Hamed Nikookalam

Table of Contents

Problem Set 1

Simulation

Using R, demonstrate that treatment and control groups are comparable when the treatment is randomly assigned. To help get you started, consider this interactive graph: https://ellaudet.github.io/graphs/random_assignment.html

You do not need to create an interactive graph or use the same variables as shown in this example. However, your response should be a simulation that does the following:

- Randomly samples *n* observations from a population with some distribution of traits
- Randomly assigns each observation to the treatment or control group with an equal probability
- Repeats this process many times
- Calculates the proportion of traits for the entire sample, treatment, and control groups for each iteration

Using this simulation, show the following:

- As *n* increases, the distribution of traits in the sample has similar proportions to the distribution in the population.

- As *n* increases, the distribution of traits in the treatment and control groups have similar proportions

You do not need to conduct any statistical tests do demonstrate that the proportions are similar. Simple tables or plots that pass an eyeball test are sufficient.

Solution

Code

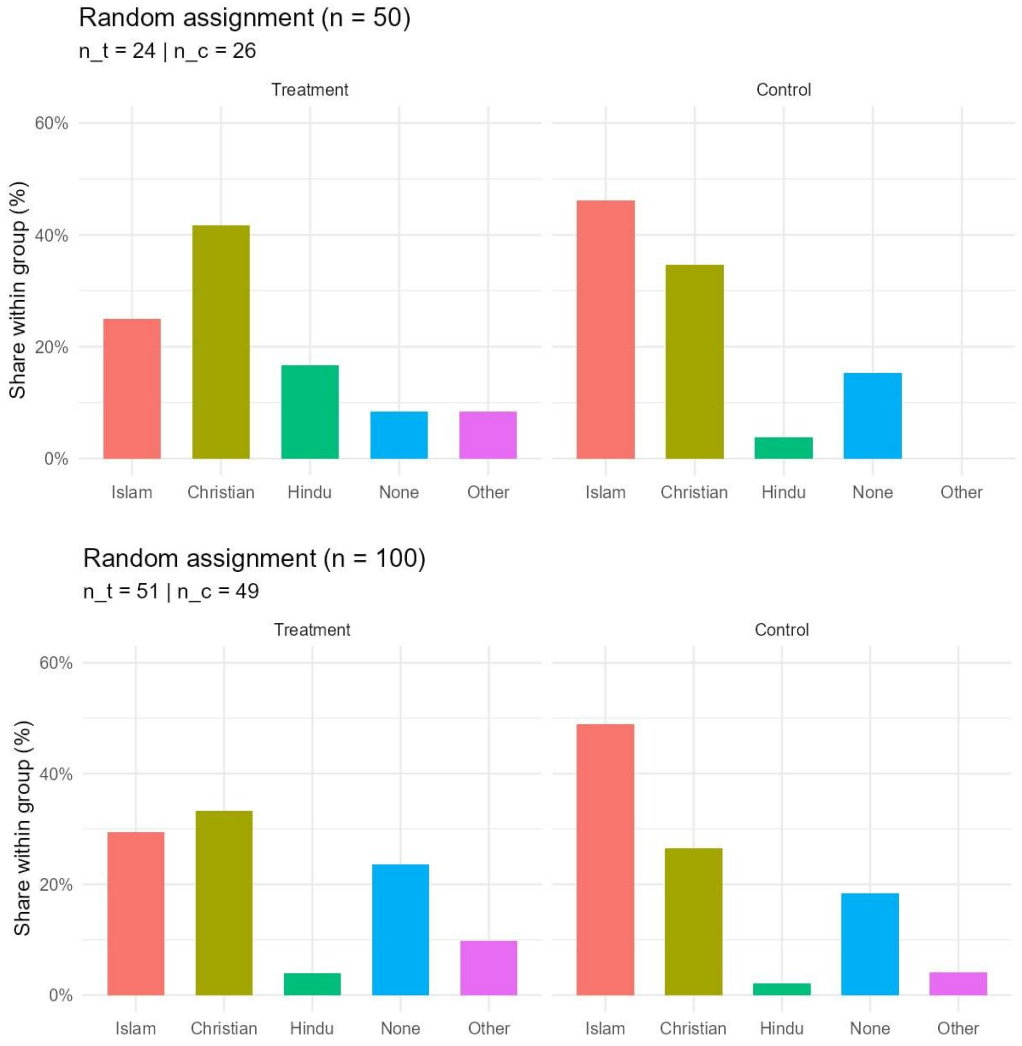
```
# --- Packages ---
library(ggplot2)
library(scales)

# ----- Set up the population once (fixed RNG, labels, shares) -----
set.seed(123)
religions <- c("Islam","Christian","Hindu","None","Other")
p_pop     <- c(0.40, 0.30, 0.10, 0.15, 0.05)
N         <- 1e6
pop       <- sample(religions, N, TRUE, p_pop)

# ----- Make a side-by-side bar plot for one sample size (n) -----
make_plot <- function(n){
  idx   <- sample.int(N, n)
  trait <- factor(pop[idx], levels = religions)
  z     <- rbinom(n, 1, 0.5)
  g     <- factor(ifelse(z==1,"Treatment","Control"),
                  levels = c("Treatment","Control"))

  tab <- prop.table(table(g, trait), margin = 1)   # within-group shares
```

