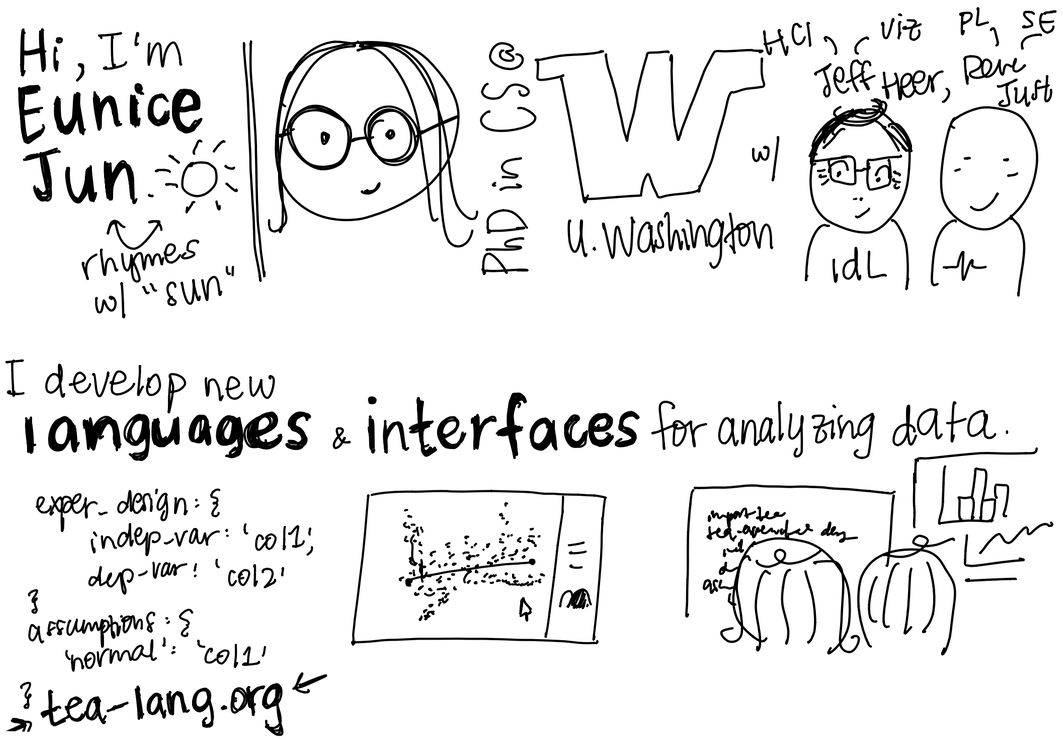
💥Stats crash course💥

Not your traditional intro to experimental design and NHST

Eunice Jun

Intro



tea

tea-lang.org

`pip install tealang`

NHST

tisane

`tisane-stats.org`

`pip install tisane`

linear modeling

Lecture norms

- Embrace that learning statistics is messy and uncomfortable!
- Ask questions! ✋, 💬

Our focus today

- Grow an understanding of experimental design and statistical analysis *terminology*
- Identify practical considerations for the *application* of experimental design and statistical methodology
- Develop a *cognitive framework* for approaching and reasoning through knowledge and gaps in knowledge

Practical considerations

- Grab a drink 🧋🍺☕
- Allot more time than you think for data analysis
- The saying about best laid plans….
- Not the community norm in HCI

Introduction

- Not always required in HCI
- May indicate a methodological maturation in the field - “breakthrough”
- Important to be conversant within and outside our field!

Class format

1. Intro
2. Why do we use statistics?
3. Experimental design
4. Break!
5. NHST basics
6. Common significance tests
7. Linear modeling
8. Connecting all the pieces

Why do we do statistics?

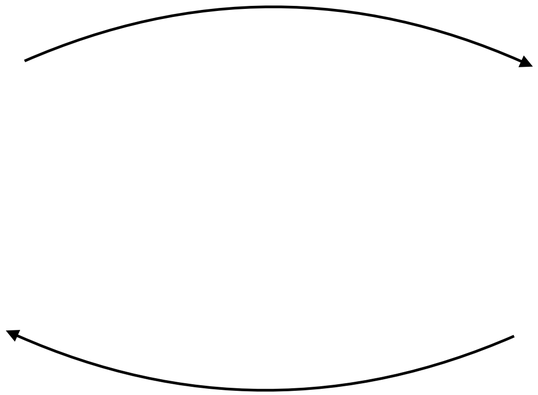
Why do we do statistics?

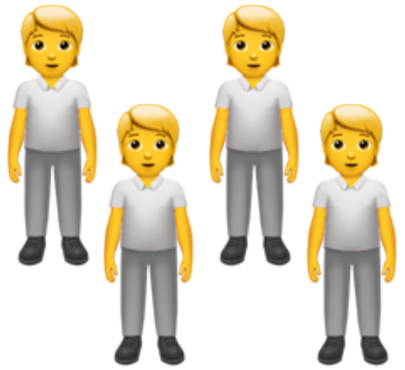
Summarization

Inference

Why do we do statistics?

population



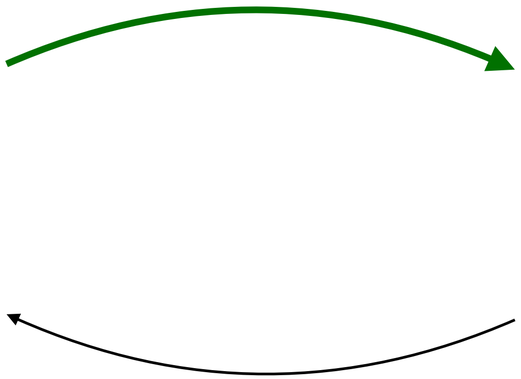


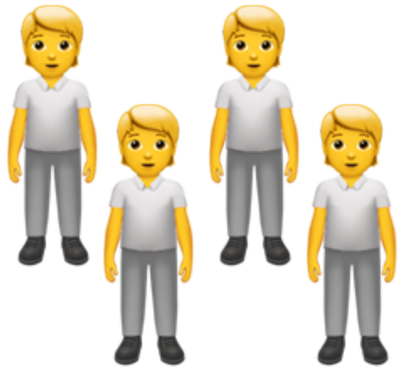
sample

Q: What should we be careful about?

Why do we do statistics?

population



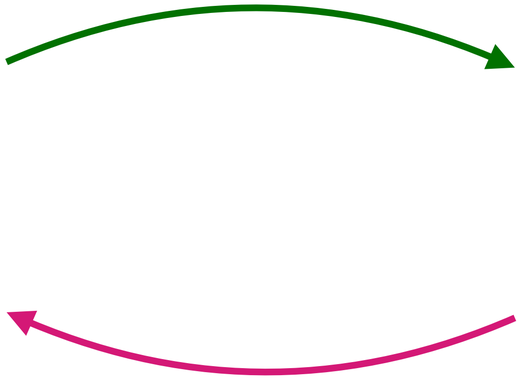


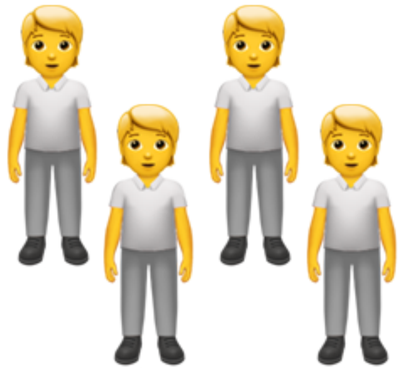
sample

data collection, sampling

Why do we do statistics?

population



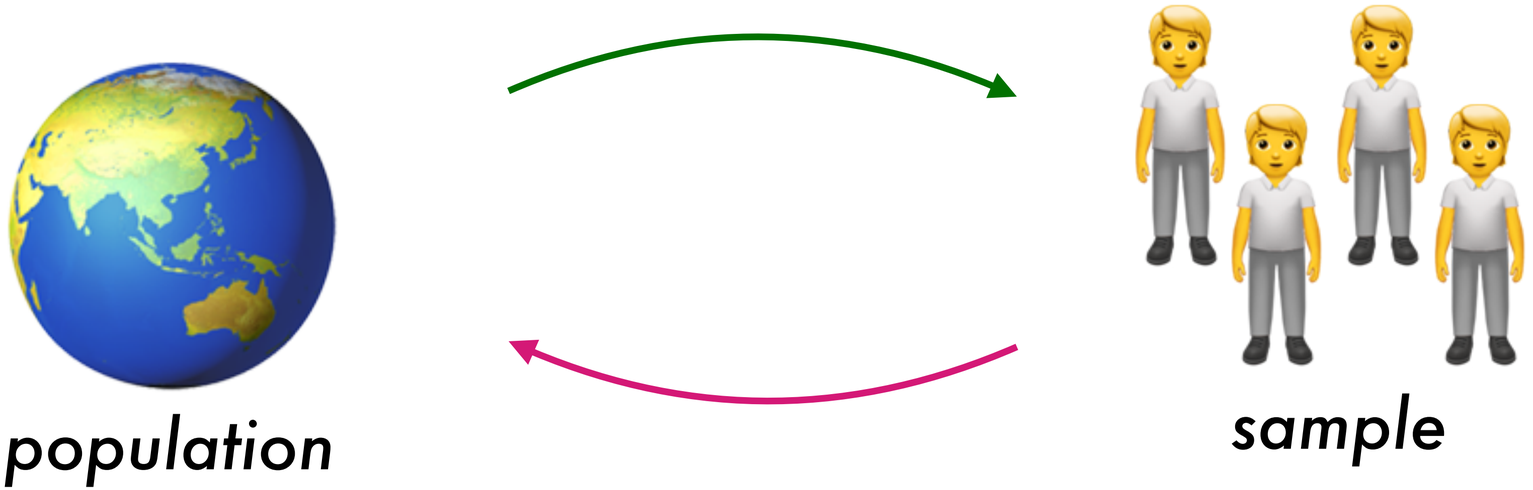


sample

data collection, sampling

data analysis methods

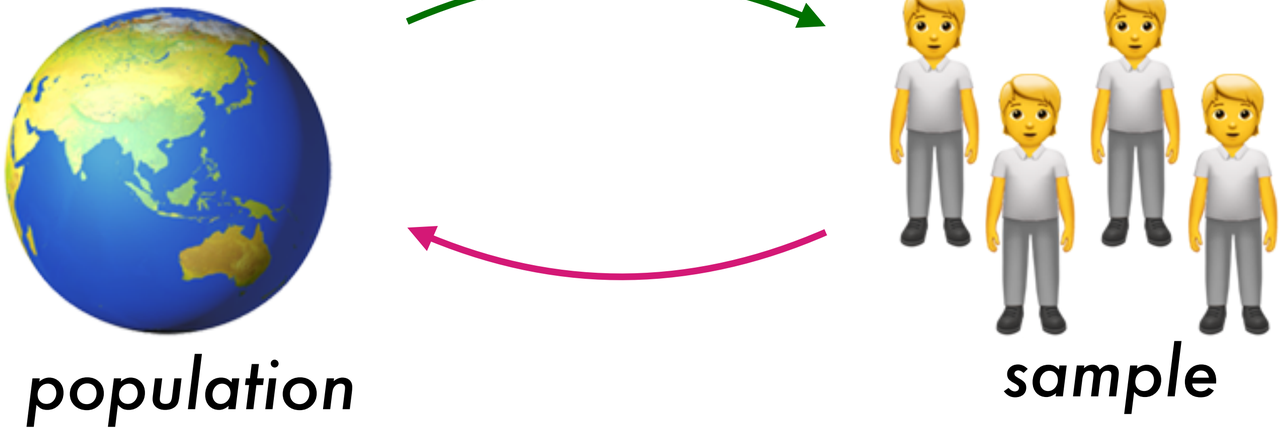
Why do we do statistics?



Experimental design: data collection, sampling

Statistical analysis: data analysis methods

Validity concerns



- **External validity:**
  - Does the experiment generalize?
  - How representative is your sample?
- **Construct validity:**
  - Does the experiment measure what it claims to measure?
  - Does the experiment use adequate proxy measures?
- **Internal validity:**
  - Does the experiment isolate the variable(s) of interest?
  - Does the experiment control for confounders and unwanted effects?
- **Statistical conclusion validity:**
  - Are the conclusions sound based on the chosen statistical test and sample size?
  - How were the conclusions made (e.g., using p values)?

Types of variables

- **Dependent variable**
  - Outcome variable: measured response
- **Independent variable**
  - Experimental variable: Systematically controlled/manipulated
- **Covariate**
  - Experimental variable: Measurable but not controlled (may not be controllable)

Types of variables

- **Categorical (Nominal)**
  - Unordered values or “levels”
  - E.g., {HCI, PLSE, Theory, Architecture}
- **Dichotomous (Binary)**
  - Manually dichotomized or “natural”
  - Categorical with exactly two possible values
  - E.g., {Day, Night}
- **Ordinal**
  - Ordered values (no assumption about equidistant values)
  - E.g., {Low, medium, high}
- **Continuous**
  - Ordered values (equidistant values)
  - E.g., [0, 100]

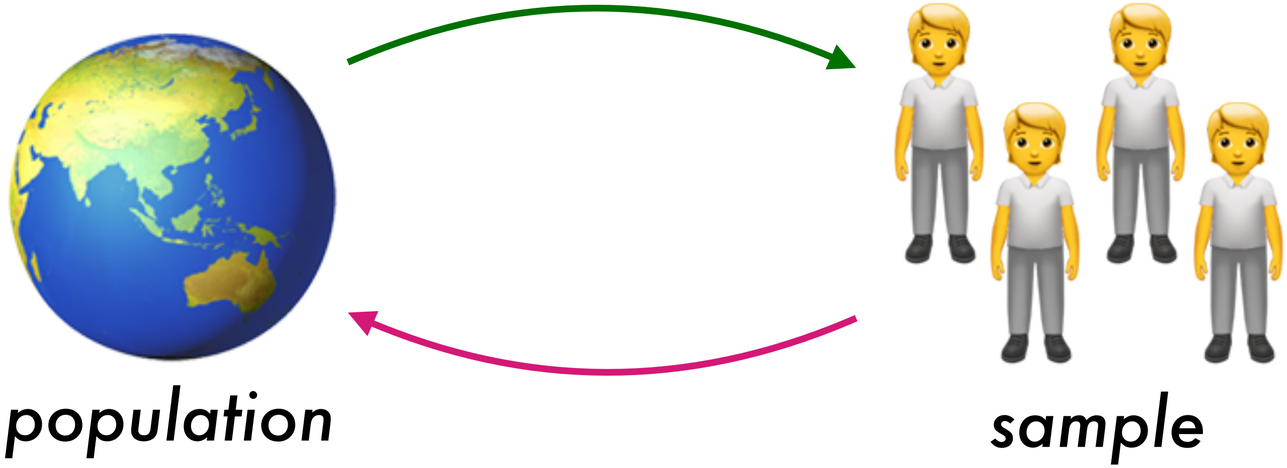
Types of studies

Small break out groups! [3-5 min]

Define and discuss validity concerns in the following:

- Experiments
- Observational studies
- Case studies

Validity concerns



- **External validity:**
  - Does the experiment generalize?
  - How representative is your sample?
- **Construct validity:**
  - Does the experiment measure what it claims to measure?
  - Does the experiment use adequate proxy measures?
- **Internal validity:**
  - Does the experiment isolate the variable(s) of interest?
  - Does the experiment control for confounders and unwanted effects?
- **Statistical conclusion validity:**
  - Are the conclusions sound based on the chosen statistical test and sample size?
  - How were the conclusions made (e.g., using p values)?

