Ratio Estimator Analysis-Reflection 4

Xingjie Yao

2024-10-03

Introduction

This document presents the analysis of the 2022 ACS dataset using Laplace's ratio estimator approach.

```
# Load necessary libraries
library(haven)
library(tidyverse)  # For data manipulation and visualization
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
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)   # For working with labelled data

# Read the data from the .dta.gz file
acs_data <- read_csv("usa_00002.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.
```

```
# Check the column names in the dataset to verify their names
colnames(acs_data)
```

```
 [1] "YEAR"     "SAMPLE"   "SERIAL"   "CBSERIAL" "HHWT"     "CLUSTER"
 [7] "STATEICP" "STRATA"   "GQ"       "PERNUM"   "PERWT"    "EDUC"
[13] "EDUCD"
```

```
# Select relevant columns and convert them to factors
acs_data <-
  acs_data |>
  select(STATEICP,EDUC, EDUCD) |>
  to_factor()
```

Filter for Doctoral Degrees

In this step, I calculated how many respondents in each state have a doctorate based on the value 11 in the EDUCD column and renamed it doctoral_degree_count.

```
# Filter rows for respondents with a doctoral degree (assuming 116 is the code for Doctoral d
doctoral_degree_counts <- acs_data |>
  filter(EDUCD == 116) |>  # Filter where 'EDUCD' equals 116 (Doctoral degree)
  group_by(STATEICP) |>  # Group by the correct state column 'STATEICP'
  summarise(doctoral_degree_count = n()) |>  # Count the number of respondents with a doctora
  ungroup()

# Display the resulting data
doctoral_degree_counts
```

```
# A tibble: 51 x 2
   STATEICP doctoral_degree_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
```

How to obtain the data.

The data for this paper is obtained from the IPUMS USA website. First, click "Get Data" and select "2022 ACS" under "Select Sample". Then click search and enter STATEICP to search again. The data for 2022 will appear. Then we go to "PERSON" and added "EDUC" and "SEX" to our cart. We click "View Cart" and then click "Create Data Extract". We change the "Data Format" to ".csv". We click "Submit Extract". Then, we will receive an email. Finally, We can download and save it locally (usa_00002.csv) and then use it in R.

A brief overview of the ratio estimators approach.

A ratio estimator is a statistical technique that uses a known relationship between two variables to estimate a population total or mean. In this paper, we use the number of respondents with a doctorate in a given state and the total number of respondents in California to estimate the total number of respondents in each state in the United States.

Compare with Actual Respondent Counts

Perform the comparison between estimated and actual total respondents.

```
# Get the total count of respondents in California
total_respondents_california <- 391171  # Given value for California

# Get the number of respondents with a doctoral degree in California
doctoral_respondents_california <- doctoral_degree_counts |>
  filter(STATEICP == "6") |>  # Assuming "6" is the code for California in STATEICP
  pull(doctoral_degree_count)

# Calculate the ratio of doctoral degree holders to total respondents in California
doctoral_ratio_california <- doctoral_respondents_california / total_respondents_california

# Get the total count of respondents in each state using the ratio estimator
estimated_total_counts <- doctoral_degree_counts |>
```

```
  mutate(estimated_total = doctoral_degree_count / doctoral_ratio_california)

# Get the actual count of respondents in each state
actual_counts <- acs_data |>
  group_by(STATEICP) |>
  summarise(actual_total = n()) |>  # Count actual number of respondents in each state
  ungroup()

# Merge the estimated totals with actual respondent counts
comparison <- doctoral_degree_counts |>
  left_join(actual_counts, by = "STATEICP") |>
  left_join(estimated_total_counts, by = "STATEICP") |>
  select(STATEICP, actual_total, estimated_total)

# Display the comparison between actual and estimated counts
comparison
```

```
# A tibble: 51 x 3
   STATEICP actual_total estimated_total
      <dbl>        <int>           <dbl>
 1        1        37369        1791623.
 2        2        14523         492696.
 3        3        73077        6013881.
 4        4        14077         728593.
 5        5        10401         528529.
 6        6         6860         391171
 7       11         9641         453878.
 8       12        93166        4293923.
 9       13       203891        8447502.
10       14       132605        4837382.
# i 41 more rows
```

Some explanation of why you think they are different.

1 The proportion of PhD holders in different states may vary according to state policies and population size, including different education levels, so it may not always be accurate to apply California's ratio to other states.

2 The sample size of each state will be relatively small, which will make the ratio estimator of these states less reliable.

3 Different geographical locations can attract different numbers of PhD respondents. For example, near economic centers or Silicon Valley, there may be more PhD respondents. Such differences will also lead to different