Randomized Experiments

An Adventure in Nonparametric Inference

Here's the packages we're using.

```
# generally useful packages
library(tidyverse)
library(modelsummary)
```

Opening Comments

Opening Comments

This week, we continue applied modeling!

1. **engines**: design-based inference!
2. **distribution**: none!
3. **quantities of interest**: the ATE!
4. **evaluating models** frequentist properties

The notation here follows Imbens and Rubin (2015). An exception is that I use ATE to represent their τ_{fs} and $\widehat{\text{ATE}}$ to represent their $\hat{\tau}_{\text{dif}}$.

Review

- For discrete random variable X, $E(X) = \sum_{\text{all } x} x \cdot \Pr(X = x)$
- $\text{Var}(X) = E\left[(X - E(X))^2\right] = E\left[X^2\right] - E(X)^2$.

Our goal is to obtain point estimates, variance estimates, and confidence intervals from randomized experiments **using only the assumptions of the design**.

ATE

ATE

We have N units in our experiment and we assign N_t units to treatment and $N_c = N - N_t$ units to control. We require that $1 < N_t < N$, so that at least one unit is assigned to treatment and control.

We use the random variable W_i to assign unit i to treatment ($W_i = 1$) or control ($W_i = 0$).

For each unit i, we have two potential outcomes: $Y_i(1)$ if assigned to treatment and $Y_i(0)$ if assigned to control.

For unit $i \in \{1, ..., N\}$, we have one observed potential outcome, denoted by Y_i^{obs}.

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

For each unit, we also have one missing potential outcome, denoted by Y_i^{mis}:

$$Y_i^{\text{mis}} = Y_i(1 - W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1. \end{cases}$$

So what is this $Y_i(\cdot)$ thing?

A Data Set of Fully Observed Potential Outcomes

Unit	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	0	1	1
2	1	2	1
3	0	0	0
4	1	3	2
5	0	1	1
6	1	2	1
7	0	0	0
8	1	3	2
9	0	1	1
10	1	2	1

Table 1: Mock Dataset with Potential Outcomes and Their Differences

Average Treatment Effect: Lets use these potential outcomes to define the average treatment effect (ATE) as
ATE $= \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1) - Y_i(0)\right] = \overline{Y}(1) - \overline{Y}(0)$, where $\overline{Y}(0)$ and $\overline{Y}(1)$ are the averages of the potential control and treated outcomes respectively, so that $\overline{Y}(0) = \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$ and $\overline{Y}(1) = \frac{1}{N}\sum_{i=1}^{N} Y_i(1)$.

In our case, we are explicitly interested in what we call today the **sample average treatment effect** (SATE). That is, were are interested in the ATE among the units *in our sample* not an existing or imagined larger population.

Two Potential Realization for the Experiment

Unit	$Y_i(0)$	$Y_i(1)$	W_i
1	?	1	1
2	1	?	0
3	?	0	1
4	1	?	0
5	0	?	0
6	?	2	1
7	0	?	0
8	?	3	1
9	0	?	0
10	?	2	1

Unit	$Y_i(0)$	$Y_i(1)$	W_i
1	0	?	0
2	?	2	1
3	0	?	0
4	?	3	1
5	?	1	1
6	1	?	0
7	?	0	1
8	1	?	0
9	?	1	1
10	1	?	0

Notice that we cannot compute the ATE from these data because we cannot never obtain $Y_i(1) - Y_i(0)$. (This is the fundamental problem of causal inference).

$\widehat{\text{ATE}}$

An Estimator for the Average Treatment Effect

Suppose that we observe data from a completely randomized experiment in which $N_t = \sum_{i=1}^{N} W_i$ units are randomly selected to be assigned to treatment and the remaining $N_c = \sum_{i=1}^{N}(1 - W_i)$ are assigned to control. Because of the randomization, a natural[1] estimator for the average treatment effect is the difference in the average outcomes between those assigned to treatment and those assigned to control, so that

$$\widehat{\mathsf{ATE}} = \overline{Y}_t^{\mathsf{obs}} - \overline{Y}_c^{\mathsf{obs}},$$

where

$$\overline{Y}_c^{\mathsf{obs}} = \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{\mathsf{obs}} \quad \text{and} \quad \overline{Y}_t^{\mathsf{obs}} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\mathsf{obs}}.$$

Recall that $Y_i^{\mathsf{obs}} = Y_i(1)$ if $W_i = 1$ and $Y_i^{\mathsf{obs}} = Y_i(0)$ if $W_i = 0$. Thus we can write $\widehat{\mathsf{ATE}}$ as

$$\widehat{\mathsf{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1 - W_i) \cdot Y_i(0)}{N_c/N} \right).$$

[1] See next slide

Why $\widehat{\text{ATE}}$?

