DS 102: Data, Inference, and Decisions

Lecture 2

Michael Jordan
University of California, Berkeley

Basics of Decision Making

- We'll start by considering the most simple of decision-making formulations
- Let's suppose that Reality is in one of two states, which we denote as 0 or 1
- We don't observe this state, but we do obtain Data that is drawn from a distribution that depends on whether the state is 0 or 1
- We make a Decision based on the Data, which we denote as 0 or 1
- We can think of the Decision as our best guess as to the state of Reality or, more generally, as an action we think is best given our guess of the state of Reality

The Basic Two-by-Two Table

Reality \ Decision	0	1
0	TN	FP
1	FN	TP

TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

The Basic Two-by-Two Table

Reality \ Decision	0	1
0	TN	FP
1	FN	TP

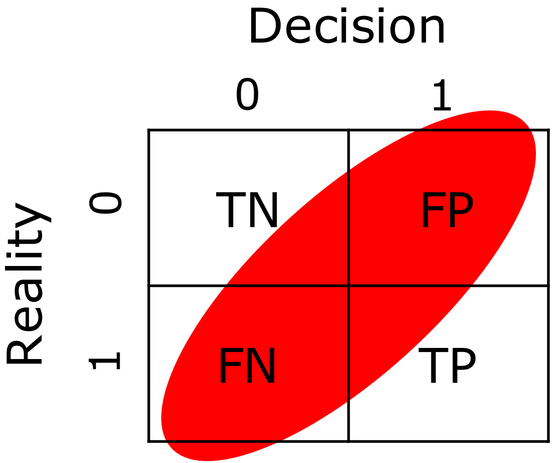
TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

The Basic Two-by-Two Table



TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

Rough goal: lots of green outcomes, few red outcomes!

Towards a Statistical Framework

- Let's now imagine that we not only make a decision, but we build a decision-making algorithm
- We want to evaluate the algorithm not just on one problem, but on a set of related problems

- Concretely, we may have a collection of hypothesis-testing problems, where we repeatedly decide whether to accept the null or accept the alternative
- Or we may have a set of classification decisions, where we repeatedly classify data points into one of two classes

Towards a Statistical Framework

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

$$N = n_{00} + n_{01} + n_{10} + n_{11}$$

Some Row-Wise Rates

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

$$\text{sensitivity} = \frac{n_{11}}{n_{10}+n_{11}}$$

aka, "true positive rate" or "recall" or "power"

Some Row-Wise Rates

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

$$\text{specificity} = \frac{n_{00}}{n_{00}+n_{01}}$$

aka, “true negative rate” or “selectivity”

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
  - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$
- As such, they are not dependent on the prevalence (i.e., the probabilities of the two states of Reality in the population)

The Bayesian Posterior

- The posterior probability of the hypothesis given the data:

$$P(\text{Reality} \mid \text{Decision}) = \frac{P(\text{Decision} \mid \text{Reality})P(\text{Reality})}{P(\text{Decision})}$$

where $P(\text{Reality})$ is the prior (the “prevalence”)

Back to Hypothesis Testing

- Let's now consider a column-wise perspective

Let's Return to our Column-Wise Rates

Decision

Reality	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

Bayes' Theorem

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

Bayes' Theorem

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

- This relates a row-wise quantity, $P(D = 1 \mid H = 0)$, to a column-wise quantity, $P(H = 0 \mid D = 1)$
- And shows that the latter depends on the prevalence: $P(H = 0) = 1 - P(H = 1)$

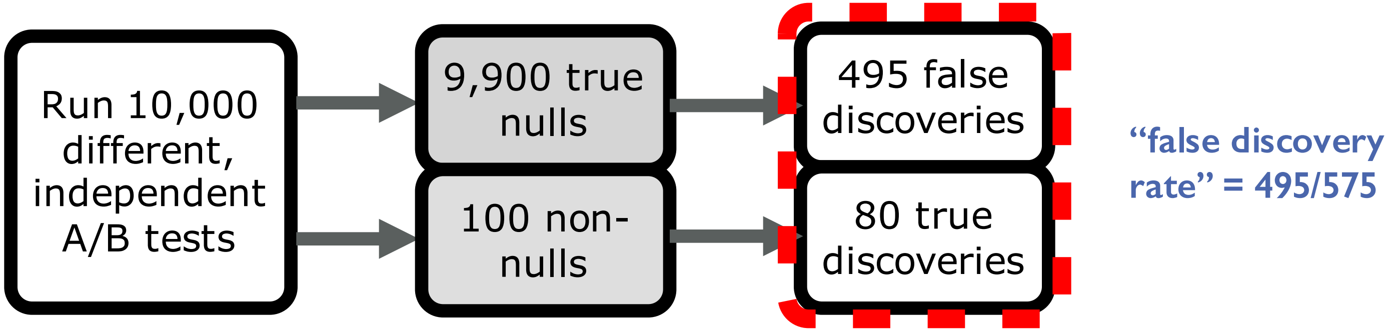
A Bayesian Calculation

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

Some Implications

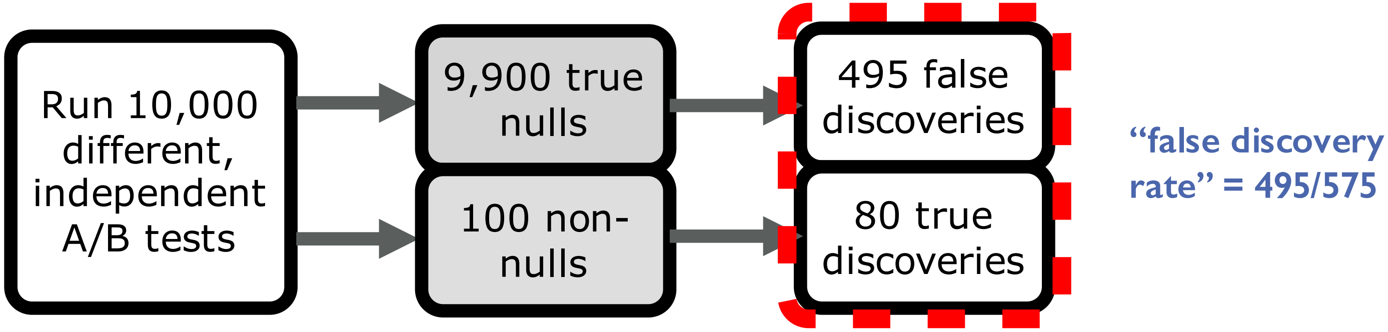
- We see that the prevalence has a major effect on the probability of an error
- Suppose that $P(D = 1 \mid H = 1) = 0.8$ and $P(D = 1 \mid H = 0) = 0.05$
- Then the ratio is 16/1, and if the prevalence was 0.5, the probability of an error would be small
- But…. if the prevalence is small, say 1/1000, then the factor $(1 - \pi_0)/\pi_0$ is tiny and it kills the 16/1
- And so the probability of error goes to one ☹

