After feedback review

MY457: Problem Set One

Peter Tiborcz

Sun/02/Feb

1 Concepts

Consider a simple study of the effect of a treatment:

$$D_i \in \{0,1\} \text{ on } Y_i \text{ for all } i \in \{1,2,3,\ldots,N\}$$

1.1

Y_{i1} represents the potential outcome for unit if they receive treatment (D=1). This situation of Y_i having the same quantity as Y_{i1} is only possible when the unit i is treated.

1.2

The difference between $E[Y_{i0} \mid D_i = 1]$ and $E[Y_{i0} \mid D_i = 0]$ is that the first one cannot be observed, while the second one we can observe.

I would expect these quantities to be equal when we randomly select the units to be treated. This is because the treatment is randomly assigned, so the potential outcomes are also randomly assigned.

We would expect these to be unequal when the treatment is not randomly assigned. This is because the potential outcomes are not randomly assigned, so the treatment is not randomly assigned.

1.3

Random assignment ensures that the treatment and control groups are statistically identical on average before treatment. This allows us to estimate the Average Treatment Effect (ATE).

Since randomization removes systematic differences between groups, any difference in outcomes can be attributed to the treatment itself rather than confounding factors.

1.2 Ideally we want to create an experiment where we know that participants were randomly assigned to treatment and control. We would like this to be the case because this way we can avoide introdu ing selection bias which if we unwaere of them can ellude our assensment of the effects of the treatment. Since we cant observe Yi0 D=1 because of the FPCI we need to infer from out control group to the potential outcoems of our treatment. if we cant do thsi reeliable due to potential selection bias then we cannot be sure what we measure

1.3 is kinda similar, which is that if we avoided introducing selection bias then we can identify the ATE by using the control group as the casuq best appricomator for the missing pontential outcomes ofthe treatment group

2 Simulations

2.1

The following R code generates a dataset with 500 observations, a treatment effect of 5000, and several characteristics:

```
##         Age Education       Y0       Y1 D        Y
## 1  36.39524         1 65384.30 70384.30 1 70384.30
## 2  39.69823         4 48902.90 53902.90 1 53902.90
## 3  57.58708         2 55114.71 60114.71 1 60114.71
## 4  42.70508         3 52139.58 57139.58 0 52139.58
## 5  43.29288         2 48138.79 53138.79 1 53138.79
## 6  59.15065         3 48796.06 53796.06 1 53796.06
## 7  46.60916         1 60128.34 65128.34 0 60128.34
## 8  29.34939         4 47985.42 52985.42 0 47985.42
## 9  35.13147         2 29623.18 34623.18 0 29623.18
## 10 37.54338         3 48041.11 53041.11 0 48041.11
```

echo = True to show code

- Age is generated from a normal distribution with a mean of 42 and a standard deviation of 10.
- Education is randomly assigned one of four categories.
- Y0 represents the baseline outcome (without treatment), drawn from a normal distribution with a mean of 50,000.
- Y1 is the potential outcome if treated, which is Y0 + tau.
- D is the treatment indicator, randomly assigned.
- Y is the observed outcome, which depends on D.

Because we simulated the data we can also simulate both potential outcomes for each observation which are the Y1 and Y0 variables. We set the seed to 123 to ensure reproducibility.

2.2

To ensure that randomization successfully balances covariates, we conduct t-tests comparing Age and Education across treatment and control groups.

```
## Effect sizes were labelled following Cohen's (1988) recommendations.
##
## The Welch Two Sample t-test testing the difference of Age by D (mean in group 0
## = 41.78, mean in group 1 = 42.98) suggests that the effect is negative,
## statistically not significant, and very small (difference = -1.20, 95% CI
## [-2.92, 0.51], t(485.80) = -1.38, p = 0.169; Cohen's d = -0.13, 95% CI [-0.30,
## 0.05])
```

```
## Effect sizes were labelled following Cohen's (1988) recommendations.
##
## The Welch Two Sample t-test testing the difference of Education by D (mean in
## group 0 = 2.48, mean in group 1 = 2.53) suggests that the effect is negative,
## statistically not significant, and very small (difference = -0.05, 95% CI
## [-0.25, 0.14], t(492.69) = -0.54, p = 0.593; Cohen's d = -0.05, 95% CI [-0.22,
## 0.13])
```

Both Age and Education are balanced across treatment (D=1) and control (D=0) groups, as there are no statistically significant differences in means. This supports the randomization assumption that the two groups are comparable.

