

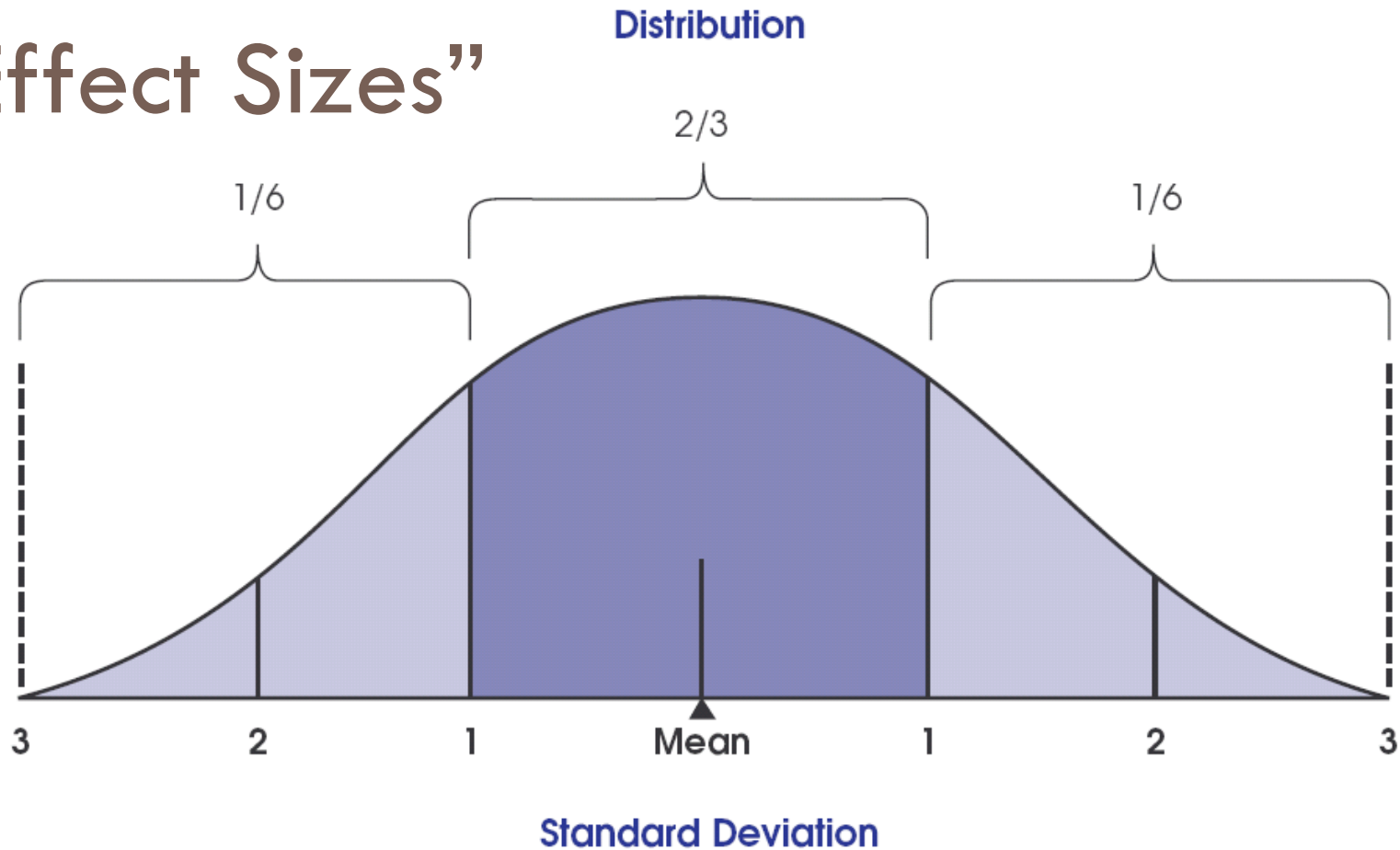
LECTURE 6: PROJECT STAR AND STATISTICAL POWER

February 18, 2010

Outline for Today

1. Policy Implications of TFA
2. Project STAR
3. Statistical power as a function of...
 - Outcome variability
 - Sample size
 - Expected impact size
4. Computer assignment #1

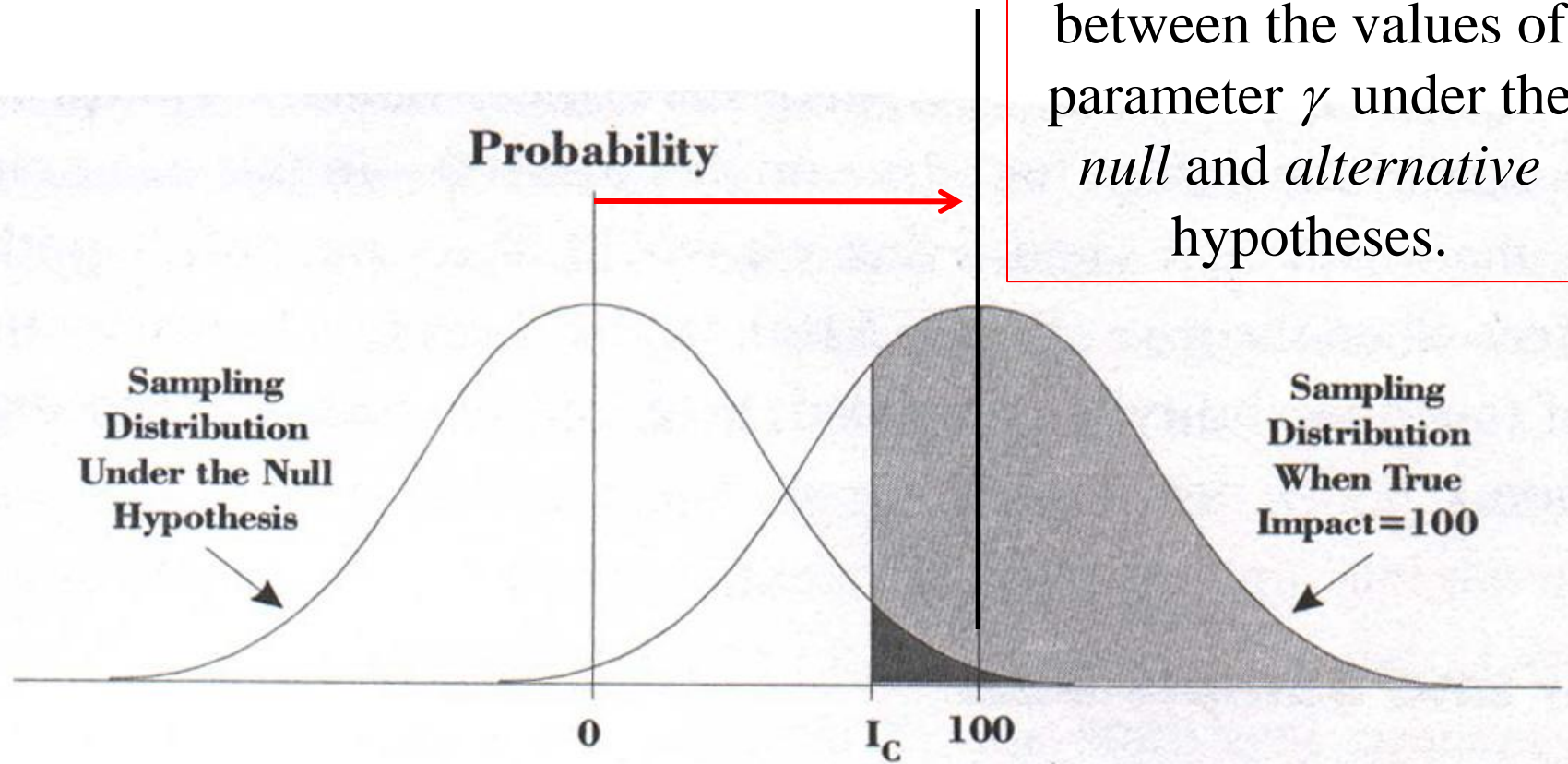
“Effect Sizes”



