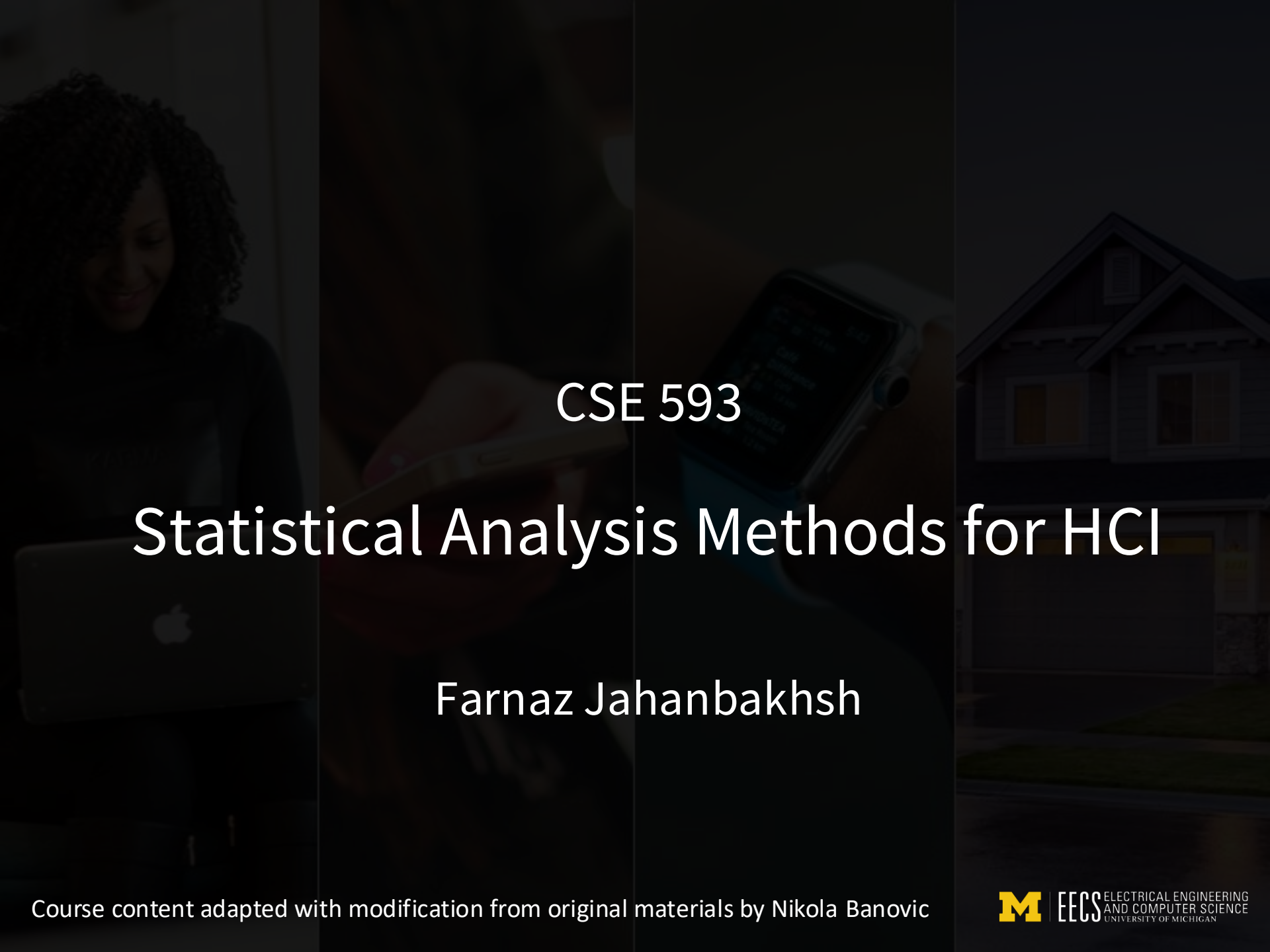


CSE 593

Statistical Analysis Methods for HCI

Farnaz Jahanbakhsh

Course content adapted with modification from original materials by Nikola Banovic

EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE UNIVERSITY OF MICHIGAN

Logistics

- Assignment 4 individual due today at 3pm
- Assignment 4 group due next week (Nov 13)
- Quiz 5 assigned today, due tomorrow at 5pm
- First special topics lecture on Tuesday!
  - Prof. Xu Wang
- Required reading next Tuesday due before lecture
- Required readings for Nov 19th and 21st posted

Goals

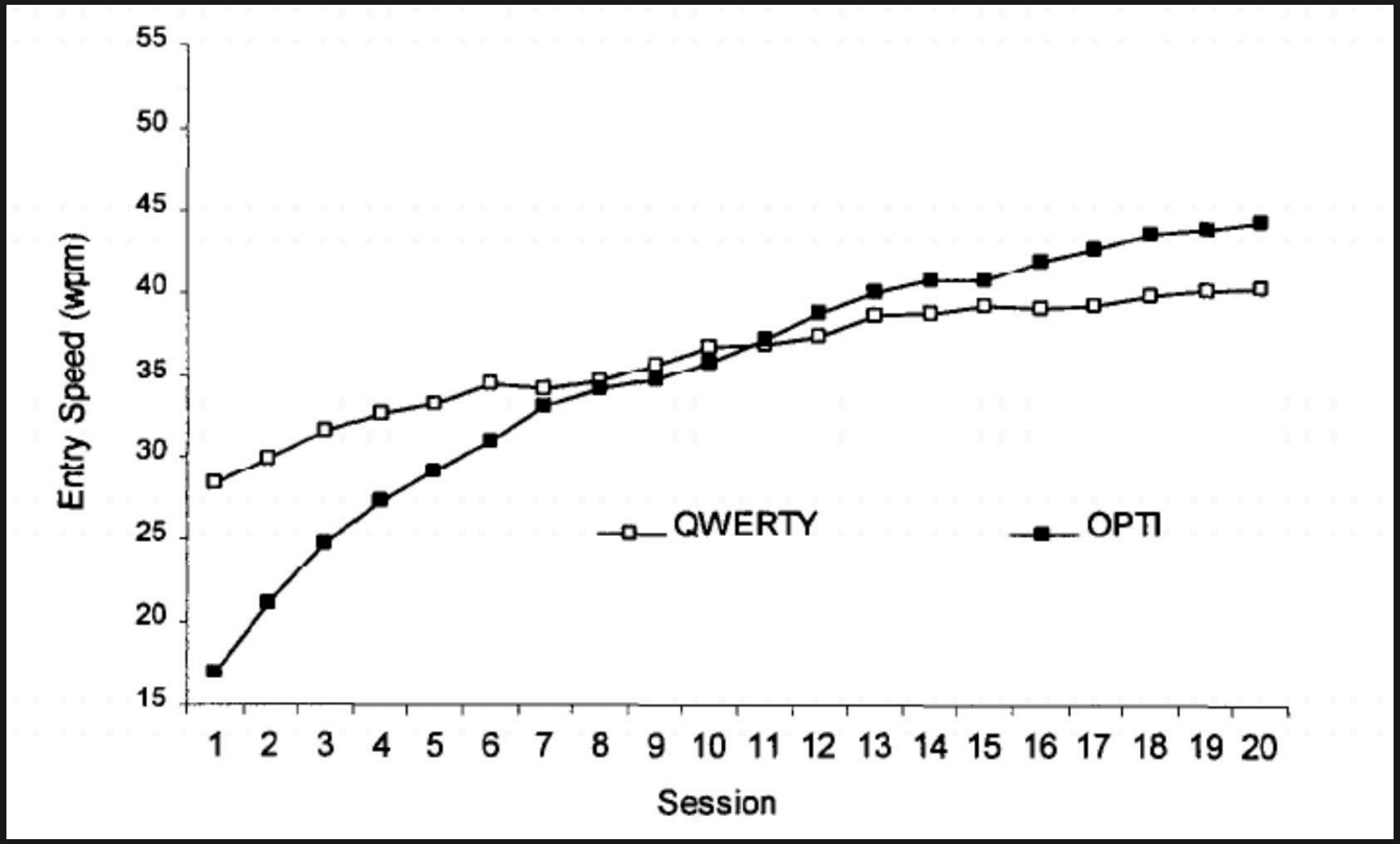
Learn about statistical analysis methods

