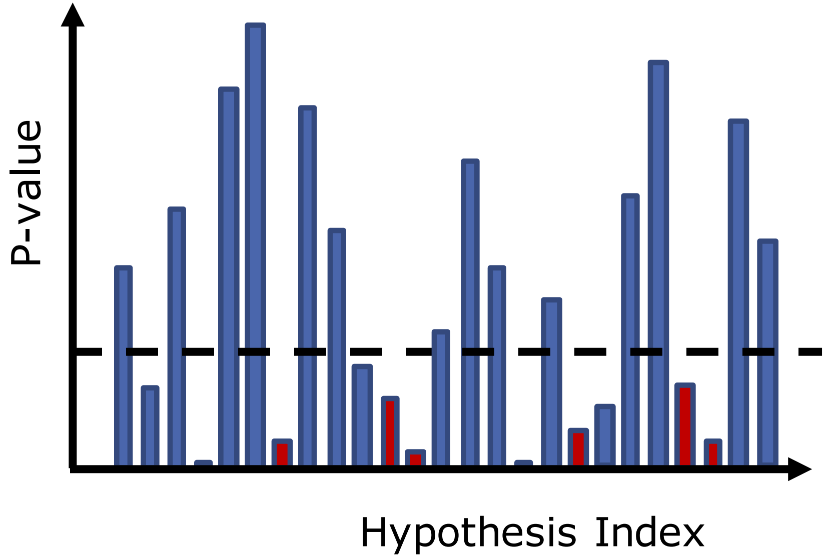
DS 102: Data, Inference, and Decisions

Lecture 4

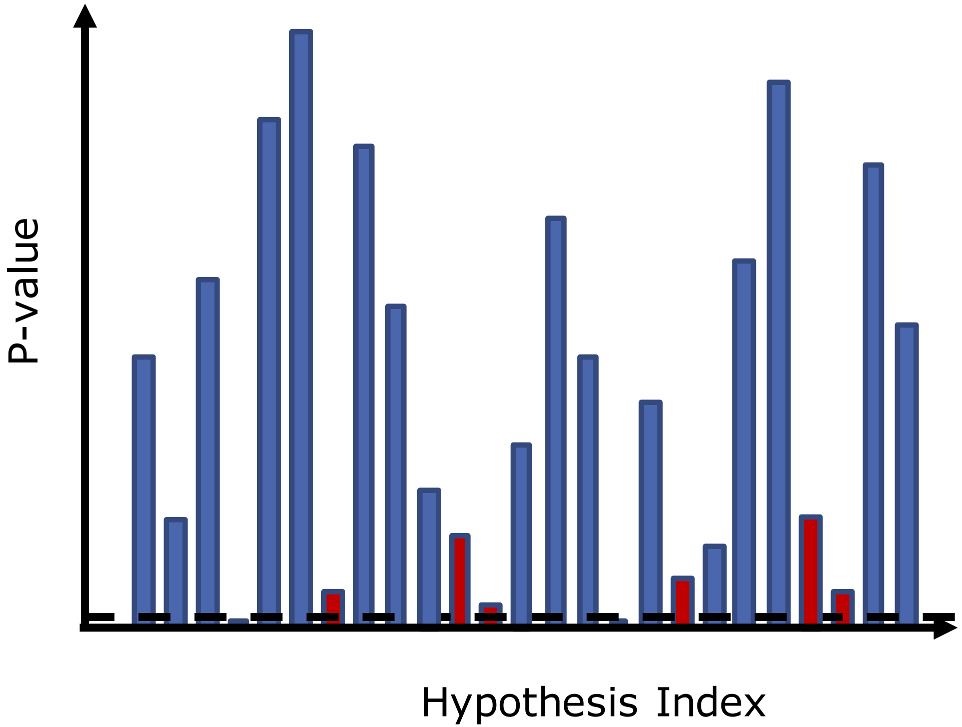
Michael Jordan
University of California, Berkeley

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!