Types of studies

Big group discussion

Experiments

Observational studies

Case studies

Types of studies

Big group discussion

Experiments

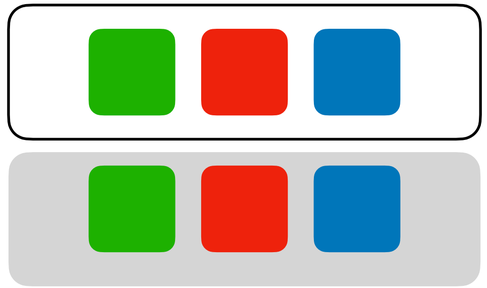
Observational studies

Case studies

Types of studies

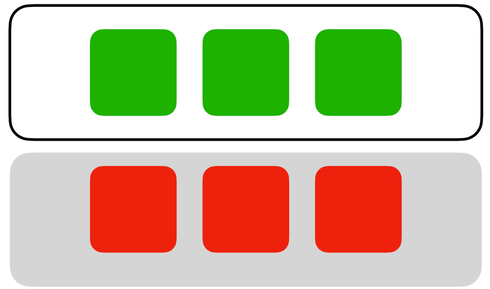
Big group discussion

Experiments



- *Randomization* is key!!
- Independent variable(s) are *directly manipulated/controlled*.
- Repeatable with a testable hypothesis.

Observational studies



- Variables are *not manipulated/controlled.*
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlation.

Case studies

- Focus on one particular subject (“deep dive”).
- Useful for *descriptive* quant. analysis and qual. analysis for context.

Practical considerations

- Clarify what is the **real question** you are after?
  - What **kind of knowledge** are you trying to create?
- Who is your **population**? How are you sampling from that population?
- Is it possible to **manipulate** what you want to manipulate? (This goes back to your motivating question.)
- Consider and write about your study choice through the lens of **validity**.

Study design

Small break out groups! [3-5 min]

Between-subjects design

Within-subjects design

Mixed design

Study design

Small break out groups! [3-5 min]

Between-subjects design

- Each participant has exactly one assignment of independent variable(s)
- E.g., Each participant uses my system or another system. No one uses both.

Within-subjects design

- Each participant has multiple assignments of independent variable(s)
- E.g., Each participant uses my system and another system. Everyone uses both.

Mixed design

- A study with both between-subjects and within-subjects variables.
- E.g., Each participant uses my system or another system. When using my system, each participant uses it in two modes.

Practical considerations

Between vs. Within

- **Sample size:**
  - Between-subjects designs require more participants.
  - Within-subjects designs require fewer participants.
- **Confounding experimental effects:**
  - Participants only provide data once in between-subjects designs.
  - Within-subjects designs run into practice or order effects.

Randomization

- Carryover effects: Learning effects, fatigue effects
- One solution: Latin squares!

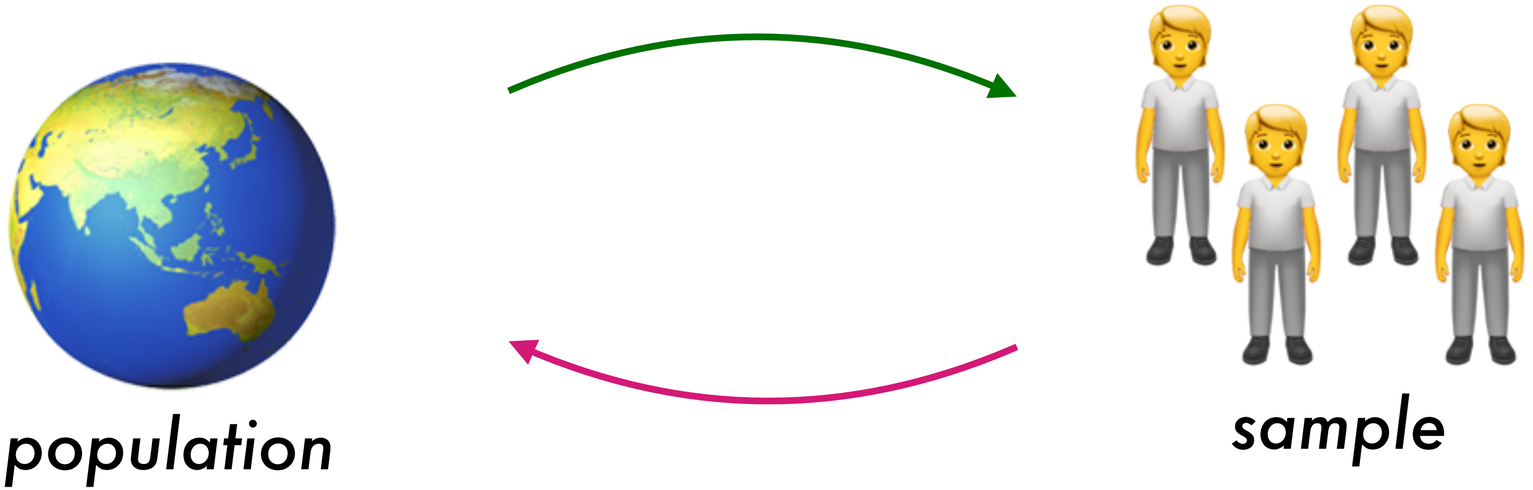
A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C

Randomization

- Carryover effects: Learning effects, fatigue effects
- One solution: Latin squares!
- Other scenarios: self-selection effects, intervention effects (e.g., epidemiology), sensor drift in physiological studies

BREAK!

Why do we do statistics?



Experimental design: data collection, sampling

Statistical analysis: data analysis methods

Null Hypothesis Significance Testing (NHST)

Devil's advocate argument

Hypothesis

- *H0 Null hypothesis*: "Our tool has *no effect* on bugs made while programming."
- *H1 Alternative hypothesis*: "Our tool *effects* the number of bugs made while programming."

Results

- *Reject null*: Not definite evidence that hypothesis is correct. Practically, want to make sure claims are commensurate with evidence (re: types of validity)
- *Fail to reject null:* No evidence that null hypothesis is true or false. (No-op)

NHST Basics

Self + Big group Madlibs

	Reject	Fail to Reject
H0 is true in reality.		
H0 is false in reality.		

α

β

Type I Error

Type II Error

Power

False Positive

True Positive

False Negative

True Negative

NHST Basics

Self + Big group Madlibs

	Reject	Fail to Reject
H0 is true in reality.	Type I Error *probability* = α	*probability* = *1-*α
H0 is false in reality.	Power *probability = 1-*β	Type II Error *probability* = β

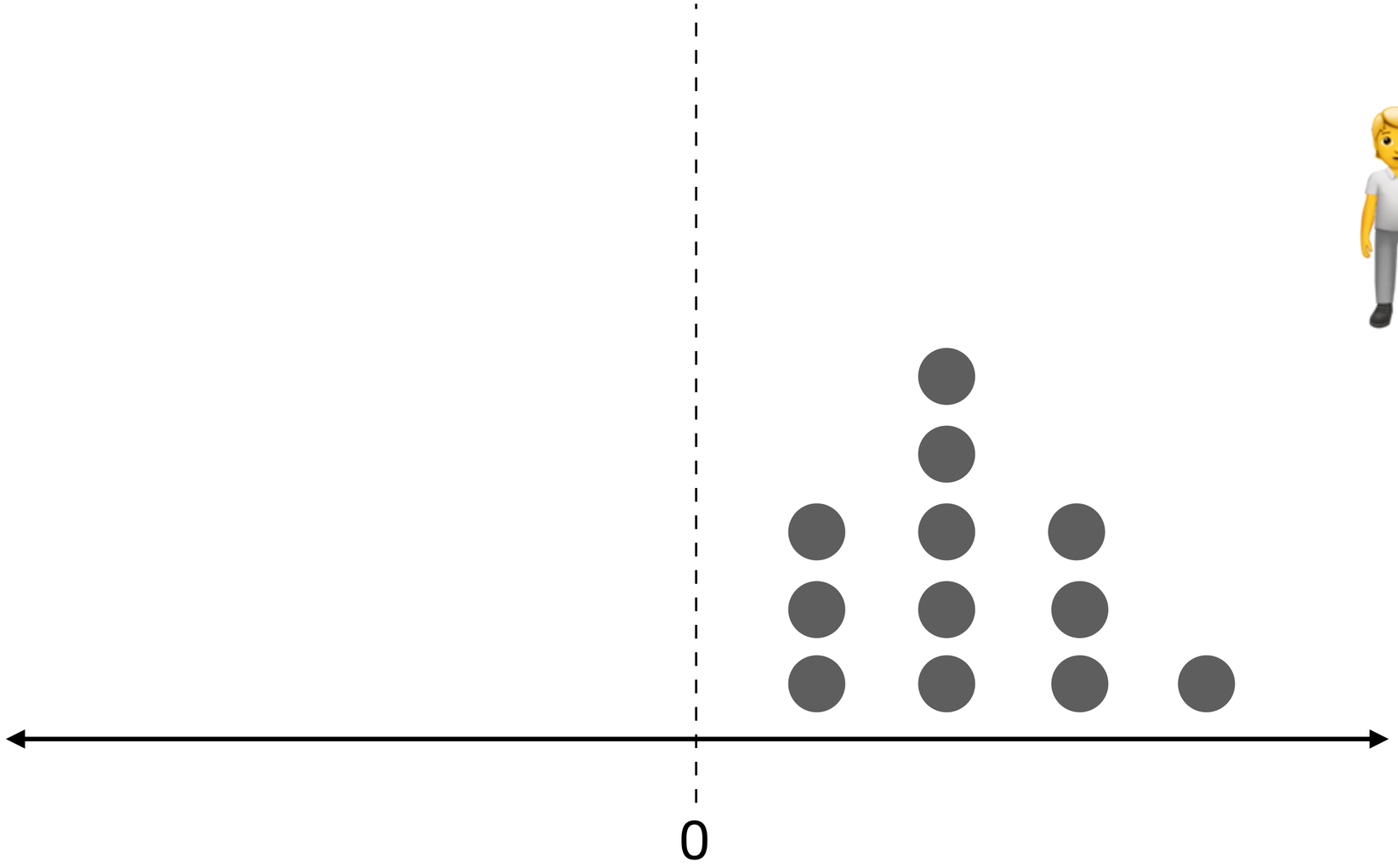
NHST Basics

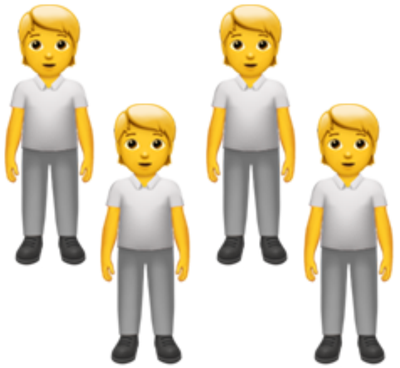
Self + Big group Madlibs

	Reject	Fail to Reject
H0 is true in reality.	Type I Error *probability* = α False Positive	*probability* = *1-*α True Negative
H0 is false in reality.	Power *probability = 1-*β True Positive	Type II Error *probability* = β False Negative

