LECTURE 2: A FRAMEWORK FOR THINKING ABOUT CAUSALITY & RANDOMIZED EXPERIMENTS

September 20, 2010

Plan for Today

1. Introductions
2. Internal and External Validity
3. Selection Bias
4. Quantitative data and correlation
5. The logic of statistical inference
6. Rubin's Causal Model
7. Randomized experiments
    1. Interpreting experimental data
    2. Limitations of experiments
8. Project STAR

Introductions

- Name, year, area of study
- Example(s) of the kind of research questions you are interested in:
  - Descriptive questions
  - Association questions
  - Causal questions
  - Why questions

Importance of Theory

- Provides guidance about the question to ask
- Informs the key constructs to measure
- Suggests the direction of relationships
- Common examples:
  - Human Capital theory
  - Market Signaling theory
  - Bourdieu's Social Reproduction theory

John Stuart Mill's 3 Critical Conditions for Causal Research

1. Cause comes before effect
2. Different levels of cause lead to different levels of effect
3. There can be no other alternative explanations for the link between cause and effect other than that X really does cause Y
   - Random assignment discounts all other possible explanations

Internal and External Validity

- ***Internal validity***: the statistical inferences about causal effects are valid for the population being studied.
  - There are no rival explanations for the statistical relationship between the treatment and the outcomes.

- ***External validity***: the statistical inferences can be generalized from the population and setting studied to other populations and settings.

Threats to internal validity

1. **Maturation:** changes in outcomes among participants as a function of the passage of time
2. **Regression to the mean:** pseudo-effects observed when participants are selected based on their (high or low) pre-treatment outcomes
3. **Testing:** the effect of taking a test upon scores of a second testing – or the effect of publishing a social indicator upon subsequent readings of that indicator
4. **Changing the measurement instrument over time**
5. **Non-random attrition:** the differential loss of members (in rate or patterns) from the groups being compared
6. **Expectancy:** bias caused by differential behavior of either the experimenter or the subjects that leads to different measures of the dependent variable than would have occurred had there been no study (e.g. “Hawthorne” or “John Henry” effects).

Threats to External Validity

- Policy changes over time
- Small sample size- not generalizable
- Geographic concentration

Compared to…what?

Common Research Designs

(*X* = intervention; *O* = Outcome; *T* = Treatment Group)

One-group post-test only

X	O_T

One-group Pretest-Posttest

I_T	X	O_T

Pre-test/ Post-test

Common Research Designs

(*X* = intervention; *O* = Outcome; *T* = Treatment Group; *C* = Comparison Group)

Two-group post-test only

X	O_T
	O_C

Two-group Pretest-Posttest

I_T	X	O_T
I_C		O_C

How Comparison Groups Matter

Orr, *Social Experiments*, Table 1.2					
Group	Pre GPA		Post GPA		Pre-Post Change
Participants, w/ program	2.0		2.6		0.6
Estimated Impact (one-group) Comparison group	2.2		2.4		0.2
					0.4
Estimated Impact (two-group) Participants w/o program	2.0		2.5		0.5
					0.1
True Impact	--		--		

The "Evaluation Problem"

- You want to know what would have happened in the absence of a program, however, it is often very difficult to identify the correct comparison group

- Specifically, if your comparison group does a poor job of estimating the counterfactual of the treatment group, your estimate of the treatment effect will be biased.

Difficulty selecting the comparison group

- Selection bias- Participants are self selected based on unobserved characteristics or selected by someone else based on non-random characteristics
- Primary sources of selection bias: Motivation/ Ability

The Identification Strategy

- Endogenous assignment- the assignment to treatment is a result of actions by participants within the system
- In most educational research, the assignment to treatment is endogenous
- *Why does this bias the results?*
- Exogenous assignment- relating to external causes- determined by investigator or some independent agency-
  - Project STAR
  - but some parents still switched their kids into smaller classes= endogenous manipulation

Potential Problems with Experiments

- Placebo effect- Participation in a study causes a change in behavior/ outcomes
- Ethical Issues- evidence of harm in offering or withholding a treatment
- Duration- Takes time to see changes in behavior
- Attrition
- Costs
- Substitution of alternative treatment- control group seeks out new treatment
- Treatment group doesn't take up the treatment- "intent to treat" effect

Data: key terms

- Variables
  - *Continuous variable*: a variable that can take on any real value within a given range. It takes on numerical values for which arithmetic operations such as adding and averaging make sense (e.g. height, IQ, earnings).
  - *Categorical variable*: places individuals into categories that cannot be quantified in a meaningful way (e.g. city of residence, ethnic background, country of birth, hair color, gender, etc.). The categories may or may not have a natural ordering.
  - *Indicator variable* (aka dummy variable or just dummy): takes on values of 0 or 1 only.
- Datasets
  - *Cross-Sectional*: Different entities in a single time period
  - *Time-Series*: Single entity in multiple time periods
  - *Panel/Longitudinal*: Different entities in multiple time periods for each entity

Data on a pre-12th grade math intervention

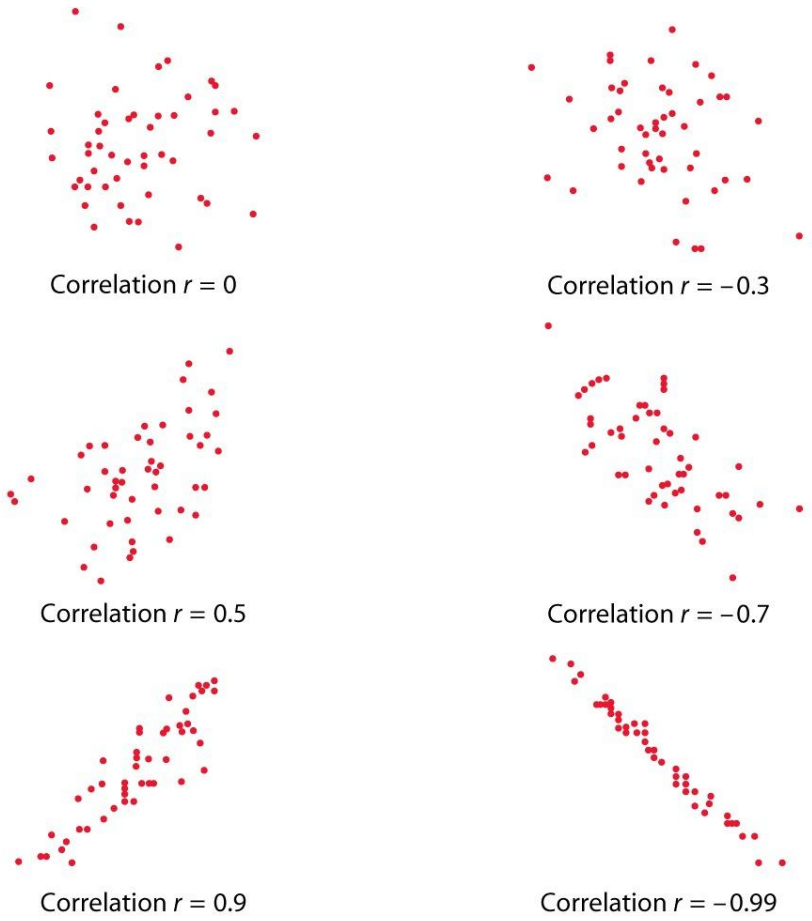
	id	ses	math8	math12	female	program
1.	124902	-.529	45.24	52.14	0	0
2.	124915	-.377	46.16	62.46	0	1
3.	124916	-.859	43.14	52.7	1	0
4.	124932	-.191	42.35	47.96	0	1
5.	124944	-.319	52.37	56.43	1	1
6.	124966	-.601	51.21	57.56	1	0
7.	124968	-.07	52.09	69.1	0	0
8.	124970	.198	43.43	50.53	0	0
9.	124972	.096	44.78	56.15	1	1
10.	124974	.136	55.52	66.02	0	0
11.	124981	.089	42.4	49.56	0	1

Correlation: a few reminders

-- ρ_{xy} is always between –1 and +1

-- does not describe curved relationships

-- not affected by units – e.g. could use inches or cm



Correlation does not imply causation!

Correlation is neither necessary nor sufficient for causation…

(some systematic relationship between the treatment and outcome variables is of course necessary for causation, but we may not be able to observe that relationship in the raw data)

…but it is often a strong hint.

