Social Science Inquiry II

Week 5: Uncertainty and inference, part I

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2022

Loading packages for this class

```
> library(ri)
> library(ggplot2)
> set.seed(60637)
```

Example adapted from:

Chattopadhyay, Raghabendra & Duflo, Esther (2004). Women as policy makers: Evidence from a randomized policy experiment in India.

as discussed in

Gerber, Alan S. & Green, Donald P. (2012). Field experiments: Design, analysis, and interpretation.

Chattopadhyay & Duflo example

- Population: 7 villages
- Treatment: $D = 1$ if female-headed council, $D = 0$ if male
- Outcome: Budget allocation to sanitation

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

$$\text{ATE} = \mathrm{E}\left[\tau_i\right] = 5$$

Chattopadhyay & Duflo example

What we actually see:

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

To produce an *estimate* of the ATE, we can compare people who received treatment 1 to people who received treatment 0.

$$\bar{Y}(1) = 22.5 \qquad \bar{Y}(0) = 16$$

$$\hat{\tau} = 22.5 - 16 = \mathbf{6.5}$$

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 τ for the ATE, μ for the mean, or σ for the standard deviation.

Estimator

An *estimator* is a function of the data we observe; it is a statistic that gives us an informed guess about the value of the estimand.

Below, the estimator is the function $g(\cdot)$.

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

We can also think of it as a recipe. Given some data, $X_1, \ldots, X_n$, follow the instructions $g(\cdot)$ to produce an **estimate**.

Estimate

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

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

estimand

estimate

Ingredients

- 150g unsalted butter, plus extra for greasing
- 150g plain chocolate, broken into pieces
- 150g plain flour
- ½ tsp baking powder
- ½ tsp bicarbonate of soda
- 200g light muscovado sugar
- 2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.

2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

estimator

(image cred: simongrund89)

Difference-in-means

Proposed estimator for $\mathrm{E}[\tau_i]$: compare people who received treatment 1 to people who received treatment 0.

This is the difference in means estimator:

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

Chattopadhyay & Duflo example

- We have produced an estimate here, by taking the difference in means.
- But what can we say about our uncertainty about this estimate? The number of observations in each group is pretty small.
- Is the estimate meaningfully different from zero?
- How likely would we be to see an effect this size just by chance?

Chattopadhyay & Duflo example

- One way to think about this is to assume the individual treatment effect for all individuals is exactly zero. Then, no matter how we randomized treatment assignment, we would see the same Y_is.
- We can then say how often we would see a treatment effect estimate of this size just by chance, under the assumption that individual treatment effects were actually zero.

Sharp null hypothesis of no effect

Assuming all treatment effects are exactly zero is called the **sharp null hypothesis of no effect.** Also referred to as "Fisher's null" after Sir Ronald Aylmer Fisher (1890-1962).

Sharp null hypothesis of no effect

We might write the sharp null hypothesis this way:

$$H_0 : \tau_i = 0, \text{ for all } i \text{ in our pop.}$$

This implies that potential outcomes are identical under treatment and control, for all individuals.

$$Y_i(0) = Y_i(1)$$

Sharp null hypothesis of no effect

- To test this hypothesis, we also need to know (or assume we know) the randomization procedure
- Here, we'll assume exactly two villages are assigned treatment.

Chattopadhyay & Duflo example

Then we can fill in this table...

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

Chattopadhyay & Duflo example

. . . as this:

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

Gerber Green example

We can re-run the randomization, and the potential outcomes and observed Y_i will not change, but the treatment effect estimate will.

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

$$\bar{Y}(1) = 15 \qquad \bar{Y}(0) = 19$$

$$\hat{\tau} = 15 - 19 = \mathbf{-4}$$

Gerber Green example

We can re-run the randomization, and the potential outcomes and observed Y_i will not change, but the treatment effect estimate will.

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

$$\bar{Y}(1) = 20 \qquad \bar{Y}(0) = 17$$

$$\hat{\tau} = 20 - 17 = \mathbf{3}$$

Gerber Green example

We can re-run the randomization, and the potential outcomes and observed Y_i will not change, but the treatment effect estimate will.

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

$$\bar{Y}(1) = 25 \qquad \bar{Y}(0) = 15$$

$$\hat{\tau} = 25 - 15 = \mathbf{10}$$

- Because we know how treatment was assigned, we know all the possible ways treatment could be assigned across the villages, and the exact probability.
- There are seven villages, and we say that exactly two will get treatment. Each village is assigned treatment with equal probability.

We can use the package `ri` to find the *exact* distribution of $\hat{\tau}_{DM}$ under the sharp null.

