Problem Set 1

Katherine Jones

2025-10-02

#Simulation

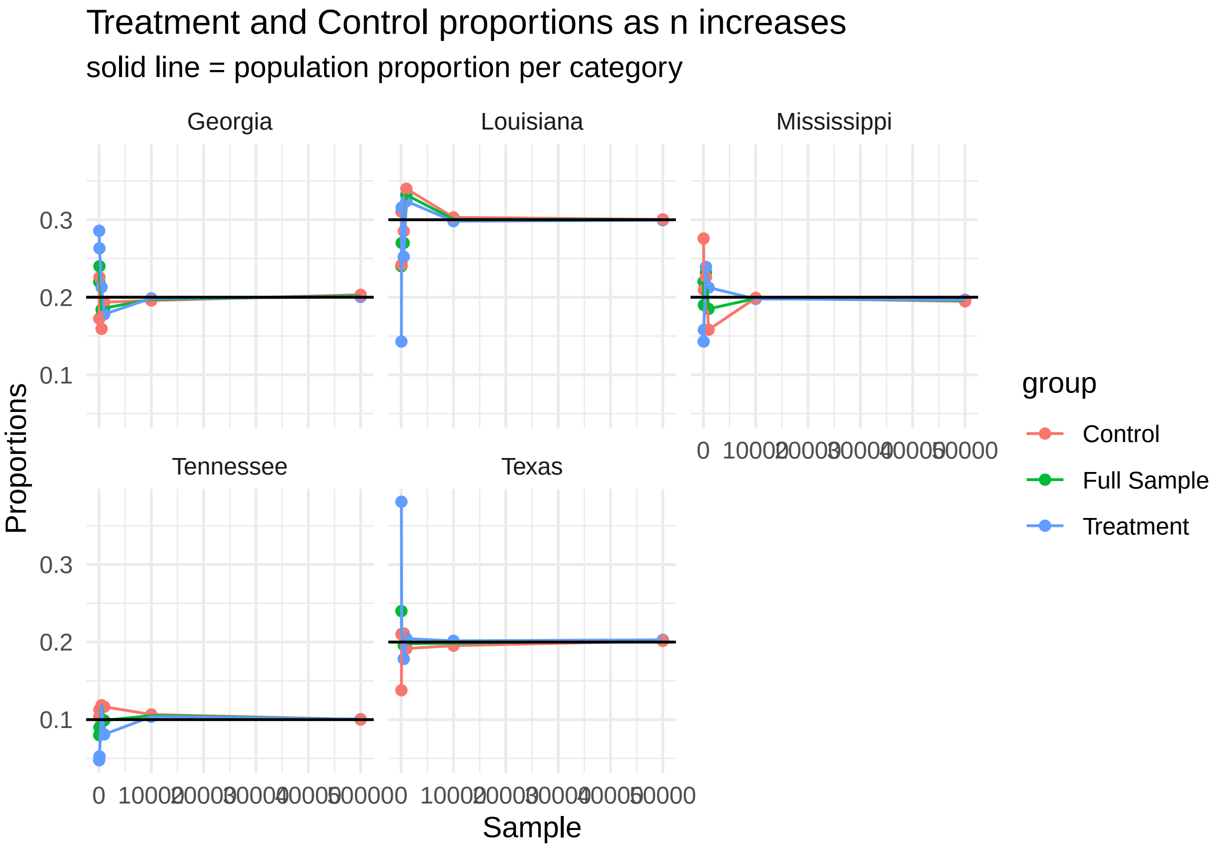
```
##Simulation Setup
set.seed(1)
library(tidyverse)
```

```
## -- 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   1.5.2
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- 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 become errors
```

