Session I
Survey Experiments in Context

Thomas J. Leeper

Government Department
London School of Economics and Political Science

Activity!

Activity!

1. Ask you to guess a number

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room
3. Group 2, close your eyes

Activity!

Group 1

Think about whether the population of Chicago is more or less than 500,000 people. What do you think the population of Chicago is?

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room
3. Group 2, close your eyes
4. Group 1, close your eyes

Activity!

Group 2

Think about whether the population of Chicago is more or less than 10,000,000 people. What do you think the population of Chicago is?

Enter your data

- Go here: `http://bit.ly/297vEdd`
- Enter your guess and your group number

Results

- True population: 2.79 million

Results

- True population: 2.79 million
- What did you guess? (See Responses)

Results

- True population: 2.79 million
- What did you guess? (See Responses)
- What's going on here?
  - An experiment!
  - Demonstrates "anchoring" heuristic

Results

- True population: 2.79 million
- What did you guess? (See Responses)
- What's going on here?
  - An experiment!
  - Demonstrates "anchoring" heuristic
- Experiments are easy to analyze, but only if designed and implemented well

1 Introductions

2 Course Outline

3 History of Experiments

4 Logic and Analysis

Who am I?

- Thomas Leeper
- Assistant Professor in Political Behaviour at London School of Economics
  - 2013–15: Aarhus University (Denmark)
  - 2008–12: PhD from Northwestern University (Chicago, USA)
  - Birth–2008: Minnesota, USA
- Interested in public opinion and political psychology
- Email: t.leeper@lse.ac.uk

Who are you?

- Introduce yourself to a neighbour
- Where are you from?
- What do you hope to learn from the course?

Quick Survey

Quick Survey

1. How many of you have worked with survey data before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?
3. How many of you have worked with experimental data before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?
3. How many of you have worked with experimental data before?
4. Of those, how many of you have *performed* an experiment before?

Course Materials

All material for the course is available at:

http://www.thomasleeper.com/surveyexpcourse/

Learning Outcomes

By the end of the week, you should be able to...

Learning Outcomes

By the end of the week, you should be able to...

1. Explain how to analyze experiments quantitatively.

Learning Outcomes

By the end of the week, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.

Learning Outcomes

By the end of the week, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.

Learning Outcomes

By the end of the week, you should be able to. . .

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.
4. Identify practical issues that arise in the implementation of experiments and evaluate how to anticipate and respond to them.

Schedule of Four Sessions

1. Survey Experiments in Context
2. Examples and Paradigms
3. External Validity
4. Practical Issues

Questions?

1 Introductions

2 Course Outline

3 History of Experiments

4 Logic and Analysis

Experiments

Oxford English Dictionary defines "experiment" as:

1. A scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact
2. A course of action tentatively adopted without being sure of the outcome

Experiments

- "Experiments" have a very long history
- Major advances in design and analysis of experiments based on agricultural and later biostatistical research in the 19th century
  - R.A. Fisher
  - Jerzy Neyman
  - Karl Pearson
  - Oscar Kempthorne

In Social Sciences

- "Experiments" emerged in psychology 19th century
  - Not randomized – more like "What if?" studies
  - Heavily laboratory-based or clinical

In Social Sciences

- "Experiments" emerged in psychology 19th century
  - Not randomized – more like "What if?" studies
  - Heavily laboratory-based or clinical
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884

In Social Sciences

- “Experiments” emerged in psychology 19th century
  - Not randomized – more like “What if?” studies
  - Heavily laboratory-based or clinical
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884
- RCTs came later to medicine (circa 1950)

In Social Sciences

- "Experiments" emerged in psychology 19th century
  - Not randomized – more like "What if?" studies
  - Heavily laboratory-based or clinical
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884
- RCTs came later to medicine (circa 1950)
- And have been a major part of the "credibility revolution" in economics
  - See, especially, LaLonde (1986)