Lean how to apply Null Hypothesis Statistical Testing (NHST)

Example: Empirical Keyboard Evaluation

MacKenzie & Zhang. 1999. The Design and Evaluation of a High-performance Soft Keyboard. In Proc. CHI '99

Reporting results



MacKenzie & Zhang. 1999. The Design and Evaluation of a High-performance Soft Keyboard. In Proc. CHI '99

Frequentist statistical analysis

Statistical Analysis in HCI

- Used to show the effects of independent variable(s) on a dependent variable (e.g., correlation, causation)

Statistical Analysis in HCI

- Used to show the effects of independent variable(s) on a dependent variable (e.g., correlation, causation)

- Requires a hypothesis (e.g., user performance with one design is better than performance with another design)

Null Hypothesis Testing

Used to establish causation “indirectly.”

Null Hypothesis Testing

Used to establish causation “indirectly.”

Assume there is no effect (null hypothesis).

- Null hypothesis: the keyboard design has not effect on user performance

Null Hypothesis Testing

Used to establish causation “indirectly.”

Assume there is no effect (null hypothesis).

Show that the null hypothesis is unlikely (*p*-value less than some significance level alpha).

- alpha = the probability of falsely rejecting the null hypothesis
- Traditionally alpha = 0.05, moving toward 0.01

p value & alpha

p value: probability of observing your data (or a more extreme difference) if the NH were true

If p < alpha, results are statistically significant

alpha: type I error (the probability of incorrectly rejecting the NH)

Null Hypothesis Testing

Used to establish causation “indirectly.”

Assume there is no effect (null hypothesis).

Show that the null hypothesis is unlikely (*p*-value less than some significance level alpha).

Therefore, conclude that there is an effect.

Example: Null Hypothesis Testing

Null hypothesis (H_0): “The users are equally fast on the two keyboards.”

Example: Null Hypothesis Testing

Null hypothesis (H_0): “The users are equally fast on the two keyboards.”

Alternative hypothesis (the one you want to show evidence for): “The users are faster on one of the two keyboards.”

Null Hypothesis Testing

Run appropriate statistical test.

If *p* less than some alpha:

- Conclude that users are unlikely to be equally fast on the two keyboards.
- Conclude that the keyboard with a faster mean time is the one that users are faster with.

Null Hypothesis Testing

Run appropriate statistical test.

If *p* greater than the set alpha:

- The test is inconclusive! Even if there is a difference in means
- You cannot accept the null hypothesis and conclude that there is no difference

Remember: You cannot prove the Null Hypothesis!

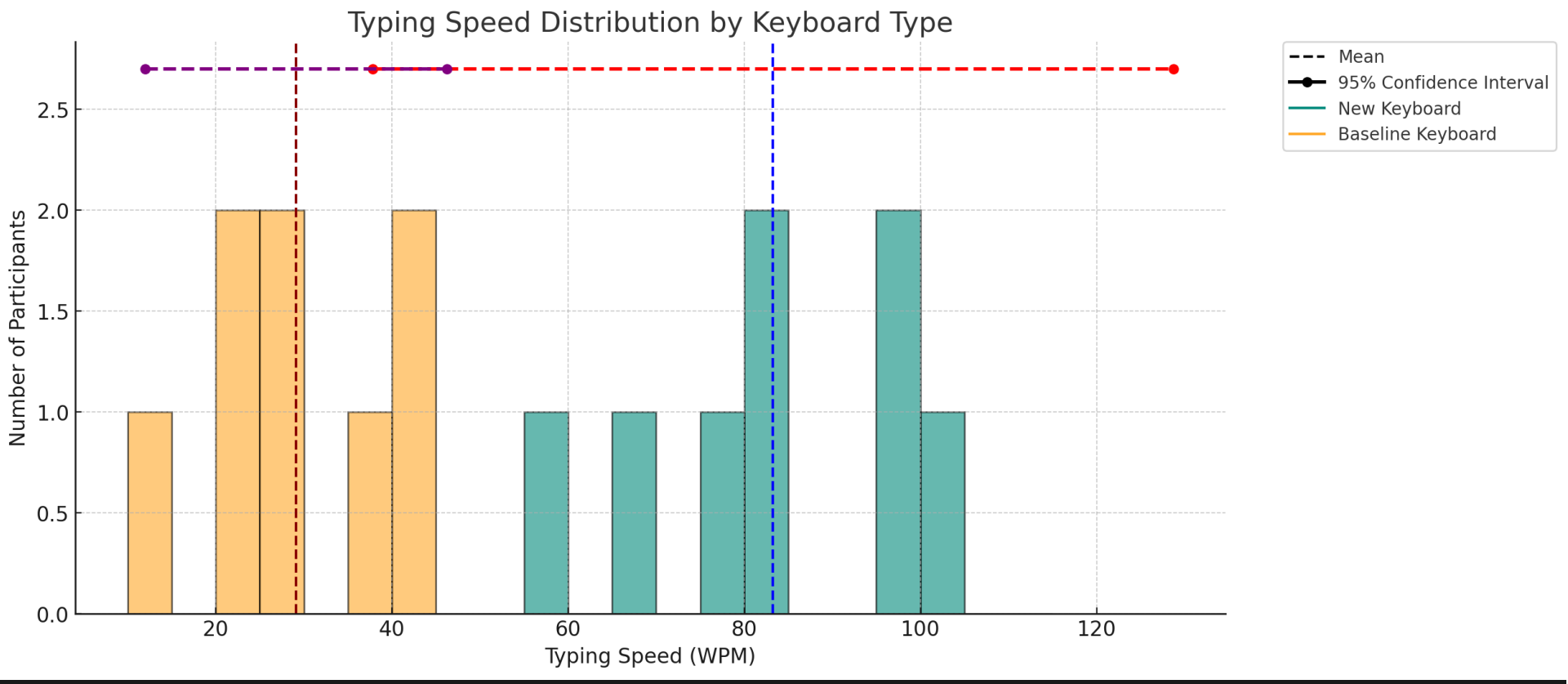
Example: Null Hypothesis Testing

Example: show that means of two measures are different

One independent variable: keyboard (baseline, intervention).

