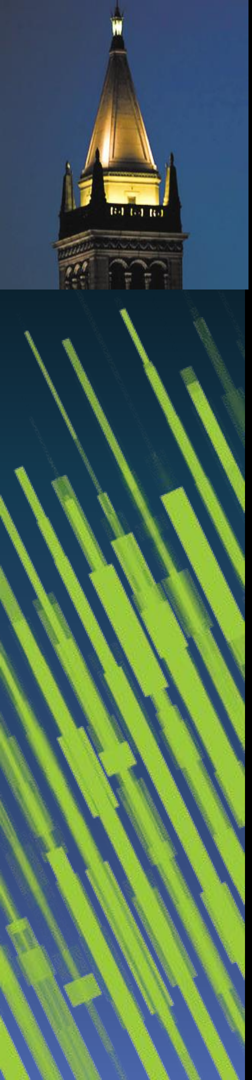


DS 102: Data, Inference, and Decisions

Lecture 5

Michael Jordan

University of California, Berkeley

Some Column-Wise Rates

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

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

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!

P-Values

- Consider a point-null hypothesis, $\theta = 0$, and $\mathbb{P}$ denote that null
- 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:

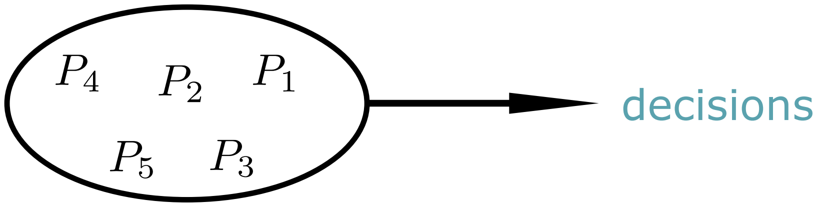
$$T_i = \begin{cases} 1, & \text{if } P_i \leq \alpha_i \\ 0, & \text{otherwise} \end{cases}$$

The Online Problem

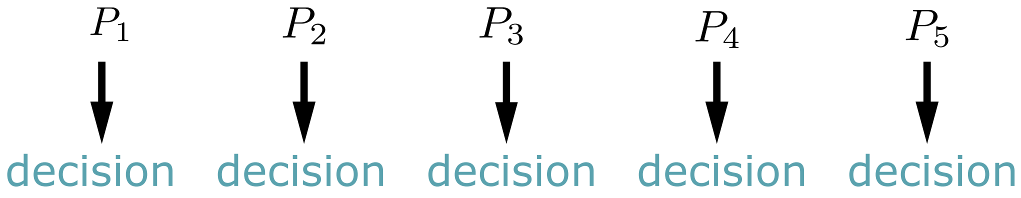
- Classical statistics, and also the Benjamini & Hochberg algorithm focused on a batch setting in which all data has already been collected
- E.g., for Benjamini & Hochberg, you need all of the p-values before you can get started
- Is is possible to consider methods that make sequences of decisions,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

Online vs Offline FDR Control

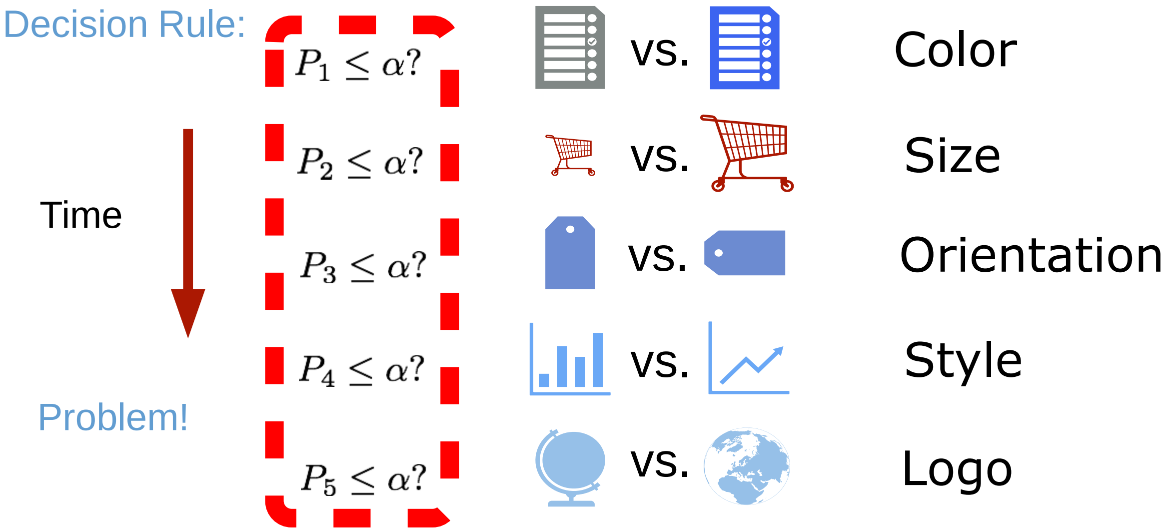
- Classical FDR procedures (such as BH) which make all decisions simultaneously are called “offline”



