SECTION 1: REVIEW SESSION

March 3, 2010

Plan for Today

- Discuss Computer Assignment 1, and in the process review Section 1
- Interpreting Regression Output
- Preview Section 2

Variance of the treatment estimator

$$SE_I = \sqrt{V_I}$$

$$V_I = \frac{V(Y)}{n_T} + \frac{V(Y)}{n_C}$$

T-statistic:

$$t = \frac{I}{\sqrt{V_I}}$$

Implications:

1. V_I increases with V(Y).
2. V_I decreases with n_t and n_c.
3. V_I decreases with the balance between n_t and n_c.

Treatment Effects

- Individual Treatment Effect (ITE)
  - ITE = $Y_{i1} - Y_{i0}$
- Average Treatment Effect
  - ATE= E[Yi(1) – Yi(0)]
- Population Average Treatment Effect

$$\widehat{ATE} = \left(\frac{\sum_{i=1}^{n_1} Y_i}{n_i}\right) - \left(\frac{\sum_{i=1}^{n_0} Y_i}{n_0}\right)$$

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…*

Minimum Detectable Effects (MDE)

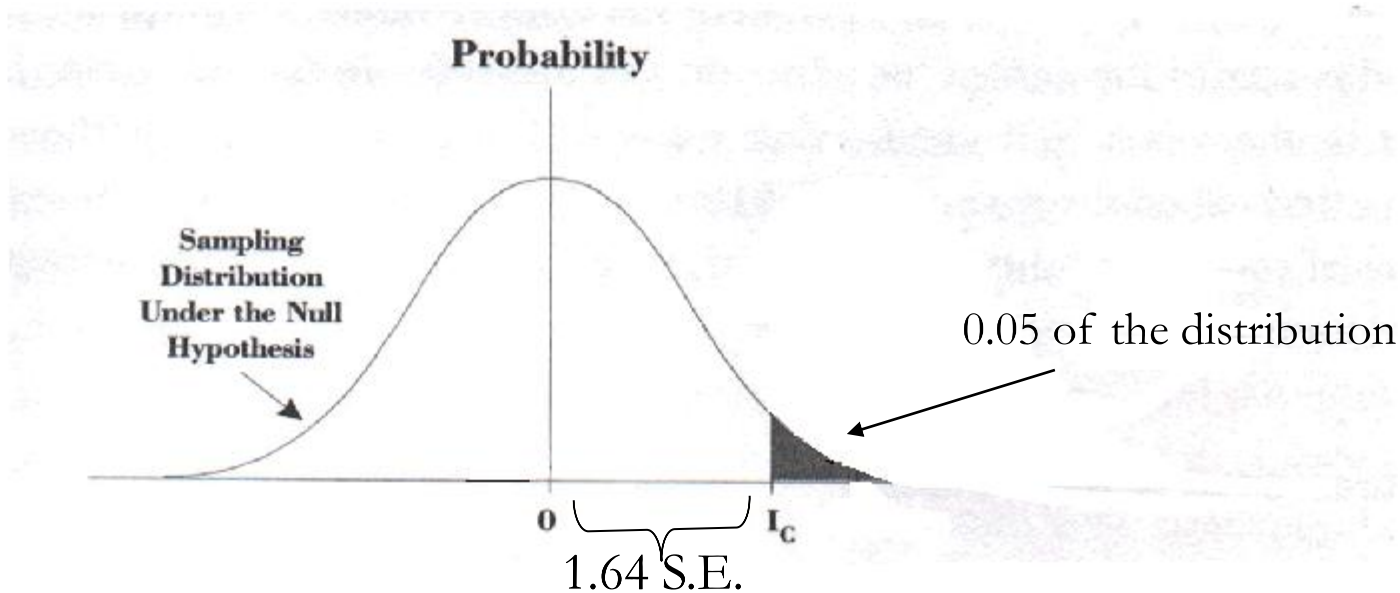
MDE: the smallest true impact that would be found to be statistically significant with a given α and β