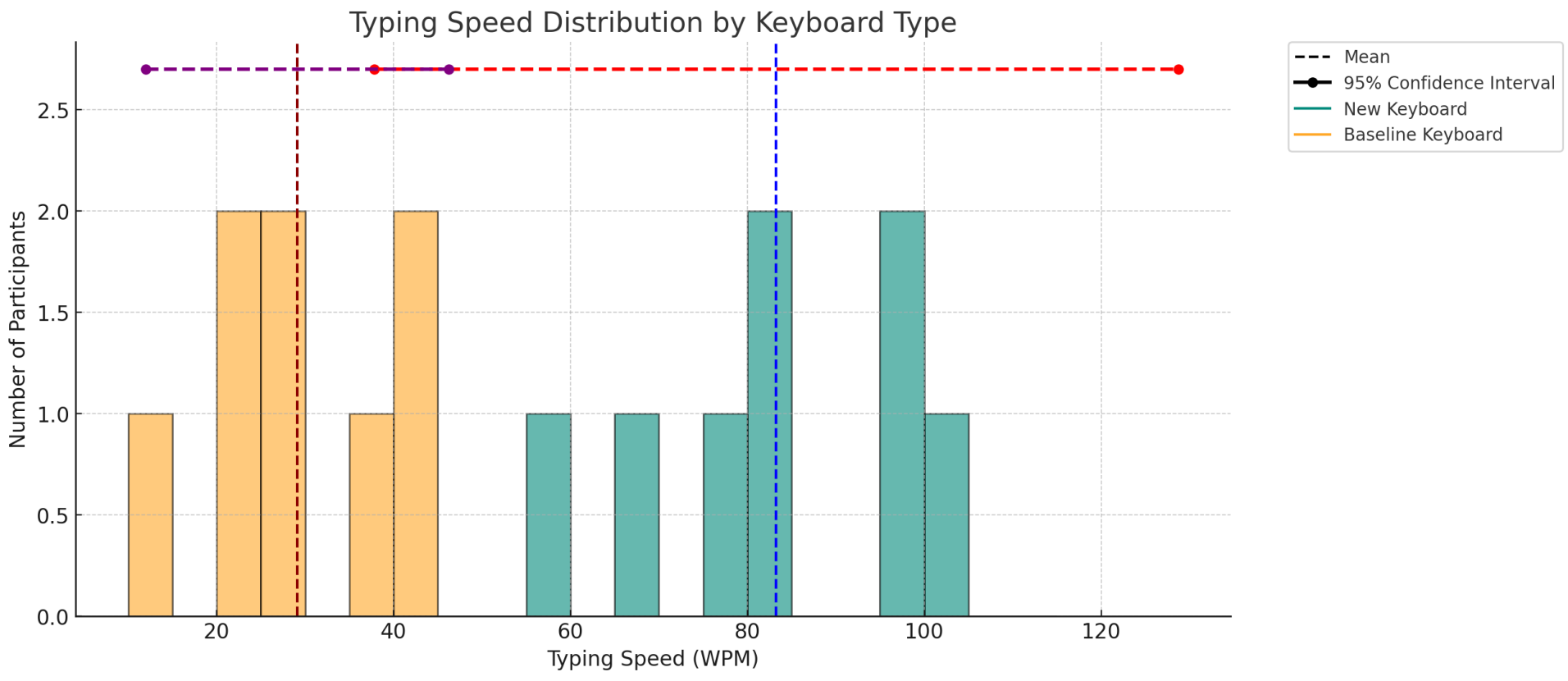
One dependent variable: typing speed.

Example: what the data looks like



Is there a difference on average?

Example: what the data looks like

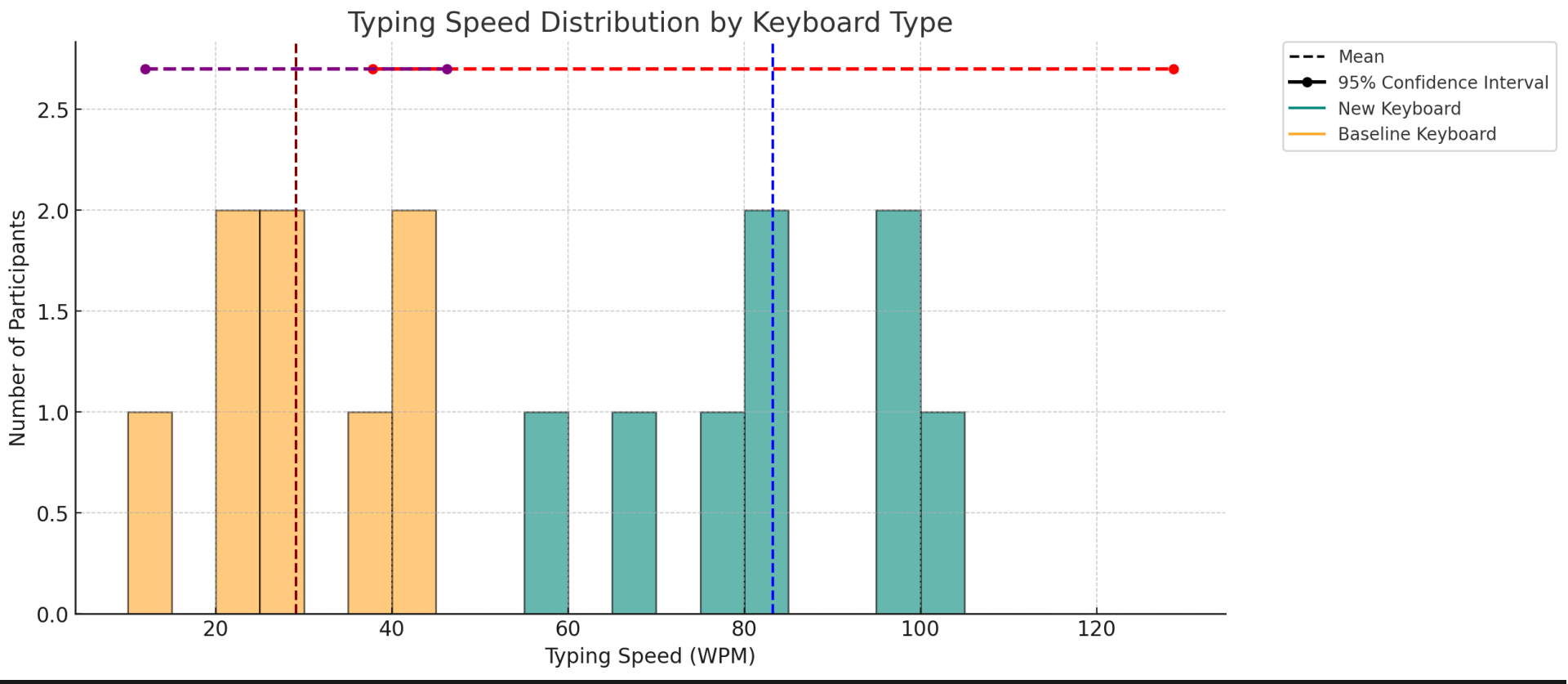


Is the mean speed of baseline < mean speed of treatment?
How confident are we that the *true* means are close to the sample means?

