Session I
Survey Experiments in Practice

Thomas J. Leeper

Government Department
London School of Economics and Political Science

30 January 2017

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. History of Experimentation
2. Logic and Analysis
3. From Theory to Design
4. Operationalization Principles
   - Common Paradigms and Examples

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?

- Where are you from?
- Have you designed a survey and/or experiment before?
- 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/`

`https://github.com/leeper/surveyexpcourse`

Learning Outcomes

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

Learning Outcomes

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

1. Explain how to analyze experiments quantitatively.

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.

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 several common survey experimental paradigms.

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 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.

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: History

- "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 (Fisher, Neyman, Pearson, etc.)
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884

Experiments: History

- "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 (Fisher, Neyman, Pearson, etc.)
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884
  - First experiment by Gosnell (1924)
  - Gerber and Green (2000) first major *field* experiment

What distinguishes a *survey* experiment from any other experiment?

1. Field Experiments
2. Laboratory Experiments
3. Survey Experiments

Difference is only about *setting* and *mode*.

1. Field Experiments
2. Laboratory Experiments
3. Survey Experiments

Difference is only about *setting* and *mode*.
Logic and methods of analysis are the same!

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Paper-and-pencil mode limited experimentation
  - Limited use of "split ballots"

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Paper-and-pencil mode limited experimentation
  - Limited use of "split ballots"
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop **CATI**

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Paper-and-pencil mode limited experimentation
  - Limited use of "split ballots"
- 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.

Survey-experiments, specifically

Survey-experiments, specifically

- 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

- A survey experiment is just an experiment that occurs in a survey context
  - As opposed to in the field or in a laboratory
- Properties:
  - 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
  - SATE is unbiased estimate of PATE

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[2]

[2] See also: LISS, Bergen's Citizen Panel, Gothenburg's Citizen Panel

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

TESS has "Open Protocols"

Protocol is the complete planning document for how to design, implement, and analyze an experiment.[3]

1. Theory/hypotheses
2. Instrumentation
   - Manipulation(s)
   - Outcome(s)
   - Covariate(s)
   - Manipulation check(s)
3. Sampling
4. Implementation
5. Analysis

[3] Thomas J. Leeper. 2011. "The Use of Protocol in the Design and Reporting of Experiments." *The Experimental Political Scientist*.

Why bother writing a protocol?

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing

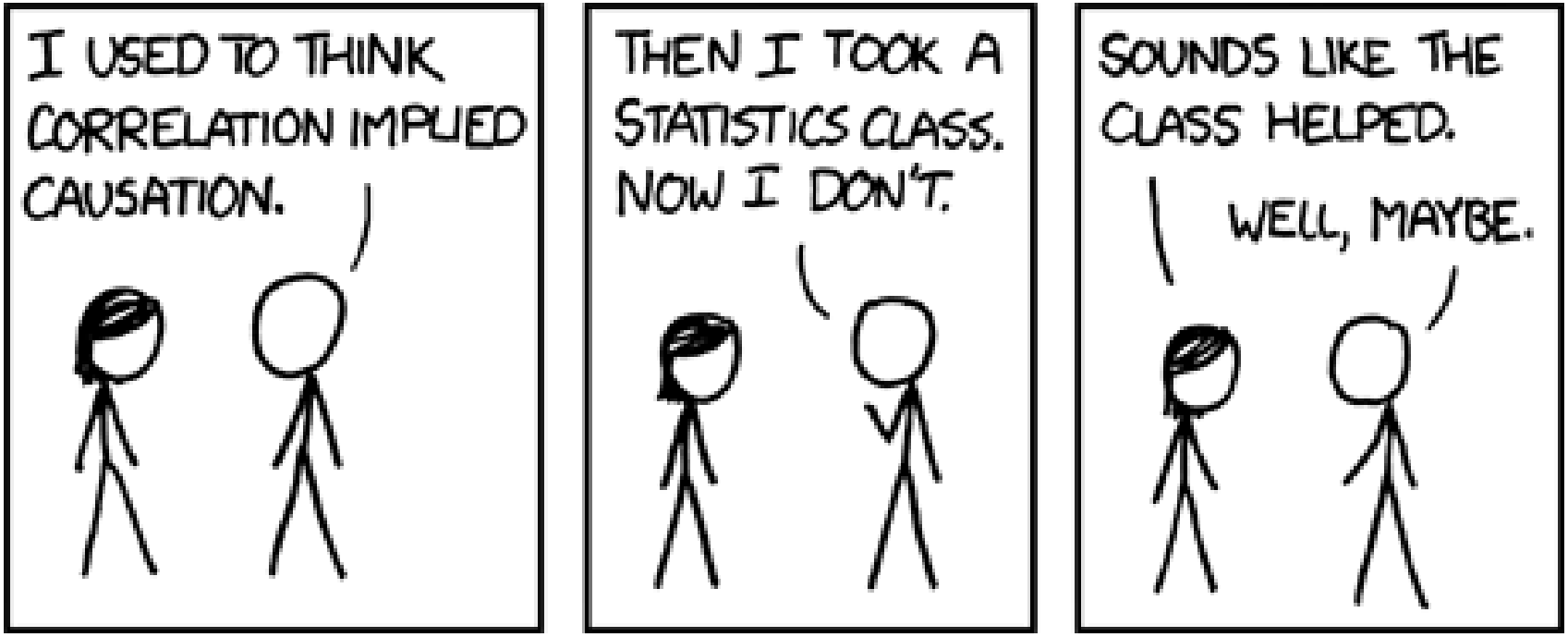
Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing
- Economize questionnaire development

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing
- Economize questionnaire development
- Study preregistration

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

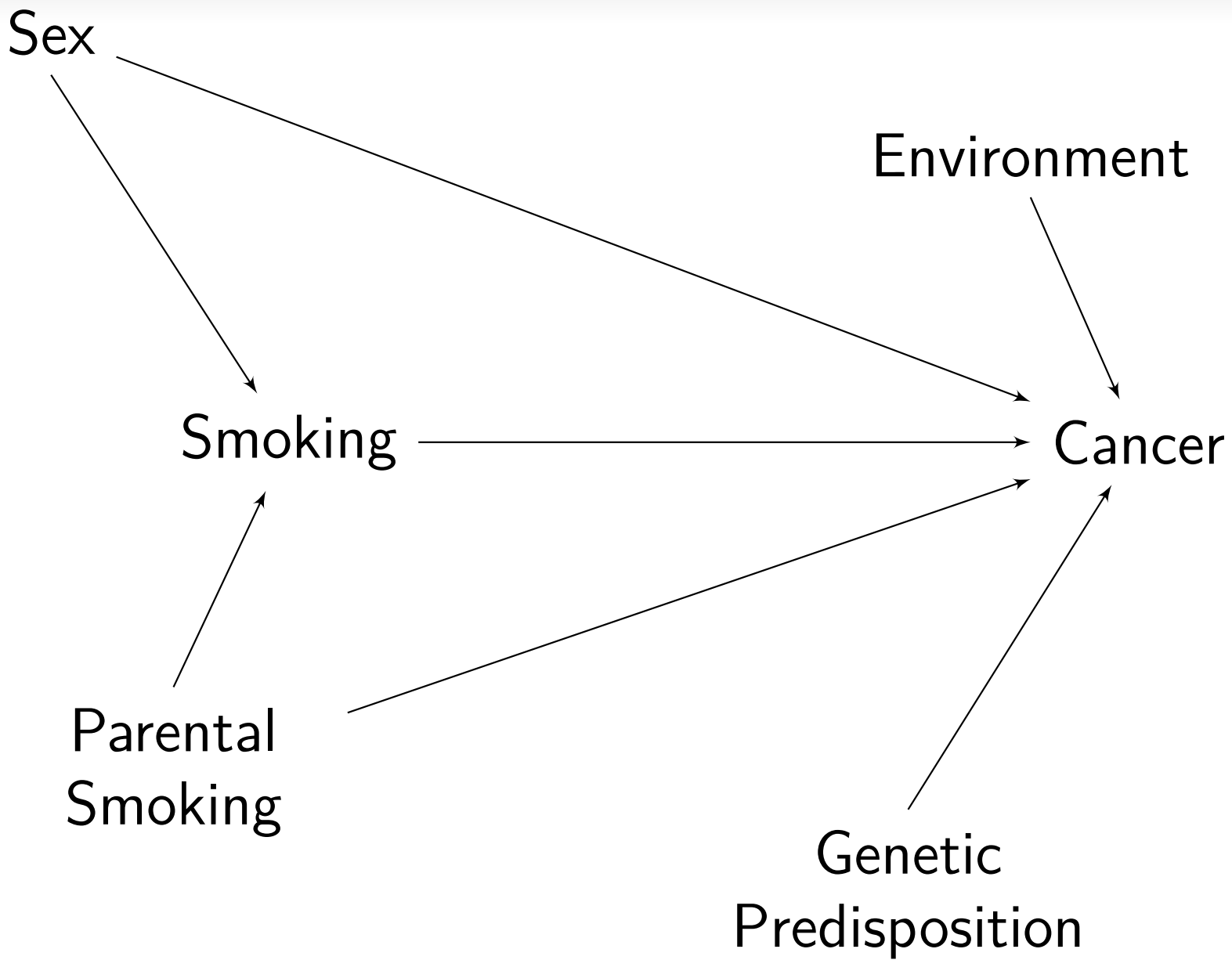
In observational research...

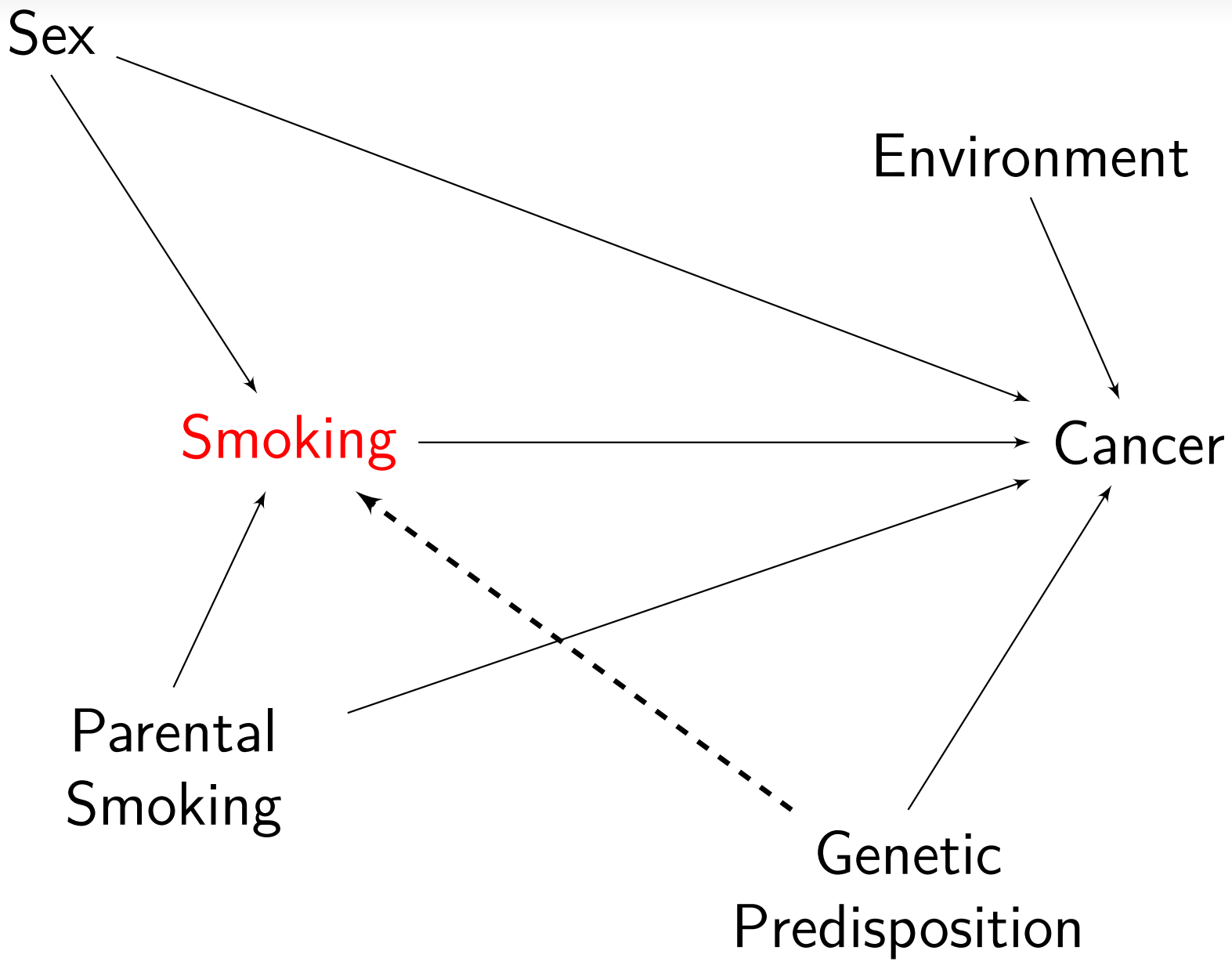
1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all possibl econfounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$

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![4]

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

[4] From Holland

Two solutions![4]

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"

[4] 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:

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

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

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

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[5]

