Top Coding in the 2010 1-Year ACS PUMS

Nicole Crimi

The American Community Survey (ACS) is a survey that has replaced the "long" form of the Census form. It is given to around 1% of the US population every year and asks about a wide range of topics, from education level to the cost of some household utilities. Along with some general information about the entire ACS survey, the Census Bureau also releases the complete responses to all questions for about 1% of the total ACS sample (about .01% of the total US population). This is called the Public Use Microdata Sample (PUMS) and it is used for private and public research and allocation of some government funds. In an effort to protect individual privacy, the Census Bureau uses several techniques to remove certain identifying characteristics (such as extreme old age or high income level).

This summer, my research has focused on one type of technique used to remove identifying information known as top coding, specifically as it relates to age. In traditional top-coding, a truncation age is chosen, and any age greater than this value is simply replaced by the truncation value. Consider this example. The truncation point is chosen to be age 90. After age 90, the true age is replaced with the age 90. This means that if someone is aged 92, their age would be published as age 90. The ACS PUMS age data is protected using a modified version of top-coding, which we'll call "mean modified top-coding". Instead of placing all the ages greater than the truncation value at the truncation value, they replace those ages with the mean of the ages above the truncation point, which we'll call the "age placement value". Using age 90 as the truncation point again, instead of simply replacing the true age with 90 for those with ages greater than 90, another step is added in. The next step would be to figure out the mean of all the ages above age 90. If that mean is found to be 95, all of the ages greater than 90 would be replaced with age 95 instead, so as not to upset the overall mean value of the ages.

In the 2010 1-year ACS PUMS data, this is done on a state-by-state basis, meaning that the age placement value is different for each state. When looking at the individual states, this is not a particularly large problem, but it turns into one when looking at the national data, which is an aggregate of the state data. An individual state is not represented at any age between its truncation point and age placement value, or at any age above its age placement value. Since Tennessee has a truncation point at 90, it is not represented in the published data for ages 90-92, and ages greater than 93. Figure 1 shows the number of states represented at each age in the published 2010 ACS 1-year PUMS data. For most of the ages, all 50 states (plus the District of Columbia) are represented, but starting at age 85 the number of states represented begins to decrease, until by age 95 there are only 4 states represented. Although this would not be a problem for those studying ages 85 and up as a whole group, there is a possibility of geographic effect for those who are trying to study individual ages above 85 years.

Issues also arise when looking at the sample sizes by age. Figures 2 and 3 show that up until age 90, the sample size generally decreases, but after age 90, there is an odd increase, which could be caused by the mean modified top coding. There is also a problem when looking at the population estimates from this data (Figure 4). Once again, after age 90, the graph exhibits some odd behavior that could also be caused by the mean modified top coding.

At this time mean modified top coding protects those who might be identified by their old age, but it also causes problems with the use of the data. Although it doesn't cause problems for those who are studying people over age 85 as a whole group, For those who are trying to study specific ages over age 85, this style of top coding renders this data inaccurate for their purposes. My hope for future research is to further study and document the effects of this style of top coding and possibly look into alternative methods for protection of privacy.