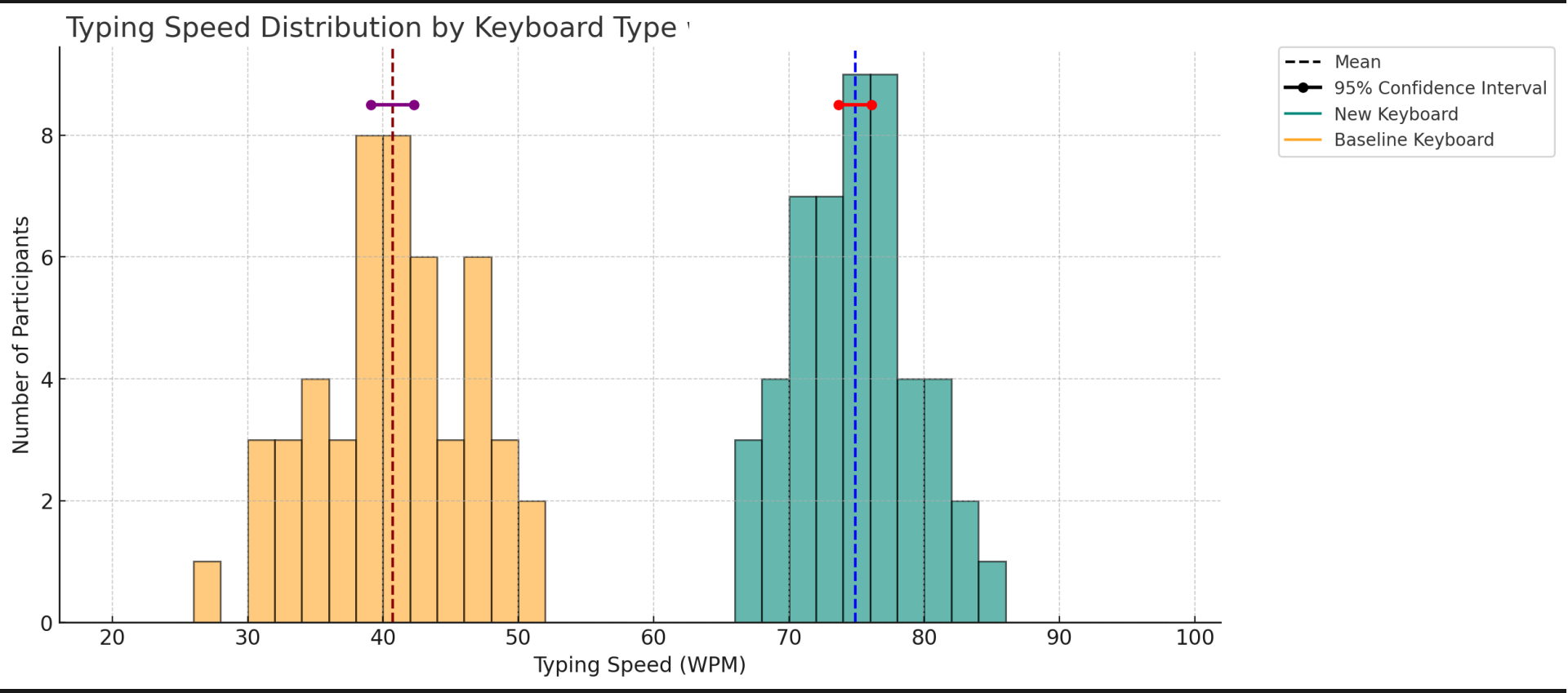
Population, sample, and CIs

- Population: all possible users who might use the keyboards
- Sample: the specific participants in the study
- We're using the sample mean to estimate the population mean
- The more users we have, the more confident we are
- Confidence intervals formalize this confidence
  - CI: the range of values likely to contain the population parameter

Example: what the data looks like

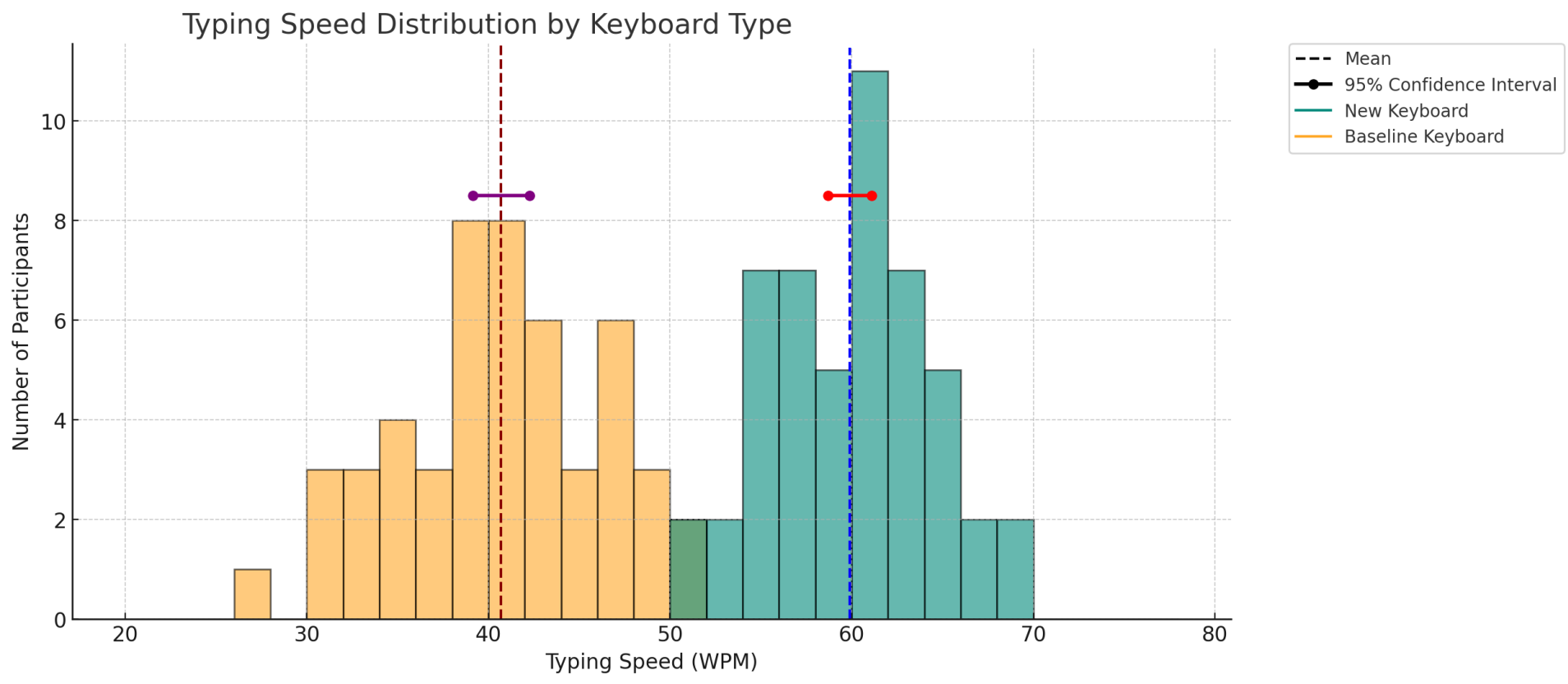


Example: what the data looks like



Larger sample size leads to narrower CIs

Example: what the data looks like



Effect size

- Intuitively captures the *magnitude* of the effect
- Depends on the statistical test
  - E.g., Cohen's d for t test: how many pooled standard deviations the means are from each other
- A large sample size increases the likelihood of detecting even very small effects as statistically significant
  - So effect size matters!