- “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 aws Benjamini-Hochberg procedure to be online

Challenges

- It's not clear how to do change batch procedures such aws 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

Decision Rule:

Time ↓

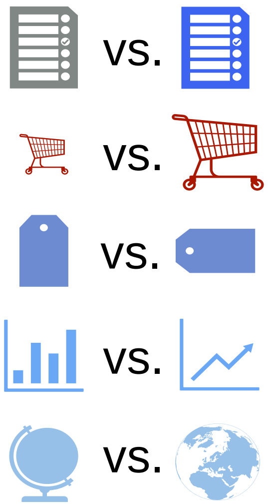
$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

$P_3 \leq \alpha_3$?

$P_4 \leq \alpha_4$?

$P_5 \leq \alpha_5$?



Color

Size

Orientation

Style

Logo

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

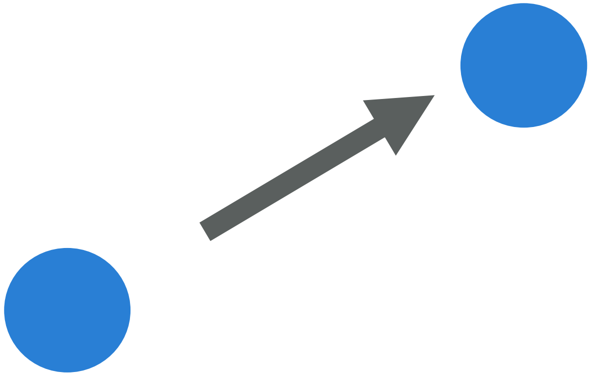
A Glimmer of Hope

- 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 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 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

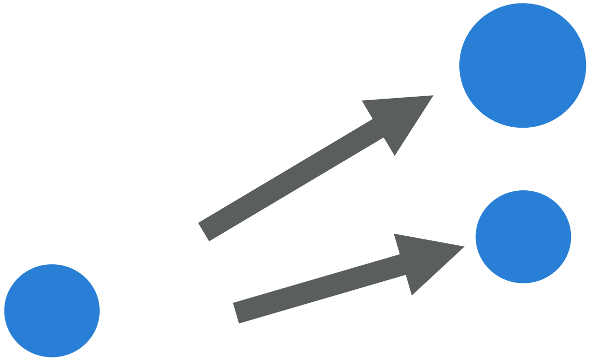
Online FDR Control : High-Level Picture



Error budget
for first test

Remaining error budget
or "alpha-wealth"