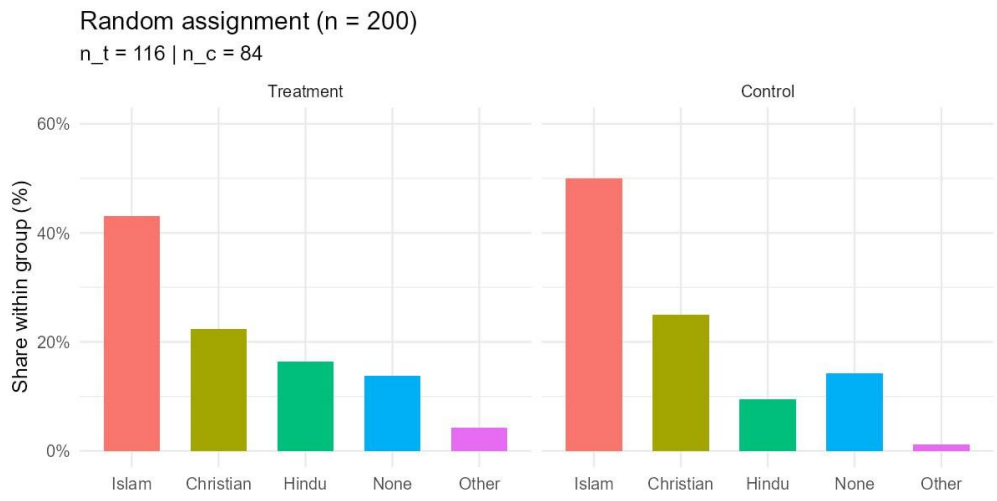
```
  df  <- as.data.frame(tab)
  names(df) <- c("Group","Religion","perc")

  n_by_g <- table(g)
  n_t <- as.integer(n_by_g["Treatment"])
  n_c <- as.integer(n_by_g["Control"])

  ggplot(df, aes(Religion, perc, fill = Religion)) +
    geom_col(width = 0.65) +
    scale_y_continuous(labels = percent_format(accuracy = 1), limits = c(0, 0.6)) +
    facet_wrap(~Group, nrow = 1) +
    labs(
      title    = sprintf("Random assignment (n = %d)", n),
      subtitle = sprintf("n_t = %d | n_c = %d", n_t, n_c),
      x = NULL, y = "Share within group (%)"
    ) +
    theme_minimal(base_size = 12) +
    theme(legend.position = "none")
}

# ----- Save bar plots next to the script -----
n_vals <- c(50, 100, 200, 1000, 5000)
for(n in n_vals){
  p <- make_plot(n)
  print(p)
  ggsave(sprintf("bars_n_%d.png", n), p, width = 8, height = 4, dpi = 150)
}
```

```
# ----- Plot how the T–C difference shrinks as n grows -----
set.seed(202)
n_grid <- c(20, 50, 100, 200, 500, 1000, 2000, 5000)

mean_gap <- function(n){
  idx   <- sample.int(N, n)
  trait <- factor(pop[idx], levels = religions)
  z     <- rbinom(n, 1, 0.5)
  g     <- factor(ifelse(z==1,"Treatment","Control"),
                  levels = c("Treatment","Control"))
  tab <- prop.table(table(g, trait), margin = 1)
  mean(abs(tab["Treatment", ] - tab["Control", ]))
}

err_df <- data.frame(n = n_grid, tc_diff = sapply(n_grid, mean_gap))

p_line <- ggplot(err_df, aes(x = n, y = tc_diff)) +
  geom_line() +
  geom_point() +
  labs(
    title = "T vs C difference by sample size",
    x = "Sample size (n)",
    y = "Mean |T - C| across religions"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none")
```