Why is $\widehat{\text{ATE}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ a "natural'' estimator of the ATE?

For both intuitive and rigorous reasons, simply replacing population moments (e.g., the population mean) with sample moments (e.g., the sample mean) produces a good estimator.

Here's a sketch of the theory. For distributions F that are "not too weird'' (i.e., meet fairly weak regularity conditions), then the empirical cdf $\hat{F}$ is a "good substitute" for F. It's a good substitute in the sense that $\hat{F}_N \to F$ as $N \to \infty$. This is an important "brute force" approach to nonparametric estimation.

In our case, we don't use it to defend the estimator, but to obtain a starting point.

What makes $\widehat{\text{ATE}}$ uncertain?

From our usual parametric modeling perspective, we *imagine* that the data are draw from a distribution $y_i \sim f(\theta_i)$.

But this approach to randomized experiments is not like that; we're not imagining the $Y_i(1)$ and $Y_i(0)$ are random. Instead, they are fixed values.

The only thing that's random in the equation below is W_i, which is the assignment to treatment and control. Because we designed that mechanism, we know its distribution. Thus we have *design-based inference* rather than model-based inference.

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1 - W_i) \cdot Y_i(0)}{N_c/N} \right).$$

Is $\widehat{\text{ATE}}$ a good estimator?

Theorem: $\widehat{\text{ATE}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is an unbiased estimator of ATE.

Proof.

We're trying to show that $E_W\left[\widehat{\text{ATE}}\right] = \text{ATE}$. *Walk through* $(2) \rightarrow (3)$ *together.*

$$\begin{aligned}
E_W\left[\widehat{\text{ATE}}\right] &= E_W\left[\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right] && (1)\\
&= E_W\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1-W_i)\cdot Y_i(0)}{N_c/N}\right)\right] && (2)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(E_W[W_i]\cdot Y_i(1)\frac{1}{N_t/N} - E_W[1-W_i]\cdot Y_i(0)\frac{1}{N_c/N}\right) && (3)\\
&= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) && (4)\\
&= \text{ATE} && (5)
\end{aligned}$$

Done! $\widehat{\text{ATE}}$ is an unbiased estimator of ATE under the assumptions of the design. No parametric model of the data needed, just a (known) model of the the design.

$\text{Var}\left(\widehat{\text{ATE}}\right)$

Some Initial Results

To calculate the variance of $\widehat{\text{ATE}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, we need $E_W[W_i^2]$ and $\text{Var}_W(W_i)$.

We have

$$E_W[W_i^2] = E_W[W_i] = \frac{N_t}{N} \quad \text{and} \quad \text{Var}_W(W_i) = \frac{N_t}{N}\left(1 - \frac{N_t}{N}\right).$$

Proof. Homework Exercise.

We also need $E_W[W_i \cdot W_{i'}]$.

$$E_W[W_i \cdot W_{i'}] = P_W(W_i = 1) \cdot P_W(W_{i'} = 1 | W_i = 1) = \frac{N_t}{N} \cdot \frac{N_t - 1}{N - 1}, \quad \text{for } i \neq j,$$

Theorem The variance of $\hat{\text{ATE}}$ is

$$\text{Var}_W(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N},$$

where S_c^2 and S_t^2 are the variances of $Y_i(0)$ and $Y_i(1)$:

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(0) - \overline{Y}(0))^2, \quad \text{and} \quad S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(1) - \overline{Y}(1))^2,$$

and S_{tc}^2 is the variance of the unit-level treatment effects:

$$S_{tc}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))\right)^2.$$

Proof. Long and tedious! So here we go...

