Improving Voting Turnout

Kahsif Khan

30 August 2022

Name:Kashif_Khan

Laws regarding the conditions of voting are being actively debated and changing right now in the US. Voter turnout rates in the US are often low by international standards. Many policies and interventions have been considered to increase voter turnout such as making voting day a national holiday (so work does not keep people from voting) or making mail-in ballots much more widely accessible. Other policies being discussed or passed may further decrease voter turnout while some claim they will decrease voter fraud (voter fraud is a topic in the US where misinformation has been common).

Some of these interventions have been tested and there is empirical evidence on how successful they are.

In this homework, we will be considering data from a 2006 randomized experiment where people in a large US city were randomly assigned to a control group that received no treatment (usual voting conditions) or to a treatment group that received a letter telling them that their neighbors would be informed about whether or not they voted (the social psychologists call this a 'social pressure intervention'). You may find out more about this study in the paper "Social Pressure and Voter Turnout: Evidence from a Large Scale Field Experiment" DOI: https://doi.org/10.1017/S000305540808009X. The researchers then recorded whether each potential voter voted in the 2006 primary election.

We have created a data set with some of the data from this experiment called `social_HW1.csv` The table below displays the names and descriptions of variables in the `social_HW1.csv` data file.

Name	Description
`age`	Age of potential voter
`Voted`	Whether or not the potential voter voted in the 2006 primary
`message`	Whether the potential voter was treated or control
`hhsize`	Size of potential voters household
`gender`	Gender of the potential voter

A note about the variables: In this data set 'gender' is measured using only two categories. It is not known how those not identifying with either of the two categories available were categorized.

Question 1 (8 points)

Load the `tidyverse` packing using the command `library(tidyverse)` and follows the steps we discussed in lab to download the data to the correct folder and then load the data into R.

How many observations (rows) are there? What does each row correspond to? How many columns are there? What is the range of ages of those in this data set? What was the overall voter turnout rate? What proportion of study participants are recorded here as female? How many people are in the control group and how many people are in the treatment group? What proportion of study participants are in the treatment group?

Answer 1 Code

```
# load library tidyverse
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Load the data...
social <- read_csv("/Users/kashifkhan/Documents/CMU/3rd Semester/Stat/Assignments/Assignment 1/data/soci
```

```
## Rows: 130603 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): voted, message, gender
## dbl (2): age, hhsize
##
## 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.
```

```
#Numeric coding the dichotomous outcome variable of voted
social$votedN <- if_else(social$voted == "yes", 1, 0)
# Categorizing the continuous variable of ages of the participants in different categories
social$agecat = case_when(social$age <= 30 ~ "under 30",
                          social$age >= 30 & social$age <= 44 ~ "30-44",
                          social$age >= 45 & social$age<= 59 ~ "45-59",
                          social$age >= 60 ~ "above 60")
#Q1
# Finding range of ages
range(social$age)
```

```
## [1]  20 106
```

```
# Finding overall turnout
proportions(table(social$voted))
```

```
##
##        no       yes
## 0.7096698 0.2903302
```

```
# Finding proportion of female
proportions(table(social$gender))
```

```
##
##    female      male
## 0.5014127 0.4985873
```

```
# Finding number of the voters in control and treatment groups
table(social$message)
```

```
##
##   control treatment
##    108833     21770
```

```
# Finding proportion of the study group in the treatment group
proportions(table(social$message))
```

```
##
##   control treatment
## 0.8333116 0.1666884
```

Answer 1 Text

There are 130603 observations i.e number of participants in this study.Each row corresponds to a unique voter who may also be called the study participant. There are five columns as there are five categorical variables. The ages of study participants are between 20 and 106. Overall turnout rate was 29.03%. Female proportion of the voters is 50.14%. In the control group there are 108833 people while in treatment group there are 21770 people. 16.66% of the total participants are in the treatment group.

Question 2 (14 points)

What is the specific causal question the researchers who carried out this randomized experiment wanted to answer? What impact do they hypothesize their intervention will have? What are the potential outcomes for one study participant before they are randomized to the treatment or control group? For those randomly assigned to the intervention group what is their factual outcome? What is their missing counterfactual outcome? As this is a randomized experiment, how will we estimate the average MCF for those in the intervention group?

