IPUM_200_California*

W4 Reflection Group100

October 3, 2024

Table of contents


```
# A tibble: 3,373,378 x 3
   stateicp educ                educd
   <fct>    <fct>               <fct>
 1 alabama  1 year of college   1 or more years of college credit, no degree
 2 alabama  grade 12            ged or alternative credential
 3 alabama  grade 5, 6, 7, or 8 grade 8
 4 alabama  n/a or no schooling no schooling completed
 5 alabama  1 year of college   1 or more years of college credit, no degree
 6 alabama  grade 12            ged or alternative credential
 7 alabama  1 year of college   1 or more years of college credit, no degree
 8 alabama  grade 10            grade 10
 9 alabama  grade 11            grade 11
10 alabama  grade 12            regular high school diploma
# i 3,373,368 more rows
```

*A GitHub Repository containing all data, R code, and other files used in this investigation is located here: https://github.com/jeno0403/IPUMS_2022-

Making use of the codebook, how many respondents were there in each state (STATEICP) that had a doctoral degree as their highest educational attainment (EDUC)? (Hint: Make this a column in a tibble.)

```
# A tibble: 51 x 2
   stateicp      doctoral_count
   <fct>                  <int>
 1 connecticut              600
 2 maine                    165
 3 massachusetts           2014
 4 new hampshire            244
 5 rhode island             177
 6 vermont                  131
 7 delaware                 152
 8 new jersey              1438
 9 new york                2829
10 pennsylvania            1620
# i 41 more rows
```

Instructions on how to obtain the data.

To obtain the data from IPUMS USA, we started by selecting "IPUMS USA" on the IPUMS website, then clicked "Get Data" and selected "2022 ACS" under "SELECT SAMPLE." We specified state-level data by selecting "HOUSEHOLD" > "GEOGRAPHIC" and added "STATEICP" to our cart. For individual-level data, we went to "PERSON" and added "EDUC" to our cart. We clicked "VIEW CART" and then "CREATE DATA EXTRACT." We changed the "DATA FORMAT" to ".dta". We clicked "SUBMIT EXTRACT." After logging in or creating an account, we received an email when the extract was ready. We then downloaded and saved it locally (e.g., "usa_00004.dta") for use in `R`.

A brief overview of the ratio estimators approach.

The ratio estimators approach is a statistical technique used to estimate population totals or means based on known ratios from a sample. This approach works by taking a ratio of a specific characteristic (e.g., number of doctoral degree holders) to the total population for a known subset (e.g., California). The ratio is then applied to other subsets to estimate totals, assuming similar relationships exist across the entire population. It is particularly useful when the exact population size is unknown but a sample provides proportional relationships that can be generalized.

Your estimates and the actual number of respondents.

```
# A tibble: 51 x 3
   stateicp      actual_total estimated_total
   <fct>                <int>           <dbl>
 1 connecticut          37369          37043.
 2 maine                14523          10187.
 3 massachusetts        73077         124340.
 4 new hampshire        14077          15064.
 5 rhode island         10401          10928.
 6 vermont               6860           8088.
 7 delaware              9641           9384.
 8 new jersey           93166          88779.
 9 new york            203891         174656.
10 pennsylvania        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 of respondents for several reasons:

- Assumption of Similarity: The ratio estimator assumes that the proportion of respondents with doctoral degrees in California is representative of the proportion in other states. However, educational attainment can vary significantly due to differences in demographics, economic opportunities, and educational infrastructure across states. This variance leads to discrepancies between the estimated and actual counts.
- Sampling Variability: If the data used in the estimation is a sample rather than a complete population census, then random sampling variability will affect the calculated ratio and the accuracy of the estimates.
- Non-Uniform Distribution: The distribution of educational attainment is not uniform across the United States. Factors like regional policies, cultural differences, and access to higher education vary, which means the California ratio might not be applicable to other states.
- Bias in the Ratio: The Laplace ratio approach works well if the relationship between the characteristic of interest and the population is consistent across all units. In this case, if the ratio of doctoral degree holders to the total population in California is not indicative of other states due to unobserved factors, the estimates will be biased.

These reasons indicate that the assumption of homogeneity used in ratio estimators often leads to differences when applied to diverse populations such as different states in the US.