Online FDR Control : High-Level Picture

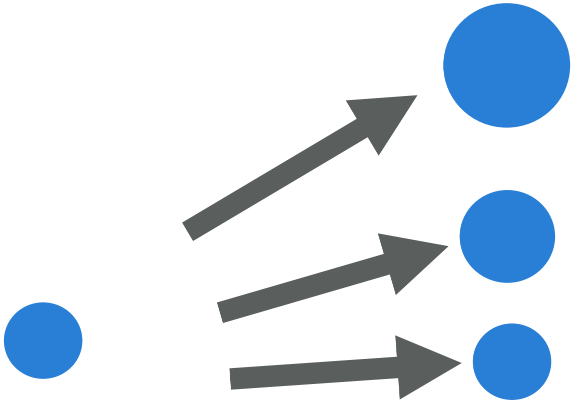


Error budget
for first test

Error budget for
second test

Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture



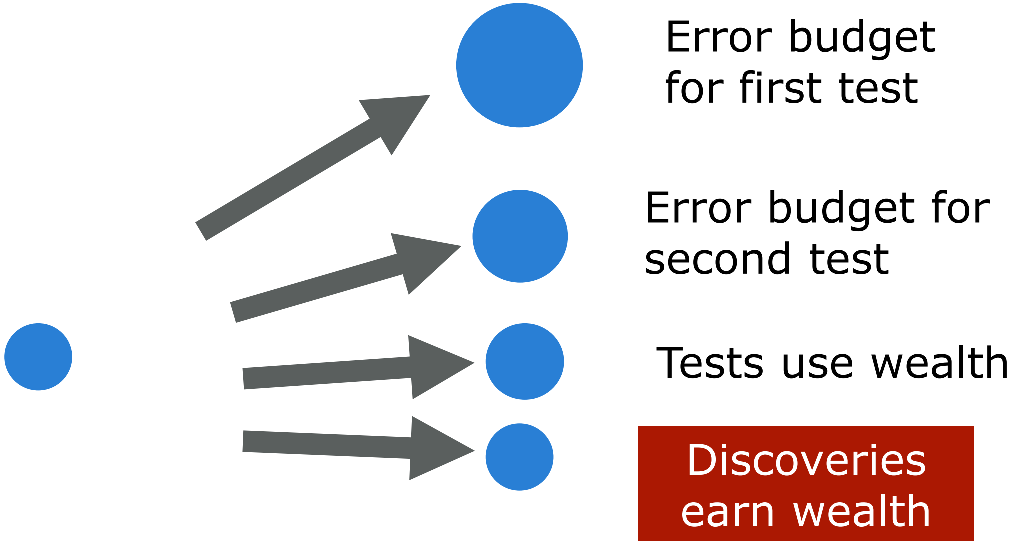
Error budget
for first test

Error budget for
second test

Tests use wealth

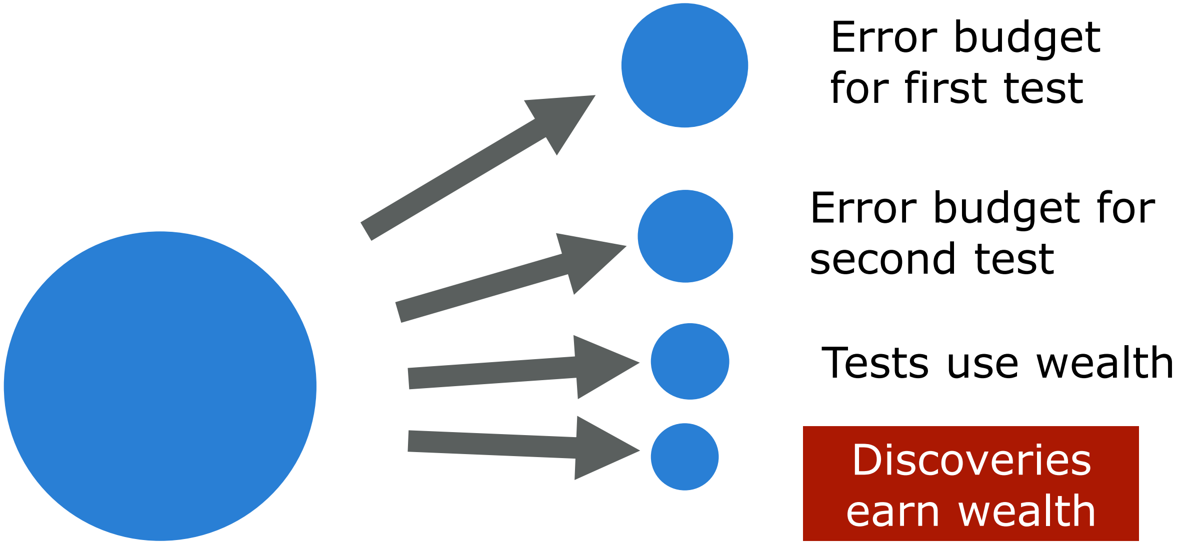
Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture



