Challenges and Limitations of Ratio Estimators in Estimating Population Characteristics Across States*

Zitong Guo

2024-10-03

Instructions on how to obtain the data

To obtain the required data from the IPUMS USA website, start by navigating to the site and accessing the "Select Data" section. Here, deselect any default options and choose only the **2022 ACS** dataset. In the **HOUSEHOLDS** section, select **STATEICP** from the **GEOGRAPHIC** category to capture state information. Next, under the **PERSON** section, select **EDUC** from the **EDUCATION** category for educational attainment and **SEX** from the **DEMOGRAPHIC** section. After selecting all necessary variables, submit your extract request. Once processed, download the dataset as a CSV file for use in R.

Making use of the codebook, number of respondents in each state (STATEICP) that had a doctoral degree as their highest educational attainment.

```
# A tibble: 51 x 2
   STATEICP doctoral_count
      <dbl>          <int>
 1        1            600
 2        2            165
 3        3           2014
 4        4            244
 5        5            177
 6        6            131
 7       11            152
 8       12           1438
```

*Code and data are available at: https://github.com/jennygzt/USA-DATA.git.

```
 9       13           2829
10       14           1620
# i 41 more rows
```

A brief overview of the ratio estimators approach.

The ratio estimator approach is a statistical technique used to estimate unknown population totals by utilizing known ratios. In this document, the approach taken involves dividing the number of respondents with a doctoral degree in California by the total number of respondents in California, resulting in a ratio of approximately 0.0162. This ratio can then be applied to other groups to estimate their total population based on the number of doctoral respondents, assuming proportionality is similar across states. This method enhances the estimates by leveraging existing, known data. Next, the data is grouped by `STATEICP`, a state identifier, and the number of doctoral degree holders is counted for each state. The California ratio is then used to estimate the total number of respondents in each state by dividing the number of doctoral degree holders by the ratio, assuming similar proportions of doctoral degree holders across all states.

Your estimates and the actual number of respondents.

```
# A tibble: 51 x 3
   STATEICP actual_total estimated_total
      <dbl>        <int>           <dbl>
 1        1        37369          37043.
 2        2        14523          10187.
 3        3        73077         124340.
 4        4        14077          15064.
 5        5        10401          10928.
 6        6         6860           8088.
 7       11         9641           9384.
 8       12        93166          88779.
 9       13       203891         174656.
10       14       132605         100015.
# i 41 more rows
```

Some explanation of why you think they are different.

The estimated total number of respondents in each state using the ratio estimators approach may differ from the actual number for several reasons. One key factor is the **assumption**

of similarity. The ratio estimator assumes that the proportion of respondents with doctoral degrees in California is representative of that in other states. However, educational attainment varies widely due to differences in demographics, economic opportunities, and educational infrastructure across states. These variations naturally lead to discrepancies between the estimated counts and the actual number of respondents in each state.

Another source of difference is **sampling variability**. If the data used for estimation is based on a sample rather than a full population census, the randomness inherent in sampling will influence the accuracy of the calculated ratio. Sampling variability can cause deviations between the estimated numbers and the actual counts, depending on the sample's representativeness and size.

Additionally, the **non-uniform distribution** of educational attainment across the United States contributes to estimation differences. Factors such as regional policies, cultural diversity, and varying levels of access to higher education mean that educational patterns in one state, like California, may not hold in others. Applying a single ratio across all states without accounting for these differences can lead to inaccurate estimates.

Finally, there may be **bias in the ratio** itself. The Laplace ratio approach assumes a consistent relationship between the characteristic of interest and the population across all units. If the ratio of doctoral degree holders to the population in California is not reflective of other states due to unobserved factors, the estimates will be skewed. These biases highlight the limitations of assuming homogeneity when using ratio estimators in a context as diverse as the U.S. states.