Homework Assignment 2

DPI 610

Introduction

You are employed at a political consulting firm that will be working on several races in Florida in 2020. Your boss has assigned you to look back over data gathered from an experiment conducted by your firm preceding the 2014 general election that (for some strange reason) was never analyzed (the data set is named `fl` and pre-loaded).

In the experiment, the firm mailed postcards to a randomly selected set of registered voters, showing the recipient their past turnout history and comparing it to the level of participation of the typical person in their state.[1] Your task is to determine what lessons, if any, can be drawn from this experiment and applied to future efforts at boosting turnout among target voters.

Question 1: Sample Population and Randomization

Question 1(a)

The 2014 American Community Survey (ACS) estimates that the demographic breakdown (in terms of Race) in Florida is the following:

Race	Pct
White	56.6
Black	15.4
Hispanic	23.3
Other	4.7

How does the demographic breakdown of the sample of voters in our `fl` data compare to the ACS estimates in terms of race? Use the `race` variable to compute the proportion of racial categories.

Briefly interpret the result.

Answer 1(a)

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 4 x 4
##   race    pct acs_pct pct_diff
##   <chr> <dbl>   <dbl>    <dbl>
```

[1] The data comes from the replication archive of the paper "The Generalizability of Social Pressure Effects on Turnout Across High-Salience Electoral Contexts: Field Experimental Evidence From 1.96 Million Citizens in 17 States" by Alan Gerber, Greg Huber, Albert Fang, and Andrew Gooch.

```
## 1 B     27.6      15.4   12.2
## 2 H     32.4      23.3    9.12
## 3 O      5.52      4.7    0.823
## 4 W     34.4      56.6  -22.2
```

The sample data from the fl dataset has around 12% more Black voters, 9% more Hispanic voters, less than 1% more Other-identifying voters, and 22% less white voters.

Question 1(b)

In terms of gender, the 2014 ACS estimates were that 51.1 percent of the population in Florida is female. How does our sample compare?

Answer 1(b)

```
##       gender pct acs_pct  diff
## 1     Female  68    51.1  16.9
## 2 Not Female  32    48.9 -16.9
```

The sample data from the fl dataset has around 17% more female voters.

Question 1(c)

In terms of turnout, the breakdown between 2006 and 2014 in Florida[2] was as follows:

Year	Pct
2006	47
2008	75
2010	49
2012	72
2014	51

How does our sample compare?

Answer 1(c)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 4
##    year   pct acs_pct  diff
##   <dbl> <dbl>   <dbl> <dbl>
## 1  2006    19      47   -28
## 2  2008    61      75   -14
## 3  2010    27      49   -22
## 4  2012    54      72   -18
## 5  2014    34      51   -17
```

[2] Available at https://dos.myflorida.com/elections/data-statistics/elections-data/voter-turnout

The sample data from the fl dataset has around 28% less 2006 voters, 14% less 2008 voters, 22% less 2010 voters, 18% less 2012 voters, and 17% less 2014 voters than the ACS data.

Question 1(d)

One reason why randomized experiments allow us to estimate the causal impact of an intervention is because, through randomization, the observable and unobservable characteristics of the subjects in the experiment are independent of, or at least uncorrelated with, treatment assignment. One consequence of this is that covariates will be "balanced" between the treatment and control group. You can check for balance by calculating the mean of a variable for the treatment group and for the control group.

Is the turnout history of subjects in the experiment between 2006 and 2012 balanced between treatment and control?

Answer 1(d)

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```
## # A tibble: 10 x 3
## # Groups:   year [5]
##    year  treat   pct
##    <chr> <int> <dbl>
##  1 06        0  18.4
##  2 06        1  18.6
##  3 08        0  61.5
##  4 08        1  61.4
##  5 10        0  26.8
##  6 10        1  27.2
##  7 12        0  53.9
##  8 12        1  54.4
##  9 14        0  32.4
## 10 14        1  33.8
```

Yes, the turnout history of subjects in the experiment between 2006 and 2012 has negligible difference in the percentages per group such that it is balanced between treatment and control.

Question 1(e)

Are gender, marriage, age and race balanced between treatment and control?

Answer 1(e)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 8
##   treat pct_female pct_married mean_age pct_white pct_black pct_hispanic
##   <int>      <dbl>       <dbl>    <dbl>     <dbl>     <dbl>        <dbl>
## 1     0       68.5        20.8     46.3      34.7      27.4         32.1
## 2     1       67.6        20.7     46.5      34.4      27.6         32.4
## # ... with 1 more variable: pct_other <dbl>
```

Yes, the turnout history of subjects in the experiment has negligible difference in the percentages per gender, marriage, age and race such that it is balanced between treatment and control.