In Social Science I

- APSA Pres. A. Lawrence Lowell (1922):
  "We are limited by the impossibility of experiment. Politics is an observational, not an experimental science. . . "

In Social Science I

- APSA Pres. A. Lawrence Lowell (1922):
  "We are limited by the impossibility of experiment. Politics is an observational, not an experimental science..."
- First experiment by Gosnell (1924)

In Social Science I

- APSA Pres. A. Lawrence Lowell (1922): *"We are limited by the impossibility of experiment. Politics is an observational, not an experimental science. . . "*
- First experiment by Gosnell (1924)
- Gerber and Green (2000) first major *field* experiment

In Social Science II

- Rise of surveys in the behavioral revolution
- Survey research was not experimental because interviewing was still mostly paper-based
    - "Split Ballots" (Schuman & Presser; Bishop)

In Social Science II

- Rise of surveys in the behavioral revolution
- Survey research was not experimental because interviewing was still mostly paper-based
    - "Split Ballots" (Schuman & Presser; Bishop)
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop CATI

In Social Science II

- Rise of surveys in the behavioral revolution
- Survey research was not experimental because interviewing was still mostly paper-based
  - "Split Ballots" (Schuman & Presser; Bishop)
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop CATI
- Mid-1980s: Paul Sniderman & Tom Piazza performed the first survey experiment[1]
  - Then: the "first multi-investigator"
  - Later: Skip Lupia and Diana Mutz created TESS

[1] Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: Harvard University Press.

TESS

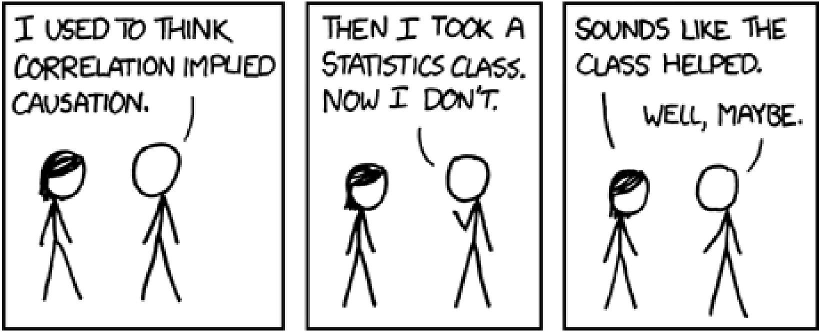
- Time-Sharing Experiments for the Social Sciences
- Multi-disciplinary initiative that provides infrastructure for survey experiments on nationally representative samples of the United States population
- Funded by the U.S. National Science Foundation
- Anyone anywhere in the world can apply

TESS-like Projects

There are some TESS-like initiatives outside the United States:

- Netherlands: LISS
- Norway: Bergen's Citizen Panel
- Sweden: Gothenburg's Citizen Panel

Questions?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Addressing Confounding

In observational research...

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)

Addressing Confounding

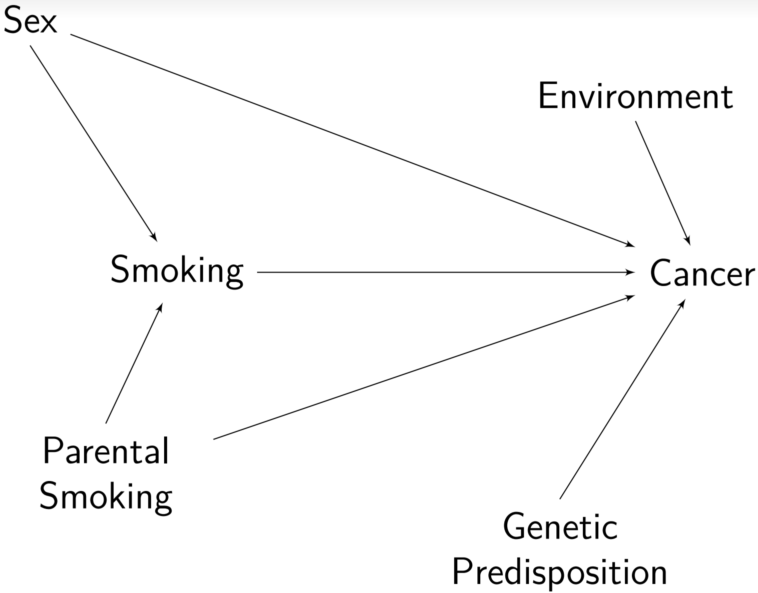
In observational research. . .