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

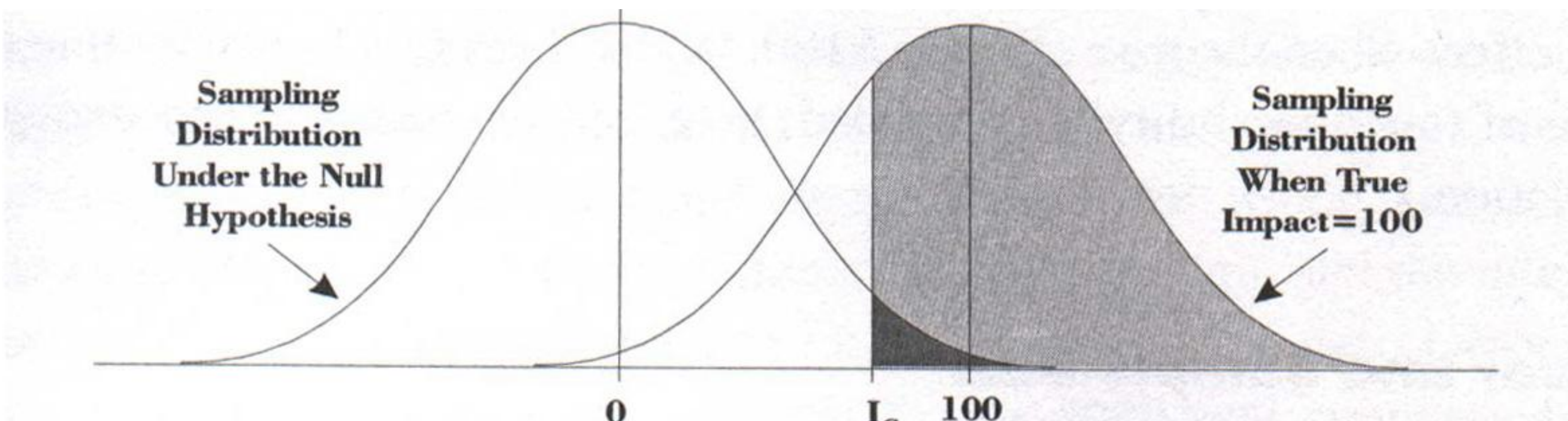
…where $k = f(\alpha,\beta)$

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

If α =.05…



To calculate the power of the design to detect effects of a given size—say, 100—we must determine the probability that the estimated impact will lie within the critical region when the true impact is 100.



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.

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.

