DS 102: Data, Inference, and Decisions

Lecture 3

Michael Jordan

University of California, Berkeley

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\}$$

$$\delta(X) \in \{0, 1\}$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

		Decision	
		0	1
Reality	0		
	1		

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: 0/1 loss

$$\theta \in \{0, 1\} \quad \text{(Reality)}$$

$$\delta(X) \in \{0, 1\} \quad \text{(Decision)}$$

		Decision 0	Decision 1
Reality	0	0	1
	1	1	0

Decision-Theoretic Framework

- Define a family of probability models for the data X, indexed by a parameter θ
- Define a procedure $\delta(X)$ that operates on the data to make a decision
- Define a loss function:

$$l(\theta, \delta(X))$$

- Example: L2 loss

$$\theta \in \mathbb{R}$$

$$\delta(X) \in \mathbb{R}$$

$$l(\theta, \delta(X)) = (\delta(X) - \theta)^2$$

Examples (on the White Board)

- The risk under the 0/1 loss
- The risk under the L2 loss

Back to Hypothesis Testing

- Let's now consider a column-wise perspective

Back to Hypothesis Testing

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

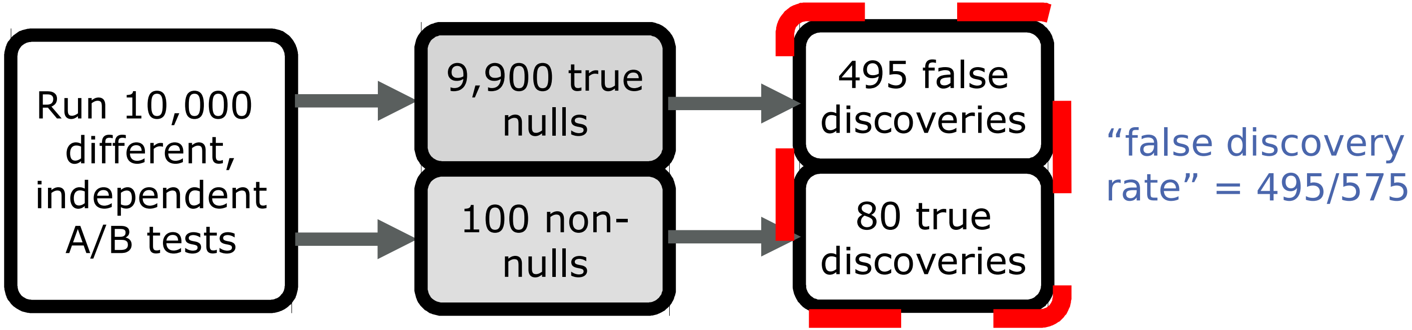
- Let's now consider a column-wise perspective

Some Column-Wise Rates

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

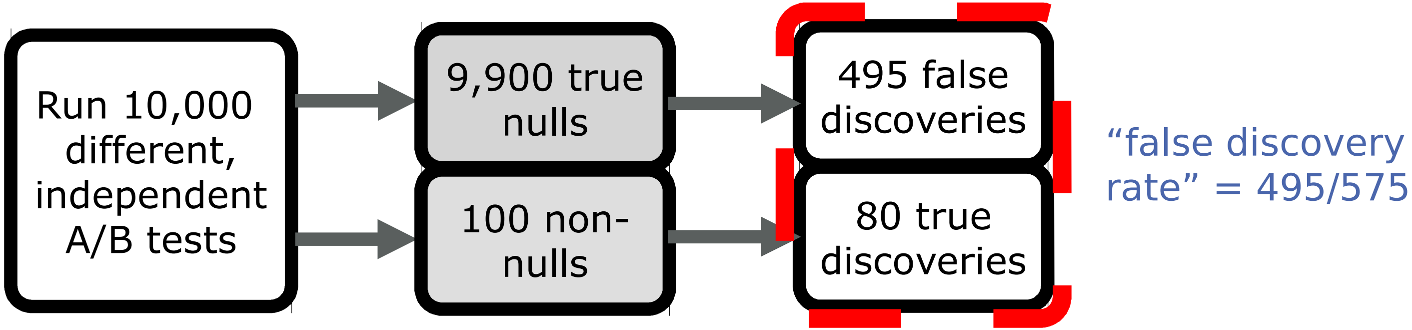
false discovery proportion $= \dfrac{n_{01}}{n_{01}+n_{11}}$

Type I error rate (per test) = 0.05



Power (per test) = 0.80

Type I error rate (per test) = 0.05



Power (per test) = 0.80

(NB: We’re again not being rigorous at this point; FDR is actually an expectation of this proportion. We’ll do it right anon.)