NHST Basics

population



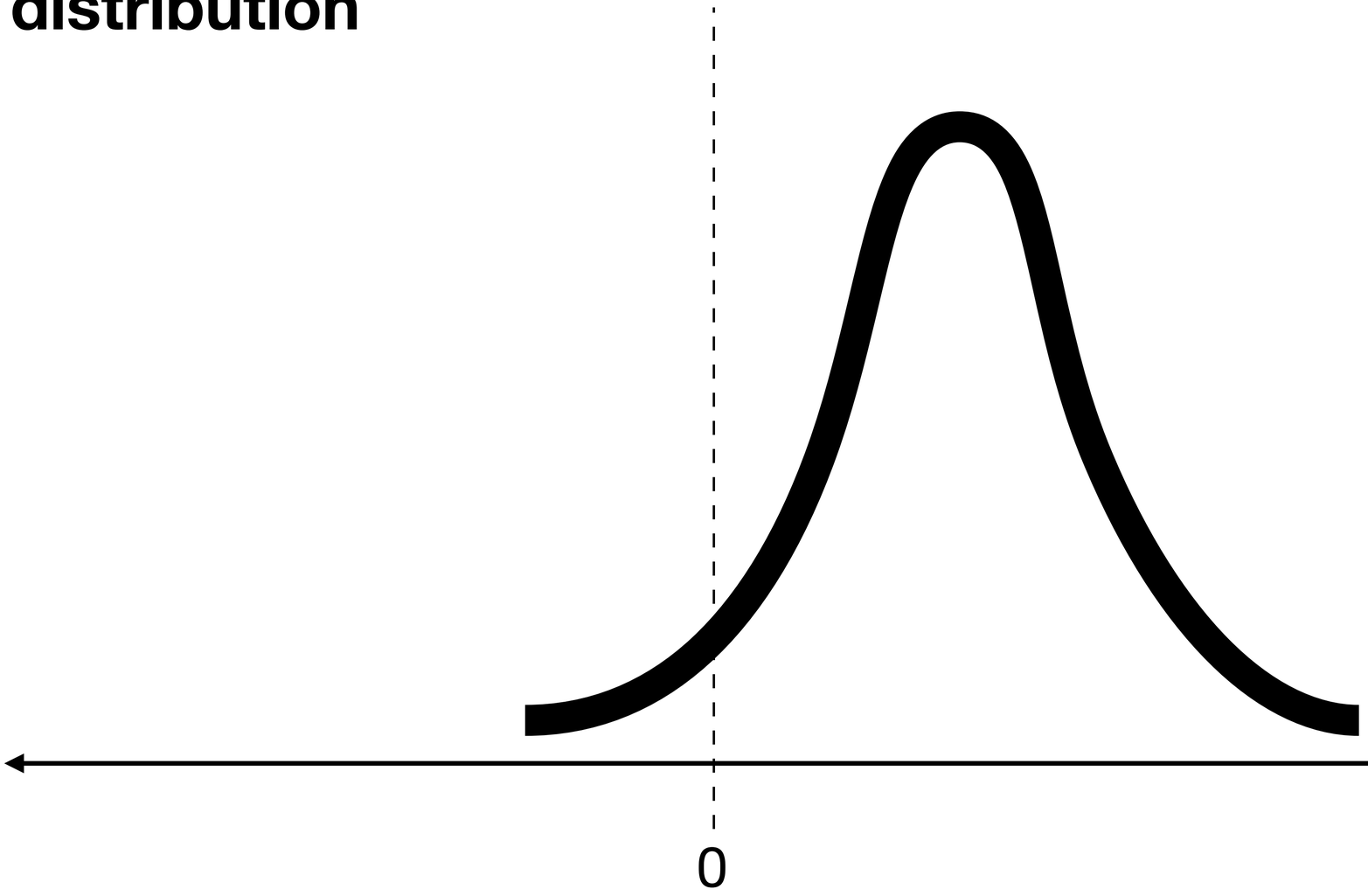


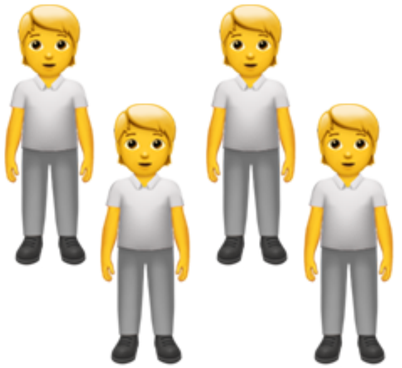
sample

NHST Basics

Sampling distribution

population



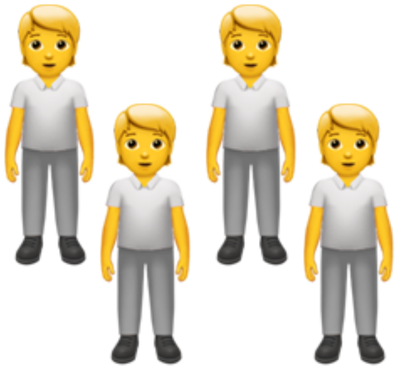


sample

NHST Basics

Sampling distribution

population



sample

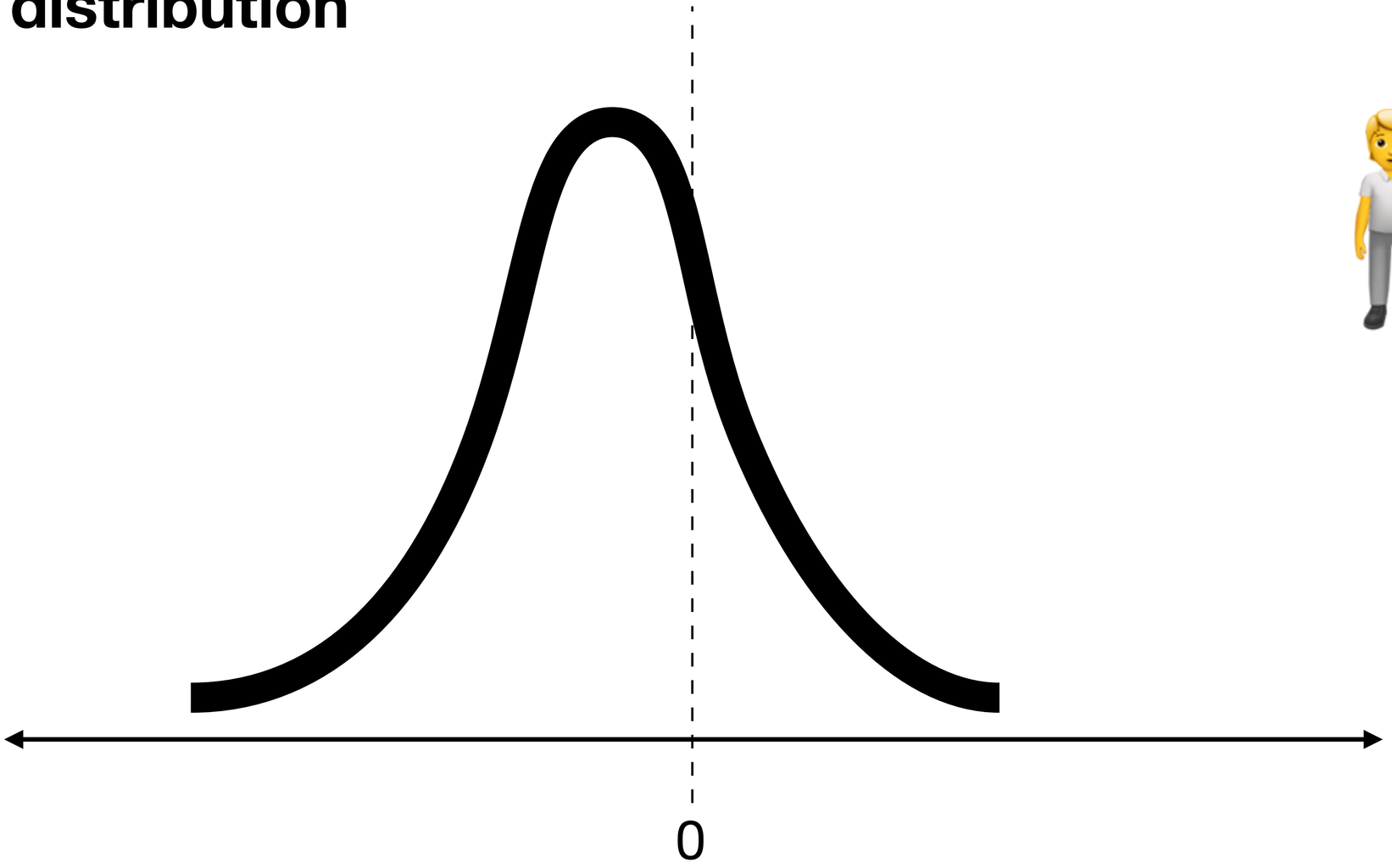
Mean

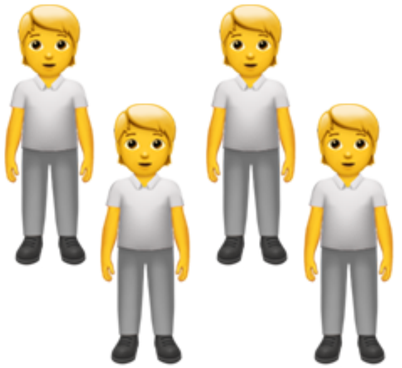
0

NHST Basics

Sampling distribution

population



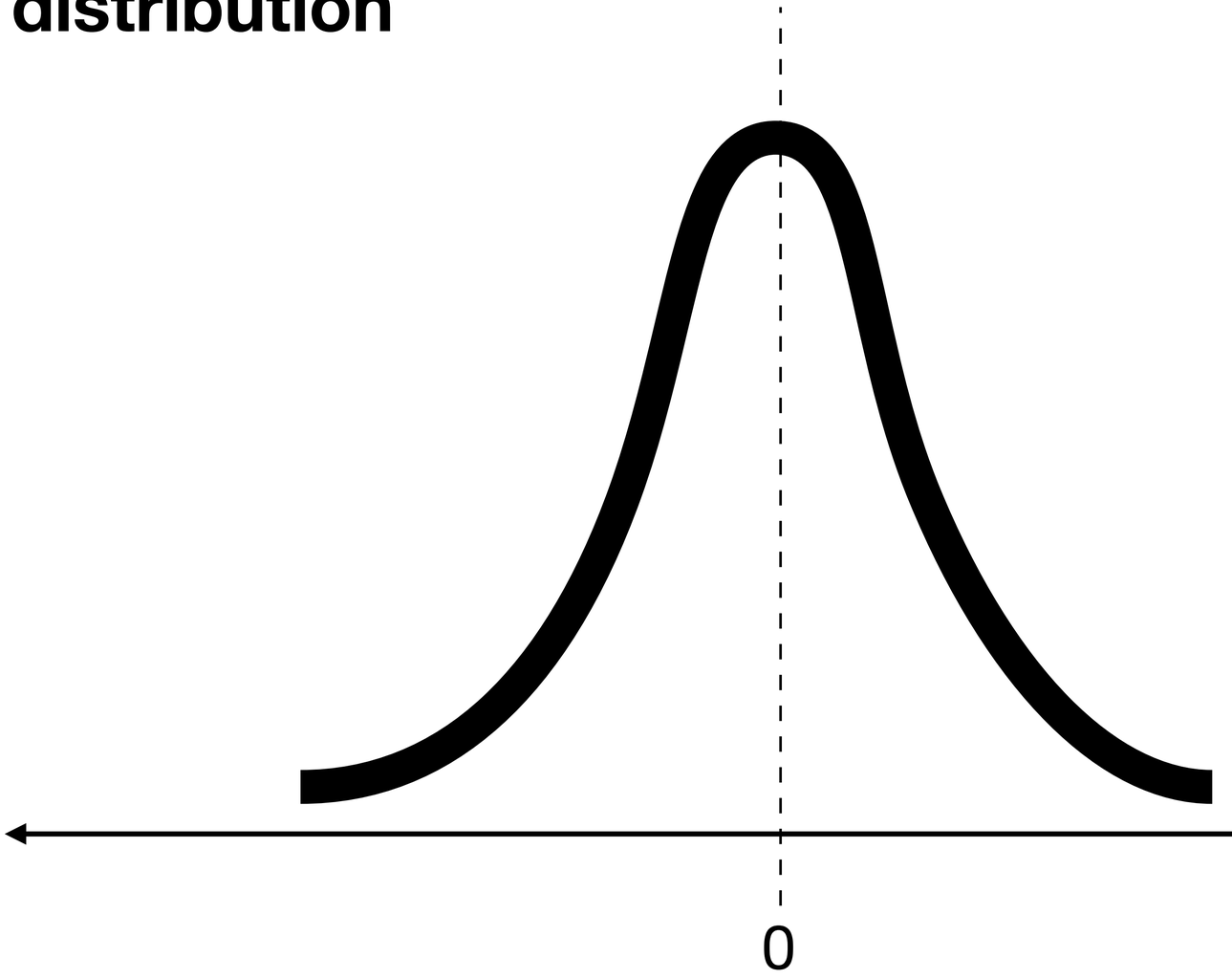


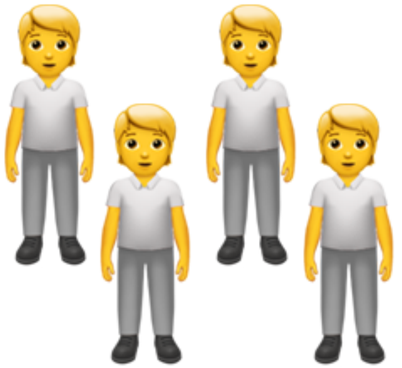
sample

NHST Basics

Sampling distribution

population





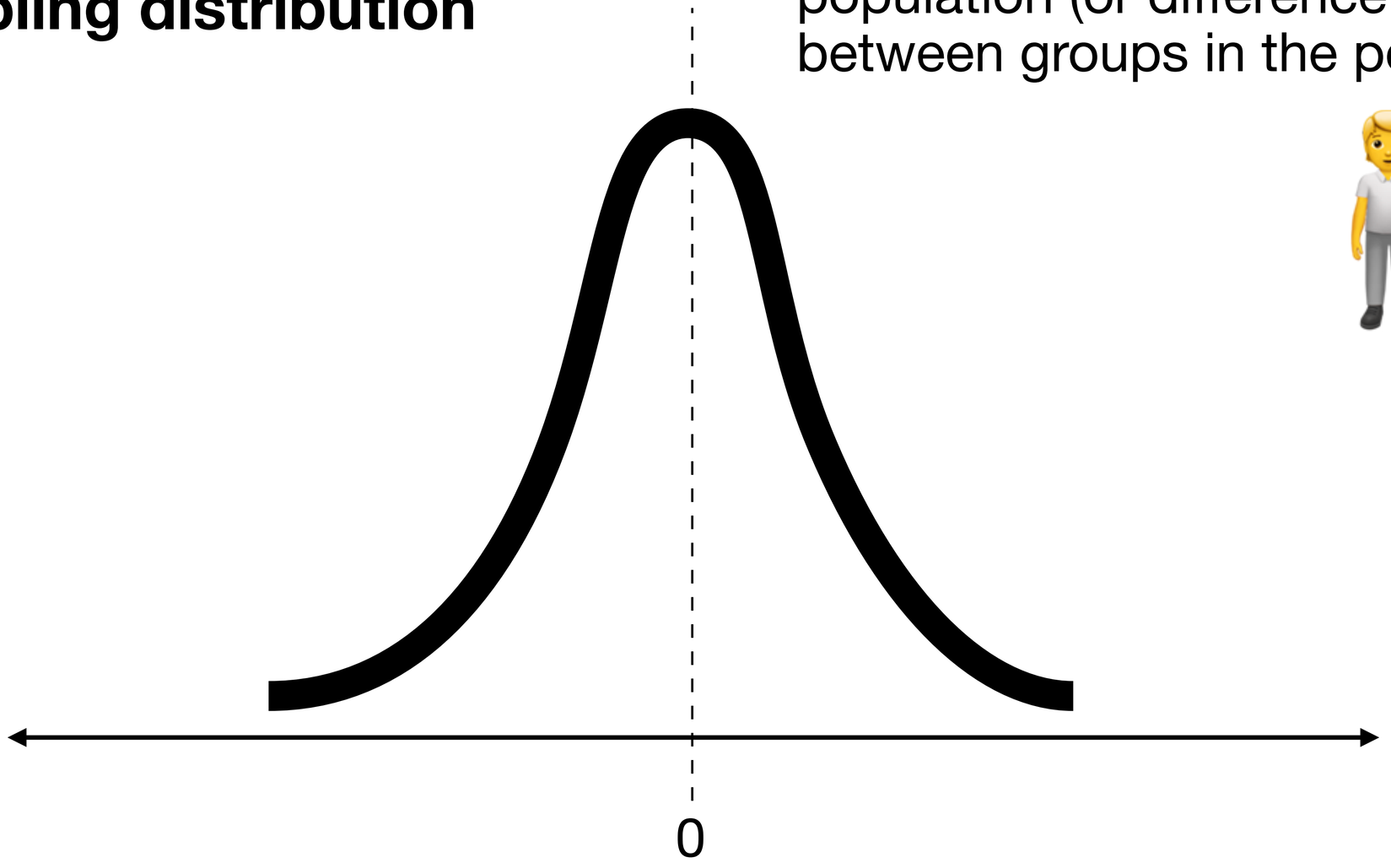
sample

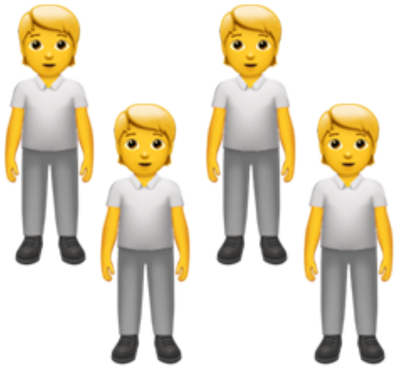
NHST Basics

Null sampling distribution

The **null hypothesis** assumes a null distribution where the mean of the population (or difference in means between groups in the population) is 0.

