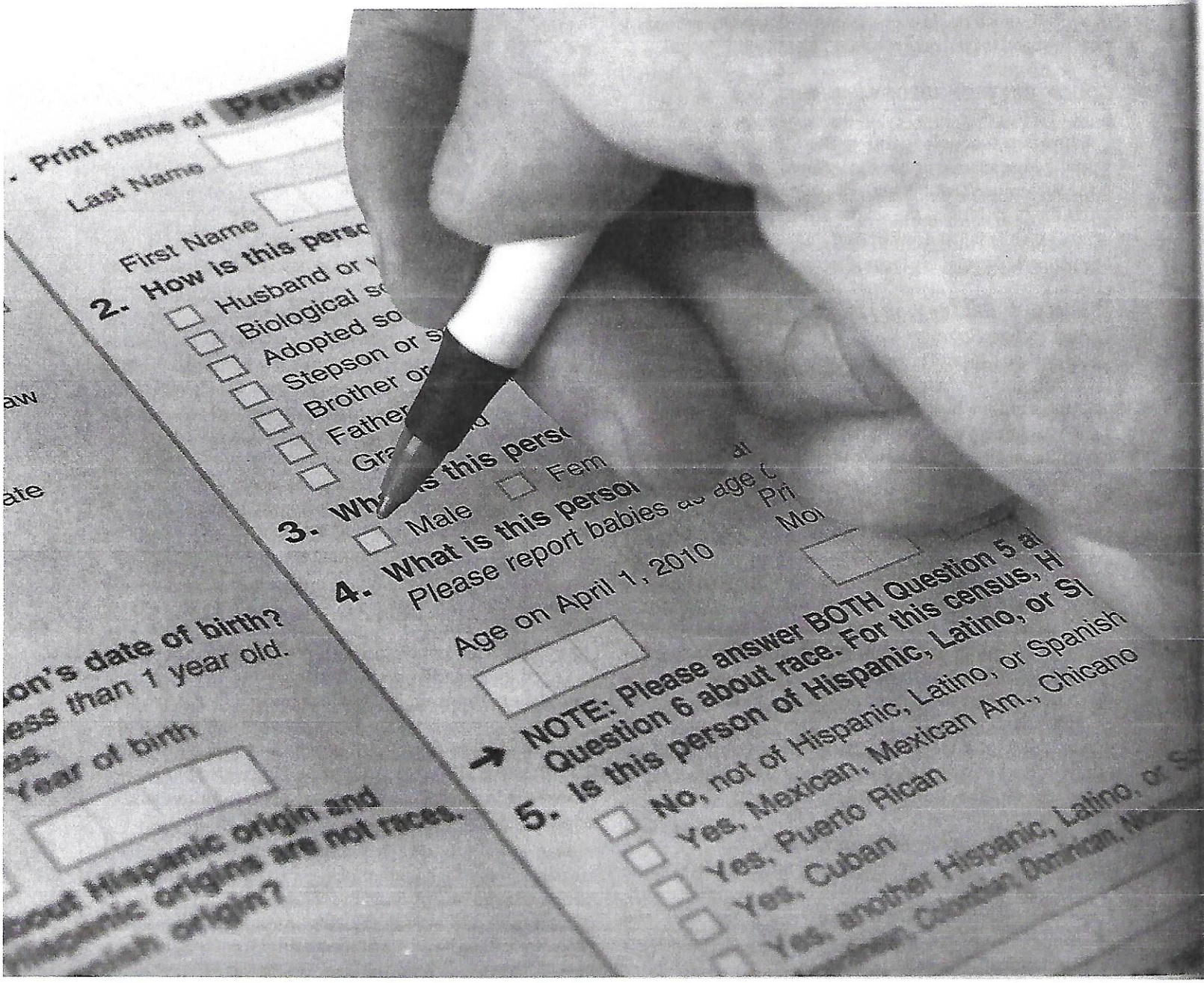
Delivering

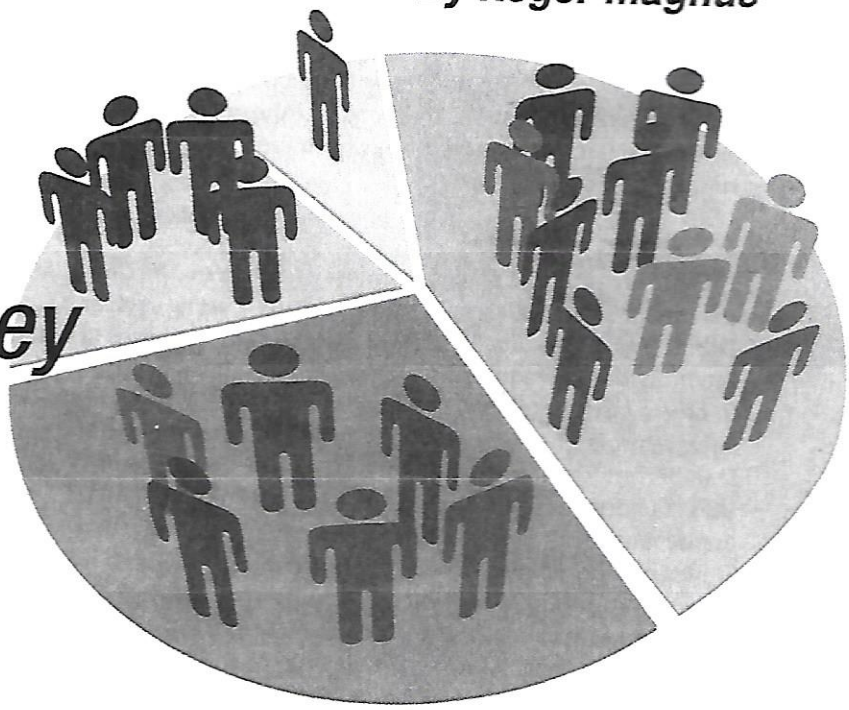
The U.S. Census Bureau's American



Data

Community Survey

By Roger Magnus



Big Data is grabbing headlines, but first you have to deliver the data. And one entity that will immerse you in data is the U.S. Census Bureau. The American Community Survey (ACS) is one of about 130 data programs it manages.

An ongoing monthly survey of households and group quarters in the United States and Puerto Rico that measures a wide variety of population, social, and economic variables, the ACS does *not* replace the official population counts measured by the Decennial Census (2000 and 2010, Summary File 1, or Short Form). It does replace the Decennial Census 2000, Summary File 3 (SF3), or Long Form, that measured detailed characteristics of the U.S. population, dating back to the 1940 Census. While the ACS and SF3 data are similar, there are several important differences:

- Data comes out more frequently (1, 3, or 5 years versus 10).
- There is a significantly smaller sample size.
- Participants are sampled continuously and not for a particular date.
- Data points are estimates with associated Margins of Error (MOEs): SF3 data had MOEs but was not published.

For searchers, the ACS can be intimidating. Why? Is it the hazy idea of having data available for multiple time periods (1, 3, and 5 years) depending on geographic level and population size? Is it the absurd notion of having a data point with not one but two numbers—an estimate plus or minus a MOE—and, more importantly, can I safely ignore the latter as I have no idea why it matters or what to do with it? How does this all of this relate to the Decennial Census anyway?

HIDDEN KNOWLEDGE GEMS

The ACS homepage (census.gov/programs-surveys/acs) contains many hidden knowledge gems about the program. For example, there is a section that details the rationale for each question on the survey, its history, and its applications for government (census.gov/acs/www/about/why-we-ask-each-question). Under the News and Updates link, there is a section for Data Releases that reveals Table and Geography changes. Click on the Data link to see the various places on the Census Bureau's website where you can find different ACS data packages and products. Guidance for Data Users contains information about survey subject areas, multiyear estimates, and advice on comparing ACS data, plus a series of helpful guidebooks.

Participation in the ACS, akin to the Decennial Census, is mandatory. Confidentiality of all responses and respondents is guaranteed by Title 13, Section 9 of the U.S. Code. For Housing Units, data is collected for 3 months from the time

a survey is first initiated and, for Group Quarters, 6 weeks. No Housing Units are sampled more than once in a 5-year period. Data is collected via four modes: mail, telephone, in-person, and internet; these vary depending on the type of recipient and the point reached in the survey collection process.

For the survey questions that request information back in time, the reference point is always from the date of the survey response or interview.

