Multiple Hypothesis Testing and Error Rate Correction

Data 102 Spring 2022

Lecture 3

Weekly Overview

- Last week: course intro, binary decision-making

- **Today: multiple hypothesis testing**
  - What can go wrong when making multiple hypothesis tests?
  - Error rates for multiple testing
  - Bonferroni correction, Benjamini-Hochberg procedure, and the LORD algorithm

- Next time: making multiple decisions while controlling error rates

Recap: Row-wise rates

- TNR: $\frac{n_{00}}{n_{00} + n_{01}}$

  specificity

- FPR: $\frac{n_{01}}{n_{00} + n_{01}}$

- TPR: $\frac{n_{11}}{n_{10} + n_{11}}$

  sensitivity
  recall

- FNR: $\frac{n_{10}}{n_{10} + n_{11}}$

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Wikipedia: Sensitivity and Specificity

Recap: Column-wise rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

False discovery proportion (FDP):

$$\frac{n_{01}}{n_{01} + n_{11}}$$

Recap: rates as conditional probabilities

- Row-wise rates condition on reality
  - TNR: P(D = 0 | R = 0)
  - FPR: P(D = 1 | R = 0)
- Column-wise rates condition on the decision
  - FDP: P(R = 0 | D = 1)
  - Conditional probability is only valid if we treat reality as random

Recap: connecting column-wise and row-wise rates

$$FDP = \frac{1}{1 + \frac{TPR}{FPR}\frac{1-\pi_0}{\pi_0}}$$

Higher prevalence ($\pi_0 \approx 0$): FDP ↓

Higher TPR: FDP ↓

Higher FPR: FDP ↑

Recap: definition of *P*-value (from Data 8)

The *P*-value is the chance,

- if the null hypothesis is true,
- that the test statistic
- is equal to the value observed in the data
- or is even further in the direction of the alternative.
    - *We'll always use "larger" in this class (WLOG)*

Row-wise and column-wise rates

TNR: $$\frac{n_{00}}{n_{00} + n_{01}}$$

FPR: $$\frac{n_{01}}{n_{00} + n_{01}}$$

TPR: $$\frac{n_{11}}{n_{10} + n_{11}}$$

FNR: $$\frac{n_{10}}{n_{10} + n_{11}}$$

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Row-wise rates:

- True/false positive rate
- True/false negative rate
- *Sensitivity, specificity, recall, etc.*

Column-wise rates:

- True/false discovery proportion/rate
- True/false omission proportion/rate
- *Precision, positive predictive value, etc.*

False omission proportion (FOP):

$$\frac{n_{10}}{n_{10} + n_{00}}$$

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

False discovery proportion (FDP):

$$\frac{n_{01}}{n_{01} + n_{11}}$$

How do we make decisions from P-values?

- Classical NHST: choose p-value threshold, see whether p-value is above/below
  - Threshold gives us a tradeoff between TPR and TNR
  - Alternative hypotheses are “vague”: hard to reason precisely about TPR
- Neyman-Pearson: choose a specific alternative, reason about TPR
- Binary classification: without p-values
  - More on this on Thursday
- Vocabulary: we’ll use these interchangeably
  - “Reject the null”
  - “Make a discovery”
  - “Decision = 1”