Our real data:

```
> df <- data.frame(
+         # our initial treatment vector
+         D = c(1, 0, 0, 0, 0, 0, 1),
+         # our initial response vector
+         Y = c(15, 15, 20, 20, 10, 15, 30),
+         # treatment assignment probability
+         probs = rep(2/7, 7)
+ )
> df
```

```
  D  Y     probs
1 1 15 0.2857143
2 0 15 0.2857143
3 0 20 0.2857143
4 0 20 0.2857143
5 0 10 0.2857143
6 0 15 0.2857143
7 1 30 0.2857143
```

And our difference in means estimate of the average treatment effect under the real data:

```
> Y1 <- df$Y[which(df$D == 1)]
> Y0 <- df$Y[which(df$D == 0)]
> (dm_hat <- mean(Y1) - mean(Y0))
```

```
[1] 6.5
```

Adding in the hypothetical data.

```
> df <-  cbind( # binds the columns together
+          df,
+          # Y(0) under the sharp null of no effect
+          Y0 = df$Y,
+          # Y(1) under the sharp null of no effect
+          Y1 = df$Y)
> df
```

```
  D  Y     probs Y0 Y1
1 1 15 0.2857143 15 15
2 0 15 0.2857143 15 15
3 0 20 0.2857143 20 20
4 0 20 0.2857143 20 20
5 0 10 0.2857143 10 10
6 0 15 0.2857143 15 15
7 1 30 0.2857143 30 30
```

Consider all the ways treatment could be assigned: (this is what `ri` is doing for us)

```
> (perms <- genperms(df$D))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	1	1	1	1	0	0	0
3	0	1	0	0	0	0	1	0	0	0	0	1	1	1
4	0	0	1	0	0	0	0	1	0	0	0	1	0	0
5	0	0	0	1	0	0	0	0	1	0	0	0	1	0
6	0	0	0	0	1	0	0	0	0	1	0	0	0	1
7	0	0	0	0	0	1	0	0	0	0	1	0	0	0

	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0
4	0	1	1	1	0	0	0
5	0	1	0	0	1	1	0
6	0	0	1	0	1	0	1
7	1	0	0	1	0	1	1

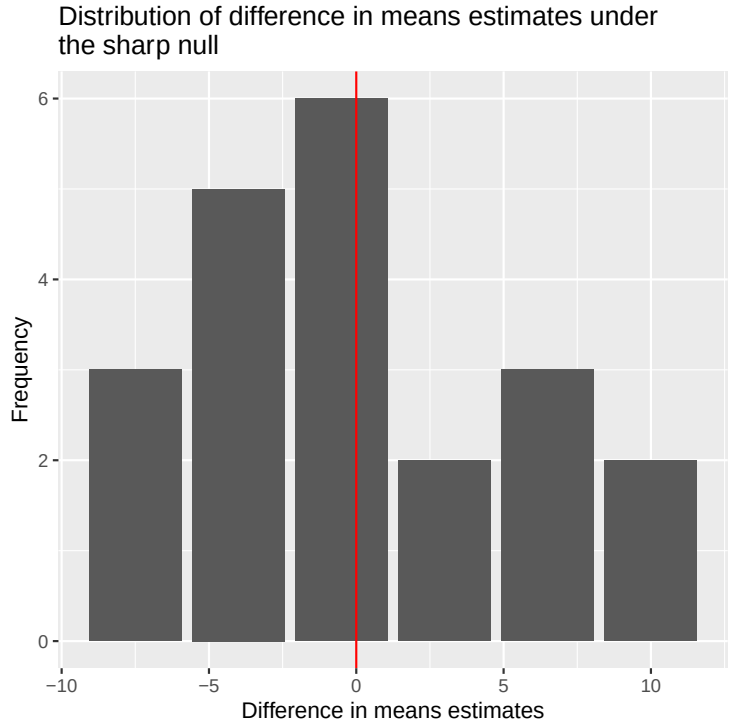
Then generate the sampling distribution of the ATE estimate under the sharp null of no effect.

```
> Ys_null <- list(
+           Y0 = df$Y0,
+           Y1 = df$Y1
+ )
> dm <- gendist(Ys_null,
+                 perms,
+                 prob=df$probs)
> dm
```

```
 [1] -4.0 -0.5 -0.5 -7.5 -4.0  6.5 -0.5 -0.5 -7.5 -4.0  6.5
[16] -4.0 -0.5 10.0 -7.5  3.0  6.5
```

The mean under our null distribution is *exactly* zero. Why?

