Experiments and Causality

Alex Hughes

Fall 2025

Contents

Live Session

This is the live session work space for the course. Our goal with this repository, is that we're able to communicate *ahead of time* our aims for each week, and that you can prepare accordingly.

Figure 1: Coffee Sorting Experiment

Course Introduction

Every research, product, or policy question that is interesting is actually a causal question.

- Do messages that are targeted to low-propensity voters cause them to turnout? (Nope!)
- Can cellphone trace data be used to successfully target aid to the world's poorest? [link]

Figure 2: McCloud River, 2025

- Does the changing climate have an impact on our wellbeing? [link], [link]
- Can design improve lives? [link]
- What do people like in the design of a house?
- Does increasing login pressure increase compliance with cybersecurity training?
- Does {this} product feature affect {that} KPI?

Deep understanding is knowing how things work when they are taken apart.

0.1 Core Questions

This course is about designing experiments that we run in the *real-world*.

- What is the value of making a causal statement?
- Why do we conduct experiments?
- This is a modern academic program – we've got, you know, LLMs and stuff... With enough data and a savvy enough model, can't we just generate a causal statement that will be right? Can't I generate a statement that converges in probability to the *correct* value?

0.2 Learning Objectives

At the end of this conversation, students will be able to

1. *Gain Access* to bCourses.
2. *Understand* the course goals, and access course assessments
3. *Understand* the course learning model
4. *Evaluate* an experiment they have run for themselves.

0.3 Student Introductions [Breakout One]

In a breakout room of between three and four students introduce yourself!

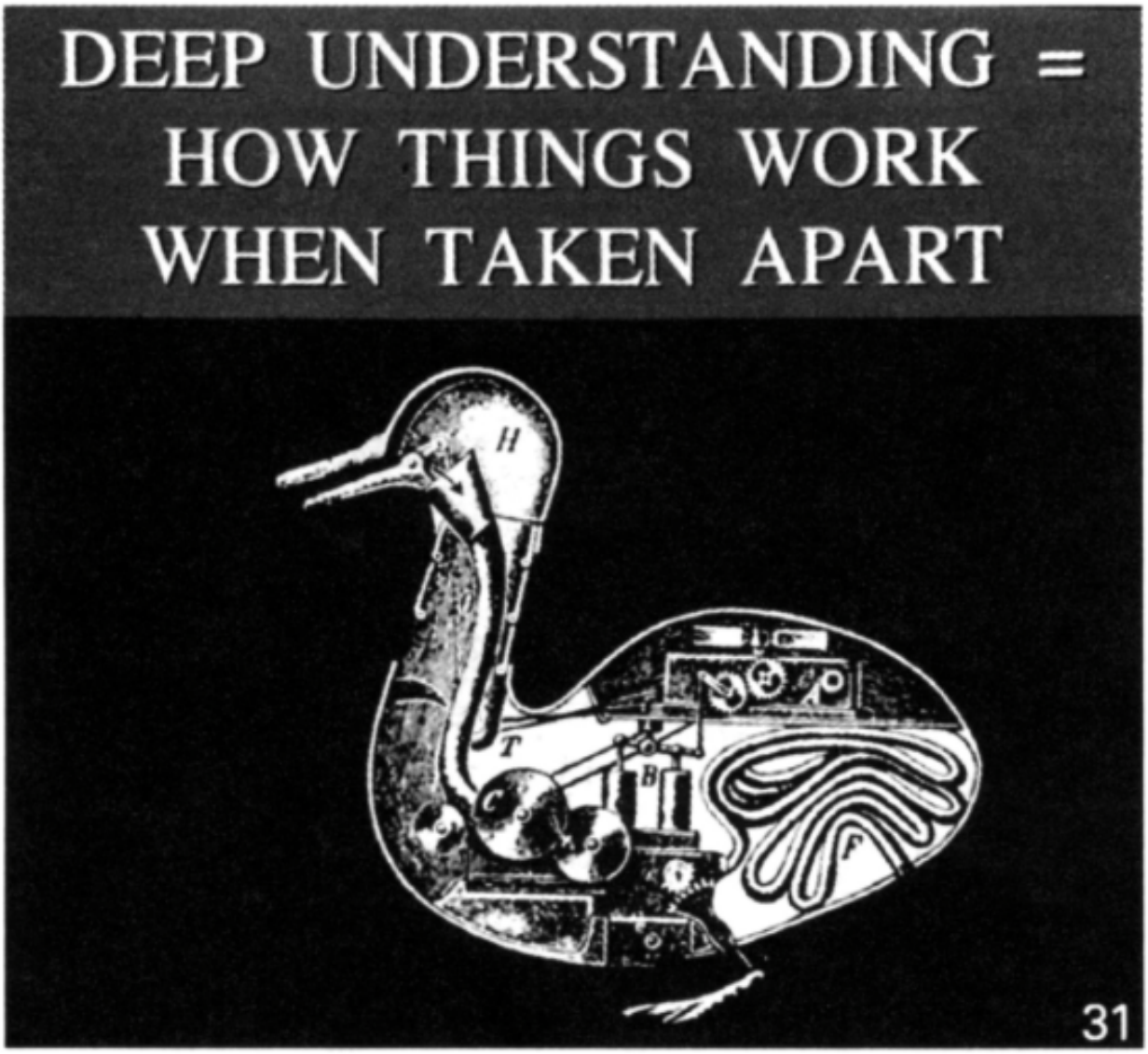


Figure 3: duck

Breakout One. A *name story* is the unique, and individual story that describes how you came to have the name that you do. While there may be many people are called the same thing, each of their name stories is unique.
Please share: *What is your name story?*

0.4 Student Introductions [Breakout Two]

In the same breakout room:

Breakout Two. Like our names, the reasons that we joined this program, our goals and our histories are different.
Please share: *What is your data science story? How did you wind up here, in this room today?*

0.5 Course Plan

The course is built out into three distinct phases

- **Part 1** Develops causal theory, potential outcomes, and a permutation-based uncertainty measurement
- **Part 2** Further develops the idea of a treatment effect, and teaches how the careful design of experiments can improve the efficiency, and easy of analysis
- **Part 3** Presents practical considerations when conducting an experiment, including problems that may arise, and how to design an experiment in anticipation of those problems.

0.6 Course Logistics

- bCourses
  - Learning Modules attached to weeks
  - Modules contain async lectures, coding exercises, and quizzes
- GitHub
  - All the course materials are available in a GitHub repository
  - We have protected the `main` branch, so you can't do anything destructive
  - Use that as empowerment! This is your class, propose changes that you would like to see!
- Github Classroom
  - Assignments will all be applied programming assignments against simulated and real data
  - All assignment code will be distributed through GitHub Classroom
- Gradescope
  - All assignments will be submitted to Gradescope where we'll read your solutions and provide scores and feedback

0.6.1 Learning model for the class

The course assignments are designed to put what we have learned in reading, async, and live session into practice in code. In our ideal version of your studying, we would have you working hard together with your classmates in a study group on the assignments, coming to office hours to talk candidly about what is and isn't working, and then *every single student* arriving at a full solution.

0.6.2 Feedback model for the class

We want to get you feedback *very* quickly after you turn your assignments.

1. We will release a solution set the day that you turn your assignment in
2. We will hold a problem set debrief office hour the Friday (i.e. next day) after the problem set is submitted
3. We will have light-feedback on your assignments within 7 days of when you submitted them.

4. You should bring your assignment to office hours after you have turned it in so that we can talk about any differences between your approach, and my approach.

0.6.3 Office hour model for the class

- I will be holding an office hour session every Monday from 11:00a-12:00p. Come by! If we find that we need structure, I'll set up a signup form.
- I will be holding a lab session every Tuesday from 11:00a - 12:00p.
  - It will also be a chance to write code: To support your learning, by doing.
- I think that we're going to be assigned a TA for this class as well. They will have office hours on Friday, the day after class so that you can talk about problem sets.

0.7 Run Our Own Experiment

There's no actual magic to running an experiment. In fact, it is so core to how we think about the world that you can assuredly conduct an experiment here, today, in the next 30 minutes that checks all the boxes.

Experiment options:

1. Take action to cause an **increase** in heart rate.
2. Take action to cause a **decrease** in heart rate.
3. Take action to **increase** the duration that we can hold our breaths.

0.7.1 Decisions to make

1. What experiment are we going to conduct?
2. What intervention are we going to choose?
3. What comparison are we going to make?
4. How are we going to randomize?
5. What are we going to measure?
6. How are we going to measure it?
7. What test are we going to conduct?
8. What evidence would be required for us to make a conclusion?
9. What could go wrong?
10. What *actually* went wrong?

Bloom's Taxonomy

An effective rubric for student understanding is attributed to Bloom (1956). Referred to as *Bloom's Taxonomy*, this proposes that there is a hierarchy of student understanding; that a student may have one *level* of reasoning skill with a concept, but not another. The taxonomy proposes to be ordered: some levels of reasoning build upon other levels of reasoning.

In the learning objective that we present in for each live session, we will also identify the level of reasoning that we hope students will achieve at the conclusion of the live session.

1. **Remember** A student can remember that the concept exists. This might require the student to define, duplicate, or memorize a set of concepts or facts.
2. **Understand** A student can understand the concept, and can produce a working technical and non-technical statement of the concept. The student can explain why the concept *is*, or why the concept works in the way that it does.
3. **Apply** A student can use the concept as it is intended to be used against a novel problem.
4. **Analyze** A student can assess whether the concept has worked as it should have. This requires both an understanding of the intended goal, an application against a novel problem, and then the ability to introspect or reflect on whether the result is as it should be.
5. **Evaluate** A student can analyze multiple approaches, and from this analysis evaluate whether one or another approach has better succeeded at achieving its goals.

6. **Create** A student can create a new or novel method from axioms or experience, and can evaluate the performance of this new method against existing approaches or methods.

1 Importance of Experimentation

Figure 4: Point Reyes National Seashore

1.1 Learning Objectives

1. *Define*, in non-technical language, what it means for an action to cause an outcome.
2. *Understand* the difference between a causal statement, and an association statement.
3. *Apply* the framework of causal thinking against a series of studies to determine whether the study has achieved the goal that it intends.

1.2 Class Introductions

1.3 Article Discussion

1.3.1 Predict or Cause

- What are a few examples that Athey raises of causal questions masquerading as prediction questions?
  1.
  2.
  3.
- Which of these examples is the most surprising to you?
- Is there something that is common to each of these examples? Is this a general phenomenon, or is Athey very clever in picking examples? Said differently, is Athey making a clever argument or is a lot of what we do as data scientists actually causal work in disguise?

1.3.2 Do the suburbs make you fat?

1. What is the causal claim being made in this article?
2. If you had to draw out this causal claim, using arrows, what would it look like?
3. Do you acknowledge the association that the authors present? Is there *actually* a difference between the BMI of people who live in cities and the suburbs?
4. If you acknowledge the association, does that compel you to believe the causal claim? Why or why not?
5. Name, and draw, five alternative *confounding* variables that might make you skeptical that the claimed relationship exists.
6. (Optional) Name, and draw two *mechanisms* that might exist between suburbs and BMI. Why does the existence (or not) of these mechanisms *not* pose a fundamental problem to the causal claim that the authors make?
7. At the conclusion of reading this paper, do you believe that there is a causal relationship between location and BMI? If so, what compels you to to believe this; if not, why are you not compelled to believe this?

1.3.3 Nike Shoes

1. What is the causal claim being made in this article?
2. If you had to draw out this causal claim, using arrows, what would it look like?
3. Do you acknowledge the association that the authors present? Is there actually a difference in the finish time between people who are running with the Nike shoes vs. other shoes?
4. If you acknowledge the association, does that compel you to believe the causal claim? Why or why not?
5. What are some of the confounding relationships that the authors identify? (Can you name four?) How do they adjust their analyses once they acknowledge the confounding problem?
6. At the conclusion of reading this paper, do you believe that there is a causal relationship between shoes and finish time? If so, what compels you to to believe this; if not, why are you not compelled to believe this?

1.3.4 What is Science: Feynman's View

In *Cargo Cult Science*, Richard Feynman poses a view of science that is about a seeking of the truth.

1. What is Feynman's view of science? What does he think makes something *scientific*?
2. What are ways that individuals fool themselves when they are working as scientists? What are ways that individuals fool themselves when they are working as data scientists?
3. How can we as (data) scientists, train ourselves not to be fooled?[1]

2 Apples to Apples

2.1 Learning Objectives

At the conclusion of this week's live session, student will be able to:

1. *Describe*, using the technical language of potential outcomes, what it means for an input to *cause* an output.
2. *Describe* the fundamental problem of causal inference.
3. *Apply* iid sampling as a method of producing an unbiased, consistent estimator of a population.
4. *Prove* that the average treatment effect estimator produces an unbiased, consistent estimator for the average treatment effect.

2.2 Revisiting Ideas of Science

Questions about epistemology are a *classic* questions. These questions are particularly relevant here at the School of Information. You're working toward being a data scientist that has a full view of not only how to

[1] This is a footnote, rendered into an html document.

Figure 5: fruit salad

build the technology, but also for how that technology will behave, alter, and affect the people who use it.

The idea of epistemology – the idea that some things are known truths, while others are merely opinions – is perhaps one of the earliest philosophical (i.e. academic questions).

*What does it mean to **know** something?*

"Do we know this is true, or do we just believe it to be true?" is more than just an academic question. In our workplaces we need to know how to take the best course of action. As data scientists we need to know that the answers we are producing stand on some justification. And, for the purposes of this course, we are attempting to separate things that *certainly* cause an outcome from those that we *think* cause an outcome.

2.2.1 From last week

Think back to the reading and discussion from last week: For Feynman, what does it mean to be "Doing science?"

- Would Feynman say that data science, as we are practicing it, is a "science"?
- Would Feynman say that 205 is a science?
- What about computer science or statistics?

2.2.2 From this week: Lakatos

For Lakatos, what does it mean for something to be a part of a science? What does it mean for something to be a part of a psuedo-science? Really, this is a question about how Lakatos draws a line between things that we *know* and things that we *believe*.

- Is Lakatos' view as simple a view as Feynman espouses?
- Where would Lakatos agree, and where would he disagree with Feynman on the "scientific" nature of the courses in the MIDS program?

What do you think produces knowledge?

Can a single conversation produce knowledge? Can a non-experimental study produce knowledge about a causal effect?

Can an experiment fail to produce knowledge? If an experiment fails to reject some null hypothesis, does that mean that it has not produced any knowledge?

2.3 This Causes That

What does it mean "to cause"?

In your own words, what does it mean for an action to *cause* an outcome? Do not focus on conducting an experiment to *measure* the cause; and don't worry about the difficulties of measurement. Just engage with the concept of what it means for something to cause.

- How would you describe the idea of "cause" to a grandparent? See if you can describe it without relying on an example. Dig deeper to find the core essence of the concept?
- How would you describe the idea of "cause" to a student who is enrolled in MIDS 203?

2.4 Working Cases of Causality

In this short section, you are going to apply your breakout group's concept of causality against a series of scenarios. Rather than defining your concept using examples or scenarios, you have built a working conceptual definition that should be able to address the ideas of causality that they are confronted. If you find your group's working definition cannot address the scenario that it is faced with, take the time to alter the definition so that it *can* be useful.

2.4.1 Damn fine coffee

Suppose that you're getting ready for class, and you want to make sure that you're at your best. So, you drink a cup of water, eat a small snack, and brew a small pot of coffee for while you're in class.

Why do you do this?

Presumably, you're doing this because you like each of these things, but also because you're interested in these things causing you to have a better class. If you framed this as as causal question, you might ask:

If I drink a cup of coffee before class, will it cause me to be more alert?

What does it mean for coffee to cause alertness?

- Does coffee cause everyone to become more alert?
- Does coffee have to affect everyone equally in order for you to say it causes alertness?
- Could coffee have no effect for some people, and you would still say it causes alertness?

2.4.2 Meditation for focus

Suppose that you're getting ready for class, and you want ensure that you're at your best. So, you find a quiet place, and set your mind at ease with whatever form of meditation you think might be helpful.

If I meditate before class, will it cause me to be more focused?

What does it mean for meditation to cause focus?

- Does meditation cause everyone to become more focused?
- Does meditation have to affect everyone equally?
- Some people are frustrated by not being able to quiet their thoughts, and actually find meditation frustrating. Can this be true, and still believe that meditation causes focus?

2.4.3 Selling coffee and meditation

Suppose that you're an enterprising soul, and you want to sell a book about brewing coffee as a meditation. You reason that there must be a niche for this approach. To get the word out, you place a few flyers with tear off phone-numbers at the local yoga studios and tech incubators (good intuition to find those MIDS students).

If shown a flyer for coffee-meditation, will it cause someone to take my training?

What does it mean for for flyers to cause people to sign-up for the training?

- Does the flyer cause everyone to take the training?
- Does the flyer affect everyone equally?

One might be a radical behaviorist who believes that, "In matters of human behavior, if I cannot see it, then I cannot reason or know about." If this is your view, then you would simply stop your investigation (and reasoning) at the conclusion of your experiment.

In many ways, experiments suited only to answer empirical, observable questions. These are the questions, and lens proposed by the radical behaviorist paradigm.

2.4.4 Limits of Behaviorist Reasoning

Are you comfortable being a radical behaviorist? Are you willing to know only what you can see and observe and measure?

Suppose that you run an experiment about whether coffee affects your alertness. And, you find that, "Yes! It does!" Then, as you're getting ready for class, and you want to be alert for the discussion, what might you do?

Suppose that you go on a coffee-bender, and as you're getting ready for live session in week three, you go to the cupboard to brew a pot, and realize, "Oh no! I've drank all the coffee in the house! Now, I'll have to scramble to find something else to make sure that I'm alert in class."

What would you go and consume to make you alert? Why do you think that this will be effective at making you more alert? Did you experiment tell you that this new substance would help make you more alert?

If you're a radical behaviorist, or in this case, just a reasonable scientist do you have any knowledge of what you should drink?

2.4.5 Reflecting on Causes

Does anything unify questions of causes?

When you think about *{this}* causing *{that}*, do you think about it at a population level, a smaller group level, or at the individual level?

2.4.6 Evaluating Value

Is this an entirely academic exercise, the discussion of *{this}* causing *{that}*? Or, is there some value to thinking about things in these terms? Susan Athey, whom we read last week, seems to think that there is value in distinguishing between associations and causes. However, hers is a view that is generated by an academic; much like the views of David and David, and all of the live session instructors. We're all academics, so maybe we're being *typical* academic pedants.

What is a case, perhaps that you read or wrote about for your first essay, mistakenly believing they had measured a causal effect? What would happen if they implemented the policy that is implicated in their study? Or, what would happen if they took action consonant with what their study purports to find?

2.4.7 Value of Theory

- Can you produce several theories (some of them might be silly) about why coffee might increase alertness in class?
  - Proposed theory #1:
  - Proposed theory #2:
  - Proposed theory #3:
- Does Feynman's approach to *Science* provide a method to adjudicate which of these theories is consonant with the evidence, and which are not consonant with the evidence?
- Does Lakatos' approach to *Science* provide such a method?

2.4.8 Evaluating Theories

- What data might you be able to produce that would allow you to "drive a wedge" between the different theories?
- This ability to proactively deign an experiment to distinguish between theories is the goal you're striving to achieve, *and it is very hard to accomplish.*

2.5 Reading Discussion: The Power of Experiments

The Power of Experiments starts the discussion of experimentation in the workplace with what is, for the course instructors, a uniquely pedestrian example, increasing contributions to taxes. In particular, Her Majesty's Revenue and Customs sends different versions of a letter to British taxpayers, and observes that different language leads to different amounts of taxes being paid.

2.5.1 Chapter One: The Power of Experiments

1. Is it *actually* a "big-deal" to increase tax compliance by 2 percentage points?
2. On page five, the book identifies five "one-liners" that HMRC chose to send to taxpayers:

3. *Nine out of ten people pay their tax on time."*
4. *Nine out of ten people in the UK pay their tax on time.*
5. *Nine out of ten people in the UK pay their tax on time. You are currently in the very small minority of people who have not paid us yet."*
6. *Paying tax means we all gain from vital public services like the NHS, roads and schools.*
7. *Not paying tax means we all lose out on vital public services like the NHS, roads, and schools.*
8. Which of these sentences would be the most effective at getting you to pay your taxes? Which do you think will be most effective, overall, at generating tax compliance? Why? How willing are you to make a million pound bet that you're correct?
9. Some of your instructors are vegetarians. None of them, however, has previously made an argument for why everyone should be vegetarian based on the example of Daniel and his study of diet and divine intervention. What about the study that Daniel conducted produces evidence that you think is useful for evaluating diet? What are the limitations that you see in this study? The book lists several, but there are other issues, along the lines of the *exclusion restriction* that *Field Experiments* identifies.
10. In order for Pasteur to be declared the winner of the vaccine argument, the observers said that every control group sheep had to die and every treatment group (i.e. vaccine-receiving) sheep had to live. Is this a fair burden of proof? Do the frequentest tests that we developed in 203 and are going to use here in 241 set a higher or lower bar than Pasteur faced? What are the merits of a relatively higher or lower bar?

2.5.2 Chapter Two: The Rise of Experimetnts in Psychology and Economics

Freud is noted as being specifically *against* experimentation. But, *PoE* then goes on to write, "[Freud's] big ideas inspired entire fields of psychological research. Including the notion that unconscious processes shape our judgement and behavior, psychological disorders are rooted in the mind rather than the body; and that sexual urges and behavior are worthy of study" (p. 19). Some of the theories that Freud promulgated were found to have evidence that was consistent with the theory; some of these theories could not produce evidence to support the theory; and many were outright contradicted by the evidence.

1. Is there value in being an "idea person"? How would you ever know if your ideas were actually right if you're unwilling to evaluate them?
2. What, if any, are the limitations of experimenting without any "big ideas" to ground your experiments?

Behaviorists (Skinner is the leading behaviorist) make a compelling argument: "One cannot directly observe what is happening in the mind of a person." A classic implication of this argument for behaviorists is that only that which is empirically observable is reasoned about. "Why does the rat avoid getting shocked? Does it really matter?" "Why does the child want a cookie? Does it really matter why?"

1. Is this position reasonable for you to take as you navigate your own life? If you spoke with a therapist or a coach and said, "I've been feeling stressed over the past several weeks," would be satisfied with a *mindful* answer like, "Well, let's acknowledge those feelings and hold them for a moment" or would you want to reason further about why you feel stressed? What are the types of things in people's heads that you think we can profitably reason about; what are the types of things in people's heads that we cannot reason about? Is there something that is common to those that we can or cannot work with?

The experiments of Milgram and Zimbardo are widely identified as the reason that human-subjects review boards no exist. These review boards serve as an external review that keeps researchers from inflicting harm to individuals that is not outweighed by societal or scientific benefits.

1. What did Milgram and Zimbardo do to their subjects?
2. By talking continuing to talk about these experiments nearly fifty years after they were conducted – even if we are talking about them negatively – are we adding to the fame of these researchers? (For those interested in inside baseball, Zimbardo was the president of the American Psychology Association in 2002, and was awarded a lifetime achievement award from his discipline.) How should we learn and react to work that shouldn't have been conducted in the first place?

Kahneman and Tversky propose that individuals think about expected values differently depending on

whether they are thinking in the domain of gains or the domain of losses. They come to this theory through the, now cringe-worthy, *Asian Disease Problem*:

In the positive frame, they ask the question:

> Imagine that the US is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
>
> - If **Program A** is adopted, 200 people will be saved.
> - If **Program B** is adopted, with a 1/3 probability 600 will be saved and with a 2/3 probability nobody will be saved.

The authors also present a countervailing pair of scenarios framed in terms of losses

- If **Program C** is adopted, 400 people will die.
- If **Program D** is adopted, there is a one-third probability that no one will die, and a 2/3 probability that everyone will die.

Clearly, all these programs have the same expected number of deaths; but, people can disagree about which of these is the program that we should pursue. Just ask as a poll in the class; and, ask people to justify their beliefs.

2.5.3 The Rise of Behavioral Experiments in Policymaking

PoE points out that experiments abound in policy making. Part of this stems from a truthful ignorance of the optimal policy to pursue. Another part of this stems from the ability of policy makers to make decisions by fiat that affect a large number of people.

1. Does this justification for experiments align with your current understeanding of the landscape in human-facing data science?