Answer 2 Text

#SCQ What is the impact of ‘social pressure intervention’ relative to ‘usual voting condition’ on the voters turnout in 2006 primary elections for the people of the large US city? #What impact do they hypothesize their intervention will have? The researcher’s hypothesis is that social pressure intervention will increase voter turnout rate of 2006 primary elections, relative to the alternative of usual voting condition. #What are the potential outcomes for one study participant before they are randomized to the treatment or control group? The potential outcomes for one study participant are: (i) Whether the study participant would vote if they get social pressure treatment. (ii) Whether the study participant would vote if they get usual voting condition. #For those randomly assigned to the intervention group what is their factual outcome? Whether they vote after they receive the social pressure treatment. #What is their missing counterfactual outcome? Whether they vote if instead of receiving social pressure treatment they had received usual voting condition. #As this is a randomized experiment, how will we estimate the average MCF for those in the intervention group? In the randomized experiment, the given set of participants are randomly allocated to any of the treatment or the control groups. The average of the outcome of the treatment group then becomes the factual outcome and the average of outcome of the control group becomes the MCF. In this study, the missing counter factual outcome is whether the people in the large US city have voted if they had not got the social pressure treatment but all else remained the same.

Question 3 (8 points)

Calculate the voter turnout rate among those assigned to the control group. For those in the control group, what are the turnout rates by gender? For those in the control group, what are the turnout rates among those younger than 30, those 30-44, those 45-59, and those 60 and older (create a new `agecat` variable with these categories)? Briefly describe what you have learned about voter turnout rates for this primary election in this city when there is no intervention (under usual voting conditions). In particular, under normal voting conditions, who in this city, for this election, is more or less likely to vote?

Answer 3 Code

```r
# Voter turnout rate among those assigned to the control group
social.filtered <- social %>%
  filter(message == "control")
mean(social.filtered$votedN)
```

```
## [1] 0.2761111
```

```r
# for those in the control group, what are the turnout rates by gender?
tapply(social.filtered$votedN, social.filtered$gender, mean)
```

```
##    female      male
## 0.2726505 0.2795826
```

```r
# turnout rates in different age categories
tapply(social.filtered$votedN, social.filtered$agecat, mean)
```

```
##     30-44     45-59  above 60  under 30
## 0.2484061 0.2828863 0.3386371 0.1571514
```

Answer 3 Text

Out of the control group 27.61% have voted. Out of the control group, 27.26% of the females and 27.95% of the males have voted. In different age categories in this group the turnouts are: under 30 is 15.72%, from 30-44 is 24.84%, between 45-59 is 28.29%, and above 60 the turnout is 33.86%. Under usual voting conditions, importantly, only 27.61% have voted (will be quite interesting to see the comparative turnout in the treatment group). Further, among the ones who have voted from this group, there is almost equal voting turnout between the males and females. The voting turnout among people under 30 is quite low compared to the people falling in other age categories whereas, in this control group, turnout is comparatively the highest in the people above 60 years of age and these are the people who are likely to vote more under normal voting conditions.

Question 4 (8 points)

Repeat Question 3 for those randomly assigned to the social pressure treatment group.

Answer 4 Code

```
# Voter turnout rate among those assigned to the treatment group
socialT.filtered <- social %>%
  filter(message == "treatment")
mean(socialT.filtered$votedN)
```

```
## [1] 0.3614148
```

```
# for those in the control group, what are the turnout rates by gender?
tapply(socialT.filtered$votedN, socialT.filtered$gender, mean)
```

```
##    female      male
## 0.3546977 0.3682551
```

```
# turnout rate in different age categories
tapply(socialT.filtered$votedN, socialT.filtered$agecat, mean)
```

```
##     30-44     45-59  above 60  under 30
## 0.3388900 0.3613310 0.4452732 0.2045550
```

Answer 4 Text

In the treatment group 36.14% have voted. Out of the treatment group, 35.47% of the females and 36.83% of the males have voted. In different age categories in this group the turnouts are: under 30 is 20.46%, from 30-44 is 33.89%, between 45-59 is 36.13%, and above 60 the turnout is 44.53%. Under treatment voting conditions, 36.14% have voted. Further, among the ones who voted from this group, there is a small difference in the voting turnouts of the males and females. The voting turnout among people under 30 is quite low as compared to the people falling in other age categories whereas, in this treatment group, turnout is very high in the people above 60 years of age and these are the people who are likely to vote more under treatment voting conditions.

