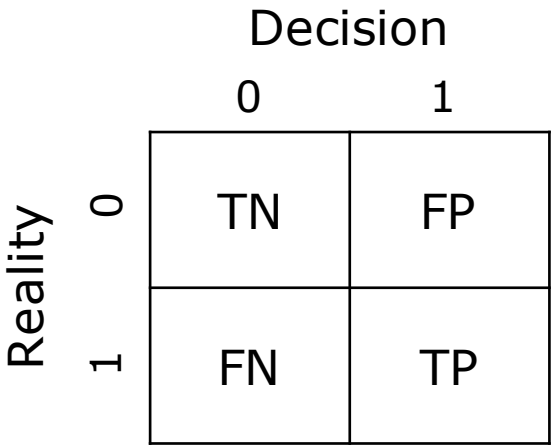
DS 102: Data, Inference, and Decisions

Lecture 3

Michael Jordan
University of California, Berkeley

The Basic Two-by-Two Table



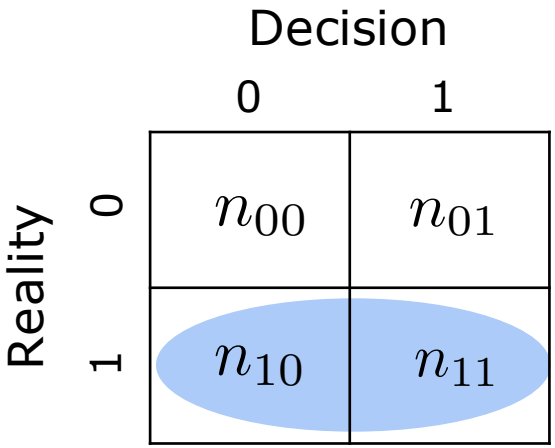
TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

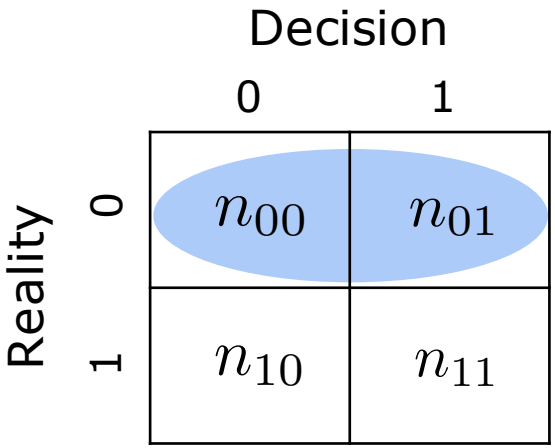
Some Row-Wise Rates



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

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

Some Row-Wise Rates



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

aka, "true negative rate"
or "selectivity"

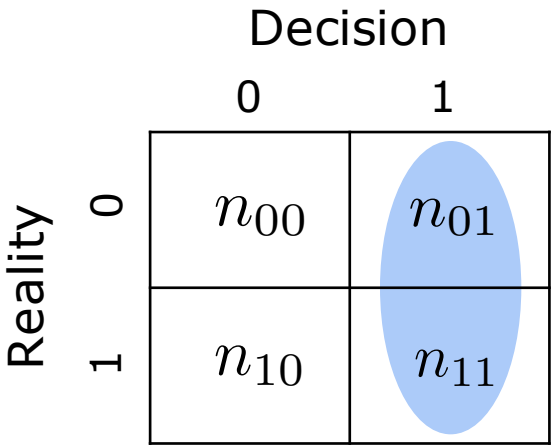
The Bayesian Posterior

- 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”)

Let's Return to our Column-Wise Rates



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

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

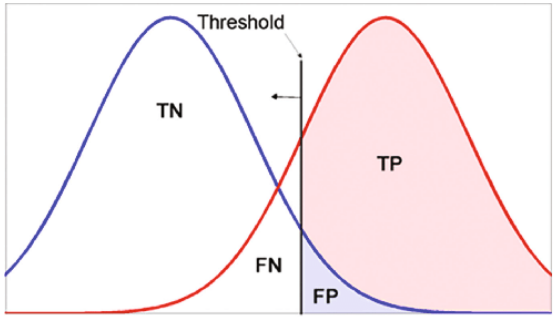
- The row-focused Neyman-Pearson paradigm turns the problem into a constrained optimization problem

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
  - i.e., make the specificity be greater than 0.95
- And to maximize the true-positive probability subject to that constraint
  - i.e., maximize the sensitivity subject to the constraint on the specificity

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 sensitivity) 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:

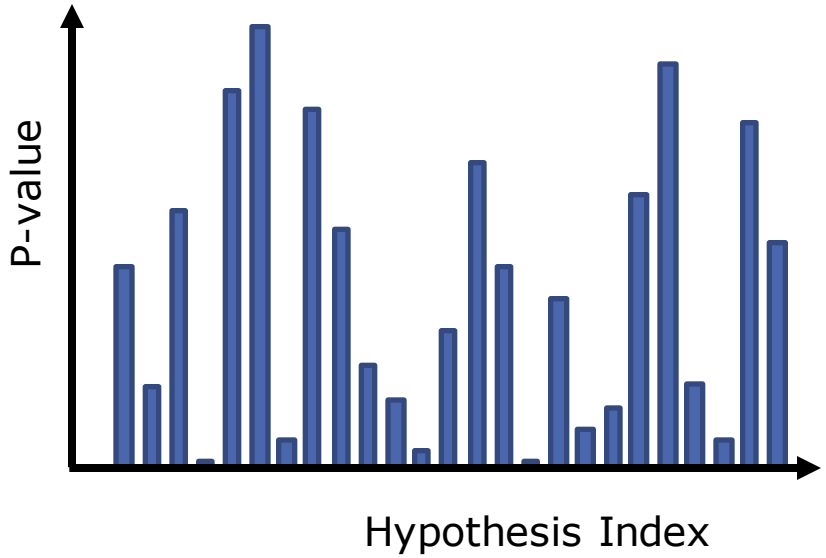
$$T_i = \begin{cases} 1, & \text{if } P_i \leq \alpha_i \\ 0, & \text{otherwise} \end{cases}$$

- 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

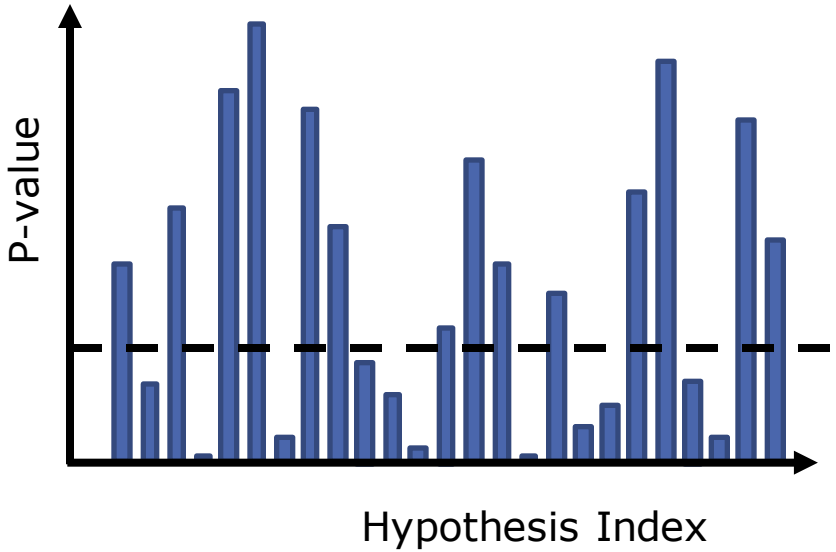
- Let's now consider multiple tests, in particular repeated tests of the same hypothesis
- The row-focused Neyman-Pearson paradigm provides a priori control over errors made in those cases in which the null hypothesis is true
  - this isn't very natural when the hypotheses are "cases" which arise randomly according to their prevalence
  - it also makes little sense when we're testing a bag of different hypothesis (cf., A/B testing)
- Our example with 10,000 A/B tests showed that Neyman-Pearson tests do not control column-wise quantities such as the FDP ☹

Example



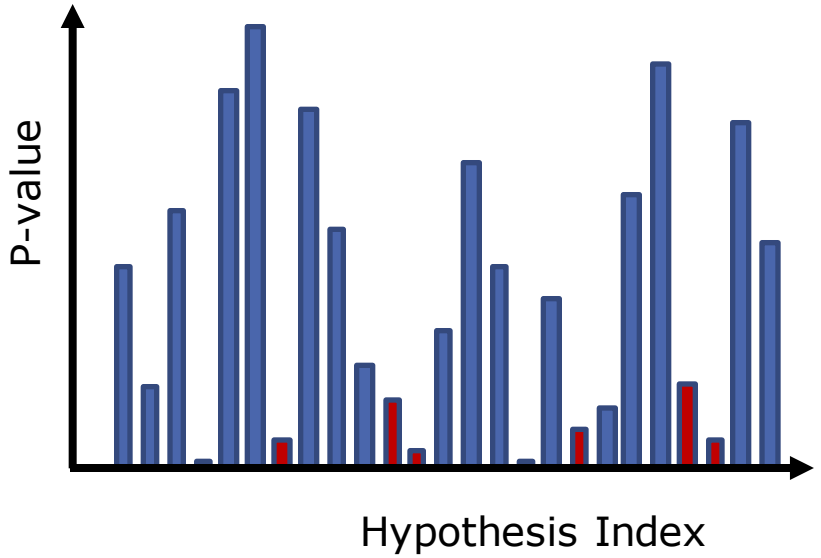
- Suppose that we obtain p-values from 25 experiments

