Lab #4 - Racial Bias in the Labor Market

Econ 224

September 4th, 2018

Introduction

In our last lab, we looked at experimental evidence for racial bias in the labor market. Today we'll look at the same question using an *observational* dataset drawn from the US Current Population Survey (CPS). The dataset is available from http://masteringmetrics.com/wp-content/uploads/2015/02/cps.dta. Download and save this file in an appropriate location on your machine before continuing. Recall from last time that we use the function `read_dta` from `haven` to open files of this format in R. But before we examing the `cps` dataset, we will briefly revisit the data from Bertrand & Mullainathan from last time.

Exercise #1

1. Visit census.gov to find information about the CPS. Answer the following questions:

- What data is the CPS primarily concerned with?
- How often is the CPS carried out?
- How many households are included in the CPS?

2. Use `dplyr` to calculate all the summary statistics you'll need to test the null hypothesis that there is no difference between callback rates for black and white-sounding names. Hint: you'll need to use the `dplyr` function called `n()` to calculate the sample size for each group. Look this up in *R for Data Science*, online, or in the R Help files to find out how it works.
3. Write R code to calculate the p-value for the test of the null hypothesis that there is no difference in callback rates across black and white-sounding names against the two-sided alternative, using the summary statistics you calculated in part 1. Do this "the hard way" i.e. not by using a built-in function like `t.test` to do it for you: I want to see that you understand all the steps in the calculation.
4. It's a pain doing tests by hand. Figure out how to carry out the test from part 2 *without* manually computing all the of the summary statistics first. Hint: read the help file for the base R function `t.test` and the `dplyr` function `pull`.
5. Compare and interpret your results from parts 3 and 4.

Solution to Exercise #1

Write your code and solutions here

1. Regarding the CPS:

- Primarily labor market data: employment and earnings; also collects demographic data
- Monthly
- 60,000 households

```
library(haven)
library(tidyverse)

# Part 2
bm <- read_dta('~/econ224/labs/lakisha_aer.dta')
summary_stats <- bm %>%
  group_by(race) %>%
  summarize(p_call = mean(call),
            sample_size = n())
# Part 3
summary_stats
```

```
# A tibble: 2 x 3
  race  p_call sample_size
  <chr>  <dbl>       <int>
1 b     0.0645        2435
2 w     0.0965        2435
```

```
p <- 0.0645
q <- 0.0965
n <- m <- 2435
SE <- sqrt(p * (1 - p) / n + q * (1 - q) / m)
test_stat <- abs(p - q) / SE
test_stat
```

```
[1] 4.111147
```

```
2 * pnorm(1 - test_stat)
```

```
[1] 0.001863624
```

```
# Part 4
call_black <- bm %>%
  filter(race == 'b') %>%
  pull(call)
call_white <- bm %>%
  filter(race == 'w') %>%
  pull(call)
t.test(call_black, call_white)
```

```

    Welch Two Sample t-test

data:  call_black and call_white
t = -4.1147, df = 4711.6, p-value = 3.943e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04729503 -0.01677067
sample estimates:
 mean of x  mean of y
0.06447639 0.09650924
```

The cps Dataset

The cps dataset contains information on employment, race, sex, education, and years of experience for 8,891 individuals living in Boston and Chicago in 2001. Note that these are the *same* cities used by Bertrand and Mullainathan in their experiment, which was carried out between 2001 and 2002. Three of the variables in this dataset are binary: `employed` equals 1 if a given individual was employed at the time of the survey, `black` equals 1 if the individual is black, and `female` equals 1 if the individual is female. The variable `education` takes on four values: 1 indicates high school dropout, 2 indicates high school graduate, 3 indicates some college, and 4 indicates a college degree. Finally, `yearsexp` gives years of experience.

Exercise #2

1. Read in the `cps` dataset and store it in a tibble called `cps`.
2. Create a dummy variable called `somecollege` that takes the value 1 if `education` equals 3 or 4 and store it in the tibble `cps`.
3. Calculate the means of `employed`, `somecollege`, and `yearsexp` separately for blacks and whites.

Solution to Exercise #2

Write your code and solutions here

```
# Part 1
cps <- read_dta('~/econ224/labs/cps.dta')
cps <- cps %>%
  mutate(somecollege = (education == 3 | education == 4))
cps %>%
  group_by(black) %>%
  summarize(mean(employed, na.rm = TRUE),
            mean(somecollege), mean(yearsexp))
```

```
# A tibble: 2 x 4
  black `mean(employed, na.rm = TRUE)` `mean(somecollege~ `mean(yearsexp)`
  <dbl>                          <dbl>              <dbl>            <dbl>
1     0                          0.795              0.642             21.0
2     1                          0.704              0.526             20.6
```

Solution to Exercise #???

Write your code and solutions here