```
> mean(dm)
[1] 0
```

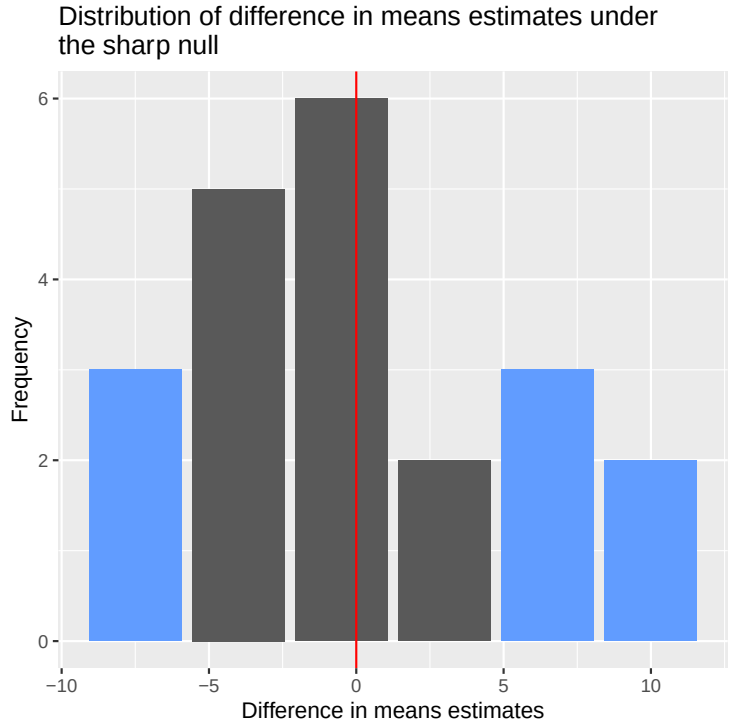



```
> prop.table(table(dm))

dm
     -7.5        -4      -0.5         3       6.5        10
0.1428571 0.2380952 0.2857143 0.0952381 0.1428571 0.0952381

> (pval <- mean(abs(dm) >= dm_hat))

[1] 0.3809524
```




```
> prop.table(table(dm))

dm
     -7.5        -4      -0.5         3       6.5        10
0.1428571 0.2380952 0.2857143 0.0952381 0.1428571 0.0952381

> (pval <- mean(abs(dm) >= dm_hat))

[1] 0.3809524
```

p-value

- Under the null distribution, if we were to re-randomize the experiment many times, we would see a value *at least as extreme as our estimate* 38.1% of the time.
- That doesn't seem **that** unlikely–more than one in three times, we would see an estimate as large as we got, just by chance.

`ri` produces the same exact p-value.

```
> dispdist(distout = dm,
+          ate = dm_hat,
+          display.plot = FALSE)$two.tailed.p.value.abs
```

```
[1] 0.3809524
```

That's because it is doing the exact same thing under the hood.

- Why is this test called an "exact" test?
- Because we know the *exact* distribution of our estimate under the specified null. We do not have to approximate the distribution.
- This will not be true for all of our hypothesis testing...

Hypotheses

- We framed our null hypothesis as below:

$$H_0 : \tau_i = 0, \text{ for all } i \text{ in our pop.}$$

- Implicitly, the alternative is that for some individual(s), the treatment effect is non-zero.

$$H_A : \tau_i \neq 0, \text{ for some } i \text{ in our pop.}$$

- In our case, we did not find strong evidence to reject the null hypothesis, i.e., our data is consistent with what we would see if the null hypothesis were true.
- Note that we do NOT say that we reject or accept the alternative hypothesis.
- We can only say that our results were not consistent with or were consistent with what we would have seen under the null–i.e., we have evidence to reject or fail to reject the null.

Distributions of estimators

- Our difference in means estimator is a function of the data we observe.
- Because there is randomness in the data, here, due to random assignment of treatment, the estimator is also a random variable.
- Just like other random variables have distributions to describe them, estimators also have distributions.
- We don't know the *true* distribution of the difference in means estimator, for the same reason that we don't know individual treatment effects.(FPoCI!)
- But we DO know what the distribution would be under the null.

Note that we are conducting inference with respect to:

- a defined population
- a defined treatment
- a defined outcome
- a known treatment assignment mechanism
- a given estimating procedure

The null

- How do we determine what the null is?
- We formalize our hypotheses in terms of the effect we are trying to find in the data. Is there a treatment effect? Is there a difference between these two groups?
- The null is (often, but not always) the case when there is no effect, or no difference.
- We can imagine other kinds of hypotheses, for example that effects are bounded away from zero and positive, or exactly .2. And we can characterize the distribution of our test statistic under the null.

The null

- Is the sharp null of no individual level effects plausible in this setting?
- Does it matter?

Getting to some real data...

Butler, Daniel M., & Broockman, David 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`

Loading the data

```
> df <- read.csv('../data/butler-broockman.csv', as.is = TRUE)
> head(df)
```

	leg_party	leg_republican	leg_black	leg_latino	reply_atall	treat_deshawn
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

	treat_demprimary	treat_repprimary	treat_noprimary	treat_group	treat_jake
1	1	0	0	5	1
2	0	1	0	2	0
3	0	0	1	6	0
4	0	0	1	6	0
5	0	0	1	0	1
6	0	1	0	2	0

	leg_notwhite	leg_white	leg_notblackotherminority	treat_primary
1	0	1	0	1
2	0	1	0	1
3	0	1	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	1

Examining the data

```
> str(df)