In a section titled **The nuance behind behavioral insights** the authors state a series of three caveats:

1. *Context matters*
2. *Design choices matter*
3. *Unintended consequences abound*

1. What do they mean when they raise the three points?
2. We are going to ask you to justify conducting experiments by staking out an extreme point of view, and asking you to convince us that this point of view is so extreme that it cannot be justified. *"Writing that context matters, design choices matter, and unanticipated consequences abound is little more than writing that experiments cannot produce any more useful insights than the theories of Freud. As a result, there is little reason to conduct any experiments because what we learn will be highly contextualized, affected by very small implementation choices, and may generate as many (or more) negative outcomes as positive outcomes."*

2.6 Potential Outcomes

Potential outcomes are a system of reasoning, and a corresponding notation, that allow us to talk about observable and un-observable characteristics of the world.

> What is your position on *ontology*? What does it mean for something to exist?

- Does *Field Experiments*, as a textbook, exist?
- Do Don Green and Alan Gerber, the authors of the textbook that we're reading, exist?
- Does David Reiley, the slower-talking Davids in the async, exist?
- Do I, your section, instructor, exist (or am I a deep fake in this room with you)?
- Can a concept exist, even if you can't hold it? Even if you haven't seen it?

2.6.1 Defining Potential Outcomes

For each of the following sets of notation: (1) Read the notation aloud, not as "Y sub i zero", but instead as "The potential outcome to control . . . ".

- $Y_i(0)$:
- $Y_i(1)$:

- $E[Y_i(0)]$:
- $E[Y_i(1)]$:
- $E[Y_i(0)|D_i = 0]$:
- $E[Y_i(1)|D_i = 1]$:
- $E[Y_i(0)|D_i = 1]$:
- $E[Y_i(1)|D_i = 0]$:
- Which of these concepts that you have just read aloud exist?
- Can a concept exist, even if you can't hold it? Even if you can't see it?

2.7 Using Independence

Suppose that you have a random variable that is defined as the function,

$$Y = \begin{cases} \frac{1}{10} & , 0 \leq y \leq 10 \\ 0 & , \text{otherwise} \end{cases}$$

- What is the expected value of this function?

$$\begin{aligned} E[Y] &= \int_0^{10} y \cdot f_y(y)\ dy \\ &= \int_0^{10} y \cdot \frac{1}{10}\ dy \\ &= \frac{1}{10} \int_0^{10} y\ dy \\ &= \frac{1}{10} \cdot \frac{1}{2} y^2 \Big|_0^{10} \\ &= \frac{1}{20} y^2 \Big|_0^{10} \\ &= \frac{1}{20} \cdot [(100) - (0)] \\ &= \frac{1}{20} \cdot 100 \\ &= \mathbf{5} \end{aligned}$$

- Why is the expected value a good characterization of a random variable?
- If you wanted to write down an estimator to produce a summary statistic for Y given a sample of data, what properties do the following estimators possess:
- $\hat{\theta}_1 = y_1$

- $\hat{\theta}_2 = \frac{1}{2}\sum_{i=1}^{2} y_i$
- $\hat{\theta}_3 = \frac{1}{n-1}\sum_{i=1}^{N} y_i$
- $\hat{\theta}_4 = \frac{1}{n}\sum_{i=1}^{N} y_i$

```
conduct_sample <- function(size) {
  runif(n=size, min=0, max=10)
}
```

```
theta_1 <- function(data) {
  # take the first element

}

theta_2 <- function(data) {
  # sum the first two elements and divide by two

}

theta_3 <- function(data) {
  # sum the sample, and divide by 1 less than the sample size

}

theta_4 <- function(data) {
  # sum the sample, and divide by the sample size
  # honestly, just use the mean call.
  # clearly, this is a silly function to write, since you're just
  # providing an alias, without modification, to an existing function.

  mean(data)

}
```

```
theta_4(conduct_sample(size=100))
```

```
## [1] 4.944307
```

- Just to put a fine point on it: **What estimator properties does the sample average provide, and why are these desirable?"

2.8 Use Randomization to Produce Independence

How can we use the independence that is induced by "random **assignment** to treatment" combined with the sample average estimator to produce an estimate of an otherwise very difficult concept to measure?

2.9 Theoretical Justification

Before we show that this very simple ATE estimator work against a sample of data, it is worth reasoning about whether we can guarantee that it works in a general case. If we can show that it works in a general case, then any specific case inherits that guarantee. However, if we can only reason thorugh the existence of a single examlpe, it is not a sufficient argument to compell us to believe that it must hold for all cases.

Here's an example, "Behold! I see a black sheep! Therefore all sheep are black." This doesn't make sense, and it is not a logically sound argument. However, if you say, "All sheep say, 'Baaah!' This is a black sheep, so it must say 'Baah!' " is a logically sound arugment, so long as the antecedent is, in fact true. When we're proving something, we're proving that the antecedent to this statement is generally true. For anyone who took a symbolic logic course in, this method of argument might be marked down as $\forall X \implies \exists X$, whereas $\exists Y \not\Rightarrow \forall Y$.

- What concepts compose τ_{David}?
- What concepts compose τ_i?
- Is there any reason to believe that $\tau_{David\ Reiley} = \tau_{David\ Broockman}$?
- Is there any reason to believe that $\tau_i = \tau_j$, where $j \neq i$?
- Could $\tau_i = \tau j$?
- What is the fundamental problem of causal inference?

The proof for this argument is also made in *Field Experiments*, on or about page 30 of the text. However, in our view, the authors don't give enough room to fully develop this proof, and so we skipped right past it the first time that we read the chapter.

Begin our proof with the statement for what a treatment effect is, τ_i.

$$
\begin{aligned}
\tau_i &= Y_i(1) - Y_i(0) && Definition \\
ATE &= E[\tau] && Definition \\
&= E[Y(1) - Y(0)] && Substitution \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&=
\end{aligned}
$$

2.10 Simulation Example

Now, let's work through an example that shows this works not only in the math, but also in the realized, i.e. sampled, world.

To begin with, lets work with a *very* simple sample that has 100 observations, potential outcomes to control are uniformly distributed between `0` and `1` and every single unit has a potential outcome to treatment that is `0.25` units larger than their potential outcomes to control.

```
make_simple_data <- function(size=100) {
  require(data.table)

  d <- data.table(id = 1:100)
```

```
  d[ , y0 := runif(.N, min = 0, max = 1)]
  d[ , y1 := y0 + .25]

  return(d)
  }

d <- make_simple_data(size=100)

d[1:5]
```

```
##       id        y0        y1
##    <int>     <num>     <num>
## 1:     1 0.7054674 0.9554674
## 2:     2 0.4235482 0.6735482
## 3:     3 0.8641622 1.1141622
## 4:     4 0.7710945 1.0210945
## 5:     5 0.9426160 1.1926160
```

In this world, we've taken a sample of 100 individuals, and at this point, each of those individuals that we've sampled has both a potential outcome to control **and also** a potential outcome to treatment. We haven't talked at all about measurement yet; we're just asserting that both of these potential outcomes exist for each person.

Essentially, this stage of creating the sample is the same as bringing people in the door to your experiment. If you were running this in the laboratory, you'd literally think of this as sitting your subjects down at their chairs, getting ready to begin their task.

> Is randomly sampling people to be a part of your experiment sufficient to ensure that your experiment produces an unbiaed, consistent estimate of the true treamtent effect?

Suppose that for each unit, you then toss a coin, placing the subject either into treatment or control based on the result of that coin flip.

- Does this coin flip ensure that you have the same number of units in treatment as control? Does this matter to you? Why or why not?
- Are there other ways that you could assign individuals to treatment an control, rather than through a simple-randomization process?
- What are the relative merits or limitations of each of the methods?
- Are some of these methods *more random* than others? Or, are all things that are random equal in their randomness?

2.10.1 Assign to Treatment and Control

```
d[ , experimental_assignment := sample(0:1, size = .N, replace = TRUE)]
d[1:5]
```

```
##       id        y0        y1 experimental_assignment
##    <int>     <num>     <num>                   <int>
## 1:     1 0.7054674 0.9554674                       0
## 2:     2 0.4235482 0.6735482                       0
## 3:     3 0.8641622 1.1141622                       0
## 4:     4 0.7710945 1.0210945                       1
## 5:     5 0.9426160 1.1926160                       1
```

As as comparison, suppose that instead of randomly assigning individuals into treatment and control we allowed individuals to select into treatment and control. And suppose that people with the lowest potential outcomes to control opt to take the treatment. You might think of this as being something like, "The people

who are the most tired are the most likely to drink a cup of coffee before they start class," if an example helps you ground this.

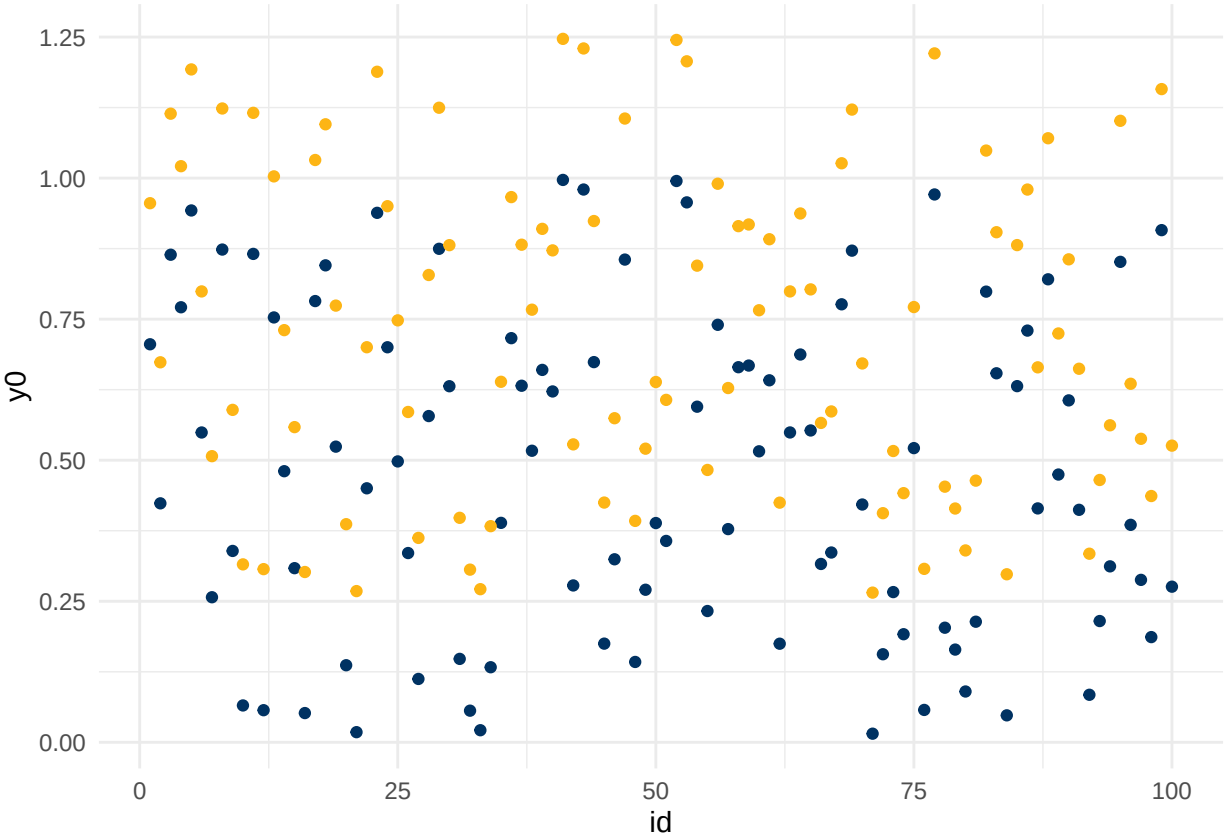
Specifically, suppose that any unit that has a potential outcome lower that `0.33` opts to take the treatment.

```
d[ , observational_selection := ifelse(y0 < .33, 1, 0)]
d[1:5]
```

```
##       id        y0        y1 experimental_assignment observational_selection
##    <int>     <num>     <num>                   <int>                   <num>
## 1:     1 0.7054674 0.9554674                       0                       0
## 2:     2 0.4235482 0.6735482                       0                       0
## 3:     3 0.8641622 1.1141622                       0                       0
## 4:     4 0.7710945 1.0210945                       1                       0
## 5:     5 0.9426160 1.1926160                       1                       0
```

These represent two different ways that you might conduct your research, each time with the same subject pool. Of course, in reality you probably would not be able to run these two studies at the same time, but since this is a simulation, we can stretch the confines of reality just a little bit.

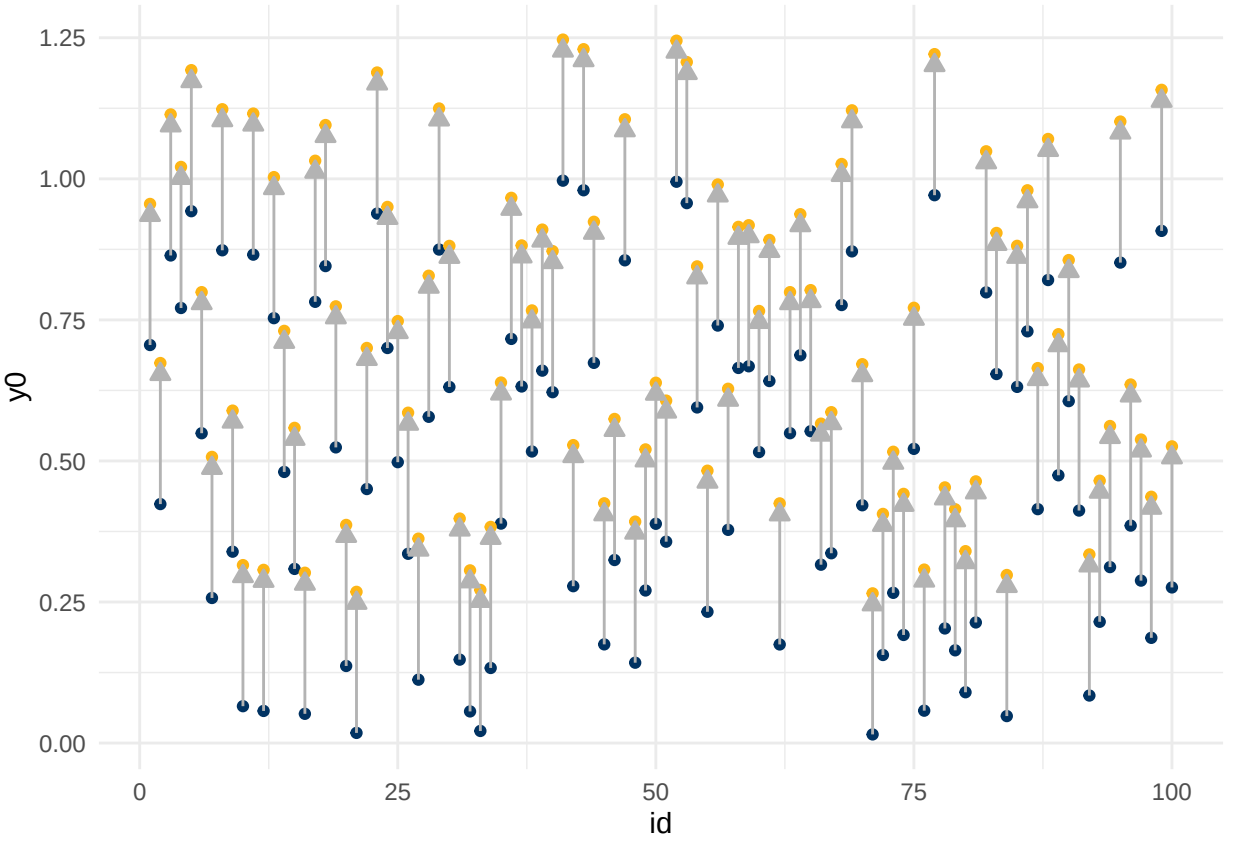
```
first_plot <- ggplot(data=d) +
  geom_point(aes(x = id, y = y0), color = blue) +
  geom_point(aes(x = id, y = y1), color = gold)
first_plot
```



What's actually happening in this? It might be more clear if we add arrows to this plot to show.

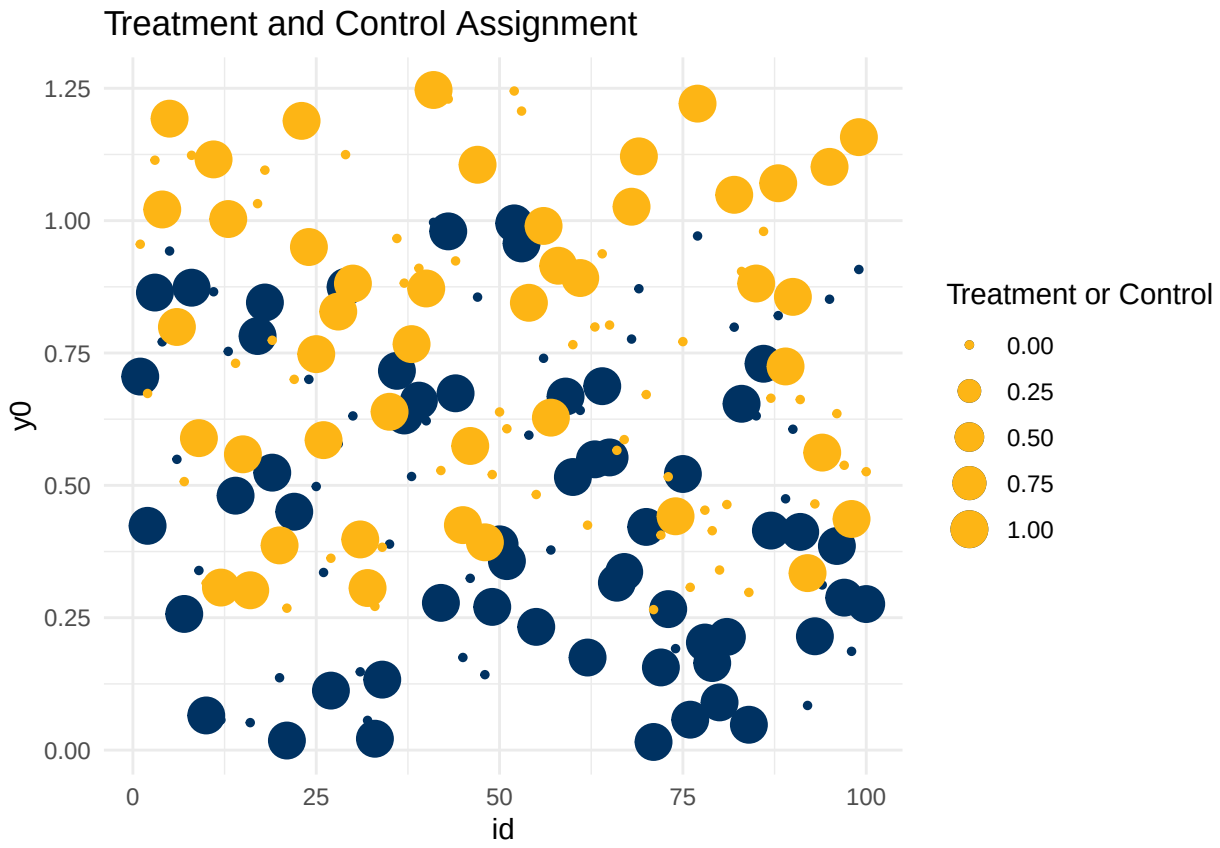
```
first_plot +
  geom_segment(
    aes(x = id, xend = id, y = y0, yend = y1),
```

```
    arrow = arrow(ends = 'last', length = unit(0.1, "inches"), type = 'closed'),
    color = 'grey70'
    )
```



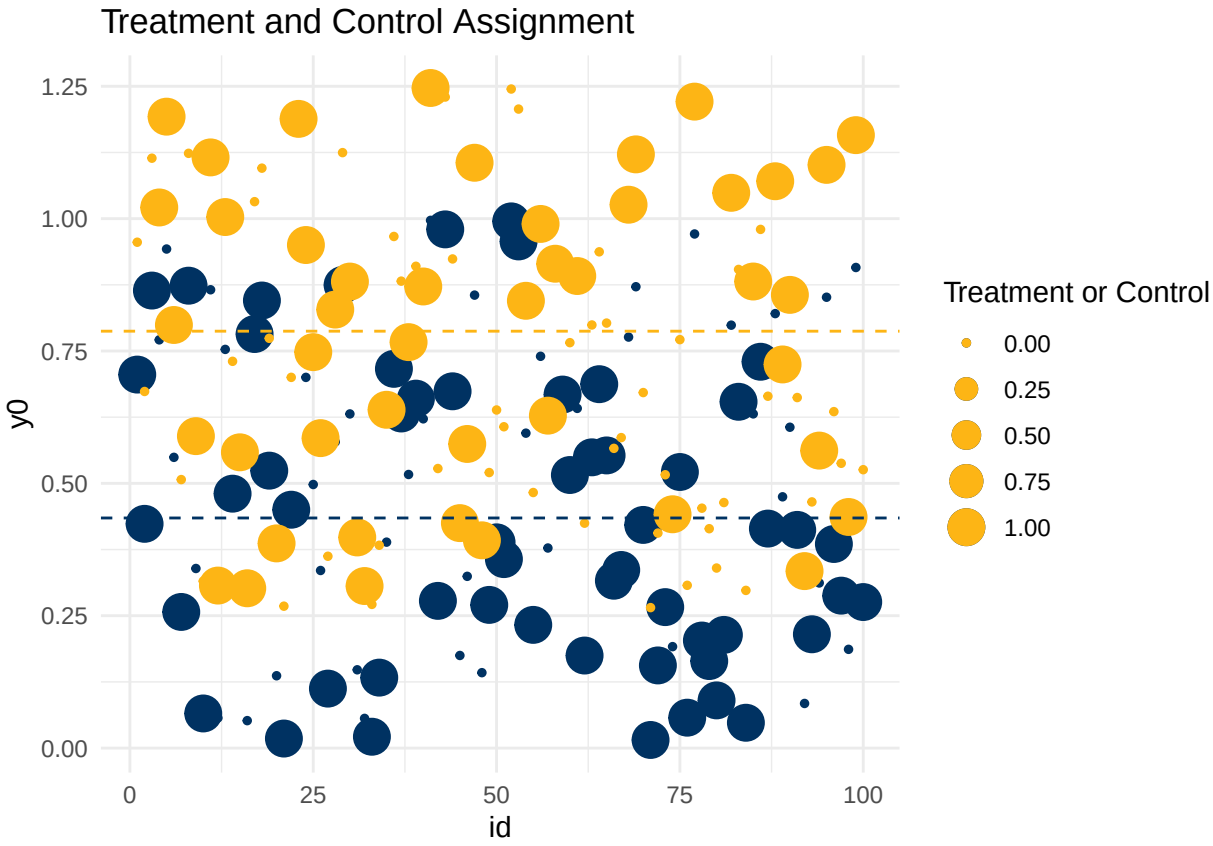
Even though these potential outcomes exist for all the units, is it possible to actually see them for all the units? How do we go about showing, and then measuring the potential outcomes to control for a set of units? How about the potential outcomes to treatment?

```
second_plot <- ggplot(data = d) +
  geom_point(
    aes(x = id, y = y0, size = 1 - experimental_assignment),
    color = blue) +
  geom_point(
    aes(x = id, y = y1, size = 0 + experimental_assignment),
    color = gold) +
  labs(
    title = 'Treatment and Control Assignment',
    size = 'Treatment or Control'
  )

second_plot
```



What are the averages of these samples that have been assigned to treatment?