ACCESSING ACS DATA

There are multiple places to access ACS data on the Census Bureau's site. Datasets for more advanced users, such as Public User Microdata Sample, or PUMS (data responses at the individual person or household level that can be used to create custom tables), are available at the FTP site (census.gov/programs-surveys/acs/data/data-via-ftp.html) and DataFerrett (dataferrett.census.gov). American FactFinder (AFF; factfinder.census.gov) is sufficient for many researchers because this versatile interface contains the Detailed Tables and other summary-level data products needed to answer many data questions.

You can search for ACS data through the AFF interface using Guided Search, Advanced Search, or the Download Center. I advise selecting Advanced Search and then Topics (for every search, the latter is used at least twice to choose the dataset and table topic). First, select the ACS dataset desired, and it will be transferred to the search box. Next, select People or Housing. For example, under People, there are 15 categories including Age and Sex, Education, Employment, Origins (includes Race and Hispanic or Latino items), and Veterans. Drill down and select the link(s) needed. Each time you do this, the search results will become more targeted and decrease in number. The default geography is the entire country. To narrow your search, click on the Geographies button and go to the List header. Then enter a geographic entity or use the drill-down menu to select geographic level and name.

Each searching action opens a new window. To navigate more easily, close it after its contents have been loaded into the search box. You can delete items in a search by clicking on the X next to the search item. You can also clear the search entirely.

An infographic from the Census Bureau explaining the American Community Survey's measurement implications.

(census.gov/content/dam/Census/library/visualizations/2015/comm/YourAnswersYourFuture_1536.png)

GEOGRAPHIC SEARCHING

Prominent in the ACS are several major Census geography types (census.gov.geo/reference/terms.html).

- In 29 states (including many in the Northeast, where county governments don't exist) and Puerto Rico, towns and cities are often called County Subdivisions or Minor Civil Divisions instead of City, Town, or Place. An easy way to determine nomenclature is to enter a street address under the Address tab, which brings up a profile of Census geography types for that location.
- Census Tracts (population 1,200–8,000), Census Block Groups (population 600–3,000), and Census Blocks (any size population, though generally smaller than Census Block Groups) are all smaller, Census Bureau-created geographic areas based on counties.
- A ZCTA (ZIP Code Tabulation Area) is not identical to a ZIP code. ZCTAs are essentially based on majority ZIP codes for individual Census Blocks that are then aggregated.

Once you've created a table, there are multiple ways to alter, retain, and download it. However, it is important first to look directly underneath the title at the table's universe, or denominator, as sometimes it is not the total population. The wrong universe can throw off an interpretation of the numerator if your analysis involves proportions. For example, employment is only for the population 16 years and older, thus, the percentage of workers 25–29 years of age does not include all workers. This universe can be mentioned in the table's title as well.

The Modify Table link enables you check and uncheck selected rows and columns to add or subtract data points from a table. For some tables, you can select Transpose Rows and Columns (or flip them), which significantly improves the table's appearance. Under Bookmark/Save, you can create a URL for the table, copy the web address to your browser, and create a bookmark for future reference. You can also download the table into several formats, including PDF, Excel, and Rich Text Format.

TYPES OF ACS QUESTIONS

The ACS questionnaire for Households contains more than 70 questions and Group Quarters almost 50. From these answers, 1,000-plus Detailed Tables are derived. Table data cover a diverse list of subject areas, such as Age, Ancestry, Disability, Education, Employment, Gender, Home Features, Housing Rent, Income, Language, Poverty, Race, and Veterans, making these data extremely valuable for many purposes.

Currently, new ACS questions are determined jointly by the Census Bureau and the Office of Management and Budget (OMB) via a subcommittee that helps to justify and prioritize them. Questions can only be changed once each year, are developed jointly by the federal agency requesting the information and the Census Bureau, and are pretested with alternative word choices and sequencing in the larger questionnaire by the Census Bureau. The proposed question must receive final approval from the OMB; the whole process takes about 2 years.

To understand the data's scope and limitation, check the subject definitions (www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2014_ACSSubjectDefinitions.pdf). Most variables are provided not only with a definition but also with the rationale for asking, a history if the wording has changed over time, and guidance for comparing it to earlier years. Definitions are supplemented by Table Changes, updated each year, that show which tables are New, Modified (and the effect on prior year comparisons), or Deleted (with alternatives available).

MULTIYEAR ESTIMATES

Multiyear estimates enlarge the sample size, which lowers measurement error and increases the reliability or repeatable consistency of these numbers. The data compiled and published for each time period represents an average, respectively, of 12, 36, or 60 months.