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

- We've introduced concepts such as false-positive rates and false-discovery rates as descriptions of performance
- We now want to use them as ways to design algorithms
- We want to give a priori guarantees that a certain algorithm will have good performance

The Neyman-Pearson Paradigm

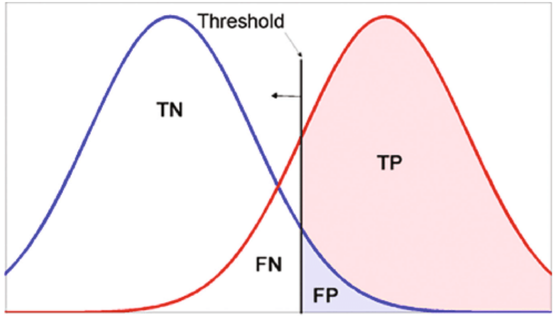
- The row-focused Neyman-Pearson paradigm turns the problem into a constrained optimization problem

The Neyman-Pearson Paradigm

- The row-focused Neyman-Pearson paradigm turns the problem into a constrained optimization problem
- The idea is to control the false-positive probability, $P(D = 1 \mid H = 0)$, to be less than some target value, say 0.05
- And to maximize the true-positive probability (the power) subject to that constraint

The Neyman-Pearson Paradigm

- The row-focused Neyman-Pearson paradigm turns the problem into a constrained optimization problem
- The idea is to control the false-positive probability, $P(D = 1 \mid H = 0)$, to be less than some target value, say 0.05
- And to maximize the true-positive probability (the power) subject to that constraint



P-Values

- Consider a simple null hypothesis $\mathbb{P}$
- Consider a statistic, $T(X)$, which has a continuous distribution under the null, and let $F(t)$ denote its tail cdf:

$$F(t) = \mathbb{P}(T > t)$$

- Define the P-value as $P = F(T)$
- The P-value has a uniform distribution under the null:

$$\mathbb{P}(P < p) = \mathbb{P}(F(T) < p) = \mathbb{P}(T > F^{-1}(p)) = F(F^{-1}(p)) = p$$

A Generic Decision Rule

- Reject H_i if the random variable T_i is equal to 1:

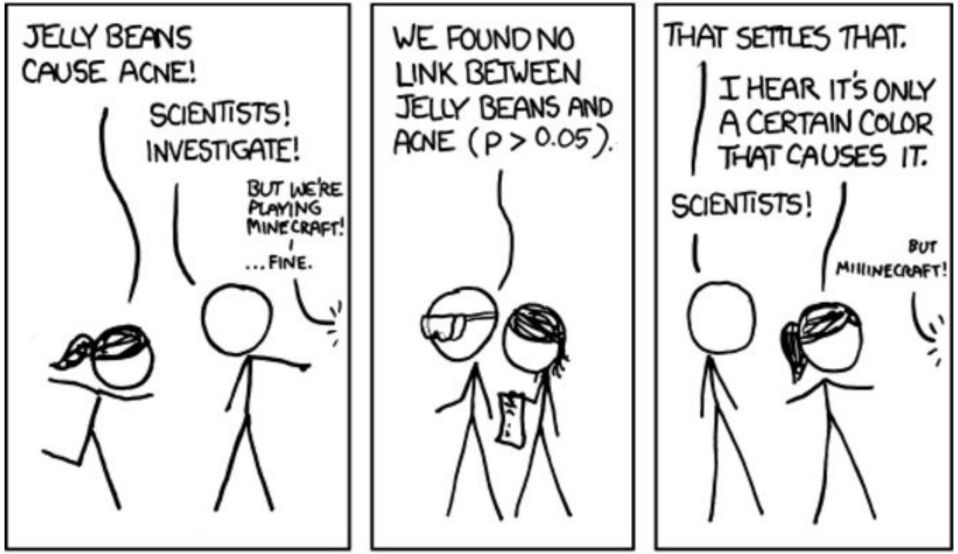
$$T_i = \begin{cases} 1, & \text{if } P_i \leq \alpha_i \\ 0, & \text{otherwise} \end{cases}$$