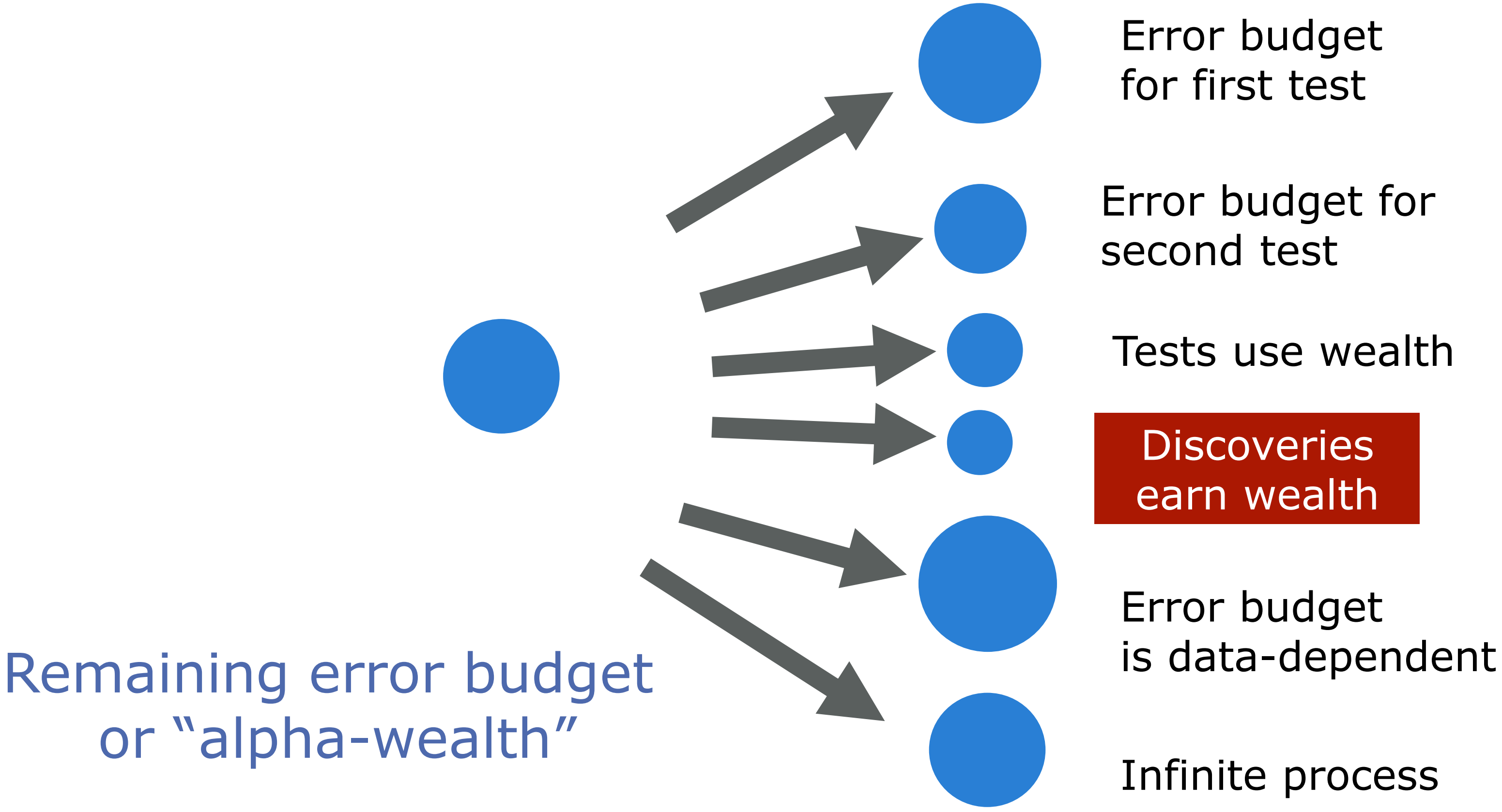
Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture



Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



Online FDR 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 assign α_t in a way that ensures

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

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 Type I error probabilities

A Stripped-Down Version of LORD

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

A Stripped-Down Version of LORD

- 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

A Stripped-Down Version of LORD

- 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

A Stripped-Down Version of LORD

- 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: $\sum_{i=1}^{t} 1\{P_i \leq \alpha_i\}$

A Stripped-Down Version of LORD

- 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 make an 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:

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

LORD's Control of mFDR (Modified FDR)

- We make an 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:

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

Further Perspective on Hypothesis Testing

- We've focused on providing guarantees that a test, or a set of tests, perform well
- Can you think of situations where one would like to guarantee the opposite---that a test cannot perform well?