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

Experimental Inference VI

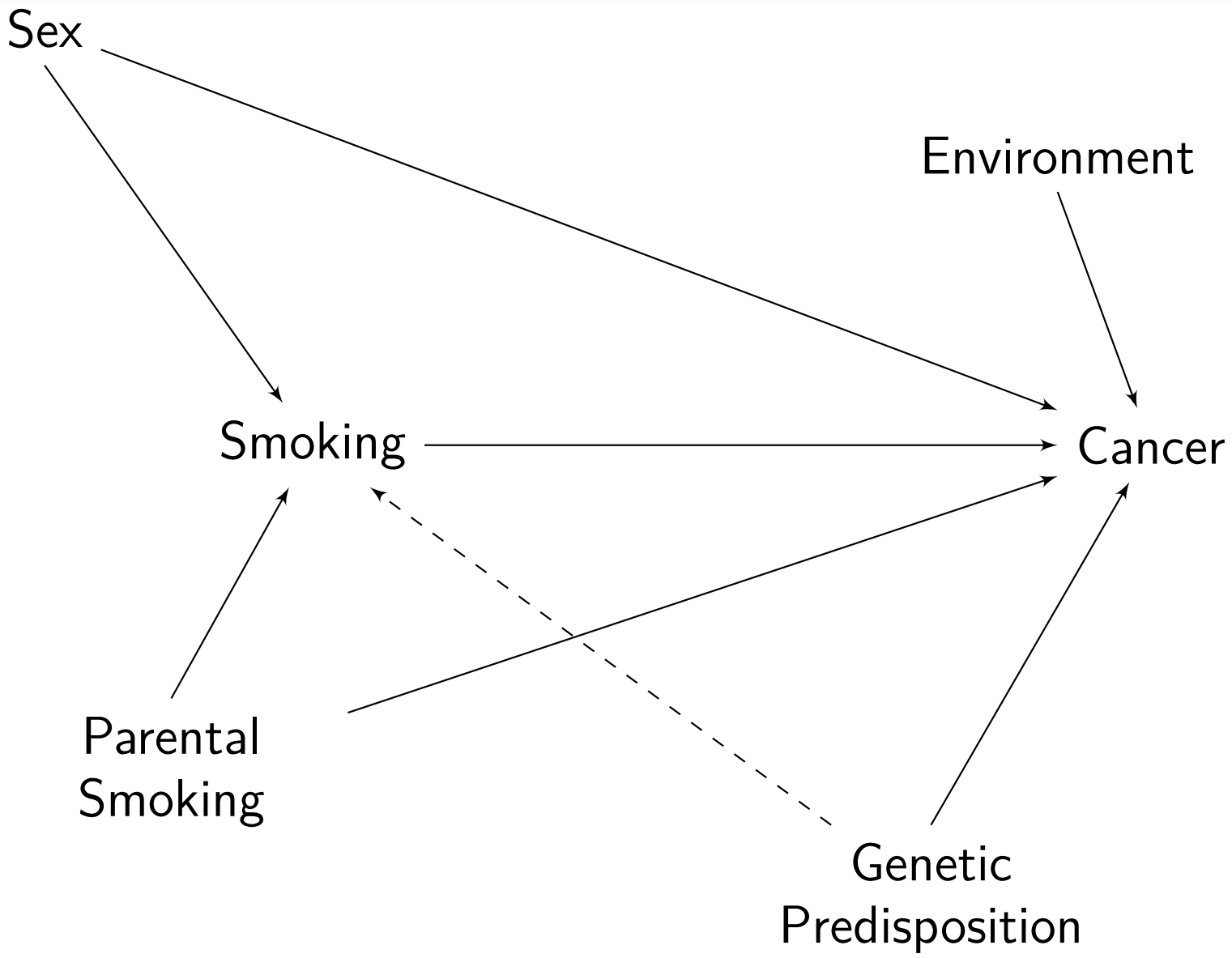
- This holds in experiments because of a *physical process of randomization*[5]
- Units differ only in side of coin that was up
  - $X_i = 1$ only because $D_i = 1$

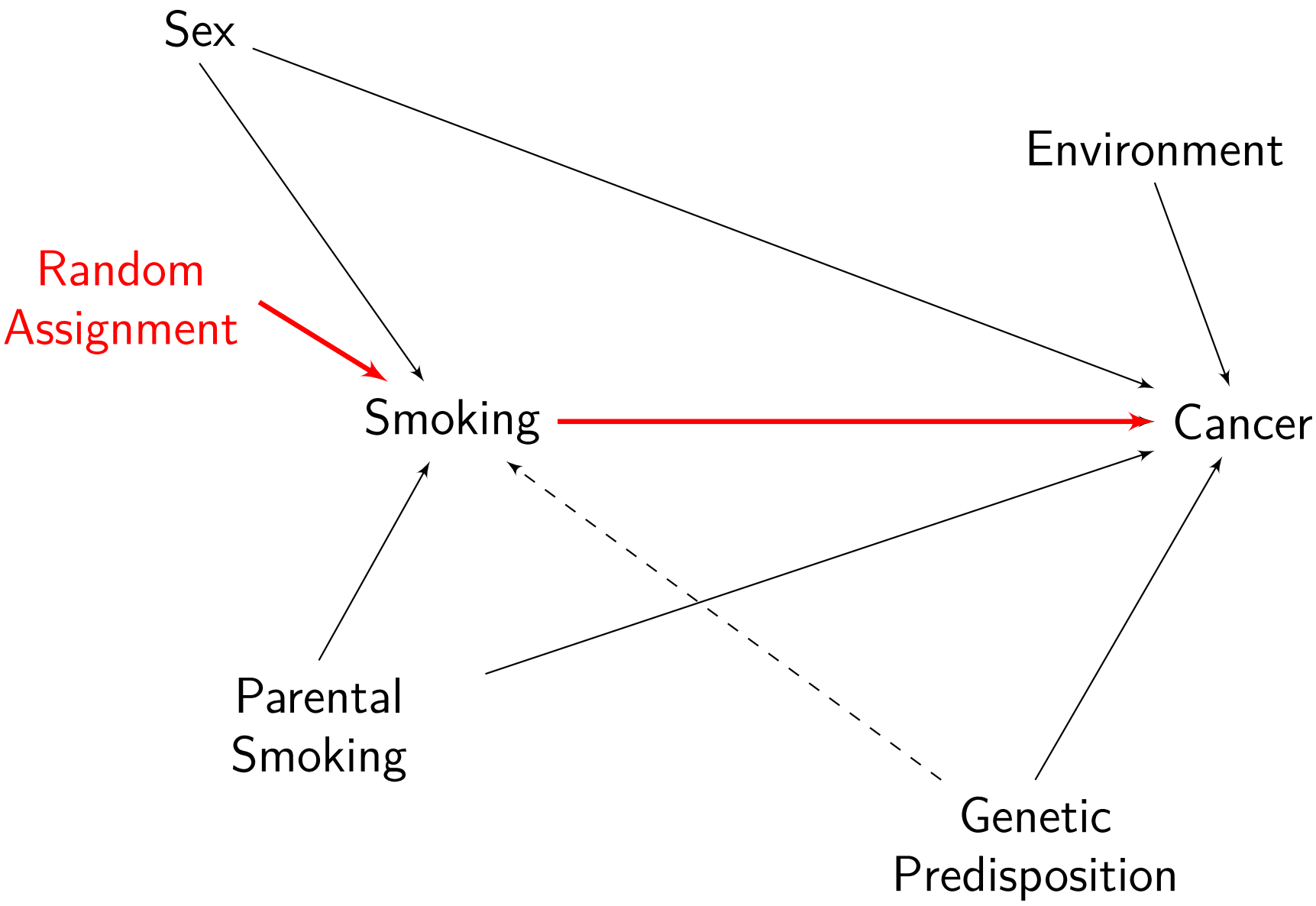
[5] Random means "known probability of treatment" not "haphazard".

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[5]
- 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)

[5] 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

Let's work in Stata!

(Balance testing!)

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:

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

[6] 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*[6]

[6] 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?

Let's work in Stata!

(Basic analysis)

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 → 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

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: $SE = 0.025$

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

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

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

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.[7]

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

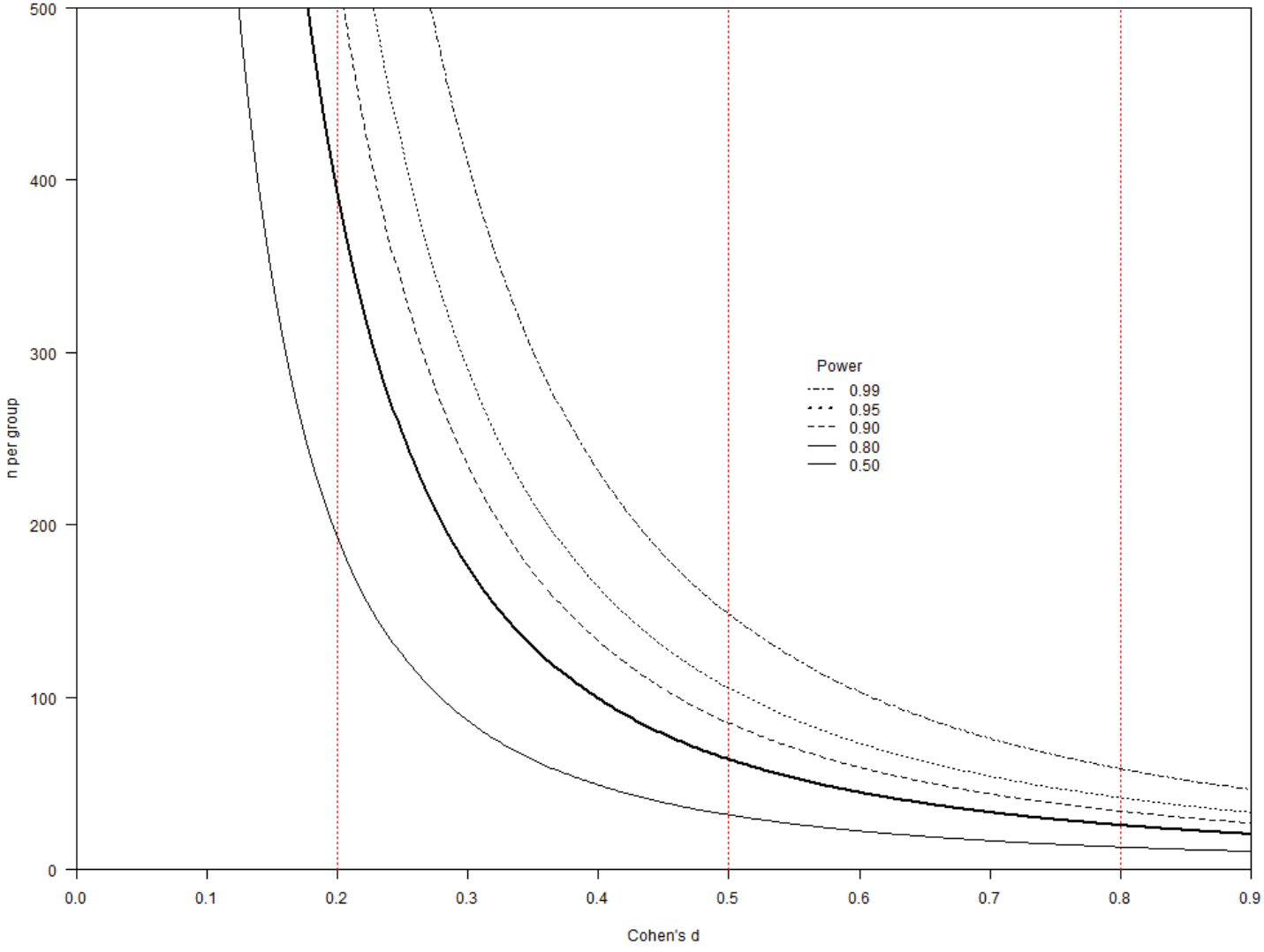
Intuition about Power

- 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

Let's work in Stata!

(Power Analysis)

Intuition about Power



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): $w_{ij} = \frac{1}{p_{ij}}$
  - Weight (Control): $w_{ij} = \frac{1}{1 - p_{ij}}$

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"

Questions?

What kinds of questions can we answer with (survey) experiments?

What kinds of questions can we answer with (survey) experiments?

- Forward causal questions
  - Can X cause Y?
  - What effects does X have?

What kinds of questions can we answer with (survey) 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?

What kinds of questions can we answer with (survey) 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

Hypothesis Testing

- From theory, we derive testable hypotheses
  - Hypotheses are expectations about differences in outcomes across levels of a putatively causal variable
  - Hypothesis must be testable by an SATE
- Manipulations are developed to create variation in that causal variable

Example: News Framing

- Theory: Presentation of news affects opinion
- Hypotheses:
  - News emphasizing free speech increases support for a hate group rally
  - News emphasizing public safety decreases support for a hate group rally
- Manipulation:
  - Control group: no information
  - Free speech group: article emphasizing rights
  - Public safety group: article emphasizing safety

Example: Partisan Identity

- Theory: Strength of partisan identity affects tendency to accept party position
- Hypotheses:
  - Strong partisans are more likely to accept their party's position on an issue
- Manipulation:
  - Control group: no manipulation
  - "Univalent" condition
  - "Ambivalent" condition

Univalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **your party**. Then think of 2 to 3 things you especially dislike about **the other party**. Now please write those thoughts in the space below.

Ambivalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **the other party**. Then think of 2 to 3 things you especially dislike about **your party**. Now please write those thoughts in the space below.

Treatments Test Hypotheses!

Treatments Test Hypotheses!

- Derive experimental design from hypotheses

Treatments Test Hypotheses!

- Derive experimental design from hypotheses
- Experimental "factors" are expressions of hypotheses as randomized groups

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

Ex.: Presence/Absence

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to a negative advertisement criticizing a party reduces support for that party.
- Manipulation:
  - Control group receives no advertisement.
  - Treatment group watches a video containing a negative ad describing a party.

Ex.: Levels/doses

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to higher levels of negative advertising criticizing a party reduces support for that party.
- Manipulation:
  - Control group receives no advertisement.
  - Treatment group 1 watches a video containing 1 negative ad describing a party.
  - Treatment group 2 watches a video containing 2 negative ads describing a party.
  - Treatment group 3 watches a video containing 3 negative ads describing a party.
  - etc.

Ex.: Qualitative variation

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to a negative advertisement criticizing a party reduces support for that party, while a positive advertisement has no effect.
- Manipulation:
  - Control group receives no advertisement.
  - Negative treatment group watches a video containing a negative ad describing a party.
  - Positive treatment group watches a video containing a positive ad describing a party.

Questions?

Activity!

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

Some possible criteria

- Significant results
- Face validity
- Coherent for respondents
- Non-obvious to respondents
- Simple
- Indirect/unobtrusive
- Validated by prior work
- Innovative/creative
- ...

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

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 (29 January 2017)

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

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

- Outcomes are affected consistent with theory

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*)

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*

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*

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*

I. Outcomes Affected

- Follows a circular logic!
- Doesn't tell us anything if we hypothesize null effects

II. Pilot Testing

- Goal: establish construct validity of manipulation
- Assess whether a set of possible manipulations affect a measure of the *independent* variable

II. Pilot Testing

- Goal: establish construct validity of manipulation
- Assess whether a set of possible manipulations affect a measure of the *independent* variable
- Example:
  - Goal: Manipulate the "strength" of an argument
  - Write several arguments
  - Ask pilot test respondents to report how strong each one was

III. Manipulation Checks

- Manipulation checks are items added post-treatment, post-outcome that assess whether the *independent* variable was affected by treatment
- We typically talk about manipulations as directly setting the value of X, but in practice we are typically manipulating something *that we think* strongly modifies X

III. Manipulation Checks

- Manipulation checks are items added post-treatment, post-outcome that assess whether the *independent* variable was affected by treatment
- We typically talk about manipulations as directly setting the value of X, but in practice we are typically manipulating something *that we think* strongly modifies X
- Example: information manipulations aim to modify knowledge or beliefs, but are necessarily imperfect at doing so

Manipulation check example[8]

1. Treatment 1: Supply Information
2. Manipulation check 1: measure beliefs
3. Treatment 2: Prime a set of considerations
4. Outcome: Measure opinion
5. Manipulation check 2: measure dimension salience

[8] Leeper & Slothuus. n.d. "Can Citizens Be Framed?" Available from: `http://thomasleeper.com/research.html`.