- Effect size of 0.25 S.D. would move the median student to the 60th percentile

Effect Size

Effect Size: *difference* between the values of parameter γ under the *null* and *alternative* hypotheses.



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

Effect Size	% of control group who would be below <u>average person</u> 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%

TFA evaluation: Results

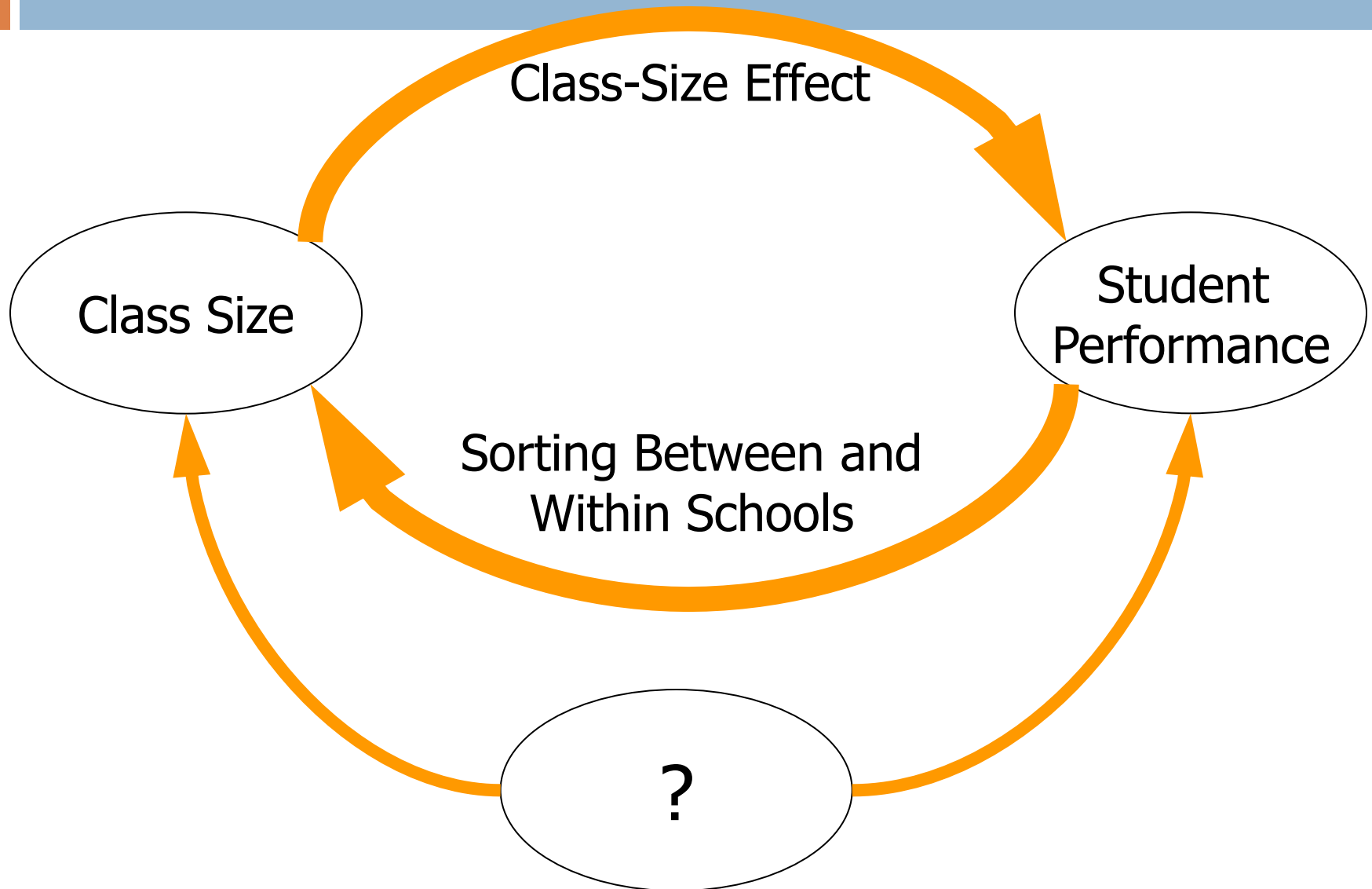
- “When adjusted using sample weights and regression methods, the difference in the growth rates [in math] was statistically significant, which demonstrates that TFA teachers generated larger math achievement gains.”
- “The impacts on both math and reading scores were reasonably similar across locations.”
- Effect size of approximately 0.15 of a standard deviation and translates into roughly 10 percent of a grade equivalent, or about one additional month of math instruction.
- Compared with their novice counterparts, new TFA teachers generated math test scores that were 0.26 standard deviations higher, on average.

➤ Policy Implications?

Class-size Reduction: The Context

- Popular
 - ▣ Parents
 - ▣ Policymakers
 - ▣ Teachers
 - ▣ Teachers Unions
- Expensive
 - ▣ Teacher compensation is the main factor in the overall cost of K-12 education
- Does it “work”?