The Goal: Control Errors A Priori

- The row-focused Neyman-Pearson paradigm, with its Type I and Type II errors, provides a priori control
  - meaning that if my assumptions about the null and alternative distributions are correct, then I can guarantee that these errors will be small (in an average, frequentist sense---over multiple draws of data)

The Goal: Control Errors A Priori

- The row-focused Neyman-Pearson paradigm, with its Type I and Type II errors, provides a priori control
  - meaning that if my assumptions about the null and alternative distributions are correct, then I can guarantee that these errors will be small (in an average, frequentist sense---over multiple draws of data)
- If I'm only testing one hypothesis, that's satisfying

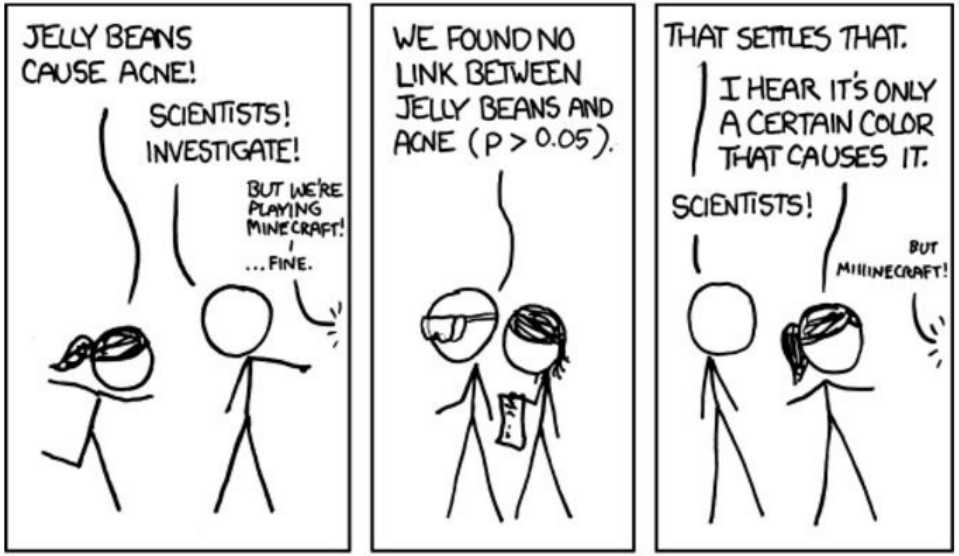
The Goal: Control Errors A Priori

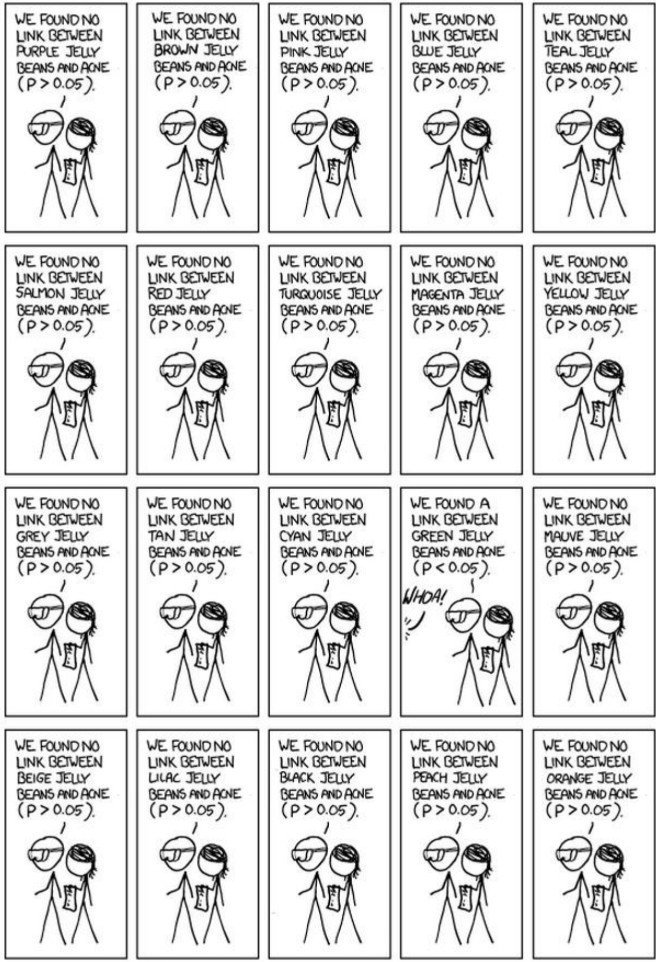
- The row-focused Neyman-Pearson paradigm, with its Type I and Type II errors, provides a priori control
  - meaning that if my assumptions about the null and alternative distributions are correct, then I can guarantee that these errors will be small (in an average, frequentist sense---over multiple draws of data)
- If I'm only testing one hypothesis, that's satisfying
- The problem that arose with our A/B testing example arose because we were doing many tests

