I still don’t have my UMD account… 😂

Canvas should have been fixed ✌️

Other Canvas problems… fixing those 💪

Piazza next week (hopefully) 🤞

Announcements

My office hour: Mon after the class

This/next week: I will be in my office on Wed

Midterm exam: take-home

Sign up for presentations

Announcements

Statistics & Uncertainty

CMSC839E: Uncertainty Communication for Decision-making

https://fumeng-yang.github.io/CMSC839E

August 28, 2024 @ UMD by Fumeng Yang

Motivation

Motivation

Many of my colleagues complain that reviewers complained about data analysis in their papers

In many HCI/VIS (sometimes HCI+ML) papers I reviewed, the statistical practices really … worth complaining about

Background knowledge and philosophy on statistics in HCI

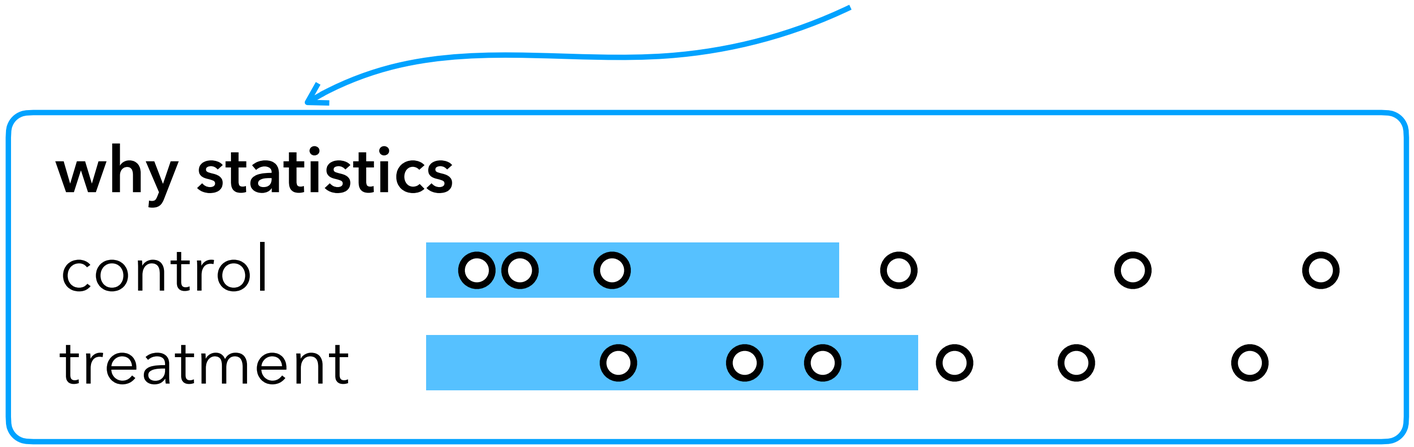
Align your knowledge levels, provide background for next weeks

This lecture

This is **not** a statistical tutorial

I think most statistics are self-learnable, and statistical courses could be misleading…
I took two stats courses in a psychology department (e.g. PSYC200)
You are welcome to discuss with me about your data analysis in the future

I assume you at least know normal distribution, control vs. treatment



Outline

Hypothesis Testing, Significance Testing, & Null Hypothesis Significance Testing

Linear regression, multilevel modeling

Confidence intervals

Bayesian data analysis

Transparent practice & preregistration

I'll also comment on **qualitative** analysis

Outline

NHST

Hypothesis Testing, Significance Testing, & Null Hypothesis Significance Testing

Linear regression, multilevel modeling

the papers will provide further thinkings, and pros and cons

Confidence intervals

Bayesian data analysis

Transparent practice & preregistration ← Assignment Q

I'll also comment on qualitative analysis

> "Don't take this the bad way, I don't think it is humanly possible to understand Bayesian stats in 1h, so this is a big endeavour…" — past course attendee

Questions?

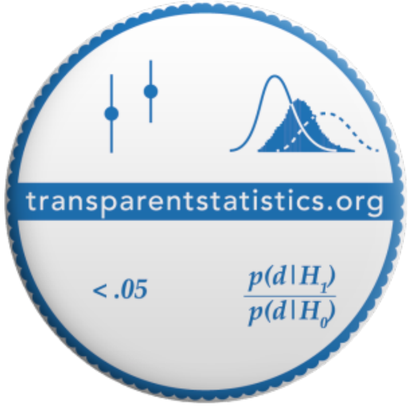
NHST
Hypothesis Testing
Significance Testing

NHST

Ubiquitous, predominate

Ingrained in the minds and current practice of most researchers, journal editors and publishers (social science, psychology, HCI, ML, …)

… probably the most misinterpreted method



Assignment Q: preference over two badges: stats and cat

	Strongly disagree	Disagree	Neither	Agree	Strongly agree
	1	2	3	4	5
I want to have this button.	◯	◯	◯	◯	◯

NHST

1. develop null and alternative hypotheses

H0: people's preference over the two badges is the same.

H1: there is a significant difference in people's preference…

NHST

1. develop null and alternative hypotheses

H0: people's preference over the two badges is the same.

H1: there is a significant difference in people's preference…

2. choose a test based on your experiment and data

I use paired t-test. R code. Input two datasets

```
t.test(df_cat$response, df_stats$response, paired=TRUE)
```

NHST

1. develop null and alternative hypotheses

H0: people's preference over the two badges is the same.

H1: there is a significant difference in people's preference…

2. choose a test based on your experiment and data

I use paired t-test.

```
t.test(df_cat$response, df_stats$response, paired=TRUE)
```

3. get results and write the reports

```
t = 4.5659, df = 16, p-value = 0.0003173
```

P < .05! YAY! Paper accepted!→

"We used a paired t-test, and found there is a significant difference in people's preference over the two designs (p < .05)."

What was "wrong"

1. develop null and alternative hypotheses

3 questions in total

	Strongly disagree	Disagree	Neither	Agree	Strongly agree
	1	2	3	4	5
I want to have this button.	○	○	○	○	○
I want to wear this button during the conference.				○	○
Having this button will make me want to talk about my experience in this tutorial.	○	○	○	○	○

I tried this first, n.s.

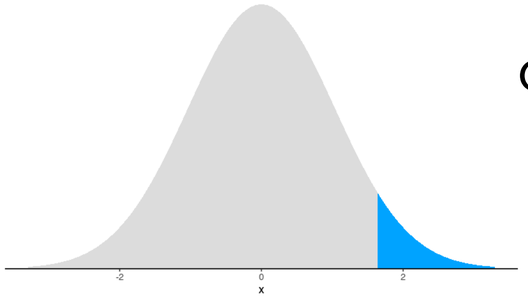
What was "wrong"

1. develop null and alternative hypotheses

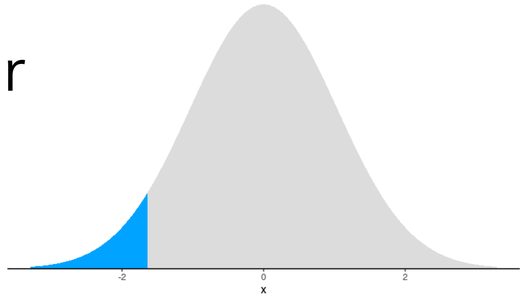
2. choose a test based on your experiment and data

One tailed vs. two tailed (R default)

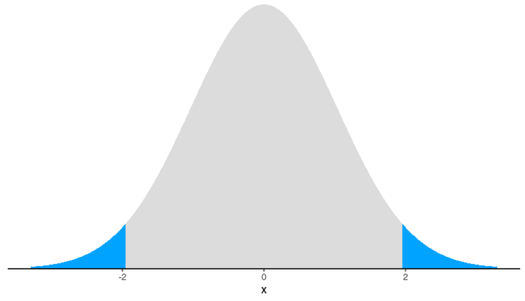
One-tailed: greater or less



or



Two-tailed: whether there is a difference



What was "wrong"

1. develop null and alternative hypotheses

2. choose a test based on your experiment and data

I totally ignored t-test's assumptions.

https://resources.nu.edu/statsresources/TestingAssumptions

- continuous scale → Hmm, Likert scale (in your assignment reading)
- random sample → Assume it's okay
- a normal distribution → Actually okay in this case

Violation of assumptions: your cherished p values might be problematic.

What was “wrong”

1. develop null and alternative hypotheses

2. choose a test based on your experiment and data

I totally ignored t-test’s assumptions.

Most people might not get very deep. But this is definitely a target.

