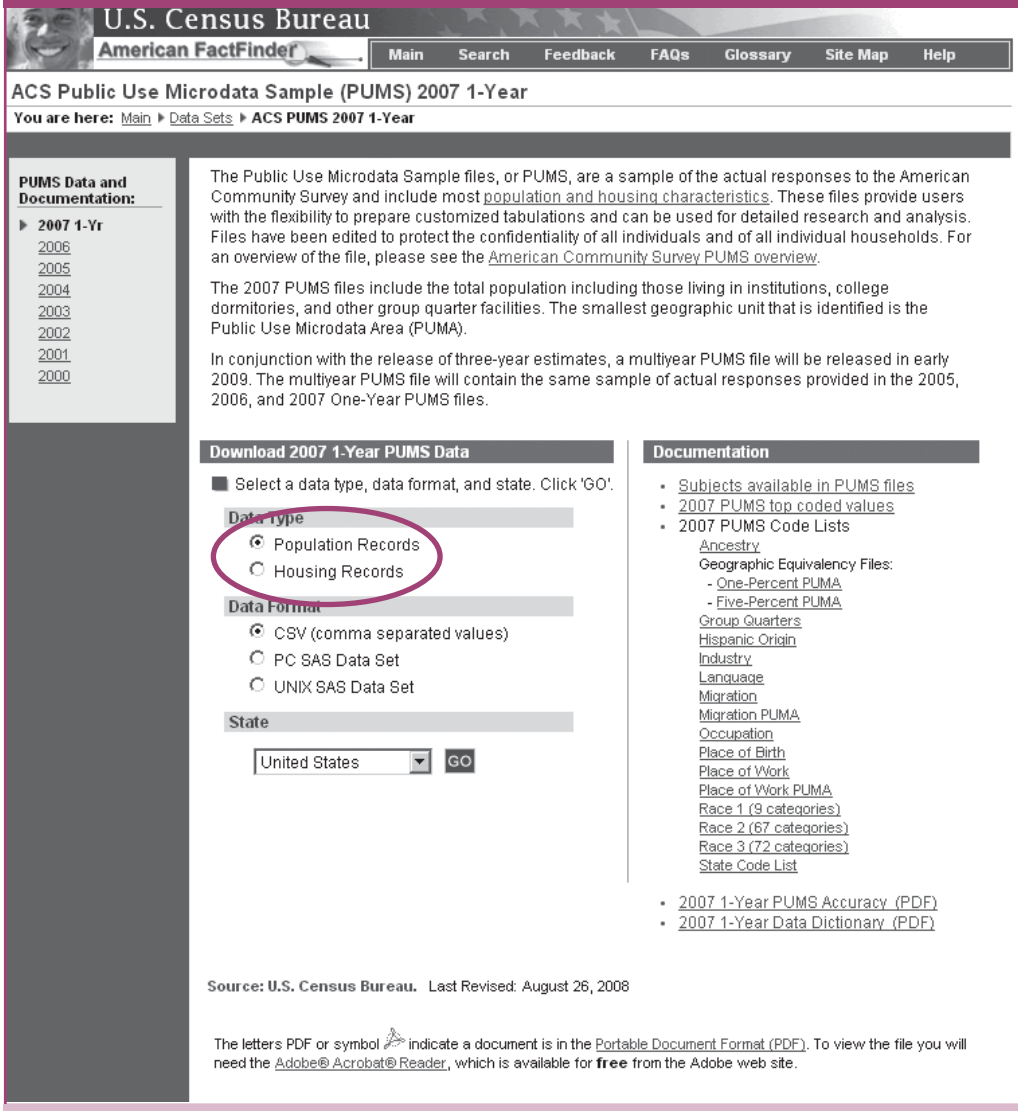
In Figure 7, please note that the ACS data products are listed with the most recent years at the top of the list. In order to select a year other than the most recent one, click on the radio button for the year desired. To get to the ACS PUMS files, click on the "Download PUMS data" menu item near the bottom of the list of options on the right. This will take you to the ACS PUMS download page shown in Figure 8.

You can also access the PUMS page through the Census Bureau's ACS Web site. The advantage of going through the ACS Web site is that the PUMS user verification files are listed. User verification files provide estimates for selected housing and population characteristics to help data users determine that they are using the weights correctly. The ACS Web site also includes a brief description of the PUMS. To access PUMS through the ACS Web site, start at the ACS Web site <www.census.gov/acs/www>. Click on the "**Access Data**" tab and under "**GET DATA**" select "**Public Use Microdata Sample (PUMS) Files**." On this page you will find links to documentation and to the PUMS files.

The ACS PUMS download page contains information about the ACS PUMS files in addition to providing access to the files themselves. This background information is on the right side of the page and includes the following:

- Lists of the subjects included in each of the housing and population record files.
- The state-specific values used in top coding the variables for confidentiality protection.

Figure 8. **ACS PUMS Download Page**



Source: U.S. Census Bureau, American FactFinder, accessed at <http://factfinder.census.gov>.

- Detailed codes for the variables that contain a large number of coded responses, such as ancestry and occupation.
- Links to the geographic equivalency files mentioned earlier. These are actually links to the Census 2000 PUMS files and documentation.
- A PDF file containing information about the accuracy of the PUMS and methods of calculating the standard errors and related measures.
- A PDF file containing the PUMS data dictionary. This has the names of each of the variables included in the PUMS files, a description of the variables' contents, the possible values for each variable, and the meanings of these values.

Some of these links are to specific files and others are to Web pages.

The middle column of the ACS PUMS download page (Figure 8) shows the options for downloading the ACS PUMS data. The first choice to make is which records to download, by choosing either population records or housing records. Table 3 shows the variables included in each of these record types for the 2007 ACS.

The next choice is which of the three formats to download. The first option is CSV (ASCII comma separated values) with the variable names in the first line. The second option is a PC SAS data set. The third option is a UNIX SAS data set. Choose the format that is easiest to import into your software. For example, when you are working with SAS on a PC, the PC SAS version is

best. When using SPSS, either the CSV file or the PC SAS is easy to read, but using the PC SAS files will provide the meaningful variable labels. The CSV format is a good choice if you plan to load the data into a relational database system such as Oracle, Access, or MySQL.

Finally, choose the geographic area. You can download the entire nation or any individual state by selecting the area and pressing the "GO" button. The national file is a single *large* data set containing all of the ACS PUMS records for the nation. If you want to look at just a few states, you can save time by downloading just those rather than the entire nation. If you want items from the housing records and the population records, you need to download both file types and merge them with your software.

The data file is sent as a compressed ZIP file. Once you save it on your computer, you will have to uncompress the file and read it into your software.

Creating PUMS Tables Using General Statistical Software

Getting Started

To demonstrate how to use the PUMS files to produce a table, we will ask the question, "What is the employment status of college students living in rental units by gross rent in Tompkins County, New York?"[5] We will

Table 3. **Topics Included in the 2007 ACS PUMS Files by Record Type**

Items in the housing record include:		Items in the person record include:	
Bedrooms	Meals included in rent	Ability to speak English	Mobility status
Condominium status	Mortgage status and selected monthly owner costs	Age	Occupation
Contract rent (monthly rent)	Plumbing facilities	Ancestry	Personal care limitations
Cost of utilities and fuels	Presence and age of own children	Citizenship	Place of birth
Family income	Presence of subfamilies in household	Class of worker	Place of work
Family, subfamily, and household relationships	Property value	Disability status	Poverty status
Farm status and value	Real estate taxes	Educational attainment	Race
Fire, hazard, and flood insurance	Residence state	Fertility	Relationship
Food stamps	Rooms	Hispanic origin	School enrollment and type of school
Fuel used	Telephone in housing unit	Income by type	Sex
Gross rent	Tenure	Industry	Time of departure for work
House heating fuel	Units in structure	Language spoken at home	Travel time to work
Household income	Vacancy status	Last week work status	Vehicle occupancy
Household type	Vehicles available	Marital status	Weeks worked
Kitchen facilities	Year householder moved into unit	Means of transportation to work	Work status
Linguistic isolation*	Year structure built	Migration	Work limitation status
		Military status, periods of active duty military service, veteran period of service	Year of entry

* Households in which no person, age 14 or over, speaks only English or speaks English very well.
Source: U.S. Census Bureau, accessed online at <www.census.gov/acs/www/Products/PUMS/PUMS3.htm>.

[5] While this example was created for this handbook, it is based on many requests the author has seen from policy analysts, local governments, and the private sector over the years.

look at how this table can be produced using the 2006 ACS PUMS file and general statistical software.

The first steps are to identify the relevant PUMAs and the variables of interest. From the map shown in Figure 1, we know that we are interested in PUMA 01700 in New York State. Given that Tompkins is by far the larger county and—with Cornell University and Ithaca College in the county—has many more college students than Seneca County, it is reasonable to use this PUMA as a proxy for just Tompkins County. If there were more college students within Seneca County, we might assume that the characteristics of the students in the two counties were the same and could estimate the number of students in Tompkins County by assigning it the county's share of college students in 2000.

Table 3 shows the topics included in the housing unit records and the population records. To answer our question, we can use data on tenure (owned or rented) and gross rent from the housing unit records and data on school enrollment and employment status from the population records. This means we need to combine both record types in order to produce the table.

Figure 9 shows the data dictionary information for the tenure variable. It includes the variable name, value length, description, value codes, and the descriptions.

From this part of the data dictionary we determine that we want the people with tenure (TEN) values of "3" (rented for cash rent) or "4" (no cash rent). Reviewing the data dictionary for the other variables, we determine that:

- We need to recode gross rent of the housing unit (GRNTP) into categories since the ACS PUMS file lists individual values. Recoding creates a new set of values for a variable by consolidating one or more values found in the data set into a new recoded value. For example, the age variable can be recoded to consolidate ages 18, 19, 20, and 21 into a new value of "18–21." Frequently, recoding is done by creating new variables because of differences in format, numeric versus text, or just to help keep things clearer when doing the analysis.

- We want to include people with a response of "6" (college undergraduate) or "7" (graduate or professional school) for the grade level attending (SCHG) variable.

- We need to recode the employment (ESR or employment status recode) variable so that the two civilian employed values are combined, as are the two Armed Forces values.

Figure 9. **Selected Variables From the Data Dictionary**

```
TEN         1
     Tenure
            b .N/A (GQ/vacant)
            1 .Owned with mortgage or loan
            2 .Owned free and clear
            3 .Rented for cash rent
            4 .No cash rent

VACS        1
     Vacancy status
            b .N/A (occupied/GQ)
            1 .For rent
            2 .Rented, not occupied
            3 .For sale only
            4 .Sold, not occupied
            5 .For seasonal/recreational/occasional use
            6 .For migratory workers
            7 .Other vacant

VAL         2
     Property value
            bb .N/A (GQ/rental unit/vacant, not for sale only)
            01 .Less than $ 10000
            02 .$ 10000 - $ 14999
            03 .$ 15000 - $ 19999
            04 .$ 20000 - $ 24999
            05 .$ 25000 - $ 29999
            06 .$ 30000 - $ 34999
```

Source: U.S. Census Bureau, 2006 ACS PUMS Data Dictionary, accessed online at <www.census.gov/acs/www/Downloads/PUMSDataDict06.pdf>.

Using General Statistical Software

Although the programming syntax varies across the different general statistical programs, there are also many similarities in terms of approaches to working with the PUMS files. SPSS produced the work shown here.

As mentioned above, first combine the housing unit and population records. To do this, merge the two data sets together to match the records on the SERIALNO variable. The SERIALNO variable is a 7-digit code that is unique across the nation. This appends the housing unit variables onto the population records.

After combining the records, limiting the number of records being processed (by selecting those of interest) will often increase processing speed. For this example, the universe of interest includes college students (grade level attending equals 6 or 7) renting apartments (tenure equals 3 or 4) in PUMA 01700. Selecting only these individuals reduces the number of records from 193,742 in the state to 60 college-student renters in Tompkins and Seneca Counties in New York.

Now that the records are limited to those of interest, recode the variables that need additional manipulation. Exactly how to do this varies with the software. In this example,

1. Recode gross rent (GRNTP) into a new variable for grouped gross rent (GGRNTP) with the categories of no cash rent, under $500, $500 to $999, and $1,000 or more.

2. Recode the employment status (ESR) variable into a new variable (EMPSTAT) with the categories of Civilian Employed (ESR equals 1 or 2), Unemployed (ESR equals 3), Armed Forces (ESR equals 4 or 5), and Not in Labor Force (ESR equals 6).

After creating these new variables, we are ready to create the data table. When producing the table, it is critical to remember that each person in the PUMS represents a different number of people in the population because of the ACS's sampling and weighting procedures. To account for this, the PUMS file contains a weighting factor for each population (PWGTP) and housing unit record (WGTP) that we will use to inflate the sample to the full population. Because we are interested in students (population), we use the PWGTP to weight the input records to estimate the total population. There are an additional 80 population and housing unit weight variables on the file, but the main purpose of these replicate weights is to calculate standard errors described later in this handbook. Again, exactly how you apply the weights and create the tables depends on the software. Table 4 shows the data produced through SPSS for our example.

Table 4. **PUMS Tabulation Results—2006 ACS**

Grouped Gross Rent and Employment Status Crosstabulation

		Employment Status			Total
		Civilian employed	Unemployed	Not in labor force	
Grouped gross rent	Under $500	36	0	279	315
	$500 to $999	2338	181	1439	3958
	$1,000 or more	1805	85	2421	4311
	No cash rent	217	0	14	231
	Total	4396	266	4153	8815

Source: U.S. Census Bureau, 2006 American Community Survey, Public Use Microdata Sample.

Creating PUMS Tables Using DataFerrett

If you do not have access to a general statistical program, it is still possible to create PUMS tabulations through the Census Bureau's DataFerrett program. DataFerrett is a tool developed by Census Bureau staff for extracting data and producing tables from a wide range of data products generated by a number of federal government agencies. DataFerrett functionality goes far beyond the simple example in this handbook. ACS data users are encouraged to work with DataFerrett to produce maps and charts and to develop complex recoding applications.

To demonstrate how to use DataFerrett to produce tables from the PUMS data, we will walk through an example of how to use the 2006 ACS PUMS to find the number and percentage of people aged 75 and older in every state with graduate or professional degrees. While it is possible to get the number of people aged

65 and older with bachelor or higher degrees from a pretabulated ACS tables, the age breakout we are interested in (75 and older) and the restriction to graduate and professional degrees are not available in any pretabulated table.

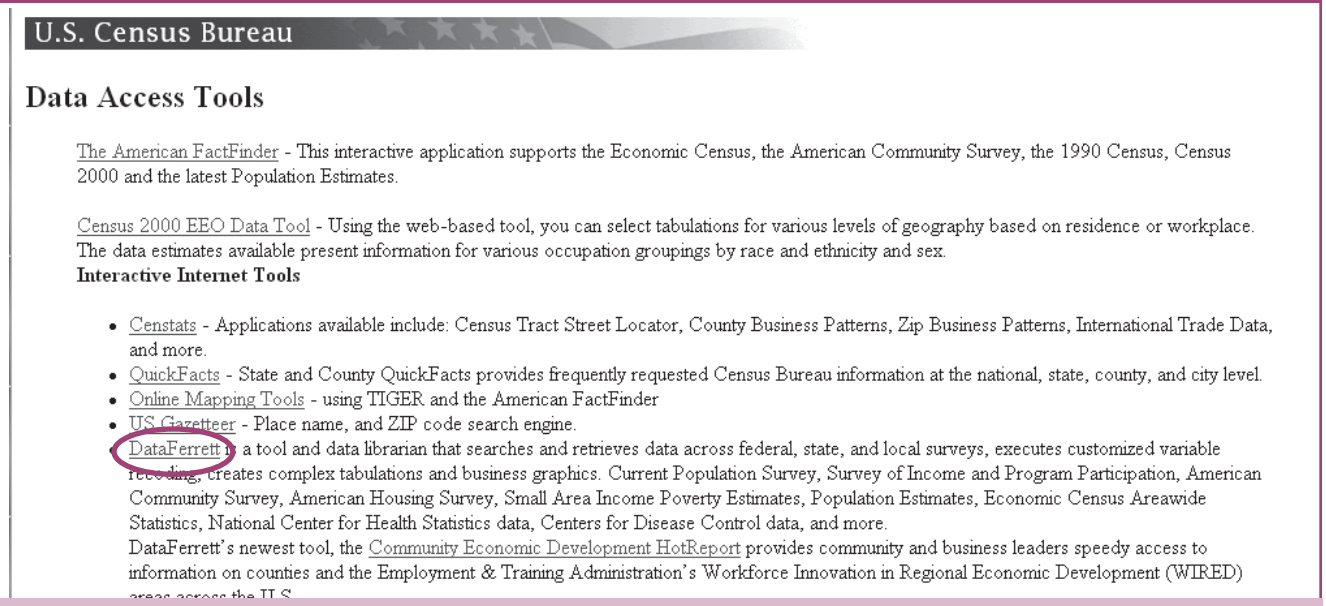
Getting Started

Since DataFerrett is a software application, download it from the Census Bureau's Web site and install it on your computer. Starting at the Census Bureau's home page <**www.census.gov**>, click on "**Data Tools**" on the left side of the screen as shown in Figure 10. This will take you to the page shown in Figure 11; click on the "**DataFerrett**" link. This will take you to the DataFerrett home page shown in Figure 12. In the upper-right hand corner of this page are a number of different versions of DataFerrett that you can download and install on your computer. The installation process is fairly standard and the prompts are clearly noted. This page also contains a number of useful tutorials and other support material about DataFerrett.

Figure 10. **Census Bureau's Home Page for Accessing DataFerrett**

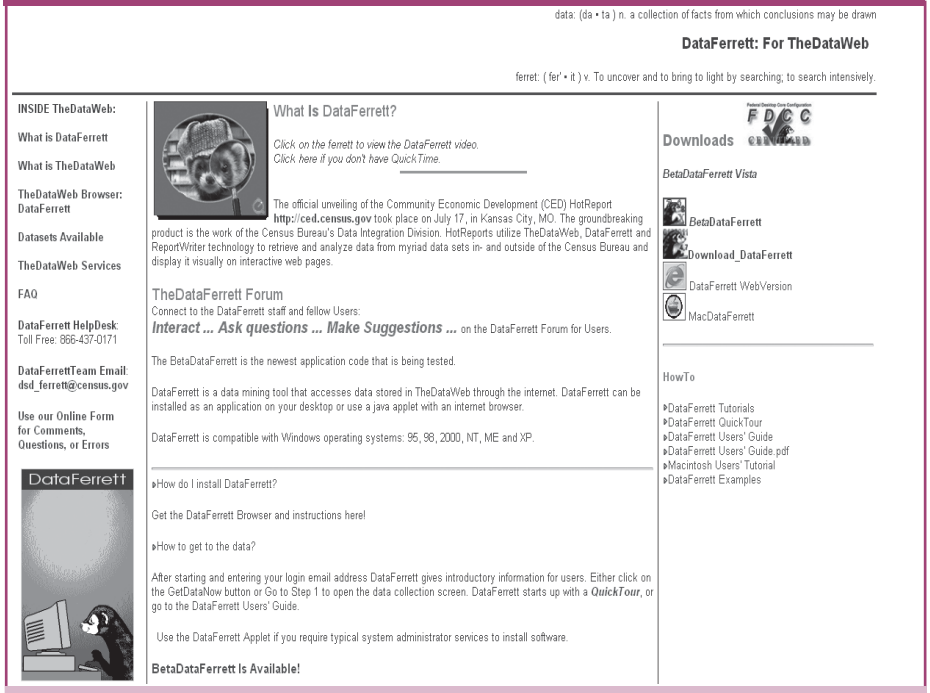
Source: U.S. Census Bureau, accessed online at <www.census.gov>.

Figure 11. **Census Bureau's Data Tools Page**



Source: U.S. Census Bureau, accessed online at <www.census.gov/main/www/access.html>.

Figure 12. **DataFerrett Home Page**

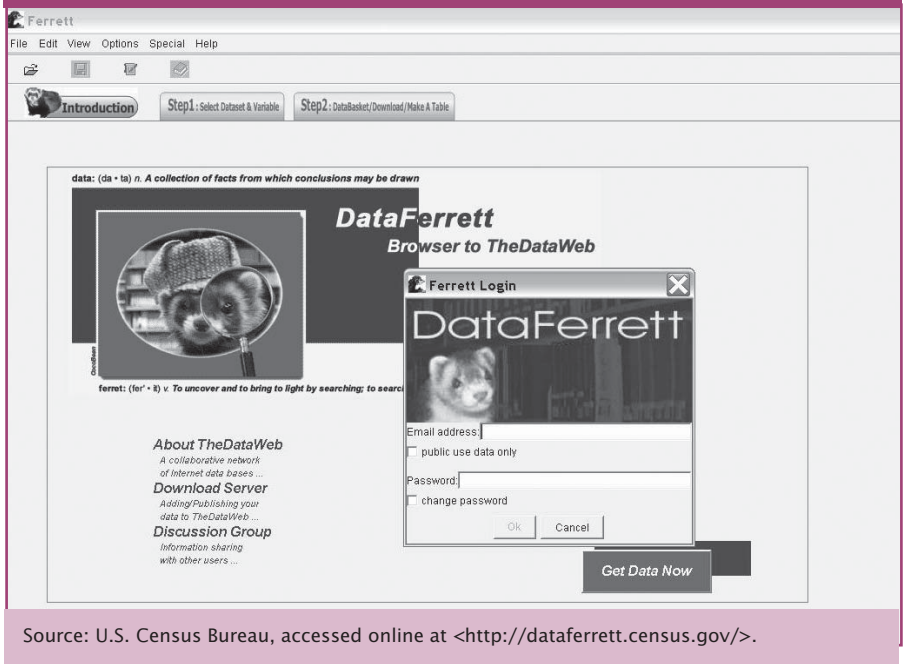


Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

After downloading and installing DataFerrett, you can run it as you would any other program; Figure 13 shows the opening screen of the program. Please note that you will need to sign onto DataFerrett with an e-mail address. After signing onto the program, you will see the screen shown in Figure 14. In order to get the data, click on the "**Get Data Now**" button.