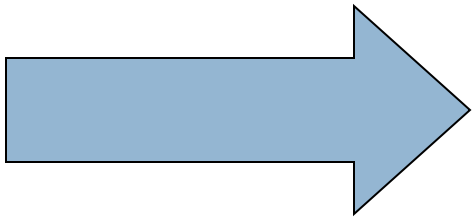
The Logic of Statistical Inference

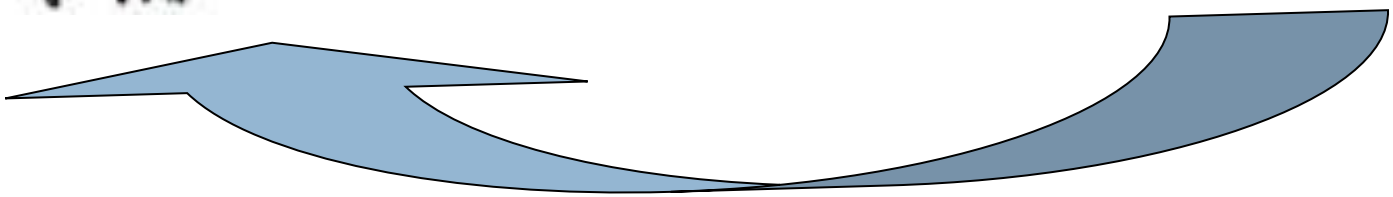
- What is the difference between descriptive and inferential statistics?

Key concepts:

- A sample vs. the population
- Sample statistics vs. population parameters
- An estimate vs. "THE TRUTH"
- Estimators and their sampling distributions

With a sample and an estimator, we can *estimate* the unseen "truth"





μ_{IQ}

$\overline{IQ}$

Population mean IQ

Mean IQ in the sample

Candidate "estimators" of the population mean IQ

- Sample mean IQ→ $\overline{IQ}$
- Sample median IQ
- The IQ of a randomly drawn individual

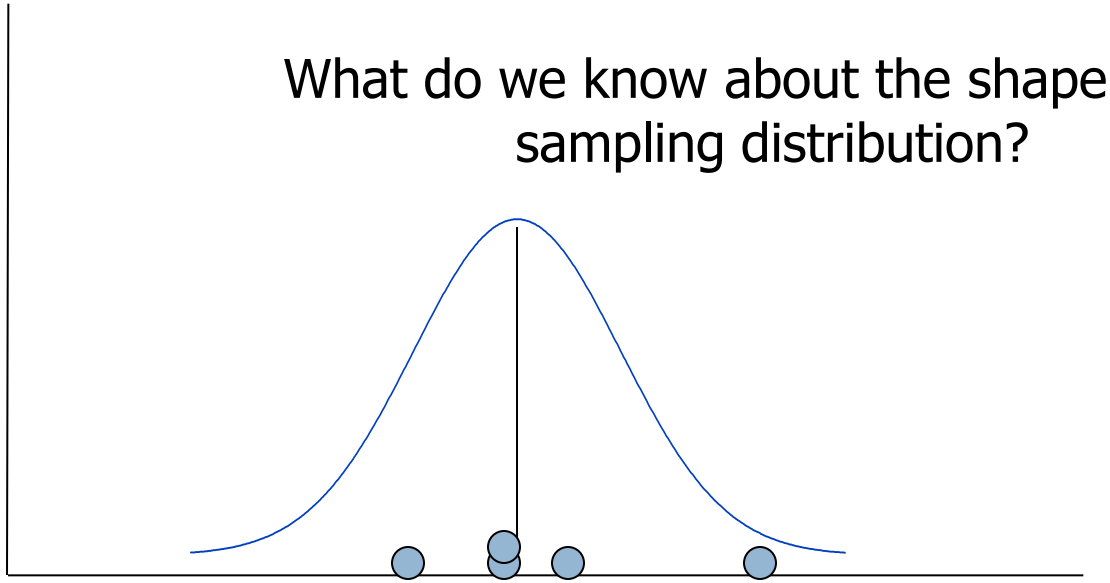
- Desirable properties of an estimator?
  - Unbiased: the expected value of the estimator is equal to the parameter (where expected value means the "long-run average" – or the average from many repeated samples of the same size). Centered around the actual value of the population parameter.
  - Precise: the values of the estimator from many repeated samples have the smallest variance of any estimator. Smallest standard error= best precision.

- Unbiasedness and precision are both claims about the *sampling distribution* of the estimator.

Sampling distribution of the sample mean

What do we know about where the sampling distribution is centered?

What do we know about the shape of the sampling distribution?



Making statistical inferences in impact evaluation

- Null Hypothesis, H_0 ($\beta_1 = 0$)

There really is **no relationship** between X and Y in the population

- Alternative Hypothesis ($\beta_1 \neq 0$)

There really is a **relationship** between X and Y in the population

Assuming the null hypothesis was true, how likely is it that we would have gotten the sample result we did?
We usually hope to reject H_0 because we usually want there to be a relationship!

Making statistical inferences in impact evaluation

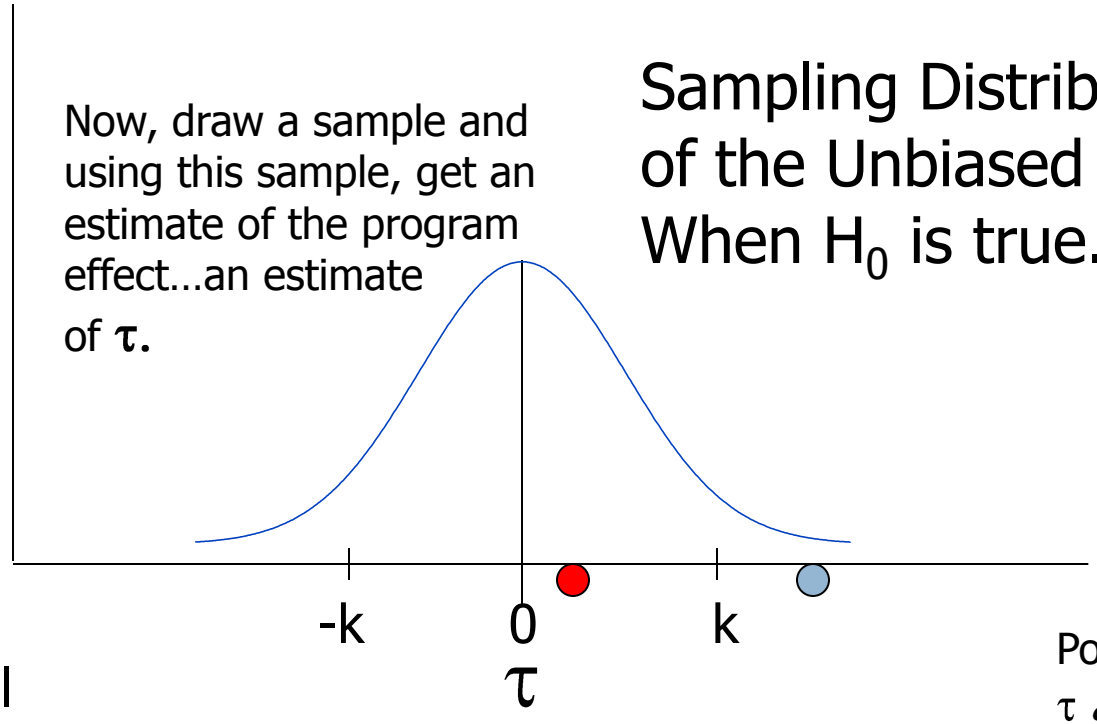
Consider some program.

Let τ be the true difference in mean outcomes between treatment and comparison groups attributable to program participation.

Null hypothesis: $\tau = 0$

Assume that we have an appropriate sample and an unbiased estimator of τ.

