Hypothesis Testing

Gareth Nellis, UC San Diego

June, 2019

Big picture

- We are interested in an intervention
- Our goal is to understand its *causal effect*
- We formulate a hypothesis about that effect; we test it; then say something about whether or not it might be true in light of the evidence
- Proceduce amenable to accurate quantification

Roadmap

Three worlds:

- 1st-best world: we know people's outcomes under treatment and control
- 2nd-best world: we get to run (infinitely) many experiments
- 3rd-best world—ours: run AN experiment

Two problems:

- Fundamenal problem of causal inference
- We generally only get to run an experiment once

Three key concepts:

- Potential outcomes
- Sampling distribution
- P-value

Section 1

First-best world

First-best world

Imagine:

1. We have a drug for curing bad eyesight
2. We have a set of patients
3. For each patient, we can measure how good or bad is their eyesight on a scale of 1 (bad) to 5 (good)
4. Say we know, for every patient, two things
   i. What would be their eyesight if they **did** get the drug (treatment)
   ii. What would be their eyesight if they **did not** get the drug (control)

These are called patients' *potential outcomes*

Let's see this in action

name	no_drug	drug
Abiola	2	4
Aga	1	2
Brice	5	5
Kamala	3	5
Edris	3	4
Ines	2	5
Lucy	2	2
Oscar	2	1
Rita	1	5
Tess	4	4

Two questions for you:

1. What is the effect of the drug for each individual?
2. What is the *average* effect of the drug for people in this group?

Calculating the effects of the drug for each person

name	no_drug	drug	difference
Abiola	2	4	??
Aga	1	2	??
Brice	5	5	??
Kamala	3	5	??
Edris	3	4	??
Ines	2	5	??
Lucy	2	2	??
Oscar	2	1	??
Rita	1	5	??
Tess	4	4	??

Calculating the effects of the drug for each person

```
POs$difference <- POs$drug - POs$no_drug
```

name	no_drug	drug	difference
Abiola	2	4	2
Aga	1	2	1
Brice	5	5	0
Kamala	3	5	2
Edris	3	4	1
Ines	2	5	3
Lucy	2	2	0
Oscar	2	1	-1
Rita	1	5	4
Tess	4	4	0

Calculating the *average* effect of the drug: method 1

name	no_drug	drug	difference
Abiola	2	4	2
Aga	1	2	1
Brice	5	5	0
Kamala	3	5	2
Edris	3	4	1
Ines	2	5	3
Lucy	2	2	0
Oscar	2	1	-1
Rita	1	5	4
Tess	4	4	0

```
POs$difference
```

```
[1] 2 1 0 2 1 3 0 -1 4 0
```

```
length(POs$difference)
```

```
[1] 10
```

Calculating the *average* effect of the drug: method 1

name	no_drug	drug	difference
Abiola	2	4	2
Aga	1	2	1
Brice	5	5	0
Kamala	3	5	2
Edris	3	4	1
Ines	2	5	3
Lucy	2	2	0
Oscar	2	1	-1
Rita	1	5	4
Tess	4	4	0

```
mean(POs$difference)
```

```
[1] 1.2
```

Calculating the average effect of the drug: method 2

name	no_drug	drug	difference
Abiola	2	4	2
Aga	1	2	1
Brice	5	5	0
Kamala	3	5	2
Edris	3	4	1
Ines	2	5	3
Lucy	2	2	0
Oscar	2	1	-1
Rita	1	5	4
Tess	4	4	0

```
mean(POs$drug)
```

```
[1] 3.7
```

```
mean(POs$no_drug)
```

```
[1] 2.5
```

Calculating the average effect of the drug: method 2

name	no_drug	drug	difference
Abiola	2	4	2
Aga	1	2	1
Brice	5	5	0
Kamala	3	5	2
Edris	3	4	1
Ines	2	5	3
Lucy	2	2	0
Oscar	2	1	-1
Rita	1	5	4
Tess	4	4	0

```
mean(POs$drug) - mean(POs$no_drug)
```

```
[1] 1.2
```

Calculating the average effect of the drug: method 2

To recap:

```
mean(POs$difference)
```

[1] 1.2

```
mean(POs$drug) - mean(POs$no_drug)
```

[1] 1.2

They're the same!

