Insights into Educational Trends Through Data Analysis*

Wei Wang Chiyue Zhuang

November 21, 2024

```
library(haven)
library(labelled)
library(tidyverse)
```

```
-- 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 becor
```

```
ipums_data <- 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.
```

*Code and data are available at: [https://github.com/zcyjn233/Week-5-reflection).

```r
ipums_data <- ipums_data |>
  select(STATEICP, EDUCD) |>
  rename(stateicp=STATEICP, educd = EDUCD) |>
  to_factor()
```

0.0.1 Task:

Using the provided codebook, determine the number of respondents from each state (STATEICP) whose highest educational attainment was a doc degree (EDUC). Create a column in a tibble reflecting this.

```r
doc_counts <- ipums_data |>
  filter(educd == 116) |>
  group_by(stateicp) |>
  summarise(doc_count = n()) |>
  ungroup()

doc_counts
```

```
# A tibble: 51 x 2
   stateicp doc_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
```

0.0.2 Instructions for Data Access:

To obtain the data from IPUMS USA:

1. Visit the IPUMS USA website and select “IPUMS USA.”
2. Click “Get Data,” and under “SELECT SAMPLE,” choose “2022 ACS.”

3. For state-level data, go to "HOUSEHOLD" → "GEOGRAPHIC" and add "STATEICP" to your cart.
4. For individual-level data, navigate to "PERSON" and add "EDUC" to the cart.
5. Review your cart and click "CREATE DATA EXTRACT." Ensure the data format is set to ".dta."
6. Submit the extract, log in, or create an account, and you'll receive an email once the data is ready.
7. Download and save the data extract locally (e.g., "usa_00004.dta") for analysis in R.

0.0.3 Your estimates and the act number of respondents

```
total_respondents_california <- 391171

doc_respondents_california <- doc_counts |>
  filter(stateicp == 71) |>
  pull(doc_count)

doc_ratio_california <- doc_respondents_california / total_respondents_california

est_total_counts <- doc_counts |>
  mutate(est_total = doc_count / doc_ratio_california)

act_cnt <- ipums_data |>
  group_by(stateicp) |>
  summarise(act_total = n()) |>
  ungroup()

contrast <- doc_counts |>
  left_join(act_cnt, by = "stateicp") |>
  left_join(est_total_counts, by = "stateicp") |>
  select(stateicp, act_total, est_total)

contrast
```

```
# A tibble: 51 x 3
   stateicp act_total est_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
```

0.0.4 Reasons for Differences Between est and act Respondent Counts

The differences between the est and act respondent counts in each state using the ratio estimators method arise due to several factors. One key reason is the assumption of similarity inherent in the ratio estimator approach. This method presumes that the proportion of individuals with doc degrees in California is representative of the proportions in other states. However, the reality is that educational attainment varies significantly across states, influenced by differing demographics, economic opportunities, and educational infrastructure. Such state-level differences result in notable discrepancies between est and act respondent numbers.

Another factor is sampling variability. Estimates derived from sample data, as opposed to a full population census, are subject to random sampling variability. This variability can affect the calculated ratio, which in turn impacts the accuracy of the est respondent counts. Since sample data inherently includes some degree of randomness, the estimates may not always align closely with the act values.

In addition, educational attainment is not uniformly distributed across the United States. Different states and regions have unique policies, cultural norms, and varying levels of access to higher education, all of which influence the number of individuals attaining doc degrees. Consequently, a ratio derived from California may not be applicable to other states, where the conditions that shape educational outcomes could be markedly different.

Furthermore, there may be bias in the ratio estimator itself. The ratio estimator is most effective when the relationship between the characteristic of interest—in this case, the prevalence of doc degrees—and the population is consistent across all units. However, if the ratio of doc degree holders to the total population in California is not indicative of the situation in other states due to unobserved factors, then the estimates produced will be biased.

These reasons collectively explain why the ratio estimators method often leads to differences between est and act respondent counts. The assumption of homogeneity, which underpins this method, can lead to inaccuracies when applied to a diverse population, such as the various states across the United States, each with its unique characteristics.