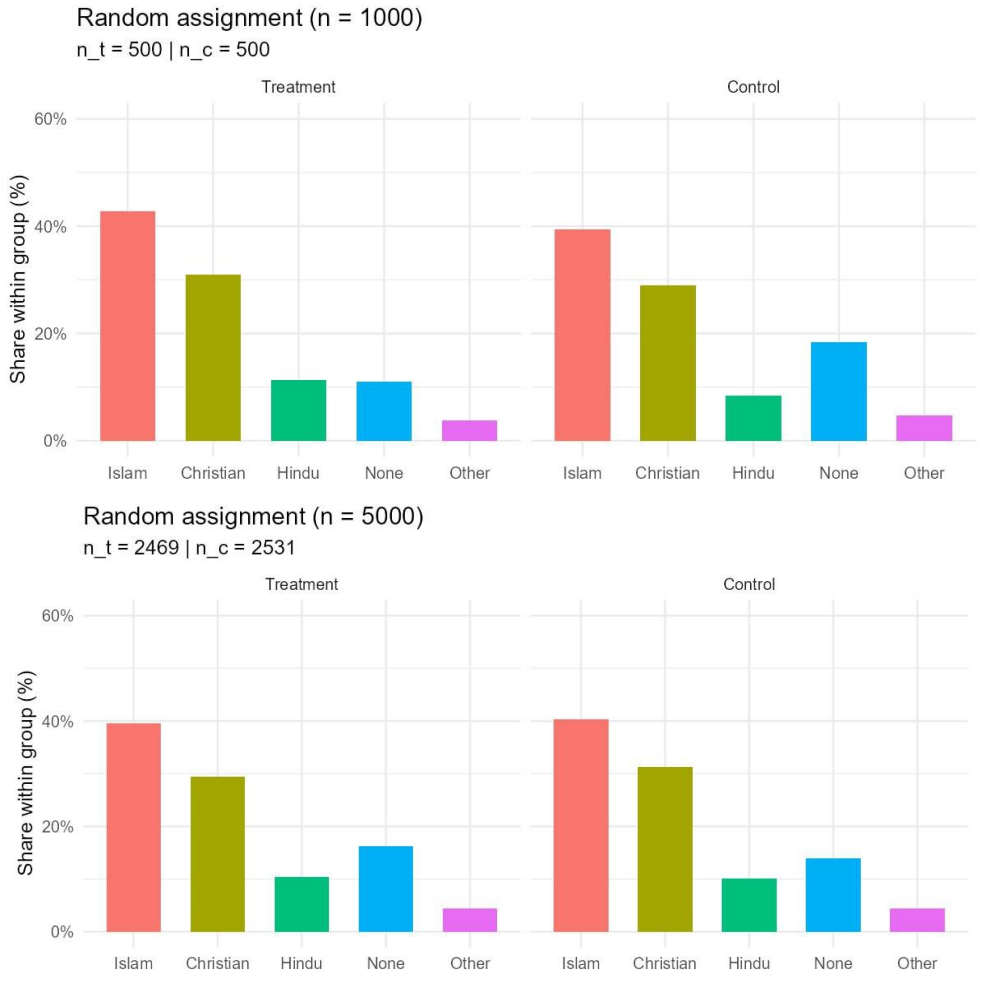
```
print(p_line)

ggsave("treat_control_diff_vs_n.png", p_line, width = 7, height = 4, dpi = 150)
```