Big insight

- The average individual-level treatment effect is equal to the difference in average outcomes under treatment and control
- **Impossible**: knowing the unit-level treatment effects; why?
- **Possible**: estimating the averages

Next section: how we estimate those averages

Section 2

Second-best world: lots of experiments

The problem: what we can actually observe

- Unit-level treatment effects are unknowable
- Intiution: you either do or don't get the drug; you can't both have it and not have it *at the same time*
- This is called the **fundamental problem of causal inference**
- We can't observe the counterfactual
- We're missing lots of data

Workaround: random sampling/assignment

- We can't measure the unti-level differences—we just don't have enough data
- But we're **good** at reliably estimating averages without all the data
- Central to this is **random sampling**

Illustration: miracle of random sampling

Taking average of a random sample of a population many times gets us VERY CLOSE to the true average

Let's focus on potential outcomes under treatment ("drug"). Recall:

name	no_drug	drug
Abiola	2	4
Aga	1	2
Brice	5	5
Kamala	3	5
Edris	3	4
Ines	2	5
Lucy	2	2
Oscar	2	1
Rita	1	5
Tess	4	4

```
mean(POs$drug)
```

```
[1] 3.7
```

Question

Can we recover this average without knowing the potential outcome for every person?

Illustration: miracle of random sampling (take 1)

Computer picks half of you at random:

name	drug	picked
Abiola		Not picked
Aga		Not picked
Brice	5	Picked
Kamala	5	Picked
Edris		Not picked
Ines		Not picked
Lucy	2	Picked
Oscar	1	Picked
Rita		Not picked
Tess	4	Picked

```
mean(RS1$drug, na.rm = TRUE)
```

```
[1] 3.4
```

Illustration: miracle of random sampling (take 2)