Sometimes you can probably find some references to defense yourself…

What was "wrong"

1. develop null and alternative hypotheses

2. choose a test based on your experiment and data

3. get results and write the reports

"We use a paired t-test, and find there is a significant difference in people's preference over the two designs ($p < .05$).

in a second

wording, transparency, dichotomous thinking

“Some solutions” I

1. develop null and alternative hypotheses

   Planning & preregistration

2. choose a test based on your experiment and data

   Figure 1.1 in “Statistical Rethinking” and Google

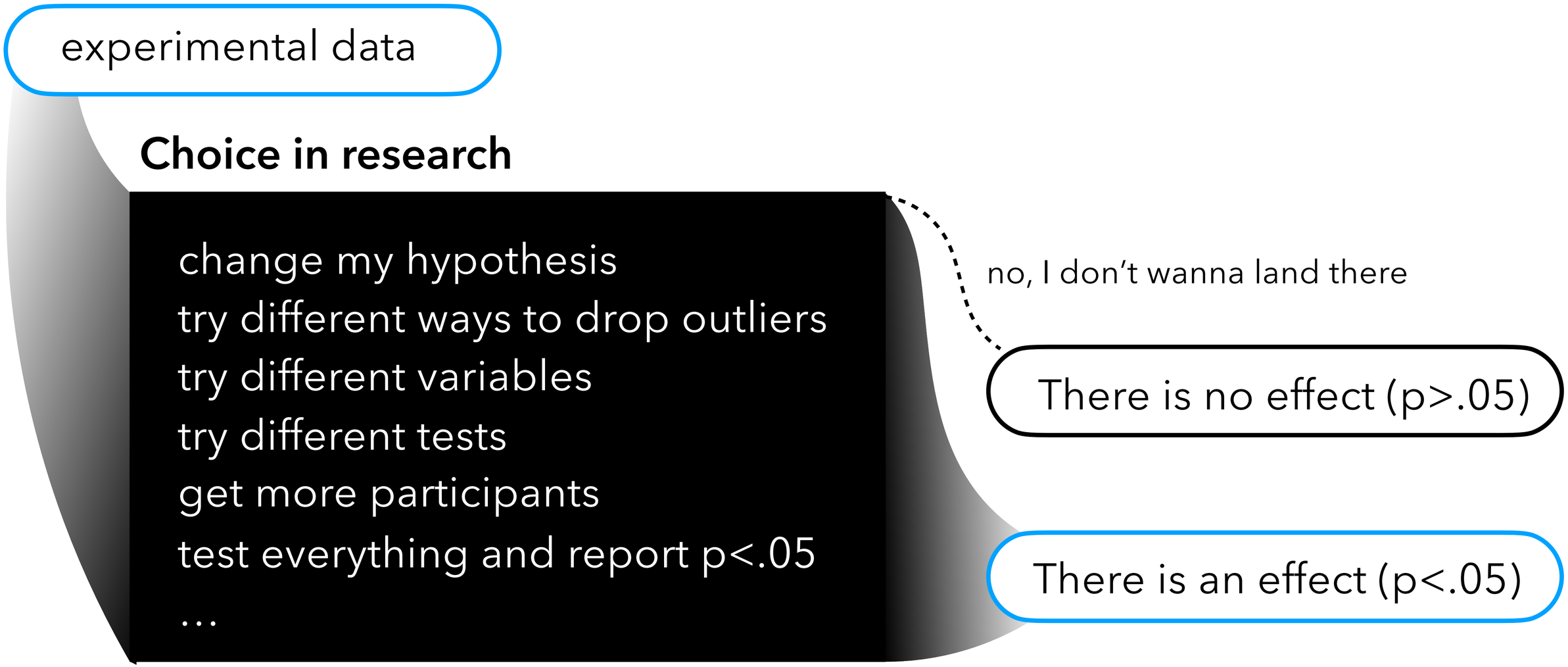
3. get results and write the reports

   How & what to report: APA style

   https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf

… probably can pass most of reviewers if you do these

The dominant mental model



Genuine questions

What does the data look like?

How big is the difference?

If high-stake, can I deploy the treatment?

…

Questions?

Traditional statistics

" .. It is based upon **a fundamental misunderstanding** of the nature of rational inference, and is seldom if ever **appropriate** to the aims of scientific research."

Rozeboom quoted by Dragicevic

William W. Rozeboom quoted in Timothy R. Levine, René Weber, Craig Hullett, Hee Sun Park and Lisa L. Massi Lindsey. 2008. A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. Human Communication Research, 34(2), 171-187.

How did it come to this point?

Significance Testing – by Fisher

Hypothesis Testing – by Neyman and Pearson

NHST is a Frankenstein of these two incomparable approaches, taking whatever is easier from both

recommended

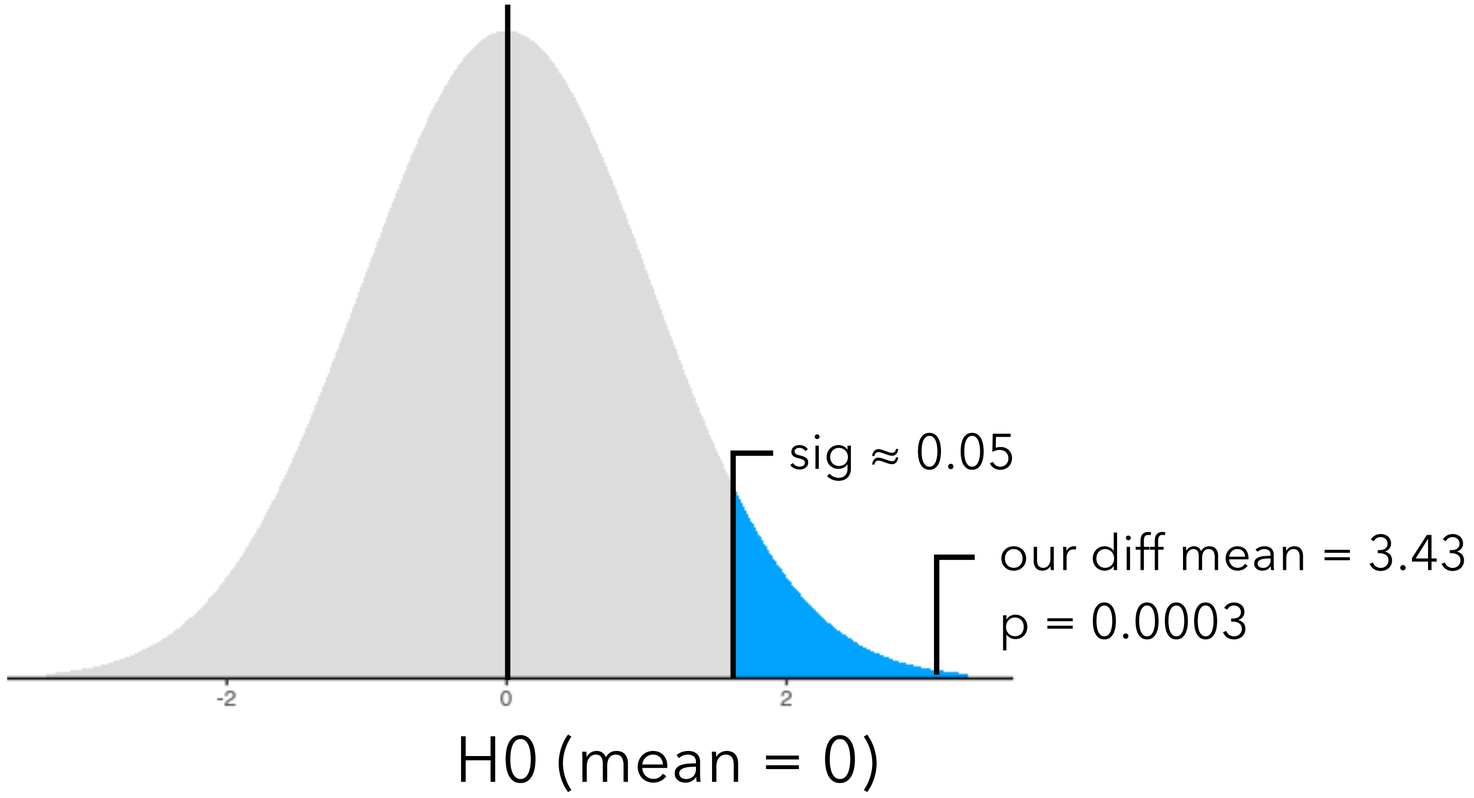
↓

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in psychology*, *6*, 223.

Significance Testing (Fisher)

How rare (significant) is our data? – e.g., normal distribution

Difference in sample means



Significance Testing (Fisher)

p-value = **evidence** against the null hypothesis

smaller = stronger evidence

Do **not** need to be rigid (e.g., 0.049 and 0.051 have about the **same** statistical significance around a convenient level of significance of 5%)

Significance Testing (Fisher)

1. Select an appropriate test
2. Set up the null hypothesis (H_0)
3. Calculate the theoretical probability of the results under H_0 (p)
4. **Assess the statistical significance of the results**
5. **Interpret the statistical significance of the results**

Too simple? Too complex?