Question 1(f)

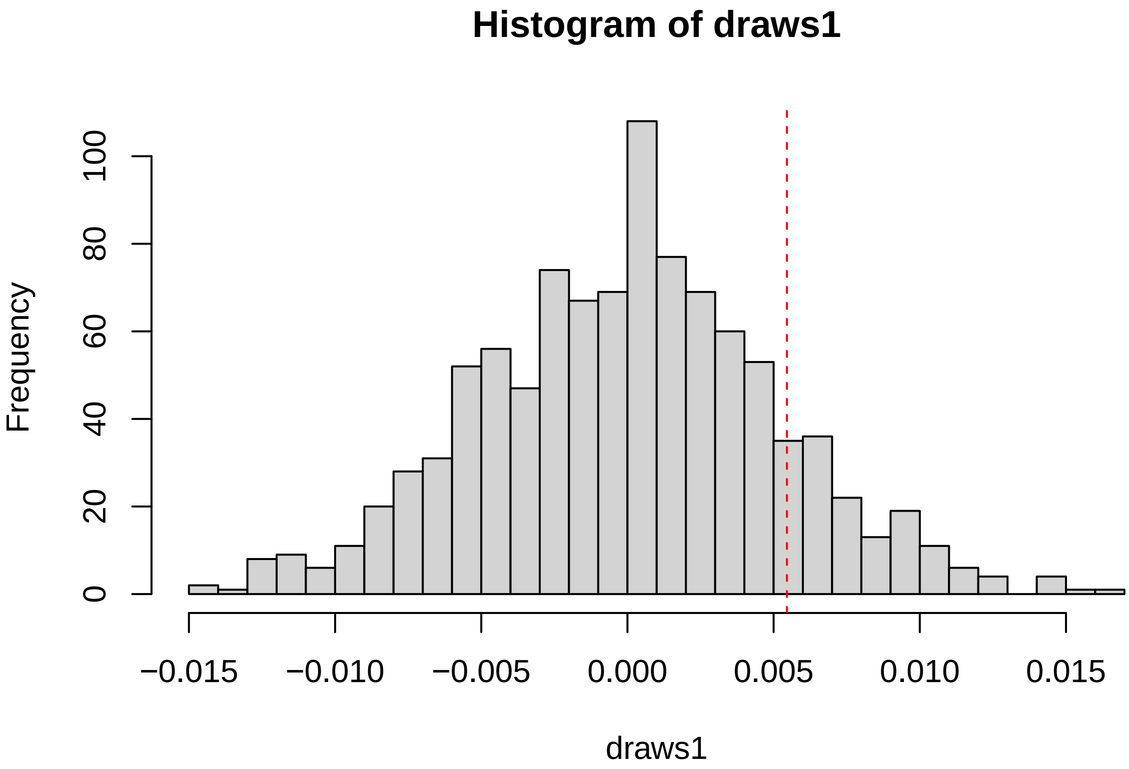
What is the probability that the observed difference in 2012 turnout rate between treatment and control would occur due to random chance? Use permutation inference to determine this (use `set.seed(1234)` so your results are replicable).

Specifically, first compute the difference-in-means on `voted12` between the treatment and control group. Then, permute the treatment assignment 1000 times, and compute the difference-in-means based on each permuted treatment assignment. Finally, plot the histogram of difference-in-means over the 1000 permutations. Show the observed difference-in-means (i.e., the actual value that we observed in the data) as the vertical line.

Compute the two-sided p-value based on the permuted difference-in-means.

Can we reject the null hypothesis of no effect on the past outcome?

Answer 1(f)



No, we cannot reject the null hypothesis of no effect on the past outcome. The actual difference in means appears significantly closely to the center of the distribution. Because the histogram shows normally distributed behavior, we can assume that the center of the distribution is also the mean of the distribution, which means that within one standard deviation of the center of the distribution will immediately capture more than two-thirds of the replicate date we've devised. Since our actual difference in means in our original observed sample is so close to the mean, there's a significant chance that this event occurs at random. Therefore, we fail to reject the null hypothesis of no effect on the past outcome. In addition, the p-value of 0.2942395 is far below our actual difference in means, which further asserts our conclusion that it is not a statistically significant result.