There are different release dates and minimum population thresholds for each:

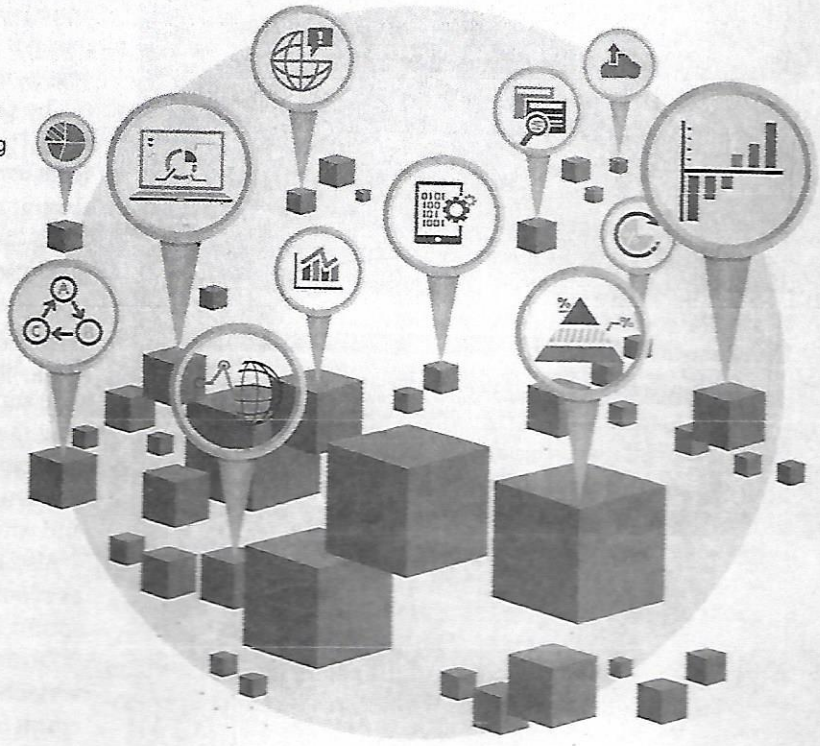
- 1-year estimates (typically in September) – 65,000 (In July 2016, the Census Bureau released, for the first time, 1-year estimates for geographies of 20,000 for 59 selected variables for 2014.)
- 3-year estimates (no longer being collected) – 20,000
- 5-year estimates (typically in December) – Census Block Groups (second smallest Census geography to Census Blocks that only have data available from the Decennial Census)

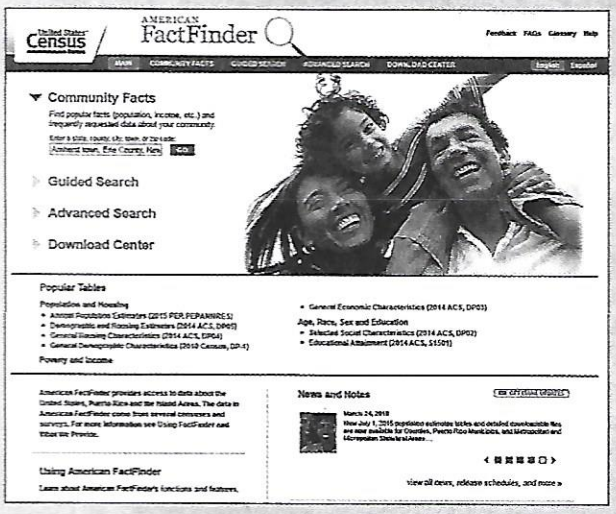
Larger places currently have both 1- and 5-year estimates collected, while in smaller places (a population of 64,999 or less), only 5-year estimates are collected.

Given a choice, should you use single-year estimates or multiyear estimates? Single-year estimates are preferred for larger geographic entities, when current data is the priority over more reliable numbers, and when the Coefficient of Variation (CV), which is the Standard Error/Estimate, is low, usually less than 15%.