- Let's see; computer will pick half of you at random

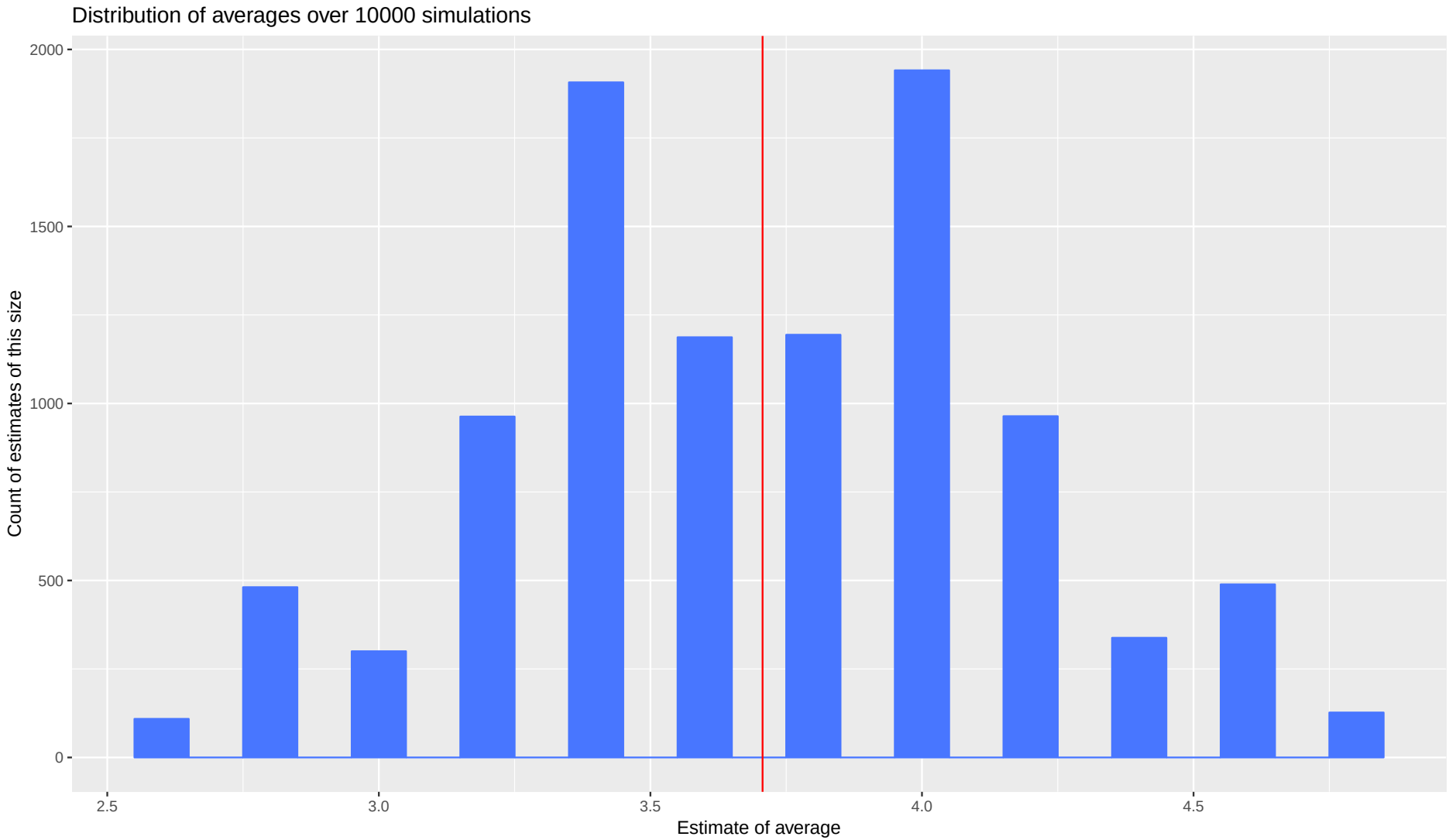
name	drug	picked
Abiola		Not picked
Aga		Not picked
Brice		Not picked
Kamala	5	Picked
Edris	4	Picked
Ines	5	Picked
Lucy		Not picked
Oscar		Not picked
Rita	5	Picked
Tess	4	Picked

```
mean(RS1$drug, na.rm = TRUE)
```

```
[1] 4.6
```

Illustration: the miracle of random sampling

What if we did it lots of times?



Confirm

Here's the average of the 10000 averages:

```
mean(averages)
```

[1] 3.7065

Here's the "real" average:

```
mean(POs$drug)
```

[1] 3.7

Big idea

- We can estimate the true average without having all the data points, by taking a random sample
- We're going to take this idea and run with it; it's the key to everything that follows
- Next step: sampling distributions of estimated average treatment effects

Random assignment

Remember: we never get to see both potential outcomes; we only get to see one of them

Random assignment

- Imagine you're a doctor and you wanted to test the effectiveness of the drug
- Let's do a random assignment, where half of you get the drug; for those who get the drug, we observe your potential outcome with the drug; for those who don't get the drug, we observe your potential outcome without the drug

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga		1	no_drug
Brice	5		drug
Kamala	5		drug
Edris	4		drug
Ines		2	no_drug
Lucy	2		drug
Oscar		2	no_drug
Rita		1	no_drug
Tess	4		drug

Estimate the average treatment effect of the drug in this assignment

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga		1	no_drug
Brice	5		drug
Kamala	5		drug
Edris	4		drug
Ines		2	no_drug
Lucy	2		drug
Oscar		2	no_drug
Rita		1	no_drug
Tess	4		drug

```
mean(RA.sim$drug, na.rm = TRUE) -
  mean(RA.sim$no_drug, na.rm = TRUE)
```

[1] 2.4

Let's do it again for a different assignment

name	drug	no_drug	treatment_status
Abiola	4		drug
Aga		1	no_drug
Brice	5		drug
Kamala	5		drug
Edris		3	no_drug
Ines	5		drug
Lucy		2	no_drug
Oscar		2	no_drug
Rita	5		drug
Tess		4	no_drug

```
mean(RA.sim$drug, na.rm = TRUE) -
  mean(RA.sim$no_drug, na.rm = TRUE)
```

```
[1] 2.4
```

And another one...