Remember,

$$\widehat{\text{ATE}} = \frac{1}{N_t}\sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_c}\sum_{i=1}^{N}(1-W_i)\cdot Y_i^{\text{obs}}.$$

Re-arranging, we have

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_t}\cdot W_i \cdot Y_i(1) - \frac{N}{N_c}\cdot(1-W_i)\cdot Y_i(0)\right).$$

It happens to be helpful to replace W_i with D_i.

$$D_i = W_i - \frac{N_t}{N} = \begin{cases} \frac{N_c}{N} & \text{if } W_i = 1, \\ -\frac{N_t}{N} & \text{if } W_i = 0. \end{cases}$$

Facts:

- $E(D_i) = 0$.
- $\text{Var}_W(D_i) = E[D_i^2] = \frac{N_t N_c}{N^2}$.

Proof. Homework exercise.

Another fact: For $i \neq j$ the distribution of $D_i \cdot D_j$ is

$$P_W(D_i \cdot D_j = d) = \begin{cases} \frac{N_t \cdot (N_t - 1)}{N \cdot (N-1)} & \text{if } d = \frac{N_c^2}{N^2}, \\ \frac{2 \cdot N_t \cdot N_c}{N \cdot (N-1)} & \text{if } d = -\frac{N_t N_c}{N^2}, \\ \frac{N_c \cdot (N_c - 1)}{N \cdot (N-1)} & \text{if } d = \frac{N_t^2}{N^2}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$E_W[D_i \cdot D_j] = \begin{cases} \frac{N_c \cdot N_t}{N^2} & \text{if } i = j, \\ -\frac{N_t \cdot N_c}{N^2 \cdot (N-1)} & \text{if } i \neq j. \end{cases}$$

Proof. Homework exercise.

Written using D_i rather than W_i, the estimate is:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_t}\cdot D_i + \frac{N_t}{N}\right)\cdot Y_i(1) - \frac{N}{N_c}\left(\frac{N_c}{N} - D_i\right)\cdot Y_i(0).$$

Re-arranging, this becomes

$$= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) + \frac{1}{N}\sum_{i=1}^{N} D_i \cdot \left(\frac{N}{N_t}\cdot Y_i(1) + \frac{N}{N_c}\cdot Y_i(0)\right).$$

Exercise: Using the equation above, show that $\widehat{\text{ATE}}$ is an unbiased estimator of ATE.

$$\widehat{\text{ATE}} = \overbrace{\frac{1}{N}\sum_{i=1}^{N}(Y_i(1)-Y_i(0))}^{\text{fixed}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} D_i \cdot \overbrace{\left(\frac{N}{N_t}\cdot Y_i(1) + \frac{N}{N_c}\cdot Y_i(0)\right)}^{Y_i^+}}_{\text{random variable}}$$

Let $Y_i^+ = (N/N_t)Y_i(1) + (N/N_c)Y_i(0)$. Then

$$\text{Var}_W(\widehat{\text{ATE}}) = \text{Var}_W\left(\frac{1}{N}\sum_{i=1}^{N} D_i \cdot Y_i^+\right) = \frac{1}{N^2}E_W\left[\left(\sum_{i=1}^{N} D_i \cdot Y_i^+\right)^2\right].$$

Expanding, we get:

$$\text{Var}_W(\widehat{\text{ATE}}) = E_W\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_i D_j Y_i^+ Y_j^+\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}(Y_i^+)^2 E_W[D_i^2] + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i} E_W[D_i \cdot D_j]\cdot Y_i^+ \cdot Y_j^+$$

$$= \frac{N_c \cdot N_t}{N^4}\sum_{i=1}^{N}(Y_i^+)^2 - \frac{N_c \cdot N_t}{N^4\cdot(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} Y_i^+ \cdot Y_j^+$$

$$= \frac{N_c \cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^{N}(Y_i^+)^2 - \frac{N_c \cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} Y_i^+ \cdot Y_j^+$$