Results: The bar charts show that with random 50/50 assignments, Treatment and Control have very similar religion compositions, and the similarity improves with larger n (law of large numbers).

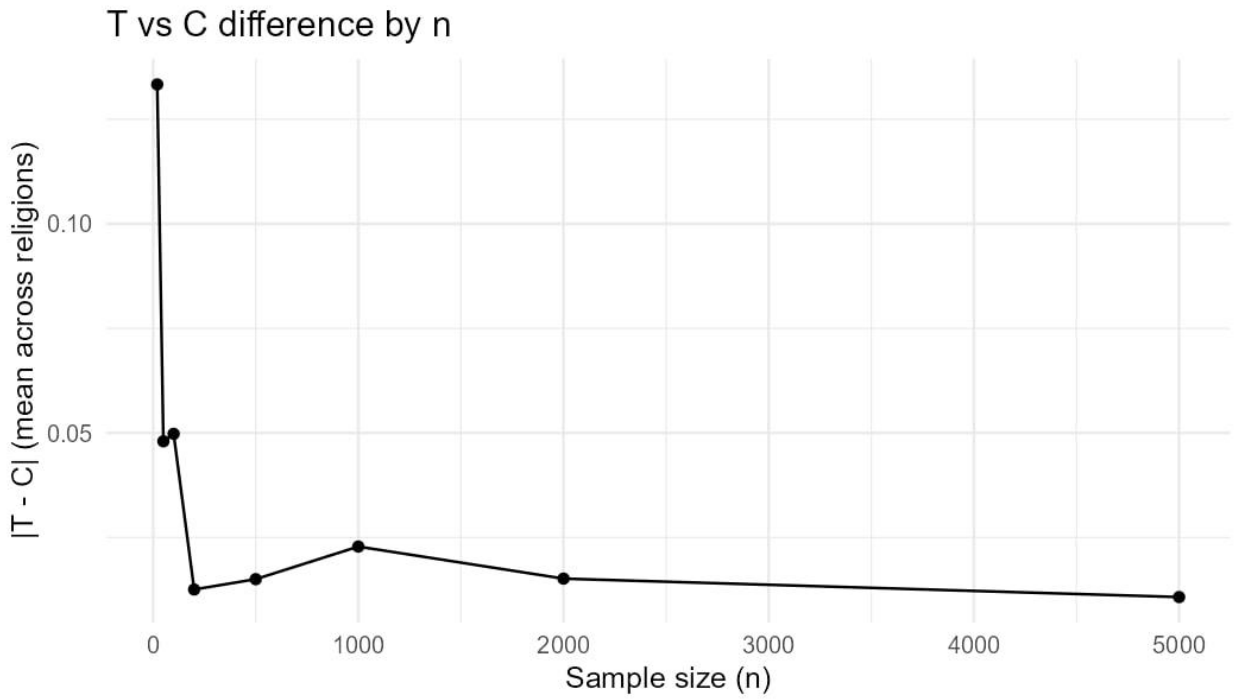




At small n (50–200), random noise creates visible imbalances (e.g., Treatment "Islam" higher than Control at n=50/100/200). These are typical sampling fluctuations, not bias.



By n = 1000 and 5000, the two panels are nearly indistinguishable and both groups track the population shares (≈40% Islam, 30% Christian, 10% Hindu, 15% None, 5% Other).



Here we have the line plot, which I use to show convergence: for each sample size n, I draw a sample, randomly assign treatment, compute the within-group religion shares, take the absolute T–C gaps across the five religions, and average them—this average gap is the y-value. If we repeat this experiment many times, the mean curve across iterations keeps trending down toward zero, with only small jitter from finite-sample noise.

