Causal Inference I

Mixtape Session

Roadmap

- Introduction to course
  - History of causal inference
  - Design versus Model
- Potential outcomes
  - Naive correlation
  - Potential outcomes notation
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
  - Example of physical experimentation: HIV status
- Randomization inference
  - Lady tasting tea
  - Fisher's sharp null
  - Alternative test statistics

Welcome to Mixtape Sessions!

- Mixtape Sessions is an educational platform designed to "democratize causal inference" at all levels by helping bridging people with teachers
- Causal inference, in my mind, is an *applied* field as much as it is a *technical* field and so learning more about it is to also learn about a range of topics not normally covered in an econometrics course
- These include econometric estimation, detailed exposition of research design elements, but also coding practices, handling of data, more detailed dives on specific topics and even advice on publishing and communicating results

Why Mixtape Sessions

- Scott Cunningham, Professor of Economics, Baylor University (about me)
- Not an econometrician; more of a run of the mill typical applied microeconomist
- Workshop is about the science but also the art of causal inference (as is the broader Mixtape Sessions platform)

Class goals

1. **Confidence**: You will feel like you have a good understanding of design-based causal inference by the end such that it doesn't feel so mysterious or intimidating
2. **Comprehension**: You will have learned a lot both conceptually but also in various specifics, particularly with regards to issues around identification and estimation
3. **Competency**: you will have had some experience working together implementing these methods using code in Stata and R, syntax, possession of programs, knowledge of packages

4-day Causal Inference Workshop

- Our workshop together is 4-days, 9am to 6pm PST, with 15 min breaks on the hour and a 1-hour lunch break at 12:00PM PST
- It will mix exposition, discussion of papers, coding exercises and discussion as best as I can
- It’s essentially a semester’s worth of material

Causal Inference table of contents

Causal Inference I				
	Day 1	**Day 2**	**Day 3**	**Day 4**
Pre-Class	*Read Thornton AER*	*Read Dehejia and Wahba RESTAT*	*Read Card Econometrica*	*Read Hansen AER*
9am	Potential outcomes and counterfactuals	Causal graphs	Instrumental variables, intuition, and 2SLS	Introduction to regression discontinuity design
10am				
11am		**Lalonde Coding Lab Part 1**	Weak instruments and 2SLS bias	Nonlinearities and estimation
12pm	Lunch			
1pm	Randomization, selection bias	Matching	Heterogenous treatment effects & LATE	RD Shiny App
2pm	**Thornton Coding Lab Part 1**			Nonparametric estimation
3pm				
4pm	Randomization inference		**IV Coding Lab**	General tips (data visualization, density tests, etc)
5pm	RI Shiny App			**Hansen Coding Lab**
6pm	**Thornton Coding Lab Part 2**	**Lalonde Coding Lab Part 2**	Judge IV Design	

Workshop (Part 1) Topics

1. Foundations: Day 1
2. Graphs and Matching: Day 2
3. IV: Day 3
4. RDD: Day 4

What is causality

> *"Causation is something that makes a difference, and the difference it makes must be a difference from what would have happened without it." – David Lewis (philosopher)*

Key idea $\rightarrow$ counterfactual. Counterfactuals are neither past nor future. They are alternative histories created by thought experiments but we use them as framing devices to decipher causality in our timeline

Causal inference is also fundamentally practically because if we know causality, then we can also know not only the past, but also the future (policy)

Different types of prediction

Traditional prediction

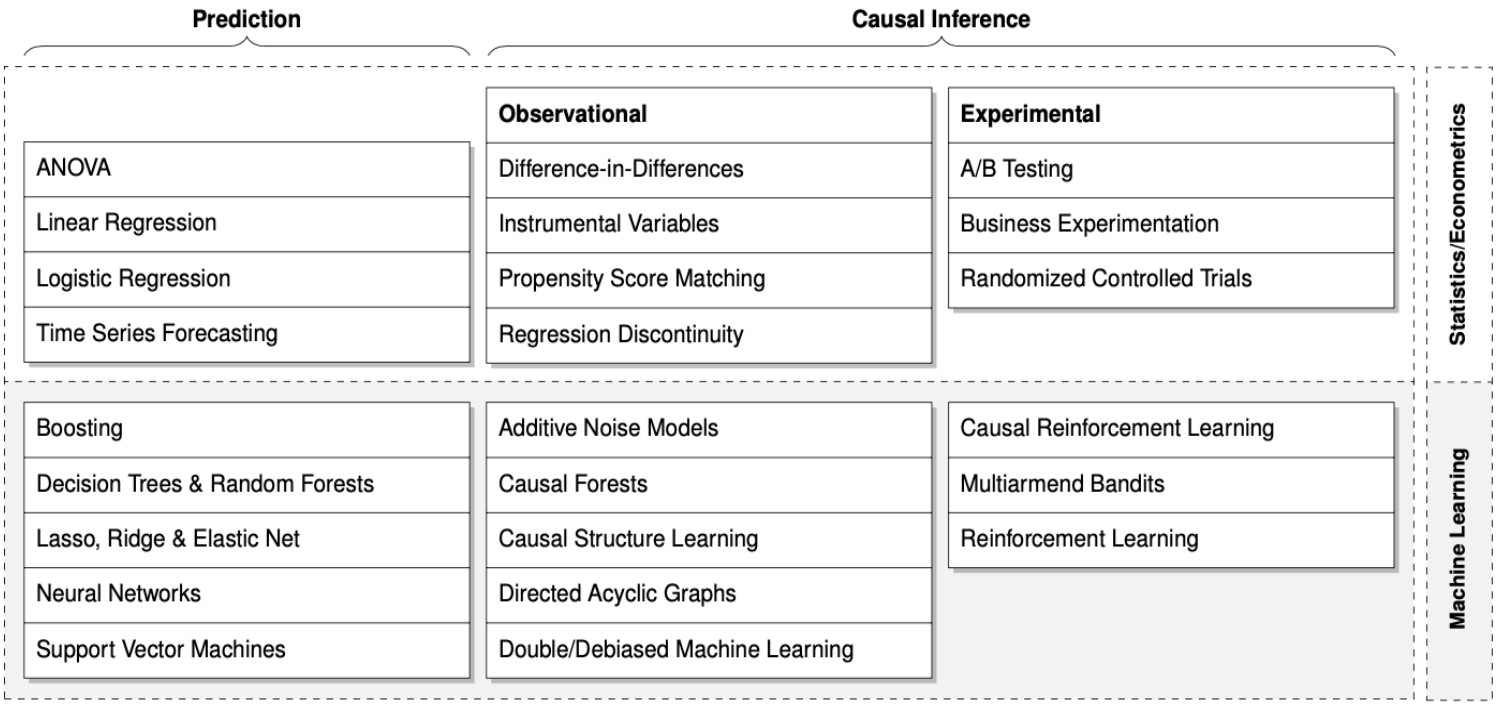
- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- “Does this person have heart disease?”, “How many books will I sell?”
- It is not predictions of what effect a choice will have, though

Causal inference

- Causal inference is also a type of prediction, but it’s a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

Identification problem

Figure 1: Examples of popular data analysis algorithms in statistics and econometrics, as well as machine learning and artificial intelligence, classified according to prediction and causal inference methods. Causal inference methods are further differentiated according to observational (based on ex-post observed data) and experimental approaches.



Lost credibility within empirical economics

- Some history can help sometimes understand where all this comes from
- Tremendous turmoil in applied economics in late 1970s and early 1980s
- Several high profile papers found serious credibility problems in many applied studies, particularly those using macro data
- This criticism eventually turned to labor economics (Lalonde 1985; Ashenfelter and Card 1985)

Job training

- Bob Lalonde (1985) evaluated an RCT of a job training program; ground truth was known
- Cleverly dropped the control group and replaced it with CPS and PSID, re-analyzed the data
- Impossible to get the answer without specification searching (“peeking”, p-hacking)

Lewis 1986 book

"After reviewing virtually every study since 1963, Lewis reached the awkward conclusion that simple OLS of union wage effects were more useful and reliable than those based on IV or endogenous selection approaches. The problem, in his view, was that researchers used arbitrary and unsupported assumptions to identify their models with little or no concern for the validity of their assumptions or the implications of their findings. This criticism was particularly salient because many of the new methods had been tested initially on the union wage effect question (e.g., Lee 1978)." – Card (2022)

Replication problems

- Deward, Thursby and Anderson (1986) suggest that applied estimates can't be replicated partly bc many authors won't share their data or their programs
- They tried to replicate the papers at the JMCB journal between 1980 and 1984 (there was a data sharing agreement in 1982 and 1/4 authors still won't respond to request; before that 2/3 failed to respond)
- Only 2 of 9 papers could be reproduced exactly; 5 had substantial errors

Take the con out of econometrics

- Famous paper by Ed Leamer – “Take the con out of econometrics”
- He says in Hendry, Leamer and Poirier (1990), “we don’t take empirical work seriously in economics. It’s not the source by which economists accumulate their opinions, by and large.”
- Hits hard at Princeton’s Industrial Relations Section where Card, Ashenfelter, Krueger, Angrist, Lalonde, and more are. They seek to fix this

Princeton IRS Model

- Ashenfelter 1970s work in government focuses on extreme self-selection into job trainings programs (“Ashenfelter dip”); Ashenfelter and Card (1985) note longitudinal studies are not designed for this
- Princeton Industrial Relations Section Model: more transparency, more credible “research design” that make explicit sources of identification and work hard to verify the legitimacy of that source
- Krueger notes that he would read NEJM and they’d often include a short description of the study’s “research design”

Natural experiments

- Richard Freeman had been pushing for natural experiments for years – “big shocks” like federal minimum wages in Puerto Rico
- Card really pushes this early on
  - → Immigration labor supply shifts from the Mariel Boatlift paper (Card 1990);
  - → Minimum wage increases from neighboring state comparisons (Card and Krueger 1994)
- Steven Levitt seems instrumental in the late 1990s for bringing natural experiments and this labor economics approach to crime, more or less cleaning house; also very good abt sharing data and programs

Theory's role

- Economists write models to help them understand the world
- Economists also use those models to guide their empirical work
- But the connection between those two steps has not always been done the same; let's review as this framework can help give you the big picture