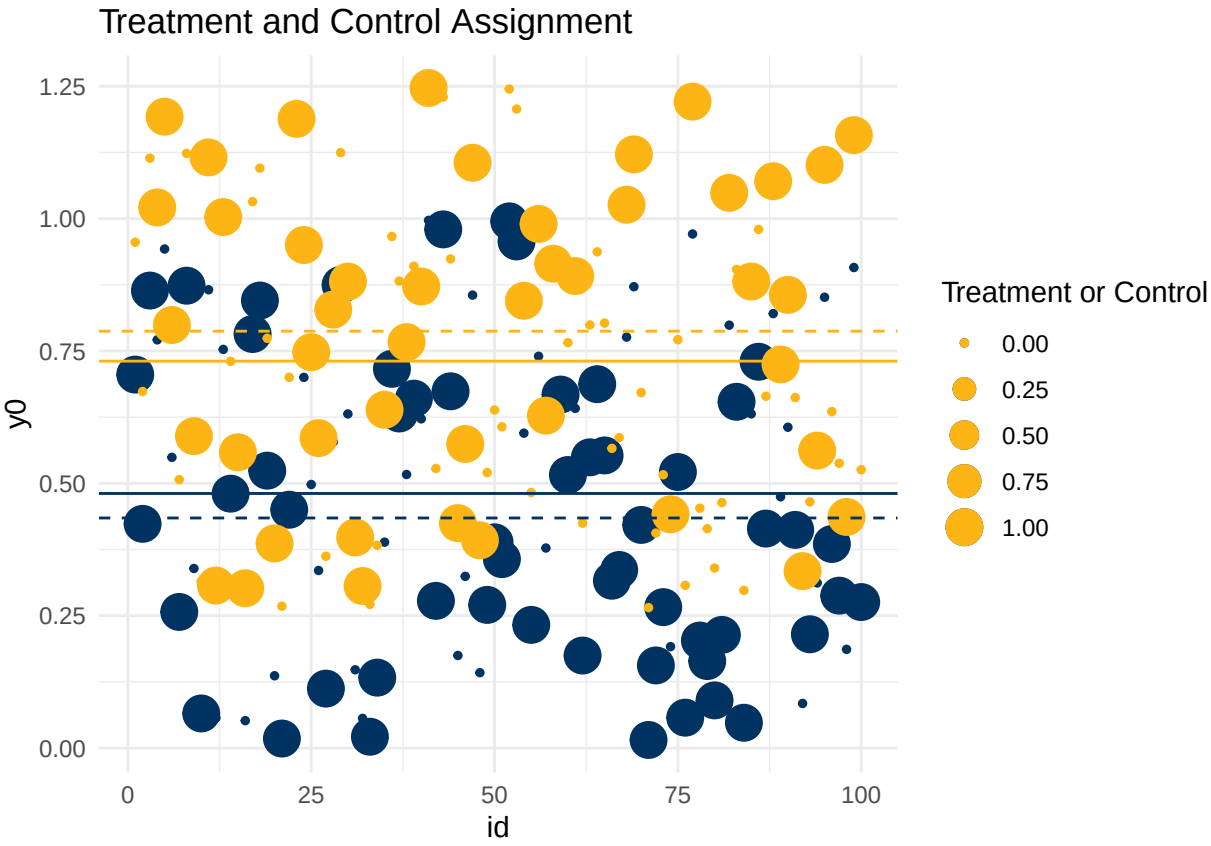
```
third_plot <- second_plot +
  geom_hline(
    yintercept = mean(d[experimental_assignment==0, y0]),
    color = blue,
    linetype = 2) +
  geom_hline(
    yintercept = mean(d[experimental_assignment==1, y1]),
    color = gold,
    linetype = 2)
third_plot
```



Even though we aren't able to see it, can we reason about what the sample average would be if we could see both of an individual's potential outcome to treatment and control?

- Is there a guarantee that the sample should be the same as the feasible realization?
- Should they be close? What property from 203 provides this guarantee?

```
third_plot +
  geom_hline(yintercept = mean(d[, y0]), color = blue, linetype = 1) +
  geom_hline(yintercept = mean(d[, y1]), color = gold, linetype = 1)
```



Put it all together, what has this little demo shown?

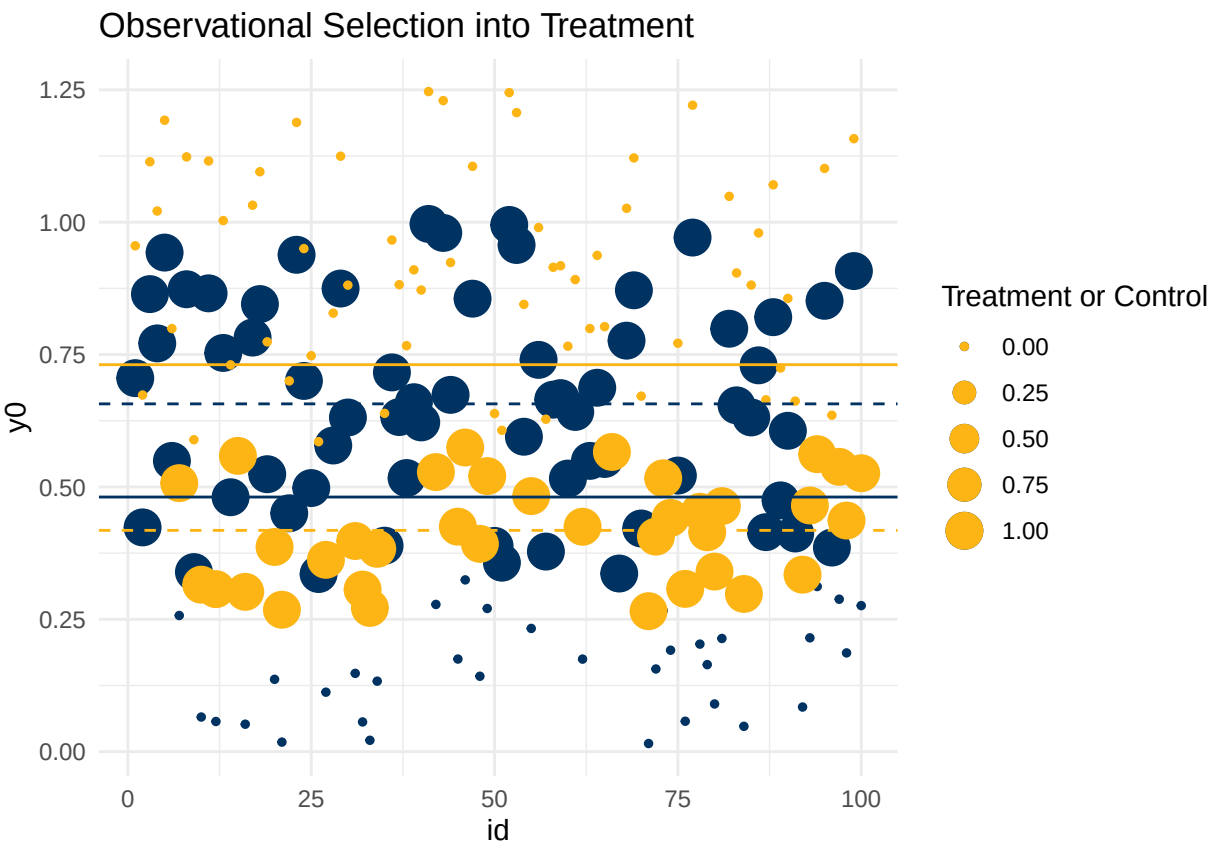
2.10.2 What if there is selection?

What if, rather than being assigned to treatment and control, instead individuals had been able to opt into treatment and control?

Produce only the last plot, but this time for the observational, or selected data.

```
selection_plot <- ggplot(d) +
  geom_point(
    aes(x = id, y = y0, size = 1 - observational_selection),
    color = blue) +
  geom_point(
    aes(x = id, y = y1, size = 0 + observational_selection), color = gold) +
  geom_hline(
    yintercept = mean(d[, y0]),
    color = blue,
    linetype = 1) +
  geom_hline(
    yintercept = d[observational_selection == 0, mean(y0)],
    color = blue,
    linetype = 2) +
  geom_hline(
    yintercept = mean(d[, y1]),
    color = gold,
    linetype = 1) +
  geom_hline(
```

```
    yintercept = d[observational_selection == 1, mean(y1)],
    color = gold,
    linetype = 2) +
  labs(
    title = 'Observational Selection into Treatment',
    size = 'Treatment or Control'
  )

selection_plot
```



2.11 Requirements of An Experiment

David Reiley makes the case that an experiment is something where we intervene in the world to produce knowledge. This is essentially providing a definition and making an argument that this is the correct definition. One difficulty with argument through definitions is that reasonable people can disagree because their definitions, through their lived experience, just disagree.

Here's the demonstrated proof:

> Who in class is from the "midwest" broadly defined? Is Chicago-style pizza, pizza *per se*? Who in class is from the east coast? Is Chicago-style pizza, pizza *per se*?
>
> Try not to make deep character judgments about your classmates.

Green and Gerber, in *Field Experiments* make additional requirements of experiments. As they argue on page 45 of the textbook, in their view, experiments require:

1. **Random Assignment**
2. **Excludability**

3. **Non-interference**

What do each of these terms mean? Why is each necessary?

- Did the experiment that Daniel conducted, described in *Power of Experiments* satisfy these three requirements? For any of these requirements that David's experiment did not satisfy, what are the consequences for the scientific knowledge that the experiment generated?

- Did the Nurses Health Study, described in the async, satisfy all these three requirements? For any that this experiment did not satisfy, what are the consequences for the scientific knowledge that the experiment generated?

2.11.1 Meta-Questions

- Can an experiment generate scientific knowledge about a causal effect, even without satisfying all of these requirements? Is it guaranteed to produce scientific knowledge about a causal effect?
- What then, justifies the use of experiments to measure causal effects?

3 Quantifying Uncertainty

3.1 Learning Objectives

At the end of this week's live session, students will be able to

1. *Understand* the sharp null, and how to apply it in an argument using randomization inference.
2. *Describe* how randomization creates uncertainty, and *assess* how this uncertainty differs from that in Frequentist paradigm
3. *Apply* the sharp null and randomization inference to data
4. *Assess* the assumptions necessary for Frequentist inference to produce nominal coverage on confidence intervals; *assess* the assumptions necessary for randomization inference to produce nominal coverage on confidence intervals; and, *evaluate* which of the two approaches is appropriate given a set of data.
5. *Describe* the concept of statistical power and what it means in the context of conducting an experiment.

3.2 Power of Experiments

3.2.1 Five Key Barriers to Experimentation

Power of Experiments identifies five key barriers to experimentation in companies:

1. **Not enough participants.** How can it be that even a huge, digital company (i.e. Uber) might not have enough participants to conduct an experiment?
2. **Randomization can be hard to implement.** This is not to be taken lightly; because in students essays this week, nearly every experiment proposed was of the form, "Randomly assign people to…". What might make it hard to randomize?
3. **Experiments require data to measure their impact.** This should ring of 201 conversations, but what is the concept that you would *ideally* like to measure about the impact of a policy? And, what instead are you able to measure? How much conceptual slippage is there between your conceptual definition and your data?

4. **Under-appreciation of decision-makers unpredictability.** Do we actually have a theory about what people will do? How sure are we that the theory is correct?

5. **Overconfidence in our ability to guess the effect of an intervention**.

3.2.2 Experimental Ethics

There is a very, *very* strong norm that academic researchers who conduct experiments need to pass their interventions, data collection, and procedures through a review board. This review board expects researchers will weigh the costs borne by the participants of an experimental study against the potential benefits to science from learning the results of this experiment.

In some cases, these boards determine that the costs are too high; nobody should be subject to those costs, no matter the scientific merits. In other cases, these boards will allow potentially costly actions to be taken, some that might even harm participants in the short-run. While it is quite unlikely that a review board would still approve either Milgram's or Zimbardo's infamous experiments, there are still many experiments that might harm participants.

- Is this OK?
- What are the tradeoffs, or goals that you would like to balance in an experiment?

A research team at Facebook (as your instructors if they have any juicy details about this case) was interested in the effects of their platform on its user's emotions. In pursuit of this question, they conducted an experiment – they intentionally manipulated the environment – to post more or fewer positive and negative posts.

- Is this OK?
- What are the tradeoffs, or goals that you would like to balance in this experiment?

3.3 Statistical Uncertainty – Randomization Inference Style

When we are working with a sample of data, estimates produced by an estimator might change – sometimes being higher than the *true* value, other times lower than the *true* value.

In Frequentist inference, we understand the variance in these estimates as *sampling based variance of the sample estimator*. In this week, we present a different inferential paradigm, **Randomization Inference**.

In randomization inference, there is no uncertainty about the parameter estimate that is generated in the experiment: The estimate that we observe is the estimate that we observe. Uncertainty, instead, comes from the acknowledgment that different *randomization* could have been realized, even from within the same sample.

3.4 Stating the sharp null

Suppose that you are evaluating the effect of coffee on students' alertness in class. You reason that drinking coffee will increase students' alertness in class.

Continue with our idea of an experiment to evaluate if coffee produces alertness in class. Here, we are going to further develop this notional experiment into something that we might actually be able to conduct.

- What is the *sharp null* hypothesis that is at risk in this investigation?
- How, if at all, does this sharp null differ from the null hypothesis you might be more familiar with?
- Is the sharp null hypothesis a concept that ever makes sense? Is the sharp null hypothesis a concept that is ever, actually, true?

3.5 Randomization Inference

3.5.1 Stating the process of Randomization Inference

Randomization inference is a method of understanding the variability of results in an experiment that you have conducted. It specifically acknowledges several facts:

1. The sample of data that you collected or used in your experiment is, quite simply, the sample of data that you collected for your experiment. There might be a larger population; there might be an infinite population; or, there might not.

Figure 6: "Damn Fine Coffee."

2. The observed outcomes that you observe are, quite simply, the outcomes that you observed. There is no uncertainty about having seen these.
3. When the experiment assigned some units to treatment and others to the control, it revealed some outcomes, for some people. Specifically, it revealed the potential outcoems to treatment, denoted $Y_i(1)$ for those who were assigned to the treatment group and the potential outcomes to control, denoted $Y_i(0)$ for those who were assigned to the control group.
4. The experimenter chose one *out of many possible* treatment assignments.
5. If the *sharp null hypothesis* were to be true (note the subjunctive verb tense there) then, the particular revelation of potential outcomes to treatment and control are inconsequential. Despite seeing only half the data (referred to as the **Fundamental Problem of Causal Inference**) we actually posess all the data. After all, if the sharp null were true, $Y_{Alex}(1) = Y_{Alex}(0)$, and $Y_{David}(1) = Y_{David}(0)$, $Y_i(1) = Y_i(0)$ for all of the $i = 1, \ldots, N$ people who are a part of the experiment.

3.5.2 Questions about Randomization Inference

- Where does uncertainty come from in an experiment that is evaluated using randomization inference?
- How is the ATE estimand defined?
- What is the feasible method that we use to write down an estimator (call it θ) for this quantity?
  - Which of the following properties does this feasible method possess?
    a. Unbiasedness: $E[\theta] = ATE$
    b. Convergence: $\theta \xrightarrow{p} ATE$, where $\xrightarrow{p}$ means converges in probability
    c. Efficiency: The mean squared error of θ is either (i) smaller than some other estimator, or (ii) as small is theoretically possible.

3.6 Applying Randomization Inference

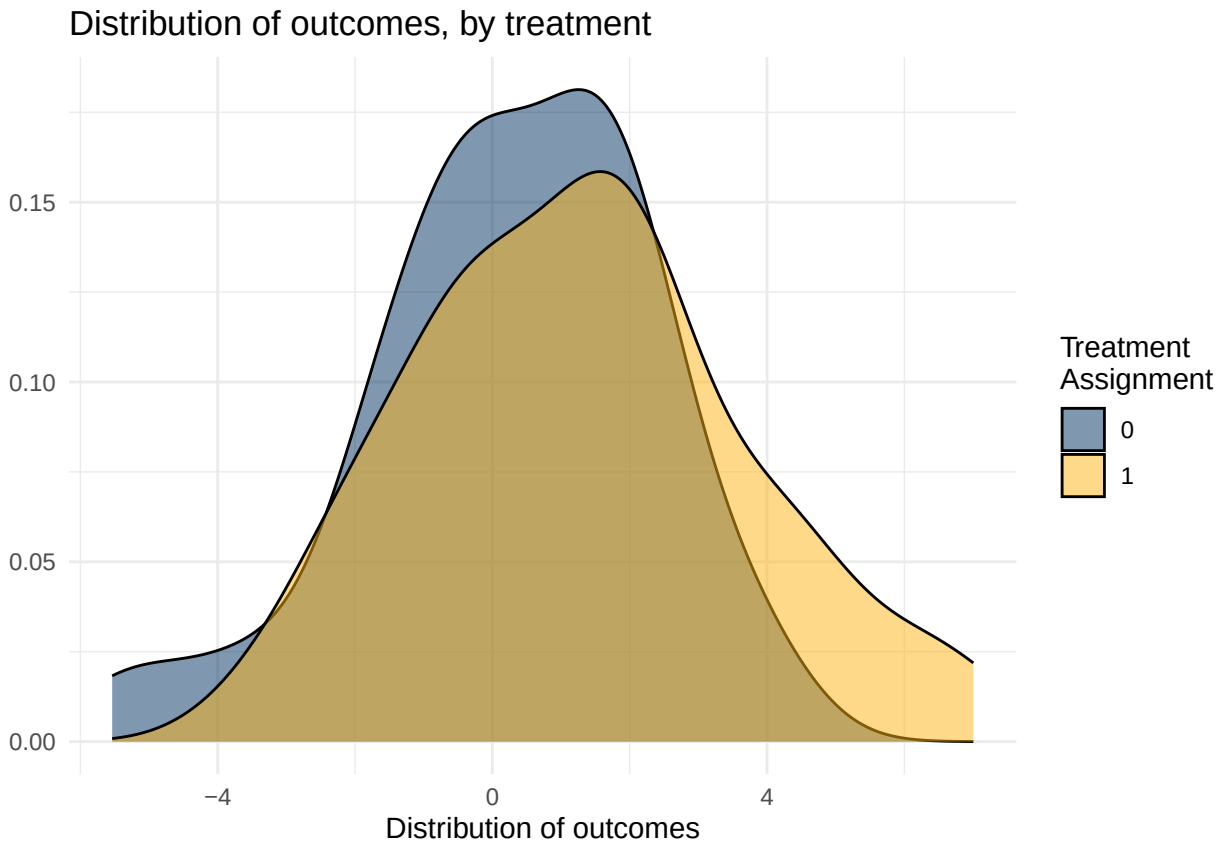
3.6.1 Make Data

```
set.seed(1)
d <- data.table(
  id = 1:100,
  D  = rep(0:1, each = 50),
  Y  = c(rnorm(n=50, mean=0, sd=2.5), rnorm(n=50, mean=1, sd=2.5))
)
```

3.6.2 Plot Data

In the following plot, are you able to assess whether there is a treatment effect simply by looking at the distributions?

```
ggplot(d) +
  aes(x=Y, fill=as.factor(D)) +
  geom_density(alpha=0.5) +
  labs(
    x        = 'Distribution of outcomes',
    y        = NULL,
    title    = 'Distribution of outcomes, by treatment',
    fill    = 'Treatment\nAssignment') +
  scale_fill_manual(
    values = c('#003262', '#FDB515')
  )
```



3.6.3 Classic Test

If you were to write a *classic* test against this data, given what you know about how it was generated, what would be the classic test? What do you learn from this test, and what is the interpretation?

```
d[ , t.test(Y ~ D)]
```

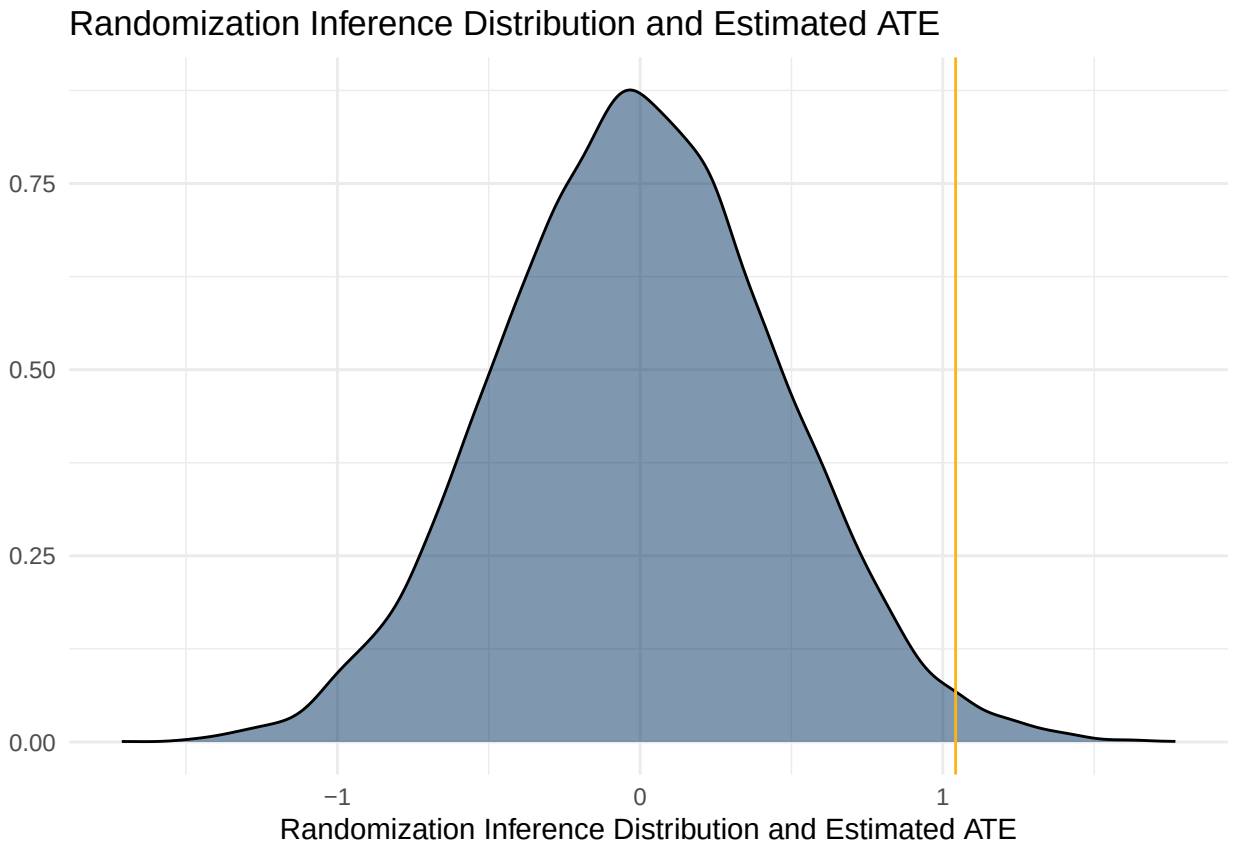
```
##
##  Welch Two Sample t-test
##
## data:  Y by D
## t = -2.309, df = 95.793, p-value = 0.02309
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.9381728 -0.1462181
## sample estimates:
## mean in group 0 mean in group 1
##       0.2511207       1.2933161
```

3.6.4 Randomization Inference Test

Now, instead suppose that you were to conduct the randomization inference. What are the steps to the algorithm for producing a result using randomization?

1. State the null hypothesis
2. Compute the statistic of interest using the observed data
3. Fill in data, under the statement of the null hypothesis
4. Permute the treatment assignment labels to generate a new sample of the treatment assignment vector, and then estimate the statistic of interest
5. Repeat the permutation and estimation (step 4) process repeatedly to sample from the randomization inference distribution of the statistic
6. Examine randomization inference distribution

```
## 1. The sharp null is that tau = 0
## 2. Compute the statistic of interest
true_ate <- d[ , .(group_mean = mean(Y)), keyby = .(D)][ , group_mean[D==1] - group_mean[D==0]]
## 3, 4, 5. Permute the treatment assignment labels and repeatedly compute the statistic of interest
ri_distribution <- replicate(
  n=10000,
  expr = d[ , .(group_mean = mean(Y)), keyby = .(ri_treatment = sample(D))][ ,
            group_mean[ri_treatment==1] - group_mean[ri_treatment==0]]
  )
# 6. Examine distribution
ggplot() +
  geom_density(aes(x=ri_distribution), fill = '#003262', alpha = 0.5) +
  geom_vline(xintercept = true_ate, color = '#FDB515') +
  labs(
    x     = 'Randomization Inference Distribution and Estimated ATE',
    y     =  NULL,
    title = 'Randomization Inference Distribution and Estimated ATE')
```



How much of the randomization inference is more extreme than the treatment effect?

```
ri_p_value <- mean(abs(ri_distribution) > abs(true_ate))
ri_p_value
```

```
## [1] 0.0226
```

Notice that 0.023 of the randomization inference distribution is more extreme than the observed treatment effect. How does this compare to the t-test p-value that we calculated above?

3.7 Comparing Randomization Inference and Frequentist Inference

If both Randomization Inference and Frequentist Inference produce similar p-values, what is utility in learning another set of methods for communicating estimator-based uncertainty?

What are the requirements (frequently referred to as "assumptions") that are necessary for the Frequentist paradigm to provide guarantees? What happens if these guarantees are not, in fact, satisfied or true in the data generating process? How do you react, respond, or address those problems?

- If data is not sampled *iid*, is it sufficient to simply note that limitation (frequently referred to as an "assumption violation") and report whatever p-value you report?
- How affected is this p-value by the violation? How do you know this?
- What does it mean for the p-value to be affected by this violation? (*Recall that a p-value is just a random variable that is produced through a series of summarizing transformations and then a comparison against a reference distribution.*)

3.7.1 Donations to a political campaign

In *Field Experiments* Green and Gerber provide some useful (hypothetical) data about donations to a political campaign. The data is defined in the following way, D is an indicator for whether the potential donor is

assigned to treatment or control, and Y is the outcome of how much the potential donor actually gave.