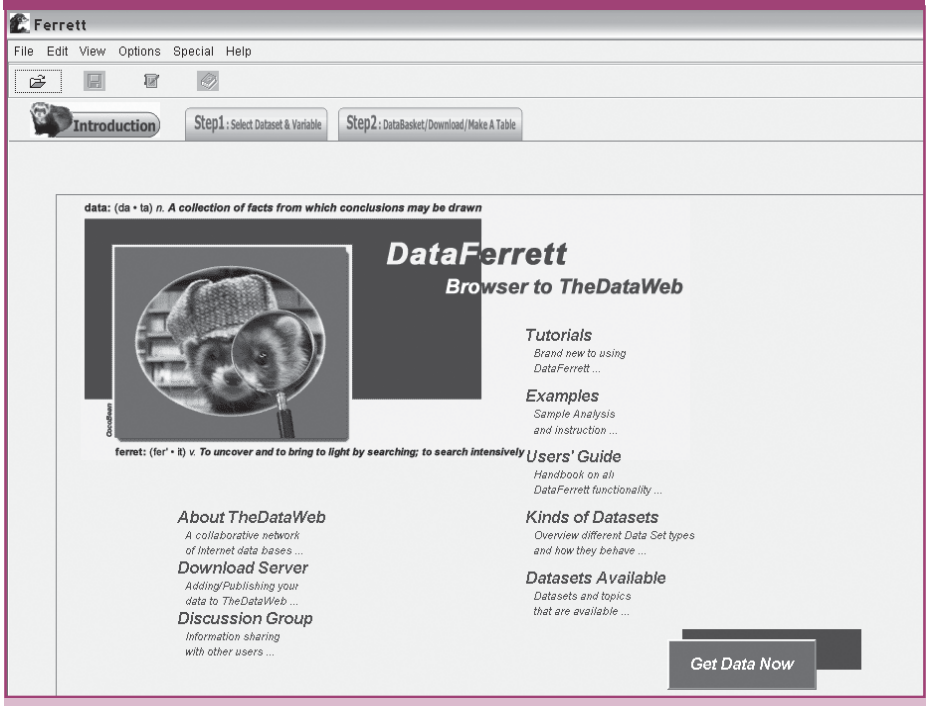
Clicking on the "**Get Data Now**" button opens DataFerrett's data selection screen shown in Figure 15. This screen allows you to find the data sets you want by type (Microdata, Aggregate Data, Longitudinal Data, Time Series Data), name of the data set (along the left side), or by the variable or subject you are interested in (along the top).

Figure 13. **DataFerrett Opening Page**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 14. **DataFerrett Introduction Page**



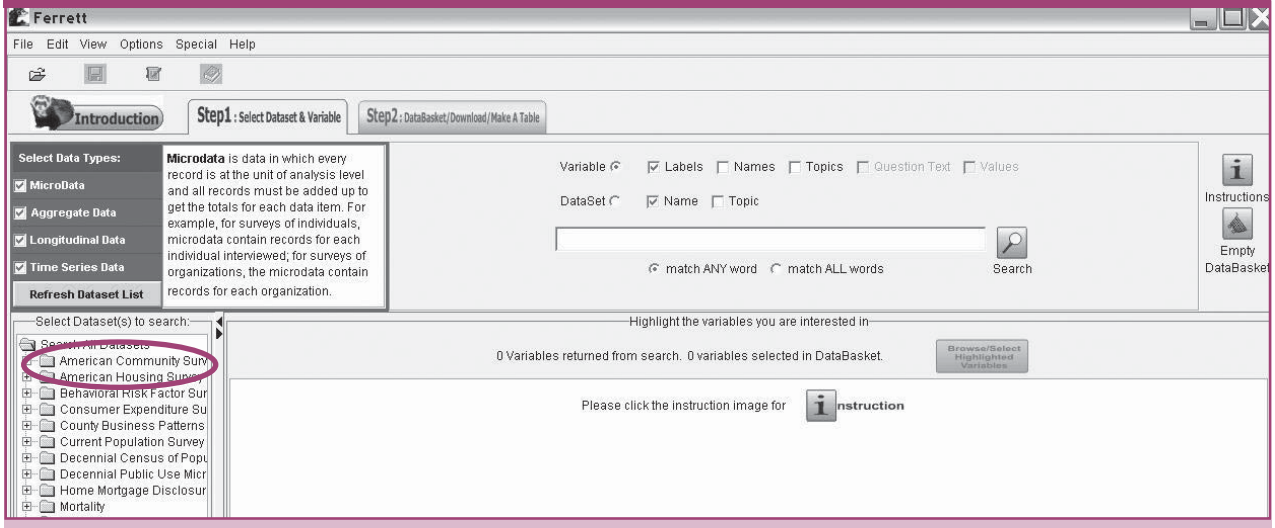
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Using DataFerrett

In this example we are using DataFerrett to produce 2006 ACS state-level tables of the number and percentage of people aged 75 and older with graduate or professional degrees. We are therefore interested in data from the 2006 ACS, specifically age and educational attainment data that are on the population record. In addition we want to have data by state.

The first step is to select our data set and topics. In the left sidebar, under "**Search All Datasets**" is an alphabetical list of data sets available to DataFerrett. Double-clicking on the "**American Community Survey**" folder (see Figure 15) reveals two options—the Public Use Microdata Sample and the Puerto Rico Public Use Microdata Sample. Opening either PUMS folder identifies the specific PUMS files that are available. For this example we will select "**Public Use Microdata Sample**" and

Figure 15. **DataFerrett Data Selection Screen**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

"**2006**." As shown in Figure 16, once we have selected "2006" we have a choice of reading a description of this data set (description) or reviewing the specific variables in the data set (view variables).

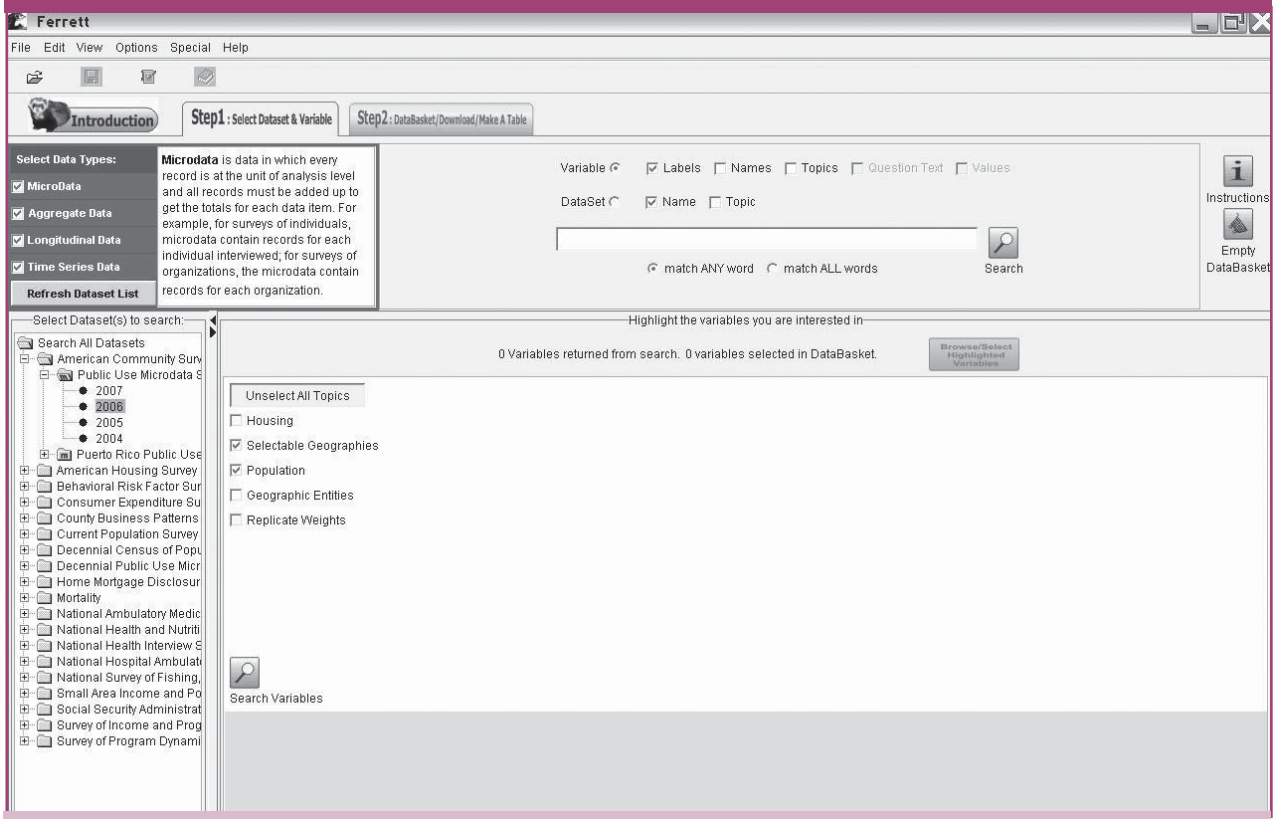
Select the "**View Variables**" button to get more information about the variables. This will open a screen similar to that shown in Figure 16. You are prompted to select the topics of interest. In this case, check the "**Selectable Geographies**" box to obtain state-level summaries and check the "**Population**" box to obtain age (AGEP) and educational attainment variables (SCHL).

A list of variables will appear for you to review, as shown in Figure 17. It is possible to sort this list by clicking on the column headings. You can select multiple variables by using the standard Window's **Ctrl-Click** method to select additional individual variables or **Shift-Click** to select a range of variables. Select the "**PUMS Age**," "**Geographic Items**," and the "**Educational Attainment**" lines using the **Ctrl-Click** method. Click the "**Search Variables**" button.

After highlighting the variables you are interested in, click on the "**Browse/Select Highlighted Variables**" button circled in Figure 17. If you select a mix of geographic selection and nongeographic variables, you will receive a warning box. Clicking "**OK**" allows you to proceed to selecting the geographic area of interest, as shown in Figure 18. Available geographies associated with the data set you selected will be displayed. For the 2006 ACS this includes states and PUMAs. For this example we would select "**FIPS State Code—ST—State of current residence**." Clicking the "**Next>**" button updates the selection box by displaying a list of states as shown in Figure 19. For this example we click on the "**Select All**" line and click on "**Next>**" to move this selection into the "**Geographies Selected**" box. After you have selected the geographic area or areas of interest, click "**Finish**." This will take you to the variable selection menu shown in Figures 20 and 21.

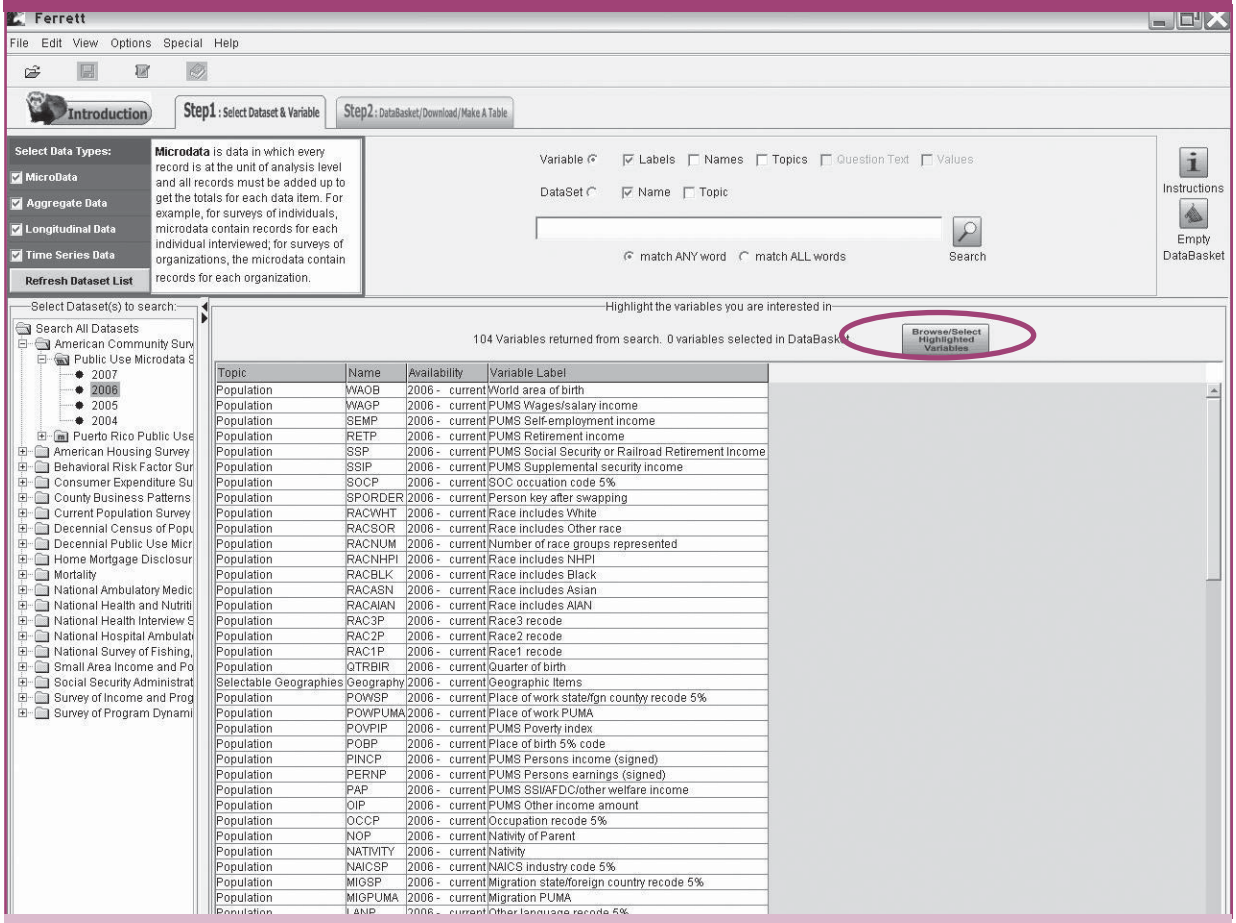
As you highlight a specific variable, details about that variable are displayed. Hightlight the "**PUMS Age**" variable. Figure 20 shows the details of the age variable. Note that the documentation states that the PUMS file includes continuous values of the variable AGEP between 0 and 99. This means that it is possible to obtain data for single years of age or to use single years of age to define specific age ranges within this interval. For this example, change the 0 to 75 in the "**Continuous values of AGEP**" limited to the universe to people aged 75 and older.

Figure 16. **2006 ACS PUMS Data Category Selection Screen**



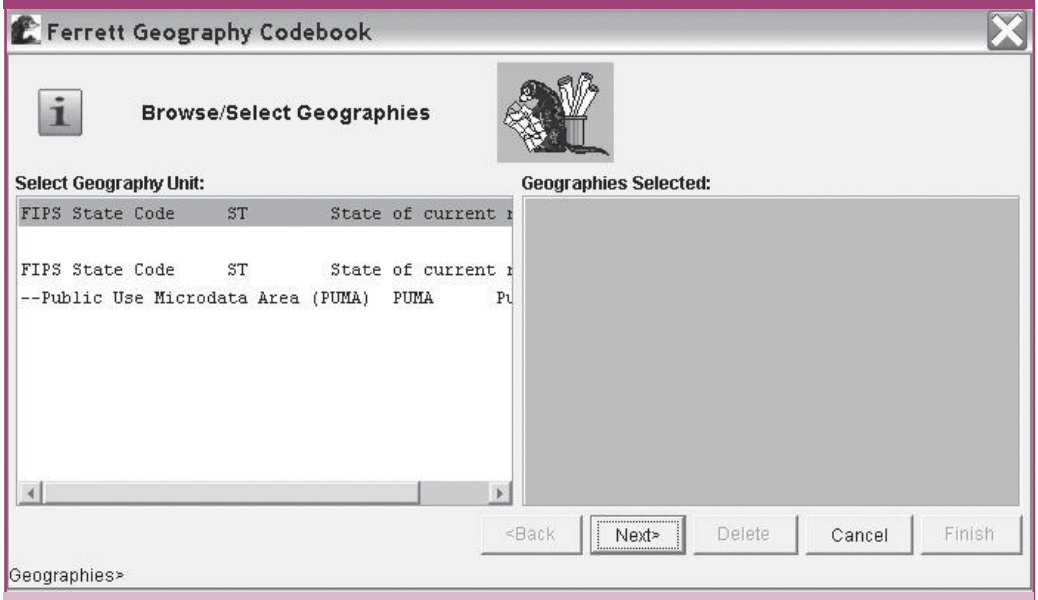
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 17. **DataFerrett Variable Selection Screen**



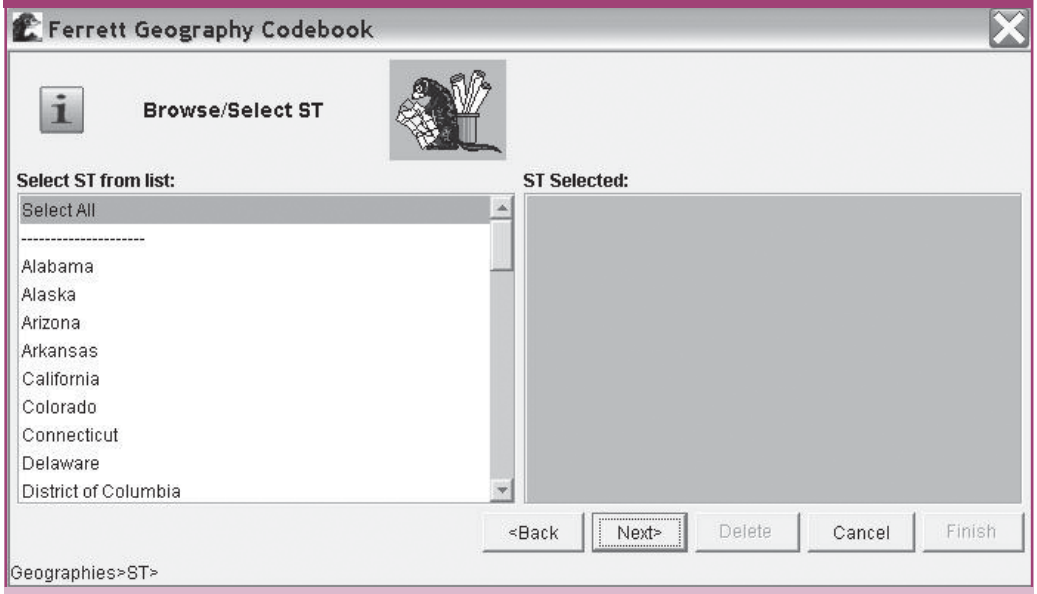
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 18. **DataFerrett Geographic Selection Dialog Box**



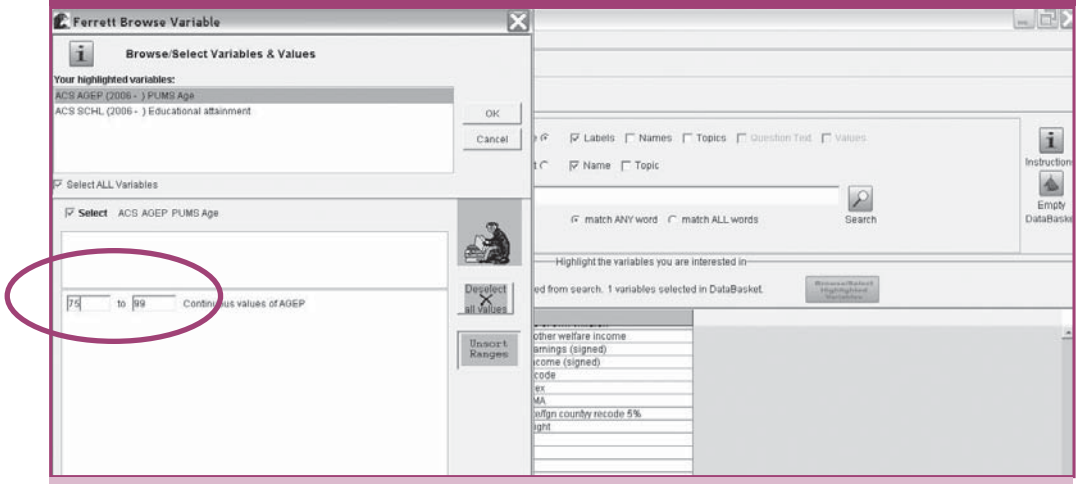
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 19. **DataFerrett State Selection Screen**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 20. **DataFerrett Variable and Value Selection Screen for a Continuous Variable—Age**