Naïve Multiple Decision-Making



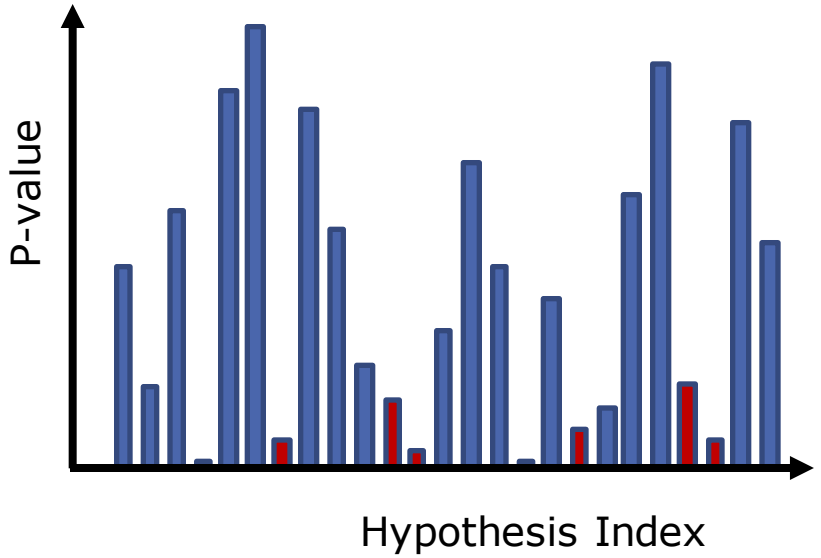
- Suppose that we simply reject each test independently if its p-value is smaller than some threshold

Naïve Multiple Decision-Making



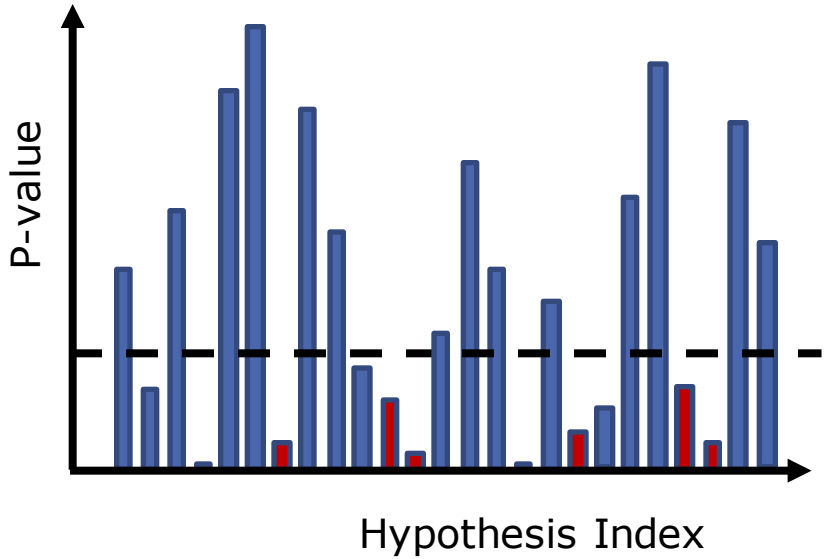
- Suppose that we simply reject each test independently if its p-value is smaller than some thresholding

Naïve Multiple Decision-Making



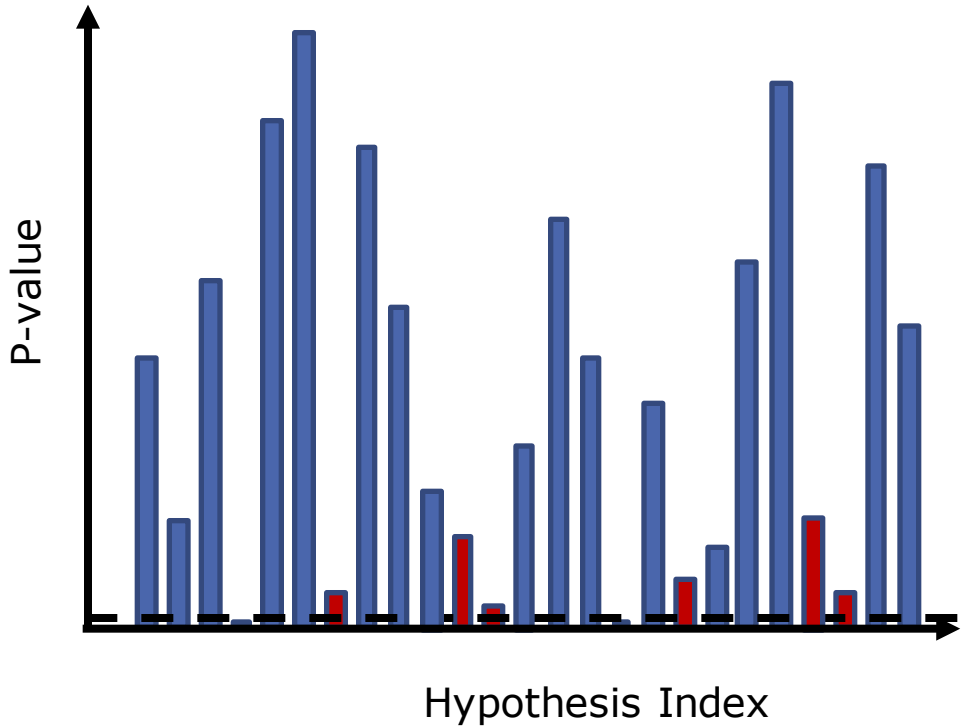
- An oracle knows the truth: that the blue-shaded bars correspond to nulls (Reality = 0) and the red-shaded bars to alternatives (Reality = 1)