Design vs Model

- **Model**: Causality exists within the framework of a theory that says “D causes Y” (e.g., Heckman)
- **Design**: Causality is design-based and can be discerned with *physical* manipulation of a treatment D (e.g., Rubin, Holland)

Approximating models

1. **Approximating models**: Consumer demand, labor supply models (e.g., Mincer 1958; 1974)
   - → Theory implies $y_i = f_i(x_i)$ with restrictions on f_i (e.g., concavity)
   - → Researcher estimates a simpler version

$$y_i = \alpha + x_i\beta + \varepsilon_i$$

Exact models

2. **Exact models**: Models gives us all causes (“complete DGP”)
   - → More structural approach to identification, less focused on physical assignment of treatments
   - → Estimate model parameters and distribution of heterogeneity
   - → Functional form, useful for welfare analysis

Working model

3. **Working model**: Called reduced form, program evaluation, more like Princeton's IRS model
   - → Model formulates questions, intuition, but does not necessarily assist with identification
   - → Focus is on physical assignment of treatments not on modeling assumptions
   - → Extreme focus on verifying the assumptions through placebos, falsifications, checks for identifying assumptions throughout

Models are useful, but not necessarily *used*, the way they are in structural approaches

Topics broaden

Dependence on the model vs freed from the model for causal inference increases topics

- **Design**: Anything goes, “economics is what economists study”, happiness, fringe stuff (e.g., sex work) (opening up topics)
- **Model**: Neoclassical topics due to needing agreed upon models (limiting topics)

Strengths of design

- Approaches are oftentimes easier to explain; identification (historically a mathematical term) becomes synonymous with research design (Imbens and Angrist 1994)
- Precise research designs, as we will see, make it much easier to evaluate the core assumptions of the model (like employing event studies in a diff-in-diff)
- Designs become portable – we can repeat these approaches in other settings (which is in many ways what this class is doing)

Weaknesses

- Limits the questions we can answer oftentimes
  - → It's very backwards looking (may be a strength insofar as it's conservative)
  - → Lacks generalizability
- Listen to Petra Todd describe these weaknesses

  https://youtu.be/m1Mpc7-b-1I?t=2776

Roadmap

- Introduction to course
  - History of causal inference
  - Design versus Model
- Potential outcomes
  - Naive correlation
  - Potential outcomes notation
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
  - Example of physical experimentation: HIV status
- Randomization inference
  - Lady tasting tea
  - Fisher's sharp null
  - Alternative test statistics

Introduction to Counterfactuals

- Aliens come and orbit earth, see people dying in hospitals and conclude “doctors are hurting people”
- They kill the doctors, unplug patients from machines, throw open the doors – many patients inexplicably die
- *We are the aliens in our research*

#1: Correlation and causality are different

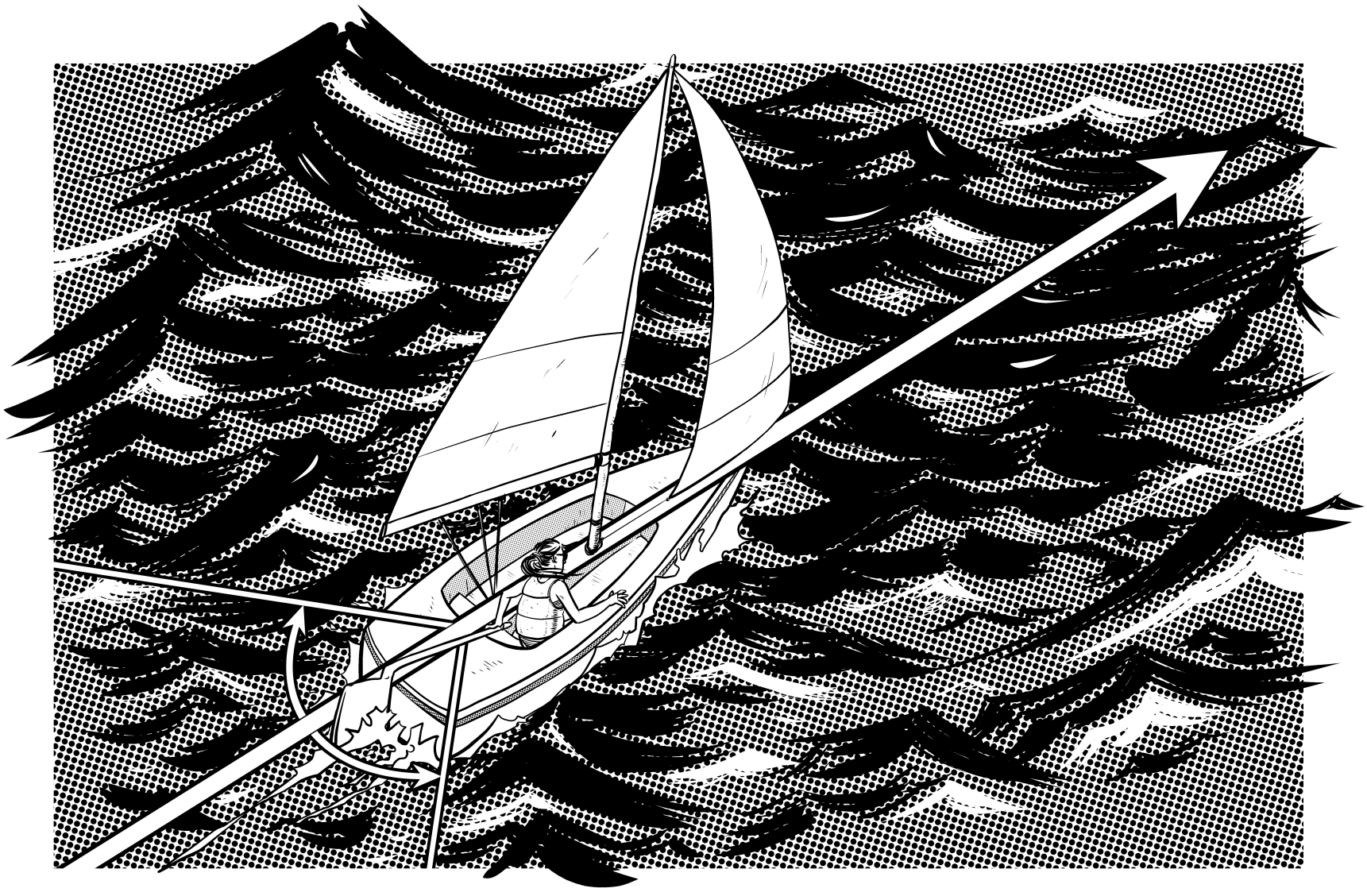
Causal is one unit, correlation is many units

- Causal question: “If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?”
- Correlation question:

$$\frac{Cov(D,Y)}{\sqrt{Var_D}\sqrt{Var_Y}}$$

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- What if cat killed the rooster?
- *Post hoc ergo propter hoc*: “after this, therefore, because of this”



#3: No correlation does not mean no causality!

- A sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- The same aliens observe from space and say “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken.” So they send her a new rudder
- They’re wrong but why are they wrong? There is, after all, no correlation
- Example: Fed and open market operations

Potential outcomes

- Conceptual framework for design based causal inference is counterfactual reasoning
- Counterfactual modeling becomes linked to potential outcomes models with Don Rubin 1970s work on propensity scores
- Potential outcomes notation was created by Jerzy Neyman (1923) and led Ronald Fisher (1925) to suggest RCTs
- Huge push by Donald Rubin, a Neyman "grandson", in the 1970s and 1980s with Paul Rosenbaum on the propensity score
- Guido Imbens, a Rubin coauthor, notes that it was crucial to his work on IV with Angrist, but it may even have made that work more appealing *outside economics*

  https://youtu.be/cm8V65AS5iU?t=1097

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if hospitalized at time } t \\ 0 \text{ if not hospitalized at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1 \text{ health if hospitalized at time } t \\ 0 \text{ health if not hospitalized at time } t \end{cases}$$

where j indexes a counterfactual state of the world

Realized vs potential outcomes

- Potential outcomes Y^1 are not the realized outcomes Y either conceptually or notationally
- Potential outcomes are hypothetical states of the world but realized outcomes are *ex post* the observed outcomes we have in our data due to treatment assignment
- This distinction is subtle, creates challenges at the introductory stage, it isn't how econometrics was historically taught, except for at the very beginning; again Imbens on this
  https://youtu.be/cm8V65AS5iU?t=1175

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, equals $Y_i^1 - Y_i^0$

Definition 3: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

Definition 2: Switching equation

An individual's observed health outcomes, Y, is determined by treatment assignment, D_i, and corresponding potential outcomes:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}$$

Missing data problem

- Causal inference is fundamentally a missing data problem requiring prediction, not of the present or the future, but of a missing past – sometimes explicitly (nearest neighbor, synthetic control), sometimes implicitly (RDD, IV)
- Aggregate parameters based on individual treatment effects are descriptions of causal effects
- Fundamental problem of causal inference holds because of the switching equation *even with big data*

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Conditional Average Treatment Effects

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Any collection of treatment effects

- Notice how in all three of these, all we did was take the defined treatment effect at the individual and aggregate
- We will see this again with IV when we introduce the “local” average treatment effect
- Just keep in mind – these parameters can be defined, but they cannot be calculated due to the switching equation

Good and bad variation

- Naive use of statistical models will often find and take advantage of all types of variation for the purpose of prediction
- But causal inference is much more cautious because it only uses *some* of the variation
- This is better seen with a story and a decomposition

Causality and comparisons

- Epistemology: what beliefs are warranted and what beliefs are not
- Without counterfactuals, we do not *know* treatment effects, but with groups of data we can sometimes obtain *estimates*
- We do this by making comparisons of groups treated and not treated
- But not all comparisons are equal – selection bias (e.g., aliens making unwarranted conclusions about causality because of failing to use design)
- We will decompose a simple estimator so we can see what *selection bias* is

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by the sample averages:

$$\begin{aligned} SDO &= E[Y^1|D=1] - E[Y^0|D=0] \\ &= E[Y|D=1] - E[Y|D=0] \end{aligned}$$

I tend to use expectation operators $E[\cdot]$ but note we are using samples $E_N[\cdot]$

SDO

- Simple difference in mean outcomes is our first estimator
- Notice that we switched from potential outcomes to observed outcomes
- This means that because the SDO is based on the switching equation, it uses data
- So when is the SDO causal and when is it not?