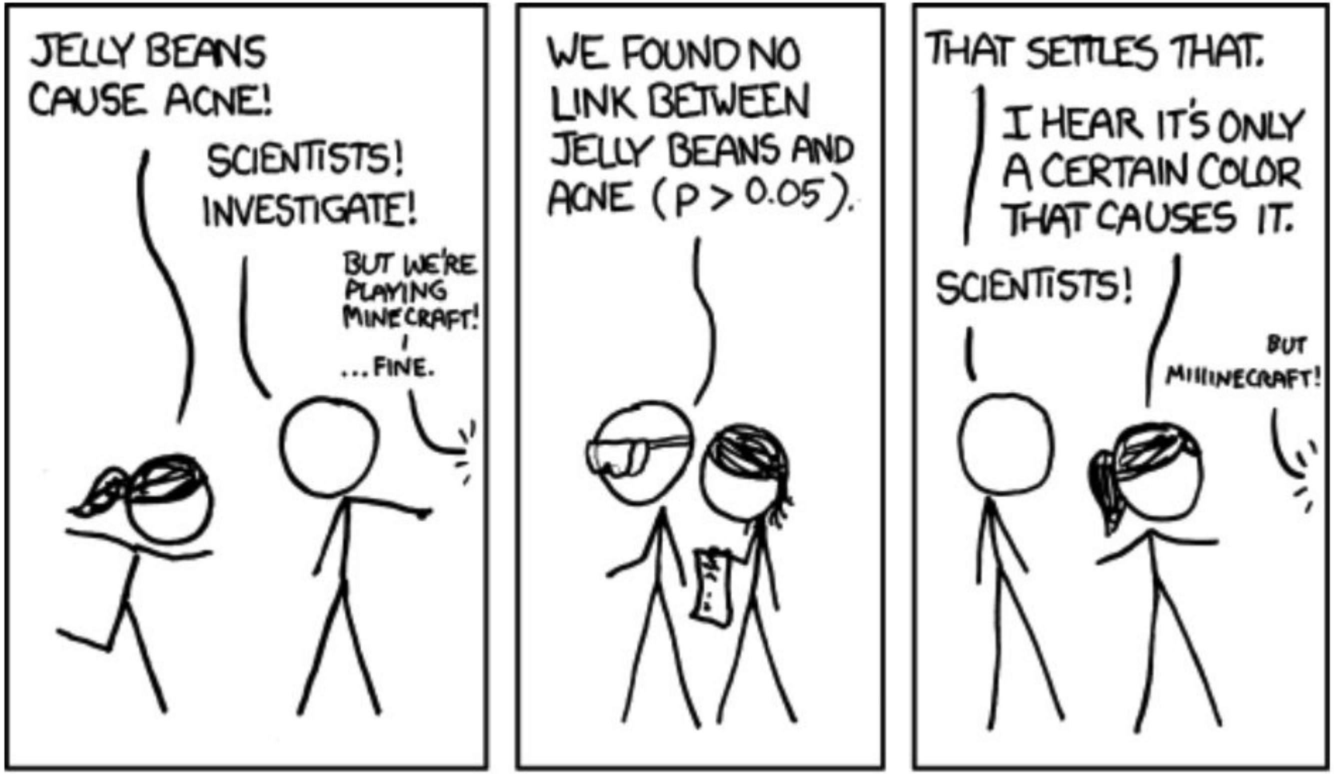
P-values: a closer look

- If we can reason about the distribution of p-values, we can use that to help us build a theory of how to control error rates involving p-values
- Process:
  - Collect data (random)
  - Compute test statistic from data (also random)
  - Compute p-value from test statistic (also random)
- The p-value is a random variable
- Under the null hypothesis, it has a **uniform distribution**
  - You'll explore this more in lab

How do we make decisions from P-values?

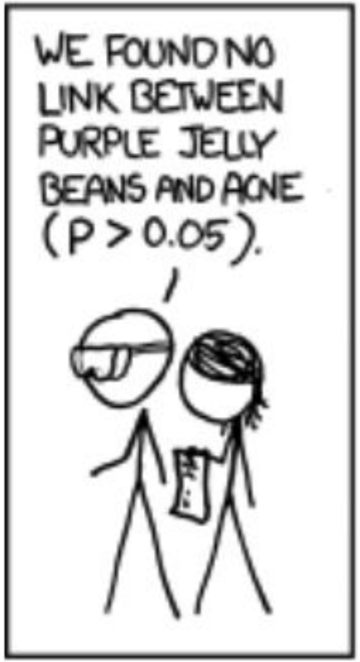
- Classical NHST: choose p-value threshold, see whether p-value is above/below
  - Threshold gives us a tradeoff between TPR and TNR
  - Alternative hypotheses are “vague”: hard to reason precisely about TPR
- Neyman-Pearson: choose a specific alternative, reason about TPR
- Binary classification: without p-values
  - More on this next time
- Vocabulary: we’ll use these interchangeably
  - “Reject the null”
  - “Make a discovery”
  - “Decision = 1”

Multiple Hypothesis Testing


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

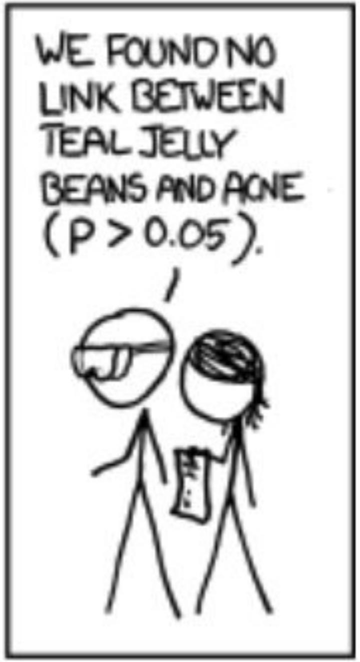
xkcd.com/882

Multiple Hypothesis Testing


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).

