Notes on American Community Survey Sampling Errors and Significance Tests

The American Community Survey (ACS) publishes a sampling error for each estimate. The Sampling error is termed Margin of Error (MOE). It is 1.645 * Standard Error or the Standard Error at the 90% Confidence Level. In many cases the standard error has to be derived by summing components. For example the estimate of population less than 5 years of age is the sum of males less than 5 and females less than 5. More details may be found in the ACS publication ***A Compass for Understanding and Using American Community Survey Data – What Researchers Need to Know***. This is available in pdf format at the Census Bureau website.

The aggregate MOE for the sum of components is:

$$MOE_{agg} = \sqrt{\sum_c MOE_c^2}$$

where:

$MOE_{agg} = aggregate\ margin\ of\ error$

$MOE_c = margin\ of\ error\ for\ c^{th}\ component$

This formula is used not only in summing components of a particular estimate, but also when summing estimates for individual Census Tract of Census Block Group data across geographies. An example is the total minority population in all Census Block Groups that have Active Landfills.

When calculating percents, all component values in the numerator are summed and all component values in the denominator are summed separately. The formula for estimated percent is:

$$\hat{p} = \frac{\hat{X}}{\hat{Y}} = \frac{\sum_c X_c}{\sum_c Y_c}$$

where:

$\hat{p} = percent\ estimate$

$X_c = c^{th}\ component\ of\ numerator$

$Y_c = c^{th}\ component\ of\ denominator$

$\hat{X} = sum\ of\ numerator\ components$

$\hat{Y} = sum\ of\ denominator\ components$

It is assumed that the numerator is a subset of the denominator.

Although this is the formula for a ratio estimator, the ACS documentation makes a distinction between a percent and a ratio. For a percent the numerator is a subset of the denominator, while for a ratio it is not.

The formula for the MOE of an estimated percent is:

$$MOE_{\hat{p}} = \frac{\sqrt{MOE_{\hat{X}}^2 - \left(\hat{p}^2 * MOE_{\hat{Y}}^2\right)}}{\hat{Y}}$$

where :

$MOE_{\hat{p}} = MOE\ of\ \ estimated\ \ percent$

$\hat{p} = estimated\ \ percent$

$\hat{Y} = sum\ of\ \ deno\min ator\ components$

$MOE_{\hat{X}}^2 = MOE\ of\ \ aggregated\ \ numerator\ components$

$MOE_{\hat{Y}}^2 = MOE\ of\ \ aggregated\ \ deno\min ator\ components$

The formula for the MOE of an estimated ratio is identical except the minus sign is replaced by a plus sign in the numerator. In actuality, the significance testing done for this project used the ACS definition of a percent, not a ratio. However, to examine a more conservative estimate of MOE, the ratio MOE was calculated and found to be slightly larger than the percent MOE. In any case, the percent MOE was used for the significance testing.

When testing significance between percents for different types of facilities, geographies, classes etc., the formula used was:

$$test\ \ statistic\ = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{MOE_1^2 + MOE_2^2}}$$

where :

$\hat{p}_1 = percent\ \ estimate\ \ for\ type, geography\ etc\ 1$

$\hat{p}_2 = percent\ \ estimate\ \ for\ type, geography\ etc\ 2$

$MOE_1 = MOE\ \ for\ type, geography\ etc\ 1$

$MOE_2 = MOE\ \ for\ type, geography\ etc\ 2$

Since the MOE's are already expressed as 1.645 * the standard error, a test statistic value greater than +1 or lesser than -1 was assumed significant at the 90% level.

The ACS documentation adds the following statement: "When comparing estimates within the same time period, the areas or groups will generally be nonoverlapping (e.g., comparing estimates for two different counties). In this case, the two estimates are independent, and the formula for testing differences is statistically correct. In some cases, the comparison may involve a large area or group and a subset of the area or group (e.g., comparing an estimate for a state with the corresponding estimate for a county within the state or comparing an estimate for all females with the corresponding estimate for Black females). In these cases, the two estimates are not independent. The estimate for the large area is partially dependent on the estimate for the subset and, strictly speaking, the formula for testing differences should account for this partial dependence. However, unless the user has reason to believe that the two estimates are strongly correlated, it is acceptable to ignore the partial dependence and use the formula for testing differences as provided in the previous section. However, if the two estimates are positively correlated, a finding of statistical significance will still be correct, but a finding of a lack of statistical significance based on the formula may be incorrect."

This study mainly compared estimates for Los Angeles County as a whole with different subsets of the County so the above discussion of dependence is relevant. However, it was mainly concerned with finding statistical

significance where it existed, and was less concerned with finding statistical significance in doubtful cases, so the ACS formula was used.

Some comments on the use of ACS formulas and procedures are:

1. The above discussion is concerned with percent estimates. In this study other estimates were also used such as Median Income and Median Year Housing Unit Structure Built. The ACS MOE for these was just used as given. However, caution should be used when interpreting results for these indicators. When aggregating data for a particular type, a simple average of medians was used and the aggregate MOE was divided by N to reflect the fact that it is based on an average rather than a sum.
2. This study did simultaneous significance testing of 150 to 200 indicators at a time, so some adjustment to the test statistic was needed. The Bonferroni approach was used where the confidence level is adjusted from the alpha=0.10 level to the 0.10 / (2 * number of tests) level. In effect the 1.645 was adjusted to the value for the 0.001 alpha level = 3.291. So the test statistic had to be greater than 2 rather than greater than 1.
3. The testing was done between two estimates of different sample size. Technically the degrees of freedom should be adjusted to account for this, but the limiting values of 1.645 and 3.291 were used.
4. Non-sampling errors were not accounted for in any additional way than those used in the ACS itself.

Due to the huge size of these data sets, a combination of Microsoft Excel, Microsoft Access and Statistica Software files were used. More detailed information is available on request.

Correlation Analysis

The correlation analysis was also done using a combination of Microsoft Excel, Microsoft Access and Statistica Software files.

Factor Analysis

The factor analysis was also done using a combination of Microsoft Excel, Microsoft Access and Statistica Software files. Actual calculations were done in Statistica. As mentioned in the presentation, non-percent indicators were not used in either the correlation analysis or the factor analysis.

Principal Components analysis was used and no adjustments were done for communalities. The factor rotation was done using the Equamax Normalized procedure.

Due to the large number of indicators used, some matrices were ill-conditioned and only 3 factors were used.