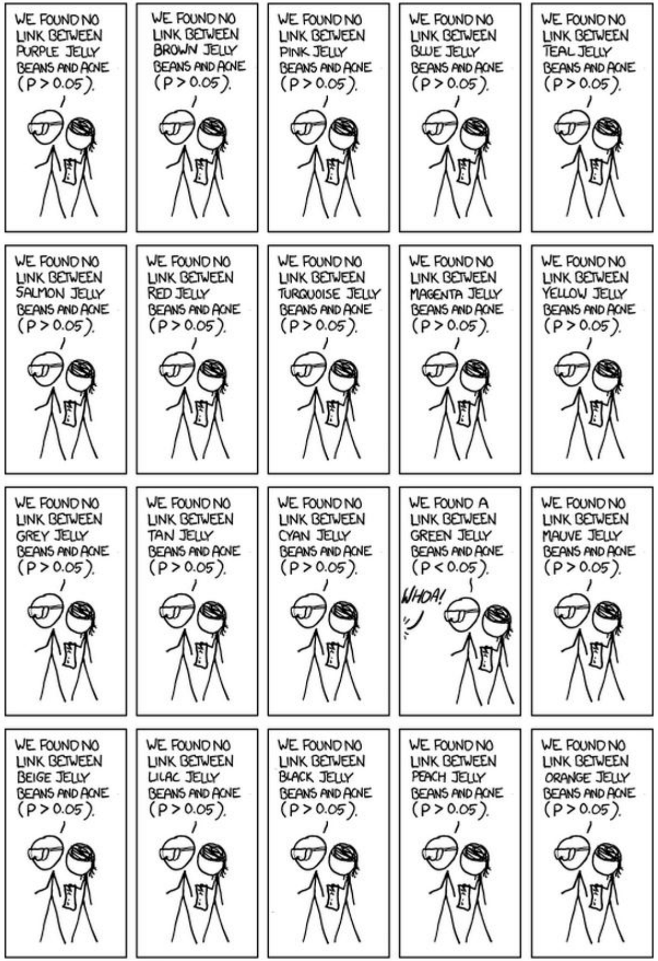
- This yields Neyman-Pearson control in the case of a single simple hypothesis (where all the H_i are the same and all the α_i are set equal to some fixed value, say 0.05)

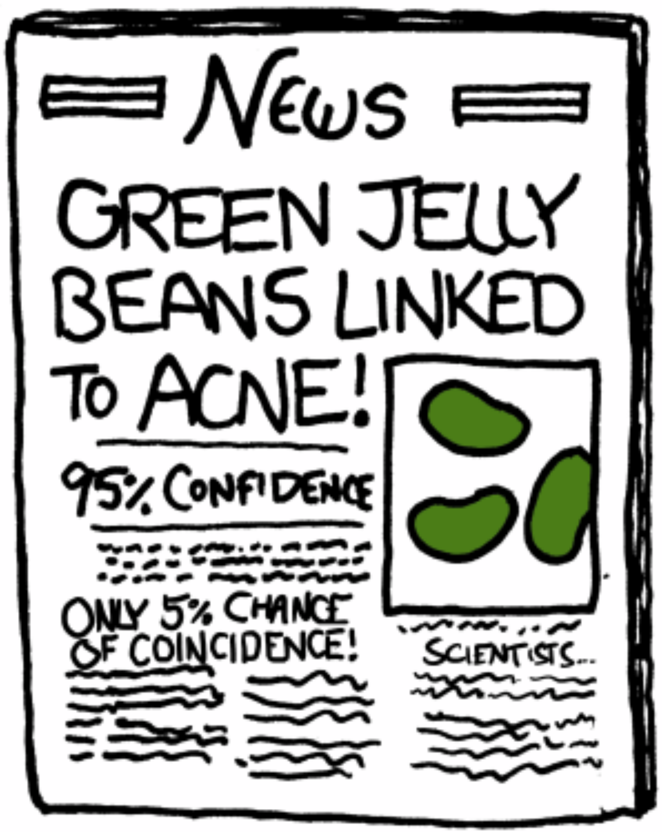
Multiple Hypothesis Testing

- Let’s now consider multiple tests, in particular repeated tests of the same hypothesis
- The row-focused Neyman-Pearson paradigm provides a priori control over errors made in those cases in which the null hypothesis is true
- This isn’t very natural when the hypotheses are “cases” which arise randomly according to their prevalence
- It also makes little sense when we’re testing a bag of different hypothesis (cf., A/B testing)

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

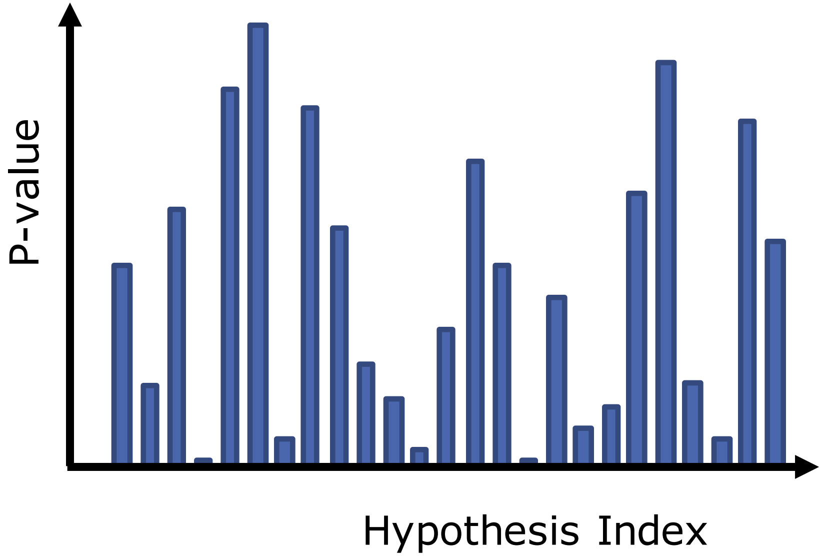
- Let's suppose that we're conducting m tests, not just one
- Let V denote the number of false-positive errors in my m tests, and let $\{E_i = 1\}$ denote the event of a false positive error on the $i\text{th}$ test
- Let's use a rejection threshold of α/m in the classical paradigm instead of α
- This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

A First Attempt: Bonferroni

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

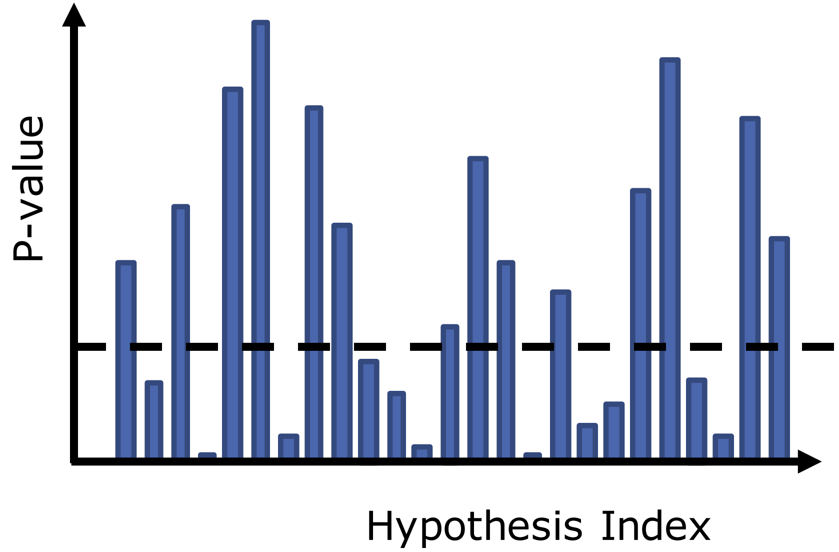
- We've controlled a quantity known as the family-wise error rate (FWER)