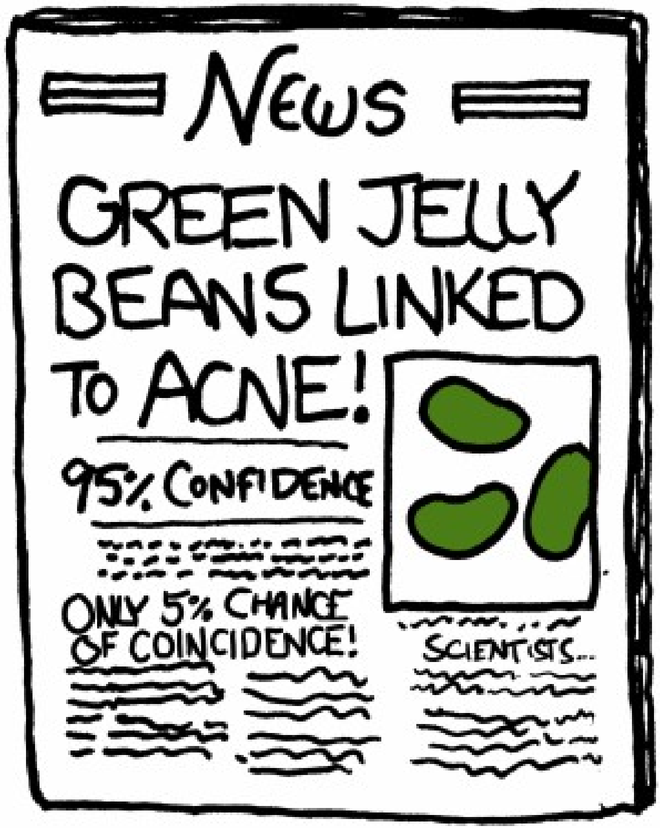
The Goal: Control Errors A Priori

- The row-focused Neyman-Pearson paradigm, with its Type I and Type II errors, provides a priori control
  - meaning that if my assumptions about the null and alternative distributions are correct, then I can guarantee that these errors will be small (in an average, frequentist sense---over multiple draws of data)
- If I’m only testing one hypothesis, that’s satisfying
- The problem that arose with our A/B testing example arose because we were doing many tests
- Can we find a way to obtain a priori control when there are many tests?

Multiple Decisions: The Statistical Problem




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


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

A First Attempt: Bonferroni

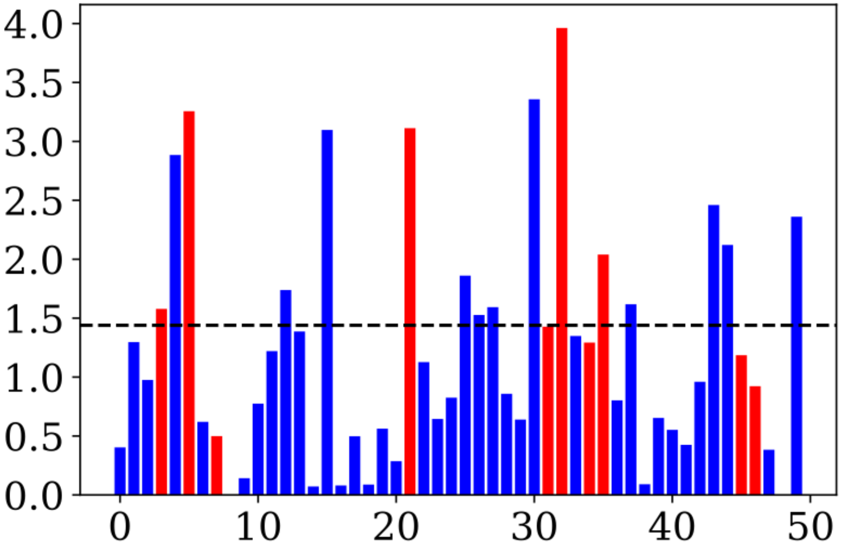
- Let's suppose that we're conducting m tests, not just one
- Let V denote the number of Type I errors in my m tests, and let $\{E_i = 1\}$ denote the event of a Type I error on the $i\mathrm{th}$ test
- Let's use a rejection threshold of α/m in the classical paradigm instead of α
- This controls a certain error rate…

