Experimental Research in Legislative Studies

Thomas J. Leeper

Government Department
London School of Economics and Political Science

18 August 2017

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

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? *Group 2*

Enter your data

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

Results

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

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Who am I?

- Thomas Leeper
- Associate 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 survey and experimental methods and political psychology
- Email: t.leeper@lse.ac.uk

Who are you?

- What's your name?
- Where are you from?
- Have you designed and/or analyzed an experiment before?

Course Materials

All material for this workshop, including required and suggested readings, are available at:

http://www.thomasleeper.com/legexpcourse/

Learning Outcomes

By the end of the day, 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 three common legislative experimental paradigms: survey experiments, field experiments, and simulations.
4. Identify practical issues that arise in the implementation of experiments and evaluate how to anticipate and respond to them.

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Questions?

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Experiments: Definition

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 have a long history

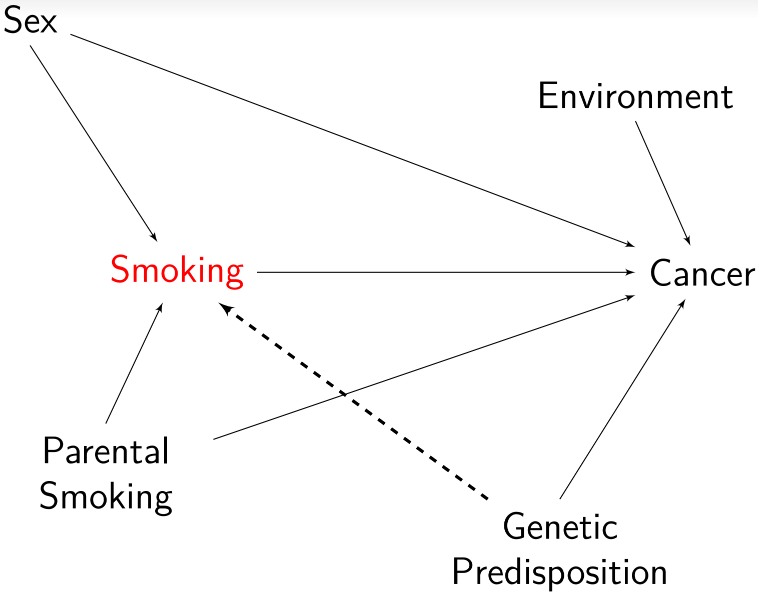
- Origins in agricultural and biostatistical research in the 19th century (Fisher, Neyman, Pearson, etc.)
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884
- First polisci experiment by Gosnell (1924)
- Survey experiments have been common since 1930s
- Gerber and Green (2000) first *major*, *modern*, *field* experiment

Legislative Experiments

- Experiments in legislative contexts fit awkwardly in that history and the dominant paradigms have very different histories
- **Simulations**
  - Originated in formal literatures on committee behavior, coalition formation, and majority rule institutions
- **Field experiments**
  - Really only emerged in the past decade
- **Survey Experiments**
  - Much more sparsely used for reasons that will become obvious

What kinds of questions can we answer with experiments?

- Forward causal questions
  - Can X cause Y?
  - What effects does X have?
- Backward causal questions
  - What causes Y?
  - How much of Y is attributable to X?
- Even though answering “forward” causal question, we start with an outcome concept


Sex
Environment
Smoking
Cancer
Parental Smoking
Genetic Predisposition

Principles of causality

1. **Correlation/Relationship**
2. **Nonconfounding**
3. **Direction ("temporal precedence")**
4. Mechanism
5. Appropriate level of analysis

Establishing Relationship

- This is fairly trivial
- Simply find value of $Corr(X, Y)$
- In causal inference we often talk about correlations in terms of *differences*
  - Difference in values of Y across values of X
  - The presence of a difference indicates a correlation

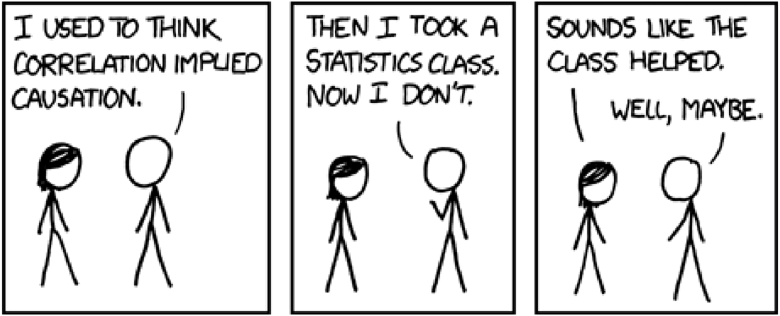
Addressing Confounding

In observational studies, we address confounding by:

