MY457: Solution Set 1 - Potential Outcomes and Randomized Experiments

WT 2025

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 12pm (noon) on Tue/11/Mar. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a simple study of the effect of a treatment $D_i \in \{0, 1\}$ on Y_i for all $i \in \{1, 2, 3, \ldots, N\}$.

1.1. Explain the notation Y_{1i}. For the same unit i, when would this quantity be equal to Y_i?

Y_{i1} or Y_{i1} or $Y_i(1)$ typically refers to the potential outcome under treatment for unit i. Y_i by contrast refers to the realized outcome of unit i. For the same unit i, we would expect Y_{1i} and Y_i to be the same when individual i got treated, hence $D_i = 1$.

1.2. What is the difference between $\mathbb{E}[Y_{0i}|D_i = 1]$ and $\mathbb{E}[Y_{0i}|D_i = 0]$? When would you expect these quantities to be equal? When would you expect them to be unequal?

$\mathbb{E}[Y_{0i}|D_i = 1]$ is the expected value of the potential outcome without treatment for the treated group. $\mathbb{E}[Y_{0i}|D_i = 0]$ is the expected value of potential outcomes without treatment for the control group.

When there are no confounders, and the randomization process is done correctly, we expect these quanitites to be equal. However, if there is some sort of selection bias, we could expect these to be different.

1.3. Explain how randomly assigning individuals into treatment ($D = 1$) and control ($D = 0$) allows for the identification of the average treatment effect (ATE).

Random assignment ensures that individuals in both treatment and control are similar on average in both observed and unobserved characteristics. This is called the balancing principle, both in terms of potential outcomes and covariates.

If balance in covariates holds, then we suppose that the expected value of potential outcomes of both groups are the same, effectively eliminating selection bias. If there is no selection bias, we can identify the ATE as:

$$ATE = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

2 Simulations

In this question will use simulated data to test some of our intuitions about randomised experiments. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Suppose we are planning a trial to test the effectiveness of a policy on the wages of workers, with what we assume is a constant average treatment effect τ_{ATE}.

The real trial data that we collect will include information on participant characteristics, specifically age and education, along with actual treatment assignment (D), and the outcome variable (Y).

To explore some of the properties of our proposed trial, we first simulate some data. Explain in words what the code below does.

```
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)
)
```

We define a set with 500 observations. We assign each observation a value for Age and for Education, as well as an expected outcome, when not treated, Y_0 with mean 50,000 and standard deviation 10,000.

Additionally, we set a treatment effect τ to be of 5,000, hence our outcome under treatment, Y_1, is 5,000 higher than Y_0. This is a homogenous and constant treatment effect.

Finally, we randomly assign individuals to treatment or control D, and observe either Y_0 or Y_1 as Y, depending on treatment.

We are simulating an RCT, where treatment status is independent of both potential outcomes and covariates.

