Binary Decision-Making: Multiple Decisions, Error Rates, and P-Values

Data 102 Spring 2022

Lecture 2

Weekly Overview

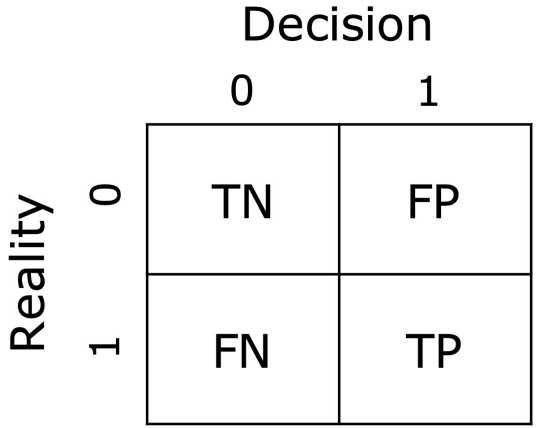
- Last week: course intro, binary decision-making

- **Today: multiple decisions**
  - Row- and column-wise error rates
  - Connecting decision-making and hypothesis testing
  - P-values and multiple hypothesis testing
  - Error rates for multiple testing

- Next time: making multiple decisions while controlling error rates

Recap: Binary Decision Making (One Decision)

- The simplest kind of decision: yes/no, 1 or 0, positive or negative.
- Setup
  - Reality is 0 or 1.
  - Make a decision (our best guess for reality) on a new instance.
  - The decision is 0 or 1 (and can be based on past experiences).
- Examples
  - COVID testing
  - Fraud detection
  - Predicting recidivism
  - Detecting underground oil wells
  - Movie/TV recommendations



Multiple Decisions

Single decision

Reality \ Decision	0	1
0	TN	FP
1	FN	TP

Multiple decisions

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

Number of times we made a "1" decision when reality was "0"

“Row-wise” rates: *what if we knew reality?*

- TNR: *specificity* $\dfrac{n_{00}}{n_{00} + n_{01}}$
- FPR: $\dfrac{n_{01}}{n_{00} + n_{01}}$
- TPR: *sensitivity recall* $\dfrac{n_{11}}{n_{10} + n_{11}}$
- FNR: $\dfrac{n_{10}}{n_{10} + n_{11}}$

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

[Wikipedia: Sensitivity and Specificity]

Pop Quiz

What is the mathematical relationship between TNR and FPR? Write an equation relating the two.

- TNR: *specificity* $\frac{n_{00}}{n_{00} + n_{01}}$
- FPR: $\frac{n_{01}}{n_{00} + n_{01}}$
- TPR: *sensitivity recall* $\frac{n_{11}}{n_{10} + n_{11}}$
- FNR: $\frac{n_{10}}{n_{10} + n_{11}}$

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

[Wikipedia: Sensitivity and Specificity]

A column-wise rate: what if we made a “1” decision?

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

False discovery proportion (FDP):

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

Interpreting row-wise and column-wise rates

- COVID testing
  - FPR: within people without COVID, how many test positive?
  - FDP: within positive tests, how many don’t have COVID?
- Movie/TV recommendations
  - FPR: within disliked shows, how many were recommended?
  - FDP: within recommendations, how many were disliked?

Question: For the fraud detection example, which of these is FDP and which is FPR?

1. Within flagged transactions, how many were valid?
2. Within valid transactions, how many are flagged for fraud?

Revisiting row-wise rates

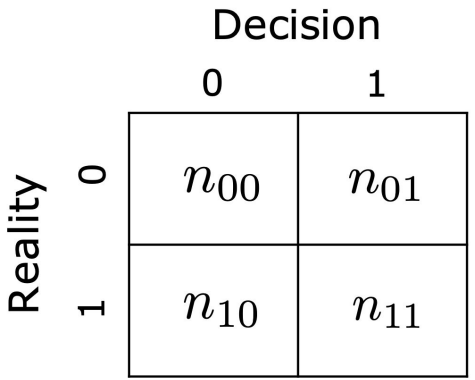
- TNR: $\frac{n_{00}}{n_{00} + n_{01}}$
  specificity
- FPR: $\frac{n_{01}}{n_{00} + n_{01}}$
- TPR: $\frac{n_{11}}{n_{10} + n_{11}}$
  sensitivity
  recall
- FNR: $\frac{n_{10}}{n_{10} + n_{11}}$

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

[Wikipedia: Sensitivity and Specificity]

