Speaker Notes:

Introduction to the Public Use Microdata Sample (PUMS) Files from the American Community Survey

Updated: 2/20/09

Slide 1

This presentation provides an introduction to the Public Use Microdata Sample files from the American Community Survey.

Slide 2

During this introductory presentation to the ACS PUMS, we will address the following questions:

1. What is the ACS Public Use Microdata Sample, or PUMS, file?
2. What geographic areas are available in the ACS PUMS?
3. How does the Census Bureau protect confidentiality in the ACS PUMS?
4. And, how do I access ACS PUMS data?

Slide 3

First, what is the ACS PUMS and how does it differ from the tables available through American FactFinder?

The ACS Public Use Microdata Sample, or PUMS, is one of the ACS data products provided to ACS users by the Census Bureau. It is a sample of population and housing unit records from the American Community Survey. The ACS PUMS files include the actual responses collected in ACS questionnaires, although some responses have been edited to protect confidentiality of respondents. The ACS PUMS file includes sample weights for each person and housing unit, which you can apply to the individual records to expand the sample to estimate totals, percentages, means, and medians of the full population.

The 1-year ACS PUMS file represents about 1-percent of the total U.S. population. The Census Bureau has also released a 3-year ACS PUMS file that combines the responses from the 2005, 2006, and 2007 1-year PUMS files. There are some differences between the 2005, 2006, and 2007 1-year PUMS files. The multiyear PUMS file resolves the differences in a consistent manner.

Slide 4

The ACS PUMS files are microdata, which are very different from the summary data that are available through American FactFinder.

Summary data and tables present measures such as counts, averages and medians for specific geographic areas, such as states or counties. Summary data are available through predefined tables, charts and graphs.

In microdata files, the unit of analysis is the individual housing unit or person, and the records include actual questionnaire responses. For example, the PUMS data show how respondents answered questions on education, occupation, housing characteristics, and so forth. When working with microdata files, the data user specifies the tables to be produced and the geographic areas that are needed.

Slide 5

ACS PUMS data are most useful for those who need to create tables that are not available through the Census Bureau's American FactFinder. It is a great resource for researchers, students, government officials, and others who need publicly available and up-to-date information about U.S. population and housing characteristics. Many in academia find the PUMS data useful for regression analysis and modeling applications. Some of these researchers previously used data from the Current Population Survey or other surveys but are now using the ACS because of its larger sample size.

The PUMS files are especially useful for studying population groups, such as those of German ancestry in Pennsylvania or U.S. scientists and engineers, for which published tables may be limited.

The ACS PUMS can also be used to show the distribution of population and housing characteristics across cities, suburbs, and rural areas. We'll talk more about geography in the ACS PUMS in a minute.

Slide 6

Here are a few sample questions that could be answered with the ACS PUMS:

1. What proportion of children ages 5 to 9 live in households that do not have telephone service?
2. What is the veteran status of college students living in Maine?
3. What proportion of low-income workers in Kentucky commute 90+ minutes to work?

As with any data collected from a sample survey instead of a census of the full population, your analysis needs to take sampling error into consideration. The smaller the geographic area or population group of interest, the more you will need to pay attention to sampling error in the PUMS data. We will talk briefly about measuring sampling error later in the presentation.

Slide 7

The geographic areas in the ACS PUMS files are limited to the nation, the 50 states, D.C., Puerto Rico, and Public Use Microdata areas, or PUMAs.

PUMAs, which we describe in the next slide, are the smallest geographic areas identified in the ACS PUMS.

Slide 8

The Census Bureau divides each state into a series of Public Use Microdata Areas, or PUMAs, each of which has a minimum population of 100,000. PUMAs are generally constructed based on county, neighborhood, and city boundaries and are not allowed to cross state lines. Typically, counties with large populations are subdivided into multiple PUMAs, while PUMAs in more rural areas are made up of groups of adjacent counties. PUMA boundaries can cross county boundaries and a PUMA can be made up of parts of several different counties.

PUMAs are especially useful for looking at characteristics in rural areas because, unlike many geographic areas such as counties, towns and places, they all meet the 65,000-population threshold that is needed to produce 1-year ACS estimates.

The PUMAs available in the ACS are the same as those used for the 5-Percent Public Use Microdata Sample in Census 2000, with one exception: Due to the population displacement in Louisiana caused by Hurricane Katrina, three PUMAs no longer met the 65,000-population threshold required for single-year estimates. Records for those three PUMAs were combined to meet confidentiality requirements.

Slide 9

PUMAs are identified by a five digit number that is unique within each state. If you are looking at PUMAs across multiple states, you will need to add state FIPS codes to the PUMA identifiers to create state-specific PUMA IDs.

The Census Bureau has also produced a series of geographic equivalency files—one for each state—that shows how PUMAs line up with other census geographies.

The existing PUMA boundaries and codes will continue to be used until after the 2010 Census, when PUMAs will be redefined.

Slide 10

Here are the boundaries for a PUMA in New York State, which is comprised of two counties: Seneca County and Tompkins County. You can find detailed maps of PUMA boundaries on the Census Bureau’s website.

Slide 11

Here are the boundaries for the PUMAs in Los Angeles County in California. If a county is larger than 200,000, generally that county will have more than one PUMA.

For those interested in producing your own PUMA maps, you can also access PUMA boundary files on the Census Bureau's website for use in GIS applications.