The curve drops quickly and then flattens near zero, showing that as $n$grows, random assignment balances observed traits in expectation. From about $n = 1000$ onward, the average imbalance is only a few percentage points, so Treatment and Control are practically indistinguishable in composition.

Data Analysis

You will be analyzing the data we discussed in class from Gerber et al.'s 2008 paper "Social Pressure and Voter Turnout: Evidence form a Large-Scale Field Experiment." The experiment sent a random selection of voters a message that pressured people to vote by promising to tell their neighbors if they voted in the upcoming election. The dataset is named voting.csv and is in the data folder of the course github repo. There are three variables in the dataset:

birth: Year of birth of registered voter

message: Whether the voter received the social pressure message

voted: Whether the voter voted in the 2006 election

Use this data to complete the following tasks/questions.

Solution

1. What is the treatment variable? Is it a discrete or continuous variable? What is the variable's data type?

> Treatment variable: **Message**
> Discrete or continuous? **Discrete (binary: yes and no).**
> Data type: Stored as **String (character)** in the CSV; analytically a **binary categorical (factor).**

2. Create a new treatment variable in your data frame that is a binary version of the existing treatment variable. Your new variable should equal 1 if the observation was treated, and 0 otherwise.

> Created in the code below (see R script)
> **df$message_binary <- ifelse((df$message) == "yes", 1L, 0L)**

3. Compute the average outcome for the treatment group and the average outcome for the control group. Interpret the results by writing 1-2 sentences about what these numbers mean substantively.

> Treatment: **0.3779 → 37.79% voted**
>
> Control: **0.2966 → 29.66% voted**

On average, the social-pressure mailing raised turnout by about 8.1 percentage points—37.8% voted among those who received it versus 29.7% in the control group.

4. Use brackets to subset the data frame and create two new data frames, one for the treatment group and one for the control group.

Created in the code below (see R script)

```
df_treat  <- df[df$message_binary == 1, ]

df_control <- df[df$message_binary == 0, ]
```

5. What is the average birth year for the treatment and control groups?

Control: 1956.186

Treatment: 1956.147

(These are essentially the same—consistent with balance from randomization.)

6. What is the estimated average causal effect for this experiment? Provide the calculated average effect and a substantive interpretation.

Estimated Average Causal Effect: 0.0813
Receiving the social-pressure mailing increased the probability of voting by about 8 percentage points on average. In practical terms, that's roughly 8 more voters out of every 100 when the message is sent.

7. Suppose we wanted to claim that the estimated causal effect is an estimated effect for the entire U.S. population. What assumption would need to hold for us to make this claim?

This claim needs the sample to be representative of U.S. voters: our study participants and context must reflect the U.S. electorate, and the mailing would be implemented similarly nationwide. If so, the effect we estimate in the sample is the effect for the population.

Code

```
# Read the file

df <- read.csv("C:/Users/hamed/OneDrive/Desktop/Texas A and M_ Courses Fall 2025/602_Quant/PS1/voting.csv")


# Q2. Binary treatment: 1 if message == "yes", else 0

df$message_binary <- ifelse((df$message) == "yes", 1L, 0L)


# Q3. The average outcome for the treatment group and the the control group.

control_mean  <- mean(df$voted[df$message_binary == 0])

treat_mean    <- mean(df$voted[df$message_binary == 1])


# Q4. Creating two new data frames

df_treat  <- df[df$message_binary == 1, ]

df_control <- df[df$message_binary == 0, ]


# Q5. The average birth year for the treatment and control groups

avg_birth_control <- mean(df$birth[df$message_binary == 0])

avg_birth_treat   <- mean(df$birth[df$message_binary == 1])

avg_birth_control; avg_birth_treat


# Q6. The estimated average causal effect

ate <- with(df, mean(voted[message_binary == 1]) - mean(voted[message_binary == 0]))

ate
```