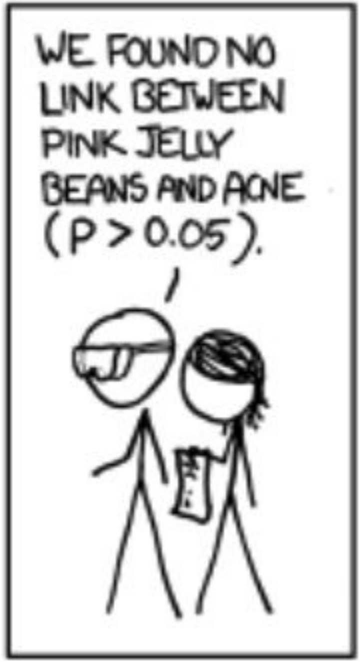
xkcd.com/882

Multiple Hypothesis Testing


WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

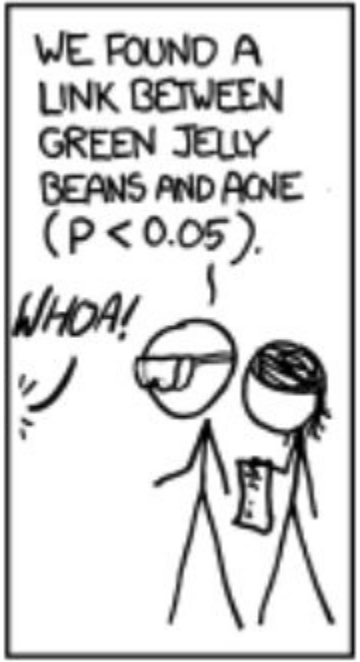
xkcd.com/882

Multiple Hypothesis Testing


WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).

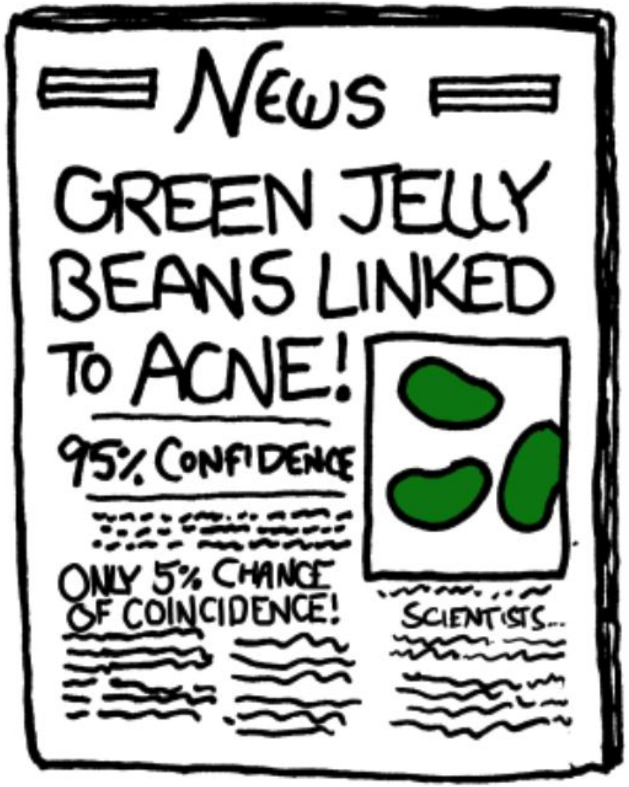
xkcd.com/882

Multiple Hypothesis Testing


WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!