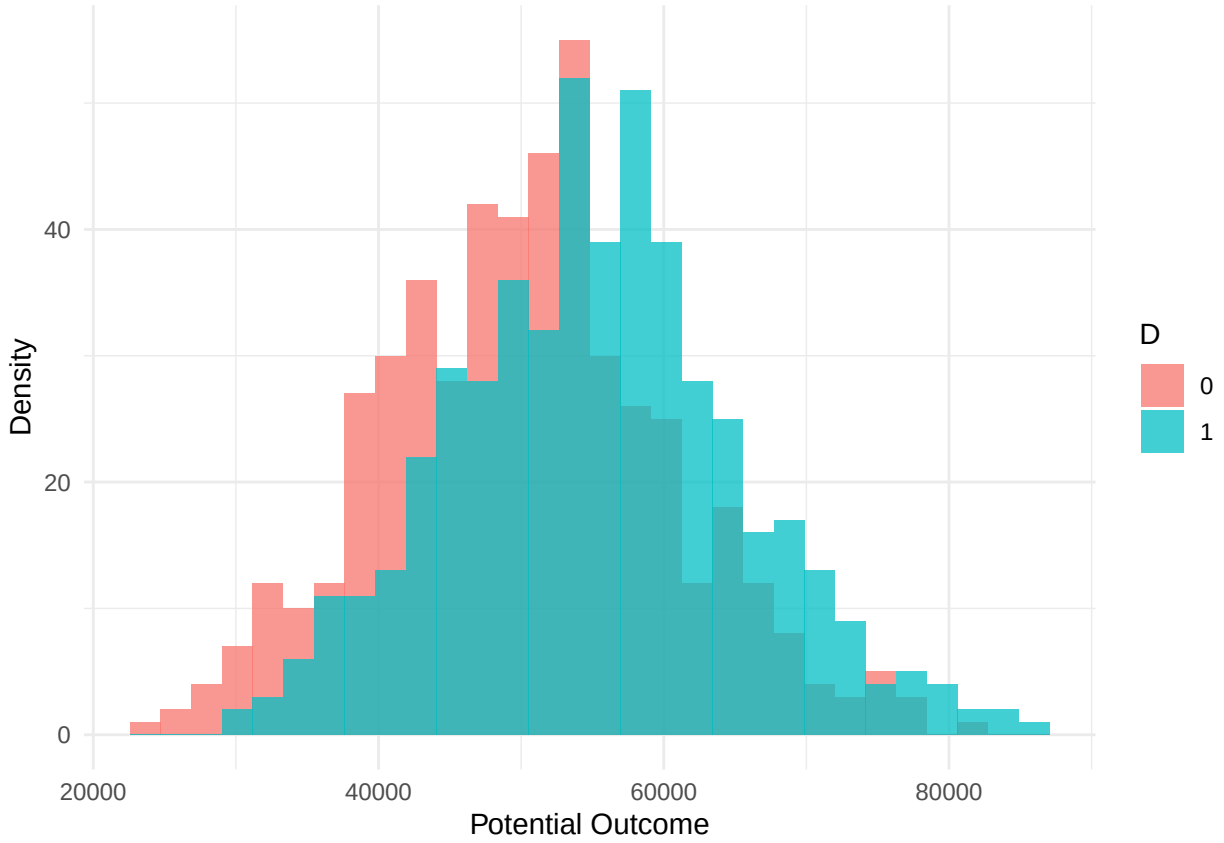
2.2. Randomisation implies the balancing principle. One way of testing for balance is by showing that the means of the two groups is statistically indistinguishable. Using a t-test, assess whether both age and education are balanced across conditions.

Variable	Control	Treatment	P.Value
Age	41.7782408172098	42.9809180589045	0.168905352781632
Education	2.48106060606061	2.48106060606061	0.592515711377063

For both Age and Education we find no evidence that the two groups have different means. On average, we expect treated and control units to be very similar between each other, this

suggests that the randomization is not conditional on any covariate and there is balance in the covariates.

2.3. Let's look at the difference in **potential outcomes** by treatment condition. One attractive way of doing this is through a histogram of our potential outcomes, as shown below.