Let us provide a little bit more back story, that is necessary for the example to work, fully. Suppose that a progressive political candidate was hosting a fundraiser in Berkeley and has to make a choice about what to serve the attendees at the fundraiser.

In the $D = 0$ group, suppose that the candidate elects to serve a hippie-vegetarian staple, tofu sauteed in Bragg's liquid aminos. (It *is* Berkeley after all.) In the $D = 1$ group, suppose that the candidate decides to be a little more, well, progressive in their vegetarian food offerings and instead serves Gado-Gado from Katzen's *The Enchanted Broccoli Forest*. (Still Berkeley... .)

After dinner, and the requisite drum-circle, attendees to this shin-dig are asked to donate to the candidates re-election efforts. Every attendee is expected to contribute something – social norms rule out failing to donate when the collection plate is passed – but the amount donated is at the discretion of the attendee.

```
d <- data.table(
  id = 1:20,
  D  = rep(0:1, each = 10),
  Y = c(500, 100, 100, 50, 25, 25, 0, 0, 0, 0, ## tofu diners
        25,  20,  15,  15, 10, 5,  5, 0, 0, 0)  ## gado gado diners
)
```

1. With this data, conduct a `t.test` to assess whether the choice of dinner affects the amount donated to the campaign. What is your null-hypothesis (be specific), what is your rejection criteria, and do you reject or fail to reject this null hypothesis under the t-test framework.

```
## Null Hypothesis:
## Rejection Criteria:

## Conduct the Test Here:

## Conclusion:
```

2. With this data, use randomization inference to assess whether the choice of dinner affects the amount donated to the campaign. What is your null-hypothesis (be specific), what is your rejection criteria, and do you reject or fail to reject this null hypothesis under the t-test framework.

```
## Null Hypothesis:
## Rejection Criteria:

## Conduct the Test Here:

## Conclusion
```

1. Characterize the distribution of the sharp-null distribution of treatment effects. Talk about what, if anything, is notable about it, and what components of the data might be leading to any patterns that you note.
2. How many of the randomization inference loops are larger than the treatment effect that you calculated? How would you use this statement to construct a one-sided test, and an associated p-value?
3. How many of the randomization inference loops are *more extreme* (:metal:) than the treatment effect that you calculated? How would you use this statement to construct a two-sided test, and an associated p-value?
4. Compare the two-sided p-value against the p-value that you generate from a two-tailed t-test. If these p-values are the same, would this be a positive or a negative characteristic of randomization inference? If these p-values are different, why would they be different? Don't go looking all over hill-and-dale for the call for a t-test, it is at `t.test`.

5. Which of the two of these inferential methods do you prefer, randomization inference or a t-test, and why? Ease of use is not an acceptable answer.

3.8 Statistical Power

- What is statistical power?
- Why is it particularly relevant to consider statistical power when you are thinking about conducting an experiment?
  - What would happen if you were to conduct an experiment that has only an achieved power of 0.1?
  - What would you learn if you were to fail to reject the sharp-null hypothesis?
  - What would you learn if you were to reject the sharp-null hypothesis?

```
make_data <- function(
  sample_size                       = 100,
  potential_outcome_to_control_mean = 10,
  potential_outcome_to_control_sd   = 2,
  treatment_effect                  = 1,
  sd_treatment                      = 2) {
  ## this is a function to make data to simulate the power of a test

  }

test_data <- function(data, treatment_indicator, outcome) {

}

## p_values <- replicate(n = 1000)
```

4 Blocking and Clustering

When assigning treatment to units, unless there are restrictions created by the researcher, any of the treatment assignment vectors are equally probable. Blocking and clustering are ways of restricting the treatment assignments to a subset of the whole schedule of possibilities.

Blocking is a method of creating "blocks", or groups, of units that are similar along one or more dimensions and then creating a full random assignment within each of those similar groups. Through careful design, blocking can generate power or nuance for an experiment without any extra marginal costs for paying for additional units of treatment.

Clustering is a circumstance that arises from a state of the world that *requires* you to assign several similar units to the same condition, be it treatment or control. Through careful design, clustering might not hamper the power of an experiment; though realizing the necessity of a clustered design is typically met with the following statement, "@#$%, we've got to cluster."

4.1 Learning Objectives

At the end of this week, student will be able to

1. **Recognize** when there is the potential to block random assign in their experiment, and **remember** why block random assignment beneficial.
2. **Recognize** when they are required to cluster random assign – either due to a pragmatic (i.e. real-world) limitation, or to avoid violating the requirement that units not interfere with one another – and **identify** ways that they can mitigate the reduction-in-power that arises from the need to cluster.
3. **Distinguish** between the circumstances that lead to blocking and clustering.
4. **Analyze** both blocked and clustered experiments using the appropriate test, and generating statements of certainty and uncertainty using *randomization inference*.

4.2 Setting terms: Blocking

- What does it mean to block randomize?
- Does the elimination of some randomization mean that the randomization is not longer, well, random?
- Relative to when treatment is administered, when are we able to block? Why are we not able to block after we've assigned treatment?

4.3 Math: Block random assignment

In equation 3.6 (on page 61) of *Field Experiments* Green and Gerber write,

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i(0))}{N-m} + \frac{\widehat{Var}(Y_i(1))}{m}}$$

When we block randomize, we're essentially creating smaller groups of units and producing an estimate of the variance within each of those smaller groups of units.

How do the authors arrive at the following formula for a block randomized standard error?

$$\widehat{SE}(\widehat{ATE}_{blocked}) = \sqrt{\sum_{j=1}^{J} \left(\frac{N_j}{N}\right)^2 * \widehat{SE}^2(\widehat{ATE}_j)}$$

- Specifically, why are we squaring the scaling parameter $\frac{N_j}{N}$?
- If you look at this summation, what has to happen to the variance within the groups, relative to the size of the groups, in order for blocking to actually increase power?
- Is it possible that you block, without increasing power, even if the blocking variable is actually useful?

Green and Gerber, in equation 3.10, write that the overall ATE of the population is:

$$ATE = \sum_{j=1}^{J} \frac{N_j}{N} ATE_j$$

- What does this equation "feel like"? Does that seem reasonable? Why or why not?
- Why might it be a good idea to have different rates of assignment to treatment within different blocks? Consider the following example:
  - Suppose that you are looking at an experiment among your whole user base, and you are considering changing the "check out flow" (we have not idea what that might mean either...) for this group.
  - Some of the users are *really* likely to purchase, while others are very unlikely to purchase.
  - Does it make sense to block randomize based on this prior purchase history?
  - Are there any, reasonable business reasons to not make the treatment assignments be 50% treatment and 50% control in both of the populations?
  - What would happen if you randomized 10% of the "high value" customers into treatment and 50% of the low value customers into treatment. But, then you forget (or lost) that table of whether they were "high" or "low" value customers.
  - *What would be the consequence to your treatment effect estimate?*

4.4 Intuition: Block Random Assignment

To discuss the idea of blocking, consider the working example that David and David present in the async lectures:

Eating too much tofu (aka the *Berkeley diet*) might increase decrease one's brain function, leading to decreased performance on cognitive tests, lower brain weight, and cause ventrical enlargment of the brain.

Don't ask your instructors what any of that medical jargon might mean. It isn't our field! But, these are real claims made by a group of researchers in an observational nutrition study titled "Brain Aging and Midlife Tofu Consumption."

Original Research

Brain Aging and Midlife Tofu Consumption

Lon R. White, MD, MPH, Helen Petrovitch, MD, G. Webster Ross, MD, Kamal Masaki, MD, John Hardman, MD, James Nelson, MD, Daron Davis, MD, and William Markesbery, MD

National Institute on Aging, NIH (L.R.W., formerly), Pacific Health Research Institute (L.R.W., H.P.), University of Hawaii at Manoa (L.R.W., H.P., G.W.R., K.M., J.H.), Department of Veterans Affairs, Honolulu, (L.R.W., G.W.R.), Kuakini Medical Center, Honolulu, (H.P., K.M.), Hawaii, Louisiana State University (J.N.), Baton Rouge, Louisiana, and the University of Kentucky (D.D., W.M.), Lexington, Kentucky

Key words: brain, aging, nutrition, soy, cognition

Objective: To examine associations of midlife tofu consumption with brain function and structural changes in late life.

Methods: The design utilized surviving participants of a longitudinal study established in 1965 for research on heart disease, stroke, and cancer. Information on consumption of selected foods was available from standardized interviews conducted 1965–1967 and 1971–1974. A 4-level composite intake index defined "low-low" consumption as fewer than two servings of tofu per week in 1965 and no tofu in the prior week in 1971. Men who reported two or more servings per week at both interviews were defined as "high-high" consumers. Intermediate or less consistent "low" and "high" consumption levels were also defined. Cognitive functioning was tested at the 1991–1993 examination, when participants were aged 71 to 93 years (n = 3734). Brain atrophy was assessed using neuroimage (n = 574) and autopsy (n = 290) information. Cognitive function data were also analyzed for wives of a sample of study participants (n = 502) who had been living with the participants at the time of their dietary interviews.

Results: Poor cognitive test performance, enlargement of ventricles and low brain weight were each significantly and independently associated with higher midlife tofu consumption. A similar association of midlife tofu intake with poor late life cognitive test scores was also observed among wives of cohort members, using the husband's answers to food frequency questions as proxy for the wife's consumption. Statistically significant associations were consistently demonstrated in linear and logistic multivariate regression models. Odds ratios comparing endpoints among "high-high" with "low-low" consumers were mostly in the range of 1.6 to 2.0.

Conclusions: In this population, higher midlife tofu consumption was independently associated with indicators of cognitive impairment and brain atrophy in late life.

Figure 7: this is your brain on tofu

Suppose that, motivated by your distaste for bunk, casual causal claims about diet, and taste for tofu, you decide to conduct a real experiment among your friends, families, and classmates to determine the actual impacts of tofu on diet.

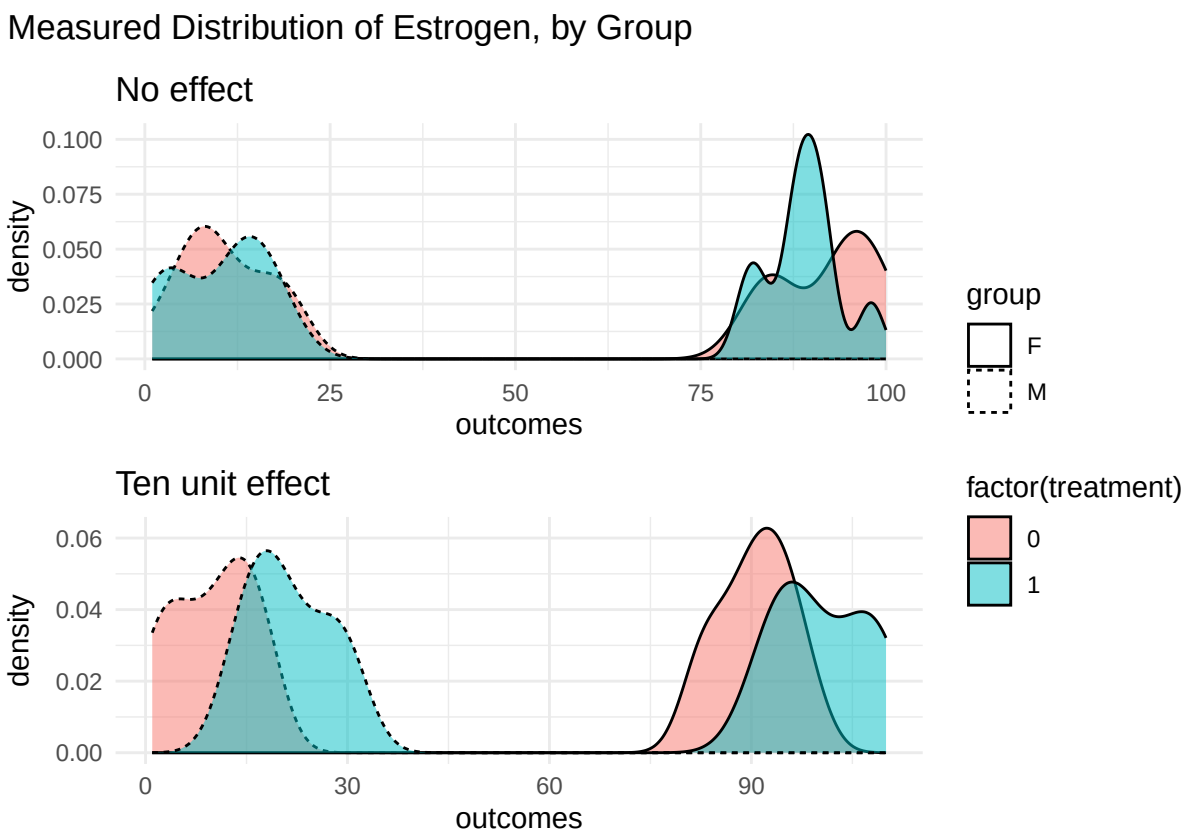
```
set.seed(1414)

sim_normal_study <- function(treatment_effect=0) {
  ## this function will create a "world" to analyze using an experiment,
  ## then, it will estimate the ate within that world
  ## it returns the ate and the number of women who are in treatment

  require(data.table)
```

```
  d <- data.table(
      group        = rep(c('M', 'F'), each = 20),
      po_control   = c(1:20, 81:100),
        ## treatment_effect = 0 --> sharp null is true
      po_treatment = c(1:20, 81:100) + treatment_effect,
      treatment = sample(1:0, size = 40, replace = TRUE))[ , ## notice we're now assigning
      outcomes := po_treatment * treatment + po_control * (1 - treatment)]

  ate <- d[ , mean(outcomes[treatment == 1]) - mean(outcomes[treatment == 0])]
  n_women_treatment = d[treatment == 1 & group == 'F', .N]

  return(list(
    data = d,
    ate = ate,
    n_women_treatment = n_women_treatment
    ))
}
```

4.5 With this data, what does the distribution of outcomes look like?

```
experiment_one <- sim_normal_study(treatment_effect = 0)
experiment_two <- sim_normal_study(treatment_effect = 10)
```

```
experiment_one_plot <- ggplot(data = experiment_one$data) +
  aes(x = outcomes, fill = factor(treatment), linetype = group) +
  geom_density(alpha = 0.5) +
  labs(title = 'No effect'
  )
experiment_two_plot <- ggplot(data = experiment_two$data) +
  aes(x = outcomes, fill = factor(treatment), linetype = group) +
  geom_density(alpha = 0.5) +
  labs(title = 'Ten unit effect'
  )

(experiment_one_plot / experiment_two_plot) +
  plot_annotation(title = 'Measured Distribution of Estrogen, by Group') +
  plot_layout(guides = 'collect')
```



In these two different cases – where there is no treatment effect on top, and when there is a large treatment effect on bottom – what are the group means? Where would they be on these plots?

Consider the formula for the SE_{ATE}.

$$SE(\tau) \approx \sqrt{\frac{V[\tau]}{N}}$$

The important parts to consider for this discussion (despite being not a full statement of the SE) is that the standard error of the difference of group averages is a ratio of the underlying variance of the treatment effect, divided by the number of observations in that group.

$$\begin{aligned} SE[\tau] &\approx \sqrt{\frac{V[Y(1)]}{n_1} + \frac{V[Y(0)]}{n_0}} \\ &\approx \sqrt{\frac{E[(Y(1) - E[Y(1)])^2]}{n_1} + \frac{E[(Y(0) - E[Y(0)])^2]}{n_0}} \end{aligned}$$

- When you examine the plot above, what are the expected values of the treatment and control groups?
- What does the expected value of the square of the deviations look like on this plot?

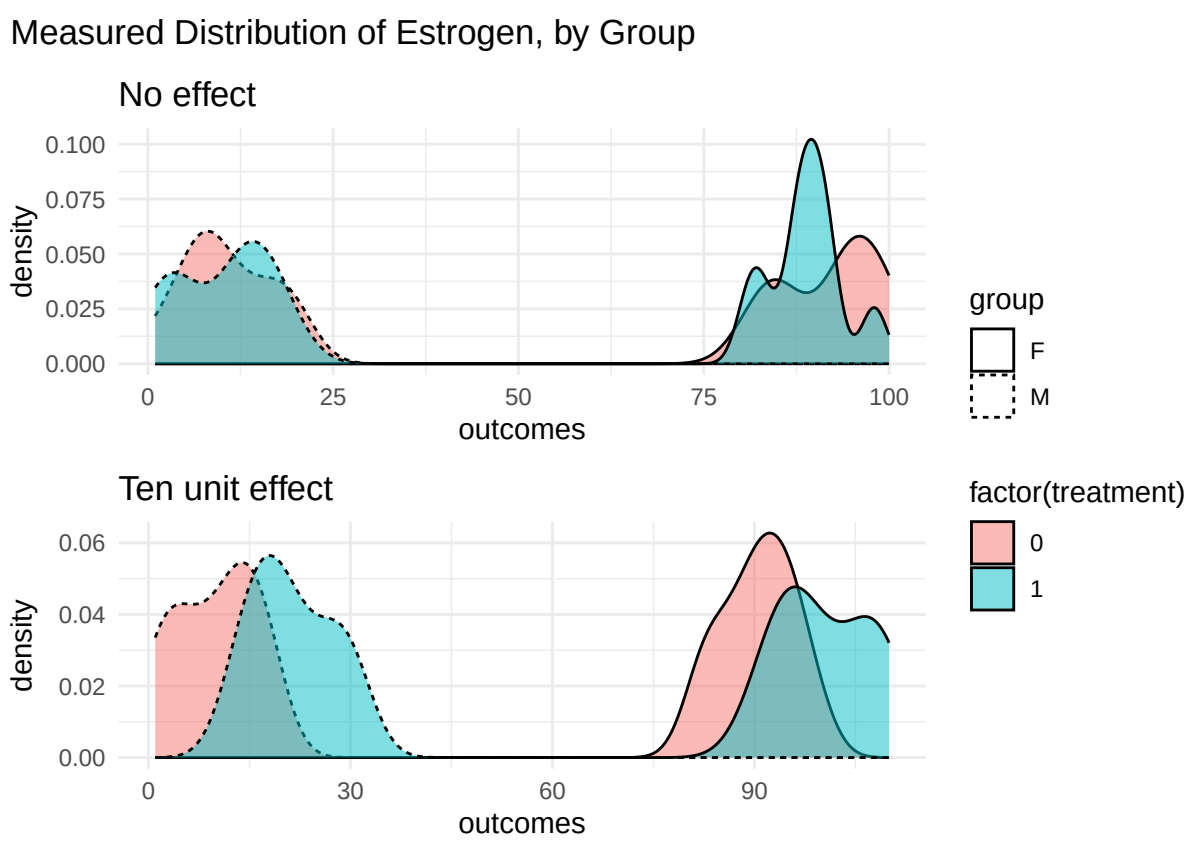
4.6 Technical Benefits of Blocking

How how does breaking this population into two smaller groups create a reduction in the calculated standard error that you observe from an experiment?

- What is (draw) the conditional expectation among the `M` group and the `F` group.
- What is (draw) the conditional variance among the `M` group and the `F` group.

- How has this change produced a reduction in the overall variance?

```
experiment_one_plot <- ggplot(data = experiment_one$data) +
  aes(x = outcomes, fill = factor(treatment), linetype = group) +
  geom_density(alpha = 0.5) +
  labs(title = 'No effect'
  )
experiment_two_plot <- ggplot(data = experiment_two$data) +
  aes(x = outcomes, fill = factor(treatment), linetype = group) +
  geom_density(alpha = 0.5) +
  labs(title = 'Ten unit effect'
  )

(experiment_one_plot / experiment_two_plot) +
  plot_annotation(title = 'Measured Distribution of Estrogen, by Group') +
  plot_layout(guides = 'collect')
```



4.7 How should we block randomize?

Let's take several discussion points, in order:

4.7.1 What makes a useful feature? (part 1)

- When we are considering a block randomization to improve the *power* of a test, what about a feature makes it a useful blocking feature? (For instructors, probably don't read each of these, but try to get the discussion to address them.)
  - Does a good blocking feature have to be associated with the treatment?
  - Does a good blocking feature have to be associated with potential outcomes?

  - Does a good blocking feature have to have a causal effect on the measured outcomes?

- Suppose that have two possible features that you could use to block in the estrogen experiment. Either, you can block randomize using:

  (a) blood-serum levels of estrogen, measured a week before the experiment begins; or (

  b) "stated form" sex (i.e. female, male, nonbinary).

4.7.2 What makes a useful feature (part 2)

- In the async, and to this point in this live session, we have spoken only about features that are categorical for blocking.
- Is it possible to block on a continuous feature?
  - What if it were measured very, very precisely, so every unit had a unique value on a continuous variable?
  - If you *could* develop a method of blocking on a continuous variable, what might be the benefits?

4.7.3 Strategies of blocking

- If there is a benefit of creating two mini-experiments through blocking – as you have proposed in the code above – could there be a benefit to creating a third mini-experiment through blocking? What about a fourth? Is there a limit that you run into?
  - What is the most blocks that you can produce in an experiment?
  - Or, alternatively, what is the smallest size block that you can produce in an experiment?
  - Is there a reason to take this strategy?
  - What if you created many blocks, but with a noisy blocking feature. Would this work well?
  - What if you created many blocks, but with a very precise blocking feature. Would this work well?
- To this point, we have discussed blocking on only a single variable. Is it possible to block on more than one variable at a time?
  - If you have already blocked on one variable, what are the characteristics that are useful for the next variable that you consider blocking on?
  - For example, suppose that you have already blocked the tofu experiment on experimental units' stated-form sex. Would it be useful to then block based on wearing glasses, or hair length, or blood-serum estrogen? Why or why not?

4.8 Clustering

- What are the circumstances in the world that make it necessary to cluster random assign?
- Are these circumstances academic? Or, are there actually examples of where this might come into play?
  - Consider the ride sharing example that we read about in *Power of Experiments*. What would happen if we gave some people really low prices to get into a rideshare, while we gave other people really high prices? What if they are standing next to eachother at the airport? What if one is at an airport in Oakland, while the other is at SFO?

4.9 Blocking or Clustering?

4.9.1 Let is snow!

Suppose we want to measure the effect of snowplowing on local retail activity. We design an experiment that plows some locations but not others. Which of the following do you prefer? Explain the relative advantages and disadvantages of each option.

- On a given street, we randomly assign which businesses we plow in front of.
- We randomly assign which streets to plow and which streets not to plow.

- We randomly assign which neighborhoods to plow and which neighborhoods not to plow.
- Do the differences above illustrate blocking, or clustering?

Returning to the snowplow example, suppose we have two wealthy neighborhoods, nine middle-class neighborhoods, and four poor neighborhoods available to experiment on. We are worried that if we put both of the wealthy neighborhoods into the treatment group, we will get an overestimate of the treatment effect of snowplowing on retail activity. We will assign treatment at the neighborhood level. Now consider blocking this experiment based on social class. Describe treatment assignment for the fifteen neighborhoods.

- Does blocking reduce bias?
- What benefit do we expect blocking to have on our ATE estimator?

4.9.2 Strolling through Berkeley

David Reiley walks through Berkeley and observes retail shops. As he goes, he takes each pair of stores he encounters, flips a coin, and goes into one store in each pair to give them a free Google ad coupon. He later observes how much each spent on Google ads in the month after.

- Why might this increase power compared to picking stores totally at random?
- Reiley does the same as above, but picks one store on every street only.
- Reiley does the same as above, but picks two stores on every street only.
- Reiley picks one side of each street to treat on many streets.

4.9.3 Always low prices?

Imagine that an executive at Walmart gives you the keys to the pricing at the store and asks you to determine how demand for goods changes depending on the pricing of those goods? Basically, does "rolling back prices" lead to increased demand? And by how much?

- What are the different levels at which you could assign different prices?
- What are the benefits and limitations of assigning different prices at those levels?

5 Covariates and Regression

Adding covariates to what we're measuring, even if those covariates are non experimental, can help us improve our measurement."

5.1 Learning Objectives

1.
2.
3.

5.2 Covariates

Covariates as we will call them in this unit are are supplemental variables that do *not* have a causal meaning but which might predict the outcome variable. Because treatment is randomly assigned in an experiment, covariates are not required in order to generate for unbiased inference in an experiment, but including covariates in our estimation of a treatment effect might *improve* the precision of estimates.

