THE AMERICAN COMMUNITY SURVEY

6.1 INTRODUCTION

The American Community Survey (ACS) is going to be your primary source for current, detailed socioeconomic characteristics of the population and the nation's housing units. Formally launched in 2005, it replaced the decennial census long form so that this data could be produced annually at a lesser cost. All the basic demographic variables that are published in the decennial census are included in the ACS, plus educational enrollment and attainment, veteran status, income, marital status, labor force status, place of origin and birth, commuting time to work, housing costs, detailed physical characteristics of housing units, and more. Each question in the ACS satisfies some requirement in federal law, and the data is used for distributing federal aid to state and local governments and individuals. Like the decennial census, ACS data is published for most legal and statistical areas in the country.

The ACS is a sample survey, and the statistics are estimates published with a range of possible values. With the decennial census, we can say that 5,716 people lived in Taos, New Mexico, on April 1, 2010. With the ACS, we can say that we are 90% confident that 5,735 people (plus or minus 249 people) lived in Taos between 2012 and 2016. The estimates are more challenging to interpret and work with, and the margin of error (MOE) can be high enough that estimates for small population groups or small geographies can be too unreliable for practical use. Creating derivatives like aggregates and percent totals is more laborious, as recomputing the MOE requires formulas that may seem complex to new users. Some researchers and the media simply ignore

the error margins and present the estimates as if they are counts, which is a terribly misleading and inappropriate use of the data.

We begin this chapter with a brief discussion of how and why the ACS was created, followed by a detailed overview of how it differs from the decennial census, and what the benefits and challenges of using it are. There are a lot of variables in the ACS, and we will cover the major categories and how you can identify the different subjects. Understanding how to interpret the estimates and create derivatives is crucial for working with this dataset, so in the exercises, we will cover the formulas for recalculating MOEs in great detail.

The program page for the ACS is continuously updated. With every new period-estimate release, there is new documentation on subjects, variables, and methodology. Compared with the other datasets, more documentation is devoted to using and interpreting the data: `https://www.census.gov/programs-surveys/acs/`

When Do You Use the ACS?

The ACS is a sample survey of 3.5 million addresses that is published annually as 1-year- and 5-year-period estimates. Estimates are published at a 90% confidence interval with MOEs provided for each value. Functionally useful data is available down to the census tract level for the 5-year dataset and for any geography that has at least 65k people in the 1-year dataset. When should you use the ACS relative to another dataset? When

- you need broad and detailed socioeconomic characteristics of the population;
- you need the most recent data that's available for these characteristics;
- you need data for small geographies like census tracts, Zip Code Tabulation Areas (ZCTAs), and subcounty divisions that is not available from other federal statistical sources; or
- you want a data source that includes many different variables that are collected using the same methodology and time frame.

The ACS will *not* be your first choice if you need the following:

- Actual counts of the population (use the decennial census instead)
- Actual counts of socioeconomic characteristics at the state or county level (search for administrative data sources from other federal agencies)

- Data for the smallest census geographies like blocks and block groups (use the decennial census instead)
- To study basic demographic variables on an annual basis (use the Population Estimates Program [PEP] instead)
- National data on a monthly basis (use the Current Population Survey [CPS] instead)

6.2 FUNDAMENTALS OF THE ACS

In this section, we'll discuss the origins of the ACS, the basics of how the survey is conducted, and the fundamentals for understanding how this dataset of sample-based estimates differs from the basic decennial census count. We will see how to interpret the estimates and MOEs and decide when to use one period estimate versus the other. Last, we will grapple with the challenges and shortcomings of the ACS, so that you can learn to interpret and use the data correctly.

Origins of the ACS

In the early 1990s, against a backdrop of undercounting controversies and a desire to shrink costs, Congress passed a bill that mandated that the Census Bureau investigate ways to improve census taking. The Census Bureau began experimenting with the continuous rolling sample method. The idea was that data could be collected from sample surveys on a monthly basis and then could be rolled up to create a pool of samples for generating different period estimates. Estimates for large geographies and heavily populated areas could be created annually from 12 months of sample data. For smaller areas, you could accumulate a larger pool of samples over a longer time period and publish estimates for the longer interval. These intervals could be updated each year by refreshing the sample pool, dropping out an older period of data while adding a newer period. Eventually, the entire population would be sampled using this process.

Experiments were conducted throughout the 1990s using the continuous rolling sample concept, and test data was compared with the 2000 census count. The Bureau and stakeholders liked the results of the tests, and Congress authorized funds to expand the ACS in the early 2000s. The first official ACS numbers were published in 2005 for geographies that had at least 65k people. In 2007, data was published in 3-year intervals for geographies that had at least 20k people, and by 2009, 5-year data was published for all geographies down to the census tract level. By the time the

2010 census was conducted, there was enough ACS data to publish detailed 5-year estimates on an annual basis, and the decennial long form was discontinued.

There are a number of factors that led to the creation and adoption of the ACS (Hayslett & Kellam, 2010; Herman, 2008). The first was the expense that was involved in conducting the decennial census. Every decade, there was a large ramp up in hiring and costs followed by a large ramp down once the census was complete. By conducting a survey on an ongoing basis, the Census Bureau could hire a smaller number of full-time experts to keep the program going, which was less expensive than hiring, training, and letting go a large body of people. The ACS also uses a much smaller sample; it captures about two to three people out of a 100, whereas the decennial long form captured an enormous one in six people. The smaller sample size alone would generate savings as there are fewer people to survey and follow up with.

Second, placing the collection of detailed socioeconomic statistics into a separate program allowed the decennial census to focus more on its original mission and temporarily satisfied concerns from conservative members in Congress that the census had become too burdensome as it asked too many questions. With the ACS, fewer people would be surveyed over a longer period of time.

Third, it satisfied the needs of many stakeholders and census data users who argued that the value of the decennial census decreased as the decade wore on, as the data became less useful for studying contemporary issues. The other census data sources were either too limited in the scope of their variables (the PEP) or in their level of geographic detail (the CPS) to serve as replacements. The ACS provided detailed variables and geography on an ongoing basis and used a similar definitional framework as the decennial census for subjects and geography.

Sampling and Estimate Creation

The ACS is a continuous rolling sample survey of approximately 292,000 residential addresses a month, adding up to 3.5 million addresses a year. The sample is stratified (broken into groups) to decrease variability in the sample selection (U.S. Census Bureau, 2017c, 2017d). The sample is assembled in stages. In the first stage, any housing units that were already sampled in the past 4 years are excluded, and 20% of all addresses that are included in the sample are "new" addresses that have never appeared in a previous extract. In the second stage, 16 strata or groupings are applied based on geographic size. The strata are sorted by the number of addresses in each county by stratum and geographic order, including tract, block, street name, and house number. A census block is assigned to a stratum proportionately based on information about the set of geographic entities that contain the block. These

entities include counties, incorporated places, minor civil divisions with functioning governments, Native American Areas, urban/rural status, and a few others. The purpose of this is to prevent creating a sample that happens to be concentrated in one part of the country—for example, a sample of addresses that is largely concentrated in California and Texas without any households from Alaska and Wyoming. In addition, approximately 2.5% of all group quarters addresses are sampled in a year and are divided into a small and large group quarters stratum based on the size of the quarters.

The questions that appear on the ACS form are largely fixed at the beginning of the decade, when subjects and questions for both the decennial census and ACS are approved. Households and group quarters that are included in a given sample are notified by mail, and they have the option of submitting a paper form or filling it out online. Completing the ACS is required by law, but in contrast to the decennial census, the Census Bureau follows up only with a sample of nonrespondents. In 2016, out of the 3,527,047 addresses that were included in the survey, only 2,229,872 (63%) were used as the base for generating the estimates. This is referred to as the coverage rate and is published in every ACS dataset in Table B98011.

After a year's worth of responses have been received, the Census Bureau uses this sample pool to create population estimates for each characteristic and geography. Similar to the decennial census, the Census Bureau uses imputation techniques to assign and allocate responses that are in error or are missing. Allocations for each characteristic are published in tables that begin with B98 and B99. The actual number of persons and housing units that are in the sample are published in Tables B00001 and B00002, respectively. This is referred to as the unweighted sample.

The Bureau assigns weights to each sample to quantify the number of people in the total population who are represented by a sampled household or individual. In their thorough review of sources of error and uncertainty in the ACS, Spielman and colleagues (2014) use the example of an Asian male who earns a certain amount of income. If that category were assigned a weight of 50, then the population for an area that includes one Asian man in that bracket in the sample would equal 50 Asian men in the same income bracket for the total population. Naturally, a good method for estimating survey weights is necessary for creating the estimates, and the process is a complex one that must be constantly updated. In 2016, weighting areas were built from collections of whole counties with similar demographic and social characteristics using data from the 2010 Census and the 2007–2011 ACS (U.S. Census Bureau, 2017c, 2017d). Additionally, the Census Bureau uses estimates from the PEP by age, sex, race, and Hispanic origin as controls for what the total population should be.

This ensures that the ACS estimates for counties, incorporated places, and minor civil divisons do not exceed these population thresholds.

Interpreting Period Estimates

Because the ACS sample size is much smaller than the old decennial long form, all ACS estimates are published as intervals at a 90% confidence level. The MOE describes the precision of an estimate at the given level of confidence—that is, the likelihood that the sample estimate is within a certain distance from the population value. The confidence level is the degree of certainty that the interval defined by the MOE contains the actual value of the characteristic. Essentially, ACS values represent likely distributions: The MOEs represent the spread of the distribution, the estimate is the center of the spread, and the confidence level indicates the likelihood that the actual population falls within the estimated distribution.

For example, we are 90% confident that the population of South Dakota's largest city, Sioux Falls, was 174,350 in 2016, ±38. The MOE of 38 people means that the actual population could be as low as 174,312 or as high as 174,388. The Census Bureau publishes estimates with MOEs as $174,350 \pm 38$. Alternatively, if the actual range of possible values is presented (174,312 174,388), this is referred to as the confidence *interval*. The confidence *level* means that there is a 90% chance that the true estimate lies within this interval and a 10% chance that it lies outside the interval.

The MOE measures the variation in the random samples due to chance, and its size is affected by three factors. First, the MOE increases as the size of a sample decreases. With a smaller sample, we cannot estimate what the true population is at the same level of precision. If we wanted a more precise estimate, we would need a bigger sample. Second, increasing the confidence level (to say 95% or 99%) while keeping the sample size the same would also increase the MOE. We would be more certain that the true population values fall within the interval, but the range of the interval would be greater. Last, the amount of variability in the population can also increase the size of the MOE. The more variable a population is, the harder it is to estimate. To decrease the MOE, we would need to increase the sample size or decrease the level of confidence. Decreasing the level of confidence is not desirable; in statistics, a confidence level of 90% is typically the lowest acceptable threshold.

Given the size of the ACS sample and a confidence level of 90%, the Census Bureau cannot estimate the population for small areas on an annual basis with an acceptable level of precision. So each year, they release two different period estimates: (1) a 1-year estimate for geographies that have a population of at least 65k and (2) a 5-year estimate for all geographies down to the census block group level. By

TABLE 6.1 ⬢ COMPARISON OF ACS ESTIMATES FOR SOUTH DAKOTA CITIES

	Sioux Falls	Rapid City	Pierre
2016, 1-Year			
Total population	174,350 ± 38	74,050 ± 26	NA
Population 16 and over	133,814 ± 1,414	58,144 ± 1,279	NA
Civilian labor force	99,531 ± 2,462	36,222 ± 1,802	NA
2012–2016, 5-Year			
Total population	167,884 ± 45	72,441 ± 103	13,959 ± 25
Population 16 and over	130,318 ± 490	57,021 ± 388	11,231 ± 126
Civilian labor force	96,330 ± 1,030	37,059 ± 765	7,795 ± 262

publishing 5-year estimates, the Census Bureau is able to use a much larger sample (60 months of data instead of 12), which it can use to more reliably estimate the population for these smaller areas. With a larger sample, the MOEs will be lower and will be acceptable for most (but not all) applications, depending on the size of the geography and population group.

The Bureau used to publish 3-year estimates (36 months) for any geographies that had at least 20k people, but they dropped this alternative beginning with the 2014 ACS and replaced it with 1-year supplemental estimates. This supplemental 1-year data is provided for all geographies that have at least 20k people and uses larger subject and interval groupings so that the MOEs are not unacceptably large. There are a smaller number of tables in this series, and their table ID numbers are prefaced with the letter "K."

