4.2 Experiments and Estimands

PLSC30500, Fall 2021

co-taught by Molly Offer-Westort & Andy Eggers

(This lecture with references to Gerber & Green (2012))

Estimation

Estimand

The *estimand* is the parameter of interest--it is the quantity that we would like to know about.

For example, when we care about causal effects, the estimand may be:

- the Average Treatment Effect (ATE)
- the Average effect of Treatment on the Treated (ATT)
- the Average effect of Treatment on the Control (ATC)

Why might these three quantities differ?

Notational aside: we often denote the estimand with the greek letter θ. Specific estimands may have conventional notations, such as μ for the mean or σ for the standard deviation.

Estimator

An *estimator* is a function of the data we observe; it is an informed guess about the value of the estimand. Below, the estimator is the function $g(\cdot)$.

$$g(X_1, \ldots, X_n)$$

Estimate

An *estimate* is what we calculate from our estimator with a specific set of data. Below, the estimate is the quantity $\hat{\theta}_n$.

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

Bias of an estimator

The bias of an estimator is the expected difference between the estimate and the true value of the estimand.

$$\mathrm{bias}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n] - \theta$$

An estimator is *unbiased* if

$$\mathrm{E}[\hat{\theta}_n] = \theta$$

The estimator is a function of the data, and so whether or not the estimator is biased for our estimand *also depends on the data we see*.

In randomized experiments, we are in a special setting where get input on what data we observe.

Randomization

Design

Gerber Green example:

	$Y_i(0)$	$Y_i(1)$	τ_i
Village 1	?	15	?
Village 2	15	?	?
Village 3	20	?	?
Village 4	20	?	?
Village 5	10	?	?
Village 6	15	?	?
Village 7	?	30	?

Design

Gerber Green example:

	$Y_i(0)$	$Y_i(1)$	τ_i
Village 1	?	15	?
Village 2	15	?	?
Village 3	20	?	?
Village 4	20	?	?
Village 5	10	?	?
Village 6	15	?	?
Village 7	?	30	?
Average	**16**	**22.5**	**?**

Target estimand: ATE

$$\mathrm{E}[\tau_i] = \mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)]$$

Under random assignment where everyone has the same probability of being assigned treatment,

$$\mathrm{E}[Y_i(1)|D_i = 1] = \mathrm{E}[Y_i(1)|D_i = 0]$$

Target estimand: ATE

$$\mathrm{E}[\tau_i] = \mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)]$$

Under random assignment where everyone has the same probability of being assigned treatment,

$$\begin{aligned}\mathrm{E}[Y_i(1)|D_i = 1] &= \mathrm{E}[Y_i(1)|D_i = 0] \\ &= \mathrm{E}[Y_i(1)]\end{aligned}$$

And,

$$\mathrm{E}[Y_i(0)|D_i = 0] = \mathrm{E}[Y_i(1)|D_i = 1]$$

Target estimand: ATE

$$\mathrm{E}[\tau_i] = \mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)]$$

Under random assignment where everyone has the same probability of being assigned treatment,

$$\begin{aligned}\mathrm{E}[Y_i(1)|D_i = 1] &= \mathrm{E}[Y_i(1)|D_i = 0] \\ &= \mathrm{E}[Y_i(1)]\end{aligned}$$

And,

$$\begin{aligned}\mathrm{E}[Y_i(0)|D_i = 0] &= \mathrm{E}[Y_i(1)|D_i = 1] \\ &= \mathrm{E}[Y_i(0)]\end{aligned}$$

We are then able to use our estimates of $\mathrm{E}[Y_i(1)|D_i = 1]$ and $\mathrm{E}[Y_i(0)|D_i = 0]$ to get an estimate of $\mathrm{E}[\tau_i]$. This is the difference in means estimator:

$$\hat{\tau}_{DM} = \frac{\sum_i^n Y_i D_i}{\sum_i^n D_i} - \frac{\sum_i^n Y_i(1 - D_i)}{\sum_i^n (1 - D_i)}$$

The difference-in-means estimator is unbiased for the average treatment effect estimand, when treatment assignment is random.