The Problem: Endogeneity / Omitted-Variable Bias



Key Implementation Issues with STAR

- Late entry
 - ▣ Problem: 45 percent entered after Kindergarten (most in 1st grade)
 - ▣ Solution: randomly assign newcomers to class types
- Noncompliance
 - ▣ Problem: 10 percent of students moved between class types
 - ▣ Solution: Use initial class type assignment as treatment variable → “intent-to-treat” estimates (not “treatment-on-treated”)
- Initial enrollment (not assignment) recorded
 - ▣ Problem: students may have lobbied to switch class types
 - ▣ Solution(?): assignment data available for some students

The Sample

Table 1. Characteristics of Students in STAR, Tennessee, and the United States

<i>Characteristic</i>	<i>STAR (1)</i>	<i>Tennessee (2)</i>	<i>United States (3)</i>
Percent minority students	33.1	23.5	31.0
Percent black students	31.7	22.6	16.1
Percent of children below poverty level	24.4	20.7	18.0
Percent of teachers with master's degree or higher	43.4	48.0	47.3
Average ACT score	19.2	19.8	21.0
Average third-grade enrollment across schools	89.1	69.5	67.1
Average current expenditures per student across schools (dollars)	3,423	3,425	4,477

11,600 students and 1,330 teachers from 79 elementary schools participated in STAR

Was randomization successful?

- In theory, random assignment ensures that treatment and control groups will be comparable
- Check by comparing their observed characteristics (including baseline measure of outcome variable)
 - ▣ Problem: No baseline achievement data available
 - ▣ Solution: Look at other observed characteristics

Table 2. Mean Characteristics, by Entry Wave

<i>Entry wave and characteristic</i>	<i>Small (1)</i>	<i>Regular (2)</i>	<i>Regular-aide (3)</i>	<i>P value (4)</i>
<i>Students who entered STAR in kindergarten</i>				
Free lunch	0.47	0.48	0.50	0.46
White or Asian	0.68	0.67	0.66	0.66
Age in 1985	5.44	5.43	5.42	0.38
Female	0.49	0.49	0.48	0.87
Attrition rate	0.49	0.52	0.53	0.01
Days absent	10.00	10.50	10.90	0.01
Class size in kindergarten	15.10	22.40	22.80	0.00
Standardized test score in kindergarten	0.17	0.00	0.00	0.00

Sample Attrition

- Probably the most difficult practical obstacle facing experimental evaluators
- If attrition is random:
 - ▣ Sample size is reduced → variance of treatment estimator increases
 - ▣ Estimated treatment effects are unbiased
- If attrition is non-random:
 - ▣ Estimated treatment effects *may be* biased

Assessing the importance of non-random attrition

1. Check for high overall attrition rate (e.g. $>30\%$)
 2. Compare baseline characteristics of “attriters” and “non-attriters” → relevant for external validity
 3. Check for differences in attrition rates between treatment and control group
 4. Compare baseline characteristics of remaining treatment and control group → relevant for internal validity
- However, none of these tests are foolproof!

Estimating Equation

$$Y_{igs} = \beta_{0g} + \beta_{1g}SMALL_{is} + \beta_{2g}AIDE_{is} + \beta_{3g}\mathbf{X}_{is} + \alpha_{sw} + \varepsilon_{igs}$$

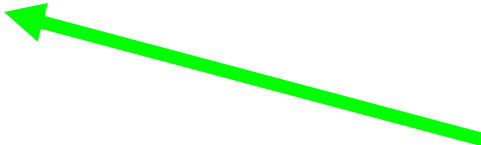
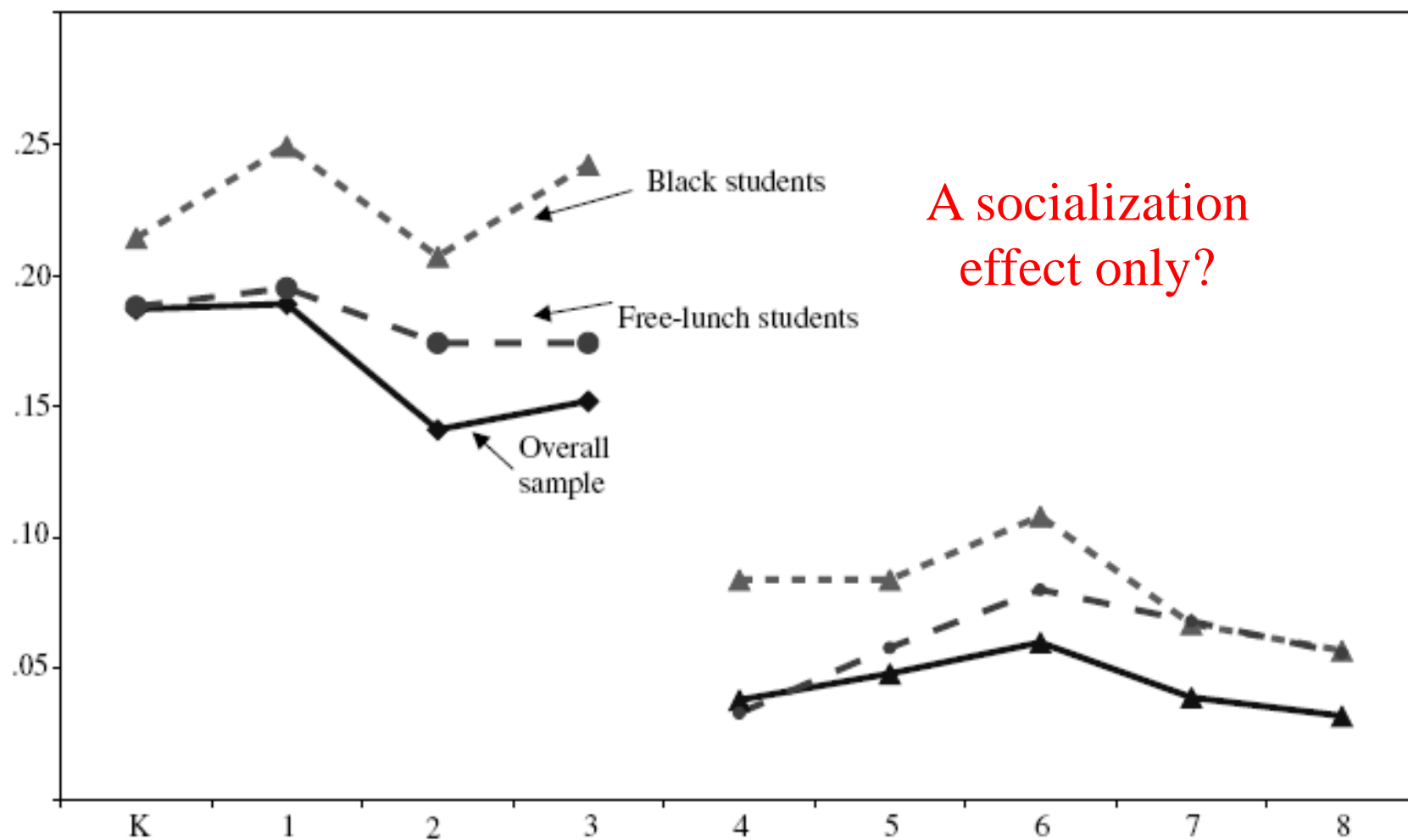
- Y = Test score
 - $Small$ = treatment indicator
 - $Aide$ = treatment indicator (for aide)
 - \mathbf{X} = vector of control variables
 - α = vector of dummy variables for each school-by-entry-wave combination
 - ε = zero-mean error term
- Impact Estimate
- 