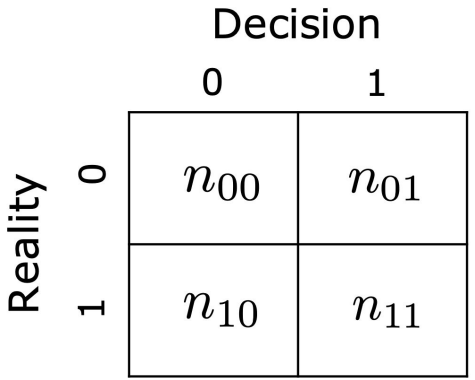
Row-wise rates as conditional probabilities

- TNR: P(D = 0 | R = 0)
- FPR: P(D = 1 | R = 0)
- ...and so on



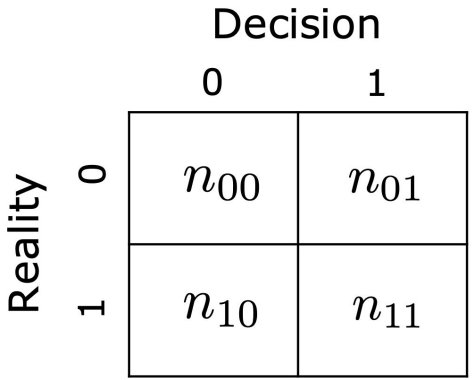
Discussion Question 1

Write the FNR (false negative rate) as a conditional probability.



Discussion Question 2

If we're using probabilities, then we're dealing with randomness. What all is random here? The data? The decision? Reality?



Probabilities and randomness

- In statistics, we always assume the *data* are random
- What about the decision?
- What about reality?

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

$$P(\ D = 1\ \mid\ R = 0\ )$$

Decision: function of data => random!

Reality: depends on whether we're being Bayesian or frequentist

Probabilities and randomness

- If reality is *random* (Bayesian mindset):
  - We'll think about column-wise rates like FDR
  - We need to specify "how" reality is random: what are P(R=0) and P(R=1)?
  - We'll call P(R=1) the prevalence, base rate, or π_1 (interchangeably)
  - "Bayesian" mindset

- If reality is *fixed* (frequentist mindset):
  - We can connect this to hypothesis testing (more on this later)
  - We want row-wise error rates like FPR and FNR as small as possible
  - In most cases, we can't have both: there are *tradeoffs*
  - The Neyman-Pearson framework (1932) gives us some useful theory
  - We can compute column-wise rates, but can't interpret them as probabilities
    - P(R = 0 | D = 0) isn't defined if R is fixed!
    - But, the fraction $\frac{n_{01}}{n_{01} + n_{11}}$ is still defined!

Relating column-wise and row-wise rates with Bayes' rule

$$
\begin{aligned}
FDP &= P(R=0|D=1) \\
\text{(using Bayes' rule)} &= \frac{P(D=1|R=0)P(R=0)}{P(D=1)} \\
\text{(Law of total probability)} &= \frac{P(D=1|R=0)P(R=0)}{P(D=1|R=0)P(R=0)+P(D=1|R=1)P(R=1)} \\
\text{(dividing by the numerator)} &= \frac{1}{1+\frac{P(D=1|R=1)}{P(D=1|R=0)}\frac{P(R=1)}{P(R=0)}} \\
\text{(applying definitions)} &= \frac{1}{1+\frac{TPR}{FPR}\frac{1-\pi_0}{\pi_0}}
\end{aligned}
$$

Frequentist: binary decision-making & hypothesis testing

- Hypothesis testing is a kind of decision making
- Reality: null vs alternative hypothesis
- Decision: whether or not we “reject” the null

How to do a hypothesis test (from Data 8)

- Figure out the viewpoint you want to test, and formulate:
  - **Null hypothesis**: Chance model under which you can simulate data
    - *(or, under which you can analytically compute the null distribution)*
  - **Alternative hypothesis**: Viewpoint from the question
  - **Test statistic**: to help you choose one viewpoint
- Compute the value of the test statistic in your data
- Simulate the test statistic under the null many times
  - (*or compute the null distribution analytically)*
- Compare the results

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)*

P-Values and null distributions

Exercise 1

Suppose the null hypothesis is true, but based on the p-value, we reject the null hypothesis. What kind of decision is this? True/false negative/positive?

