Problem Set 1

Neeraj Tom Savio

2025-10-13

```
## Introductory Code and Setting Up

rm(list = ls()) #Clear the environment

setwd("~/Desktop/POLS 602/R Files") #Setting Working Directory

#Load tidyverse as we need dplyr for binding rows and ggplot 2 for creating..
#..plots.
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     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
```

```
set.seed(3300) #For reproductibility
```

Simulation

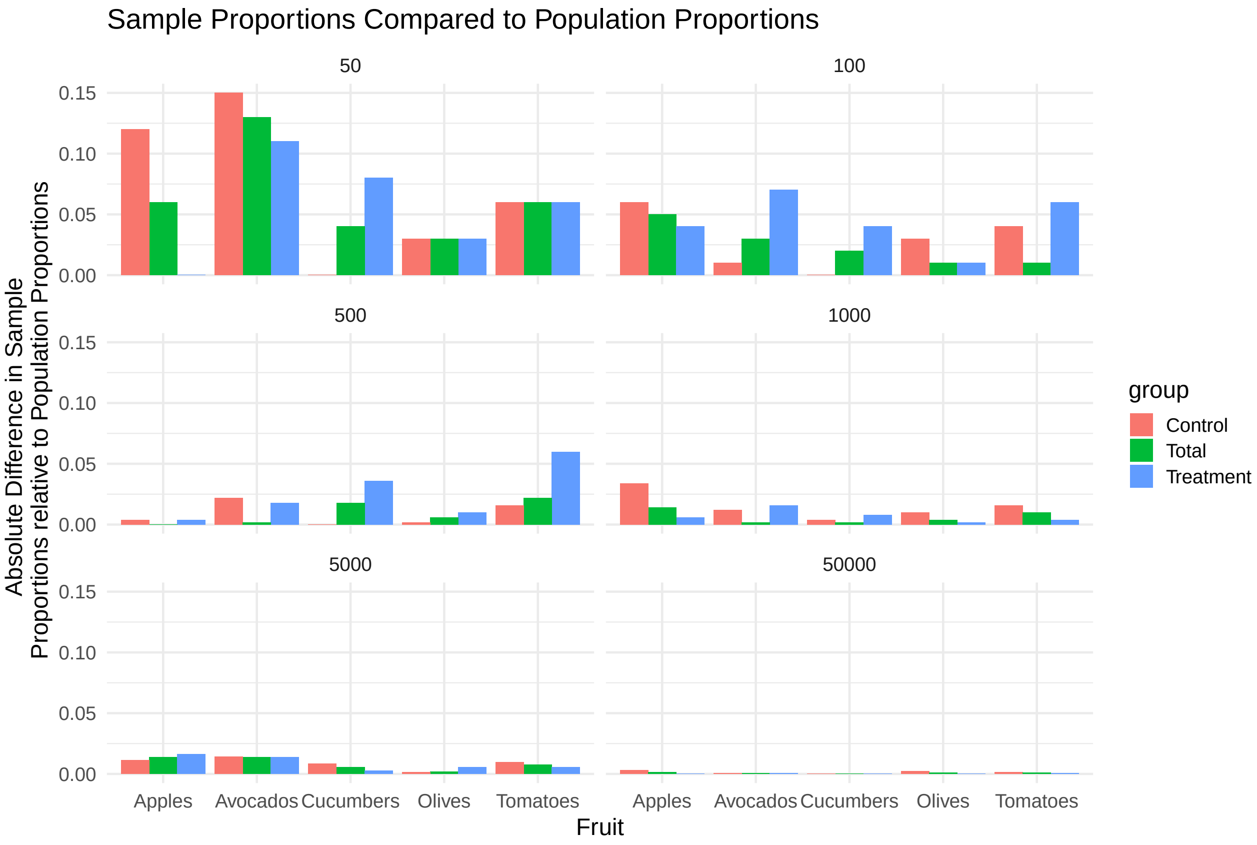
```
## Creating an empty container to hold our results for different sample sizes
res <- data.frame(n = integer(),
                  group = character(),
                  fruit = character(), prop = numeric())

## Designating fruit names
fruit_names <- c(
  "Apples", "Tomatoes", "Olives", "Avocados", "Cucumbers")
#Yes, they are all fruits

## Designating Sample Sizes
sample_sizes <- c(50, 100, 500, 1000, 5000, 50000)

## Designating Population Proportions
pop_prop <- c(0.2, 0.3, 0.05, 0.25, 0.2)
```

```r
## Creating a FOR loop to reiterate the code over each sample size
for (i in sample_sizes){
## Creating population
fruits <- sample(fruit_names,
                 size = i, replace = TRUE,
                 prob = pop_prop)

## Creating a randomly sampled index of treatment observations from the sample
tre_index <- sample(seq_along(fruits), size = i %/% 2, replace = FALSE)

tre <- fruits[tre_index] #Subsetting from the sample  using the treatment index

con <- fruits[-tre_index] #Creating a control subset using the treatment index

## Finding out the proportions of each group
prop_total <- prop.table(table(factor(fruits, levels = fruit_names)))
prop_con <- prop.table(table(factor(con, levels = fruit_names)))
prop_tre <- prop.table(table(factor(tre, levels = fruit_names)))

## Joining these proportions to the original empty result table
res <- bind_rows(
    res,
    data.frame(n = i, group = "Total",
               fruit = names(prop_total), prop = as.numeric(prop_total)),
    data.frame(n = i, group = "Control",
               fruit = names(prop_con), prop = as.numeric(prop_con)),
    data.frame(n = i, group = "Treatment",
               fruit = names(prop_tre), prop = as.numeric(prop_tre)),
  )
}

## Calculating the Difference in Sample Proportions to Population Proportions

#Creating a table containing population proportions
pop_df <- data.frame(fruit = fruit_names,
                     population_proportions = pop_prop)

#Merging the population proportions table into our original results table
res <- merge(res, pop_df, by = "fruit")

#Calculating the absolute difference.
#We are interested in the magnitude of the difference, not in its direction.
res$diff_prop <- abs(res$population_proportions - res$prop)
```