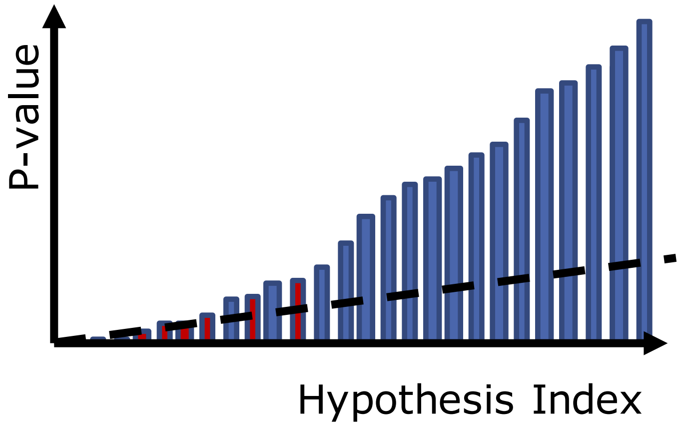
Bonferroni



- 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!



A Bayesian Derivation

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

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

- 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?
  - in the multiple hypothesis testing problem it's easy to estimate $P(D = 1)$ directly from the data!

Controlling the FDR

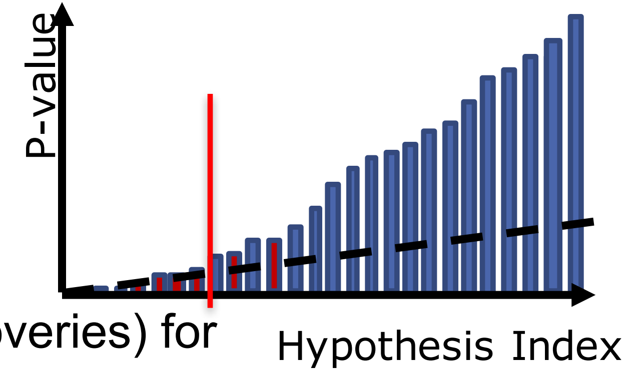
- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain P-values P_i , and sort them from smallest to largest, denoting the sorted P-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

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

Controlling the FDR

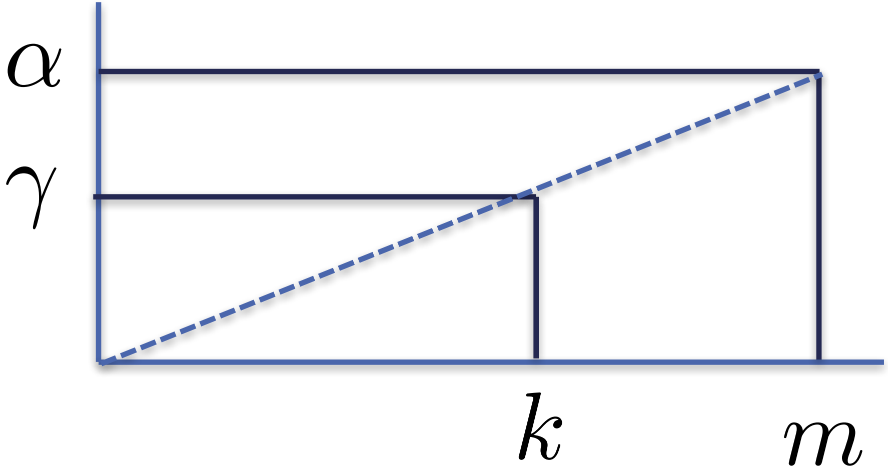
- Benjamini & Hochberg (1995) proposed an algorithm that does it
- Given m tests, obtain P-values P_i , and sort them from smallest to largest, denoting the sorted P-values as $P_{(k)}$
  - the small ones are the safest to reject
- Now, find the largest k such that:

