Problem Set 1

Ellie Choe

2025-10-11

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
set.seed(123)
```

—- Simulation —-

```
# Population
ages <- c("10s", "20s", "30s", "40s", "50s")
p_pop <- c(0.15, 0.30, 0.25, 0.20, 0.10)
names(p_pop) <- ages
```

```
# Sample Sizes
n_vals <- c(50, 100, 500, 1000)

# Create an Empty Data Frame
res <- data.frame(
  n = integer(),
  group = character(),
  age = character(),
  prop = numeric()
)

# Simulation Loop
for (n in n_vals) {
```

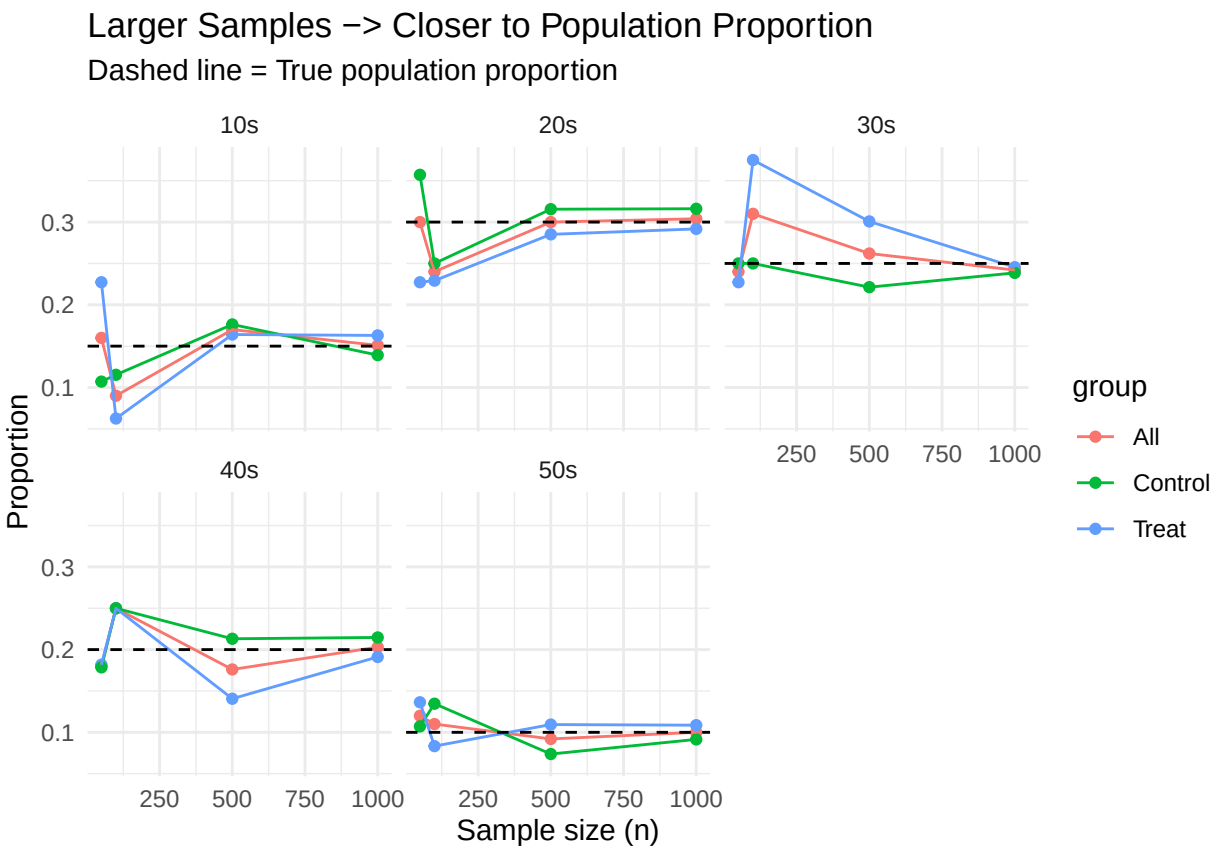
```r
  trait <- sample(ages, size = n, replace = TRUE, prob = p_pop) # Draw a random sample
  Z <- rbinom(n, 1, 0.5) # Randomly assign each obervation to Treatment (1) or Control (0)
  prop_all <- as.numeric(table(factor(trait, levels = ages))) / n # Calculate proportions for the entir
  prop_treat <- as.numeric(table(factor(trait[Z == 1], levels = ages))) / sum(Z == 1) # Calculate propo
  prop_control <- as.numeric(table(factor(trait[Z == 0], levels = ages))) / sum(Z == 0) # Calculate pro
res <- rbind(res, data.frame(n = n, group = "All", age = ages, prop = prop_all), data.frame(n = n, grou
}