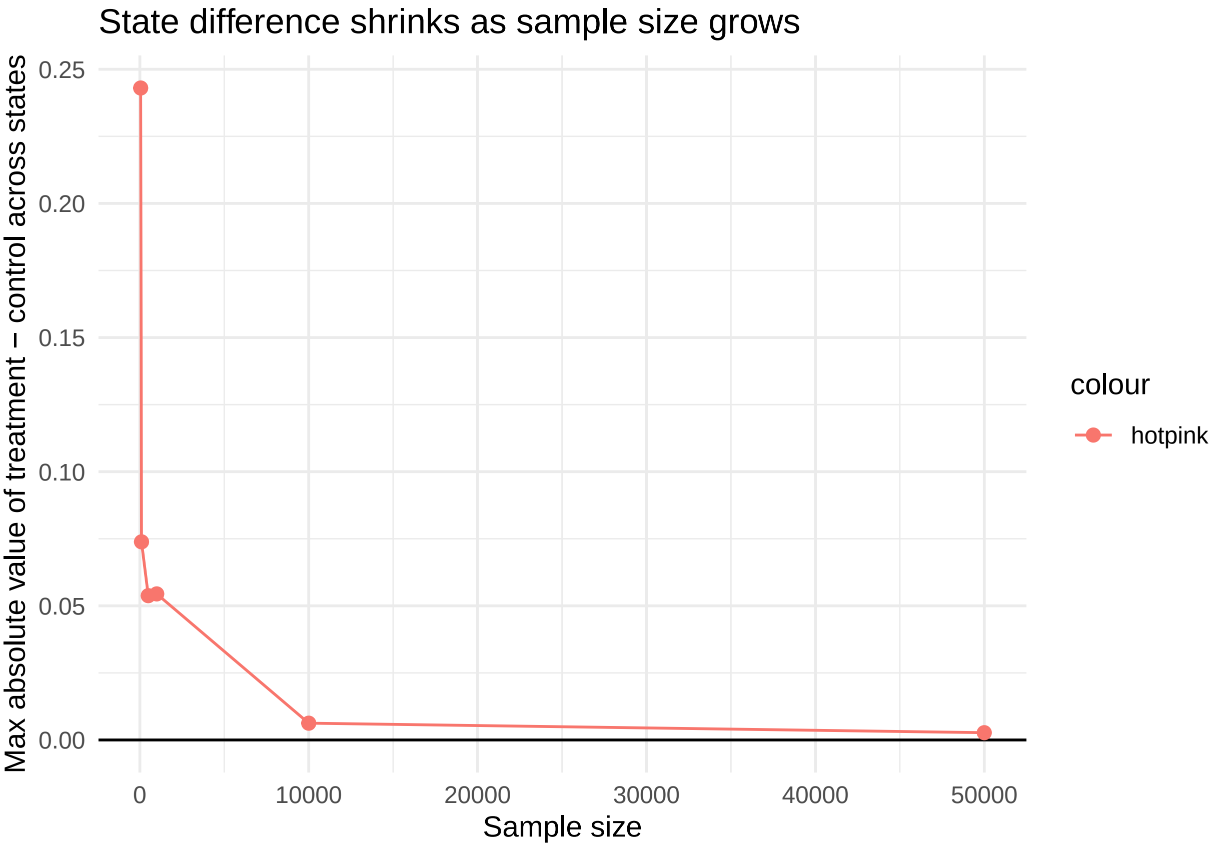
```
states <- c("Louisiana", "Texas", "Mississippi", "Georgia", "Tennessee") #traits of my population
prop_states <- c(0.3, 0.2, 0.2, 0.2, 0.1) #proportion of each state in the population
names(prop_states) <- states #assign my states a proportion

n_ob <- c(50, 100, 500, 1000, 10000, 50000) #all my n values in a container

results <- data.frame(n = integer(), group = character(), state = character(), prop = numeric() ) #need

###Actual Simulation
for(n in n_ob) {
  observation <- sample(states, size = n, replace = TRUE, prob = prop_states) #okay we sampled the stat
  Z <- rbinom(n, 1, 0.5) #okay sets treatment and control groups with EQUAL PROBABILITY
  prop_all <- as.numeric(table(factor(observation, levels = states))) / n #proportions for all
  n_treatment <- sum(Z == 1) #amount of the sample assigned to a treatment group
  n_control <- n-n_treatment #basically the other part of the sample not assigned to the treatment grou

  n_treatment + n_control #checking to make sure that no one was left behind
  # proportion time
  prop_treatment <- as.numeric(table(factor(observation[Z==1], levels = states))) / n_treatment
  prop_control <- as.numeric(table(factor(observation[Z==0], levels = states))) / n_control

  results <- bind_rows(results, #okay everything comes together here
                       data.frame(n = n, group = "Full Sample", state = states, prop = prop_all),
                       data.frame(n = n, group = "Treatment", state = states, prop = prop_treatment),
                       data.frame(n = n, group = "Control", state = states, prop = prop_control))
}
```

```
results_full <- results %>% left_join(data.frame(state = states, prop_pop = prop_states), by = "state")
  mutate(difference = prop_pop - prop) #attaching results to data frame and caluclating difference
```

```
#Simulation Results,  Plot 1

results_table <- tibble(state = states, prop_pop = prop_states)

ggplot(results, aes(x = n, y = prop, color = group)) + #line plot
  geom_point() + geom_line() +
  geom_hline(data = results_table, aes(yintercept = prop_pop), linetype = "solid") +
  facet_wrap( ~ state, nrow = 2) +
  labs(x = "Sample", y = "Proportions",
       title = "Treatment and Control proportions as n increases",
       subtitle = "solid line = population proportion per category") +
  theme_minimal()
```



```
#Simulation Results, Plot 2

difference <- results_full %>%
  filter(group %in% c("Treatment", "Control")) %>%
  select(n, state, group, prop) %>%
  pivot_wider(names_from = group, values_from = prop) %>%
  mutate(abs_diff = abs(Treatment - Control)) %>%
  group_by(n) %>%
  summarise(
    max_abs_diff = max(abs_diff),
    l1_sum_diff = sum(abs_diff)
  )

ggplot(difference, aes(x = n, y = max_abs_diff, col = "hotpink")) + #another line plot
  geom_hline(yintercept = 0, linetype = "solid") +
  geom_point(size = 2) + geom_line() +
  labs(x = "Sample size", y = "Max absolute value of treatment - control across states",
       title = "State difference shrinks as sample size grows") +
  theme_minimal()
```