'data.frame':        4859 obs. of  15 variables:
 $ leg_party                : chr  "R" "D" "R" "R" ...
 $ leg_republican           : int  1 0 1 1 0 0 1 1 1 1 ...
 $ leg_black                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ leg_latino               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ reply_atall              : int  1 1 0 0 0 1 1 1 1 1 ...
 $ treat_deshawn            : int  0 1 1 1 0 1 1 1 0 1 ...
 $ treat_demprimary         : int  1 0 0 0 0 0 1 0 1 1 ...
 $ treat_repprimary         : int  0 1 0 0 0 1 0 1 0 0 ...
 $ treat_noprimary          : int  0 0 1 1 1 0 0 0 0 0 ...
 $ treat_group              : int  5 2 6 6 0 2 4 2 5 4 ...
 $ treat_jake               : int  1 0 0 0 1 0 0 0 1 0 ...
 $ leg_notwhite             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ leg_white                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ leg_notblackotherminority: int  0 0 0 0 0 0 0 0 0 0 ...
 $ treat_primary            : int  1 1 0 0 0 1 1 1 1 1 ...
```

Data

- Where does it come from/how is it generated?
- What is the sample population?
- What is being measured?

Recall that 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
```

We're going to manipulate our data so it takes the format $Y \sim D$.

```
> df$D <- df$treat_deshawn
> df$Y <- df$reply_atall
```

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

```
> Y1 <- df$Y[which(df$D == 1)]
> Y0 <- df$Y[which(df$D == 0)]

> (dm_hat <- 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.

Note that again, the population that we're taking inference over is legislators–all of whom are included in our experiment. We're not assuming we're sampling from some other distribution. The only source of randomness is how treatment is assigned.

- There are $\binom{4859}{2428}$ different ways treatment could be assigned–this is too many to generate the whole matrix of permutations and get the *exact* sampling distribution.
- Instead, we'll simulate the sampling process many times to find the *approximate* sampling distribution of $\hat{\tau}_{DM}$ under the sharp null.

```
> # randomization inference function
> my_ri <- function(df){
+         df_ri <- df
+         df_ri$newD <- sample(df$D)
+         Y1_ri <- df$Y[which(df_ri$newD == 1)]
+         Y0_ri <- df$Y[which(df_ri$newD == 0)]
+         ate_hat <- mean(Y1_ri)-mean(Y0_ri)
+   return(ate_hat)
+ }
>
```

Let's try it.

```
> my_ri(df)
```

```
[1] -0.007122444
```

And again.

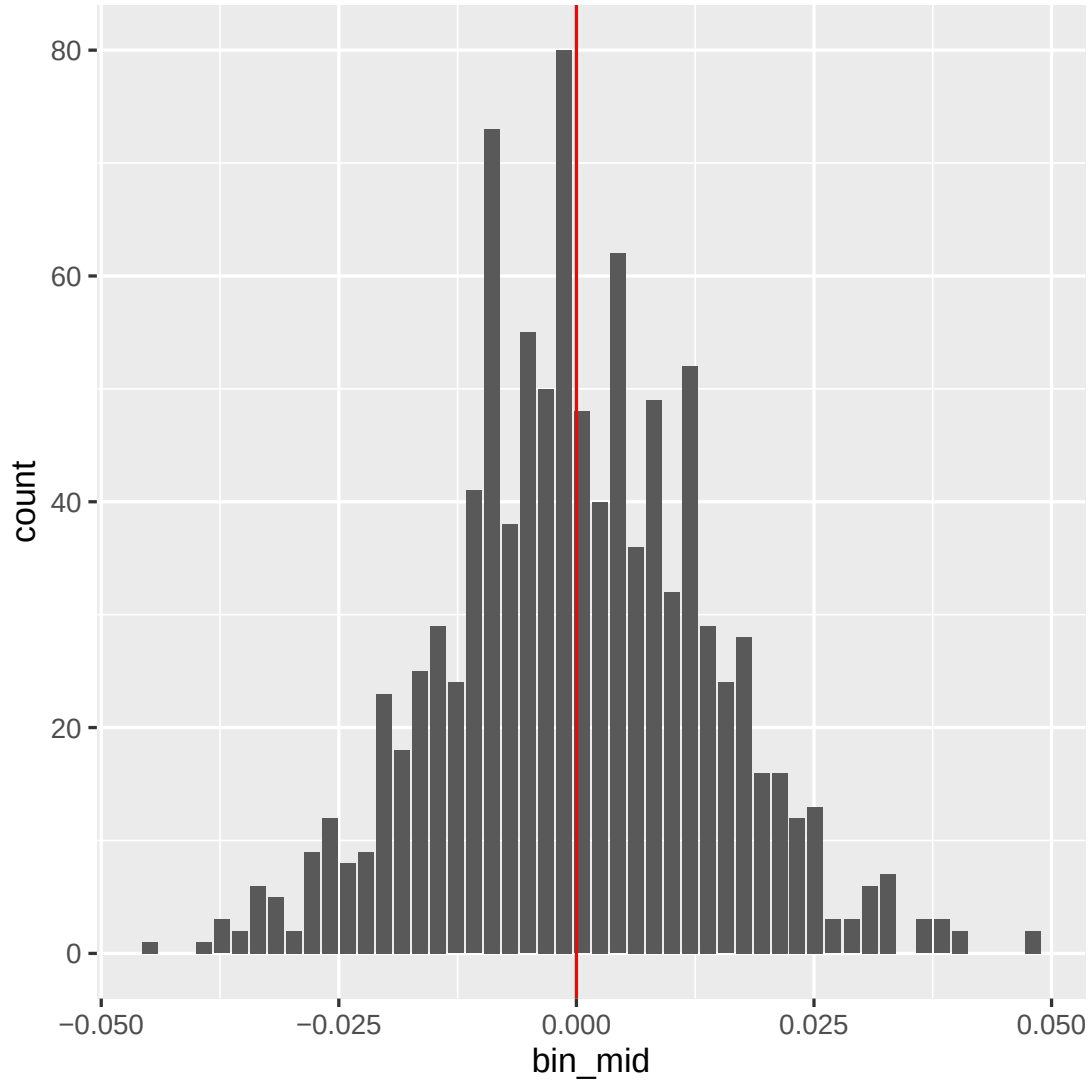
```
> my_ri(df)
```