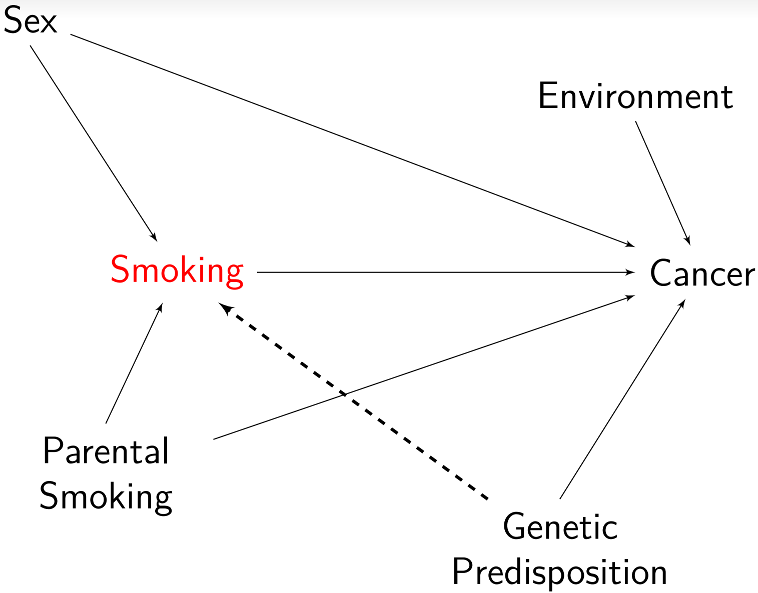
1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all possibl econfounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all possibl econfounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$
4. Basically: $Y = \beta_0 + \beta_1 X + \beta Z + \epsilon$


Sex
Environment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition


Sex
Environment
Smoking
Cancer
Parental Smoking
Genetic Predisposition

Experiments are different

Experiments are different

1. Draw causal inferences through *design* not *analysis*

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias
3. We don't need to "control" for anything

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias
3. We don't need to "control" for anything
4. We see "causal effects" in the comparison of experimental groups

Mill's Method of Difference

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or an necessary part of the cause, of the phenomenon.

Mill's Method of Difference

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, **have every circumstance save one in common**, that one occurring only in the former; **the circumstance in which alone the two instances differ, is the** effect, or **cause**, or an necessary part of the cause, **of the phenomenon**.

Definitions

Definitions

Unit: A physical object at a particular point in time

Definitions

Treatment: An intervention, whose effect(s) we wish to assess relative to some other (non-)intervention

Definitions

Potential outcomes: The outcome for each unit that we would observe if that unit received each treatment

- Multiple potential outcomes for each unit, but we only observe one of them

Definitions

Causal effect: The comparisons between the unit-level potential outcomes under each intervention

The Experimental Ideal

A randomized experiment, or randomized control trial is:

> *The observation of units after, and possibly before, a randomly assigned intervention in a controlled setting, which tests one or more precise causal expectations*

This is Holland's "statistical solution" to the fundamental problem of causal inference

Two solutions![2]

1. Scientific Solution
    - All units are identical
    - Each can provide a perfect counterfactual
    - Common in, e.g., agriculture, biology

[2] From Holland

Two solutions![2]

1. Scientific Solution
   - All units are identical
   - Each can provide a perfect counterfactual
   - Common in, e.g., agriculture, biology
2. Statistical Solution
   - Units are not identical
   - Random exposure to a potential cause
   - Effects measured on average across units
   - Known as the "Experimental ideal"

[2] From Holland

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables
- Thus experiments are a "gold standard" of causal inference

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables
- Thus experiments are a "gold standard" of causal inference
- Basically: $Y = \beta_0 + \beta_1 X + \epsilon$

Neyman–Rubin Potential Outcomes Framework

If we are interested in some outcome Y, then for every unit i, there are numerous "potential outcomes" $Y*$ only one of which is visible in a given reality. Comparisons of (partially unobservable) potential outcomes indicate causality.

