ECON 626: Applied Microeconomics

Lecture 1:

Selection Bias and the Experimental Ideal

Professors: Pamela Jakiela and Owen Ozier

Department of Economics
University of Maryland, College Park

Potential Outcomes

Do Hospitals Make People Healthier?

Your health status is: excellent, very good, good, fair, or poor?

	Hospital	No Hospital	Difference
Health status	3.21	3.93	-0.72^{***}
	(0.014)	(0.003)	
Observations	7,774	90,049	

A simple comparison of means suggests that going to the hospital makes people worse off: those who had a hospital stay in the last 6 months are, on average, less healthy than those that were not admitted to the hospital

- What's wrong with this picture?

Potential Outcomes

We are interested in the relationship between "**treatment**" and some outcome that may be impacted by the treatment (eg. health)

Outcome of interest:

- Y = outcome we are interested in studying (e.g. health)
- Y_i = value of outcome of interest *for individual i*

For each individual, there are two **potential outcomes**:

- $Y_{0,i}$ = i's outcome if she **doesn't** receive treatment
- $Y_{1,i}$ = i's outcome if she **does** receive treatment

Potential Outcomes

For any individual, we can only observe one potential outcome:

$$Y_i = \begin{cases} Y_{0i} & if D_i = 0 \\ Y_{1i} & if D_i = 1 \end{cases}$$

where D_i is a treatment indicator (equal to 1 if i was treated)

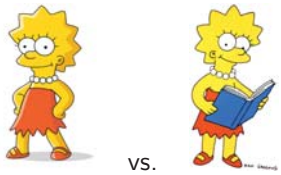
- Each individual either participates in the program or not
- The causal impact of program (D) on i is: $Y_{1i} - Y_{0i}$

We observe i's actual outcome:

$$Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{impact} D_i$$

Establishing Causality

In an ideal world (research-wise), we could clone each treated individual and observe the impacts of the treatment on their lives



What is the impact of giving Lisa a book on her test score?

- Impact = Lisa's score with a book - Lisa's score without a book

In the real world, we either observe Lisa with a book or without

- We never observe the **counterfactual**

Establishing Causality

To measure the causal impact of giving Lisa a book on her test score, we need to find a comparison group that did not receive a book


vs.


Our estimate of the impact of the book is then the difference in test scores between the **treatment group** and the **comparison group**

- Impact = Lisa's score with a book - Bart's score without a book

As this example illustrates, finding a good comparison group is hard

Selection Bias

When we compare means for participants and non-participants:

$$\begin{aligned}\textbf{Difference in group means} &= E[Y_i|D_i=1] - E[Y_i|D_i=0] \\ &= E[Y_{1,i}|D_i=1] - E[Y_{0,i}|D_i=0]\end{aligned}$$

Adding in $\underbrace{-E[Y_{0,i}|D_i=1] + E[Y_{0,i}|D_i=1]}_{=0}$, we get:

Difference in group means

$$= \underbrace{E[Y_{1,i}|D_i=1] - E[Y_{0,i}|D_i=1]}_{\text{average causal effect on participants}} + \underbrace{E[Y_{0,i}|D_i=1] - E[Y_{0,i}|D_i=0]}_{\text{selection bias}}$$

How Do We Estimate Causal Impacts?

Quasi-experimental approaches:

- Conditional Independence Assumption (CIA) approaches
  - "$\hat{\theta}_{hfb}$" – Associate Professor Bryan Graham, UC Berkeley
- Difference-in-difference estimation
  - Requirement: common trends in treatment, comparison groups
- Instrumental variables
  - Requirement: a valid instrument (satisfying the exclusion restriction)
- Regression discontinuity
  - Requirement: the existence of discontinuity

How Do We Estimate Causal Impacts?