Table 6.1 illustrates the differences between different ACS periods and population subgroups for the two largest cities in South Dakota and the state capital. Since Sioux Falls and Rapid City have populations above the 65k threshold, they are included in both the 1-year- and 5-year-period estimates, while Pierre falls below the threshold and is only available in the 5-year ACS. Notice that the MOEs for the total populations for all cities is pretty small for both period estimates. This is because total population for counties and incorporated places is controlled in the ACS using PEP data, so the values do not exceed those estimates by a large degree. The MOEs become much larger once you start looking at subgroups of the population. As we move

TABLE 6.2 ⬢ DISTINGUISHING FEATURES BETWEEN ACS PERIOD ESTIMATES

1-Year Estimates	1-Year Supplemental Estimates	5-Year Estimates
12 months of data	12 months of data	60 months of data
Data for areas with populations of 65,000+	Data for areas with populations of 20,000+	Data for all areas down to block groups
Smallest sample size	Smallest sample size	Largest sample size
Less reliable than 5-year	Less reliable than 5-year	Most reliable
Most current data	Most current data	Least current data
Released 2005–present	Released 2014–present	Released 2009–present
Use when currency is more important than precision, analyzing large populations	Use when currency is more important than precision, analyzing smaller populations, examining smaller geographies not available in the standard 1-year estimates	Use when precision is more important than currency, analyzing very small populations, and examining tracts and other smaller geographies not available in the 1-year estimates

Source: Derived from `https://www.census.gov/programs-surveys/acs/guidance/estimates.html`
Note: When comparing geographies, use a period estimate that includes them all.

from the total population, to the population that's over 16, to the population over 16 in the civilian labor force, the MOEs increase. If we look at the same characteristic in two different period estimates, there is a big difference in the size of the intervals. In 2016, the civilian labor force for Sioux Falls was estimated to be 99,531 ± 2,462, but in 2012–2016, the estimate was 96,330 ± 1,030. Since the 5-year data is based on a larger sample, we can get an estimate that has a lower MOE.

When should you use the 1-year estimate versus the 5-year estimate? If you were making comparisons between places and not all the places were included in the 1-year dataset, then it's better to use the 5-year dataset as mixing and matching places from two different periods is not ideal. If we were comparing all three of these cities, we should use the 5-year dataset, as Pierre is not included in the 1-year dataset. What if we were comparing just Sioux Falls and Rapid City? Then we would have a choice. If we wanted our estimate to be more precise in terms of the size of the interval, we would use the 5-year period as the MOE is much lower. If we want our estimate to be more precise in terms of the time frame (we want data that's most current), then we would use the 1-year estimate. Table 6.2 provides guidance on which period estimate to choose.

Every year, the Census Bureau releases new 1-year-period estimates at the beginning of the fall and new 5-year-period estimates at the end of the fall. There is a lag of about 9 months between the time data collection for a year ends and the time data is published; data collection for 2016 was finished by January 2017, and the estimates were released in September 2017, while the 5-year 2012–2016 data was released in December 2017.

For the 5-year estimates, the estimate is recomputed by dropping the oldest year of data from the sample and adding the latest year. Geographies that cross the threshold of 65k will appear in the 1-year estimates for the year they cross the threshold, and geographies that fall below the threshold will be dropped. The statistical geographies for the ACS are based on the decennial census geographies that were current during the latest year of the estimate. All ACS data from 2010 to 2019 is based on the 2010 census tracts, block groups, ZCTAs, census-designated places, and so on (Public Use Microdata Areas [PUMAs] are a notable exception, as they are delineated a couple of years after the decennial census). For 5-year estimates, the samples are recalculated to fit the latest statistical geography: While the 2015–2019 ACS will use 2010 census geography, the 2016–2020 ACS will use 2020 geography. Legal areas (states, counties, incorporated places) are different, in that ACS data is tabulated based on the latest boundaries from the Boundary and Annexation Survey. Dollar values in the 5-year ACS are based on the Consumer Price Index for the most recent year in the sample.

Challenges and Caveats

While the ACS offers a number of benefits over the decennial long form (it's more timely, less expensive to conduct), there are also a number of disadvantages. These challenges include greater complexity in interpreting, understanding, and working with the estimates and less reliability in the precision of the statistics, especially for small geographies and population groups. Many of these issues and approaches for addressing them were identified soon after the release of the first ACS datasets (National Research Council, 2007).

Table 6.3 shows the 2010 census count and the 5-year 2008–2012 ACS population estimates by race for a census tract in midtown Manhattan. The population for the tract was 7,614 people on April 1, 2010, and was between 5,566 and 7,200 (with 90% confidence) during the period from 2008 to 2012. The difference between these two values is rather stark, in the first case, we have a simple total for a single point in time, and in the second case, we have a fuzzier range of values over a longer period. The ACS estimate may be lower than the actual count simply due to the methodological differences between the count and the survey or due to a longer time range (before or after 2010) when the population may have been lower. This underscores

TABLE 6.3 ⬢ COMPARISON OF 2010 CENSUS AND ACS ESTIMATE FOR CENSUS TRACT 68, NEW YORK COUNTY

	2010 Census	2008–2012 ACS
Total population	7,614	6,383 ± 817
White	5,414	4,716 ± 735
Black	330	297 ± 187
Native American	9	0 ± 17
Asian	1,080	725 ± 295
Pacific Islander	3	0 ± 17
Some Other Race	30	85 ± 137
Multiracial	188	91 ± 75
Hispanic/Latino	560	469 ± 186

Note: Data by race is presented for people of that race alone who are not Hispanic, while Hispanic / Latino includes all Hispanics regardless of their race.

that the decennial census is designed to get an actual count, while the ACS attempts to generally characterize an area over a period of time.

The size of the MOE is a big problem for the smaller population groups. Since white people represent a majority of this census tract and their overall population is high, the MOE for this group is lower. There are 4,716 whites, within a range of 3,981 and 5,451. The MOE for black people is much higher relative to the estimate: The estimate is 297 but could be 110 or 484. The MOE for Some Other Race is actually higher than the estimate itself, so the true value could be zero! A statistic called the coefficient of variation measures the relative amount of sampling error associated with an estimate and is typically used to determine whether a particular estimate is reliable enough to use. You can compute the coefficient of variation (CV) for an ACS estimate with this formula (which we will do in the exercises):

$$\text{CV} = \frac{\text{MOE}/1.645}{\text{ACS Estimate}} * 100$$

What is an acceptable CV? Opinions vary. In a national study, the Census Bureau examined CVs at the county level for key ACS variables and grouped the findings into three categories: (1) estimates with CVs of 0 to 12 were highly reliable, (2) 12 to 34 were of medium reliability, and (3) 35 and above were of low reliability (Heimel, 2014). In its web applications and published tables, the Population Division of the New York City (NYC) Department of City Planning

flags all estimates with a CV higher than 20 as unreliable. For this particular census tract, the only reliable estimates based on any standard would be the total and the white populations that have a CV of 8 and 10, respectively. The CV for the Hispanic population is 24, the Asian population is 25, and the black population is 38.

What can you do when estimates are unreliable? The solution is to increase the size of the sample. For end users working with published census data, this means either aggregating categories into larger groups or geographies into larger areas. In some cases, aggregating groups is acceptable; for instance, the number of households classified by income brackets can be grouped into a smaller number of brackets with larger intervals. In this example, our options are limited. We could aggregate the smaller racial groups into another category if we were studying this one census tract, but census tracts are seldom studied individually. If we were studying all of Manhattan or NYC, there would be areas where the black and Hispanic populations are higher and the white and Asian populations are lower, so we can't aggregate these groups if we want to make comparisons. The other option would be to aggregate the census tracts into larger areas (which we will do in our first exercise) or use a larger statistical geography like PUMAs. This would increase the accuracy of the estimate by reducing the MOE, but the trade-off is that the geography is less detailed.

For example, in 2008–2012, neighboring Census Tract 64 had a total population of $7,439 \pm 621$ with a CV of 5. If we combine its population with Census Tract 68, we have a total population of 13,822. To calculate the new MOE, we take the square root of the sum of the squares for each of the MOEs:

$$\text{MOE} = \sqrt{\text{MOE1}^2 + \text{MOE2}^2}$$

This gives us a total population of $13,822 \pm 1,026$. The CV for this new estimate is 4.5, which is a little lower than the CVs for the total population for each individual tract (8 for Tract 68 and 5 for Tract 64). We can get an even better estimate if we aggregate additional tracts.

NYC's Population Division has created aggregates of census tracts called neighborhood tabulation areas to facilitate the use of ACS data. However, the aggregation of tracts in smaller cities and towns may not make sense for local planners and policy makers, as tracts can cross the town's boundaries and jurisdictional areas (Salvo & Lobo, 2010). Even within large cities, aggregating tract-level data can mask small subgroups of the population that researchers are trying to study. For researchers who are studying marginalized populations and poverty, this data may be visible at the tract level but becomes diluted if you look at a larger area like a ZCTA or PUMA,

as this smaller population gets combined with a larger population with different characteristics (Bazuin & Fraser, 2013).

The amount of uncertainty in the ACS relative to the old census long form has exceeded the Census Bureau's expectations (Spielman et al., 2014). MOEs in the ACS are higher due to the smaller sample size, the decision to follow up with only a sample of nonresponding households, and the use of population controls that are not collected in conjunction with the ACS. In their review of the patterns and causes of uncertainty in the ACS, Spielman and colleagues (2014) state that block group and tract-level data in the ACS "fail to meet even the loosest standards of data quality" (p. 147). The sample sizes of the ACS simply aren't large enough to provide reasonable certainty for estimating the characteristics of census tracts (Salvo et al., 2007). To some degree, this is because the sample itself is too small, but it's also due to the degree of variability that exists for several demographic characteristics at that scale. Samples of the same size can yield different estimates based on variations in the composition of populations in census tracts.

Bazuin and Fraser (2013) conducted a case study of a largely black census tract in Nashville, Tennessee, whose poverty declined sharply between the 2000 census and 2005–2009 ACS. They administered their own survey of the area and compared their results with the 2000 decennial census count and with estimates from the ACS. They discovered that the small sample size, the variability in characteristics of the tract, and nonresponse rates were skewing the poverty rate lower. The ACS was capturing both the elderly black population—homeowners who were more likely to be retired, be home during the day, own landline telephones, and respond to surveys—and a small younger group of more affluent whites. It was missing poorer black renters who had children and was overestimating the number of vacant homes and underestimating average household size.

The U.S. Government Accountability Office (2016) found that there is a lack of data for measuring income for small communities, as the ACS is either unreliable at that scale or is tabulated for geographic areas that do not conform to the project area's boundaries. The greater complexity of interpreting the ACS has created barriers to applying it correctly. A survey of three federal agencies that use ACS data to fulfill regulatory requirements for distributing funds found that none of them take the MOE into account when using the data (Nesse & Rahe, 2015). Regulators were concerned about the complexity of incorporating calculations to account for the MOE and were worried that stakeholders would be confused by what it was.

The Census Bureau has taken steps to improve the accuracy of the ACS, with better nonresponse follow-up, adjustments in creating weights, and more publicity

to encourage people to respond. The sample size was originally fixed at 3 million addresses, but as the U.S. population grew, the sample represented a smaller portion of the population. The sample size was increased to 3.5 million beginning with the 2012 ACS and has increased slightly each year. In 2016, 3.54 million addresses were included in the sample, but it's important to remember that this number represents the addresses that were selected for inclusion. The actual number that was finally interviewed excludes the following addresses: determined to be nonexistent or commercial, not selected in the subsample for personal visit follow-up, and not interviewed due to refusals or other reasons. In 2016, 2.23 million addresses were used to actually compute the estimates. For the group quarters population, just over 200k people are included in the initial sample and about 160k are used to create the estimates. The Census Bureau is transparent about the ACS methodology, which you can read in detail at `https://www.census.gov/programs-surveys/acs/methodology.html`.