Slide 12

The Census Bureau has to take steps to protect the confidentiality of ACS respondents, as required under Title 13, because the PUMS data provide access to very detailed information about individuals and households. This is accomplished through several means:

- All personal identification, such as name and address, is stripped from the records.
- A small number of records are switched with similar records from neighboring areas through a process called "data swapping."
- Answers to open-ended questions, such as age, income, or housing unit value are top- or bottom-coded. For example, in the 2007 ACS, people who earned $10 million or more are grouped into a single income category. The same type of procedure is used to protect the identity of those at the bottom of the income scale.
- The Census Bureau also protects confidentiality by limiting the number of records in the PUMS file. The 2007 ACS PUMS file included about 1.2 million housing units, which represents a subset of the 2 million housing units that were interviewed.
- In exchange for the rich data available in the PUMS, the Census Bureau limits the geographic detail that is available. Data users can produce customized estimates for small population groups, but not for small geographic areas.

Slide 13

ACS PUMS files can be accessed through American FactFinder and analyzed using statistical software such as SAS or SPSS. In general, these files are too large to work with in spreadsheets, and such databases are not well suited to cross-tabulating data.

You can also access ACS PUMS data through the Census Bureau's DataFerrett system, which does not require knowledge or use of statistical software.

We will provide a brief overview of both of these options, but regardless of which method you use, you will need to refer to the ACS PUMS data dictionary. The data dictionary, available on the Census Bureau's website, is an indispensable tool that shows what is available, how each of the ACS variables has been coded, and where they are located in the raw data files.

Slide 14

Here is a detailed list of the housing topics covered in the ACS PUMS files.

Slide 15

Here is a list of the person topics covered in the ACS PUMS. The 2008 ACS PUMS—to be released in the fall of 2009—will also include new variables on health insurance coverage, marital history, and military service-connected disability.

Slide 16

If you are planning to use statistical software to analyze the data, the first thing you need to do is download the data.

To access the PUMS download page in the American FactFinder, you can go through the Data Sets page for the American Community Survey. Links are available in two places on the right side of the screen.

Slide 17

ACS population and housing data are available for download in American FactFinder as **two separate files**. If you plan to look at the population and housing unit data together, then you need to download both file types and then merge them together. For example, if you are interested in the percent of Hispanics living in rental units in the Bronx, you will need both the population and the housing data files.

The files are available in two basic formats: as ASCII text files with comma-delimited values) and as two versions of SAS datasets (PC-SAS files and UNIX files). Most statistical software can read files in at least one of these formats.

Finally, you need to choose the area for which you want the files. You can download data for the entire nation or any individual state. The national file is two *large* data sets 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 states that you need.

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

Slide 18

In addition to the data dictionaries found in the column labeled "Documentation," the American FactFinder provides links to several other useful resources on the ACS PUMS, including:

1. The state-specific values used in top- and bottom-coding the variables that required this type of confidentiality protection.

2. Detailed codes for the variables that contain a large number of coded responses, such as ancestry and occupation.

3. Links to the geographic equivalency files mentioned earlier. These are actually links to the Census 2000 PUMS files and documentation.

4. A PDF file containing information about the accuracy of the PUMS and methods of calculating sampling error and related measures.

Slide 19

If you are using both housing unit and person records, then the next step is to merge these two data sets together. You need to use the SERIALNO variable to merge the two files. The values of this variable are unique within a state and across states.

After combining the records, limiting the number of records you are processing by selecting those of interest will often increase processing speed.

Then, you can apply person or housing weights to your variables of interest to weight the individual records to the total population.

Statistical software such as SAS or SPSS will assist you with measuring sampling error and producing standard errors for the estimates in your analysis.

Slide 20

As we mentioned earlier, sampling error occurs when estimates are derived from a sample survey rather than from a census of the full population. Sampling error is measured using statistical techniques. Standard error is a measurement of sampling error. Each calculated estimate has an associated standard error that indicates the extent to which the estimate can be expected to deviate from the value you would get if a census of the full population had been taken. The underlying variability of the estimate itself and the geographic area under consideration determine the relative size of the standard error.

There are two basic methods to calculate standard errors in the ACS PUMS. The first method, using a generalized variance formula, is relatively simple but may produce less accurate results.

The second method, which uses a series of replicate weights, was first available with the release of the 2005 ACS PUMS. It is more complex but may produce more accurate results.

Both methods for calculating standard errors are described in detail in the *PUMS Accuracy of the Data* document provided on the Census Bureau's website. In calculating standard errors, you should

always refer to the documentation for the specific ACS dataset that you are using, because procedures can change from year to year.

Slide 21

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.

Data users need to provide an email address in order to use DataFerrett. The software can then be downloaded and installed on your computer. There is a tutorial available on the DataFerrett website for those who are interested in getting more information.

One drawback of using DataFerrett, as opposed to downloading the raw data, is that the software currently does not allow the user to calculate direct standard errors using replicate weights. A forthcoming version of DataFerrett will include this capability.

Slide 22

Here is a screen shot of the DataFerrett website. You can access the DataFerrett tool by going to http://dataferrett.census.gov.

On the right-hand side of the page, you will find links to download the DataFerrett application, as well as access the online tutorials and user guides.

Slide 23

ACS PUMS data have been available each year starting with the 2000 survey. The 2007 ACS 1-Year PUMS data were released in September 2008, and a 2005-2007 ACS 3-year PUMS file was released in January 2009.

Slide 24

You can find more information about the ACS PUMS on the American Community Survey website. Staff are developing an additional presentation on the ACS PUMS that will address more technical issues such as how to calculate standard errors using various methods.

One of the handbooks in *The ACS Compass Products* focuses on the PUMS. For more information, you can access the handbook titled *What Public Use Microdata Sample (PUMS) Data Users Need to Know* from the Internet address listed on the slide.

Please feel free to contact the Census Bureau if you have questions or need further information. If you have questions that are not answered by the Web site, please call 1-800-923-8282 or email acso.users.support@census.gov.