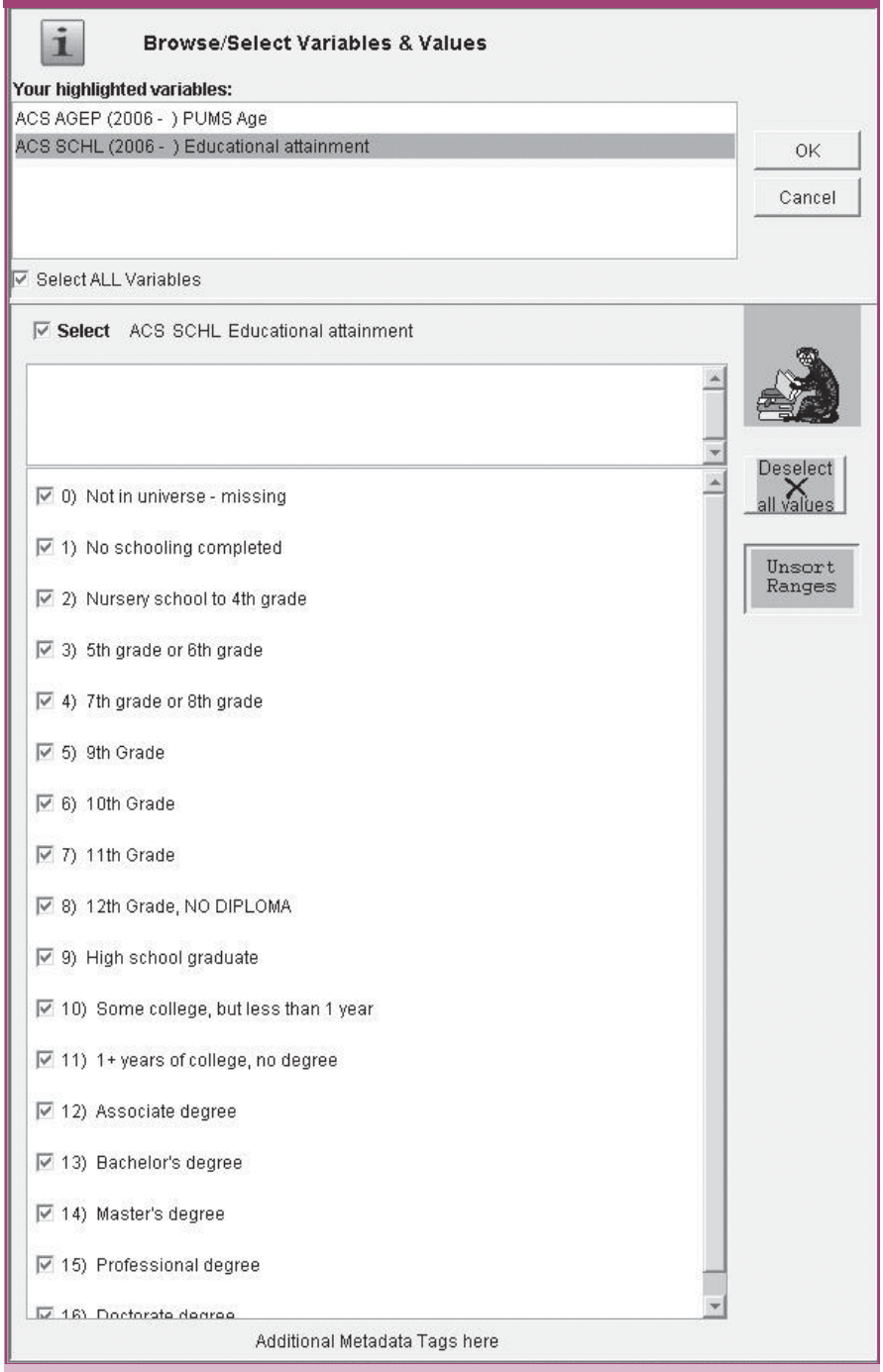
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Highlight the "**Educational attainment**" variable. Figure 21 shows the selection screen for the variable SCHL that has categorical or discrete values. Here you can choose selected values from a list of options by checking the boxes next to the values you want to include. Since our interest is finding the percentage of people aged 75 and older with a graduate or professional degree, we need to include all of the values in order to get the total population. So, we leave all of the choices checked. After you have selected the variables

and values you want to include, check the "**Select ALL Variables**" box and click "**OK**." You will then receive a confirmation about how many variables you have put into your data basket.

At this point, you are ready to download the data or create a data table. You can do this by clicking on the "**Step 2: DataBasket/Download/Make A Table**" tab at the top of the screen as shown in Figure 17. This will take you to a screen similar to that shown in Figure 22. First you need to determine if you want to combine existing categories into new categories. This is called "recoding." Highlighting the "**SCHL—educational attainment**" variable and click on the "**Recode**

Figure 21. **DataFerrett Variable and Value Selection Screen for a Discrete Variable—Educational Attainment**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Variable(s)" button on the right will bring up a recode dialog box like the one shown in Figure 23.

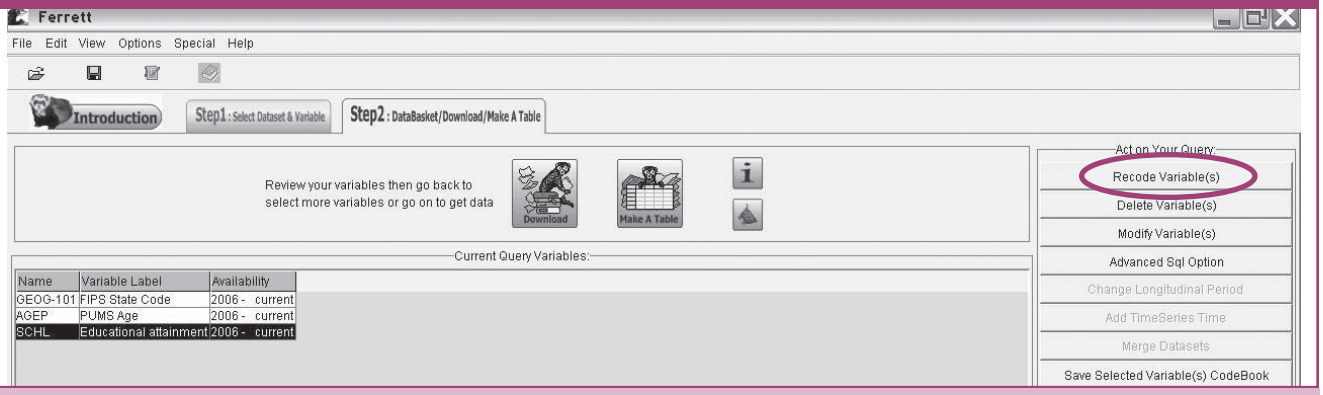
This dialog box allows you to recode the values of individual variables into new categories. The exact look of this box will vary depending on the variable you are recoding. Double-clicking on a label allows you to edit the label itself. Once you have recoded a variable, click "OK." Then you can recode another variable. The "Modify Variable(s)" button in Figure 22 allows you to change selected values of existing variables and labels of recode variables.

For this example you should recode the SCHL variable to identify people with a master's degree, professional degree, or a doctorate degree. Highlight the SCHL variable and click the "**Recode Variable(s)**" button.

A default label will be provided for this recode. It is best to rename it so you can remember it later. For this example, double-click on the default "**Recode1**" label and type "Grad School." Use the "**Shift**" key to select multiple values to highlight the values 14, 15, and 16 and then click the "**Recode**" button creates a new value that is defined as the sum of these three values. By default the remainder is assigned a value of 2. You should also edit the labels of this new value.

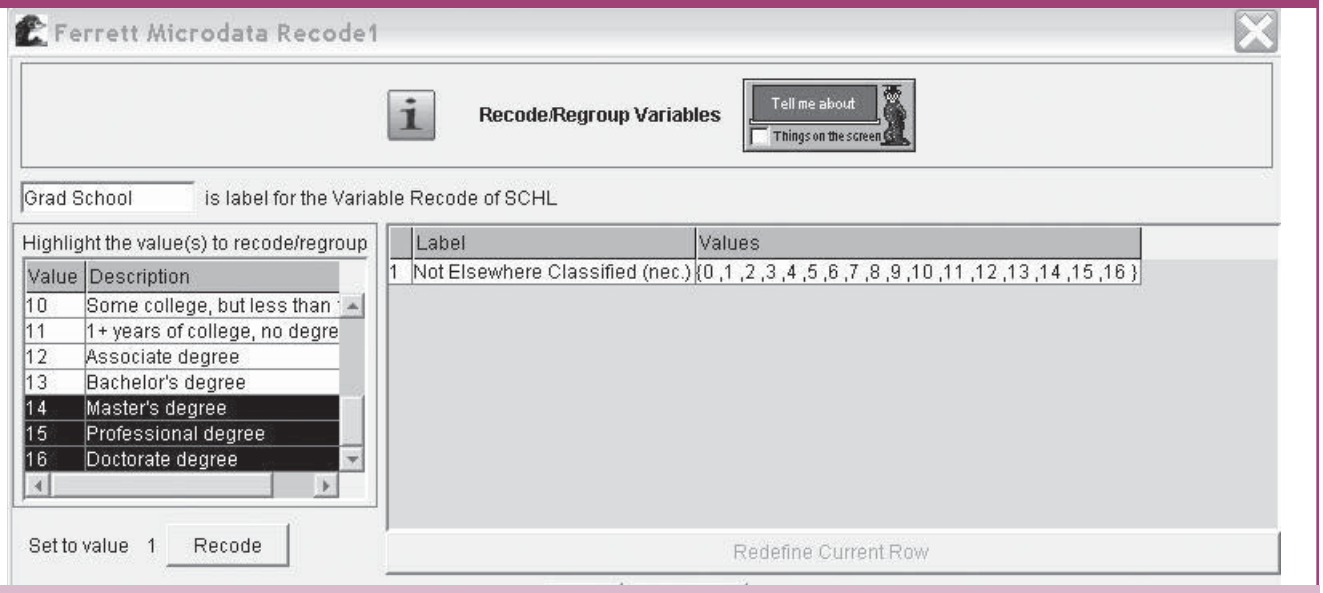
Once you have your variables the way you want them, you can create the table by clicking on the "**Make A Table**" button shown in Figure 22. This will open a table window like the one shown in Figure 24. In order to set up the table, simply drag and drop the variables you want in the table. Specifically for this example you would drag the variable that you want in the rows

Figure 22. **DataFerrett DataBasket Screen**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 23. **DataFerrett Recode Dialog Box**



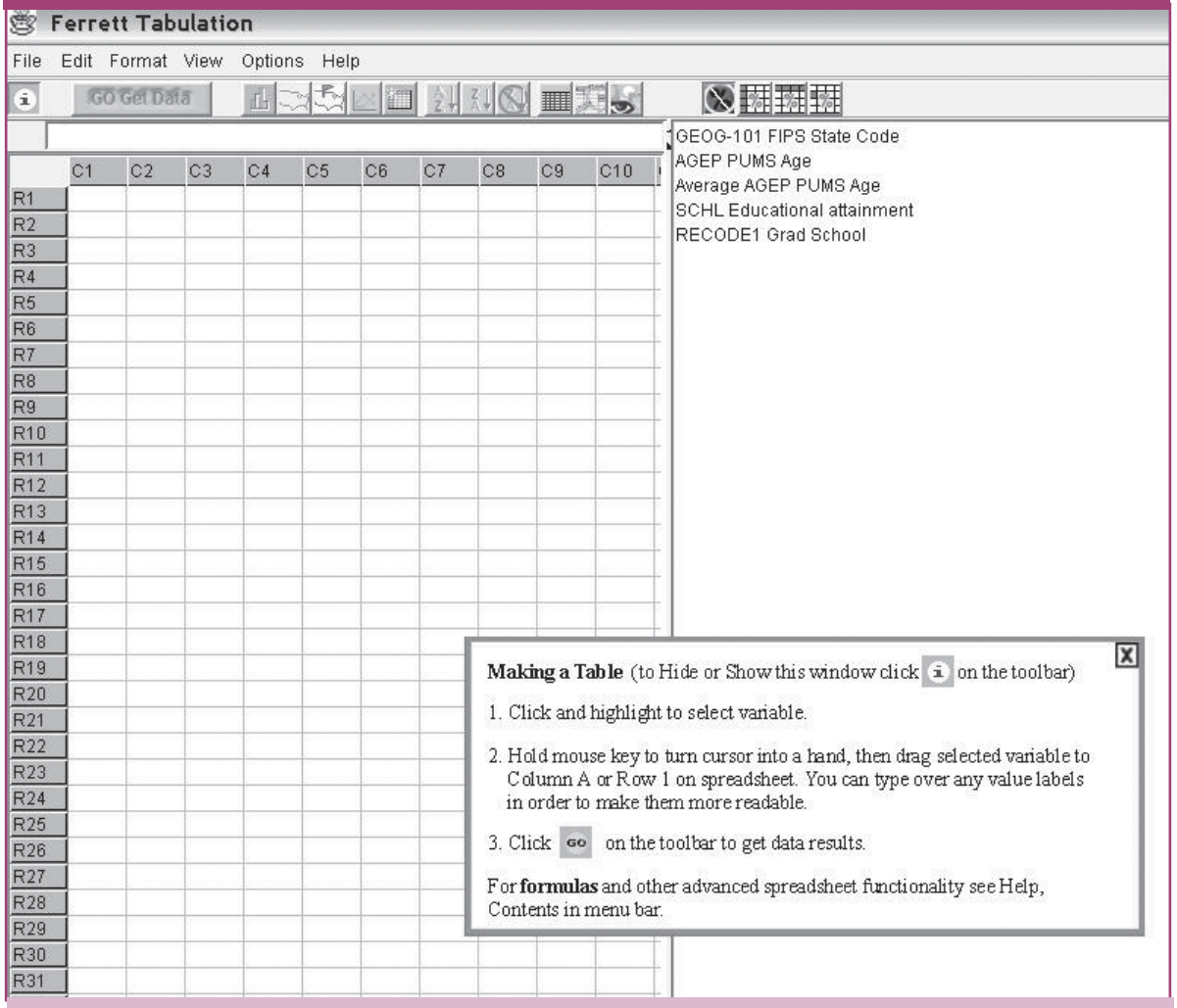
Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

(**GEOG-101**) and drop it in cell **R1/C1**. The list of states will appear. Then drag the variable that you want in the columns (**Record 1 Grad School**) and drop it in cell **R1/C2**. The resulting table is shown in Figure 25.

Once you have your data shell set up the way you want it, simply press the "**GO Get Data**" button. This will create the table shown in Figure 26. Recall that we had restricted our universe to the population aged 75 and older. This table therefore provides us with the values we need to determine the percentage of the total population aged 75 and older with professional or graduate degrees. From this table, we see, for example, that California has about 1,925,000 people aged 75 and older. Of these, 154,178 have graduate or professional degrees. To convert these values to percentages, select the "Percent of first column" option from the tool bar. By selecting different options from the tool bar, it is also possible to sort on one of these variables, display graphs and maps, and more.

DataFerrett automatically chooses either the population weight (PWGTP) or the housing unit weight (WGTP) based on the variables you have chosen for your tabulations. It is often helpful to check the weight chosen by DataFerrett to ensure that the correct weight is being used for your analysis. For this example, click on the "**Options**" pull-down menu, select "**Weighting**" from the drop-down menu and then select "**Unweighted**" before running your tabulation. Click on "**Go get data.**" If you think you should use a different one, you can simply click on that one.

Figure 24. **DataFerrett Tabulation Shell**



Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 25. **DataFerrett Tabulation Shell Ready for Data Tabulation**

	C1	C2	C3	C4
R1		Total RECODE1	RecodeValue_1	Not Elsewhere C
R2	Total GEOG-101	?	?	
R3	Alabama	?	?	
R4	Alaska	?	?	
R5	Arizona	?	?	
R6	Arkansas	?	?	
R7	California	?	?	
R8	Colorado	?	?	
R9	Connecticut	?	?	
R10	Delaware	?	?	
R11	District of Columbia	?	?	
R12	Florida	?	?	
R13	Georgia	?	?	
R14	Hawaii	?	?	
R15	Idaho	?	?	
R16	Illinois	?	?	
R17	Indiana	?	?	
R18	Iowa	?	?	
R19	Kansas	?	?	
R20	Kentucky	?	?	
R21	Louisiana	?	?	
R22	Maine	?	?	
R23	Maryland	?	?	
R24	Massachusetts	?	?	
R25	Michigan	?	?	
R26	Minnesota	?	?	
R27	Mississippi	?	?	
R28	Missouri	?	?	
R29	Montana	?	?	
R30	Nebraska	?	?	
R31	Nevada	?	?	
R32	New Hampshire	?	?	
R33	New Jersey	?	?	
R34	New Mexico	?	?	
R35	New York	?	?	
R36	North Carolina	?	?	
R37	North Dakota	?	?	
R38	Ohio	?	?	
R39	Oklahoma	?	?	
R40	Oregon	?	?	
R41	Pennsylvania	?	?	
R42	Rhode Island	?	?	
R43	South Carolina	?	?	
R44	South Dakota	?	?	
R45	Tennessee	?	?	

Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Figure 26. **DataFerrett Tabulated Results**

Ferrett Tabulation

File Edit Format View Options Help

GO Get Data

(C1R7)

	C1	C2	C3	C4
R1		Total RECODE1	RecodeValue_1	Not Elsewhere C
R2	Total GEOG-101	18,317,548	1,199,228	
R3	Alabama	286,332	11,742	
R4	Alaska	18,232	969	
R5	Arizona	388,346	32,541	
R6	Arkansas	191,133	7,745	
R7	California	1,925,329	154,178	
R8	Colorado	222,073	17,909	
R9	Connecticut	245,202	20,237	
R10	Delaware	56,835	3,447	
R11	District of Columbia	34,541	6,353	
R12	Florida	1,574,850	119,564	
R13	Georgia	405,896	23,059	
R14	Hawaii	95,664	5,712	
R15	Idaho	81,325	4,445	
R16	Illinois	765,763	44,190	
R17	Indiana	381,011	23,165	
R18	Iowa	228,178	10,442	
R19	Kansas	185,380	9,778	
R20	Kentucky	251,064	11,615	
R21	Louisiana	245,940	14,580	
R22	Maine	93,437	5,603	
R23	Maryland	304,166	31,444	
R24	Massachusetts	444,306	33,710	
R25	Michigan	619,252	35,930	
R26	Minnesota	321,612	17,237	
R27	Mississippi	171,434	7,898	
R28	Missouri	388,454	19,013	
R29	Montana	65,022	4,928	
R30	Nebraska	119,726	4,469	
R31	Nevada	121,501	7,688	
R32	New Hampshire	80,858	7,045	
R33	New Jersey	571,812	36,004	

Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Data Quality in PUMS

Because data produced from the PUMS are based on a sample, there is little chance that the numbers shown in these tables are the exact numbers that would be obtained if everyone in the population were counted. For example, the number of college students in Tompkins and Seneca Counties paying gross rent less than $500 and employed could be 36, as shown in Table 4, but it is also possible that there is only one person in this situation and that one person just happened to be in the sample with a weight of 36.

One way to quickly check for results based on an extremely small sample is to reproduce the tables without using weights. Table 5 shows the unweighted sample counts for the data shown in Table 4. From this table, you see that in fact only one student was paying less than $500 in gross rent and was employed. You also see that several cells are based on samples of one to six people. This suggests that you should use these results with great care. If the cells of interest are those with 12, 13, or 21 students in the sample, then this tabulation may be more useful. The best way to determine that is to look at the sampling error for your estimates. One way to improve the quality of this analysis would be to use either the 3- or 5-year PUMS files for the area (when available). Another method would be to look at these data for a larger geographic area.

Looking at the unweighted counts for the graduate degree example produced using DataFerrett, shown in Figure 27; we see that this tabulation is based on a very robust sample of 214,072 people, of whom 14,253 have graduate or professional degrees. The larger sample size indicates that the results are generally going to be fairly accurate. However, we also see that in Alaska only 10 people are aged 75 and older with graduate or professional degrees. This suggests that we need to be a bit more careful about conclusions drawn for this particular state.

