Problem_Set_1

Santiago Vidal

2025-10-23

```r
# Simulation Portion

# Set seed for reproducibility
set.seed(123)

# Define population trait distribution
traits <- c("A", "B", "C")
pop_probs <- c(0.2, 0.5, 0.3)

# Function to simulate one round
simulate_once <- function(n) {
  sample_traits <- sample(traits, n, replace = TRUE, prob = pop_probs)
  group <- sample(c("Treatment", "Control"), n, replace = TRUE)

  sample_prop <- prop.table(table(sample_traits))
  group_prop <- prop.table(table(sample_traits, group), margin = 2)

  list(sample = sample_prop, group = group_prop)
}

# Small sample simulation (n = 30)
sim_30 <- simulate_once(30)
print(sim_30$sample)
```

```
## sample_traits
##         A         B         C
## 0.2666667 0.3666667 0.3666667
```

```r
print(sim_30$group)
```

```
##              group
## sample_traits   Control Treatment
##             A 0.2727273 0.2631579
##             B 0.3636364 0.3684211
##             C 0.3636364 0.3684211
```

```r
# Large sample simulation (n = 1000)
sim_1000 <- simulate_once(1000)
print(sim_1000$sample)
```

```
## sample_traits
##     A     B     C
## 0.194 0.507 0.299
```

```r
print(sim_1000$group)
```

```
##              group
## sample_traits   Control Treatment
##             A 0.2004132 0.1879845
##             B 0.5227273 0.4922481
##             C 0.2768595 0.3197674
```

```r
# Data Analysis Portion

# Load voting.csv from your Desktop R folder
voting <- read.csv("/Users/santividal5/Desktop/R/voting.csv")

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

# The treatment variable is 'message'
# > It shows whether the person received the social pressure message ("yes" or
# >"no")
# > This is a discrete variable (two categories)
# > The data type is a 'character'

class(voting$message)
```

```
## [1] "character"
```

```r
# 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.

# Create a binary version of the treatment variable
# 1 if message == "yes", 0 if message == "no"
voting$treat_binary <- ifelse(voting$message == "yes", 1, 0)

# Check the result
head(voting)
```

```
##   birth message voted treat_binary
## 1  1981      no     0            0
## 2  1959      no     1            0
## 3  1956      no     1            0
## 4  1939     yes     1            1
## 5  1968      no     0            0
## 6  1967      no     0            0
```

```r
# 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.

# Average voting rate (outcome) for treatment group
mean(voting$voted[voting$treat_binary == 1])  # Treated group
```

```
## [1] 0.3779482
```

```r
# Average voting rate for control group
mean(voting$voted[voting$treat_binary == 0])  # Control group
```

```
## [1] 0.2966383
```

```r
# This tells us how many people voted in each group on average.
# > For example, if treated = 0.378 and control = 0.297,
# > then treated voters turned out about 8.1 percentage points more than the
# > control group.
```

```r
# 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.

# Create two new data frames using bracket notation
treatment_group <- voting[voting$treat_binary == 1, ]
control_group <- voting[voting$treat_binary == 0, ]

# Check their sizes
nrow(treatment_group)
```

```
## [1] 38201
```

```r
nrow(control_group)
```

```
## [1] 191243
```

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

# Average birth year of treated voters
mean(treatment_group$birth)
```

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

```r
# Average birth year of control voters
mean(control_group$birth)
```

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

```r
# 6. What is the estimated average causal eLect for this experiment? Provide the
# > calculated average eLect and a substantive interpretation.

# ATE = difference in average outcomes between treatment and control
ate <- mean(treatment_group$voted) - mean(control_group$voted)
ate
```

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

```r
# If ate = 0.08130991, then the message increased voter turnout by about 8.139
# > percentage points.
```

```r
# 7. Suppose we wanted to claim that the estimated causal eLect is an estimated
  # > eLect for the entire U.S. population. What assumption would need to hold
  # > for us to makethis claim?

# To generalize this effect to the U.S. population,
# we need to assume external validity:
# The sample and treatment effect must be representative of the broader
# > population.
```

```
# > The assumption is that the experimental sample and context are similar
# > enough to the general U.S. voting population that the same treatment
# > effect would apply.
```