Study method design

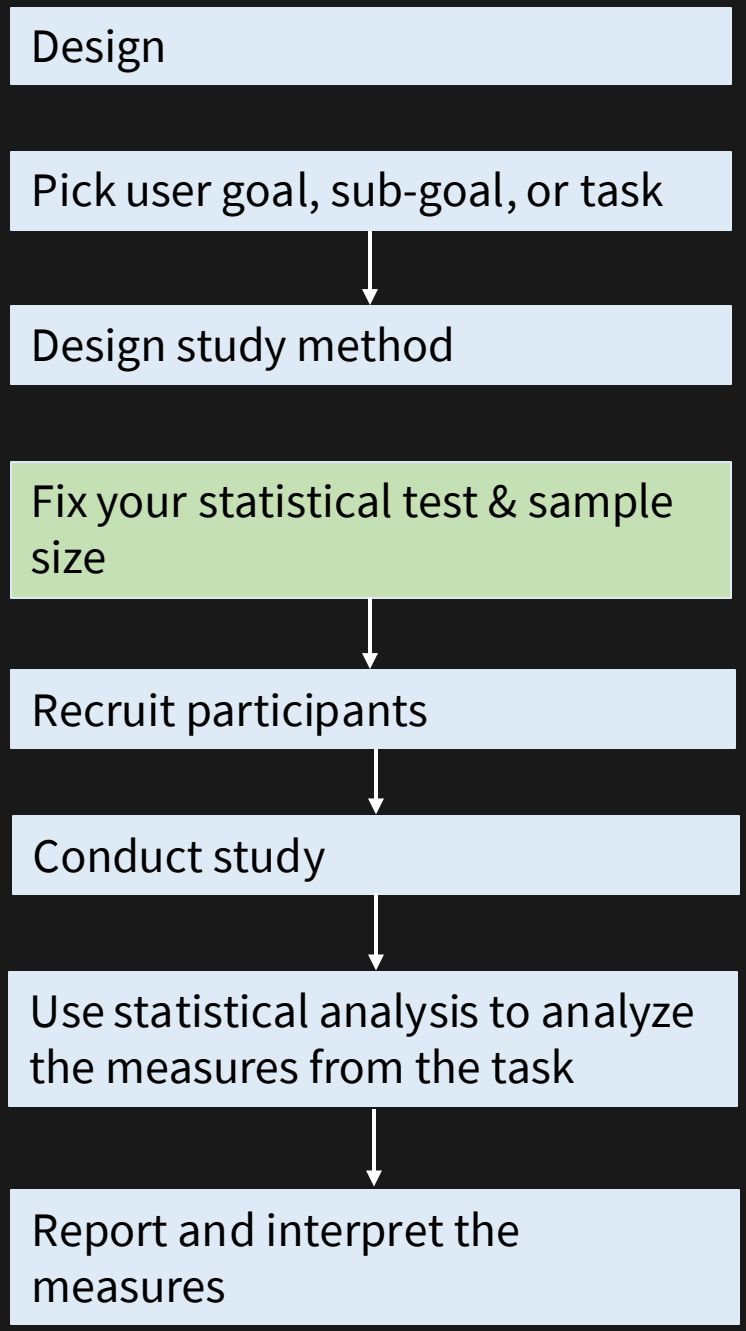
Pick independent variables – i.e., conditions (e.g., baseline, intervention)

Decide on type of study (e.g., between subjects, within subjects, mixed design, repeated measures)

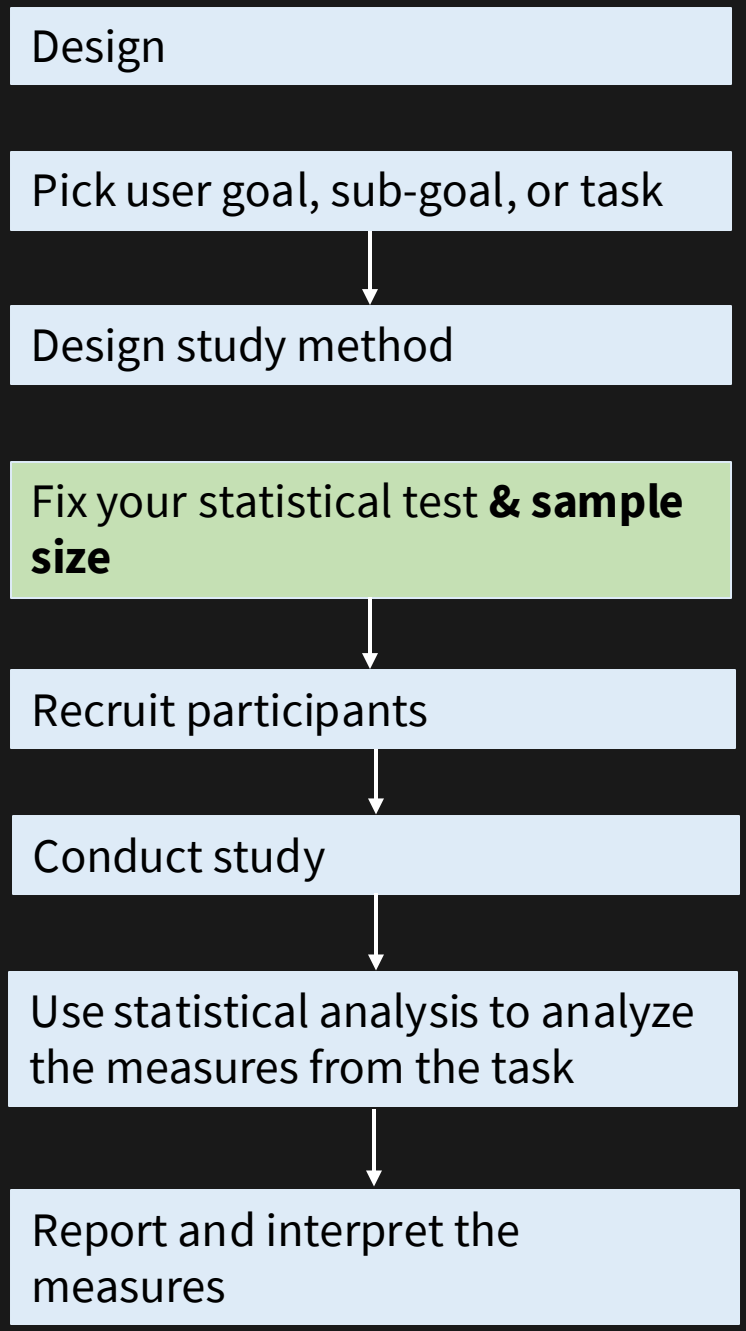
Pick dependent variables (e.g., performance)

Pick a statistical test

The (simplified) anatomy of an experiment



The (simplified) anatomy of an experiment



Power analysis

Shows the relationship between statistical power, sample size, effect size, & alpha

- Alpha: threshold for rejecting the null hypothesis
- Sample size: number of participants
- Effect size: magnitude of the effect
- Statistical power: the probability of correctly rejecting the null hypothesis when it is false
  - Power= 1 - β
  - β = type II error (false negative)

Errors

Question from the past: why not reduce alpha as much as possible?

Alpha: Type I error (false positive)

Beta: Type II error (false negative)

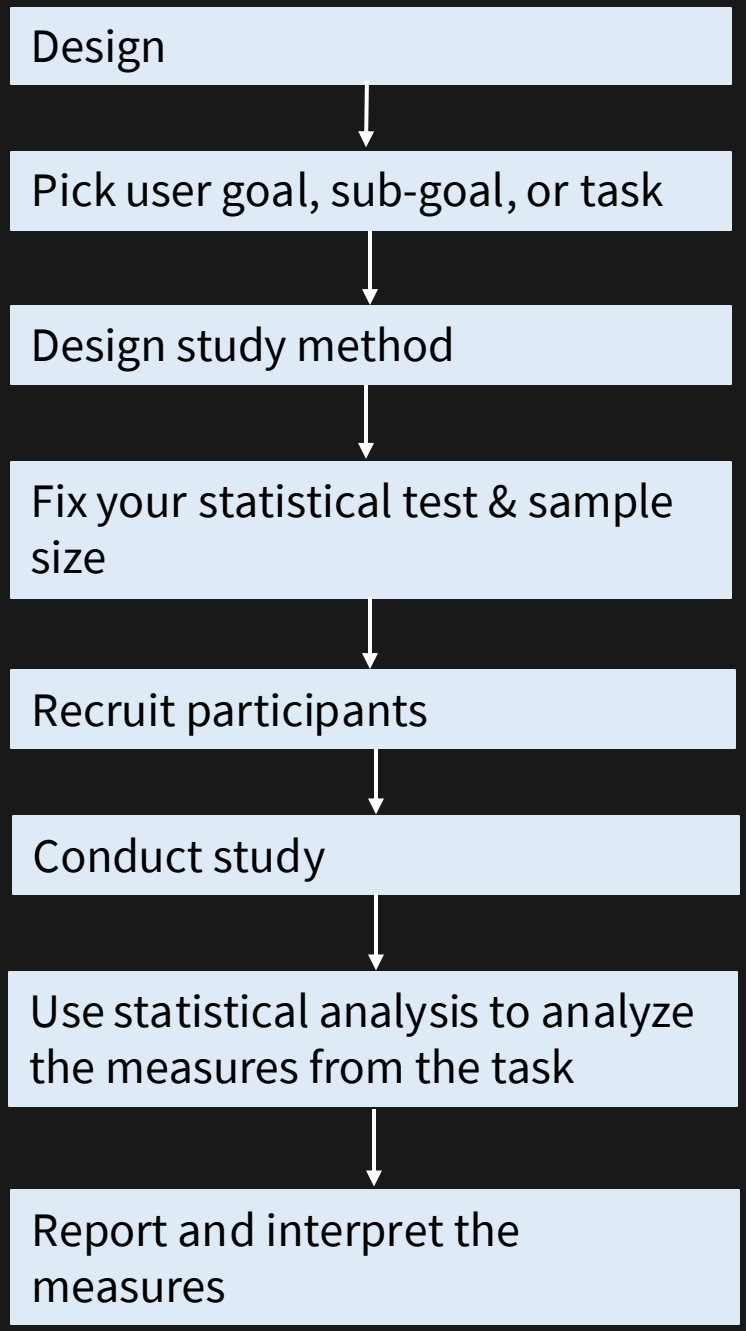
- Lowering α increases β
- For lowering β (increasing power), we need:
  - A larger sample size, or
  - Higher α

