Analysis of Ratio Estimators and State Respondent Counts

Group 19

Instructions for Obtaining the Data

To download the 2022 ACS PUMS dataset, follow these steps:

1. Visit the IPUMS USA website: [https://usa.ipums.org/usa/%5D(https://usa.ipums.org/usa/).
2. Create an account or log in if you already have one.
3. Navigate to "SELECT SAMPLES", then select ACS data for 2022, click "SUBMIT SAMPLE ELECTIONS"
4. Under "HOUSEHOLD" drop-down tab, select "GEOGRAPHIC", then add "STATEICP" to data cart.
5. Under "PERSON" drop-down tab, select "SEX" under "DEMOGRAPHIC", and "EDUC" under "EDUCATION", add both variables to data cart.
6. View data cart, then click "CREATE DATA EXTRACT", then "SUBMIT EXTRACT"
7. Wait for the data to process then download to local folder.

Make sure to store the dataset in your working directory as `"usa_00003.csv"`.

Overview of Ratio Estimators Approach

The **ratio estimator** approach is a statistical technique used when you have a sample from which a ratio of two variables can be calculated, and you apply this ratio to the population. In our case, we use California as a reference because we know the total number of respondents in California. By calculating the ratio of doctoral degree holders to total respondents in California, we can estimate the total population of respondents in other states based on the number of doctoral degree holders in those states.

The formula we use is as follows:

$$\text{Estimated Total Respondents in State} = \frac{\text{Doctoral Count in State}}{\text{Doctoral/Total Ratio for California}}$$

Code and Estimates

Below is the R code used to calculate the ratio estimator and compare it to the actual number of respondents in each state.

```r
# Load necessary libraries
library(dplyr)
```

```


Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(readr)
library(tibble)

# Load the dataset
data <- read_csv("usa_00003.csv")
```

```
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.
```

```
# Define the code for a doctoral degree (EDUCD = 116)
doctoral_educd_code <- 116

# 1. Calculate the actual number of respondents in each state using PERWT
actual_respondents <- data %>%
 group_by(STATEICP) %>%
 summarise(actual_total_respondents = sum(PERWT)) %>%
 as_tibble()

# 2. Calculate the number of respondents with doctoral degrees in each state using EDUCD
doctoral_respondents <- data %>%
 filter(EDUCD == doctoral_educd_code) %>%
 group_by(STATEICP) %>%
 summarise(doctoral_count = sum(PERWT)) %>%
 as_tibble()

# Set the correct STATEICP code for California (71)
california_state_code <- "71"

# Ensure that we have doctoral degree holders in California
california_doctoral_respondents <- doctoral_respondents %>%
 filter(STATEICP == california_state_code) %>%
 pull(doctoral_count)

# Debugging step: Check if california_doctoral_respondents has a valid value
print(california_doctoral_respondents)
```

```
[1] 516430
```

```
# Check for empty result and stop if necessary
if (length(california_doctoral_respondents) == 0) {
 stop("Error: No doctoral respondents found for California. Check STATEICP or dataset.")
}

# Define the total number of respondents in California
california_total_respondents <- 391171

# Calculate the ratio for California (doctoral degree holders / total respondents)
california_ratio <- california_doctoral_respondents / california_total_respondents

# 3. Use this ratio to estimate the total number of respondents in each state
```

```r
state_estimates <- doctoral_respondents %>%
 mutate(estimated_total_respondents = doctoral_count / california_ratio)

# 4. Merge the estimated totals with the actual totals
comparison <- state_estimates %>%
 left_join(actual_respondents, by = "STATEICP") %>%
 mutate(difference = estimated_total_respondents - actual_total_respondents,
percentage_error = (difference / actual_total_respondents) * 100)

# Display the comparison result in a more readable format
print(comparison)
```

```
# A tibble: 51 x 6
   STATEICP doctoral_count estimated_total_respondents actual_total_respondents
      <dbl>          <dbl>                       <dbl>                    <dbl>
 1        1          49333                      37367.                  3626205
 2        2          15786                      11957.                  1385340
 3        3         169872                     128670.                  6981974
 4        4          21948                      16625.                  1395231
 5        5          15986                      12109.                  1093734
 6        6          11300                       8559.                   647064
 7       11          14619                      11073.                  1018396
 8       12         118843                      90018.                  9261699
 9       13         258461                     195772.                 19677151
10       14         159941                     121148.                 12972008
# i 41 more rows
# i 2 more variables: difference <dbl>, percentage_error <dbl>
```

```r
# Results and Comparison

# Displaying the final comparison with relevant columns
comparison %>%
 select(STATEICP, doctoral_count, estimated_total_respondents, actual_total_respondents, dif
```

```
# A tibble: 51 x 6
   STATEICP doctoral_count estimated_total_respondents actual_total_respondents
      <dbl>          <dbl>                       <dbl>                    <dbl>
 1        1          49333                      37367.                  3626205
 2        2          15786                      11957.                  1385340
 3        3         169872                     128670.                  6981974
 4        4          21948                      16625.                  1395231
 5        5          15986                      12109.                  1093734
 6        6          11300                       8559.                   647064
 7       11          14619                      11073.                  1018396
 8       12         118843                      90018.                  9261699
 9       13         258461                     195772.                 19677151
10       14         159941                     121148.                 12972008
# i 41 more rows
# i 2 more variables: difference <dbl>, percentage_error <dbl>
```

Explanation of Differences

The estimates derived from the ratio estimator approach will differ from the actual respondent counts due to:

1. Variability in educational attainment across states
2. Differences in the sampling weights used in the ACS dataset
3. Population size and unique demographic characteristics across states