A First Attempt: Bonferroni

$$
\begin{aligned}
P(V \geq 1) &= P(\cup_{i=1}^{m} \{E_i = 1\}) \\
&\leq \sum_{i=1}^{m} P(\{E_i = 1\}) \\
&\leq \sum_{i=1}^{m} \alpha/m \\
&= \alpha
\end{aligned}
$$

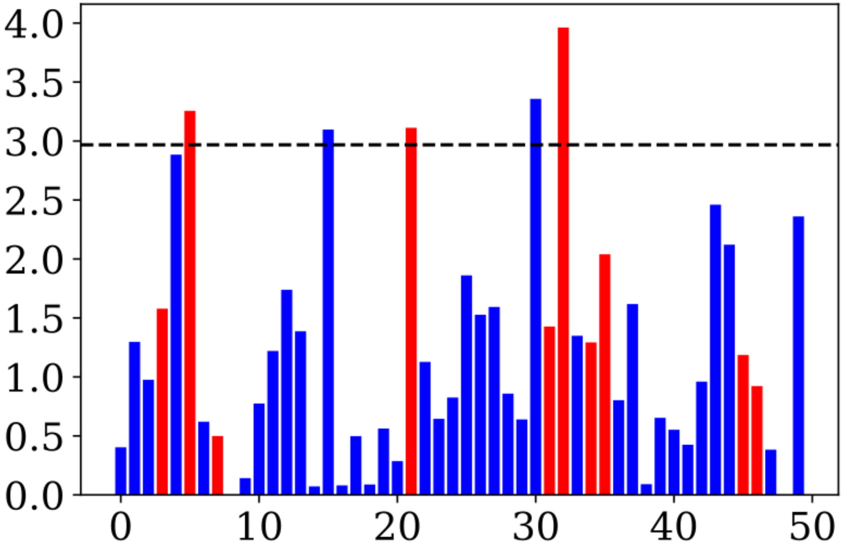
- We've controlled a quantity known as the family-wise error rate (FWER)

Naïve Multiple Hypothesis Testing



- This is the kind of mess that we've alluded to earlier; how about Bonferroni?

Bonferroni



- Bonferroni is overly stringent---it prevents us from making many discoveries

Let's Return to our Column-Wise Rates

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

$$\text{false discovery proportion} = \frac{n_{01}}{n_{01}+n_{11}}$$

Comments on the Column-Wise Rates

- They can be thought of as estimates of conditional probabilities
- They are dependent on the prevalence (i.e., the probabilities of the two states of Reality in the population), via Bayes' Theorem
  - as such, they are more Bayesian
  - this is arguably a good thing

- Notation: let H denote Reality, and let D denote the decision

A Bayesian Calculation

$$P(H = 0 \mid D = 1) = \frac{P(H = 0, D = 1)}{P(D = 1)}$$

A Bayesian Calculation

$$P(H = 0 \mid D = 1) = \frac{P(H = 0, D = 1)}{P(D = 1)}$$

$$= \frac{P(D = 1 \mid H = 0)P(H = 0)}{P(D = 1)}$$

$$= \frac{P(\text{Type I error}) \cdot \pi_0}{P(D = 1)}$$

- We could upper bound π_0 with 1, and so the numerator can be controlled; what about the denominator?