Some Best Practices

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable
- Generally, measure post-outcome

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable
- Generally, measure post-outcome
- Measure both what you wanted to manipulate *and* what you didn't want to manipulate
  - Most treatments are *compound*!

IV. Placebos

- Include an experimental condition that *does not* manipulate the variable of interest (but might affect the outcome)

IV. Placebos

- Include an experimental condition that *does not* manipulate the variable of interest (but might affect the outcome)
- Example:
  - Study whether risk-related arguments about climate change increase support for a climate change policy
  - Placebo condition: control article with risk-related arguments about non-environmental issue (e.g., terrorism)

V. Non-equivalent outcomes

- Measures an outcome that *should not* be affected by independent variable

V. Non-equivalent outcomes

- Measures an outcome that *should not* be affected by independent variable
- Example:
  - Assess effect of some treatment on attitudes toward group A
  - Focal outcome: attitudes toward group A
  - Non-equivalent outcome: attitudes toward group B

Aside: Demand Characteristics

- "Demand characteristics" are features of experiments that (unintentionally) imply the purpose of the study and thereby change respondents' behavior (to be consistent with theory)

[9] But, consider the ethics of not doing so

Aside: Demand Characteristics

- "Demand characteristics" are features of experiments that (unintentionally) imply the purpose of the study and thereby change respondents' behavior (to be consistent with theory)
- Implications:
  - Design experimental treatments that are non-obvious
  - Do not disclose the purpose of the study up front[9]
  - Be careful about using manipulation checks and pre-outcome measures

[9] But, consider the ethics of not doing so

Question Wording Designs

- Kahneman and Tversky used a lot of "question wording" experiments
- Hypothesized difference in outcomes according to the decision being faced
  - Risky or not risky
  - Gains or losses
- Manipulation operationalizes this by asking two different questions
- Outcome is the answer to the question

"Framing" or "Priming" Experiments

Example: Schuldt et al. "'Global Warming' or 'Climate Change'? Whether the Planet is Warming Depends on Question Wording."

What's this study about?

You may have heard about the idea that the world's temperature may have been **going up** over the past 100 years, a phenomenon sometimes called **global warming**. What is your personal opinion regarding whether or not this has been happening?

- Definitely has not been happening
- Probably has not been happening
- Unsure, but leaning toward it has not been happening
- Not sure either way
- Unsure, but leaning toward it has been happening
- Probably has been happening
- Definitely has been happening

You may have heard about the idea that the world's temperature may have been **changing** over the past 100 years, a phenomenon sometimes called **climate change**. What is your personal opinion regarding whether or not this has been happening?

- Definitely has not been happening
- Probably has not been happening
- Unsure, but leaning toward it has not been happening
- Not sure either way
- Unsure, but leaning toward it has been happening
- Probably has been happening
- Definitely has been happening

Another framing example[10]

Today, tests are being developed that make it possible to detect serious genetic defects **before a baby is born**. But so far, it is impossible either to treat or to correct most of them. If (you/your partner) were pregnant, would you want (her) to have a test to find out if the **baby** has any serious genetic defects? (Yes/No)

Suppose a test shows the **baby** has a serious genetic defect. Would you, yourself, want (your partner) to have an abortion if a test shows the **baby** has a serious genetic defect? (Yes/No)

[10] Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example[10]

Today, tests are being developed that make it possible to detect serious genetic defects **in the fetus during pregnancy**. But so far, it is impossible either to treat or to correct most of them. If (you/your partner) were pregnant, would you want (her) to have a test to find out if the **fetus** has any serious genetic defects? (Yes/No)

Suppose a test shows the **fetus** has a serious genetic defect. Would you, yourself, want (your partner) to have an abortion if a test shows the **fetus** has a serious genetic defect? (Yes/No)

[10] Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example[11]

Do you favor or oppose the death penalty for persons convicted of murder?

[11] Bobo & Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." Du Bois Review 1(1): 151–180.

Another framing example[11]

Blacks are about 12% of the U.S. population, but they were half of the homicide offenders last year. Do you favor or oppose the death penalty for persons convicted of murder?

[11] Bobo & Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." Du Bois Review 1(1): 151–180.

Another framing example[12]

Concealed handgun laws have recently received national attention. Some people have argued that law-abiding citizens have the right to protect themselves. What do you think about concealed handgun laws?

[12] Haider-Markel & Joslyn. 2001. "Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames." *Journal of Politics* 63(2): 520–543.

Another framing example[12]

Concealed handgun laws have recently received national attention. Some people have argued that laws allowing citizens to carry concealed handguns threaten public safety because they would allow almost anyone to carry a gun almost anywhere, even onto school grounds. What do you think about concealed handgun laws?

[12] Haider-Markel & Joslyn. 2001. "Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames." *Journal of Politics* 63(2): 520–543.

Question testing

Use question wording designs to select which survey measures we want to use

- Select possible question wordings
- Select some criterion(-ia) for assessing which is better
- Pilot test and then use the item that performs better

Aside: Experimentation vs. Other Pretesting Methods

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods
- Specific value added of an experiment: optimize questions or other survey features against a specific criterion, e.g.:
  - (Non-)Response or drop-off rates
  - "Don't know" rates
  - Item characteristics
  - Reading times or response latencies

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods
- Specific value added of an experiment: optimize questions or other survey features against a specific criterion, e.g.:
  - (Non-)Response or drop-off rates
  - "Don't know" rates
  - Item characteristics
  - Reading times or response latencies
- But! Power considerations...

Classic question testing experiment[13]

Some people feel that The 1975 Public Affairs Act should be repealed-do you agree or disagree with this idea?

[13] Bishop, G.F., Tuchfarber, A. & Oldendick, R.W. 1986. "Opinions on Fictitious Issues: The Pressure to Answer Survey Questions." *Public Opinion Quarterly* 50(2): 240–250.

Classic question testing experiment[13]

Some people feel that The 1975 Public Affairs Act should be repealed-do you agree or disagree with this idea, or haven't you thought much about this issue?

[13] Bishop, G.F., Tuchfarber, A. & Oldendick, R.W. 1986. "Opinions on Fictitious Issues: The Pressure to Answer Survey Questions." *Public Opinion Quarterly* 50(2): 240–250.

An example[14]

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you–did you vote in the elections this November?

[14] Holbrook & Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77: 106–123.

An example[14]

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. Which of the following statements best describes you?

- One, I did not vote in the November 3 election
- two, I thought about voting this time but didn't
- three, I usually vote but didn't this time
- four, I am sure I voted

[14] Holbrook & Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77: 106–123.

An Instructional Manipulation[15]

For the next few questions, I am going to read out some statements, and for each one, please tell me if it is true or false. If you don't know, just say so and we will skip to the next one.

1. Britain's electoral system is based on proportional representation.
2. MPs from different parties are on parliamentary committees.
3. The Conservatives are opposed to the ratification of a constitution for the European Union.

[15] Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation[15]

For the next few questions, I am going to read out some statements, and for each one, please tell me if it is true or false. If you don't know, please just give me your best guess.

1. Britain's electoral system is based on proportional representation.
2. MPs from different parties are on parliamentary committees.
3. The Conservatives are opposed to the ratification of a constitution for the European Union.

[15] Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation + [16]

In the next part of this study, you will be asked 14 questions about politics, public policy, and economics. Many people don't know the answers to these questions, but it is helpful for us if you answer, even if you're not sure what the correct answer is. We encourage you to take a guess on every question. At the end of this study, you will see a summary of how many questions you answered correctly.

[16] Prior & Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American journal of Political Science* 52(1): 169–183.

An Instructional Manipulation +[16]

We will pay you for answering questions correctly. You will earn $1 for every correct answer you give. So, if you answer 3 of the 14 questions correctly, you will earn $3. If you answer 7 of the 14 questions correctly, you will earn $7. The more questions you answer correctly, the more you will earn.

[16] Prior & Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American journal of Political Science* 52(1): 169–183.

Question Order Designs

- Manipulation of pre-outcome questionnaire

Question Order Designs

- Manipulation of pre-outcome questionnaire
- Example:
  - Goal: assess influence of value salience on support for a policy
  - Manipulate by asking different questions:
    - Battery of 5 "rights" questions, or
    - Battery of 5 "life" questions
  - Measure support for legalized abortion

Question Order Designs

- Manipulation of pre-outcome questionnaire
- Example:
  - Goal: assess influence of value salience on support for a policy
  - Manipulate by asking different questions:
    - Battery of 5 "rights" questions, or
    - Battery of 5 "life" questions
  - Measure support for legalized abortion
- If answers to manipulated questions matter, can measure rest post-outcome

Ex. Question-as-treatment[17]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[17] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[17]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[17] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[17]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[17] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[17]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[17] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex.: Knowledge and Political Interest

1. Do you happen to remember anything special that your U.S. Representative has done for your district or for the people in your district while he has been in Congress?
2. Is there any legislative bill that has come up in the House of Representatives, on which you remember how your congressman has voted in the last couple of years?
3. Now, some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say that you follow what's going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all?

Ex.: Knowledge and Political Interest

1. Now, some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say that you follow what's going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all?
2. Do you happen to remember anything special that your U.S. Representative has done for your district or for the people in your district while he has been in Congress?
3. Is there any legislative bill that has come up in the House of Representatives, on which you remember how your congressman has voted in the last couple of years?

Vignettes

- A "vignette" is a short paragraph of text describing a situation
- Vignettes are probably the most common survey experimental paradigm, after question wording designs
- Take many forms and increasingly encompass non-textual stimuli
- Basically limited to web-based mode

A classic vignette[18]

Now think about a **(black/white)** woman in her early thirties. She is a high school **(graduate/drop out)** with a ten-year-old child, and she has been on welfare for the past year.

- How likely is it that she will have more children in order to get a bigger welfare check? (1 = Very likely, . . . , 7 = Not at all likely)
- How likely do you think it is that she will really try hard to find a job in the next year? (1 = Very likely, . . . , 7 = Not at all likely)

[18] Gilens, M. 1996. "'Race coding' and white opposition to welfare. *American Political Science Review* 90(3): 593–604.

Newer vignette[19]

Imagine that you were living in a village in another district in Uttar Pradesh and that you were voting for candidates in **(village/state/national)** election. Here are the two candidates who are running against each other: The first candidate is named **(caste name)** and is running as the **(BJP/SP/BSP)** party candidate. **(Corrupt/criminality allegation)**. His opponent is named **(caste name)** and is running as the **(BJP/SP/BSP)** party candidate. **(Opposite corrupt/criminality allegation)**. From this information, please indicate which candidate you would vote for in the **(village/state/national)** election.

[19] Banerjee et al. 2012. "Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India." Working paper.

Longer texts[20]

We are testing materials for use in a study **of the structure of sentences people use when writing news editorials**. Along these lines, we would like you to read a series of paragraphs, taken from recent major newspaper editorials.

[20] Druckman & Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–896.

Longer texts[20]

We are testing materials for use in a study **that is related to the kinds of opinions people form about public policies**. Along these lines, we would like you to read a series of paragraphs, taken from recent major newspaper editorials.

[20] Druckman & Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–896.

Please read the following paragraphs and, for each, rate **how *dynamic* you think it is. A paragraph is more "dynamic" when it uses more vivid action words. For example, a statement like, "He sped up and raced through the light before crashing into the swerving truck," seems more dynamic than, "He went faster to get through the light before having an accident." The action words in the first sentence (which we have highlighted in bold) seem more dynamic or vivid than those contained in the second sentence.** There are no right or wrong opinions and your responses to all questions are completely confidential.

Please read the following paragraphs and, for each, rate **the extent to which it decreases or increases your support for the Patriot Act. In subsequent surveys we will ask you for your overall opinion about the state-run casino (i.e., the extent to which you oppose or support the state-run casino).** There are no right or wrong opinions and your responses to all questions are completely confidential.

Please read the paragraphs carefully and, after each one, rate **the extent to which you think it is *dynamic***.

With the passage of the Patriot Act in 2001, the FBI can now enter your home, search around, and doesn't ever have to tell you it was there. You could be perfectly innocent, yet federal agents can go through your most personal effects. When considering new laws, a test of the impact on liberty should be required. On that test, the Patriot Act fails. At a massive 342 pages, it potentially violates at least six of the ten original amendments known as the Bill of Rights — the First, Fourth, Fifth, Sixth, Seventh and Eighth Amendments — and possibly the Thirteenth and Fourteenth as well.

Please read the paragraphs carefully and, after each one, rate **the extent to which it decreases or increases your support for the Patriot Act**.

With the passage of the Patriot Act in 2001, the FBI can now enter your home, search around, and doesn't ever have to tell you it was there. You could be perfectly innocent, yet federal agents can go through your most personal effects. When considering new laws, a test of the impact on liberty should be required. On that test, the Patriot Act fails. At a massive 342 pages, it potentially violates at least six of the ten original amendments known as the Bill of Rights — the First, Fourth, Fifth, Sixth, Seventh and Eighth Amendments — and possibly the Thirteenth and Fourteenth as well.

Example[21]

Fears of Future Terror Attacks Warranted

By Andrew Tardaca
Published: January 17, 2009

U.S. citizens are bracing for another 9/11 type terrorist attack, according to a variety of reports. A recent Gallup poll finds that 87% of the American public is highly concerned about the possibility of a terrorist attack at home. According to new information from several international sources, these fears are well supported.

A raid on a London terrorist hideout on November 9, 2008 resulted in the capture of computer files that identified numerous U.S. financial districts, cultural centers, and transportation systems on a list of future Al Qaeda targets. According to a recent overseas intelligence report, "al Qaeda already has several cells operating in the U.S. that may be on the verge of mounting a large-scale terrorist attack."

On September 11, 2001, Al Qaeda's attacks killed nearly 3,000 men, women, and children, and injured over 6,000 more. Since September 11th, Al Qaeda and groups affiliated with Al Qaeda have waged attacks in countries such as Egypt, Indonesia, Kenya, Morocco, Saudi Arabia, Spain, Turkey, the United Kingdom, and most recently India. U.S. security officials are warning that current terrorist plots include plans for attacks on U.S. soil at least twice the magnitude of 9/11. An anonymous source reported that recent intelligence documents contain "sobering information" concerning the magnitude of future terrorist attacks.

Warnings issued by extremist groups such as Al Qaeda to "attack U.S. interests and allies on its soil" are even more alarming given the state of preparedness for future incidents. Experts have issued warnings about

[21] Merolla & Zechmeister. 2013. "Evaluating Political Leaders in Times of Terror and Economic Threat: The Conditioning Influence of Politician Partisanship." *Journal of Politics* 75(3): 599–712.