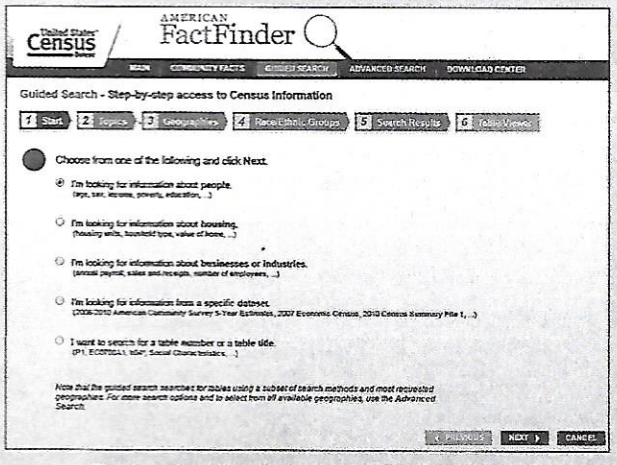
HISTORY OF THE ACS

The notion of "continuous measurement" was first suggested to the Census Bureau around 1990, and different proposals were considered in the early 1990s. Survey testing began with four U.S. counties in 1994 and was expanded over time. Finally, in 2005, the first full nationwide implementation was completed with a survey sample of 2.9 million households in the U.S. and 36,000 more in Puerto Rico. Group Quarters were not included until 2006, when 20,000 Group Quarters facilities and their residents were also surveyed. In 2011, the sample size for U.S. households was increased to 3.5 million. Multiyear estimates were first released in 2008 (2005–2007 3-year estimates). In 2010 came the first 5-year estimates (2005–2009). In 2015, it was announced that the 3-year estimates would stop being published (the 2011–2013 estimates were the last to be released), so that currently, only 1-year and 5-year estimates are compiled. In 2013, the Census Bureau added the internet as a response option and decided to survey college dormitories, a type of Group Quarters, only during the non-summer months when most students would be living there.

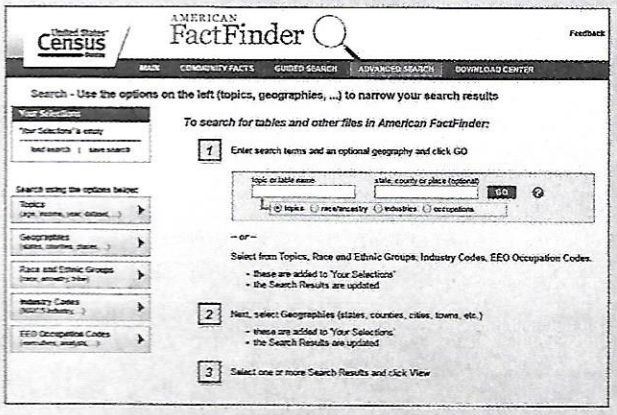




American FactFinder homepage



The American FactFinder Guided Search interface



The American FactFinder Advanced Search interface

Multiyear estimates are preferred for comparing smaller geographic entities to both other smaller and larger areas; for comparing a single area's non-overlapping time periods; when a larger geography's data has a high CV (often for a small subpopulation) and an area's data currency is less important than reliability; and for showing longer-term trends (1-year estimates can show more variable peaks and valleys).

DATA COMPARISONS

Comparing ACS data with other data can be a complex task influenced by many factors and multilevel considerations. On a macro level, two single-year ACS estimates or two non-overlapping multiple year estimates of the same number of years can be compared. However, even the straightforward, non-overlapping 5-year estimate comparisons between 2005–2009 and 2010–2014 have several added complexities:

- The addition of Group Quarters in 2006 in the second year of the first 5-year estimate
- Changes in legal geographic boundaries (counties, incorporated municipalities, Congressional Districts) between 2009 and 2013 and Statistical Geographies (such as Census Tracts) that are created by the Census Bureau for each Decennial Census
- Different Census years (2000 and 2010) for the population estimate controls on which the ACS total population data is based, meaning these population numbers (which have no MOEs) are not comparable

And all of the above do not include smaller level issues such as how coverage of the subject areas surveyed changes across time because of question changes such as altered phrasing, added or deleted options, new questions that cannot be compared back in time, and different time spans, different universes, residence rules, etc.

The Census Bureau publishes two guides about comparing data: "Comparing 2014 American Community Survey Data" (census.gov/programs-surveys/acs/guidance/comparing-acs-data.html/2014) and "Comparing 2005–2009 ACS 5-Year and 2010–2014 ACS 5-Year" (census.gov/programs-surveys/acs/guidance/comparing-acs-data/2014/5-year-comparison.html).