Typically, covariate adjustment happens through the use of a regression. Blocking (discussed last week) is doing the mechanically the same thing as regression, but blocking possesses the beneficial guarantees that all blocks will have good random assignment.

One important point about covariates: For the appropriate use of covariates in an analysis of an experiment, the covariates *must* not change as a consequence of treatment assignment. If they change,then they are a

down-stream consequence of the experiment, and therefore are a "result" of the experiment. (In the future, we will talk about why these are 'bad controls'.)

5.3 Rescaling Outcomes

Suppose that in your design, you are able to measure every unit twice, once before they are exposed to treatment, and again after they are exposed to treatment.

Suppose that we have the following grammar, or notation to describe the experiment:

- `R` is a indicator for a randomization process.
- `N` is an indicator for a non-randomization process.
- `X` is an indicator that we have provided treatment to a unit.
- `O` is an indicator that we have provided control to a unit.
- `Y` is an indicator that we have made a measurement of a unit.

With these operators set up, we can think about three different experiment designs.

5.3.1 Design One: Two Group Post-Test

To this point, and in nearly every essay proposed for the first assignment in the class, student had in mind a two group, post-test only design. In this experiment design, we randomize an experimental population into two groups, assign treatment to one of these groups, and then observe outcomes. In many ways, the one group pre-post design is the simplest design to implement.

```
R X Y
R O Y
```

5.3.2 Design Two: One Group, Pre-test Post-test

In this case, we take the units that are a part of our experiment, expose them to control and measure these units outcomes, and then expose them to treatment and measure these units outcomes.

We might write out this design in the following way:

```
N O Y1 X Y2
```

Does this meet the base definition of an experiment that you've written about in your homework? Would David Reiley think that this is an experiment? Would Green and Gerber think that this is an experiment?

5.3.3 Evaluate the strengths of the two designs

Under what circumstances would you prefer to one or another of these two designs?

- Suppose that you are attempting to learn what part of your code on problem set 2 is leading to a Latex compile error. Which of the experiments would you propose to undertake?
- Suppose that you are attempting to learn the effects of giving a birthday gift to twins where measurement is magically easy.
- Suppose that you are attempting to learn about the effect of coffee on alertness, measured as the number of characters written down while attending async lectures.
- Suppose that you are attempting to learn about the effect of coffee on alertness, measured through galvanic skin conductance?

Are there general principles, or circumstances that lead you to go one way or another?

5.4 Combining Designs?

By combining the two previous designs, it is possible to develop a new design that contains the benefits of each.

5.4.1 Design Three: Two Group, Pre-test Post-test

In this case, we randomize into two-groups, but we also measure each unit more than once.

```
R O Y1 X Y2
R O Y1 O Y2
```

This design has the benefit of the apples to apples comparison created through randomization, but additionally adds the improvement in measurement that are possible by re-scaling the outcome variable into a difference score. If we redefine the outcome to be $\delta = Y_2 - Y_1$, and if there is a covariance between Y_1 and Y_2, which seems reasonable for many cases where the unit has "sticky" behaviors, then we are able to produce estimates of δ that are more precise because they use this "stickyness" (i.e. covariance).

Even in the case when we don't know *why* outcomes are correlated through time, we can still us this relationship profitably to produce estimates with smaller standard errors.

5.5 Working with simple data

In *Field Experiments* on page 74, Green and Gerber provide a table of potential outcomes for community public works projects. In the `Village` varaible is an index from 1-14 of the village id; in the `Y` variable is the outcome if assigned to control; in the `D` variable is the outcome if assigned to treatment; and in the `Block` variable is a variable that indicates the block where the unit was located.

```
d <- fread('http://hdl.handle.net/10079/cf1a6ba7-1603-4b36-ab18-1a7e81a63990')
head(d)
```

```
##    Village     Y     D Block
##      <int> <int> <int> <int>
## 1:       1     0     0     1
## 2:       2     1     0     1
## 3:       3     2     1     1
## 4:       4     4     2     1
## 5:       5     4     0     1
## 6:       6     6     0     1
```

Although this will produce numbers that are different than are reported in the book (because R implements sample variance and covariance, and the book instead uses population variance and covariance) compute the following.

5.5.1 Without blocking

1. Compute the variance of the potential outcomes to control, the variance in the potential outcomes to treatment, and the covariance between treatment the potential outcomes to treatment and control.
2. With these values, then, compute the standard error of the ATE.

5.5.2 With blocking

1. Compute the variance of potential outcomes to control within each block, the variance in the potential outcomes to treatment in each block, and the covariance between the treatment and control potential outcomes.
2. With these values, then compute the standard error of the blocked ATE.

5.6 Using Measurements to Diagnose Problems

6 Regression and Multifactor Experiments

```
theme_set(theme_minimal())
```

6.1 Learning Objectives

At the end of today's session, student will be able to

1. **Understand** the difference between good and bad controls, and **evaluate** whether a control variable is one or another.
2. **Articulate** the importance of asking "why" and how this enables search for multifactor experiments.
3. **Analyze** multifactor experiments using best-in-class linear models.
4. **Appreciate** that the model does not generate interpretation; design does.

6.2 Design Notation

This week, you read three very short chapters in a book by Trochim and Donelly. This reading begins with a series of one-group "threats" to causal inference, which we will enumerate again here:

- *History threat*
- *Maturation threat*
- *Testing threat*
- *Instrumentation threat*
- *Mortality threat*
- *Regression threat*

Many of these contain a plain language statement of a problem that might arise from an experiment design. For example, a maturation threat might mean that as your subjects get older or more experienced through the experiment, they may do better (or worse) at the task that they are being asked to undertake. This isn't an academic-only concern, this is something that is actually likely to happen if you measure performance over a long period of time.

The author then moves on to describe several multiple-group threats. Notice that each of these multi-group threats are simply "selection-" version of the threads that we have already enumerated.

> How to do we ensure that we do not witness any of the problems created by these *selection-*threats?

6.2.1 Design Notation

Finally, the authors introduce us to the real point of this week: design notation whereby they provide us with a constrained set of actions that can be taken.

```
R
O
Y
N
X
```

6.3 Good Controls

6.4 Bad Controls

What goes wrong with bad controls? Everything!

6.5 A Very Simple Example

6.6 Make Data

Let's make some data in just the same way that we typically make data. We will produce a vector of potential outcomes to control, and then two outcomes that are affected by treatment. One we will consider the outcome that we are interested in understanding as a causal effect, the other, we're going to call the "bad control".

```
make_data <- function(n_rows=1000) {

  d <- data.table(
    id = 1:n_rows,
    key = 'id'
    )

  d[, ':='(
    ## each of these are independent of all others
    y0           = runif(min=-10, max=10, n=.N),
    tau          = rnorm(n=.N, mean=4),
    epsilon      = rnorm(n=.N, mean=0, sd=2),
    D            = sample(x=0:1, size=.N, replace = TRUE)
  )]

  ## send 1/2 the effect through the bad control
  ## and the other 1/2 through a direct channel
  d[, bad_control := .5*tau*D + .2*epsilon]
  d[, Y := y0 + bad_control + .5*tau*D + .8*epsilon]

  }

d <- make_data(n_rows=10000)
```

6.7 What is the causal model we hold?

When we are thinking about the causal model here, we're saying, "I think that the conditional expectation of Y depends on the treatment status". Or, even more simply, "Treatment affects outcomes."

But, maybe I think I want to also control for the variable `bad_control`, despite the scary name that it has in the dataset.

In fact, we can estimate a reliable causal effect for *either* `Y ~ D` *or* `bad_control ~ D`, but not the two together.

```
model_1 <- d[ , lm(Y ~ D)]
model_2 <- d[ , lm(Y ~ bad_control )]
model_3 <- d[ , lm(Y ~ D + bad_control)]
```

```
stargazer(
  model_1, model_2, model_3,
  type = 'text',
  se = list(rse(model_1), rse(model_2), rse(model_3)),
  omit.stat = c('ser', 'f')
)
```

```
##
## ============================================
##                   Dependent variable:
##               ---------------------------
##                             Y
##                  (1)       (2)       (3)
## --------------------------------------------
## D            3.962***            -3.350***
##               (0.123)             (0.248)
```

```
##
## bad_control            2.354***  3.659***
##                         (0.052)   (0.110)
##
## Constant       0.022   -0.357***   0.002
##               (0.085)   (0.078)   (0.081)
##
## ----------------------------------------
## Observations  10,000    10,000    10,000
## R2             0.094     0.169     0.184
## Adjusted R2    0.093     0.169     0.184
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

6.7.1 Do the estimates match the world?

When you look at what the models have estimated, do they match the data that we created above?

6.8 Robust Standard Errors

David R. makes the good point in the async material that if we don't have a good reason to assume that the variance is the same between different groups, or really across all values of our explanatory variables, then these variances might, in fact be different! As a consequence we might have overly optimistic estimates of our standard errors.

Why would this be bad? As we've said in the past, if we only want to falsely reject the null hypothesis in 5% of cases due just to chance (roughly an equivalent thought to a 95% confidence interval), then if our standard errors are wrong, there is the possibility that we falsely reject the null more frequently.

So, we think we're only making this type of mistake in 5% of cases to to random chance, but perhaps we're actually making this type of mistake in 20% of cases. Why would this be bad? Remind yourself?

Luckily, it is pretty easy to estimate robust standard errors. In fact, acknowledging heteroskedasticiy does not have ANY effect on the location of our estimates of the relationships between variables. What does this mean? It means that the estimated β_1 that you pull off of some regression is the same whether you are using homoskedastic or heteroskedastic-consistent standard errors.

What is actually happening when we compute HCE? Well, rather than presuming that all the residuals are the same, instead we're actually calculating those residuals from from the regression line. What is the penalty we pay for this? Well, in the case of homoskedastic error, we have a slightly less efficient estimator (which makes our findings more conservative when they don't need to be). And because we're estimating things, we're burning a few degrees of freedom.

Otherwise though, there isn't really *that* strong a penalty to pay.

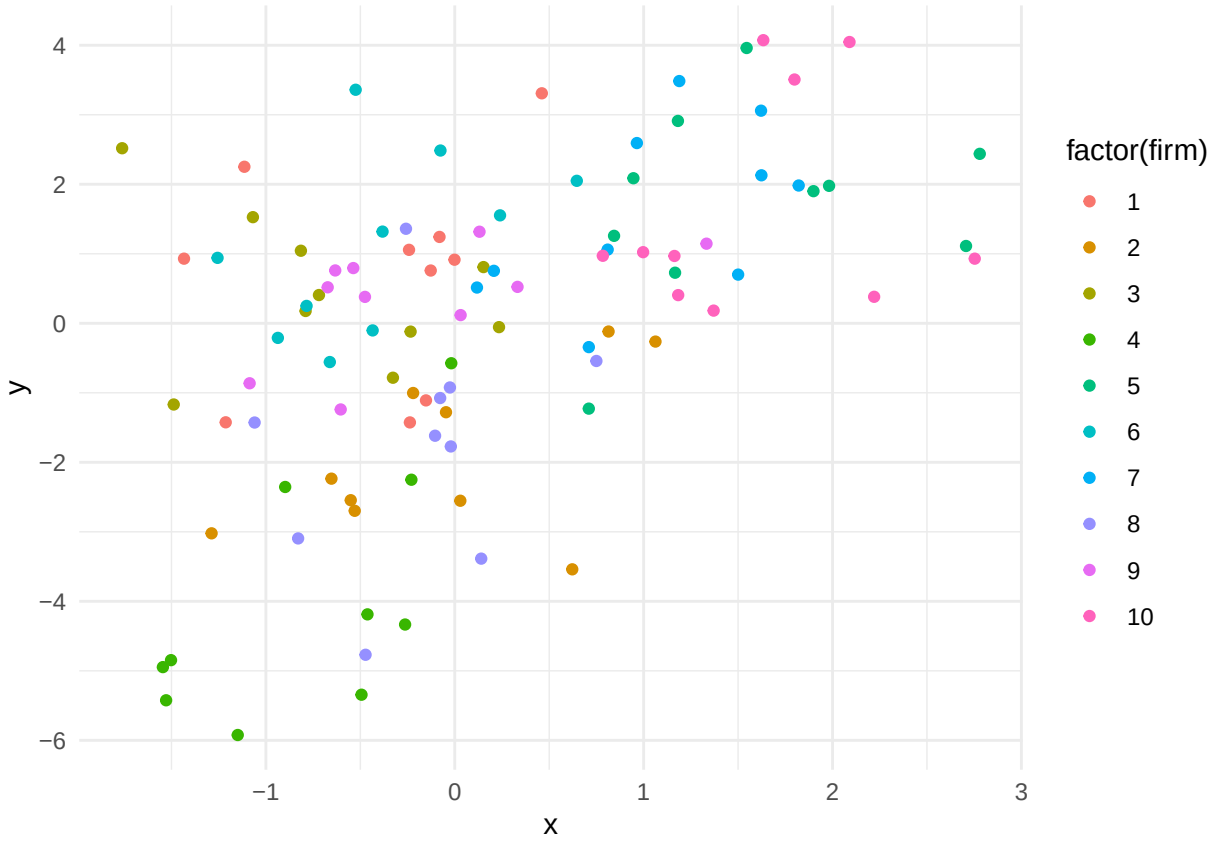
We're going to load data that has a clustering structure. This data is due to a simulation, initially written by Petersen (2009). In this simulation, there are repeated observations of firms for ten years. This post on R-bloggers, replicates the data, in case you're curious about the specific clustering strucutre.

We're using this data because (a) it has robust standard error considerations; and (b) it has clustering considerations.

```
library(sandwich) # estimates RSE easily
library(lmtest)   # sets up t-test easily

data('PetersenCL', package = 'sandwich')
pcl <- data.table(PetersenCL)
head(pcl)
```

```
##     firm  year           x          y
##    <int> <int>       <num>      <num>
## 1:     1     1 -1.11397266  2.2515347
## 2:     1     2 -0.08085376  1.2423458
## 3:     1     3 -0.23760724 -1.4263762
## 4:     1     4 -0.15248568 -1.1093940
## 5:     1     5 -0.00142616  0.9146864
## 6:     1     6 -1.21273661 -1.4246863
```

```
ggplot(pcl[firm <= 10]) +
  aes(x=x, y=y, color = factor(firm)) +
  geom_point()
```




```
model_1 <- pcl[ , lm(y ~ x)]

## since i have the lmtest loaded; i can call:
coeftest(model_1, vcov = vcovHC(model_1, type = 'const'))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.029680   0.028359  1.0466   0.2954
## x           1.034833   0.028583 36.2041   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## to estimate a robust se is a one line solution
coeftest(model_1, vcov = vcovHC(model_1, type = 'HC3'))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.029680   0.028366  1.0463   0.2955
## x           1.034833   0.028412 36.4223   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These two packages are recommended packages and are **extremely** well used in R. I've been harping on `data.table` as abig deal, and it is. Lots of people use the frameowrk and it is great. But these two packages – `sandwich` and `lmtest`, are **core**. There is no disputing that.

There is a specific relationship between the variance-covariance matrix and the standard error. in fact, it is very much like the relationship between the variance and standard error in any other application we've examined so far.

This relationship is the following:

$$SE(\hat{\beta}) = \sqrt{diag(vcov)}$$

So, all we're really doing is making a post-estimation correction to the variance covariance matrix, and then dividing by this new standard error. Quite straightforward. Why would you want to know this little bit? If you're going to run the test yourself, you will want to be able to pull off the SEs from the `vcovHC` object.

```
t.numerator   <- coef(m2)
t.denominator <- sqrt(diag(vcov(m2)))
t.denominator.robust <- sqrt(diag(vcovHC(m2, type = "HC1")))

# t.ratios:
# not robust:
t.numerator / t.denominator
# robust
t.numerator / t.denominator.robust
```

But, like as I showed earlier, we can wrap all this up with the `lmtest` package's call `coeftest`.

```
# coeftest(m2, vcov(m2))
# coeftest(m2, vcovHC(m2))
```

What if we wanted to pretty-print ourselves a table? If we are using stargazer, or other packages, we will need the SEs off that model.

```
# m2.se  <- sqrt(diag(vcov(m2)))
# m2.rse <- sqrt(diag(vcovHC(m2, type = "HC1")))
#
# stargazer(m2, m2, se = list(m2.se, m2.rse),
#           type = "latex", header = FALSE)
```

6.9 What about clustered standard errors?

Ok, now we're a little deeper down the rabbit hole.

As we've talked about, clustered standard errors acknowledge that you've got treatment assigned at the cluster level, and that there may be significant covariance in potential outcomes at that cluster level. If this is the case, then we have functionally fewer observations than we have nominally, and we also have less power to detect an effect.

To estimate clustered standard errors, we use `sandwich::vcovCL`.

```
model_1 <- lm(y ~ x, data = pcl)

## without clustering
coeftest(model_1, vcovHC(model_1, type = 'const'))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.029680   0.028359  1.0466   0.2954
## x           1.034833   0.028583 36.2041   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## when we cluster
coeftest(model_1, vcovCL(model_1, ~ firm, type = 'HC3'))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.029680   0.067143   0.442   0.6585
## x           1.034833   0.050816  20.364   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pretty print that.

```
stargazer(m1, m1, m1,
          se = list(sqrt(diag(vcov(m1))),
                    sqrt(diag(vcovCL(m1, ~ firm))),
                    sqrt(diag(vcovCL(m1, ~ firm + year)))),
          type = 'text',
          header = FALSE)
```

6.10 Treatment by Treatment Interaction

Summarized data is... a drag.

In the book, we're provided data about responses to questions from purported constituents. These people who are writing letters are either names "Colin" or "José" and who are either writing with "good" or "bad" grammar.

But, the book gives us data of the form:

	Colin	Colin	José	José
% Received Reply	52	29	37	34
(N)	(100)	(100)	(100)	(100)

Can you run inference against a table that is structured like that? How would you run a model against that form of "data"?

6.10.1 Recreate Data

In order to get a model running against this data, we're going to make a dataset that has the same information in it, but that we can actually run a model against.

1. To begin with, what are the units of observation?
2. What are the features in the dataset?
3. What is the outcome in the dataset and how is it coded?

Once we've been able to name these, we're able to make a `data.table` that matches this format.

```
dat <- data.table(y       = rep(NA, 400),
                  name    = rep(NA, 400),
                  grammar = rep(NA, 400) )

dat[ , y := c(rep(1, 52), rep(0, 48),
              rep(1, 29), rep(0, 71),
              rep(1, 37), rep(0, 63),
              rep(1, 34), rep(0, 66) )]

dat[ , name    := as.factor(c(rep("C", 200), rep("J",200)))]
dat[ , grammar := as.factor(rep(c("G", "B", "G", "B"), each = 100) )]

dat[ , ':='(cg = I(name == "C") * I(grammar == "G"),
           cb = I(name == "C") * I(grammar == "B"),
           jg = I(name == "J") * I(grammar == "G"),
           jb = I(name == "J") * I(grammar == "B"))
   ]

dat
```

```
##          y   name grammar     cg     cb     jg     jb
##      <num> <fctr>  <fctr> <AsIs> <AsIs> <AsIs> <AsIs>
##   1:     1      C       G      1      0      0      0
##   2:     1      C       G      1      0      0      0
##   3:     1      C       G      1      0      0      0
##   4:     1      C       G      1      0      0      0
##   5:     1      C       G      1      0      0      0
##  ---
## 396:     0      J       B      0      0      0      1
## 397:     0      J       B      0      0      0      1
## 398:     0      J       B      0      0      0      1
## 399:     0      J       B      0      0      0      1
## 400:     0      J       B      0      0      0      1
```

If we've got a `data.table` that matches the format, can we then estimate models that correspond to the models that are written in the book?

The first of these models, which the book and async refer to as a "saturated model" have the following form:

$$Y_i = b_1 L_i \overset{Non-Hispanic}{Good\ Grammar} + b_2 L_i \overset{Hispanic}{Good\ Grammar} + b_3 L_i \overset{Non-Hispanic}{Bad\ Grammar} + b_4 L_i \overset{Hispanic}{Bad\ Grammar} + u_i$$

How would you write a regression that produces this output?

6.10.2 A more common estimating strategy

More commonly, but equivalently, we estimate this same form with a treatment-by-treatment interaction model. This has the functional form:

$$Y_i = \beta_0 + \beta_1 J_i + \beta_2 G_i + \beta_3 J_i \times G_i + u_i$$

Where, in this model J_i is an indicator for the sender signing "José", and G_i is an indicator for the sender using "good grammar". Finally, the $J_i \times G_i$ is meant to imply that these two indicator are interacted with on another.

How would you write a regression that matches this form?

The book claims that this facilitates testing of nuanced hypotheses, specifically like, "Is the effect of being named Colin or José different if there is poor grammar compared to if there is good grammar? How would you test this?

6.11 Pre-Test, Post-Test

The Pre-Test, Post-Test model is the *most core* :metal: experiment design that is out there.

1. How would you represent this in the "Grammar of Experiments" that we talked about last week?
2. What threats to validity are present in this design? To what extent are they present?
3. What type of model would you estimate in order to match the efficiency of the *design* with the efficiency of the model?

6.11.1 The Difference in Differences Model

The Difference in Differences model, sometimes referred to as the D-in-D model, or the D&D model, is the regression equivalent of the paired t-test.

1. The D-in-D model is *extensible* meaning that you can include extra information in the model, as "good controls" to try to improve the model performance.
2. The D-in-D model is *efficient* because it uses all the possible information from the experiment design in the model. You could also, always analyze your experiment with a simple between-subjects comparison, but it will be noisier if there is variability between subjects.

```
library(data.table)
library(ggplot2)
library(here)
```

```
## here() starts at /Users/dhughes/teaching/241/info-241-live-session
```

7 Heterogeneous Treatment Effects

When we consider heterogeneous treatment effects, we acknowledge that it is very unlikely that every single person reacts to treatment in exactly the same way. But, are there certain *types* of people who always react more strongly to treatment? Are there certain *types* of people who always react less strongly?

How should we go about (a) looking for HTEs; (b) testings for HTEs with nominal p-value coverages; and (c) reporting HTEs to individuals in a way that make sense?

7.1 Learning Objectives

At the end of this week, students will be able to

1. Understand what an HTE is, and what it is not.
2. Conduct, test, and interpret models with interaction terms as specific tests of hypotheses.

7.2 Reading and Discussion: Goodson

1. Why do we call them A/B tests, rather than experiments? Are you uncomfortable with the idea that companies are experimenting on you? Are you uncomfortable experimenting in your subjects? If there is a gap between your feeling about being experimented **on** compared to how you feel when you are **doing** the experimenting, why does this gap exist?
2. What is an A/B test in Goodson's estimation?
3. How should we know when a test should be determined to be complete? How should we determine that one experience is *causing* different behaviors in our reference population than another experience?

- Can we decide this while we are working through the experiment?
- Can we make this choice when we see the outcome that we thought we were going to see?
- Can we be *sure* that we have set the correct stopping rules ahead of time?

4. What are the consequences of peaking, and then stopping early, once we have seen the results?

7.3 Coding and Demo: The Californians

```
library(data.table)
library(stargazer)

make_data <- function(sim_size) {
    d <- data.table(
      id              = 1:sim_size,
      cal_stanford    = rep(c('cal', 'stanford'), each = sim_size/2),
      affluence       = c(
        sample(1:7, size = sim_size/2, replace = T, prob = c(.1, .2, .2, .2, .2, .05, .05)),
        sample(1:7, size = sim_size/2, replace = T, prob = c(.05, .05, .05, .15, .2, .2, .3))),
      founder_motivation = rnorm(sim_size, mean = 100, sd = 7),
      treatment_group = sample(c(0,1), sim_size, replace = TRUE)
      )

    d[ , capital_access := rnorm(sim_size, mean = d$affluence, sd = 2)]
    d[ , tau            := rnorm(sim_size, mean = 5 + 5*I(cal_stanford == 'cal') + affluence)]
    d[ , founding_prob  :=  founder_motivation + tau*treatment_group]

    return(d)
}
```

```
d <- make_data(sim_size = 1000)
```

7.3.1 Overall Treatment Effect

What is the (unobserved) true average treatment effect? Estimate a model that includes only the treatment effect and interpret all the coefficients.

```
model_0 <- d[ , lm(founding_prob ~ treatment_group)]
```

7.3.2 Founder Motivation

Estimate a model that includes the (nearly impossible to measure) variable about `motivation`. What happens to your estimates? Why does this happen?