Example[21]

Economic Recession Projected to Deepen

By Andrew Tardaca
Published: January 17, 2009

U.S. citizens are bracing for a drastic deepening of the current economic recession. A recent Gallup poll finds that 87% of the American public is highly concerned about economic conditions in the country. The report further states "The economic mood is grimmer than it has been since 1992."

On September 16, failures of large financial institutions in the United States, such as Lehman Brothers and AIG, rapidly evolved into a global crisis resulting in bank failures across the U.S. and Europe. In the United States alone, 15 banks failed in 2008, while several others were rescued through government intervention or acquisitions by other banks. These events led to sharp reductions in the value of stocks and commodities worldwide. Over the past year, the Dow Jones Industrial Average lost 33.8%, the third worst loss in our nation's history. On October 11, 2008, the head of the International Monetary Fund (IMF) warned that the world financial system is teetering on the "brink of systemic meltdown".

The bank failures and subsequent market collapse were tied to sub-prime loans and credit default swaps. Increasing interest rates on loans hit the housing market particularly hard, as individuals were unable to keep up with mortgage payments. 2008 witnessed a record number of foreclosures, leading to the worst housing crisis, banking failure, and market collapse since the Great Depression.

Future projections are looking even grimmer. Experts predict that the housing market will not recover for at least a decade, especially now that banks are hesitant to make loans. The downturn in the economy has led to

[21] Merolla & Zechmeister. 2013. "Evaluating Political Leaders in Times of Terror and Economic Threat: The Conditioning Influence of Politician Partisanship." *Journal of Politics* 75(3): 599–712.

Some vignette considerations

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability

Some vignette considerations

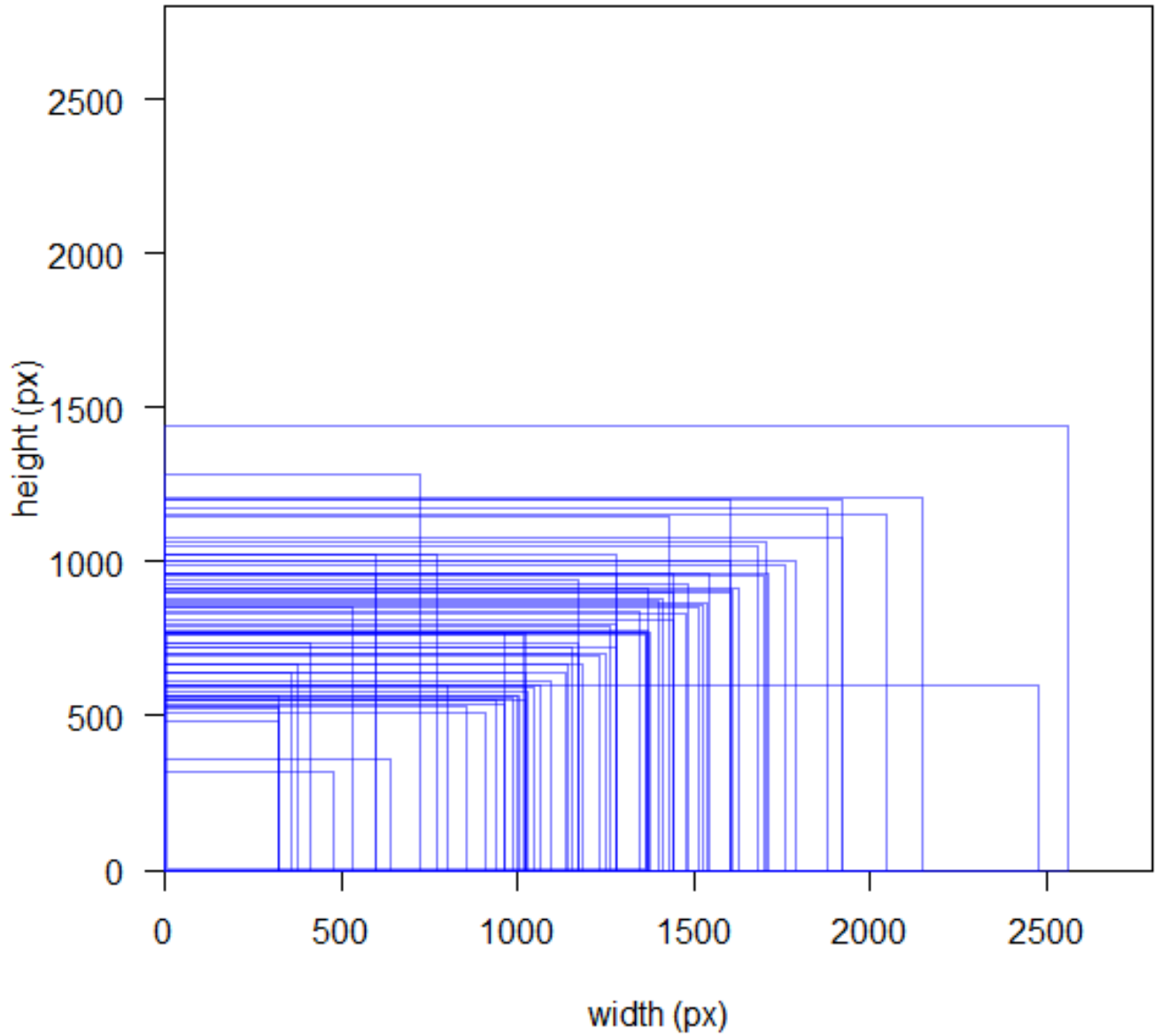
- Comparability across conditions
  - Length
  - Readability
- Language proficiency

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability
- Language proficiency
- Length
  - Timers
  - Forced exposure
  - Mouse trackers

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability
- Language proficiency
- Length
  - Timers
  - Forced exposure
  - Mouse trackers
- Devices
  - Browser-specificity
  - Device sizes (e.g., mobile)


height (px)
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
width (px)

Aside: Unique features of online studies

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.
- Complex randomization

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.
- Complex randomization
- Panel data
- Synchronous, multi-person designs

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question

[22] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[22] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags

[22] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[22] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags
  - Subliminal primes possible, depending on software

[22] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[22] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags
  - Subliminal primes possible, depending on software
  - Lots of recent examples of facial manipulation

[22] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Example[23]

Light Complexion

Original

Dark Complexion

[23] Iyengar et al. 2010. "Do Explicit Racial Cues Influence Candidate Preference? The Case of Skin Complexion in the 2008 Campaign." Working paper.

Example[24]



[24] Laustsen & Petersen. 2016. "Winning Faces vary by Ideology." *Political Communication* 33(2): 188–211.

Example[25]

[25] Bailenson et al. 2006. "Transformed Facial Similarity as a Political Cue: A Preliminary Investigation." *Political Psychology* 27(3): 373–385.

Audio & Video manipulations

- Problematic for same reasons as long texts

[26] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[27] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

Audio & Video manipulations

- Problematic for same reasons as long texts
- Best practices
  - Keep it short
  - Have the video play automatically
  - Disallow survey progression
  - Control and validate

[26] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[27] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

Audio & Video manipulations

- Problematic for same reasons as long texts
- Best practices
  - Keep it short
  - Have the video play automatically
  - Disallow survey progression
  - Control and validate
- Examples
  - Television Advertisements[26]
  - News Programs[27]

---

[26] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[27] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings
- Most common example: writing something

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings
- Most common example: writing something
- Can be problematic:
  - Time-intensive
  - Invites drop-off
  - Compliance problems

Univalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **your party**. Then think of 2 to 3 things you especially dislike about **the other party**. Now please write those thoughts in the space below.

Ambivalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **the other party**. Then think of 2 to 3 things you especially dislike about **your party**. Now please write those thoughts in the space below.

Questions?

Sensitive Item Designs

- Experiments can also be used to measure something
- Goal here is not necessarily causal inference, though the causal insight of the experiment provides *descriptively* useful information
- Paradigms
  - List experiments
  - Endorsement experiments

List Experiments [28]

Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me *how many* of them upset you. I don't want to know which ones. just *how many*.

1. the federal government increasing the tax on gasoline
2. professional athletes getting million-dollar salaries
3. large corporations polluting the environment

[28] Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402–419.

List Experiments [28]

Now I'm going to read you three things that sometimes make people angry or upset. After I read all four, just tell me *how many* of them upset you. I don't want to know which ones. just *how many*.

1. the federal government increasing the tax on gasoline
2. professional athletes getting million-dollar salaries
3. large corporations polluting the environment
4. **a black family moving in next door**

[28] Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402–419.

Endorsement experiments[29]

A recent proposal calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

[29] Lyall, Blair, & Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4): 679–705.

Endorsement experiments[29]

A recent proposal **by the Taliban** calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

[29] Lyall, Blair, & Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4): 679–705.

Questions?

Let's work in Stata!

(Analysis of Example Experiments)

Homework!

- Get a sense of what can be studied survey-experimentally
- Visit Time-Sharing Experiments for the Social Sciences
  - `http://tessexperiments.org`
- Pick two studies from TESS
- We will share them in tomorrow

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative
- Representative

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative
- Representative
  - Unrepresentative

Discuss in Pairs!

What does it mean for a "sample" to be representative of a population?

Different conceptualizations

- **Design-based**: A sample is representative because of how it was drawn (e.g., randomly)
- **Model-based**: A sample is representative because it resembles in the population with respect to certain variables (e.g., same proportion of women in sample and population, etc.)
- **Expert judgement**: A sample is representative as judged by an expert who deems it "fit for purpose"

Obtaining Representativeness

Obtaining Representativeness

- Census

Obtaining Representativeness

- Census
- Convenience/Purposive samples

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- Simple random sampling

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- Simple random sampling
- Complex survey designs

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- **Simple random sampling**
- Complex survey designs