xkcd.com/882

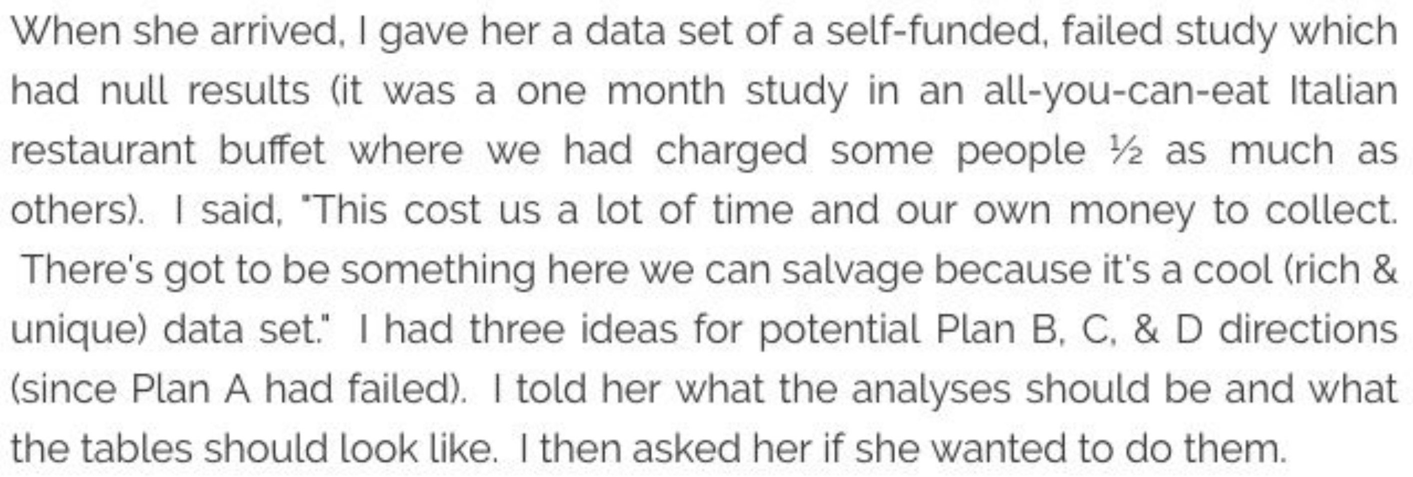
Multiple Hypothesis Testing


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com/882

This happens in the real world!

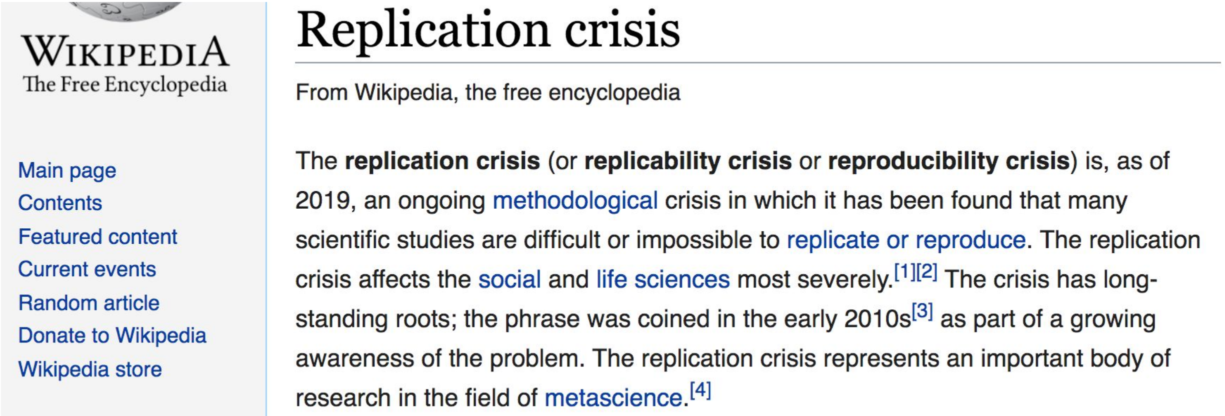
Ex-professor B. Wansink bragging about asking a student to commit statistical malpractice:



When she arrived, I gave her a data set of a self-funded, failed study which had null results (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people ½ as much as others). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

[Buzzfeed News], [Wansink's own blog]