Table 4. Small Class-Size Effects on Test Scores

<i>Subgroup</i>	<i>Kindergarten</i> <i>(1)</i>	<i>First grade</i> <i>(2)</i>	<i>Second grade</i> <i>(3)</i>	<i>Third grade</i> <i>(4)</i>
<i>Overall</i>	0.187 (0.039)	0.189 (0.035)	0.141 (0.034)	0.152 (0.030)
<i>Race</i>				
Black	0.214 (0.074)	0.249 (0.063)	0.207 (0.054)	0.242 (0.060)
White	0.172 (0.042)	0.161 (0.040)	0.105 (0.042)	0.115 (0.034)
<i>Free-lunch eligibility</i>				
Free lunch	0.188 (0.046)	0.195 (0.042)	0.174 (0.041)	0.174 (0.039)
No free lunch	0.177 (0.051)	0.194 (0.047)	0.126 (0.047)	0.118 (0.041)
<i>Gender</i>				
Male	0.209 (0.041)	0.192 (0.040)	0.144 (0.039)	0.172 (0.400)
Female	0.157 (0.049)	0.180 (0.040)	0.132 (0.042)	0.122 (0.040)

Figure 1. Small-Class Impacts on Test Scores

Effect size (standard deviation units)



Expectancy Effects 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

Follow-up Studies

- Outcomes studied:
 - ▣ Test scores through 8th grade
 - ▣ SAT/ACT participation
 - ▣ SAT/ACT scores (conditional on participation)
 - ▣ Criminal arrest data
 - ▣ Non-cognitive skills (engagement, initiative, and disruptive behavior)
- Cost-benefit analyses based on earnings benefits of improved cognitive skills suggest modest (positive) returns

Other Research using Project STAR Data

- The fact that both teachers and students were randomly assigned to classes is very useful!

- A few of the issues studied:
 - ▣ Teacher effectiveness
 - ▣ Peer effects
 - ▣ Relative age of students
 - ▣ Having a teacher of the same race

Statistical Power and Sample Size

- How many participants do you need to select into the sample? How many is enough?
- To answer this, you conduct a statistical power analysis
- Statistical Power definition: the probability that we will reject the null hypothesis of zero impact ($H_o: \mu_t - \mu_c = 0$).
- OR, the probability that, assuming the true impact is as large as expected, the estimated impact will be significantly different from zero

Key Terms


- α -level: significance level used to test null hypothesis
- Type I (α) Error: Rejecting a true null hypothesis
- β -level: probability of failing to reject the null hypothesis when the true impact equals K [where $K \neq 0$]
- Type II (β) Error: Not rejecting a false null hypothesis

$$\text{Power} = 1 - \beta$$

(β depends on α , K , sample size, and outcome variability)

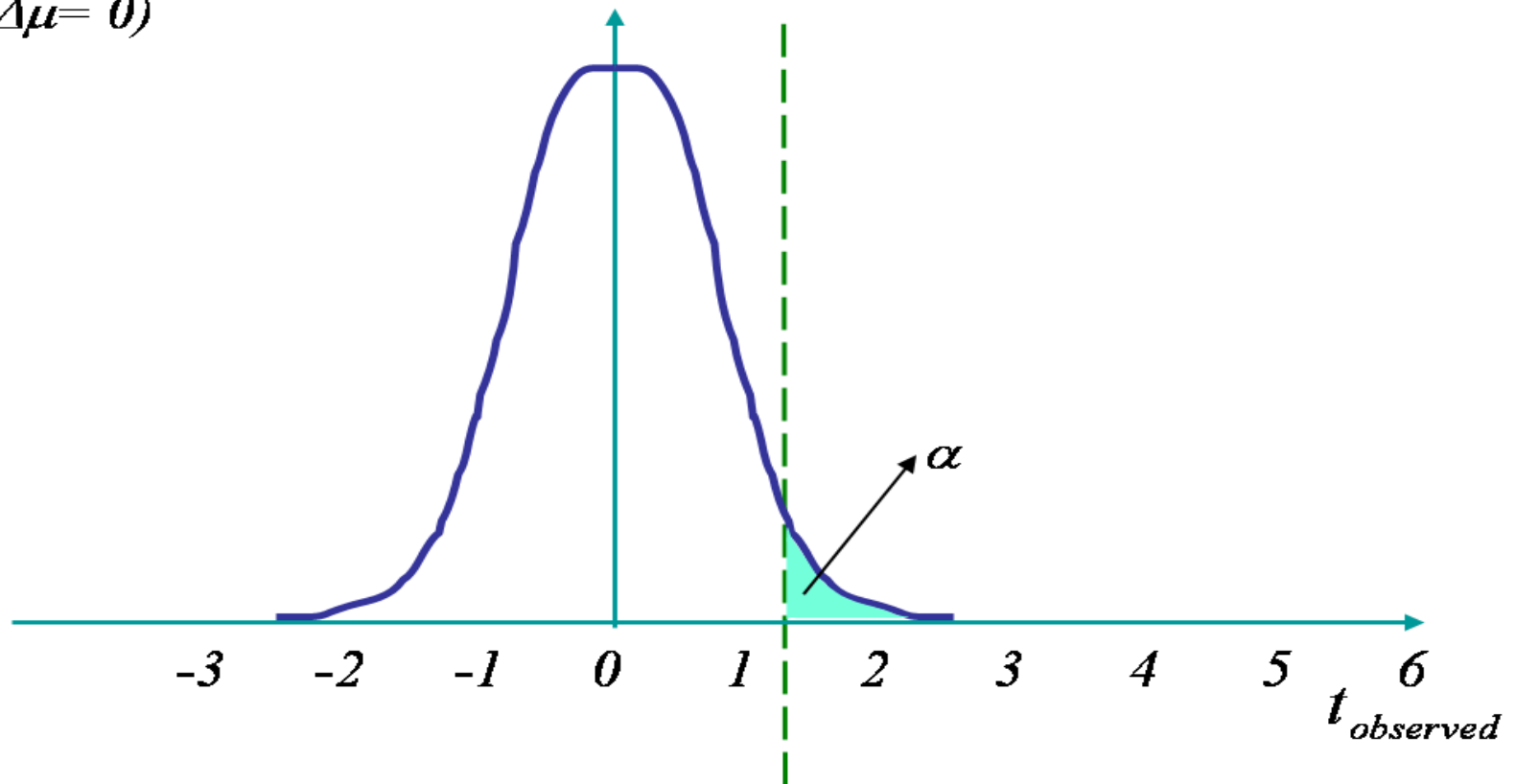


value of the true impact

 = statistical difference
at 0.05 level

0.05 α -level

Hypothetical Null Population
($H_0: \Delta\mu = 0$)