Naïve Multiple Decision-Making



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

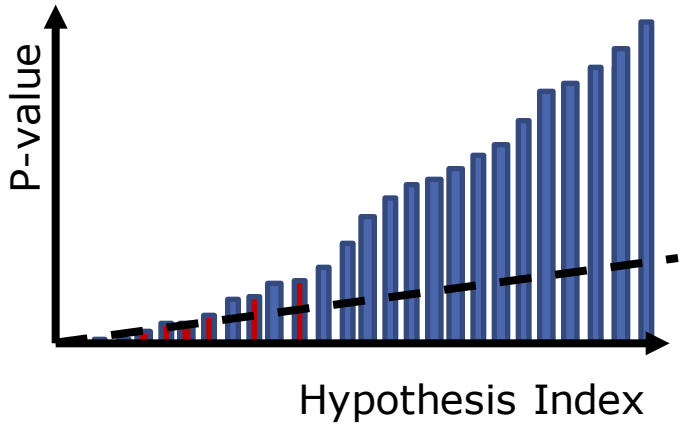
Bonferroni



- 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!



Recall Our Bayesian Calculation

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

$$= \frac{P(\text{false positive})\pi_0}{P(D = 1)}$$

Recall Our Bayesian Calculation

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

$$= \frac{P(\text{false positive})\pi_0}{P(D = 1)}$$

- We can (quite reasonably) upper bound π_0 with 1, and upper bound $P(\text{false positive})$ using Neyman-Pearson thinking
- 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:

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

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

A Bayesian Calculation

- Using the law of total probability, we have:

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

  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! In the multiple hypothesis testing problem 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

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 -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

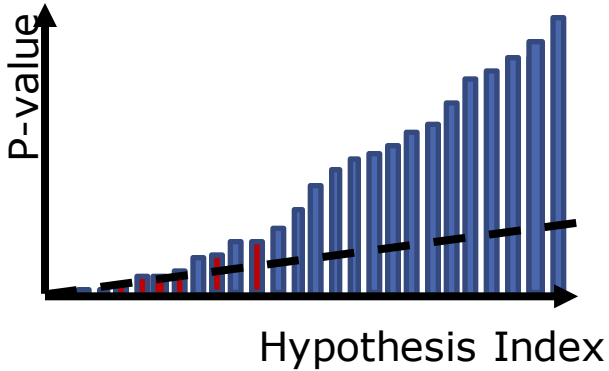
- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain P-values P_i , and sort them from smallest to largest, denoting the sorted P-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

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

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:

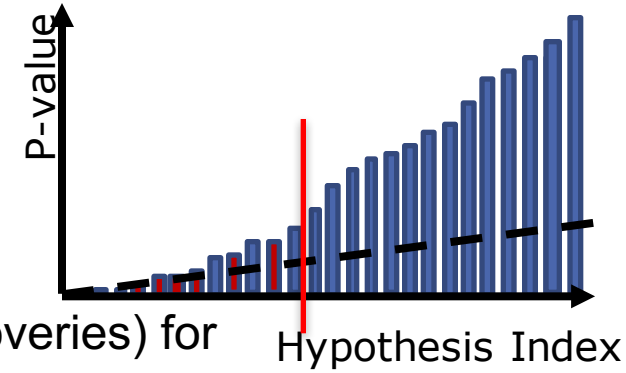
$$P_{(k)} \leq \frac{k}{m}\alpha$$



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:

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

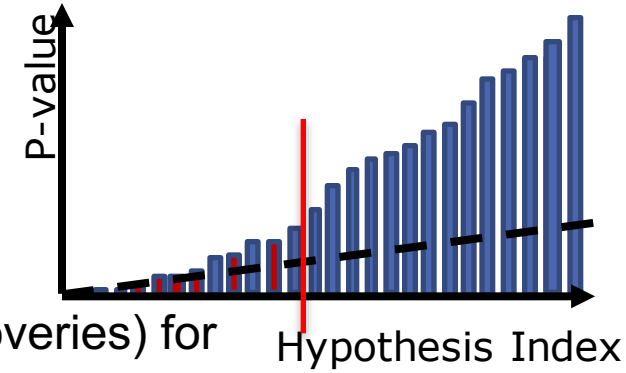


- Reject the null hypothesis (i.e., declare discoveries) for all hypotheses H_i such that $i \leq k$

Controlling the FDR

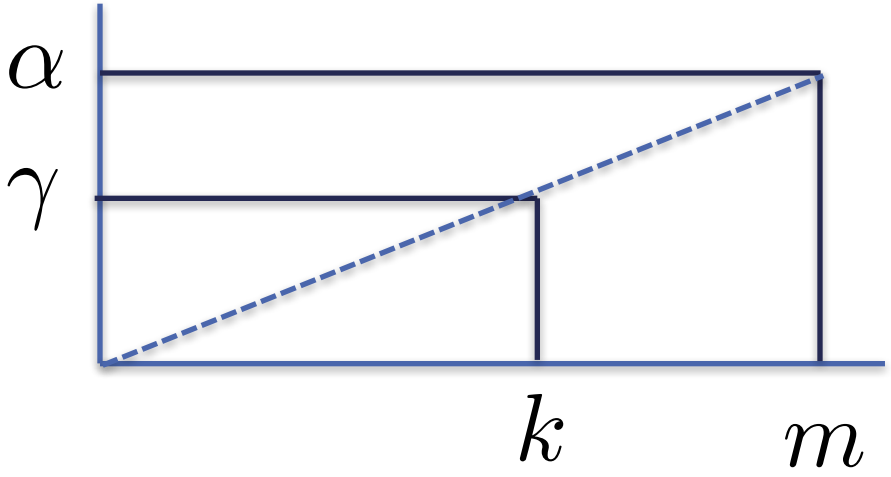
- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain P-values P_i , and sort them from smallest to largest, denoting the sorted P-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