Potentially biased comparisons

Decomposition of the SDO

The SDO can be decomposed into the sum of three parts:

$$
\begin{aligned}
E[Y^1|D=1] - E[Y^0|D=0] \quad &= \quad ATE \\
&\quad + E[Y^0|D=1] - E[Y^0|D=0] \\
&\quad + (1-\pi)(ATT - ATU)
\end{aligned}
$$

Seeing is believing so let's work through this identity!

Use LIE to decompose ATE into the sum of four conditional average expectations

$$\begin{aligned}
\text{ATE} &= E[Y^1] - E[Y^0] \\
&= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\
&\quad -\{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\}
\end{aligned}$$

Substitute letters for expectations

$$\begin{aligned}
E[Y^1|D=1] &= a \\
E[Y^1|D=0] &= b \\
E[Y^0|D=1] &= c \\
E[Y^0|D=0] &= d \\
\text{ATE} &= e
\end{aligned}$$

Rewrite ATE

$$e = \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\}$$

Move SDO terms to LHS

$$
\begin{aligned}
e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\
e &= \pi a + b - \pi b - \pi c - d + \pi d \\
e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\
0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)
\end{aligned}
$$

Rewrite from previous slide

$$\mathbf{a} - \mathbf{d} \quad = \quad e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)(a - c) - (1 - \pi)(b - d)$$

Substitute conditional means

$$\begin{aligned}
E[Y^1|D = 1] - E[Y^0|D = 0] \quad &= \quad \text{ATE} \\
&\quad +(E[Y^0|D = 1] - E[Y^0|D = 0]) \\
&\quad +(1 - \pi)(\{E[Y^1|D = 1] - E[Y^0|D = 1]\} \\
&\quad -(1 - \pi)\{E[Y^1|D = 0] - E[Y^0|D = 0]\}) \\
E[Y^1|D = 1] - E[Y^0|D = 0] \quad &= \quad ATE \\
&\quad +(E[Y^0|D = 1] - E[Y^0|D = 0]) \\
&\quad +(1 - \pi)(ATT - ATU)
\end{aligned}$$

Decomposition of difference in means

$$\underbrace{E_N[y_i|d_i=1] - E_N[y_i|d_i=0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}} + \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

Using the switching equation, we get $E_N[Y|D=1] \rightarrow E[Y^1|D=1]$, $E_N[Y|D=0] \rightarrow E[Y^0|D=0]$ and $(1-\pi)$ is the share of the population in the control group.

Selection bias

- Notice this term “selection bias”

$$E[Y^0|D=1] \neq E[Y^0|D=0]$$

- Selection bias was the problem that the aliens failed to overcome – without treatment Y^0, COVID patients on vents $(D=1)$ would likely have been different from those with COVID not on vents $(D=0)$

What is selection bias?

Let's put this into words.

1. Put $E[Y^0|D=1] \neq E[Y^0|D=0]$ to your best friend who hasn't taken this course
2. If people choose a treatment, $D=1$, or control, $D=0$, because they expect it benefits them, $Y^1 - Y^0 > 0$, or doesn't $Y^1 - Y^0 \leq 0$ then what do you suspect is true about the mean value of Y^0 for the treatment and control groups?

Illustrating selection bias with spreadsheets

- Chronic PTSD has historically been treated with cognitive behavior therapies like mindfulness, but recent work shows therapist assisted MDMA (street name: ecstasy), are effective too
- Ongoing work in psychopharmacology has begun experimenting with long dormant approaches in the psychedelics and empathogens fot treating mental illness, including PTSD
- MAPS organization has been funding RCTs in compliance with FDA trials to study MDMA's effect on PTSD
  https://www.nature.com/articles/s41591-021-01336-3

Illustrating selection bias with spreadsheets

- Perfect Doctor can accurately determine whether mindfulness practices or MDMA is more beneficial for treating a patient's chronic PTSD ($Y^1 - Y^0$ is positive or negative), and makes treatment assignments ($D = 1$ or 0) depending on its impact
- We will go through an exercise together (copy this google sheet) analyzing the implications of the perfect doctor's choices on a range of statistics, followed by discussion

  https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing

Humans always make bias

- People make choices because they think their live will be better as a result (i.e., based on Y^1 or Y^0)
  1. I chose to get a PhD because I didn't like my life – i.e., Y^0 maybe was different for me than others
  2. I chose to get a PhD because I thought it would help me – i.e., Y^1 maybe was different for me than others
- When humans make choices based on expected gains, it introduces "selection bias" and heterogeneous treatment effect bias
- Rational choice is why correlations in non-experimental data stop illustrating causal effects

Selection bias

- For many of us, we have heard the word “selection bias” before but it was with respect to “non-random samples”
- In causal inference, that isn’t what we mean. We mean $E[Y^0|D=1] \neq E[Y^0|D=0]$ for two groups of people
- But notice, only one of those quantities via the switching equation can be seen, but if they are different, then two groups can look very different and it not reflect a causal effect
- Selection bias was the problem that the aliens failed to overcome – people were on vents because $E[Y^0]$ was much lower than those not on it, and because it probably would help them more than it would other COVID patients $E[Y^1]$

Group Questions for Engagement

1. Some workers work from home and others work at the office and we observe differences in productivity for the ones who work from home. Why might $E[Y^0|WFH]$ be different from the ones who work at the physical office?

Goal of causal inference

Our goal in all of causal inference is to estimate aggregate causal parameters by modeling treatment assignment by *imputing* missing counterfactuals

This imputation process happens sometimes explicitly (nearest neighbor matching) and sometimes implicitly (RCTs)

Let's look what happens in an RCT *and why* this addresses selection bias term $E[Y^0|D=1]$ and $E[Y^0|D=0]$

Independence

Independence assumption

Treatment is assigned to a population independent of that population's potential outcomes

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

This is random or quasi-random assignment and ensures mean potential outcomes for the treatment group and control group are the same. Also ensures other variables are distributed the same for a large sample.

$$\begin{aligned} E[Y^0|D=1] &= E[Y^0|D=0] \\ E[Y^1|D=1] &= E[Y^1|D=0] \end{aligned}$$

Random Assignment Solves the Selection Problem

$$\underbrace{E_N[y_i|d_i=1] - E_N[y_i|d_i=0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}} + \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

- If treatment is independent of potential outcomes, then swap out equations and **selection bias** zeroes out:

$$E[Y^0|D=1] - E[Y^0|D=0] = 0$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D=1] - E[Y^0|D=1]$$
$$\text{ATU} = E[Y^1|D=0] - E[Y^0|D=0]$$

- Rewrite the third row bias after $1-\pi$:

$$\begin{aligned} ATT - ATU &= \textbf{E[Y}^1 \mid \textbf{D=1]} - E[Y^0|D=1] \\ &\quad -\textbf{E[Y}^1 \mid \textbf{D=0]} + E[Y^0|D=0] \\ &= 0 \end{aligned}$$

- If treatment is independent of potential outcomes, then:

$$\begin{aligned} E_N[y_i|d_i=1] - E_N[y_i|d_i=0] &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

SUTVA

- Potential outcomes model places a limit on what we can measure: the “stable unit-treatment value assumption”
    1. **S**: ***s****table*
    2. **U**: across all ***u****nits*, or the population
    3. **TV**: ***t****reatment-value* (“treatment effect”, “causal effect”)
    4. **A**: ***a****ssumption*
- Largely about spillovers, poorly defined treatments and scale

SUTVA: No spillovers to other units

- What if we impose a treatment at one neighborhood but not a contiguous one?
- Treatment may spill over causing $Y = Y^1$ even for the control units because of spillovers from treatment group
- Informs the design stage

SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable” (i.e., notice that the super scripts contain either 0 or 1, not 0.55, 0.27)
- If we are estimating the effect of aspirin on headaches, we assume treatment is 200mg per person in the treatment
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*; you may have to think of it as multiple arms

SUTVA: Scale can affect stability of treatment effects

Easier to imagine this with a different example.

- Let's say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?
- Scaling up a policy can be challenging to predict if there are rising costs of production
- What if expansion requires hiring lower quality teachers just to make classes?
- That's a general equilibrium effect; we only estimated a partial equilibrium effect (external versus internal validity)

Econometrica, Vol. 83, No. 1 (January, 2015), 155–174

CONSUMER HETEROGENEITY AND PAID SEARCH EFFECTIVENESS: A LARGE-SCALE FIELD EXPERIMENT

BY THOMAS BLAKE, CHRIS NOSKO, AND STEVEN TADELIS[1]

Internet advertising has been the fastest growing advertising channel in recent years, with paid search ads comprising the bulk of this revenue. We present results from a series of large-scale field experiments done at eBay that were designed to measure the causal effectiveness of paid search ads. Because search clicks and purchase intent are correlated, we show that returns from paid search are a fraction of non-experimental estimates. As an extreme case, we show that brand keyword ads have no measurable short-term benefits. For non-brand keywords, we find that new and infrequent users are positively influenced by ads but that more frequent users whose purchasing behavior is not influenced by ads account for most of the advertising expenses, resulting in average returns that are negative.

KEYWORDS: Advertising, field experiments, causal inference, electronic commerce, return on investment, information.