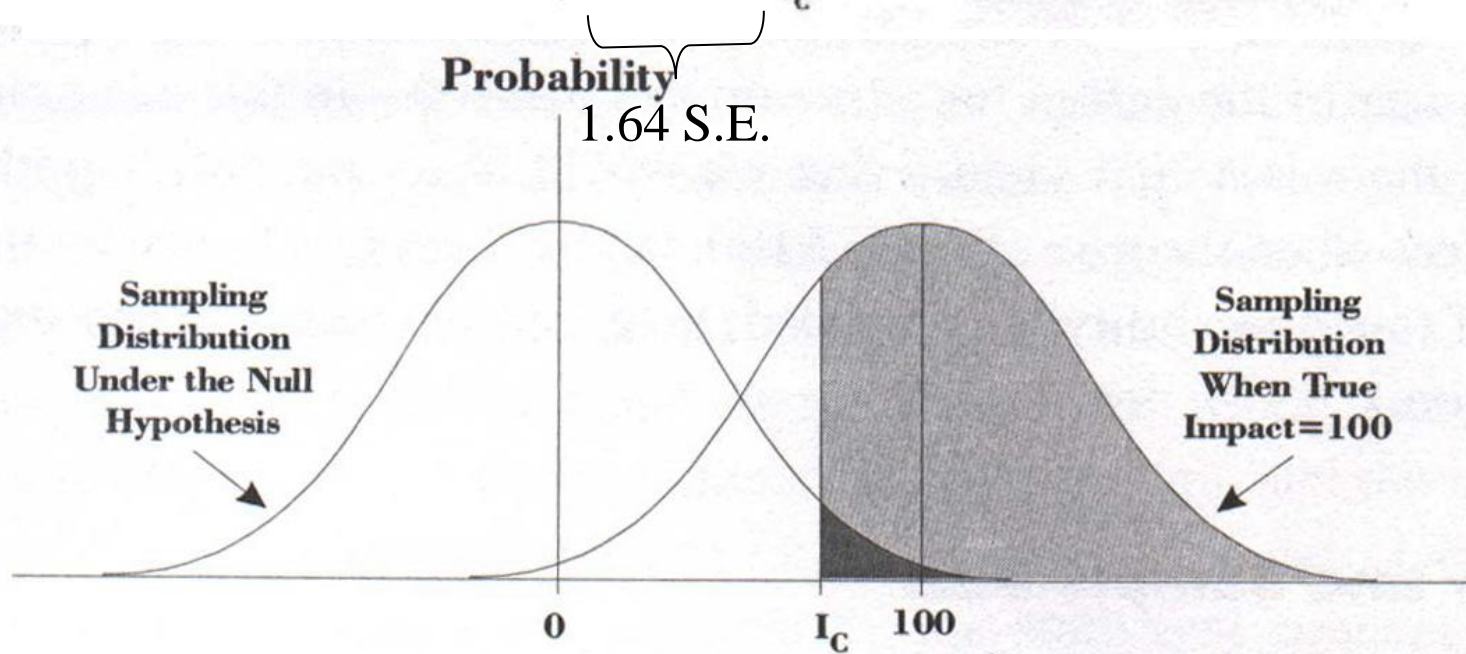
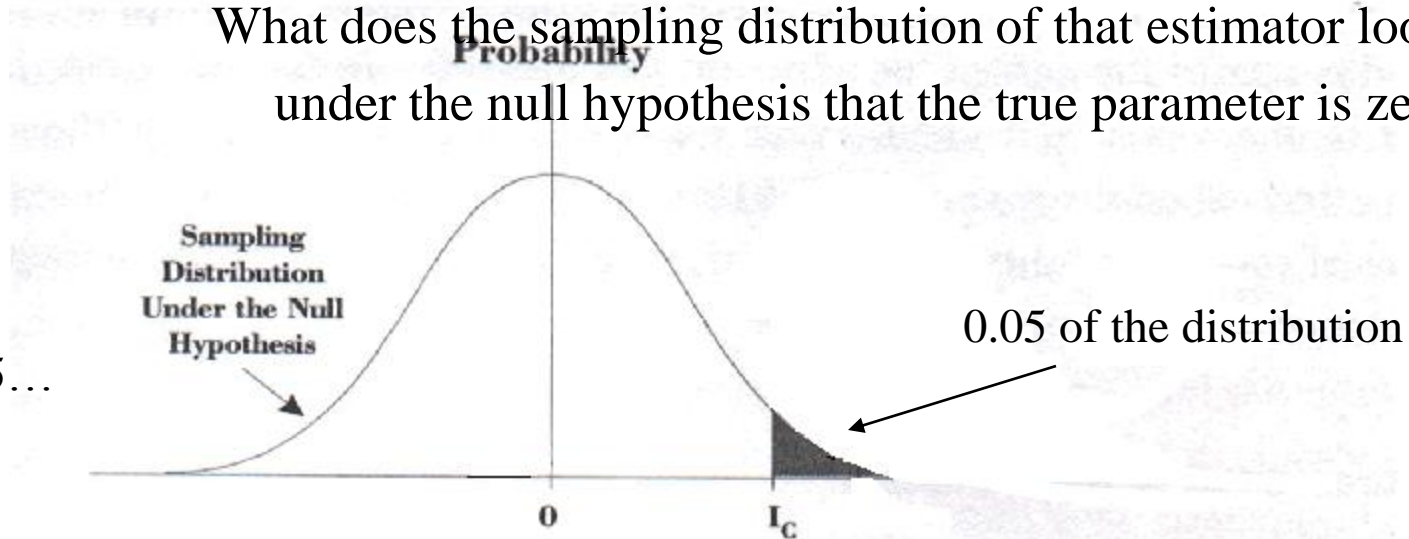
One-sided and Two-Sided T-tests

- One-sided test should only be used if you are certain of the direction in which the treatment affects the outcome.
- Two-sided test: No assumed directionality. Treatment could effect outcome either positively or negatively.