Neyman–Rubin Potential Outcomes Framework

Concisely, we typically discuss two potential outcomes:

- Y_{0i}, the *potential* outcome *realized* if $X_i = 0$ (b/c $D_i = 0$, assigned to control)
- Y_{1i}, the *potential* outcome *realized* if $X_i = 1$ (b/c $D_i = 1$, assigned to treatment)

Historical Aside

- The history of the potential outcomes framework is contested
- Most people attribute it to Donald Rubin
- Paul Holland was the first to link to the philosophical discussions of causality
- Donald Rubin attributes this to Jerzy Neyman (1923)
- James Heckman denies all of this and attributes it Andrew Roy (1951)

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high
1	?	?
2	?	?
3	?	?
4	?	?

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high	control
1	?	?	?
2	?	?	?
3	?	?	?
4	?	?	?

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high	control	etc.
1	?	?	?	...
2	?	?	?	...
3	?	?	?	...
4	?	?	?	...

Experimental Inference II

- We cannot see individual-level causal effects

Experimental Inference II

- We cannot see individual-level causal effects
- We can see *average causal effects*
  - Ex.: Average difference in cancer between those who do and do not smoke

Experimental Inference II

- We cannot see individual-level causal effects
- We can see *average causal effects*
  - Ex.: Average difference in cancer between those who do and do not smoke
- We want to know: $TE_i = Y_{1i} - Y_{0i}$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average: $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average: $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
- But we still only see one potential outcome for each unit:

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0]$$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average: $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
- But we still only see one potential outcome for each unit:

  $ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0]$
- Is this what we want to know?

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

- Are the following statements true?
  - $E[Y_{1i}] = E[Y_{1i}|X = 1]$
  - $E[Y_{0i}] = E[Y_{0i}|X = 0]$

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

- Are the following statements true?
  - $E[Y_{1i}] = E[Y_{1i}|X = 1]$
  - $E[Y_{0i}] = E[Y_{0i}|X = 0]$
- Not in general!

Experimental Inference V

- Only true when both of the following hold:

$$E[Y_{1i}] = E[Y_{1i}|X = 1] = E[Y_{1i}|X = 0] \qquad (3)$$
$$E[Y_{0i}] = E[Y_{0i}|X = 1] = E[Y_{0i}|X = 0] \qquad (4)$$

- In that case, potential outcomes are *independent* of treatment assignment
- If true (e.g., due to randomization of X), then:

$$\begin{aligned} ATE_{naive} &= E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \qquad (5) \\ &= E[Y_{1i}] - E[Y_{0i}] \\ &= ATE \end{aligned}$$

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[3]

[3] Random means “known probability of treatment” not “haphazard”.