1. INTRODUCTION

ADVERTISING EXPENSES ACCOUNT for a sizable portion of costs for many companies across the globe. In recent years, the Internet advertising industry has grown disproportionately, with revenues in the United States alone totaling $36.6 billion for 2012, up 15.2 percent from 2011. Of the different forms of Internet advertising, paid search advertising, also known in industry as "search engine marketing" (SEM), remains the largest advertising format by revenue, accounting for 46.3 percent of 2012 revenues, or $16.9 billion, up 14.5 percent from $14.8 billion in 2010. Google Inc., the leading SEM provider, registered $46 billion in global revenues in 2012, of which $43.7 billion, or 95 percent, were attributed to advertising.[2]

Internet advertising facts

- In 2012, revenues from Internet advertising was $36.6 billion and has only grown since
- Paid search (“search engine marketing”) is the largest format by revenue (46.3% of 2012 revenues, or $16.9 billion)
- Google is leading provider (registered $46 billion in global revenues in 2012 of which 95% was attributed to advertising)

Selection bias

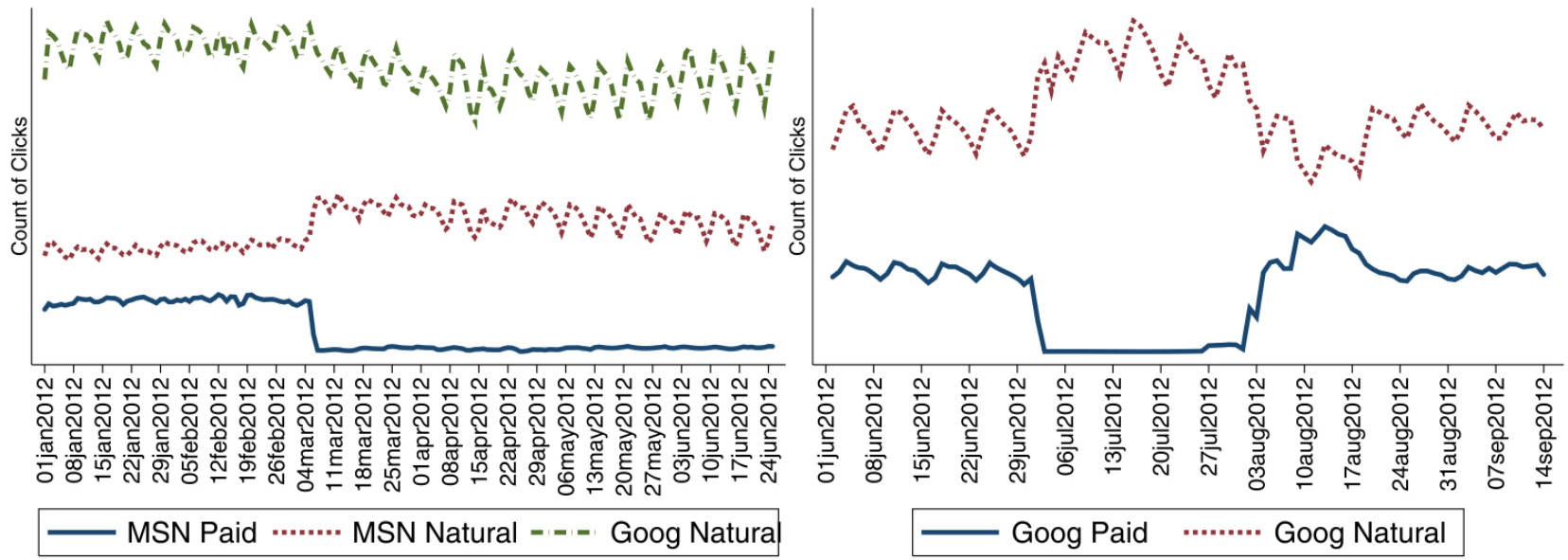
- Treatment was targeted ads at particular people conducting particular types of keyword search
- Consumers who choose to click on ads are loyal and already informed about products with high likelihood to buy already
- Problem is ads are targeting people at the end of their search, so the question is whether they would've found it already (i.e., $E[Y^0|D=1] \neq E[Y^0|D=0]$)

Selection bias

- Estimated return on investment using OLS found ROI of over 1600%
- Compared this to experimental methods and found ROI of -63% with a 95% CI of $[-124\%, -3\%]$, rejecting the hypothesis that the channel yielded short-run positive returns
- Think back to perfect doctor – Even without the treatment (Y^0), the treated group observationally would've still found a way

Natural experiment

- Study began with a naturally occurring and somewhat fortuitous event at eBay
- eBay halted SEM queries for brand words (i.e., queries that included the term eBay) on Yahoo! and Microsoft but continued to pay for these terms on Google
- Blake, Nosky and Tadelis (2015) showed almost all of the foregone click traffic and attributed sales were captured by natural search
- Substitution between paid and unpaid traffic was nearly one to one complete



(a) MSN Test

(b) Google Test

FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for 'ebay' terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

Interpretation of natural experiment

"The evidence strongly supports the intuitive notion that for brand keywords, natural search is close to a perfect substitute for paid search, making brand keyword SEM ineffective for short-term sales. After all, the users who type the brand keyword in the search query intend to reach the company's website, and most likely will execute on their intent regardless of the appearance of a paid search ad."

Selection bias

Observational data masked causal effect (recall the decomposition of the any non-designed estimation strategy)

> *“Advertising may appear to attract these consumers, when in reality they would have found other channels to visit the company’s web-site. We overcome this endogeneity challenge with our controlled experiments.”*

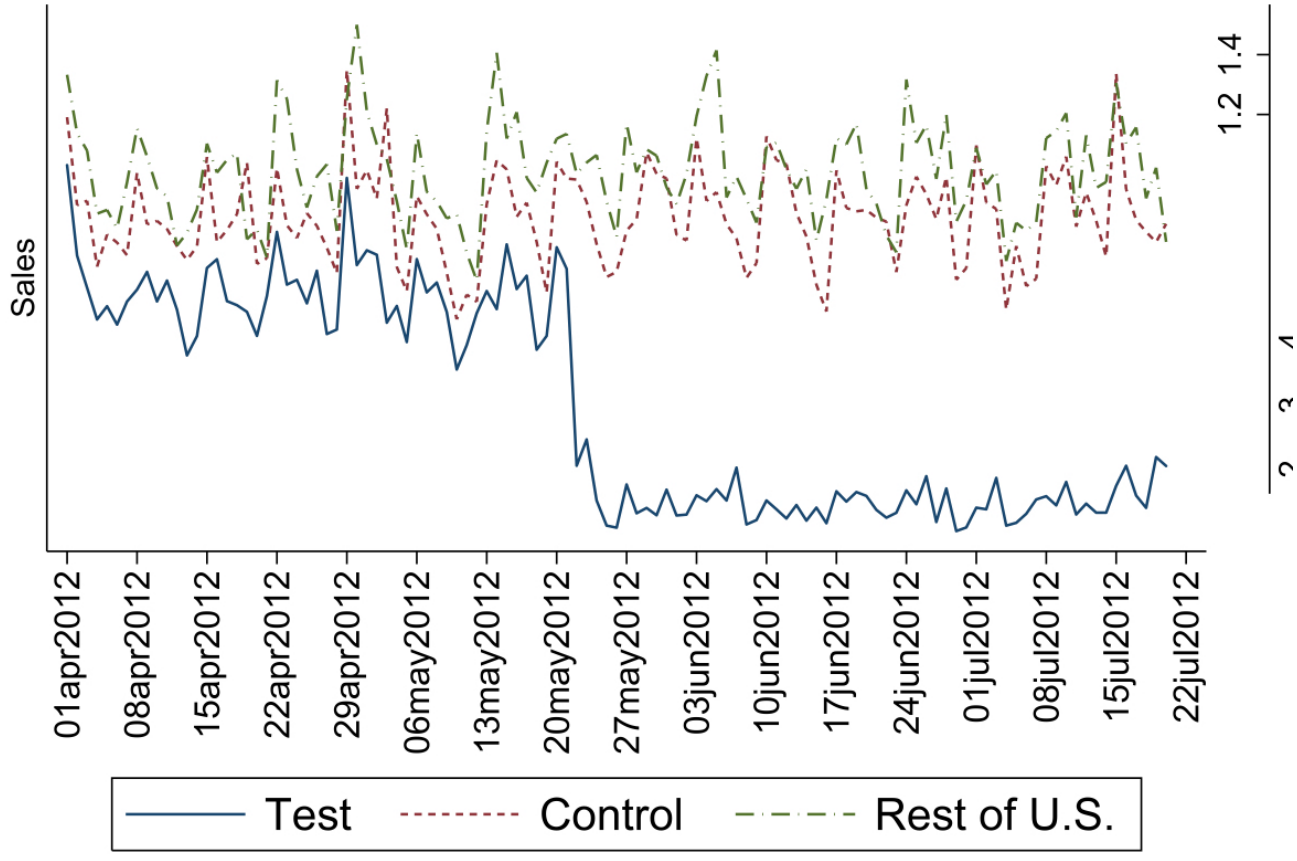
RCT

Natural experiment was valuable, but eBay could run a large scale RCT.

Use this finding of a nearly one-to-one substitution once paid search was dropped to convince eBay to field a large scale RCT discontinuing non-band key words