A Bayesian Calculation

- Using the law of total probability, we have:

$$P(D=1) = P(D=1 \mid H=0)P(H=0) + P(D=1 \mid H=1)P(H=1)$$

A Bayesian Calculation

- Using the law of total probability, we have:

$$
\begin{aligned}
P(D=1) &= P(D=1 \mid H=0)P(H=0) + P(D=1 \mid H=1)P(H=1) \\
&= \pi_0 P(D=1 \mid H=0) + (1-\pi_0)P(D=1 \mid H=1)
\end{aligned}
$$

- So we see that $P(D=1)$ depends on the prior π_0

A Bayesian Calculation

- Using the law of total probability, we have:

$$\begin{aligned} P(D=1) &= P(D=1 \mid H=0)P(H=0) + P(D=1 \mid H=1)P(H=1) \\ &= \pi_0 P(D=1 \mid H=0) + (1-\pi_0)P(D=1 \mid H=1) \end{aligned}$$

- So we see that $P(D=1)$ depends on the prior π_0
- Is this a problem?
  - i.e., do we have to either decide to be Bayesian and supply the prior, or decide to be frequentist and abandon this approach?

- No! Note that it's easy to estimate $P(D=1)$ directly from the data!

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain p-values P_i , and sort them from smallest to largest, denoting the sorted p-values as $P_{(k)}$

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain p-values P_i, and sort them from smallest to largest, denoting the sorted p-values as $P_{(k)}$
  - the small ones are the safest to reject

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain p-values P_i, and sort them from smallest to largest, denoting the sorted p-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

$$P_{(k)} \leq \frac{k}{m}\alpha$$

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain p-values P_i , and sort them from smallest to largest, denoting the sorted p-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

$$P_{(k)} \leq \frac{k}{m}\alpha$$

- Reject the null hypothesis (i.e., declare discoveries) for all hypotheses H_i such that $i \leq k$

Controlling the FDR

- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain p-values P_i, and sort them from smallest to largest, denoting the sorted p-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

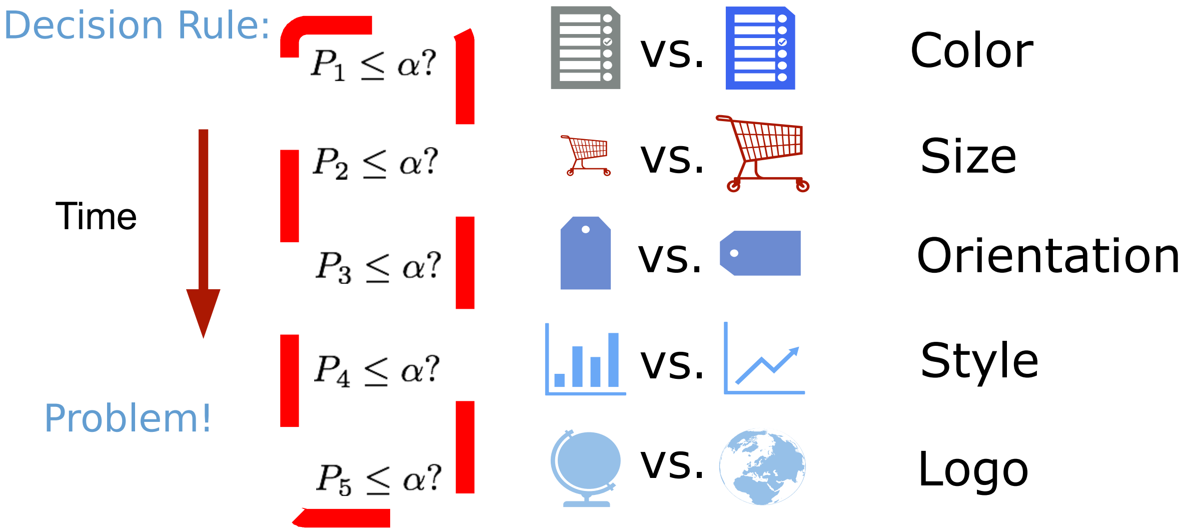
$$P_{(k)} \leq \frac{k}{m}\alpha$$

- Reject the null hypothesis (i.e., declare discoveries) for all hypotheses H_i such that $i \leq k$
- This controls the FDR!

The Online Problem

- Classical statistics, and also the Benjamini & Hochberg algorithm focused on a batch setting in which all data has already been collected
- E.g., for Benjamini & Hochberg, you need all of the p-values before you can get started
- Is is possible to consider methods that make sequences of decisions, and provide FDR control at any moment in time
- Is it conceivable that one can achieve lifetime FDR control?