Inference from Sample to Population

- We want to know pop. parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

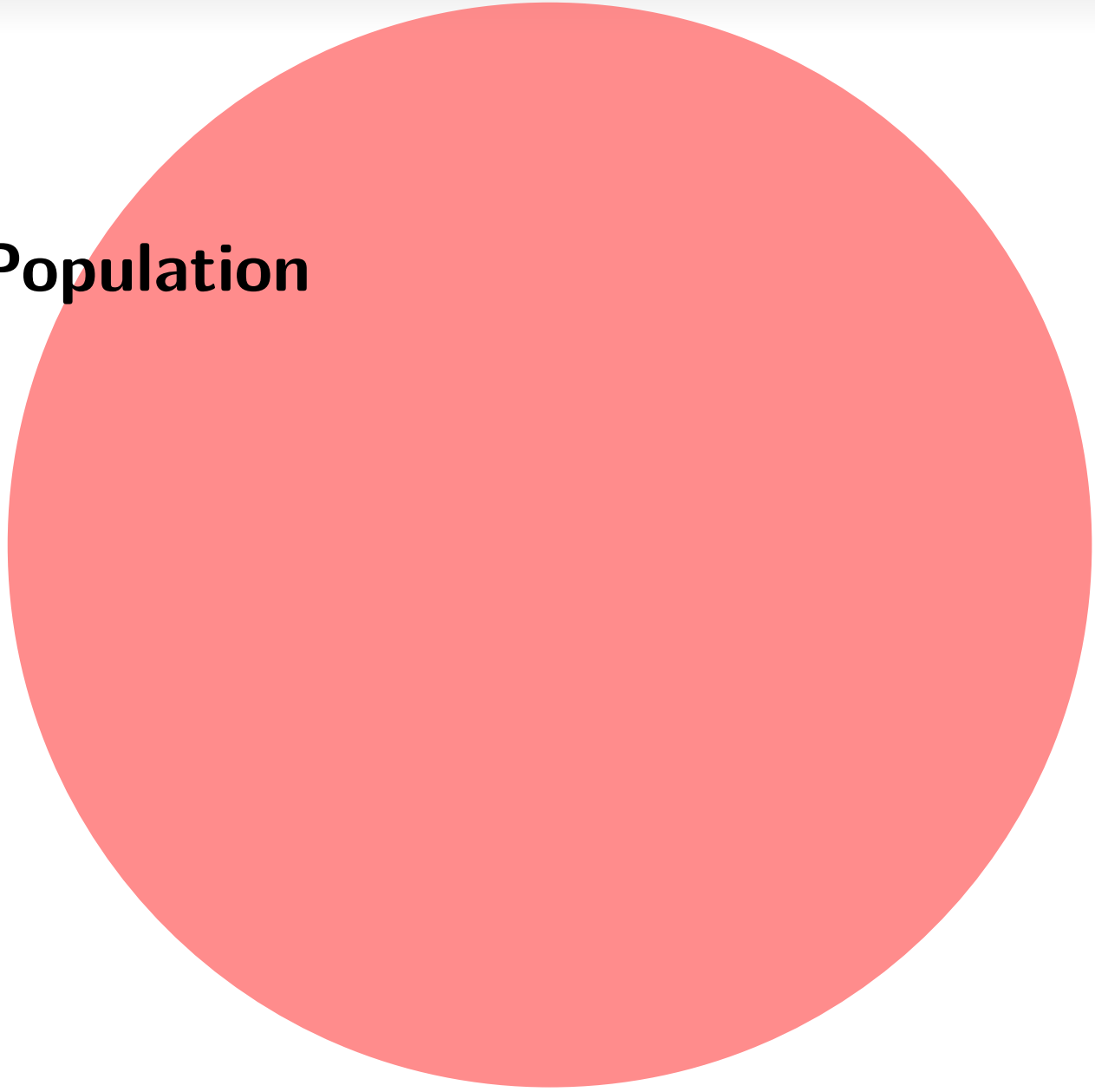
Inference from Sample to Population

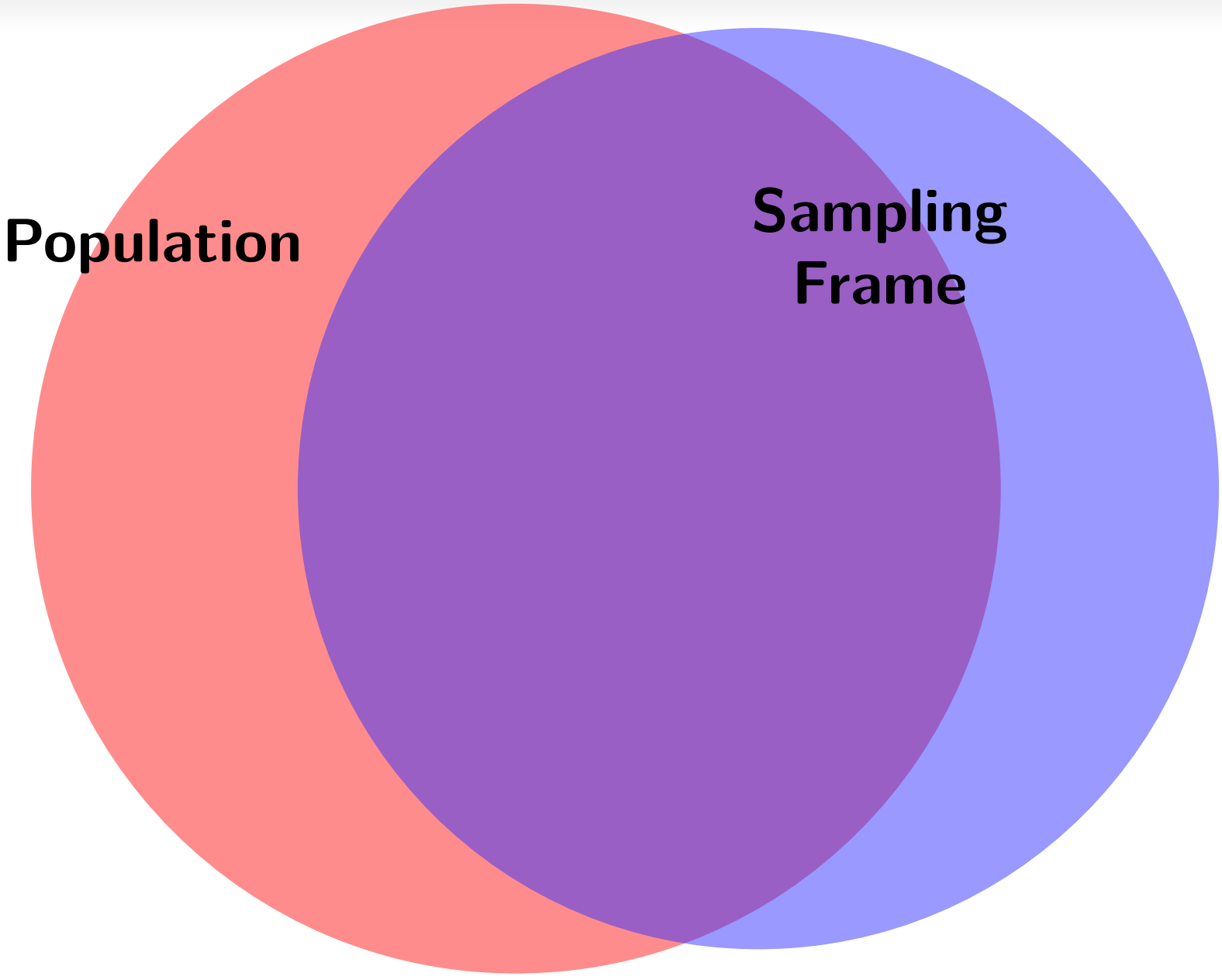
- We want to know pop. parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

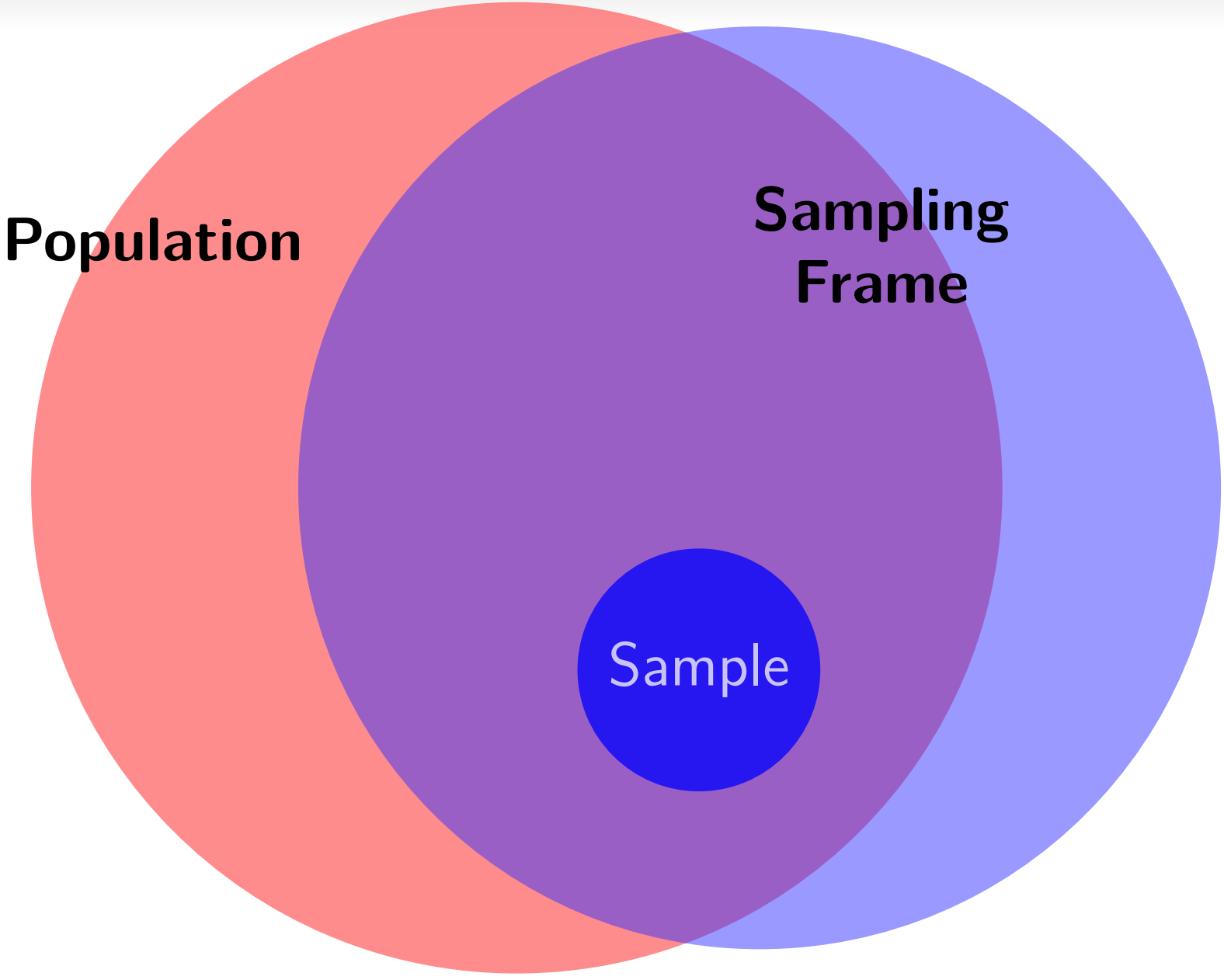
- What range of values for θ does our $\hat{\theta}$ imply?

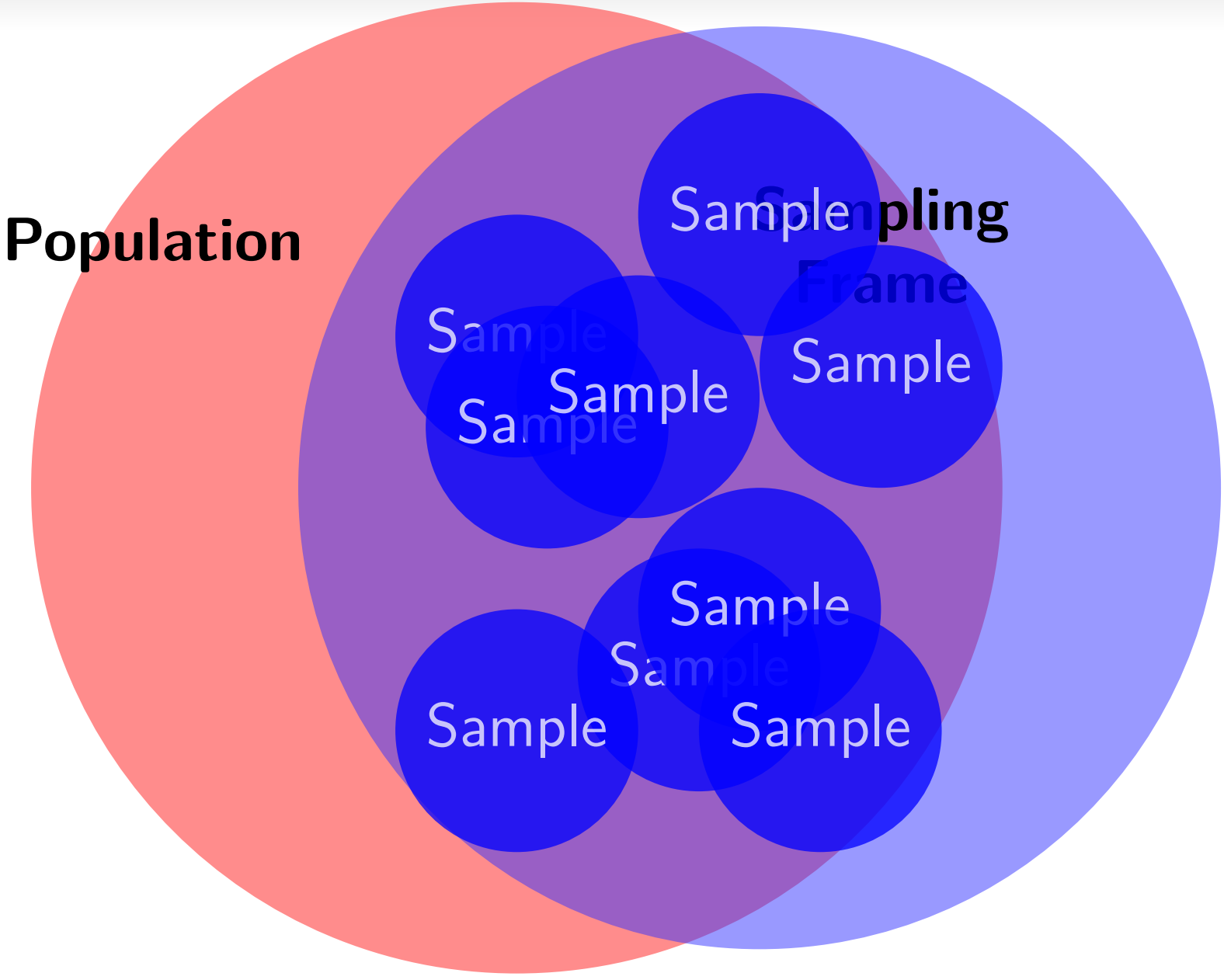
Simple Random Sampling

1. Define target population
2. Create "sampling frame"
3. Each unit in frame has equal probability of selection
4. Collect data on each unit
5. Calculate sample *statistic*
6. Draw an inference to the population


Population


Population
Sampling
Frame


Population
Sampling Frame
Sample


Population
Sampling Frame
Sample
Sample
Sample
Sample
Sample
Sample
Sample
Sample

Simple Random Sampling

1. Define target population
2. Create "sampling frame"
3. Each unit in frame has equal probability of selection
4. Collect data on each unit
5. Calculate sample *statistic*
6. Draw an inference to the population

Statistical Inference I

To calculate a sample mean (or proportion):

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{9}$$

where y_i = value for a unit, and

n = sample size

Statistical Inference II

- If we calculate $\bar{y}$ in our *sample*, what does this tell us about the $\bar{Y}$ in the *population*?

Statistical Inference II

- If we calculate $\bar{y}$ in our *sample*, what does this tell us about the $\bar{Y}$ in the *population*?
- The sample *estimate* is our guess at the value of the population *parameter* within some degree of uncertainty

Law of Large Numbers

- Definition: The *mean* of the $\hat{\theta}$ from each of a number of samples will converge on the population θ, as the number of samples increases

Sampling Variance

- The $\hat{\theta}$ in any particular sample can differ from the population value θ
- This variation is calling "sampling variance" or "sampling error"
- The standard error describes the average amount of variation of the $\hat{\theta}$'s around θ

How Uncertain Are We?

- Our uncertainty depends on sampling procedures
- Most importantly, *sample size*
  - As $n \to \infty$, uncertainty $\to 0$
- We typically summarize our uncertainty as the *standard error*

Standard Errors (SEs)

- Definition: "The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each."

Standard Errors (SEs)

- Definition: "The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each."
- Square root of the sampling variance

Sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{10}$$

where y_i = value for a unit, and
n = sample size

$$SE_{\bar{y}} = \sqrt{(1 - f)\frac{s^2}{n}} \tag{11}$$

where f = proportion of population sampled,
s^2 = sample (element) variance, and
n = sample size

SATE

$$\hat{SATE} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i,1} - \frac{1}{n_0}\sum_{i=1}^{n_1} y_{i,0} \tag{12}$$

where $y_{i,1}$ = value for a treatment group unit, and
$y_{i,0}$ = value for a control group unit, and
n_1, n_0 = group sample sizes

$$\widehat{Var}(SATE) = \frac{\widehat{Var}(y_1)}{n_1} + \frac{\widehat{Var}(y_0)}{n_0} \tag{13}$$

where $Var(y_0) = \sum_{i=1}^{n_0}(y_{i,0} - \bar{y}_0)^2$,
$Var(y_1) = \sum_{i=1}^{n_1}(y_{i,1} - \bar{y}_1)^2$, and
n_1, n_0 = group sample sizes

Questions?

Response Rates

- Why do we care?

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias
- Sample size calculations (and design effects) are based on completed interviews

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias
- Sample size calculations (and design effects) are based on completed interviews
- Cost, time, and effort

Response Rates

- Imagine we need $n = 1000$
- How many attempts to obtain that sample:

Response Rate	Needed Attempts
1.00	1000
0.75	1333
0.50	2000
0.25	4000
0.10	10,000

Response Rate

- Interviews divided by eligibles
- $RR = \frac{I}{E}$
- Challenges
  - Unknown eligibility
  - Partial interviews
  - Non-probability samples
  - Complex survey designs
- Cooperation Rate (I's divided by contacts)

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)

What is a refusal?

- How do categorize a respondent as a refusal?

What is a refusal?

- How do categorize a respondent as a refusal?
- When can we try to convert an apparent refusal?

What is a refusal?

- "I don't want to participate."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."
- "How did you get my number?"

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."
- "How did you get my number?"

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)
- Unknowns (U)
- Ineligibles

Eligibility

- Why would an ineligible unit be in our sample?

Eligibility

- Why would an ineligible unit be in our sample?
- How do we determine ineligibility?

Eligibility

- Why would an ineligible unit be in our sample?
- How do we determine ineligibility?
- What do we do with “unknowns”?