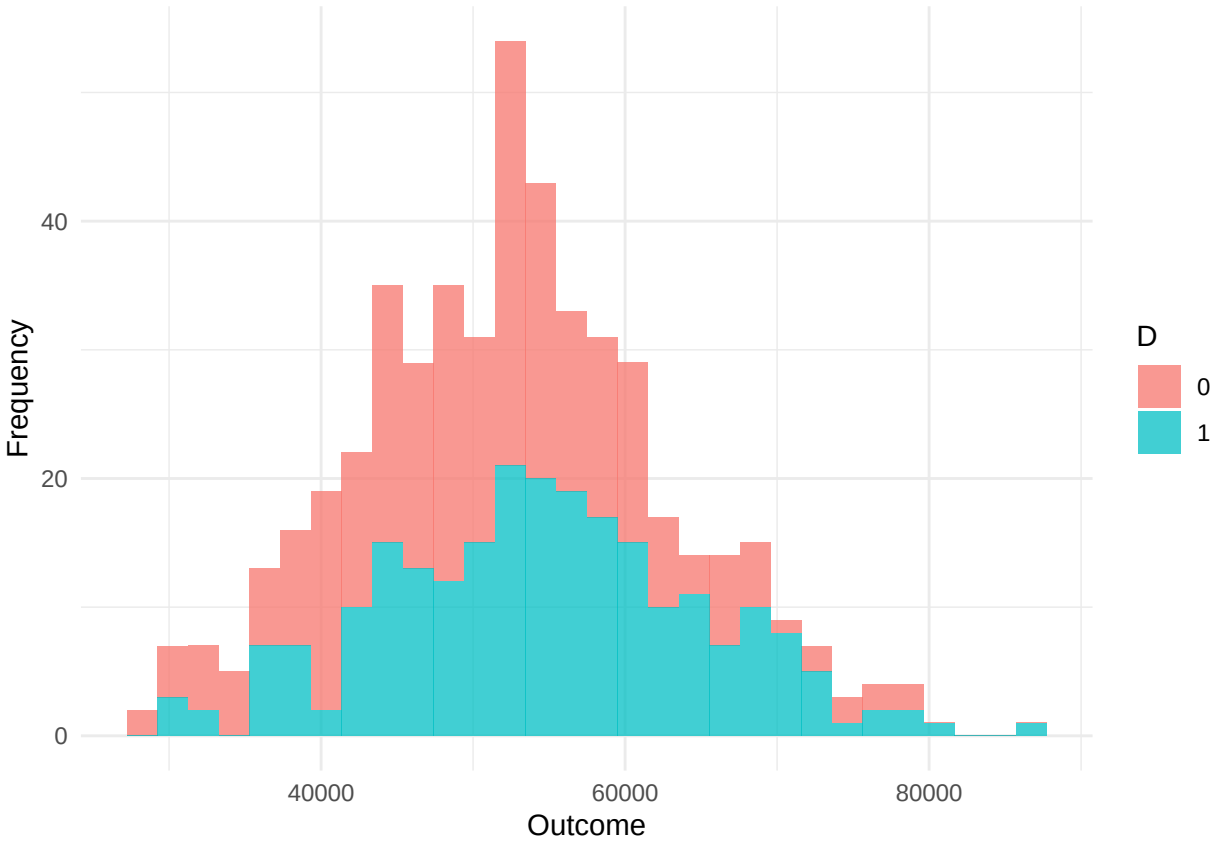


Calculate the difference between the potential outcomes as a difference-in-means (do not worry about statistical inference). What do you find? Is this surprising or unsurprising?

Diff.
5000

When we look at the difference in means, we have a difference of exactly 5,000. This is not surprising since we set τ to be exactly 5,000, and when we use potential outcomes there is no 'sampling error' (in terms of which potential outcome is revealed as observed data – there may be sampling error if we are concerned about population-level inferences).

2.4. Let's assess the effect of our intervention on the **realised outcome** Y. First, generate a plot that shows the distribution of Y conditional on treatment status D. Here is a link to an introduction to ggplot for plotting. You can use any other package to generate this plot if you wish, including base `R`.



The plot shows that the distribution of Y for individuals under treatment is slightly higher than that of individuals under control.

Second, estimate the average treatment effect (ATE). Since we have a randomized controlled trial, one obvious estimator is the difference-in-means between the two groups. An alternative estimator would be linear regression (OLS). Implement both estimators. What do you notice? How does your result compare to the 'ground truth' answer you calculated in Question 2.3?