population



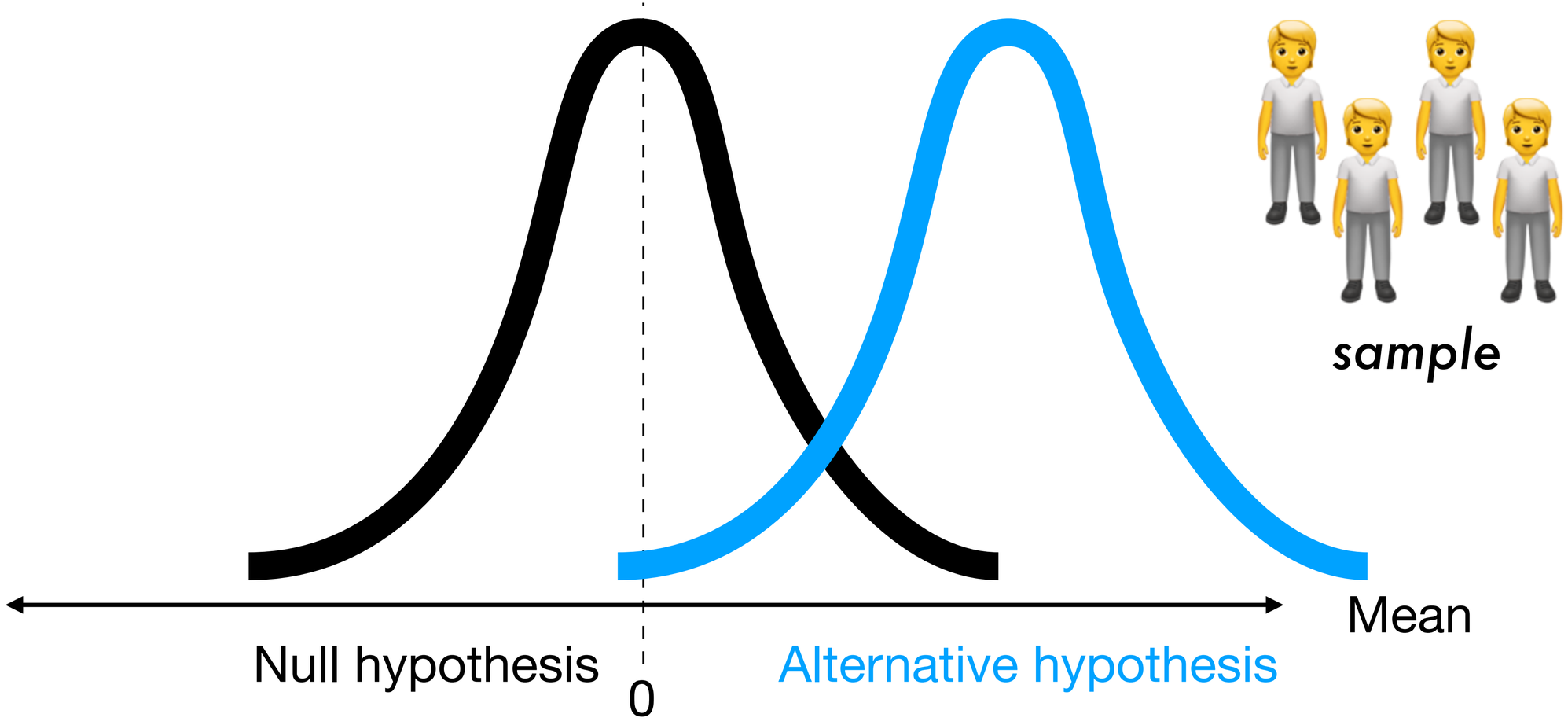


sample

NHST Basics

Assuming the null hypothesis is true, what is **the likelihood of seeing a sample mean just as "extreme"** (or more "extreme") than what we see in our sample?

population



sample

NHST Basics

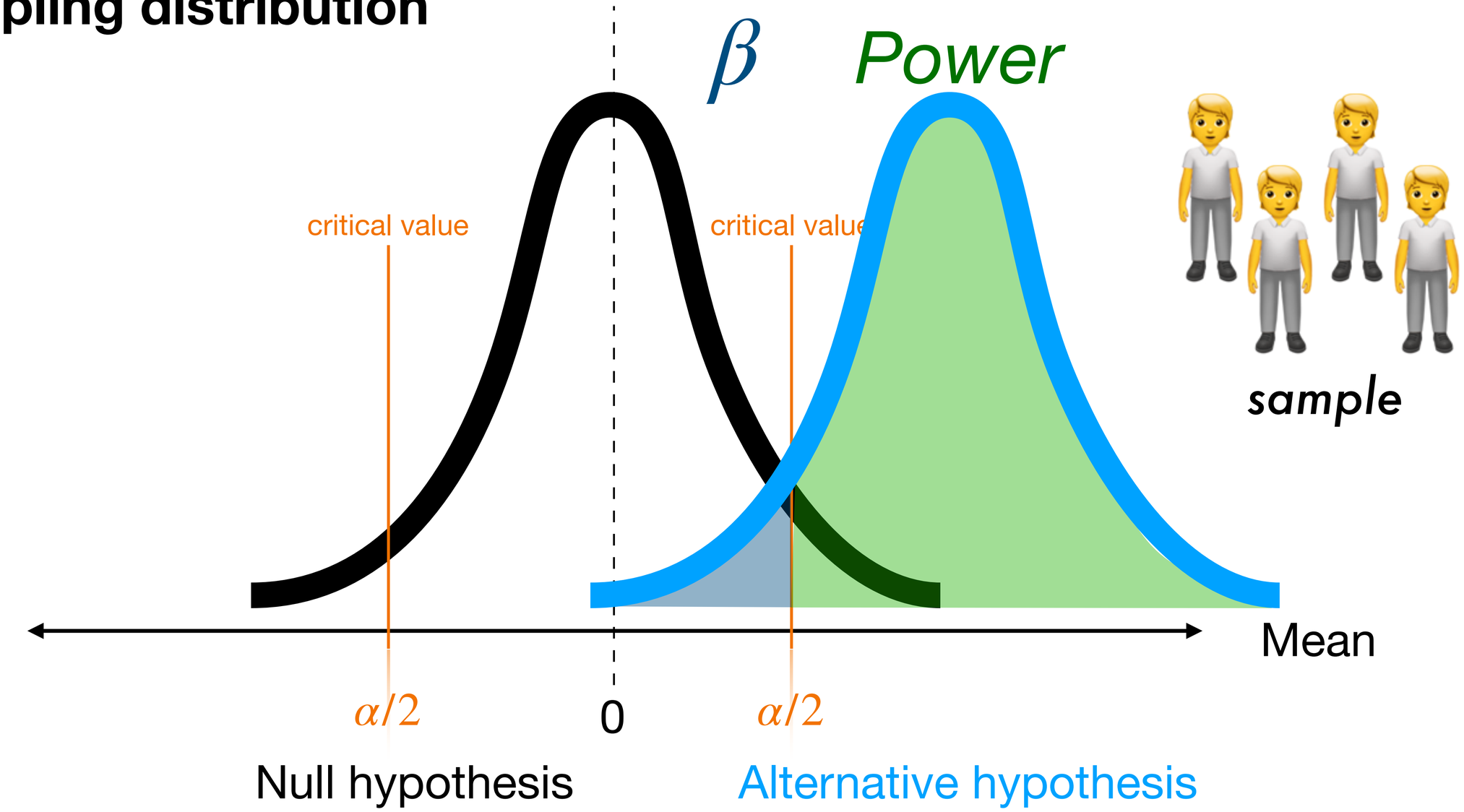
What is a p-value?

- p-value = .03
- “Assuming that the system has no effect on programming productivity, you’d obtain the observed difference or more in 3% of studies due to random sampling error.”
- NOT “If you reject the null hypothesis, there’s a 3% chance that you’re making a mistake.”

p-values as a measure of Shannon entropy - Rafi & Greenland, 2020.