Experimental approach:

- **Random assignment to treatment:** eligibility for program is literally determined at random, e.g. via pulling names out of hat

The **law of large numbers** tells us that a sample average can be brought as close as we like to the population average just by enlarging the sample

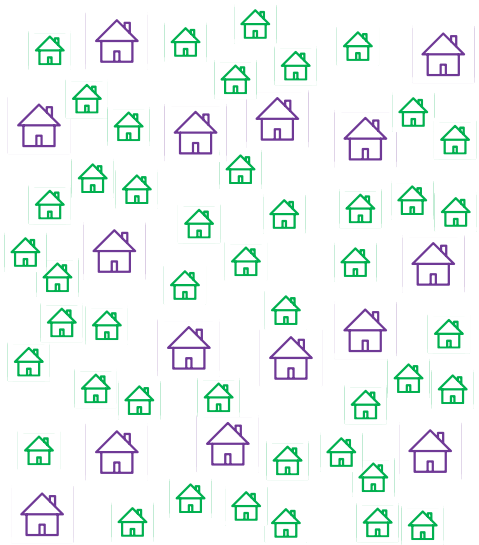
When treatment is randomly assigned,
the treatment, control groups are random samples of a single population (e.g. the population of all eligible applicants for the program)

$$\Rightarrow E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected outcomes are the same in the absence of the program

Random Assignment & the Law of Large Numbers

Population of eligible households



25% purple households

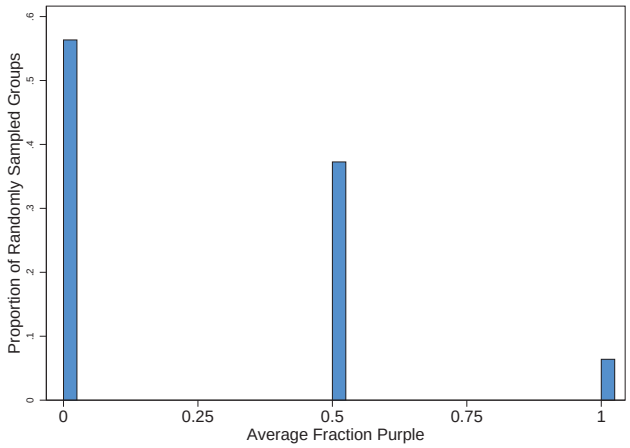
If you chose one at random, probability it is purple: 0.25

However, any one house (chosen at random) is either purple or green.

What if you chose 2 HHs?

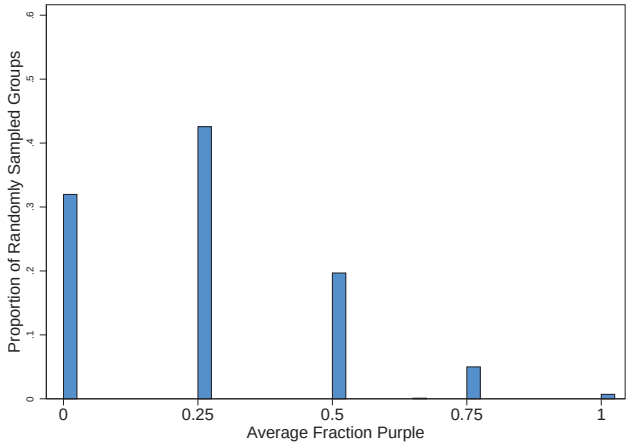
Random Assignment & the Law of Large Numbers

When you randomly sample groups of 2:



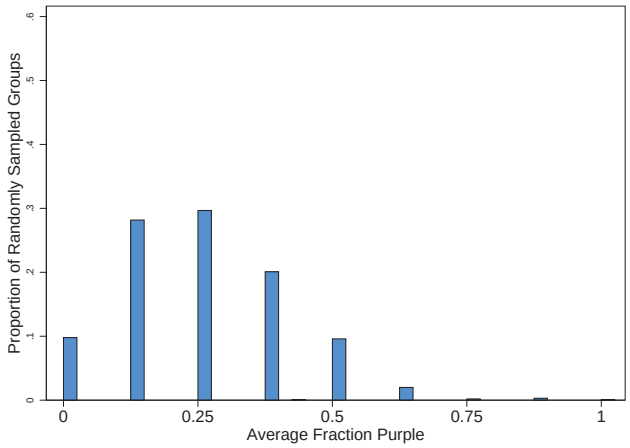
Random Assignment & the Law of Large Numbers

When you randomly sample groups of 4:



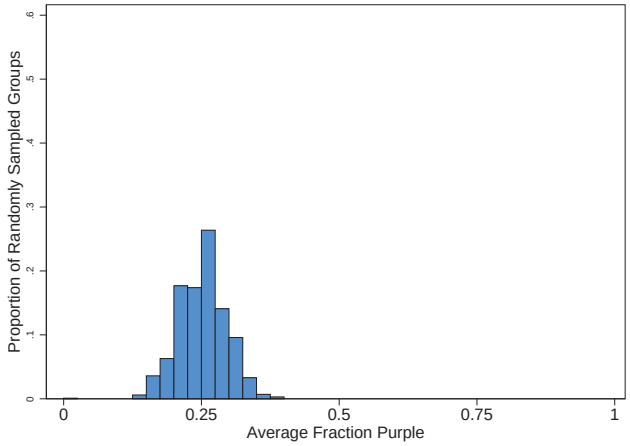
Random Assignment & the Law of Large Numbers