In these guides, the Census Bureau makes three suggestions: Compare, Compare with Caution, and Don't Compare. For Compare with Caution, the Census Bureau does not provide further explanation. My recommendation would be to read the comparison guides and the Subject Definition Guide to decipher the reasons why this is a gray area, and, if the comparison is made, explain in a note that it was done with Caution and why.

Also, for 5-year estimates, the geographic boundaries used are as of Jan. 1 of the final year to include new places that may have sprung up in the 5 years since data collection began.

On an even deeper level, income and other monetary-based variable comparisons require additional computations to account for inflation before making year-to-year comparisons. While income is automatically adjusted for the national infla-

tion rate over an average of 12 months from January through December for 1 year of data, inflation is not adjusted for other monetary variables such as rent or home value, nor are local and regional inflation rates provided that will differ from the national Consumer Price Index (CPI). Thankfully, there are instructions for making these complex, income-related comparisons (census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf).

WHY MARGINS OF ERROR ARE SO IMPORTANT

Other than total population numbers, which are updated each year by the Census Bureau's Population Estimates Program and used as controls for these ACS data points, all ACS data estimates are based on statistical samples and are accompanied by a plus/minus MOE. Don't ignore these MOEs or accompanying Standard Errors (SEs), as they are crucial to making data comparisons, deriving aggregated estimates, and computing proportions and products. From these MOEs, Confidence Intervals (Cis) can be computed that provide a range of values for an estimate.

The MOE explains the variation due to sampling errors (caused by the whole population not being surveyed) for any sampled estimate at a certain confidence level (the Census Bureau uses 90%). Interestingly, the MOE would be larger if it could measure non-sampling error, but this group of errors, which includes undercoverage or overcoverage of the sample, non-response, inaccurately measured responses, and post-data collection miscues, is not included in the calculation.

Sampling error is measured by the SE, which is the MOE/Z Score, and used for calculations. The MOE measures the extent of SE, and MOEs must be converted to SEs and then back to MOEs for the computations involving ACS data below. The Z Score measures in number of Standard Deviations (1.64 for the 90% confidence level), which is the square root of variance, the distance between a sample score in the range and the population mean. A higher confidence level (95%) has a larger MOE and associated Cis as well as a higher Z score (1.96).

A data comparison between a variable's estimates is made to see if there is a statistically significant difference either among different geographies (for example, number of 18–24-year-olds in two or more counties or cities), between single-year or multiyear estimates (number of 18–24-year-olds in X county for 2014 and 2013), between different cells of the same table for the same geography (for example, number of 18–24-year-old males and females in the same county in 2014), or between ACS estimates and Decennial Census counts (if comparable). Here is the simplest formula to use for estimates and MOEs:

Z = |est 1 – est 2|/ Sq. Root
(MOE est 1 squared + MOE est 2 squared)

For example, if you want to know if the total population for Amherst, Mass., is statistically different in the 2010–2014 5-year estimates (39,260 +/-39) from the 2005–2009 5-year estimates (35,971 +/-31), follow the steps below. Note: 5-year estimates, unlike 1-year estimates, are not controlled to the Population Estimates Program and do have associated MOEs. Also, technically, there are issues with this population comparison, so it is for illustration purposes only.

1. *Find the absolute value of the estimates:*
   39,260-35,971 = 3,289

2. *Square the MOEs and add up:*
   39 Squared + 31 Squared = 1,521+961 = 2,482

3. *Compute the Square Root of Step 2: 49.8197*

4. *Divide Step 1 by Step 3: 3,289/49.8197 = 66.018*

5. *If Step 4 >1, then the two estimates are statistically different; yes, as 66.018 >1.*

You can also compute statistical difference between two estimates using SEs and Z score, although this process is more involved:

Z = A - B/Sq. Root (SE A squared + SE B squared),
if Z>1.645 or <-1.645, then statistically significant.

There are several anomalies involving statistical comparisons:

- Controlled Estimates to the Population Estimates Program, where the MOE is 0 and indicated by three asterisks in the MOE.
- There are estimates that equal zero (very common) with a MOE. Use the MOE or, if multiple zero estimates, use the largest MOE once. This happens when a characteristic in the area sampled received no response but a different sample of the same area provided data. Thus, a MOE is estimated.
- Issues with median data for the table where there are not enough observations, etc., as indicated by *** in the MOE. These estimates cannot be compared.