Statistical Inference When You Have an ***Unbiased*** Estimator



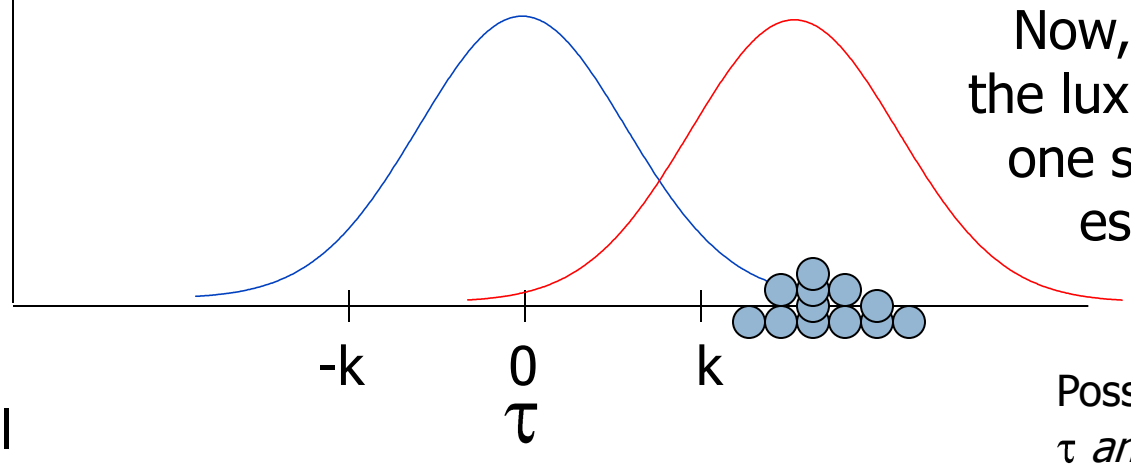
Statistical Inference in the Real World

With the luxury of many samples, the following interpretations both fit the data.

Probability of obtaining any given estimate of τ

1) Your unbiased estimator provides evidence against the null hypothesis.

2) You have a biased estimator.



Now, what if we had the luxury of more than one sample to obtain estimates of τ?

The null hypothesis H_0: $\tau = 0$

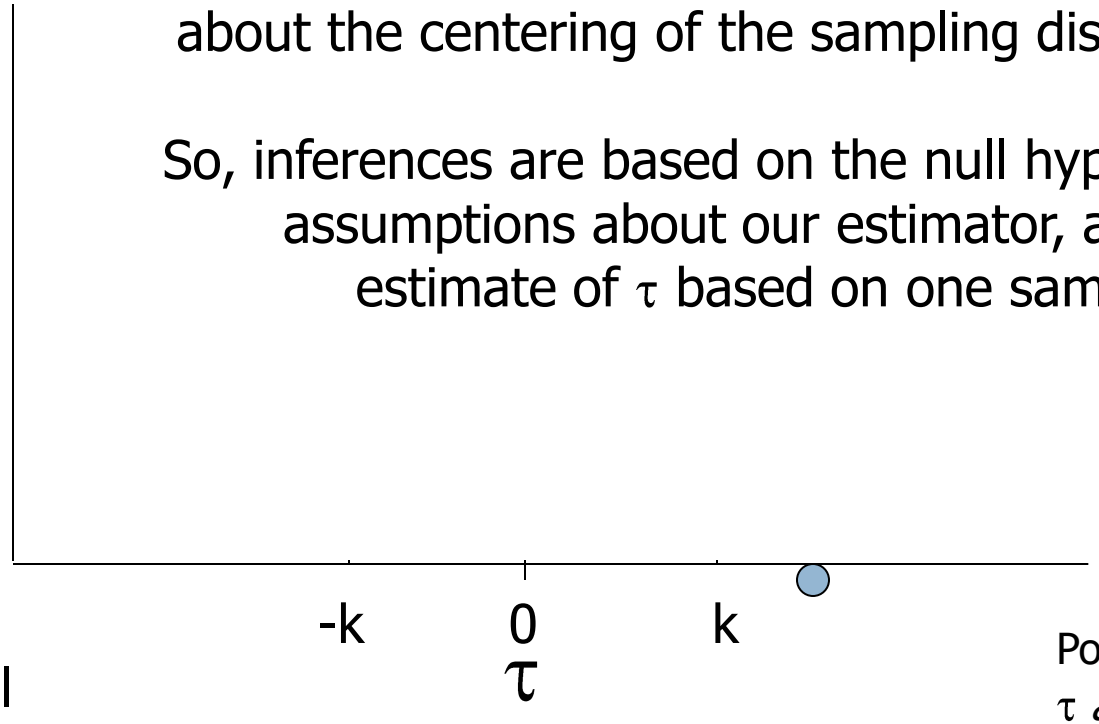
Let's get an estimate of τ from a sample.

Possible values of τ *and* of any given estimate of τ

Statistical Inference in the Real World

but we can usually only hypothesize about the centering of the sampling distribution...

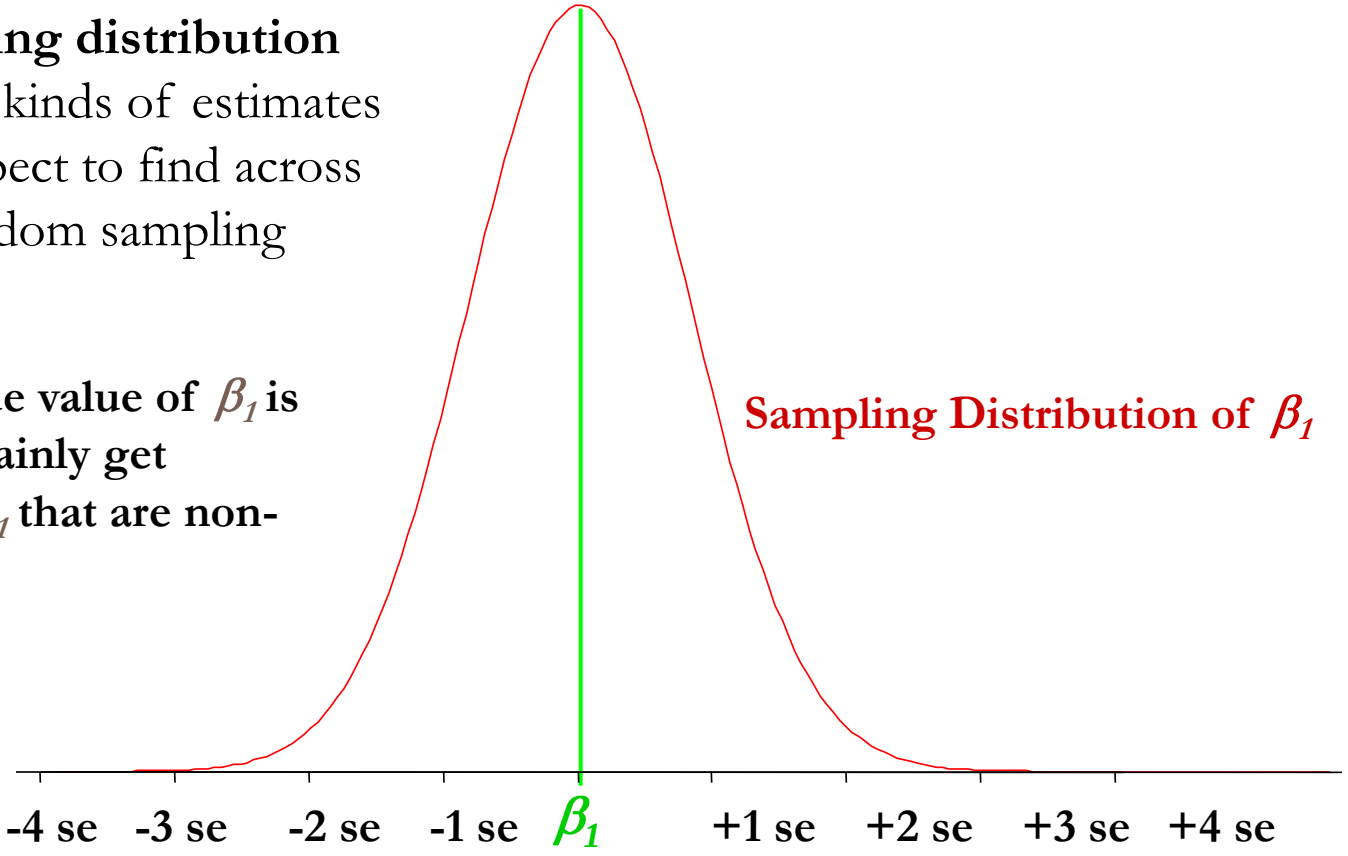
So, inferences are based on the null hypothesis, our assumptions about our estimator, and our estimate of τ based on one sample.