Continuous improvements to the Master Address File should boost the number of addresses that make it to the final interview. However, without a substantive increase in the sample size, reliability issues with the smallest geographies will remain. Data at the block group level is so poor that users should disregard it entirely. Data at the census tract level varies in quality, and you should scrutinize the MOE and calculate CVs to determine reliability. Despite these challenges, the ACS is still a valuable dataset that is broad in scope and geographically detailed. In the exercises, we will cover how you can assess the quality of the estimates and use formulas to create aggregates. InfoBox 6.1 contains suggestions for tools that you can use to make your work with ACS formulas a bit easier for either automating the process or checking your work.

6.3 ACS VARIABLES

The ACS has a wealth of variables compared with the current decennial census. It includes all the questions asked in the current decennial census, most of the questions asked on the old decennial long form last used in 2000, and new questions on topics such as computer and internet access and health insurance coverage that were added due to new federal laws.

There are a few key conceptual differences between the decennial census and ACS. First, the concept of residency for the ACS is based on "current residence," which is where the respondents are currently living at the time they receive the ACS questionnaire. If they are living there at that time and intend to stay there for at least

INFOBOX 6.1 TOOLS FOR WORKING WITH ACS FORMULAS

Cornell PAD ACS Calculator: One calculator for comparing two values to test for significant difference and another for computing a new value (estimate and MOE) for two estimates. Great for checking your own work. https://pad.human.cornell.edu/acscalc/

Census Bureau's Statistical Testing Tool: A spreadsheet that allows you to compare estimates for more than 3,000 pairs of values to determine whether they are significantly different or not. https://census.gov/programs-surveys/acs/guidance/statistical-testing-tool.html

Fairfax County VA ACS Tools: As part of their Research Tools, the county has developed individual spreadsheets for calculating specific formulas like CVs, statistical difference, percent change, percent total, and aggregates. https://www.fairfaxcounty.gov/demographics/research-tools

The American Community Survey Statistical Analyzer: Created at the University of South Florida, this spreadsheet is a sophisticated collection of macros and formulas that includes formulas not only for the published ACS estimates but also for ACS PUMS (Public Use Microdata Sample) and Census Transportation Planning Products. It goes beyond the basics and includes formulas for calculating other derivatives like means, medians, and frequencies. http://www.nctr.usf.edu/abstracts/abs77802.htm

Map Reliability Calculator: Developed by the Population Division of NYC City Planning, this spreadsheet calculator is designed for thematic mappers of ACS data who want to measure the reliability of classification schemes. http://www1.nyc.gov/site/planning/data-maps/nyc-population/geographic-reference.page

acs–R: A package for the open source statistical programing language R that includes specific modules and functions for working with ACS data. https://cran.r-project.org/web/packages/acs/index.html

tidycensus: Another R package for working with ACS and decennial census data. https://walkerke.github.io/tidycensus/

the next 2 months, then that address counts as their current residence. In contrast, the decennial census employs the concept of "usual residence," which is the place where the household lives and sleeps most of the time as of April 1 of the census year.

Second, the concept of time in the ACS is not constant for all people who are filling out the form, since the ACS is a rolling sample and a couple of hundred thousand addresses are receiving the form each month. Questions like "Last month what was the cost of electricity in this house?" and "Did this person live in this house or apartment a year ago?" are going to have different points of reference for households who filled the form out in February versus August. The ACS is used to provide general

characteristics of the population over a given time period, not precise measurements at a fixed point in time.

The Census Bureau publishes its list of subjects for the ACS each year (U.S. Census Bureau, 2017b) and categorizes them into four groups for Social, Economic, Housing, and Demographic topics (see Table 6.4). As you recall, the ACS publishes data profile tables for each of these categories (DP02 through DP05), which are convenient for seeing what's included in the ACS at a glance. As we saw in Chapter 2, each of the tables has its own ID code with the following prefixes: S for subject (collections and summaries of data), B for base, and C for collapsed (detailed tables focused on a narrow subject), followed by a numeric ID. The first two digits of the numeric ID for the detailed tables indicates what the table's subject is (Table 6.5).

TABLE 6.4 ⬢ SUBJECTS INCLUDED IN THE ACS

Social	**Housing**
Ancestry	Bedrooms
Citizenship status	Computer and internet use
Disability status	House heating fuel
Educational attainment	Kitchen facilities
Fertility	Occupancy/vacancy status
Grandparents as caregivers	Occupants per room
Language spoken at home	Plumbing facilities
Marital history	Rent
Marital status	Rooms
Migration/residence 1 year ago	Selected monthly owner costs
Place of birth	Telephone service available
School enrollment	Tenure (owner/renter)
Undergraduate field of degree	Units in structure
Veteran status; period of military service	Value of home
Year of entry	Vehicles available
Economic	Year householder moved into unit
Class of worker	Year structure built

TABLE 6.4 ⬢ CONTINUED

Commuting and place of work	**Demographic**
Employment status	Age; sex
Supplemental Nutrition Assistance Program (SNAP)	Group quarters population
Health insurance coverage	Hispanic or Latino origin
Income and earnings	Race
Industry and occupation	Relationship to householder
Poverty status	Total population
Work status last year	

TABLE 6.5 ⬢ ACS TABLE SUBJECTS AND ID CODES

00	Unweighted Count (of the Sample)
01	Age and Sex
02	Race
03	Hispanic Origin
04	Ancestry
05	Foreign Born; Citizenship; Year or Entry; Nativity
06	Place of Birth
07	Residence 1 Year Ago; Migration
08	Journey to Work; Workers' Characteristics; Commuting
09	Children; Household Relationship
10	Grandparents; Grandchildren
11	Household Type; Family Type; Subfamilies
12	Marital Status and History
13	Fertility
14	School Enrollment
15	Educational Attainment
16	Language Spoken at Home and Ability to Speak English
17	Poverty

TABLE 6.5 ⬢ CONTINUED	
18	Disability
19	Income (Households and Families)
20	Earnings (Individuals)
21	Veteran Status
22	Transfer Programs (Public Assistance)
23	Employment Status; Work Experience; Labor Force
24	Industry; Occupation; Class of Worker
25	Housing Characteristics
26	Group Quarters
27	Health Insurance
98	(Data) Quality Measures
99	Imputations

The way data is presented varies based on the subject. Most of the estimates are published as total values that are subdivided into other categories and cross tabulated with basic subject characteristics like age, sex, race, Hispanic origin, household and family status, occupancy, and tenure. As we have seen, some of the estimates are calculated as subsets of the population; that is, educational attainment is measured for people 18 and over and 25 and over, veteran status is measured for civilians who are 18 and over, and so on.

Interval variables like income, earnings, home value, and rent are presented as estimates for households, families, and individuals who are divided into brackets for a certain range of values. The last value in the bracket is always top-coded, measuring items of a certain value to some unspecified upper limit. The Census Bureau also provides summary measures for interval data such as a mean, median, per capita value, and aggregate value (i.e., the sum of all rents or incomes). The aggregate value is useful if you need to combine subject groups or geographies and need to compute new summary measures.

Like the decennial census, variables in the ACS may appear in several tables, presented in different ways. For example, Table 6.6 shows gross rent in Charlotte, North Carolina, from the 2012–2016 ACS, as published in the data profile table DP04 for housing. Gross rent is the amount of contract rent plus the estimated average monthly cost of utilities and fuels if these are paid for by the renter. In the data profile table,

TABLE 6.6 ⬢ GROSS RENT IN CHARLOTTE, NORTH CAROLINA, 2012–2016

	Estimate	Margin of Error	Percent	Percent Margin of Error
Occupied units paying rent	141,768	± 1,973	141,768	(X)
Less than $500	7,070	± 561	5.0%	± 0.4
$500 to $999	69,911	± 1,447	49.3%	± 0.9
$1,000 to $1,499	50,009	± 1,476	35.3%	± 0.9
$1,500 to $1,999	11,125	± 733	7.8%	± 0.5
$2,000 to $2,499	2,189	± 341	1.5%	± 0.2
$2,500 to $2,999	524	± 145	0.4%	± 0.1
$3,000 or more	940	± 203	0.7%	± 0.1
Median (dollars)	$966	± 7	(X)	(X)
No rent paid	3,084	± 373	(X)	(X)

Source: Data Profile Table DP04 2012–2016 ACS.

occupied housing units that pay rent are summarized: in rent brackets with a range of $500 and a top-coded value of $3,000 or more, by median rent, and with an estimate of those paying no rent. In the detailed tables, Gross Rent (B25063), Median Gross Rent (B25064), and Aggregate Gross Rent (B25065) are published separately. The detailed gross rent table provides estimates of households in smaller brackets that range from $50 at lower values to $500 at higher values, with a top-coded value of $3,500 or more. The trade-off is that the MOEs are higher, since the brackets are smaller interval ranges. Gross rent and median gross rent are cross tabulated in other tables with number of bedrooms, year structure built, and as a percentage of household income, and the components of gross rent (contract rent, costs of utilities and fuel) are published in separate tables.

Some of the variables are derived from user responses and defined based on specific measurement criteria. For example, the ACS does not ask whether or not a person is in poverty. The ACS includes questions on income and earnings and on the relationship status of every person in the household. The Census Bureau uses this data and follows the Office of Management and Budget's Statistical Policy Directive 14 to determine whether or not a household or family is in poverty based on family size, composition, and a set of money income thresholds. The official poverty

definition uses money income before taxes and does not include capital gains (from selling stocks or bonds) or noncash benefits (e.g., public housing, Medicaid, and food stamps). The poverty thresholds are updated for inflation using the Consumer Price Index and are published annually. The thresholds vary based on size of the household, number of children, and age above or below 65. In 2016, the threshold for four people of whom two are children was $24,339.

Some poverty researchers argue that the thresholds are too low and that the methodology is outdated (Klass, 2012, pp. 157–170). The Census Bureau and the Bureau of Labor Statistics produce a Supplemental Poverty Measure that expands the definitions beyond families, accounts for a wider array of expenditures, adjusts for geography and different housing costs, and uses a different mix of income. This data is derived from the CPS and is published annually at the national and state level with an accompanying report (Fox, 2017).

Many of the economic variables on occupation, industry, and the labor force are classified into industries using the North American Industrial Classification System. We will explore this classification system in Chapter 8 when we cover data for businesses. Variables on migration and commuting are also available in special tabulations in different formats that allow users to examine the flows of people from one place to another:

Migration: `https://www.census.gov/topics/population/migration.html`

Commuting: `https://www.census.gov/topics/employment/commuting.html`

While the ACS captures many characteristics, it doesn't capture everything. InfoBox 6.2 provides a sample of what's not available in any census datasets. The Census Bureau conducts a number of smaller sample surveys that are of interest to researchers in specific fields, such as the American Housing Survey and the Survey of Income and Program Participation. These surveys ask more detailed questions related to their subject matter but can only be summarized for large geographic areas like states and metropolitan areas. We will provide a summary of these in Chapter 12 when we discuss the CPS.

6.4 EXERCISES

Working with ACS data requires you to interpret how reliable the estimates are and to calculate the MOE for any new derivatives you create, such as aggregates or percent totals. In these exercises, you will learn several of the formulas for measuring

INFOBOX 6.2 WHAT'S NOT IN THE CENSUS

While the census covers a lot of ground, it doesn't include the three topics that you should avoid discussing with your family during holidays, plus other information that has nothing to do with people, houses, or businesses.

Sex: As we saw in Chapter 4, the census does not include questions on gender identity or sexual preference, other than asking whether someone is in a same-sex partnership or marriage. Public sources for this data are scattered across a number of federal and local datasets. A professor from Drexel University has compiled a list at `http://www.lgbtdata.com/`.

Politics: The census does not provide data on election results, party affiliation, or campaign donations. The Federal Election Commission provides a detailed database on donations and woefully inadequate data on election results (compiled at the state level). Elections in the United States are local affairs, and detailed data are kept at the state level. The Census Bureau does publish national and state-level data on voter eligibility, participation, and registration every 2 years following a federal election in the CPS and tabulates the number of people who are eligible to vote in the ACS.

Religion: During the 19th century, religious bodies were counted in the decennial census, but in the early 20th century, these questions were moved to a separate survey that was defunded in the 1950s. During the 1970s, Congress amended the laws governing the census and prohibited any mandatory questions on religion citing a separation of church and state. The Association of Religion Data Archives publishes a decennial census of religion that counts adherents at the state and county level: `http://www.thearda.com/`.