```
model_1 <- d[ , lm(founding_prob ~ treatment_group + founder_motivation)]
```

Print these two modeles next to one another and describe what is happening and why.

- Are the intercept terms the same? Why or why not?
- Are the standard errors teh same? Why or why not?
- Are the estimate treatment effects the same? Why or why not?

7.3.3 Subset Models

Subset the data into two groups based on `cal_stanford` and estimate a model that only includes the treatment effects.

Print these two models side by side, and tell me what is going on.

```
model_cal      <- d[cal_stanford == 'cal', lm(founding_prob ~ treatment_group)]
model_stanford <- d[cal_stanford == 'stanford', lm(founding_prob ~ treatment_group)]

stargazer(
  model_cal, model_stanford,
  type = 'text'
)
```

```
##
## ===============================================================
##                                   Dependent variable:
##                               ----------------------------
##                                       founding_prob
##                                   (1)            (2)
## ---------------------------------------------------------------
## treatment_group                 13.359***      9.839***
##                                  (0.636)        (0.661)
##
## Constant                        99.804***     100.031***
##                                  (0.438)        (0.466)
##
## ---------------------------------------------------------------
## Observations                       500            500
## R2                                0.470          0.308
## Adjusted R2                       0.468          0.307
## Residual Std. Error (df = 498)    7.104          7.386
## F Statistic (df = 1; 498)       440.848***     221.810***
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Based on this, would you conclude that these are different? Use the `confint()` function on each of these models to inform this discussion – this is a total trap, because there is very weak statistical basis for what you're about to say, but do it anyways.

7.3.4 Interaction Model To Test

The models that you estimated above are extremely intuitive to talk about, and are probably the right models to report to collaborators, especially those who aren't read into this class. **But**, they don't include a formal statisical test.

To conduct this test, we're going to rely on the **same model form** as we used for treatment-by-treatment investigations – the difference in differences model.

Estimate a model that interacts the treatment indicator with the `cal_stanford` indicator, and report what you see from this model.

```r
model_interaction <- d[ , lm(founding_prob ~ treatment_group * cal_stanford)]

summary(model_interaction)
```

```
##
## Call:
## lm(formula = founding_prob ~ treatment_group * cal_stanford)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7099  -4.8896  -0.2101   4.9203  24.4802
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         99.8039     0.4468 223.361  < 2e-16 ***
## treatment_group                     13.3588     0.6490  20.583  < 2e-16 ***
## cal_stanfordstanford                 0.2271     0.6394   0.355 0.722564
## treatment_group:cal_stanfordstanford -3.5197     0.9172  -3.837 0.000132 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.246 on 996 degrees of freedom
## Multiple R-squared:  0.3989, Adjusted R-squared:  0.3971
## F-statistic: 220.3 on 3 and 996 DF,  p-value: < 2.2e-16
```

7.4 ? anova

7.5 Finally, use the results from model 5 to tell me what the treatment

7.6 effect is for males and for californians.

7.7

7.8 AT HOME:

7.9 Work to examine what including the other affluence and literacy

7.10 triggers does to your estimates.

7.11

7.12 Coding and Discussion: Tips at a Restaurant

```r
## Green and Gerber: Question 9.6
## a, b, and c.

d <- fread('http://hdl.handle.net/10079/cd6be01a-a827-4312-a2fa-74329ce7f96d')

## a. (Probably skip this one)
##    Suppose that you ignored the gender of the server and simply analyzed whether
##    the happyface treatment has and effect (and/or) a heterogeneous effects. Use randomization infere
##    to test whether the Variance of \tau = 0 using randomization inference by
##    comparing the variance of potential outcomes in treatment and control.

## b. Write down a regression model that depicts the effect of the gender of
##    the waitstaff, whether they put a happyface on the bill, and the interaction
```

```
##    of these factors.
##

## c. Estimate the regression model that you wrote down in (b) and test the
##    interaction between waitstaff and the happyface treatment.
##    Is the interaction significant.

## d. Waiting tables in the time of covid: Suppose that you're on the waitstaff
##    at this restaurant, and while you're waiting tables you're FULLY garbed
##    up: facemask, face-shield, full operating gown, and so on.
##    What this means is that you have the choice to reveal a gender identity
##    that is either "Male" or "Female" to the patrons.
##
##    - Is there one gender identity that receives higher tips in this restaurant?
##    - Is there a gender that has a higher treatment effect? What is the
##      test that you would run to assess this?
```

7.13 Sleeeeeeeeeeeep. . .

Suppose that you've conducted an experiment to evaluate the effectiveness of meditation prior to sleeping. Some people are free to do what they want, while others are assigned to a 10 minute guided mindfulness exercise before they go to bed. Treatment is randomly assigned at the individual-level, and people are placed into their groups (and maintained in those groups) for 10 days; then, after two weeks, the groups are flipped.

1. What does the design of this experiment look like?

   R R

2. Suppose that you also possess some data about the individuals. Specifically, you collect:
3. Their age;
4. Their coffee drinking habits;
5. Their tea drinking habits;
6. The number of people in their bed on a typical weeknight;
7. The number of cats they own;
8. The number of dogs they own.
9. First, is there an effect of treatment? Combine the data that you have on hand to write the first, best model.
10. Are there are parts of the population that this is especially effective (or ineffective)? How do you know?

```
load(file = here("data", "sleep_study.Rdata"))
```

8 Treatment Noncompliance

This begins a section of the class where we are going to evaluate what happens when problems creep into the actual experiments that we are conducting. We are first going to look at what happens when we instruct people to take treatment, but they choose not to. Or, when we instruct people to take control, but they choose to take treatment instead.

It might seem, at first, like we should just proceed by analyzing according to the treatment condition that they actually received. However, because we haven't experimentally assigned this condition, this creates an unprincipled estimator.

This doesn't mean that all is lost however. We can redefine the causal quantity that we are estimating, and produce a reliable estimate of this new concept.We're going to present two such concepts this week. The first concept is the idea of the intent to treat effect (the ITT). The second concept is the idea of the treatment effect among compliers, which we will call the CACE.

8.1 Learning Objectives

1. **Recognize** when experimental units have not complied with the treatment assignments they were given, and **appreciate** that this causes problems for our two-group estimator.
2. **Recover** causal estimators, but for sub-populations of the overall population.
3. **Utilize** a new class of model, the two-stage least squares model, or 2SLS, which is the appropriate model choice when we are dealing with either one- or two-sided non-compliance.

8.2 Starting conversation

Life on campus is exciting! Whether students are involved in affinity groups, advocacy groups, protest groups, or just party groups, student life on campus is exciting. We're not to be left out! We're not to be denied the chance to make our voices heard.

What is there to complain about? How about that "*god awful*" sound of the bell-tower ringing every hour on the hour.

Suppose that we are to discuss this *very. important. issue.* before a panel of the deans and University administration. And, further suppose that in light of global events of the past several years, they're actually amenable to what we're proposing – cutting off those bells, and providing the Berkeley Carillon player a generous retirement. University Carillonist Video.

However, there's a catch. They are worried that taking such an action would be detrimental to the student experience on campus, and they would like to measure the causal effect of playing vs. not playing the Carillon while students are changing classes.

In breakout rooms, please design an experiment that would be able to measure the difference in student experience. You will have to propose a measurement, a timescale for that measurement, and a feasible randomization that *could actually* occur given the real-world constraints that what is at question are loud sounds emanating from a 300 foot tower in the middle of a large, busy campus.

If there are any limitations to what you design, please voice those concerns and talk about why they arise, relative to an *ideal* experiment (that you are probably unable to conduct).

8.3 Non-compliance Discussion

8.3.1 Setting Terms of Understanding

- What is does the concept of the *intent to treat* effect mean? When is this ITT measurable? When is the ITT an interesting quantity to estimate? When is it uninteresting?

- What is the compliance with treatment assignment? How does someone measure ITT_D? Why do Green and Gerber choose to pick such arcane notation? Why is it necessary to know the compliance rate?

- What is the compliers average causal effect (CACE)? Under what conditions is this CACE guaranteed to be *exactly* the same as the ITT? Under what conditions is this CACE larger than the ITT? Under what conditions is it smaller than the ITT?
- Is it possible to estimate the CACE without knowing *specifically* who complied? Is it possible to estimate the CACE without knowing anything about compliance rates?

8.3.2 Where Does Noncompliance Occur

Is all of this concern about non-compliance actually a concern? Or, is this just another example of academics getting ahead of themselves and worrying about things that are not actually a concern?

In three distinct breakout groups, please talk about one of the following scenarios.

1. **You are a MIDS instructor writing online content to cause students to be their best possible data scientist.** In this role, you write curriculum, record lectures and assign readings, and create homework assignments for students to work on. How might compliance issues affect curriculum choices?
2. **You are a non-profit interested in reducing litter at your local surf spot.** In this role, you take steps to raise awareness through signs. How does compliance affect what you are likely to learn in any trial or evaluation of your work? How would you know if someone complies?
3. **You are a publisher seeking to sell more copies of the newest, and best causal estimators textbook.** You propose to use pre-roll advertisements on video streaming services to get the word out about the new book. What are all the ways that you can imagine measuring compliance? Which would you propose to use, and why? How might compliance issues affect what you're able to estimate?

When we come back, each student group will have 5 minutes to talk the other two groups through their scenario, including major risks, opportunities, and learning that they took away from the scenario.

8.4 Estimating with Non-compliance

8.4.1 Estimating with non-compliance

```
## install.packages("AER")      # this has a nice wrapper for iv regression
                                # but we can do it by hand with VERY little work

nrows = 1000

d <- data.table(
  id  = 1:nrows,
  y0  = rnorm(nrows, mean=10),
  tau = rnorm(nrows, mean=5)
)

## create a treatment effect
d[ , y1 := y0 + tau]

## create an assignment vector and a treatment vector with everybody initially
## set to be untreated.
d[ , assigned := sample(rep(c(0,1), each=nrows/2))] # z in the book
d[ , treated := 0]

## then, among the people who are assigned to receive treatment, actually
## treat some of them at random.
d[
  assigned == 1,
  treated  := sample(
    x=1:0, size=.N, replace=TRUE, prob=c(.7, .3)
    )]

## finally, observe:
##   - the potential outcomes to treatment for the treatment group; and,
##   - the potential outcomes to control for the control group.
```

```
d[treated == 1, Y := y1]
d[treated == 0, Y := y0]
```

With this data created, we can confirm that the different between potential outcomes to treatment and control still produces the treatment effect. If it doesn't we've got bigger problems than compliance!

```
d[ , mean(y1 - y0)]
```

```
## [1] 4.990989
```

But, the data that we have created to this point has all the data in the science table. In real life, we won't get access to all this data; instead, we get access to observing some potential outcomes for one group and some other potential outcomes for another group. Let's create this restricted set of data to ensure that our estimator can recover the population parameter that we're looking for, even though it only has access to a sample of data from that population.

```
d_observed <- d[ , .(assigned, treated, Y)]
```

Does our estimator for the ATE produce an unbiased estimator of the population parameter, given this sample of data? We know that the population parameter has a treatment effect of 5.

```
## how would you code a simple two-group ATE estimator?
```

8.4.2 Build Estimators

Suppose that you cannot /actually/ observe whether someone was treated or not. This will require that you suspend reality for a moment, to suppose that we do not have access to the `treateed` variable.

In this case, what concept **are** you able to estimate? What concept are you **not** able to estimate?

8.4.2.1 Estimate the Intent to Treat Effect

Use a linear model to estimate the intent to treat effect. What variables do you need to produce this estimate, and what subset of the data (up to, and including the full set of data) do you use to produce this estimate?

- What should be the magnitude of your ITT estimator, relative to the actual population average treatment effect that we encoded (i.e. 5)? Why do you anticipate that it will be at this level?

8.4.3 Estimate the Compliance Rate

Using the full set of data, estimate the compliance rate in two different ways:

1. Estimate using a `.N` counter, or `mean` or some other such simple transformation on the data.table.
2. Estimate using a linear model, interpreting the coefficient of that model appropriately.

What are the trade offs to these two different methods?

8.4.4 Compute the CACE

Because you have the ITT and the compliance rate, estimate the CACE. Once again, compute this compliance rate in several ways.

1. Using a `.N` counter, or `mean` or some other such simple transformation on the **entire** data.table. Once again, you might notice that this does not have a sampling based uncertainty estimate associated with it. You need not code this now, but talk about how you would produce something akin to a standard error for the mean given this data.
2. Using a `.N` counter, or `mean` or some other such simple transformation on a reasonable **subset** of the data.table.

- Will the estimate that you produce on this subset of data be larger, smaller, or about the same as the estimate that you produced on the full data?

- Will the estimate of the standard error for the mean be larger, smaller, or about the same as the estimate that you produced on the full data?

3. Finally, estimate the CACE using two linear models. How, if at all, would you produce an estimate of the sampling based uncertainty of these estimates?

```
cace_one   <- 'fill this in'
cace_two   <- 'fill this in'
cace_three <- 'fill this in'
```

How do you feel about the sampling based uncertainty that you can produce with these estimators?

8.5 Two Stage Least Squares

In order to estimate with a reliable standard error, we can turn to two stage least squares.

Two-stage least squares estimators have the benefits of

1. Doing *exactly* the same thing that the $CACE = ITT/ITT_d$; but,
2. Doing it in a way that has known standard errors that are quickly and easily computable.

8.5.1 First Stage

In the first stage we:

- Estimate the proportion of people who are receive treatment as a function of being assigned to treatment.
- In the case of one-sided non compliance this is *exactly* the same thing as estimating the ITT_{d}, right?

```
first <- 'fill this in'
```

```
## calculate the fitted values from this regression
##   - that is, just multiply the coefficients that you estimate from the
##     first stage times the data values. In the event that the exclusion
##     restriction holds, then these predictions are just orthagonal to every
##     thing that is not modeled in your data!
```

8.5.2 Second stage

In the second stage we:

- Estimate the relationship between the predicted values and the outcome.
- This will just tell you how the outcome changes in response to the amount of change that your treatment assignment is able to produce.

```
second <- 'fill this in'
# coeftest(second, vcovHC(second, type = "const")) ## these ses might be wrong.
```

```
## I'll note that the standard errors from this "hand-rolled" 2SLS will not
## be correct (due to some accounting issues in the variance between the predictions
## in the first stage and the second stage.
##
## we can fix this by hand -- though I wouldn't -- or we can use a library that will
## do the accounting for us, from the library AER

# library(AER)
# iv.model <- ivreg(Y ~ treated | assigned, data = d2)
#
# coeftest(iv.model, vcov = vcovHC(iv.model, type = "const"))
```

9 Spillover and Interference

At the outset of the course, we enumerate three hard-core requirement of an experiment design. In addition to intervening in the world, to produce an unbiased estimator of a treatment effect, we require that an experiment:

1. Assign that intervention to experimental units at random to eliminate the possibility of confounding due to selection bias;
2. That one, and only one difference exists between two comparison groups, thereby allowing us to exclude all other possible causes *but for* the feature that we have experimentally assigned; and,
3. That the treatment experienced by one experimental unit does not "interfere" with the potential outcomes of another unit.

In previous weeks, we've engaged with how to evaluate whether a treatment has been successfully randomized. In this week's materials, we are going to examine what, if anything, we may do in response to interference between units.

There are two possibilities. First, we might design our experiment to minimize the effects of interference between units by re-designing or measuring differently. In doing so, we endeavor to maintain the measurement of an individual-level treatment effect, through a multi-group experiment. Second, we might acknowledge the existence of interference and expand our thinking about what *is* a treatment effect.

9.1 Learning Objectives

At the conclusion of this week, student will be able to:

1. **Articulate** in clear terms what circumstances *are*, and what circumstances *are not* interference events.
2. **Appreciate**, and **evaluate** the extent that interference between units changes both the concept of a treatment effect, and also how a multi-group measurement's estimates change in response to interference between units.
3. **Identify** common situations where interference is likely to occur, and anticipate some methods of mitigating, ameliorating, or designing in response to this interference.

9.2 Defining Terms

- What does it mean for one unit to interfere with another unit?
  - If two units communicate with one another, is this interference?
  - If three units are all genetically related to one another, is this interference?
  - If ten units all work in the same building, is this interference?
  - If two partners share a tablet for browsing the internet, is this interference?
- Now, be *very* precise with your language: Using the term "potential-outcomes" how do Green and Gerber define interference?

9.3 Defining Notation

9.3.1 Identify concepts

Until this week, we have used two concepts to describe treatment assignment and application:

- Z is the assignment to treatment; and,
- D is the dose received of treatment.

What concept is identified in the following notation:

- $E[Y_i(1)|D_i = 1]$? Is this measurable?
- $E[Y_i(1)]$
- Interpret the expression $Y_i(\mathbf{d}) = Y_i(d)$ and explain how it conveys the non-interference assumption.

9.3.2 Classroom Assignments

(From Green and Gerber, p. 283): Sometimes researcher are reluctant to randomly assign individual students in elementary classrooms because they are concerned that treatments administered to some students are likely to spill over to untreated students in the same classroom.

In an attempt to get around possible violations of the non-interference assumption, they assign classrooms as clusters to treatment and control, and administer the treatment to all students in a classroom.

1. State the interference event that commonly leads researcher to assign an entire classroom to a condition.
2. State the interference assumption that is implicitly made when classrooms are cluster random assigned. Where, if anywhere does the researcher assume that spillover exists? Where, if anywhere, does the researcher assume that spillover **not** exist?
3. An *estimand* is the concept that an estimator is attempting to estimate. For example, the ATE estimator produces an unbiased, consistent estimate of the individual-level causal effect. What causal estimand does the clustered design identify? Does this estimand include or exclude spillovers within classrooms? What about spillovers between classrooms? What about spillovers between schools?

9.3.3 Working with a simple example

Suppose that we are conducting an experiment where we examine the effects of releasing solution sets early to some students in the 241 classroom.

- What form of interference is possible?
- Suppose that *Abby*, *Bobby*, *Cathy* and *David* are all on a project team together. Furthermore, suppose that all members of the team work well together, have an ambitious class project that they are working on, and talk regularly.
  - If every one of the students were to be assigned to the control group, name values that are plausible for their completion time on problem set three.
  - Suppose that Abby, Bobby and Cathy are assigned to the control group, but that David is assigned to the treatment group. What do you think will happen in their daily project meeting?
    - Suppose that, no matter the empirical reality, you assume that there is no interference within this group. What would you call the value that you observe for Abby, given this assumption? Consistent with what you have said will happen in their daily project meeting, what values are you actually seeing for Abby?
    - What are the consequences for your estimated treatment effect?
  - Suppose that Abby and Bobby are assigned to the control group and that Cathy and David are assigned to the treatment group. What do you think will happen in their daily project meeting?
  - What would you call the value that you observe for Abby, given this assumption? Is it different when both Cathy and David are assigned to treatment compared to when only David is assigned to treatment?
- Given what you have stated about this small world, how many treatment assignment conditions do you have to be aware of?

9.3.4 Working with a more complex example

Suppose that we are conducting an experiment where we examine the effects of releasing solution sets early to some students in **a law school** classroom. Law school is notoriously competitive, and outside one's immediate group of friends, there is little collaboration.

- Suppose that A, B, C and D are again friends.
- Suppose that W, X, Y and Z are also friends.
- But suppose that the two groups are not friends between groups.
- If A receives an exam solution, but W, X, Y and Z do not, what would you call the values observed for W, X, Y and Z?

- If A, B, and C receive an exam solution, but D, W, X, Y and Z do not, what would you call the values observed for D, W, X, Y and Z?
- Given this example, are the potential outcomes for un-curved exam score different for A if W receives or does not receive a solution?
- Given this example, are the potential outcomes for curved exam score different for A if W receives or does not receive a solution?

9.4 Within subjects experiments

Earlier in the course, we talked about two-group pre-test/post-test experiments. These experiments are exceptionally strong against a large series of threats to identification. And, they form the basis of the expanded topic of a *within-subject* experiment.

1. What *is* a within subject experiment?
2. When might you propose that a within subject experiment would be advisable? Why? What is the benefit of a within subject experiment?
3. When are within subject experiments difficult to conduct?

Green and Gerber, and the async identify two requirements of within-subjects experiments:

- **No anticipation**
- **No persistence**

What do these two assumptions mean in terms of what you are measuring at the individual-time level?

- Suppose that you were worried that your experimental units might either anticipate being put into treatment or that the treatment they take might have long-run effects. How might you design a test to see if either of these concerns are present in your design?

9.4.1 Survey Experiments

(From Green and Gerber, p.285): Concerns about interference between units sometimes arise in survey experiments. For example surveys sometimes administer a series of *vignettes* involving people with different attributes. A respondent might be told about a low-income person who is randomly described as white or black; after hearing the description, the respondent is asked to rate whether this person deserves public assistance. The respondent is the presented with a vignette about a second person, again randomly described as white or Black, and asked about whether *this* person deserves public assistance.

This design creates four experiment groups:

1. Two vignettes describing Black beneficiaries;
2. Two vignettes describing white beneficiaries;
3. A vignette describing a Black beneficiary first, followed by a white beneficiary; and,
4. A vignette describing a white beneficiary first, followed by a Black beneficiary.

Suppose that each respondent provides a rating after each vignette.

Questions to answer:

1. Propose a model of potential outcomes that reflects the ways that subjects might respond to the treatment and the sequences in which they are presented. How might you represent this using the `R O X Y` grammar?
2. Using your model of potential outcomes, define all of the ATE or ATEs that a researcher might seek to estimate.
3. Suggest an experiment design that could estimate this/these causal estimand(s) using observed data.
4. Suppose a researcher analyzing this experiment estimates the average *race effect* by comparing the average evaluation of the white recipient to the average evaluation of the black recipient. Is this a sound approach? Why or why not?

9.5 Discussing the reading: Blake and Coey (2014)

Here is a link to the reading.

- What is the treatment, and how does treatment assignment work?
- What is the outcome, and how is it measured?
- How does this experimental setup generate spillovers within an auction?
- What is the naive research strategy that produces a biased estimate in the presence of the spillover?
- Tell a story to explain why the within-auction spillovers might give you upward bias in the measured treatment effect.
- (Optional; harder) How does the experiment generate spillovers between auctions?
- Tell a story to explain why you might get downward bias from between-auction spillovers.
- What is the proposed empirical analysis strategy to reduce the bias?
- What would be a better experimental design to conduct in the first place?
- Do you see an example of a stepped-wedge design in this article? Explain.

9.6 Discussing the reading: Miguel and Kremer (2004)

Here is a link to the reading.

- What question are Miguel and Kremer trying to answer? Why is this important?

- What is the spillover problem in this setting?
- How did doctors get the wrong answer in randomized trials before Miguel and Kremer addressed the spillover problem? (The article refers to this as a double penalty.)
- When not taken into account correctly, did the spillovers to cause underestimation or overestimation of the treatment effect? Explain why.
- Which feature do the authors choose to make their experiment less vulnerable to this spillover problem?
- How do the authors still have a (smaller) spillover problem despite this design decision?
- What was the compliance rate for those whom the researchers intended to treat in 1998?
- Name two kinds of noncompliance described in the article, and say which one was largest. Due to noncompliance, we can only measure the CACE rather than the ATE. Why is the CACE just fine for the policy question asked in the article?
- Do you see an example of a stepped-wedge design in this article? Explain.

10 Causality from Observational Data

```
theme_set(theme_minimal())
berkeley_blue   <- '#003262'
california_gold <- '#FDB515'
```

What happens if we cannot run an experiment? Perhaps we don't have the budget or time, perhaps the context is too fraught to conduct an experiment. Should we walk away and learn nothing?

10.1 Learning Objectives

At the end of this weeks *extensive* content, students should be able to

1. **Describe** a series of techniques that have been proposed to estimate causal effects even when a randomized experiment has not been conducted;
2. **Evaluate** whether a particular technique matches with the data generating context;

Figure 8: punkin belly

3. **Analyze** whether an observational data technique is likely to identify a treatment effect; and,
4. **Communicate** the risks and limitations that are brought about when using observational data to make causal claims.

10.2 The Experimental Ideal

If you've been through this once, you've been through it one-hundred times this semester, but it might be worth re-stating what we get out of conducting a randomized experiment.

Why do we conduct experiments? What guarantees exist as a result of a well-run experiment?

10.3 A Continuum of Plausibility

As we are talking today, consider the fully-randomized, full-compliance, full-reporting, high-powered field experiment to be the high-water mark of credibility. Under such a scenario, we can think of any analysis that we undertake as producing a highly-credible, highly-reliable estimate of a treatment effect.

Through our discussion this week, we hope to name where we think other techniques and data generating processes fall relative to this high-water mark. Some, as we will see, might actually produce estimates that are *very nearly* as credible as the experimental ideal. Others are ghastly in their performance.