$$
\begin{aligned}
&= \frac{N_c \cdot N_t}{N^3 \cdot (N-1)} \sum_{i=1}^{N} \left(\frac{N}{N_t} \cdot Y_i(1) - \frac{N}{N_t} \cdot \overline{Y}(1) \right)^2 \\
&\quad + \frac{N_c \cdot N_t}{N^3 \cdot (N-1)} \sum_{i=1}^{N} \left(\frac{N}{N_c} \cdot Y_i(0) - \frac{N}{N_c} \cdot \overline{Y}(0) \right)^2 \\
&\quad + \frac{2 \cdot N_c \cdot N_t}{N^3 \cdot (N-1)} \sum_{i=1}^{N} \left(\frac{N}{N_t} \cdot Y_i(1) - \frac{N}{N_t} \cdot \overline{Y}(1) \right) \cdot \left(\frac{N}{N_c} \cdot Y_i(0) - \frac{N}{N_c} \cdot \overline{Y}(0) \right) \\
&= \frac{N_c}{N \cdot N_t \cdot (N-1)} \sum_{i=1}^{N} (Y_i(1) - \overline{Y}(1))^2 + \frac{N_t}{N \cdot N_c \cdot (N-1)} \sum_{i=1}^{N} (Y_i(0) - \overline{Y}(0))^2 \\
&\quad + \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N} (Y_i(1) - \overline{Y}(1)) \cdot (Y_i(0) - \overline{Y}(0)).
\end{aligned}
$$

Recall the definition of S_{tc}^2, which implies that

$$\begin{aligned}
S_{tc}^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1)-\overline{Y}(1)-(Y_i(0)-\overline{Y}(0))\right)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1)-\overline{Y}(1)\right)^2+\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0)-\overline{Y}(0)\right)^2 \\
&\quad -\frac{2}{N-1}\sum_{i=1}^{N}\left(Y_i(1)-\overline{Y}(1)\right)\cdot\left(Y_i(0)-\overline{Y}(0)\right) \\
&= S_t^2+S_c^2-\frac{2}{N-1}\sum_{i=1}^{N}\left(Y_i(1)-\overline{Y}(1)\right)\cdot\left(Y_i(0)-\overline{Y}(0)\right).
\end{aligned}$$

Hence, the expression in (Result 1) is equal to

$$\begin{aligned}
\text{Var}_W(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}) &= \frac{N_c}{N\cdot N_t}\cdot S_t^2+\frac{N_t}{N\cdot N_c}\cdot S_c^2 \\
&\quad +\frac{1}{N}\left(S_t^2+S_c^2-S_{tc}^2\right) \\
&= \frac{S_t^2}{N_t}+\frac{S_c^2}{N_c}-\frac{S_{tc}^2}{N}.
\end{aligned}$$

$$\widehat{\mathrm{Var}}\left(\widehat{\mathrm{ATE}}\right)$$

$\widehat{\mathrm{Var}}\left(\widehat{\mathrm{ATE}}\right)$

For a given set of potential outcomes, we now know the variance. But to make inferences, we need to be able to *estimate* this variance using the observed data Y_i^{obs}.

An unbiased estimator for S_c^2 is

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i(0) - \overline{Y}_c^{\text{obs}}\right)^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right)^2.$$

Similarly, an unbiased estimator for S_c^2 is

$$s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(Y_i(1) - \overline{Y}_t^{\text{obs}}\right)^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}}\right)^2.$$

The third term, S_{tc}^2, is the tricky one. This is population variance of the *unobservable* unit-level treatment effects—we cannot estimate this quantity.

We can only make assumptions about this quantity.

Assumption 1 (Sharp Null): If the treatment effect $Y_i(1) - Y_i(0)$ equals zero, then an unbiased estimator of the variance of $\widehat{ATE}$ is

$$\widehat{\text{Var}}(\widehat{ATE}) = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}. \qquad \text{(Neyman Variance Estimator)}$$

Assumption 2 (Constant Effects): If the treatment effect $Y_i(1) - Y_i(0)$ is constant, then an unbiased estimator of the variance of $\widehat{ATE}$ is also Neyman's variance estimator.

Without an Assumption: For any $Y_i(1)$ and $Y_i(0)$, then expectation of Neyman's variance estimator is *at least* as large as the actual variance.

$$\text{Var}_W(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}) = \frac{S_t^2}{N_t} + \frac{S_c^2}{N_c} - \frac{S_{tc}^2}{N}.$$

Asymptotics

A CI with Conservative Coverage Asymptotically

In Section 3 (pp. 1763-1763), Li and Deng (2017) provide a finite version of the CLT, so long as the potential outcomes are "not too weird.' '

A *conservative* $(1-\alpha) \times 100\%$ CI is $\widehat{\text{ATE}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \cdot \sqrt{\widehat{\text{Var}}\left(\widehat{\text{ATE}}\right)}$.

