

# **PSY 503: Foundations of Statistical Methods in Psychological Science**

## **Power, Effect Size**

Suyog Chandramouli

311 PSH (Princeton University)

27th October, 2025

# Final Project Overview

# Course Project

- Due at end of semester (Dec 19<sup>th</sup>) \*
- Submit on Canvas, a link to the project on Github/OSF
- Goal:
  - Apply statistical knowledge from the class
  - Get started with analyzing your data (if applicable)
  - Develop ability to conduct a reproducible analysis

# Course Project: 3 components

- Reproducible Report
- APA-style Paper using papaja \*
- Simulation-based Power Analysis

\* - or apaquarto

# Part 1 - Reproducible Report

- Options:
  - Analyze your own data
  - Use existing public dataset
  - Reproduce analysis from published paper with open data
- Use regression methods from class or any other relevant analyses
- Report should be fully reproducible (e.g. R Markdown)

# Part 2 – APA paper

- Use papaja package in R Markdown
- Include R code chunks for analysis
- Compile to PDF in APA format
- Benefits
  - Have your text and code in the same place

# Part 3 - Power Analysis

- Add simulation-based power analysis to the APA paper
- Include:
  - R code for power analysis
  - Power curve graph
  - Discussion of results and implications for design

# Key Requirements

- Entire paper must be reproducible
- Have a classmate test reproducibility
- Focus on demonstrating skills, not full-length paper
- Cite sources appropriately

# Assignment (due Nov 1)

- Think about what project you'd like to work on
- Send a plan on Canvas
- If you are considering multiple options, outline them in this draft / contact me before that to discuss and narrow it down

Power

# Recap

- Null Hypothesis  
Alternative Hypothesis
- Type-I error  
Type-II error

# In-class exercise

A large nationwide poll recently showed an unemployment rate of 9% in the US. The mayor of a local town wonders if this national result holds true for her constituency. So, she plans on taking a sample of her residents to see if the unemployment rate is significantly different than 9% in her town.

Let  $p$  represent the unemployment rate in her town.  
Here are the hypotheses she'll use:

$$H_0: p = 0.09$$

$$H_a: p \neq 0.09$$

- When would the mayor commit a Type I error? Type II error?
  - A) She concludes the town's unemployment rate is not 9% when it actually is.
  - B) She concludes the town's unemployment rate is not 9% when it actually is not.
  - C) She concludes the town's unemployment rate is 9% when it actually is.
  - D) She concludes the town's unemployment rate is 9% when it actually is not.

# Types of Errors

Assume  $H_0$  is the “Null hypothesis”

False **positive** =  
Assuming  $H_0$  is false when it is true.

	Retain $H_0$	Reject $H_0$
$H_0$ is true	Correct decision	Type I Error ( $\alpha$ )
$H_0$ is false	Type II Error ( $\beta$ )	Correct decision

False **negative** =  
Assuming  $H_0$  is true when it is not.

***Positive/negative are in relation to  
the research hypothesis/  
alternate hypothesis***

# Recap

- Null Hypothesis  
Alternative Hypothesis

- Type-I error  
Type-II error

- Type-I error rate ( $\alpha$ )  
Type-II error rate ( $\beta$ )
  - Power:  $1 - \beta$



These are all probability  
values

# Recap

- Null Hypothesis  
Alternative Hypothesis

- Type-I error  
Type-II error

- Type-I error rate ( $\alpha$ )  
Type-II error rate ( $\beta$ )

- Power:  $1 - \beta$

- P(Detecting an effect when it really exists)

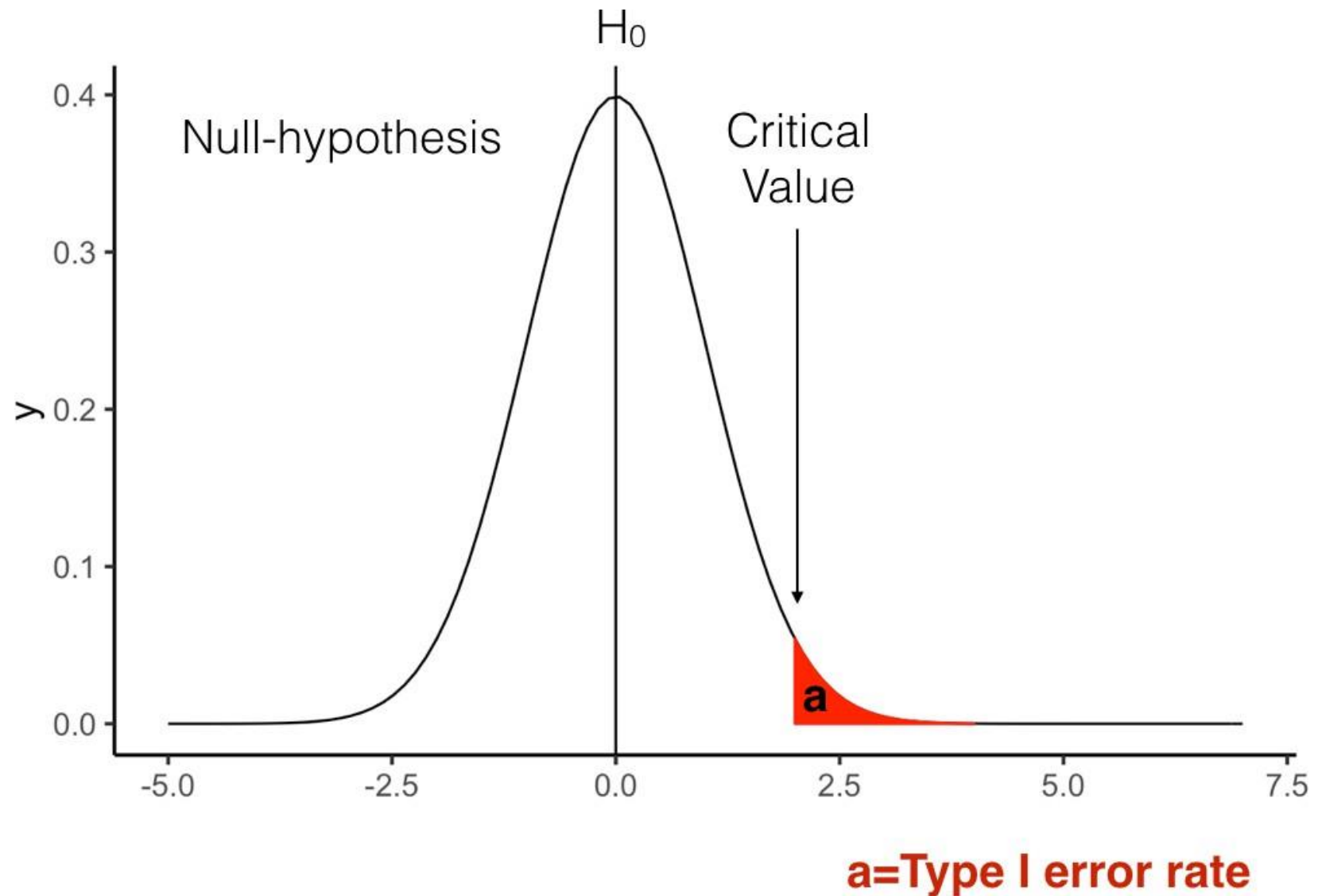


These are all probability  
values

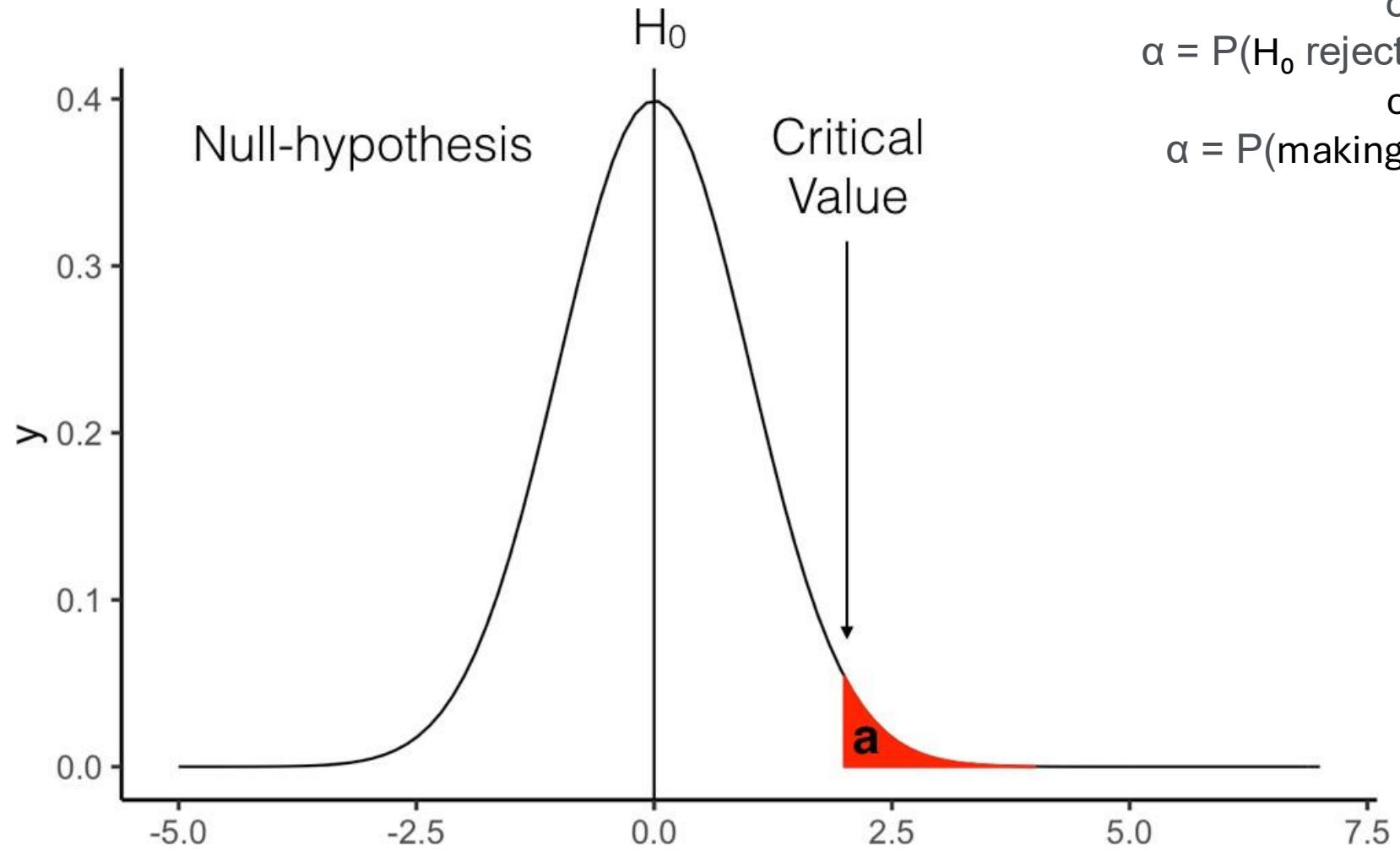
# Statistical Inference

- From the distribution of all possible samples from a DGP, we are looking at where our observed sample lies.
  - i.e. where does our sample lie in the sampling distribution
- If the probability of the sample  $< \alpha$  , e.g.  $\alpha = 0.05$ , we reject the Null
  - There is always a possibility of error
    - Wrongly, rejecting the Null (Type I)
    - Wrongly, failing to reject the Null (Type II)