Good answer but rewrite the output of the report pavkage or reference the cited lit

2.3 - the answer should be the difference in means aka mean(y1)-mean(y0)

2.3

To figure out the difference in means between the potential outcomes Y_0 and Y_1 we can use the following code:

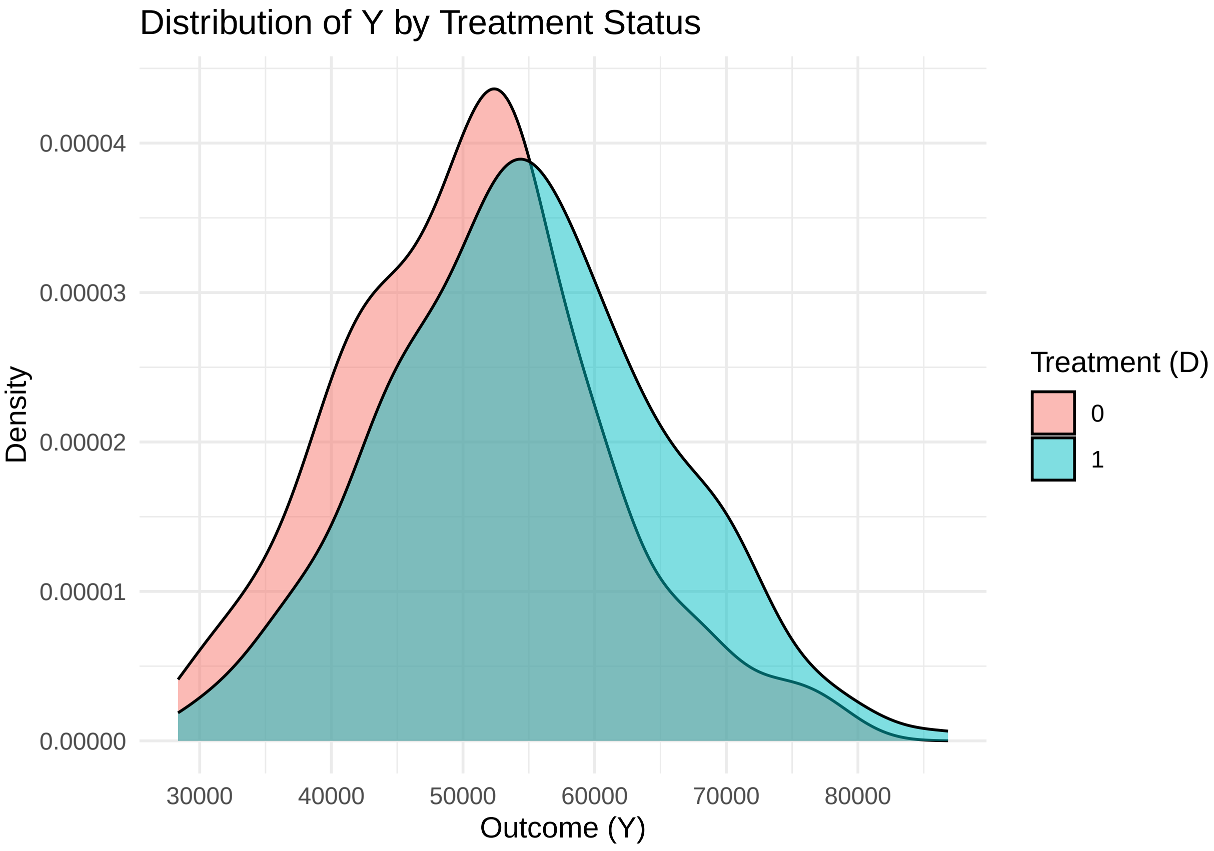
```
## [1] 4181.018
```

Given that the treatment effect (tau) was set to 5000, but the estimated effect is 4181.018, the estimate is lower than the true effect. The discrepancy may be due to randomization variability or sampling variability.

```
## Effect sizes were labelled following Cohen's (1988) recommendations.
##
## The Welch Two Sample t-test testing the difference of Y by D (mean in group 0 =
## 50612.03, mean in group 1 = 54793.05) suggests that the effect is negative,
## statistically significant, and small (difference = -4181.02, 95% CI [-5974.54,
## -2387.50], t(484.78) = -4.58, p < .001; Cohen's d = -0.42, 95% CI [-0.60,
## -0.24])
```

2.4

We visualize the outcome distribution across treatment groups:



We then estimate the ATE using two approaches:

```
## [1] 4181.018
```

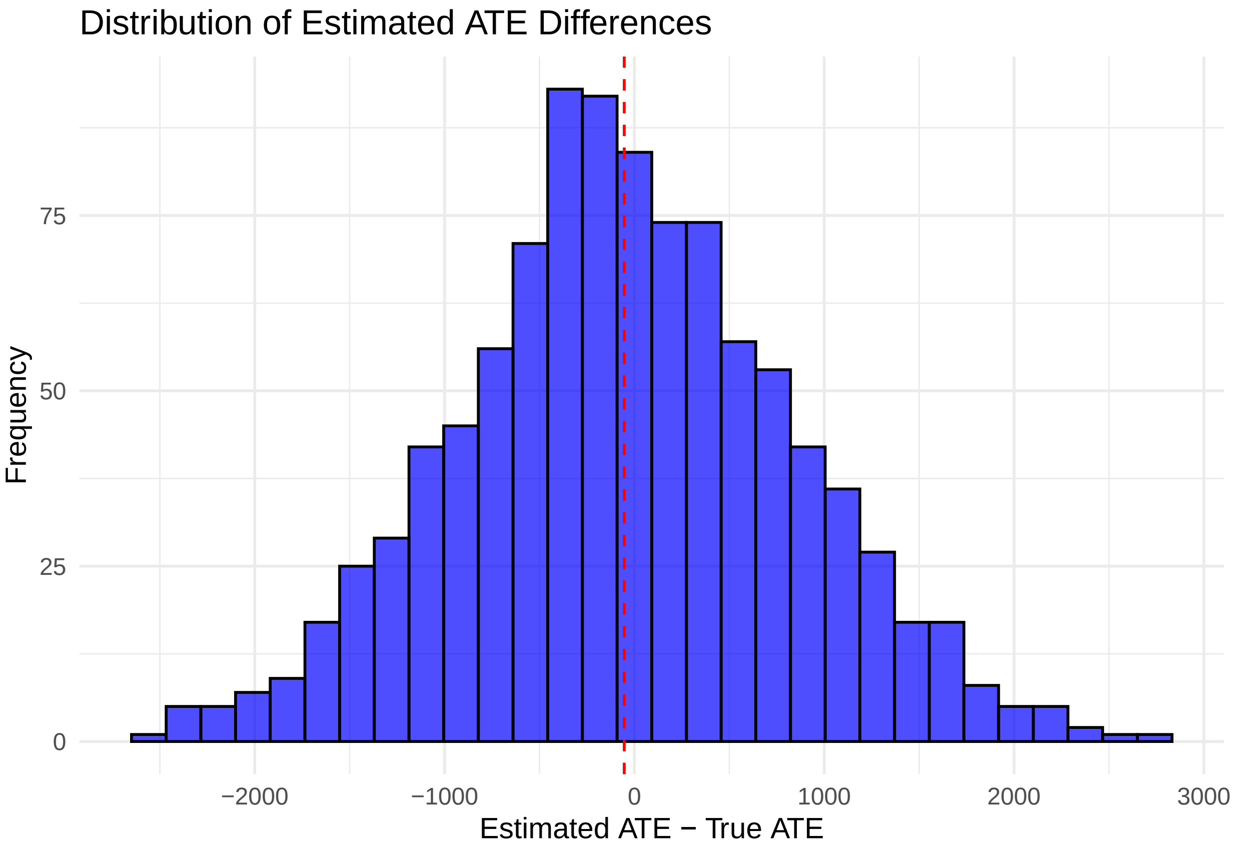
```
##
## Call:
## lm(formula = Y ~ D, data = data)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -25286  -7123    -52   6148  32047
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  50612.0      625.2  80.947 < 0.0000000000000002 ***
## D             4181.0      910.1   4.594           0.00000551 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10160 on 498 degrees of freedom
## Multiple R-squared:  0.04066,    Adjusted R-squared:  0.03873
## F-statistic: 21.11 on 1 and 498 DF,  p-value: 0.000005514
```