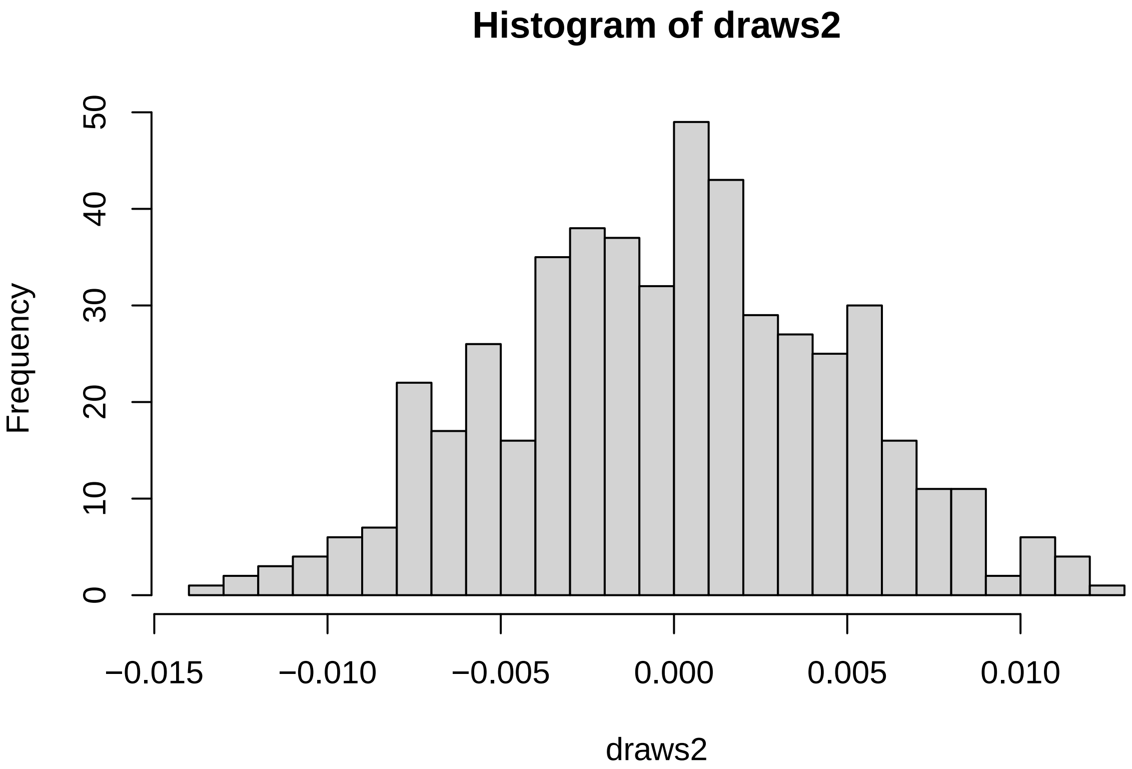
Question 2: Estimating Treatment Effects

In Question 2, we estimate the treatment effect using the `fl` data.

Question 2(a)

What is the overall 2014 turnout rate for people in the sample? What is the 2014 turnout rate for people in the treatment group? In the control group? What is the estimate of the average treatment effect (ATE) from the experiment? Using permutation inference, determine the statistical significance of the difference in turnout rates for people in the treatment group versus the control group (use 500 draws and `set.seed(02138)` so your results are replicable).

Answer 2(a)



The overall 2014 turnout rate from the sample is 0.337539. For the treatment group, the 2014 turnout rate is 0.338247; for the control group, it's 0.3242014. The estimated ATE is 0.0140456. Since the p-value is 0.2942395, the difference in turnout rates for people in the treatment group versus the control group is statistically significant.

Question 2(b)

The intervention involved sending a postcard with someone's past voting history on it and telling the recipient if he or she voted more than the typical person in the state. Would you hypothesize that the ATE is: (1)

larger for people with a history of regular voting, (2) larger for people with a history of not voting, or (3) the same for both types? Why?