# Type I error



# Type I error

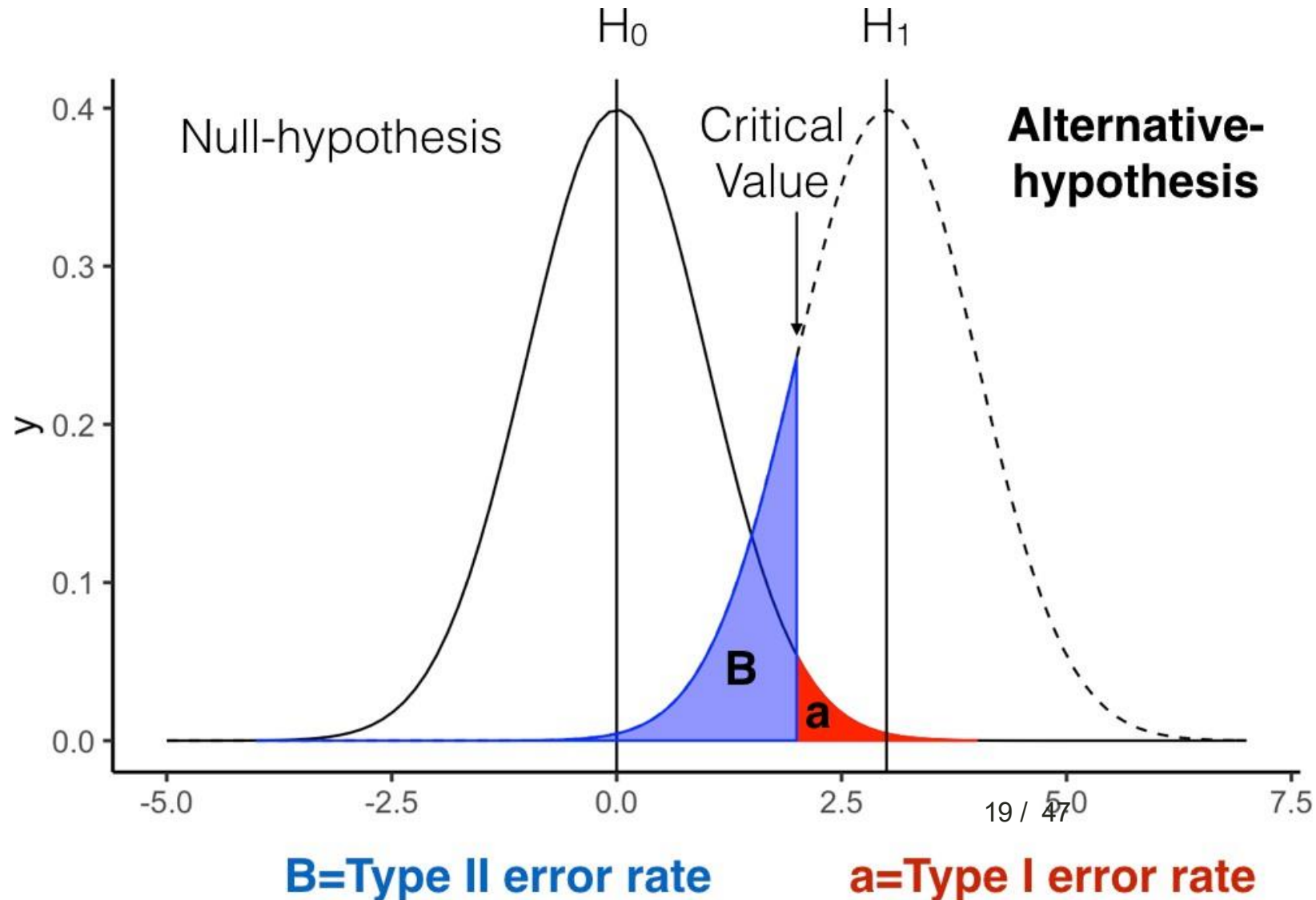


$\alpha = P(\text{wrongly rejecting the Null})$   
or  
 $\alpha = P(H_0 \text{ rejected} | \text{Truth} = H_0)$   
or  
 $\alpha = P(\text{making a type I error})$