```
[1] -0.01370816
```

We can do this many times to find the distribution of $\hat{\tau}_{DM}$ under the sharp null.

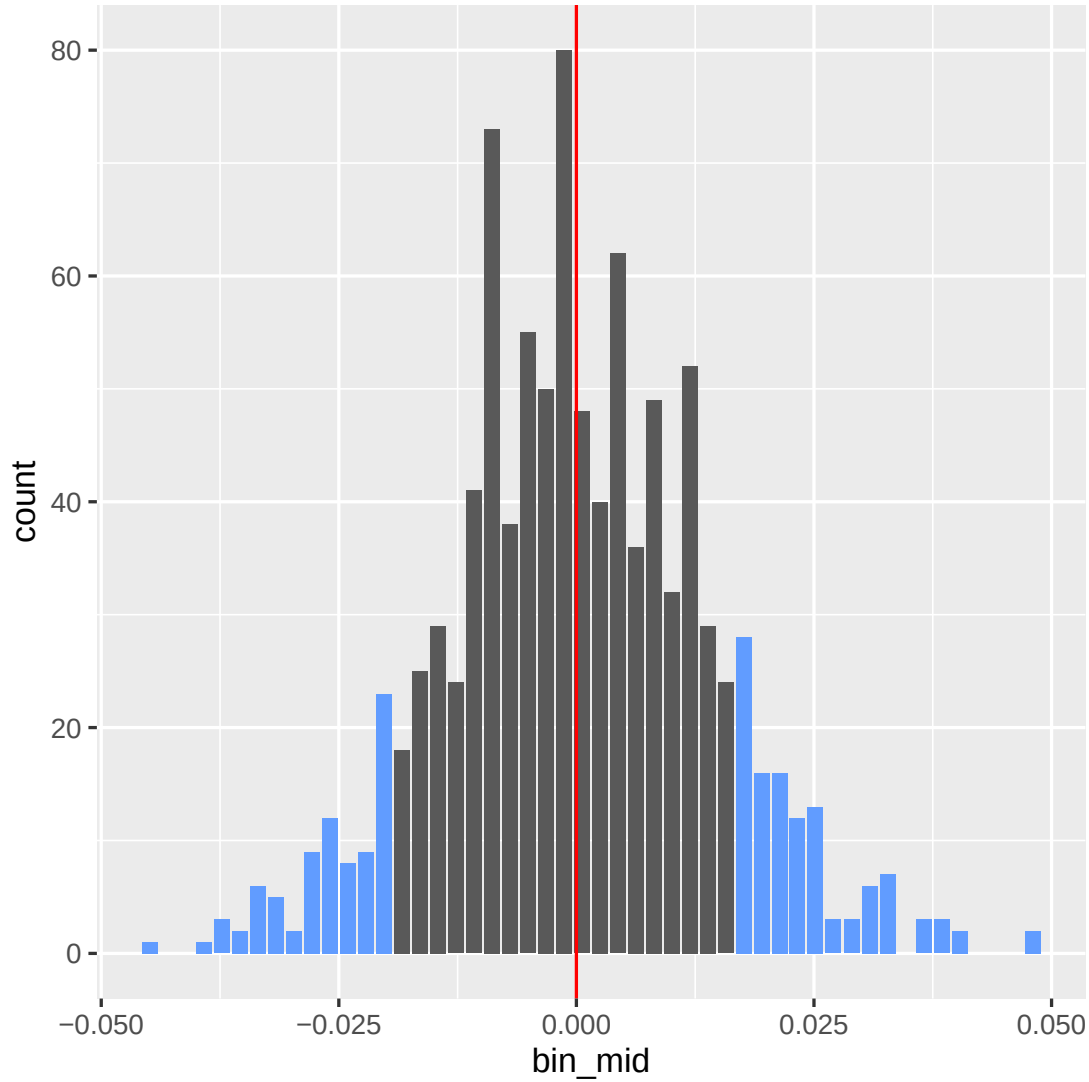
```
> # number of iterations
> n_iter <- 1000
> # replicate does the same (random) function many times
> dm <- replicate(n = n_iter, my_ri(df))
> head(dm)

[1]  0.009341855 -0.009592089  0.028275799 -0.002183155  0.005225780
[6] -0.003006370
```