Privacy and Data Analysis

- Individuals are not generally willing to allow their personal data to be used without control on how it will be used and now much privacy loss they will incur
- “Privacy loss” can be quantified via differential privacy
- We want to trade privacy loss against the value we obtain from data analysis
- The question becomes that of quantifying such value and juxtaposing it with privacy loss
- We’ll have an entire section on privacy later in the course, but let’s make some initial comments here

Privacy

query

↓

database

Privacy

query

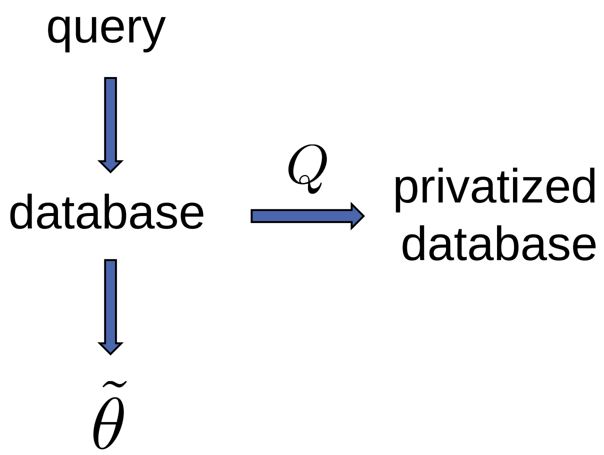
↓

database

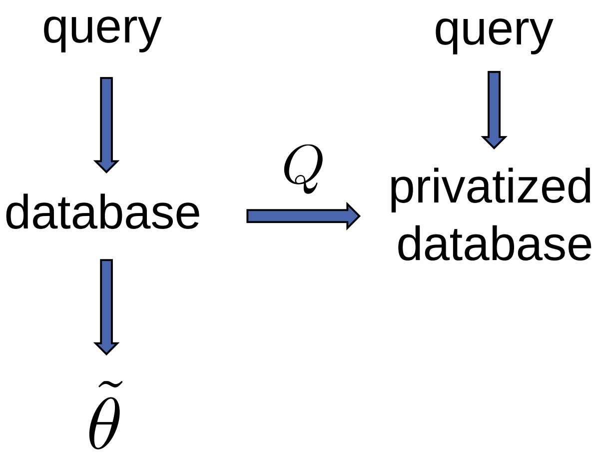
↓

$\tilde{\theta}$

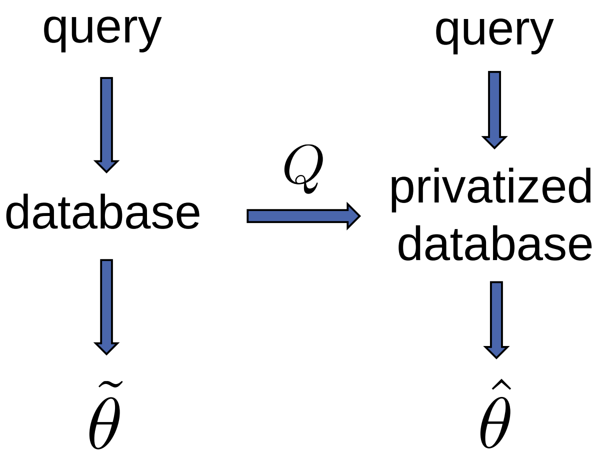
Privacy



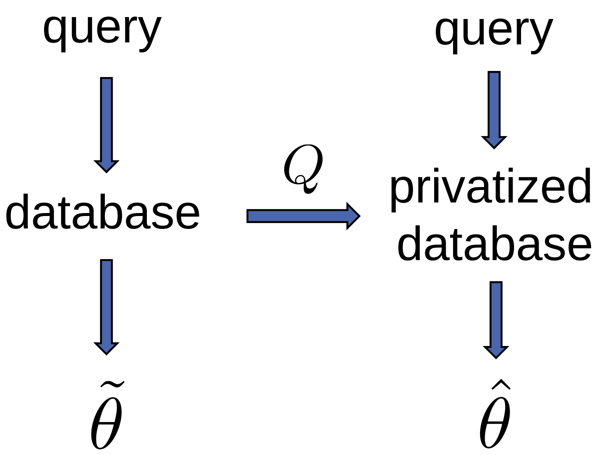
Privacy



Privacy


query
query
database
Q
privatized database
$\tilde{\theta}$
$\hat{\theta}$

Privacy



Q is a “noisy channel”