Consider an estimator of some parameter your interested in.

What does the sampling distribution of that estimator look like under the null hypothesis that the true parameter is zero?

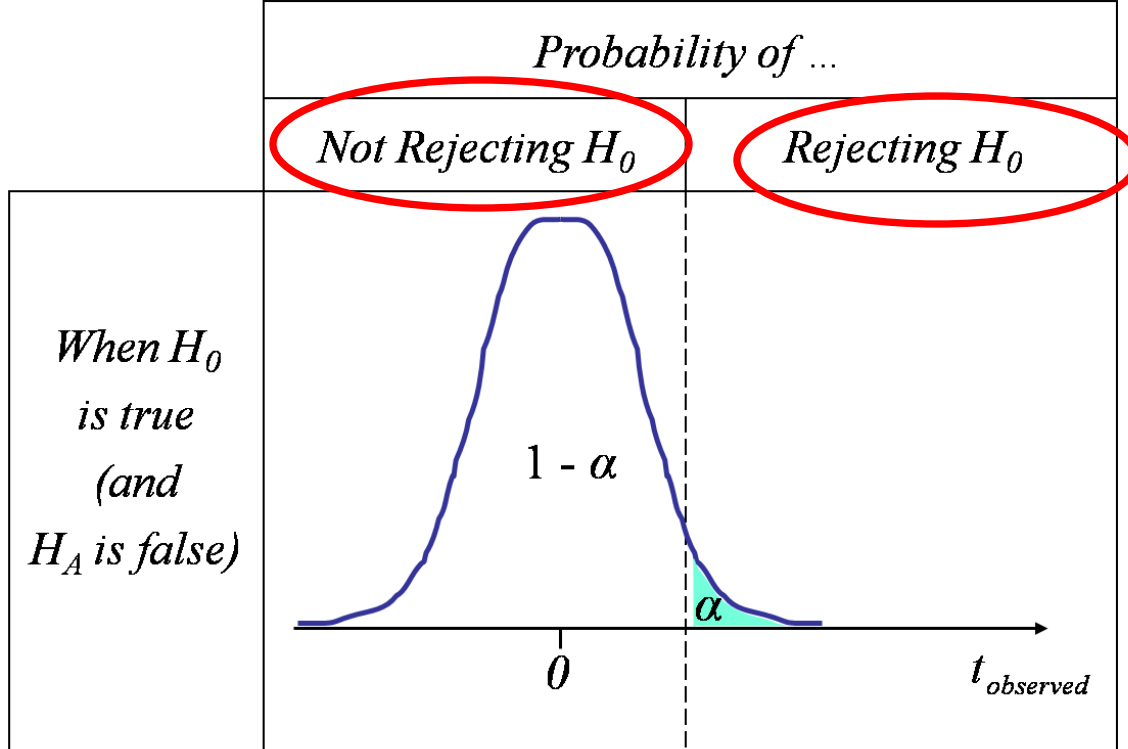
If $\alpha = .05 \dots$



The power of the design (for $K=100$) is the probability that the estimate will fall within the shaded area under the sampling distribution on the right.

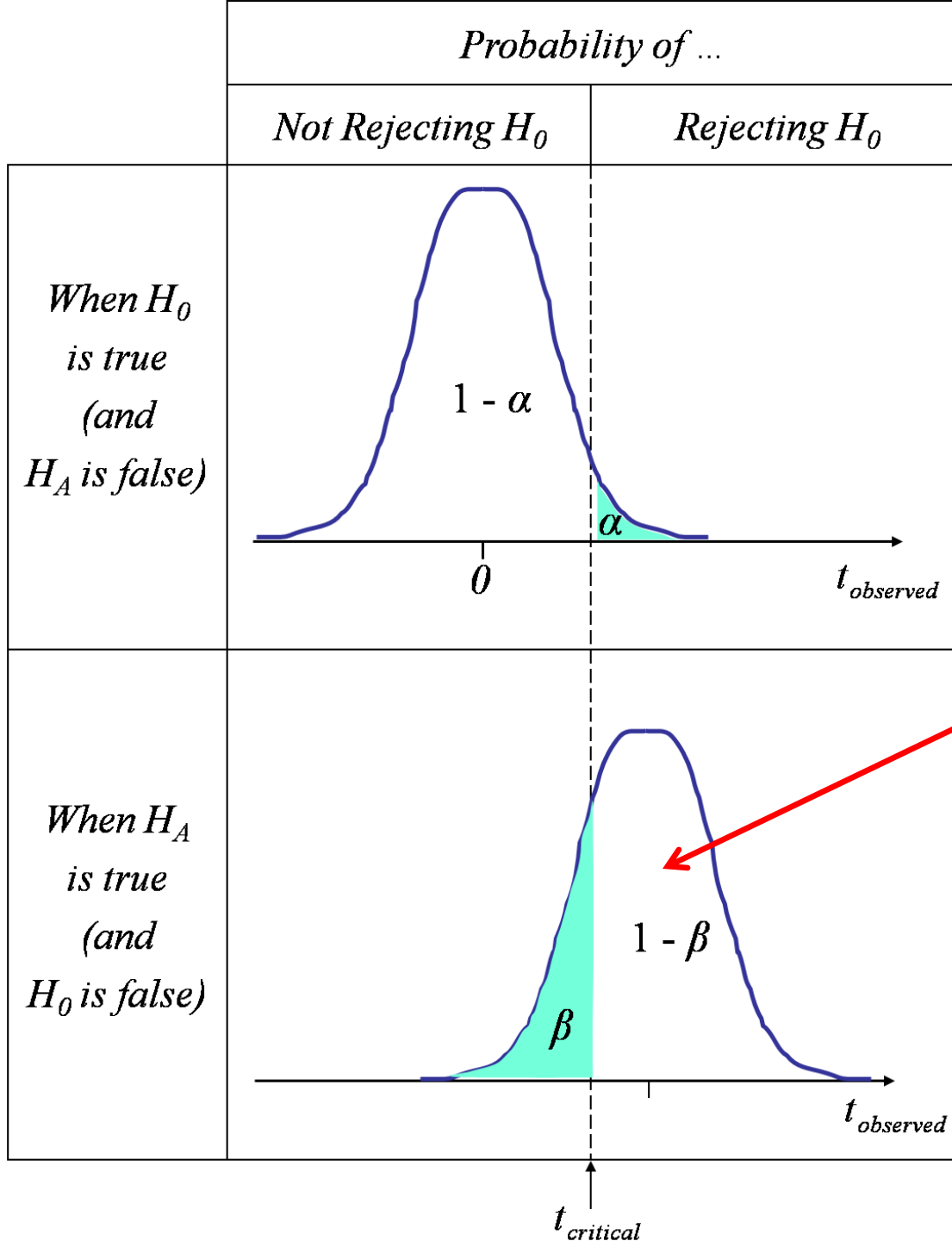
Type I and II Errors

		Observed in Sample	
		<i>Fail to Reject H_0</i>	<i>Reject H_0</i>
True State of Affairs	H_0 is True	Correct decision (1- α)	Type I Error (α)
	H_A is True	Type II Error (β)	Correct decision (1- β)