Making statistical inferences in impact evaluation

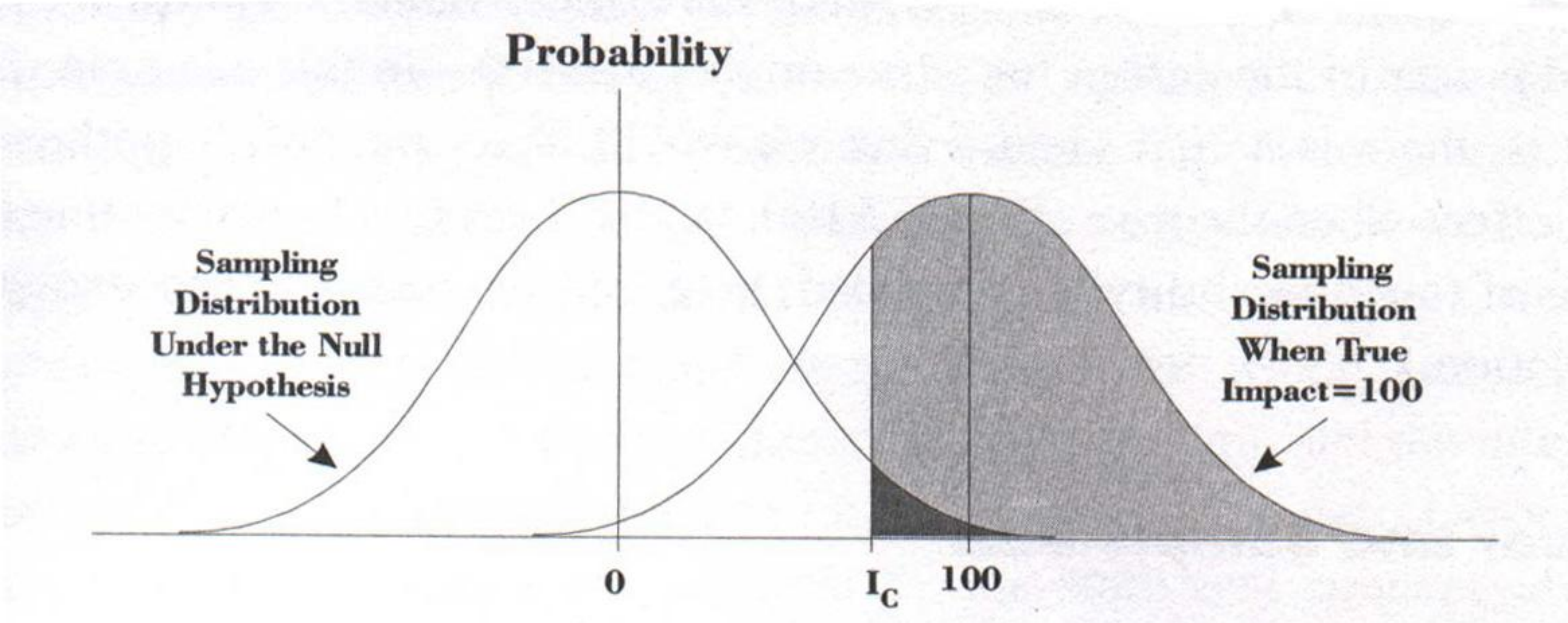
- Null Hypothesis, H_0 ($\beta_1 = 0$)