When you randomly sample groups of 8:



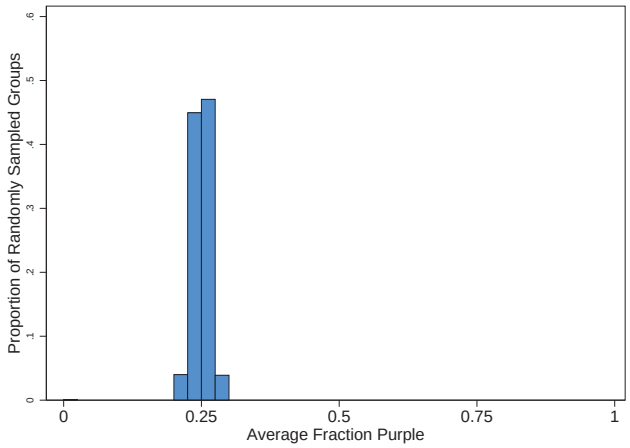
Random Assignment & the Law of Large Numbers

When you randomly sample groups of 100:



Random Assignment & the Law of Large Numbers

When you randomly sample groups of 1000:



Random Assignment Eliminates Selection Bias

If treatment is random and $E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$

- In practice, we should be wary of small samples

The difference in means estimator gives us the average treatment effect:

Difference in group means

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1] + E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{average treatment effect on participants}} + \underbrace{E[Y_{0,i}] - E[Y_{0,i}]}_{=0}$$

$$= \underbrace{E[Y_{1,i}] - E[Y_{0,i}]}_{\text{ATE}}$$

Internal Validity

Excellent news: random assignment eliminates selection bias*

*Some restrictions apply

The **Stable Unit Treatment Value Assumption (SUTVA)**:

- "The potential outcomes for any unit do not vary with the treatments assigned to other units."

  Source: Imbens and Rubin (2015)

When is SUTVA likely to be violated?

SUTVA Violations

Causal effects in the presence of spillovers:

- What is the appropriate unit of randomization?
  - Cluster-randomized trials make sense when spillovers are anticipated
- When can we use additional assumptions to measure the direct and indirect effects of treatment (e.g. via multi-level randomization)?
- When can we anticipate the direction of bias?

Internal Validity: Additional Assumptions?

Imbens and Rubin include a second component of SUTVA:

- "There are no different forms or versions of each treatment level which lead to different potential outcomes."

This terminology is not standard, and the assumption is often violated

- Treatments often vary across locations or strata
- Cox (1958) proposes an alternative: "either only average treatment effects are required, or that the treatment effects are constant"
  - In other words, we'll always have internal validity
  - External validity is another matter

Gerber and Green (2012) highlight an add'l assumption, **excludability**: the treatment shouldn't be confounded (well, duh, right?)

Randomization: A History of Thought

Randomization: A Timeline

1885 Pierce and Jastrow use randomization in a psychology experiment (varying order in which different stimuli are presented to subjects)

1898 Johannes Fibiger conducts a trial of an anti-diphtheria serum in which every other subject is assigned to treatment (or control)

1923 Neyman suggests the idea of potential outcomes

1925 Fisher suggests the explicit randomization of treatments (in the context of agriculture experiments)

1926 Amberson *et al* study of sanocrysin treatments for TB: 24 patients divided into two comparable groups; coin flipped to determine which group of 12 receives treatment and which group serves as controls

1942 Launch of Cambridge-Somerville Youth Study of at-risk boys

1948 Randomized trial of streptomycin treatment for TB conducted by the Medical Research Council of Great Britain

1962 Perry preschool experiment in Ypsilanti, MI

1974 Rubin introduces the concept of potential outcomes (as we know it)

The Lady Tasting Tea

Chapter II of Fisher's *The Design of Experiments* begins:

> *"A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup."*

Critical lesson to take away from this anecdote:
Caffeine breaks with colleagues are critical to the advancement of science