Hypothesis Testing (Neyman-Pearson)

We set up expectations for our data, **then** collect data, and decide which hypothesis is true (a cutoff)

1. Expected effect size
2. α - Type I error (prob of H_{main} is true but rejected)
3. β - Type II error (prob of $H_{alternative}$ is true but rejected; 1 - β = power)

Hypothesis Testing (Neyman-Pearson)

1. Set up the expected effect size in the population (a priori).

Roughly how large you want the difference is. e.g., d = 0.7

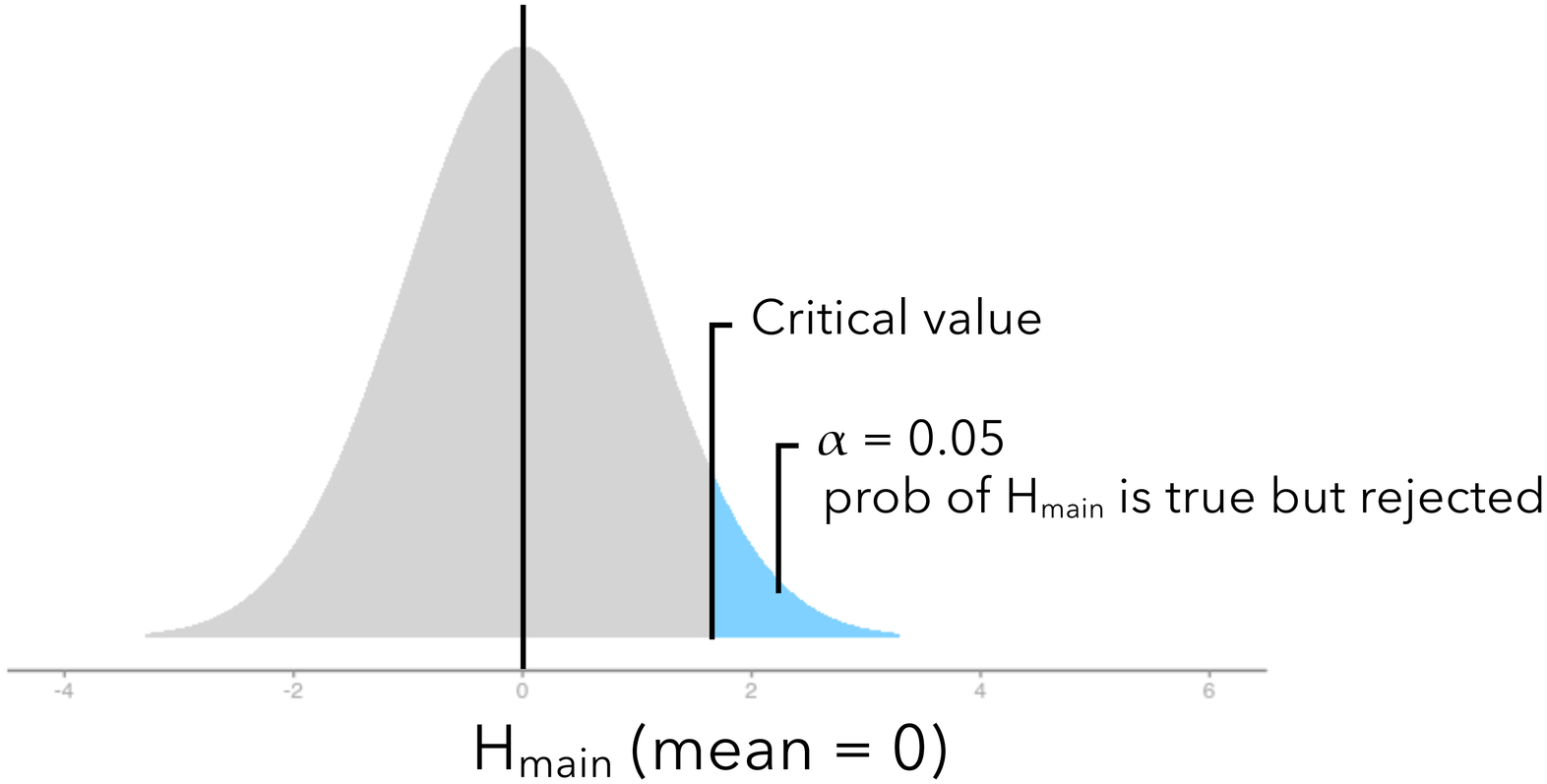
Hypothesis Testing (Neyman-Pearson)

2. Select an optimal test (a priori).

e.g., t-test

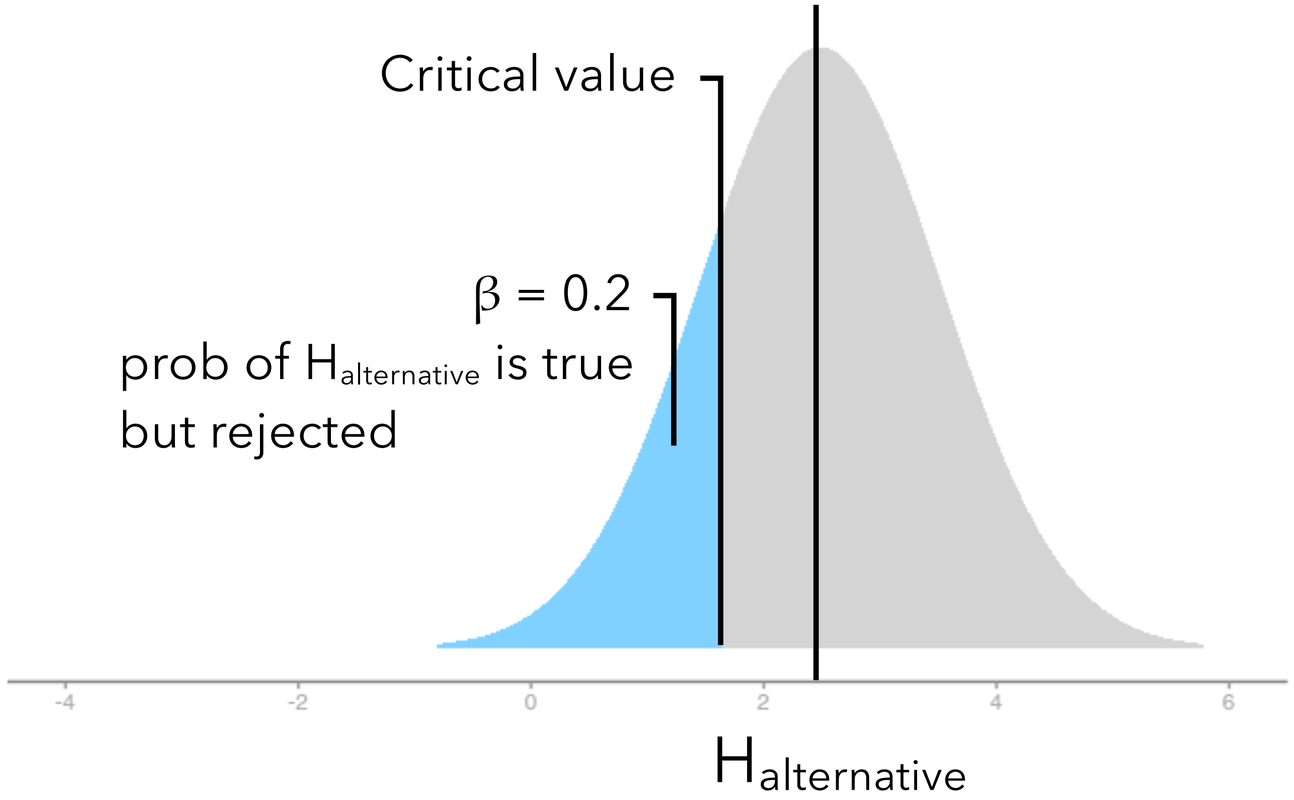
Hypothesis Testing (Neyman-Pearson)