```
> (pval <- mean(abs(dm) >= abs(dm_hat)))
```

```
[1] 0.193
```




```
> (pval <- mean(abs(dm) >= abs(dm_hat)))
```

```
[1] 0.193
```

The types of hypotheses we've considered are two-sided hypotheses: we look at effects in either direction from zero.

$$H_0 : \tau_i = 0, \text{ for all } i \text{ in our pop.}$$

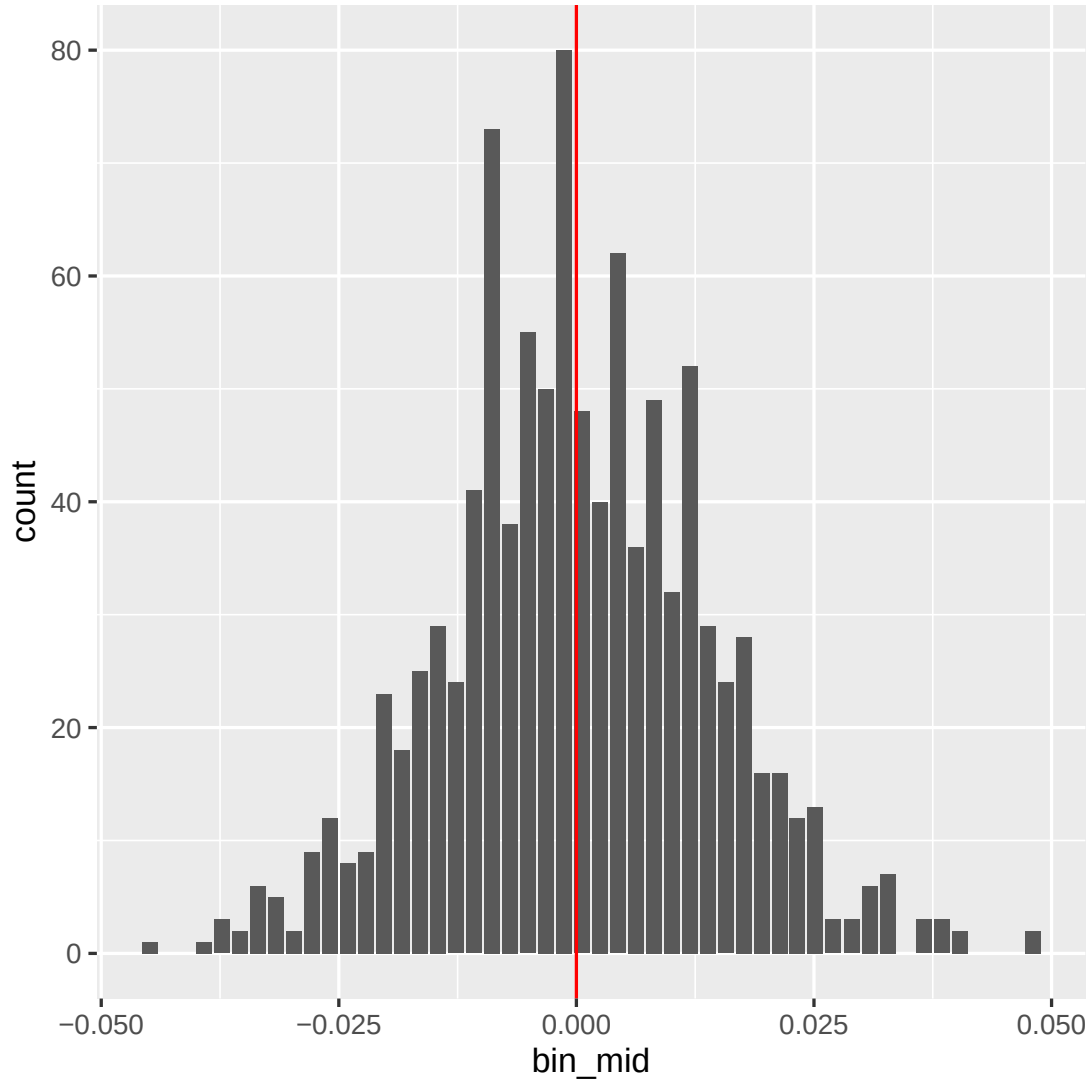
$$H_A : \tau_i \neq 0, \text{ for some } i \text{ in our pop.}$$

We can also consider other alternative hypotheses. For example, the hypothesis that treatment effects are less than zero. This is called a one-sided hypothesis.