To demonstrate that when treatment and control group are randomly assigned, they are comparable, I created a simulation that random assigns samples from a population that has a distribution of traits. For this simulation, my distribution of traits represents the state that sample observation is from. These traits were then assigned a proportion within the population. Once the states had their proportion, these observations were then randomly assigned with equal probability to either the treatment or control group. This process was then repeated six times across an increasing n value, which represents the observation size, to show that as the value of n increases, that the distribution within the sample observations will have similar proportions to the population distribution that I designed in the set up. In fact, both plots 1 and 2 show that as the sample size increase, my distribution of traits within my groups (treatment, control, and sample), have increasing similar distributions to population distribution. Plot 1 shows this distribution across the five states and for the treatment, control, and sample groups. Plot 2 shows the difference in the maximum absolute value of the difference in the population distribution to the different groups within my simulation. This plot also shows that as the sample size increases, the differences in the population distribution and the sample distribution decreases. Overall, the simulation and plots successfully demonstrates that as sample increases, the distribution of this sample is similar to the population distribution due to random sample assignment.

#Data Analysis

```
#Data Analysis Prep
voting <- read.csv("voting.csv")
```

1. The treatment variable is the message variable. This is a discrete variable as it is only able to take on the value “yes” or “no.” The message variable’s data type is character.

```
##Question 1 code
str(voting) #shows the variable names along with data type
```

```
## 'data.frame':    229444 obs. of  3 variables:
##  $ birth  : int  1981 1959 1956 1939 1968 1967 1941 1969 1967 1961 ...
##  $ message: chr  "no" "no" "no" "yes" ...
##  $ voted  : int  0 1 1 1 0 0 1 0 1 1 ...
```

2. I created a new binary variable for the treatment variable in the data frame.

```
#Question 2
#New treatment variable

voting$treatment <- ifelse(voting$message == "yes", 1, 0)
```

3. On average, 16.65% of those who recieve the treatment effect voted, while 31.01% of the full sample, including the treatment group, voted.

```
#Question 3 code

mean(voting$treatment) #mean for treatment group
```

```
## [1] 0.1664938
```

```
mean(voting$voted) #mean for full sample
```

```
## [1] 0.3101759
```

4. By subsetting the dataframe, I created two new data frames. One data frame represents the treatment group and the other represents the control group.

```
#Question 4 code

treatment <- voting[voting$treatment == 1, ] #first new subset
control <- voting[voting$treatment == 0, ] #second new subset

##no one needs to be left behind
nrow(treatment) + nrow(control)
```

```
## [1] 229444
```

```
nrow(voting)
```

```
## [1] 229444
```

5. The average birth year for both the treatment and control groups is 1956.

```
#Question 5 code

mean(control$birth) #mean control birth year
```

```
## [1] 1956.186
```

```
mean(treatment$birth) #mean treatment birth year
```

```
## [1] 1956.147
```

6. The calculated average causal effect is 0.0813. There was a 8.13 percent point increase in voter turnout from the treatment group compared to the control group.

```r
#Question 6 code

mean(treatment$voted) - mean(control$voted) #difference in means estimator
```

```
## [1] 0.08130991
```

7. For this assumption to hold to make this claim, the estimated causal effect would have to be both internally and externally validated. While the claim are internally validated as this is an experiment, it is not externally validated due to systematic exclusion of population traits. In Gerber et. al's 2008 paper, when they are collecting their data, they systematically exclude voters who live in apartments, voters who live on rural mailing routes, as well as voters who are more likely to vote in the Democratic primary. For external validation, and for our assumption to hold, these voters cannot be systematically excluded because there are many people within the U.S. population to which these excluded traits apply to. Therefore, the assumption that this estimated causal effect is an estimated causal effect for the entire U.S. population cannot hold.

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. American Political Science Review, 102(1), 33–48. https://doi.org/10.1017/S000305540808009X