3. Set up the main hypothesis (H_{main}) (a priori)



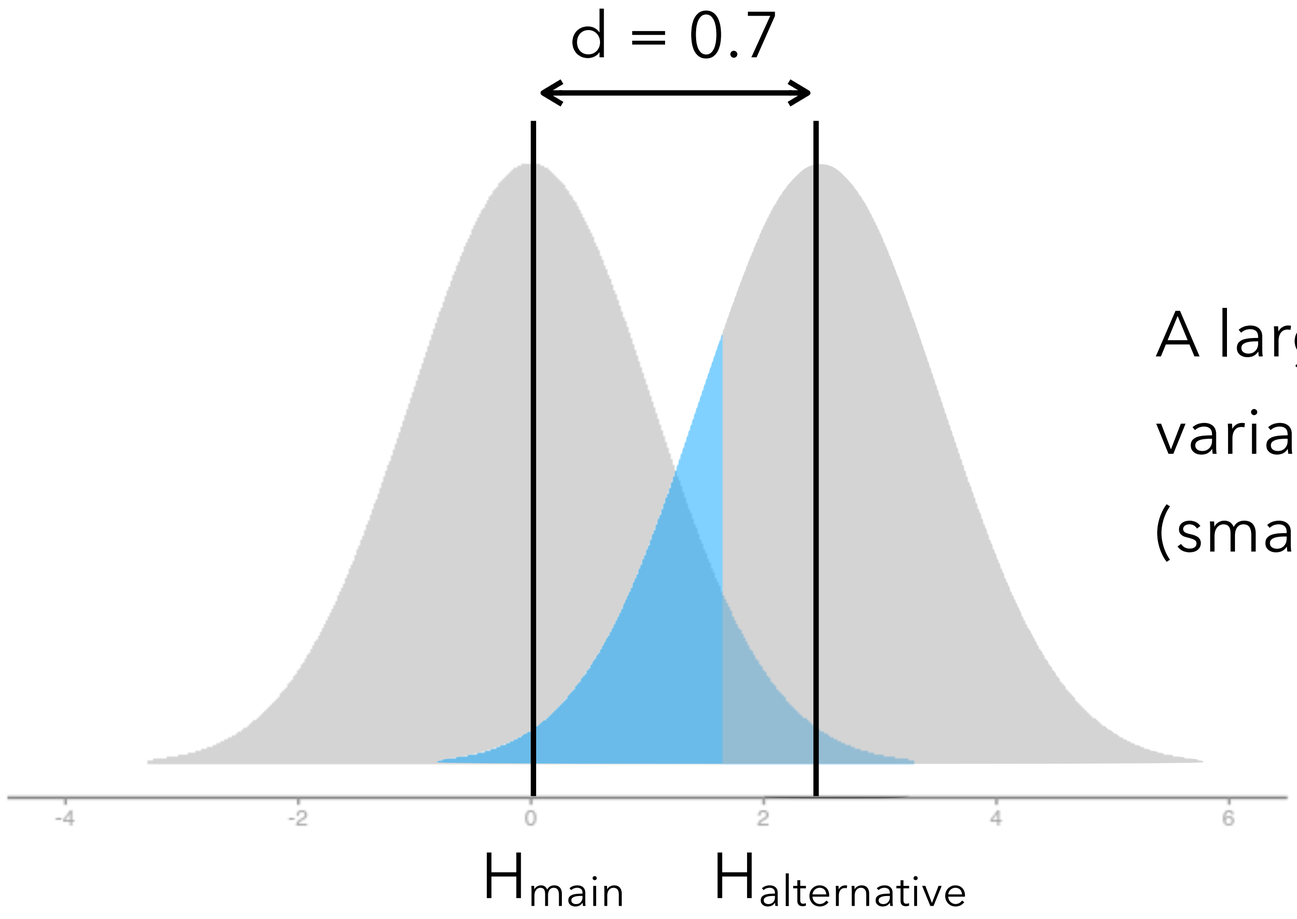
Hypothesis Testing (Neyman-Pearson)

3. Set up the alternative hypothesis ($H_{alternative}$) (a priori)



Hypothesis Testing (Neyman-Pearson)

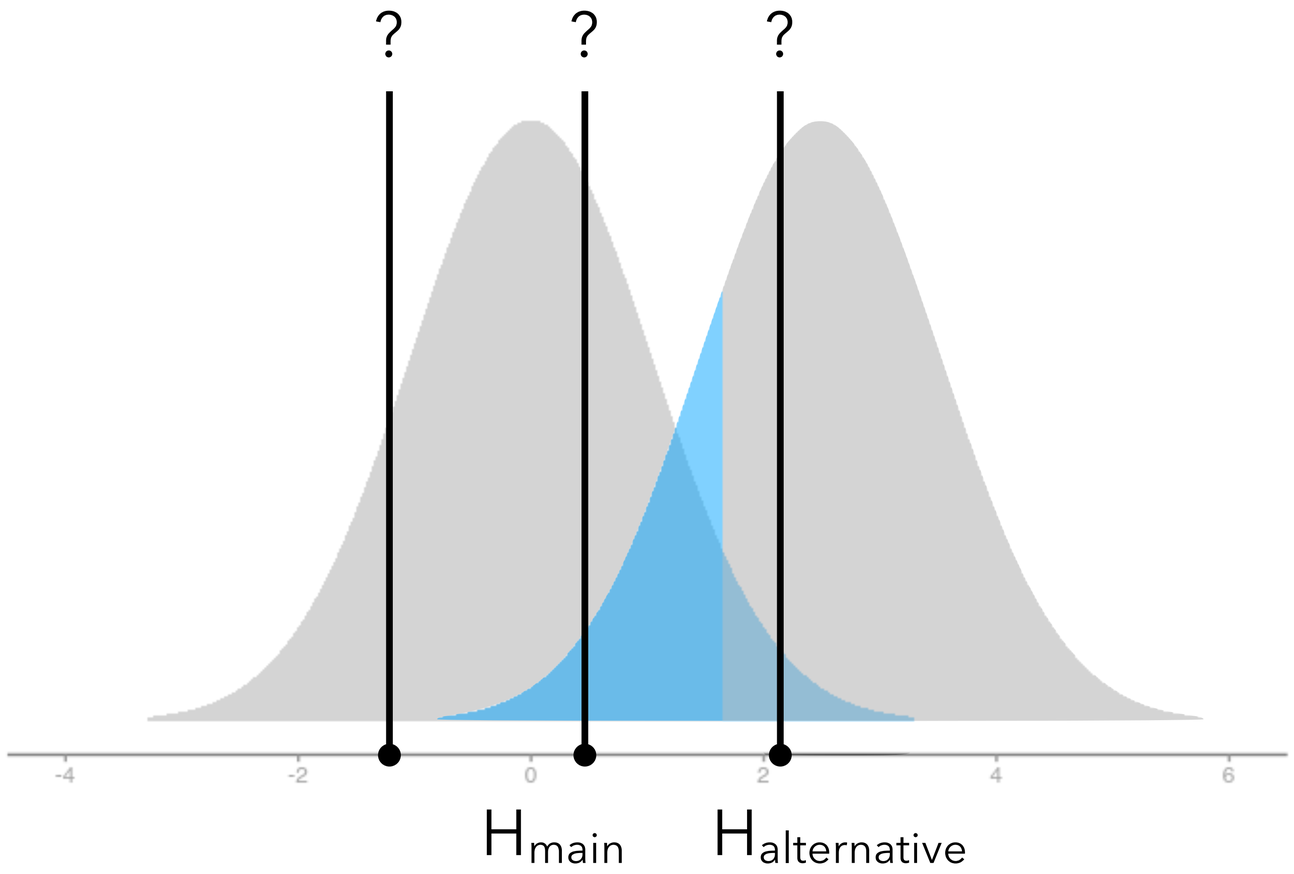
4. Calculate the sample size (N) - formula omitted (a priori)



A larger N usually means less variance and a larger power (smaller β).

Hypothesis Testing (Neyman-Pearson)