$$H_0 : \tau_i = 0, \text{ for all } i \text{ in our pop.}$$

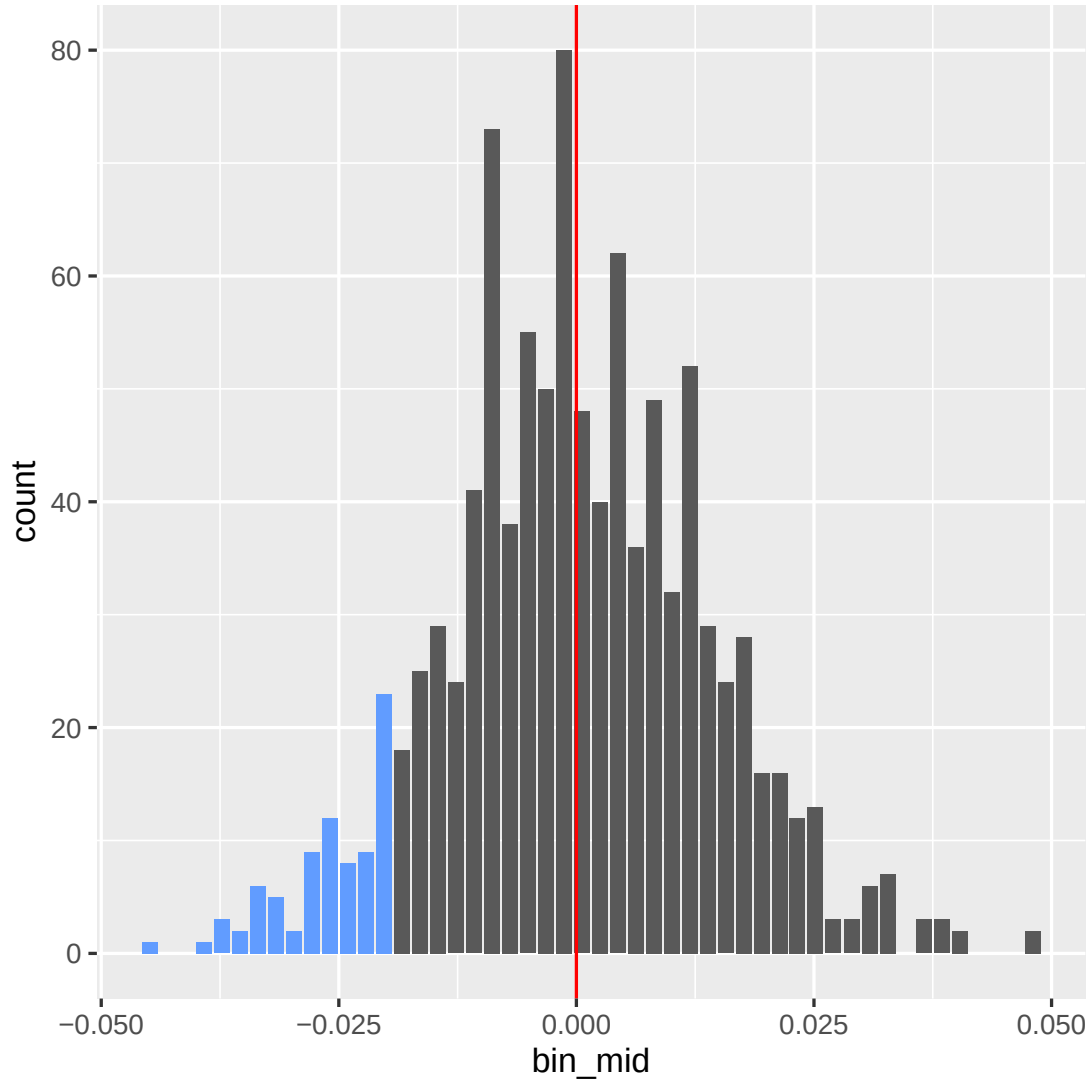
$$H_A : \tau_i < 0, \text{ for some } i \text{ in our pop.}$$

- The alternative hypothesis that we consider is a consequence of the social science theory we're trying to test.
- Here, we want to see if legislators are *less likely* to respond to a constituent named DeShawn Jackson, as compared to Jake Mueller.
- When we test a one-sided hypothesis, we want to check how likely we would be to observe statistics at least as large as the test statistic that we actually observe, in the direction of our hypothesis.