Answer 2(b)

I would hypothesize that the ATE is larger for people with a history of not voting because they would see that they participate less than the typical person in their state, which could induce social pressure to be a more active citizen that can contribute to increased voter turnout.

Question 2(c)

Create a new variable called `vote_history` and store the individual-level propensity to vote by computing the proportion of turnout for each person between 2006 and 2012 (use `voted06`, `voted08`, `voted10`, and `voted12`).

Then, determine the average past turnout in the state according to this measure by computing the mean of `vote_history`. Create a second new variable that indicates people with lower than average turnout in the Florida data.

Answer 2(c)

```
## # A tibble: 191,886 x 3
## # Groups:   id [191,886]
##         id vote_history lower
##      <int>        <dbl> <dbl>
##  1 1833919         0.75     0
##  2 1296025         0        1
##  3 1267933         0.75     0
##  4  169993         0        1
##  5 2111289         0.75     0
##  6  742314         0        1
##  7 1321908         0.25     1
##  8  598131         0.25     1
##  9 1008697         0.5      0
## 10  332225         1        0
## # ... with 191,876 more rows
```

The average past turnout in the state according to a voter's voter history is 0.404100611821602.

Question 2(d)

Estimate the average treatment effects for (1) those with lower past turnout than average in the state and (2) those with past turnout greater than or equal to the average in the state. Which effect is larger? Does this support your hypothesis from Question 2(b)?

Answer 2(d)

```
## `summarise()` regrouping output by 'lower' (override with `.groups` argument)
## # A tibble: 4 x 3
## # Groups:   lower [2]
##   lower treat  mean
##   <dbl> <int> <dbl>
## 1     0     0 0.504
## 2     0     1 0.521
## 3     1     0 0.113
## 4     1     1 0.119
```

Regardless of whether or not the participant has a lower level of voter turnout than the average for the state, the treatment makes the mean turnout for 2014 higher than the control group's mean turnout. This supports my hypothesis because it shows that the treatment can possibly actively increase turnout.

Question 2(e)

What do these results suggest about the ability of social-pressure based GOTV efforts to turn out voters who have not voted in the past?

Answer 2(e)

The results suggest that there is a social pressure to participate more actively in politics, most likely in order to gain respect as an individual civic agent in society, such that when reminded of it, an individual is influenced to consider voting more often—hence, the 2014 election's higher voter turnout.

Question 2(f)

Estimate the difference in treatment effects for (1) men versus women and (2) people whose age is young versus middle versus older (18–29, 30-64 and 65+). Briefly interpret the result.

Answer 2(f)

```
## # A tibble: 2 x 4
## # Groups:   female [2]
##   female Control Treated   diff
##    <int>   <dbl>   <dbl>  <dbl>
## 1      0   0.281   0.299 0.0178
## 2      1   0.344   0.357 0.0131
```

```
## # A tibble: 3 x 6
## # Groups:   young, middle, older [3]
##   young middle older Control Treated    diff
##   <dbl>  <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1     0      0     1   0.408   0.426 0.0182
## 2     0      1     0   0.338   0.347 0.00929
## 3     1      0     0   0.208   0.233 0.0248
```

For gender, the treatment increased both observed genders' voter performances by 1.78% for male voters and 1.31% for female voters. For age groups, the treatment increased all observed gender groups' voter performances by 2.48% for young voters, 0.93% for middle-aged voters, and 1.82% for older voters.

Question 3: Optional Challenge Question

In the experiment conducted in 2014, the subjects' party registration statuses were not available. Suppose you would like to learn whether untargeted efforts at boosting turnout in Florida through the social pressure mailing helped the prospects of the Democratic Party or the Republican Party overall. Use the survey results from the CCES conducted in Florida to predict who leans towards the Republican party based on age, gender, race, and marital status. You may use stated party leanings (based on the variable `pid3`) as the outcome. Then, make predictions about who is likely to vote Republican (versus Democrat or Independent) in the experimental sample and calculate treatment effects for these groups. What do you find?

Answer 3

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## # A tibble: 3 x 4
##   group          control treatment difference
##   <chr>            <dbl>     <dbl>      <dbl>
## 1 All in Sample     32.4      33.8       1.40
## 2 Not Republican    30.8      32.0       1.15
## 3 Republicans       38.0      40.3       2.22
```