Miscellaneous: The Census Bureau primarily counts people, housing units, and businesses. If you are looking for data on point incidents (crime, traffic accidents, home sales, air travel, pollution, the weather), visit the federal agency responsible for collecting that data. While the Census Bureau collects some health-related data (insurance and recent births in the ACS, births and deaths in the PEP), it does not specialize in this area. The National Center for Health Statistics at the Centers for Disease Control and Prevention is the primary source. The 500 Cities Project provides detailed health data for small areas: `https://www.cdc.gov/500cities/`. There are some federal agencies that provide alternate sources to some census variables (i.e., the Department of Education publishes data on education). This data tends to be from administrative sources and is usually published at the state and county level. See *The Reference Guide to Data Sources* for suggestions (Bauder, 2014).

reliability and creating estimates. The Census Bureau explains these formulas in the annual technical documentation they release for each period estimate (U.S. Census

Bureau, 2017c, 2017d), and they published a guidebook for researchers with many examples (U.S. Census Bureau, 2018c). We will use Calc to translate these formulas from statistical notation into spreadsheet formulas.

In the first exercise, we'll cover many of these formulas as we aggregate data for a neighborhood. In the second exercise, we will learn how to use the MCDC's Dexter tool to build customized extracts of ACS data. In terms of complexity and flexibility, this tool sits inbetween what we have covered thus far, as it's a bit more powerful than **data.census.gov** and less involved than downloading data in bulk from the FTP (File Transfer Protocol) site. At the end of the exercise, we'll return to SQLite with some brief examples of writing queries that account for the fuzziness of ACS estimates. There's also an online supplemental exercise that demonstrates additional formulas and the potential impact that interpreting estimates has on policy decisions.

Visit the publisher's website for the data we will be using, or download it from the source:

B07204 Geographical Mobility In The Past Year For Current Residence: 2012–2016 ACS, Cincinnati city, Ohio (place in state) and Census Tracts 9, 10, 16, and 17, Hamilton County, Ohio (census tracts)—`https://data.census.gov/`

Exercise 1: Aggregating Tract-level Estimates to Neighborhoods With Calc

The reliability of ACS data at the tract level can be quite low. In this exercise, we'll calculate CVs to quantify reliability, and then we'll aggregate tracts into a larger area to get a better estimate. We'll also compute percent totals for each category. For the aggregate and the percentages, we'll calculate new MOEs and will assess how the estimates have improved as a result of aggregating them.

When aggregating small geographies like tracts into larger ones, you must establish some criteria so that you are not arbitrarily grouping areas together. You could create areas that have an equal population size or that have similar demographic characteristics. Better yet, you could combine the tracts to approximate areas that have some existing social, economic, or political meaning. Neighborhoods are areas that are defined locally and are a historic product of the physical, social, and political landscape. Ideally, they are defined by the people who live there, but neighborhoods are also contested spaces where outsiders may seek to influence what the neighborhood is to suit their own purposes. For example, real estate agents or developers may try to

"re-brand" neighborhoods to make them more attractive to affluent home buyers, often in opposition to current residents.

The question of what constitutes a neighborhood is fraught with issues (Nicotera, 2007; Sperling, 2012), and there are few official "definitions" that you can apply and certainly none that are national in scope. Do some research on the internet for your area of study and see what exists. In particular, look at the local city and regional planning agencies, neighborhood organizations, universities, and the local open data movement. Some cities have taken census tracts and combined them either to create or approximate neighborhoods that local residents would recognize or to create statistical areas for studying communities, or both. Some cities have areas like districts, wards, or community planning areas that fulfill some legal or administrative function. These might be related or constructed from census geographies, or they might not. If they are related, you can use their data to identify which census tracts were used for constructing the area, and then you can reconstruct it using the same tracts. If they are not related, you may have to compare resources like local maps to the census geography (e.g., in TIGERweb) to approximate the areas using census tracts. Geographic information system (GIS; covered in Chapter 10) can be used to help make these decisions.

In this example, we will look at four census tracts that make up the Over the Rhine neighborhood, which is just north of downtown Cincinnati, Ohio (Figure 6.1). Originally a working class neighborhood, the area has a large concentration of historic buildings and has recently become more gentrified. Like many city planning agencies, the Department of City Planning in Cincinnati has a section of its website dedicated to Plans, Maps, and Data, and another section dedicated to Census and Demographics. The department has aggregated census tracts and, in some cases, block groups to create Statistical Neighborhood Approximations: `https://www.cincinnati-oh.gov/planning/reports-data/census-demographics/`. The city publishes reports with detailed maps displaying the boundaries along with 2010 census and 2006–2010 ACS data. They do not publish the MOEs for the ACS data for either the individual tracts or the aggregated values.

We will look at Table B07204: Geographical Mobility in the Past Year for Current Residence–State, County and Place Level. This data shows us how many people lived in their current homes during the past year. For residents who didn't, we can see if they previously lived in a different city, county, state, or outside the United States. This data is useful for illustrating residential stability versus dynamism: Have most people lived in the neighborhood for a long time, or is there a large influx of newcomers?

FIGURE 6.1 ⬢ CENSUS TRACTS IN THE OVER THE RHINE NEIGHBORHOOD, CINCINNATI, OHIO

Source: Base map is from the OpenStreetMap: `https://www.openstreetmap.org/`

Interpreting this requires a point of reference, so we will compare the neighborhood's data with the entire city.

Before we begin to work with the formulas, we will need to sift through all the variables in this table and whittle them down to just the ones we need. Given the large number of variables in the ACS, it's important to distinguish what represents a category versus a subcategory versus a subsubcategory in order to ensure that you are comparing apples with apples. We will spend some time dissecting the mobility and residency table to understand how the pieces fit.

1. **Import the residency table and save**. Launch Calc, go to Sheet—Insert Sheet From File, and from Chapter 6 Exercise 1 folder import the file ACSDT5Y2016_B07204_with_ann.csv. Hit OK to get through the import screen. Save the file as a Calc spreadsheet in Chapter 6 Exercise 1 folder. As you look at this file, you'll see there are five rows of data, one for the City of Cincinnati and four for each of the census tracts that are part of the Over the Rhine neighborhood (Figure 6.2). There are a lot of columns: In ACS tables, columns always come in pairs, one for the estimate and one for the MOE for that estimate.

2. **Transpose the data and study the attributes**. Add a new worksheet and name it OTR (for Over the Rhine). Copy the data in the original sheet and do a Paste Special—Transpose into this new sheet, so that the attribute names are now rows and the geographies are in columns. This will make it easier for us to analyze. Make the attribute column wider so you can read the labels. Save your work. The data in this table is broken down into a number of subcategories or facets, and the facets are indicated with repeating labels for the upper category followed by a dash and the label for the subcategory (Figure 6.3). At the top is the total estimate and its MOE, which for this dataset is the population aged 1 year and over. Under that is the number of people who lived in the Same House 1 year ago and its MOE. This category has no additional facets. But for the number who lived in a Different house in the United States 1 year ago, there are

FIGURE 6.2 ⬢ RESIDENTIAL MOBILITY FOR CINCINNATI AND OVER THE RHINE CENSUS TRACTS

	A	B	C	D	E
1	GEO.id	GEO.display-label	ESTIMATE#HD01_VDIM#VD01	ESTIMATE#HD02_VDIM#VD01	ESTIMATE#HD01_VDIM#VD02
2	Id	${dim.label}	Estimate; Total:	Margin of Error; Total:	Estimate; Total: - Same house 1 year ago
3	0600000US3906115000	Cincinnati city, Hamilton County, Ohio	293772	451	220219
4	1400000US39061000900	Census Tract 9, Hamilton County, Oh	1588	246	1033
5	1400000US39061001000	Census Tract 10, Hamilton County, O	1399	225	975
6	1400000US39061001600	Census Tract 16, Hamilton County, O	1004	196	677
7	1400000US39061001700	Census Tract 17, Hamilton County, O	1074	142	759

FIGURE 6.3 ⬢ TABLE TRANSPOSED, FACETING OF VARIABLES

	A	B	C	D	E	F	G
1	GEO.id	Id	0600000US3	1400000US3	1400000US3	1400000US3	1400000US3906
2	GEO.display	${dim.label}	Cincinnati cit	Census Tract	Census Tract	Census Tract	Census Tract 17
3	ESTIMATE#	Estimate; Total:	293772	1588	1399	1004	1074
4	ESTIMATE#	Margin of Error; Total:	451	246	225	196	142
5	ESTIMATE#	Estimate; Total: - Same house 1 year ago	220219	1033	975	677	759
6	ESTIMATE#	Margin of Error; Total: - Same house 1 year ago	2593	202	165	167	188
7	ESTIMATE#	Estimate; Total: - Different house in United States 1 year ago:	71169	555	419	327	315
8	ESTIMATE#	Margin of Error; Total: - Different house in United States 1 year ago:	2501	150	203	97	143
9	ESTIMATE#	Estimate; Total: - Different house in United States 1 year ago: - Same city or town:	41198	332	211	253	207
10	ESTIMATE#	Margin of Error; Total: - Different house in United States 1 year ago: - Same city or town:	2226	116	110	77	125
11	ESTIMATE#	Estimate; Total: - Different house in United States 1 year ago: - Same city or town: - Same county	41198	332	211	253	207
12	ESTIMATE#	Margin of Error; Total: - Different house in United States 1 year ago: - Same city or town: - Same county	2226	116	110	77	125

several facets: Different house—Same City or town, then Different House—Same City or town—same county, and so on. It's important to interpret this correctly to make proper comparisons without double counting or omitting values that we need.

You can view a more human-readable version of this table in **data.census.gov** or view the table structure either in the Census Bureau's technical documentation or on an alternative site like the Census Reporter `https://censusreporter.org/`, where you can type the table ID number into the Explore Box and get the table structure as a result (Figure 6.4). Based on the indentation, we can clearly see that under the Total the top level categories are "Same house 1 year ago," "Different house in the United States 1 year ago," and "Abroad 1 year ago." If we add up these three values, we would get the total value. If we summed the subcategories for Same city or town and Elsewhere, we would get the total for Different house in United States 1 year ago. Alternatively, if we summed Same city or town with Same county and Different county under Elsewhere, we would also get the total for Different house. Paying attention to the indentation or faceting is the key to understanding the relationship between the variables.

In deciding which variables to include in our summary, we want to make sure our data will sum to the total and that we can capture enough essential or interesting information about residency and migration. We would omit subcategories where the values are too small to be reliable or are not meaningful for the place we are studying.

3. **Delete attributes**. Starting from the bottom and working our way up, delete the following rows:

   (a) Rows 35 to 40, abroad subcategories. These values are too small to be meaningful for our area.

   (b) Rows 25 to 32, different state regional subcategories. These values are also too small for our area.

   (c) Rows 19 to 20, elsewhere—different county. We'll keep the subcategories for this total instead.

   (d) Rows 15 to 16, elsewhere. We're keeping several subcategories of this category.

   (e) Rows 11 to 14, same city or town subcategories. These don't apply to our area, since Cincinnati does not cross county boundaries.

   (f) Rows 7 to 8, different house in United States. We're keeping several subcategories of this category.

FIGURE 6.4 ⬢ READABLE TABLE STRUCTURE FOR ACS TABLE B07204

Table universe: Population 1 Year and Over in the United States

Colums in this table

Total:
- Same house 1 year ago
- Different house in United States 1 year ago:
  - Same city or town:
    - Same county
    - Different county (same state)
  - Elsewhere:
    - Same county
    - Different county:
      - Same state
      - Different state:
        - Northeast
        - Midwest
        - South
        - West
- Abroad 1 year ago:
  - Puerto Rico
  - U.S. Island Areas
  - Foreign country

Source: The Census Reporter `https://censusreporter.org/tables/B07204/`

When you are finished, we should have 16 rows remaining (Figure 6.5). We have enough variables to indicate whether residents live in the same house, or in a different house in the same city, outside the city but in the same county, in a different county in the same state, in a different state, or abroad. Pick a row for one of the tracts, and verify that each of the estimates sums to the total to ensure that you've saved the right columns. Save your work.

