ACS2022

Fatima Yunusa, Michelle Ji, Jiaqi Lai, Lexi Knight, and Aliza Mithwani

2024-10-03

Obtaining ACS 2022 Data

In order to obtain data for the 2022 American Community Survey:

- Login to the IPUMS USA site
- click "get data"
- go to "select sample"
- Un- select "default sample from each year"
- Select "2022 ACS"
- Then "submit sample selections"
- Select "household" then "geographic" then select 'Stateicp'
- Select "person" then "education" then select 'educ'
- After selecting the desired variables, click "View cart" and then click "create data extract".
- Once satisfied with the dimensions of the data click "submit extract" and wait for your data to finish downloading.

The data can be obtained at IPUMS USA, here

Ratio Estimators Approach

The Ratio estimators were first used by Pierre-Simon Laplace in order to create a good estimate for the overall population of France. This was based on the numbered ratio of registered births to the number of inhabitants. Generally, the ratio estimator of a given population parameter is the ratio of two means. A prevalent variant of the ratio estimator that is used in ecology is capture and recapture. Here, the sample is captured, marked and then released. The researchers come back afterwards and capture another sample.

Ratio Calculation

The calculated Laplace ratio is around 0.01619. We used this to estimate the total respondents for each state.

```
Warning: package 'ggplot2' was built under R version 4.3.2

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
Rows: 3373378 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (14): YEAR, SAMPLE, SERIAL, CBSERIAL, HHWT, CLUSTER, STATEICP, STRATA, G...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

# A tibble: 51 x 2
   STATEICP doctoral_respondents
      <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
 9       13                 2829
10       14                 1620
# i 41 more rows

# A tibble: 51 x 3
```

```
   STATEICP doctoral_respondents estimated_total_respondents
      <dbl>                <int>                        <dbl>
 1        1                  600                       37043.
 2        2                  165                       10187.
 3        3                 2014                      124340.
 4        4                  244                       15064.
 5        5                  177                       10928.
 6        6                  131                        8088.
 7       11                  152                        9384.
 8       12                 1438                       88779.
 9       13                 2829                      174656.
10       14                 1620                      100015.
# i 41 more rows
```

Potential Reasons for Variance

As we are assuming that every state has the same proportion of respondents with doctoral degrees as California (a random state), we are ignoring the demographic variation between different states. Some states may have a higher average income level and therefore, higher levels of educational attainment. California could have more or less popular universities for doctoral degrees compared to other states; therefore, there might be a disproportional number of doctoral degree holders in California due to this external factor. By sampling a random state (California), we are introducing sampling bias because certain types of people that live in California might disproportionally exhibit certain characteristics. Since we are selecting just 1 state for simplicity and ease, we are also introducing selection bias and convenience bias. One state is an unreasonably small sample size to represent all 50 U.S. states accurately.