Plugging in from our example above,

	$Y_i(0)$	$Y_i(1)$	τ_i
Village 1	?	15	?
Village 2	15	?	?
Village 3	20	?	?
Village 4	20	?	?
Village 5	10	?	?
Village 6	15	?	?
Village 7	?	30	?
Average	**16**	**22.5**	**6.5**

$$\hat{\tau}_{DM} = \frac{\sum_i^n Y_i D_i}{\sum_i^n D_i} - \frac{\sum_i^n Y_i (1 - D_i)}{\sum_i^n (1 - D_i)}$$

$$= \frac{15 + 30}{2} - \frac{15 + 20 + 20 + 10 + 15}{5}$$

$$= 6.5$$

Considering some real data, we'll look at

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators

Data is available at the Yale ISPS data archive: isps.yale.edu/research/data

```
df <- read_csv('../data/legislators_email/Butler_Broockman_AJPS_2011_public_csv.cs
```

```
head(df)
```

```
## # A tibble: 6 x 15
##   leg_party leg_republican leg_black leg_latino reply_atall treat_deshawn treat_demprima
##   <chr>              <dbl>     <dbl>      <dbl>       <dbl>         <dbl>            <db
## 1 R                      1         0          0           1             0
## 2 D                      0         0          0           1             1
## 3 R                      1         0          0           0             1
## 4 R                      1         0          0           0             1
## 5 D                      0         0          0           0             0
## 6 D                      0         0          0           1             1
## # … with 7 more variables: treat_noprimary <dbl>, treat_group <dbl>, treat_jake <dbl>, l
## #   leg_white <dbl>, leg_notblackotherminority <dbl>, treat_primary <dbl>
```

Description

Emails are sent to state legislators. We signaled the race of the email sender by randomizing whether the email was signed by and sent from an email account with the name Jake Mueller or the name DeShawn Jackson.

Treatment is 1 if the sender was DeShawn Jackson, and 0 if Jake Mueller.

```
table(df$treat_deshawn)
```

```
##
##    0    1
## 2431 2428
```

The primary outcome is whether legislators replied at all.

```
table(df$reply_atall)
```

```
##
##    0    1
## 2112 2747
```

To get the difference-in-means estimate of the ATE,

```
Y1 <- filter(df, treat_deshawn == 1) %>% pull(reply_atall)
Y0 <- filter(df, treat_deshawn == 0) %>% pull(reply_atall)

head(Y1)
```

```
## [1] 1 0 0 1 1 1
```

```
head(Y0)
```

```
## [1] 1 0 1 1 1 1
```

```
mean(Y1) - mean(Y0)
```

```
## [1] -0.01782424
```

Legislators were 1.7 percentage points less likely to reply to an email if the sender was identified as DeShawn Jackson as compared to Jake Mueller.

The `estimatr` package produces a difference-in-means estimate for us; the standard errors give us information about how precise we think the point estimate is--we'll come back to these later.

```
library(estimatr)

difference_in_means(reply_atall ~ treat_deshawn, data = df)
```

```
## Design:  Standard
##                  Estimate Std. Error   t value  Pr(>|t|)    CI Lower   CI Upper       DF
## treat_deshawn -0.01782424  0.0142235 -1.253154 0.2102099 -0.04570874 0.01006026 4856.827
```

We can also look at conditional treatment effects. Here, we can condition on whether the legislator receiving the email was a Democrat vs. Republican.

$$\mathrm{E}[Y_i(1) - Y_i(0) | \text{party} = D]$$

```
Y1D <- filter(df, treat_deshawn == 1, leg_party == 'D') %>% pull(reply_atall)
Y0D <- filter(df, treat_deshawn == 0, leg_party == 'D') %>% pull(reply_atall)

mean(Y1D) - mean(Y0D)
```

```
## [1] 0.015963
```

$$\mathrm{E}[Y_i(1) - Y_i(0) | \text{party} = R]$$

```
Y1R <- filter(df, treat_deshawn == 1, leg_party == 'R') %>% pull(reply_atall)
Y0R <- filter(df, treat_deshawn == 0, leg_party == 'R') %>% pull(reply_atall)

mean(Y1R) - mean(Y0R)
```

```
## [1] -0.05954107
```

Democrats are *more* likely to respond to an email if the sender was identified as DeShawn Jackson as compared to Jake Mueller; whereas Republicans were *less* likely to respond.