Next, an aggregated estimate is created by combining categories (geographies, age ranges, income ranges, etc.) to form a new, larger estimate and MOE. This can also be useful if a MOE for one table cell (often a small geography) exceeds its estimate, as this aggregation will lower the overall table's CV.

For instance, if instead, you want to find out the combined populations from the 2010–2014 5-year estimates for Amherst, Mass. (39,260 +/- 39), and Northampton, Mass. (28,637 +/-27), from the 2010–2014 5-year estimates, derive a new number from both municipalities' Estimates and MOEs (converting to SEs and then back to MOEs)

1. *Add up the estimates: 39,260 + 28,637 = 67,897*

2. *Compute the SE for each place:*
   Amherst (39/1.645 = 23.708) and Northampton (27/1.645 = 16.413)

3. *Compute the aggregate SE:*
   Sq. Root (SE 1 squared + SE 2 squared) = Sq. Root (23.708 squared + 16.413 squared) = Sq. Root (562 + 269) = Sq. Root (831) = 28.827

4. *Derive the aggregate MOE by multiplying the aggregate SE by the Z score of 1.645: 28.827 * 1.645 = 47.42*

5. *For the final answer, add up the new estimate (Step 1) and use the aggregated MOE (Step 4): 67,897+/- 47*

You can use many of the same steps to compute a difference in estimates. For example, compute the 2010–2014 5-year estimate of the population in Amherst minus the population in the South Amherst CDP (5,267 +/-604) to find out the population for the rest of the town. All the steps are the same except for 1 and 5:

1. *Find the absolute value of the difference in estimates: 39,260 - 5,267 = 33,993*

2. *Compute the SE for each place:*
   Amherst = 23.708; South Amherst CDP: 604/1.645 = 367.173

3. *Compute the aggregate SE:*
   Sq. Root (562 + 134,816) = Sq. Root (135,378) = 367.938

4. *Compute the aggregate MOE: 1.645 * 367.938 = 685.258*

5. *Subtract the new estimate (Step 1) and use the aggregate MOE (Step 4): 33,993 +/- 685*

Note that the aggregate MOE went up mainly due to the smaller area's much higher MOE.

There are two types of proportions or percentages that can be calculated with ACS MOEs:

- Numerator is a subset of the denominator
- Numerator is independent of the denominator.

In both cases, the data is made up of raw numbers and not percentages. If the numerator is a percentage of the denominator, then we will want to compute a Product below.

For the first example, we want to know using the 2010–2014 5-year estimates the percent of male 18–19-year-olds (4,206, +/- 368) in Amherst, Mass., to the total population:

*SE (proportion) = 1/Y * Sq. Root (SE X Squared) - X Squared/Y Squared * (SE Y Squared))*

1. *Calculate the proportion of the two estimates: 4,206/39,260 = .107 or 10.7%*

2. *Calculate the SE for 18–19-year-olds in Amherst: MOE/Z = 368/1.645 = 223.708*