Example



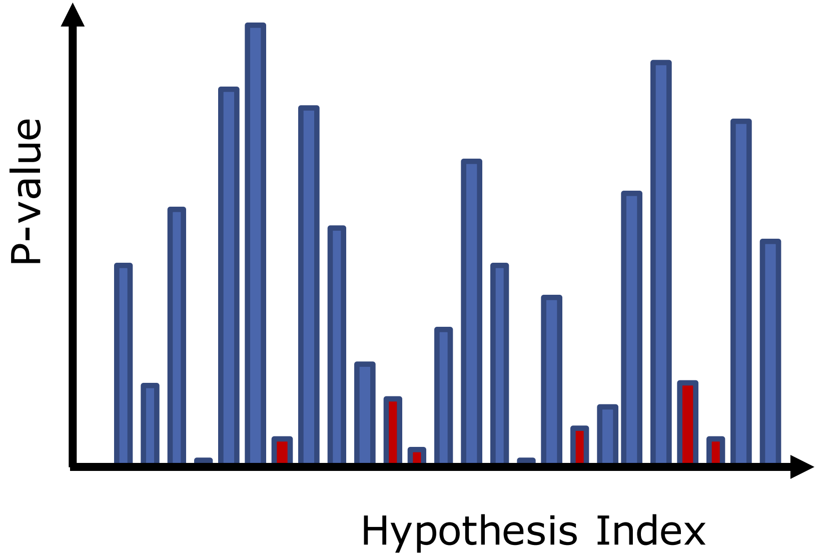
- Suppose that we obtain p-values from 25 experiments

Naïve Multiple Decision-Making



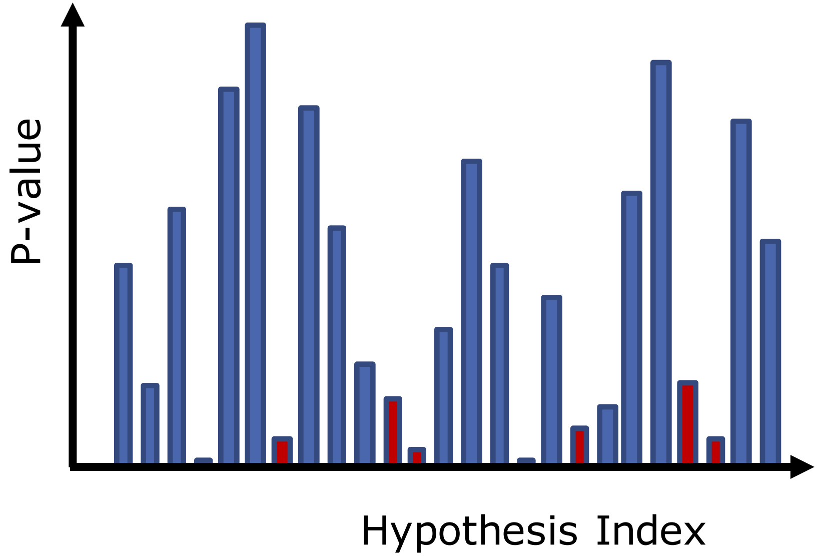
- Suppose that we simply reject each test independently if its p-value is smaller than some thresholding

Naïve Multiple Decision-Making



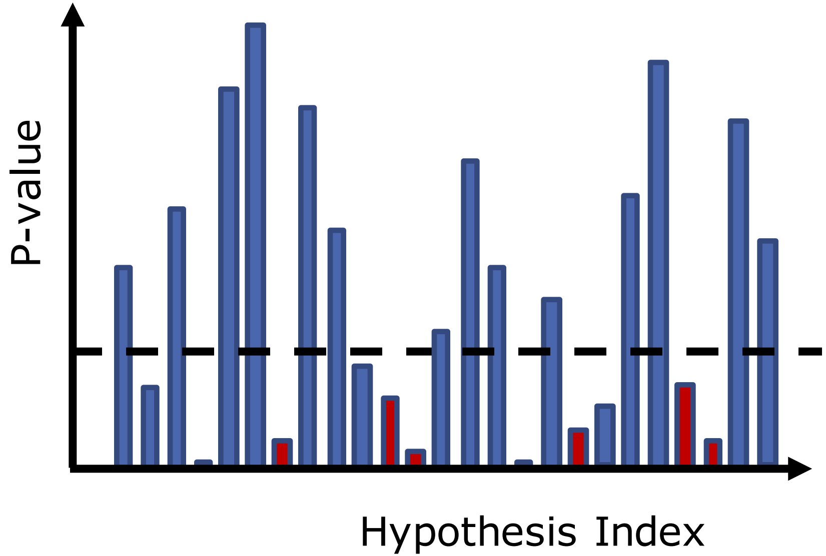
- Suppose that we simply reject each test independently if its p-value is smaller than some thresholding

