

# (Sample) Size Matters: Visualizing How Sample Size Affects Sampling Error

## Census 2000 Summary File 3

Elaine Hallisey, MA; Barry Flanagan, PhD; Brian Lewis, BS; Caitlin Mertzluft, MPH  
Geospatial Research, Analysis & Services Program CDC/ATSDR/DTHHS

### Abstract

At the Centers for Disease Control and Prevention (CDC) we regularly use US Census data in analyses relating to population health and safety. For variables unavailable in the decennial Census 100% counts, we used sample estimates collected via the Census long form, Summary File 3 (SF3). SF3 data, collected at a single point in time and based on a sample size of approximately one in six households, were last collected in 2000. Now we depend on estimates from the Census's American Community Survey (ACS), conducted over various time periods ranging from one to five years. The ACS sample size is roughly one in 12 households for five-year sample sets. Here we examine the effects of the difference between SF3 and ACS sample size at three scales: state, county, and census tract level. We look at the coefficient of variation (CV), a relative measure of sampling error for a critical variable in many health studies, percentage of persons below the poverty level.

## Persons Below the Poverty Level Estimate Reliability\*

- High (CV <=12%)
- Medium (CV 13 to 40%)
- Low (CV >40%)
- Unavailable\*\*

\*Classes are based on Esri reliability threshold ranges. The National Research Council recommends a CV no higher than 12%.  
\*\* When a poverty percentage estimate is 0, the CV cannot be calculated because of a 0 denominator in the equation.

### Percentage of Enumeration Units in Each Reliability Class

| Reliability | SF3 2000 |        |     |             |
|-------------|----------|--------|-----|-------------|
|             | High     | Medium | Low | Unavailable |
| States      | 100      | 0      | 0   | 0           |
| Counties    | 99       | 1      | 0   | 0           |
| Tracts      | 40       | 55     | 4   | 1           |

| Reliability | ACS 2006-2010 |        |     |             |
|-------------|---------------|--------|-----|-------------|
|             | High          | Medium | Low | Unavailable |
| States      | 100           | 0      | 0   | 0           |
| Counties    | 71            | 29     | 1   | <1          |
| Tracts      | 1             | 71     | 26  | 2           |

## American Community Survey 2006-2010

### Method

We downloaded American Community Survey 2006-2010 and SF3 2000 poverty estimates for US states, counties, and census tracts. In ArcGIS 10, we joined each data set to its associated spatial data layer, i.e. state, county, or tract. The ACS data include a poverty estimate for each enumeration unit as well as a margin of error (MOE) for each estimate. We used the ACS Toolbox, developed in CDC/ATSDR/DTHHS/GRASP, to calculate the standard error (SE) and coefficient of variation (CV) for each estimate. Formulas are:

$$SE = MOE/1.645 \text{ for data at the 90\% confidence level (the Census standard)}$$

$$CV = (SE/Estimate) * 100$$

The SF3 2000 data do not include a MOE or SE for estimates. To calculate SEs for SF3 data, we used the method described in the Census 2000 Summary File 3 Technical Documentation. The formula for calculating an unadjusted SE for percentages (in this case poverty percentage), is:

$$SE = \sqrt{5/(\text{base of estimated percentage}) * \text{estimated percentage} * (100 - \text{estimated percentage})}$$

We then multiplied each unadjusted SE by a design factor the Census provides in a look up table. The design factor is based on the variable and percent-in-sample. The adjusted SE was then entered into the CV formula above to obtain the relative sampling error for each of the SF3 estimates.

### Results & Discussion

The CV maps on the left are the primary focus. These maps show relative sampling error for poverty percentage estimates. The CV maps are sized so the map reader can distinguish the increasingly smaller enumeration units as we move from states, to counties, to census tracts. The actual poverty percentage values are shown in the smaller map series below.

The National Research Council (NRC), which provides independent advice to the government on science and technology, recommends a CV threshold of no higher than 12%. Esri, a GIS software company, uses a less stringent set of ranges, indicating high (CV <=12%), medium (CV 13 to 40%), and low (CV >40%) estimate reliability.

