An Analysis of 2022 ACS Data from IPUMS

Ricky Yuan

```
library("haven")
library("tidyverse")
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- 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 beco
```

```
library("labelled")
library("dplyr")


ipums_extract <- read_csv("usa_00004.csv.gz")
```

```
Rows: 3373378 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (13): 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.
```

```r
ipums_extract <-
  ipums_extract %>%
  select(STATEICP, EDUCD) %>%
  as_factor()
```

Using the data from 2022 ACS, showing the respondents in each state of STATEICP that had a doctoral degree as their highest educational attainment of EDUC.

```r
doctoral_counts <- ipums_extract |>
  filter(EDUCD == 116) |>
  group_by(STATEICP) |>
  summarise(doctoral_count = n()) |>
  ungroup()
doctoral_counts
```

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

Instructions on how to obtain the data.

The data from the website of IMPUS of USA, and the I creates the account that can help me to download the data. After log in, selecting data and click on "SELECT DATA" in the top navigation bar to access the data selection page. In the Sample Selection section, select data for the 2022 ACS. Next, click on "SELECT VARIABLES" on the left side of the page. In the search bar, look for "STATEICP" (for state) and add it to the data cart, then look for "EDUCD" (for educational attainment). After selecting your variables, click View Cart at the top right of the page and then click Create Data Extract. Next, name the extract and submit the request, which will be processed in the background. Check the status in My Data Extracts when the extraction is complete. Finally, download the .csv file.

A brief overview of the ratio estimators approach.

Use ratio estimation to estimate population totals or means from known ratios derived from a sample. When there is a known number of a particular characteristic in one population and a similar number is estimated for the rest of the population. In this method, a ratio is calculated by dividing the number of people with the characteristic in question by the total population in the sample and applying that ratio to the rest of the population. When the exact population size is not known, but the ratio can be estimated from the sample, a generalized estimate can be made of the total number of other populations or areas.

Your estimates and the actual number of respondents.

```
total_respondents_california <- 391171

doctoral_respondents_california <- doctoral_counts |>
  filter(STATEICP == 71) |>
  pull(doctoral_count)


doctoral_ratio_california <- doctoral_respondents_california / total_respondents_california


estimated_total_counts <- doctoral_counts |>
  mutate(estimated_total = doctoral_count / doctoral_ratio_california)


actual_counts <- ipums_extract |>
  group_by(STATEICP) |>
  summarise(actual_total = n()) |>
  ungroup()

comparison <- doctoral_counts |>
  left_join(actual_counts, by = "STATEICP") |>
  left_join(estimated_total_counts, by = "STATEICP") |>
  select(STATEICP, actual_total, estimated_total)

comparison
```

```
# 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.

Differences between actual and estimated totals occur when using the ratio estimation method. This method is consistent based on the ratios observed in California. However, this assumption may not hold true. Education levels, demographics, and population composition can vary widely among states. For example, some states may have higher or lower proportions of doctoral degree holders due to economic conditions, access to higher education, or immigration patterns. Thus, applying California's rates to other states may lead to overestimates or underestimates, especially in states with markedly different educational profiles. These differences highlight the limitations of ratio estimates when the distribution of demographic characteristics is not uniform across regions.