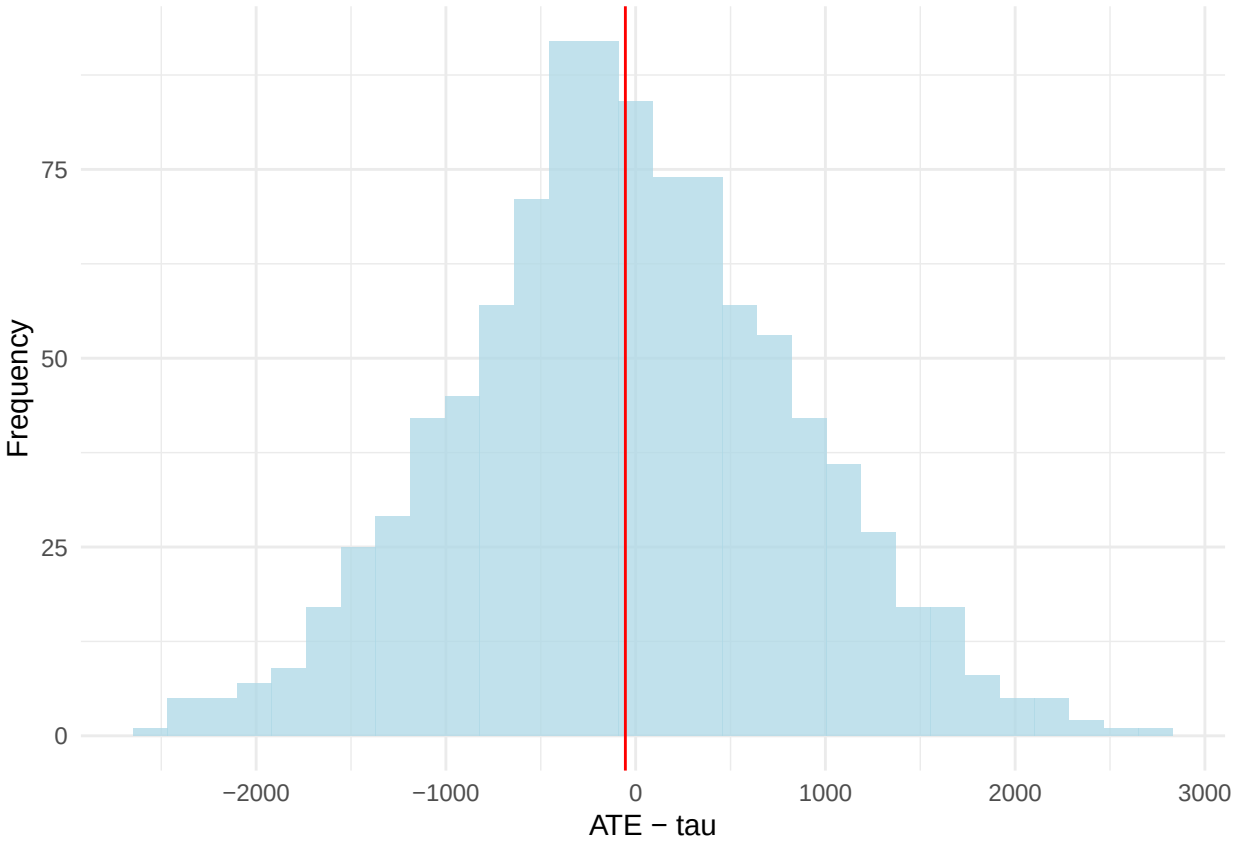
Diff.
4181.018

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50612.031	625.2493	80.946961	0.0000000
D	4181.018	910.0856	4.594093	0.0000055

Both estimators give the same result. We find a difference of 4,181, lower than the true ATE.

The difference between the true ATE and our estimated ATE is probably due to chance.

2.5. (Extra credit): Show that the answer to Question 2.4 was not due to chance (a 'lucky draw'). Over at least 1,000 repeat simulations (note: you will want to remove the `set.seed` step), generate a fresh draw of the data and estimate the sample-specific ATE with observed data. Calculate the difference between this and what you know to be the true (fixed) value of τ_{ATE} and store that difference. Finally, produce a histogram that shows the distribution of that difference over your repeated samples, along with its mean. What do you conclude?



When we replicate the exercise with random selections of the data, we find that the difference between the true ATE and the one we estimate converges to 0 on average. This suggests that our previous finding (of a large difference between the ATE and the estimated ATE was indeed due to random chance given our particular draw from the data generating process.

3 Replication

In this question we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *How to Elect More Women: Gender and Candidate Success in a Field Experiment*.

In the USA, women tend to be underrepresented in legislative bodies. The authors designed a field experiment to test whether messages from party leaders can affect women's electoral success. The general belief is that there are two factors that explain why few women are elected. The first factor is related to the so-called 'supply side', with fewer female candidates vying for office. However, particularly among conservatives, voters' biases, also called the 'demand side', may play an important role. In the experiment messages are sent to leaders of precinct-level caucus meetings to see if tackling the demand side, the supply side, or both jointly, can increase the number of women who are elected.