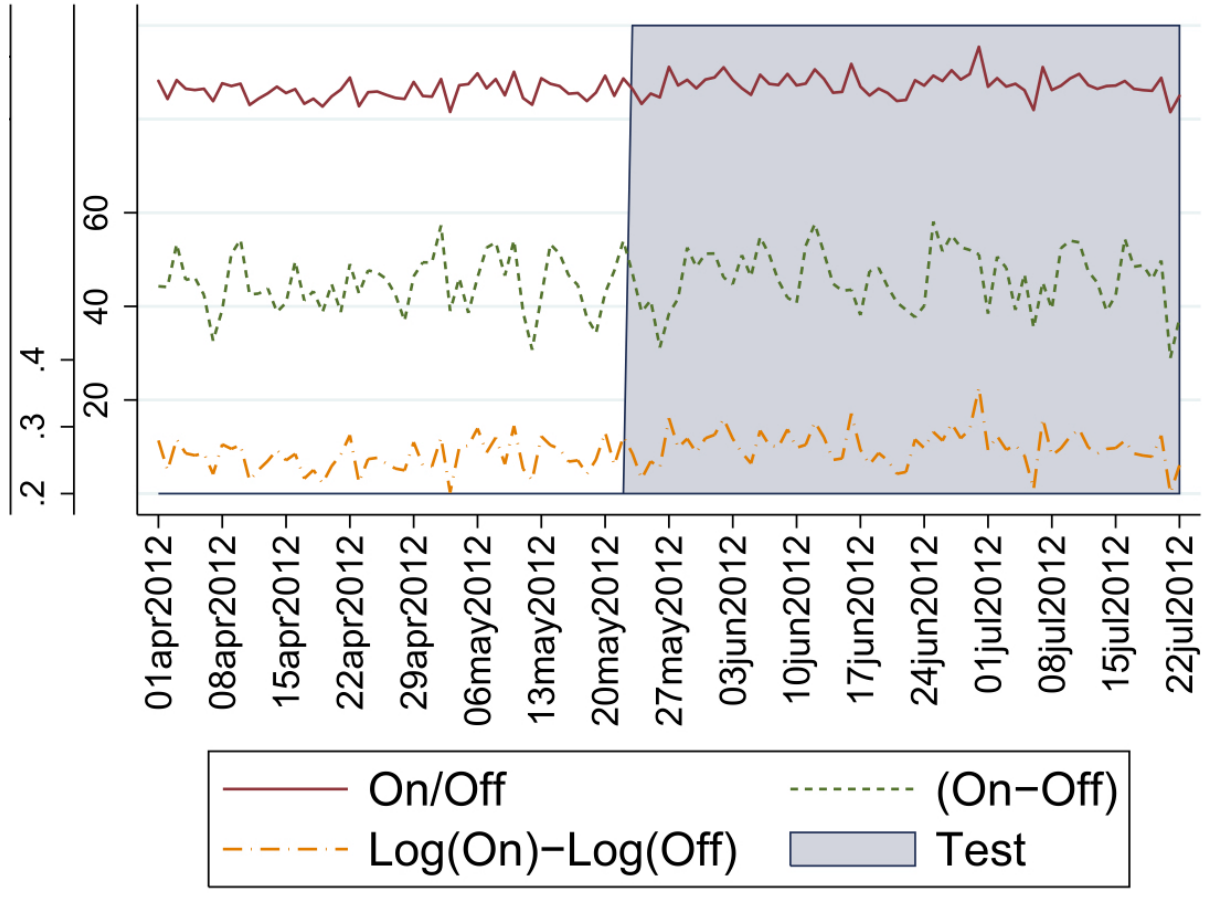
Design of the experiment

- Randomly assigned 30 percent of eBay's US traffic to stop all bidding for all non-brand keywords for 60 days
- Some random group of users, in other words, were exposed to ads; a control group did not see the ads
- Used Google's geographic bid feature that can accurately identify geographic market of the user conducting the search
- Ads were suspended in 30 percent of markets to reduce the scope of the test and minimize the potential cost and impact to the business



(a) Attributed Sales by Region

Figure: Attributed sales due to clicking on a Google link (treatment group)



(b) Differences in Total Sales

Figure: Differences in total sales by market (treatment to control)

	OLS	
	(1)	(2)
Estimated Coefficient	0.88500	0.12600
(Std Err)	(0.0143)	(0.0404)
DMA Fixed Effects		Yes
Date Fixed Effects		Yes
N	10,500	10,500
$\Delta \ln(Spend)$ Adjustment	3.51	3.51
$\Delta \ln(Rev)$ (β)	3.10635	0.44226
Spend (Millions of $)	$51.00	$51.00
Gross Revenue (R′)	2,880.64	2,880.64
ROI	4,173%	1,632%
ROI Lower Bound	4,139%	697%
ROI Upper Bound	4,205%	2,265%

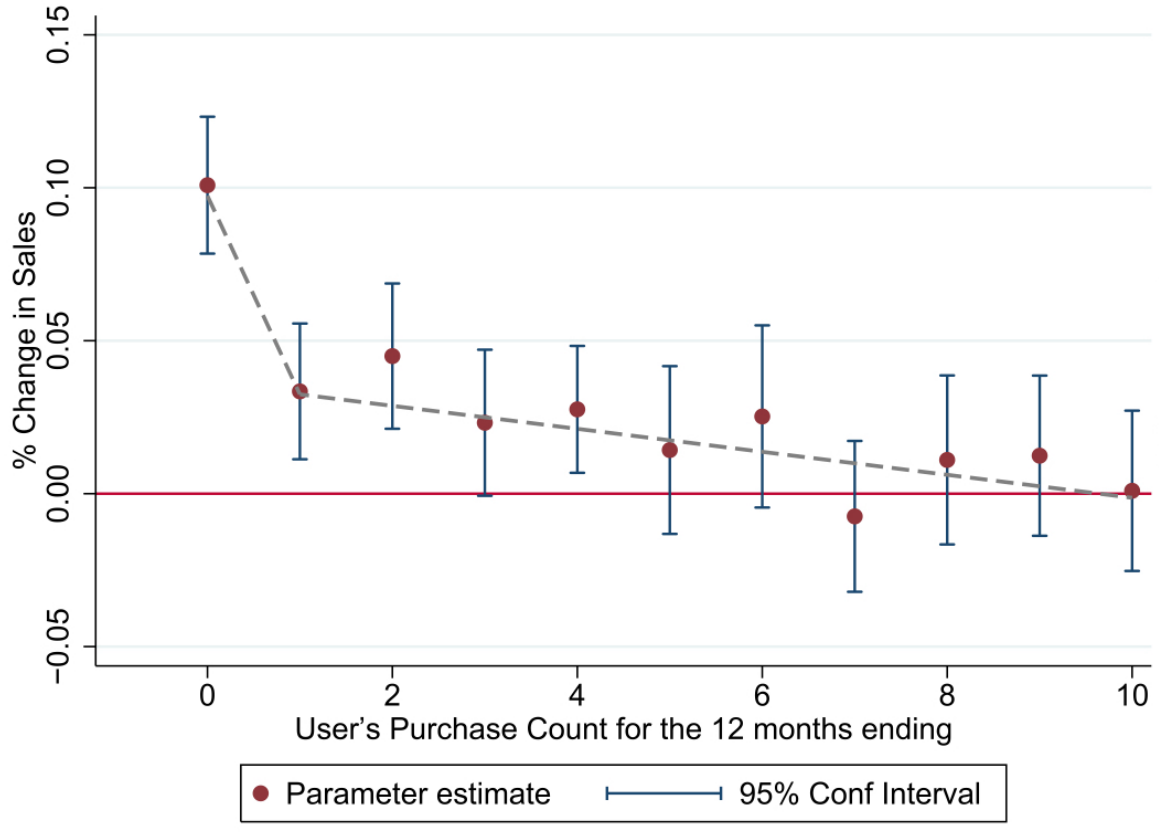
Figure: Spending effect on revenue using OLS but not the randomization. Effects are gigantic.

	(5)
Estimated Coefficient	0.00659
(Std Err)	(0.0056)
DMA Fixed Effects	Yes
Date Fixed Effects	Yes
N	23,730
$\Delta \ln(Spend)$ Adjustment	1
$\Delta \ln(Rev)$ (β)	0.00659
Spend (Millions of $)	$51.00
Gross Revenue (R′)	2,880.64
ROI	−63%
ROI Lower Bound	−124%
ROI Upper Bound	−3%

Figure: Spending effect on revenue using the randomization. Effects are negative.

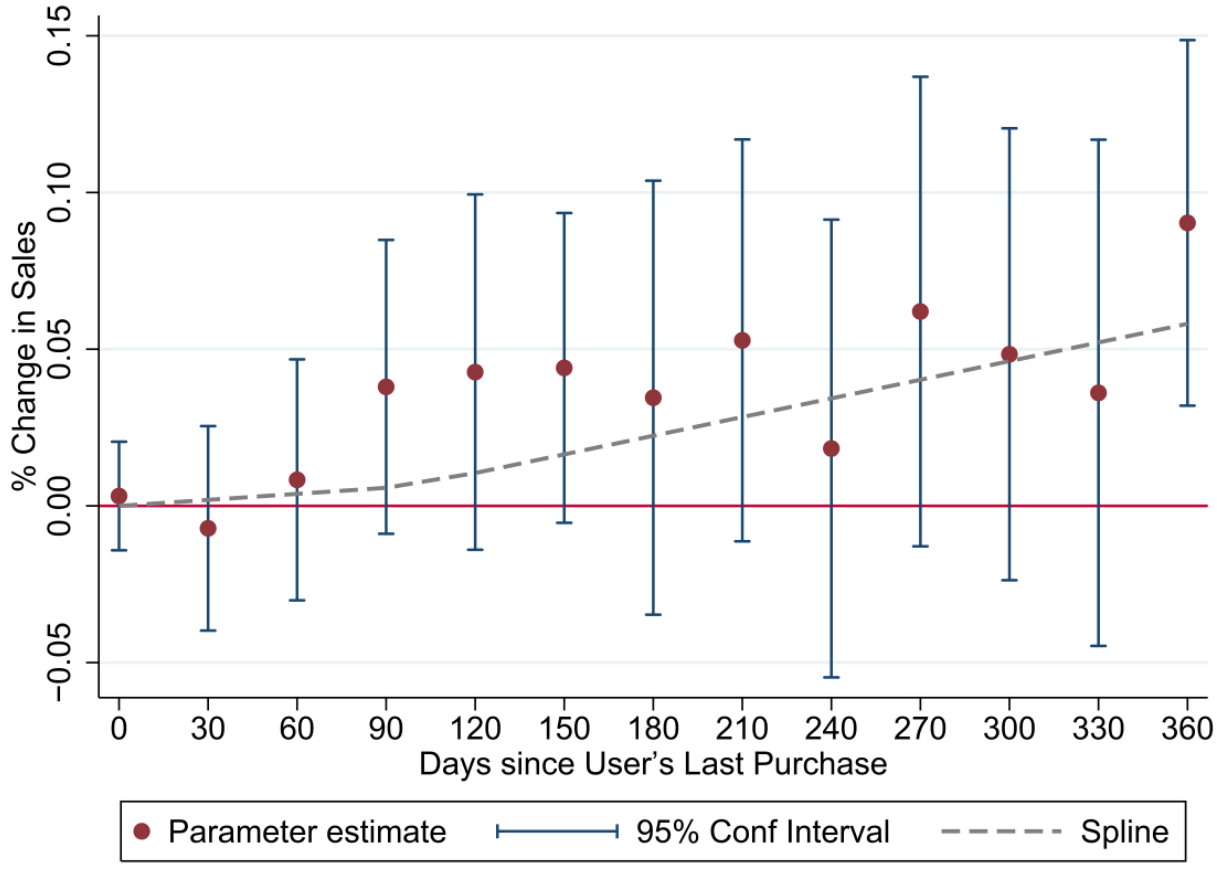
Heterogenous treatment effects

- Recall how the potential outcomes model explicitly models individual treatment effects could be unique and that the perfect doctor showed selection on gains masked treatment effects, perhaps even reversing sign
- Search advertising in this RCT only worked if the consumer had no idea that the company had the desired product
- Large firms like eBay with powerful brands will see little benefit from paid search advertising because most consumers already know that they exist, as well as what they have to offer



(a) User Frequency

Figure: Effects on new users are positive and large, but not others.



