Analyzing an Experiment in R

STSCI/INFO/ILRST 3900: Causal Inference

September 11, 2024

Reminders and Announcements

- Peer Reviews

- See Ed Discussion for more instructions and details!

- Office Hours throughout the week (see Syllabus or website)

  - Filippo: Monday 11am-12pm in Comstock 1187

  - Shira: Wednesday 3-4pm in in Comstock 1187

  - See Ed Discussion for Zoom links/info

Agenda for Today

- Analyzing an experiment in R

- Technical/Homework/Peer-Review Questions

Get out and Vote Experiment

- Why do people vote?

- One long-standing theory: People vote due to social norms (civic duty)

- Empirical evidence for this theory was extremely thin

- **Research Question:** to what extent do social norms cause voter turnout?

Get out and Vote Experiment

- **Research Question:** to what extent do social norms cause voter turnout?

- Article: "Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment" in *American Political Science Review*

- Authors: Alan S. Gerber, Donald P. Green, and Christopher W. Larimer

- We'll be analyzing their experiment today!

http://tinyurl.com/mut4e9dj

Experimental Design

- Approximately 80k Michigan households were randomly assigned 1 of 4 mailings encouraging them to vote

  1. Simply reminded them that voting is a civic duty

  2. Told that researchers would be studying their turnout based on public records

  3. Received record of voting turnout *within* their household

  4. Received record of voting turnout within their household *and* their neighbors' households.

- Third and fourth treatment arms were told that their turnout would be revealed as well

Goal for Today
Replicate Something Similar to Tables 1 and 2 from the article

TABLE 1. Relationship between Treatment Group Assignment and Covariates (Household-Level Data)

	Control	Civic Duty	Hawthorne	Self	Neighbors
	Mean	Mean	Mean	Mean	Mean
Household size	1.91	1.91	1.91	1.91	1.91
Nov 2002	.83	.84	.84	.84	.84
Nov 2000	.87	.87	.87	.86	.87
Aug 2004	.42	.42	.42	.42	.42
Aug 2002	.41	.41	.41	.41	.41
Aug 2000	.26	.27	.26	.26	.26
Female	.50	.50	.50	.50	.50
Age (in years)	51.98	51.85	51.87	51.91	52.01
N =	99,999	20,001	20,002	20,000	20,000

Note: Only registered voters who voted in November 2004 were selected for our sample. Although not included in the table, there were no significant differences between treatment group assignment and covariates measuring race and ethnicity.

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Resources for Markdown

- Hadley Wickham's R for Data Science [Chapter 27](#)

- [RMarkdown cheat sheet](#) from RStudio

- [Data Wrangling and Analyses with Tidyverse](#) by Bookdown

- [RMarkdown for Scientists](#) by Nicholas Tierney

- If you can't figure out how to do something, try Googling it first!

- Also feel free to ask a classmate or ask me :)

- For homework sets, don't forget about Ed Discussion!

Step 1: Download the Markdown File

- Go to the course website: https://causal3900.github.io/

- Navigate to "Discussion 3. Analyzing an experiment in R"

- Download the .Rmd file and open it up on your computer

- Start by running the code in **Section "Necessary packages"**

  - If you get an error, you may need to **install** the package

  - Hint: Look for the "Tools" tab and click "Install Packages"

Step 2: Import and Clean the Data

- Calculate the ages of everyone in our dataset

- Replace the numeric labels of treatment (0-4) with word labels ("Control", "Civic Duty", "Hawthorne", "Self", and "Neighbors")

- When you run glimpse(gotv), you should see something like this ->

```
## Rows: 344,084
## Columns: 17
## $ sex          <dbl> 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, …
## $ yob          <dbl> 1941, 1947, 1951, 1950, 1982, 1981, …
## $ g2000        <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, …
## $ g2002        <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, …
## $ g2004        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ p2000        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ p2002        <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, …
## $ p2004        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, …
## $ treatment    <chr> "Civic Duty", "Civic Duty", "Hawthor…
## $ cluster      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ voted        <dbl> 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, …
## $ hh_id        <dbl> 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, …
## $ hh_size      <dbl> 2, 2, 3, 3, 3, 3, 3, 3, 2, 2, 1, 2, …
## $ numberofnames <dbl> 21, 21, 21, 21, 21, 21, 21, 21, 21, …
## $ p2004_mean   <dbl> 0.09523810, 0.09523810, 0.04761905, …
## $ g2004_mean   <dbl> 0.8571429, 0.8571429, 0.8571429, 0.8…
## $ age          <dbl> 65, 59, 55, 56, 24, 25, 47, 50, 38, …
```

Step 3: Table 1
Is the data balanced on covariates?

- We want to check that the treatment groups are *balanced on covariates*

- For each treatment arm/group, calculate the mean for each of the designated covariates

- Your table should look like this

	treatment	sex	age	g2000	g2002	p2000	p2002	p2004	hh_size
1	Civic Duty	0.5001832	49.65904	0.8417238	0.8111099	0.2535716	0.3888482	0.3994453	2.189126
2	Control	0.4989411	49.81355	0.8433773	0.8108950	0.2518837	0.3893737	0.4003388	2.183667
3	Hawthorne	0.4990053	49.70480	0.8444142	0.8129515	0.2503665	0.3943304	0.4032300	2.180138
4	Neighbors	0.5000654	49.85294	0.8416534	0.8113400	0.2511976	0.3865867	0.4066647	2.187770
5	Self	0.4995813	49.79251	0.8403893	0.8114763	0.2511120	0.3919096	0.4024805	2.180805

Step 4: Table 2
What are the results of the experiment?

- For each treatment group, calculate the percent that voted and the total number of individuals in that group

- Your table should look like this

	treatment	Percentage_Voting	num_of_individuals
1	Civic Duty	0.3145377	38218
2	Control	0.2966383	191243
3	Hawthorne	0.3223746	38204
4	Neighbors	0.3779482	38201
5	Self	0.3451515	38218