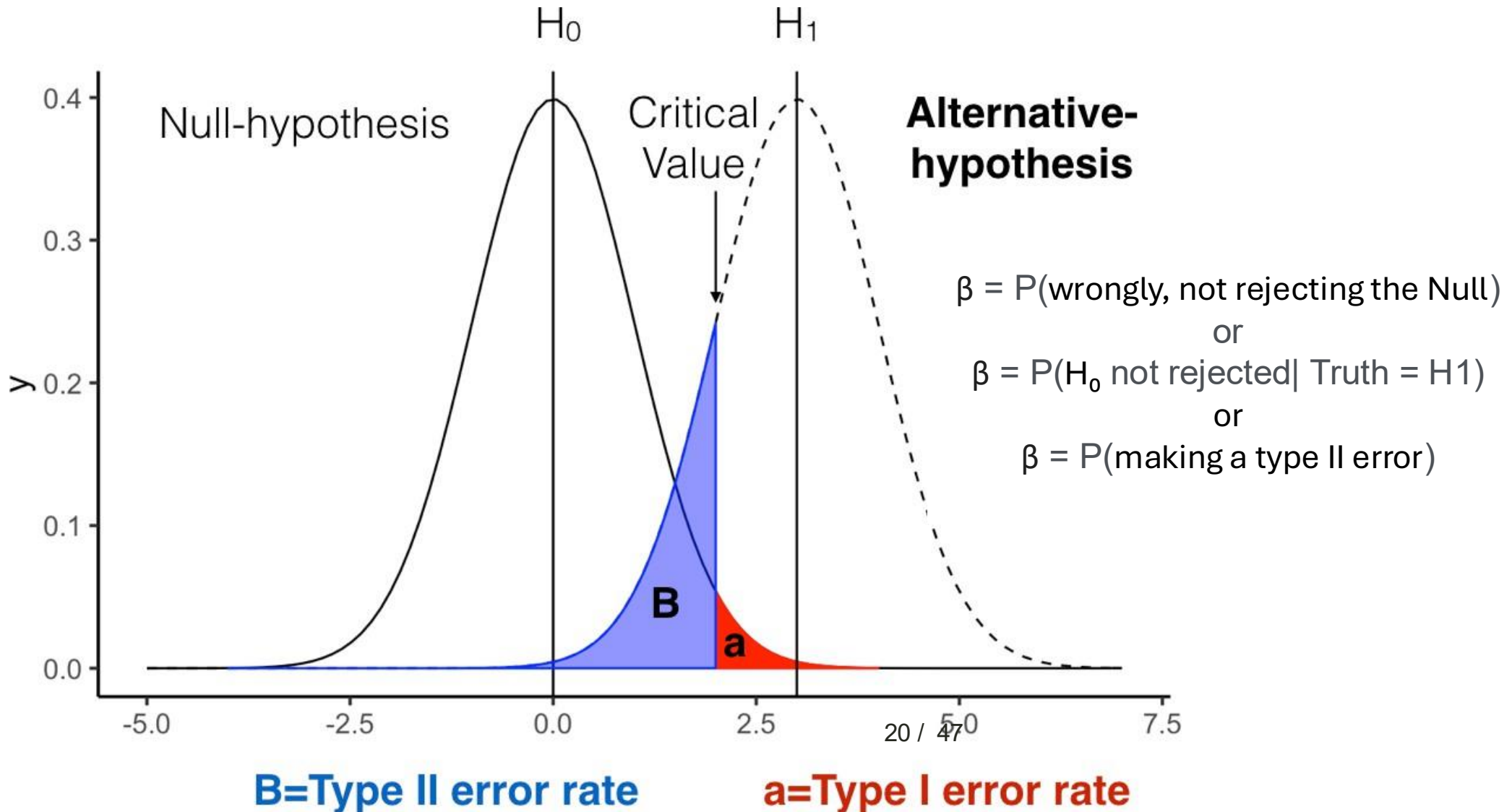
**a=Type I error rate**

# $\beta$ = type II error rate

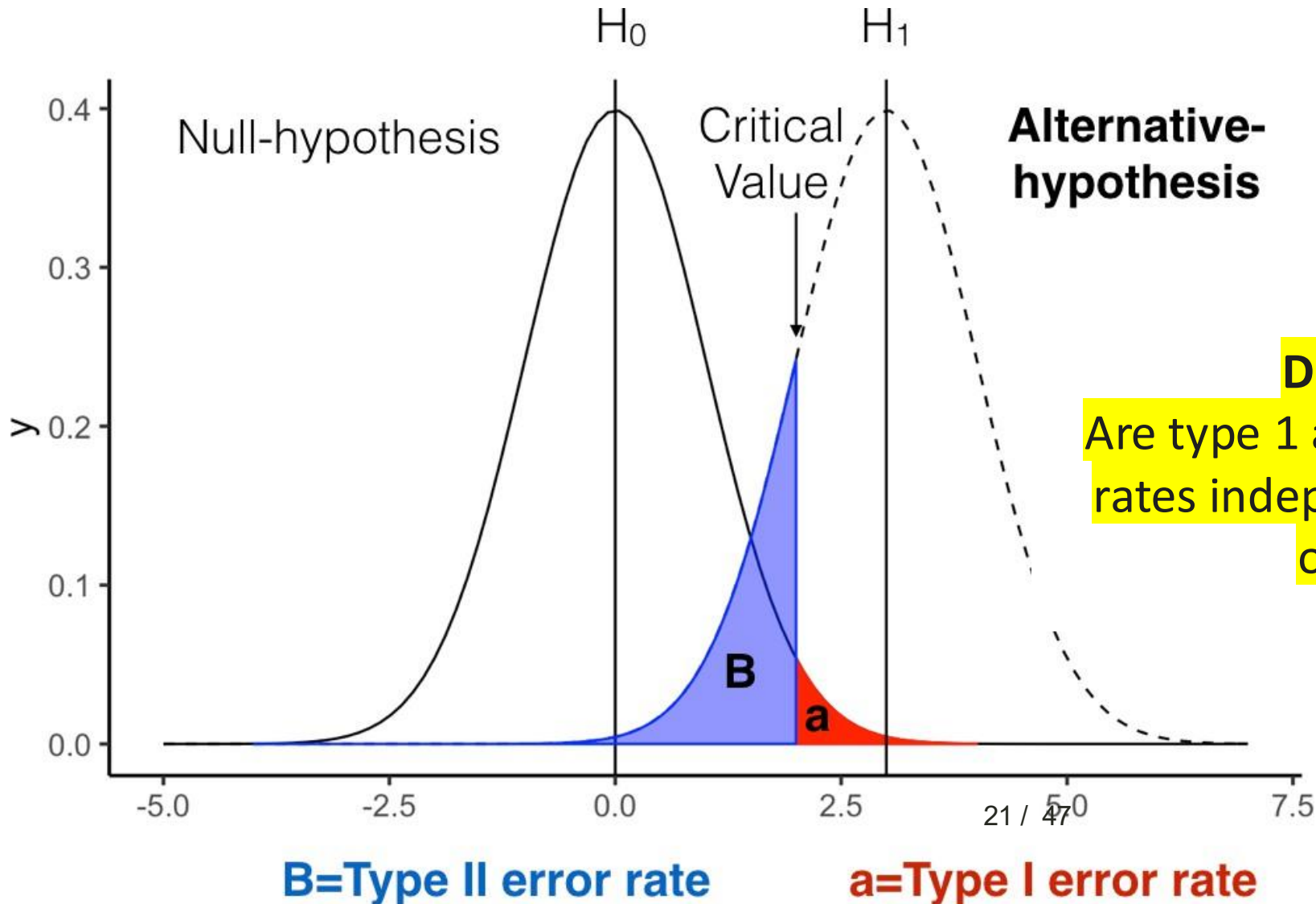
$$\beta = P(H_0 \text{ rejected} | \text{Truth} = H_0)$$



# Type-II error



# Type-II error

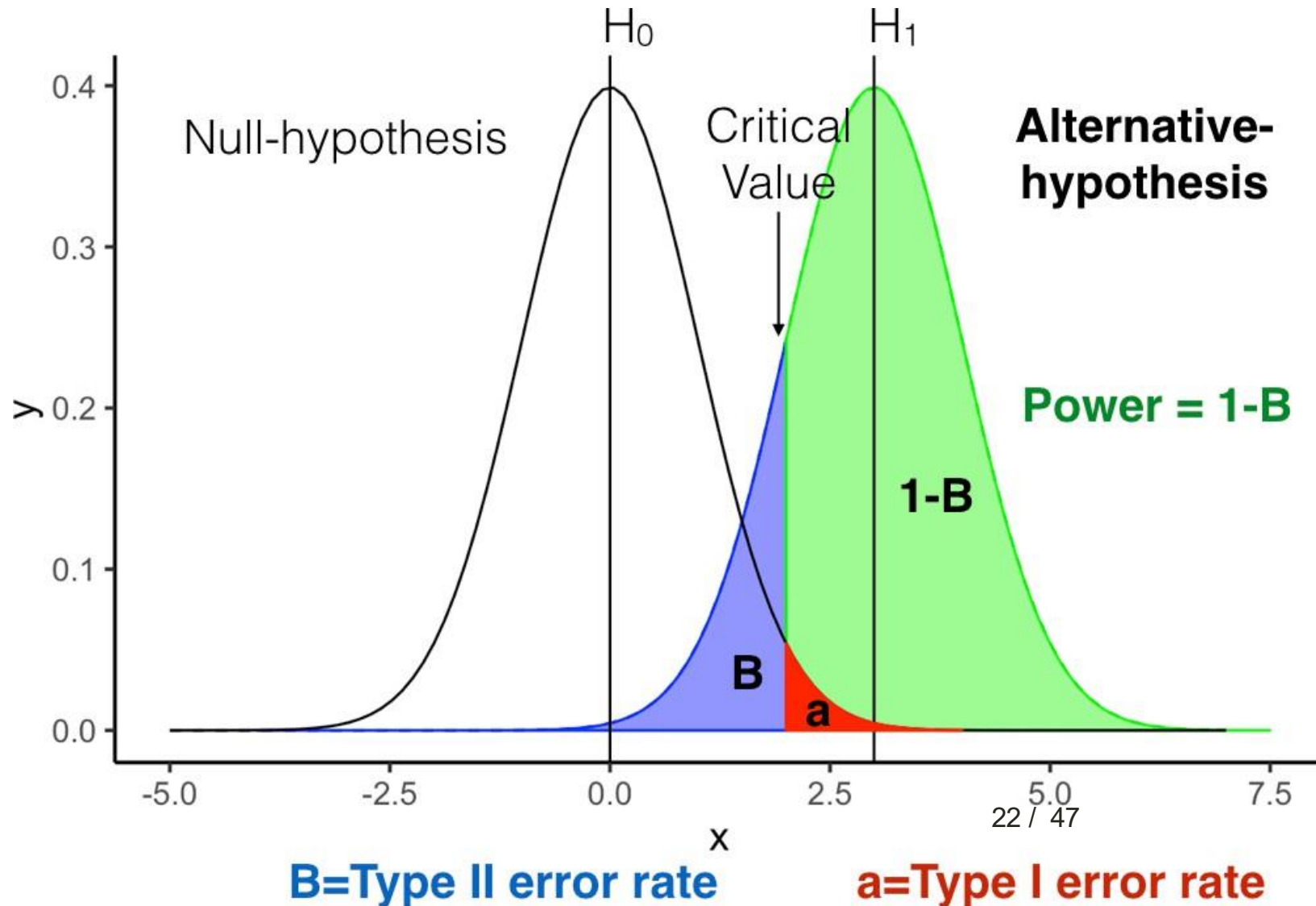


**Discuss:**

Are type 1 and type 2 error rates independent of each other?

# Power = 1-B

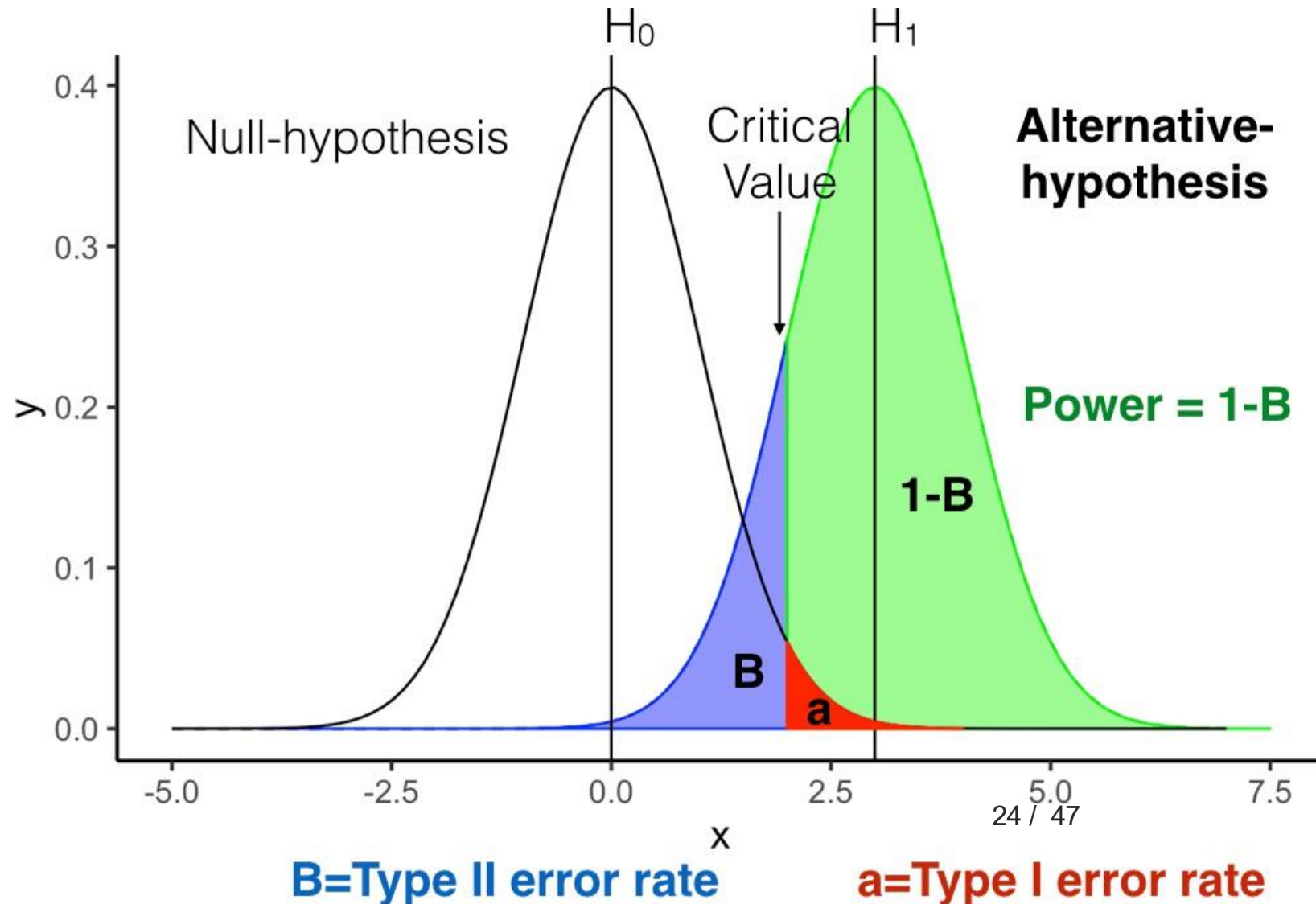
$$\text{Power} = 1 - \beta$$
$$= P(H_0 \text{ rejected} | \text{Truth} = H_1)$$



# Understanding Power

- Power is
  - $P(\text{Rejecting } H_0 | H_0 \text{ is False})$ 
    - Ability to find an effect (in the population) that actually exists
  - $1 - \beta$ 
    - $1 - P(\text{Not Rejecting } H_0 | H_0 \text{ is False})$
    - $P(\text{not making a type II error})$
- We ideally want to maximize the power of our test.
- Being a probability, Power values range between 0 and 1 (or 0 to 100%)
  - Power = .2,
    - You will reject the null-hypothesis 20% of the time (20/100 experiments)
  - Power = .8
    - (considered high power), You will reject the null-hypothesis 80% of the time (80/100 experiments)

# Discuss: Under what setting is power high? Why?



# Factors that impact power

- $\alpha$ , the significance level
  - Increases power
  - But increases type I error
- Variability of sampling distribution
  - $N$ 
    - Increases power
- Variability in data
  - Experimental design, good measurements
- Distance of the true parameter from  $H_0$ 
  - E.g. effect size,  $r$

# General info about power

1. Increasing sample-size, increases power
2. Increasing effect-size, increases power
3. Lowering alpha (making it easier to reject null), increases power
  - Power is about  $P(\text{rejecting null} \mid \text{null is false})$

# Power is a property of a design

- Every design has its own **Power** to detect effects of different sizes.
- The power of a design depends on:
  - sample-size ( $n$ )
  - Effect-size ( $d$ )
  - alpha-criterion

# Discuss

- Assume that we are dealing with coin tosses.
- We are hypothesizing about the fairness of a coin.
  - $H_0$  : the coin is fair ( $\theta = 0.5$ )
- For what type of coin is our hypothesis test the most powerful?