Many enterprises run thousands of different (independent) A/B tests over time



What we will do instead:

Decision Rule:

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

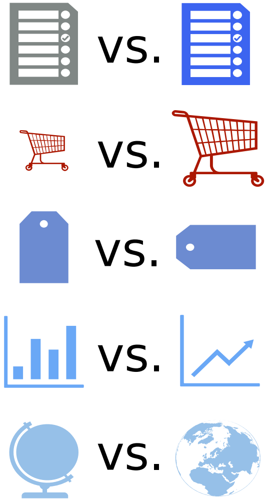
$P_3 \leq \alpha_3$?

$P_4 \leq \alpha_4$?

$P_5 \leq \alpha_5$?

Time

How do we set each error target to control FDR at any time?



Color

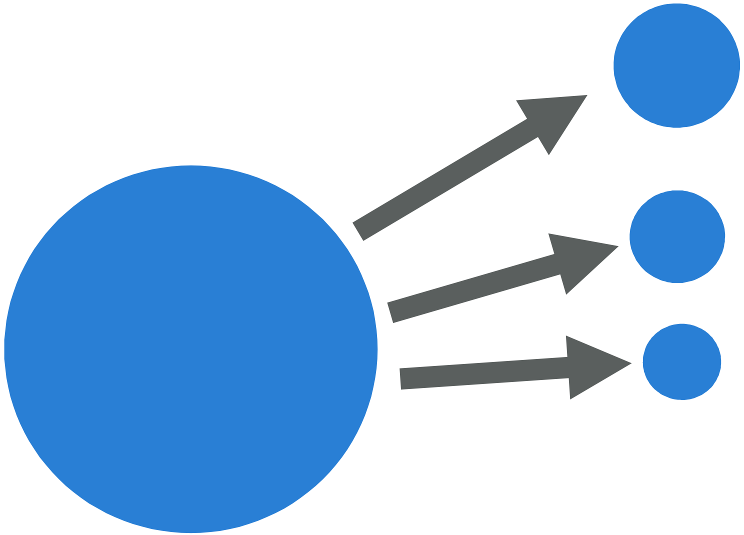
Size

Orientation

Style

Logo

Online FDR control : high-level picture



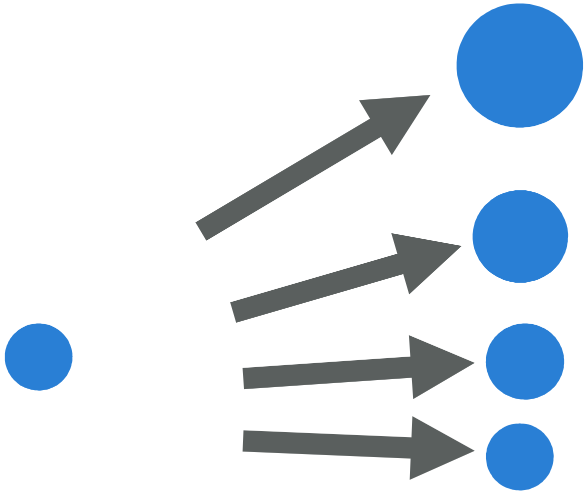
Error budget
for first test

Error budget for
second test

Tests use wealth

Remaining error budget
or “alpha-wealth”