- The lady in question was biologist Muriel Bristol, who worked with Fisher at the Rothamsted Experimental Station in Harpenden, UK
- H_0: Fisher believes that Dr. Bristol cannot taste the difference
- A test of the hypothesis: *"Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order."*

The Lady Tasting Tea: Experimental Design

Rule #1: do not confound your own treatment

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, then she will choose 4 cups at random
- Fisher points out that the experimenter could screw this up:

> *"If all those cups made with the milk first had sugar added, while those made with the tea first had none, a very obvious difference in flavour would have been introduced which might well ensure that all those made with sugar should be classed alike."*

- Gerber and Green refer to this as **excludability**

The Lady Tasting Tea: Experimental Design

Rule #1B: do not accidentally confound your own treatment

- Fisher, in (perhaps) the earliest known scientific subtweet:

> *"It is not sufficient remedy to insist that 'all the cups must be exactly alike' in every respect except that to be tested. For this is a totally impossible requirement."*

- To minimize the likelihood of accidentally confounding your treatment, the best approach is to constrain yourself by randomizing
  - Randomization minimizes the likelihood of unfortunate coincidences
  - This was a highly controversial position at the time, and it is still debated in some circles; the alternative is to force balance (on observables, and then just hope that unobservables don't matter)

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

Suppose Dr. Bristol correctly identified all 4 "treated" cups

- How likely is it that this outcome could have occurred by chance?
  - There are $\binom{8}{4} = 70$ possible ways to choose 4 of 8 cups
  - Only one is correct; a subject with no ability to discriminate between treated and untreated cups would have a $1/70$ chance of success
  - The p-value associated with this outcome is $1/70 \approx 0.014$, which is less than the cutoff for the "standard level of significance" of 0.05

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

Suppose Dr. Bristol correctly identified 3 "treated" cups

- How likely is it that this outcome could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 of 8 cups
    - There are 17 ways to choose **at least** 3 correct cups
  - The p-value associated with this outcome is $17/70 \approx 0.243$
  - We should not reject the null hypothesis

The only experimental result that would lead to the rejection of the null hypothesis was correct identification of all 4 treated cups

- In the actual experiment, the null hypothesis was rejected

Fisher's Exact Test

	Identified by Dr. Bristol?	
	No	**Yes**
Milk poured first	a	b
Tea poured first	c	d

Is Dr. Bristol more likely to select cups where the milk was poured first?

$$\text{probability} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

The p-value is the sum of the probabilities of outcomes that are at least as extreme (i.e. contrary to H_0) as the observed outcome

The Lady Tasting Tea: Size and Power

The size of a test is the likelihood of rejecting a true null

- Fisher asserts that tests of size 0.05 are typical

Alternative experiment: what if we had treated 3 out of 6 cups of tea?

- There are $\binom{6}{3} = 20$ possible ways to choose 3 of 6 cups
- Best possible p-value is therefore 0.05

Alternative experiment: what if we had treated 3 out of 8 cups of tea?

- There are $\binom{8}{3} = 56$ possible ways to choose 3 of 8 cups
- Best possible p-value is therefore 0.017

$\Rightarrow$ **Optimal to have equal numbers of treated, untreated cups**

The Lady Tasting Tea: Size and Power

An alternate experiment: an unknown number of treated cups

- Under the null, the probability of getting 8 right is 1 in 2^8
- Probability of getting 7 right is $8/256 = 0.03125$

This design would achieve higher power with the same number of trials

- Possible to reject the hypothesis that the lady tasting tea cannot tell the difference even when her ability to discriminate is imperfect

Ronald Fisher's Contributions to Statistics

1. Introduced the modern randomized trial
2. Introduced the idea of permutation tests
3. Reminded us of the importance of caffeine

Fisher's permutation-based approach to inference is not the norm in economics; our default is regression analysis and classical statistics

The Analysis of Experiments

Regression Analysis of Experiments

Suppose we estimate an OLS regression of the form:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

When D_i is a dummy variable,

$$E[\hat{\beta}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

When the true model is one of constant (i.e. homogeneous) effects,

$$Y_{1,i} = \delta + Y_{0,i}$$

It is clear that $E[\hat{\beta}] = \delta$ when treatment is randomly assigned

Constant Treatment Effects? Really?

Consider the hospitalization example?