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



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

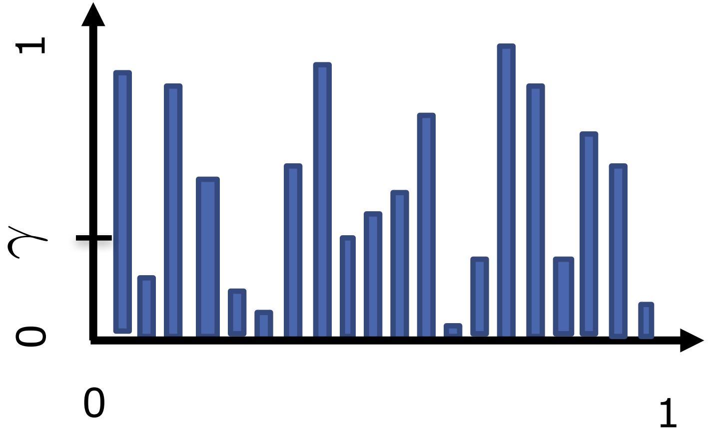
Heuristic Argument



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

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

Recall that P-Values are Uniform Under the Null

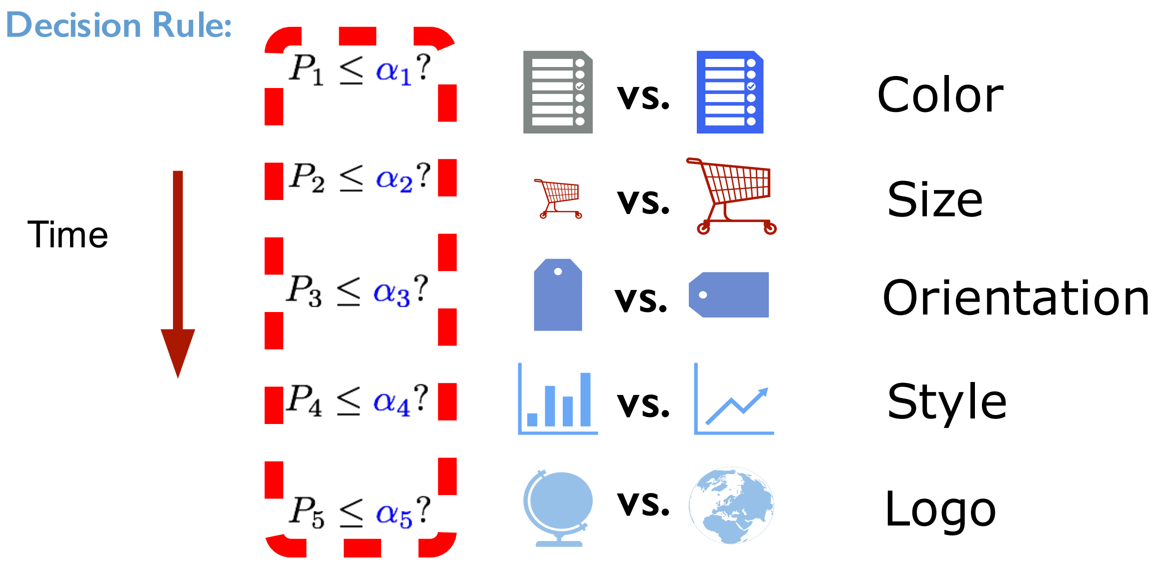


- If there are m_0 such P-values, then there are approximately γm_0 P-values in the interval $(0, \gamma)$, for any γ

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?

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
- 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

The Tower Property of Conditional Expectation

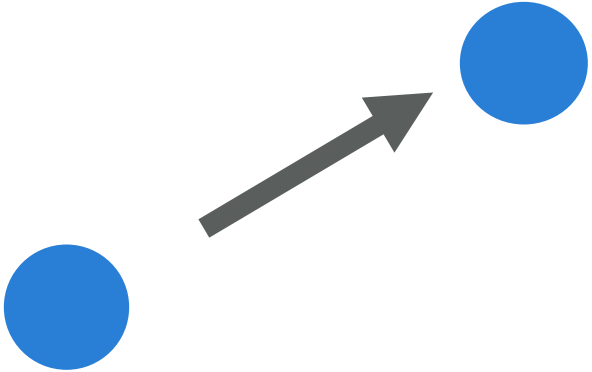
- A really important theorem from probability theory:

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right]$$

"the average of an average is an average"

- Note that $\mathbb{E}[X \mid Y]$ is a random variable
  - roughly, it averages over X in any region in the sample space where Y is a constant, yielding something like a "step function" over the sample space
  - and the outer expectation averages over those averages, weighting them appropriately