Difference-in-means and OLS regression should produce similar estimates if randomization is successful.

2.5

We run 1,000 simulations to verify that our ATE estimate is not due to chance:



After simulating 1000 data generations, the estimated ATE averages really close to 0 suggesting an unbiased estimation.

3 Replication

3.1

We begin by loading the dataset and creating a new binary variable sd_onefem2014, which takes the value 1 if at least one woman was elected in 2014, and 0 otherwise:

```
##  [1] "unique_id"                "random"
##  [3] "condition"                "county"
##  [5] "age2014"                  "attendees2014"
##  [7] "sd2014"                   "prop_fem_attend2014"
##  [9] "prop_sd_fem2014"          "receiveletter"
## [11] "content_recruit"          "content_para"
## [13] "recruit_num"              "para_read"
## [15] "women_more"               "women_ideas"
## [17] "women_menbetter"          "women_toofar"
## [19] "gender"                   "yearborn"
## [21] "religion"                 "race"
## [23] "income"                   "age"
## [25] "agecat"                   "chair_survey"
## [27] "distance"                 "distance_100"
## [29] "age2012"                  "attendees2012"
## [31] "sd2012"                   "prop_fem_attend2012"
## [33] "prop_sd_fem2012"          "pcfemale_2014"
## [35] "letter"                   "Total_Running"
## [37] "Total_Female_Running"     "Total_Gave_Speach_Female"
## [39] "observed"                 "sd_onefem2014"
```

This transformation allows us to assess how treatment conditions impact the likelihood of electing at least one woman. We can printed out the column names to ensure that the new variable was created.

3.2

To verify the distribution of precincts across conditions, we compute the count and proportion of precincts in each group:

```
## # A tibble: 5 x 4
##   condition     precinct_count sum_fem2014 proportion_fem2014
##   <fct>                  <int>       <dbl>              <dbl>
## 1 Control                  541      111.              0.233
## 2 Supply                   539      123.              0.256
## 3 Demand                   538      116.              0.243
## 4 Supply+Demand            538      126.              0.265
## 5 <NA>                      11        1.17            0.00244
```

The summary table confirms whether randomization achieved balanced group sizes and provides insight into how treatment affected women's representation.