Sampling distribution of an estimated regression slope (β_1)

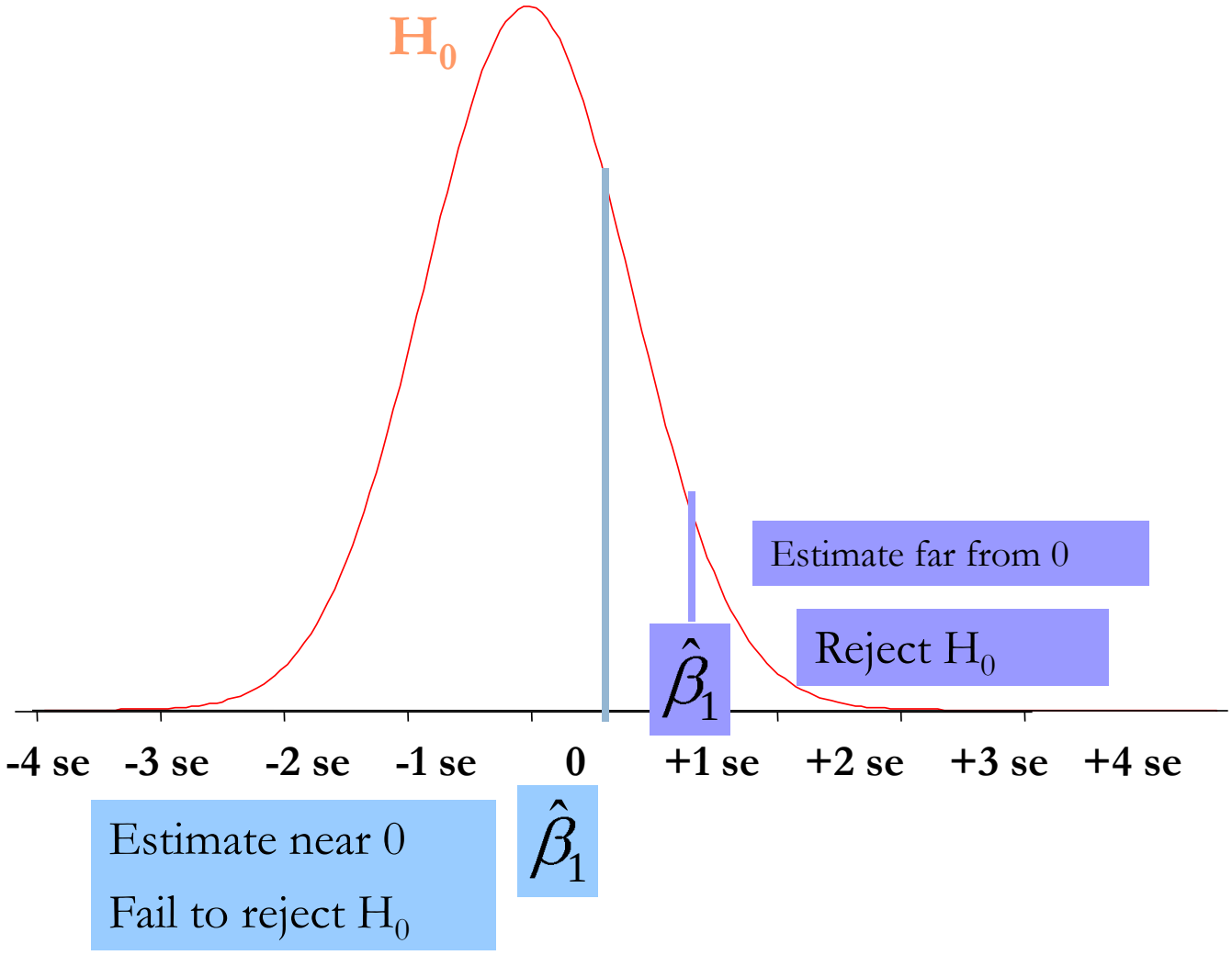
This sampling distribution tells us what kinds of estimates we could expect to find across repeated random sampling

Even if the true value of β_1 is 0, we will certainly get estimates of β_1 that are non-zero

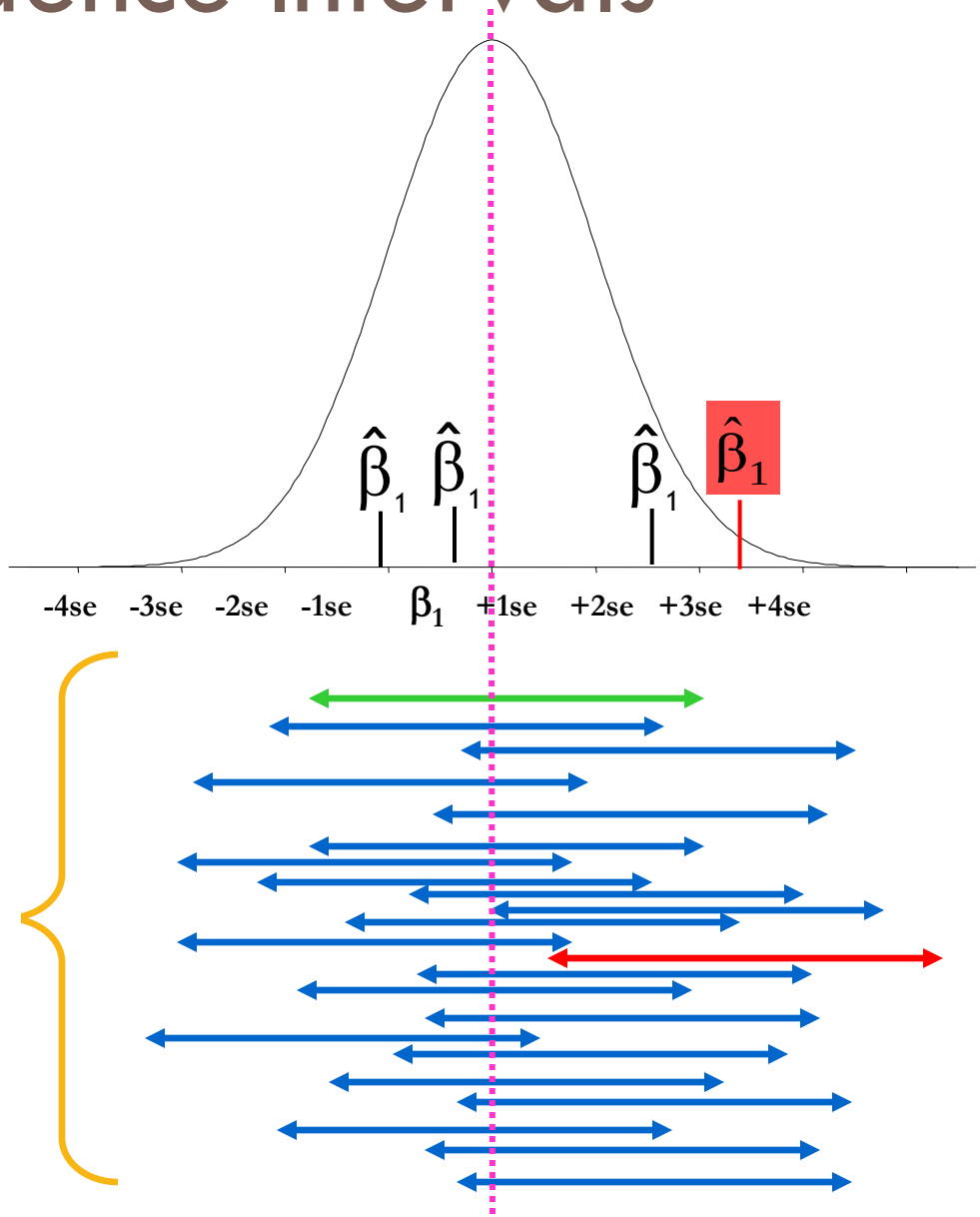


The standard deviation of a sampling distribution is known as a standard error (se)

Reject or Fail to Reject H_0?