```
> (pval <- mean(dm <= dm_hat))

[1] 0.099
```




```
> (pval <- mean(dm <= dm_hat))

[1] 0.099
```

P-values

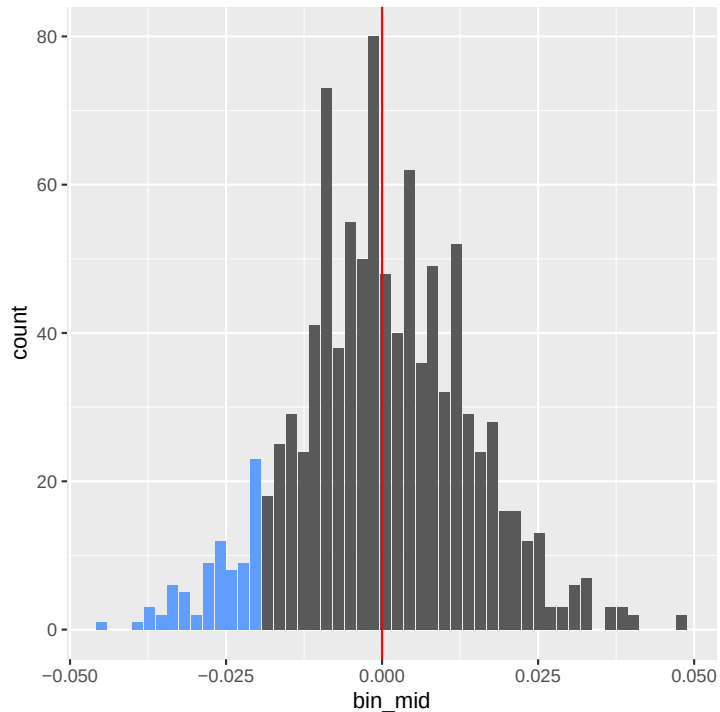
Suppose $\hat{\theta}$ is the general form for an estimate produced by our estimator, and $\hat{\theta}^*$ is the value we have actually observed.

P-values

- A lower one-tailed p-value under the null hypothesis is

$$p = ¶_0[\hat{\theta} \leq \hat{\theta}^*]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ that is less than or equal to what we saw from the data.

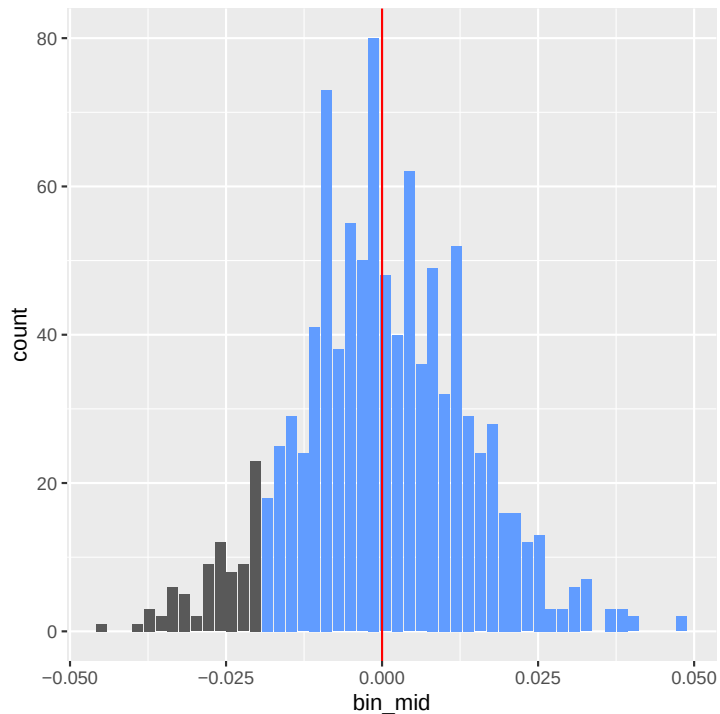


P-values

- An upper one-tailed p-value under the null hypothesis is

$$p = \P_0[\hat{\theta} \geq \hat{\theta}^*]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ that is greater than or equal to what we saw from the data.

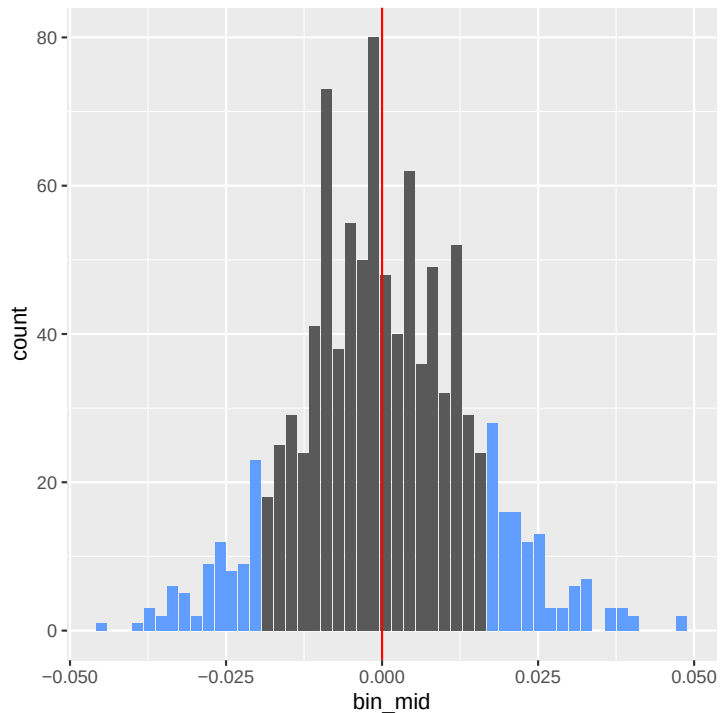


P-values

- A two-tailed p-value under the null hypothesis is

$$p = ¶_0[|\hat{\theta}| \geq |\hat{\theta}^*|]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ as or more extreme as what we saw from the data.



References I

Butler, D. M. and Broockman, D. E. (2011). Do politicians racially discriminate against constituents? a field experiment on state legislators. American Journal of Political Science, 55(3):463–477.

Chattopadhyay, R. and Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in india. Econometrica, 72(5):1409–1443.

Gerber, A. S. and Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. WW Norton.