Power analysis

Shows the relationship between statistical power, sample size, effect size, & alpha

Plug in 3, get the last one

The (simplified) anatomy of an experiment

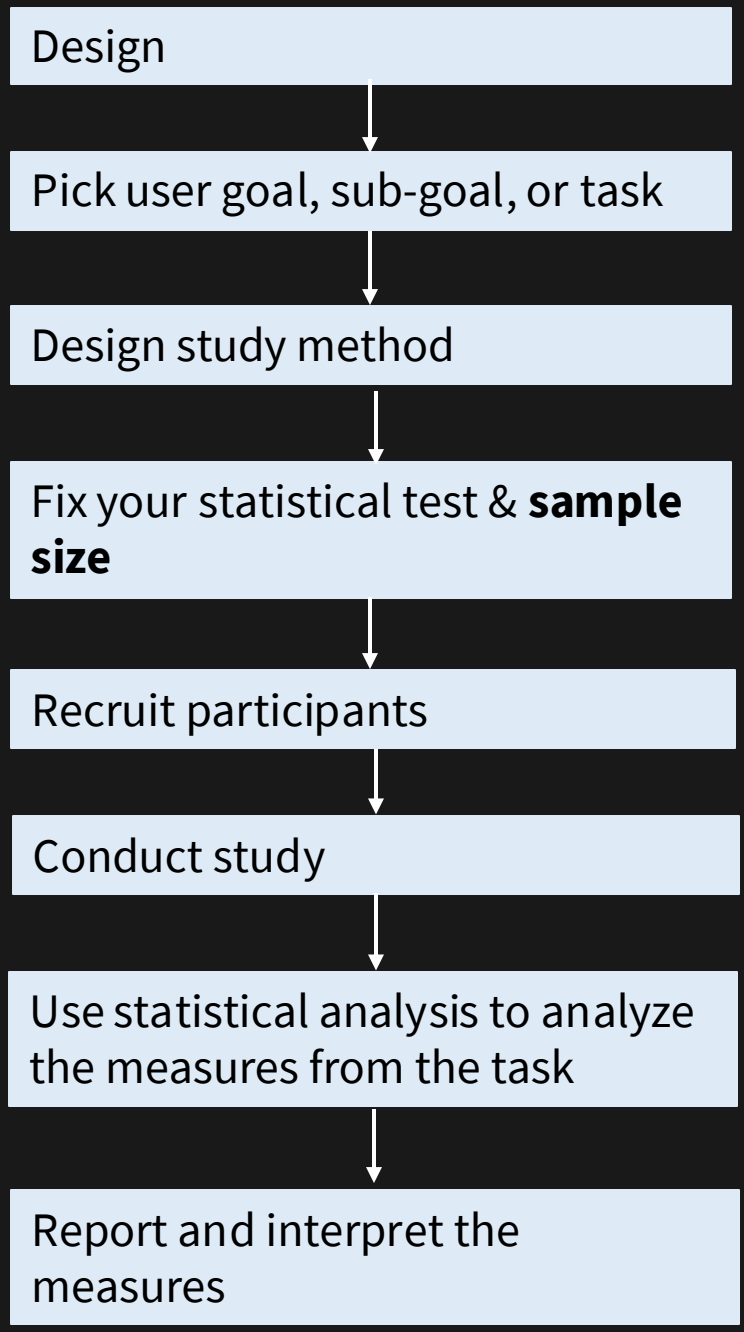


The (simplified) anatomy of an experiment

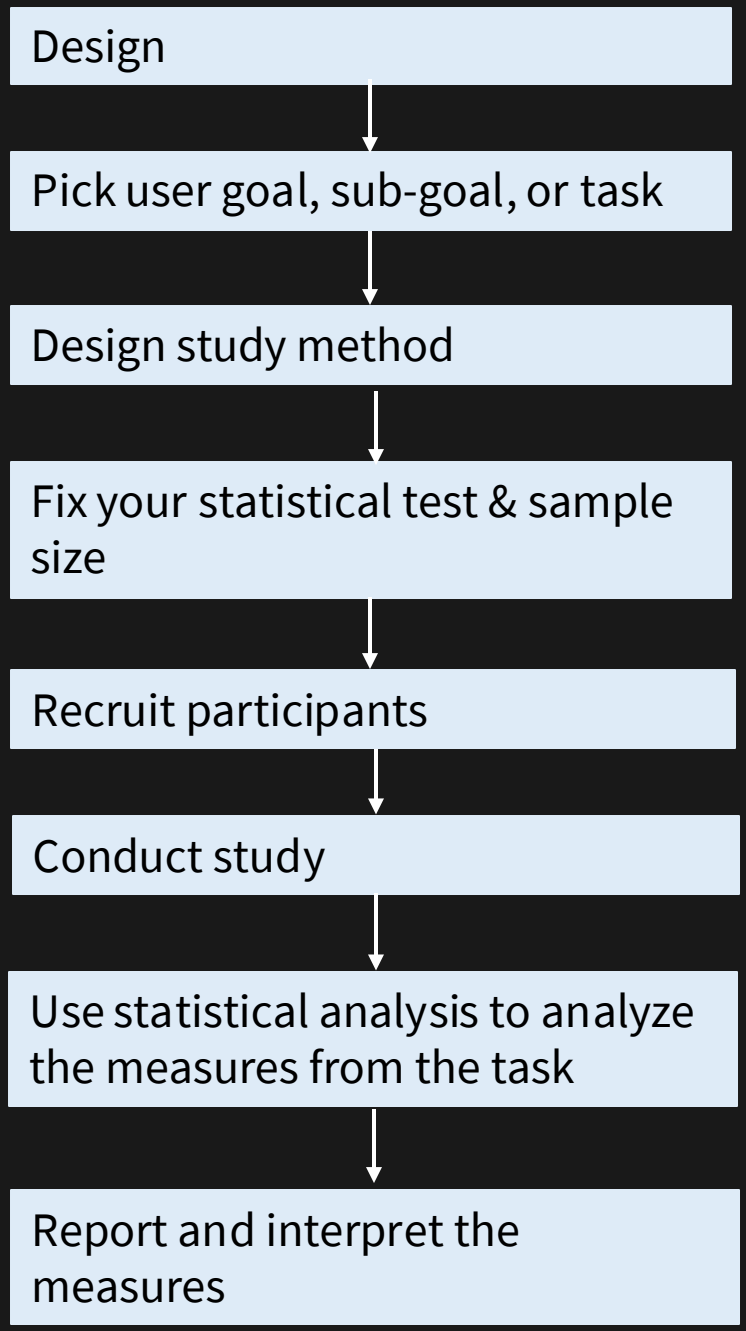
First problem:

We need the effect size for sample size calculation

How do we know that a priori?