head(res)
```

```
##    n group age      prop
## 1 50   All 10s 0.1600000
## 2 50   All 20s 0.3000000
## 3 50   All 30s 0.2400000
## 4 50   All 40s 0.1800000
## 5 50   All 50s 0.1200000
## 6 50 Treat 10s 0.2272727
```

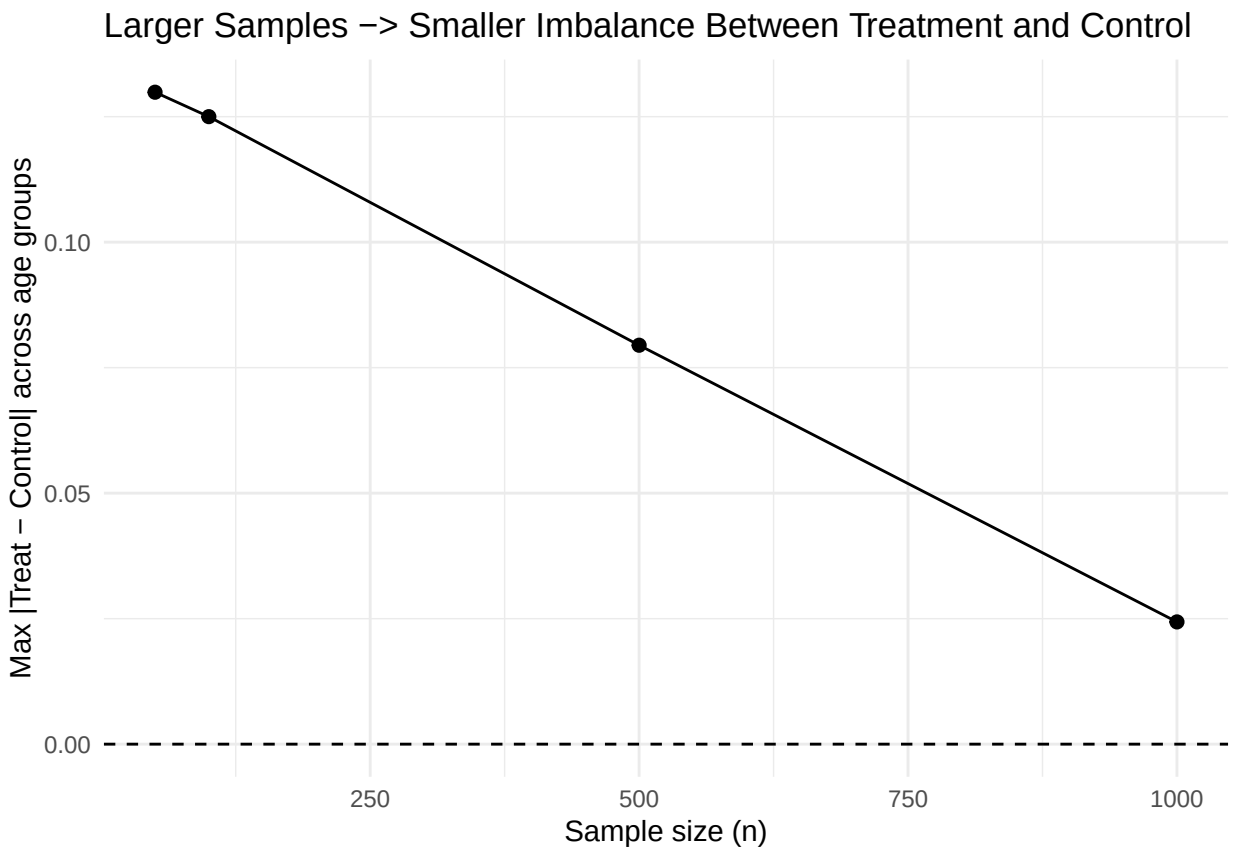
```r
# Convergence to the Population Distribution
pop_tbl <- data.frame(age = ages, pop_prop = p_pop)

ggplot(res, aes(x = n, y = prop, color = group)) + geom_point() + geom_line() + geom_hline(data = pop_t
```




```r
# Imbalance Between Treatment and Control
imbalance <- res %>% filter(group %in% c("Treat", "Control")) %>% pivot_wider(names_from = group, value
```

```
ggplot(imbalance, aes(x = n, y = max_abs_diff)) + geom_hline(yintercept = 0, linetype = 2) + geom_point
```



—- data analysis —-

```
df <- read.csv("voting.csv")
```

1. Treatment variable

variable: message type: discrete data type: character

2. create a binary variable

```
df$treat <- ifelse(df$message == "yes", 1, 0)
```

3. Compute the average outcome for the treatement and control groups

```
avg_treat <- mean(df$voted[df$treat == 1]) # average voting rate among treated voters
avg_ctrl <- mean(df$voted[df$treat == 0]) # average voting rate among control voters. If avg_treat > av
```

4. Subset the data frame

```
treat_df <- df[df$treat == 1, ]
ctrl_df <- df[df$treat == 0, ]
```

5. Average birth year

```
mean(treat_df$birth)
```

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

```
mean(ctrl_df$birth)
```

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

6. Estimated Average Causal Effect

```
ate <- avg_treat - avg_ctrl
ate # it means treated voters were 8.1 percentage points more likely to vote.
```

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

7. Assumption for generalization

The sample must be representative of the US population, and the treatment effect must be homogeneous across subgroups.