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

Null and Alternative Hypotheses

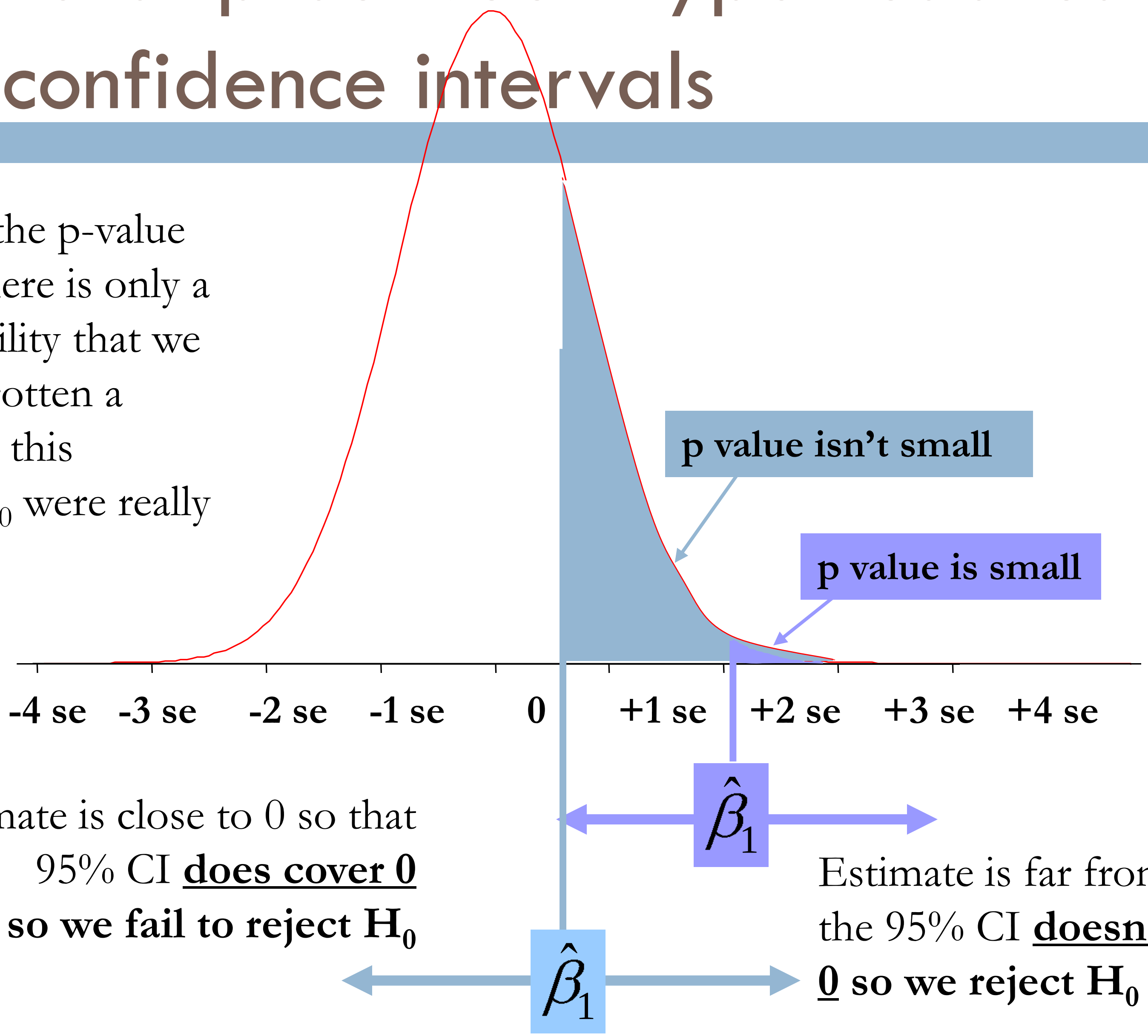


Null
$\beta_1 = 0$

Alternative
$\beta_1 = 100$

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).