Naïve Multiple Decision-Making



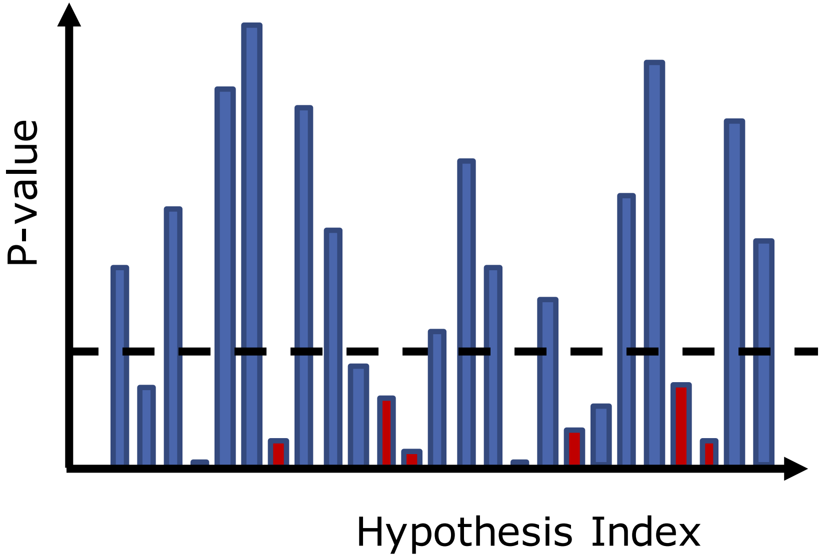
- An oracle knows the truth: that the blue-shaded bars correspond to nulls (Reality = 0) and the red-shaded bars to alternatives (Reality = 1)

Naïve Multiple Decision-Making



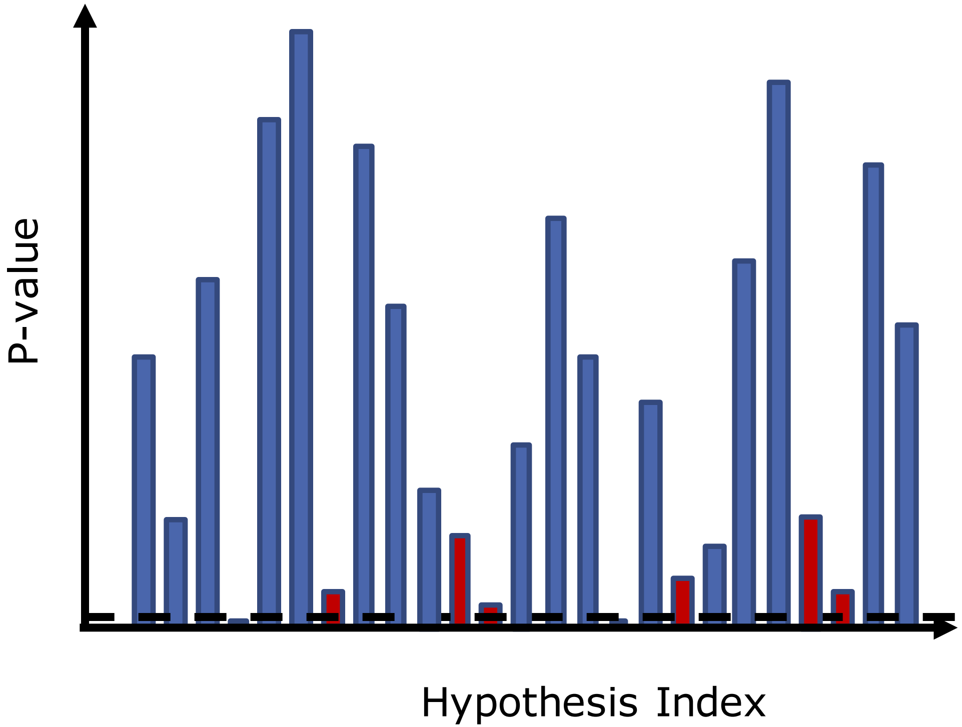
- We see that the decision-maker is avoiding false negatives, but its false discovery proportion is 4/11; pretty bad!

Naïve Multiple Decision-Making



- We see that the decision-maker is avoiding false negatives, but is making a lot of false positives, and its false discovery proportion is 4/11; pretty bad!

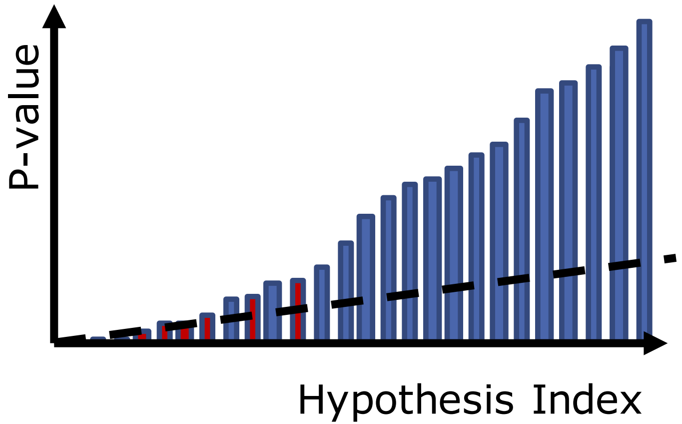
Bonferroni



- Bonferroni avoids those false positives, but is making a lot of false negatives, and its false discovery proportion is 1/2; even worse!

Is There Something Else We Can Do?

- It's not clear that any fixed threshold will work, and it's not how to set such a threshold without knowing the truth
- We have to think out of the box: we'll be developing a procedure that works with sorted p-values, and compares them to a line with a positive slope, not a horizontal line!



Is There Something Else We Can Do?

- It’s not clear that any fixed threshold will work, and it’s not how to set such a threshold without knowing the truth
- We have to think out of the box: we’ll be developing a procedure that works with sorted p-values, and compares them to a line with a positive slope, not a horizontal line!

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:

$$P(D=1) = P(D=1\,|\,H=0)P(H=0) + P(D=1\,|\,H=1)P(H=1)$$
$$= \pi_0 P(D=1\,|\,H=0) + (1-\pi_0)P(D=1\,|\,H=1)$$

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

A Bayesian Calculation

- Using the law of total probability, we have:

$$P(D=1) = P(D=1 \mid H=0)P(H=0) + P(D=1 \mid H=1)P(H=1)$$
$$= \pi_0 P(D=1 \mid H=0) + (1-\pi_0)P(D=1 \mid H=1)$$

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!