3.3

To check if treatment groups were balanced, we conduct pairwise t-tests on two pre-treatment variables: agecat (age category) and race.

```
##
## T-test for agecat between Control and Demand
##
##  Welch Two Sample t-test
```

```
##
## data:  agecat by condition
## t = -0.66307, df = 260.56, p-value = 0.5079
## alternative hypothesis: true difference in means between group Control and group Demand is not equal
## 95 percent confidence interval:
##  -0.3690888  0.1831356
## sample estimates:
## mean in group Control  mean in group Demand
##              4.646154              4.739130
##
##
## T-test for agecat between Control and Supply
##
##  Welch Two Sample t-test
##
## data:  agecat by condition
## t = -1.0271, df = 270.81, p-value = 0.3053
## alternative hypothesis: true difference in means between group Control and group Supply is not equal
## 95 percent confidence interval:
##  -0.4327653  0.1360319
## sample estimates:
## mean in group Control  mean in group Supply
##              4.646154              4.794521
##
##
## T-test for agecat between Control and Supply+Demand
##
##  Welch Two Sample t-test
##
## data:  agecat by condition
## t = -0.91979, df = 265.63, p-value = 0.3585
## alternative hypothesis: true difference in means between group Control and group Supply+Demand is no
## 95 percent confidence interval:
##  -0.4158762  0.1510410
## sample estimates:
##       mean in group Control mean in group Supply+Demand
##                    4.646154                    4.778571
##
##
## T-test for race between Control and Demand
##
##  Welch Two Sample t-test
##
## data:  race by condition
## t = -0.90273, df = 272.67, p-value = 0.3675
## alternative hypothesis: true difference in means between group Control and group Demand is not equal
## 95 percent confidence interval:
##  -0.13880211  0.05152823
## sample estimates:
## mean in group Control  mean in group Demand
##              4.978261              5.021898
##
##
## T-test for race between Control and Supply
```

```
##
##  Welch Two Sample t-test
##
## data:  race by condition
## t = -1.0296, df = 282.57, p-value = 0.3041
## alternative hypothesis: true difference in means between group Control and group Supply is not equal
## 95 percent confidence interval:
##  -0.16036078  0.05021585
## sample estimates:
## mean in group Control  mean in group Supply
##              4.978261              5.033333
##
##
## T-test for race between Control and Supply+Demand
##
##  Welch Two Sample t-test
##
## data:  race by condition
## t = -0.30434, df = 281.5, p-value = 0.7611
## alternative hypothesis: true difference in means between group Control and group Supply+Demand is no
## 95 percent confidence interval:
##  -0.11119576  0.08141613
## sample estimates:
##       mean in group Control mean in group Supply+Demand
##                    4.978261                    4.993151
```

No t test is significant, so we can say that the randomization was successful.

3.4

To measure the effect of each intervention, we estimate the ATE using separate regressions:

```
##
## ATE Estimate for Demand vs. Control:
##
## Call:
## lm(formula = sd_onefem2014 ~ treatment, data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4327 -0.4327 -0.3753  0.5673  0.6247
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.37528    0.02304  16.289 <0.0000000000000002 ***
## treatment    0.05746    0.03271   1.757              0.0793 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4904 on 897 degrees of freedom
##   (180 observations deleted due to missingness)
## Multiple R-squared:  0.003428,   Adjusted R-squared:  0.002317
## F-statistic: 3.086 on 1 and 897 DF,  p-value: 0.07932
##
##
```

```
## ATE Estimate for Supply vs. Control:
##
## Call:
## lm(formula = sd_onefem2014 ~ treatment, data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4340 -0.4340 -0.3753  0.5660  0.6247
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.37528    0.02305  16.281 <0.0000000000000002 ***
## treatment    0.05877    0.03230   1.819              0.0692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4906 on 921 degrees of freedom
##   (157 observations deleted due to missingness)
## Multiple R-squared:  0.003581,   Adjusted R-squared:  0.002499
## F-statistic:  3.31 on 1 and 921 DF,  p-value: 0.06918
##
##
## ATE Estimate for Supply+Demand vs. Control:
##
## Call:
## lm(formula = sd_onefem2014 ~ treatment, data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4537 -0.4537 -0.3753  0.5463  0.6247
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.37528    0.02309  16.249 <0.0000000000000002 ***
## treatment    0.07845    0.03284   2.388              0.0171 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4915 on 894 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.006341,   Adjusted R-squared:  0.005229
## F-statistic: 5.705 on 1 and 894 DF,  p-value: 0.01713
```