Experimental Inference VI

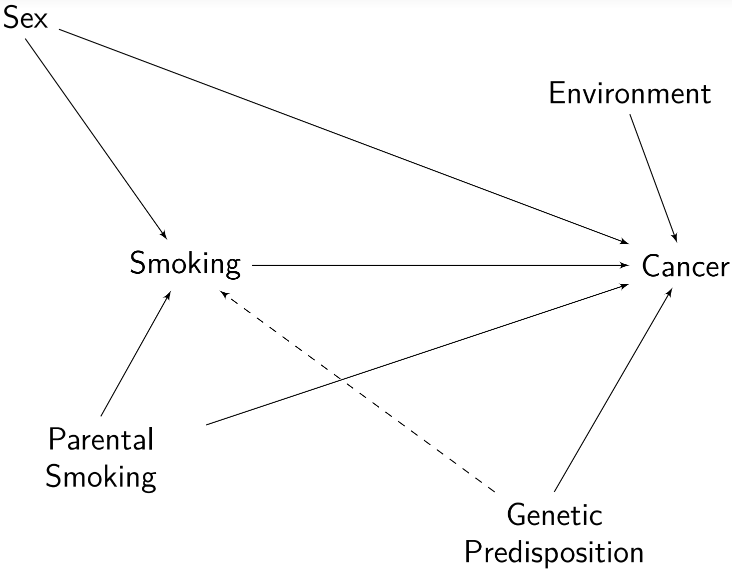
- This holds in experiments because of a *physical process of randomization*[3]
- Units differ only in side of coin that was up
  - $X_i = 1$ only because $D_i = 1$

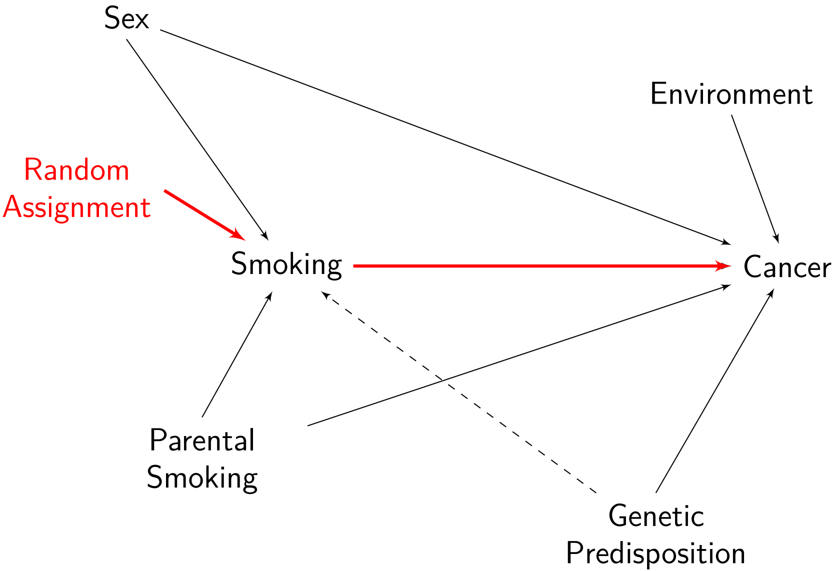
[3] Random means "known probability of treatment" not "haphazard".

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[3]
- Units differ only in side of coin that was up
  - $X_i = 1$ only because $D_i = 1$
- Implications:
  - Covariate balance
  - Potential outcomes balanced and independent of treatment assignment
  - No confounding (selection bias)

[3] Random means "known probability of treatment" not "haphazard".


Sex
Environment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition


Sex
Environment
Random
Assignment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition

Questions?

Does randomization *guarantee balance*?
Does it work every time?

Does randomization *guarantee balance*?
Does it work every time?

What happens if there is imbalance? How would we know?

Balance Testing I

- Analysis of experiments assumes that randomization produces covariate balance

Balance Testing I

- Analysis of experiments assumes that randomization produces covariate balance
- But this is only true *in expectation*

Balance Testing I

- Analysis of experiments assumes that randomization produces covariate balance
- But this is only true *in expectation*
- If we find covariate imbalance, we can:
    - Ignore it
    - Condition on imbalanced covariates

Balance Testing II

There are three basic ways to detect covariate imbalance:

1. Regressing treatment assignment on covariates
2. Conducting t-tests for each covariate across experimental groups
3. Examining covariate means visually

Experimental Analysis

- The statistic of interest in an experiment is the *sample average treatment effect* (SATE)
- If our sample is *representative*, then this provides an estimate of the population average treatment (PATE)
- This boils down to being a mean-difference between two groups:

$$SATE = \frac{1}{n_1}\sum Y_{1i} - \frac{1}{n_0}\sum Y_{0i} \tag{5}$$

[4] But not medians, etc.