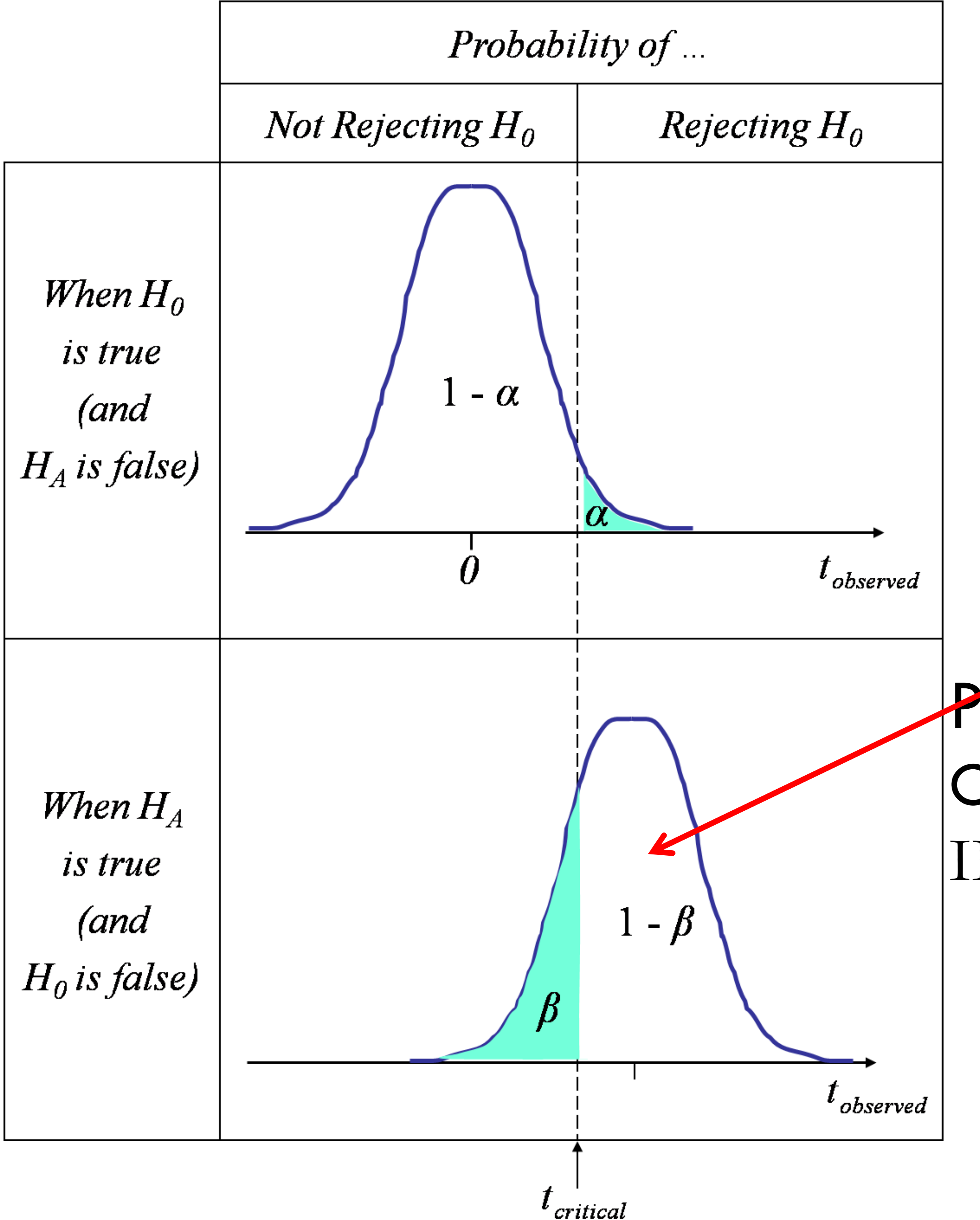


Estimate is close to 0 so that 95% CI **<u>does cover 0</u>** **so we fail to reject H_0**

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

Type I and II Errors

		Observed in Sample	
		Fail to Reject H_0	*Reject H_0*
True State of Affairs	*H_0 is True*	Correct decision $(1-\alpha)$	Type I Error (α)
	H_A is True	Type II Error (β)	Correct decision $(1-\beta)$