Classical problem in differential privacy: show that $\hat{\theta}$ and $\tilde{\theta}$ are close under constraints on Q

Inference

query

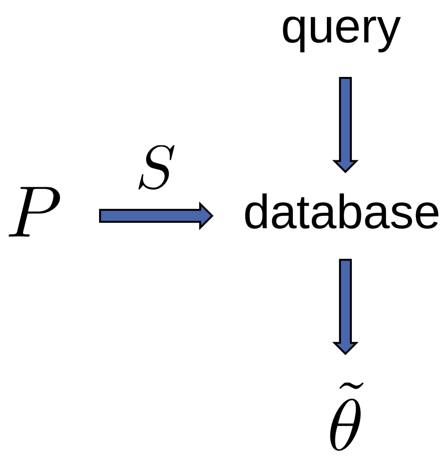
↓

database

↓

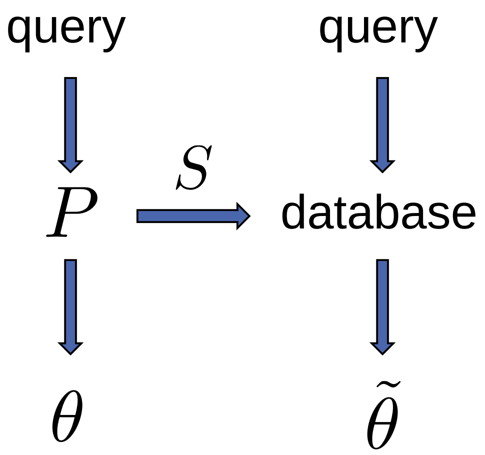
$\tilde{\theta}$

Inference

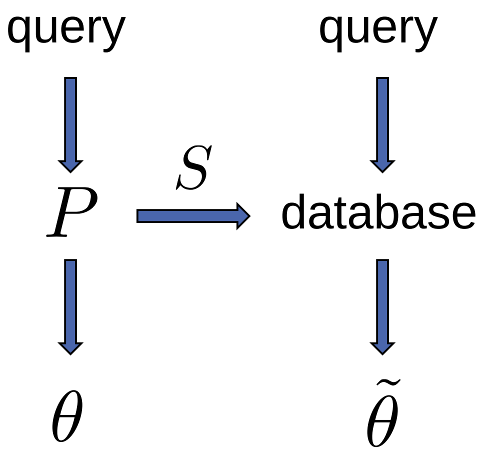


S is the sampling process

Inference

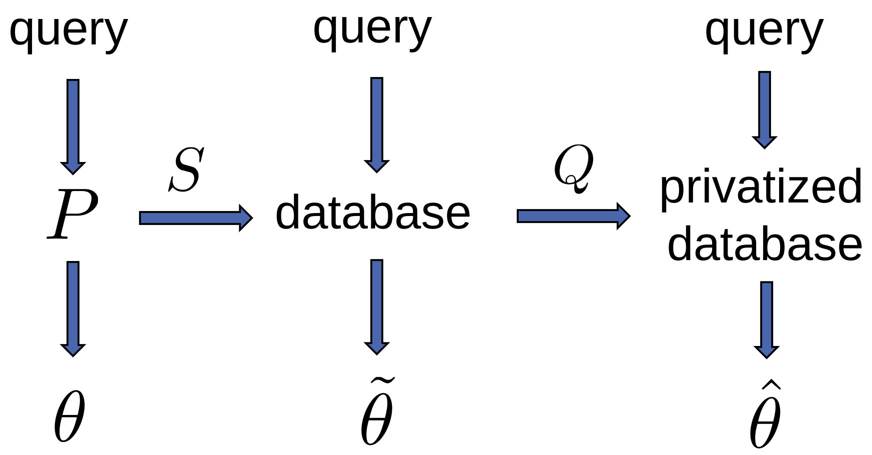


Inference



Classical problem in statistical theory: show that $\tilde{\theta}$ and θ are close under constraints on S