# Understanding Power

- Power varies based on the true parameter value
  - i.e. the parameters of the true 'DGP' that is producing the data
    - Hence, it's a characteristic of different scenarios for DGPs rather than the test
- Power depends on how far the true value is from the null hypothesis.

# The Power function

- Shows power as a function of different 'true' scenarios / true parameter values
- Properties: Power increases as parameter moves away from  $H_0$ 
  - Power decreases near the null hypothesis value
  - Forms a characteristic curve shape

# Power function for all thetas (for a given N)

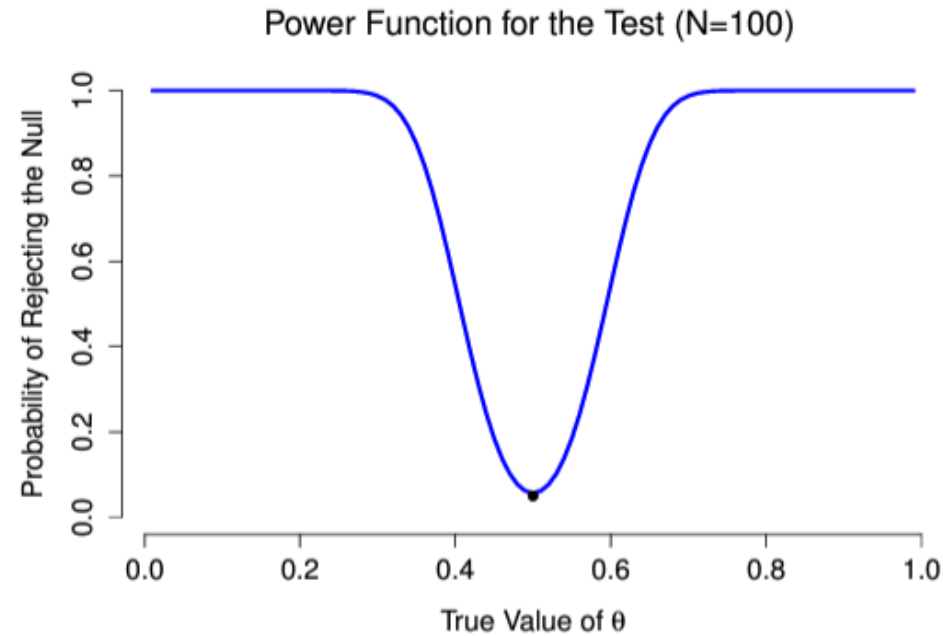
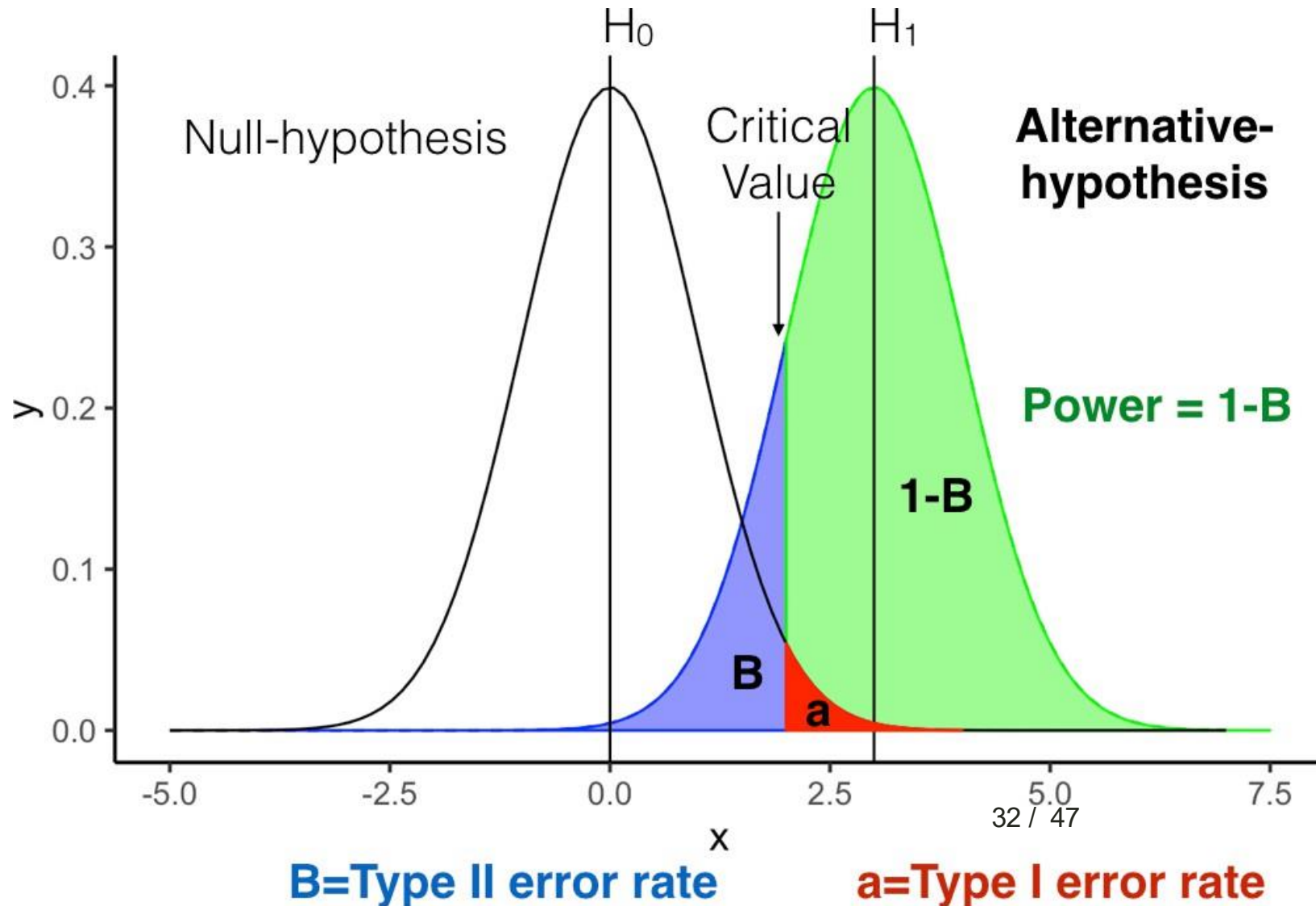


Figure 11.6: The probability that we will reject the null hypothesis, plotted as a function of the true value of  $\theta$ . Obviously, the test is more powerful (greater chance of correct rejection) if the true value of  $\theta$  is very different from the value that the null hypothesis specifies (i.e.,  $\theta = .5$ ). Notice that when  $\theta$  actually is equal to .5 (plotted as a black dot), the null hypothesis is in fact true: rejecting the null hypothesis in this instance would be a Type I error.

# Power = 1-B

$$\text{Power} = 1 - \beta$$
$$= P(H_0 \text{ rejected} | \text{Truth} = H_1)$$



# The Power function

- Shows power as a function of different 'true' scenarios / true parameter values
- Properties: Power increases as parameter moves away from  $H_0$ 
  - Power decreases near the null hypothesis value
  - Forms a characteristic curve shape

# Discuss

- Assume that we are dealing with coin tosses.
- We are hypothesizing about the fairness of a coin.
  - $H_0$  : the coin is fair ( $\theta = 0.5$ )
- ~~• For what type of coin is our hypothesis test the most powerful?~~
- ~~• How many coin tosses are required to confidently reject the null model?~~
- Say, you are working with alternate hypothesis of
  - $H_1$  : The coin is always Heads
  - $H_1'$  : The coin has more Heads than Tails
- Does this have a bearing on power?