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga	2		drug
Brice		5	no_drug
Kamala	5		drug
Edris		3	no_drug
Ines		2	no_drug
Lucy		2	no_drug
Oscar	1		drug
Rita	5		drug
Tess	4		drug

```
mean(RA.sim$drug, na.rm = TRUE) -
  mean(RA.sim$no_drug, na.rm = TRUE)
```

```
[1] 0.6
```

And another one...

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga	2		drug
Brice		5	no_drug
Kamala		3	no_drug
Edris		3	no_drug
Ines	5		drug
Lucy	2		drug
Oscar	1		drug
Rita		1	no_drug
Tess	4		drug

```
mean(RA.sim$drug, na.rm = TRUE) -
  mean(RA.sim$no_drug, na.rm = TRUE)
```

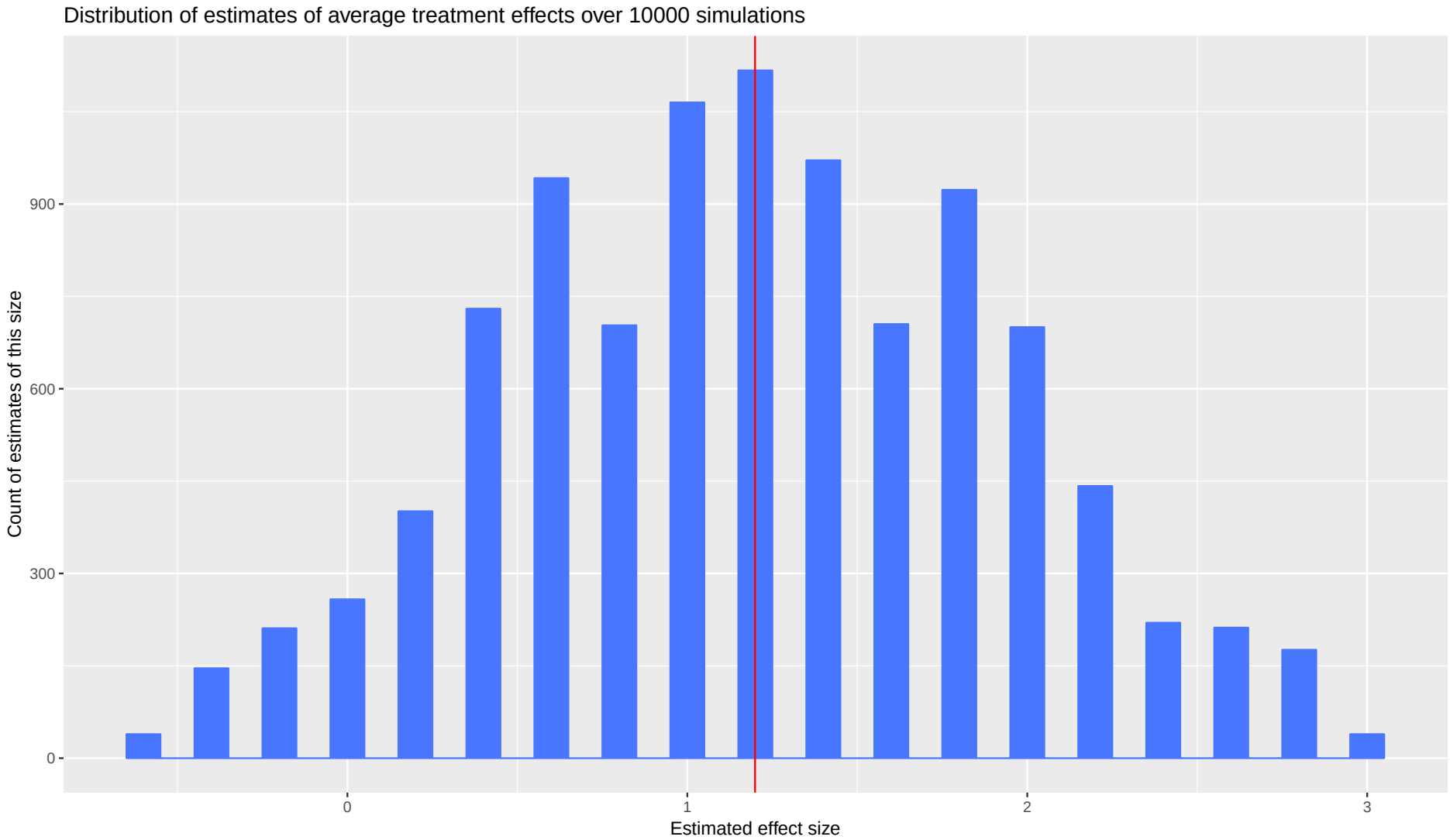
```
[1] 0
```

Sampling distributions

- We'll do this 10000 times[1]
- Each time we'll randomly assign half of you to "drug" and half to "no_drug"
- We'll calculate the average outcomes for those in both groups, then take the difference of those means

[1] For such a small sample, this isn't necessary; but nice for illustration

Sampling distributions



Sampling distribution

Average of the 10000 estimates of the treatment effect

```
mean(sim.averages)
```

[1] 1.19914

REAL treatment effect:

```
mean(POs$difference)
```

[1] 1.2

- Wowza! We're super close!

Lessons

- In a world where we could run 10000 experiments, we could do a pretty excellent job of estimating the true effect of the drug
- Sometimes too high, sometimes too low, but on average almost dead on
- Why don't we just run 0000s of experiments?

Section 3

Third-best world / our world

Estimation vs hypothesis testing

- We **could** run one experiment and say that's our best bet of what the effect is, and leave it at that
- Problem? A lot of the time we'll be quite far off in an unknown direction
- We usually take a more conservative approach: hypothesis testing

Hypotheses: an overview

What are hypotheses?

Hypotheses: an overview

Statements of the about the world that you seek to *reject*

Good hypotheses:

- They are possibly TRUE or FALSE
- They are *falsifiable*
- They are statements about the world, not your analysis
- They are simple
- They involve clear concepts
- They are **few**
- They are **contested**: you are not sure if they are true or false, & therefore you will learn something from the experiment

Some hypotheses

- Education is very important