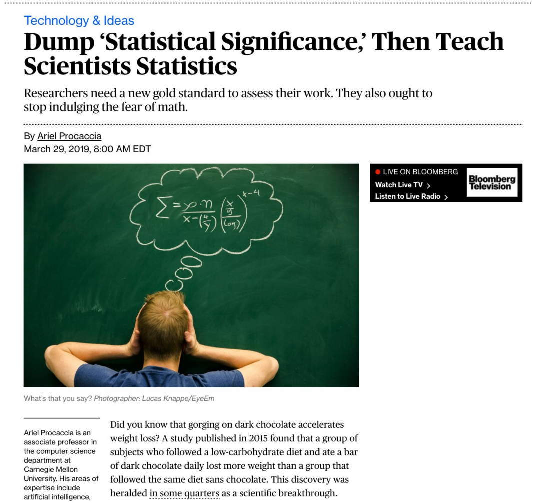
Replication Crisis In Science



Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis** or **reproducibility crisis**) is, as of 2019, an ongoing methodological crisis in which it has been found that many scientific studies are difficult or impossible to replicate or reproduce. The replication crisis affects the social and life sciences most severely.[1][2] The crisis has long-standing roots; the phrase was coined in the early 2010s[3] as part of a growing awareness of the problem. The replication crisis represents an important body of research in the field of metascience.[4]



Technology & Ideas

Dump 'Statistical Significance,' Then Teach Scientists Statistics

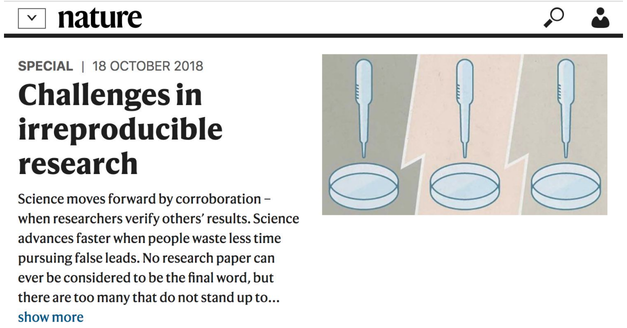
Researchers need a new gold standard to assess their work. They also ought to stop indulging the fear of math.

By Ariel Procaccia
March 29, 2019, 8:00 AM EDT

What's that you say? *Photographer: Lucas Knappe/EyeEm*

Ariel Procaccia is an associate professor in the computer science department at Carnegie Mellon University. His areas of expertise include artificial intelligence,

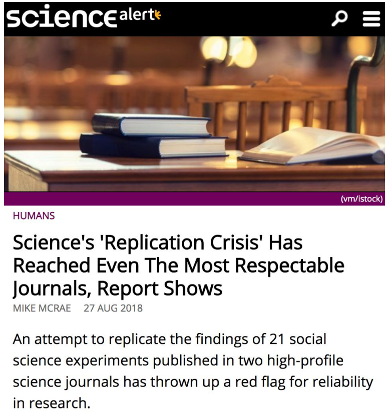
Did you know that gorging on dark chocolate accelerates weight loss? A study published in 2015 found that a group of subjects who followed a low-carbohydrate diet and ate a bar of dark chocolate daily lost more weight than a group that followed the same diet sans chocolate. This discovery was heralded in some quarters as a scientific breakthrough.



SPECIAL | 18 OCTOBER 2018

Challenges in irreproducible research

Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to...
show more



(vm/istock)

HUMANS

Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows

MIKE MCRAE 27 AUG 2018

An attempt to replicate the findings of 21 social science experiments published in two high-profile science journals has thrown up a red flag for reliability in research.

Multiple Hypothesis Testing

- Multiple tests/decisions can arise in several ways:
  - Same dataset, multiple questions
  - Multiple datasets, same question
  - Multiple datasets, multiple questions

- Sometimes, we need to do multiple tests: how can we avoid making errors?
  - Solution: define an error rate we’re interested in, and adjust decisions to control it
  - Example: GWAS studies
  - Example: fMRI studies

Multiple Hypothesis Testing: Avoiding the Jellybean Problem

- For one test, we can try to limit the FPR of the test
- For multiple tests, we need an error rate that’s related to **all** the tests
  - Even if FPR = 0.01, we’re very likely to see some FPs if we do 100s of tests
- Suppose we do m tests
- Two rates that control this
  - FWER: P(any of the m tests is a false positive)
  - FDR: like FDP, but averaged over the randomness in the sequence of p-values