Towards an Algorithm

- We will plug in an estimate of $P(D = 1)$ into the Bayesian posterior probability
  - this is called empirical Bayesian
- And we will use the empirical Bayesian estimate to set a threshold
- Let's consider

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:

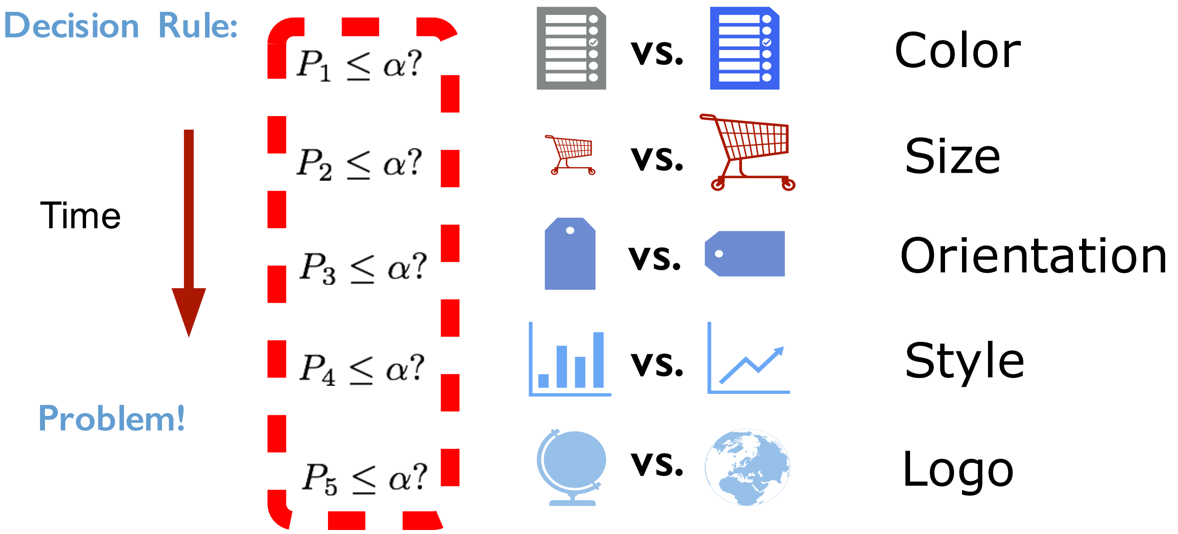
$$P_{(k)} \leq \frac{k}{m}\alpha$$

- 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



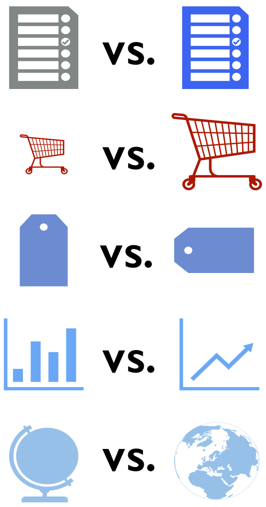
What we will do instead:

Decision Rule:

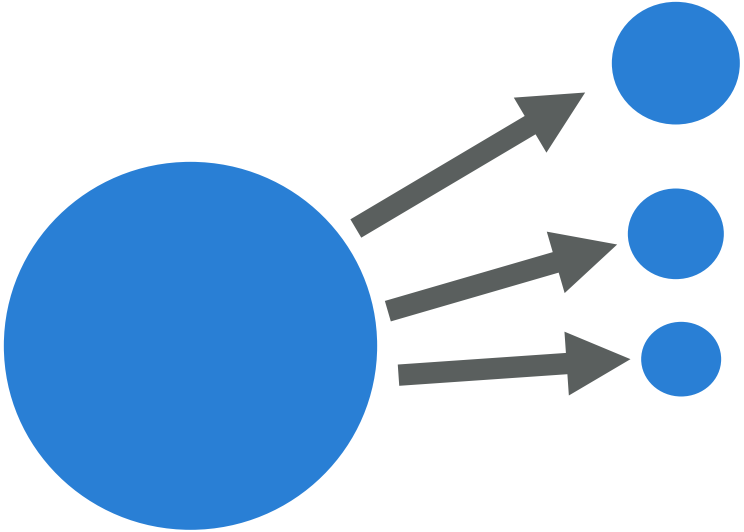
Time ↓

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

Decision Rule		
$P_1 \leq \alpha_1$?	vs.	Color
$P_2 \leq \alpha_2$?	vs.	Size
$P_3 \leq \alpha_3$?	vs.	Orientation
$P_4 \leq \alpha_4$?	vs.	Style
$P_5 \leq \alpha_5$?	vs.	Logo