Experimental Analysis

- The statistic of interest in an experiment is the *sample average treatment effect* (SATE)
- If our sample is *representative*, then this provides an estimate of the population average treatment (PATE)
- This boils down to being a mean-difference between two groups:

$$SATE = \frac{1}{n_1}\sum Y_{1i} - \frac{1}{n_0}\sum Y_{0i} \tag{5}$$

- The Neyman–Rubin logic only works for *means*[4]

[4] But not medians, etc.

Computation of Effects

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent

Computation of Effects

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
    - Disciplinary norms

Computation of Effects

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
  - Disciplinary norms
  - Ease of interpretation

Computation of Effects

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
    - Disciplinary norms
    - Ease of interpretation
    - Flexibility for >2 treatment conditions

Experimental Data Tidying

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Experimental Data Tidying

Sometimes it looks like this instead, which is bad:

unit	treatment	outcome0	outcome1
1	0	13	.
2	0	6	.
3	0	4	.
4	0	5	.
5	1	.	3
6	1	.	1
7	1	.	10
8	1	.	9

Experimental Data Tidying

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Experimental Data Tidying

Sometimes it looks like this instead, which is even worse:

unit	treatment	outcome0	outcome1
1	.	13	.
2	.	6	.
3	.	4	.
4	.	5	.
5	.	.	3
6	.	.	1
7	.	.	10
8	.	.	9

Experimental Data Tidying

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Experimental Data Tidying

Sometimes it looks like this instead, which is even more worse:

unit	treatment	outcome0	outcome1	order
1	.	13	6	0,1
2	.	6	8	0,1
3	.	4	2	0,1
4	.	5	1	0,1
5	.	9	3	1,0
6	.	4	1	1,0
7	.	2	10	1,0
8	.	8	9	1,0

Experimental Data Tidying

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Computation of Effects in Stata

Stata:

```
ttest outcome, by(treatment)
reg outcome i.treatment
```

R:

```
t.test(outcome ~ treatment, data = data)
lm(outcome ~ factor(treatment), data = data)
```

Questions?

SATE Variance Estimation

- We don't just care about the size of the SATE. We also want to know whether it is significantly different from zero (i.e., different from no effect/difference)
- To know that, we need to estimate the *variance* of the SATE
- The variance is influenced by:
    - Total sample size
    - Variance of the outcome, Y
    - Relative size of each treatment group

SATE Variance Estimation

- Formula for the variance of the SATE is:
  $$\widehat{Var}(SATE) = \frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}$$
  - $\widehat{Var}(Y_0)$ is control group variance
  - $\widehat{Var}(Y_1)$ is treatment group variance
- We often express this as the *standard error* of the estimate:
  $$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Intuition about Variance

- Bigger sample $\rightarrow$ smaller SEs
- Smaller variance $\rightarrow$ smaller SEs
- Efficient use of sample size:
  - When treatment group variances equal, equal sample sizes are most efficient
  - When variances differ, sample units are better allocated to the group with higher variance in Y

Important considerations

- Required sample size depends on $SATE$ and $Var(Y)$

Important considerations

- Required sample size depends on $SATE$ and $Var(Y)$
- In large populations, population size is irrelevant

Important considerations

- Required sample size depends on *SATE* and $Var(Y)$
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled

Important considerations

- Required sample size depends on $SATE$ and $Var(Y)$
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled
- In anything other than an SRS, sample size calculation is more difficult

Important considerations

- Required sample size depends on $SATE$ and $Var(Y)$
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled
- In anything other than an SRS, sample size calculation is more difficult
- Most research assumes SRS even though a more complex design is actually used

Important considerations

- Required sample size depends on *SATE* and *Var*(*Y*)
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled
- In anything other than an SRS, sample size calculation is more difficult
- Most research assumes SRS even though a more complex design is actually used
- Sample size needed to obtain a precise estimate

Estimating sample size

What precision (margin of error) do we want?