Type I and II Errors

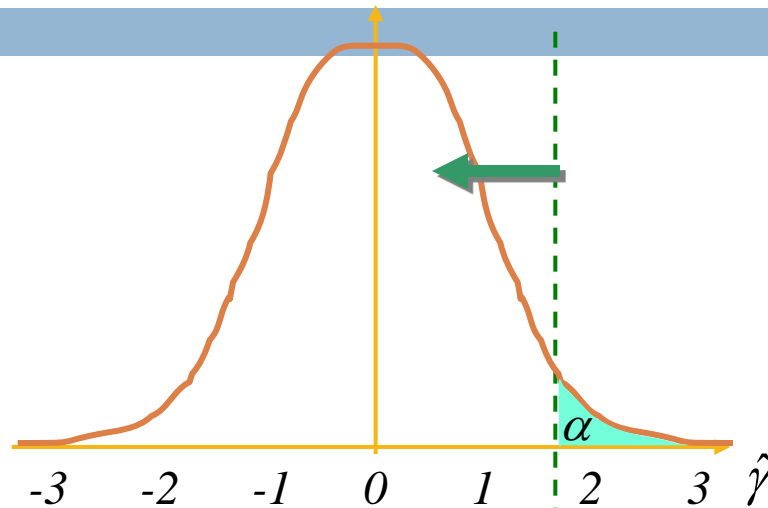
		Observed in Sample	
		<i>Fail to Reject H_0</i>	<i>Reject H_0</i>
True State of Affairs	<i>H_0 is True</i>	Correct decision (1- α)	Type I Error (α)
	<i>H_A is True</i>	Type II Error (β)	Correct decision (1- β)



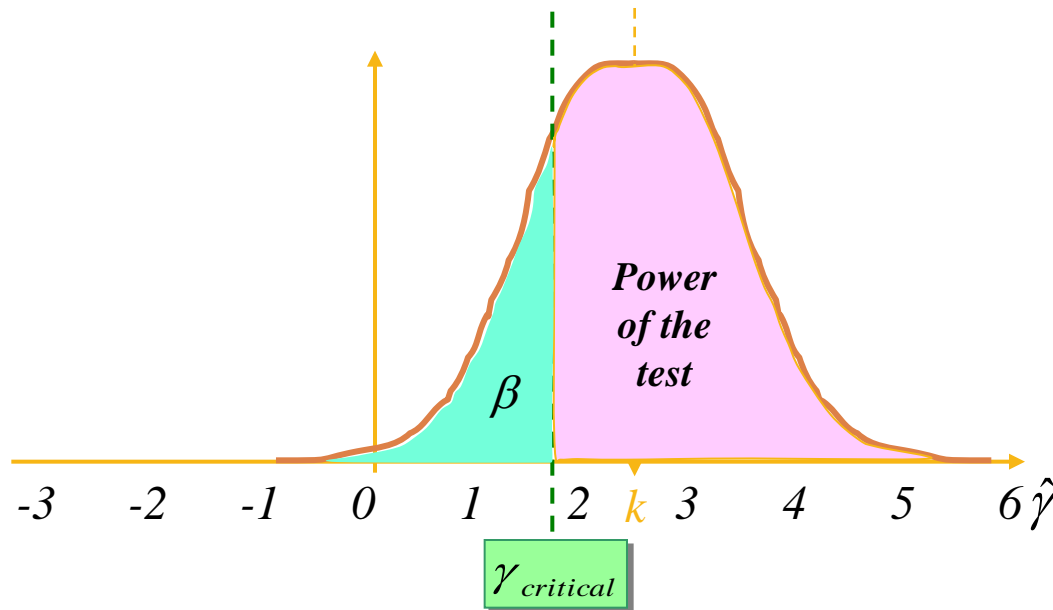
Question:
Which do
you think is
worse? Type
I error or
Type II error
and why?

Power = $1 - \beta$
Or, 1 minus Type
II error

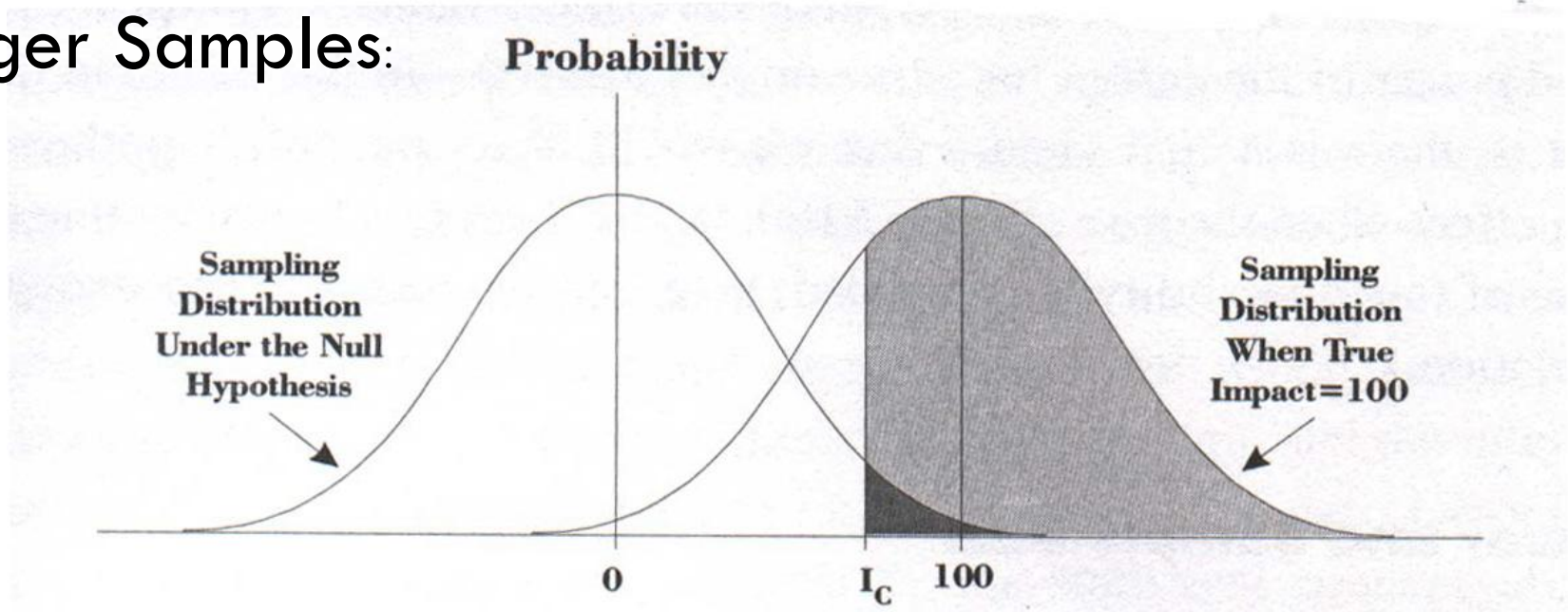
Higher α levels...