- Education increases your income
- Education either increases, decreases, or has no effect on your income
- Education is good for you because it strengthens your character in very fundamental ways that you could never measure
- Just one of these is a good hypothesis. Which one?

Now back to hypothesis testing...

Nulls hypotheses (tricky)

Because of an unusual convention, social scientists often describe hypotheses in terms of what they **expect** but then *test* the null hypothesis of no effect

eg:

- H1: Education increases income
- H-null: Education has no effect on income

Test: how likely is the data given the null

Hypothesis testing: the steps

1. We set up a null hypothesis, and **assume that it is true**
2. We generate the sampling distribution under the null
3. We gather data from a real-world experiment that is relevant to the hypothesis
4. We make a determination about the null hypothesis, based on the idea of "how likely is our data given the null hypothesis?"

Let's follow these steps with our drug trial

Set up null hypothesis

What would be a null hypothesis about the drug in our case?

1. Set up null hypothesis

H-null: the drug has no effect on eyesight

2. Run an experiment

Suppose we ran one experiment, and assigned patients to T and C as follows:

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga	2		drug
Brice		5	no_drug
Kamala	5		drug
Edris	4		drug
Ines	5		drug
Lucy	2		drug
Oscar		2	no_drug
Rita		1	no_drug
Tess		4	no_drug

Estimate the average treatment effect and store the result

2. Run an experiment

Estimate the average treatment effect and store the result:

```
mean(RA.sim$drug, na.rm = TRUE) -
  mean(RA.sim$no_drug, na.rm = TRUE)
```

```
[1] 0.8
```

3. Generate sampling distribution under the null hypothesis

Now for the trick

Suppose: if the drug **really has no effect for any individual** what values could we put in the empty cells?

name	drug	no_drug	treatment_status
Abiola		2	no_drug
Aga	2		drug
Brice		5	no_drug
Kamala	5		drug
Edris	4		drug
Ines	5		drug
Lucy	2		drug
Oscar		2	no_drug
Rita		1	no_drug
Tess		4	no_drug

3. Generate sampling distribution under the null hypothesis

Suppose: if the drug **really has no effect for any individual** what values could we put in the empty cells?