Online FDR control : high-level picture



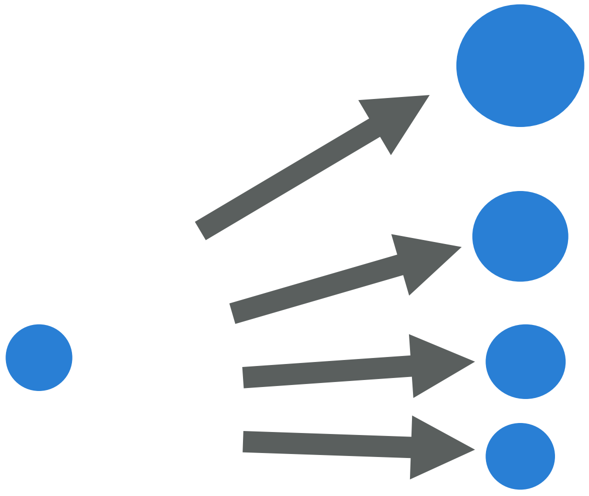
Error budget
for first test

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

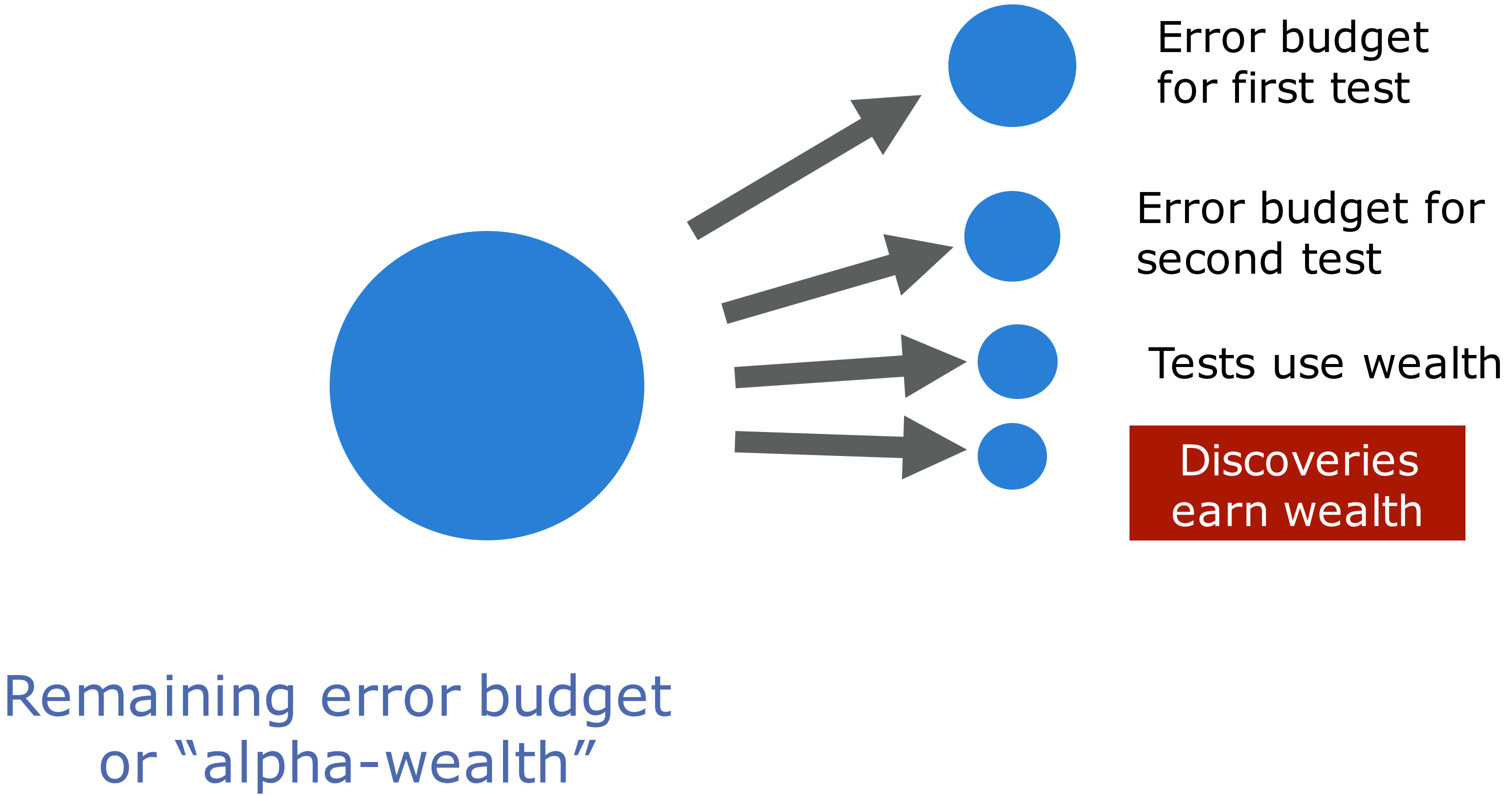
Error budget for second test

Tests use wealth

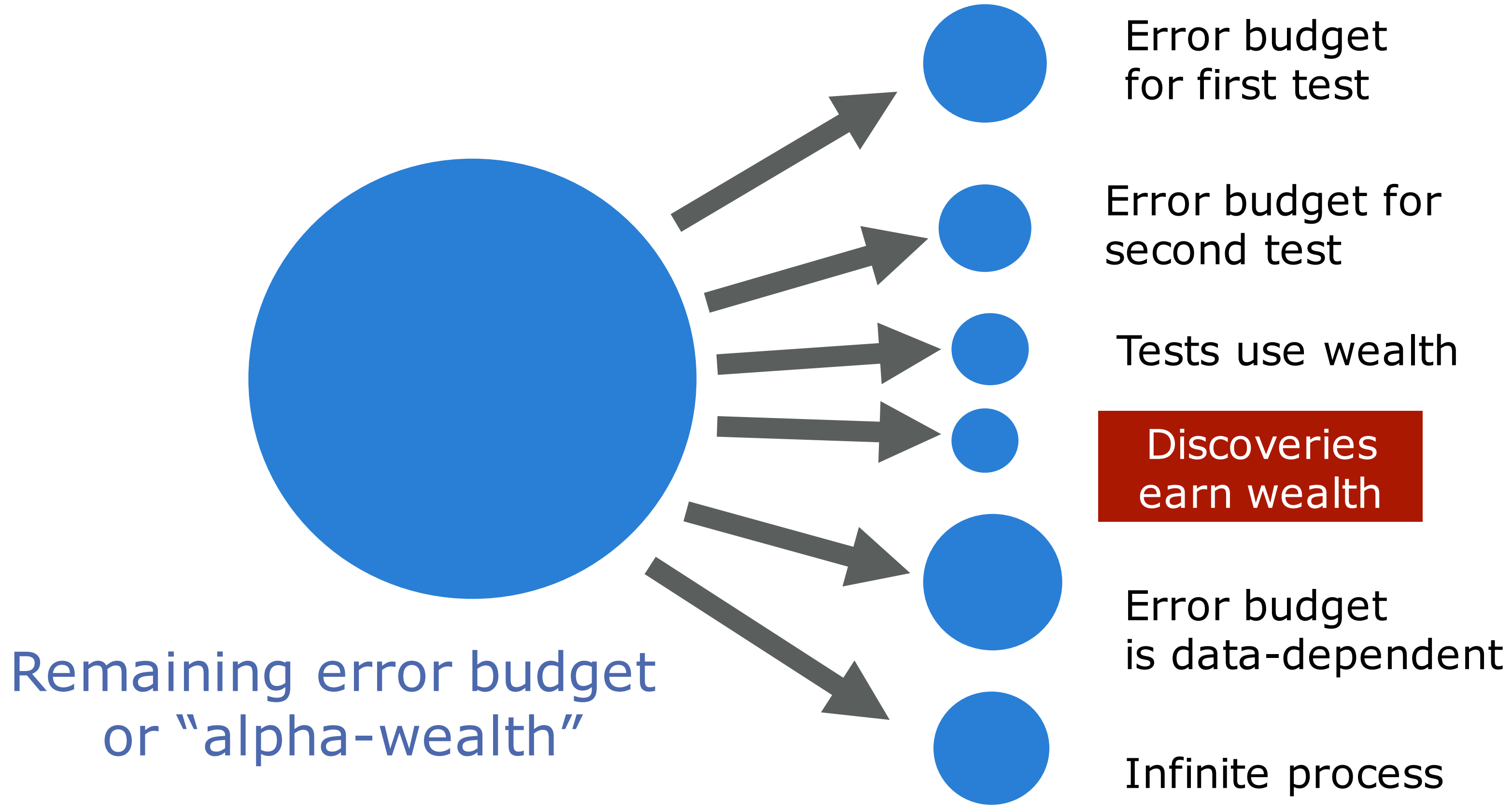
Discoveries earn wealth

Remaining error budget or "alpha-wealth"

Online FDR control : high-level picture



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

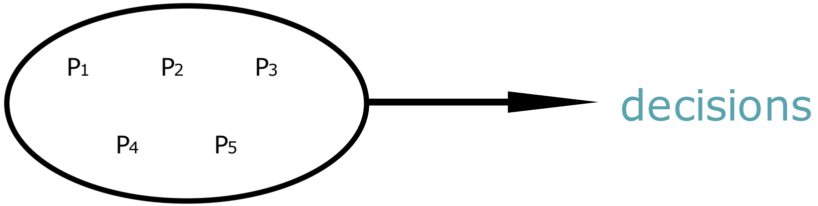
- instead, a possibly infinite sequence of tests (i.e. p-values) arrives *sequentially*

- 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*

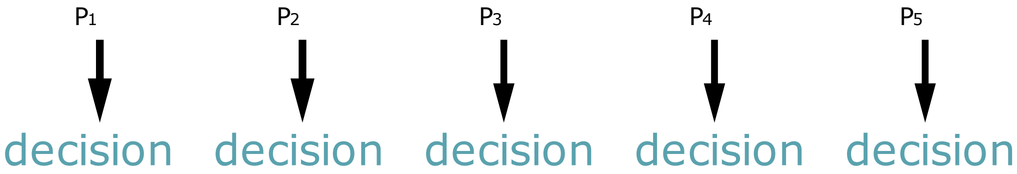