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



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

Heuristic Argument



- Letting m_0 denote the number of true nulls, we have (very roughly):

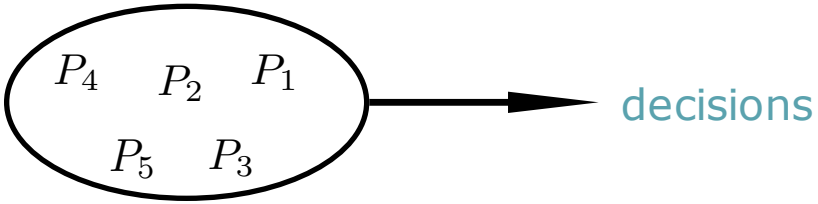
$$\text{FDR} \leq \frac{\gamma m_0}{k} = \frac{\frac{\alpha k}{m} m_0}{k} = \frac{\alpha m_0}{m} \leq \alpha$$

The Online Problem

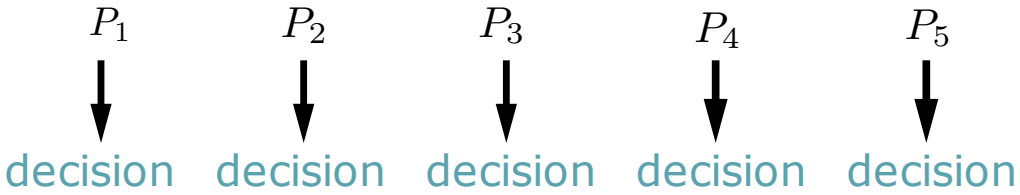
- Classical statistics, and also the Benjamini & Hochberg framework, focused on a batch setting in which all data has already been collected
- E.g., for Benjamini & Hochberg, you need all of the p-values before you can get started
- Is is possible to consider methods that make sequences of decisions,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

Offline vs. Online FDR Control

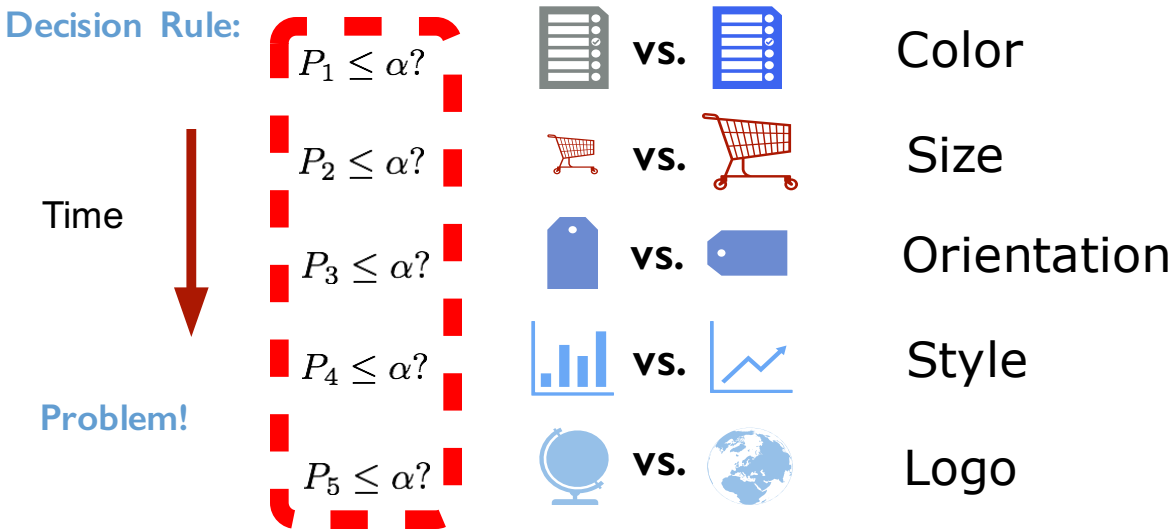
- Classical FDR procedures (such as BH) which make all decisions simultaneously are called “offline”