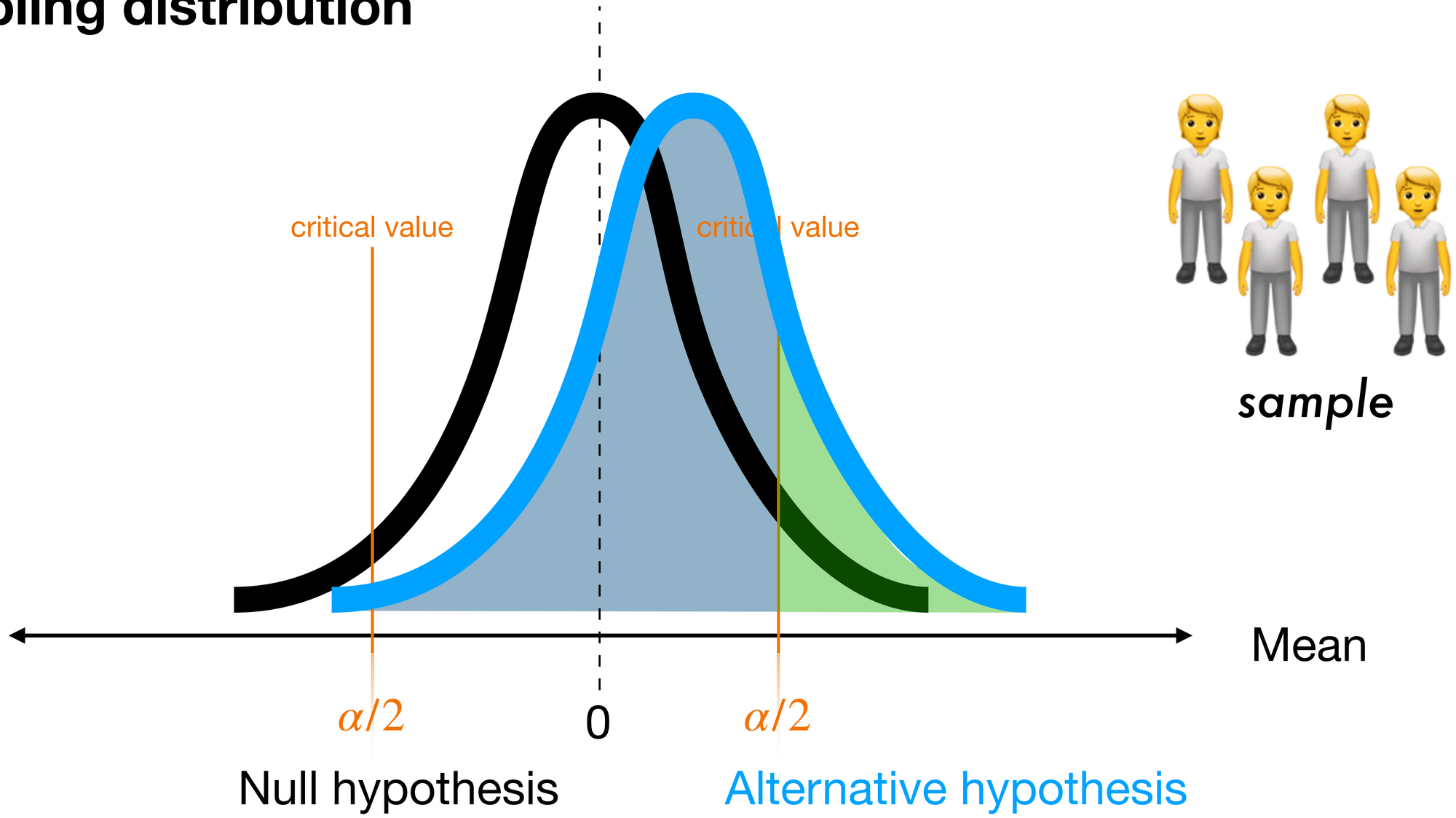
NHST Basics

Null sampling distribution



NHST Basics

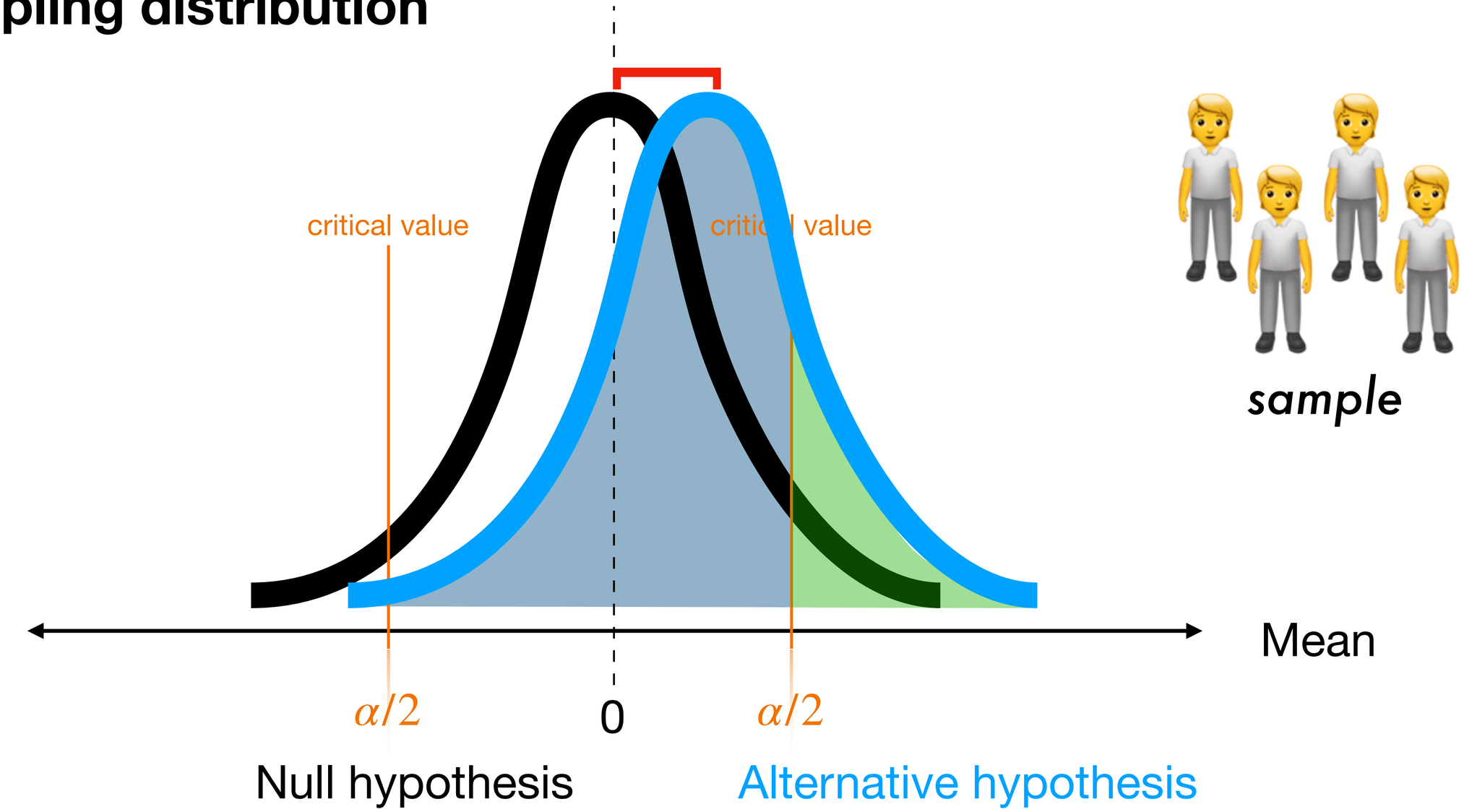
Null sampling distribution



sample

NHST Basics

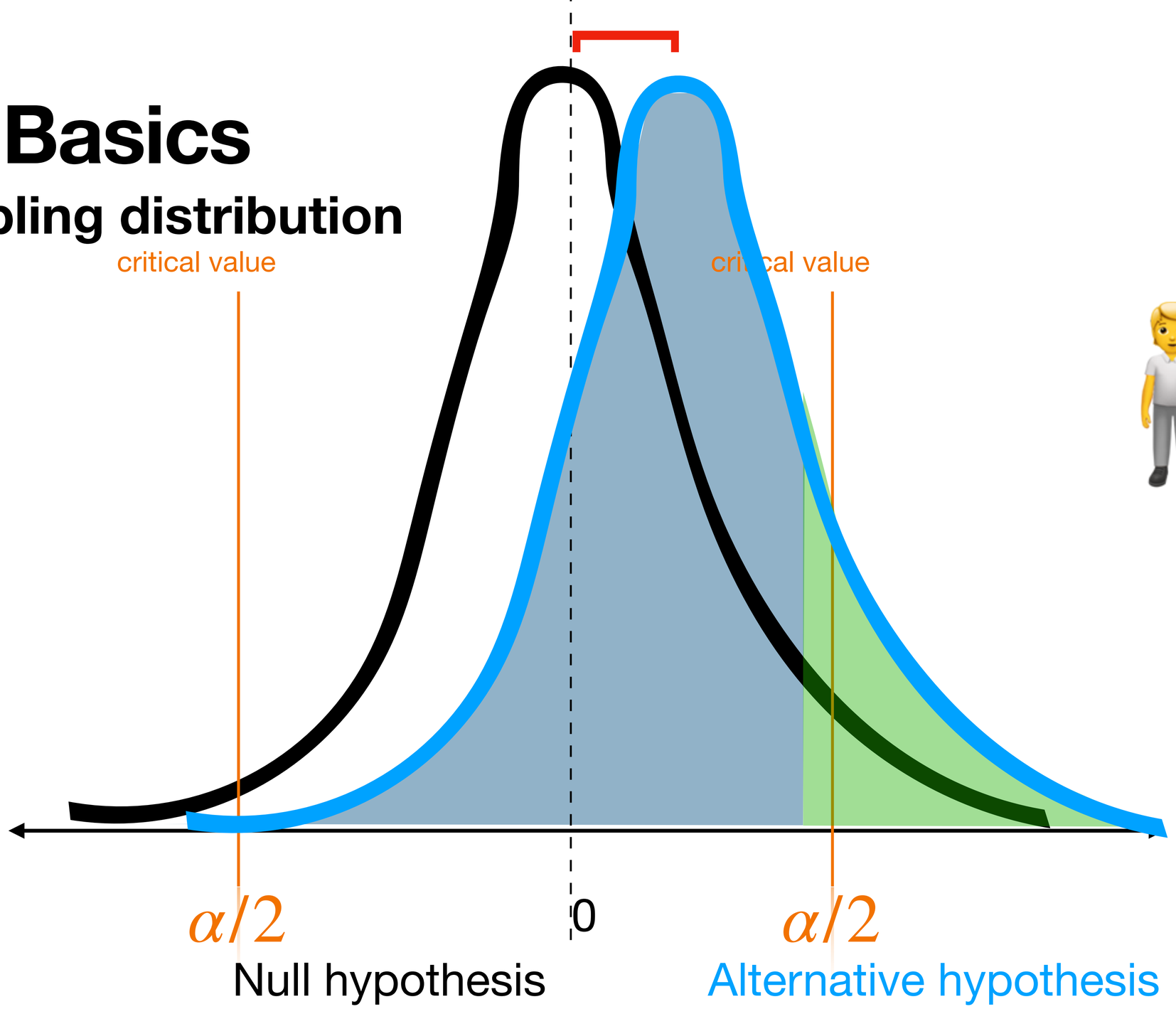
Null sampling distribution

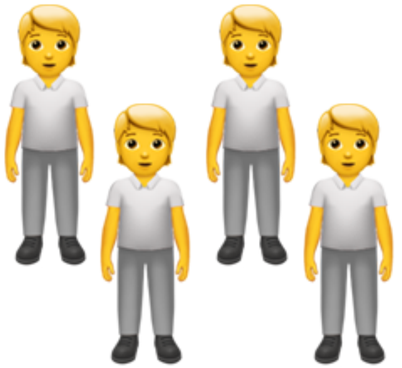


sample

NHST Basics

Null sampling distribution





sample

NHST Basics

Every test tries to account for/get rid of Sampling Error!

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

Student's t-test

$$W = \Sigma[sgn(x_{2,i} - x_{1,i}) * R_i]$$

Wilcoxon signed-rank test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Welch's t-test

$$H = \frac{\Sigma[n_g(M_g - M_{all})^2]}{\frac{N(N+1)}{12}}$$

Kruskal-Wallis test

$$F = \frac{\frac{(SSE_1 - SSE_2)}{m}}{\frac{SSE_2}{n-k}}$$

ANOVA/F-test

$$\chi^2 = \Sigma\frac{(O_i - E_i)^2}{E_i}$$

Chi-Square test

Practical considerations

- **Avoid dichotomous thinking** and writing.
  - ❌ “We found statistically significant results, so our system is better.”
  - ❌ “We did not find statistically significant results, so our system doesn’t work.”
  - ❌ “Our findings are more statistically significant than prior work, so our system is better.”
- Go back to **your research question**!
  - Incorporate qual and quant data to help contextualize the results beyond statistical significance
- Statistical significance != **practical significance**
  - Effect size
  - Power analyses (power = 0.80 in psych)

Common Significance Tests

ROGUE edition

“Everything is regression.”

T-tests

Compare 2 means

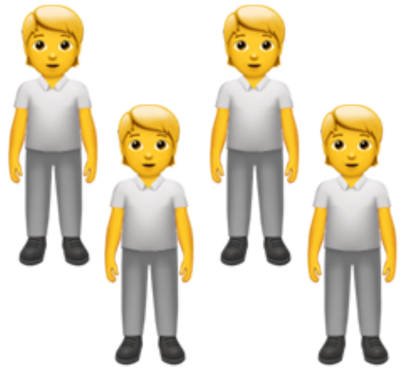
- Student’s t-test
- Welch’s t-test
- Wilcoxon signed-rank test
- Wilcoxon rank-sum test

T-tests

Compare 2 means

- “One sample test” Compare sample mean with known population mean.

=



\+ Error

Population mean Sample mean

$$Y = \text{Intercept} + \text{Error}$$

T-tests

Compare 2 means

- “Two sample test” Compare two groups: Are they from the same population?

control

your system

(X)

System

0=control

1=your system

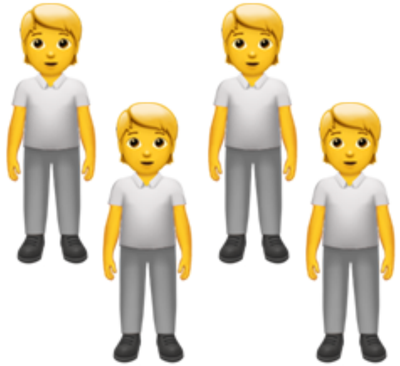
“group”

T-tests

Compare 2 means

- “Two sample test” Compare two groups: Are they from the same population?

=



+

+ Error

All programmers

Sample mean
(“reference mean”)

System
0=control
1=your system
“group”

Y = Intercept + X + Error

T-tests

Compare 2 means

- “Two sample test” Compare two groups: Are they from the same population?

=

+

+ Error

All programmers

Sample mean
(“reference mean”)

System
0=control
1=your system
“group”

Y = Intercept + BX + Error

- “Two sample test” Compare two groups: Are they from the same population?

Control

=

#

+ 0 +

Error

All programmers

Sample mean
(“reference mean”)

- “Two sample test” Compare two groups: Are they from the same population?

Control

=

\+ 0 + Error

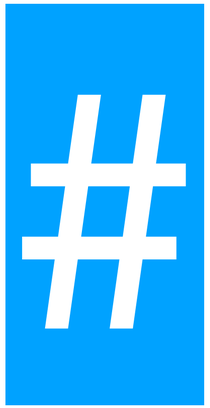
All programmers

Sample mean
(“reference mean”)

Your system

=

\+



\+ Error

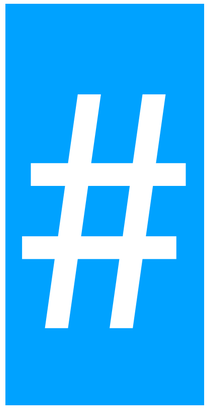
All programmers

Sample mean
(“reference mean”)

T-tests

Compare 2 means

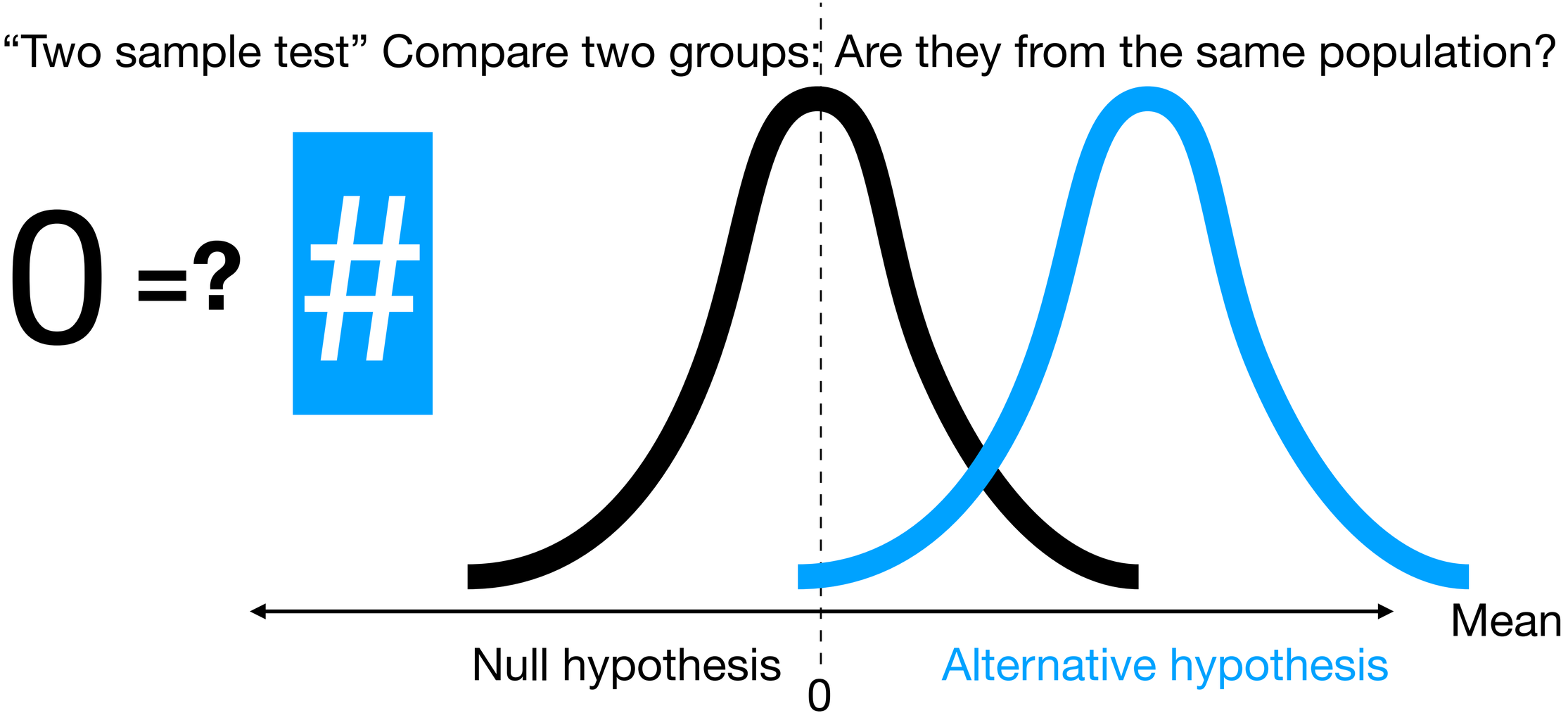
- “Two sample test” Compare two groups: Are they from the same population?