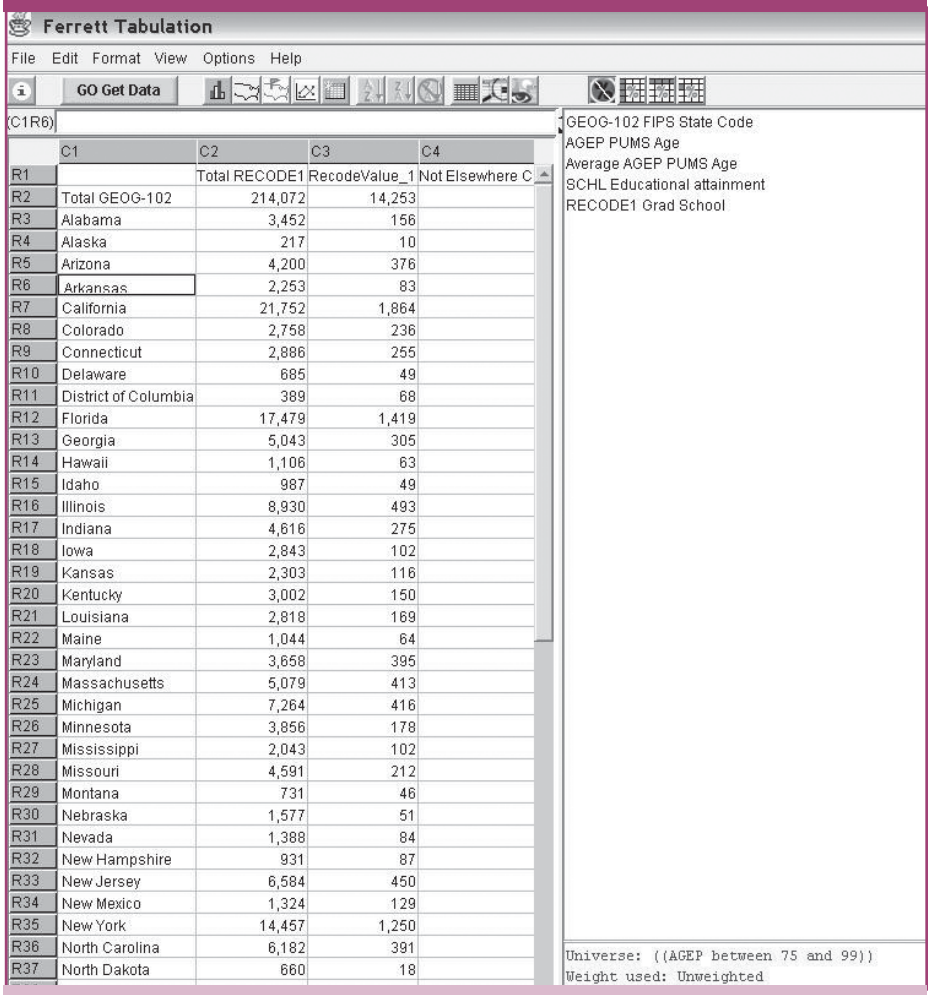
Table 5. **Unweighted Sample Counts for Tompkins-Seneca County Example**

Group Gross Rent and Employment Status Crosstabulation

		Employment Status			
		Civilian employed	**Unemployed**	**Not in labor force**	**Total**
Grouped gross rent	Under $500	1	0	3	4
	$500 to $999	21	1	13	35
	$1,000 or more	6	1	12	19
	No cash rent	1	0	1	2
	Total	29	2	29	60

Source: U.S. Census Bureau, 2006 American Community Survey, Public Use Microdata Sample.

Figure 27. **Unweighted Counts for the Graduate-Professional Degree Example**



Ferrett Tabulation

File Edit Format View Options Help

GO Get Data

(C1R6)

	C1	C2	C3	C4
R1		Total RECODE1	RecodeValue_1	Not Elsewhere C
R2	Total GEOG-102	214,072	14,253	
R3	Alabama	3,452	156	
R4	Alaska	217	10	
R5	Arizona	4,200	376	
R6	Arkansas	2,253	83	
R7	California	21,752	1,864	
R8	Colorado	2,758	236	
R9	Connecticut	2,886	255	
R10	Delaware	685	49	
R11	District of Columbia	389	68	
R12	Florida	17,479	1,419	
R13	Georgia	5,043	305	
R14	Hawaii	1,106	63	
R15	Idaho	987	49	
R16	Illinois	8,930	493	
R17	Indiana	4,616	275	
R18	Iowa	2,843	102	
R19	Kansas	2,303	116	
R20	Kentucky	3,002	150	
R21	Louisiana	2,818	169	
R22	Maine	1,044	64	
R23	Maryland	3,658	395	
R24	Massachusetts	5,079	413	
R25	Michigan	7,264	416	
R26	Minnesota	3,856	178	
R27	Mississippi	2,043	102	
R28	Missouri	4,591	212	
R29	Montana	731	46	
R30	Nebraska	1,577	51	
R31	Nevada	1,388	84	
R32	New Hampshire	931	87	
R33	New Jersey	6,584	450	
R34	New Mexico	1,324	129	
R35	New York	14,457	1,250	
R36	North Carolina	6,182	391	
R37	North Dakota	660	18	

GEOG-102 FIPS State Code
AGEP PUMS Age
Average AGEP PUMS Age
SCHL Educational attainment
RECODE1 Grad School

Universe: ((AGEP between 75 and 99))
Weight used: Unweighted

Source: U.S. Census Bureau, accessed online at <http://dataferrett.census.gov/>.

Measuring Statistical Accuracy

Most researchers want more formal measures of the impact of sampling on the results. The most commonly used measures are the standard error, the margin of error, and the confidence interval. Both the margin of error and the confidence interval are based on the standard error. The relationships between these measures are described in more detail in Appendix 3.

There are two ways to calculate the standard error. The first is a generalized standard error formula provided by the Census Bureau. The second is through the use of the replicate weights provided by the Census Bureau as part of the PUMS file.

Generalized Standard Error Formula Method

The Census Bureau provides a number of formulas to approximate the standard error for most of the situations PUMS users are likely to encounter. These are included in the "Accuracy of the PUMS" provided for each year's PUMS files. The most commonly used formulas are those for totals and percentages.

To find the standard error for the estimated 2,338 employed college students paying between $500 and $999 a month on rental housing, use the standard error formula for the totals. This formula is:

$$SE(\hat{Y}) = DF * \sqrt{99 * \hat{Y}(1 - \frac{\hat{Y}}{N})}$$

where:

DF = design factor

N = size of the geographic area (total population or housing units)

$\hat{Y}$ = estimated value.

The design factor is found in a state-specific table in the "Accuracy Statement" and varies for the different characteristics being considered, such as tenure, employment status, and gross rent. When more than one characteristic is involved in the analysis, it is best to use the largest design factor for the factors being considered. In this example, that would be the design factor for gross rent, which in New York State is 1.8.

The estimate ($\hat{Y}$) from Table 4 is 2,338. The area size is the combined population of Seneca and Tompkins Counties, because we are looking at people (college students). This value (N) from Table 2 is 131,869.

Using these numbers in the above equation shows the standard error of this estimated value to be 858.

Replicate Weights Method

Another method of producing the standard error takes advantage of the 80 replicate weights provided for each population and each housing unit record in the PUMS. While this method is a bit more accurate than the generalized formula above, it is also more computationally intense.

The first step is to produce the estimate. In the example above, this is 2,338. Then you would produce this same estimate 80 times, using each of the 80 different replicate weights. Once you have these 81 estimated values, you can calculate the standard error using the following formula:

$$SE(X) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} \left(X_r - X\right)^2}$$

where:

X = the estimate based on the original weight, e.g., 2,338

X_r = the 80 individual estimates based on each of the replicate weights.

The ease with which the 80 replicated estimates can be produced depends largely on the software used to produce data from the PUMS files.

Margin of Error and Confidence Intervals

Two measures that are particularly useful in looking at the quality of the PUMS estimates are the margin of error (MOE) and the confidence interval. The Census Bureau reports these for a 90-percent confidence interval. Once you have the standard error, the margin of error is very easy to calculate. It is the standard error times 1.645, for the 90-percent confidence level. In our example, where the standard error was 858, the margin of error is 1,411.

The confidence interval is the estimate plus or minus the margin of error. In this case, the 90-percent confidence level would run from 927 to 3,749. In other words, there is a 90 percent chance that this interval would contain the average estimate of employed college students paying between $500 and $999 a month for rent, taken over all possible samples. How useful this estimated 2,338 answer is depends on the sensitivity of the decision being made to variations in the estimate.

Care is required when reporting values when the confidence interval drops below zero or above the area's total population. In these cases, consider those values as logical limits when reporting the confidence intervals.

Summary

The ACS PUMS files are designed to allow data users to produce their own tabulations of ACS data without the expense or time required when the Census Bureau produces custom tabulations. Producing PUMS tabulations or doing more sophisticated modeling of the data requires specialized software, such as SPSS, SAS, another statistical program, a relational database management program, or the Census Bureau's DataFerrett program.

While the ACS PUMS files provide great flexibility, the need to protect the confidentiality of the respondents has imposed some limitations on the data. These include the use of a limited portion of ACS responses and a limited set of geographic areas with populations of 100,000 or more.

The smaller sample size of the PUMS increases the need to calculate measures of uncertainty, such as margins of error, around the estimates. As described above, there are two ways to do this. The general method is less accurate than the replicate weights method, but it is less cumbersome for the data user. The choice of method is one that has to be guided by balancing these factors.

Glossary

Accuracy. One of four key dimensions of survey quality. Accuracy refers to the difference between the survey estimate and the true (unknown) value. Attributes are measured in terms of sources of error (for example, coverage, sampling, nonresponse, measurement, and processing).

American Community Survey Alert. This periodic electronic newsletter informs data users and other interested parties about news, events, data releases, congressional actions, and other developments associated with the ACS. See <http://www.census.gov/acs/www/Special/Alerts/Latest.htm>.

American FactFinder (AFF). An electronic system for access to and dissemination of Census Bureau data on the Internet. AFF offers prepackaged data products and user-selected data tables and maps from Census 2000, the 1990 Census of Population and Housing, the 1997 and 2002 Economic Censuses, the Population Estimates Program, annual economic surveys, and the ACS.

Block group. A subdivision of a census tract (or, prior to 2000, a block numbering area), a block group is a cluster of blocks having the same first digit of their four-digit identifying number within a census tract.

Census geography. A collective term referring to the types of geographic areas used by the Census Bureau in its data collection and tabulation operations, including their structure, designations, and relationships to one another. See <http://www.census.gov/geo/www/index.html>.

Census tract. A small, relatively permanent statistical subdivision of a county delineated by a local committee of census data users for the purpose of presenting data. Census tract boundaries normally follow visible features, but may follow governmental unit boundaries and other nonvisible features; they always nest within counties. Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants.

Coefficient of variation (CV). The ratio of the standard error (square root of the variance) to the value being estimated, usually expressed in terms of a percentage (also known as the relative standard deviation). The lower the CV, the higher the relative reliability of the estimate.

Comparison profile. Comparison profiles are available from the American Community Survey for 1-year estimates beginning in 2007. These tables are available for the United States, the 50 states, the District of Columbia, and geographic areas with a population of more than 65,000.

Confidence interval. The sample estimate and its standard error permit the construction of a confidence interval that represents the degree of uncertainty about the estimate. A 90-percent confidence interval can be interpreted roughly as providing 90 percent certainty that the interval defined by the upper and lower bounds contains the true value of the characteristic.

Confidentiality. The guarantee made by law (Title 13, U.S. Code) to individuals who provide census information, regarding nondisclosure of that information to others.

Consumer Price Index (CPI). The CPI program of the Bureau of Labor Statistics produces monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services.

Controlled. During the ACS weighting process, the intercensal population and housing estimates are used as survey controls. Weights are adjusted so that ACS estimates conform to these controls.

Current Population Survey (CPS). The CPS is a monthly survey of about 50,000 households conducted by the Census Bureau for the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population.

Current residence. The concept used in the ACS to determine who should be considered a resident of a sample address. Everyone who is currently living or staying at a sample address is considered a resident of that address, except people staying there for 2 months or less. People who have established residence at the sample unit and are away for only a short period of time are also considered to be current residents.

Custom tabulations. The Census Bureau offers a wide variety of general purpose data products from the ACS. These products are designed to meet the needs of the majority of data users and contain predefined

sets of data for standard census geographic areas, including both political and statistical geography. These products are available on the American FactFinder and the ACS Web site.

For users with data needs not met through the general purpose products, the Census Bureau offers "custom" tabulations on a cost-reimbursable basis, with the American Community Survey Custom Tabulation program. Custom tabulations are created by tabulating data from ACS microdata files. They vary in size, complexity, and cost depending on the needs of the sponsoring client.

Data profiles. Detailed tables that provide summaries by social, economic, and housing characteristics. There is a new ACS demographic and housing units profile that should be used if official estimates from the Population Estimates Program are not available.

Detailed tables. Approximately 1,200 different tables that contain basic distributions of characteristics. These tables provide the most detailed data and are the basis for other ACS products.

Disclosure avoidance (DA). Statistical methods used in the tabulation of data prior to releasing data products to ensure the confidentiality of responses. See Confidentiality.

Estimates. Numerical values obtained from a statistical sample and assigned to a population parameter. Data produced from the ACS interviews are collected from samples of housing units. These data are used to produce estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology.

File Transfer Protocol (FTP) site. A Web site that allows data files to be downloaded from the Census Bureau Web site.

Five-year estimates. Estimates based on 5 years of ACS data. These estimates reflect the characteristics of a geographic area over the entire 5-year period and will be published for all geographic areas down to the census block group level.

Geographic comparison tables. More than 80 single-variable tables comparing key indicators for geographies other than states.

Geographic summary level. A geographic summary level specifies the content and the hierarchical relationships of the geographic elements that are required to tabulate and summarize data. For example, the county summary level specifies the state-county hierarchy. Thus, both the state code and the county code are required to uniquely identify a county in the United States or Puerto Rico.

Group quarters (GQ) facilities. A GQ facility is a place where people live or stay that is normally owned or managed by an entity or organization providing housing and/or services for the residents. These services may include custodial or medical care, as well as other types of assistance. Residency is commonly restricted to those receiving these services. People living in GQ facilities are usually not related to each other. The ACS collects data from people living in both housing units and GQ facilities.

Group quarters (GQ) population. The number of persons residing in GQ facilities.

Item allocation rates. Allocation is a method of imputation used when values for missing or inconsistent items cannot be derived from the existing response record. In these cases, the imputation must be based on other techniques such as using answers from other people in the household, other responding housing units, or people believed to have similar characteristics. Such donors are reflected in a table referred to as an allocation matrix. The rate is percentage of times this method is used.

Margin of error (MOE). Some ACS products provide an MOE instead of confidence intervals. An MOE is the difference between an estimate and its upper or lower confidence bounds. Confidence bounds can be created by adding the MOE to the estimate (for the upper bound) and subtracting the MOE from the estimate (for the lower bound). All published ACS MOE are based on a 90-percent confidence level.

Multiyear estimates. Three- and five-year estimates based on multiple years of ACS data. Three-year estimates will be published for geographic areas with a population of 20,000 or more. Five-year estimates will be published for all geographic areas down to the census block group level.

Narrative profile. A data product that includes easy-to-read descriptions for a particular geography.

Nonsampling error. Total survey error can be classified into two categories—sampling error and nonsampling error. Nonsampling error includes measurement errors due to interviewers, respondents, instruments, and mode; nonresponse error; coverage error; and processing error.

Period estimates. An estimate based on information collected over a period of time. For ACS the period is either 1 year, 3 years, or 5 years.

Point-in-time estimates. An estimate based on one point in time. The decennial census long-form estimates for Census 2000 were based on information collected as of April 1, 2000.

Population Estimates Program. Official Census Bureau estimates of the population of the United States, states, metropolitan areas, cities and towns, and counties; also official Census Bureau estimates of housing units (HUs).

Public Use Microdata Area (PUMA). An area that defines the extent of territory for which the Census Bureau releases Public Use Microdata Sample (PUMS) records.

Public Use Microdata Sample (PUMS) files. Computerized files that contain a sample of individual records, with identifying information removed, showing the population and housing characteristics of the units, and people included on those forms.

Puerto Rico Community Survey (PRCS). The counterpart to the ACS that is conducted in Puerto Rico.

Quality measures. Statistics that provide information about the quality of the ACS data. The ACS releases four different quality measures with the annual data release: 1) initial sample size and final interviews; 2) coverage rates; 3) response rates, and; 4) item allocation rates for all collected variables. The ACS Quality Measures Web site provides these statistics each year. In addition, the coverage rates are also available for males and females separately.

Reference period. Time interval to which survey responses refer. For example, many ACS questions refer to the day of the interview; others refer to "the past 12 months" or "last week."

Residence rules. The series of rules that define who (if anyone) is considered to be a resident of a sample address for purposes of the survey or census.

Sampling error. Errors that occur because only part of the population is directly contacted. With any sample, differences are likely to exist between the characteristics of the sampled population and the larger group from which the sample was chosen.

Sampling variability. Variation that occurs by chance because a sample is surveyed rather than the entire population.

Selected population profiles. An ACS data product that provides certain characteristics for a specific race or ethnic group (for example, Alaska Natives) or other population subgroup (for example, people aged 60 years and over). This data product is produced directly from the sample microdata (that is, not a derived product).

Single-year estimates. Estimates based on the set of ACS interviews conducted from January through December of a given calendar year. These estimates are published each year for geographic areas with a population of 65,000 or more.

Standard error. The standard error is a measure of the deviation of a sample estimate from the average of all possible samples.

Statistical significance. The determination of whether the difference between two estimates is not likely to be from random chance (sampling error) alone. This determination is based on both the estimates themselves and their standard errors. For ACS data, two estimates are "significantly different at the 90 percent level" if their difference is large enough to infer that there was a less than 10 percent chance that the difference came entirely from random variation.

Subject tables. Data products organized by subject area that present an overview of the information that analysts most often receive requests for from data users.

Summary files. Consist of detailed tables of Census 2000 social, economic, and housing characteristics compiled from a sample of approximately 19 million housing units (about 1 in 6 households) that received the Census 2000 long-form questionnaire.

Thematic maps. Display geographic variation in map format from the geographic ranking tables.

Three-year estimates. Estimates based on 3 years of ACS data. These estimates are meant to reflect the characteristics of a geographic area over the entire 3-year period. These estimates will be published for geographic areas with a population of 20,000 or more.

Appendix 1.

Understanding and Using ACS Single-Year and Multiyear Estimates

What Are Single-Year and Multiyear Estimates?

Understanding Period Estimates

The ACS produces period estimates of socioeconomic and housing characteristics. It is designed to provide estimates that describe the average characteristics of an area over a specific time period. In the case of ACS single-year estimates, the period is the calendar year (e.g., the 2007 ACS covers January through December 2007). In the case of ACS multiyear estimates, the period is either 3 or 5 calendar years (e.g., the 2005–2007 ACS estimates cover January 2005 through December 2007, and the 2006–2010 ACS estimates cover January 2006 through December 2010). The ACS multiyear estimates are similar in many ways to the ACS single-year estimates, however they encompass a longer time period. As discussed later in this appendix, the differences in time periods between single-year and multiyear ACS estimates affect decisions about which set of estimates should be used for a particular analysis.

While one may think of these estimates as representing average characteristics over a single calendar year or multiple calendar years, it must be remembered that the 1-year estimates are not calculated as an average of 12 monthly values and the multiyear estimates are not calculated as the average of either 36 or 60 monthly values. Nor are the multiyear estimates calculated as the average of 3 or 5 single-year estimates. Rather, the ACS collects survey information continuously nearly every day of the year and then aggregates the results over a specific time period—1 year, 3 years, or 5 years. The data collection is spread evenly across the entire period represented so as not to over-represent any particular month or year within the period.

Because ACS estimates provide information about the characteristics of the population and housing for areas over an entire time frame, ACS single-year and multiyear estimates contrast with "point-in-time" estimates, such as those from the decennial census long-form samples or monthly employment estimates from the Current Population Survey (CPS), which are designed to measure characteristics as of a certain date or narrow time period. For example, Census 2000 was designed to measure the characteristics of the population and housing in the United States based upon data collected around April 1, 2000, and thus its data reflect a narrower time frame than ACS data. The monthly CPS collects data for an even narrower time frame, the week containing the 12th of each month.

Implications of Period Estimates

Most areas have consistent population characteristics throughout the calendar year, and their period estimates may not look much different from estimates that would be obtained from a "point-in-time" survey design. However, some areas may experience changes in the estimated characteristics of the population, depending on when in the calendar year measurement occurred. For these areas, the ACS period estimates (even for a single-year) may noticeably differ from "point-in-time" estimates. The impact will be more noticeable in smaller areas where changes such as a factory closing can have a large impact on population characteristics, and in areas with a large physical event such as Hurricane Katrina's impact on the New Orleans area. This logic can be extended to better interpret 3-year and 5-year estimates where the periods involved are much longer. If, over the full period of time (for example, 36 months) there have been major or consistent changes in certain population or housing characteristics for an area, a period estimate for that area could differ markedly from estimates based on a "point-in-time" survey.

An extreme illustration of how the single-year estimate could differ from a "point-in-time" estimate within the year is provided in Table 1. Imagine a town on the Gulf of Mexico whose population is dominated by retirees in the winter months and by locals in the summer months. While the percentage of the population in the labor force across the entire year is about 45 percent (similar in concept to a period estimate), a "point-in-time" estimate for any particular month would yield estimates ranging from 20 percent to 60 percent.