- “Online” FDR procedures make decisions one at a time



Example: Many Enterprises Run Thousands of So-Called A/B Tests Each Day



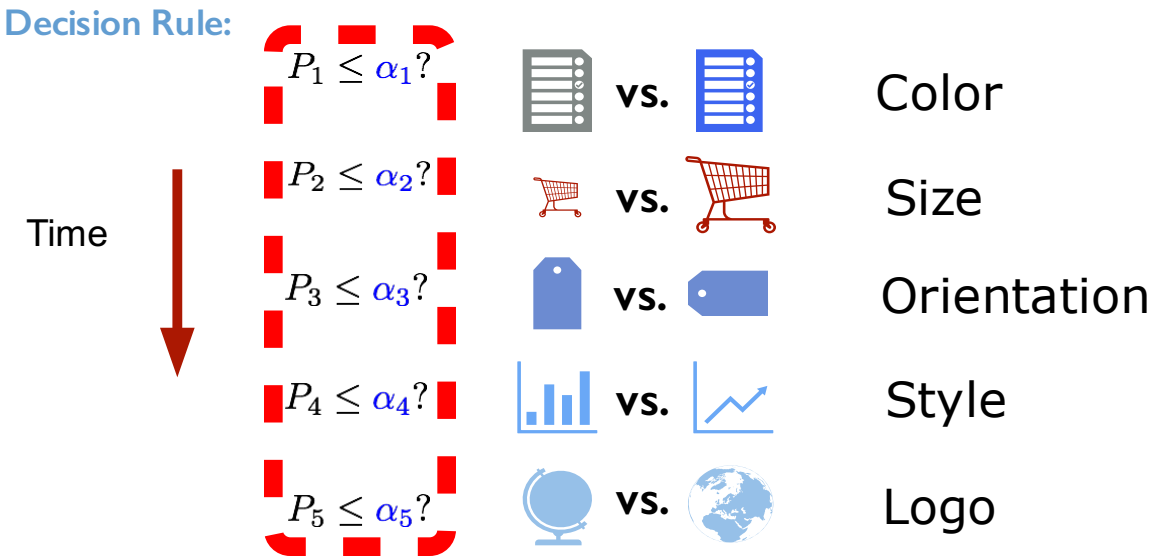
Challenges

- It’s not clear how to do change batch procedures such as Benjamini-Hochberg procedure to be online

Challenges

- It's not clear how to do change batch procedures such as Benjamini-Hochberg procedure to be online
- We might retreat to Bonferroni, which would allow us to set α to $0.05/n$ and thereby have a FWER of 0.05 after n tests
  - but what do we do on the $(n+1)th$ test?
  - we eventually can't do any more tests
  - we've used up our "alpha wealth"

A More General Approach: Time-Varying Alpha



More Challenges

- We want to keep going for an arbitrary amount of time, so we need $\sum_{t=1}^{\infty} \alpha_t = 1$, and $\sum_{t=1}^{T} \alpha_t < 1$ for any fixed T
- An example: $\alpha_t = 2^{-t}$
- But now we have less and less power to make discoveries over time, and eventually we may as well quit
- Is there any way out of this dilemma?

A Glimmer of Hope

- Recall that the FDP is a ratio of two counts
- We can make a ratio small in one of two ways:
  - make the numerator small
  - make the denominator big

A Glimmer of Hope

- Recall that the FDP is a ratio of two counts
- We can make a ratio small in one of two ways:
  - make the numerator small
  - make the denominator big
- The numerator has the false-positive rate in it, and so we're back to the same problem of controlling sums of α_i values

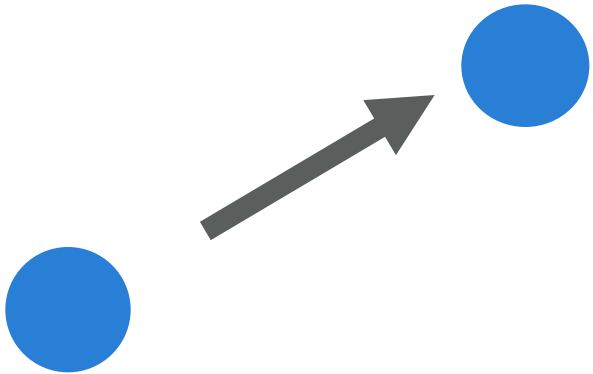
A Glimmer of Hope

- Recall that the FDP is a ratio of two counts
- We can make a ratio small in one of two ways:
  - make the numerator small
  - make the denominator big
- The numerator has the false-positive rate in it, and so in terms of controlling the numerator we're back to the same problem of controlling sums of α_i values
- The denominator can be made large by making lots of discoveries

A Glimmer of Hope

- Recall that the FDP is a ratio of two counts
- We can make a ratio small in one of two ways:
  - make the numerator small
  - make the denominator big
- The numerator has the false-positive rate in it, and so in terms of controlling the numerator we're back to the same problem of controlling sums of α_i values
- The denominator can be made large by making lots of discoveries
- Perhaps we can earn a bit of alpha whenever we make a discovery, to be invested and used for false discoveries later