Online FDR control : high-level picture



Error budget for first test

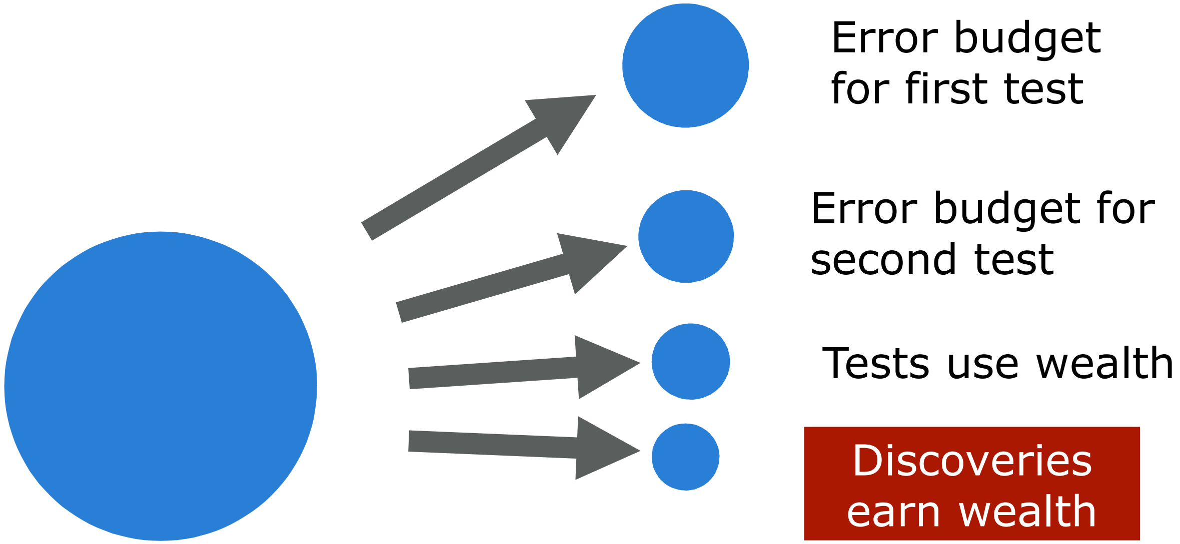
Error budget for second test

Tests use wealth

Discoveries earn wealth

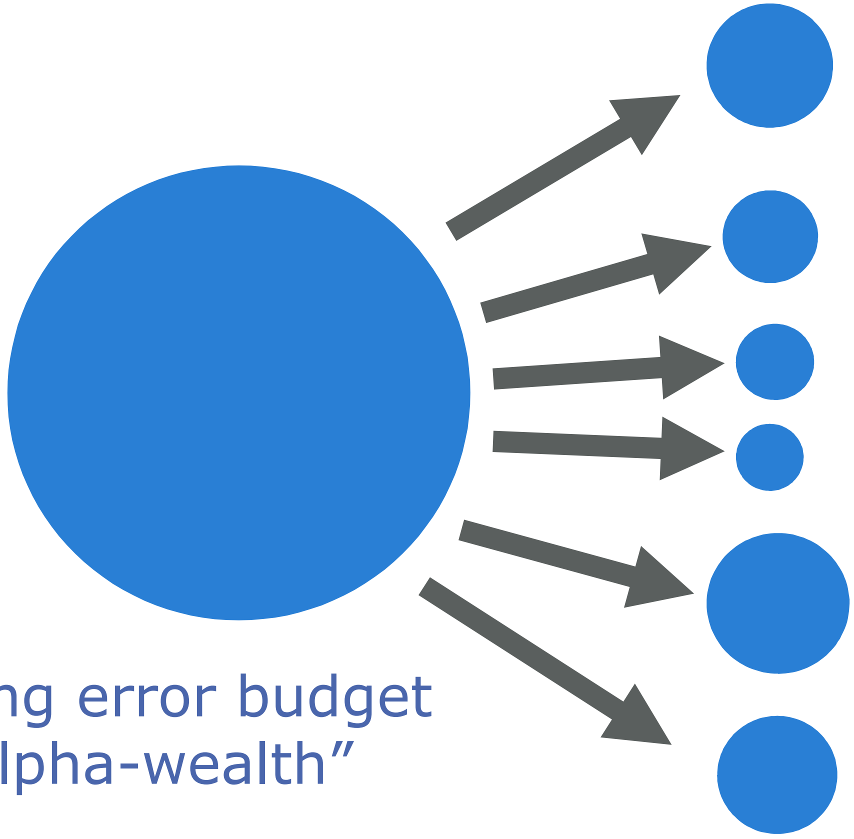
Remaining error budget or “alpha-wealth”

Online FDR control : high-level picture



Remaining error budget
or “alpha-wealth”