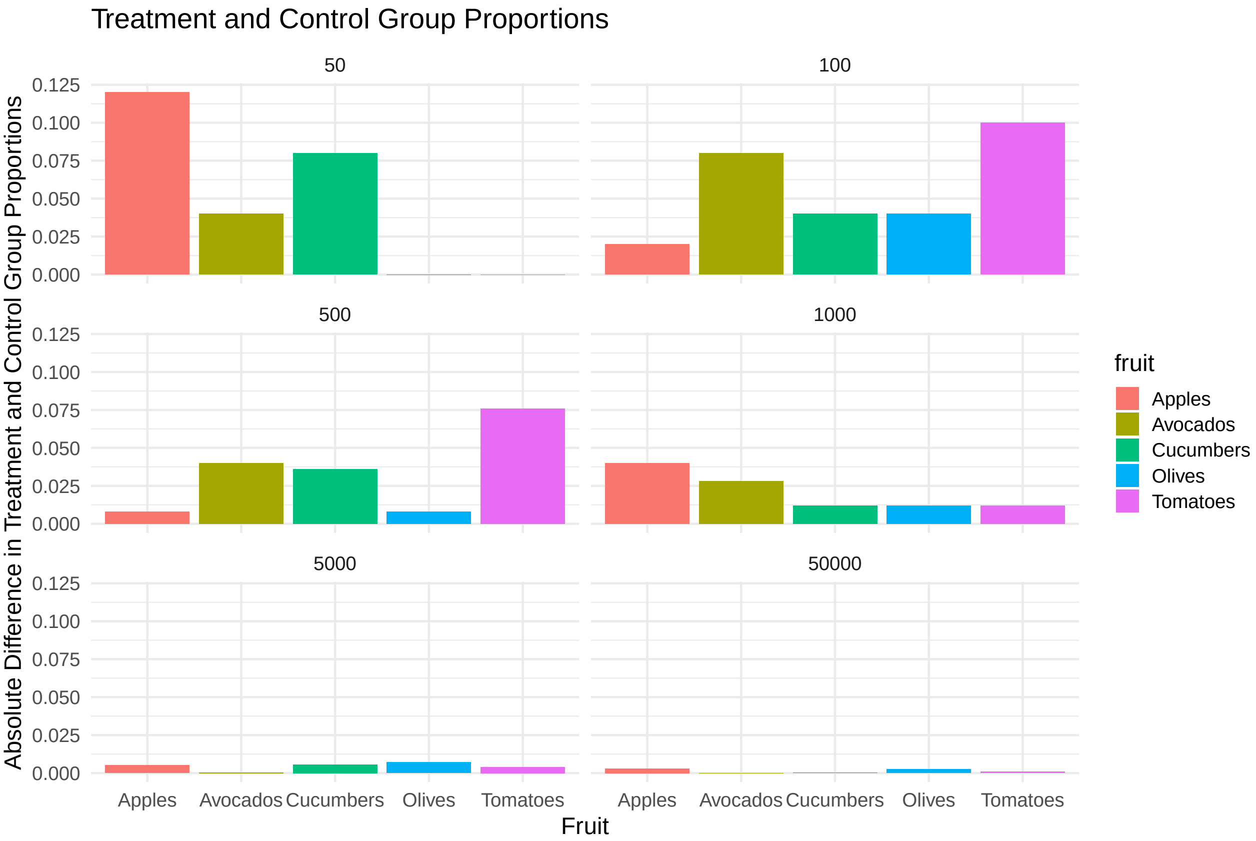
We also need to calculate the difference in proportions between the control and treatment groups to shows that they become more equal in all observed and unobserved characteristics as the sample size grows.

```r
## Creating a new dataframe for only treatment and control group proportions.
res1 <- res[res$group %in% c("Control", "Treatment"),
            c("n", "fruit", "group", "prop")]

#The control and treatment proportions ned to be separated into 2 separate columns.
#This will help us compute the absolute difference between the two
res1 <- reshape(res1, timevar = "group",
                idvar = c("n" , "fruit"), direction = "wide")

#Calculating the absolute difference in proportions
res1$prop_diff <- abs(res1$prop.Treatment - res1$prop.Control)
```

```
## Creating Plots

# Visualizing the difference between Sample and Population Proportions
ggplot(res, aes(x=fruit, y = diff_prop, fill = group)) +
  geom_col(position = "dodge") +
  facet_wrap(~ n, nrow = 3) +
  labs(
    title = "Sample Proportions Compared to Population Proportions",
    x= "Fruit",
    y = "Absolute Difference in Sample
    Proportions relative to Population Proportions"
  ) +
  theme_minimal(base_size = 16)
```



```
# Visualizing the difference between Treatment and Control Group Proportions
ggplot(res1, aes(x=fruit, y=prop_diff, fill = fruit)) +
  geom_col(position = "dodge") +
  facet_wrap(~ n, nrow = 3) +
  labs(
    title = "Treatment and Control Group Proportions",
    x = "Fruit",
    y = "Absolute Difference in Treatment and Control Group Proportions" ) +
  theme_minimal(base_size = 16)
```