Response Rate 1[30]

- $$RR1 = \frac{I}{(I+P)+(R+NC)+U}$$

[30] Note: Simplified slightly

Response Rate 2[31]

- $RR2 = \frac{I+P}{(I+P)+(R+NC)+U}$

[31] Note: Simplified slightly

Response Rates 3 and 4[32]

- $RR3 = \frac{I}{(I+P)+(R+NC)+(e*U)}$
- $RR4 = \frac{I+P}{(I+P)+(R+NC)+(e*U)}$
- e is estimated proportion eligible among unknowns

[32] Note: Simplified slightly

Refusal Rates

- Related to response rate
- Numerator is refusals
- E.g., $REF1 = \frac{R}{(I+P)+(R+NC)+U}$

Complex Survey Designs

- Stratified Sampling (unequal allocation)
  - Sums of codes weighted by $\frac{1}{p}$
  - p is probability of selection
  - May want to report stratum-specific rates
- Multi-stage sampling (e.g., cluster sampling)
  - RR is product of cluster cooperation and within-cluster response rate

Internet Surveys

- For *probability-based samples*, RR is a product of:
  - Recruitment Rate (RR for panel enrollment)
  - Completion Rate (RR for specific survey)
  - Profile Rate (in some cases)
  - E.g., if Recruitment Rate is 30% and Completion Rate is 80%, $RR = 0.3 * 0.8 = 24\%$
- For *non-probability samples*, RR is undefined
  - No sampling involved (so no denominator)
  - If from panel, report Completion Rate
  - If fully opt-in, there's nothing you can do

Differential Response Rates

- Experimental inference breaks if treatment causes breakoff or item nonresponse
- If known in advance, you should:
  - Change the experiment
  - Differential incentives
  - Mode differences
- If discovered in field, you can:
  - Add/modify incentives
  - Refusal conversion

Uses of Regression

1. Description
2. Prediction
3. Causal Inference

Descriptive Inference

1. We want to understand a *population* of cases
2. We cannot observe them all, so:
    1. Draw a *representative* sample
    2. Perform mathematical procedures on sample data
    3. Use assumptions to make inferences about population
    4. Express uncertainty about those inferences based on assumptions

Parameter Estimation

- We want to observe population *parameter* θ
- If we obtain a representative sample of population units:
  - Our sample statistic $\hat{\theta}$ is an unbiased estimate of θ
  - Our sampling procedure dictates how uncertain we are about the value of θ

Causal Inference

Causal Inference

1. Everything that goes into descriptive inference

Causal Inference

1. Everything that goes into descriptive inference
2. Plus, philosophical assumptions

Causal Inference

1. Everything that goes into descriptive inference
2. Plus, philosophical assumptions
3. Plus, randomization *or* perfectly specified model

Relationship

- Covariance:

$$Cov(X, Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Relationship

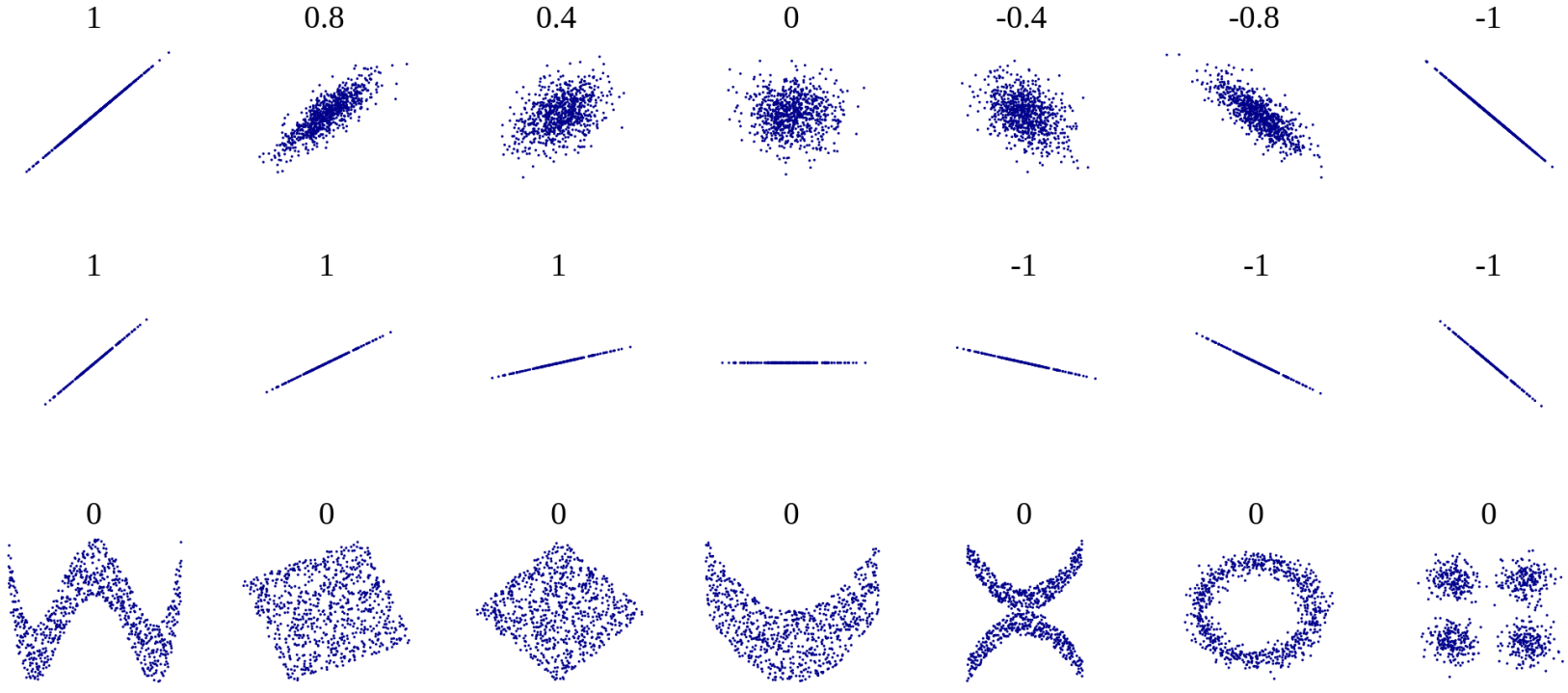
- Covariance:

$$Cov(X, Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- Pearson's Correlation:

$$Corr(X, Y) = r_{x,y} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)s_x s_y}$$

Correlation is linear!



Source: Wikimedia

Analyzing Complex Surveys

- There's a saying: "Every simple random survey is simple in the same way, but every complex survey is complex in its own way."

Analyzing Complex Surveys

- There's a saying: "Every simple random survey is simple in the same way, but every complex survey is complex in its own way."
- Statistics courses will almost always assume simple random sampling

Analyzing Complex Surveys

- There's a saying: "Every simple random survey is simple in the same way, but every complex survey is complex in its own way."
- Statistics courses will almost always assume simple random sampling
- Any sample that is not self-weighting requires more complicated *estimators* that account for varying weights

Analyzing Complex Surveys

- There's a saying: "Every simple random survey is simple in the same way, but every complex survey is complex in its own way."
- Statistics courses will almost always assume simple random sampling
- Any sample that is not self-weighting requires more complicated *estimators* that account for varying weights
- Don't try to do this by hand
  - Stata `svy` module
  - R `survey` package

Ways of Thinking About OLS

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)
4. Line (or surface) of best fit

Bivariate Regression I

- Y is continuous
- X is a randomized treatment indicator/dummy $(0, 1)$
- How do we know if the treatment X had an effect on Y?

Bivariate Regression I

- Y is continuous
- X is a randomized treatment indicator/dummy $(0, 1)$
- How do we know if the treatment X had an effect on Y?
- Look at mean-difference: $E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$

Three Equations

1. Population: $Y = \beta_0 + \beta_1 X \ (+\epsilon)$
2. Sample estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
3. Unit:

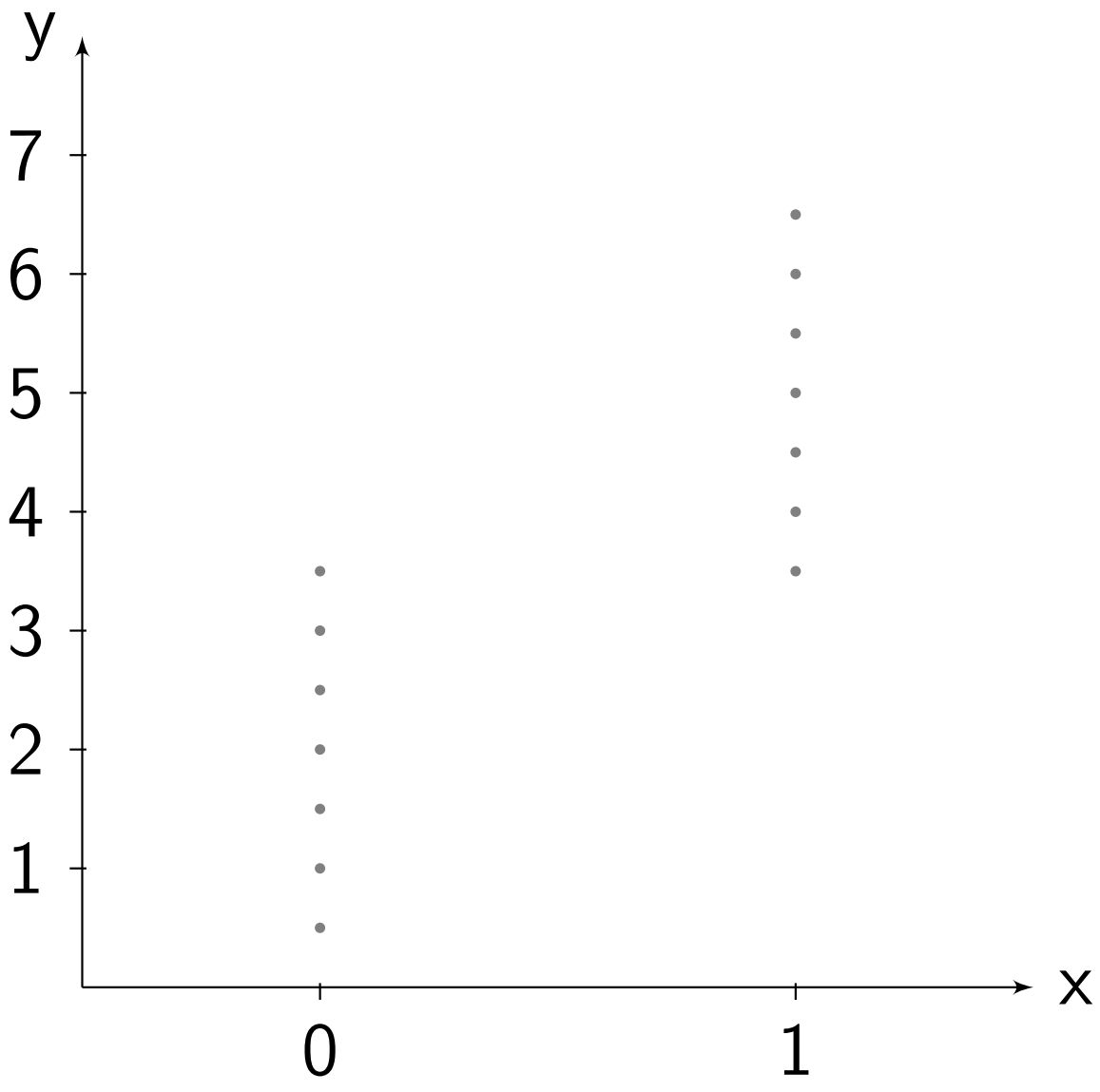
$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \\ &= \bar{y}_{0i} + (y_{1i} - y_{0i})x_i + (y_{0i} - \bar{y}_{0i}) \end{aligned}$$