The messages were divided into 4 categories: 1) Placebo control, 2) Supply messages, 3) Demand messages, and 4) Supply+Demand messages. In the data, group 1 is the control group receiving a placebo message unrelated to the aforementioned factors. Groups 2 to 4 represent the different treatment groups with messages relating to both factors.

3.1. Read into `R` the replication data file `how_to_elect_more_women.dta`. These data are at the precinct level. Using `prop_sd_fem2014`, the proportion of state delegates elected from the precinct in 2014 who were women, create a new dummy variable called `sd_onefem2014` that takes a value of 1 if at least one women was elected within the precinct in 2014, and 0 otherwise. This will be our outcome variable.

```r
data <- read_dta("./how_to_elect_more_women.dta")

data$sd_nofem2014 <- ifelse(data$prop_sd_fem2014 == 0, 1, 0) # Create variable for counties with 0 W el

data$sd_onefem2014 <- 1 - data$sd_nofem2014 # Create dummy for at least 1 W elected

data <- data %>%
  select(sd_onefem2014, condition, age, gender, yearborn, religion,
         race, income)

head(na.omit(data)) %>% kable()
```

sd_onefem2014	condition	age	gender	yearborn	religion	race	income
1	3	39	1	19	3	5	4
0	1	64	1	44	3	5	4
1	2	55	1	35	3	5	5
0	1	45	1	25	3	5	4
0	3	48	2	28	3	5	4
0	1	66	1	46	3	5	6

3.2. Show the proportion of precincts at each treatment/control group.

Group	Frequency
1	25.09276
2	25.00000
3	24.95362
4	24.95362

We have around 25% of total observations in each of the categories.

3.3. Take two pre-treatment variables of your choice that are not the outcome (`sd_onefem2014`) and test whether there is balance between the treatment groups.

Group.1.vs.	P.value
2	0.4467990
3	0.6043137
4	0.5634717

Groups	P.value
2 - 3	0.791385462964775
2 - 4	0.854318981692281
3 - 4	0.93992761133711

We first look at Age. Comparing the control vs. the different treatment groups, we find no evidence for a statistical difference in the group means. This holds when we compare treatment groups.

Group.1.vs.	P.value
2	0.6422008
3	0.6588426
4	0.6279856

Groups	P.value
2 - 3	0.35928482683312
2 - 4	0.981254139125381
3 - 4	0.350554660928727

Comparing the control vs. the different treatment groups for Gender, we find no evidence for a statistical difference in the group means. This holds when we compare treatment groups.

3.4. Estimate the ATE of the different treatments. Hint: you can use three separate linear regressions, where you subset the data to just the control condition and one of the treatments, to estimate the effect of each treatment level. What do you find? Is there any effect of treatment. Why do you think this is happening?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3752759	0.0230496	16.281242	0.0000000
factor(condition)2	0.0587666	0.0323009	1.819347	0.0691831

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3752759	0.0230391	16.288681	0.0000000
factor(condition)3	0.0574595	0.0327097	1.756647	0.0793192

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3752759	0.0230948	16.249396	0.0000000
factor(condition)4	0.0784487	0.0328447	2.388471	0.0171253

We do not find a treatment effect of treatment status 2 and 3 compared to the control group that is statistically significant at conventional levels. However, we do find an effect for treatment status 4. However, it is worth noting that the substantive point estimates are not that different between the three conditions, suggesting that it is possible our first two results are false negatives, or our third result is a false positive. We should carefully consider whether our tests were appropriately powered. Taking the results at face value (demand and supply don't work independently, but do work together, we might argue that:

The lack of an effect in individuals' efforts may come from a setting where there is asymmetric information. Suppose you only send messages trying to tackle the supply part, you could promote more women running for office. However, if there are biases towards voting for men (demand side), more female candidates will likely not result in more elected candidates.

If one only tackles the demand side, the opposite happens, even though more voters are more sensitive to the topic and want to elect more women, if there are no canditates this is not possible.

Tackling both factors can effectively increase the number of women running for post and reduce the biases resulting in an increase of elected women.