1. Correlating a "putative" cause (X) and an outcome (Y)
2. Identifying all possible confounds ($\mathbf{Z}$)
3. "Conditioning" on all confounds
   - Calculating correlation between X and Y at each combination of levels of $\mathbf{Z}$

Temporal Precedence

- Even if an observational design identifies a relationship and credibly addresses sources of confounding, it still may not be a credible causal inference
- "Reverse causality" is vague, referring to:
    - Ambiguity about causal ordering, or
    - Sequentially reinforcing causality between X and Y
- Causation is strictly forward moving in time
- X must precede Y in time for X to cause Y
    - X can be *measured* after Y as long as it comes before it

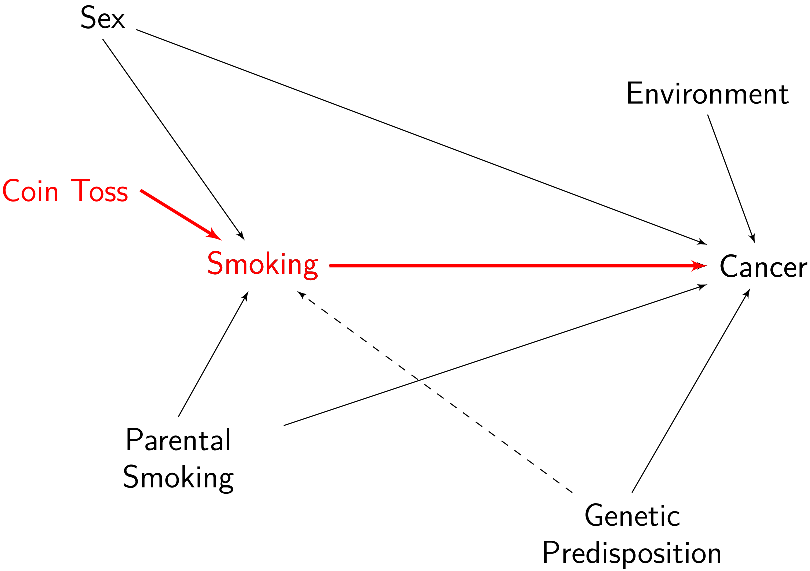

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Experiments!

- A randomized experiment, or randomized control trial (RCT) 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

- If we manipulate the thing we want to know the effect of (X), and control (i.e., hold constant) everything we do not want to know the effect of (Z), the only thing that can affect the outcome (Y) is X.


Sex
Environment
Coin Toss
Smoking
Cancer
Parental Smoking
Genetic Predisposition

Questions?

Definitions

Unit: A physical object at a particular point in time
Treatment: An intervention, whose effect(s) we wish to assess relative to some other (non-)intervention
Outcome: The variable we are trying to explain
Potential outcomes: The outcome value 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
Causal effect: The comparisons between the unit-level potential outcomes under each intervention

"The Perfect Doctor"

Unit	Y_0	Y_1
1	?	?
2	?	?
3	?	?
4	?	?
5	?	?
6	?	?
7	?	?
8	?	?
Mean	**?**	**?**

"The Perfect Doctor"

Unit	Y_0	Y_1
1	?	14
2	6	?
3	4	?
4	5	?
5	6	?
6	6	?
7	?	10
8	?	9
Mean	**5.4**	**11**

“The Perfect Doctor”

Unit	Y_0	Y_1
1	13	14
2	6	0
3	4	1
4	5	2
5	6	3
6	6	1
7	8	10
8	8	9
Mean	**7**	**5**

Experimental Inference I

- 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 II

- We want to know: $TE_i = Y_{1i} - Y_{0i}$
- We can average: $ATE = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
- 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 III

- 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 IV

- Only true when both of the following hold:

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

- In that case, potential outcomes are *independent* of treatment assignment

- If true, then:

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

Experimental Inference V

- This holds in experiments because of randomization, which is a special, physical process of unpredictable sorting[1]
    - Units differ only in what side of coin was up
    - Experiments randomly reveal potential outcomes
    - Randomization balances Z *in expectation*
- Matching/regression/etc. attempts to eliminate those confounds, such that:

$$E[Y_{1i}|Z] = E[Y_{1i}|X = 1, Z] = E[Y_{1i}|X = 0, Z]$$
$$E[Y_{0i}|Z] = E[Y_{0i}|X = 1, Z] = E[Y_{0i}|X = 0, Z]$$

[1] Not "random" in the casual, everyday sense of the word

Why an 'Experimental Ideal'?