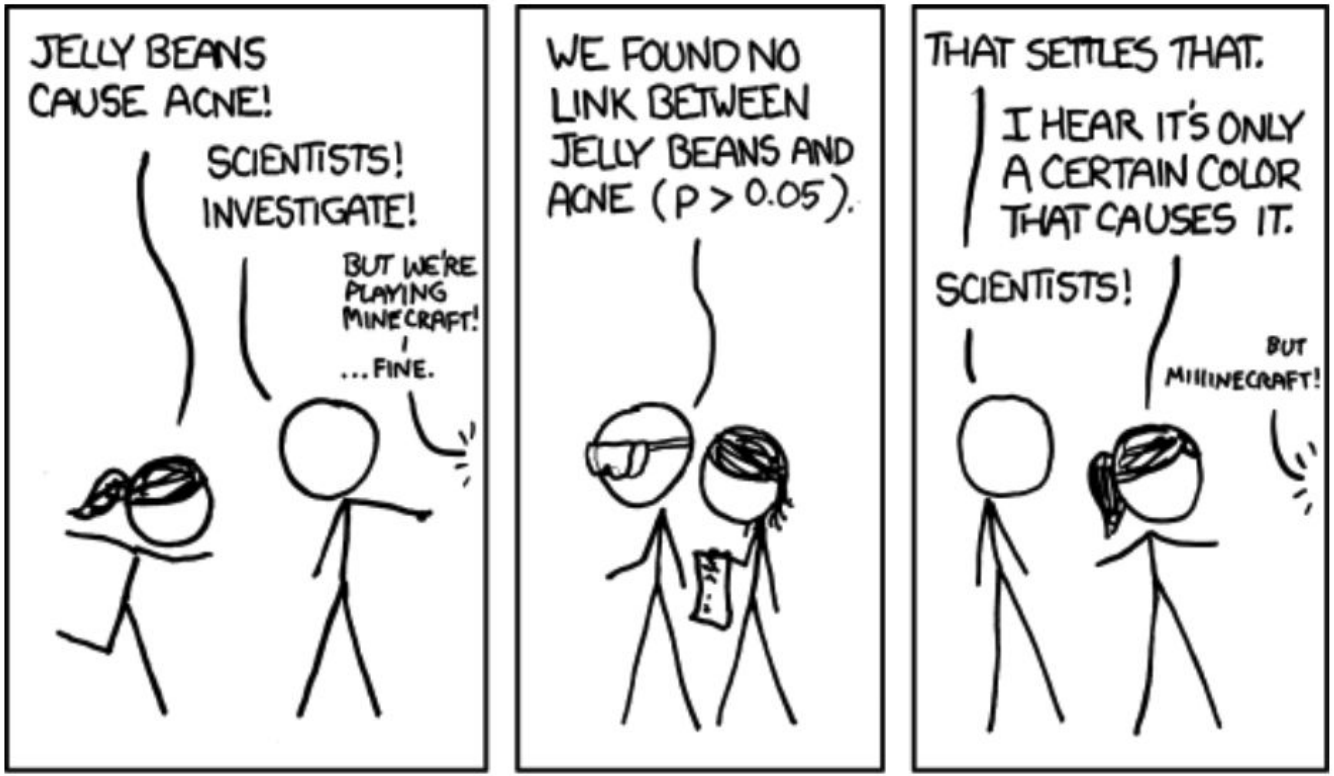
Exercise 2

If the null hypothesis is true, what is the probability of obtaining a p-value less than or equal to 0.05?

P-values: a closer look

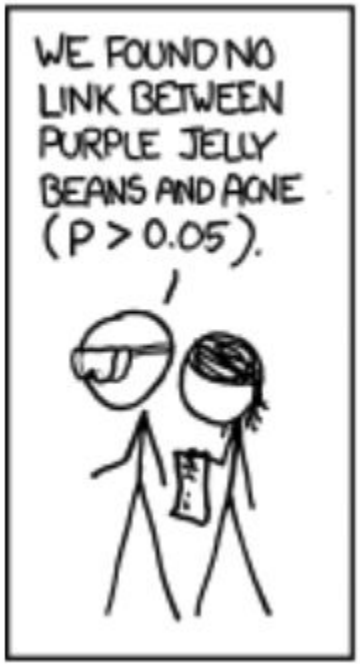
- If we can reason about the distribution of p-values, we can use that to help us build a theory of how to control error rates involving p-values
- Process:
  - Collect data (random)
  - Compute test statistic from data (also random)
  - Compute p-value from test statistic (also random)
- The p-value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