3. *Calculate the SE for total population in Amherst: 23.708*

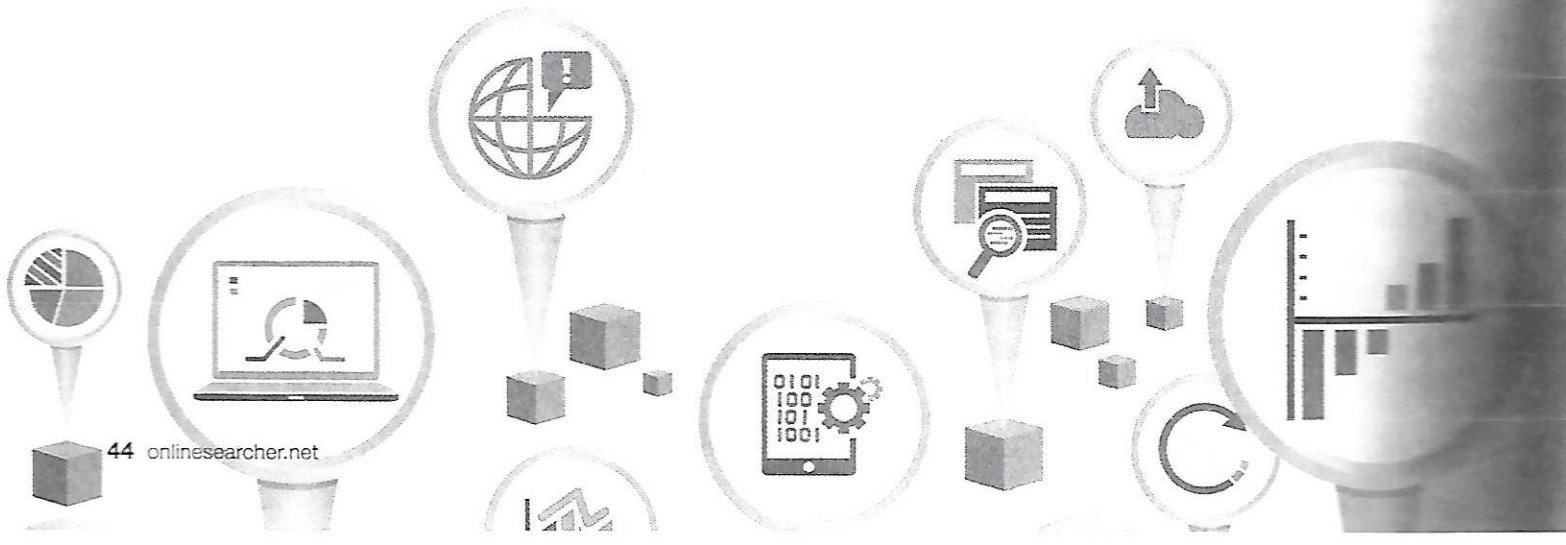
4. *Calculate the Aggregate SE:*
   *Aggregate SE (10.7%) = 100% * (1/39,260) * Sq. Root (223.708 Squared - .107 Squared * 23.708 Squared) = 100% * .0000254 * Sq. Root (50,045 - .011*562); =100% *.0000254 * Sq. Root (50,045-6.182); = 100% * .0000254 * 223.693 = .568*

5. *Compute Aggregate MOE: Multiply Aggregate SE * Z Score = .568 * 1.645 = .934 = 9.34%*

6. *Final answer: 10.7% +/-9.34%*

For the second example, we want to know the number of 18–19-year-old males in Amherst, Mass., compared to the number of 18–19-year-old-females. The formula used is almost identical and the process is exactly the same:

*SE (proportion) = 1/Y * Sq. Root (SE X Squared) + X Squared/Y Squared * (SE Y Squared))*



There is one more special case. If you want to compute a numerator when it is a subset of a denominator, but only have a percentage instead of a raw number, compute a numerator's Product by multiplying the percent by the denominator. For example, you would do this to compute the number of 20–24-year-olds in Amherst where you only know the percentage of this population segment, not the total population's estimate and MOE.

Finally, as a test of data quality for each table that is part of the 1- and 3-year estimates, the Census Bureau computes the CV (SE/Estimate) for each cell. If the median percent for all cells exceeds 61%, the table is collapsed into larger subgroups or geographies and the test is run again. If the collapsed table also exceeds this threshold, the table is not published. Although the Census Bureau does not provide exact guidance on what is a "good" overall CV for a table, most statistical experts suggest no more than 15%. If a table is above this threshold, it is best to combine data points, aggregate geography, or use a different table. Here is one additional test: If more than half of the estimates in any table equal zero, the table is not published.

For more information on any of these formulas, go to www2.census.gov/programs-surveys/acs/tech_docs/accuracy/ACS_Accuracy_of_Data_2014.pdf and refer to pages 26–30.

DATA LITERACY

While ACS data is fraught with conceptual difficulties (small sample size, multiyear estimates, complexities in making comparisons, the difficulties of incorporating MOEs for testing real difference in numbers and making computations, etc.) that sometimes seemingly require a team of statisticians to untangle, the topic area breadth, geographic depth, and currency of ACS data make learning how to use and understand these nuances a must for researchers needing to find, interpret, and explain these numbers for their patrons and clients.

However, like any data repository, it is important not to become so immersed in the "small details" that you lose sight of the "big picture" of the ACS's limitations—what it doesn't cover or what other datasets cover this topic better. Living in a Big Data world, the ACS pushes us to our data literacy outer limits and beyond. However, this ultimately makes you a better data researcher. You can apply and expand on what you know from using ACS and Census Bureau data whenever you encounter new datasets and how to make sense of them to inform the question or issue at hand.

Roger Magnus *(roger@rogermagnusresearch.com) owns Roger Magnus Research.*

Comments? Email the editor-in-chief (marydee@xmission.com).