5. DATA QUALITY IN THE ACS PUMS

The weighted estimates calculated using the Public Use Microdata Sample (PUMS) files are estimates of the entire population. The degree of uncertainty associated with American Community Survey (ACS) estimates, known as sampling error, tends to be large when the sample size is small. Indeed, PUMS estimates are subject to additional sampling error because the PUMS data consist of a subsample of the full ACS sample. Researchers using the PUMS files must calculate their own measures of uncertainty in addition to producing their own estimates.

One way to quickly evaluate your results from the ACS PUMS is to reproduce the estimates without using weights. By examining the unweighted counts, you can see if there are any cells based on just a few sample cases, which are less likely to yield statistically reliable weighted estimates.

However, most researchers want to calculate more formal measures of sampling error such as standard errors, margins of error, and confidence intervals. Both the margin of error and the confidence interval can be calculated based on the standard error. For more information about sampling error in the ACS, see the section on "Understanding Error and Determining Statistical Significance" in the U.S. Census Bureau's handbook on *Understanding and Using American Community Survey Data: What All Data Users Need to Know*.[29]

There are two ways to calculate standard errors for ACS PUMS estimates. The first is a generalized variance formula (GVF) that uses design factors. The second is a successive difference replicate (SDR) method that uses the replicate weights. The Census Bureau uses the SDR method to calculate margins of error for published ACS tables.

Generalized Standard Error Formula

Generalized standard errors are model-based. They are, therefore, considered less accurate than direct standard errors calculated with the replicate weight method, although they may be easier to calculate for some data users.

The Census Bureau provides formulas to approximate GVF standard errors in the Accuracy of the PUMS document provided with each ACS PUMS data release.[30] To calculate GVF standard errors, design factors are applied to reflect the effects of the actual ACS sample design and estimation procedures.[31] Prior to 2017, the design factors for PUMS subject groups and state are available in the PUMS accuracy document. Beginning with 2017 data, the design factors are provided in a comma separated value (CSV) file.

ACS estimates (and their corresponding standard errors) may be generated from two or more variables representing different subject areas. When more than one subject is involved in the analysis, use the largest design factor for the factors being considered. For example, an estimate of the population under the age of 21 living below poverty is derived from ACS variables on age and poverty status. The design factor for "Poverty Status (Person)," at 1.9, is larger than the design factor for "Age," at 1.0, so the design factor for poverty should be used to calculate the standard error for this estimate. The only exception to this rule is for items cross-tabulated by race or Hispanic origin. For those items, use the largest design factor not including the race or Hispanic origin design factor.

Successive Difference Replicate Formula (Replicate Weights)

One benefit of using the SDR method is that a single formula can be used to calculate standard errors for many different types of ACS estimates, such as counts, aggregates, percentages, and ratios. However, the SDR method may be inconvenient for some data users because of the computational requirements.

The SDR method uses the 80 replicate weights to calculate 80 replicate estimates. These replicate estimates use the replicate weights. For the PUMS person files, they are called PWGTP1 through PWGTP80. For the housing files, the replicate weights are WGTP1 through WGTP80. Note that these weights are used solely for calculating uncertainty. They should never be used to calculate an estimate.

The first step is to generate an ACS estimate of interest using the PUMS weight (PWGTP or WGTP). Next,

[29] U.S. Census Bureau, *Understanding and Using American Community Survey Data: What All Data Users Need to Know*, <www.census.gov/programs-surveys/acs/guidance/handbooks/general.html>.

[30] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[31] U.S. Census Bureau, American Community Survey (ACS), Code Lists, Definitions, and Accuracy, 2016 ACS 1-year Accuracy of the Data, <www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2016.html>.

generate this estimate 80 times, using each of the 80 different replicate weights. Once you have these 81 estimated values, you can calculate the standard error by using the following formula:

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

where:

X = the estimate based on the PUMS weight (PWGTP or WGTP).

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

More details about calculating standard errors using the replicate weight method can be found in the Accuracy of the PUMS documentation.[32] The PUMS Estimates for User Verification include examples of standard errors that were calculated based on the replicate weight method.[33]

Data users interested in worked examples based on the replicate weight method can also consult the documentation for the ACS Variance Replicate Tables.[34] These tables are intended for advanced users who are aggregating pretabulated ACS data and want to calculate exact margins of error.

[32] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[33] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[34] U.S. Census Bureau, American Community Survey (ACS), Variance Replicate Tables Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html>.