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

Questions?

Experimental Analysis I

- The statistic of interest in an experiment is the *sample average treatment effect* (SATE)
- This boils down to being a mean-difference between two groups:

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

- In practice we often estimate this using:
  - t-tests
  - OLS regression
- Experiments do not require “controlling for” anything, if randomization occurred successfully

Why use regression?

1. Coefficient estimates are directly interpretable as estimated SATEs
2. Basically no functional form or specification assumptions involved
3. Flexibly accommodates experiments with > 2 conditions
   - *n*-condition experiments
   - *Factorial* designs

Two ways to *parameterize* factorial designs

Dummy variable regression (i.e., treatment–control CATEs):
$Y = \beta_0 + \beta_1 X_{0,1} + \beta_2 X_{1,0} + \beta_3 X_{1,1} + \epsilon$

Interaction effects (i.e., treatment–treatment CATEs):
$Y = \beta_0 + \beta_1 X1_1 + \beta_2 X2_1 + \beta_3 X1_1 * X2_1 + \epsilon$

Use `margins` to extract marginal effects

Computation of Effects in Stata/R

Stata:

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

R:

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

Experimental Data Structures

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 Structures

Sometimes it looks like this instead, which is bad:

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

Experimental Data Structures

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

Questions?

Experimental Analysis II

- 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

Experimental Analysis III

- 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 → smaller SEs
- Smaller variance → 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

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

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

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.

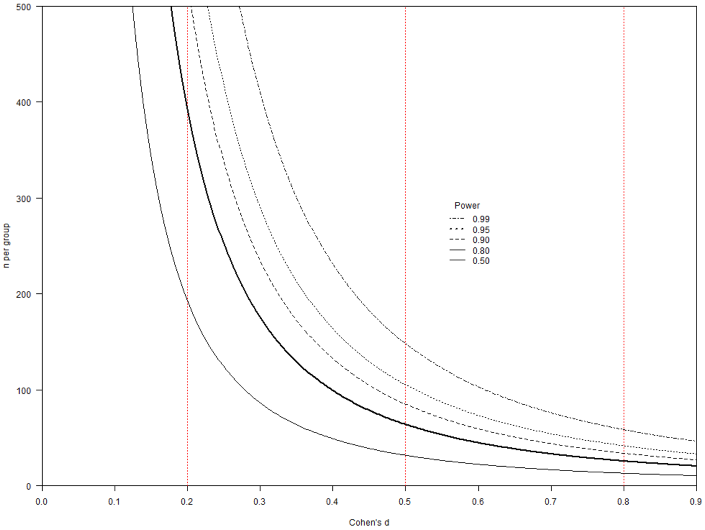
In essence: some non-zero effect sizes are not detectable by a study of a given sample size.[2]

[2] 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



Power in Legislative Experiments

- Legislatures are small!!!
- Because N is fixed, limited capacity to increase n, so power has to be maximized by:
  - Reducing item variance in outcome measures
  - Studying treatments with bigger effects
  - Expanding scope of studies
  - Creative research design

Questions?

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Experimental Hypothesis Testing

- From theory, we derive testable hypotheses
  - Hypotheses are expectations about differences in outcomes across levels of a putatively causal variable
  - In an experiment, an hypothesis must be testable by an SATE

- The experimental manipulations induce variation in the causal variable that enable tests of the hypotheses

Example: Framing and Attention[3]

- Theory: Presentation of information affects politicians' attention
- Hypothesis:
  - Information framed as a conflict draws more attention from political elites than information not framed as a conflict.
- Manipulation:
  - Control group: Presentation of headline information
  - Treatment group: Same information presented as conflict
- Outcome:
  - How likely are legislators to read full article

[3] Walgrave, Sevenans, Van Camp, Loewen (2017) – "What Draws Politicians' Attention? An Experimental Study of Issue Framing and its Effect on Individual Political Elites"

Ex.: Presence/Absence

- Theory: Legislators vote in line with constituents' preferences
- Hypothesis: Exposure to a poll of constituent views shifts legislative votes.
- Manipulation:
  - Control group receives no polling information.
  - Treatment group receives a letter containing polling information.
- Outcome:
  - How legislators vote on relevant piece of legislation

