Ratio Estimators Analysis

Chenming Zhao

1. Data Acquisition

The dataset used in this analysis is available [https://usa.ipums.org/usa/index.shtml]. We use IPUMS to access the 2022 ACS and focus on each state (STATEICP) that had a doctoral degree as their highest educational attainment (EDUC).

2. Overview of Ratio Estimators

Ratio estimators provide a method for estimating population parameters by using auxiliary information. Here's a brief summary…

```
# Load the 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
```

```
# Read the dataset (assuming it's a CSV)
data <- read.csv("/Users/dragon/Desktop/usa_00001.csv.gz")
```

```r
state_41_data <- data %>%
  filter(STATEICP == 41)

total_respondents_state_41 <- state_41_data %>%
  summarise(Total_Respondents = n())

#Calculate the number of respondents with EDUC == 6 (doctoral degree) in state 41
educ_respondents_state_41 <- state_41_data %>%
  filter(EDUC == 6) %>%
  summarise(educ_Respondents = n())

#Calculate the ratio of doctoral respondents to total respondents in state 41
ratio_state_41 <- educ_respondents_state_41$educ_Respondents / total_respondents_state_41$Tot

#Apply the ratio to estimate the total number of respondents with EDUC == 6 in each state
# Group by state and calculate the number of doctoral respondents (EDUC == 6) in each state
statewise_educ_respondents <- data %>%
  filter(EDUC == 6) %>%
  group_by(STATEICP) %>%
  summarise(educ_Respondents = n())

#Estimate the total number of respondents with a doctoral degree in each state using the rati
statewise_estimates <- statewise_educ_respondents %>%
  mutate(Estimated_Total_Respondents = educ_Respondents / ratio_state_41)

#View the results
head(statewise_estimates)
```

```
# A tibble: 6 x 3
  STATEICP educ_Respondents Estimated_Total_Respondents
     <int>            <int>                       <dbl>
1        1            10225                      30293.
2        2             4937                      14627.
3        3            17946                      53168.
4        4             4103                      12156.
5        5             2938                       8704.
6        6             2078                       6156.
```

3. Estimates and Actual Number of Respondents

Here are the estimates and the actual numbers based on the analysis:

```r
actual_state_Respondents <- data %>%
  group_by(STATEICP) %>%
  summarise(Actual_Total_responses = n())

head(actual_state_Respondents)
```

```
# A tibble: 6 x 2
  STATEICP Actual_Total_responses
     <int>                  <int>
1        1                  37369
2        2                  14523
3        3                  73077
4        4                  14077
5        5                  10401
6        6                   6860
```

```r
comparison_table <- merge(statewise_estimates %>% select(STATEICP, Estimated_Total_Respondent
                          actual_state_Respondents, by = "STATEICP")

#View the comparison table (showing both estimated and actual respondents)
head(comparison_table)
```

```
  STATEICP Estimated_Total_Respondents Actual_Total_responses
1        1                   30293.251                  37369
2        2                   14626.678                  14523
3        3                   53167.989                  73077
4        4                   12155.815                  14077
5        5                    8704.310                  10401
6        6                    6156.418                   6860
```

4. Explanation of Differences

The differences between the estimates and the actual numbers could be due to several factors, including: 1. We assumed the relationship between the number of respondents with EDUC = 6 degrees and the total number of respondents in one state (in this case, state 41) applies similarly to all other states, but this could vary by different culture and education background. 2. Each state has its own economic and cultural factors that affect education levels. 3. The dataset from state 41 may not represent the broader national trend due to sampling bias. If the respondents in state 41 are not reflective of the national population, the ratio derived from this state will introduce bias when applied to other states. 4. Ratio estimators rely on

a linear relationship between the number of doctoral degree holders and the total population, which may not be valid across all states. In reality, the relationship may be more complex, with certain factors (e.g., urbanization, economic development) influencing education levels in nonlinear ways.

quarto render document.qmd –to pdf