Privacy and Inference



The privacy-meets-inference problem: show that θ and $\hat{\theta}$ are close under constraints on Q and on S

Estimating the Null Distribution

- What if we don't have a well-specified null distribution in mind?
- In the classical single-hypothesis-testing paradigm, we are more or less stuck
- In the modern multiple-hypothesis-testing paradigm, if all of the null hypotheses are the same, then we have many draws from the null distribution at hand
  - we don't know which ones are null, but in the case of particular interest, when π_0 is large, we can assume that most of the data points corresponding to large p-values are from the null
  - and so we can estimate the null, using some form of density estimation

Relationship to Permutation Testing

- Remember permutation testing from Data 8?
- Permutation testing allows us to effectively obtain multiple draws from the null, and each draw has the same underlying probability, if we work in the appropriate conditional distribution
  - we don’t know that probability, but we know that it’s constant
  - which is enough to be able to specify a conditional null that’s easy to work with
  - let’s flesh this out…

A Data 102 Explanation of Permutation Testing

- In Data 8 we explained the permutation test intuitively
- Let's try to do a bit better now that we're at the Data 102 level
- First, we define the notion of exchangeability:
  - an infinite collection of random variables, $(X_1, X_2, \ldots)$, is exchangeable if for any n and any permutation π, the distribution of $(X_{\pi_1}, X_{\pi_2}, \ldots, X_{\pi_n})$ is the same as the distribution of $(X_1, X_2, \ldots, X_n)$
  - i.e., the order of the variables doesn't matter
  - this is a deeper concept than "independent and identically distributed"

Permutation Testing (Cont)

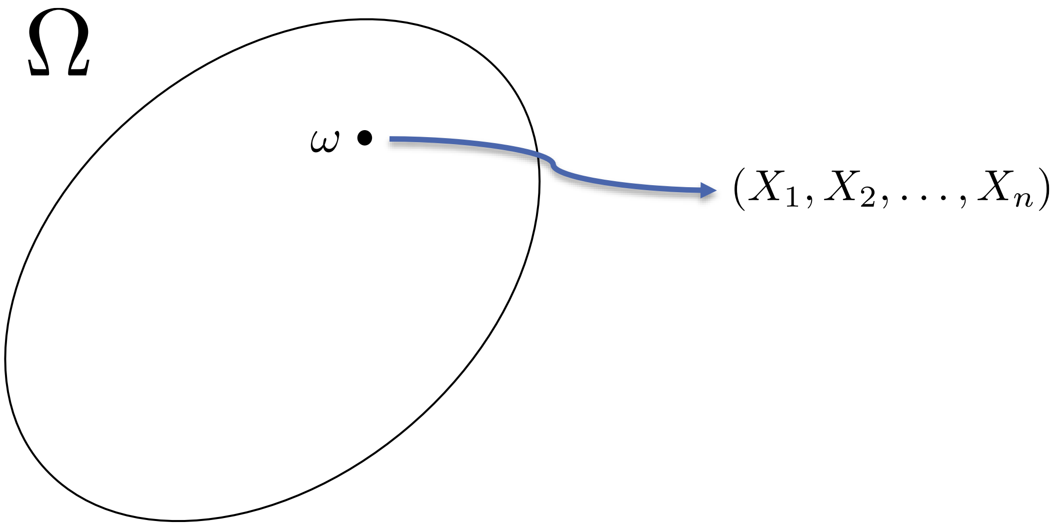
- Let $\tilde{X}$ denote the unordered set of variables $(X_1, X_2, \ldots, X_n)$, under an exchangeability assumption for the null
- Given a statistic T that is an indicator of a rejection region, consider the conditional expectation