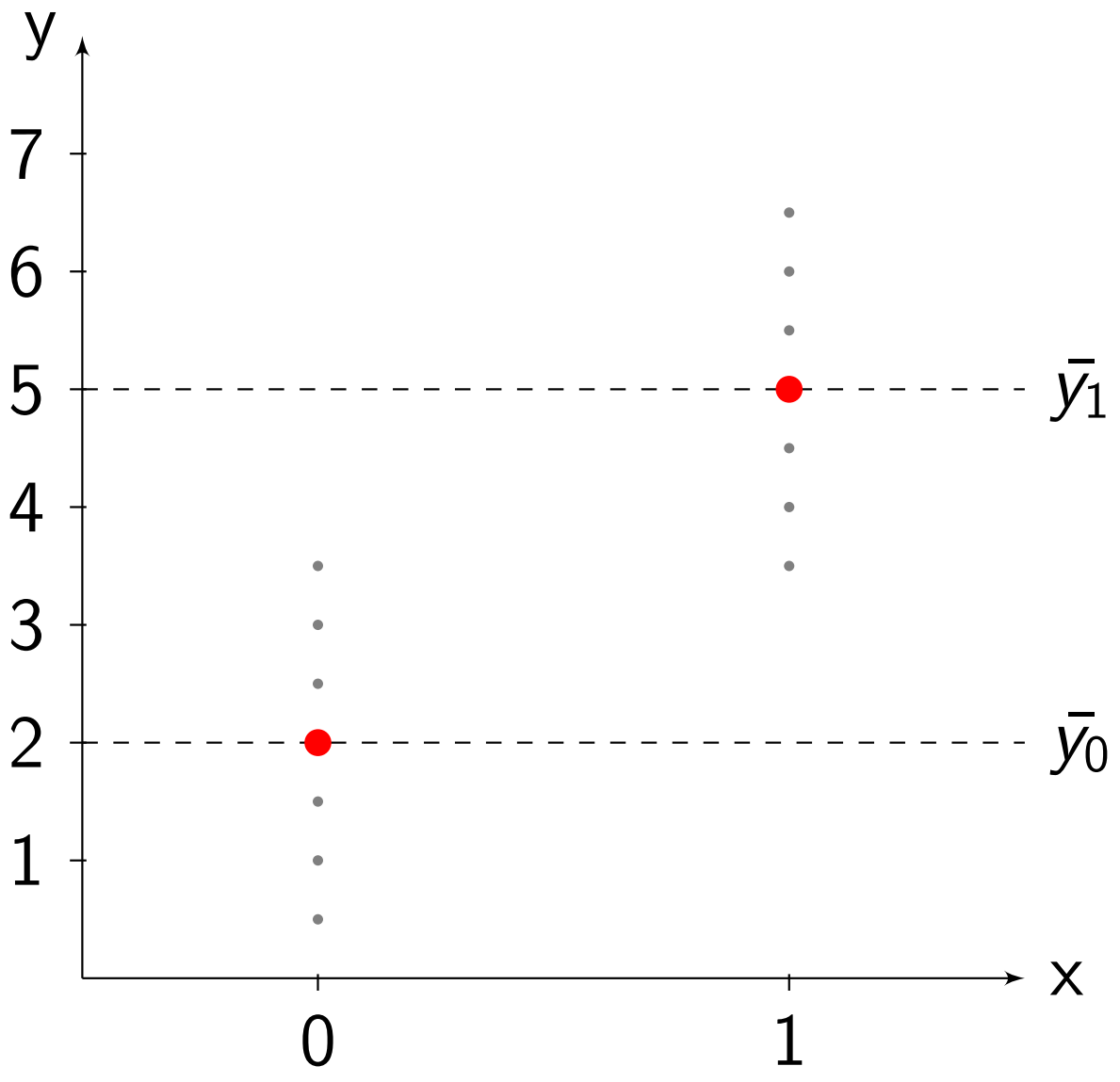
Bivariate Regression I

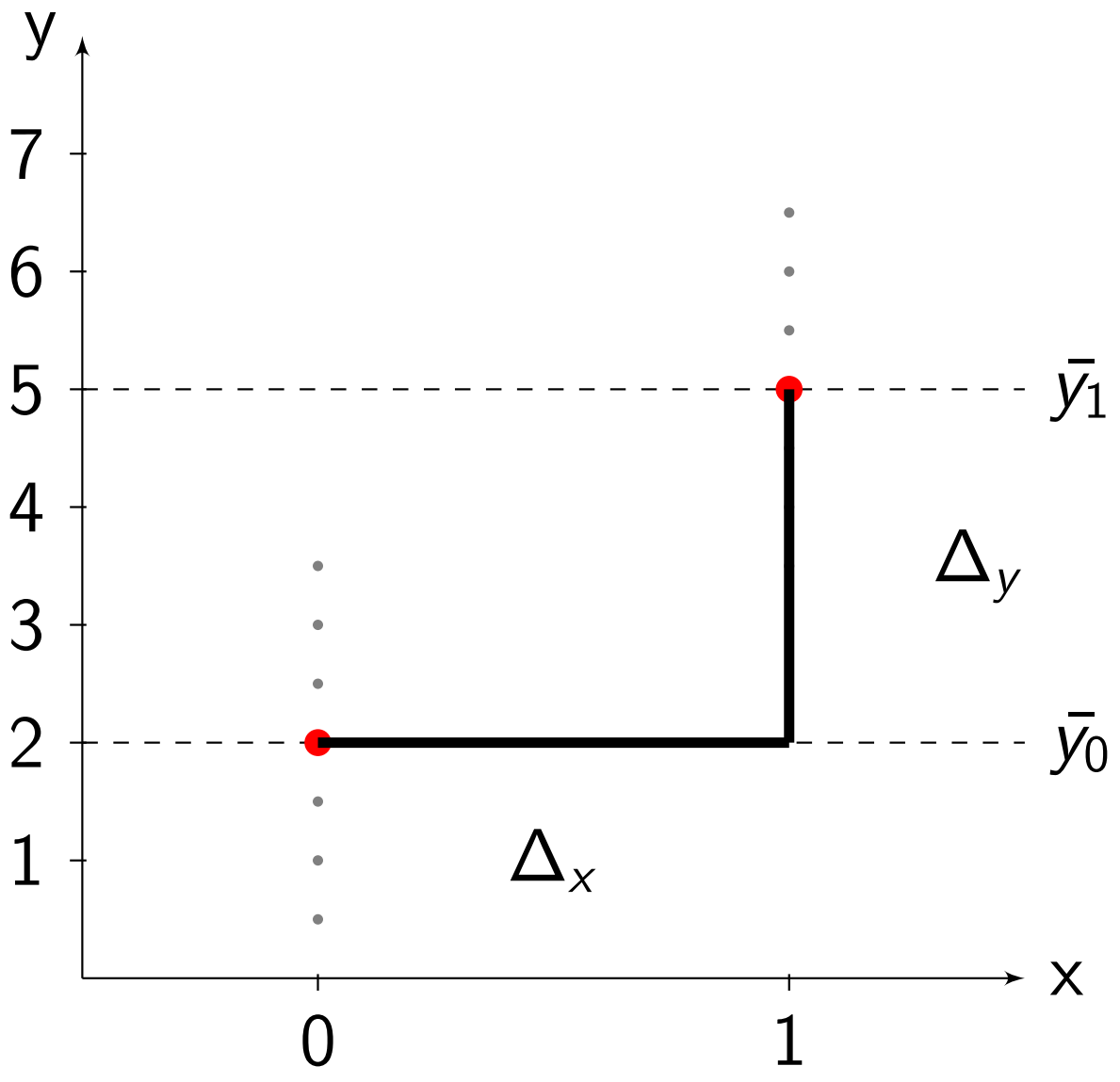
- Mean difference ($E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$) is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$

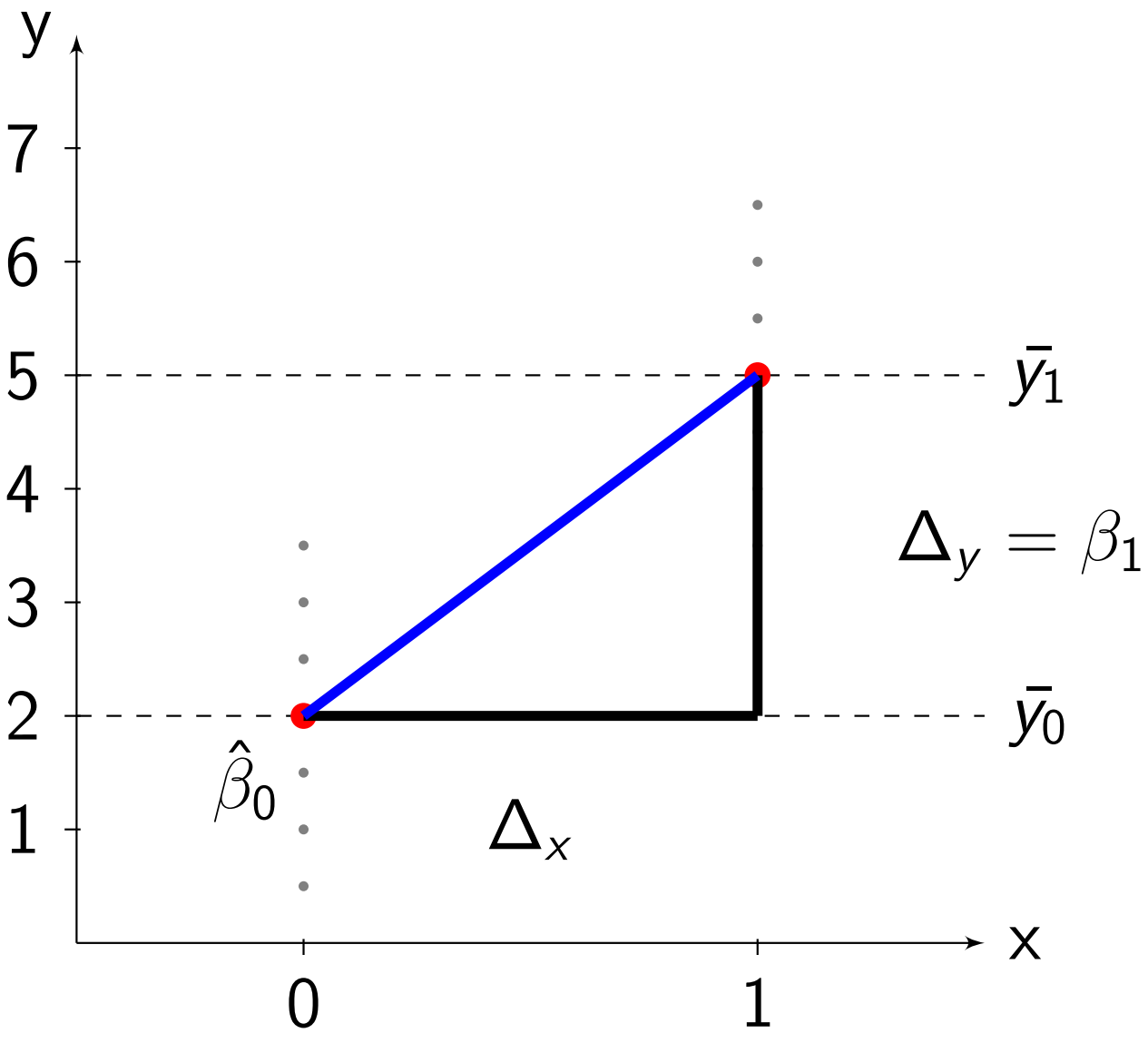
Bivariate Regression I

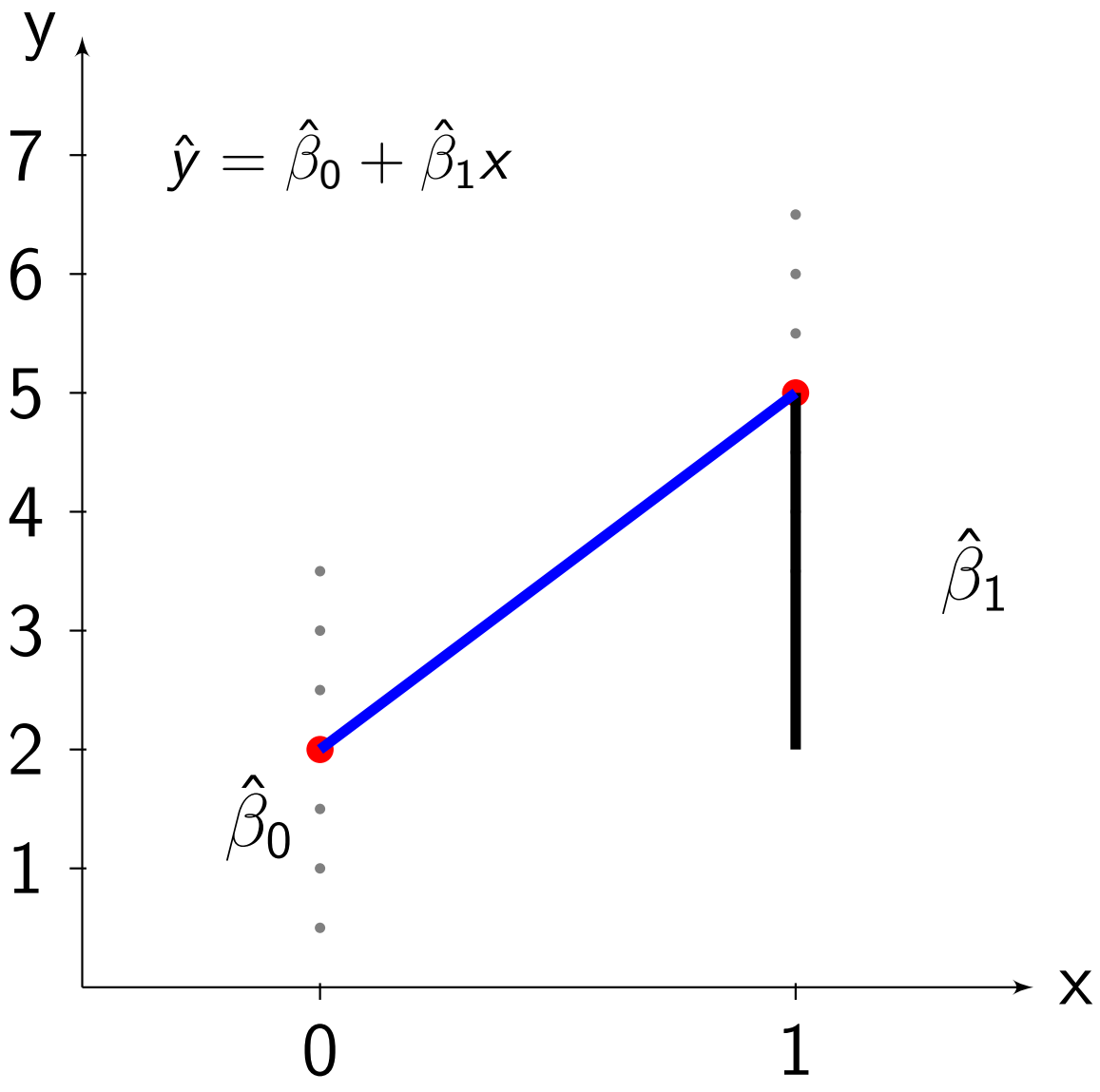
- Mean difference ($E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$) is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$
  - $\Delta Y = E[Y_i|X = 1] - E[Y_i|X = 0]$
  - $\Delta X = 1 - 0 = 1$