Online FDR control : high-level picture

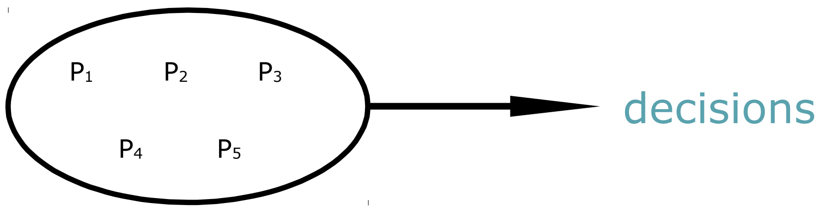


Online FDR control

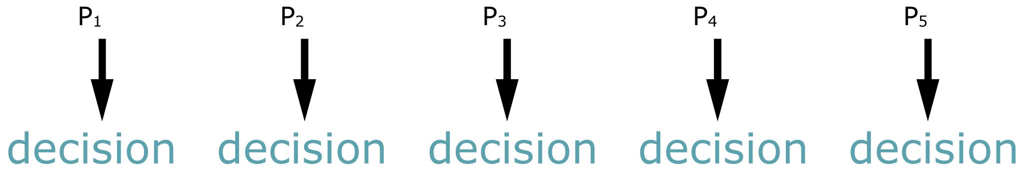
- classical FDR literature assumes that the data for all hypotheses is collected at once, and only after all the p-values are available, one can decide which of the hypotheses should be proclaimed discoveries
- in modern testing we often do not know how many hypotheses we want to test in advance
- instead, a possibly infinite sequence of tests (i.e. p-values) arrives *sequentially*
- we have to make decisions *online*, with no knowledge of future tests, in a way that guarantees FDR control under a pre-specified level α *at any given time*
- motivating examples: A/B testing, large-scale clinical trials…

Online vs offline FDR control

- classical FDR procedures (like BH) which make all decisions simultaneously are called “offline”

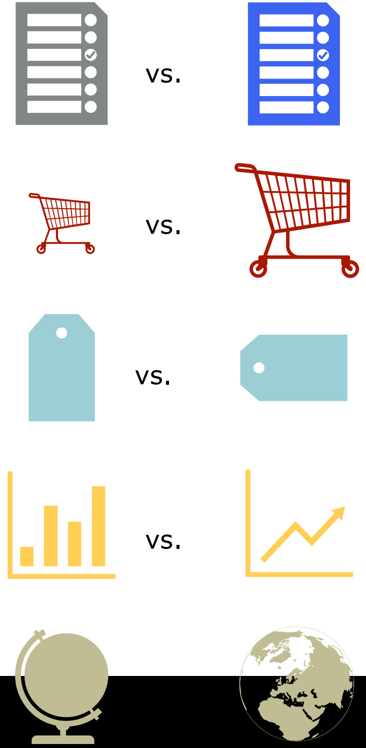


- online FDR procedures make decisions one at a time



Example: A/B testing

- online FDR algorithms pick significance level α_t adaptively



Color — $P_1 \leq \alpha_1$?

Size — $P_2 \leq \alpha_2$?

Orientation — $P_3 \leq \alpha_3$?

Style — $P_4 \leq \alpha_4$?

[redacted]

Online FDR algorithm

- the first online FDR algorithm was due to Foster and Stine (2008)
- a more recent (and simpler) online FDR algorithm is due to Javanmard and Montanari, and is called LORD
- its basic idea is to assign α_t in a way that ensures $\widehat{\mathrm{FDP}}(t) := \frac{\sum_{i=1}^{t} \alpha_i}{\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}} \leq \alpha$

- Why ensuring $\widehat{\mathrm{FDP}}(t) := \frac{\sum_{i=1}^t \alpha_i}{\sum_{i=1}^t 1\{P_i \le \alpha_i\}} \le \alpha$ controls FDR:

$$\mathrm{FDR} \approx \frac{\mathbb{E}[\sum_{i\le t, i \text{ null}} 1\{P_i \le \alpha_i\}]}{\mathbb{E}[\sum_{i \le t} 1\{P_i \le \alpha_i\}]}$$

, and we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i\le t, i \text{ null}} 1\{P_i \le \alpha_i\}\right] &= \sum_{i\le t, i \text{ null}} \mathbb{E}[\mathbb{E}[1\{P_i \le \alpha_i\}|\alpha_i]] = \sum_{i\le t, i \text{ null}} \mathbb{E}[\mathbb{P}\{P_i \le \alpha_i|\alpha_i\}] \\ &= \sum_{i\le t, i \text{ null}} \mathbb{E}[\alpha_i] \le \mathbb{E}[\sum_{i\le t} \alpha_i] \le \alpha\mathbb{E}[\sum_{i\le t} 1\{P_i \le \alpha_i\}]\end{aligned}$$

$$\mathrm{FDR} \le \alpha$$

so

Back to Inference

- Can we develop general frameworks that allow us to control column-wise quantities like the false-discovery rate (FDR)?
  - in a similar way as Neyman-Pearson controls the false-positive rate
- To be continued…