This method isolates the effect of each intervention. The results show that each treatment increased the likelihood of elect

3.5

Instead of separate regressions, we now estimate all treatment effects simultaneously using a single regression model:

```
##
## Call:
## lm(formula = sd_onefem2014 ~ treatment_demand + treatment_supply +
##     treatment_supply_demand, data = electwomen_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4537 -0.4340 -0.3753  0.5660  0.6247
##
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              0.37528    0.02320  16.174 <0.0000000000000002 ***
## treatment_demand         0.05746    0.03294   1.744              0.0813 .
## treatment_supply         0.05877    0.03252   1.807              0.0709 .
## treatment_supply_demand  0.07845    0.03300   2.377              0.0175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4938 on 1808 degrees of freedom
##   (355 observations deleted due to missingness)
## Multiple R-squared:  0.003499,   Adjusted R-squared:  0.001845
## F-statistic: 2.116 on 3 and 1808 DF,  p-value: 0.09624
```

This model allows us to compare the effects of all treatments while accounting for potential shared variance across groups. After simultaneously estimating the ATE of the different treatments we found that that the coefficients of each individual treatment are the exactly the same as when we estimated them individually. The intercept is also the same. Similarly we found that the t test also resulted in the same results as before. Doing them simultaneously did not resulst any changes in the results.