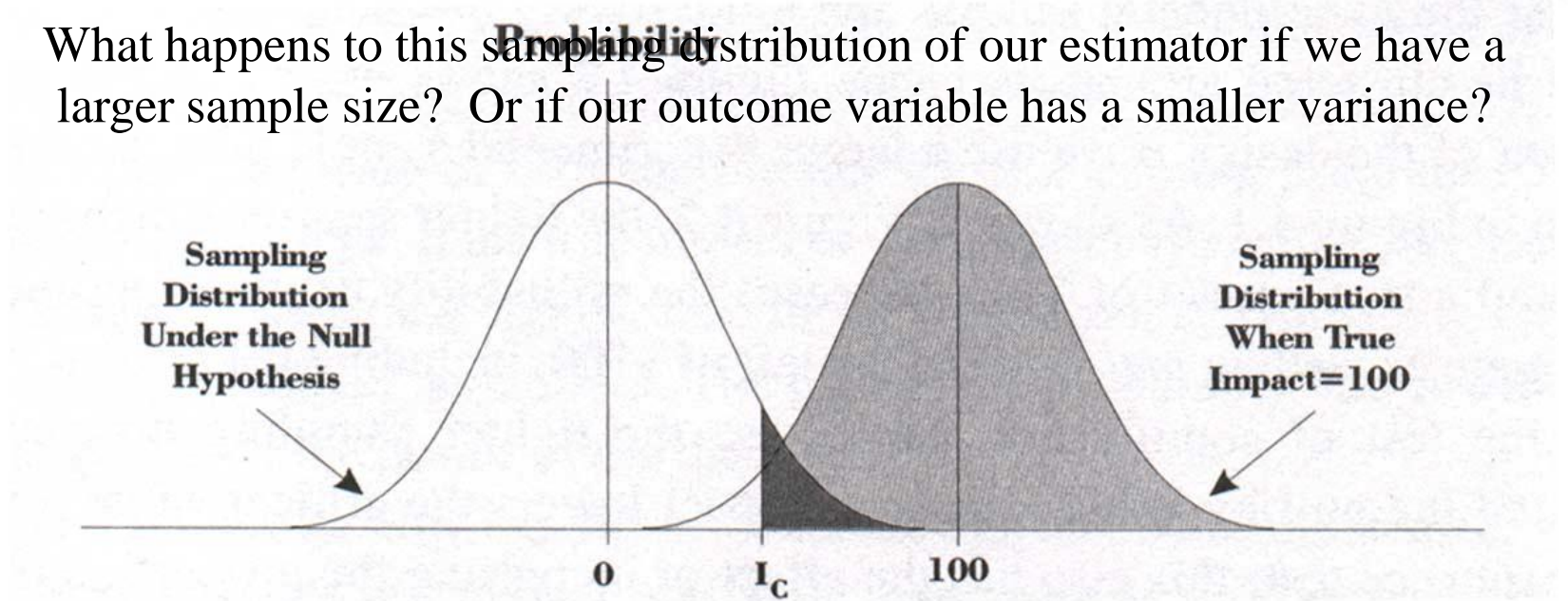
- More Type I error, but also
- More Power (less Type II error)



Larger Samples:



What happens to this sampling distribution of our estimator if we have a larger sample size? Or if our outcome variable has a smaller variance?



Power and Sample Size

TABLE 8.2. HOW MANY STUDENTS SHOULD YOU SELECT? SOME BALLPARK ESTIMATES OF TOTAL SAMPLE SIZE.

Type of effect size	Statistical test used	Statistical power	Anticipated effect size		
			Small	Medium	Large
Correlation coefficient	Pearson correlation	.90	1,047	113	37
		.80	783	85	28
		.70	616	67	23
Standardized mean difference	Two-group t-test	.90	1,052	170	68
		.80	786	128	52
		.70	620	100	40

Note: Two-

If you want *more power*, you need a *bigger sample*.

†Total sample size

If you are trying to detect a *smaller effect*, you need a *bigger sample*.

Minimum Detectable Effects (MDE)

MDE: the smallest true impact that would be found to be statistically significant with a given α and β

$$MDE = z \sqrt{\frac{V_Y}{n_t} + \frac{V_Y}{n_c}}$$

...where $z = f(\alpha, \beta)$

If MDE is smaller than an effect you would consider to be important, then it is important to increase the evaluation's power.

Ways to boost statistical power

1. Set higher alpha-level
 2. Increase sample size
 - Or improve treatment-control balance
 3. Reduce variance of outcome
 - Improve measurement
 - Use regression analysis
- Is too much power ever a bad thing?

Selecting the Sample

- Simple Random Sampling- Assign a random *ID* to each person in the sampling frame and draw a sample of n people.
- Stratified Random Sampling- Cross-tabulate members of the target population into strata by dimensions, like gender and SES. Draw a simple random sample within each stratum. Over- and under-sample selected strata as needed.
- Multi-site Cluster Sampling- If subjects are clustered naturally within identifiable sites, you first draw a random sample of sites. Then, within each site, draw a simple (or a stratified) random sample of participants.

Computer Assignment #1

- Simulated Data on housing voucher program
- Work on your own or in groups of 2-3 (no more than 3 people in a group)
- Key tasks:
 - ▣ Analyze treatment-control balance
 - ▣ Estimate program effects using mean differences and regression
 - ▣ See how the power of an evaluation varies with both sample size and the variance of the outcome measure
- Assignment, data, and Stata guide available on website
- Due Tuesday, March 2nd

Preview for Next Week



Tuesday= No Class

Thursday= Extensions of OLS

Reading:

Stock and Watson Chapter 11 or any basic
econometrics text with a section on probit, logit
regression