(b) User Recency

Figure: Effects are largest for "least active" customers.

Why are causal effects small?

- They suggest that the brand query tests found small causal returns because users simply substituted from the paid search clicks to the natural search clicks
- If that's the case, then it's explicitly a selection bias story

$$E[Y^0|D=1] \neq E[Y^0|D=0]$$

where D is being shown the branded advertisement based on search (i.e., they were already going there)

- They weren't using branded search for information; they were using to *navigate*

Self selection based on gains

- Potential outcomes is the foundation of the physical experiment because the physical experiment assigns units to treatments *independent* of potential outcomes, Y^0, Y^1
- This is important because outside of the physical experiment, we expect people select those important treatments based on whether, subjectively, they think $Y^1 > Y^0$ or $Y^1 \leq Y^0$.
- Rational actors almost by definition are thought to “self-select into treatment” making non-designed comparisons potentially misleading – sometimes by a little, sometimes by a lot

Discussion

- What's a correlation that you have heard of or seen that you think is misinterpreted because of selection bias?
- If you had all the money in the world and complete discretion, what RCT would you run to test it?

Demand for Learning HIV Status

- Rebecca Thornton implemented an RCT in rural Malawi for her job market paper at Harvard in mid-2000s
- At the time, it was an article of faith that you could fight the HIV epidemic in Africa by encouraging people to get tested; but Thornton wanted to see if this was true
- She randomly assigned cash incentives to people to incentivize learning their HIV status
- Also examined whether learning changed sexual behavior.

Experimental design

- Respondents were offered a free door-to-door HIV test
- Treatment is randomized vouchers worth between zero and three dollars
- These vouchers were redeemable once they visited a nearby voluntary counseling and testing center (VCT)
- Estimates her models using OLS with controls

Why Include Control Variables?

To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the SDO, we might estimate:

$$Y_i = \alpha + \delta D_i + \gamma X_i + \eta_i$$

Why Control Variables?

- There are 2 main reasons for including additional controls in the regression models:
  1. Conditional random assignment. Sometimes randomization is done *conditional* on some observable (e.g., gender, school, districts)
  2. Exogenous controls increase precision. Although control variables X_i are uncorrelated with D_i, they may have substantial explanatory power for Y_i. Including controls thus reduces variance in the residuals which lowers the standard errors of the regression estimates.
- Ongoing work by econometricians is investigating this more carefully

Table: Impact of Monetary Incentives and Distance on Learning HIV Results

	1	**2**	**3**	**4**	**5**
Any incentive	0.431***	0.309***	0.219***	0.220***	0.219 ***
	(0.023)	(0.026)	(0.029)	(0.029)	(0.029)
Amount of incentive		0.091***	0.274***	0.274***	0.273***
		(0.012)	(0.036)	(0.035)	(0.036)
Amount of incentive^2			-0.063***	-0.063***	-0.063***
			(0.011)	(0.011)	(0.011)
HIV	-0.055*	-0.052	-0.05	-0.058*	-0.055*
	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Distance (km)				-0.076***	
				(0.027)	
Distance^2				0.010**	
				(0.005)	
Controls	Yes	Yes	Yes	Yes	Yes
Sample size	2,812	2,812	2,812	2,812	2,812
Average attendance	0.69	0.69	0.69	0.69	0.69

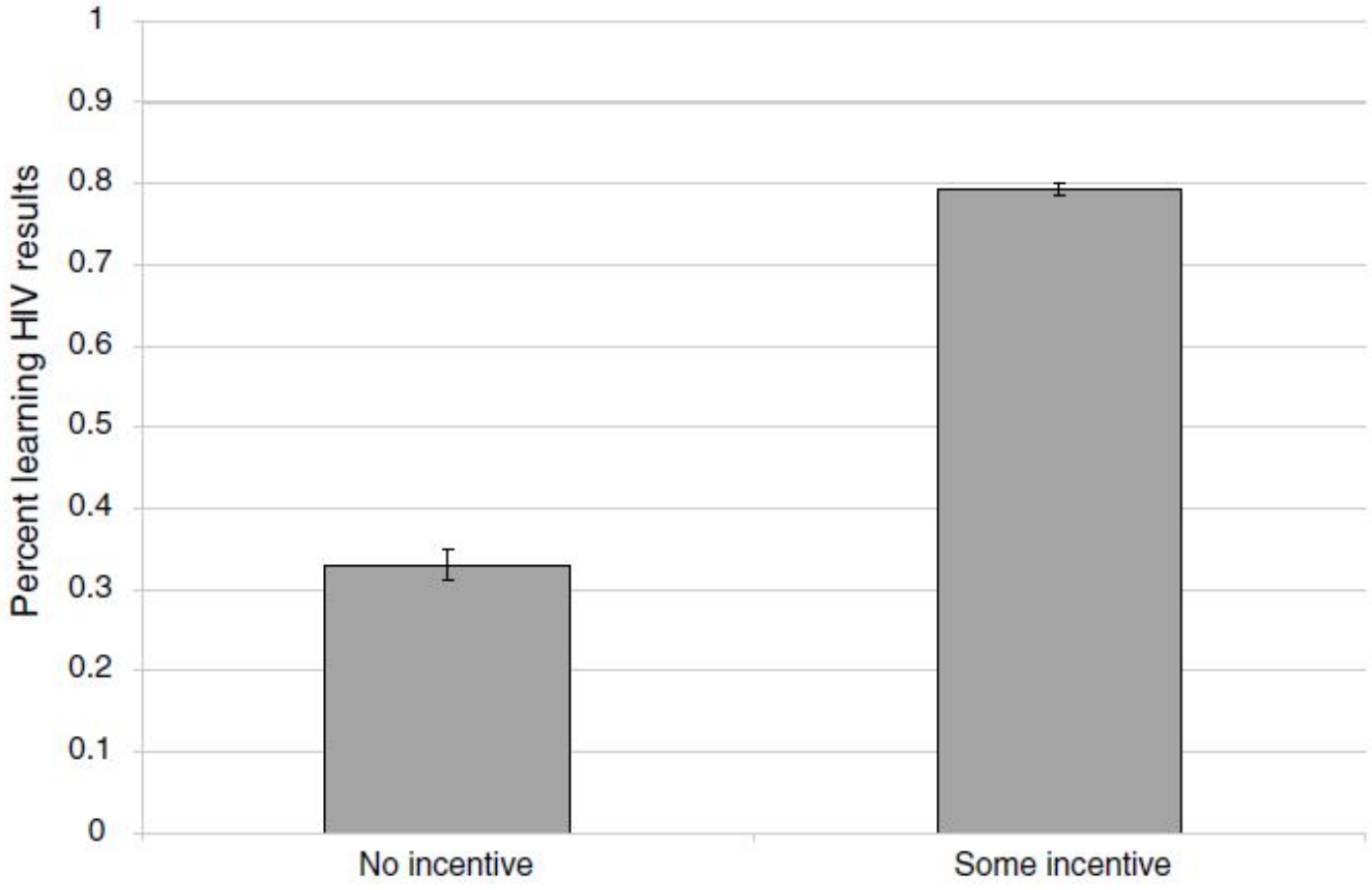


Figure: Visual representation of cash transfers on learning HIV test results.

Results

- Even small incentives were effective
- Any incentive increases learning HIV status by 43% compared to the control (mean 34%)
- Next she looks at the effect that learning HIV status has on risky sexual behavior

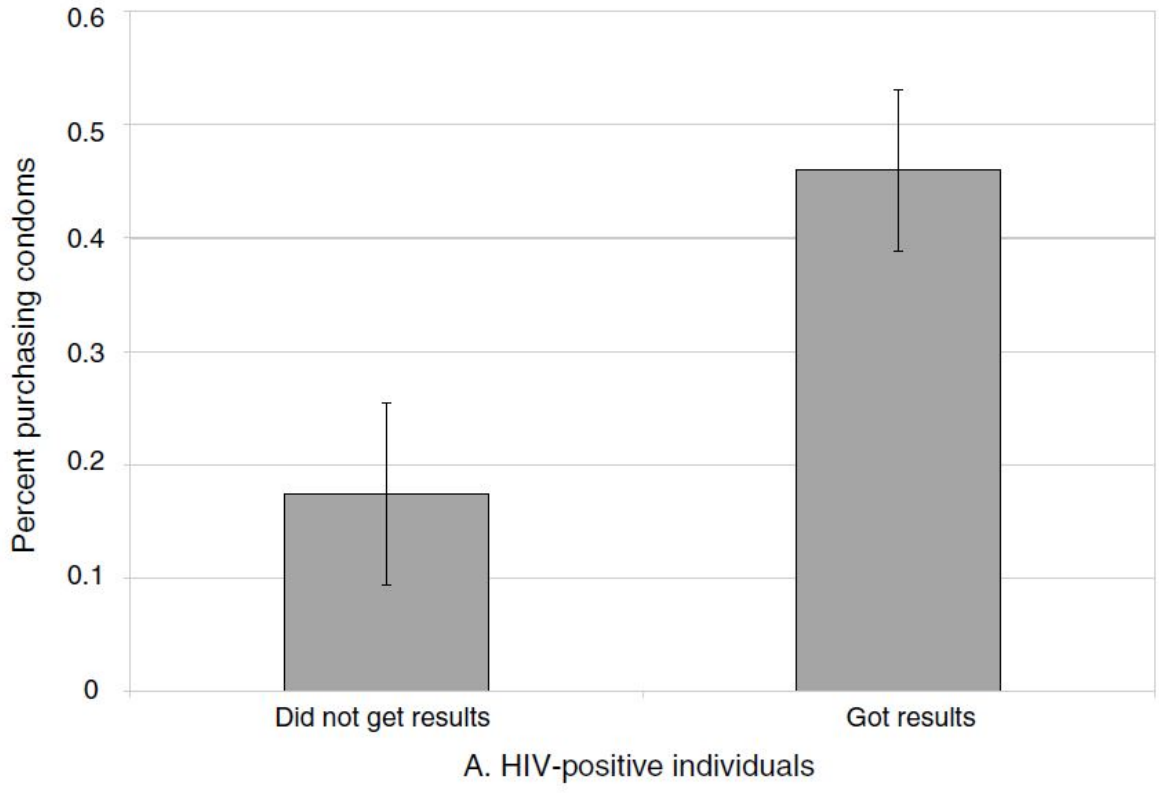


Figure: Visual representation of cash transfers on condom purchases for HIV positive individuals.