It is evident from the first plot that the sample proportions get closer to the population proportions in each group(treatment, control, and total) as the sample size increases.

The second plot shows that the treatment and control groups become more equal in terms of proportion of characteristics as the sample size grows.

Data Analysis

```
df <- read.csv("Datasets/voting.csv") #Importing dataframe
```

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

```
## Figure out the data structure and type.
str(df$message)
```

```
##  chr [1:229444] "no" "no" "no" "yes" "no" "no" "no" "no" "no" "no" "no" ...
```

```
class(df$message)
```

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

According to Gerber et al.'s 2008 paper, the treatment variable is whether the individual received the message or not. It is a discrete variable. The data type is characters.

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.

```
## Recoding DV as 1 and 0s.
df$tc <- ifelse(df$message == "yes", 1, 0) # Creating a new treatment column
```

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.

```
## Finding average outcome for treatment and control groups
avg_con <- mean(df$voted[df$tc == 0])
avg_tre <- mean(df$voted[df$tc == 1])

avg_outcome <- data.frame(
  Group = c("Control", "Treatment"),
  Average_Outcome = c(avg_con, avg_tre)
)

avg_outcome
```

```
##       Group Average_Outcome
## 1   Control       0.2966383
## 2 Treatment       0.3779482
```

The average outcome for the control group shows that approximately 30% of the control group voted (i.e., those who didn't receive a message) whereas the average outcome for the treatment group (i.e., those who did receive a message prior to voting) shows that approximately 38% of the treatment group voted.

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.

```
#Creating separate data frames
df_con <- df[df$tc == 0, ] #Control group dataframe
df_tre <- df[df$tc == 1, ] #Treatment group dataframe
```

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

```
m_con <- mean(df_con$birth)
m_tre <- mean(df_tre$birth)

#Putting results in a table for better presentation
avg_year <- data.frame(
  Group = c("Control", "Treatment"),
  Average_Birth_Year = c(m_con, m_tre)
)

avg_year
```

```
##       Group Average_Birth_Year
## 1   Control           1956.186
## 2 Treatment           1956.147
```

The average birth year for both treatment and control groups is 1956. This demonstrates that randomization yields groups that are approximately equal in observed and unobserved characteristics.

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

$$Average\widehat{Causal}Effect = \bar{Y}_{treatment} - \bar{Y}_{control}$$

```
avg_causal_effect <- avg_tre - avg_con
avg_causal_effect
```

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

The average causal effect is 0.08130991, which means that there is a difference in voting of approximately 8 percentage points between the treatment and control groups. In other words, in the sample group, those who received the message(treatment group) were on average 8 percentage points more likely to vote than those who didn't receive the message(control group).

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?

The Gerber et al. 2008 paper used data sampled from 180002 households in Michigan. If we were to claim that that the estimated causal effect could be generalized to the entire U.S. population, we would have to assume that individuals across the United States and across different type of living situations were sampled.i.e., there should be no systematic differences between the US population and those who were sampled.