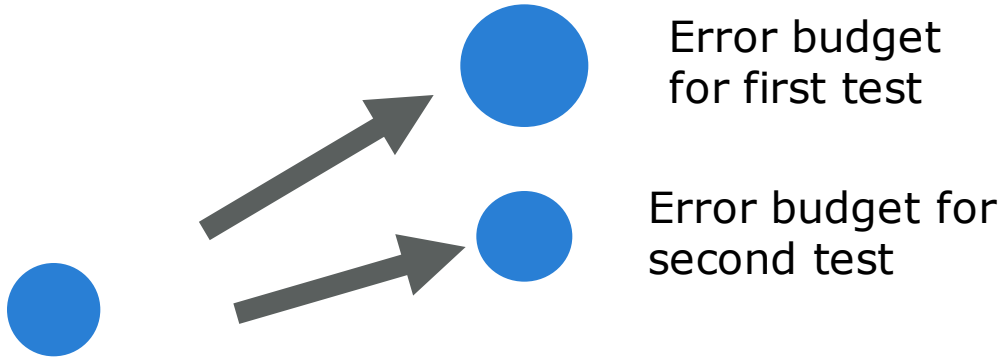
Online FDR Control : High-Level Picture



Error budget
for first test

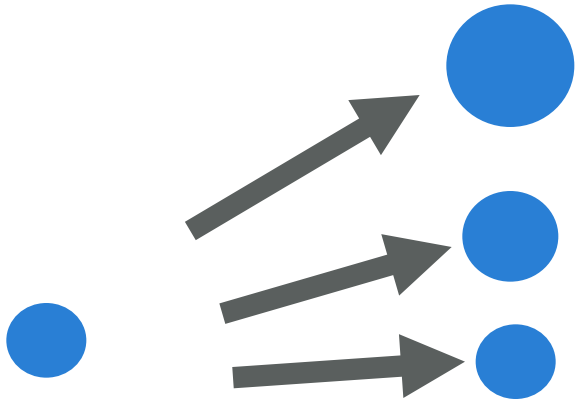
Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture



Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture



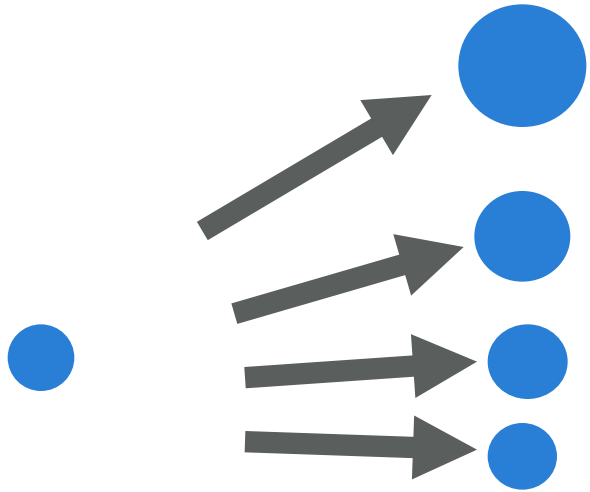
Error budget for first test

Error budget for second test

Tests use wealth

Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



Error budget for first test

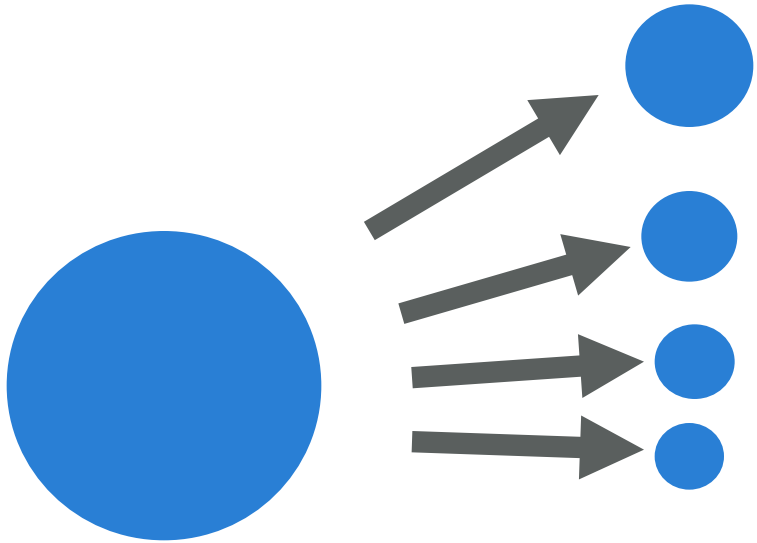
Error budget for second test

Tests use wealth

Discoveries earn wealth

Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



Error budget for first test

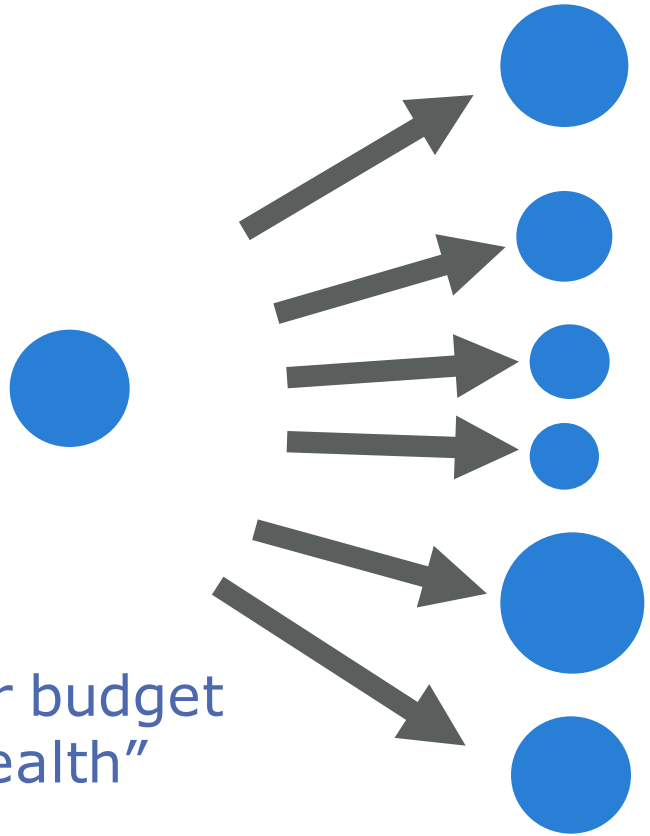
Error budget for second test

Tests use wealth

Discoveries earn wealth

Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



Error budget for first test

Error budget for second test

Tests use wealth

Discoveries earn wealth

Error budget is data-dependent

Infinite process

Remaining error budget or "alpha-wealth"

Online FDR Algorithms

- The first online FDR algorithm was known as “alpha investing” and is due to Foster and Stine (2008)
- A more recent (and simpler) online FDR algorithm is due to Javanmard and Montanari, and is called “LORD”
- The basic idea is to renew the alpha wealth every time a discovery (i.e., rejection) is made, and decay that wealth forward in time
- The current wealth is the sum of all of the decayed values of the past wealth increments

Algorithm 1 The LORD Procedure

input: FDR level α, non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} \gamma_t = 1$, initial wealth $W_0 \leq \alpha$

Set $\alpha_1 = \gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

p-value P_t arrives

if $P_t \leq \alpha_t$, reject P_t

$$\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-\tau_1}(\alpha - W_0)\mathbf{1}\{\tau_1 < t\} + \alpha \sum_{j=1}^{\infty} \gamma_{t+1-\tau_j} \mathbf{1}\{\tau_j < t\},$$

where τ_j is time of j-th rejection $\tau_j = \min\{k : \sum_{l=1}^{k} \mathbf{1}\{P_l \leq \alpha_l\} = j\}$

end

A Stripped-Down Version of LORD

- Only consider the most recent rejection
- This renews the wealth, which further decays
- Why does such an approach provide control over the FDR?

t

A Stripped-Down Version of LORD

- Only consider the most recent rejection
- This renews the wealth, which further decays
- Why does such an approach provide control over the FDR?

- Return to the Bayesian perspective, and consider the following estimate (an upper bound) of the FDP:

$$\widehat{\mathrm{FDP}}(t) := \frac{\sum_{i=1}^{t} \alpha_i}{\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}}$$

- The denominator is just the number of rejections until time t, and the numerator is an upper bound on the probability of one or more false-positive errors

Analysis

- Break up the sum $\sum_{i=1}^{t} \alpha_i$ into “episodes” between the rejections

Analysis

- Break up the sum $\sum_{i=1}^{t} \alpha_i$ into “episodes” between the rejections
- In each episode, the sum is upper bounded by $\alpha \sum_{i=1}^{t'} \gamma_{i+1-\tau}$, by the definition of (simplified) LORD, where t' is the episode length and τ is the time of the most recent rejection

Analysis

- Break up the sum $\sum_{i=1}^{t} \alpha_i$ into “episodes” between the rejections
- In each episode, the sum is upper bounded by $\alpha \sum_{i=1}^{t'} \gamma_{i+1-\tau}$, by the definition of (simplified) LORD, where t' is the episode length and τ is the time of the most recent rejection
- This sum is less than α by the definition of the $\{\gamma_i\}$ sequence

Analysis

- Break up the sum $\sum_{i=1}^{t} \alpha_i$ into “episodes” between the rejections
- In each episode, the sum is upper bounded by $\alpha \sum_{i=1}^{t'} \gamma_{i+1-\tau}$, by the definition of (simplified) LORD, where t' is the episode length and τ is the time of the most recent rejection
- This sum is less than α by the definition of the $\{\gamma_i\}$ sequence
- The number of episodes is: $\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}$

Analysis

- Break up the sum $\sum_{i=1}^{t} \alpha_i$ into “episodes” between the rejections
- In each episode, the sum is upper bounded by $\alpha \sum_{i=1}^{t'} \gamma_{i+1-\tau}$, by the definition of (simplified) LORD, where t' is the episode length and τ is the time of the most recent rejection
- This sum is less than α by the definition of the $\{\gamma_i\}$ sequence
- The number of episodes is: $\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}$
- And so we conclude:

$$\widehat{\mathrm{FDP}}(t) := \frac{\sum_{i=1}^{t} \alpha_i}{\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}} \leq \alpha$$

And Now We Connect to the FDR

- We can write the FDR in the following nice form:

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

And Now We Connect to the FDR

- We can write the FDR in the following nice form:

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

- To simplify our derivation, we will make an approximation (the “modified 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 Obtain an Actual Proof

- We make the mFDR approximation:

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

And We Obtain an Actual Proof

- We make the mFDR approximation:

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

and then compute:

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

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

where the last line uses:

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

And We Obtain an Actual Proof

- We make the mFDR approximation:

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

and then compute:

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

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

where the last line uses:

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

- This establishes:

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