5. Collect data, then calculate variance and test value (a posteriori)

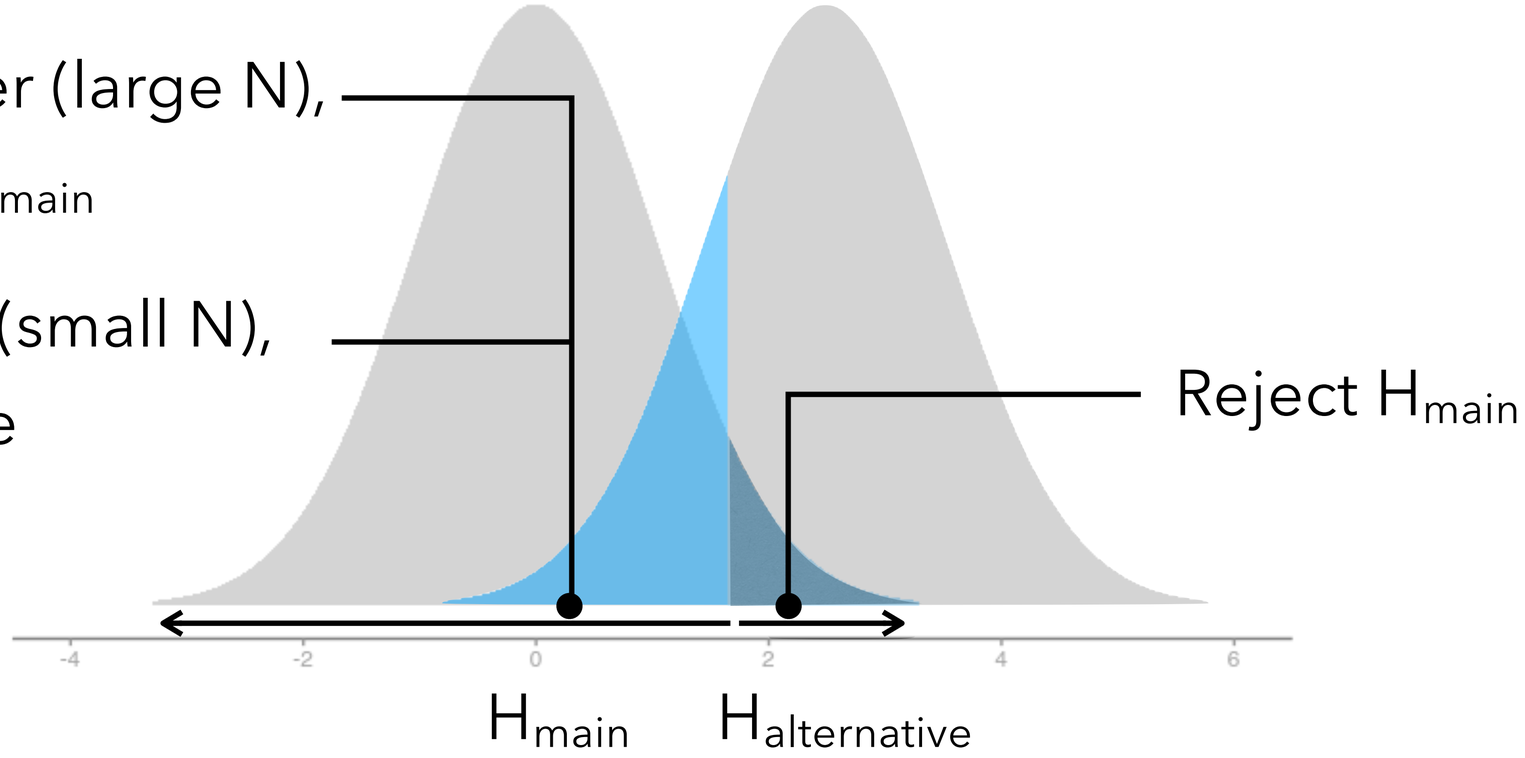


Hypothesis Testing (Neyman-Pearson)

6. Make the decision

Good power (large N), accepted H_{main}

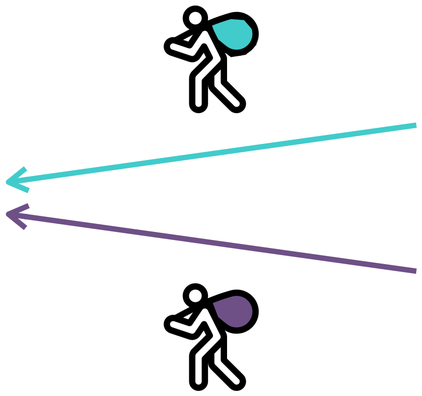
Low power (small N), inconclusive



Too simple? Too complex?

Hmm...

NHST



Significance Testing (Fisher)

The calculation part is much easier!

Hypothesis Testing (Neyman-Pearson)

The cutoff part is much easier!

recommended

↓

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in psychology*, *6*, 223.

P is highly unreliable

The dance of p-values I

```r
set.seed(12345) # for reproducibility
samples = rnorm(15000, mean = 0, sd = 1) # get a large sample
n = 50 # simulate different experiments
pvalues = c() # a container
ns = c() # a container
```

The ground truth, p should be 1

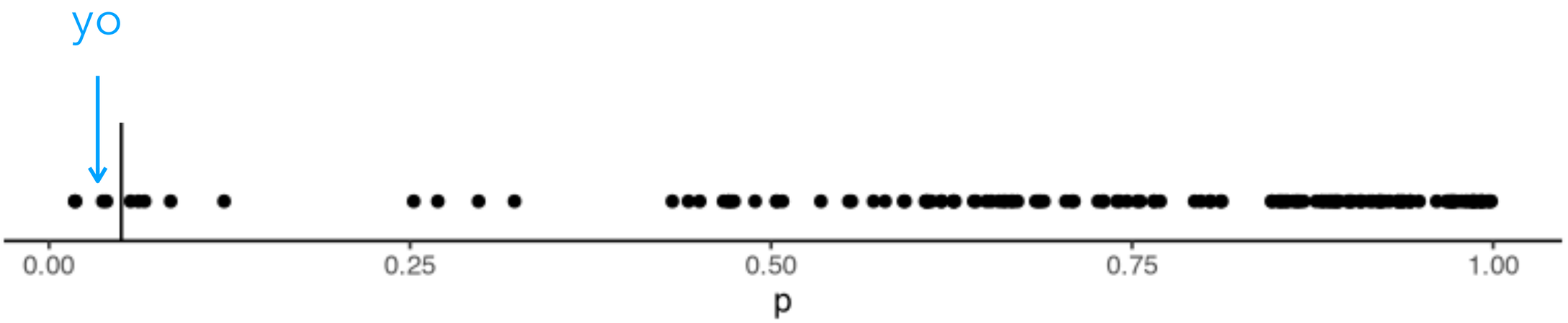
The dance of p-values I

```
repeat{
  test = t.test(samples[1:n]) # t-test
  pvalues = c(pvalues, test$p.value)  # record
  ns = c(ns, n)
  if(n > length(samples))
    break
  n = n+100  # if I see first 150, 250, 350, … samples
}
```

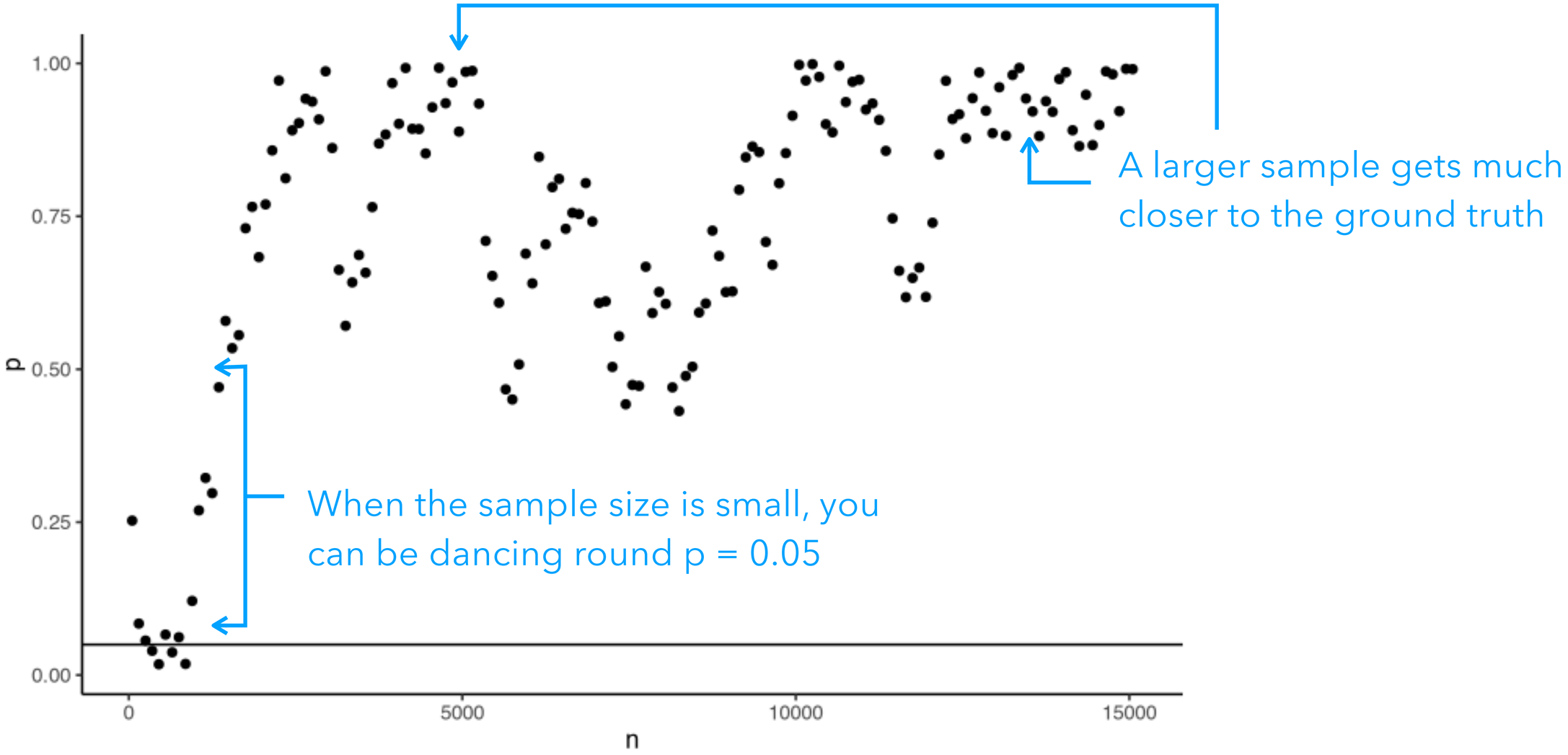
It's all about uncertainty I

output

0.25252205 0.08419887 0.05645780 0.03976305 0.01777029 0.06619174 ...