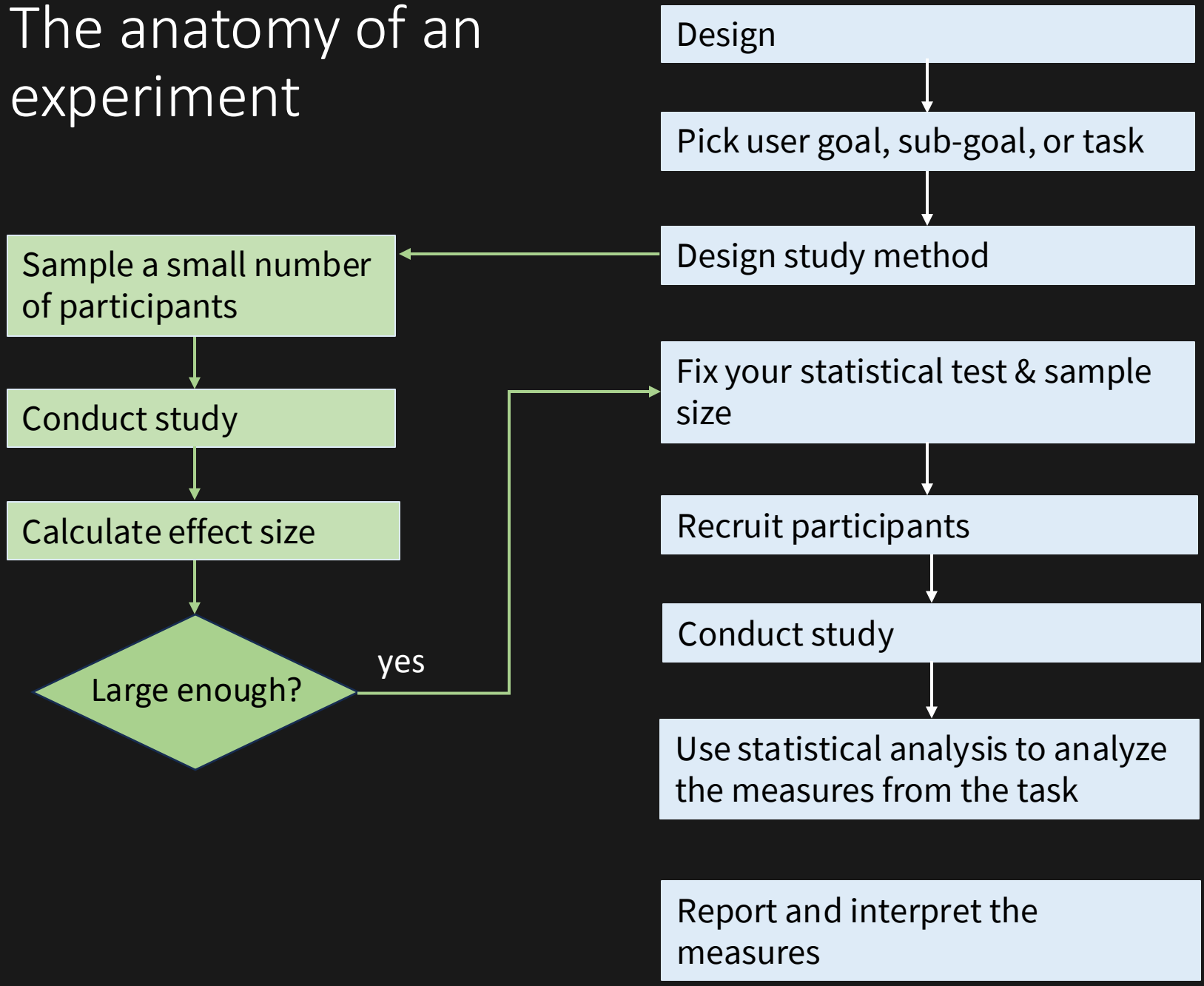
The (simplified) anatomy of an experiment



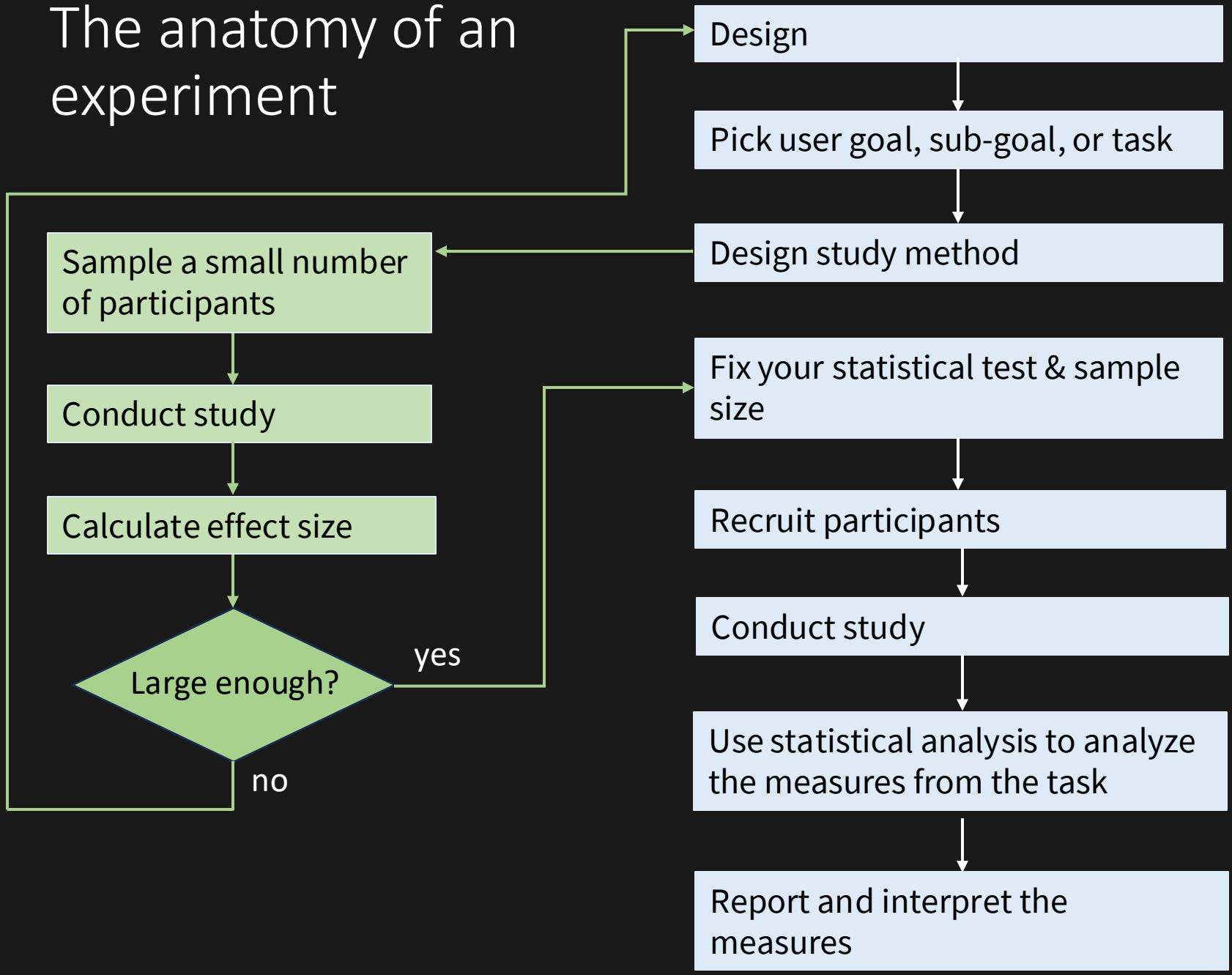
Second problem:

What if the experiment flops?