Fill in the potential outcomes under the sharp null hypothesis:

name	drug	no_drug	treatment_status
Abiola	2	2	no_drug
Aga	2	2	drug
Brice	5	5	no_drug
Kamala	5	5	drug
Edris	4	4	drug
Ines	5	5	drug
Lucy	2	2	drug
Oscar	2	2	no_drug
Rita	1	1	no_drug
Tess	4	4	no_drug

3. Generate sampling distribution under the null hypothesis

Next step: analyze the null hypothesis data under many different possible treatment assignments (preferably all of them!)

When there are 10 units, how many ways are there to assign 5 units to treatment and 5 to control?

```
choose(10, 5)  # binomial coefficient
```

[1] 252

3. Generate sampling distribution under the null hypothesis

Here are what some of the permutations look like:

name	outcome_under_null	V1	V2	V3	V4	V5	V6	V7	V8	V9
Abiola	2	1	1	1	1	1	1	1	1	1
Aga	2	1	1	1	1	1	1	1	1	1
Brice	5	1	1	1	1	1	1	1	1	1
Kamala	5	1	1	1	1	1	1	0	0	0
Edris	4	1	0	0	0	0	0	1	1	1
Ines	5	0	1	0	0	0	0	1	0	0
Lucy	2	0	0	1	0	0	0	0	1	0
Oscar	2	0	0	0	1	0	0	0	0	1
Rita	1	0	0	0	0	1	0	0	0	0
Tess	4	0	0	0	0	0	1	0	0	0

3. Generate sampling distribution under the null hypothesis

Calculate the estimated average treatment effect under sharp null for all permutations; e.g. for first permutation:

name	outcome_under_null	V1	V2	V3	V4	V5	V6	V7	V8	V9
Abiola	2	1	1	1	1	1	1	1	1	1
Aga	2	1	1	1	1	1	1	1	1	1
Brice	5	1	1	1	1	1	1	1	1	1
Kamala	5	1	1	1	1	1	1	0	0	0
Edris	4	1	0	0	0	0	0	1	1	1
Ines	5	0	1	0	0	0	0	1	0	0
Lucy	2	0	0	1	0	0	0	0	1	0
Oscar	2	0	0	0	1	0	0	0	0	1
Rita	1	0	0	0	0	1	0	0	0	0
Tess	4	0	0	0	0	0	1	0	0	0

```
mean(showperms$outcome_under_null[showperms$V1==1]) -
mean(showperms$outcome_under_null[showperms$V1==0])
```

[1] 0.8

3. Generate sampling distribution under the null hypothesis

Calculate the estimated average treatment effect under sharp null for all permutations; e.g. for second permutation:

name	outcome_under_null	V1	V2	V3	V4	V5	V6	V7	V8	V9
Abiola	2	1	1	1	1	1	1	1	1	1
Aga	2	1	1	1	1	1	1	1	1	1
Brice	5	1	1	1	1	1	1	1	1	1
Kamala	5	1	1	1	1	1	1	0	0	0
Edris	4	1	0	0	0	0	0	1	1	1
Ines	5	0	1	0	0	0	0	1	0	0
Lucy	2	0	0	1	0	0	0	0	1	0
Oscar	2	0	0	0	1	0	0	0	0	1
Rita	1	0	0	0	0	1	0	0	0	0
Tess	4	0	0	0	0	0	1	0	0	0

```
mean(showperms$outcome_under_null[showperms$V5==1]) -
mean(showperms$outcome_under_null[showperms$V5==0])
```