## Question 5 (8 points)

In this problem we will comparing the results in Question 3 to the results of Question 4 to learn about treatment effects of the social pressure intervention. Calculate the difference in voter turnout rates between the intervention and control groups (treatment - control). Explain whether the social pressure intervention seemed to be successful overall in improving voter turnout rates. Then calculate these differences by subgroup and state whether the impact (if present) differed by gender or age group - and if so for whom the impacts were larger/smaller. Note that at this stage we are not running statistical significance tests to formally test whether the intervention was successful (whether differences of these sizes are likely to occur by chance with these sample sizes). For this HW problem just state the magnitude of the differences in voter turnout that you observe in the data.

Answer 5 Code

```
# Differences in mean overall (treatment - control)
mean(socialT.filtered$votedN) - mean(social.filtered$votedN)
```

```
## [1] 0.08530368
```

```
# Differences in mean by gender
tapply(socialT.filtered$votedN, socialT.filtered$gender, mean) - tapply(social.filtered$votedN, social.
```

```
##     female       male
## 0.08204720 0.08867259
```

```
# Differences in mean by age categories
tapply(socialT.filtered$votedN, socialT.filtered$agecat, mean) - tapply(social.filtered$votedN, social.
```

```
##      30-44      45-59   above 60   under 30
## 0.09048396 0.07844469 0.10663611 0.04740367
```

Answer 5 Text

Overall, the social pressure intervention caused an almost 8.5% increase in the voter turnout rate (relative to the normal voting conditions control group).Similar voting turnout increase is reflected in the gender groups where it is increased by 8.2% in the females falling in the treatment group and 8.87% increase in turnout rate of the males of this group. Among the people in different age categories, the impact differed with age, the turnout rate is highest in the people above 60 (10.7%) whereas lowest in the people under age 30 (4.7%) while in the age category of 30-44 it is 9% high and in the category of 45-59, it is 10.7% high as compared to the control group.

Question 6 (8 points)

The way you just estimated the MCF should be unbiased because registered voters were randomly assigned to the intervention and control groups. While unlikely given the randomization and large sample sizes, if (due to an unlucky randomization) the groups strongly differ on baseline covariates that are associated with the outcome, then this estimate of the MCF could be problematic. Gender, age, and household size are baseline covariates in this experiment (baseline because they are features that existed prior to treatment randomization). Compare the gender, age, age category, and household size distributions of the intervention and control groups. For dichotomous or categorical variables compare proportions, for continuous or count variables compare means. Comment on whether or not these baseline covariates suggest problems with your estimate of the average MCF for the intervention group. Again we are not carrying out statistical inference tests on differences at this point in the class, just state the magnitude of any differences you see.

Answer 6 Code

```
# gender comparison
# Both age and household size are continuous or count variables so we apply tapply  and take mean of th
tapply(social$age, social$message, mean)
```

```
##   control treatment
##  50.49160  50.48085
```

```
tapply(social$hhsize, social$message, mean)
```

```
##   control treatment
##  2.173486  2.179972
```

```
# Both age categories and gender are categorical variables so will compare proportions for these
proportions(table(social.filtered$agecat))
```

```
##
##     30-44     45-59  above 60  under 30
## 0.2219364 0.4107853 0.2592872 0.1079911
```

```
proportions(table(socialT.filtered$agecat))
```

```
##
##     30-44     45-59  above 60  under 30
## 0.2259531 0.4003215 0.2648140 0.1089113
```

```
proportions(table(social.filtered$gender))
```

```
##
##    female      male
## 0.5007856 0.4992144
```

```
proportions(table(socialT.filtered$gender))
```

```
##
##    female      male
## 0.5045475 0.4954525
```

Answer 6 Text

On the basis of age both treatment and control groups are almost similar as mean ages of the people in the control and treatment groups are 50.49 and 50.48 respectively. The household size comparisons both of the groups also reveal similarity in the mean for both of the groups where control group has mean size of 2.17 and the treatment group has mean size of 2.18. Comparison on the basis of age categories between the two groups also reveal homogeneity as almost equal percentages of people from each of the age categories fall in both of the treatment and control groups. Similarly, gender proportions between the two groups also reinforce the similarity that is resulting from randomization. In short, all these baseline characteristics are unlikely to bias the treatment effect estimated in this randomized experiment.