It's all about uncertainty I



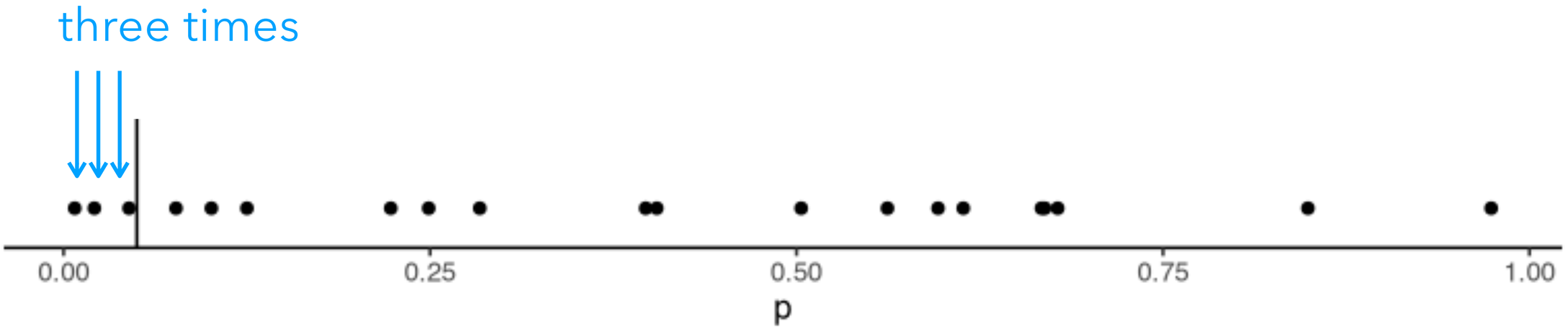
The dance of p-values II

```
n = 50 # sample size
for(i in 1:20){ # repeat the same experiment 20 times
  set.seed(i) # for reproducibility
  samples = rnorm(n, mean = 0, sd = 1)
  test = t.test(samples)
  pvalues = c(pvalues, test$p.value)
}
```

It's all about uncertainty II

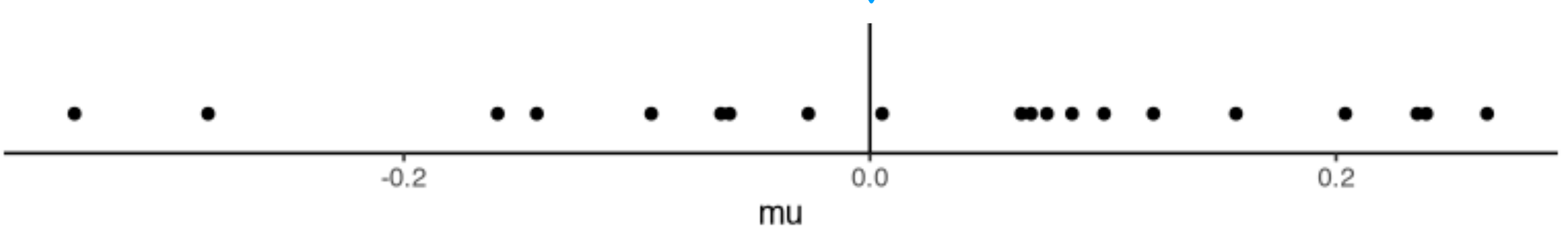
output

```
0.397085241 0.667089582 0.613750277 0.076697206 0.669336386 0.561997280 0.100805027
0.678159564 0.503181225 0.007579664 0.020994575 0.249053832 0.848857704 0.125083913
0.044664855 0.223177021 0.596531930 0.973907807 0.404662921 0.283896859
```



It's all about uncertainty II

The mean value is all over the place while the true mean is 0



Questions?

The results (?) of NHST-driven analysis

P-values hacking

e.g., remove outliers or add more participants until significant

HARKing (Hypothesizing After Results are Known)

Try everything; presenting exploratory findings as confirmatory

…

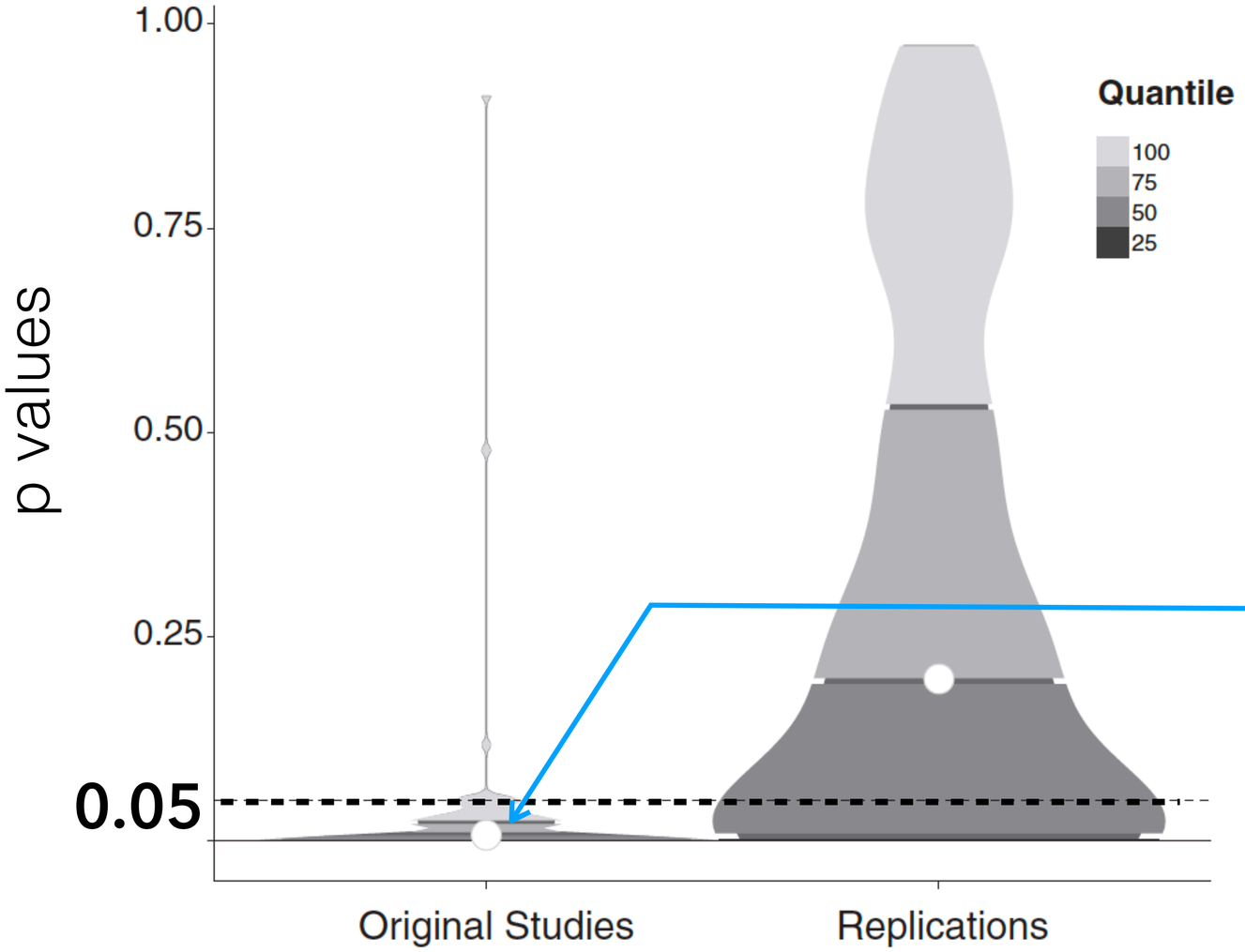
→ **Replication crisis**

Replication crisis

“We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available”