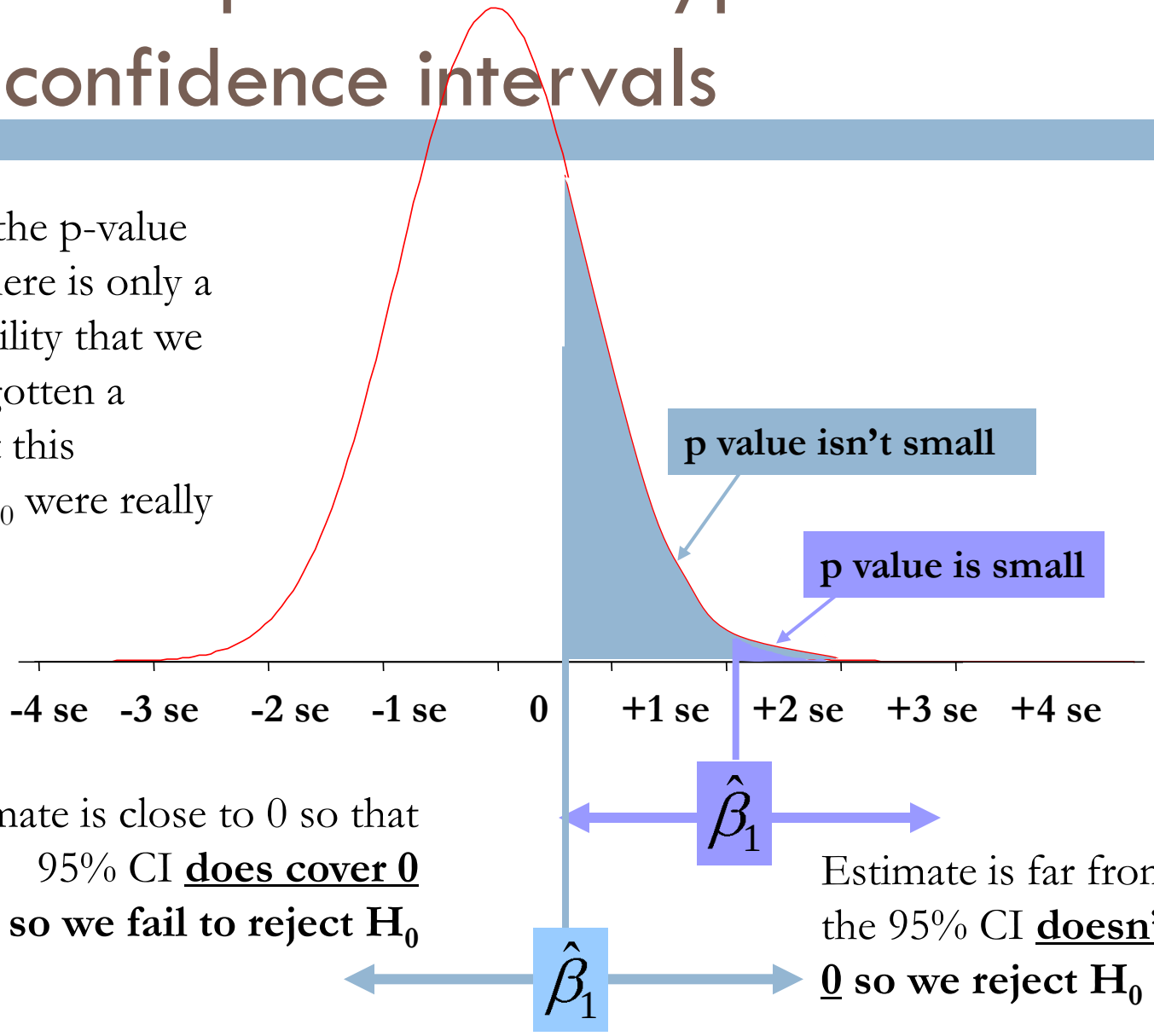
Confidence Intervals



For every 20 intervals we construct, we estimate that an average of 1 won't cover the true value of β_1

Relationship between hypothesis tests and confidence intervals

Reject H_0 if the p-value is small (if there is only a small probability that we would have gotten a sample result this extreme if H_0 were really true).



Estimate is close to 0 so that 95% CI **does cover 0** **so we fail to reject H_0**

Estimate is far from 0 so the 95% CI **doesn't cover 0 so we reject H_0**

Causal Inference and the Counterfactual

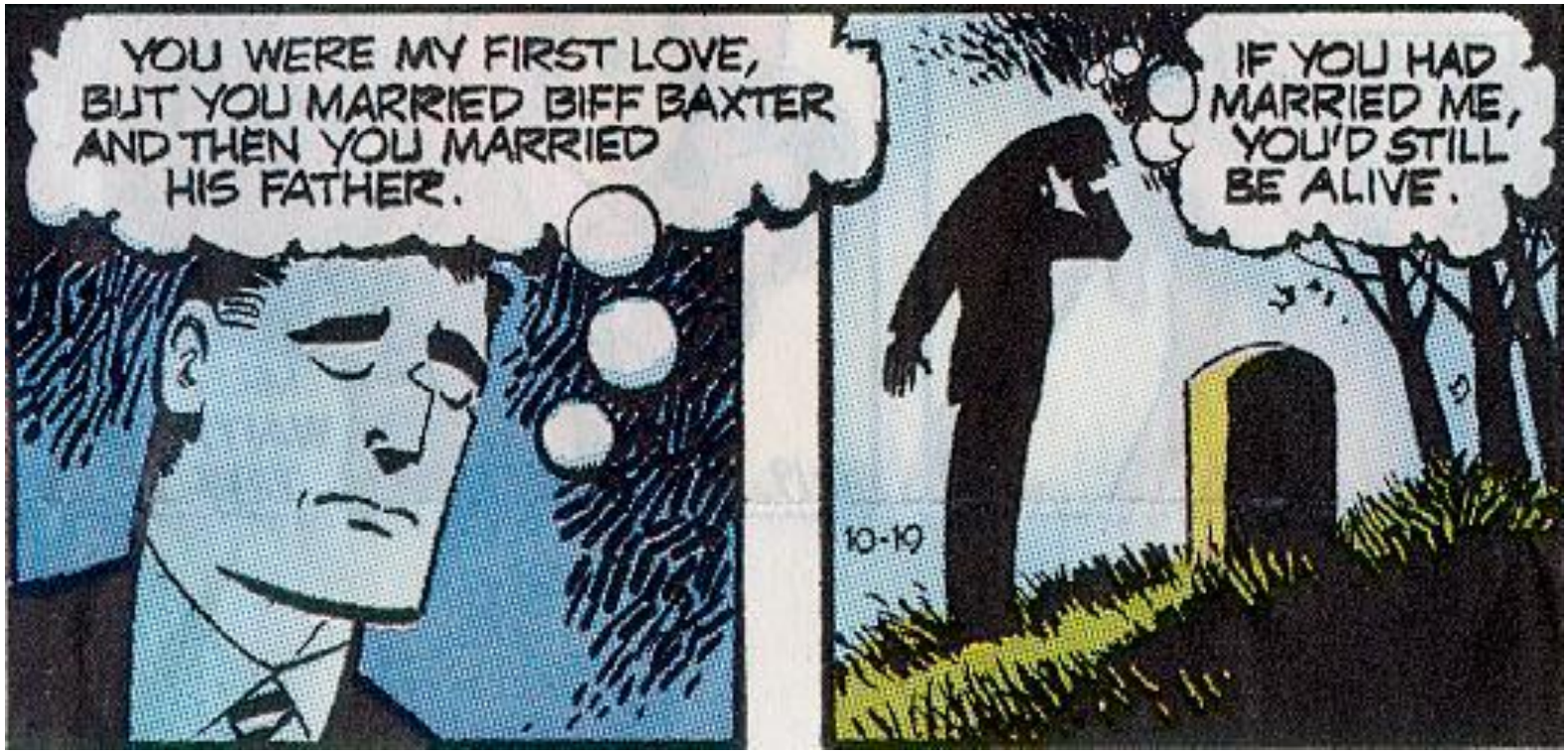
To measure the causal impact of an educational program on any given student, we want to know…

--what happened to a student who was in the program …vs.

--what would have happened to that same student had the student NOT participated in the program

This latter, unobserved outcome is the COUNTERFACTUAL.