$$\mathbb{E}(T \,|\, \tilde{X})$$

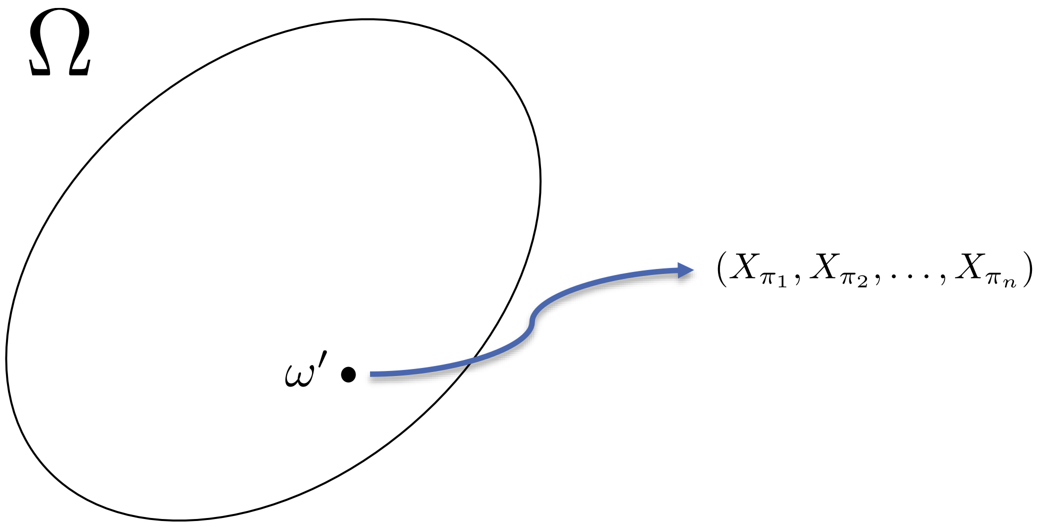
which is the probability of a Type I error

- Can we compute this conditional expectation? What is the distribution obtained by conditioning on $\tilde{X}$?

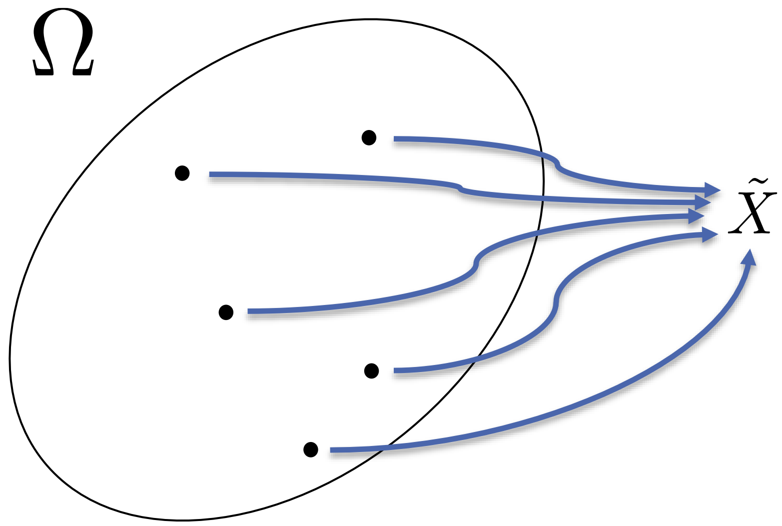
Permutation Testing (Cont)



Permutation Testing (Cont)



Permutation Testing (Cont)



Permutation Testing (Cont)

- What is the distribution obtained by conditioning on $\tilde{X}$?
- It's the uniform distribution on the orbit induced by exchangeability
  - we thereby avoid the complexities associated with knowing actual probabilities of points in the sample space
  - we can then compute $\mathbb{E}(T \,|\, \tilde{X})$ by enumerating (or, more realistically, uniformly sampling) the permutations
  - so it's easy to ensure $\mathbb{E}(T \,|\, \tilde{X}) \leq \alpha$ for the null (i.e., we get Type I error control, conditionally)
- And now the magic happens:

$$\mathbb{E}(T) = \mathbb{E}\left[\mathbb{E}(T \,|\, \tilde{X})\right] \leq \mathbb{E}[\alpha] = \alpha$$

Permutation Testing (Cont)

- What is the distribution obtained by conditioning on $\tilde{X}$?
- It's the uniform distribution on the orbit induced by exchangeability
  - we thereby avoid the complexities associated with knowing actual probabilities of points in the sample space
  - we can then compute $\mathbb{E}(T \mid \tilde{X})$ by enumerating (or, more realistically, uniformly sampling) the permutations
  - so it's easy to ensure $\mathbb{E}(T \mid \tilde{X}) \leq \alpha$ for the null (i.e., we get Type I error control, conditionally)