Multiple Hypothesis Testing



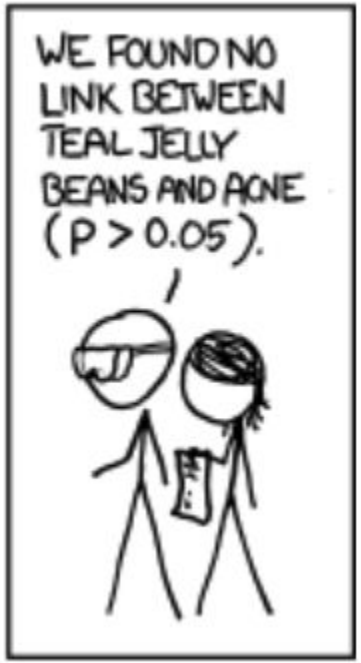
xkcd.com/882

Multiple Hypothesis Testing



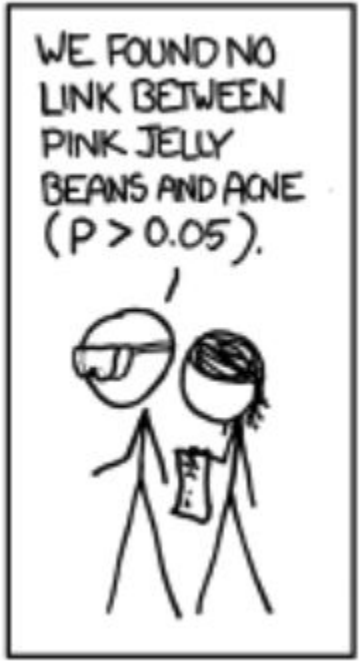
xkcd.com/882

Multiple Hypothesis Testing



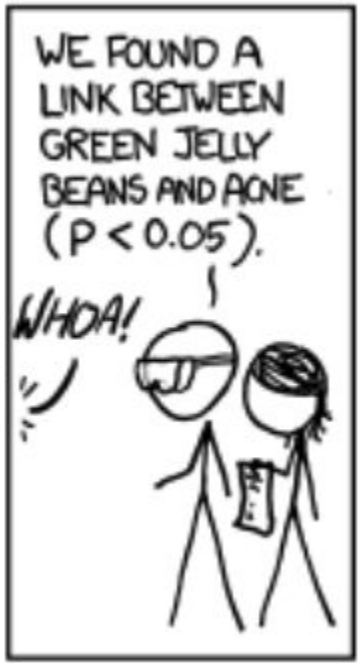
xkcd.com/882

Multiple Hypothesis Testing



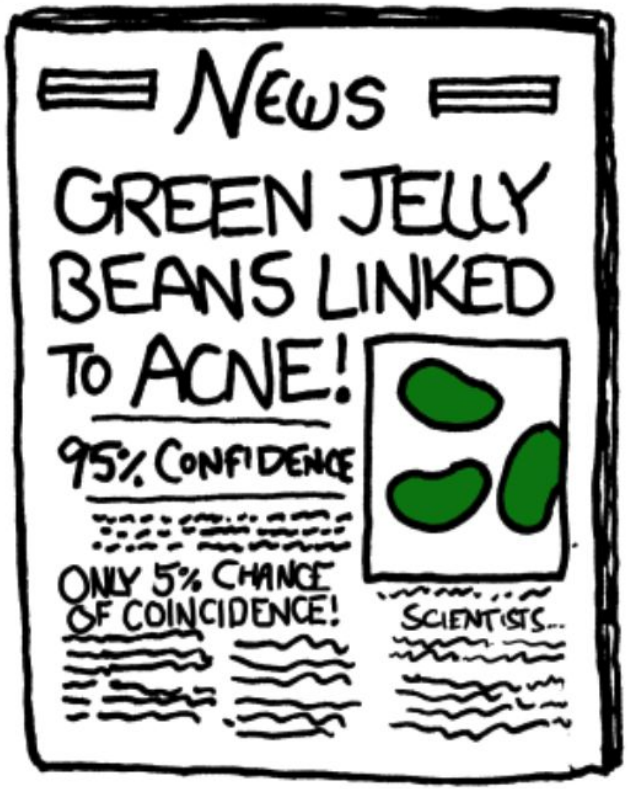
xkcd.com/882

Multiple Hypothesis Testing



xkcd.com/882

Multiple Hypothesis Testing



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

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Replication crisis

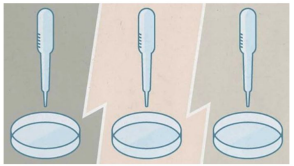
From Wikipedia, the free encyclopedia

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]

nature

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

ScienceAlert

HUMANS

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

MIKE MCRAE 27 AUG 2018

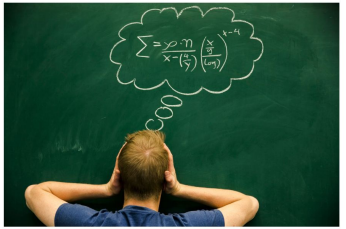
An attempt to replicate the findings of 21 social science experiments published in two high-profile science journals has thrown up a red flag for reliability in research.

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*

LIVE ON BLOOMBERG
Watch Live TV >
Listen to Live Radio >
Bloomberg Television

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.

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

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”