Rubin's Causal Model (and the World of Potential Outcomes)



Rubin's Causal Model

- Highlights the key challenge in conducting impact evaluations: estimating the counterfactual.
- Definition of bias: the difference between the observed outcomes of the comparison group and the counterfactual.
- Suggests the "gold-standard" criterion for evaluating impact evaluations: how well does comparison group estimate the counterfactual?

Conditional expectations & random assignment

- **Conditional expectations** account for systematic differences between groups (e.g., E[y | x]).
- If there are no systematic differences between groups, then ***conditional*** and ***unconditional*** expectations are equal.

- Randomization assures us that there are no systematic differences between groups. If assignment to d is random, the expected outcomes of the d=0 students are the same as those of the d=1 students.

In other words…

- When assignment is random, all factors other than treatment status will tend to be distributed equally between participants in the treatment and control groups.
- Due to random sampling and random assignment, potential members of the treatment and control groups will be identical on all observed and unobserved characteristics **on average** in the population.

Developing a Causal Model

- An example: What is the impact of providing students with laptops on math achievement?
- Consider two *potential* outcomes for each student:
  - y_{1i} = test scores of the ith student if he received a laptop
  - y_{0i} = test scores of the ith student if he did not receive a laptop

- Random Assignment- each individual has a specified probability of being assigned to each group

We are interested in $y_{i1} - y_{i0}$

What would $E[y_{i1} - y_{i0}]$ yield?

The average treatment effect (ATE)

Now let d=1 if student received a laptop and d=0 if not

What does $E[y_1 - y_0|d=1]$ represent?

(suppressing the i subscripts for convenience)

The average effect of treatment <u>on the treated</u> (ATT)

$E[y_1 - y_0|d=1]$ can be rewritten as…

$$E[y_1|d=1] - E[y_0|d=1]$$

Notice that an estimate of the first term could be obtained from the data

…but information on the second is never available in the data…regardless of the program we are evaluating.

So, how will we ever be able to estimate a causal effect if it requires impossible data?

Average effect of treatment on the treated:

$E[y_1 - y_0|d=1] = E[y_1|d=1] - E[y_0|d=1]$

What do we get in a simple comparison of the test scores of students with and without laptops?

$E[y|d=1] - E[y|d=0]$

$E[y_1|d=1] - E[y_0|d=0]$

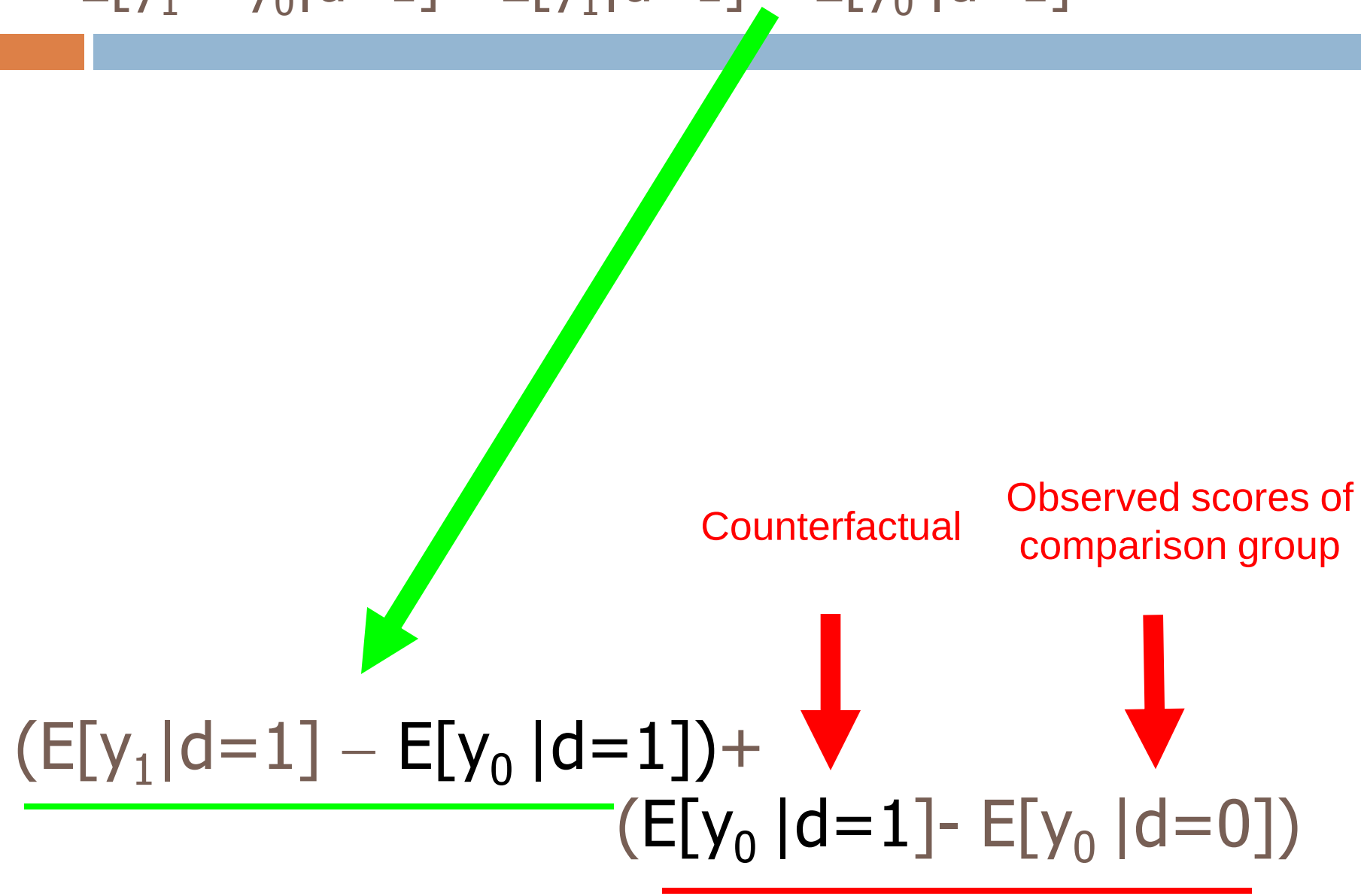
Transform these observed outcomes into potential outcomes…

…and subtract/add $E[y_0|d=1]$

$$(E[y_1|d=1] - E[y_0|d=1]) + (E[y_0|d=1] - E[y_0|d=0])$$

Average effect of treatment on the treated:

$E[y_1 - y_0|d=1] = E[y_1|d=1] - E[y_0|d=1]$



Average effect of treatment on the treated:
$E[y_1 - y_0|d=1] = E[y_1|d=1] - E[y_0|d=1]$

This is the "evaluation problem": The fact that the comparison group may not yield an unbiased estimate the counterfactual.

Bias term

$$(E[y_1|d=1] - E[y_0|d=1]) + (E[y_0|d=1] - E[y_0|d=0])$$

Summary

- Intuitively, you have an evaluation problem in a given impact study if the comparison group is inappropriate.
- Rubin's Model provides an explicit framework for posing and thinking about this problem.
- Specifically, it shows that if your comparison group does a poor job of estimating the counterfactual of the treatment group, your estimate of the treatment effect will be biased.
- So a naïve contrast of the test scores of students with and without laptops yields…
  - The average effect of treatment on the treated…
  - …plus an extra term → a bias term.
- The size of this bias term reflects how well or how poorly the comparison group estimates the counterfactual.