We see highly reliable estimates for all enumeration units at state level, for both 2000 SF3 and 2006-2010 ACS. We expect this given the large population size at state level; even the much smaller sampling rate of the ACS (~1 in 12 households) versus SF3 (~1 in 6 households) does not increase the CV beyond the NRC threshold. Once below state level, however, our maps show the ACS estimates to be much less reliable than the SF3 estimates. For both SF3 and ACS, estimate reliability decreases as the enumeration unit - and therefore the population - decreases. However, unlike SF3, ACS demonstrates large numbers of unreliable estimates, particularly at the census tract level.

### Conclusions

For most surveys, the larger the sample size, the more reliable the estimate. The small sample sizes for ACS data, particularly below the county level, result in estimates that may be unsuitable for many analyses. When working with ACS data for small areas, the analyst/cartographer should, at minimum, include a caveat indicating substantial error may exist. Ideally, the analyst should map the coefficients of variation to show error for each enumeration unit. The ACS implemented improvements during 2011 data collection to decrease sampling error, but the benefits will not be apparent until 2016. In the meantime, researchers may investigate using local ancillary data to improve ACS estimates.

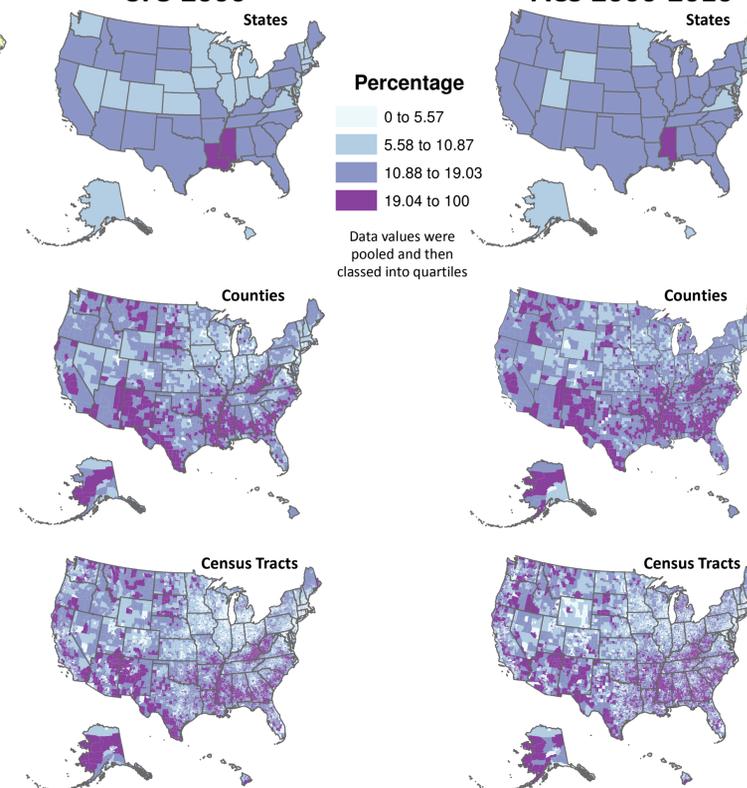
### References

- Esri White Paper (2011). The American Community Survey. <http://www.esri.com/library/whitepapers/pdfs/the-american-community-survey.pdf>
- National Research Council. (2007). Using the American Community Survey: Benefits and Challenges. [http://www.nap.edu/catalog.php?record\\_id=11901](http://www.nap.edu/catalog.php?record_id=11901)
- Census 2000 Summary File 3. Technical Documentation/prepared by the U.S. Census Bureau, 2002. <http://www.census.gov/prod/cen2000/doc/sf3.pdf>
- U.S. Census Bureau, A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. <http://www.census.gov/acs/www/Downloads/handbooks/ACSGeneralHandbook.pdf>
- U.S. Census Bureau, 2011 American Community Survey Improvements [http://www.census.gov/acs/www/about\\_the\\_survey/2011\\_acs\\_improvements/](http://www.census.gov/acs/www/about_the_survey/2011_acs_improvements/)

## Persons Below Poverty Level

### SF3 2000

### ACS 2006-2010



Data Source: American Factfinder. <http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml>  
Software Tools: ArcGIS 10. Esri.  
ACS Toolbox. Geospatial Research, Analysis & Services Program CDC/ATSDR/DTHHS.