Ex.: Levels/doses

- Theory: Legislators vote in line with constituents' preferences
- Hypothesis: Exposure to a poll of constituent views shifts legislative votes.
- Manipulation:
    - Control group receives no polling information.
    - Treatment group 1 receives a letter containing polling information.
    - Treatment group 2 receives two letters containing polling information.
    - etc.
- Outcome:
    - How legislators vote on relevant piece of legislation

Ex.: Qualitative variation

- Theory: Legislators vote in line with constituents' preferences
- Hypothesis: Exposure to a poll of constituent views shifts legislative votes.
- Manipulation:
  - Control group receives no polling information.
  - Treatment group 1 receives a letter containing polling information suggesting public support.
  - Treatment group 2 receives a letter containing polling information suggesting public opposition.
- Outcome:
  - How legislators vote on relevant piece of legislation

Treatments Test Hypotheses!

- Derive experimental design from hypotheses
- Experimental “factors” are expressions of hypotheses as randomized groups
- What intervention each group receives depends on hypotheses
  - presence/absence
  - levels/doses
  - qualitative variations

But how do we know that the experiment worked?

The best criterion for evaluating the quality of an experiment is whether it manipulated the intended independent variable and controlled everything else by design.

–Thomas J. Leeper (18 August 2017)

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory
- Before the study using *pilot testing* (or *pretesting*)
- During the study, using *manipulation checks*
- During the study, using *placebos*
- During the study, using *non-equivalent outcomes*

These may not all be possible and all are incompletely informative.

How do we know we controlled what we think we controlled?

- Measure characteristics across groups to test for *covariate balance*, but imbalance does not necessarily imply experimental failure
- In field experiments, measure whether legislators were actually treated (i.e., actually received and *complied* with their assigned treatment)
- Measure whether there were spillovers between experimental conditions if possible

These may not all be possible and all are incompletely informative.

Questions?

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Three Major Paradigms

1. Field Experiments
   - Ex. Broockman (2013) – "Black Politicians Are More Intrinsically Motivated to Advance Blacks' Interests"
2. Survey Experiments
   - Ex. Renshon, Yarhi-Milo, and Kertzer (2016) – "Democratic Leaders, Crises and War: Paired Experiments on the Israeli Knesset and Public"
3. Simulations
   - Ex. Frechette, Kagel, Lehrer (2003) – "Bargaining in Legislatures: An Experimental Investigation of Open versus Closed Amendment Rules"

Paradigm 1: Field Experiments

- Basic idea: randomly expose legislators *in situ* to some experience and measure an outcome that might be affected by it
- Two "flavours"
    1. Orchestrated by the researcher(s)
    2. "Natural" experiments not orchestrated by the researcher(s)
- Tend to be simple in terms of design due to practical difficulty of exposing legislators' to treatment and measuring outcomes
- "Natural" experiments are limited by randomized institutions being rare (e.g., committee assignments, office locations, proposal rights/order, etc.)

Paradigm 1: Field Experiments

- Example of Flavour A
  - Broockman (2013)
  - Treatment: Form of contact from a prospective constituent
  - Outcome: Whether a response is received
  - Effect: Difference in response rates by treatment
- Example of Flavour B
  - Kellermann, Shepsle (2009)
  - Treatment: Freshmen legislators are randomly ordered in determining committee assignments
  - Outcome: Various metrics of leadership and legislative activity
  - Effect: Difference in those outcomes between higher- and lower-ranked legislators

Paradigm 2: Survey experiments

- Basic idea: conduct interviews with legislators (in-person or through another mode), where features of questionnaire are randomized

- Recruiting legislators into interviews tends to be extremely difficult, thus:
    - Almost unavoidably underpowered
    - Can only study legislators who agree to participate
    - Necessarily simplistic designs with treatment and outcome measured in a single interview[4]
    - Survey experiments on legislators tend to be rare

[4] Can be generalized to allow field treatments with survey measures, or survey treatments with field measures

Paradigm 2: Survey Experiments

Example:

- Butler and Dynes (2016)
- Treatment: State legislators completing a survey read a hypothetical constituent letter with varying stated opinions
- Outcome: Measures of perceptions of constituent characteristics (e.g., knowledge)
- Effect: Difference in perceptions b/w constituents with similar/dissimilar views to legislator