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Replication crisis

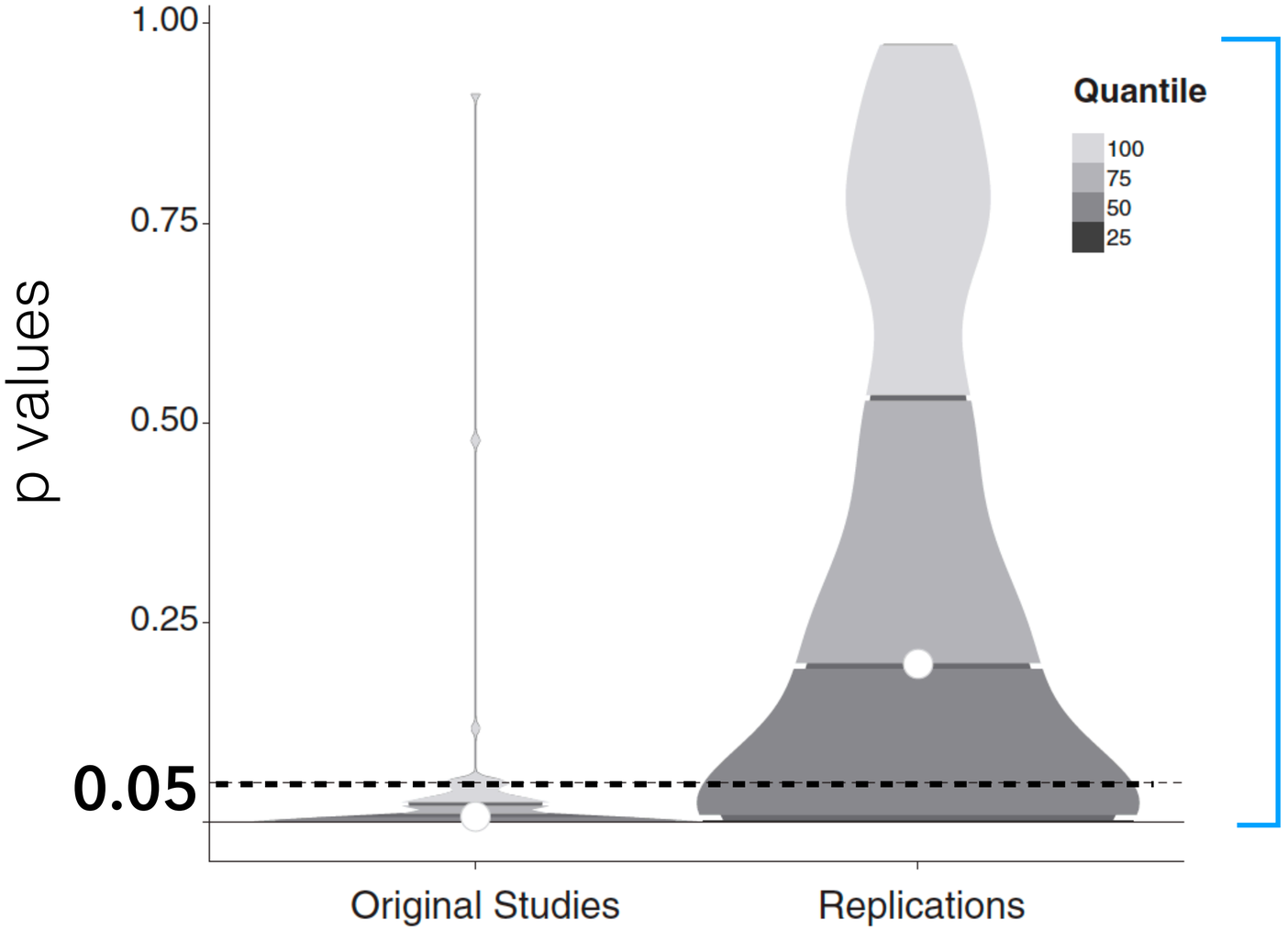


Original studies

Crowd around p<0.05

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

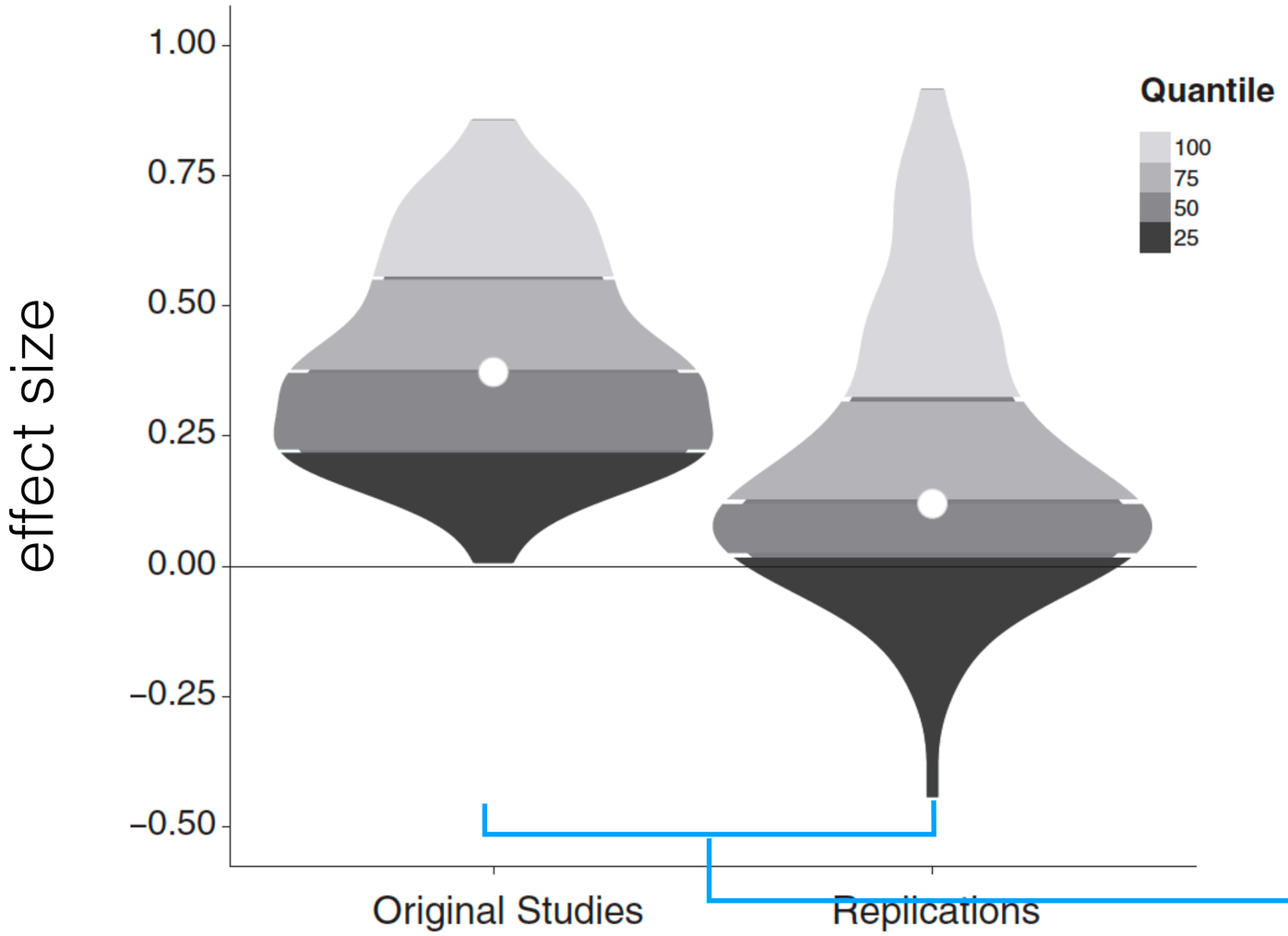
Replication crisis



Replications are all over the place

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Replication crisis



Better (differences are smaller) but not too much better

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Machine learning?

Same or maybe worse

Random initialization

Stochastic gradient descent

Cross-fold validation

Different implementations

Underspecified - too many parameters

…

“Some solutions” II

1. Confidence intervals & Bayesian data analysis

   Details in next two weeks

2. Open practice & preregistration

   Now & Assignment Q.1

3. Regression, a method (c.f. regression in ML)

   If time; get rid of some assumptions

Questions?

Open practice & preregistration

Open practice / Transparency

Sharing as much as possible about planning, analysis, materials, data, etc. through supplementary materials

Make a research easier to evaluate and replicate

Help people (e.g., reviewers, peers) evaluate the research quality

shift the focus towards the quality, not just the results (helpful for null results)

not a guarantee of quality, but it is necessary to evaluate quality.

Preregistration - planning

Specify all hypotheses & methodological choices in writing prior to data collection:

Reduces researcher degrees of freedom

Can’t p-hack

Can’t HARK (hypothesize after results are known)

Preregistration (AsPredicted template)

- What's the **main question** being asked or **hypothesis** being tested in this study?
- Describe the key **dependent variable**(s) specifying how they will be measured.
- Specify exactly **which analyses** you will conduct to examine the main question/hypothesis.
- Describe exactly **how outliers will be defined and handled**, and your precise rule(s) for excluding observations.

And more! (see https://aspredicted.org)

And visualize your data!