[1] -0.4

3. Generate sampling distribution under the null hypothesis

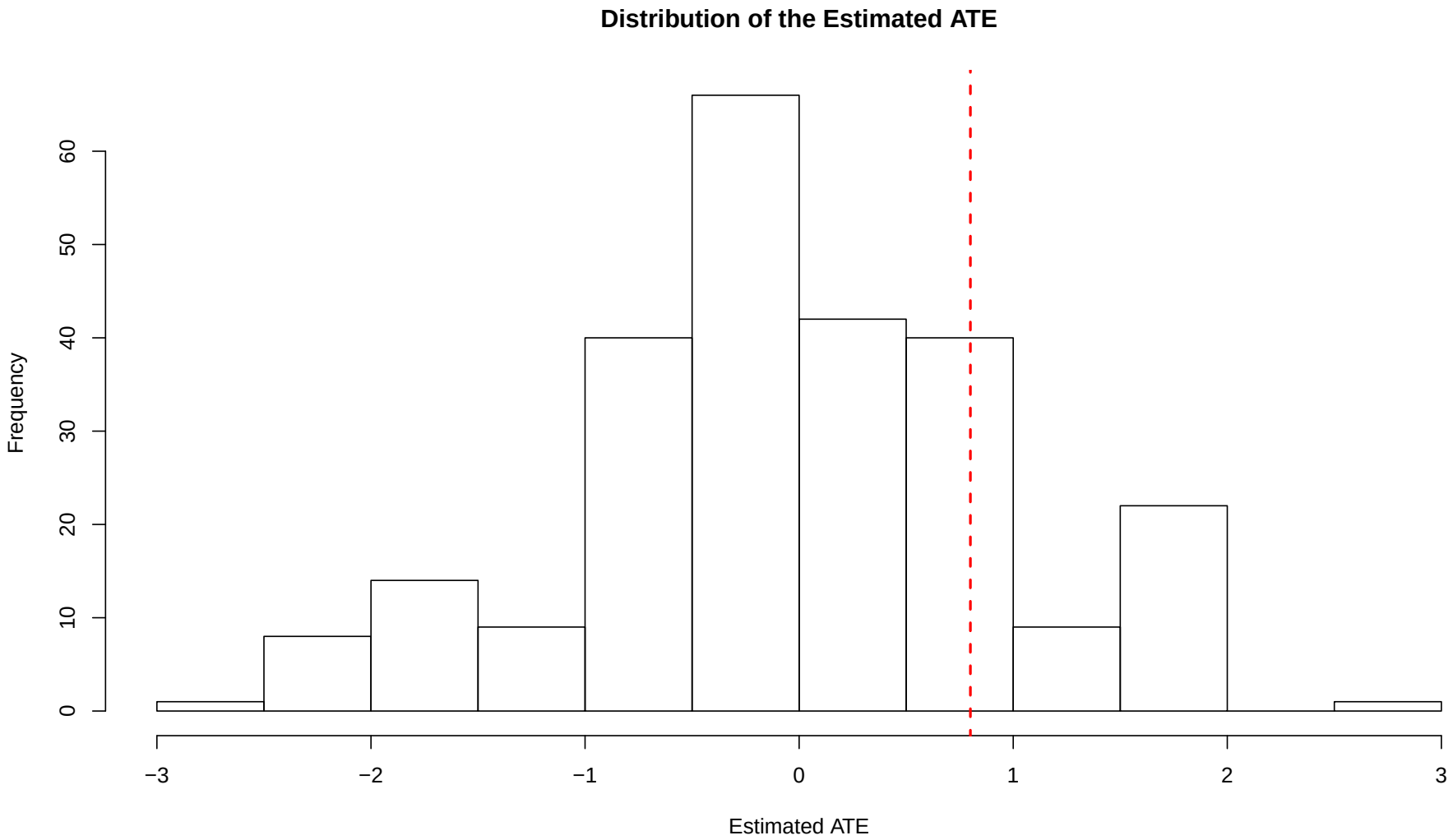
Calculate the estimated average treatment effect under sharp null for all permutations; e.g. for second permutation:

name	outcome_under_null	V1	V2	V3	V4	V5	V6	V7	V8	V9
Abiola	2	1	1	1	1	1	1	1	1	1
Aga	2	1	1	1	1	1	1	1	1	1
Brice	5	1	1	1	1	1	1	1	1	1
Kamala	5	1	1	1	1	1	1	0	0	0
Edris	4	1	0	0	0	0	0	1	1	1
Ines	5	0	1	0	0	0	0	1	0	0
Lucy	2	0	0	1	0	0	0	0	1	0
Oscar	2	0	0	0	1	0	0	0	0	1
Rita	1	0	0	0	0	1	0	0	0	0
Tess	4	0	0	0	0	0	1	0	0	0

```
mean(showperms$outcome_under_null[showperms$V9==1]) -
mean(showperms$outcome_under_null[showperms$V9==0])
```

[1] -0.4

4. Plot null distribution & see where our ACTUAL estimate falls (in red)



5. Calculate the p-value

Probability of seeing the actual data or data more extreme given that the null hypothesis is true

Simple:

1. How many estimates in the null hypothesis distribution are equal to or greater than our ACTUAL estimate?
2. How many estimates are there overall in the null distribution?
3. Divide the first number by the second number and that is our p-value

```
mean(distout>=ate) # p-value
```

[1] 0.2857143

Notes on p-value

- Low p-value: it's very unlikely that you'd see such a result in a world in which the null hypothesis is true
- High p-value: it's perfectly possible—indeed, quite likely, that such a result would be seen in a world in which the null hypothesis is true

P-values: a review

Consider this: "I estimate the treatment increased income by \$10, with a p-value of 0.05."

Which of these statements is correct (just one!):

1. The probability that treatment increased income by \$10 is just 5%
2. The probability that treatment increased income by \$10 is 95%
3. The probability that treatment increases income is 95%
4. The probability that treatment does **not** increase income is just 5%
5. The probability that we would treatment increases income is 95%
6. The probability that we would estimate an effect of \$10 if the true effect were 0 is 5%
7. The probability that we would estimate an effect of \$10 if the true effect were positive is 95%

The rejection decision

Remember: the big question is whether we can reject the null hypothesis given the data we observe; that requires a low p-value. How low is low?

The most common standard in social science is $p \leq 0.05$. That means that you reject the hypothesis if you get a p value below 0.05.

Here 0.05 is a cutoff, sometimes called the **alpha level** of the test. There are advantages and disadvantages of choosing different alpha levels.

- Say we set alpha $= 1$. This means that we will ALWAYS say that effects are significant. This has the really great advantage of guaranteeing that if there is a real effect we will always reject the null of no effect. What is the disadvantage?
- Say we set alpha $= 0$. This means that we will NEVER say that effects are significant. This has the really great advantage of guaranteeing that if really there is no effect we will not mistakenly say that the null of no effect is incorrect. What is the disadvantage?

The p-value: One sided and two sided

- p values are sometimes based on "one sided" or "two sided" tests.
- These are very similar ideas, the key difference is:
  - For a one sided test ask: What is the probability that you would get such a large estimate if there were no true effect. e.g. If I estimate "5" then: what is the probability that I would get 5 or larger by chance?"
  - For a two sided test ask: What is the probability that you would get such a large estimate in absolute magnitude if there were no true effect. e.g. If I estimate "5" then: what is the probability that I would get 5 or larger OR -5 or smaller by chance?"

Addenda

- The method I've shown for calculating p-values is called randomization inference or Fisher's exact test
- It requires very minimal assumptions
- Usually, though, we make assumptions about the shape of the data that allow us to analyze experiments using t-tests and regression
- But the core ideas about hypothesis testing are the same

Round up

- Formal hypothesis testing enormously valuable when combined with random assignment
- Allows us to accurately characterize likelihood that a null hypothesis is true or false given some generated data
- It's at the heart of analyzing experiments

Thank you!

gnellis@ucsd.edu

Credits

- Several slides borrowed from EGAP Learning Days training materials (Abu Dhabi & Malawi)