- motivating examples: A/B testing, large-scale clinical trials…

Online vs offline FDR control

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

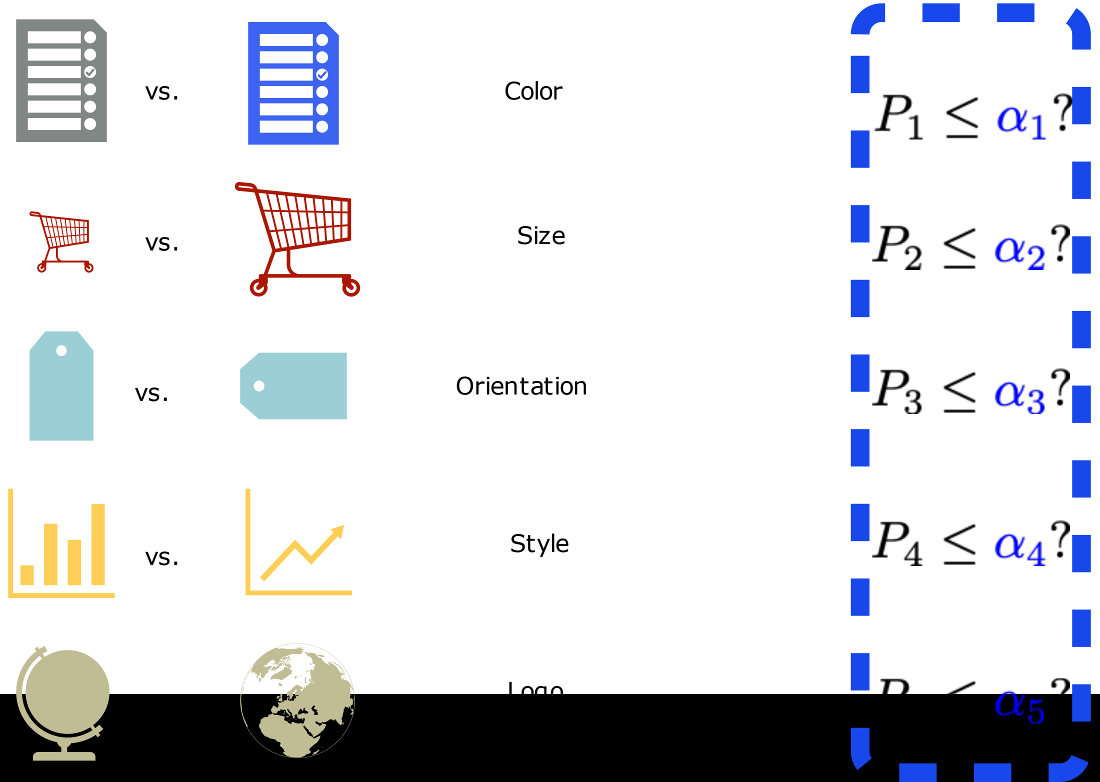


- online FDR procedures make decisions one at a time



Example: A/B testing

- online FDR algorithms pick significance level α_t adaptively



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

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

$$= \sum_{i \le t, i \text{ null}} \mathbb{E}[\alpha_i] \le \mathbb{E}[\sum_{i \le t} \alpha_i] \le \alpha \mathbb{E}[\sum_{i \le t} 1\{P_i \le \alpha_i\}]$$

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

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…