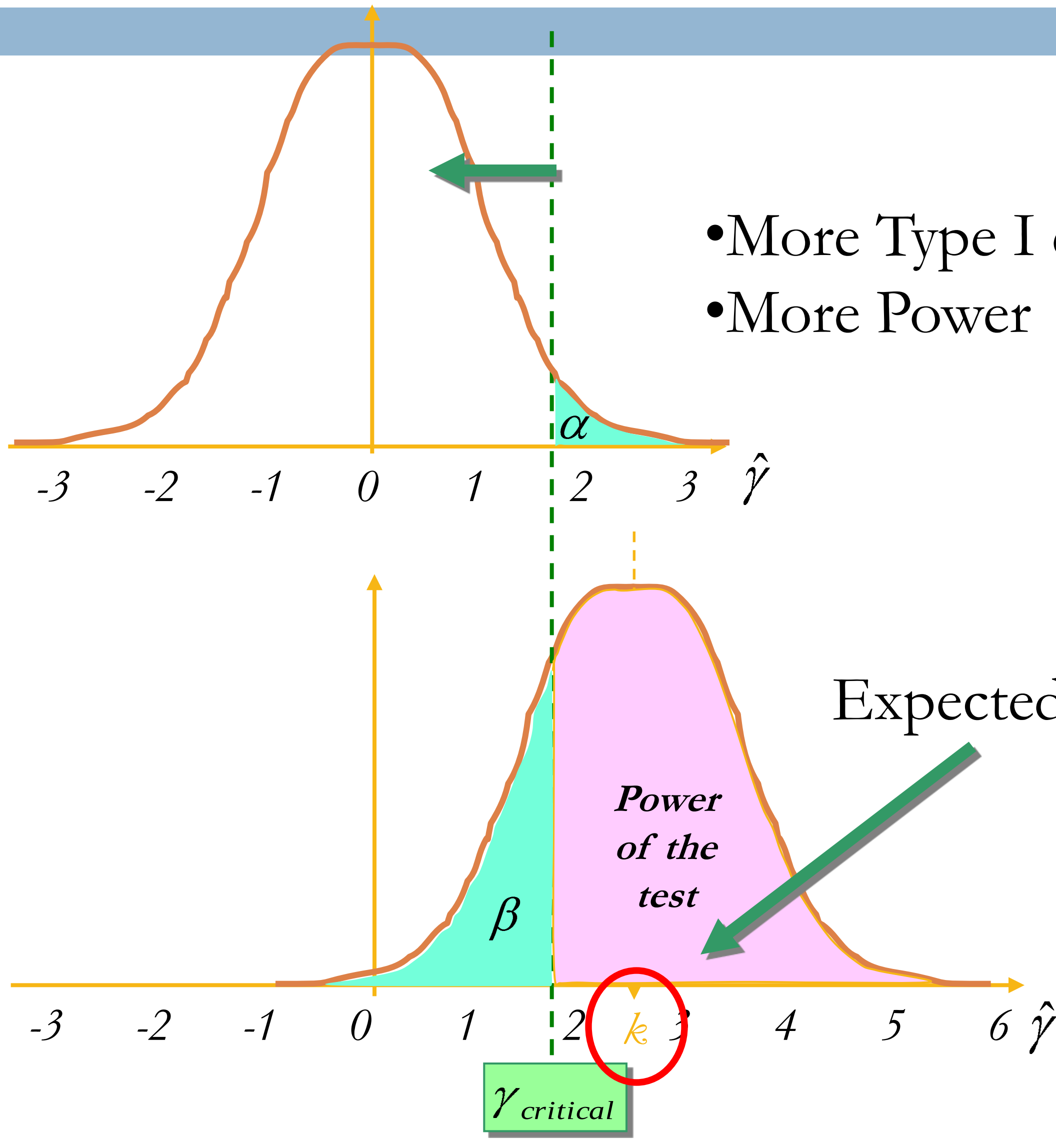


Power= 1-β

Or, 1 minus Type II error

Higher α levels…

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



Multivariate Regression: the relationship between x and y ***controlling*** for other, related factors

- "The purpose of incorporating relevant covariates into an analysis of experimental data is to reduce residual variation, decrease standard errors, and increase statistical power.
- Include covariates which:
  - Do not vary over time
  - Are measured prior to random assignment (cannot include variables that may have been affected by participation in the experiment- these are exogenous)

Adding covariates to the model

- Before:

$$Y_i = \beta_0 + \beta_1 T_i + \boxed{\varepsilon_i}$$

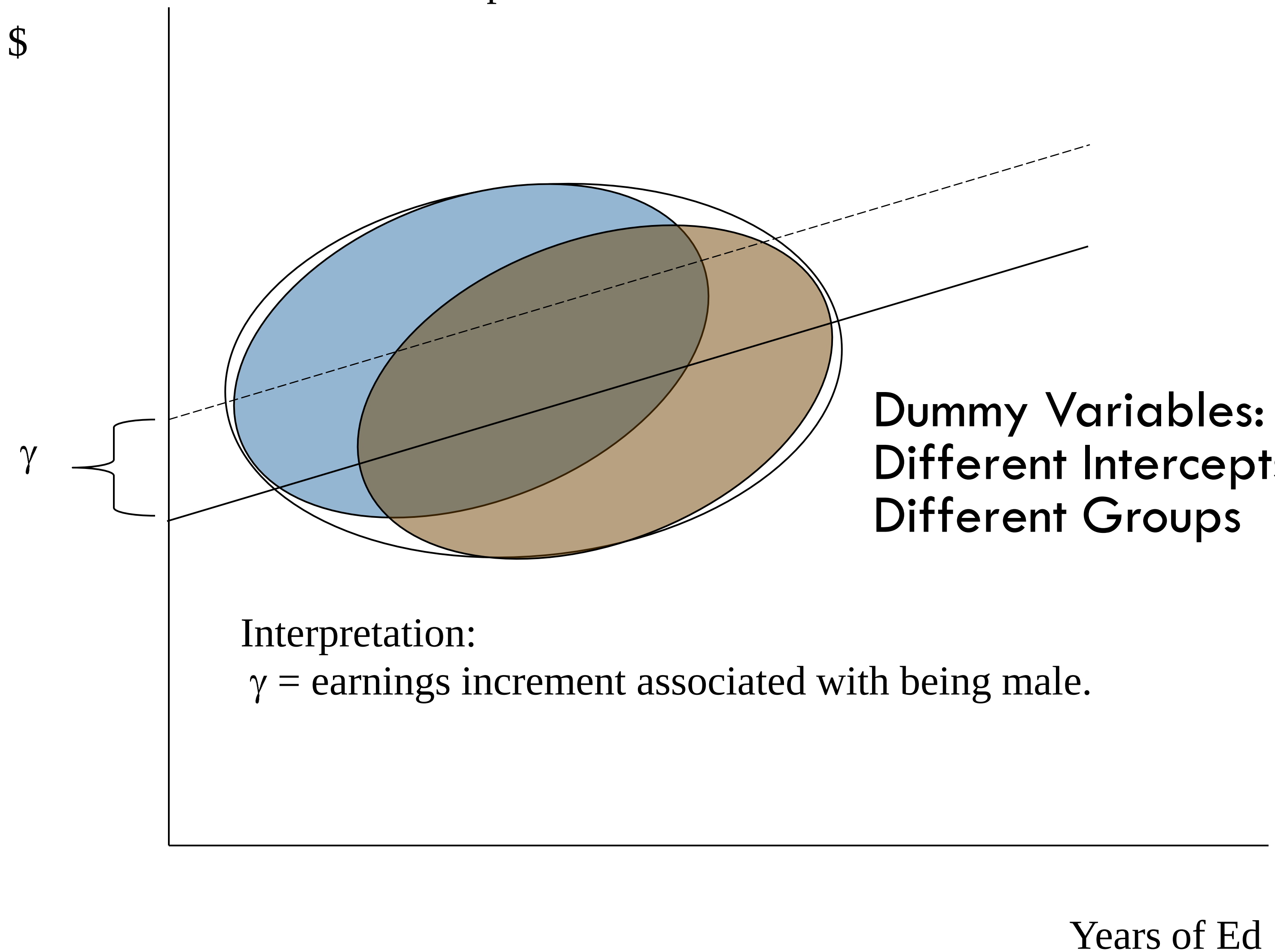
When you add control predictor, Z, the part of Y that is now being explained by Z must have been contained, before, in the residual, ε. Adding a successful covariate causes the residual to shrink

- After:

$$Y_i = \beta_0 + \beta_1 T_i + \boxed{\beta_2 Z_i + \varepsilon_i}$$

Dummy Variables: Allowing Different Intercepts for Different Groups

The correct specification, then, is:
$y = \alpha + \beta x + \gamma \text{male} + \varepsilon$, where male = 1 if the ith person is a male and 0 if female.