4. **Clean up**. Delete column A (contains variable IDs) and delete row 1 (contains geographic IDs). Give the variables and the geographies the shorter names

FIGURE 6.5 ⬢ COLUMNS REMAINING IN TABLE B07204 AFTER DELETIONS

	A	B	C	D	E	F	G
1	GEO_id	Id	0600000US3	1400000US3	1400000US3	1400000US3	1400000US39
2	GEO_display	${dim_label}	Cincinnati cit	Census Tract	Census Tract	Census Tract	Census Tract 1
3	ESTIMATE##	Estimate; Total:	293772	1588	1399	1004	1074
4	ESTIMATE##	Margin of Error; Total:	451	246	225	196	142
5	ESTIMATE##	Estimate; Total: - Same house 1 year ago	220219	1033	975	677	759
6	ESTIMATE##	Margin of Error; Total: - Same house 1 year ago	2593	202	165	167	188
7	ESTIMATE##	Estimate; Total: - Different house in United States 1 year ago: - Same city or town:	41198	332	211	253	207
8	ESTIMATE##	Margin of Error; Total: - Different house in United States 1 year ago: - Same city or town:	2226	116	110	77	125
9	ESTIMATE##	Estimate; Total: - Different house in United States 1 year ago: - Elsewhere: - Same county	9618	32	100	9	31
10	ESTIMATE##	Margin of Error; Total: - Different house in United States 1 year ago: - Elsewhere: - Same county	1056	28	155	10	42
11	ESTIMATE##	Estimate; Total: - Different house in United States 1 year ago: - Elsewhere: - Different county: - Same sta	10525	91	33	23	36
12	ESTIMATE##	Margin of Error; Total: - Different house in United States 1 year ago: - Elsewhere: - Different county: - San	676	72	21	22	43
13	ESTIMATE##	Estimate; Total: - Different house in United States 1 year ago: - Elsewhere: - Different county: - Different s	9828	100	75	42	41
14	ESTIMATE##	Margin of Error; Total: - Different house in United States 1 year ago: - Elsewhere: - Different county: - Diff	902	44	43	46	34
15	ESTIMATE##	Estimate; Total: - Abroad 1 year ago:	2384	0	5	0	0
16	ESTIMATE##	Margin of Error; Total: - Abroad 1 year ago:	455	11	8	11	11

FIGURE 6.6 ⬢ RENAME ROWS AND COLUMNS IN TABLE B07204

	A	B	C	D	E	F
1	Geography	Cincinnati	Tract 9	Tract 10	Tract 16	Tract 17
2	Total Residents	293772	1588	1399	1004	1074
3	TR MOE	451	246	225	196	142
4	Same house	220219	1033	975	677	759
5	SH MOE	2593	202	165	167	188
6	Dif House Same City	41198	332	211	253	207
7	DHSC MOE	2226	116	110	77	125
8	Dif House Dif City Same County	9618	32	100	9	31
9	DHDCSC MOE	1056	28	155	10	42
10	Dif House Dif County Same State	10525	91	33	23	36
11	DHDCSS MOE	676	72	21	22	43
12	Dif House Dif State	9828	100	75	42	41
13	DHDS MOE	902	44	43	46	34
14	Abroad	2384	0	5	0	0
15	A MOE	455	11	8	11	11

that appear in Figure 6.6. Since we are just working with this small amount of data in a spreadsheet, we can use human-readable names with spaces. If this were a larger dataset, or we were importing this data into a database or GIS, we would need to use shorter, cryptic names without spaces or simply keep the original variable ID numbers and use a lookup table for the names. Save your work.

5. **Copy and paste special—transpose**. Select all the data cells, copy, click in cell G1 and do paste special—transpose. Then delete columns A through F. This gets our finished and formatted data back to the original structure, where the geographies are in rows (Figure 6.7). This simply makes the data easier to read, as you can see the MOE beside each value. Save your work.

6. **Calculate CVs for total residents**. The MOEs for many of the tract values look high. For example, in Tract 17, 207 people ±125 lived in a different

FIGURE 6.7 ⬢ TRANSPOSE CLEANED DATA BACK TO THE ORIGINAL STRUCTURE

	A	B	C	D	E	F	G
1	Geography	Total Resider	TR MOE	Same house	SH MOE	Dif House Sa	DHSC MOE
2	Cincinnati	293772	451	220219	2593	41198	2226
3	Tract 9	1588	246	1033	202	332	116
4	Tract 10	1399	225	975	165	211	110
5	Tract 16	1004	196	677	167	253	77
6	Tract 17	1074	142	759	188	207	125

FIGURE 6.8 ⬢ CALCULATE COEFFICIENT OF VARIATION (CV) FOR TOTAL RESIDENTS

D2 f_x Σ = =ROUND(((C2/1.65)/B2)*100)

	A	B	C	D	E
1	Geography	Total Resider	TR MOE	TR CV	Same house
2	Cincinnati	293772	451	0	220219
3	Tract 9	1588	246	9	1033
4	Tract 10	1399	225	10	975
5	Tract 16	1004	196	12	677
6	Tract 17	1074	142	8	759

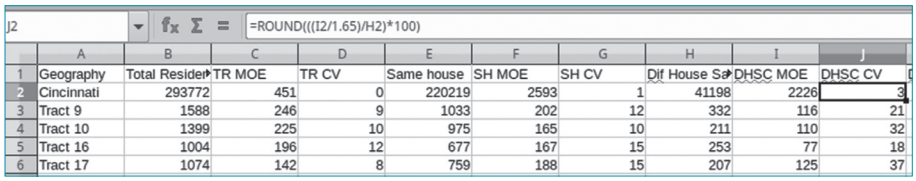
house in the same city a year ago. Let's calculate CVs for each variable. The formula is as follows:

$$\text{CV} = \frac{\text{MOE}/1.645}{\text{ACS estimate}} * 100$$

We divide the MOE by the *Z* value, then divide that result by the estimate, then multiply by 100. In statistics, *Z* values are constants that represent the critical value for the specific confidence interval being used. Since all ACS estimates are published at a 90% confidence interval, the *Z* value will always be 1.645.

Select column C, right-click, and Insert Column Right. In cell D1, type the label TR CV. In cell D2, type the formula =ROUND(((C2/1.65)/B3)*100). Paste the formula down the column. The CVs for total population are pretty low, so these estimates are of high reliability (Figure 6.8). Remember that the definition of reliability varies; we could say 0 to 15 is low, 16 to 34 is medium,

FIGURE 6.9 ⬢ CALCULATE COEFFICENT OF VARIATION (CV) FOR OTHER VARIABLES



J2 =ROUND(((I2/1.65)/H2)*100)

	A	B	C	D	E	F	G	H	I	J
1	Geography	Total Reside	TR MOE	TR CV	Same house	SH MOE	SH CV	Dif House Sa	DHSC MOE	DHSC CV
2	Cincinnati	293772	451	0	220219	2593	1	41198	2226	3
3	Tract 9	1588	246	9	1033	202	12	332	116	21
4	Tract 10	1399	225	10	975	165	10	211	110	32
5	Tract 16	1004	196	12	677	167	15	253	77	18
6	Tract 17	1074	142	8	759	188	15	207	125	37

and 35 and above is high. Or we could say anything below 20 is acceptable and everything above is not.

7. **Calculate CVs for remaining values.** Insert columns to the right of each of the MOE columns. In the first cell of each of these columns, enter a label for the column that mimics the label for the MOE column. For example, for Same house, the MOE column is SH MOE. Name the CV column SH CV. Then copy the five formulas in column D and paste them into each of the new CV columns. The formulas will auto increment, so they draw values from the estimates and MOE columns that are to their immediate left (Figure 6.9). In the final column for Abroad, you will get some error messages; this is normal because the estimates are zero, and you cannot divide zero by another number. The CVs for the census tracts are extremely high for many of the variables. For different house, same city (DHSC), most of the values are of medium reliability (16–34) except for Tract 17, which is of low reliability ($\geq$ 35). The values for different house, different city, same county (DHDCSC) are wholly unreliable, with CVs between 53 and 94. It's clear that we should aggregate to increase reliability. Save your work.

8. **Calculate sum and MOE for total residents**. In cell A7, type the label Over the Rhine. In cell B7, type the formula =SUM(B3:B6). This returns the total population of the neighborhood by combining the four tracts: 5,065. To calculate the MOE for a sum, you take the square root of the sum of the squares for each MOE:

$$\text{MOE} = \sqrt{\text{MOE1}^2 + \text{MOE2}^2}$$

In cell C7, type the formula =ROUND(SQRT(SUMSQ(C3:C6))). The SUMSQ function is convenient because we can give it an array of values, instead of having to square each individual value and sum it. The MOE for the neighborhood is 412 (Figure 6.10). Last, copy the CV formula in cell D6 and paste it into cell D7. The CV for the Over the Rhine estimate of total

FIGURE 6.10 ⬢ CALCULATE MARGIN OF ERROR FOR POPULATION OF OVER THE RHINE

C7 | fx Σ = | =ROUND(SQRT(SUMSQ(C3:C6)))

	A	B	C	D	E
1	Geography	Total Residen	TR MOE	TR CV	Same house
2	Cincinnati	293772	451	0	220219
3	Tract 9	1588	246	9	1033
4	Tract 10	1399	225	10	975
5	Tract 16	1004	196	12	677
6	Tract 17	1074	142	8	759
7	Over the Rhine	5065	412	5	

residents is 5. This is a good improvement; the CVs for the individual tracts ranged from 8 to 12. Save your work.

9. **Calculate the summaries for the other variables**. Select the three values in cells B7 to cell D7 by selecting cell B7 and dragging your mouse to cell D7. Right-click and copy. With the cells highlighted, click on the black box in the lower right-hand corner of cell D7, and drag it slowly to the right. As you do this, a purple outline appears around the cells. Drag the box until you reach the last cell V7 and then release. You just copied these three formulas across the row as a set. Spot check the formulas and values to verify they are correct. Values for Same House in Over the Rhine should be 3,444 residents with an MOE of 362 and a CV of 6. Residents who were previously in a different house but in the same city are $1,003 \pm 217$ with a CV of 13. By aggregating the values for this category, we now have an estimate that's highly reliable compared with the individual tracts that had estimates of medium reliability. In most cases, the improvements are good, but in some, the values are still too small to produce reliable estimates. The neighborhood estimate for Different house, different city, same county has a high CV of 57. Save your work.

10. **Adjust the formulas for estimates of zero**. Scroll over to the last variable, residents who lived abroad a year ago. For three of the tracts the estimate is zero, but there's still a MOE that suggests the value could be 0 ± 11. When summing estimates that include values of zero, we only incorporate the largest MOE for the zero values *once*. Click in cell U7, and modify the formula to this: =ROUND(SQRT(SUMSQ(U3:U4))). This incorporates the MOE of

just one of the zero values (Tract 9) and the one tract that had a nonzero value (Tract 10). Once you modify the formula, the MOE decreases from 21 to 14, and the CV value will decline as well.

11. **Calculate percent totals**. We'll calculate percent totals for the city and the neighborhood. In cell A9, add a label for Cincinnati, and in cell A10, one for Over the Rhine. In cell E9, enter the formula =E2/$B$2. In cell E10, enter the formula =E7/$B$7. This calculates the percentage of Cincinnati's and Over the Rhine's residents who lived in the same house a year ago. It's important that we don't multiply the result by 100 to get whole values; the MOE formula depends on the value being represented as a fractional proportion. So we can read the values more easily, select rows 9 and 10, right-click, choose Format Cells, and under Category select the option for Percent.

12. **Calculate MOE for percent totals**. Calculating the MOE for the percent total is more involved: You square the percent total and the MOE for the total population, multiply them, subtract that result from the square of the MOE for the subset population (the numerator), take the square root of that result and divide it by the total population (the denominator):

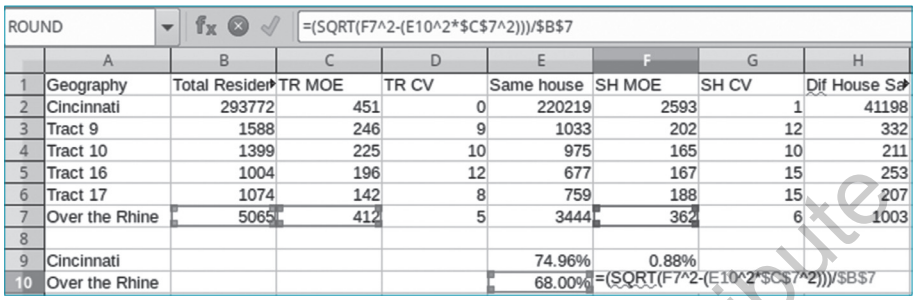
$$\text{MOE} = \frac{\sqrt{\text{MOE subset}^2 - (\text{PCT total}^2 * \text{MOE total}^2)}}{\text{Total}}$$

In cell F9, type the formula =(SQRT(F2^2 - (E9^2*$C$2^2)))/$B$2. In cell F10, type the formula =(SQRT(F7^2 - (E10^2*$C$7^2)))/$B$7.