Controlling FWER: Bonferroni correction

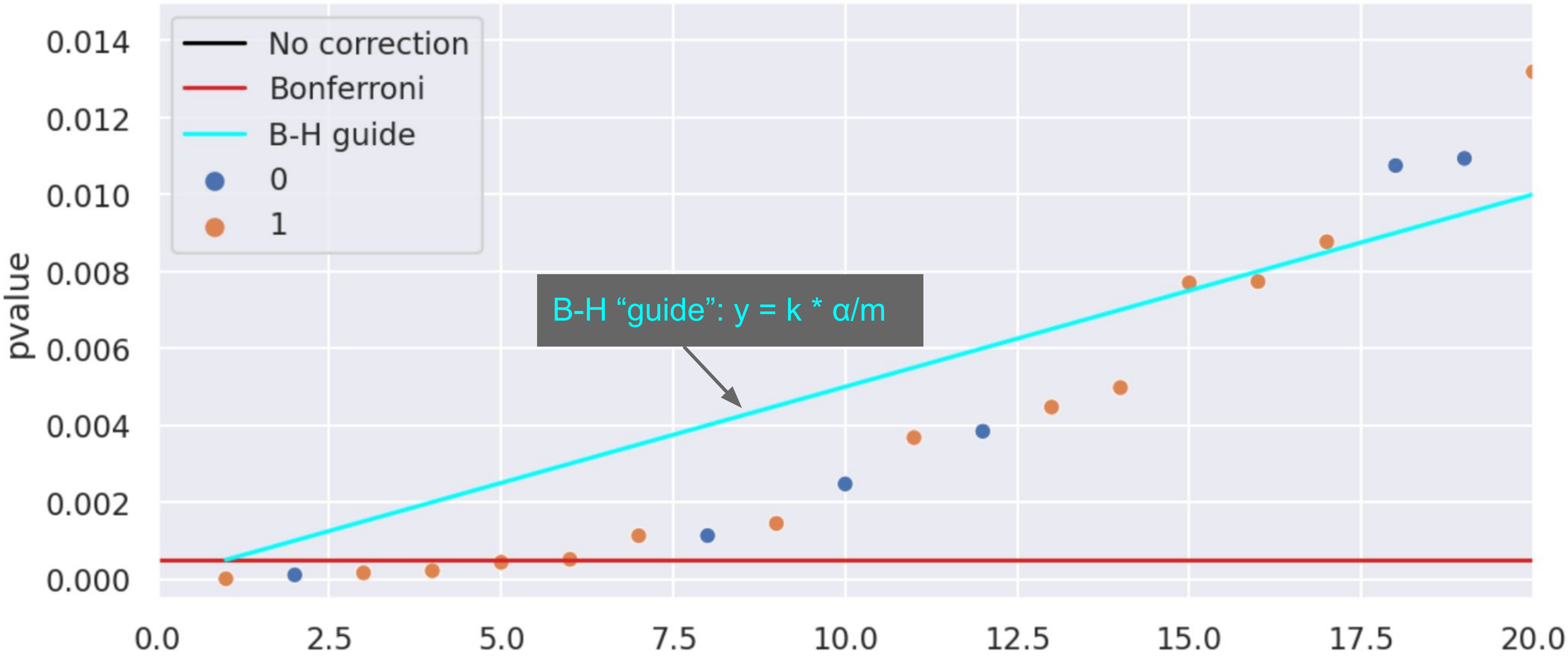
Randomness in FWER and FDR

Controlling FDR: Benjamini-Hochberg

- FDP: proportion of discoveries that were wrong *in our particular decisions*
  - Column-wise error rate
- FDR: expected value of FDP
  - Expectation is taken with respect to the randomness in our decisions
- Benjamini-Hochberg: how it works