For a 90% confidence interval, this would be $\widehat{\text{ATE}} \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}\left(\widehat{\text{ATE}}\right)}$.

Asymptotically, this interval has coverage of *at least* $(1-\alpha) \times 100\%$.

Best Practices

Best Practices

Estimation

- It turns out that OLS with robust standard errors (HC2) is equivalent to the different-in-means and Neyman variance estimator we proposed above. See Samii and Aronow (2016).
- In practice, we obtain estimates with something like the following:

```
fit <- lm(y ~ treatment_indicator, data = df)
V_hat <- sandwich::vcovHC(fit, type = "HC2")
```

Adjustment

- Some authors are critical of regression adjustment for experimental data (Freedman 2008).
- However, in my view the dangers are mostly theoretical. In real world problems, regression adjust will almost always give you better estimates, sometimes *much* better estimates
  - Simply include controls.
  - Include mean-centered controls fully interacted with treatment.
  - Select controls with LASSO.

Example

Example

The Coversight Experiment

```r
# load packages
library(tidyverse)

# read cleaned data from dropbox
coversight <- read_rds("https://www.dropbox.com/s/9grn8kkb5yzwagx/data.rds?raw=1")%>%
  select(pres_overall, amplify, failure, pid_strength, passed_mvc1) %>%
  filter(passed_mvc1) %>%  # remove respondents who failed MVC 1
  mutate(pid_strength_rs = arm::rescale(pid_strength)) # center partisan strength

# quick look
glimpse(coversight)
```

```
Rows: 1,462
Columns: 6
$ pres_overall    <dbl> 0, 1, -3, 0, 2, 3, 2, -3, -3, 0, --~
$ amplify         <fct> Amplify, Ignore, Amplify, Ignore, ~
$ failure         <fct> Failure, Failure, Failure, Failure~
$ pid_strength    <dbl> 3, 3, 3, 0, 2, 2, 3, 3, 2, 3, 3, 3~
$ passed_mvc1     <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
$ pid_strength_rs <dbl> 0.5027, 0.5027, 0.5027, -0.9731, 0~
```

```r
# use only data from amplify condition (failure/success varies)
cs_failure <- coversight %>%
  filter(amplify == "Amplify")

table(cs_failure$failure)
```

```
Success Failure
    366     364
```

```r
# make y for convenience
y <- cs_failure$pres_overall
treat <- as.numeric(cs_failure$failure == "Failure")

# compute point estimate
mean(y[treat == 1]) - mean(y[treat == 0])
```

```
[1] -1.23
```

```r
# compute variance estimate
var(y[treat == 1])/sum(treat) + var(y[treat == 0])/sum(1 - treat)
```

```
[1] 0.0147
```

```r
# ols with robust ses
fit <- lm(y ~ treat)
coef(fit)["treat"]
```

```
treat
-1.23
```

```r
diag(sandwich::vcovHC(fit, type = "HC2"))["treat"]
```

```
 treat
0.0147
```

```r
fit1 <- lm(pres_overall ~ failure, data = cs_failure)
fit2 <- lm(pres_overall ~ failure + I(pid_strength_rs + 100), data = cs_failure)
fit3 <- lm(pres_overall ~ failure*pid_strength_rs, data = cs_failure)

modelsummary(list("No Controls" = fit1,
                  "One Control" = fit2,
                  "One Control, Interacted" = fit3),
             vcov = "HC2",
             gof_map = NA,
             escape = TRUE,
             output = "latex")
```

	No Controls	One Control	One Control, Interacted
(Intercept)	−0.650	17.973	−0.657
	(0.099)	(12.081)	(0.099)
failureFailure	−1.234	−1.224	−1.225
	(0.121)	(0.121)	(0.121)
I(pid_strength_rs + 100)		−0.186	
		(0.121)	
pid_strength_rs			−0.156
			(0.199)
failureFailure × pid_strength_rs			−0.061
			(0.241)