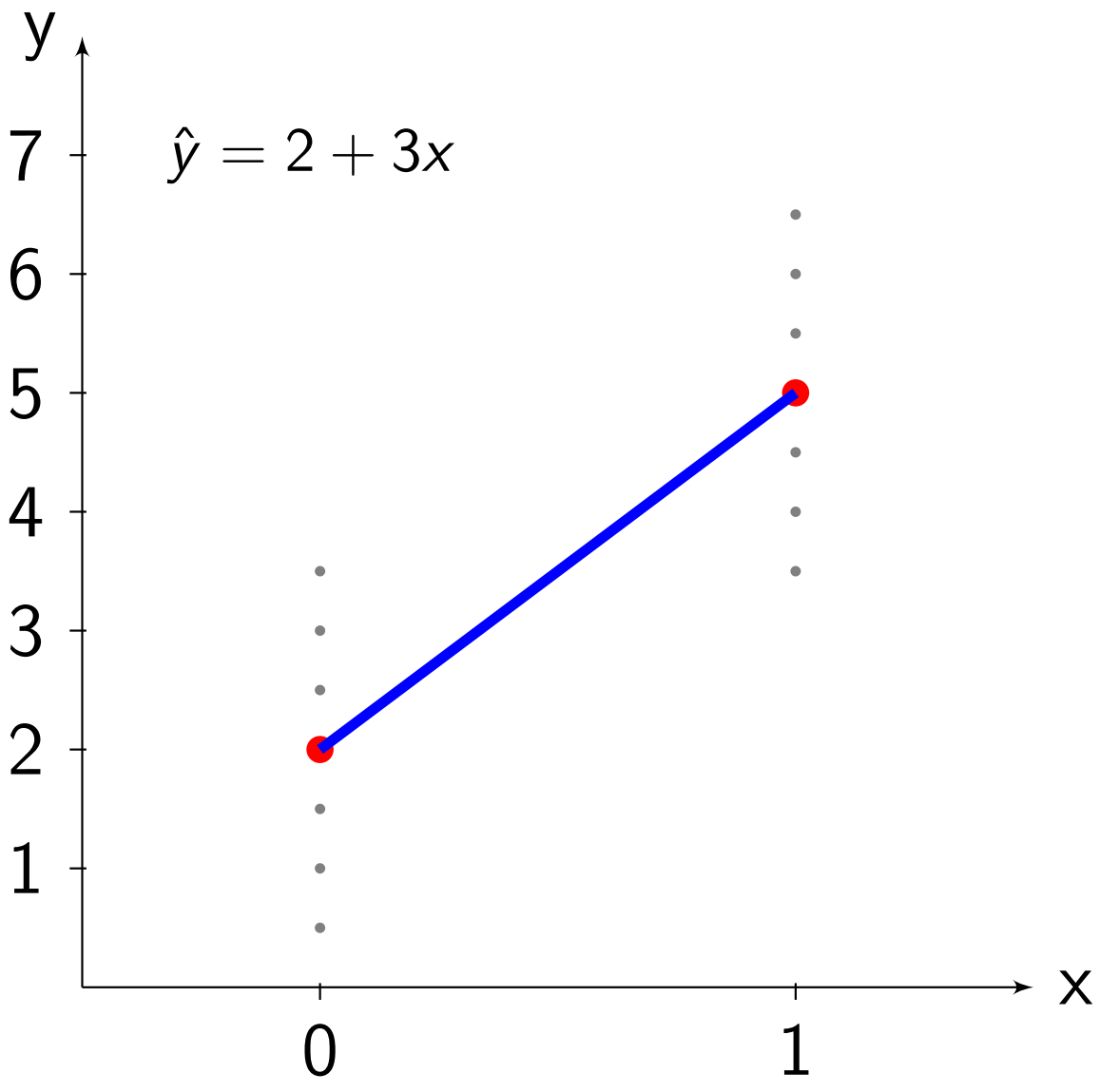

y
7
6
5
4
3
2
1
x
0
1

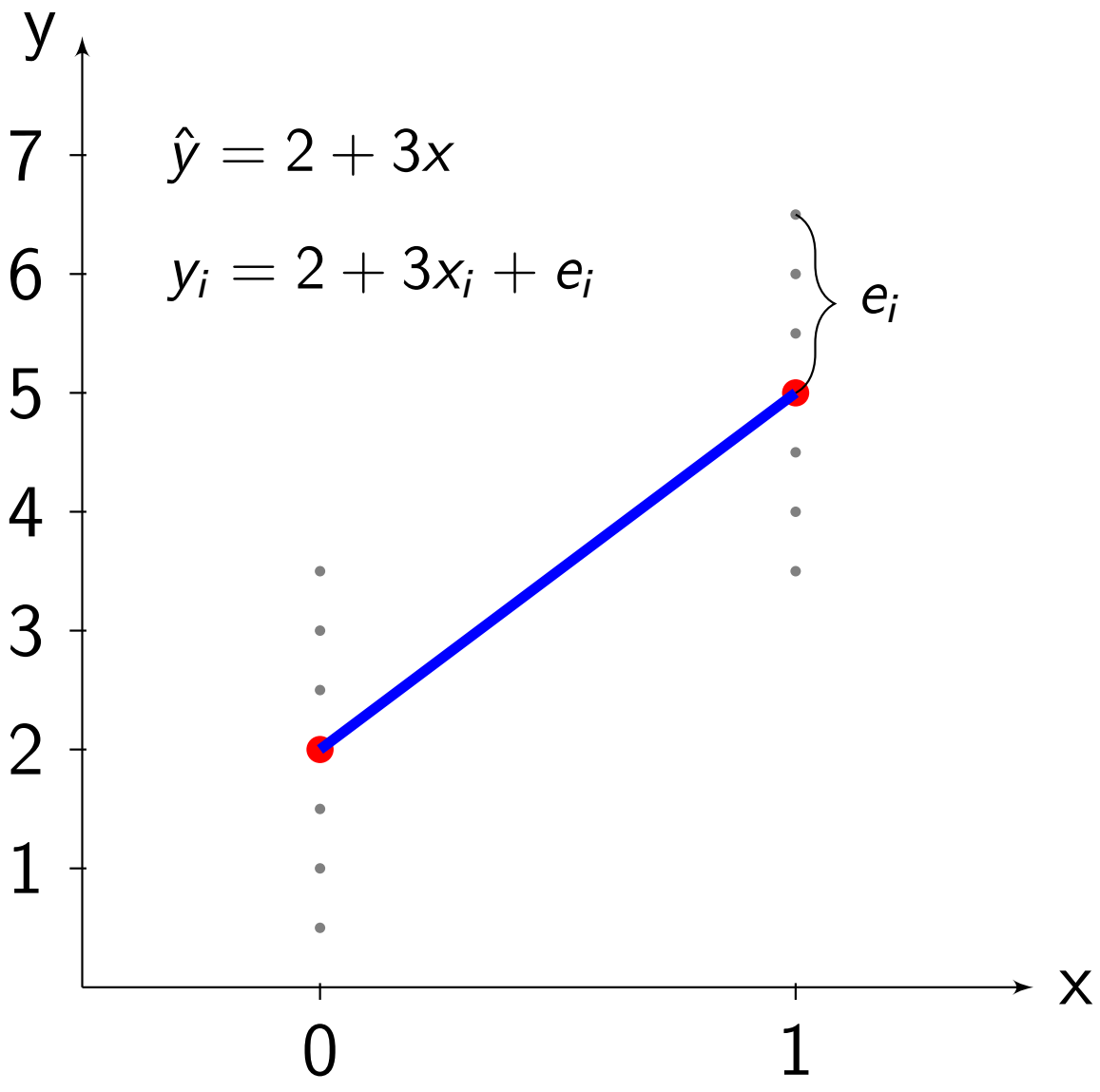

y
7
6
5
4
3
2
1
$\bar{y}_1$
$\bar{y}_0$
x
0
1


y
7
6
5
4
3
2
1
$\bar{y}_1$
$\bar{y}_0$
Δ_y
Δ_x
x
0
1


y
7
6
5
4
3
2
1
$\bar{y}_1$
$\Delta_y = \beta_1$
$\bar{y}_0$
$\hat{\beta}_0$
Δ_x
x
0
1


y
7
6
5
4
3
2
1
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{\beta}_1$
$\hat{\beta}_0$
x
0
1


y
7
6
5
4
3
2
1
$\hat{y} = 2 + 3x$
x
0
1


y
7
6
5
4
3
2
1
$\hat{y} = 2 + 3x$
$y_i = 2 + 3x_i + e_i$
e_i
x
0
1

Systematic versus unsystematic component of the data

- Systematic: Regression line (slope)
  - Linear regression estimates the conditional means of the population data (i.e., $E[Y|X]$)
- Unsystematic: Error term is the deviation of observations from the line
  - The difference between each value y_i and $\hat{y}_i$ is the *residual*: e_i
  - OLS produces an estimate of the relationship between X and Y that minimizes the *residual sum of squares*

Why are there residuals?

Why are there residuals?

- Omitted variables
- Measurement error
- Fundamental randomness

Evaluative questions

- Name an object of evaluation
- Possibly describe that object
- Ask for a transformation of the evaluation onto a set of responses

Question templates

- Ratings
  - Several varieties of rating scales
- Scales/Thermometers
- Agree-disagree
- Forced choices
- Open-ended
- Rankings (note: need alternatives to rank against)

Extended Example

- Public opinion survey in Great Britain
- Construct: Opinion toward UK involvement in air strikes on Islamic State militants in Iraq and Syria
- Think about strengths and weaknesses of each question

Example: Rating (bipolar)

Do you support or oppose Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

- Strongly support
- Somewhat support
- Neither support nor oppose
- Somewhat oppose
- Strongly oppose

Example: Rating (branching)

Do you support or oppose Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

- Support
- Neither support nor oppose
- Oppose

Would you say that you strongly [support|oppose] or somewhat [support|oppose] Great Britain's participation?

- Strongly
- Somewhat

Example: Rating (bipolar)

Are you favourable or unfavourable toward Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

- Very favourable
- Somewhat favourable
- Neither favourable nor unfavourable
- Somewhat unfavourable
- Strongly unfavourable

Example: Rating (unipolar)

To what extent do you support Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

- Strongly
- Moderately
- Somewhat
- Not at all

Example: Rating (unipolar)

How favourable are you toward Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

- Extremely favourable
- Very favourable
- Moderately favourable
- Somewhat favourable
- Not at all favourable

Example: Numbered Scale

On a scale from 1 to 5, with 1 being "strongly oppose" and 5 being "strongly support," to what extent do you support Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

1. Strongly oppose
2.
3.
4.
5. Strongly support

Example: Thermometer

We would like to get your feelings toward some of political policies. Please rate your support for the policy using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favourable and warm toward the policy. Ratings between 0 degrees and 50 degrees mean that you don't feel favourable toward the policy. You would rate the policy at the 50 degree mark if you don't feel particularly favourable or unfavourable toward.

Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria.

- 0–100 slider

Example: Agree/Disagree (bipolar)

To what extent do you agree with the following statement: I support Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

Example: Agree/Disagree (unipolar)

To what extent do you agree with the following statement: I support Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria.

- Agree completely
- Agree to a large extent
- Agree to a moderate extent
- Agree a little bit
- Agree not at all

Example: Forced choice

When thinking about Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria, which of the following comes closer to your opinion:

- Great Britain should participate in air strikes
- Great Britain should not participate in air strikes

Example: Open-ended

In your own words, how would you describe your opinion on Great Britain's participation in U.S.-led air strikes on Islamic State (IS) in Iraq and Syria?

Additional Considerations

- How many response categories?
- Middle category (presence and label)
- "no opinion" and/or "don't know" options
- Probe if "no opinion" or "don't know"?
  - Encourage guessing?
  - Clarify/describe object of evaluation?
- Branching format?
- Order of response categories
- Changes based on survey mode

History: Key Moments

1. Tuskegee (1932-1972) and Guatemala (1946-1948) Studies
2. Nuremberg Code (1947)
3. Helsinki Declaration (1964)
4. U.S. 45 CFR 46 (1974) and "Common Rule" (1991)
5. The Belmont Report (1979)
6. EU Data Protection Directive (1995; 2012)
    - UK Data Protection Act (1998)

Helsinki Declaration

- Adopted by the World Medical Association in 1964[33]
- Narrowly focused on medical research
- Expanded the Nuremberg Code
  - Relaxed consent requirements
  - Risks should not exceed benefits
  - Institutionalization of ethics oversight

[33] http://www.bmj.com/content/2/5402/177

Helsinki Declaration

- Adopted by the World Medical Association in 1964[33]
- Narrowly focused on medical research
- Expanded the Nuremberg Code
  - Relaxed consent requirements
  - Risks should not exceed benefits
  - Institutionalization of ethics oversight
- Do these rules apply to non-medical research?

[33] http://www.bmj.com/content/2/5402/177

The Belmont Report

- Commissioned by the U.S. Government in 1979[34]
- Three overarching principles:
  1. Respect for persons
  2. Beneficence
  3. Justice
- Three policy implications:
  - Informed consent
  - Assessment of risks/benefits
  - Care for vulnerable populations

[34] http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Benefits and Harm

- What is a “benefit”?
- What is a “harm”?
- How do we balance the two?

Ethical Considerations

- Most ethical issues are not unique to *experimental* social science
- Some especially important issues:
  1. Randomization
  2. Informed consent
  3. Privacy
  4. Deception
  5. Publication bias

I. Randomization

- Is it ethical to randomize?

II. Informed Consent

- Persons must consent to being a research subject

II. Informed Consent

- Persons must consent to being a research subject
- What this means in practice is complicated
  - What is consent?
  - What is "informed" consent?
  - What exactly do they have to consent to?

II. Informed Consent

- Persons must consent to being a research subject
- What this means in practice is complicated
  - What is consent?
  - What is "informed" consent?
  - What exactly do they have to consent to?
- Cross-national variations
  - Consent forms required in U.S.
  - Not required in UK

III. Privacy

- Under EU Data Protection Directive (1995), data can be processed when:
  - Consent is given
  - Data are used for a "legitimate" purpose
  - Anonymous or confidential
- Data cannot leave the EU except under conditions

III. Privacy

- Experimental might be additionally sensitive

III. Privacy

- Experimental might be additionally sensitive
- Answers reflect "manipulated" attitudes, behaviors, perceptions, etc. that respondents may not have given in another setting

IV. Deception

- Major distinction between psychology tradition and economics tradition[35]
  - Purpose of the study
  - Purpose of specific items or tasks
  - Order or length of questionnaire

[35] Dickson, E. 2011. "Economics versus Psychology Experiments." *Cambridge Handbook of Experimental Political Science*.

IV. Deception

- Major distinction between psychology tradition and economics tradition[35]
  - Purpose of the study
  - Purpose of specific items or tasks
  - Order or length of questionnaire
- Psychologists focus on *debriefing*

[35] Dickson, E. 2011. "Economics versus Psychology Experiments." *Cambridge Handbook of Experimental Political Science*.

IV. Deception

- Major distinction between psychology tradition and economics tradition[35]
  - Purpose of the study
  - Purpose of specific items or tasks
  - Order or length of questionnaire
- Psychologists focus on *debriefing*
- Within economics, norms about *acts of omission* versus *acts of commission*

[35] Dickson, E. 2011. "Economics versus Psychology Experiments." *Cambridge Handbook of Experimental Political Science*.

IV. Deception

- Major distinction between psychology tradition and economics tradition[35]
  - Purpose of the study
  - Purpose of specific items or tasks
  - Order or length of questionnaire
- Psychologists focus on *debriefing*
- Within economics, norms about *acts of omission* versus *acts of commission*
  - Omission: In a multi-round trust game, an additional round is added

[35] Dickson, E. 2011. "Economics versus Psychology Experiments." *Cambridge Handbook of Experimental Political Science*.

IV. Deception

- Major distinction between psychology tradition and economics tradition[35]
  - Purpose of the study
  - Purpose of specific items or tasks
  - Order or length of questionnaire
- Psychologists focus on *debriefing*
- Within economics, norms about *acts of omission* versus *acts of commission*
  - Omission: In a multi-round trust game, an additional round is added
  - Commission: Telling respondents it is a dictator game, but it is actually a trust game

[35] Dickson, E. 2011. "Economics versus Psychology Experiments." *Cambridge Handbook of Experimental Political Science*.

V. Publication Bias

- Publication bias not typically discussed as an ethical question

V. Publication Bias

- Publication bias not typically discussed as an ethical question
- If studies are meant to policy or practical implications, then we care about PATE or a set of CATEs, including whether their effects are positive, negative, or zero.

V. Publication Bias

- Publication bias not typically discussed as an ethical question
- If studies are meant to policy or practical implications, then we care about PATE or a set of CATEs, including whether their effects are positive, negative, or zero.
- Publication bias (toward "significant" results) invites wasting resources on treatments that actually don't work

Lots of Other Ethical Questions

Lots of Other Ethical Questions

1. Funding

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization
3. Vulnerable populations (e.g. children, sick)

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization
3. Vulnerable populations (e.g. children, sick)
4. Incentives

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization
3. Vulnerable populations (e.g. children, sick)
4. Incentives
5. Cross-national research

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization
3. Vulnerable populations (e.g. children, sick)
4. Incentives
5. Cross-national research
6. End uses/users of research

Lots of Other Ethical Questions

1. Funding
2. Independence and Politicization
3. Vulnerable populations (e.g. children, sick)
4. Incentives
5. Cross-national research
6. End uses/users of research
7. Others...

Questions?