0 =? 

T-tests

Compare 2 means

- “Two sample test” Compare two groups: Are they from the same population?



ANOVA

Compare 3+ means: Are they from the same population?

- One-way ANOVA
- Two-way ANOVA
- Factorial ANOVA
- Kruskal-Wallis test
- …

ANOVA

Compare 3+ means: Are they from the same population?

Control

System A

System B

Control

System A

System B

System A

System B

	System A	System B
Control	0	0
System A	1	0
System B	0	1

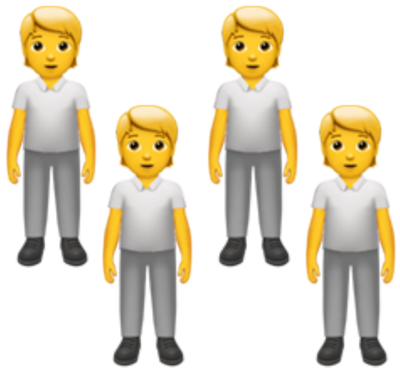
Dummy coding or “one hot encoding”

ANOVA

Compare 3+ means: Are they from the same population?

All programmers

=



Sample mean
("reference mean")

+

System A

+

System B

+ Error

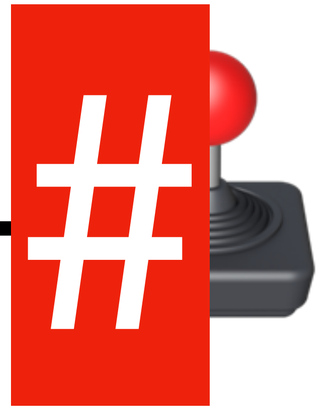
Y = Intercept + X1 + X2 + Error

ANOVA

Compare 3+ means: Are they from the same population?

=

+



+

+ Error

All programmers

Sample mean
("reference mean")

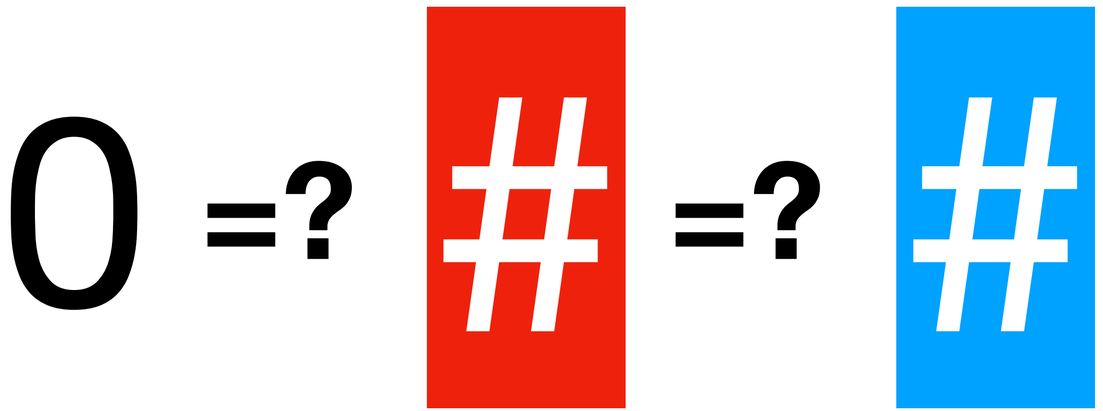
System A

System B

$$Y = \text{Intercept} + B_1X_1 + B_2X_2 + \text{Error}$$

ANOVA

Compare 3+ means: Are they from the same population?



Linear modeling

=

+

+ Error

All programmers

Sample mean
(“reference mean”)

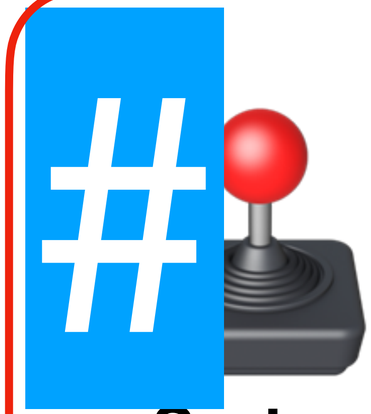
System
0=control
1=your system
“group”

Y = Intercept + BX + Error

Linear modeling: Main effects

=

+



+ Error

All programmers

Sample mean
(“reference mean”)

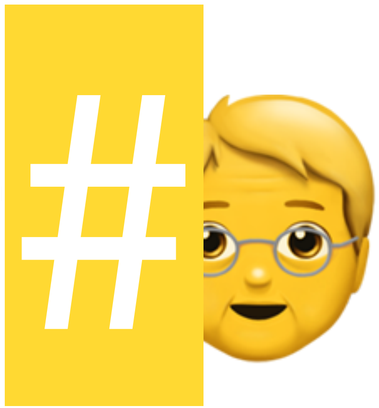
System
0=control
1=your system
“group”

Y = Intercept + BX + Error

Linear modeling: Main effects


=

+

+

+ Error

All programmers

Sample mean
(“reference mean”)

System

0=control

1=your system
“group”

Experience

0 = Low
1 = High

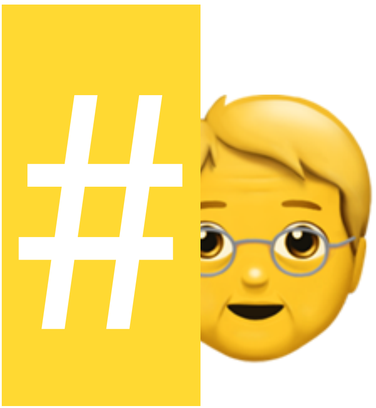
Y = Intercept + B1X1 + B2X2 + Error

Linear modeling: Interaction effects

=

+

+



+ Error

All programmers

Sample mean
(“reference mean”)

System

0=control

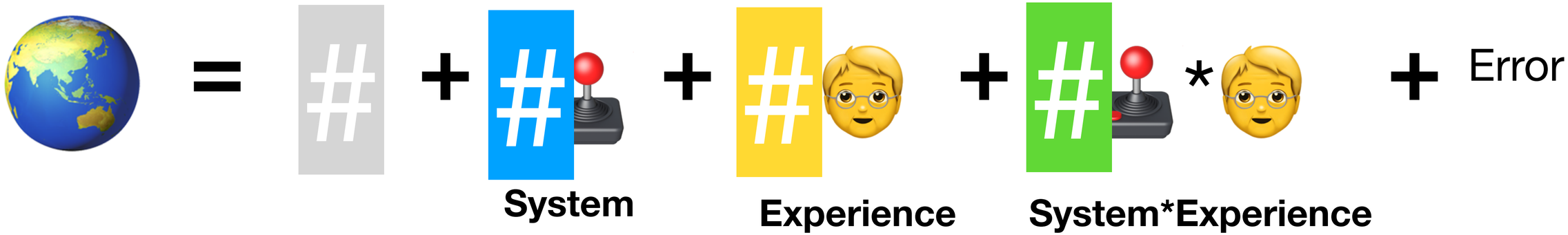
1=your system
“group”

Experience

0 = Low
1 = High

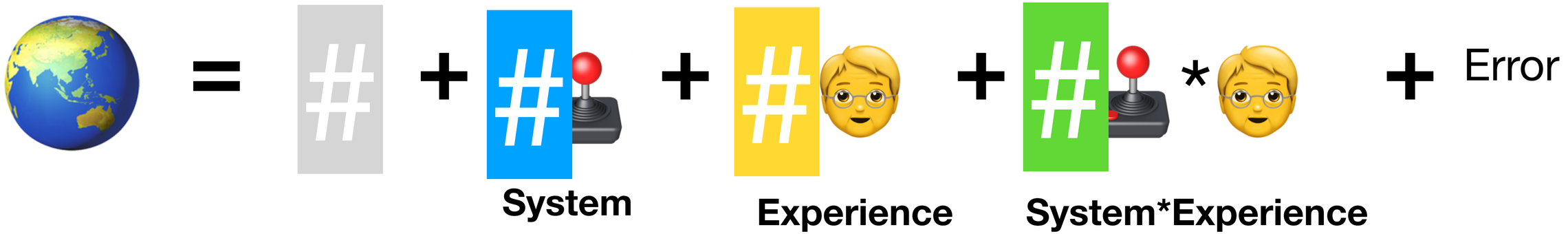
Y = Intercept + B1X1 + B2X2 + Error

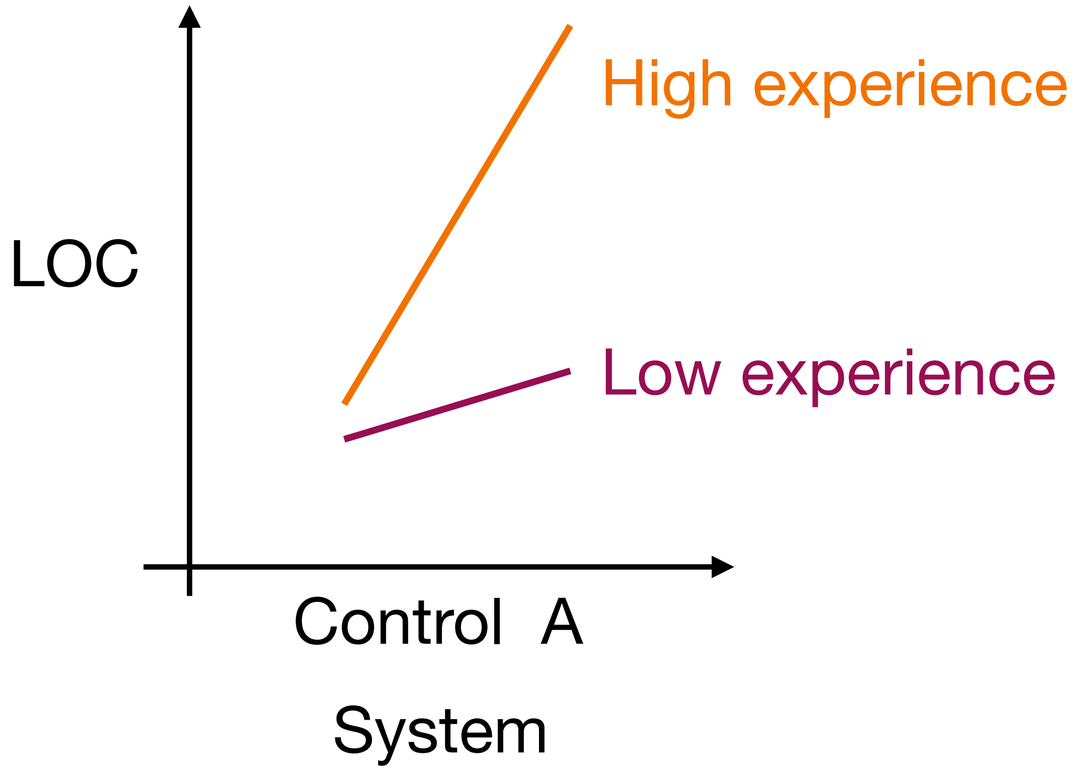
Linear modeling: Interaction effects

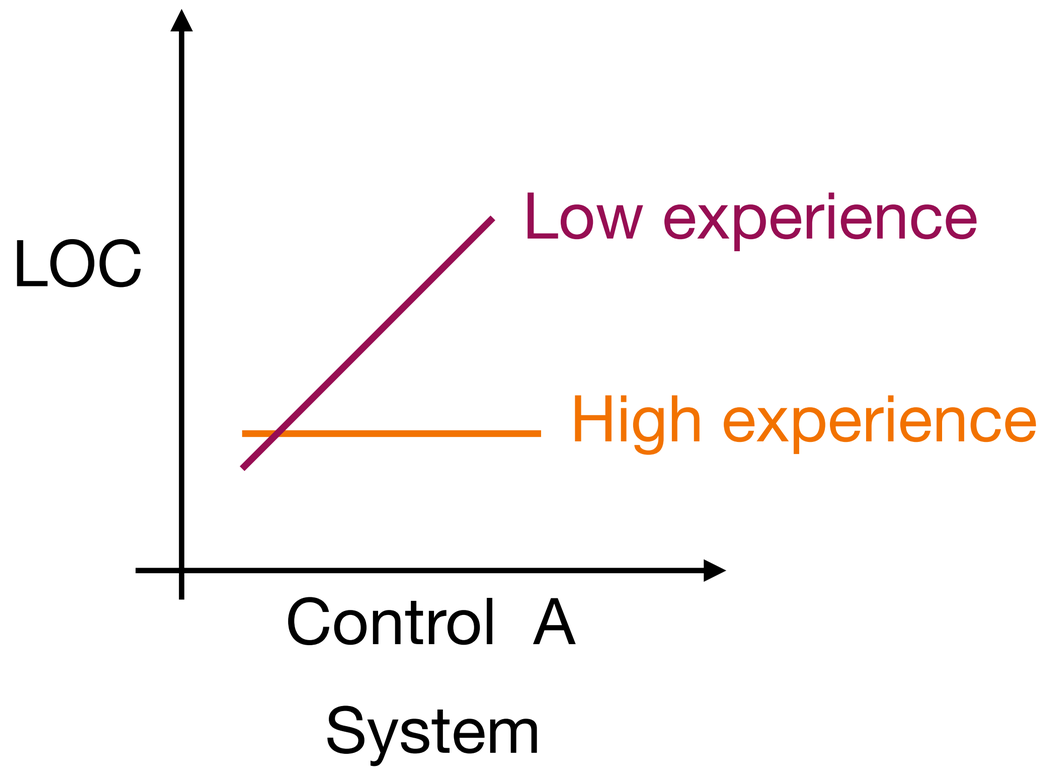


Y = Int. + B1X1 + B2X2 + B3X1*X2 + Error

Linear modeling: Interaction effects







Mixed-effects models

- Consider **sampling**: Does your data have inherent structure! (Intentional or accidental)
- For example: Repeated measures, hierarchical clustering, or non-compositional nesting (e.g., stimuli have 1:1 mapping to conditions)
- Visualization not enough
- **Random effects** (terminology is overloaded: https://statmodeling.stat.columbia.edu/2005/01/25/why_i_dont_use/)
- Maximal random effects for optimal external validity (see Dale Barr)

Generalized linear modeling / Generalized linear mixed-effects modeling

- Regression is a specific use case of GLM (Gaussian family function, Identity link function)
- Depends on dependent variable’s data type + model’s residual distribution
- Takeaway: If your data don’t look as you expect, don’t try to shoe horn into a regression if the assumptions do not hold.