- p +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \qquad (6)$$

Estimating sample size

What precision (margin of error) do we want?

- p +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{6}$$

- p +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{7}$$

Estimating sample size

What precision (margin of error) do we want?

- p +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{6}$$

- p +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{7}$$

- p +/- 0.5 percentage points: $SE = 0.0025$

$$n = \frac{0.25}{0.00000625} = 40,000 \tag{8}$$

Statistical Power

- Power analysis to determine sample size
- Type I and Type II Errors
  - True positive rate is power
  - False negative rate is the significance threshold (α)

	H_0 True	H_0 False
Reject H_0	Type 1 Error	**True positive**
Accept H_0	False negative	Type II error

Doing a Power Analysis

- μ, Treatment group mean outcomes
- N, Sample size
- σ, Outcome variance
- α Statistical significance threshold
- ϕ, a sampling distribution

$$Power = \phi\left(\frac{|\mu_1-\mu_0|\sqrt{N}}{2\sigma} - \phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)$$

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, "true" effect size, variance of outcome, power, and α.

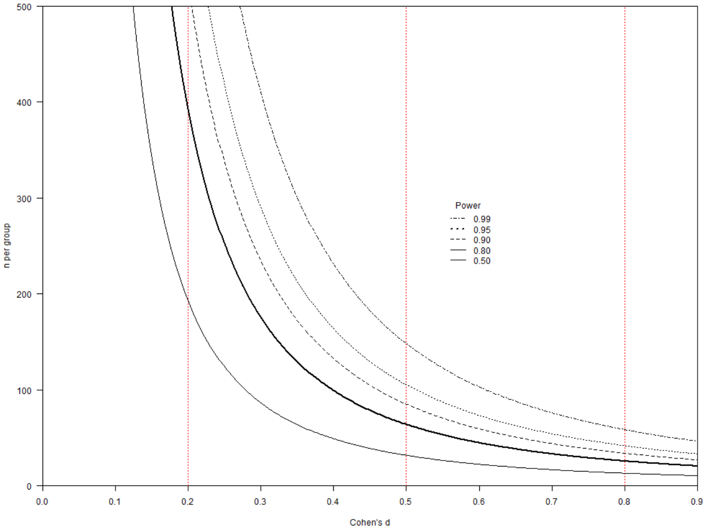
In essence: some non-zero effect sizes are not detectable by a study of a given sample size.[5]

[5] Gelman, A. and Weakliem, D. 2009. "Of Beauty, Sex and Power." *American Scientist* 97(4): 310–16

Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- Cohen's d:
  $d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, where $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1+n_0-2}}$
- Intuition: How large is the effect in standard deviations of the outcome?
  - Know if effects are large or small
  - Compare effects across studies
- Small: 0.2; Medium: 0.5; Large: 0.8

Intuition about Power



Aside: Complex Designs

- An experiment can have any number of conditions
  - Up to the limits of sample size
  - More than 8–10 conditions is typically unwieldy
- Typically analyze complex designs using ANOVA or regression, but we are still ultimately interested in pairwise comparisons to estimates SATEs
  - Treatment–treatment, or treatment-control
  - Without control group, we don't know which treatment(s) affected the outcome

Survey-experiments, specifically

- Everything so far applies to any kind of experiment

Survey-experiments, specifically

- Everything so far applies to any kind of experiment
- A survey experiment is just an experiment that occurs in a survey context
    - As opposed to in the field or in a laboratory

Survey-experiments, specifically

- Everything so far applies to any kind of experiment
- A survey experiment is just an experiment that occurs in a survey context
    - As opposed to in the field or in a laboratory
- Sometimes a distinction is made between survey and online experiments

Survey-experiments, specifically

- Everything so far applies to any kind of experiment
- A survey experiment is just an experiment that occurs in a survey context
    - As opposed to in the field or in a laboratory
- Sometimes a distinction is made between survey and online experiments
- Lots of common paradigms for survey experiments (tomorrow)