# Power vs **sample size** (for a given true theta)

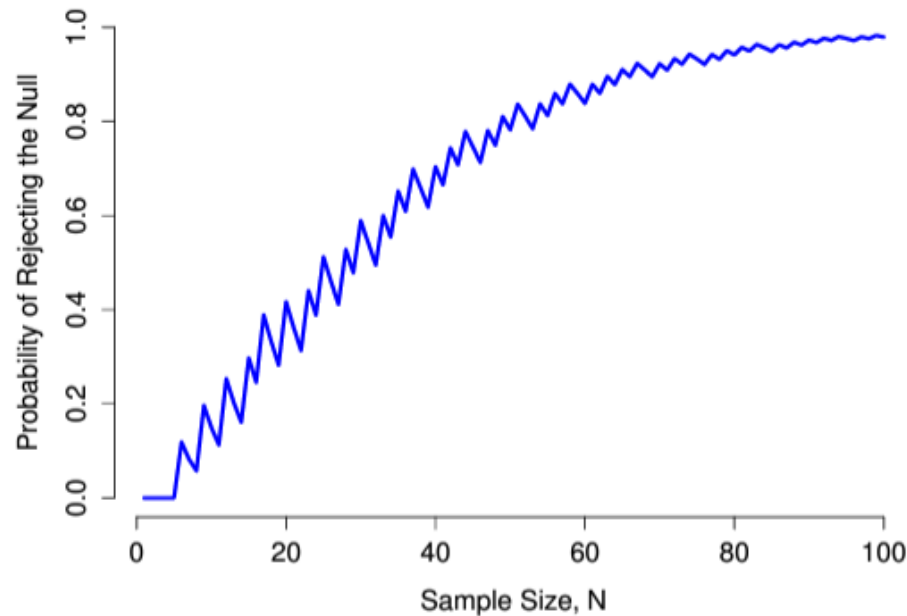
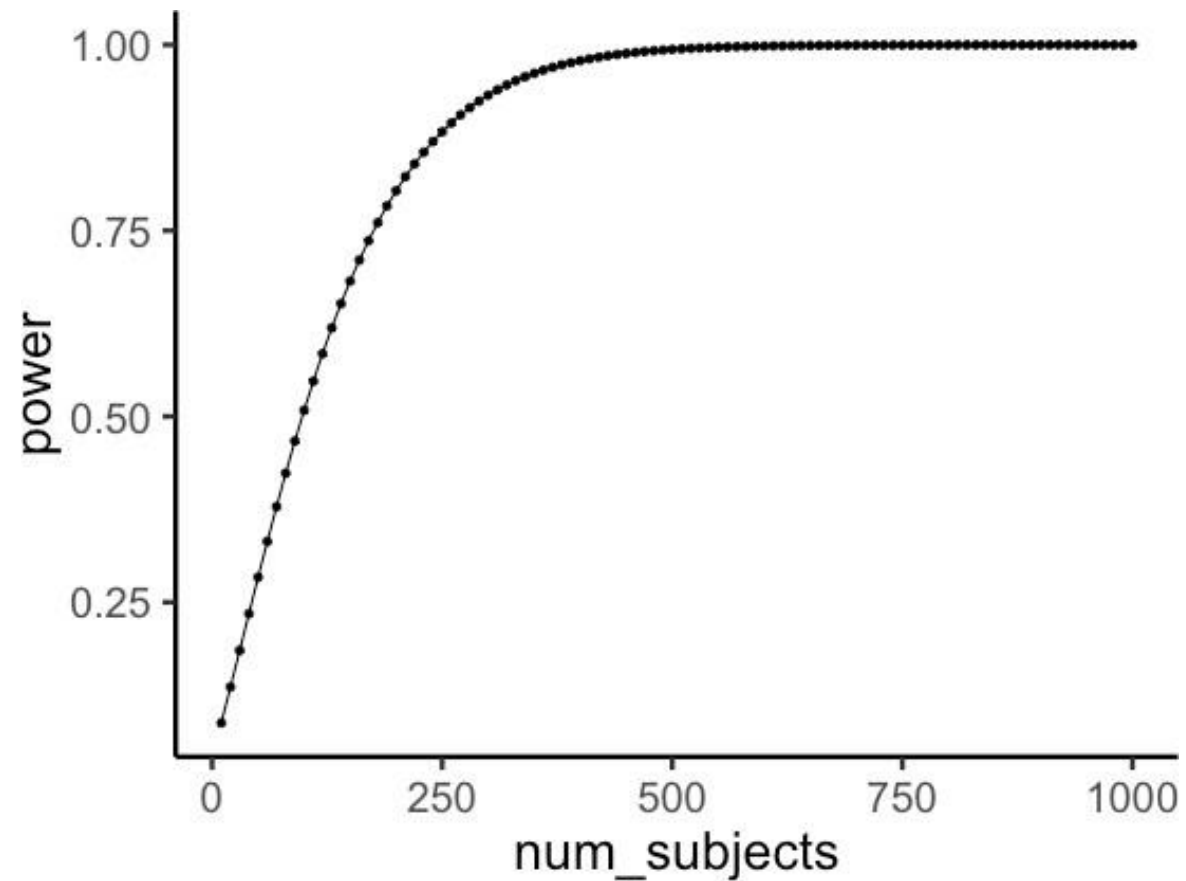


Figure 11.7: The power of our test, plotted as a function of the sample size  $N$ . In this case, the true value of  $\theta$  is 0.7, but the null hypothesis is that  $\theta = 0.5$ . Overall, larger  $N$  means greater power. (The small zig-zags in this function occur because of some odd interactions between  $\theta$ ,  $\alpha$  and the fact that the binomial distribution is discrete; it doesn't matter for any serious purpose)

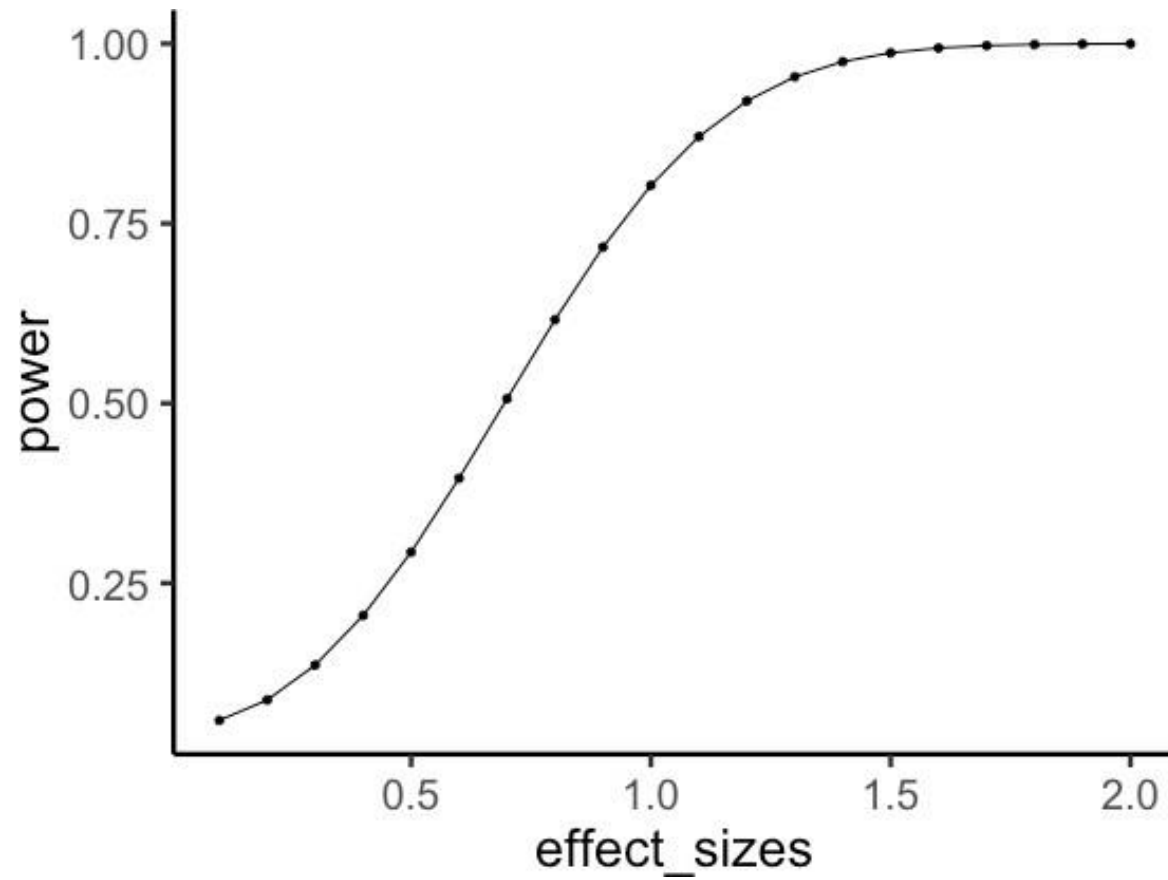
# Power as a function of n



# Power curve

- Any specific design has different levels of power to detect effects of different sizes. This can be shown on a power curve.

# Power curve, t.test, n=10



# Sample-size planning

**How many subjects do you need for your experiment?**

1. Establish a minimum effect-size of interest
2. Conduct a power-analysis, to show how power changes as a function of sample-size to detect the minimum effect size of interest.
3. Pick based on desired power

# Don't run an experiment that is designed to fail

- How?
  - Power analysis
  - Effect-size, and
  - Sample-size planning
  - (Power analysis)

# Other reasons to care about power

- Efficient use of resources
  - Power analyses tell us if our planned sample size ( $n$ ) is:
    - Large enough to be able to find what we're looking for (e.g. power  $\geq 0.8$ )
    - Not too large that we're collecting more data than necessary
- Reduces p-hacking
  - By reducing flexibility in data collection
    - E.g. stopping experiments once p-value  $< 0.05$
    - Encourages pre-registration
  - “underpowered” studies are more prone to p-hacking

# Other reasons to care about power

- More robust Null results
  - Absence of evidence  $\neq$  evidence for absence
    - but high powered studies can make us more sure (e.g. coin with  $\theta = 0.5$  vs coin with  $\theta = 0.55$ )
- Granting agencies want them now
  - Don't want to fund a study with low probability of showing anything

Effect Size

# Effect size

*“The amount of anything that’s of research interest”  
(Cumming & Calin-Jageman, 2017, p.111)*

- **Effect**

- When we run an experiment, we are interested in whether the **manipulation** caused a difference in our **measurement**
- If, our **manipulation** causes a difference in our **measurement**, then there will be an **effect**.

- **P-value** tells us if there is a difference in our measurements (an effect)  
**Effect-size** refers to how big or small the effect is.

- They’re not the same

# Measures of effect size

- There are many different measures of effect size. Consider the simplest measure for two groups, A and B.
- **Mean difference**
  - The difference between the mean of A, and the mean of B, is a measure of the effect size.
  - Large mean difference is a large effect
  - Small mean difference is a small effect

# Relative to what?

- Mean differences can be interpreted if we know what the difference is relative to.
- mean A = 1000, mean B = 1050
  - difference=50
  - 5% increase, not so big
- mean A = 1, mean B = 2
  - difference=1
  - 100% increase, pretty big

# Cohen's d

- Cohen's D express a mean difference between two samples in terms of standard deviation units (like a z-score). This allows us to know something about the relative size.
- $D = .1$  (mean difference is shifted by .1 SD)
- $D = 1$  (mean difference is shifted by 1 SD)
- $D = 2$  (mean difference is shifted by 2 SD)

# Cohen's d

- The general idea is:

$$d = \frac{\text{MeanA} - \text{MeanB}}{SD}$$

i.e. standardized mean difference – number of SDs between two means.

Where SD refers to pooled standard deviation across groups:

- equal size:  $(\sqrt{[(SD_1^2 + SD_2^2)/2]})$
- Unequal size  $(\sqrt{[(n_1-1) \times SD_1^2 + (n_2-1) \times SD_2^2]/(n_1 + n_2 - 2)])$

# Cohen's d

- The general idea is:

$$d = \frac{\text{MeanA} - \text{MeanB}}{SD}$$

i.e. standardized mean difference – number of SDs between two means.