Table 1. **Percent in Labor Force—Winter Village**

Month											
Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
20	20	40	60	60	60	60	60	60	50	30	20

Source: U.S. Census Bureau, Artificial Data.

The important thing to keep in mind is that ACS single-year estimates describe the population and characteristics of an area for the full year, not for any specific day or period within the year, while ACS multiyear estimates describe the population and characteristics of an area for the full 3- or 5-year period, not for any specific day, period, or year within the multiyear time period.

Release of Single-Year and Multiyear Estimates

The Census Bureau has released single-year estimates from the full ACS sample beginning with data from the 2005 ACS. ACS 1-year estimates are published annually for geographic areas with populations of 65,000 or more. Beginning in 2008 and encompassing 2005–2007, the Census Bureau will publish annual ACS 3-year estimates for geographic areas with populations of 20,000 or more. Beginning in 2010, the Census Bureau will release ACS 5-year estimates (encompassing 2005–2009) for all geographic areas —down to the tract and block group levels. While eventually all three data series will be available each year, the ACS must collect 5 years of sample before that final set of estimates can be released. This means that in 2008 only 1-year and 3-year estimates are available for use, which means that data are only available for areas with populations of 20,000 and greater.

New issues will arise when multiple sets of multiyear estimates are released. The multiyear estimates released in consecutive years consist mostly of overlapping years and shared data. As shown in Table 2, consecutive 3-year estimates contain 2 years of overlapping coverage (for example, the 2005–2007 ACS estimates share 2006 and 2007 sample data with the 2006–2008 ACS estimates) and consecutive 5-year estimates contain 4 years of overlapping coverage.

Table 2. **Sets of Sample Cases Used in Producing ACS Multiyear Estimates**

Type of estimate	Year of Data Release				
	2008	2009	2010	2011	2012
	Years of Data Collection				
3-year estimates	2005–2007	2006–2008	2007–2009	2008–2010	2009–2011
5-year estimates	Not Available	Not Available	2005–2009	2006–2010	2007–2011

Source: U.S. Census Bureau.

Differences Between Single-Year and Multiyear ACS Estimates

Currency

Single-year estimates provide more current information about areas that have changing population and/or housing characteristics because they are based on the most current data—data from the past year. In contrast, multiyear estimates provide less current information because they are based on both data from the previous year and data that are 2 and 3 years old. As noted earlier, for many areas with minimal change taking place, using the "less current" sample used to produce the multiyear estimates may not have a substantial influence on the estimates. However, in areas experiencing major changes over a given time period, the multiyear estimates may be quite different from the single-year estimates for any of the individual years. Single-year and multiyear estimates are not expected to be the same because they are based on data from two different time periods. This will be true even if the ACS single year is the midyear of the ACS multiyear period (e.g., 2007 single year, 2006–2008 multiyear).

For example, suppose an area has a growing Hispanic population and is interested in measuring the percent of the population who speak Spanish at home. Table 3 shows a hypothetical set of 1-year and 3-year estimates. Comparing data by release year shows that for an area such as this with steady growth, the 3-year estimates for a period are seen to lag behind the estimates for the individual years.

Reliability

Multiyear estimates are based on larger sample sizes and will therefore be more reliable. The 3-year estimates are based on three times as many sample cases as the 1-year estimates. For some characteristics this increased sample is needed for the estimates to be reliable enough for use in certain applications. For other characteristics the increased sample may not be necessary.

Table 3. **Example of Differences in Single- and Multiyear Estimates—Percent of Population Who Speak Spanish at Home**

Year of data release	1-year estimates		3-year estimates	
	Time period	Estimate	Time period	Estimate
2003	2002	13.7	2000–2002	13.4
2004	2003	15.1	2001–2003	14.4
2005	2004	15.9	2002–2004	14.9
2006	2005	16.8	2003–2005	15.9

Source: U.S. Census Bureau, Artificial Data.

Multiyear estimates are the only type of estimates available for geographic areas with populations of less than 65,000. Users may think that they only need to use multiyear estimates when they are working with small areas, but this isn't the case. Estimates for large geographic areas benefit from the increased sample resulting in more precise estimates of population and housing characteristics, especially for subpopulations within those areas.

In addition, users may determine that they want to use single-year estimates, despite their reduced reliability, as building blocks to produce estimates for meaningful higher levels of geography. These aggregations will similarly benefit from the increased sample sizes and gain reliability.

Deciding Which ACS Estimate to Use

Three primary uses of ACS estimates are to understand the characteristics of the population of an area for local planning needs, make comparisons across areas, and assess change over time in an area. Local planning could include making local decisions such as where to locate schools or hospitals, determining the need for services or new businesses, and carrying out transportation or other infrastructure analysis. In the past, decennial census sample data provided the most comprehensive information. However, the currency of those data suffered through the intercensal period, and the ability to assess change over time was limited. ACS estimates greatly improve the currency of data for understanding the characteristics of housing and population and enhance the ability to assess change over time.

Several key factors can guide users trying to decide whether to use single-year or multiyear ACS estimates for areas where both are available: intended use of the estimates, precision of the estimates, and currency of the estimates. All of these factors, along with an understanding of the differences between single-year and multiyear ACS estimates, should be taken into consideration when deciding which set of estimates to use.

Understanding Characteristics

For users interested in obtaining estimates for small geographic areas, multiyear ACS estimates will be the only option. For the very smallest of these areas (less than 20,000 population), the only option will be to use the 5-year ACS estimates. Users have a choice of two sets of multiyear estimates when analyzing data for small geographic areas with populations of at least 20,000. Both 3-year and 5-year ACS estimates will be available. Only the largest areas with populations of 65,000 and more receive all three data series.

The key trade-off to be made in deciding whether to use single-year or multiyear estimates is between currency and precision. In general, the single-year estimates are preferred, as they will be more relevant to the current conditions. However, the user must take into account the level of uncertainty present in the single-year estimates, which may be large for small subpopulation groups and rare characteristics. While single-year estimates offer more current estimates, they also have higher sampling variability. One measure, the coefficient of variation (CV) can help you determine the fitness for use of a single-year estimate in order to assess if you should opt instead to use the multiyear estimate (or if you should use a 5-year estimate rather than a 3-year estimate). The CV is calculated as the ratio of the standard error of the estimate to the estimate, times 100. A single-year estimate with a small CV is usually preferable to a multiyear estimate as it is more up to date. However, multiyear estimates are an alternative option when a single-year estimate has an unacceptably high CV.

Table 4 illustrates how to assess the reliability of 1-year estimates in order to determine if they should be used. The table shows the percentage of households where Spanish is spoken at home for ACS test counties Broward, Florida, and Lake, Illinois. The standard errors and CVs associated with those estimates are also shown.

In this illustration, the CV for the single-year estimate in Broward County is 1.0 percent (0.2/19.9) and in Lake County is 1.3 percent (0.2/15.9). Both are sufficiently small to allow use of the more current single-year estimates.

Single-year estimates for small subpopulations (e.g., families with a female householder, no husband, and related children less than 18 years) will typically have larger CVs. In general, multiyear estimates are preferable to single-year estimates when looking at estimates for small subpopulations.

For example, consider Sevier County, Tennessee, which had an estimated population of 76,632 in 2004 according to the Population Estimates Program. This population is larger than the Census Bureau's 65,000-population requirement for publishing 1-year estimates. However, many subpopulations within this geographic area will be much smaller than 65,000. Table 5 shows an estimated 21,881 families in Sevier County based on the 2000–2004 multiyear estimate; but only 1,883 families with a female householder, no husband present, with related children under 18 years. Not surprisingly, the 2004 ACS estimate of the poverty rate (38.3 percent) for this subpopulation has a large standard error (SE) of 13.0 percentage points. Using this information we can determine that the CV is 33.9 percent (13.0/38.3).

For such small subpopulations, users obtain more precision using the 3-year or 5-year estimate. In this example, the 5-year estimate of 40.2 percent has an SE of 4.9 percentage points that yields a CV of 12.2 percent (4.9/40.2), and the 3-year estimate of 40.4 percent has an SE of 6.8 percentage points which yields a CV of 16.8 percent (6.8/40.4).

Users should think of the CV associated with an estimate as a way to assess "fitness for use." The CV threshold that an individual should use will vary based on the application. In practice there will be many estimates with CVs over desirable levels. A general guideline when working with ACS estimates is that, while data are available at low geographic levels, in situations where the CVs for these estimates are high, the reliability of the estimates will be improved by aggregating such estimates to a higher geographic level. Similarly, collapsing characteristic detail (for example, combining individual age categories into broader categories) can allow you to improve the reliability of the aggregate estimate, bringing the CVs to a more acceptable level.

Table 4. **Example of How to Assess the Reliability of Estimates—Percent of Population Who Speak Spanish at Home**

County	Estimate	Standard error	Coefficient of variation
Broward County, FL	19.9	0.2	1.0
Lake County, IL	15.9	0.2	1.3

Source: U.S. Census Bureau, Multiyear Estimates Study data.

Table 5. **Percent in Poverty by Family Type for Sevier County, TN**

	2000–2004	2000–2004		2002–2004		2004	
	Total family type	Pct. in poverty	SE	Pct. in poverty	SE	Pct. in poverty	SE
All families	21,881	9.5	0.8	9.7	1.3	10.0	2.3
With related children under 18 years	9,067	15.3	1.5	16.5	2.4	17.8	4.5
Married-couple families	17,320	5.8	0.7	5.4	0.9	7.9	2.0
With related children under 18 years	6,633	7.7	1.2	7.3	1.7	12.1	3.9
Families with female householder, no husband	3,433	27.2	3.0	26.7	4.8	19.0	7.2
With related children under 18 years	1,883	40.2	4.9	40.4	6.8	38.3	13.0

Source: U.S. Census Bureau, Multiyear Estimates Study data.

Making Comparisons

Often users want to compare the characteristics of one area to those of another area. These comparisons can be in the form of rankings or of specific pairs of comparisons. Whenever you want to make a comparison between two different geographic areas you need to take the type of estimate into account. It is important that comparisons be made within the same estimate type. That is, 1-year estimates should only be compared with other 1-year estimates, 3-year estimates should only be compared with other 3-year estimates, and 5-year estimates should only be compared with other 5-year estimates.

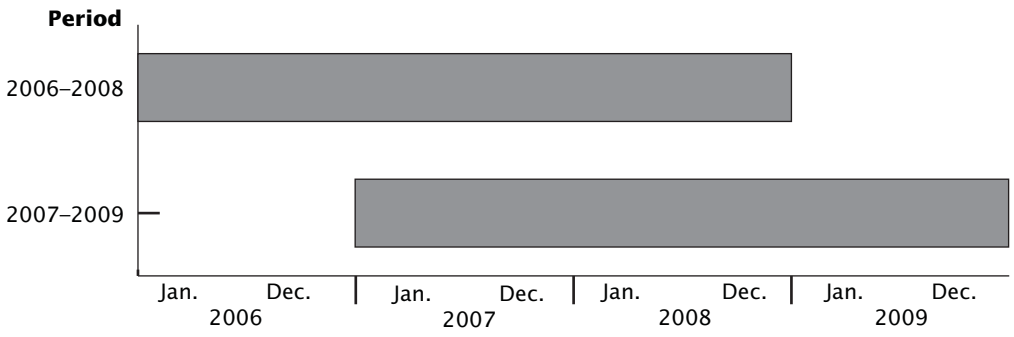
You certainly can compare characteristics for areas with populations of 30,000 to areas with populations of 100,000 but you should use the data set that they have in common. In this example you could use the 3-year or the 5-year estimates because they are available for areas of 30,000 and areas of 100,000.

Assessing Change

Users are encouraged to make comparisons between sequential single-year estimates. Specific guidance on making these comparisons and interpreting the results are provided in Appendix 4. Starting with the 2007 ACS, a new data product called the comparison profile will do much of the statistical work to identify statistically significant differences between the 2007 ACS and the 2006 ACS.

As noted earlier, caution is needed when using multiyear estimates for estimating year-to-year change in a particular characteristic. This is because roughly two-thirds of the data in a 3-year estimate overlap with the data in the next year's 3-year estimate (the overlap is roughly four-fifths for 5-year estimates). Thus, as shown in Figure 1, when comparing 2006–2008 3-year estimates with 2007–2009 3-year estimates, the differences in overlapping multiyear estimates are driven by differences in the nonoverlapping years. A data user interested in comparing 2009 with 2008 will not be able to isolate those differences using these two successive 3-year estimates. Figure 1 shows that the difference in these two estimates describes the difference between 2009 and 2006. While the interpretation of this difference is difficult, these comparisons can be made with caution. Users who are interested in comparing overlapping multiyear period estimates should refer to Appendix 4 for more information.

Figure 1. **Data Collection Periods for 3-Year Estimates**

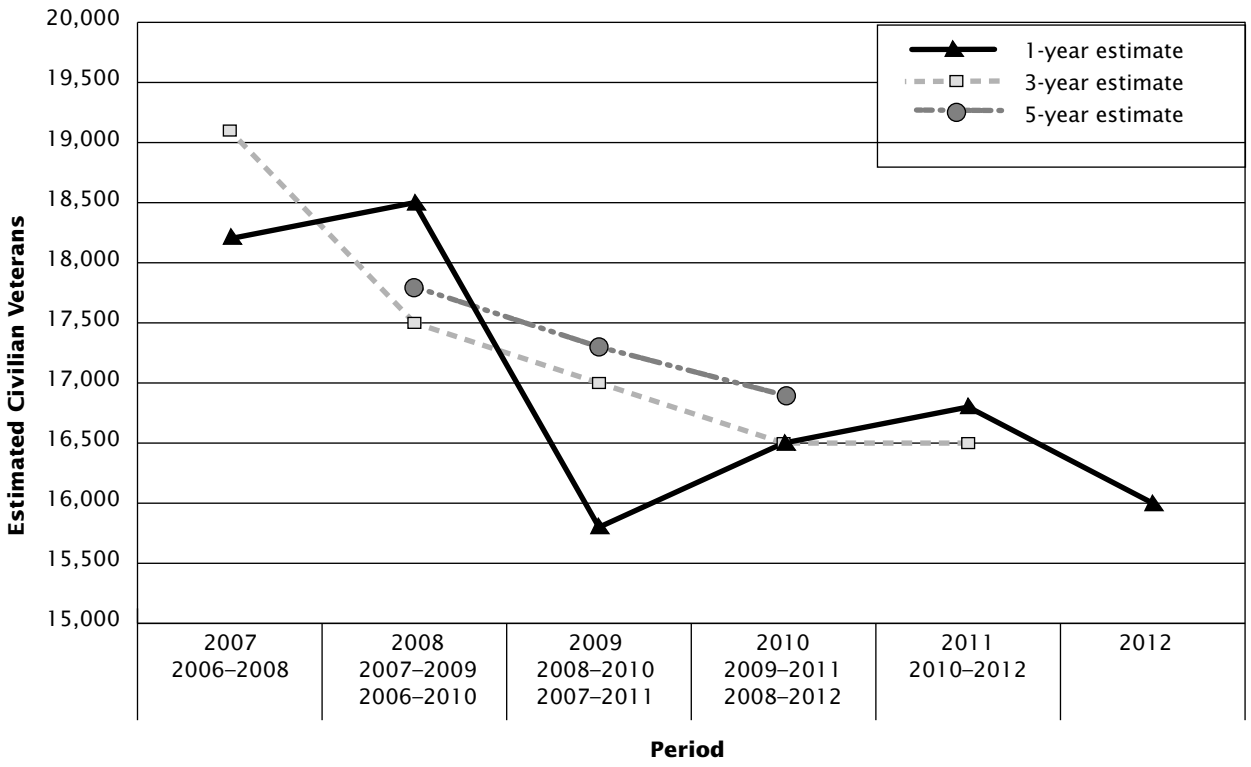


Source: U.S. Census Bureau.

Variability in single-year estimates for smaller areas (near the 65,000-publication threshold) and small subgroups within even large areas may limit the ability to examine trends. For example, single-year estimates for a characteristic with a high CV vary from year to year because of sampling variation obscuring an underlying trend. In this case, multiyear estimates may be useful for assessing an underlying, long-term trend. Here again, however, it must be recognized that because the multiyear estimates have an inherent smoothing, they will tend to mask rapidly developing changes. Plotting the multiyear estimates as representing the middle year is a useful tool to illustrate the smoothing effect of the multiyear weighting methodology. It also can be used to assess the "lagging effect" in the multiyear estimates. As a general rule, users should not consider a multiyear estimate as a proxy for the middle year of the period. However, this could be the case under some specific conditions, as is the case when an area is experiencing growth in a linear trend.

As Figure 2 shows, while the single-year estimates fluctuate from year to year without showing a smooth trend, the multiyear estimates, which incorporate data from multiple years, evidence a much smoother trend across time.

Figure 2. **Civilian Veterans, County X Single-Year, Multiyear Estimates**



Source: U.S. Census Bureau. Based on data from the Multiyear Estimates Study.

Summary of Guidelines

Multiyear estimates should, in general, be used when single-year estimates have large CVs or when the precision of the estimates is more important than the currency of the data. Multiyear estimates should also be used when analyzing data for smaller geographies and smaller populations in larger geographies. Multiyear estimates are also of value when examining change over nonoverlapping time periods and for smoothing data trends over time.

Single-year estimates should, in general, be used for larger geographies and populations when currency is more important than the precision of the estimates. Single-year estimates should be used to examine year-to-year change for estimates with small CVs. Given the availability of a single-year estimate, calculating the CV provides useful information to determine if the single-year estimate should be used. For areas believed to be experiencing rapid changes in a characteristic, single-year estimates should generally be used rather than multiyear estimates as long as the CV for the single-year estimate is reasonable for the specific usage.

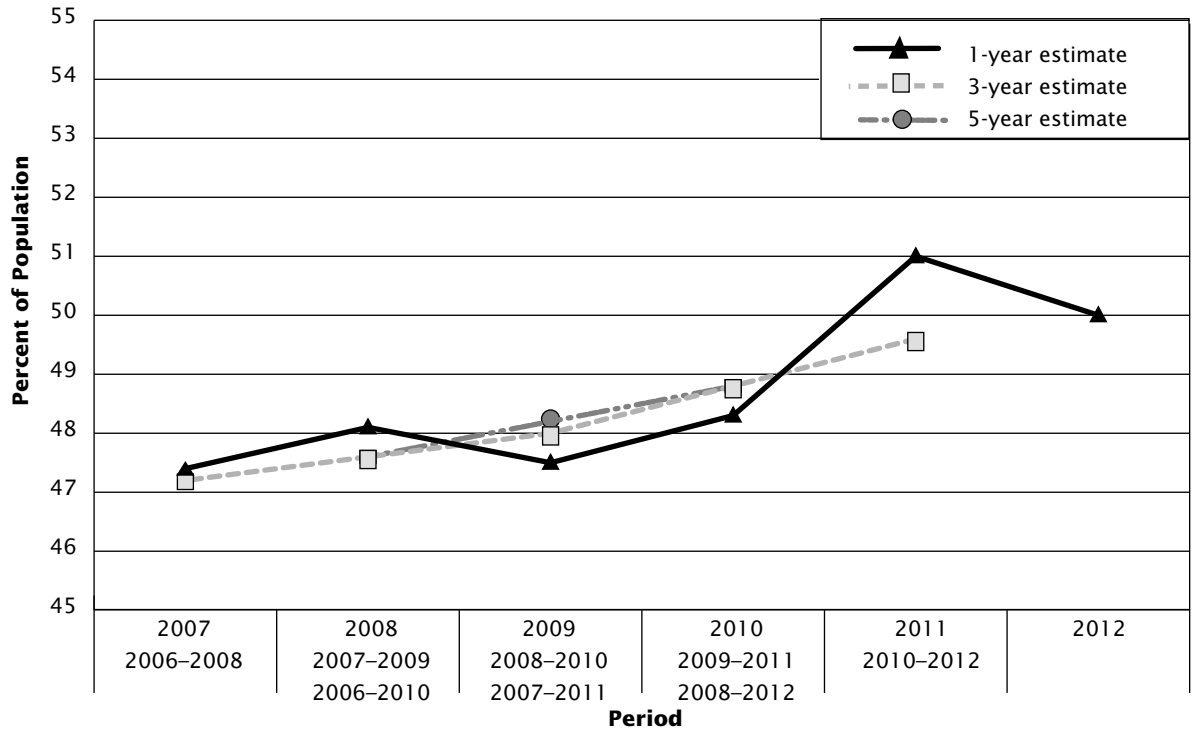
Local area variations may occur due to rapidly occurring changes. As discussed previously, multiyear estimates will tend to be insensitive to such changes when they first occur. Single-year estimates, if associated with sufficiently small CVs, can be very valuable in identifying and studying such phenomena. Graphing trends for such areas using single-year, 3-year, and 5-year estimates can take advantage of the strengths of each set of estimates while using other estimates to compensate for the limitations of each set.

Figure 3 provides an illustration of how the various ACS estimates could be graphed together to better understand local area variations.