Review: Rubin's Causal Model

- $y_{i1} - y_{i0}$
- $E[y_{i1} - y_{i0}]$ ← **Average treatment effect**
- $E[y_{i1} - y_{i0}|d=1]$ ← **Average effect of treatment on the treated.** *What we would like to estimate…*

Thus, with random assignment we have…

$$E[y|d=1] - E[y|d=0] =$$

$$E[y_1 - y_0|d=1] + (E[y_0|d=1] - E[y_0|d=0]) =$$
$$E[y_1 - y_0|d=1] + (E[y_0] - E[y_0]) =$$
$$E[y_1 - y_0|d=1] + 0 =$$
$$E[y_1 - y_0|d=1]$$

Virtues of Experiments

1. If well designed and implemented, they guarantee internal validity

2. Results are easy to explain to policymakers and the public

Limitations of Experiments

1. External validity may be low due to sample/site selection
2. Little information on why a program works
3. SUTVA (Stable-Unit-Treatment-Value-Assumption)- potential outcomes for each child cannot depend on the group to which other children were assigned (ex- peer groups)
4. Experiments take time and cost money
5. Implementation is often imperfect
6. Are not feasible when it is impossible to exclude the control group from treatment
7. You can't always answer the question of interest:
    a. Partial vs. general equilibrium effects
    b. Total vs. partial treatment effects

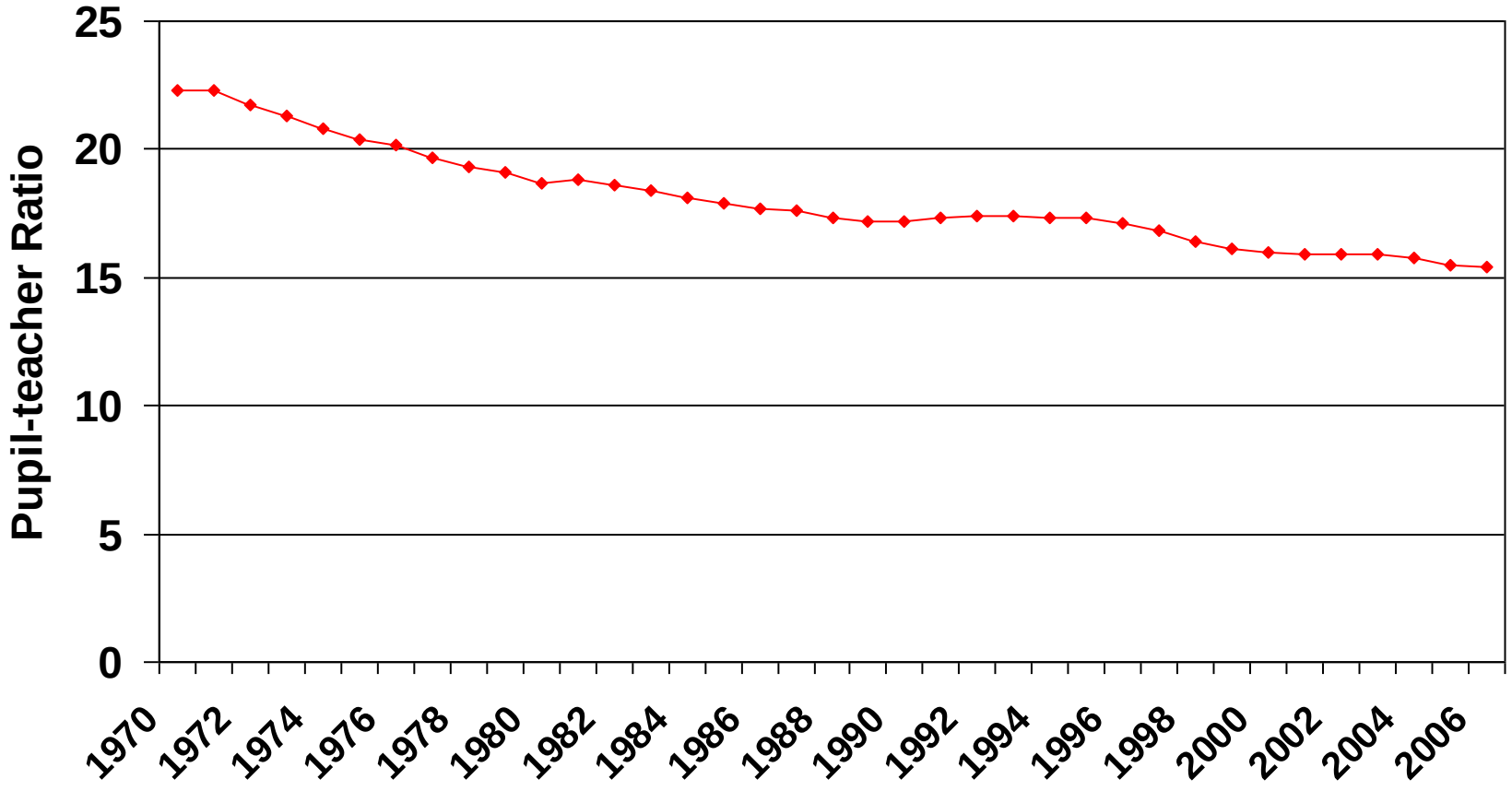
Mosteller (1995) on Project STAR

- "Because we have all gone to school, we each have ideas about how to improve the system."

Class-size Reduction: The Context

- Popular
  - Parents
  - Policymakers
  - Teachers
  - Teachers Unions
- Expensive: Teacher compensation is the dominant factor in the overall cost of K-12 education
- Mosteller: “The effects of class size on children’s learning have been studied, usually without reaching definitive conclusions.”

Pupil-Teacher Ratio in U.S. Public Schools

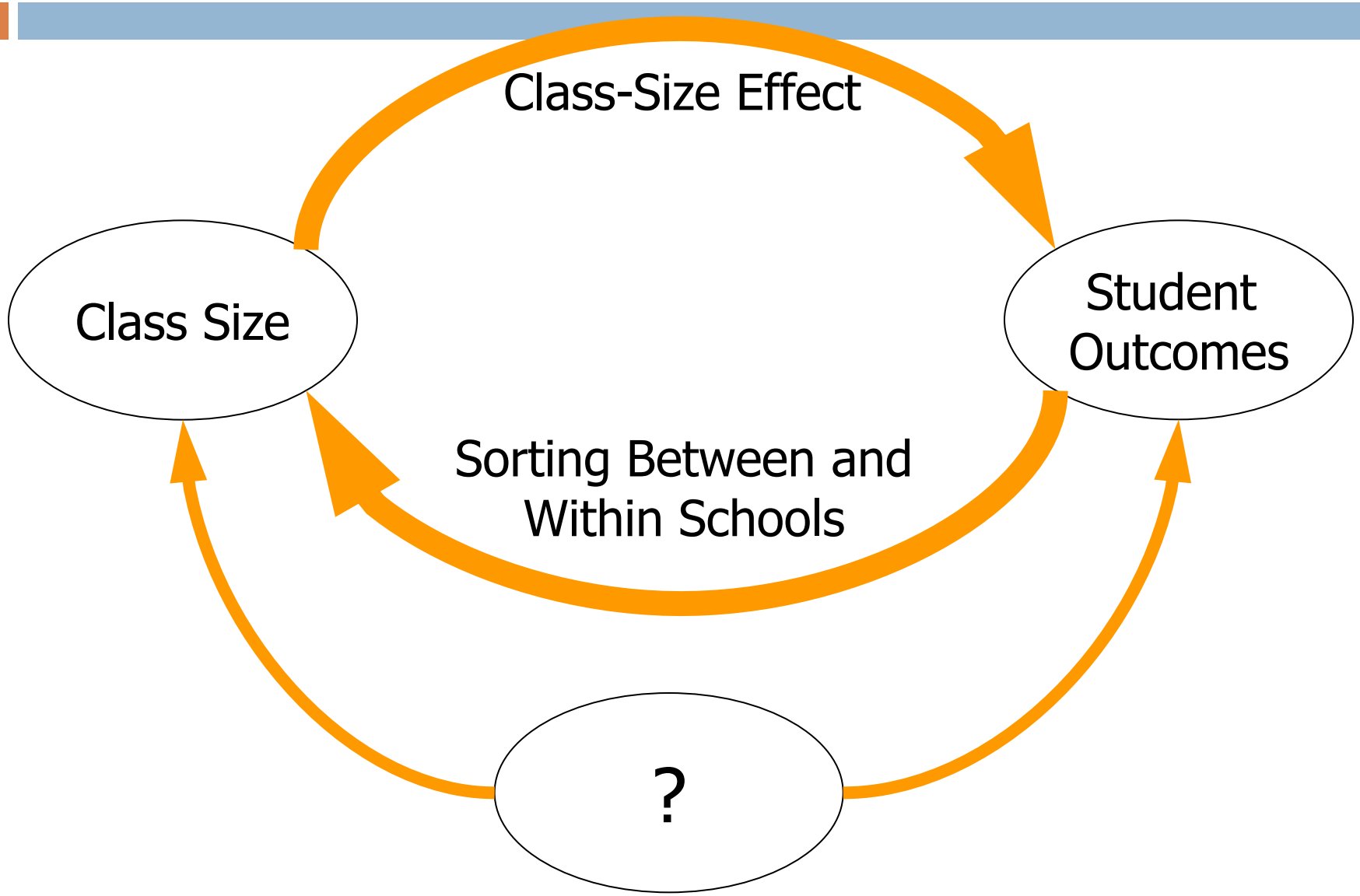


Project STAR

- Research question:
  - What is the causal effect of class size reduction in grades K-3 on student achievement?