Where SD refers to **pooled standard deviation** across groups:

- equal size:  $(\sqrt{[(SD_1^2 + SD_2^2)/2]})$
- Unequal size  $(\sqrt{[(n_1-1) \times SD_1^2 + (n_2-1) \times SD_2^2]/(n_1 + n_2 - 2)])$

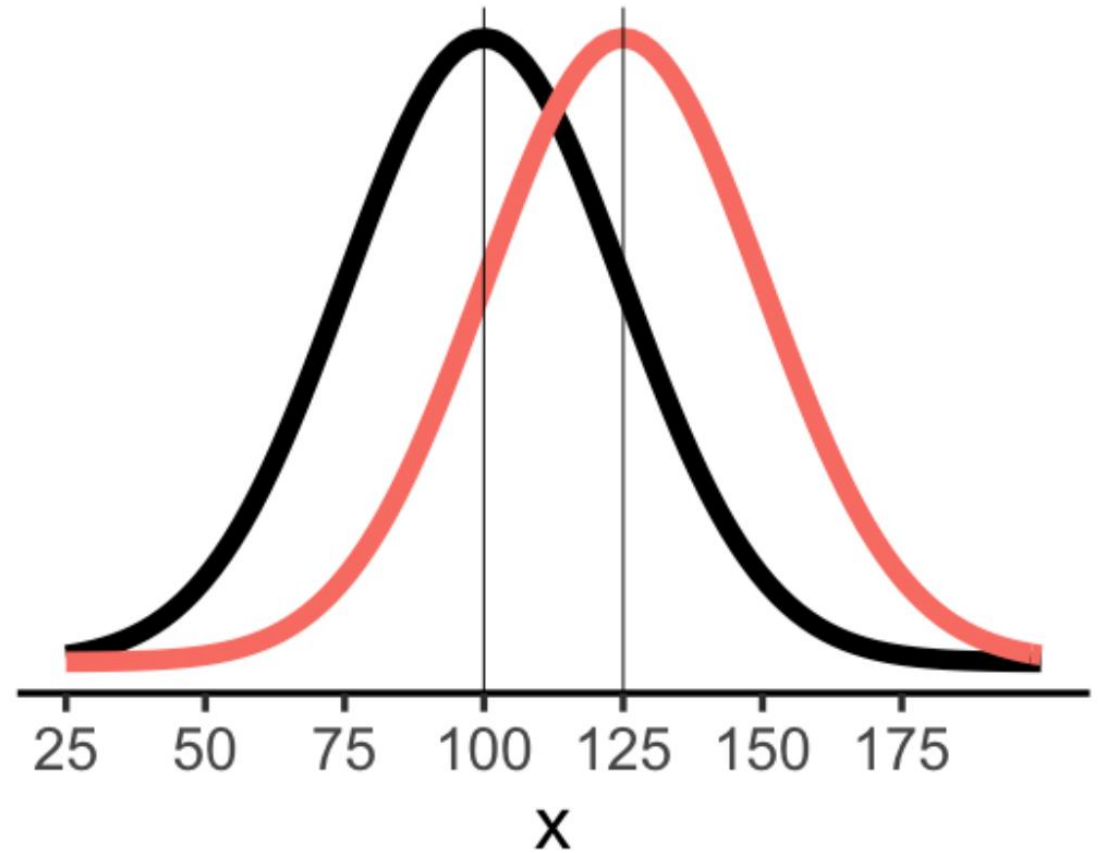
$M=0, SD=1, D=1$

A = Black, B = Red, Cohen's  $D = 1$



$M=100$ ,  $SD=25$ ,  $D=1$

A = Black, B = Red, Cohen's  $D = 1$

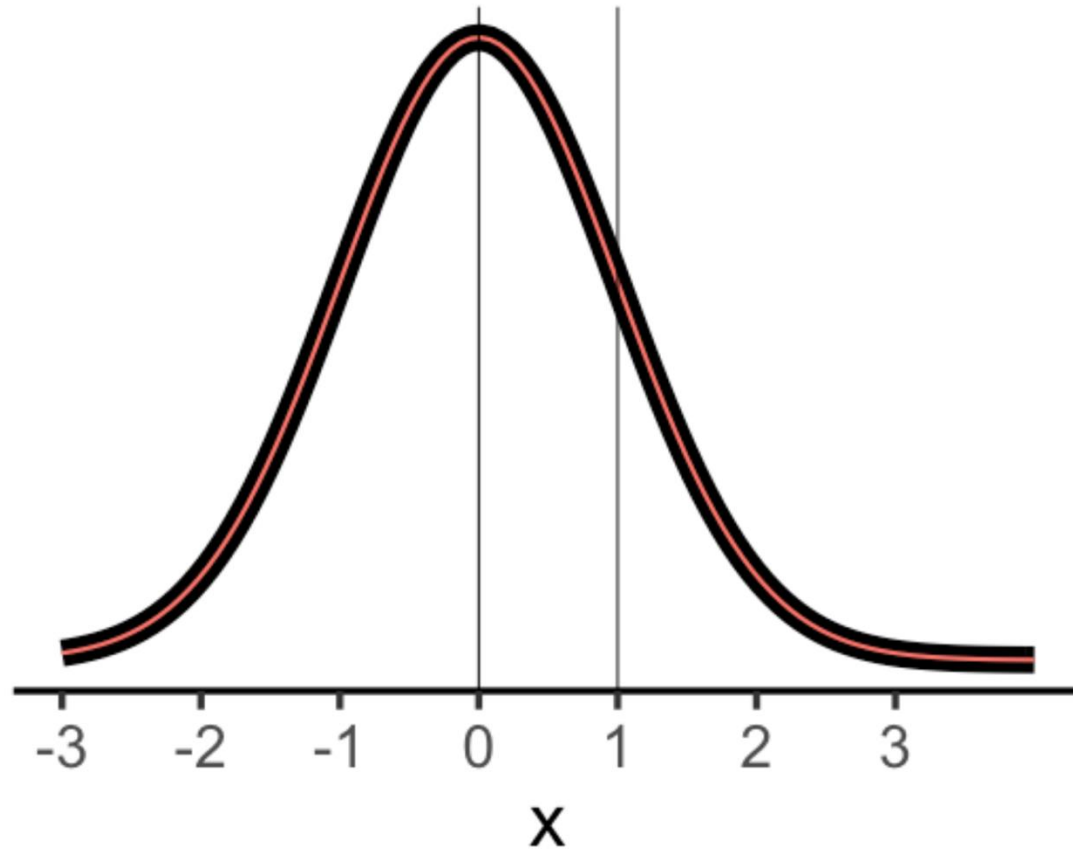


# No-difference

- If there is no difference, how big is Cohen's D?
- What do the distribution for A and B look like?

# No-difference

A and B come from the same distribution, no difference



# Interpreting Cohen's d

- Cohen gives these recommendations:
  - **Small:**  $d = .2$
  - **Medium:**  $d = .5$
  - **Large:**  $d \geq .8$
- Note d's larger than 1 are really big, they shift the whole distribution by a whole standard deviation, that's a lot!

# D's in Psychology

- Many effects in Psychology are **small**, with **d around .2**.
- One reason is that we measure people, and people are highly variable.

# Discuss

- Assume that we are dealing with coin tosses.
- We are hypothesizing about the fairness of a coin.
  - $H_0$  : the coin is fair ( $\theta = 0.5$ )
- ~~• For what type of coin is our hypothesis test the most powerful?~~
- How many coin tosses are required to confidently reject the null model?

# Power and Effect Size - Experiment Planning

# Power analysis

- Usually done a-priori, but can also be done afterward (controversial!)

# What is needed for a power analysis

- sample size (N)
- effect size
  - E.g. Cohen's d, r
- $\alpha$ , the significance level
- power

3 out of these 4

# Use cases of power analysis

1) Before collecting data, determine *necessary **sample size** ( $n$ )* to achieve sufficient power:

- Specify:
  - $\alpha$ , the significance level
  - power ( $1 - \beta$ )
  - Effect size ---- *best guess*
- Find N

## (In previous lab)

- You tried to determine a good sample size for
  - Detecting 2% (meaningful) difference
    - Most of the time (power is close enough to 1)
- Power curve => you calculate X % of detection given true effect
  - Plot it for varying N

# Use cases of power analysis

2) Before collecting data, determine *minimally detectable **effect size (MDES)*** for a range of sample sizes:

- Specify:
  - $\alpha$ , the significance level
  - power ( $1 - \beta$ )
  - sample size (N)
- Find MDES

# Use cases of power analysis

3) After collecting data, determine whether the design/ analysis and sample size had *sufficient **power*** to detect relationships if they were to exist:

- Specify:
  - $\alpha$ , the significance level
  - sample size (N)
  - Effect size
- Find Power

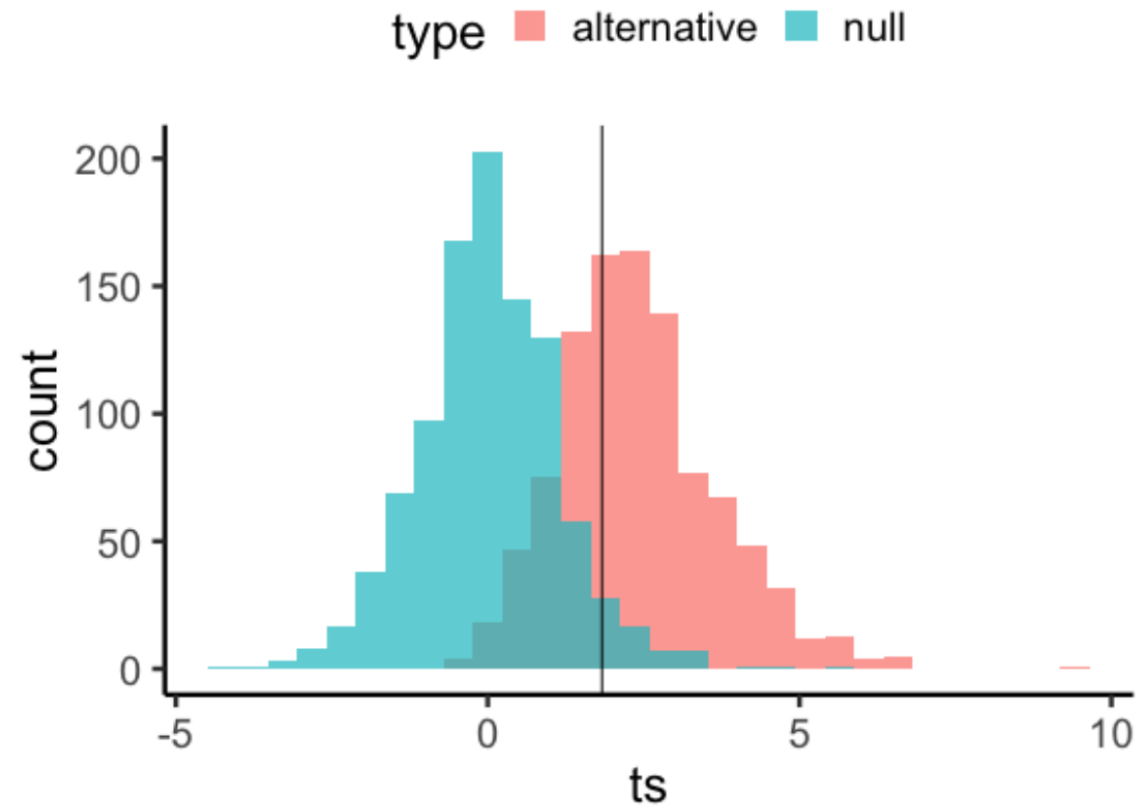
# How to carry out power analyses?