The multiyear estimates provide a smoothing of the upward trend and likely provide a better portrayal of the change in proportion over time. Correspondingly, as the data used for single-year estimates will be used in the multiyear estimates, an observed change in the upward direction for consecutive single-year estimates could provide an early indicator of changes in the underlying trend that will be seen when the multiyear estimates encompassing the single years become available.

We hope that you will follow these guidelines to determine when to use single-year versus multiyear estimates, taking into account the intended use and CV associated with the estimate. The Census Bureau encourages you to include the MOE along with the estimate when producing reports, in order to provide the reader with information concerning the uncertainty associated with the estimate.

Figure 3. **Proportion of Population With Bachelor's Degree or Higher, City X Single-Year, Multiyear Estimates**



Source: U.S. Census Bureau. Based on data from the Multiyear Estimates Study.

Appendix 2.
Differences Between ACS and Decennial Census Sample Data

There are many similarities between the methods used in the decennial census sample and the ACS. Both the ACS and the decennial census sample data are based on information from a sample of the population. The data from the Census 2000 sample of about one-sixth of the population were collected using a "long-form" questionnaire, whose content was the model for the ACS. While some differences exist in the specific Census 2000 question wording and that of the ACS, most questions are identical or nearly identical. Differences in the design and implementation of the two surveys are noted below with references provided to a series of evaluation studies that assess the degree to which these differences are likely to impact the estimates. As noted in Appendix 1, the ACS produces period estimates and these estimates do not measure characteristics for the same time frame as the decennial census estimates, which are interpreted to be a snapshot of April 1 of the census year. Additional differences are described below.

Residence Rules, Reference Periods, and Definitions

The fundamentally different purposes of the ACS and the census, and their timing, led to important differences in the choice of data collection methods. For example, the residence rules for a census or survey determine the sample unit's occupancy status and household membership. Defining the rules in a dissimilar way can affect those two very important estimates. The Census 2000 residence rules, which determined where people should be counted, were based on the principle of "usual residence" on April 1, 2000, in keeping with the focus of the census on the requirements of congressional apportionment and state redistricting. To accomplish this the decennial census attempts to restrict and determine a principal place of residence on one specific date for everyone enumerated. The ACS residence rules are based on a "current residence" concept since data are collected continuously throughout the entire year with responses provided relative to the continuously changing survey interview dates. This method is consistent with the goal that the ACS produce estimates that reflect annual averages of the characteristics of all areas.

Estimates produced by the ACS are not measuring exactly what decennial samples have been measuring. The ACS yearly samples, spread over 12 months, collect information that is anchored to the day on which the sampled unit was interviewed, whether it is the day that a mail questionnaire is completed or the day that an interview is conducted by telephone or personal visit. Individual questions with time references such as "last week" or "the last 12 months" all begin the reference period as of this interview date. Even the information on types and amounts of income refers to the 12 months prior to the day the question is answered. ACS interviews are conducted just about every day of the year, and all of the estimates that the survey releases are considered to be averages for a specific time period. The 1-year estimates reflect the full calendar year; 3-year and 5-year estimates reflect the full 36- or 60-month period.

Most decennial census sample estimates are anchored in this same way to the date of enumeration. The most obvious difference between the ACS and the census is the overall time frame in which they are conducted. The census enumeration time period is less than half the time period used to collect data for each single-year ACS estimate. But a more important difference is that the distribution of census enumeration dates are highly clustered in March and April (when most census mail returns were received) with additional, smaller clusters seen in May and June (when nonresponse follow-up activities took place).

This means that the data from the decennial census tend to describe the characteristics of the population and housing in the March through June time period (with an overrepresentation of March/April) while the ACS characteristics describe the characteristics nearly every day over the full calendar year.

Census Bureau analysts have compared sample estimates from Census 2000 with 1-year ACS estimates based on data collected in 2000 and 3-year ACS estimates based on data collected in 1999–2001 in selected counties. A series of reports summarize their findings and can be found at <http://www.census.gov/acs/www/AdvMeth/Reports.htm>. In general, ACS estimates were found to be quite similar to those produced from decennial census data.

More on Residence Rules

Residence rules determine which individuals are considered to be residents of a particular housing unit or group quarters. While many people have definite ties to a single housing unit or group quarters, some people may stay in different places for significant periods of time over the course of the year. For example, migrant workers move with crop seasons and do not live in any one location for the entire year. Differences in treatment of these populations in the census and ACS can lead to differences in estimates of the characteristics of some areas.

For the past several censuses, decennial census residence rules were designed to produce an accurate

count of the population as of Census Day, April 1, while the ACS residence rules were designed to collect representative information to produce annual average estimates of the characteristics of all kinds of areas. When interviewing the population living in housing units, the decennial census uses a "usual residence" rule to enumerate people at the place where they live or stay most of the time as of April 1. The ACS uses a "current residence" rule to interview people who are currently living or staying in the sample housing unit as long as their stay at that address will exceed 2 months. The residence rules governing the census enumerations of people in group quarters depend on the type of group quarter and where permitted, whether people claim a "usual residence" elsewhere. The ACS applies a straight de facto residence rule to every type of group quarter. Everyone living or staying in a group quarter on the day it is visited by an ACS interviewer is eligible to be sampled and interviewed for the survey. Further information on residence rules can be found at <http://www.census.gov/acs/www/AdvMeth/CollProc/CollProc1.htm>.

The differences in the ACS and census data as a consequence of the different residence rules are most likely minimal for most areas and most characteristics. However, for certain segments of the population the usual and current residence concepts could result in different residence decisions. Appreciable differences may occur in areas where large proportions of the total population spend several months of the year in what would not be considered their residence under decennial census rules. In particular, data for areas that include large beach, lake, or mountain vacation areas may differ appreciably between the census and the ACS if populations live there for more than 2 months.

More on Reference Periods

The decennial census centers its count and its age distributions on a reference date of April 1, the assumption being that the remaining basic demographic questions also reflect that date, regardless of whether the enumeration is conducted by mail in March or by a field follow-up in July. However, nearly all questions are anchored to the date the interview is provided. Questions with their own reference periods, such as "last week," are referring to the week prior to the interview date. The idea that all census data reflect the characteristics as of April 1 is a myth. Decennial census samples actually provide estimates based on aggregated data reflecting the entire period of decennial data collection, and are greatly influenced by delivery dates of mail questionnaires, success of mail response, and data collection schedules for nonresponse follow-up. The ACS reference periods are, in many ways, similar to those in the census in that they reflect the circumstances on the day the data are collected and the individual reference periods of questions relative to that date. However, the ACS estimates represent the average characteristics over a full year (or sets of years), a different time, and reference period than the census.

Some specific differences in reference periods between the ACS and the decennial census are described below. Users should consider the potential impact these different reference periods could have on distributions when comparing ACS estimates with Census 2000.

Those who are interested in more information about differences in reference periods should refer to the Census Bureau's guidance on comparisons that contrasts for each question the specific reference periods used in Census 2000 with those used in the ACS. See <http://www.census.gov/acs/www/UseData/compACS.htm>.

Income Data

To estimate annual income, the Census 2000 long-form sample used the calendar year prior to Census Day as the reference period, and the ACS uses the 12 months prior to the interview date as the reference period. Thus, while Census 2000 collected income information for calendar year 1999, the ACS collects income information for the 12 months preceding the interview date. The responses are a mixture of 12 reference periods ranging from, in the case of the 2006 ACS single-year estimates, the full calendar year 2005 through November 2006. The ACS income responses for each of these reference periods are individually inflation-adjusted to represent dollar values for the ACS collection year.

School Enrollment

The school enrollment question on the ACS asks if a person had "at any time in the last 3 months attended a school or college." A consistent 3-month reference period is used for all interviews. In contrast, Census 2000 asked if a person had "at any time since February 1 attended a school or college." Since Census 2000 data were collected from mid-March to late-August, the reference period could have been as short as about 6 weeks or as long as 7 months.

Utility Costs

The reference periods for two utility cost questions—gas and electricity—differ between Census 2000 and the ACS. The census asked for annual costs, while the ACS asks for the utility costs in the previous month.

Definitions

Some data items were collected by both the ACS and the Census 2000 long form with slightly different definitions that could affect the comparability of the estimates for these items. One example is annual costs for a mobile home. Census 2000 included installment loan costs in

the total annual costs but the ACS does not. In this example, the ACS could be expected to yield smaller estimates than Census 2000.

Implementation

While differences discussed above were a part of the census and survey design objectives, other differences observed between ACS and census results were not by design, but due to nonsampling error—differences related to how well the surveys were conducted. Appendix 6 explains nonsampling error in more detail.

The ACS and the census experience different levels and types of coverage error, different levels and treatment of unit and item nonresponse, and different instances of measurement and processing error. Both Census 2000 and the ACS had similar high levels of survey coverage and low levels of unit nonresponse. Higher levels of unit nonresponse were found in the nonresponse follow-up stage of Census 2000. Higher item nonresponse rates were also found in Census 2000. Please see <http://www.census.gov/acs/www/AdvMeth/Reports.htm> for detailed comparisons of these measures of survey quality.

Appendix 3.
Measures of Sampling Error

All survey and census estimates include some amount of error. Estimates generated from sample survey data have uncertainty associated with them due to their being based on a sample of the population rather than the full population. This uncertainty, referred to as sampling error, means that the estimates derived from a sample survey will likely differ from the values that would have been obtained if the entire population had been included in the survey, as well as from values that would have been obtained had a different set of sample units been selected. All other forms of error are called nonsampling error and are discussed in greater detail in Appendix 6.

Sampling error can be expressed quantitatively in various ways, four of which are presented in this appendix—standard error, margin of error, confidence interval, and coefficient of variation. As the ACS estimates are based on a sample survey of the U.S. population, information about the sampling error associated with the estimates must be taken into account when analyzing individual estimates or comparing pairs of estimates across areas, population subgroups, or time periods. The information in this appendix describes each of these sampling error measures, explaining how they differ and how each should be used. It is intended to assist the user with analysis and interpretation of ACS estimates. Also included are instructions on how to compute margins of error for user-derived estimates.

Sampling Error Measures and Their Derivations

Standard Errors

A standard error (SE) measures the variability of an estimate due to sampling. Estimates derived from a sample (such as estimates from the ACS or the decennial census long form) will generally not equal the population value, as not all members of the population were measured in the survey. The SE provides a quantitative measure of the extent to which an estimate derived from the sample survey can be expected to deviate from this population value. It is the foundational measure from which other sampling error measures are derived. The SE is also used when comparing estimates to determine whether the differences between the estimates can be said to be statistically significant.

A very basic example of the standard error is a population of three units, with values of 1, 2, and 3. The average value for this population is 2. If a simple random sample of size two were selected from this population, the estimates of the average value would be 1.5 (units with values of 1 and 2 selected), 2 (units with values of 1 and 3 selected), or 2.5 (units with values of 2 and 3 selected). In this simple example, two of the three samples yield estimates that do not equal the population value (although the average of the estimates across all possible samples do equal the population value). The standard error would provide an indication of the extent of this variation.

The SE for an estimate depends upon the underlying variability in the population for the characteristic and the sample size used for the survey. In general, the larger the sample size, the smaller the standard error of the estimates produced from the sample. This relationship between sample size and SE is the reason ACS estimates for less populous areas are only published using multiple years of data: to take advantage of the larger sample size that results from aggregating data from more than one year.

Margins of Error

A margin of error (MOE) describes the precision of the estimate at a given level of confidence. The confidence level associated with the MOE indicates the likelihood that the sample estimate is within a certain distance (the MOE) from the population value. Confidence levels of 90 percent, 95 percent, and 99 percent are commonly used in practice to lessen the risk associated with an incorrect inference. The MOE provides a concise measure of the precision of the sample estimate in a table and is easily used to construct confidence intervals and test for statistical significance.

The Census Bureau statistical standard for published data is to use a 90-percent confidence level. Thus, the MOEs published with the ACS estimates correspond to a 90-percent confidence level. However, users may want to use other confidence levels, such as 95 percent or 99 percent. The choice of confidence level is usually a matter of preference, balancing risk for the specific application, as a 90-percent confidence level implies a 10 percent chance of an incorrect inference, in contrast with a 1 percent chance if using a 99-percent confidence level. Thus, if the impact of an incorrect conclusion is substantial, the user should consider increasing the confidence level.

One commonly experienced situation where use of a 95 percent or 99 percent MOE would be preferred is when conducting a number of tests to find differences between sample estimates. For example, if one were conducting comparisons between male and female incomes for each of 100 counties in a state, using a 90-percent confidence level would imply that 10 of the comparisons would be expected to be found significant even if no differences actually existed. Using a 99-percent confidence level would reduce the likelihood of this kind of false inference.

Calculating Margins of Error for Alternative Confidence Levels

If you want to use an MOE corresponding to a confidence level other than 90 percent, the published MOE can easily be converted by multiplying the published MOE by an adjustment factor. If the desired confidence level is 95 percent, then the factor is equal to 1.960/1.645.[1] If the desired confidence level is 99 percent, then the factor is equal to 2.576/1.645.

Conversion of the published ACS MOE to the MOE for a different confidence level can be expressed as

$$MOE_{95} = \frac{1.960}{1.645} MOE_{ACS}$$

$$MOE_{99} = \frac{2.576}{1.645} MOE_{ACS}$$

where MOE_{ACS} is the ACS published 90 percent MOE for the estimate.

Factors Associated With Margins of Error for Commonly Used Confidence Levels
90 Percent: 1.645 95 Percent: 1.960 99 Percent: 2.576
Census Bureau standard for published MOE is 90 percent.

For example, the ACS published MOE for the 2006 ACS estimated number of civilian veterans in the state of Virginia is ±12,357. The MOE corresponding to a 95-percent confidence level would be derived as follows:

$$MOE_{95} = \frac{1.960}{1.645}(\pm 12{,}357) = \pm 14{,}723$$

Deriving the Standard Error From the MOE

When conducting exact tests of significance (as discussed in Appendix 4) or calculating the CV for an estimate, the SEs of the estimates are needed. To derive the SE, simply divide the positive value of the published MOE by 1.645.[2]

Derivation of SEs can thus be expressed as

$$SE = \frac{MOE_{ACS}}{1.645}$$

where MOE_{ACS} is the positive value of the ACS published MOE for the estimate.

For example, the ACS published MOE for estimated number of civilian veterans in the state of Virginia from the 2006 ACS is ±12,357. The SE for the estimate would be derived as

$$SE = \frac{12{,}357}{1.645} = 7{,}512$$

Confidence Intervals

A confidence interval (CI) is a range that is expected to contain the average value of the characteristic that would result over all possible samples with a known probability. This probability is called the "level of confidence" or "confidence level." CIs are useful when graphing estimates to display their sampling variabilites. The sample estimate and its MOE are used to construct the CI.

Constructing a Confidence Interval From a Margin of Error

To construct a CI at the 90-percent confidence level, the published MOE is used. The CI boundaries are determined by adding to and subtracting from a sample estimate, the estimate's MOE.

For example, if an estimate of 20,000 had an MOE at the 90-percent confidence level of ±1,645, the CI would range from 18,355 (20,000 – 1,645) to 21,645 (20,000 + 1,645).

For CIs at the 95-percent or 99-percent confidence level, the appropriate MOE must first be derived as explained previously.

Construction of the lower and upper bounds for the CI can be expressed as

$$L_{CL} = \hat{X} - MOE_{CL}$$

$$U_{CL} = \hat{X} + MOE_{CL}$$

where $\hat{X}$ is the ACS estimate and

MOE_{CL} is the positive value of the MOE for the estimate at the desired confidence level.

The CI can thus be expressed as the range

$$CI_{CL} = (L_{CL}, U_{CL}).$$[3]

[1] The value 1.65 must be used for ACS single-year estimates for 2005 or earlier, as that was the value used to derive the published margin of error from the standard error in those years.

[2] If working with ACS 1-year estimates for 2005 or earlier, use the value 1.65 rather than 1.645 in the adjustment factor.

[3] Users are cautioned to consider logical boundaries when creating confidence intervals from the margins of error. For example, a small population estimate may have a calculated lower bound less than zero. A negative number of persons doesn't make sense, so the lower bound should be set to zero instead.

For example, to construct a CI at the 95-percent confidence level for the number of civilian veterans in the state of Virginia in 2006, one would use the 2006 estimate (771,782) and the corresponding MOE at the 95-percent confidence level derived above (±14,723).

$$L_{95} = 771{,}782 - 14{,}723 = 757{,}059$$
$$U_{95} = 771{,}782 + 14{,}723 = 786{,}505$$

The 95-percent CI can thus be expressed as the range 757,059 to 786,505.

The CI is also useful when graphing estimates, to show the extent of sampling error present in the estimates, and for visually comparing estimates. For example, given the MOE at the 90-percent confidence level used in constructing the CI above, the user could be 90 percent certain that the value for the population was between 18,355 and 21,645. This CI can be represented visually as

Coefficients of Variation

A coefficient of variation (CV) provides a measure of the relative amount of sampling error that is associated with a sample estimate. The CV is calculated as the ratio of the SE for an estimate to the estimate itself and is usually expressed as a percent. It is a useful barometer of the stability, and thus the usability of a sample estimate. It can also help a user decide whether a single-year or multiyear estimate should be used for analysis. The method for obtaining the SE for an estimate was described earlier.

The CV is a function of the overall sample size and the size of the population of interest. In general, as the estimation period increases, the sample size increases and therefore the size of the CV decreases. A small CV indicates that the sampling error is small relative to the estimate, and thus the user can be more confident that the estimate is close to the population value. In some applications a small CV for an estimate is desirable and use of a multiyear estimate will therefore be preferable to the use of a 1-year estimate that doesn't meet this desired level of precision.

For example, if an estimate of 20,000 had an SE of 1,000, then the CV for the estimate would be 5 percent ([1,000 /20,000] x 100). In terms of usability, the estimate is very reliable. If the CV was noticeably larger, the usability of the estimate could be greatly diminished.

While it is true that estimates with high CVs have important limitations, they can still be valuable as building blocks to develop estimates for higher levels of aggregation. Combining estimates across geographic areas or collapsing characteristic detail can improve the reliability of those estimates as evidenced by reductions in the CVs.

Calculating Coefficients of Variation From Standard Errors

The CV can be expressed as

$$CV = \frac{SE}{\hat{X}} \times 100$$

where $\hat{X}$ is the ACS estimate and SE is the derived SE for the ACS estimate.

For example, to determine the CV for the estimated number of civilian veterans in the state of Virginia in 2006, one would use the 2006 estimate (771,782), and the SE derived previously (7,512).

$$CV = \frac{7{,}512}{771{,}782} \times 100 = 0.1\%$$

This means that the amount of sampling error present in the estimate is only one-tenth of 1 percent the size of the estimate.

The text box below summarizes the formulas used when deriving alternative sampling error measures from the margin or error published with ACS estimates.

Deriving Sampling Error Measures From Published MOE

Margin Error (MOE) for Alternate Confidence Levels

$$MOE_{95} = \frac{1.960}{1.645} MOE_{ACS}$$
$$MOE_{99} = \frac{2.576}{1.645} MOE_{ACS}$$

Standard Error (SE)

$$SE = \frac{MOE_{ACS}}{1.645}$$

Confidence Interval (CI)

$$CI_{CL} = \left(\hat{X} - MOE_{CL}, \hat{X} + MOE_{CL}\right)$$

Coefficient of Variation (CV)

$$CV = \frac{SE}{\hat{X}} \times 100$$

Calculating Margins of Error for Derived Estimates

One of the benefits of being familiar with ACS data is the ability to develop unique estimates called derived estimates. These derived estimates are usually based on aggregating estimates across geographic areas or population subgroups for which combined estimates are not published in American FactFinder (AFF) tables (e.g., aggregate estimates for a three-county area or for four age groups not collapsed).

ACS tabulations provided through AFF contain the associated confidence intervals (pre-2005) or margins of error (MOEs) (2005 and later) at the 90-percent confidence level. However, when derived estimates are generated (e.g., aggregated estimates, proportions, or ratios not available in AFF), the user must calculate the MOE for these derived estimates. The MOE helps protect against misinterpreting small or nonexistent differences as meaningful.

MOEs calculated based on information provided in AFF for the components of the derived estimates will be at the 90-percent confidence level. If an MOE with a confidence level other than 90 percent is desired, the user should first calculate the MOE as instructed below and then convert the results to an MOE for the desired confidence level as described earlier in this appendix.

Calculating MOEs for Aggregated Count Data

To calculate the MOE for aggregated count data:

1) Obtain the MOE of each component estimate.
2) Square the MOE of each component estimate.
3) Sum the squared MOEs.
4) Take the square root of the sum of the squared MOEs.

The result is the MOE for the aggregated count. Algebraically, the MOE for the aggregated count is calculated as:

$$MOE_{agg} = \pm\sqrt{\sum_{c} MOE_{c}^{2}}$$

where MOE_c is the MOE of the c^{th} component estimate.

The example below shows how to calculate the MOE for the estimated total number of females living alone in the three Virginia counties/independent cities that border Washington, DC (Fairfax and Arlington counties, Alexandria city) from the 2006 ACS.