3.5. (Extra credit): Simultaneously estimate the ATE of the different treatments using a single linear regression. Do any of your conclusions change?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3752759	0.0232029	16.173663	0.0000000
factor(condition)2	0.0587666	0.0325158	1.807326	0.0708777
factor(condition)3	0.0574595	0.0329424	1.744243	0.0812867
factor(condition)4	0.0784487	0.0329985	2.377339	0.0175415

Running a regression with the three treatment groups simultaneously yields the same results as the three separate regressions. However, we see a difference in the standard deviation of the effect. Note that these regressions can sometimes be subject to 'contamination bias.'

3.1 Appendix

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, linewidth=60)

# you can include your libraries here:
library(tidyverse)
```

```r
library(knitr)
library(haven)

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

# 2 ------
## 2.2
# Age
t.test(data$Age[data$D == 1], data$Age[data$D == 0])

# Education
t.test(data$Education[data$D == 1], data$Education[data$D == 0])

# 2.3
data %>%
  ggplot() +
  geom_histogram(aes(x = Y0, fill = "0"), alpha = 0.75) +
  geom_histogram(aes(x = Y1, fill = "1"), alpha = 0.75) +
  labs(fill = "D") +
  xlab("Potential Outcome") +
  ylab("Density") +
  theme_minimal()

# Difference in means
mean(data$Y1) - mean(data$Y0)

# 2.4
data %>%
  ggplot() + aes(x = Y, fill = factor(D)) +
  geom_histogram(alpha = 0.75) +
  labs(fill = "D") +
  xlab("Outcome") +
  ylab("Frequency") +
  theme_minimal()

# Difference in means
mean(data$Y[data$D == 1]) - mean(data$Y[data$D == 0])

# OLS
summary(lm(Y~D, data = data))

# 2.5
ATE <- NULL

set.seed(123)
for (i in 1:1000) {
  n  <- 500

  tau <- 5000

  data <- data.frame(
    Age       = rnorm(n, mean = 42, sd = 10),
```

```r
    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)
  )


  ATE[i] <- mean(data$Y[data$D == 1]) - mean(data$Y[data$D == 0])
}

data.frame(ATE) %>%
  ggplot() + aes(x = ATE - tau) +
  geom_histogram(fill = "lightblue", alpha = 0.75) +
  geom_vline(xintercept = mean(ATE - tau), color = "red") +
  ylab("Frequency") +
  xlab("ATE - tau") +
  theme_minimal()

# 3 -----
data <- read_dta("./how_to_elect_more_women.dta")

data$sd_nofem2014 <- ifelse(data$prop_sd_fem2014 == 0, 1, 0) # Create variable for counties with 0 W el

data$sd_onefem2014 <- 1 - data$sd_nofem2014 # Create dummy for at least 1 W elected

data <- data %>%
  select(sd_onefem2014, condition, age, gender, yearborn, religion,
         race, income)

# 3.2
prop.table(table(data$condition)) * 100

# 3.3
# Age
## Control vs Treatment
t.test(data$age[data$condition == 1], data$age[data$condition == 2])
t.test(data$age[data$condition == 1], data$age[data$condition == 3])
t.test(data$age[data$condition == 1], data$age[data$condition == 4])

## Different tretament status
t.test(data$age[data$condition == 2], data$age[data$condition == 3])
t.test(data$age[data$condition == 2], data$age[data$condition == 4])
t.test(data$age[data$condition == 3], data$age[data$condition == 4])

# Gender
## Control vs Treatment
t.test(data$gender[data$condition == 1], data$gender[data$condition == 2])
t.test(data$gender[data$condition == 1], data$gender[data$condition == 3])
t.test(data$gender[data$condition == 1], data$gender[data$condition == 4])
```

```r
## Different tretament status
t.test(data$gender[data$condition == 2], data$gender[data$condition == 3])
t.test(data$gender[data$condition == 2], data$gender[data$condition == 4])
t.test(data$gender[data$condition == 3], data$gender[data$condition == 4])

# 3.4
summary(lm(sd_onefem2014~factor(condition), data = data[data$condition == 1 | data$condition == 2, ]))

summary(lm(sd_onefem2014~factor(condition), data = data[data$condition == 1 | data$condition == 3, ]))

summary(lm(sd_onefem2014~factor(condition), data = data[data$condition == 1 | data$condition == 4, ]))

# 3.5
summary(lm(sd_onefem2014~factor(condition), data = data))
```