Participants' estimates for the two versions of the (two) scatterplots (A and B) slightly differ from ground truth, but generally, they were quite accurate. They reported an average correlation of 0.315 (SD = 0.19, CI = [0.294, 0.337]) and 0.301 (SD = 0.21, CI = [0.277, 0.325]) for the two versions of scatterplots A. For scatterplots B, participants reported an average correlation of 0.658 (SD = 0.19, CI = [0.737, 0.779]) and 0.625 (SD = 0.18, CI = [0.704, 0.746]).

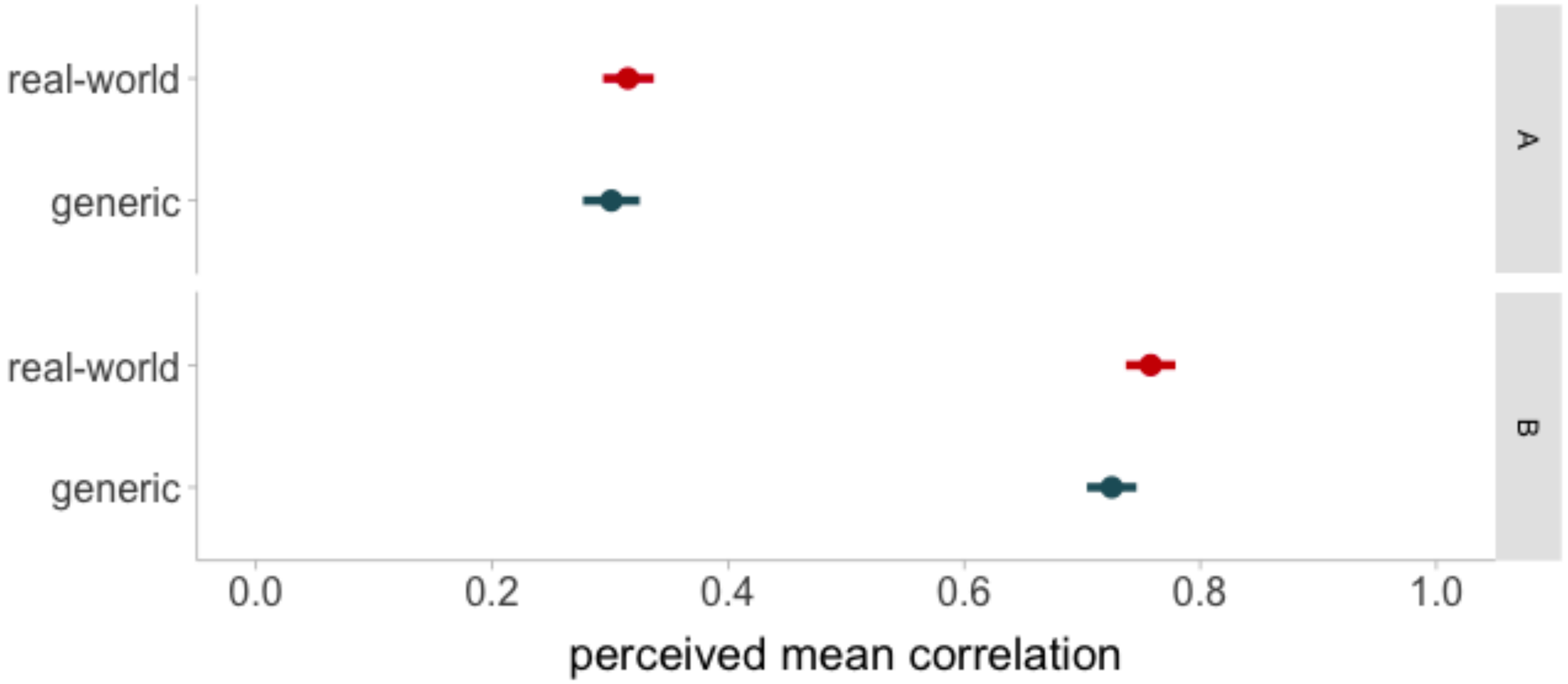
And visualize your data!

Using a table is better

Conditions	A	B
real-world names	0.315 [0.294, 0.337]	0.758 [0.737, 0.779]
generic names	0.301 [0.277, 0.325]	0.725 [0.704, 0.746]

And visualize your data!

Much better!



Assignment Q

Q.1 Preregistration

Fill in the AsPredicted template

Q.2 Analysis, report, & supplementary materials

SM: like the html I generated (maybe better document)

Commentary

“Some solutions” III

System/prototype evaluation

Hard to get more participants

Hard to design a control condition

Hard to design a quantitative experiment (speed & accuracy may not us me too much)

“Some solutions” III

System/prototype evaluation

Personally I believe an in-depth **qualitative** study is
the best solution … though some reviewers may disagree

The “flawless” example I’ve seen is to do **a lite** quant + **a deep** qual

Another discussion might be if you need to let your
users use them for a longer time (1 month, 1 year, …)

Questions?

Next week

No class on Monday

Read papers!!

Sign up for presentations

Submit questions by the noon

To be continued…

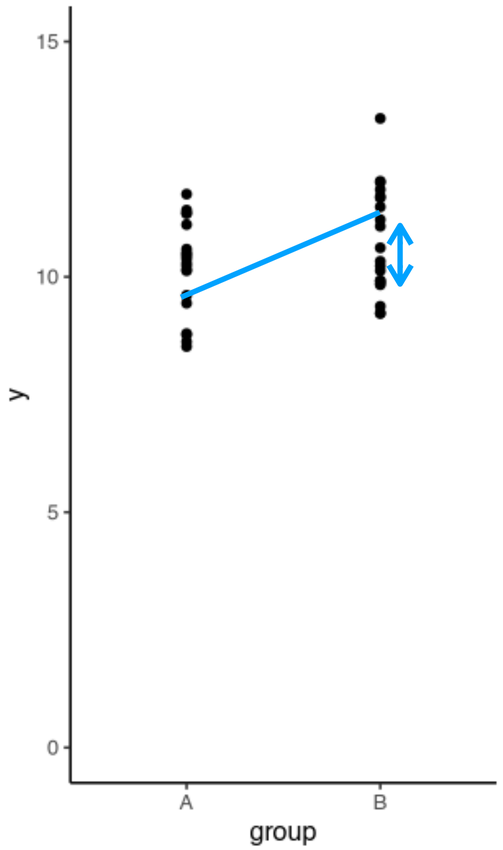
A bite of regression (as a method)

T-test vs. its regression version

t-test subtracts two groups, get the difference in mean, and compare

A generalized linear regression $y = \beta_1 x + \beta_0$

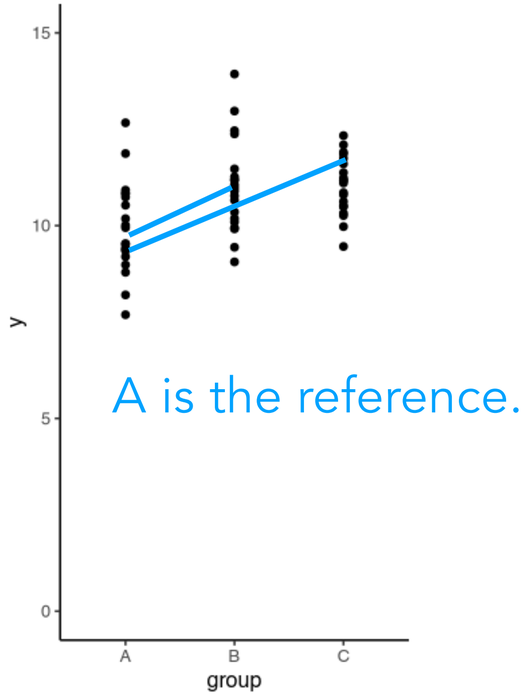
The slope is basically what's being evaluated in a t-test



One-way ANOVA vs. its regression version

One-way ANOVA = more than 2 levels (e.g., A, B, C)

Roughly, compare variance between groups to variance within groups. If this different is large enough, one group must be different from others = sig.



A is the reference.

A generalized linear regression

$$y = \beta_1 x + \beta_0$$

The calculation of global mean (intercept) captures "variance between groups"
The calculation of slopes (you have two slopes for three groups) captures "variance within groups"
And the slopes refer to the global mean.
So, they are very similar processes.

Pros and cons

Mixed Modeling has fewer assumptions:

Robust to homogeneity of variance

More flexible for complex experiments

Don’t need equal sample size for each cell

So … sometimes they have different results

Readings

Random and mixed effects models

https://people.math.ethz.ch/~meier/teaching/anova/random-and-mixed-effects-models.html

Understanding how ANOVA relates to regression

https://statmodeling.stat.columbia.edu/2019/03/28/understanding-how-anova-relates-to-regression

Two-way ANOVA in R

https://statsandr.com/blog/two-way-anova-in-r/