We use dollar signs to lock the total resident estimate and MOE in place, since this will be the base for all our percentages. About 75% ± .88% of Cincinnati's residents lived in their homes a year ago compared with 68% ± 4.53% of Over the Rhine's residents, suggesting that there is more housing turnover in the neighborhood compared with the city as a whole. Pay attention to your cell references and the order of parentheses, as it's easy to make a mistake. After you enter a formula, if you double-click in the formula cell, all the cells you are referencing will be highlighted (Figure 6.11). To exit this feature, hit the escape key (don't click in another cell, otherwise you'll be modifying your formula). Save your work.

13. **Check your work**. To verify that your formula is correct, visit the ACS calculator at the Program for Applied Demographics at Cornell University: `https://pad.human.cornell.edu/acscalc/`. Scroll down

FIGURE 6.11 ⬢ CALCULATING THE MARGIN OF ERROR FOR PERCENT TOTALS FOR THE CITY AND NEIGHBORHOOD



	A	B	C	D	E	F	G	H
1	Geography	Total Residen	TR MOE	TR CV	Same house	SH MOE	SH CV	Dif House Sa
2	Cincinnati	293772	451	0	220219	2593	1	41198
3	Tract 9	1588	246	9	1033	202	12	332
4	Tract 10	1399	225	10	975	165	10	211
5	Tract 16	1004	196	12	677	167	15	253
6	Tract 17	1074	142	8	759	188	15	207
7	Over the Rhine	5065	412	5	3444	362	6	1003
8								
9	Cincinnati				74.96%	0.88%		
10	Over the Rhine				68.00%	=(SQRT(F7^2-(E10^2*$C$7^2)))/$B$7		

to the second calculator for "Computing a new value from two existing values." For the first value, enter the estimate and MOE for Over the Rhine's Same house: 3,444 and 362. For the second value, enter the total residents: 5,065 and 412 (Figure 6.12). Hit the proportion button, and the result is calculated below. You're on the right track if this result matches yours.

14. **Apply formulas to other values**. Select cells 9 and 10 in columns E, F, *and* G: the four formulas *and* the two blank cells to the right of them. As we did earlier, copy these cells, drag across to the end, and release to paste. The references for the subset MOE and percentage should auto increment, while the two total values should be locked in place with the dollar signs. The value for different house, same city should be 14.02% ± 0.76% for the city and 19.80% ± 3.97% for the county (Figure 6.13). Save your work.

15. **Are the differences between values significant?** Based on these estimates, a higher percentage of the city's residents lived in the same house a year ago compared with Over the Rhine's residents (75% vs. 68%), and a higher percentage of Over the Rhine's residents lived in a different house in Cincinnati a year ago compared with citywide residents as a whole (almost 20% vs. 14%). This suggests that there's been a small influx of new residents to Over the Rhine, as people are moving from other parts of the city to the neighborhood (to measure change over time, see InfoBox 6.3).

    It's important to remember that these are estimates that fall within a range of values; 19.8% of Over the Rhine residents lived in a different house in Cincinnati a year ago, ± 3.97%. The actual percentage could be as low as 16% or as high as 24%, and there's a 10% chance that the true value could fall outside this range. Is the value for the neighborhood significantly different from the city's?

FIGURE 6.12 ⬢ CORNELL PAD ACS CALCULATOR

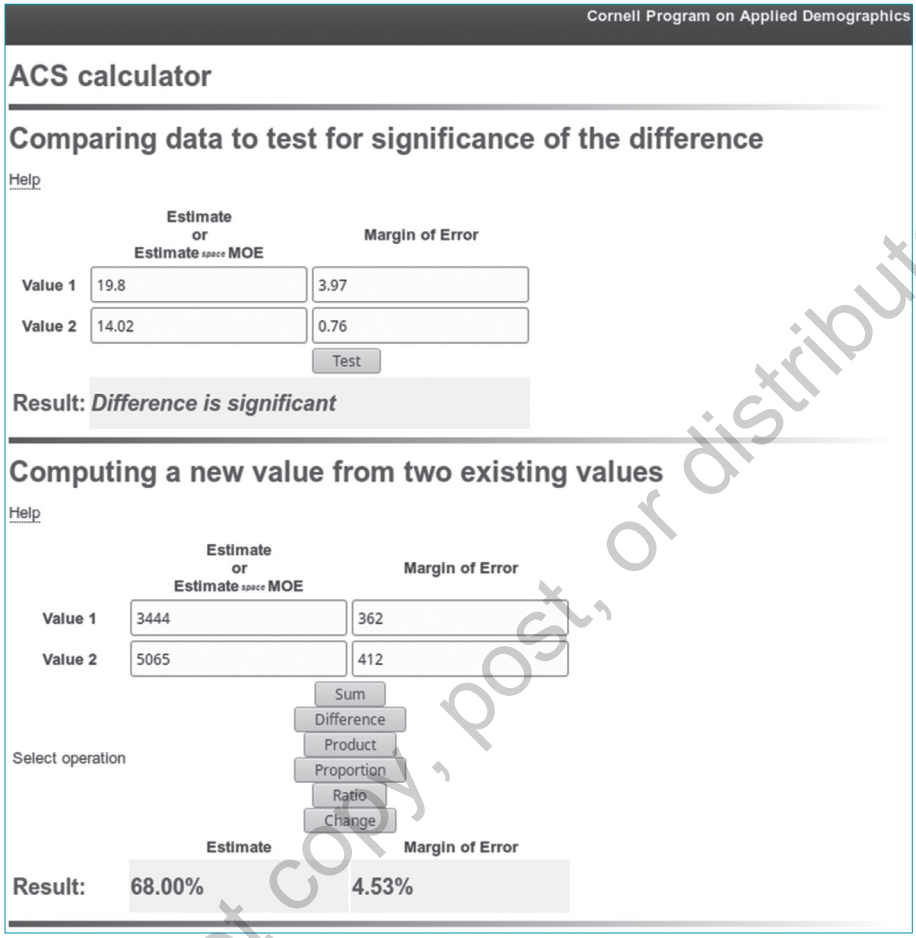


FIGURE 6.13 ⬢ CALCULATE PERCENT TOTALS AND MARGINS OF ERROR FOR REMAINING VARIABLES

I10 =(SQRT(I7^2-(H10^2*$C$7^2)))/$B$7

	A	B	C	D	E	F	G	H	I	J	K
1	Geography	Total Resider	TR MOE	TR CV	Same house	SH MOE	SH CV	Dif House Sa	DHSC MOE	DHSC CV	Dif House Di
2	Cincinnati	293772	451	0	220219	2593	1	41198	2226	3	9618
3	Tract 9	1588	246	9	1033	202	12	332	116	21	32
4	Tract 10	1399	225	10	975	165	10	211	110	32	100
5	Tract 16	1004	196	12	677	167	15	253	77	18	9
6	Tract 17	1074	142	8	759	188	15	207	125	37	31
7	Over the Rhine	5065	412	5	3444	362	6	1003	217	13	172
8											
9	Cincinnati				74.96%	0.88%		14.02%	0.76%		3.27%
10	Over the Rhine				68.00%	4.53%		19.80%	3.97%		3.40%

INFOBOX 6.3 CALCULATING CHANGE OVER TIME

There are two issues to bear in mind when using ACS estimates to measure change over time. First, if you are using a 5-year-period estimate, you should only compare estimates that cover nonoverlapping periods of time. For example, you could compare 2012–2016 with 2007–2011 because these estimates were generated from two different sample pools. It would not make sense to compare 2012–2016 with 2011–2015 because these two periods have significant overlap; four fifths of these estimates were generated from the same pool of samples.

Second, if you are using 1-year-period estimates, it may not make sense to calculate annual change or to compare changes from one year with the next. With the exception of rapidly growing or declining places, any change from one year to the next is likely due to sampling variability. If you want to study population change over time in total or by gender, age, or race, it's better to use data from the PEP.

The population for Over the Rhine was $3,163 \pm 510$ in 2007–2011 and $5,065 \pm 412$ in 2012–2016, an increase of 1,902. The formula for calculating change (difference) is the same one used for calculating aggregates (sums):

$$\text{MOE} = \sqrt{510^2 + 412^2} = 656$$

The population growth rate for Over the Rhine for this time period was 60.1% (1,902/3,163). The formula for calculating percent change is the same one used for calculating a ratio. It incorporates the MOE for the most recent population estimate, the ratio between the most recent and oldest estimate, the MOE from the oldest estimate, and the oldest estimate. Be careful and make sure that your parentheses in the spreadsheet formula are correct.

$$\text{MOE} = \frac{\sqrt{\text{MOE newpop}^2 + (\text{ratio}^2 * \text{MOE oldpop}^2)}}{\text{oldpop}}$$

$$\text{MOE} = \frac{\sqrt{412^2 + ((5065/3163)^2 * 510^2)}}{3163} = .290$$

Spreadsheet: =(SQRT(412^2+((5065/3163)^2*510^2)))/3163

This is a rapidly growing neighborhood. The Population of Over the Rhine grew by $1,902 \pm 656$ between 2007–2011 and 2012–2016, a growth rate of $60.1\% \pm 29.0\%$.

Let's go back to Cornell's ACS calculator. In the top calculator, "Comparing data to test for significance of the difference," enter the neighborhood's percentage and MOE as Value 1: 19.8 and 3.97. Enter the city's as Value 2: 14.02 and 0.76 (refer back to Figure 6.12). Hit test. The result is a significant difference, so we can say that these two estimates are indeed different from each

other. If we looked at some of the other variables to the right in our spreadsheet, it's likely that many of the differences would not be significant, as the estimates are closer to one another and many of the intervals (indicated by the MOE) overlap. In those cases, any difference in the estimate may be the result of random chance.

Instead of using the Cornell calculator, you can calculate statistical difference yourself using this formula:

$$SD = \left| \frac{\text{EST1} - \text{EST2}}{\sqrt{\text{SE1}^2 + \text{SE2}^2}} \right|$$

You subtract the second estimate from the first estimate and divide it by the square root of the sum of squares for the standard errors of each estimate, and take the absolute value of the result. The standard error is a measure of the variability of the sample mean. We calculate it by dividing the MOE by the Z value for the 90% confidence level, 1.645. If the test value (the result of this formula) is greater than the Z value of 1.645, then the differences between the values is significant. Otherwise, if the result is lower than 1.645, the difference is not significant. Here's what the formula looks like in Calc, with values from the last step of the exercise hard coded in =ABS(19.8-14.02)/SQRT((3.97/1.645)^2+(0.76/1.645)^2)

The result is 2.352, indicating significant difference since the value is higher than 1.645. If you want practice using this formula, try the supplemental online exercise at the end of this chapter.

There are two important caveats to the exercise we just did. First, in some circumstances, it is possible that the formula for calculating the MOE for the percent total will fail. This happens when the value under the square root is negative (you can't take the square root of a negative number). In this instance, and for those specific values, you would use the formula for calculating a ratio instead of a proportion. The ratio (in this case) is the total population divided by the subset (as opposed to the subset divided by the total for a proportion), and it gets added under the square root instead of subtracted:

$$\text{MOE} = \frac{\sqrt{\text{MOE subset}^2 + (\text{Ratio}^2 * \text{MOE total}^2)}}{\text{Total}}$$

Since it would be cumbersome to compute all the ratios in advance (like we did the proportions), we would do the ratio calculation as part of the formula. In our example, the formula would look like this: =(SQRT(F7^2 + (($B$7/E7)^2*$C$7^2)))/$B$7, where ($B$7/E7) is the ratio calculation.