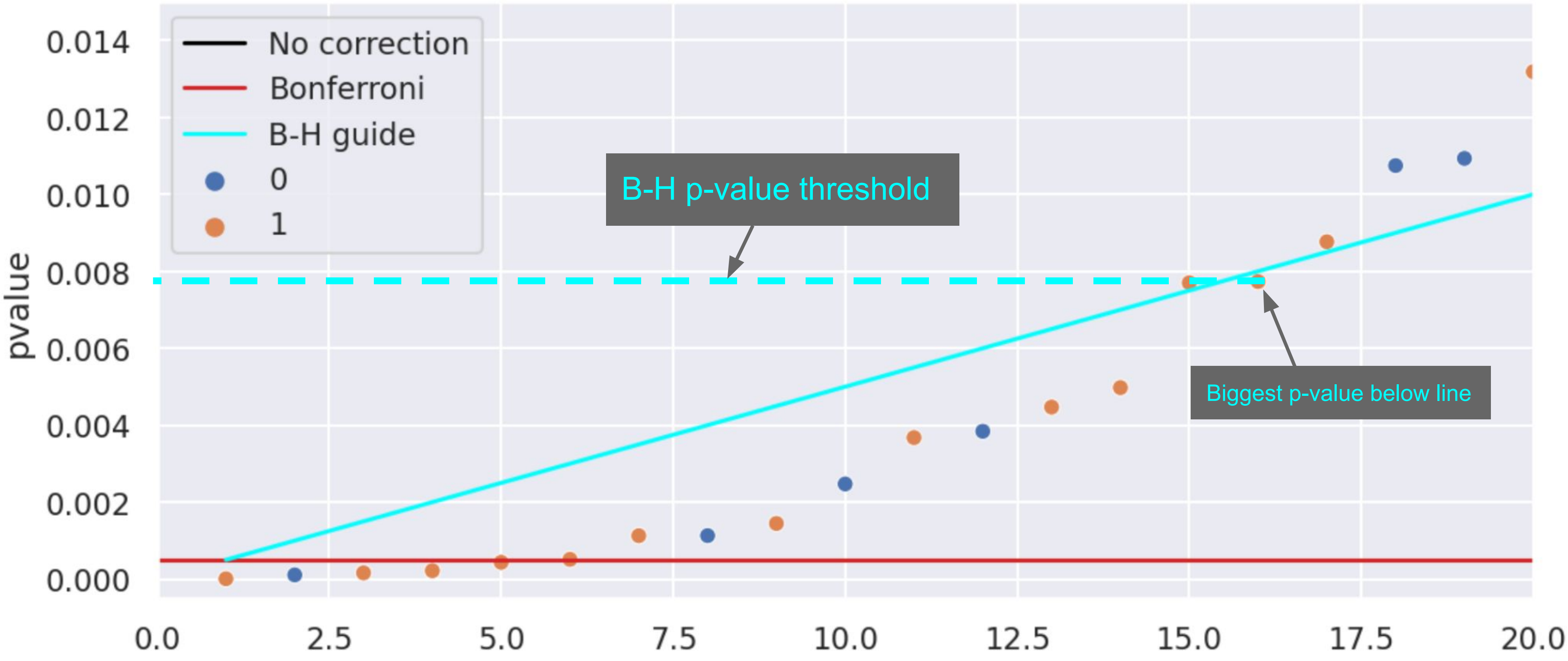
(1) sort p-values (indexed by k)
(2) draw the line y = k * α/m
(3) find the largest p-value that’s under the line, and
(4) use that p-value as the threshold

Benjamini-Hochberg



B-H “guide”: y = k * α/m

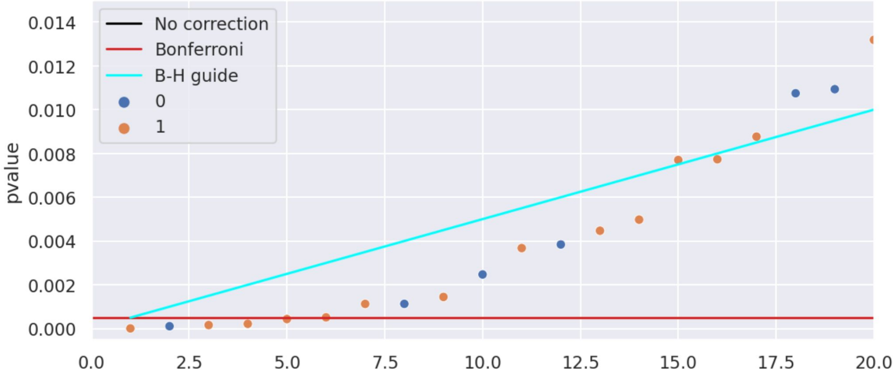
Benjamini-Hochberg



B-H p-value threshold

Biggest p-value below line

Why does B-H control FDR?


No correction
Bonferroni
B-H guide
● 0
● 1
pvalue