The anatomy of an experiment



The anatomy of an experiment



Fixing sample size

If sample size = N, should we collect:

- *At least*
- *At most*
- *Exactly*

N responses?

Can we peek at the data and collect more until we reach significance?

Fixing statistical tests

Can we collect the data first, look at various dependent variables until we find one that is significant?

Fixing statistical tests

- Suppose $\alpha = 0.05$
- For a single test, the probability of type I error = 0.05
- *Can we peek at the data and collect more until we reach significance?*
  - Suppose you do this twice
  - The probability of not making a type I error overall:
    - 0.95 X 0.95 = 0.9025
  - The probability of making at least one type I error:
    - 1 – 0.9025 = 0.0975
- Multiple comparisons increase error rate

Pre-registration

Decide on your analysis plan and sample size

Post it to pre-registration frameworks (e.g., OSF)

The document will be timestamped

You may add your hypotheses as well

Pre-registering hypotheses

ARE EMILY AND GREG MORE EMPLOYABLE
THAN LAKISHA AND JAMAL?
A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION

Marianne Bertrand
Sendhil Mullainathan

How to pick a test?

- https://yatani.jp/teaching/doku.php?id=hcistats:start#what_statistical_test_should_i_use
- https://www.coursera.org/learn/designexperiments
- Applied Linear Statistical Models, Kutner et al.

Repeated measures design

- 2 keyboards: baseline, treatment
- 100 paragraphs, sample 10 for each participant
- Each participant does 10 tasks (writes 10 paragraphs) using the assigned keyboard
- The 10 datapoints from each participant are not *independent*
- Datapoints collected on the same paragraph are not independent either
- Need to use tests that account for repeated measures

Example: empirical keyboard evaluation

- Standard transcription task empirical study
- Two 30 minute typing sessions; one for each keyboard (number of phrases depends on typing speed)
- Measuring typing speed and error rate

What kind of data is it?

Mean typing speed is usually normally distributed on the interval

What kind of data is it?

Safe to assume participants' mean typing speed on the two keyboards will have similar variance

Please answer this question in Canvas

What test would you run to analyze typing speed? Select all that apply.

- ☐ One-way ANOVA
- ☐ T-test
- ☐ Pairwise t-test
- ☐ Mann-Whitney
- ☐ Wilcoxon
- ☐ Other (be ready to tell us which one)

You have 120 seconds…

DONE!

What kind of data is it?

Error rate is usually not normally distributed on the interval

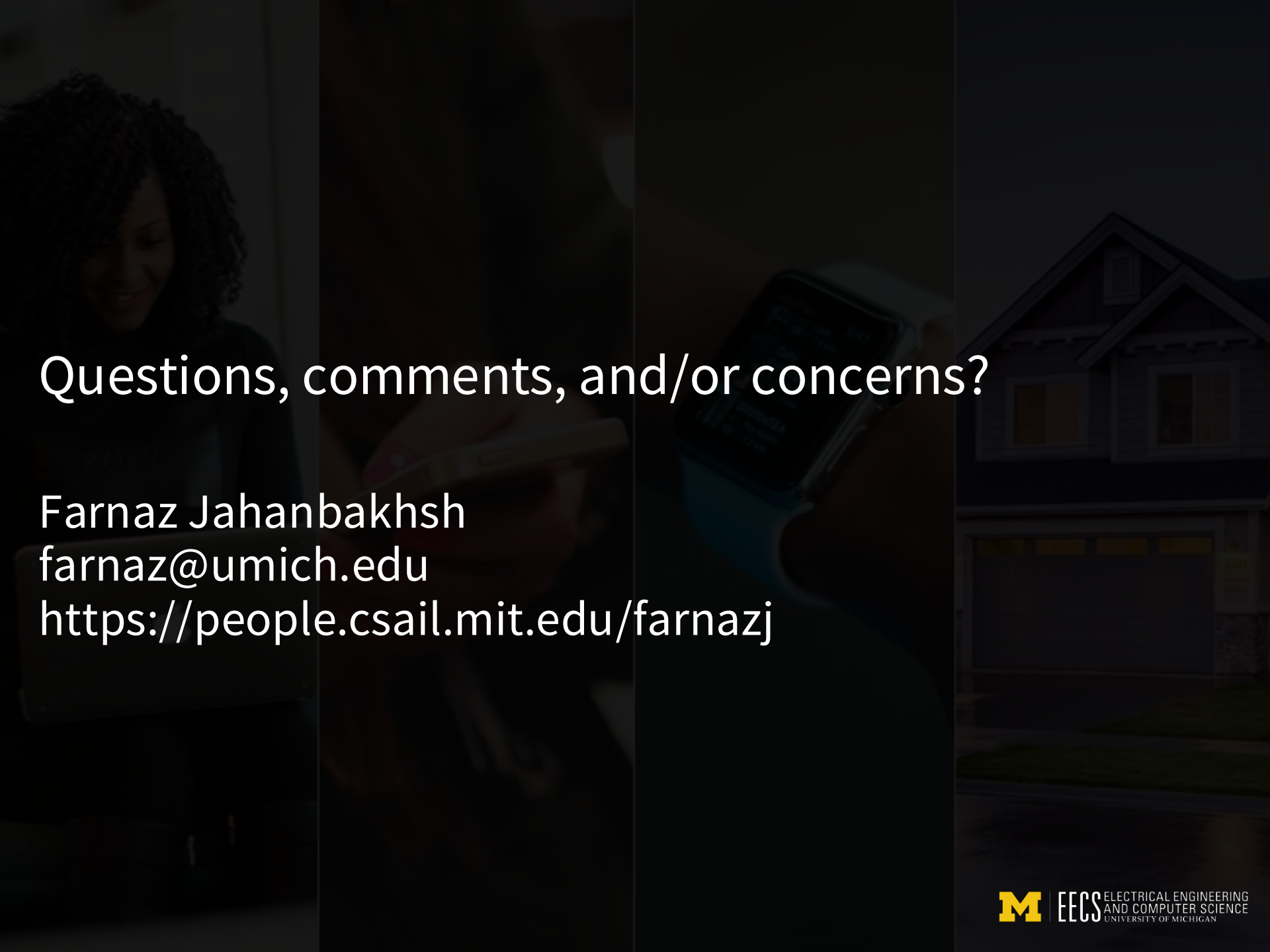
Please answer this question in Canvas

What test would you run to analyze error rate? Select all that apply.

- ☐ One-way ANOVA
- ☐ T-test
- ☐ Pairwise t-test
- ☐ Mann-Whitney
- ☐ Wilcoxon
- ☐ Other (be ready to tell us which one)

You have 120 seconds…

DONE!



Questions, comments, and/or concerns?

Farnaz Jahanbakhsh
farnaz@umich.edu
https://people.csail.mit.edu/farnazj