Paradigm 3: Simulations

- Basic idea: Derive theoretical expectations about legislative behavior and test those predictions in a *stylized* legislative context using non-legislators as participants

- These are historically much more common than paradigms 1 or 2

- Unique considerations:
  - Tend to be based in formal theories of legislatures
  - Sample sizes limited by resources
  - Historically in labs, but increasingly common online
  - Tend to lack face validity given context and participants

Paradigm 3: Simulations

Example:

- Wilson (1986)
- Treatment: “Legislators” vote under open or closed amendment rules
- Outcome: The final “policy” adopted by the “legislature”
- Effect: Difference in “policy” adopted by the legislature

Questions?

15-minute Activity!

1. Divide into three groups
2. Groups discuss one of the texts:
   - Group 1: Broockman (2013)
   - Group 2: Renshon, Yarhi-Milo, and Kertzer (2016)
   - Group 3: Frechette, Kagel, Lehrer (2003)
3. Discuss:
   - What is the experiment? How does it work?
   - What do the authors find? What is the effect?
   - What are the practical challenges/issues raised?

Activity!

- How do we know if an experiment is any good?
- Write for 3 minutes to yourself
- Talk with a partner for about 3 minutes
- Try to develop some criteria that allow you to evaluate “what makes for a good experiment?”

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Many Challenges, Too Little Time

1. Nonresponse and Noncompliance
2. Spillover
3. What can be randomized?
4. Ethics

Nonresponse

- In survey experiments, nonresponse may introduce challenges:
  - Underpowered designs
  - Response biases that affect generalizability
  - Nonresponse may be due to treatment
  - Nonresponse may be due to attrition
- The only way to avoid nonresponse is to try to incentivize response or minimize effort involved in a study
- Real risk: more surveys might create common pool resource problems!

Compliance

- Compliance is when individuals receive and accept the treatment to which they are assigned, as opposed to:
    - Receiving the wrong treatment (cross-over)
    - Failing to receive any treatment
- This causes problems for our analysis because factors other than randomization explain why individuals receive their treatment
- Possible responses to noncompliance:
    - "As treated" analysis
    - "Intention to treat" analysis
    - Estimate a LATE

Analyzing Noncompliance

- If noncompliance only occurs in one group, it is *asymmetric* or *one-sided*
- We can ignore non-compliance and analyze the "intention to treat" effect, which will underestimate our effects because some people were not treated as assigned: $ITT = \overline{Y}_1 - \overline{Y}_0$
- We can use "instrumental variables" to estimate the "local average treatment effect" (LATE) for those that complied with treatment: $LATE = \frac{ITT}{\%Compliant}$
- If noncompliance is symmetric, analysis much more complicated.

Questions?

Spillover

- A key assumption of experimental analysis is that units are *independent*
- This assumption may be implausible in legislatures because units are in regular communication and may "share" some of their treatment with others in the group
- What can be done?
  - Try to avoid it by design!
  - Exclude individuals affected by spillovers, if observable
  - More complicated procedures

What can be randomized?

- In theory almost anything can be randomized, but not everything
    - Intrinsic characteristics
    - Institutional features (outside of simulations)
    - Contextual factors
- Anything that is "information-like" can easily and obviously be randomized[5]
- If you want to study factors that are not information-like:
    - Look for "natural" experiments
    - Run simulations
    - Run field or survey experiments that attempt to modify the *salience* of those factors

[5] Messages, contact, personal interactions, etc.

Research Ethics

- Researchers have obligations to attempt to:
  - minimize risk to participants
  - to maximize benefits to human knowledge
  - to protect the privacy of personal data
  - to fairly and objectively report their research
- These rules vary to some extent across contexts
- But a major question is whether these "standard" ethical rules also apply to politicians. What do you think?

Questions?

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Student presentations!

1 Causal Inference

2 From Theory to Experimental Design

3 Paradigms and Examples

4 Challenges of Legislative Experiments

5 Student Presentations

6 Conclusion

Learning Outcomes

By the end of the day, 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 three common legislative experimental paradigms: survey experiments, field experiments, and simulations.
4. Identify practical issues that arise in the implementation of experiments and evaluate how to anticipate and respond to them.

In Conclusion

- Experiments are mostly about design, not analysis
- Experiments are underutilized in legislative contexts, in part because conducting them effectively is extremely difficult
- This means that careful but often simple design can generate potentially powerful and novel insights into legislative behavior