- Remember that power is a probability value.
- If we have the appropriate sampling distributions of the Null Hypothesis and Alternative Hypothesis and the significance level, it can be manually computed (using simulations).



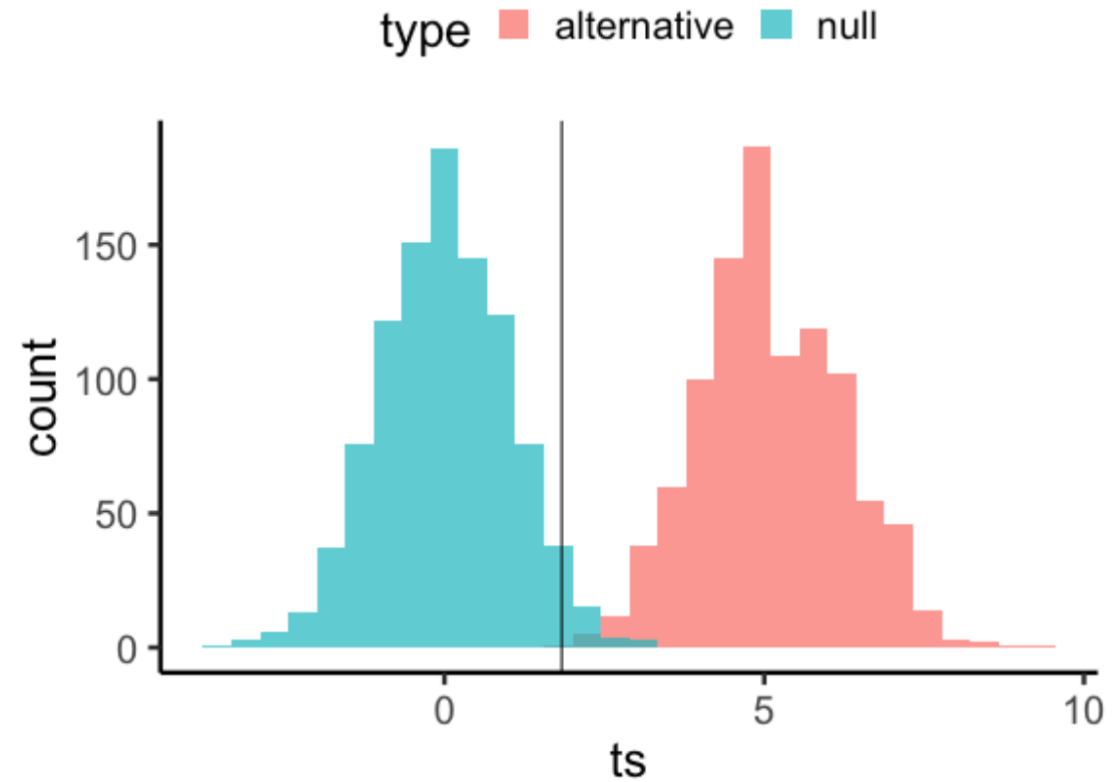
# Look at the distributions

- power = 0.665 to detect  $d=1$



# When N=50

- power = 0.999 to detect  $d=1$



# How to carry out power analysis?

- Might need guesses of effect sizes, etc
- Many tools
- R-packages
  - Pwr
    - T.tests
    - One-way anova, etc.
  - Bmem
    - Uses monte carlo simulations (more complex tests, etc.)

# How to carry out power analysis?

- Might need guesses of effect sizes, etc

- Many tools

- R-packages

- Pwr

- T.tests
    - One-way anova, etc.

- Bmem

- Uses monte carlo simulations (more complex tests, etc.)

```
install.packages("pwr")  
library(pwr)  
  
pwr.t.test(n = NULL, d = .5, sig.level = .05, power = .8, alternative = "two.sided", type = "two.sample")
```

# Two approaches to power analyses

- **Formula-based (analytic)**
  - Math formulas
  - Fast and Exact answers
  - Works for standard tests
  - pwr package

# Two approaches to power analyses

- **Formula-based (analytic)**

- Math formulas
- Fast and Exact answers
- Works for standard tests
- pwr package

- **Simulation-based (Monte Carlo)**

- Generate fake data
- Count successes
- Approximate answers
- Works for ANY design
  - Complex designs, mixed models, custom-tests,...
- Your assignment method

# Note: post-hoc power analysis is controversial

- You use your results to calculate power for detecting your results.

# post-hoc power analysis is controversial

- **Example:**

- You find  $d = 0.3$ ,  $p = 0.08$  (not significant).
- Post-hoc power analysis: "We only had 30% power."

# post-hoc power analysis is controversial

- **Example:**

- You find  $d = 0.3$ ,  $p = 0.08$  (not significant).
- Post-hoc power analysis: "We only had 30% power."

- **But so what?**

- You already know you didn't find it.
- The p-value told you that.
- Power adds nothing new.

# post-hoc power analysis is controversial

- **Issue is the circularity:**
  - Found significant result → observed effect is big → post-hoc power is high
  - Didn't find significance → observed effect is small → post-hoc power is low
  - You're just restating your p-value in different words.

# post-hoc power analysis is controversial

- **Common misuse:**

- "We didn't find an effect, but we had low power, so maybe it's really there."

# post-hoc power analysis is controversial

- **Common misuse:**

- "We didn't find an effect, but we had low power, so maybe it's really there."
- No.
- Low power means your observed effect was small. Which you already knew.

# post-hoc power analysis is controversial

- **Better approach:**

- Pre-specify power based on meaningful effect size.
- Not on what you happened to observe.

- **When post-hoc helps:**

- Evaluating if your design *could have worked*.
- Not using observed effect.
- Using theoretical or prior effect sizes.
- "We wanted to detect  $d = 0.5$ . Did we have enough  $n$  for that?"
- That's legitimate.

# Overpowered studies & practical significance

- **The overpowered study problem:**
  - $n = 1,000,000$
  - Find: Coffee improves test scores by 0.01%
  - $p < 0.001$  (highly significant!)
  - $d = 0.001$  (trivial)

# Overpowered studies & practical significance

- **The overpowered study problem:**

- $n = 1,000,000$
- Find: Coffee improves test scores by 0.01%
- $p < 0.001$  (highly significant!)
- $d = 0.001$  (trivial)

- **What happened?**

- Detected real effect.
- But who cares?
- 0.01% means nothing.

# Overpowered studies & practical significance

- **The controversy:**
  - **Statistical significance  $\neq$  practical significance**
  - Huge n detects everything.
  - Even noise becomes "significant."
  - Misleading.

# Overpowered studies & practical significance

- **The controversy:**

- **Statistical significance  $\neq$  practical significance**
- Huge  $n$  detects everything.
- Even noise becomes "significant."
- Misleading.

- **Problems:**

- **Waste of resources**
- **Publication distortion**
  - "Significant" sounds important, but  $d = 0.001$  isn't
- **Policy misuse** (see  $p < 0.05$ , ignore effect size, make bad decisions)

# Overpowered studies & practical significance

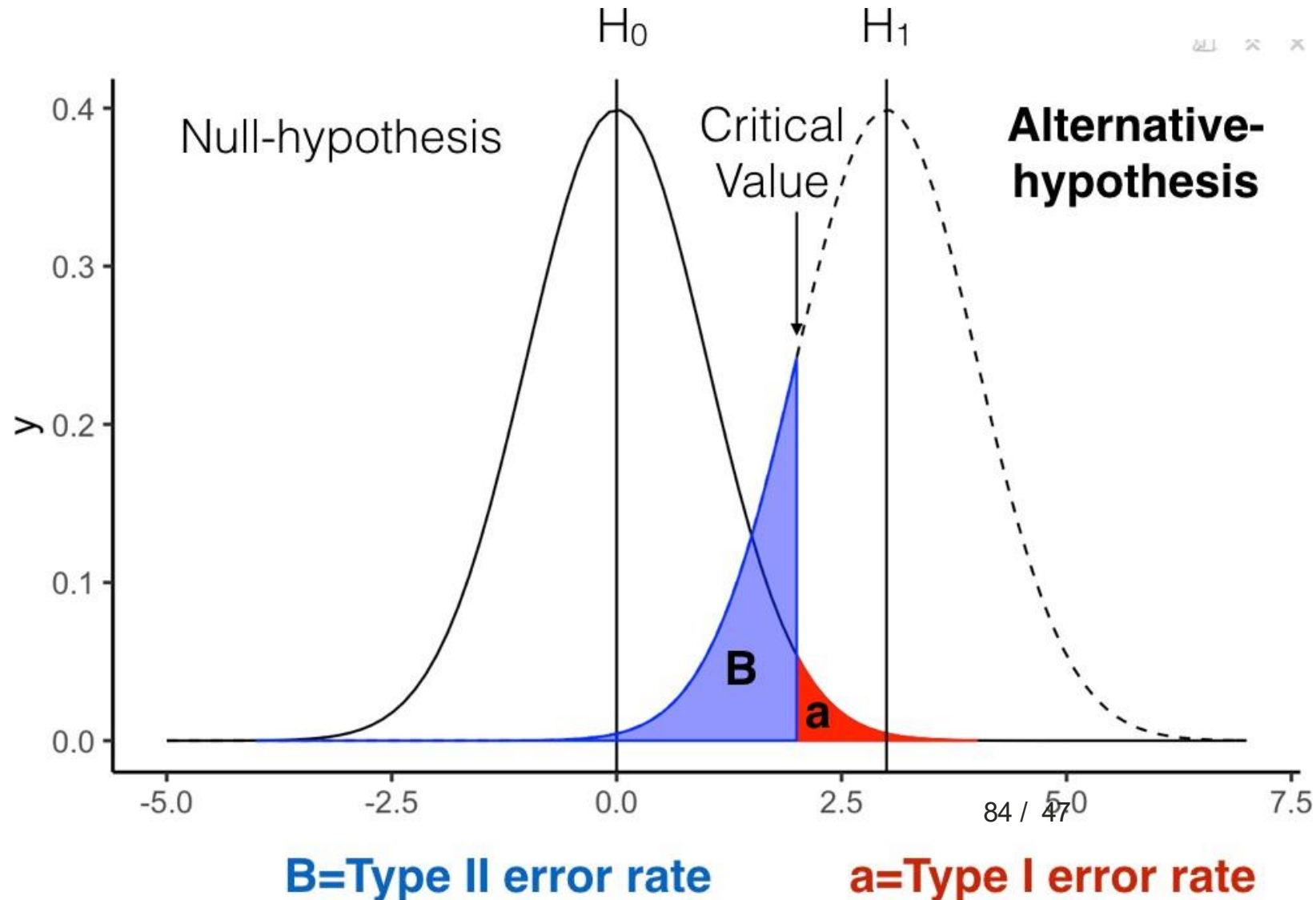
- **The balance:**

- Too little power → miss real effects
- Too much power → find meaningless effects