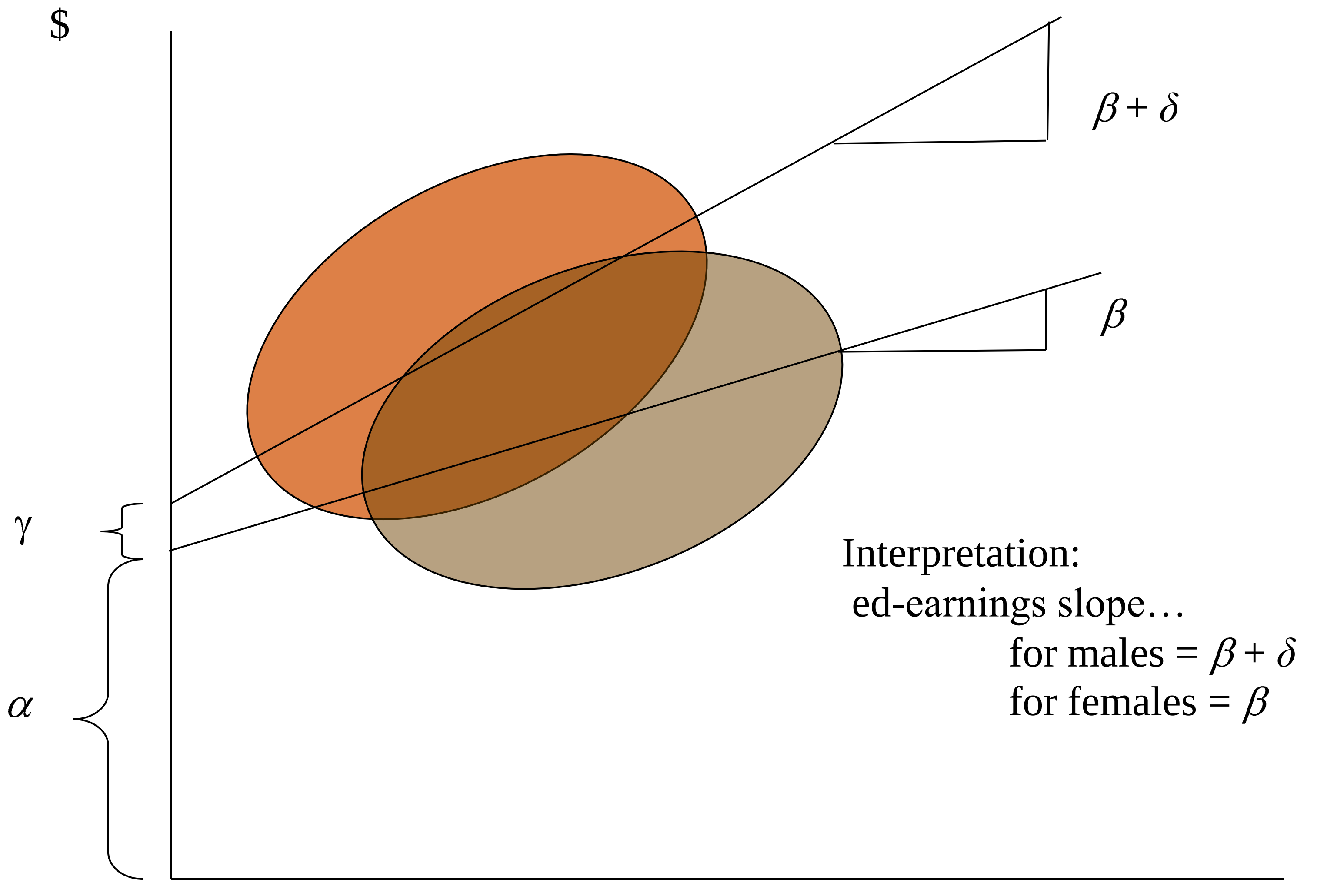


Interpretation:
γ = earnings increment associated with being male.

Interactions: Allowing Different Slopes for Different Groups

What if the relationship between education and earnings differs for *males* and females?

To allow for different slopes by gender, include a *male*-by-*x* interaction term in your regression model.



Interpretation:
ed-earnings slope…
for males = $\beta + \delta$
for females = β

$$y = \alpha + \beta x + \gamma \text{male} + \delta \text{male} \bullet x + \varepsilon$$

Preview of Section 2

Quasi-Experimental Design

- Differences-in-Differences
- Fixed Effects
- Instrumental Variables
- Regression Discontinuity
- Propensity Score Matching

Exam #1

- Available at 5pm today on mycourses
- Must be submitted by no on Monday (March 8th)
- You may use any notes or statistics texts to help you with this exam, but you may <u>not</u> consult outside materials or anyone else about the exam from the time you access it until you submit your final answers.