However, as data scientists who *have to get work done* we need to be able to produce the best possible statement about a treatment effect, and if we have any misgivings about those statements, be able to provide a clear statement about the risk that is attendant to using them.

10.4 Natural Experiments

Natural experiments are experiments that have been conducted by someone *other* than the researcher. If you remember back to Problem Set 1, consider the case of the early childhood education that is provided by the state. When the state *chose* who to provide education to based on need, this was clearly not an experiment because it isn't possible to fully understand the selection criteria used by the state, and so it is not possible to make a strong statement that any estimate produced from a two-group estimator wouldn't be possibly subject to confounding.

But, what about the case where the state *randomly assigned* some kids to get the treatment? Is there any reason that we should discount this simply because it wasn't us to do the assigning? What hubris!

10.4.1 Questions to consider

- What are the hallmarks of a natural experiment?
- How would you propose to structure your search for natural experiments?
- How will you know when you've actually *found* something that is a natural experiment?

10.4.2 Breakout activity

- What are the things in the past year of your lives that have seemed to *arrive at random*? How would you know if they actually **are** at random?
- After the members of the team have spent a few moments thinking about things that might be random, ask yourselves, "What might we be able to learn downstream from this experiment?" What is the most plausible thing that you might learn? What is the longest, most extreme possibility that you might learn?

10.4.3 How does one analyze a natural experiment?

If a natural experiment is just a randomization conducted by someone else – is there anything different that we need to do in order to analyze it? Why or why not?

We talk, with some specificity this week, about estimating using two-stage least squares regression. What is this technique, what does it promise to us, and how does it work?

Consider simulated data that is created in the following way:

- `ability`, `family_income`, and `lottery` winning to get into a "magnet" school are all random
- However, suppose that `schooling` which is the indicator that someone actually got schooling at a magnet school is correlated with ability, with family income, and with lottery.

What would be the consequence of estimating an eventual outcome, using a naive regression?

```
make_iv_data <- function(instrument_strength=1.0) {
  set.seed(2)

  d <- data.table(id = 1:1000)

  ## These are all independent of one another.
  d[ , ':='(
    ability       = rnorm(.N, mean=10, sd=1),
    family_income = rnorm(.N, mean=20, sd=2),
    lottery       = rnorm(.N, mean=10, sd=1) > 10 ## Win if larger than 10
    )]

  ## Create a schooling indicator. This has the following characteristics:
  ##   - It is related to winning the lottery (which was random)
  ##   - It is related to ability
  ##   - It has some "white noise" just so that the model will estimate.
  d[ , schooling := instrument_strength * lottery + ability + rnorm(.N, mean=0, sd=1)]

  ## Create the outcome, which might be earnings. This has the following characteristics:
  ##   - It is affected by schooling (Yay!)
  ##   - It is affected by ability (Yay!)
  ##   - It is affected by family income :|
```

```
  d[ , earnings := 2 * schooling + ability + family_income +  rnorm(.N, mean=0, sd=1)]
  return(d)
}

d <- make_iv_data(instrument_strength = 1.0)

model_wrong <- d[ , lm(earnings ~ schooling)]
coeftest(model_wrong, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 24.112215   0.570027  42.300 < 2.2e-16 ***
## schooling    2.566756   0.053039  48.394 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How close, or far from the truth is this estimate? How sure are you that this is different from zero?
- What relationship would you have to change in this data generating process in order to flip the bias of the estimate from estimating a value that is higher than the truth, to estimate a value that is lower than the truth?

10.5 Can we fix this estimate?

The promise of two stage-least squares is that it produces unbiased estimates so long as we're able to find something that is random.

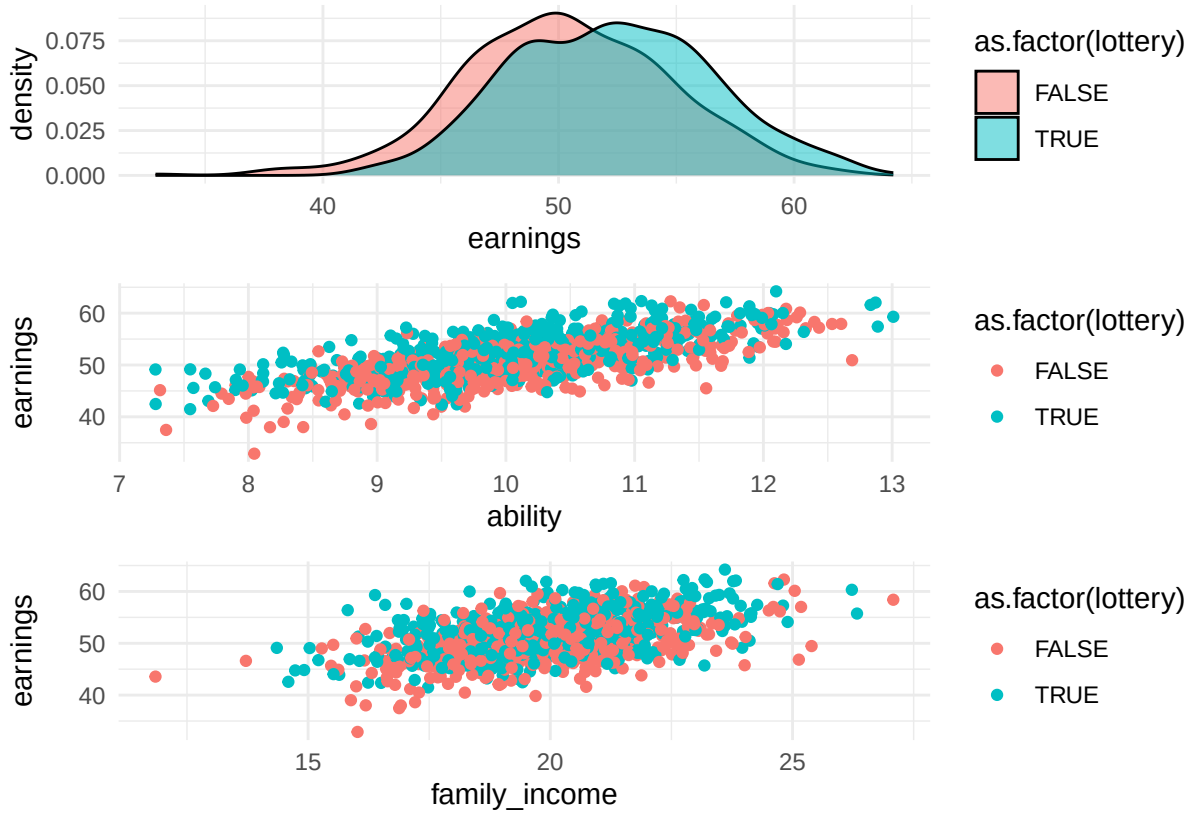
```
first_stage <- d[ , lm(schooling ~ lottery)]
d[ , schooling_hat:= predict(first_stage)]

d[ , schooling_10 := schooling > 10]

# d[ , mean(ability), by = .(schooling_10)]
# d[ , mean(ability), by = .(schooling_hat)]
# d[ , mean(ability), by = .(lottery)]

second_stage <- d[ , lm(earnings ~ schooling_hat)]
coeftest(second_stage, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   30.46240    2.90018 10.5036 < 2.2e-16 ***
## schooling_hat  1.96857    0.27245  7.2254 9.913e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

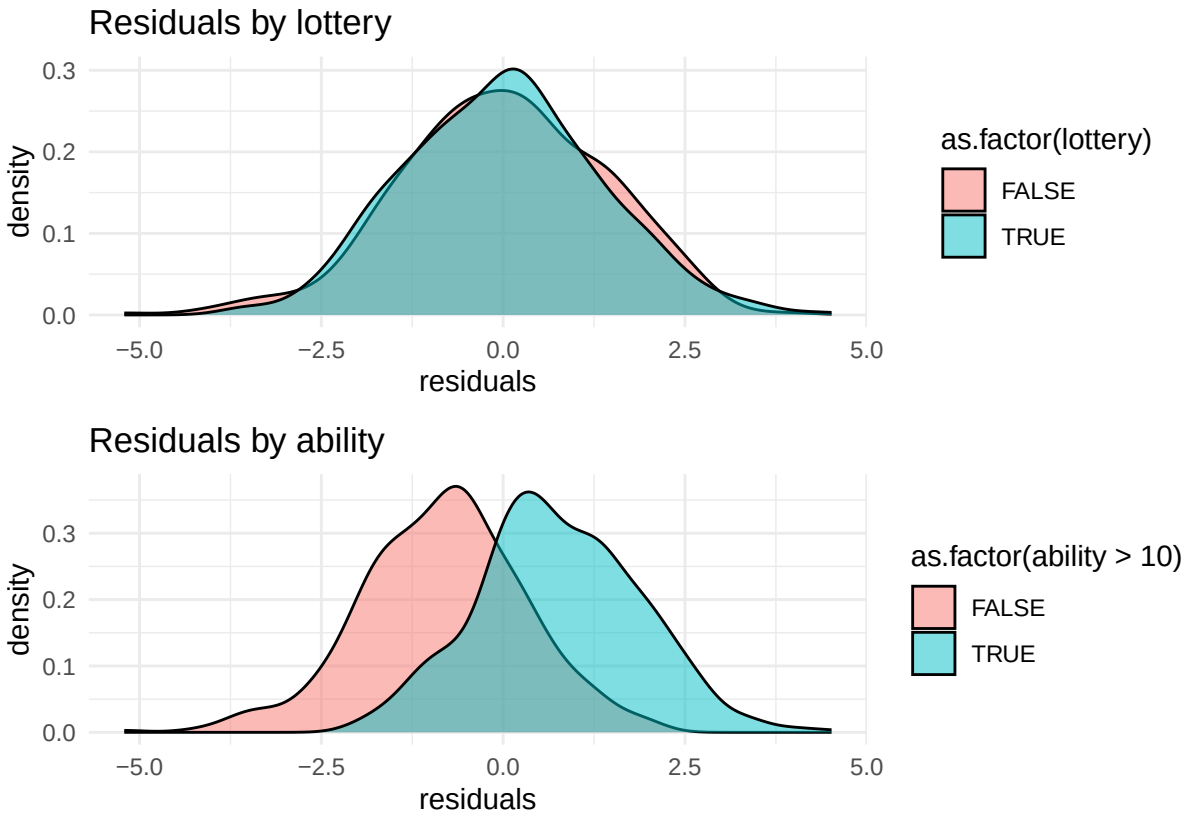
- How or why does this work?
- How does simply making predictions from the first stage regression generate eventual estimates that are unbaised?
- Consider looking at the residuals from the first stage regression

```
d[ , residuals := resid(first_stage)]
```

```r
earnings_by_lottery <-
  ggplot(d) +
  aes(x = earnings, fill = as.factor(lottery)) +
  geom_density(alpha = 0.5)
earnings_by_ability <-
  ggplot(d) +
  aes(y = earnings, x = ability, color = as.factor(lottery)) +
  geom_point()
earnings_by_family_income <-
  ggplot(d) +
  aes(y = earnings, x = family_income, color = as.factor(lottery)) +
  geom_point()

earnings_by_lottery / earnings_by_ability / earnings_by_family_income
```

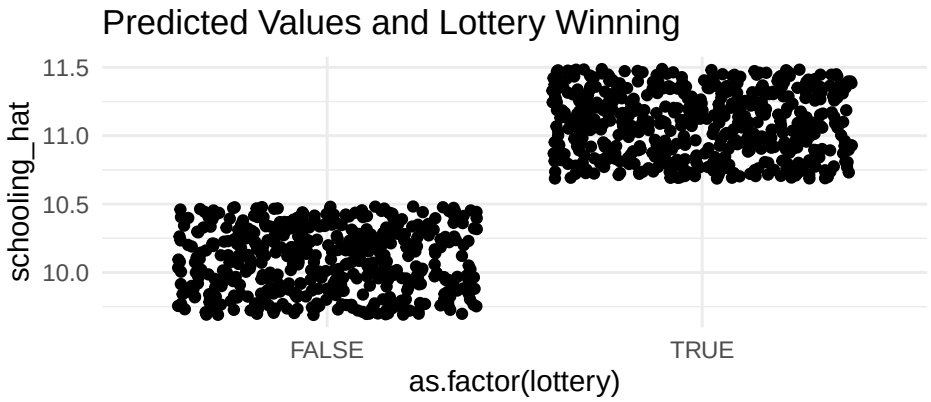


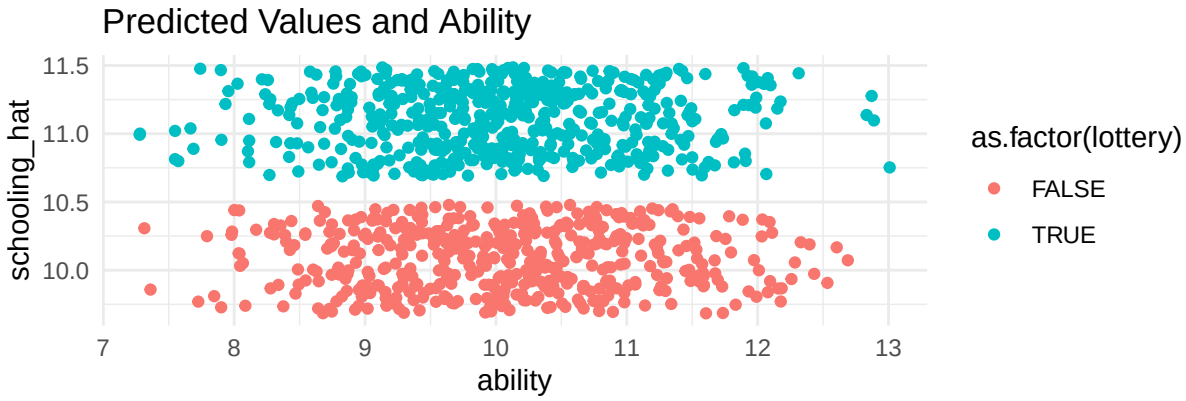

```r
residuals_lottery <- ggplot(d) +
  aes(x = residuals, fill = as.factor(lottery)) +
  geom_density(alpha = 0.5) +
  labs(title = 'Residuals by lottery')

residuals_ability <- ggplot(d) +
  aes(x = residuals, fill = as.factor(ability > 10)) +
  geom_density(alpha = 0.5) +
  labs(title = 'Residuals by ability')

residuals_lottery / residuals_ability
```



Another way to think about this is in terms of how the predicted values are associated different features. Specifically, consider: Are the predicted values associated with having won the lottery? Are the predicted values associated with ability?

```
predicted_lottery <- ggplot(d) +
  aes(x = as.factor(lottery), y = schooling_hat) +
  geom_jitter() +
  labs(title = 'Predicted Values and Lottery Winning')

predicted_ability <- ggplot(d) +
  aes(x = ability, y = schooling_hat, color = as.factor(lottery)) +
  geom_jitter() +
  labs(title = 'Predicted Values and Ability')

predicted_lottery / predicted_ability
```





Notice that this entire system works because there is **some** randomness in the instrument. If there were no randomness in the instrument, or if the instrument had only a small effect on the causal feature that we care about, we will have bias creep back into the estimate.

> Said another way: This isn't a magic trick that works every time. You **actually** need randomness, and a strong relationship in order for the the two-stage least squares estimate to work to identify a causal effect.

```
d_weak_instrument <- make_iv_data(instrument_strength = .1)

biased_estimate <- d_weak_instrument[ , lm(earnings ~ schooling)]

first_stage <- d_weak_instrument[ , lm(schooling ~ lottery)]
d_weak_instrument[ , schooling_hat := predict(first_stage)]

second_stage <- d_weak_instrument[ , lm(earnings ~ schooling_hat)]
```

```
stargazer::stargazer(
  biased_estimate, second_stage,
  type = "text"
)
```

```
##
## ============================================================
##                                   Dependent variable:
##                               ---------------------------
##                                         earnings
##                                    (1)            (2)
## ------------------------------------------------------------
```

```
## schooling                           2.645***
##                                     (0.053)
##
## schooling_hat                                          1.695
##                                                       (2.635)
##
## Constant                           23.587***          33.220
##                                     (0.543)          (26.712)
##
## ------------------------------------------------------------
## Observations                         1,000             1,000
## R2                                   0.713            0.0004
## Adjusted R2                          0.713            -0.001
## Residual Std. Error (df = 998)       2.304             4.302
## F Statistic (df = 1; 998)         2,483.207***         0.414
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

10.6 Regression Discontinuity

Regression discontinuity is a *really* clever idea, that when the data presents itself and the analysis is done correctly, provides a *very* compelling argument for having captured a causal effect.

The key insight in the case of regression discontinuity is that we might *not* need something that is actually random in order to produce a credible treatment effect. All we need is treatment assignment mechanism that is not correlated with potential outcomes. And, the argument for regression discontinuity is that if you make a comparison set similar enough along a scoring variable, then it would be very hard for people on one-side or the other-side of an arbitrary point in the scoring variable to be different.

10.6.1 Why do RDD designs "work"?

- What part of the RDD is producing an unbiased causal estimate?
- Why is this part of he design/data generating process able to produce this unbiased causal estimate?
- What is a "forcing" variable?
- How do I identify *where* the cut-point in the forcing variable is located?

10.6.2 Just how common are opportunities for RDD

Here's a controversial point of view: Everything that we do as data scientists is to make low-dimensional representations of higher dimensional space. Let's have a jam-session where the class and instructors take turns naming places where a RDD could be run. We'll start with:

- Revolving line of credit – credit scoring models bring in disparate streams of information, produce a low-dimensional 0-800 (or something like that...) rating and provide revolving lines of credit to different parts of the distribution.
- Now you...

10.6.3 Working with Regression Discontinuity Designs

Let's look at see what is happening when we're working with RDD designs. To start, let's build some data. Read through each of the lines below, and note what is happening with the data being created. (*Notice that we are chaining together the data.table after we create* `y0` *and again after we create* `y1`.)

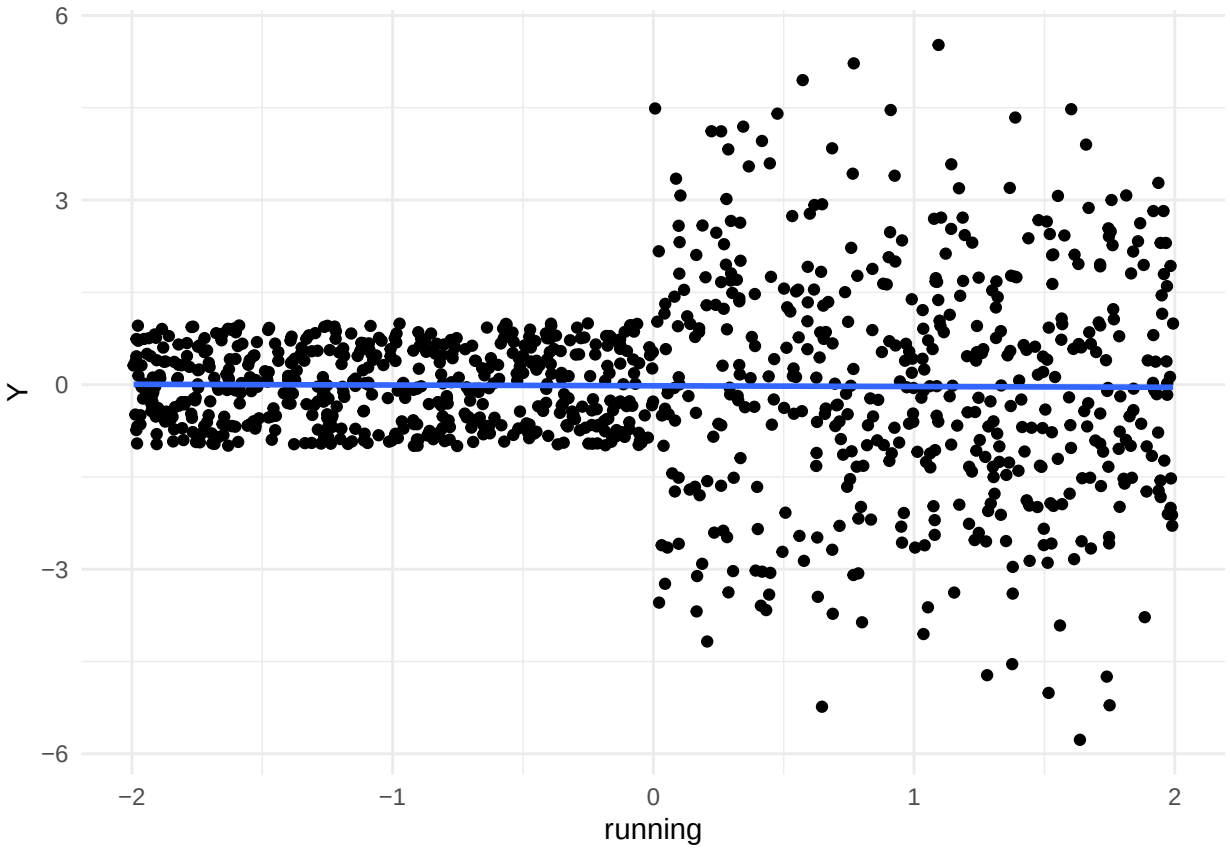
```
N <- 1000

d <- data.table(id=1:N)
d[ , ':='(
  tau     = rnorm(.N, mean=0, sd=2),
  running = runif(.N, min=-2, max=2),
  y0      = runif(.N, min=-1, max=1)) ][ ,
  y1     := y0 + tau ][ ,
  Y      := ifelse(running > 0, y1, y0)]
```

With the data created, let's quickly look at what we are working with. Does the following plot seem to capture the idea of a treatment effect?

- If so, why?
- If not, why not and how would you propose to change the plot?

```
ggplot(d) +
  aes(x=running, y=Y) +
  geom_point() +
  stat_smooth(method = 'lm', se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We're going to start doing some wacky stuff, subsetting data, and fitting models to subsets of that data. This is behavior that `ggplot2` takes an opinionated stance against; as such, we're going to work in base plots. Don't worry about the plotting syntax as much as what you're seeing in them.