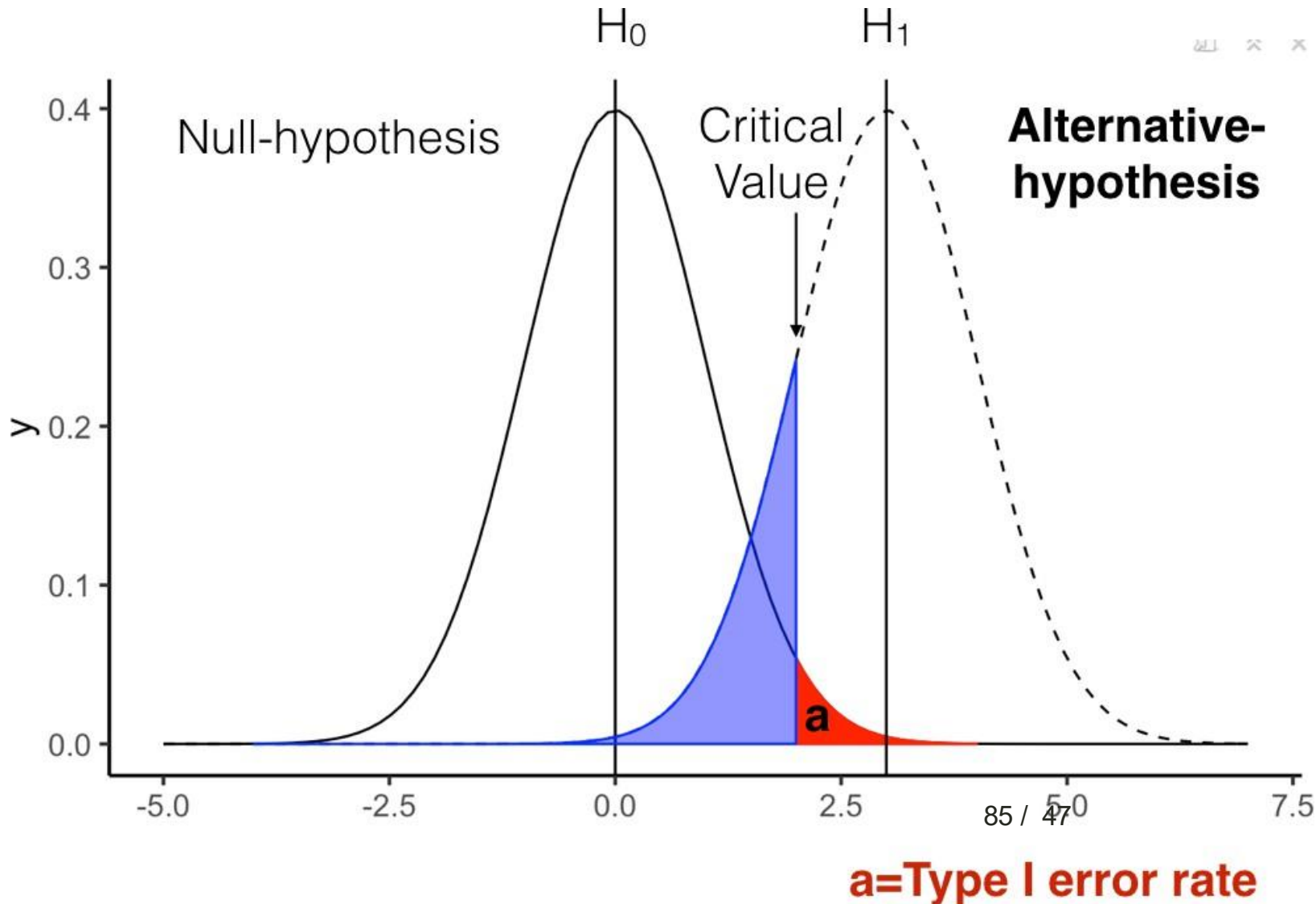
- **Good practices**

- Always report effect size.
- Never report p-value alone.
- Ask: Is this difference meaningful?
- Not just: Is it detectable?

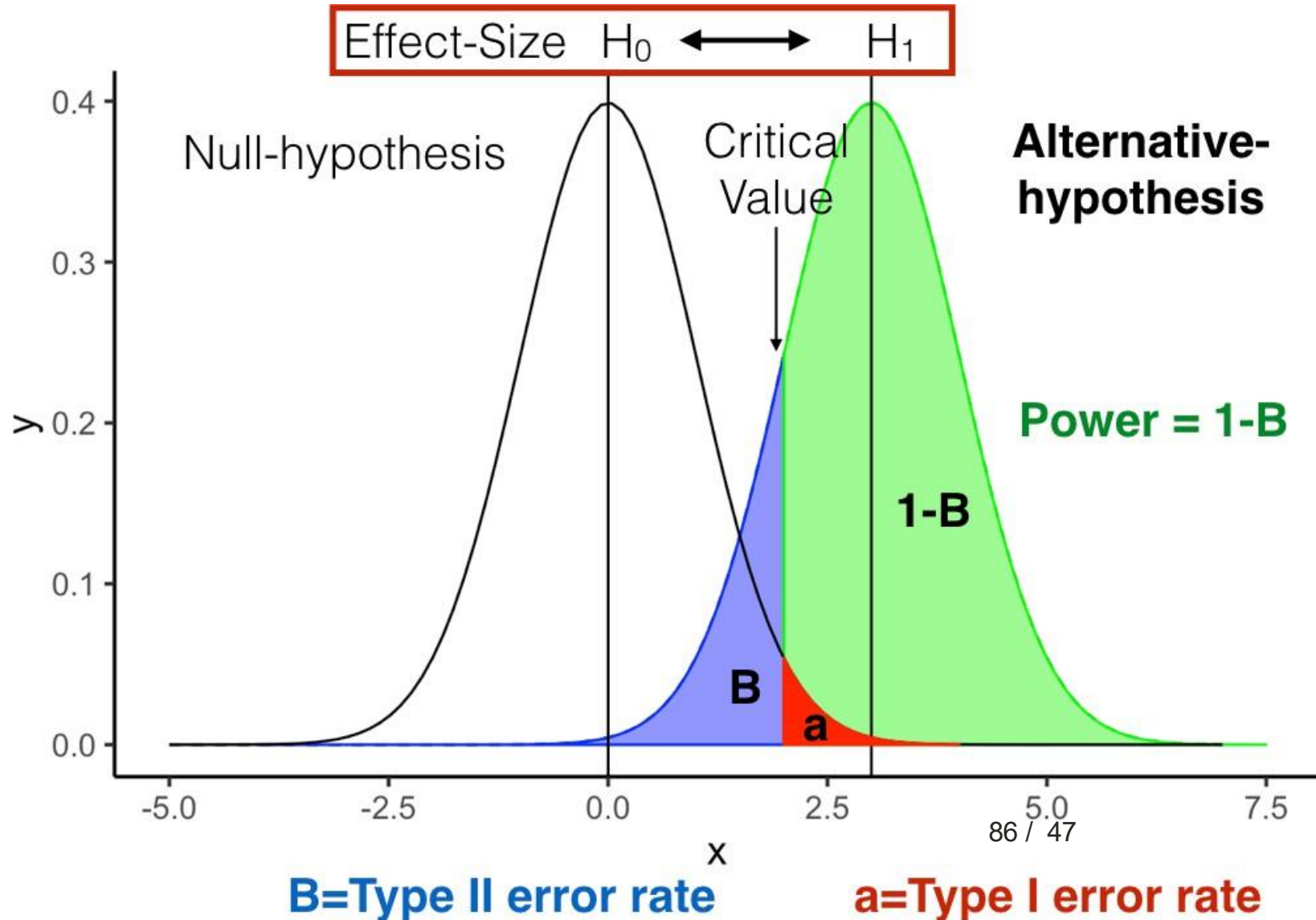
# B = type II error



# Alternative Hypothesis ( $d > 0$ )



# Power and effect-size





# Assignment (due Nov 1)

- Think about what project you'd like to work on
- Send a plan on Canvas
- If you are considering multiple options, outline them in this draft / contact me before that to discuss and narrow it down

# James Uanhoro – Assistant Professor, University of North Texas

---

Oct 28, 2025, 11:00 am – 12:00 pm 

A03 Princeton Neuroscience Institute 

Zoom 

## Speaker

James Uanhoro 

University of North Texas 

## Details

### Want Clear Effect Sizes? Use Ordinal Regression (Even for Continuous Data)

Psychology researchers rely on effect sizes to interpret the results of quantitative data analysis and communicate them to others. However, the most frequently reported effect sizes -- correlations, variance explained measures, or standardised mean differences -- are abstract and difficult to interpret without translation into meaningful terms. Common-language effect sizes (CLES) are an intuitive alternative, but they are challenging to compute for designs beyond simple comparisons. Ordinal regression solves this problem by accurately estimating the probability of superiority, a CLES. It works across diverse study designs like other regression approaches, but unlike linear regression, it remains valid across an extensive range of data types. Hence, researchers should adopt ordinal regression as a default regression approach. Doing so yields effect sizes that are statistically robust and easy to explain to students, practitioners, and the public.