- Research design:
  - Students and teachers in grades K-3 randomly assigned within schools to small (13-17 students), regular (22-25), or regular-with-aide classes
  - Conducted in Tennessee in 1985-89
  - 11,600 students and 1,330 teachers from 79 elementary schools participated
  - Students entering Project STAR schools after year (45 percent) randomly assigned to class types upon entry

The Problem: Endogeneity / Selection Bias



Treatment & Control Differences in 1st Grade

Table 2

Gains in Effect Sizes from Small Classes

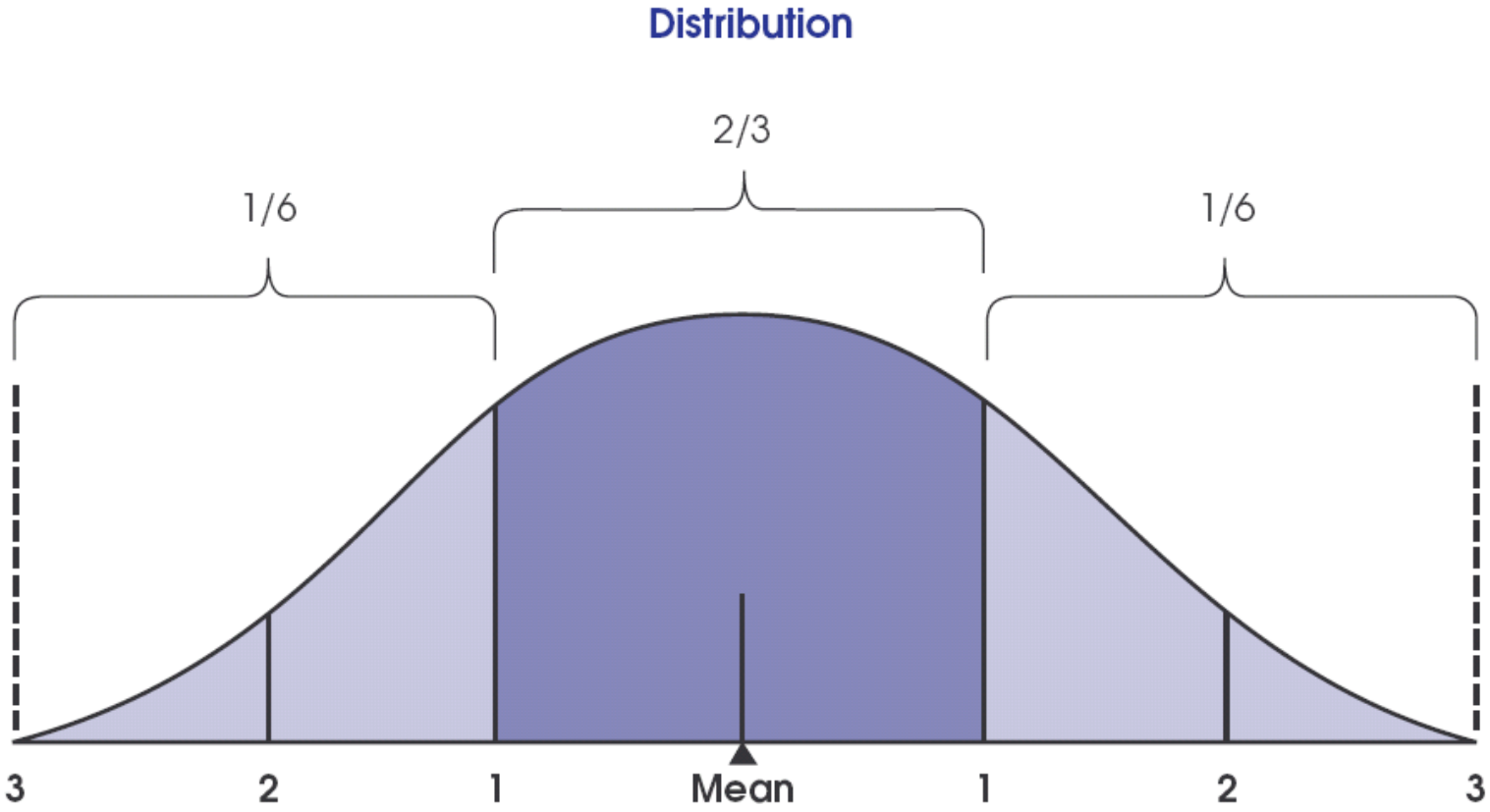
Gains in effect sizes from small classes in first grade compared with all regular-sized classes and from regular-sized classes with an aide compared with regular-sized classes without an aide

	SAT Reading	BSF Reading	SAT Math	BSF Math
The effect size on performance in small classes compared with performance in regular-sized classes with or without an aide	.23	.21	.27	.13
The effect size on performance in regular-sized classes with an aide compared with regular-sized classes without an aide	.14	.08	.10	.05

Source: Finn, J.D., and Achilles, C.M. Answers and questions about class size: A statewide experiment. *American Educational Research Journal* (1990) 27,3:557–77, Table 5.

"Effect Sizes"

- What problem do they solve?
  - Comparing treatment effects on outcomes measured in different units (usually different tests)
- How are they calculated?
  - Option 1: Divide estimated effect by one standard deviation of the treatment variable
  - Option 2: "Standardize" the outcome variable prior to estimation by subtracting its mean and dividing by its standard deviation (to produce a mean of zero and a standard deviation of 1)
- What do they mean? What is a "big" effect size?



- Effect size of 0.25 S.D. would move the median student to the 60th percentile
- The same effect size would move the student at the, say, 90th percentile a much smaller amount (as measured by percentile ranks)

Effect Size= $\frac{\text{Mean of experimental- Mean of Control}}{\text{Standard Deviation}}$

Effect Size	% of control group who would be below average person in experimental group
0	50%
.1	54%
.2	58%
.25	60%
.3	62%
.4	66%
.5	69%
.6	73%
.7	76%
.8	79%
.9	82%
1	84%

Some common points of comparison

- Black-white test score gap in 8th grade (NAEP):

  ~1 S.D.

- Difference in performance between 4th and 8th graders (NAEP):

  ~1 S.D.

- A very rough rule of thumb: Effect sizes of 0.1 or more are worth talking about

Lingering concerns

- Noncompliance: 10 percent of students moved between class types
- Initial enrollment (not assignment) recorded: students may have lobbied to switch class types
- Study attrition: about half of students left the experiment prior to grade 3

A Cautionary Tale

- California's 1996 statewide CSR policy "mandated" classes of 20 or fewer in grades K-3 by providing an extra $800/student enrolled in a small class
- Unintended consequences:
  - Most funds initially went to suburban districts
  - Dramatic increase in the share of inexperienced and uncertified teachers, especially in urban districts
  - Facilities crunch in urban districts
- No definitive evaluation (due to lack of randomization), but observational studies suggest disappointing results
- But the reform remains very popular with parents!

Expectancy Bias in Experimental Research

- "Hawthorne Effects"
  - Participants may temporarily increase their productivity when they are being evaluated
- "John Henry Effects"
  - Participants in control condition may increase effort to compensate for bad luck
- "Incentive Effects"
  - Participants may believe that future resources depend on experimental results

Preview of Next Class

- OLS regression with observational data
- Randomized experiments in practice

Readings:

1. Murnane & Willett (2011), Chapter 5
2. Schanzenbach, D. W. (2007). "What researchers have learned from Project STAR". Brookings papers on education policy: 2006-07: pp. 205-228.
3. Decker, P., Mayer, D.P., & Glazerman, S. (2004). "The effects of Teach for America on students: Findings from a national evaluation". *Report prepared by Mathematica Policy Research, Inc.* Princeton, NJ.
4. Review your notes/text on OLS