4 Code appendix

```r
# this chunk contains code that sets global options for the entire .Rmd.
# we use include=FALSE to suppress it from the top of the document, but it will still appear in the app
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, linewidth=60)

# you can include your libraries here:
library(tidyverse)
#install.packages("report")
library(report)

# and any other options in R:
options(scipen=999)

set.seed(123)
n <- 500
tau <- 5000

data <- data.frame(
Age = rnorm(n, mean = 42, sd = 10),
Education = sample(1:4, n, replace = TRUE),
Y0 = rnorm(n, mean = 50000, sd = 10000)
)
data <- data %>% mutate(
Y1 = Y0 + tau,
D = sample(c(0, 1), n, replace = TRUE),
Y = ifelse(D== 1, Y1, Y0)
)
```

```r
head(data, 10)
# Print results using report package
report::report(t.test(Age ~ D, data = data))
report::report(t.test(Education ~ D, data = data))
# Calculate the difference in means between Y1 and Y0
diffINmeans <- mean(data$Y[data$D==1]) - mean(data$Y[data$D==0])
diffINmeans
report::report(t.test(Y ~ D, data = data))
# Plot the distribution of Y conditional on D
ggplot(data, aes(x = Y, fill = as.factor(D))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Y by Treatment Status",
       x = "Outcome (Y)",
       y = "Density",
       fill = "Treatment (D)") +
  theme_minimal()
diffInMeans <- mean(data$Y[data$D == 1]) - mean(data$Y[data$D == 0])
print(diffInMeans)

lm_model <- lm(Y ~ D, data = data)
summary(lm_model)

# Set parameters
n_simulations <- 1000
tau <- 5000
n <- 500

# Initialize vector to store deviations
ate_differences <- numeric(n_simulations)

# Run 1,000 simulations
for (i in 1:n_simulations) {
  # Generate new random dataset
  data <- data.frame(
    Age = rnorm(n, mean = 42, sd = 10),
    Education = sample(1:4, n, replace = TRUE),
    Y0 = rnorm(n, mean = 50000, sd = 10000)
  )

  # Apply intervention effects
  data <- data %>%
    mutate(
      Y1 = Y0 + tau,
      D = sample(c(0, 1), n, replace = TRUE),  # Randomly assign treatment
      Y = ifelse(D == 1, Y1, Y0)               # Observed outcome
    )

  # Estimate ATE (difference-in-means)
  estimated_ate <- mean(data$Y[data$D == 1]) - mean(data$Y[data$D == 0])

  # Store the difference from true ATE
  ate_differences[i] <- estimated_ate - tau
}
```

```
# Compute the mean of the differences
mean_diff <- mean(ate_differences)

# Plot histogram of ATE differences
ggplot(data.frame(ate_differences), aes(x = ate_differences)) +
  geom_histogram(color = "black", fill = "blue", bins = 30, alpha = 0.7) +
  geom_vline(aes(xintercept = mean_diff), color = "red", linetype = "dashed") +
  labs(
    title = "Distribution of Estimated ATE Differences",
    x = "Estimated ATE - True ATE",
    y = "Frequency"
  ) +
  theme_minimal()
#install.packages("readstata13")
library(readstata13)

# Load the dta data
electwomen_data <- read.dta13("how_to_elect_more_women.dta")

# Create sd_onefem2014 dummy variable from prop_sd_fem2014
electwomen_data <- electwomen_data %>%
  mutate(sd_onefem2014 = ifelse(prop_sd_fem2014 > 0, 1, 0))

colnames(electwomen_data)
# Ensure there are no missing values
data <- electwomen_data %>% mutate(prop_sd_fem2014 = ifelse(is.na(prop_sd_fem2014), 0, prop_sd_fem2014)

# Group by 'condition' and calculate proportions correctly
condition_summary <- electwomen_data %>%
  group_by(condition) %>%
  summarise(
    precinct_count = n(),  # Number of precincts in each condition group
    sum_fem2014 = sum(prop_sd_fem2014, na.rm = TRUE),  # Sum of elected women proportion
    proportion_fem2014 = sum_fem2014 / sum(data$prop_sd_fem2014, na.rm = TRUE)  # Proportion per condit
  )

# Print results
print(condition_summary)
#test balance in pre-treatment variables of "agecat" and "race" betwen different conditions
# Define pre-treatment variables
pre_treatment_vars <- c("agecat", "race")

# Define unique conditions (excluding NA)
conditions <- c("Control", "Demand", "Supply", "Supply+Demand")

# Function to perform pairwise t-tests
perform_t_test <- function(var, group1, group2) {
  test_result <- electwomen_data %>%
    filter(condition %in% c(group1, group2)) %>%
    t.test(reformulate("condition", var), data = .)

  cat("\nT-test for", var, "between", group1, "and", group2, "\n")
  print(test_result)
```

```r
}

# Run t-tests for each variable between control and each treatment group
for (var in pre_treatment_vars) {
  for (group in conditions[conditions != "Control"]) {
    perform_t_test(var, "Control", group)
  }
}
# Function to run linear regression for ATE estimation
estimate_ate <- function(treatment_group) {
  # Subset to only include Control and the current treatment group
  subset_data <- electwomen_data %>%
    filter(condition %in% c("Control", treatment_group)) %>%
    mutate(treatment = ifelse(condition == treatment_group, 1, 0))  # Convert to binary

  # Run linear regression
  model <- lm(sd_onefem2014 ~ treatment, data = subset_data)

  # Print results
  cat("\nATE Estimate for", treatment_group, "vs. Control:\n")
  print(summary(model))
}

# Run regressions for each treatment group
treatment_groups <- c("Demand", "Supply", "Supply+Demand")

for (treatment in treatment_groups) {
  estimate_ate(treatment)
}
# Create dummy variables for each treatment group
electwomen_data <- electwomen_data %>%
  mutate(
    treatment_demand = ifelse(condition == "Demand", 1, 0),
    treatment_supply = ifelse(condition == "Supply", 1, 0),
    treatment_supply_demand = ifelse(condition == "Supply+Demand", 1, 0)
  )

# Run a single regression with all treatment groups
full_model <- lm(sd_onefem2014 ~ treatment_demand + treatment_supply + treatment_supply_demand, data = 

# Print results
summary(full_model)
# this chunk generates the complete code appendix.
# eval=FALSE tells R not to re-run (``evaluate'') the code here.
```