Second, the placement of data and formulas in the spreadsheet should vary based on how you intend to use it. If you are working with a small amount of data and your goal is presentation, then you can do what we did in this exercise and place the percentages on their own rows in the same sheet. You could even place them in a separate worksheet and rearrange the end result so it's more readable. On the other hand, if this was a large dataset or one that was eventually going into a database, stats package, or GIS, you would need to follow strict rules of keeping geography in rows and attributes in columns. You could arrange the columns in sets of four: (1) estimate, (2) MOE, (3) percent total, and (4) MOE for percent total. You might even include the CVs. This format makes it harder to read, but the idea is you would be using those other packages to pull and analyze the data.

Exercise 2: Creating Custom Extracts With Dexter and Using ACS Data in SQLite

This exercise demonstrates how to access the MCDC's Public Data Archive using the Uexplore Dexter tool. This tool allows you to create customized data extracts and to download data in bulk. After creating and downloading an extract, we'll briefly summarize how to process the data in order to prepare it for loading into a database. Since we have covered this material in the last exercise, we will not go into step-by-step detail for this part. We'll conclude with considerations and examples of querying ACS data in SQLite. SQL queries of ACS data have to be constructed to account for the fact that ACS estimates represent intervals, and some values may not be statistically different from other values.

MCDC's Uexplore/Dexter for Creating Extracts

In this exercise, we'll use the Dexter tool at the MCDC to create a customized extract. The MCDC has loaded all the summary files into its databases and has created an extract program where users specify criteria in a series of web forms to pull data. Using this tool requires familiarity with how the summary files are structured, working knowledge of summary levels and GEOID codes, and knowledge of what's available in the census. While the tool may look daunting at first glance, it's really not difficult to use *if* you are familiar with how census data is structured. It requires just a few inputs, as much of the form is optional. The MCDC provides a quick start guide as well as a video and detailed instructions at `http://mcdc.missouri.edu/help/uexplore-dexter/`. Our exercise will cover the basics.

Data on voter identification shows that areas tend to identify with one political party versus the other based on certain socioeconomic characteristics (Pew Research Center,

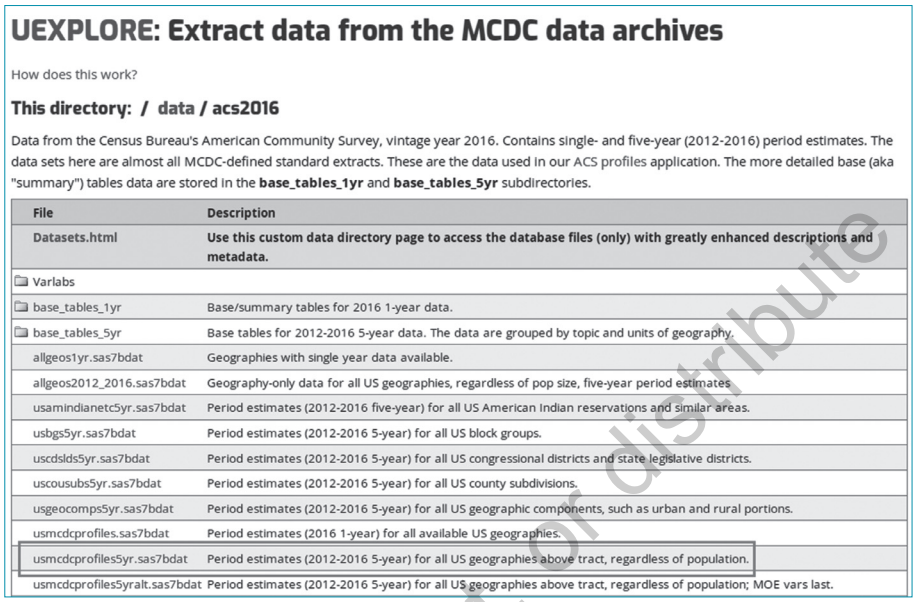
2018). Let's say, we are interested in studying voting patterns and we want to identify all the counties in the United States that have less than 250k population and are below the national average for median income and the percentage of the population with a bachelor's degree. Once we locate these counties, we also want to know what the median age is. Dexter will allow us to select just the variables we need in a single extract.

1. **Scan through the data profiles**. Since the data profiles contain a broad cross section of data, it makes sense to look there first to see if we can obtain our variables from there instead of the individual detailed tables. The MCDC creates its own customized versions of the DP02 to DP05 profile tables with some additions and deletions. We've looked at these profiles in earlier chapters. Go to the ACS profiles menu at `https://census.missouri.edu/acs/profiles/`. Select 2012–2016 as the period and counties as the type. Choose any state and any county. Generate the report for all four subjects, and browse or search through it to find our variables. It just so happens that all of them are included in these profiles, so we can get our data from this source.

2. **Launch Dexter**. Now that we have identified the variables, launch Uexplore Dexter at `http://mcdc.missouri.edu/applications/uexplore.html`. First, we select the dataset: At the top under American Community Survey Data, click 2016, then choose acs2016—American Community Survey Data, 2016 vintage. On the next page, we choose whether we want the base tables for the 1-year or 5-year ACS at the top, or one of the MCDC profiles. This list allows us to filter by geography at the outset, or we can simply select an option with all geographies and filter later. Remember, we want data on counties, and many counties are not in the 1-year ACS as their population is less than 65k people. So click on the link for usmcdcprofiles5yr.sas7bdat, Period estimates (2012–2016 5-year) for all U.S. geographies above tract, regardless of population (Figure 6.14).

3. **Enter Dexter criteria in Sections I and II**. At the top of the page, we see there are almost 580k rows in this dataset (for every piece of geography for which a profile is published) and more than 1,000 columns. We'll use Dexter to narrow that down. Section I in the tool allows you to choose a variety of output formats that include data-friendly (delimited files and database files) and presentation-friendly (listing/reports in PDF or HTML) formats. Choose CSV. Section II is where we apply filters to the rows. Under Variable/Column in the first box, we'll select SumLev, under the Operator, we'll choose Equal to (=), and in the Value box we'll type 050, which is the summary-level code for counties (see Chapter 3). The other options in the drop-down

FIGURE 6.14 ⬢ MCDC UEXPLORE MENU FOR 2016 ACS DATA



UEXPLORE: Extract data from the MCDC data archives

How does this work?

This directory: / data / acs2016

Data from the Census Bureau's American Community Survey, vintage year 2016. Contains single- and five-year (2012-2016) period estimates. The data sets here are almost all MCDC-defined standard extracts. These are the data used in our ACS profiles application. The more detailed base (aka "summary") tables data are stored in the **base_tables_1yr** and **base_tables_5yr** subdirectories.

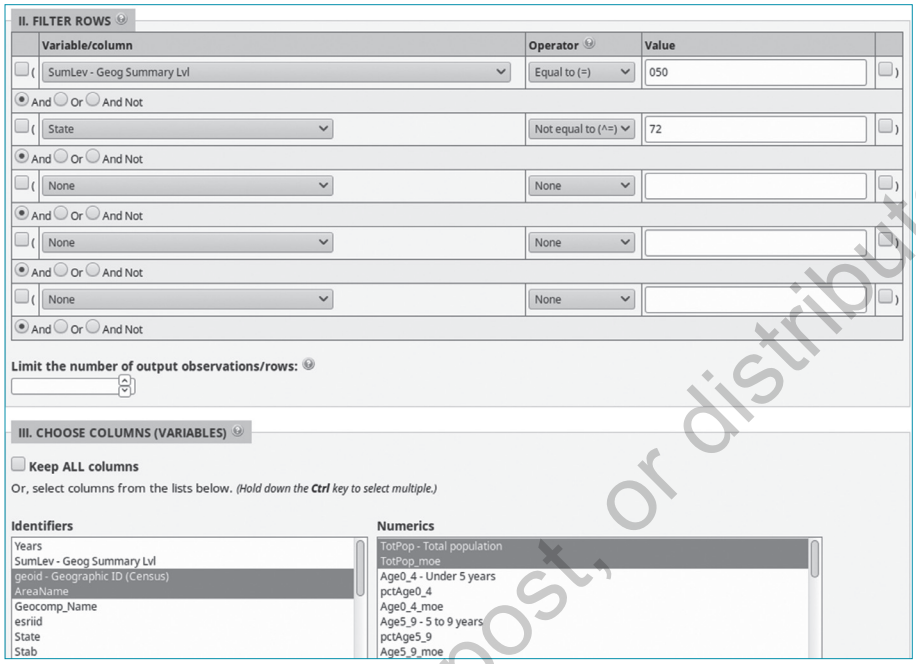
File	Description
Datasets.html	**Use this custom data directory page to access the database files (only) with greatly enhanced descriptions and metadata.**
Varlabs	
base_tables_1yr	Base/summary tables for 2016 1-year data.
base_tables_5yr	Base tables for 2012-2016 5-year data. The data are grouped by topic and units of geography.
allgeos1yr.sas7bdat	Geographies with single year data available.
allgeos2012_2016.sas7bdat	Geography-only data for all US geographies, regardless of pop size, five-year period estimates
usamindianetc5yr.sas7bdat	Period estimates (2012-2016 five-year) for all US American Indian reservations and similar areas.
usbgs5yr.sas7bdat	Period estimates (2012-2016 5-year) for all US block groups.
uscdslds5yr.sas7bdat	Period estimates (2012-2016 5-year) for all US congressional districts and state legislative districts.
uscousubs5yr.sas7bdat	Period estimates (2012-2016 5-year) for all US county subdivisions.
usgeocomps5yr.sas7bdat	Period estimates (2012-2016 5-year) for all US geographic components, such as urban and rural portions.
usmcdcprofiles.sas7bdat	Period estimates (2016 1-year) for all available US geographies.
usmcdcprofiles5yr.sas7bdat	Period estimates (2012-2016 5-year) for all US geographies above tract, regardless of population.
usmcdcprofiles5yralt.sas7bdat	Period estimates (2012-2016 5-year) for all US geographies above tract, regardless of population; MOE vars last.

menu include column names, which would allow you to eliminate rows using criteria-based extracts. For our purposes, we'll take all the data and make those decisions later.

In the second Variable/Column box, change the value from None to State. Under Operator, select the radio button for And, and in the drop-down menu, choose Not Equal to (≙). In the Value box, we'll type 72 (Figure 6.15). This is the FIPS (Federal Information Processing Series) code for Puerto Rico; Puerto Rico is always included in all national data extracts, but since they don't participate in federal elections, we'll remove them from our extract (review Chapter 3 or visit `https://census.missouri.edu/geocodes/` to look up geographic codes).

4. **Choose columns in Section III**. Section III is where we select the specific columns/variables we want. We need to select our identifiers on the left and our variables/numerics on the right. When selecting multiple variables in each menu, you need to hold down the Control key (Ctrl) while making the selections. If you select something, take your finger off the Ctrl key, and select something else, you will undo your previous selection. For the Identifiers, select geoid and AreaName. There's no need to select county, as we're filtering for counties and they will appear in the geoid and area name. Under Numerics,

FIGURE 6.15 ⬢ MCDC DEXTER DATA EXTRACTOR: CHOOSE OUTPUT FORMATS AND ROWS



the columns are designated with names created by the MCDC for their profile tables, and they are listed in approximately the order in which they appear in the profiles. Each variable appears in twos (estimate, MOE) and in some cases threes (percentage) if the value is not a total. Scroll through and select *all* the relevant columns (estimate, MOE, and percentage when available) for each of these variables (the number in parentheses represents whether there are 2 or 3 columns):

- TotPop—Total Population (2)
- Median Age—Median Age in Years (2)
- Over25—25 years and over (3)
- MedianHHInc—Median Household Income (2)
- Bachelorsormore—Bachelor degree or higher (3)

We need to take the over-25 population, as this is the population from which the Bachelor's or more population is measured. Once you've made the

selections, hit the Extract Data button at the bottom of Section III. Be patient while the application creates the extract; it could take a few seconds or a few minutes depending on how big the extract is. Eventually, you'll be presented with a Data Extraction Output screen where you can view a summary log of the request and the actual delimited file.

5. **View the log file and the extract**. Click on the Summary log. In the log, verify that you selected 16 variables (2 identifiers and 14 numerics) and scrutinize the list; if you're missing anything, hit the back button and modify your request (don't worry—you don't have to start from scratch). You should have 3,142 observations, one for each county. If the summary looks good, click the Delimited file link. This either prompts you to download it (if so, save the file), or it opens it in the browser (if so, right-click anywhere on the page and choose the option to save). Save it in Chapter 6 Exercise 3 folder as xtract.csv.