Practical considerations

Statistical computing

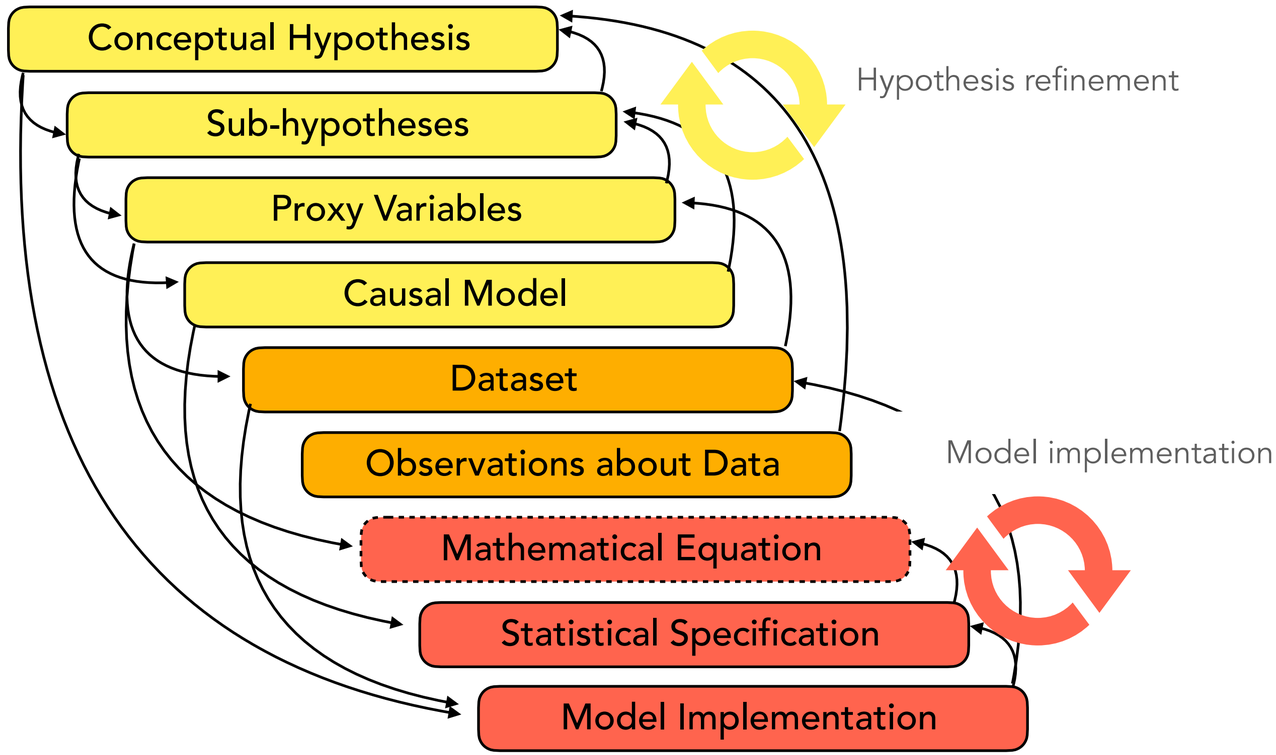
- Equations I have shown you are **different from the R formulae** you program!
- You can use a **linear model** rather than run a “specialty test function.”

Practical considerations

Reporting

- Report using APA conventions.
  - You can look up DF calculations, no need to memorize :)
  - ❌ “We performed a Student’s t-test and found that our system did better ($p < .05$).”
  - Better: “We performed a Student’s t-test. The 25 participants who used our system (M=400 LOC, SD=50.3) compared to the 25 participants in the control group(M=300 LOC, SD=78) wrote significantly more lines of code, t(48)=2.1, p=.04.”

What is it that we're *really* doing?

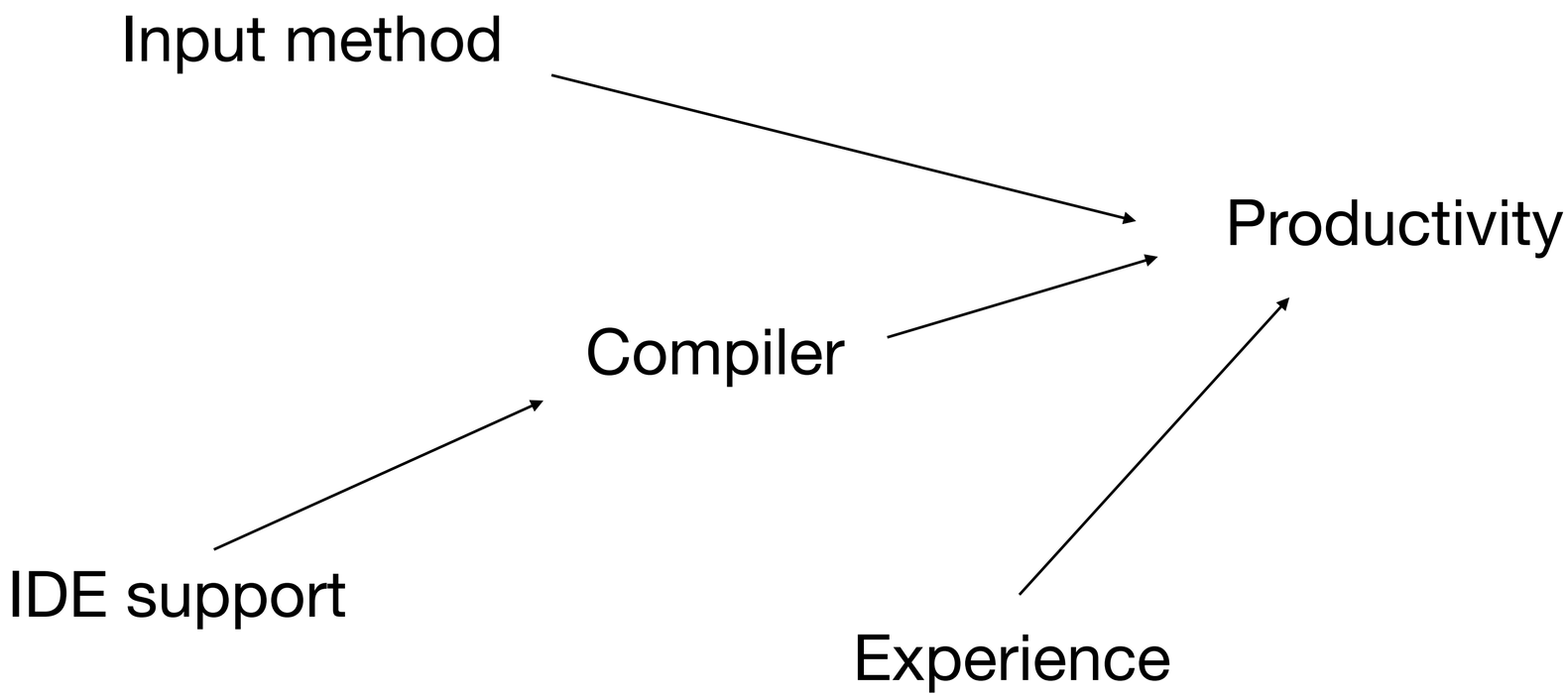


Hypothesis formalization

[TOCHI 2021]

Start with your domain knowledge.

How to pick variables to design for, collect, and analyze?



Example from CSCW 2017

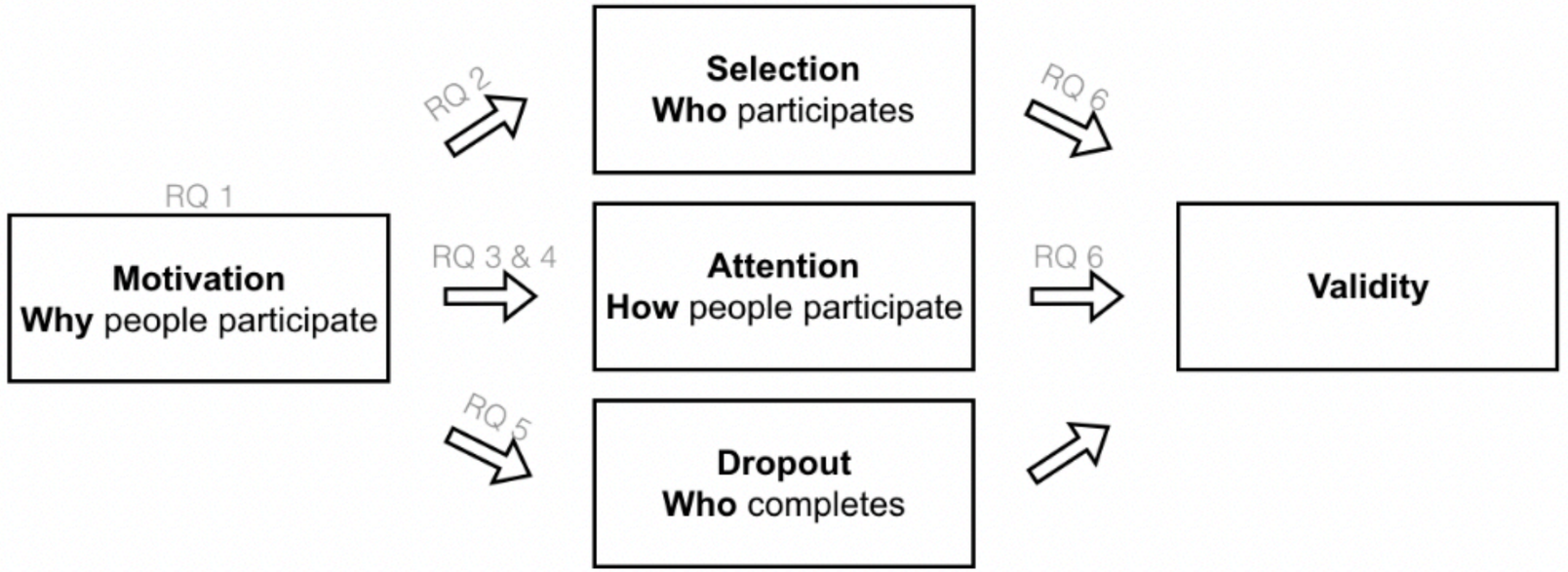
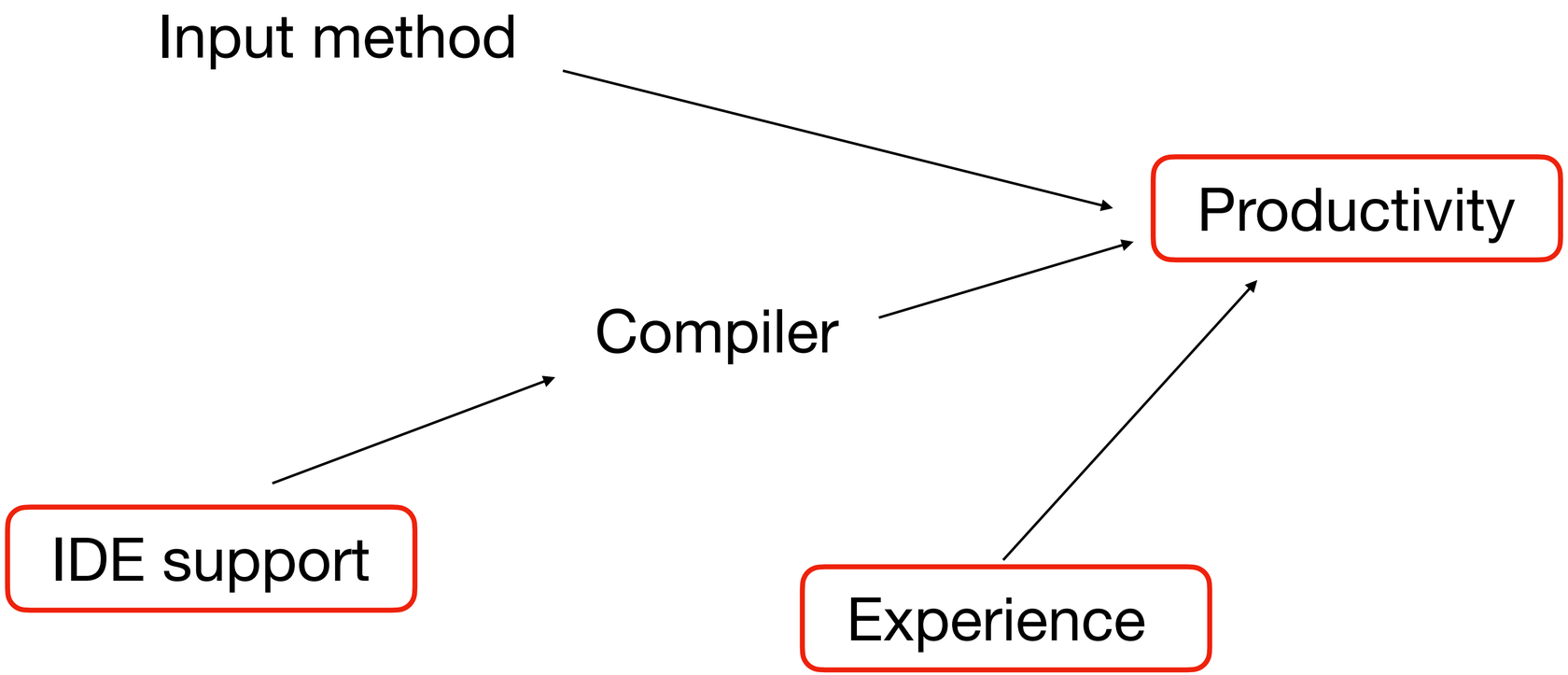


Fig. 1. Theoretical framing of how motivation affects validity. The arrows are labeled with the specific research question that addresses the links between the concepts.