Combining Survey Design and Experimental Design

- Sample is representative of population in every respect (in expectation)
- Sample Average Treatment Effect (SATE) is the average of the sample's individual-level treatment effects
    - Unbiased estimate of PATE

Combining Survey Design and Experimental Design

- Sample is representative of population in every respect (in expectation)
- Sample Average Treatment Effect (SATE) is the average of the sample's individual-level treatment effects
    - Unbiased estimate of PATE
- Says nothing about effect heterogeneity
    - Design is optimized for estimating SATE
    - Discuss this on Wednesday

Questions?

Randomization Distribution

```
# theoretical randomizations
onedraw <- function(eff=FALSE, dat = d) {
  r <- replicate(nrow(dat), sample(1:2,1))
  dat[cbind(1:nrow(dat),r)] <- NA
  if (eff) {
    return(mean(dat[,'y1'], na.rm=TRUE) -
           mean(dat[,'y0'], na.rm=TRUE) )
  } else {
    return(dat)
  }
}

onedraw() # one randomization
onedraw(TRUE) # one effect estimate

# simulate 2000 experiments from these data
x1 <- replicate(2000, onedraw(TRUE))
hist(x1, col=rgb(1,0,0,.5), border='white')
abline(v=-2, lwd=3, col='red') # true effect
```

One way to avoid covariate imbalance and improve statistical power is **block randomization**.

Block Randomization I

Stratification:Sampling::Blocking:Experiments

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment
- CATE is estimate for each stratum; aggregated to SATE

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment
- CATE is estimate for each stratum; aggregated to SATE
- Why?
  - Eliminate chance imbalances
  - Optimized for estimating CATEs
  - More precise SATE estimate

Exp.	Control				Treatment			
1	M	M	M	M	F	F	F	F
2	M	M	M	F	M	F	F	F
3	M	M	F	F	M	M	F	F
4	M	F	F	F	M	M	M	F
5	F	F	F	F	M	M	M	M

Obs.	X_{1i}	X_{2i}	D_i
1	Male	Old	0
2	Male	Old	1
3	Male	Young	1
4	Male	Young	0
5	Female	Old	1
6	Female	Old	0
7	Female	Young	0
8	Female	Young	1

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
    - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
  - Most valuable in small samples

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
    - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
    - Most valuable in small samples
    - Not valuable if all blocks have similar potential outcomes

Statistical Properties I

Complete randomization:

$$SATE = \frac{1}{n_1}\sum Y_{1i} - \frac{1}{n_0}\sum Y_{0i}$$

Block randomization:

$$SATE_{blocked} = \sum_{1}^{J}\left(\frac{n_j}{n}\right)(\widehat{CATE}_j)$$

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	3
8	Female	Young	1	9	

SATE Estimation

$$
\begin{aligned}
SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\
&= 3.75
\end{aligned}
$$

SATE Estimation

$$
\begin{aligned}
SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\
&= 3.75
\end{aligned}
$$

The blocked and unblocked estimates are the same here because $Pr(Treatment)$ is constant across blocks and blocks are all the same size.

SATE Estimation

- We can use weighted regression to estimate this in an OLS framework
- Weights are the inverse prob. of being treated w/in block
  - Pr(Treated) by block: $p_{ij} = Pr(D_i = 1 | J = j)$
  - Weight (Treated): $w_{ij} = \dfrac{1}{p_{ij}}$
  - Weight (Control): $w_{ij} = \dfrac{1}{1 - p_{ij}}$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(SATE_j)}$$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(SATE_j)}$$

When is the blocked design more efficient?

Practicalities

- Blocked randomization only works in exactly the same situations where stratified sampling works
  - Need to observe covariates pre-treatment in order to block on them
  - Work best in a panel context
- In a single cross-sectional design that might be challenging
  - Some software can block "on the fly"