Table 1. **Data for Example 1**

Characteristic	Estimate	MOE
Females living alone in Fairfax County (Component 1)	52,354	±3,303
Females living alone in Arlington County (Component 2)	19,464	±2,011
Females living alone in Alexandria city (Component 3)	17,190	±1,854

The aggregate estimate is:

$$\hat{X} = \hat{X}_{Fairfax} + \hat{X}_{Arlington} + \hat{X}_{Alexandria} = 52{,}354 + 19{,}464 + 17{,}190 = 89{,}008$$

Obtain MOEs of the component estimates:

$$MOE_{Fairfax} = \pm 3{,}303,$$
$$MOE_{Arlington} = \pm 2{,}011,$$
$$MOE_{Alexandria} = \pm 1{,}854$$

Calculate the MOE for the aggregate estimated as the square root of the sum of the squared MOEs.

$$MOE_{agg} = \pm\sqrt{(3{,}303)^2 + (2{,}011)^2 + (1{,}854)^2} = \pm\sqrt{18{,}391{,}246} = \pm 4{,}289$$

Thus, the derived estimate of the number of females living alone in the three Virginia counties/independent cities that border Washington, DC, is 89,008, and the MOE for the estimate is ±4,289.

Calculating MOEs for Derived Proportions

The numerator of a proportion is a subset of the denominator (e.g., the proportion of single person households that are female). To calculate the MOE for derived proportions, do the following:

1) Obtain the MOE for the numerator and the MOE for the denominator of the proportion.
2) Square the derived proportion.
3) Square the MOE of the numerator.
4) Square the MOE of the denominator.
5) Multiply the squared MOE of the denominator by the squared proportion.
6) Subtract the result of (5) from the squared MOE of the numerator.
7) Take the square root of the result of (6).
8) Divide the result of (7) by the denominator of the proportion.

The result is the MOE for the derived proportion. Algebraically, the MOE for the derived proportion is calculated as:

$$MOE_p = \frac{\pm\sqrt{MOE_{num}^2 - (\hat{p}^2 * MOE_{den}^2)}}{\hat{X}_{den}}$$

where MOE_{num} is the MOE of the numerator.

MOE_{den} is the MOE of the denominator.

$\hat{p} = \frac{\hat{X}_{num}}{\hat{X}_{den}}$ is the derived proportion.

$\hat{X}_{num}$ is the estimate used as the numerator of the derived proportion.

$\hat{X}_{den}$ is the estimate used as the denominator of the derived proportion.

There are rare instances where this formula will fail—the value under the square root will be negative. If that happens, use the formula for derived ratios in the next section which will provide a conservative estimate of the MOE.

The example below shows how to derive the MOE for the estimated proportion of Black females 25 years of age and older in Fairfax County, Virginia, with a graduate degree based on the 2006 ACS.

Table 2. **Data for Example 2**

Characteristic	Estimate	MOE
Black females 25 years and older with a graduate degree (numerator)	4,634	±989
Black females 25 years and older (denominator)	31,713	±601

The estimated proportion is:

$$\hat{p} = \frac{\hat{X}_{gradBF}}{\hat{X}_{BF}} = \frac{4{,}634}{31{,}713} = 0.1461$$

where $\hat{X}_{gradBF}$ is the ACS estimate of Black females 25 years of age and older in Fairfax County with a graduate degree and $\hat{X}_{BF}$ is the ACS estimate of Black females 25 years of age and older in Fairfax County.

Obtain MOEs of the numerator (number of Black females 25 years of age and older in Fairfax County with a graduate degree) and denominator (number of Black females 25 years of age and older in Fairfax County).

$$MOE_{num} = \pm 989,\ MOE_{den} = \pm 601$$

Multiply the squared MOE of the denominator by the squared proportion and subtract the result from the squared MOE of the numerator.

$$MOE_{num}^2 - (\hat{p}^2 * MOE_{den}^2) =$$
$$(989)^2 - [(0.1461)^2 * (601)^2] =$$
$$978{,}121 - 7{,}712.3 = 970{,}408.7$$

Calculate the MOE by dividing the square root of the prior result by the denominator.

$$MOE_p = \frac{\pm\sqrt{970{,}408.7}}{31{,}373} = \frac{\pm 985.1}{31{,}373} = \pm 0.0311$$

Thus, the derived estimate of the proportion of Black females 25 years of age and older with a graduate degree in Fairfax County, Virginia, is 0.1461, and the MOE for the estimate is ±0.0311.

Calculating MOEs for Derived Ratios

The numerator of a ratio is not a subset (e.g., the ratio of females living alone to males living alone). To calculate the MOE for derived ratios:

1) Obtain the MOE for the numerator and the MOE for the denominator of the ratio.
2) Square the derived ratio.
3) Square the MOE of the numerator.
4) Square the MOE of the denominator.
5) Multiply the squared MOE of the denominator by the squared ratio.
6) Add the result of (5) to the squared MOE of the numerator.
7) Take the square root of the result of (6).
8) Divide the result of (7) by the denominator of the ratio.

The result is the MOE for the derived ratio. Algebraically, the MOE for the derived ratio is calculated as:

$$MOE_R = \frac{\pm\sqrt{MOE_{num}^2 + (\hat{R}^2 * MOE_{den}^2)}}{\hat{X}_{den}}$$

where MOE_{num} is the MOE of the numerator.

MOE_{den} is the MOE of the denominator.

$\hat{R} = \frac{\hat{X}_{num}}{\hat{X}_{den}}$ is the derived ratio.

$\hat{X}_{num}$ is the estimate used as the numerator of the derived ratio.

$\hat{X}_{den}$ is the estimate used as the denominator of the derived ratio.

The example below shows how to derive the MOE for the estimated ratio of Black females 25 years of age and older in Fairfax County, Virginia, with a graduate degree to Black males 25 years and older in Fairfax County with a graduate degree, based on the 2006 ACS.

Table 3. Data for Example 3

Characteristic	Estimate	MOE
Black females 25 years and older with a graduate degree (numerator)	4,634	±989
Black males 25 years and older with a graduate degree (denominator)	6,440	±1,328

The estimated ratio is:

$$\hat{R} = \frac{\hat{X}_{gradBF}}{\hat{X}_{gradBM}} = \frac{4{,}634}{6{,}440} = 0.7200$$

Obtain MOEs of the numerator (number of Black females 25 years of age and older with a graduate degree in Fairfax County) and denominator (number of Black males 25 years of age and older in Fairfax County with a graduate degree).

$$MOE_{num} = \pm 989,\ MOE_{den} = \pm 1{,}328$$

Multiply the squared MOE of the denominator by the squared proportion and add the result to the squared MOE of the numerator.

$$MOE_{num}^2 + (\hat{R}^2 * MOE_{den}^2) =$$
$$(989)^2 + [(0.7200)^2 * (1{,}328)^2] =$$
$$978{,}121 + 913{,}318.1 = 1{,}891{,}259.1$$

Calculate the MOE by dividing the square root of the prior result by the denominator.

$$MOE_R = \frac{\pm\sqrt{1{,}891{,}259.1}}{6{,}440} = \frac{\pm 1{,}375.2}{6{,}440} = \pm 0.2135$$

Thus, the derived estimate of the ratio of the number of Black females 25 years of age and older in Fairfax County, Virginia, with a graduate degree to the number of Black males 25 years of age and older in Fairfax County, Virginia, with a graduate degree is 0.7200, and the MOE for the estimate is ±0.2135.

Calculating MOEs for the Product of Two Estimates

To calculate the MOE for the product of two estimates, do the following:

1) Obtain the MOEs for the two estimates being multiplied together.
2) Square the estimates and their MOEs.
3) Multiply the first squared estimate by the second estimate's squared MOE.
4) Multiply the second squared estimate by the first estimate's squared MOE.
5) Add the results from (3) and (4).
6) Take the square root of (5).

The result is the MOE for the product. Algebraically, the MOE for the product is calculated as:

$$MOE_{A\times B} = \pm\sqrt{A^2 \times MOE_B^2 + B^2 \times MOE_A^2}$$

where *A* and *B* are the first and second estimates, respectively.

MOE_A is the MOE of the first estimate.

MOE_B is the MOE of the second estimate.

The example below shows how to derive the MOE for the estimated number of Black workers 16 years and over in Fairfax County, Virginia, who used public transportation to commute to work, based on the 2006 ACS.

Table 4. Data for Example 4

Characteristic	Estimate	MOE
Black workers 16 years and over (first estimate)	50,624	±2,423
Percent of Black workers 16 years and over who commute by public transportation (second estimate)	13.4%	±2.7%

To apply the method, the proportion (0.134) needs to be used instead of the percent (13.4). The estimated product is 50,624 × 0.134 = 6,784. The MOE is calculated by:

$$MOE_{A\times B} = \pm\sqrt{50{,}624^2 \times 0.027^2 + 0.134^2 \times 2{,}423^2}$$
$$= \pm 1{,}405$$

Thus, the derived estimate of Black workers 16 years and over who commute by public transportation is 6,784, and the MOE of the estimate is ±1,405.

Calculating MOEs for Estimates of "Percent Change" or "Percent Difference"

The "percent change" or "percent difference" between two estimates (for example, the same estimates in two different years) is commonly calculated as

$$Percent\ Change = 100\% * \frac{\hat{X}_2 - \hat{X}_1}{\hat{X}_1}$$

Because $\hat{X}_2$ is not a subset of $\hat{X}_1$, the procedure to calculate the MOE of a ratio discussed previously should be used here to obtain the MOE of the percent change.

The example below shows how to calculate the margin of error of the percent change using the 2006 and 2005 estimates of the number of persons in Maryland who lived in a different house in the U.S. 1 year ago.

Table 5. **Data for Example 5**

Characteristic	Estimate	MOE
Persons who lived in a different house in the U.S. 1 year ago, 2006	802,210	±22,866
Persons who lived in a different house in the U.S. 1 year ago, 2005	762,475	±22,666

The percent change is:

$$Percent\ Change = 100\% * \frac{\hat{X}_2 - \hat{X}_1}{\hat{X}_1} =$$

$$100\% * \left(\frac{802{,}210 - 762{,}475}{762{,}475} \right) = 5.21\%$$

For use in the ratio formula, the ratio of the two estimates is:

$$\hat{R} = \frac{\hat{X}_2}{\hat{X}_1} = \frac{802{,}210}{762{,}475} = 1.0521$$

The MOEs for the numerator ($\hat{X}_2$) and denominator ($\hat{X}_1$) are:

$$MOE_2 = +/-22{,}866,\ MOE_1 = +/-22{,}666$$

Add the squared MOE of the numerator (MOE_2) to the product of the squared ratio and the squared MOE of the denominator (MOE_1):

$$MOE_2^2 + (\hat{R}^2 * MOE_1^2) =$$

$$(22{,}866)^2 + [(1.0521)^2 * (22{,}666)^2] =$$

$$1{,}091{,}528{,}529$$

Calculate the MOE by dividing the square root of the prior result by the denominator ($\hat{X}_1$).

$$MOE_R = \frac{\pm\sqrt{1{,}091{,}528{,}529}}{762{,}475} = \frac{\pm 33{,}038.3}{762{,}475} = \pm 0.0433$$

Finally, the MOE of the percent change is the MOE of the ratio, multiplied by 100 percent, or 4.33 percent.

The text box below summarizes the formulas used to calculate the margin of error for several derived estimates.

Calculating Margins of Error for Derived Estimates

Aggregated Count Data

$$MOE_{agg} = \pm\sqrt{\sum_c MOE_c^2}$$

Derived Proportions

$$MOE_p = \frac{\pm\sqrt{MOE_{num}^2 - (\hat{p}^2 * MOE_{den}^2)}}{\hat{X}_{den}}$$

Derived Ratios

$$MOE_R = \frac{\pm\sqrt{MOE_{num}^2 + (\hat{R}^2 * MOE_{den}^2)}}{\hat{X}_{den}}$$

Appendix 4.
Making Comparisons

One of the most important uses of the ACS estimates is to make comparisons between estimates. Several key types of comparisons are of general interest to users: 1) comparisons of estimates from different geographic areas within the same time period (e.g., comparing the proportion of people below the poverty level in two counties); 2) comparisons of estimates for the same geographic area across time periods (e.g., comparing the proportion of people below the poverty level in a county for 2006 and 2007); and 3) comparisons of ACS estimates with the corresponding estimates from past decennial census samples (e.g., comparing the proportion of people below the poverty level in a county for 2006 and 2000).

A number of conditions must be met when comparing survey estimates. Of primary importance is that the comparison takes into account the sampling error associated with each estimate, thus determining whether the observed differences between estimates are statistically significant. Statistical significance means that there is statistical evidence that a true difference exists within the full population, and that the observed difference is unlikely to have occurred by chance due to sampling. A method for determining statistical significance when making comparisons is presented in the next section. Considerations associated with the various types of comparisons that could be made are also discussed.

Determining Statistical Significance

When comparing two estimates, one should use the test for significance described below. This approach will allow the user to ascertain whether the observed difference is likely due to chance (and thus is not statistically significant) or likely represents a true difference that exists in the population as a whole (and thus is statistically significant).

The test for significance can be carried out by making several computations using the estimates and their corresponding standard errors (SEs). When working with ACS data, these computations are simple given the data provided in tables in the American FactFinder.

1) Determine the SE for each estimate (for ACS data, SE is defined by the positive value of the margin of error (MOE) divided by 1.645).[4]
2) Square the resulting SE for each estimate.
3) Sum the squared SEs.
4) Calculate the square root of the sum of the squared SEs.
5) Calculate the difference between the two estimates.
6) Divide (5) by (4).
7) Compare the absolute value of the result of (6) with the critical value for the desired level of confidence (1.645 for 90 percent, 1.960 for 95 percent, 2.576 for 99 percent).
8) If the absolute value of the result of (6) is greater than the critical value, then the difference between the two estimates can be considered statistically significant at the level of confidence corresponding to the critical value used in (7).

Algebraically, the significance test can be expressed as follows:

If $\left|\frac{\hat{X}_1 - \hat{X}_2}{\sqrt{SE_1^2 + SE_2^2}}\right| > Z_{CL}$, then the difference between estimates $\hat{X}_1$ and $\hat{X}_2$ is statistically significant at the specified confidence level, CL

where $\hat{X}_i$ is estimate i (=1,2)

SE_i is the SE for the estimate i (=1,2)

Z_{CL} is the critical value for the desired confidence level (=1.645 for 90 percent, 1.960 for 95 percent, 2.576 for 99 percent).

The example below shows how to determine if the difference in the estimated percentage of households in 2006 with one or more people of age 65 and older between State A (estimated percentage =22.0, SE=0.12) and State B (estimated percentage =21.5, SE=0.12) is statistically significant. Using the formula above:

$$\left|\frac{\hat{X}_1 - \hat{X}_2}{\sqrt{SE_1^2 + SE_2^2}}\right| = \left|\frac{22.0 - 21.5}{\sqrt{(0.12)^2 + (0.12)^2}}\right| =$$

$$\left|\frac{0.5}{\sqrt{0.015 + 0.015}}\right| = \left|\frac{0.5}{\sqrt{0.03}}\right| = \left|\frac{0.5}{0.173}\right| = 2.90$$

Since the test value (2.90) is greater than the critical value for a confidence level of 99 percent (2.576), the difference in the percentages is statistically significant at a 99-percent confidence level. This is also referred to as statistically significant at the alpha = 0.01 level. A rough interpretation of the result is that the user can be 99 percent certain that a difference exists between the percentages of households with one or more people aged 65 and older between State A and State B.

[4] NOTE: If working with ACS single-year estimates for 2005 or earlier, use the value 1.65 rather than 1.645.

By contrast, if the corresponding estimates for State C and State D were 22.1 and 22.5, respectively, with standard errors of 0.20 and 0.25, respectively, the formula would yield

$$\left|\frac{\hat{X}_1 - \hat{X}_2}{\sqrt{SE_1^2 + SE_2^2}}\right| = \left|\frac{22.5 - 22.1}{\sqrt{(0.20)^2 + (0.25)^2}}\right| =$$

$$\left|\frac{0.4}{\sqrt{0.04 + 0.0625}}\right| = \left|\frac{0.4}{\sqrt{0.1025}}\right| = \left|\frac{0.4}{0.320}\right| = 1.25$$

Since the test value (1.25) is less than the critical value for a confidence level of 90 percent (1.645), the difference in percentages is not statistically significant. A rough interpretation of the result is that the user cannot be certain to any sufficient degree that the observed difference in the estimates was not due to chance.

Comparisons Within the Same Time Period

Comparisons involving two estimates from the same time period (e.g., from the same year or the same 3-year period) are straightforward and can be carried out as described in the previous section. There is, however, one statistical aspect related to the test for statistical significance that users should be aware of. When comparing estimates within the same time period, the areas or groups will generally be nonoverlapping (e.g., comparing estimates for two different counties). In this case, the two estimates are independent, and the formula for testing differences is statistically correct.

In some cases, the comparison may involve a large area or group and a subset of the area or group (e.g., comparing an estimate for a state with the corresponding estimate for a county within the state or comparing an estimate for all females with the corresponding estimate for Black females). In these cases, the two estimates are not independent. The estimate for the large area is partially dependent on the estimate for the subset and, strictly speaking, the formula for testing differences should account for this partial dependence. However, unless the user has reason to believe that the two estimates are strongly correlated, it is acceptable to ignore the partial dependence and use the formula for testing differences as provided in the previous section. However, if the two estimates are positively correlated, a finding of statistical significance will still be correct, but a finding of a lack of statistical significance based on the formula may be incorrect. If it is important to obtain a more exact test of significance, the user should consult with a statistician about approaches for accounting for the correlation in performing the statistical test of significance.

Comparisons Across Time Periods

Comparisons of estimates from different time periods may involve different single-year periods or different multiyear periods of the same length within the same area. Comparisons across time periods should be made only with comparable time period estimates. Users are advised against comparing single-year estimates with multiyear estimates (e.g., comparing 2006 with 2007–2009) and against comparing multiyear estimates of differing lengths (e.g., comparing 2006–2008 with 2009–2014), as they are measuring the characteristics of the population in two different ways, so differences between such estimates are difficult to interpret. When carrying out any of these types of comparisons, users should take several other issues into consideration.

When comparing estimates from two different single-year periods, one prior to 2006 and the other 2006 or later (e.g., comparing estimates from 2005 and 2007), the user should recognize that from 2006 on the ACS sample includes the population living in group quarters (GQ) as well as the population living in housing units. Many types of GQ populations have demographic, social, or economic characteristics that are very different from the household population. As a result, comparisons between 2005 and 2006 and later ACS estimates could be affected. This is particularly true for areas with a substantial GQ population. For most population characteristics, the Census Bureau suggests users make comparisons across these time periods only if the geographic area of interest does not include a substantial GQ population. For housing characteristics or characteristics published only for the household population, this is obviously not an issue.

Comparisons Based on Overlapping Periods

When comparing estimates from two multiyear periods, ideally comparisons should be based on nonoverlapping periods (e.g., comparing estimates from 2006–2008 with estimates from 2009–2011). The comparison of two estimates for different, but overlapping periods is challenging since the difference is driven by the nonoverlapping years. For example, when comparing the 2005–2007 ACS with the 2006–2008 ACS, data for 2006 and 2007 are included in both estimates. Their contribution is subtracted out when the estimate of differences is calculated. While the interpretation of this difference is difficult, these comparisons can be made with caution. Under most circumstances, the estimate of difference should not be interpreted as a reflection of change between the last 2 years.

The use of MOEs for assessing the reliability of change over time is complicated when change is being evaluated using multiyear estimates. From a technical standpoint, change over time is best evaluated with multiyear estimates that do not overlap. At the same time,

many areas whose only source of data will be 5-year estimates will not want to wait until 2015 to evaluate change (i.e., comparing 2005–2009 with 2010–2014).

When comparing two 3-year estimates or two 5-year estimates of the same geography that overlap in sample years one must account for this sample overlap. Thus to calculate the standard error of this difference use the following approximation to the standard error:

$$SE(\hat{X}_1 - \hat{X}_2) \cong \sqrt{(1-C)}\sqrt{SE_1^{\,2} + SE_2^{\,2}}$$

where C is the fraction of overlapping years. For example, the periods 2005–2009 and 2007–2011 overlap for 3 out of 5 years, so C=3/5=0.6. If the periods do not overlap, such as 2005–2007 and 2008–2010, then C=0.

With this SE one can test for the statistical significance of the difference between the two estimates using the method outlined in the previous section with one modification; substitute $\sqrt{(1-C)}\sqrt{SE_1^{\,2} + SE_2^{\,2}}$ for $\sqrt{SE_1^{\,2} + SE_2^{\,2}}$ in the denominator of the formula for the significance test.

Comparisons With Census 2000 Data

In Appendix 2, major differences between ACS data and decennial census sample data are discussed. Factors such as differences in residence rules, universes, and reference periods, while not discussed in detail in this appendix, should be considered when comparing ACS estimates with decennial census estimates. For example, given the reference period differences, seasonality may affect comparisons between decennial census and ACS estimates when looking at data for areas such as college towns and resort areas.