Implications on experimental design

Ask: What do you care about? What do you have control over?



Implication: What is it that you can say?

- Validity
  - External
  - Construct
  - Internal
  - Statistical conclusion
- Confounders
- Randomization
- Causation vs. Correlation

Which statistical method?

- Tendency is to pick a research question based on the methods you know.
- As much as possible, start with the research question. **You may have to learn new statistical methods.**
- Breathe. Consider assumptions about statistical tests.
- **May need multiple experiments/analyses.**

Programming

- Figure out how to express it in your tool :P
  - Idiosyncratic: E.g., All different taxonomies
- Be clear on what the default settings are in R.

False positive discoveries

- Planned vs. unplanned comparisons
- Multiple comparisons corrections
- Fishing:
  - What is guiding your analysis?
  - Specify analyses a priori and what deviations you can reasonably justify by looking someone in the eye!

Pre-registration

- Document analysis plans to narrow or minimize explosion of “forking paths”/“researcher degrees of freedom”
- Common practice in medicine, psychology
- Growing in HCI
- Open Science Foundations: osf.io

Bayesian Statistics

- Explicitly model researcher beliefs as **priors**
- Create **posterior** distributions that update **priors** based on **data**
- Emphasize probability distributions of results
- Move away from statistical significance (i.e., p-values)
- **Interpretations** may be more intuitive
- See Phelan et al. CHI 2019 (https://dl.acm.org/doi/10.1145/3290605.3300709)

Resources at UW

- 599K Empirical Research Methods (Rene Just)
- HCDE and Information School courses:
  - HCDE 544 counts for CSE quals
  - INSC 571 counts for post-quals
  - INSC 570 (general research methods) counts for quals
- Jake Wobbrock's independent study: http://depts.washington.edu/acelab/proj/ps4hci/index.html
- CSSS consulting hours: https://csss.uw.edu/

Resources online

- UCLA Psychology in Action (great for conceptual questions)
- UCLA IDRE (great for conceptual and programming questions)
- Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC medical research methodology*, *20*(1), 1-13. (P-values and Shannon entropy)
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge. (in the Quantitative Psychology canon, my personal go-to)
- Martin, D. (2007). Doing psychology experiments. Nelson Education. (I have not read this myself, but it is a classic experimental psych textbook with an entire chapter on between- vs. within-subjects studies.)
- Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*, 328. (Great discussion of how modeling choices affect generalizability of findings and theory.)
- In general, to get ideas for experimental design, I have read child development/infant studies papers. Studies with babies have to be clever since babies can’t do much! :P

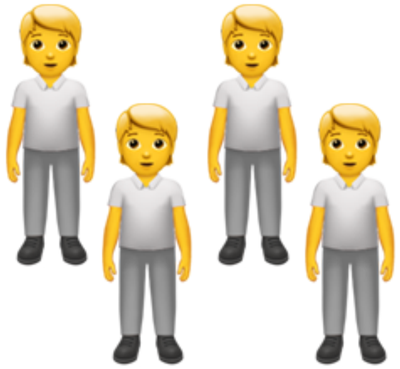
A few last practical considerations

- What is your ***real* research question**?
- What is the **population** (e.g., users) you care about? How will you **sample** that population?
- What kind of **study** is appropriate to understand that sample and population?
- What are your **null** and **alternative hypotheses**?
- How will you **interpret** your statistical results?

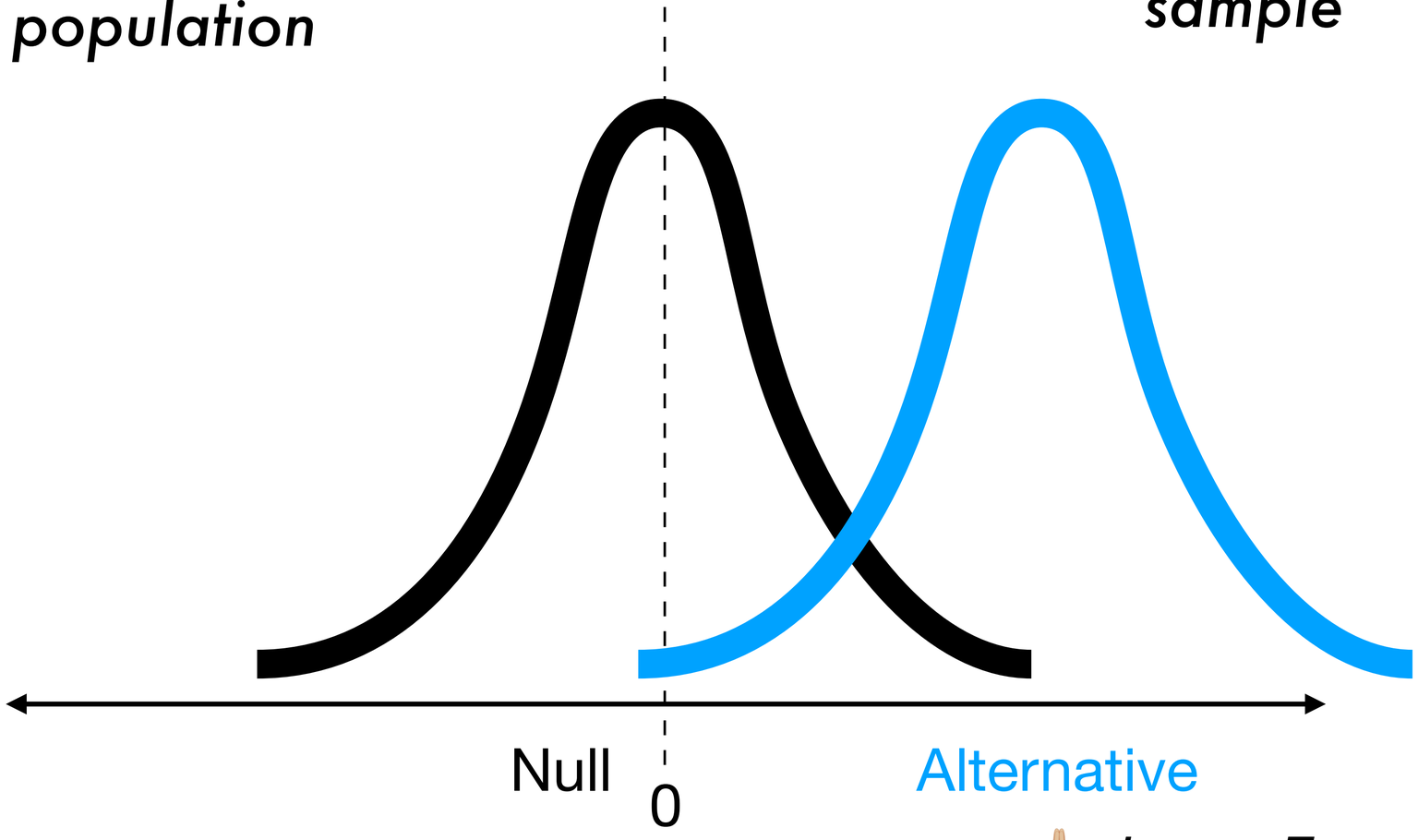
population

Experimental design

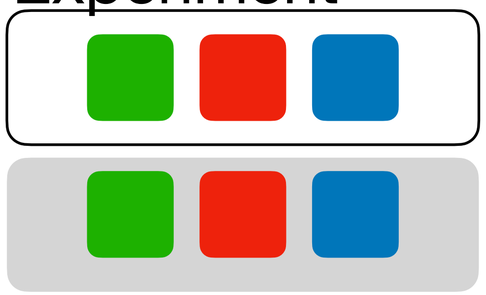
Statistical analysis



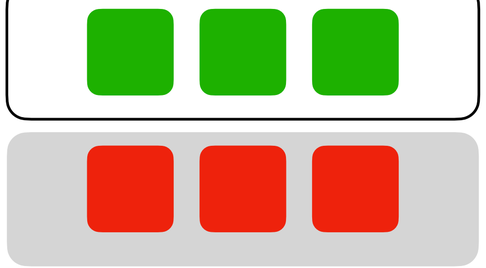
sample



Experiment



Observational



Case study

🦤 @eunicemjun

✉️ emjun@cs.washington.edu

🙏 *James Fogarty, Anant Mittal, Alex Kale, Rene Just*

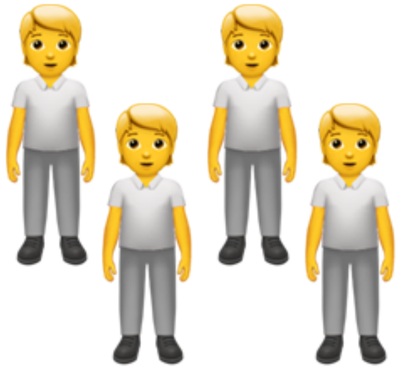
Personal takeaways [4 minutes]

- Terminology: What is one new term that you have learned or thought about in a new way?
- Application: What is one concept in experimental design or statistical analysis that you will apply to your next project?
- Cognitive framework: What is a step that you often overlook when reading about or authoring analyses?
- Question: What remains unclear to you? What more do you want to learn about?

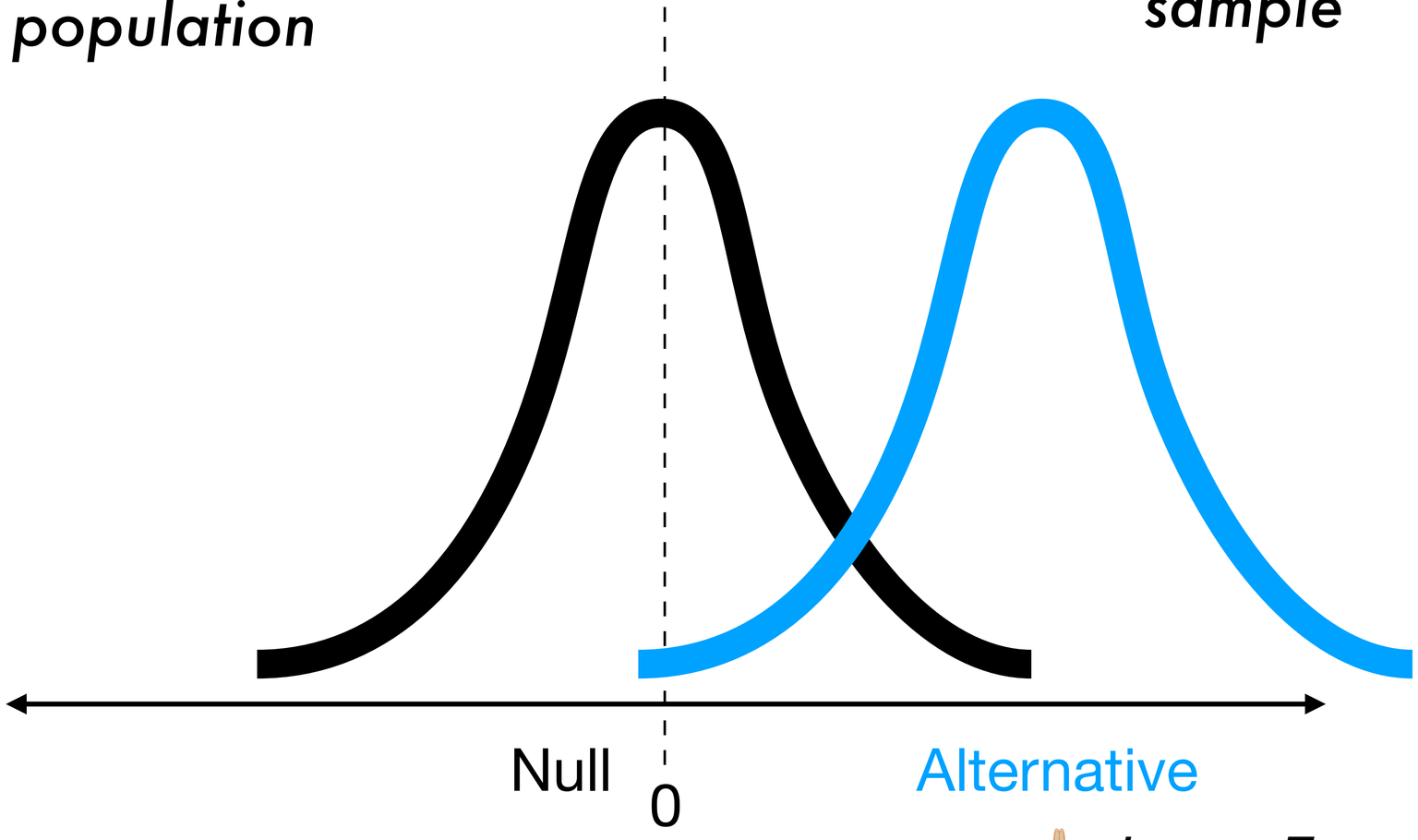
population

Experimental design

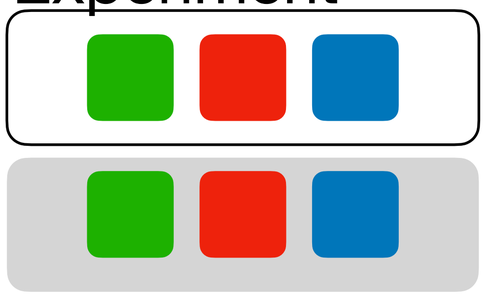
Statistical analysis



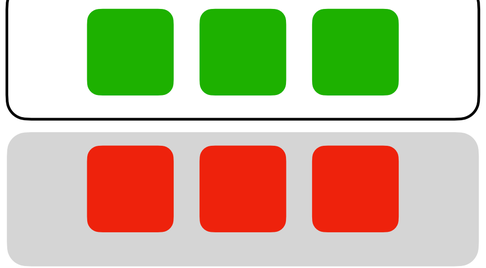
sample



Experiment



Observational



Case study

🦤 @eunicemjun

✉️ emjun@cs.washington.edu

🙏 *James Fogarty, Anant Mittal, Alex Kale, Rene Just*

Topics we didn’t get a chance to cover…

- Working with data frames in Python and R
- More complex methods (Bayesian analysis, causal graph analysis)
- More critical discussion of NHST, consider alternatives