- Is it reasonable to assume that treatment effects are homogeneous?
- **No.** Clearly, people go to the hospital when they are sick

A more interesting thought experiment:

- $z = i$'s health if she doesn't get sick
- $s =$ the reduction in health associated with sickness
- $b =$ benefit a **sick** person receives from treatment
- $c =$ the reduction in health from going to the hospital

Reasonable to assume that $b > c > 0$

Potential Outcomes: Hospital Example

	$Y_{0,i}$	$Y_{1,i}$
Sick	$z - s$	$z - s + b - c$
Not sick	z	$z - c$

What happens without random assignment?

- Do healthy people go to the hospital?
- Do sick people go to the hospital?

Life without Random Assignment

Let S_i be an indicator for being sick

- $E[S_i|D_i = 1] = ?$
- $E[S_i|D_i = 0] = ?$

What do we learn from a comparison of means?

$$\begin{aligned}\textbf{difference in means} &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\ &= z - s + b - c - z \\ &= b - c - s\end{aligned}$$

Difference in means is the treatment effect on those who choose to take up the treatment (i.e. on the sick) plus selection bias

Random Assignment: Entire Population

Suppose, absurdly, we randomize who goes to the hospital such that:

$$\lambda = E[S_i|D_i = 1] = E[S_i|D_i = 0] = E[S_i]$$

Randomization breaks the link between illness and going to the hospital

What does the difference in means tell us?

$$\begin{aligned}
\textbf{difference in means} &= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\
&= \underbrace{z - E[S_i|D_i = 1]\,(s - b) - c}_{=E[Y_{1,i}|D_i=1]} - \underbrace{\{z - E[S_i|D_i = 0]s\}}_{=E[Y_{0,i}|D_i=0]} \\
&= z - \lambda s + \lambda b - c - (z - \lambda s) \\
&= \lambda b - c
\end{aligned}$$

Difference in means = ATE of hospitalization on the population

Random Assignment: Sick People

Suppose we randomize treatment assignment among the sick:

$$E[S_i|D_i = 1] = E[S_i|D_i = 0] = 1$$

What does the difference in means tell us?

$$\begin{aligned}
\textbf{difference in means} &= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\
&= \underbrace{z - s + b - c}_{=E[Y_{1,i}|D_i=1]} - \underbrace{\{z - s\}}_{=E[Y_{0,i}|D_i=0]} \\
&= b - c
\end{aligned}$$

Difference in means = ATE of hospitalization on the sick

Is this the ideal experiment? Why or why not?

Random Assignment: Endogenous Take-Up

We might consider randomizing **access** to treatment:

- Let T_i be an indicator for random assignment to a treatment group that is allowed to choose whether or not to go to the hospital
- Those in the control group cannot use the hospital

Q: Who will choose to go to the hospital?

- **A:** People who get sick during the study
- $E[D_i|T_i = 1] = ?$

When take-up is endogenous, we (usually) have **imperfect compliance**

- With one-sided non-compliance: compliers vs. never-takers

Random Assignment: Endogenous Take-Up

What does the difference in means tell us in this case?

$$\begin{aligned}
\textbf{difference in means} &= E[Y_{1,i}|T_i = 1] - E[Y_{0,i}|T_i = 0] \\
&= \underbrace{z + E[S_i|T_i = 1]\,(-s + b - c)}_{=E[Y_{1,i}|T_i=1]} - \underbrace{\{z - E[S_i|T_i = 0]s\}}_{=E[Y_{0,i}|T_i=0]} \\
&= z - \lambda s + \lambda b - \lambda c - (z - \lambda s) \\
&= \lambda\,(b - c)
\end{aligned}$$

Difference in means = ATE of access to hospitalization

- The ATE is the **intent-to-treat** effect
- ITT = compliance $\times$ effect of **treatment on the treated**

External Validity

Three randomized evaluations, three average treatment effects

- How much can we learn from a single study?
- How much can we learn without a model?

A more realistic evaluation scenario would have considered:

- A broader range of heterogeneous treatment effects
- Two-sided non-compliance
  - Encouragement designs may increase take-up among the healthy

None of these problems is specific to randomized evaluations

External Validity

In many early randomized evaluations, the ATE of interest was clear

- The impact of new seed varieties on crop yields
- The impact of medical treatments on patients with specific ailments

Economists consider a very broad range of "treatments"

- The impact of access to credit
- The impact of having two children of the same gender
- The impact of going on the Hajj
- The impact of sunshine on the 4^{th} of July during childhood

A good research idea requires (1) identification and (2) a model