6. **Import the extract into Calc**. Figure 6.16 displays the extract in calc. Some processing is required before you can use this data in a database or GIS, such as deleting the extra header row. Any dollar values like median income are saved as text because a dollar sign is embedded in the value. You would need to use the Calc VALUE function to create a numeric value in a new column in order to work with this variable as a number. Note that all percent totals lack an MOE. You would need to calculate these yourself using the percent total formula, or, in cases where the formula fails, the ratio formula, as demonstrated in this chapter's first exercise. Save the extract as a Calc spreadsheet.

FIGURE 6.16 ⬢ MCDC EXTRACT RESULT IN CALC

	A	B	C	D	E	F
1	geoid	AreaName	TotPop	TotPop_moe	MedianAge	MedianAge_moe
2	Geographic ID (Census)		Total population	TotPop_moe	Median age in years	MedianAge_moe
3	05000US01001	Autauga County, Alabama	55049	0	37.8	0.5
4	05000US01003	Baldwin County, Alabama	199510	0	42.3	0.2999999523
5	05000US01005	Barbour County, Alabama	26614	0	38.7	0.5999999046
6	05000US01007	Bibb County, Alabama	22572	0	40.2	0.8999996185
7	05000US01009	Blount County, Alabama	57704	0	40.8	0.3999998569
8	05000US01011	Bullock County, Alabama	10552	0	39.2	1.6999998093
9	05000US01013	Butler County, Alabama	20280	0	40.6	0.3999998569
10	05000US01015	Calhoun County, Alabama	115883	0	39.1	0.3999998569
11	05000US01017	Chambers County, Alabama	34018	0	43.1	0.2999999523
12	05000US01019	Cherokee County, Alabama	25897	0	45.7	0.3999998569
13	05000US01021	Chilton County, Alabama	43817	0	38.7	0.6999998093
14	05000US01023	Choctaw County, Alabama	13287	0	45.1	0.0999999642
15	05000US01025	Clarke County, Alabama	24847	0	42	0.8999996185
16	05000US01027	Clay County, Alabama	13483	0	43.7	0.5

For the sake of demonstration, let's see what would be involved in getting detailed tables through Dexter, as opposed to the data profiles. Go back to `http://mcdc.missouri.edu/applications/uexplore.html`. Click 2016 under American Community Survey, then choose acs2016/American Community Survey Data, 2016 vintage. On the next page, choose the link for base_tables_5yr. The following page displays how the summary files are divided: There are separate files for different types of geography. Within these groups, the ACS tables are split across several files that contain sequences of tables, and then estimates are stored in one file and MOEs in another. The files at the top of this list (Table Shells and Table Number Lookup) can be used for identifying table subjects and variables.

For example, if we wanted data on citizenship status by county, we can find the table number and variables in the documentation—it's B05001. This table is located in usstcnty00_07.sas7bdat, which contains base tables from B00001 to B07413 for all states and counties. We can tell based on the file name: 00 to 07 includes all tables within this range of table prefixes, while stcnty indicates the geography. We would click on this file and go through the same interface as before and then would go back to mostcnty00_07.sas7bvew to get the MOEs for our estimates. Variable names and descriptions are stored in a series of metadata files at the bottom of the file listing.

It's not worth using the Dexter application if we just need a couple of variables from one table for one type of geography (**data.census.gov** is easier to use for that purpose), but it is invaluable if we need to create a selection of many variables from many different tables for one or more types of geography.

Working With ACS Estimates in a Database

The fact that ACS estimates are fuzzy intervals introduces uncertainty when comparing statistics and ranking variables. In this part, we will see how to account for this when writing SQL queries. Rather than following a step-by-step exercise, we will simply demonstrate how to construct some sample queries based on the data we extracted from the MCDC. For a SQL refresher, refer back to the exercises in Chapter 5.

Before you can load the MCDC extract into SQLite, you will need to perform some basic processing in Calc. You can practice making these edits yourself on the extract you've downloaded or use a cleaned version called county_xtract.csv that's stored in Chapter 6 Exercise 2 folder. Since SQLite is a "light" database, it doesn't include a lot of advanced mathematical functions, such as calculating a square root. Use Calc to calculate MOEs for the percent totals, alter these formulas to calculate ratios for MOEs in the event the percent total calculation fails, convert median income to a numeric value, round all values to a sensible number of decimal places, delete the

extra header row, and, if you wish, modify column names and the order in which they appear.

In querying county values that are above or below a national threshold, we also want to calculate whether the value for a particular county is statistically different from the national value. If it's not, we will want to exclude it from our results. In Calc, we would calculate statistical difference for income and the population with a bachelor's degree or higher using the statistical difference formula introduced in this chapter's first exercise and covered in more detail in the online supplemental exercise. We would incorporate the national values into the formula:

Median Income: $55,322 \pm 120$

Population Over 25 With Bachelor's Degree or Higher: $30.3\% \pm 0.1\%$

Import the finalized extract (saved as a CSV) into a new SQLite database. For column data types, the geoid and name would be text values, median age, all percentages, and the scores from the statistical difference formula would be reals (as they are decimal numbers), and all other values would be integers. After import, we would modify the table so geoid is designated as the primary key.

For readability, I've modified the names of the original variables in the MCDC extract. A basic query that selects counties with fewer than a quarter of a million people and with median income and college attainment lower than the national average would look like this:

```
SELECT geoid, areaname, totpop, totpop_moe, medinc,
medinc_moe, pctbach, pctbach_moe
FROM county_xtract
WHERE totpop < 250000 AND medinc < 55322
AND pctbach < 30.3
ORDER BY totpop DESC;
```

This query would return 2,277 counties out of 3,142. By adding the DESC qualifier to ORDER BY, we sort the population values from the largest to the smallest. This query fails to account for the fact that the estimates are intervals that have a possible range of values. Median household income in the United States could be $120 higher or lower than $55,322, and of course, each county estimate also has a MOE. To account for this, we subtract the MOE from each county estimate to get the *bottom* threshold for the estimate and compare it with the *top* threshold for the nation. By doing so, we ensure that we are returning all counties whose possible values fall below the highest possible values for the United States:

```
SELECT geoid, areaname, totpop, totpop_moe, medinc,
medinc_moe, pctbach, pctbach_moe
FROM county_xtract
WHERE (totpop - totpop_moe) < 250000
AND (medinc - medinc_moe) < (55322 + 120)
AND (pctbach - pctbach_moe) < (30.3 + 0.1)
ORDER BY totpop DESC;
```

This returns 2,445 records. We have more results now that we're including the full, possible range of values. What if a county value meets the criteria of being lower than the national value, but its estimate is really not statistically different from the national estimate? We omit these records from our results by selecting only counties where the statistical difference score is greater than the *Z* value of 1.645:

```
SELECT geoid, areaname, totpop, totpop_moe,
medinc, medinc_moe,
pctbach, pctbach_moe
FROM county_xtract
WHERE (totpop - totpop_moe) < 250000
AND (medinc - medinc_moe) < (55322 + 120)
AND (pctbach - pctbach_moe)< (30.3 + 0.1)
AND sd_medinc > 1.645 AND sd_bach > 1.645
ORDER BY totpop DESC;
```

This drops the result down to 2,013 rows. By incorporating statistical difference, we have omitted more than 400 counties whose estimates are not statistically different from the national estimate.

Now that we have a final result set, we can add additional variables of interest to the query like median age (since we weren't using median age as a filter or for interpretation, we omitted it from previous queries). Then add this line to the top of the statement and execute it to save it as a view:

```
CREATE VIEW county_below_us_avg AS
```

If we added additional variables to our criteria and those variables were strongly correlated with one another (i.e., low income, high unemployment, and high poverty), we probably would need a different approach where we count the number of variables that are statistically different from the national average and use that count as criteria, otherwise our result set might become too small. Ultimately, the rationale is to find counties that fit the general criteria; since the estimates are fuzzy, our criteria may also need to be a bit fuzzy.

We could also take the extra step of calculating CVs. Each estimate could have a CV in a dedicated column, and we could use that criteria to omit unreliable estimates (i.e., only includes values where the CV is less than 35). Since the CV formula involves basic arithmetic, we could do that calculation either in the database or beforehand in a spreadsheet.

Supplemental Exercise: Ranking ACS Data and Testing for Statistical Difference

Census data is commonly used to categorize and rank places, often for the purpose of distributing state or federal aid to communities. In their case study of three federal programs, Nesse and Rahe (2015) found that none of the programs incorporated the MOE for ACS estimates into their calculations. In this supplemental exercise, we will explore the impact the MOE has on rankings, and we will get some more practice with aggregating values. We will use the U.S. Department of Agriculture's Supplemental Nutrition Assistance Program as an example, as it was one of the programs included in that study. Visit **study.sagepub.com/census** to do this additional exercise.

6.5 REVIEW QUESTIONS AND PRACTICE EXERCISES

1. Describe the various components of ACS estimates—the estimate, MOE, and confidence level—and explain how these estimates are different from the decennial census count.

2. When would you use a 5-year-period estimate instead of a 1-year-period estimate?

3. Use **data.census.gov** to download housing units by tenure for the four census tracts in the Over the Rhine neighborhood and the city of Cincinnati, Ohio (Hamilton County), from the 2012–2016 ACS. Similar to what we did in Exercise 1, calculate the CV for the city and each of the four tracts for these three variables: (1) total housing units, (2) owner-occupied units, and (3) renter-occupied units.

4. Using the data from Question 3, aggregate the housing data for the four tracts, calculate percent totals, and calculate new MOEs for the total and percent totals.

5. Use the UExplore Dexter tool to create an extract for all PUMAs—summary level 795 in Wisconsin. Use the 5-year 2012–2016 ACS MCDC data profiles just as we did in Exercise 2. Select the following estimates and MOEs: renter-occupied units, median rent, and rental vacancy rate. Import the extract into Calc and do the following: Convert the median rent values from text to values, calculate CVs for median rent and rental vacancy rate, and calculate statistical difference for median rent and the vacancy rate against the *state's* 5-year values (look them up using the MCDC ACS Profile tool).

Supplementary Digital Content: Find datasets and supplemental exercises at the companion website at `http://study.sagepub.com/census`.

salaries, and occupation by industry. We'll briefly cover these datasets at the end of the chapter to see what they capture and how they differ from the census datasets.

Persons and housing units are the primary units of data collection in the decennial census and American Community Survey (ACS). For most of the business datasets that we'll cover here, the primary unit is a business establishment, which is a single physical location where business is conducted or where services or industrial operations are performed. While some of these datasets include statistics on multiunit establishments (companies and firms), these are exceptions and not the rule.

Because the business datasets count or survey establishments, they are summarizing data geographically based on where people *work*. This is fundamentally different from the demographic datasets we've covered, which primarily summarize population and labor force characteristics based on where people *live*. Our discussion of the BLS datasets at the end of the chapter will also include the labor force statistics generated from the Current Population Survey (CPS) and the ACS. Since these datasets are surveys of households and people, they are able to capture additional statistics like unemployment and labor force participation.

Using any of these datasets requires an understanding of how businesses are classified into industries. An industry is a group of businesses that produce similar products or provide similar services. The North American Industrial Classification System (NAICS—pronounced "nakes") is a hierarchical system of categories that is used to classify business establishments and the labor force in broad groups and detailed subdivisions that describe what industries produce. We'll begin this chapter with an introduction to the NAICS codes and will cover issues that are pertinent to working with all datasets that are classified using this system. Then we'll explore the Business Patterns and Economic Census and will discuss the variables and issues that are unique to each and how this data can be used to study local economies. We'll summarize the BLS datasets at the end of the chapter, and the exercise will give you the opportunity to work with industry data classified using NAICS.

When Do You Use the Business Datasets?

The Census Bureau and the BLS publish a variety of datasets that count business establishments and measure the labor force, often by industry. A business establishment is a single physical location where business is conducted, and an industry is a group of businesses that produce similar products or provide similar services. These datasets are useful for measuring the geographic distribution of businesses and comparing the economies of different places. The data can also be used in a nongeographic sense to study specific industries as a whole or to measure broad trends in the national economy.