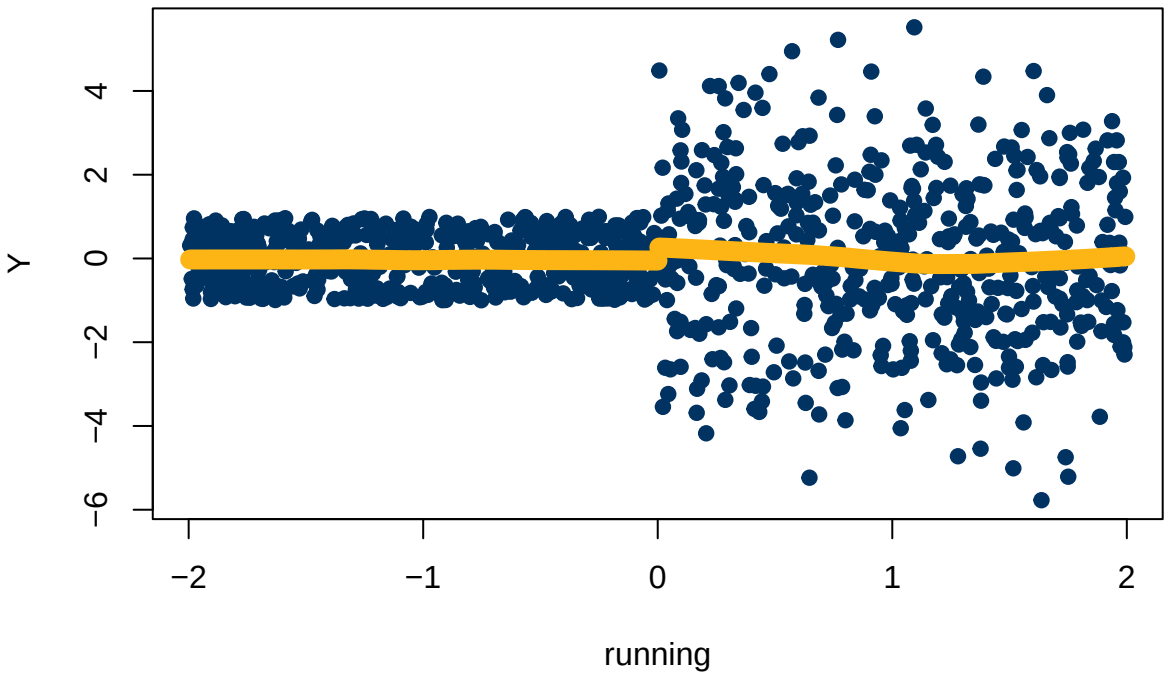
```
d[ , plot(x=running, y=Y, pch = 19, col=berkeley_blue)]
```

```
## NULL
```

```
  d[running < 0, lines(lowess(running,Y), lwd=10, col=california_gold)]
```

```
## NULL
```

```
  d[running > 0, lines(lowess(running,Y), lwd=10, col=california_gold)]
```

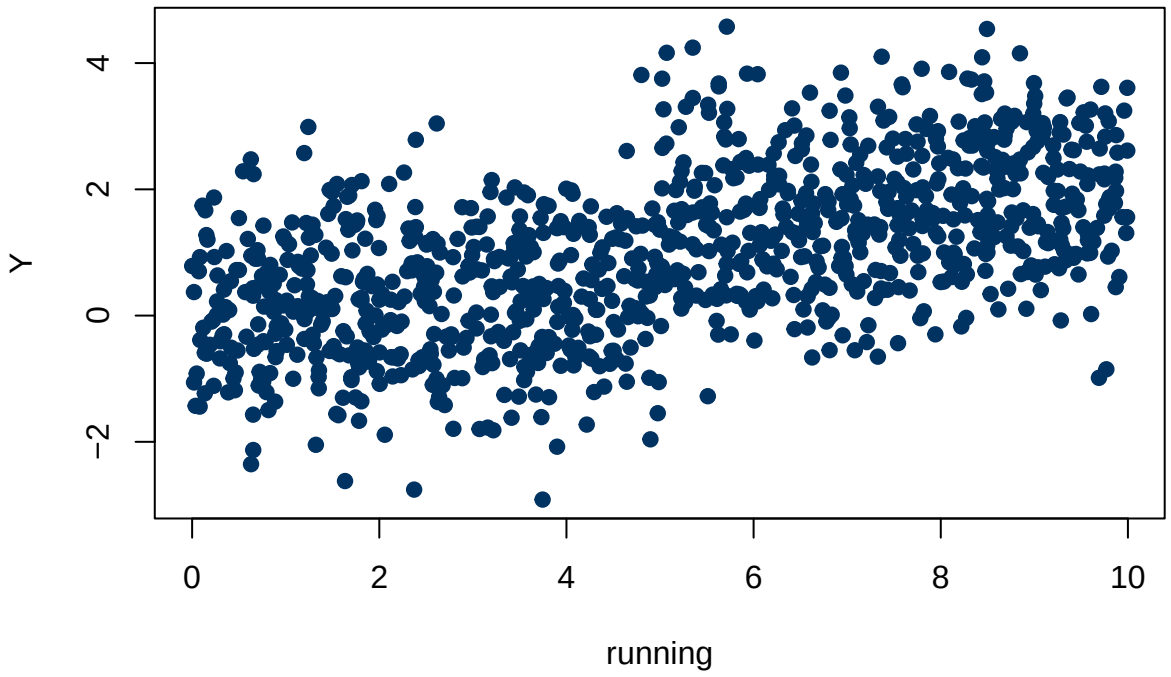



```
## NULL
```

This has been *very* fortunate data. There is no trend across the running variable, and things seem mostly linear on both sides. Naturally, the real world is not so tidy.

10.6.4 More realistic data

```
d <- data.table(id=1:N)

d[ , ':='(
  running = runif(n=.N, min=0, max=10),
  cov1    = rnorm(n=.N)) ][ ,
  Y := running * 0.1 - 0.2 * cov1 + 1 * I(running > 5) + rnorm(n=.N)]

d[ , plot(x=running, y=Y, col=berkeley_blue, pch=19)]
```



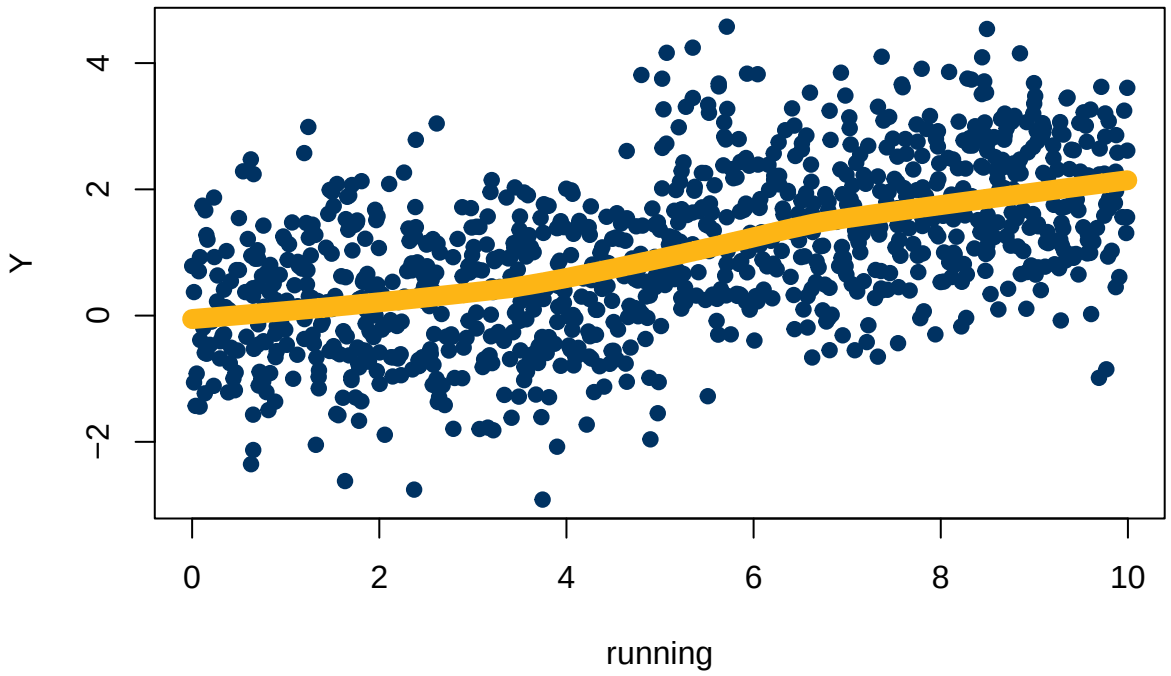

```
## NULL
```

- Is it clear if there is, or is not an effect in this data simply by looking at it?
- What if you put a smoother through the data?

```
d[ , plot(x=running, y=Y, col=berkeley_blue, pch=19)]
```

```
## NULL
```

```
  d[ , lines(lowess(running, Y), col=california_gold, lwd=10)]
```




```
## NULL
```

- What if you break that smoother at the policy point?

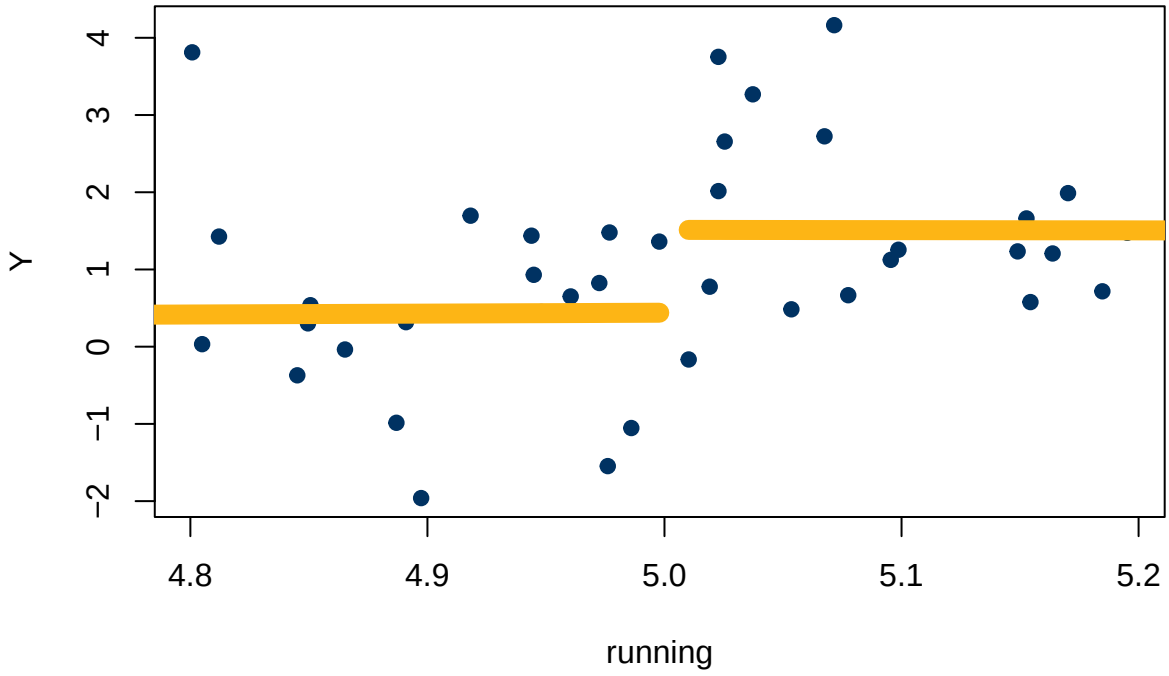
```
d[running < 5.2 & running > 4.8 , plot(x=running, y=Y, col=berkeley_blue, pch=19)]
```

```
## NULL
```

```
    d[running < 5 , lines(lowess(running, Y), col=california_gold, lwd=10)]
```

```
## NULL
```

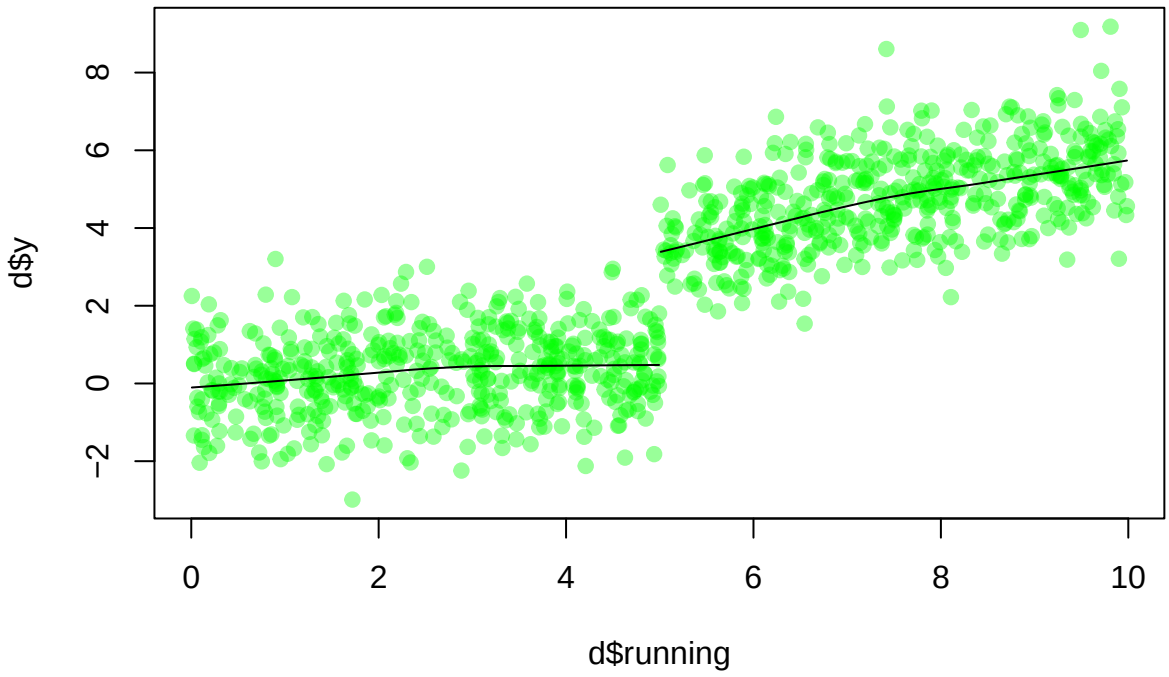
```
    d[running > 5 , lines(lowess(running, Y), col=california_gold, lwd=10)]
```




```
## NULL
```

10.6.5 What about even more challenging data?

```
d <- data.frame(running = runif(1000, min = 0, max = 10),
                cov1    = rnorm(1000))
d$y <- d$running * 0.1 - .2 * d$cov1 + 1 * I(d$running > 5) +
    .4 * d$running * I(d$running > 5) + rnorm(1000)

plot(x = d$running, d$y, pch = 19, col = rgb(0,1,0, .4))
lines(lowess(d$running[d$running < 5], d$y[d$running < 5]))
lines(lowess(d$running[d$running > 5], d$y[d$running > 5]))
```



- What model would you fit against this data?

11 Problems and Diagnostics

Figure 9: houston, we have a problem

11.1 Learning Objectives

At the conclusion of this week's live session, students will be able to:

1. **Recall** and enumerate the several types of problems that might arise in the process of conducting a *real-world* experiment.
2. **Identify** and **diagnose** when these problems are present in *their* experiment.
3. **Analyze** the consequences of these problems, and communicate to an audience how these problems affect the experiment's ability to produce an unbiased causal estimator.
4. **Propose** a way forward – how to modify the experiment or modify the estimator.

11.2 Goals of an Expermient

Recall that the goal of any experiment – same as the goal for *your* experiment – is to identify a causal effect that we have a **guarantee** is unbiased. If we cannot accomplish that, why take the time to conduct the experiment?

Conducting these experiments is **easy**! At least in a perfect world. The reality is that this world that we live in, and that your data generating process functions within, is far from "perfect". And so, as a consequence,

machines that we tell to record data fail; people that we tell to take treatment don't actually take it; doses that we think will be sufficient are not; and a litany of other small concerns!

This week's task is to acknowledge that nothing will *ever* be perfect, and then to proactively design our data collection system so that we can diagnose problems when they arise.

11.3 Problems with Randomization

Suppose that you are going to conduct an experiment that evaluates the effectiveness of encouragements to complete homework and you have designed the operational details in the following way:

1. You have randomized the list of currently enrolled MIDS 241 students into three treatment groups. Group A will receive **Always Encouragement**, their instructor will Slack them every day and remind them that Problem Set 4 is coming due; Group B will receive **Baseline Encouragement**, where they will receive the normal communication from the class; Group C will receive **Conditional Encouragement**, where students will only be contacted if their grades are below the class average.
2. Although you've designed the experiment, contact for students will require that the individual classroom instructors take action to send the encouragements.

1. **How could this randomization break?** In this breakout, list as many specific ways that you think that the list of people who are assigned to Group A, B, and C might not match those who actually are in the groups. How many ways can you imagine?
2. **What are the consequences?** After listing all the ways that the randomization *could* breakdown, what are the three ways that you think would be more damaging to generating an unbiased causal estimate?
3. **How would you detect?** For these three problems that you've identified, how would you know if the randomization has broken down in this particular way?

11.4 Placebo Test

Earlier in the semester, we talked about Placebo Designs, where we intentionally randomize a group of participants and give them a treatment that we do not think will affect outcomes. This placebo treatment, which you might think of as a sugar pill, is designed to allow compliers (and non-compliers) to identify themselves to the experimenting team, producing more efficient estimates of the compliers average causal effect when there is a lot of treatment non-compliance.

This week, we introduce the concept of the *Placebo Test* which, importantly, is different than a *Placebo Design*.

What is a placebo test? How is the concept similar to, and how is it different from, a placebo design?

- How do you know if a particular feature is a good candidate for a placebo test?
- How are placebo tests different from covariate balance checks?
- If you fail a placebo test, why is there more uncertainty about how to fix your experiment?

11.5 Manipulation Check

What happens if you don't use a *strong enough* treatment?

Example 11.1. Suppose that you're conducting an experiment to learn the market-value effect of holding a MIDS degree from the School of Information. (Gosh, I sure hope it is positive!) You decide that your experiment is going to send two versions of a resume.

1. **Version 1 (Control)**: The resume contains the candidate's name, work experience, skills, and a nice statement of their purpose searching for a job. But, the resume does not contain any information about a job candidates master's degree.

2. **Version 2 (Treatment)**: The resume contains the candidate's name, work experience, skills, and a nice statement of their purpose searching for a job. Then, **after the other materials** the resume lists education at the top of the second page.

- What concerns do you have about the strength of this manipulation? How would you address these concerns?
- What is the maximum manipulation that you can imagine using in this experiment? Are you concerned that there might be *too* much manipulation?

11.6 Advocating for Experimentation

Getting experiments done at work is *almost alwasys* an uphill battle! We won't enumerate the reasons (because we will ask you to do so in a moment), but in our experience, there are some serious impediemnts to getting experiments done.

The good news, from the point of view of your company's leadership: A/B Testing is Dead.

Using the tools of this class and the program, in breakout rooms evaluate this claim. All breakout rooms will read and learn about this claim.

One set will argue in support of the claim: Using new techniques like reinforcement learning and machine generated text, it is possible to determine the most effective messaging without needing to conduct randomized experiments.

The other set of breakout rooms will argue against the claim: While these new techniques might afford some benefits, they cannot replace the work that we've been building over the course of the semester.

11.6.1 Reasons for and against Experiments

Think about getting an experiment conducted either at work or in your lab.

- What are the reasons that you can identify *in support of* conducting an experiment?
- What are the limitations to, or reasons that you might not conduct an experiment?

12 Attrition, Mediation, and Generalizability

The theme for this week, as we mention in the async, is that these are **hard** problems – in fact, each of these problems are so hard that we do not have an ability to place a clear, numerical answer on *any* of them.

12.1 Learning Objectives

At the conclusion of this week, students will be able to

1. **Recognize** attrition, **distinguish** the differences between attrition and compliance; **design** an experimental protocol to minimize the amount of attrition that is present in their data; and **analyze** an experiment that has experienced attrition to provide best-possible, defensible estimates of treatment.
2. **Reason** about *why* one things causes another; **reason** about how this affects the ways that they design an experiment or treatment; but, also **communicate** why it is so difficult to produce clear evidence about *why* something has an effect.
3.

12.2 Why doesn't mediation analysis work?

Here's a classic case, that is actually very recent. Gaesser et al (2020) present subjects with a short text that describes a stranger in need, for example, someone who has fallen off a motorcycle on the freeway.

- **Treatment Group** members were asked to imagine helping the person who had fallen of the motorcycle.
- **Control Group** members were asked to critique the the writing style of the text that they read.

- **Both Groups** were shown the same text.

Unsurprisingly, the authors found that the episodic simulation treatment increased individuals willingness to help the stranger in need. But why?

The authors suppose that there are three possible reasons why episodic simulation might work differently.

1. It might work differently depending on how well someone can visualize the scene (*scene vividness*) measured by response to the question, "The imagined scene of helping in your mind was [1. not coherent ... 7. coherent]."
2. It might work differently depending on how well someone can visualize the person (*person vividness*) measured by response to the question, "Did you visualize the person in your mind?" [1. No, not at all ... 7. vividly, as if currently there].
3. It enables empathetic thought (*perspective taking*) measured by response to the question, "Did you consider the other person's thoughts and feelings?" [1. No, not at all ... 7. Strongly considered.]

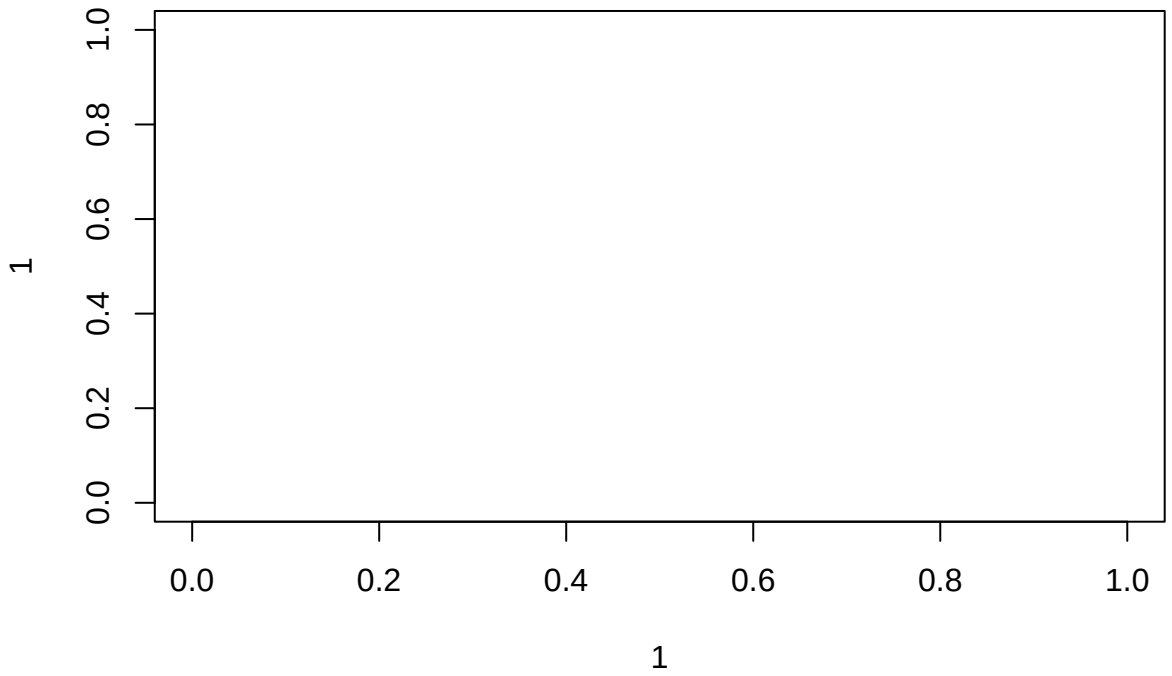
Implicit mediation analysis works in the following way:

$$
\begin{aligned}
lm(M \sim \alpha_1 + aX_i + \epsilon_1) \\
lm(Y \sim \alpha_2 + cX_i + \epsilon_2) \\
lm(Y \sim \alpha_3 + bM_i + c'X_i + \epsilon_3)
\end{aligned}
$$

Where people talk about c as the "total effect" of X on Y, and the "direct effect" of X on Y as the estimate that is reported in c'.

Can you draw this system out in the way that we did in 203? Use circles to represent concepts that you're measuring, and directed arrows to represent causal relationships between these concepts.

```
plot(x=1,y=1, xlim = c(0, 1), ylim = c(0,1), type = 'n')
```



Once you've written out these pathways, what could go wrong in this analysis?

12.3 Endless Chain of Why?

Return back to the example that we talked about at the *very* first week of class, that living in a suburban environment causes a measurable increase in people's BMI. In a five-minute breakout room, produce an enumerated list of theories about *why* living in the suburbs might increase someone's BMI.

1. The team that lists the greatest number possible causes gets a gold star.
2. The team that lists the most hilarious possible cause also gets a gold star.

12.4 Design an experiment to evaluate these possible causes

Now that we've got the list of causes created, and discussed, let's pick a smaller set of the possible causes, and have each group go back into their breakout room for five more minutes to **specifically** design an experiment that would produce evidence in support of (or in contrast to) their specific theory. Here's the thing: in creating your test, you're trying **as best as possible** to isolate one, and only one mechanism. So, an experiment that is able to change only a single mechanism is preferred to an experiment that tests who mechanisms at once.

When we come back from this breakout, each team will spend three minutes presenting the design that they produced to test their theory, and the other groups will reason about whether there are other mechanisms that *could* be at play in producing differences in outcomes.

12.5 Generalizability

Recall the Arizona towel example that we read in *Field Experiments*. It goes something like this, "There is a door hanger that goes into the bathroom of a Best Western that asks individuals to reuse their towels in an effort to lower environmental impact. There is a large effect in the first period of the study, but there is a smaller, and statistically insignificant effect in the second period of the study.

Bates and Glennerster suggest four misguided approaches that might better be called, ways that other people think about generalizability, but the headings are rather misleading. Recast the headings into four more descriptive sentences instead. I'll do the first for you:

1. An effect learned in a particular context (or location) can never be informative of another location.
2.
3.
4.

Bates and Glenerster suggest a second four-item way to instead reason about generalizability. What are these four steps?

1.
2.
3.
4.

Describe what each step means?

Now, suppose that you're the decision-maker who has to decide whether to run the experiment signs about towel re-use in Arizona (now for a third trial). How woud you use the four-step framework to evaluate whether to run another experiment?

Throughout the async, David Broockman highlights the extreme difficulty in generating data that tests mechanisms. So, isn't the Bates and Glennerster argument tantamount to saying, "Just think about this impossible thing that you're never going to be able to measure?" Or, can you use their framework profitably to generalize to other contexts?

13 Applications of Experiments

13.1 Learning Objectives

1.
2.
3.

14 Review of the Course

14.1 Learning Objectives

1.
2.
3.