Table: Reactions to Learning HIV Results among Sexually Active at Baseline

Dependent variables:	**Bought condoms**		**Number of condoms bought**	
	OLS	**IV**	**OLS**	**IV**
Got results	−0.022	−0.069	−0.193	−0.303
	(0.025)	(0.062)	(0.148)	(0.285)
Got results × HIV	0.418***	0.248	1.778***	1.689**
	(0.143)	(0.169)	(0.564)	(0.784)
HIV	−0.175**	−0.073	−0.873	−0.831
	(0.085)	(0.123)	(0.275)	(0.375)
Controls	Yes	Yes	Yes	Yes
Sample size	1,008	1,008	1,008	1,008
Mean	0.26	0.26	0.95	0.95

Results

- For those who were HIV+ and got their test results, 42% more likely to buy condoms (but shrinks and becomes insignificant at conventional levels with IV).
- Number of condoms bought – very small. HIV+ respondents who learned their status bought 2 more condoms

Discussion

- What's in your field a causal question you find interesting that you wish you could answer?
- Describe the way you would conduct the RCT by explaining the following:
  - → What's the treatment? Express it as a binary variable.
  - → How will you assign this so that SUTVA holds and independence is achieved?
  - → What is the outcome you are interested in?
- Describe the steps you would take to do this if you had all the money in the world

Roadmap

- Introduction to course
  - History of causal inference
  - Design versus Model
- Potential outcomes
  - Naive correlation
  - Potential outcomes notation
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
  - Example of physical experimentation: HIV status
- Randomization inference
  - Lady tasting tea
  - Fisher's sharp null
  - Alternative test statistics

Randomization inference and causal inference

- “In randomization-based inference, uncertainty in estimates arises naturally from the random assignment of the treatments, rather than from hypothesized sampling from a large population.” (Athey and Imbens 2017)
- Athey and Imbens is part of growing trend of economists using randomization-based methods for doing causal inference
- Unclear (to me) why we are hearing more and more about randomization inference, but we are.
- Could be due to improved computational power and/or the availability of large data instead of samples?

Lady tasting tea experiment

- Ronald Aylmer Fisher (1890-1962)
  - → Two classic books on statistics: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), as well as a famous work in genetics, *The Genetical Theory of Natural Science*
  - → Developed many fundamental notions of modern statistics including the theory of randomized experimental design.

Lady tasting tea

- Muriel Bristol (1888-1950)
  - → A PhD scientist back in the days when women weren't PhD scientists
  - → Worked with Fisher at the Rothamsted Experiment Station (which she established) in 1919
  - → During afternoon tea, Muriel claimed she could tell from taste whether the milk was added to the cup before or after the tea
  - → Scientists were incredulous, but Fisher was inspired by her strong claim
  - → He devised a way to test her claim which she passed using randomization inference

Description of the tea-tasting experiment

- Original claim: Given a cup of tea with milk, Bristol claims she can discriminate the order in which the milk and tea were added to the cup
- Experiment: To test her claim, Fisher prepares 8 cups of tea – 4 **milk then tea** and 4 **tea then milk** – and presents each cup to Bristol for a taste test
- Question: How many cups must Bristol correctly identify to convince us of her unusual ability to identify the order in which the milk was poured?
- Fisher's sharp null: Assume she can't discriminate. Then what's the likelihood that random chance was responsible for her answers?

Choosing subsets

- The lady performs the experiment by selecting 4 cups, say, the ones she claims to have had the tea poured first.

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

- "8 choose 4" – $\binom{8}{4}$ – ways to choose 4 cups out of 8
  - → Numerator is $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order.
  - → Denominator is $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups.

Choosing subsets

- There are 70 ways to choose 4 cups out of 8, and therefore a 1.4% probability of producing the correct answer by chance

$$\frac{24}{1680} = 1/70 = 0.014.$$

- For example, the probability that she would correctly identify all 4 cups is $\frac{1}{70}$

Statistical significance

- Suppose the lady correctly identifies all 4 cups. Then . . .
  1. Either she has no ability, and has chosen the correct 4 cups purely by chance, or
  2. She has the discriminatory ability she claims.
- Since choosing correctly is highly unlikely in the first case (one chance in 70), the second seems plausible.
- Bristol actually got all four correct
- I wonder if seeing this, any of the scientists present changed their mind

Null hypothesis

- In this example, the null hypothesis is the hypothesis that the lady has no special ability to discriminate between the cups of tea.
- We can never prove the null hypothesis, but the data may provide evidence to reject it.
- In most situations, rejecting the null hypothesis is what we hope to do.

Null hypothesis of no effect

- Randomization inference allows us to make probability calculations revealing whether the treatment assignment was “unusual”
- Fisher’s sharp null is when entertain the possibility that no unit has a treatment effect
- This allows us to make “exact” p-values which do not depend on large sample approximations
- It also means the inference is not dependent on any particular distribution (e.g., Gaussian); sometimes called nonparametric

Sidebar: bootstrapping is different

- Sometimes people confuse randomization inference with bootstrapping
- Bootstrapping randomly draws a percent of the total observations for estimation; “uncertainty over the sample”
- Randomization inference randomly reassigns the treatment; “uncertainty over treatment assignment”

(Thanks to Jason Kerwin for helping frame the two against each other)

6-step guide to randomization inference

1. Choose a sharp null hypothesis (e.g., no treatment effects)
2. Calculate a test statistic (T is a scalar based on D and Y)
3. Then pick a randomized treatment vector $\tilde{D}_1$
4. Calculate the test statistic associated with $(\tilde{D}, Y)$
5. Repeat steps 3 and 4 for all possible combinations to get $\tilde{T} = \{\tilde{T}_1, \ldots, \tilde{T}_K\}$
6. Calculate exact p-value as $p = \frac{1}{K} \sum_{k=1}^{K} I(\tilde{T}_k \geq T)$

Pretend experiment

Table: Pretend DBT intervention for some homeless population

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

For concreteness, assume a program where we pay homeless people $15 to take dialectical behavioral therapy (DBT). Outcomes are some measure of mental health 0-20 with higher scores being improvements

Step 1: Sharp null of no effect

Fisher's Sharp Null Hypothesis

$$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \; \forall i$$

- Assuming no effect means any test statistic is due to chance
- Neyman and Fisher test statistics were different – Fisher was exact, Neyman was not
- Neyman's null was no average treatment effect (ATE=0). If you have a treatment effect of 5 and I have a treatment effect of -5, our ATE is zero. This is not the sharp null even though it also implies a zero ATE

More sharp null

- Since under the Fisher sharp null $\delta_i = 0$, it means each unit's potential outcomes under both states of the world are the same
- We therefore know each unit's missing counterfactual
- The randomization we will perform will cycle through all treatment assignments under a null well treatment assignment doesn't matter because all treatment assignments are associated with a null of zero unit treatment effects
- We are looking for evidence *against* the null

Step 1: Fisher's sharp null and missing potential outcomes

Table: Missing potential outcomes are no longer missing

Name	D	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Fisher sharp null allows us to **fill in** the missing counterfactuals bc

Step 2: Choosing a test statistic

Test Statistic

A test statistic $T(D, Y)$ is a scalar quantity calculated from the treatment assignments D and the observed outcomes Y

- By scalar, I just mean it's a number (vs. a function) measuring some relationship between D and Y
- Ultimately there are many tests to choose from; I'll review a few later
- If you want a test statistic with high statistical power, you need large values when the null is false, and small values when the null is true (i.e., *extreme*)

Simple difference in means

- Consider the absolute SDO from earlier

$$\delta_{SDO} = \left| \frac{1}{N_T} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_C} \sum_{i=1}^{N} (1 - D_i) Y_i \right|$$

- Larger values of δ_{SDO} are evidence *against* the sharp null
- Good estimator for constant, additive treatment effects and relatively few outliers in the potential outcomes

Step 2: Calculate test statistic, $T(D, Y)$

Table: Calculate T using D and Y

Name	D	Y	Y^0	Y^1	δ_i
Andy	**1**	**10**	10	10	0
Ben	**1**	**5**	5	5	0
Chad	**1**	**16**	16	16	0
Daniel	**1**	**3**	3	3	0
Edith	**0**	**5**	5	5	0
Frank	**0**	**7**	7	7	0
George	**0**	**8**	8	8	0
Hank	**0**	**10**	10	10	0

We'll start with this simple the simple difference in means test statistic, $T(D, Y)$: $\delta_{SDO} = 34/4 - 30/4 = 1$

Steps 3-5: Null randomization distribution

- Randomization steps reassign treatment assignment for every combination, calculating test statistics each time, to obtain the entire distribution of counterfactual test statistics
- The key insight of randomization inference is that under Fisher's sharp null, the treatment assignment shouldn't matter
- Ask yourself:
  - → if there is no unit level treatment effect, can you picture a distribution of counterfactual test statistics?
  - → and if there is no unit level treatment effect, what must average counterfactual test statistics equal?

Step 6: Calculate “exact” p-values

- Question: how often would we get a test statistic as big or bigger as our “real” one if Fisher’s sharp null was true?
- This can be calculated “easily” (sometimes) once we have the randomization distribution from steps 3-5
  - → The number of test statistics ($t(D, Y)$) bigger than the observed divided by total number of randomizations

$$Pr(T(D, Y) \geq T(\tilde{D}, Y|\delta = 0)) = \frac{\sum_{D \in \Omega} I(T(D, Y) \leq T(\tilde{D}, Y)}{K}$$

First permutation (holding N_T fixed)

Name	$\tilde{D}_2$	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

$$\tilde{T}_1 = |36/4 - 28/4| = 9 - 7 = 2$$

Second permutation (again holding N_T fixed)

Name	$\tilde{D}_3$	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	1	8	8	8
Hank	0	10	10	10

$$T_{rank} = |36/4 - 27/4| = 9 - 6.75 = 2.25$$

Sidebar: Should it be 4 treatment groups each time?