The Census Bureau subject matter specialists have reviewed the factors that could affect differences between ACS and decennial census estimates and they have determined that ACS estimates are similar to those obtained from past decennial census sample data for most areas and characteristics. The user should consider whether a particular analysis involves an area or characteristic that might be affected by these differences.[5]

When comparing ACS and decennial census sample estimates, the user must remember that the decennial census sample estimates have sampling error associated with them and that the standard errors for both ACS and census estimates must be incorporated when performing tests of statistical significance. Appendix 3 provides the calculations necessary for determining statistical significance of a difference between two estimates. To derive the SEs of census sample estimates, use the method described in Chapter 8 of either the Census 2000 Summary File 3 Technical Documentation <http://www.census.gov/prod/cen2000/doc/sf3.pdf> or the Census 2000 Summary File 4 Technical Documentation <http://www.census.gov/prod/cen2000/doc/sf4.pdf>.

A conservative approach to testing for statistical significance when comparing ACS and Census 2000 estimates that avoids deriving the SE for the Census 2000 estimate would be to assume the SE for the Census 2000 estimate is the same as that determined for the ACS estimate. The result of this approach would be that a finding of statistical significance can be assumed to be accurate (as the SE for the Census 2000 estimate would be expected to be less than that for the ACS estimate), but a finding of no statistical significance could be incorrect. In this case the user should calculate the census long-form standard error and follow the steps to conduct the statistical test.

Comparisons With 2010 Census Data

Looking ahead to the 2010 decennial census, data users need to remember that the socioeconomic data previously collected on the long form during the census will not be available for comparison with ACS estimates. The only common variables for the ACS and 2010 Census are sex, age, race, ethnicity, household relationship, housing tenure, and vacancy status.

The critical factor that must be considered when comparing ACS estimates encompassing 2010 with the 2010 Census is the potential impact of housing and population controls used for the ACS. As the housing and population controls used for 2010 ACS data will be based on the Population Estimates Program where the estimates are benchmarked on the Census 2000 counts, they will not agree with the 2010 Census population counts for that year. The 2010 population estimates may differ from the 2010 Census counts for two major reasons—the true change from 2000 to 2010 is not accurately captured by the estimates and the completeness of coverage in the 2010 Census is different than coverage of Census 2000. The impact of this difference will likely affect most areas and states, and be most notable for smaller geographic areas where the potential for large differences between the population controls and the 2010 Census population counts is greater.

Comparisons With Other Surveys

Comparisons of ACS estimates with estimates from other national surveys, such as the Current Population Survey, may be of interest to some users. A major consideration in making such comparisons will be that ACS

[5] Further information concerning areas and characteristics that do not fit the general pattern of comparability can be found on the ACS Web site at <http://www.census.gov/acs/www/UseData/compACS.htm>.

estimates include data for populations in both institutional and noninstitutional group quarters, and estimates from most national surveys do not include institutional populations. Another potential for large effects when comparing data from the ACS with data from other national surveys is the use of different questions for measuring the same or similar information.

Sampling error and its impact on the estimates from the other survey should be considered if comparisons and statements of statistical difference are to be made, as described in Appendix 3. The standard errors on estimates from other surveys should be derived according to technical documentation provided for those individual surveys.

Finally, the user wishing to compare ACS estimates with estimates from other national surveys should consider the potential impact of other factors, such as target population, sample design and size, survey period, reference period, residence rules, and interview modes on estimates from the two sources.

Appendix 5.
Using Dollar-Denominated Data

Dollar-denominated data refer to any characteristics for which inflation adjustments are used when producing annual estimates. For example, income, rent, home value, and energy costs are all dollar-denominated data.

Inflation will affect the comparability of dollar-denominated data across time periods. When ACS multiyear estimates for dollar-denominated data are generated, amounts are adjusted using inflation factors based on the Consumer Price Index (CPI).

Given the potential impact of inflation on observed differences of dollar-denominated data across time periods, users should adjust for the effects of inflation. Such an adjustment will provide comparable estimates accounting for inflation. In making adjustments, the Census Bureau recommends using factors based on the All Items CPI-U-RS (CPI research series). The Bureau of Labor Statistics CPI indexes through 2006 are found at <http://www.bls.gov/cpi/cpiurs1978_2006.pdf>. Explanations follow.

Creating Single-Year Income Values

ACS income values are reported based on the amount of income received during the 12 months preceding the interview month. This is the income reference period. Since there are 12 different income reference periods throughout an interview year, 12 different income inflation adjustments are made. Monthly CPI-U-RSs are used to inflation-adjust the 12 reference period incomes to a single reference period of January through December of the interview year. Note that there are no inflation adjustments for single-year estimates of rent, home value, or energy cost values.

Adjusting Single-Year Estimates Over Time

When comparing single-year income, rent, home value, and energy cost value estimates from two different years, adjustment should be made as follows:

1) Obtain the All Items CPI-U-RS Annual Averages for the 2 years being compared.

2) Calculate the inflation adjustment factor as the ratio of the CPI-U-RS from the more recent year to the CPI-U-RS from the earlier year.

3) Multiply the dollar-denominated data estimated for the earlier year by the inflation adjustment factor.

The inflation-adjusted estimate for the earlier year can be expressed as:

$$\hat{X}_{Y1,Adj} = \frac{CPI_{Y2}}{CPI_{Y1}} \hat{X}_{Y1}$$

where CPI_{Y1} is the All Items CPI-U-RS Annual Average for the earlier year (Y1).

CPI_{Y2} is the All Items CPI-U-RS Annual Average for the more recent year (Y2).

$\hat{X}_{Y1}$ is the published ACS estimate for the earlier year (Y1).

The example below compares the national median value for owner-occupied mobile homes in 2005 ($37,700) and 2006 ($41,000). First adjust the 2005 median value using the 2005 All Items CPI-U-RS Annual Average (286.7) and the 2006 All Items CPI-U-RS Annual Average (296.1) as follows:

$$\hat{X}_{2005,Adj} = \frac{296.1}{286.7} \times \$37{,}700 = \$38{,}936$$

Thus, the comparison of the national median value for owner-occupied mobile homes in 2005 and 2006, in 2006 dollars, would be $38,936 (2005 inflation-adjusted to 2006 dollars) versus $41,000 (2006 dollars).

Creating Values Used in Multiyear Estimates

Multiyear income, rent, home value, and energy cost values are created with inflation adjustments. The Census Bureau uses the All Items CPI-U-RS Annual Averages for each year in the multiyear time period to calculate a set of inflation adjustment factors. Adjustment factors for a time period are calculated as ratios of the CPI-U-RS Annual Average from its most recent year to the CPI-U-RS Annual Averages from each of its earlier years. The ACS values for each of the earlier years in the multiyear period are multiplied by the appropriate inflation adjustment factors to produce the inflation-adjusted values. These values are then used to create the multiyear estimates.

As an illustration, consider the time period 2004–2006, which consisted of individual reference-year income values of $30,000 for 2006, $20,000 for 2005, and $10,000 for 2004. The multiyear income components are created from inflation-adjusted reference period income values using factors based on the All Items CPI-U-RS Annual Averages of 277.4 (for 2004), 286.7 (for 2005), and 296.1 (for 2006). The adjusted 2005 value is the ratio of 296.1 to 286.7 applied to $20,000, which equals $20,656. Similarly, the 2004 value is the ratio of 296.1 to 277.4 applied to $10,000, which equals $10,674.

Adjusting Multiyear Estimates Over Time

When comparing multiyear estimates from two different time periods, adjustments should be made as follows:

1) Obtain the All Items CPI-U-RS Annual Average for the most current year in each of the time periods being compared.
2) Calculate the inflation adjustment factor as the ratio of the CPI-U-RS Annual Average in (1) from the most recent year to the CPI-U-RS in (1) from the earlier years.
3) Multiply the dollar-denominated estimate for the earlier time period by the inflation adjustment factor.

The inflation-adjusted estimate for the earlier years can be expressed as:

$$\hat{X}_{P1,Adj} = \frac{CPI_{P2}}{CPI_{P1}} \hat{X}_{P1}$$

where CPI_{P1} is the All Items CPI-U-RS Annual Average for the last year in the earlier time period (P1).

CPI_{P2} is the All Items CPI-U-RS Annual Average for the last year in the most recent time period (P2).

$\hat{X}_{P1}$ is the published ACS estimate for the earlier time period (P1).

As an illustration, consider ACS multiyear estimates for the two time periods of 2001–2003 and 2004–2006. To compare the national median value for owner-occupied mobile homes in 2001–2003 ($32,000) and 2004–2006 ($39,000), first adjust the 2001–2003 median value using the 2003 All Items CPI-U-RS Annual Averages (270.1) and the 2006 All Items CPI-U-RS Annual Averages (296.1) as follows:

$$\hat{X}_{2001-2003,Adj} = \frac{296.1}{270.1} \times \$32{,}000 = \$35{,}080$$

Thus, the comparison of the national median value for owner-occupied mobile homes in 2001–2003 and 2004–2006, in 2006 dollars, would be $35,080 (2001–2003 inflation-adjusted to 2006 dollars) versus $39,000 (2004–2006, already in 2006 dollars).

Issues Associated With Inflation Adjustment

The recommended inflation adjustment uses a national level CPI and thus will not reflect inflation differences that may exist across geographies. In addition, since the inflation adjustment uses the All Items CPI, it will not reflect differences that may exist across characteristics such as energy and housing costs.

Appendix 6.
Measures of Nonsampling Error

All survey estimates are subject to both sampling and nonsampling error. In Appendix 3, the topic of sampling error and the various measures available for understanding the uncertainty in the estimates due to their being derived from a sample, rather than from an entire population, are discussed. The margins of error published with ACS estimates measure only the effect of sampling error. Other errors that affect the overall accuracy of the survey estimates may occur in the course of collecting and processing the ACS, and are referred to collectively as nonsampling errors.

Broadly speaking, nonsampling error refers to any error affecting a survey estimate outside of sampling error. Nonsampling error can occur in complete censuses as well as in sample surveys, and is commonly recognized as including coverage error, unit nonresponse, item nonresponse, response error, and processing error.

Types of Nonsampling Errors

Coverage error occurs when a housing unit or person does not have a chance of selection in the sample (undercoverage), or when a housing unit or person has more than one chance of selection in the sample, or is included in the sample when they should not have been (overcoverage). For example, if the frame used for the ACS did not allow the selection of newly constructed housing units, the estimates would suffer from errors due to housing undercoverage.

The final ACS estimates are adjusted for under- and overcoverage by controlling county-level estimates to independent total housing unit controls and to independent population controls by sex, age, race, and Hispanic origin (more information is provided on the coverage error definition page of the "ACS Quality Measures" Web site at <http://www.census.gov/acs/www/UseData/sse/cov/cov_def.htm>). However, it is important to measure the extent of coverage adjustment by comparing the precontrolled ACS estimates to the final controlled estimates. If the extent of coverage adjustments is large, there is a greater chance that differences in characteristics of undercovered or overcovered housing units or individuals differ from those eligible to be selected. When this occurs, the ACS may not provide an accurate picture of the population prior to the coverage adjustment, and the population controls may not eliminate or minimize that coverage error.

Unit nonresponse is the failure to obtain the minimum required information from a housing unit or a resident of a group quarter in order for it to be considered a completed interview. Unit nonresponse means that no survey data are available for a particular sampled unit or person. For example, if no one in a sampled housing unit is available to be interviewed during the time frame for data collection, unit nonresponse will result.

It is important to measure unit nonresponse because it has a direct effect on the quality of the data. If the unit nonresponse rate is high, it increases the chance that the final survey estimates may contain bias, even though the ACS estimation methodology includes a nonresponse adjustment intended to control potential unit nonresponse bias. This will happen if the characteristics of nonresponding units differ from the characteristics of responding units.

Item nonresponse occurs when a respondent fails to provide an answer to a required question or when the answer given is inconsistent with other information. With item nonresponse, while some responses to the survey questionnaire for the unit are provided, responses to other questions are not obtained. For example, a respondent may be unwilling to respond to a question about income, resulting in item nonresponse for that question. Another reason for item nonresponse may be a lack of understanding of a particular question by a respondent.

Information on item nonresponse allows users to judge the completeness of the data on which the survey estimates are based. Final estimates can be adversely impacted when item nonresponse is high, because bias can be introduced if the actual characteristics of the people who do not respond to a question differ from those of people who do respond to it. The ACS estimation methodology includes imputations for item nonresponse, intended to reduce the potential for item nonresponse bias.

Response error occurs when data are reported or recorded incorrectly. Response errors may be due to the respondent, the interviewer, the questionnaire, or the survey process itself. For example, if an interviewer conducting a telephone interview incorrectly records a respondent's answer, response error results. In the same way, if the respondent fails to provide a correct response to a question, response error results. Another potential source of response error is a survey process that allows proxy responses to be obtained, wherein a knowledgeable person within the household provides responses for another person within the household who is unavailable for the interview. Even more error prone is allowing neighbors to respond.

Processing error can occur during the preparation of the final data files. For example, errors may occur if data entry of questionnaire information is incomplete

or inaccurate. Coding of responses incorrectly also results in processing error. Critical reviews of edits and tabulations by subject matter experts are conducted to keep errors of this kind to a minimum.

Nonsampling error can result in random errors and systematic errors. Of greatest concern are systematic errors. Random errors are less critical since they tend to cancel out at higher geographic levels in large samples such as the ACS.

On the other hand, systematic errors tend to accumulate over the entire sample. For example, if there is an error in the questionnaire design that negatively affects the accurate capture of respondents' answers, processing errors are created. Systematic errors often lead to a bias in the final results. Unlike sampling error and random error resulting from nonsampling error, bias caused by systematic errors cannot be reduced by increasing the sample size.

ACS Quality Measures

Nonsampling error is extremely difficult, if not impossible, to measure directly. However, the Census Bureau has developed a number of indirect measures of nonsampling error to help inform users of the quality of the ACS estimates: sample size, coverage rates, unit response rates and nonresponse rates by reason, and item allocation rates. Starting with the 2007 ACS, these measures are available in the B98 series of detailed tables on AFF. Quality measures for previous years are available on the "ACS Quality Measures" Web site at <http:/www.census.gov/acs/www/UseData/sse/>.

Sample size measures for the ACS summarize information for the housing unit and GQ samples. The measures available at the state level are:[6]

Housing units
- Number of initial addresses selected
- Number of final survey interviews

Group quarters people (beginning with the 2006 ACS)
- Number of initial persons selected
- Number of final survey interviews

Sample size measures may be useful in special circumstances when determining whether to use single-year or multiyear estimates in conjunction with estimates of the population of interest. While the coefficient of variation (CV) should typically be used to determine usability, as explained in Appendix 3, there may be some situations where the CV is small but the user has reason to believe the sample size for a subgroup is very small and the robustness of the estimate is in question.

For example, the Asian-alone population makes up roughly 1 percent (8,418/656,700) of the population in Jefferson County, Alabama. Given that the number of successful housing unit interviews in Jefferson County for the 2006 ACS were 4,072 and assuming roughly 2.5 persons per household (or roughly 12,500 completed person interviews), one could estimate that the 2006 ACS data for Asians in Jefferson County are based on roughly 150 completed person interviews.

Coverage rates are available for housing units, and total population by sex at both the state and national level. Coverage rates for total population by six race/ethnicity categories and the GQ population are also available at the national level. These coverage rates are a measure of the extent of adjustment to the survey weights required during the component of the estimation methodology that adjusts to population controls. Low coverage rates are an indication of greater potential for coverage error in the estimates.

Unit response and nonresponse rates for housing units are available at the county, state, and national level by reason for nonresponse: refusal, unable to locate, no one home, temporarily absent, language problem, other, and data insufficient to be considered an interview. Rates are also provided separately for persons in group quarters at the national and state levels.

A low unit response rate is an indication that there is potential for bias in the survey estimates. For example, the 2006 housing unit response rates are at least 94 percent for all states. The response rate for the District of Columbia in 2006 was 91 percent.

Item allocation rates are determined by the content edits performed on the individual raw responses and closely correspond to item nonresponse rates. Overall housing unit and person characteristic allocation rates are available at the state and national levels, which combine many different characteristics. Allocation rates for individual items may be calculated from the B99 series of imputation detailed tables available in AFF.

Item allocation rates do vary by state, so users are advised to examine the allocation rates for characteristics of interest before drawing conclusions from the published estimates.

[6] The sample size measures for housing units (number of initial addresses selected and number of final survey interviews) and for group quarters people cannot be used to calculate response rates. For the housing unit sample, the number of initial addresses selected includes addresses that were determined not to identify housing units, as well as initial addresses that are subsequently subsampled out in preparation for personal visit nonresponse follow-up. Similarly, the initial sample of people in group quarters represents the expected sample size within selected group quarters prior to visiting and sampling of residents.

Appendix 7.
Implications of Population Controls on ACS Estimates

As with most household surveys, the American Community Survey data are controlled so that the numbers of housing units and people in categories defined by age, sex, race, and Hispanic origin agree with the Census Bureau's official estimates. The American Community Survey (ACS) measures the characteristics of the population, but the official count of the population comes from the previous census, updated by the Population Estimates Program.

In the case of the ACS, the total housing unit estimates and the total population estimates by age, sex, race and Hispanic origin are controlled at the county (or groups of counties) level. The group quarters total population is controlled at the state level by major type of group quarters. Such adjustments are important to correct the survey data for nonsampling and sampling errors. An important source of nonsampling error is the potential under-representation of hard-to-enumerate demographic groups. The use of the population controls results in ACS estimates that more closely reflect the level of coverage achieved for those groups in the preceding census. The use of the population estimates as controls partially corrects demographically implausible results from the ACS due to the ACS data being based on a sample of the population rather than a full count. For example, the use of the population controls "smooths out" demographic irregularities in the age structure of the population that result from random sampling variability in the ACS.

When the controls are applied to a group of counties rather than a single county, the ACS estimates and the official population estimates for the individual counties may not agree. There also may not be agreement between the ACS estimates and the population estimates for levels of geography such as subcounty areas where the population controls are not applied.

The use of population and housing unit controls also reduces random variability in the estimates from year to year. Without the controls, the sampling variability in the ACS could cause the population estimates to increase in one year and decrease in the next (especially for smaller areas or demographic groups), when the underlying trend is more stable. This reduction in variability on a time series basis is important since results from the ACS may be used to monitor trends over time. As more current data become available, the time series of estimates from the Population Estimates Program are revised back to the preceding census while the ACS estimates in previous years are not. Therefore, some differences in the ACS estimates across time may be due to changes in the population estimates.

For single-year ACS estimates, the population and total housing unit estimates for July 1 of the survey year are used as controls. For multiyear ACS estimates, the controls are the average of the individual year population estimates.

Appendix 8.
Other ACS Resources

Background and Overview Information

American Community Survey Web Page Site Map: <http://www.census.gov/acs/www/Site_Map.html> This link is the site map for the ACS Web page. It provides an overview of the links and materials that are available online, including numerous reference documents.

What Is the ACS? <http://www.census.gov/acs/www/SBasics/What/What1.htm> This Web page includes basic information about the ACS and has links to additional information including background materials.

ACS Design, Methodology, Operations

American Community Survey Design and Methodology Technical Paper: <http://www.census.gov/acs/www/Downloads/tp67.pdf> This document describes the basic design of the 2005 ACS and details the full set of methods and procedures that were used in 2005. Please watch our Web site as a revised version will be released in the fall of 2008, detailing methods and procedures used in 2006 and 2007.

About the Data (Methodology: <http://www.census.gov/acs/www/AdvMeth/> This Web page contains links to information on ACS data collection and processing, evaluation reports, multiyear estimates study, and related topics.

ACS Quality

Accuracy of the Data (2007): <http://www.census.gov/acs/www/Downloads/ACS/accuracy2007.pdf> This document provides data users with a basic understanding of the sample design, estimation methodology, and accuracy of the 2007 ACS data.

ACS Sample Size: <http://www.census.gov/acs/www/SBasics/SSizes/SSizes06.htm> This link provides sample size information for the counties that were published in the 2006 ACS. The initial sample size and the final completed interviews are provided. The sample sizes for all published counties and county equivalents starting with the 2007 ACS will only be available in the B98 series of detailed tables on American FactFinder.

ACS Quality Measures: <http://www.census.gov/acs/www/UseData/sse/> This Web page includes information about the steps taken by the Census Bureau to improve the accuracy of ACS data. Four indicators of survey quality are described and measures are provided at the national and state level.

Guidance on Data Products and Using the Data

How to Use the Data: <http://www.census.gov/acs/www/UseData/> This Web page includes links to many documents and materials that explain the ACS data products.

Comparing ACS Data to other sources: <http://www.census.gov/acs/www/UseData/compACS.htm> Tables are provided with guidance on comparing the 2007 ACS data products to 2006 ACS data and Census 2000 data.

Fact Sheet on Using Different Sources of Data for Income and Poverty: <http://www.census.gov/hhes/www/income/factsheet.html> This fact sheet highlights the sources that should be used for data on income and poverty, focusing on comparing the ACS and the Current Population Survey (CPS).

Public Use Microdata Sample (PUMS): <http://www.census.gov/acs/www/Products/PUMS/> This Web page provides guidance in accessing ACS microdata.