- In this experiment, I've been using the same N_T *under the assumption* that N_T had been fixed when the experiment was drawn.
- But if the original treatment assignment had been generated by something like a Bernoulli distribution (e.g., coin flips over every unit), then you should be doing a complete permutation that is also random in this way
- This means that for 8 units, sometimes you'd have 1 treated, or even 8
- Correct inference requires you know the original data generating process

Randomization distribution

Assignment	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	$\lvert T_i \rvert$
True D	1	1	1	1	0	0	0	0	1
$\tilde{D}_2$	1	0	1	1	0	1	0	0	2
$\tilde{D}_3$	1	0	1	1	0	0	1	0	2.25
...									

Step 2: Other test statistics

- The simple difference in means is fine when effects are additive, and there are few outliers in the data
- But outliers create more variation in the randomization distribution
- A good test statistic is the one that best fits your data.
- Some test statistics will have weird properties in the randomization as we'll see in synthetic control.
- What are some alternative test statistics?

Transformations

- What if there was a constant multiplicative effect: $Y_i^1 / Y_i^0 = C$?
- Difference in means will have low power to detect this alternative hypothesis
- So we transform the observed outcome using the natural log:

$$T_{log} = \left| \frac{1}{N_T} \sum_{i=1}^{N} D_i ln(Y_i) - \frac{1}{N_C} \sum_{i=1}^{N} (1 - D_i) ln(Y_i) \right|$$

- This is useful for skewed distributions of outcomes

Difference in medians/quantiles

- We can protect against outliers using other test statistics such as the difference in quantiles
- Difference in medians:

$$T_{median} = |median(Y_T) - median(Y_C)|$$

- We could also estimate the difference in quantiles at any point in the distribution (e.g., 25th or 75th quantile)

Rank test statistics

- Basic idea is rank the outcomes (higher values of Y_i are assigned higher ranks)
- Then calculate a test statistic based on the transformed ranked outcome (e.g., mean rank)
- Useful with continuous outcomes, small datasets and/or many outliers

Rank statistics formally

- Rank is the domination of others (including oneself):

$$\tilde{R} = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} I(Y_j \leq Y_i)$$

- Normalize the ranks to have mean 0

$$\tilde{R}_i = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} I(Y_j \leq Y_i) - \frac{N+1}{2}$$

- Calculate the absolute difference in average ranks:

$$T_{rank} = |\overline{R}_T - \overline{R}_C| = \left| \frac{\sum_{i:D_i=1} R_i}{N_T} - \frac{\sum_{i:D_i=0} R_i}{N_C} \right|$$

- Minor adjustment (averages) for ties

Randomization distribution

Name	D	Y	Y^0	Y^1	Rank	R_i
Andy	1	10	**10**	10	6.5	2
Ben	1	5	**5**	5	2.5	-2
Chad	1	16	**16**	16	8	3.5
Daniel	1	3	**3**	3	1	-3.5
Edith	0	5	5	**5**	2.5	-2
Frank	0	7	7	**7**	4	-0.5
George	0	8	8	**8**	5	0.5
Hank	0	10	10	**10**	6.5	2

$$T_{rank} = |0 - 0| = 0$$

Effects on outcome distributions

- Focused so far on “average” differences between groups.
- Kolmogorov-Smirnov test statistics is based on the difference in the distribution of outcomes
- Empirical cumulative distribution function (eCDF):

$$\widehat{F}_C(Y) = \frac{1}{N_C} \sum_{i:D_i=0} 1(Y_i \leq Y)$$

$$\widehat{F}_T(Y) = \frac{1}{N_T} \sum_{i:D_i=1} 1(Y_i \leq Y)$$

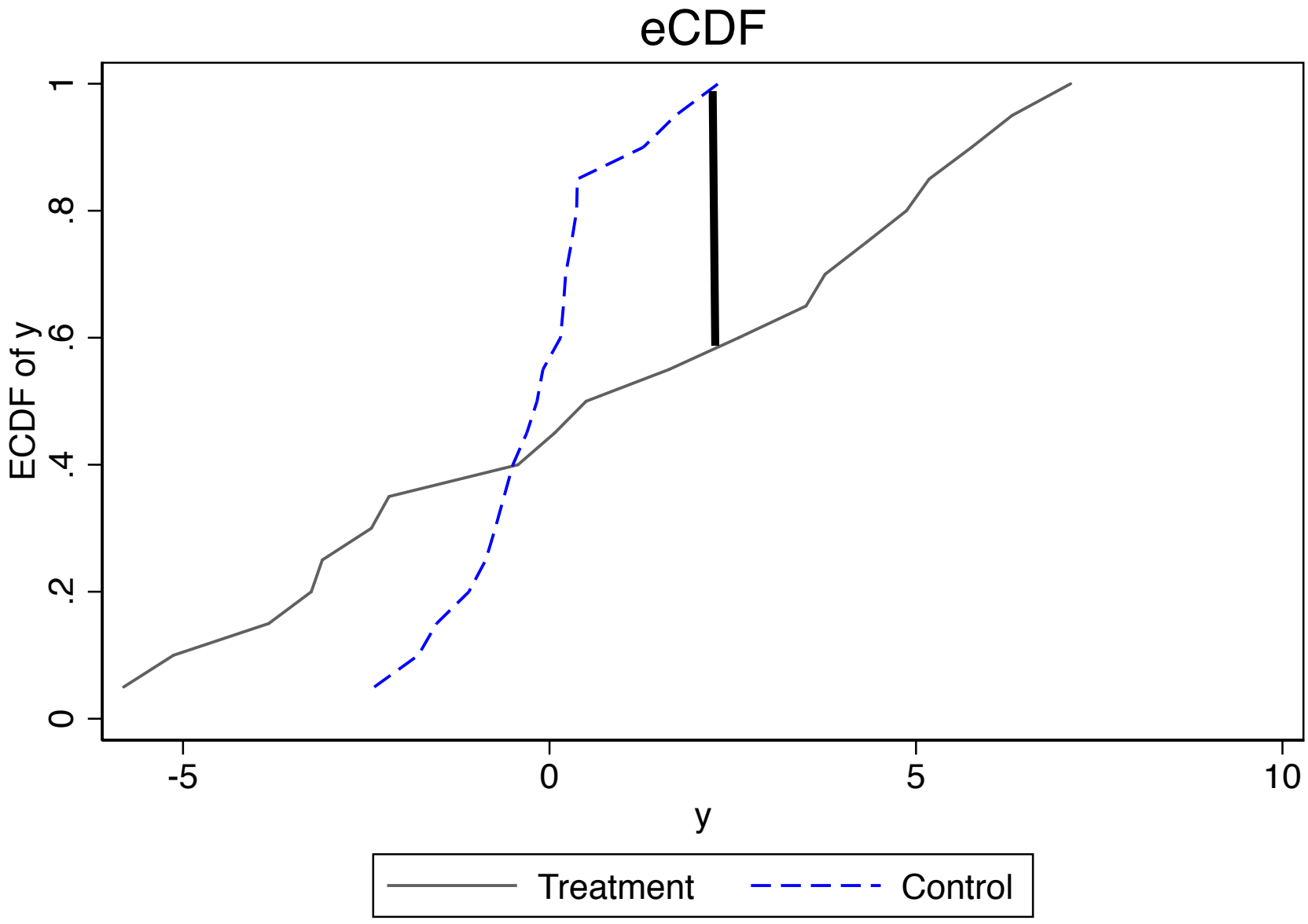
- Proportion of observed outcomes below a chosen value for treated and control separately
- If two distributions are the same, then $\widehat{F}_C(Y) = \widehat{F}_T(Y)$

Kolmogorov-Smirnov statistic

- Test statistics are scalars not functions
- eCDFs are functions, not scalars
- Solution: use the maximum discrepancy between the two eCDFs:

$$T_{KS} = max|\widehat{F}_T(Y_i) - \widehat{F}_C(Y_i)|$$

eCDFs by treatment status and test statistic



Small vs. Modest Sample Sizes are non-trivial

Computing the exact randomization distribution is not always feasible (Wolfram Alpha)

- $N = 6$ and $N_T = 3$ gives us 20 assignment vectors
- $N = 8$ and $N_T = 4$ gives us 70 assignment vectors
- $N = 10$ and $N_T = 5$ gives us 252 assignment vectors
- $N = 20$ and $N_T = 10$ gives us 184,756 assignment vectors
- $N = 50$ and $N_T = 25$ gives us 1.2641061×10^{14} assignment vectors

Exact p calculations are not realistic bc the number of assignments explodes at even modest size

Approximate p-values

These have been “exact” tests when they use every possible combination of D

- When you can’t use every combination, then you can get *approximate* p-values from a simulation (TBD)
- With a rejection threshold of α (e.g., 0.05), randomization inference test will falsely reject less than $100 \times \alpha\%$ of the time

Approximate p values

- Use simulation to get approximate p-values
  - → Take K samples from the treatment assignment space
  - → Calculate the randomization distribution in the K samples
  - → Tests no longer exact, but bias is under your control (increase K)
- Imbens and Rubin show that p values converge to stable p values pretty quickly (in their example after 1000 replications)

Thornton's experiment

ATE	Iteration	Rank	p	no. trials
0.45	1	1	0.01	100
0.45	1	1	0.002	500
0.45	1	1	0.001	1000

Table: Estimated p-value using different number of trials.

Including covariate information

- Let X_i be a pretreatment measure of the outcome
- One way is to use this as a gain score: $Y^{d'} = Y_i^d - X_i$
- Causal effects are the same $Y^{1i} - Y^{0i} = Y_i^1 - Y_i^0$
- But the test statistic is different:

$$T_{gain} = \left| (\overline{Y}_T - \overline{Y}_C) - (\overline{X}_T - \overline{X}_C) \right|$$

- If X_i is strongly predictive of Y_i^0, then this could have higher power
  - $\rightarrow$ Y_{gain} will have lower variance under the null
  - $\rightarrow$ This makes it easier to detect smaller effects

Regression in RI

- We can extend this to use covariates in more complicated ways
- For instance, we can use an OLS regression:

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon$$

- Then our test statistic could be $T_{OLS} = \widehat{\delta}$
- RI is justified even if the model is wrong
  - → OLS is just another way to generate a test statistic
  - → The more the model is "right" (read: predictive of Y_i^0), the higher the power T_{OLS} will have
- See if you can do this in Thornton's dataset using the loops and saving the OLS coefficient (or just use `ritest`)

Concluding remarks

- Randomization inference is very common, particularly useful you don’t want to make strong assumptions (parametric free)
- We’ll now explore its use in a popular observational method – the synthetic control