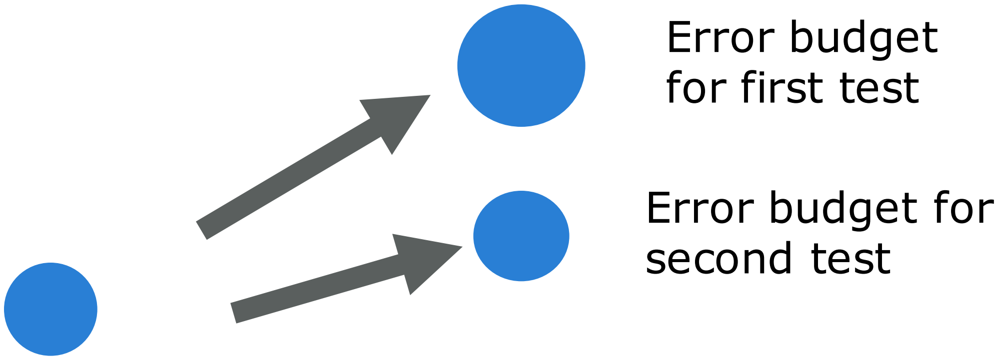
Online FDR Control : High-Level Picture



Error budget
for first test

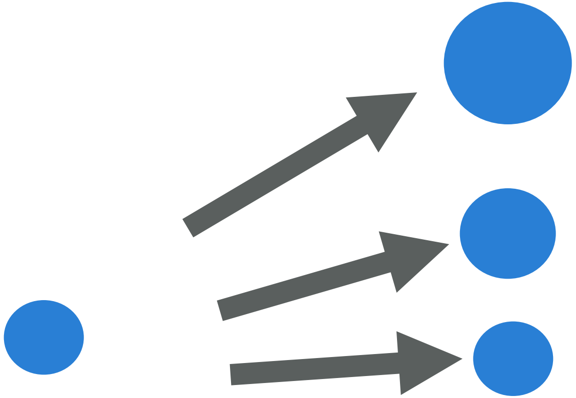
Remaining error budget
or "alpha-wealth"

Online FDR Control : High-Level Picture



Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



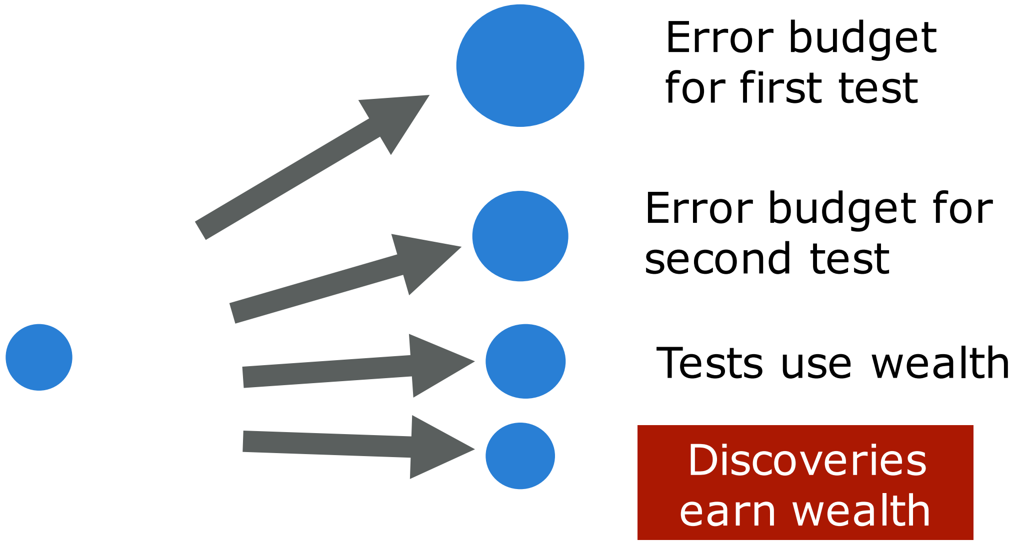
Error budget
for first test

Error budget for
second test

Tests use wealth

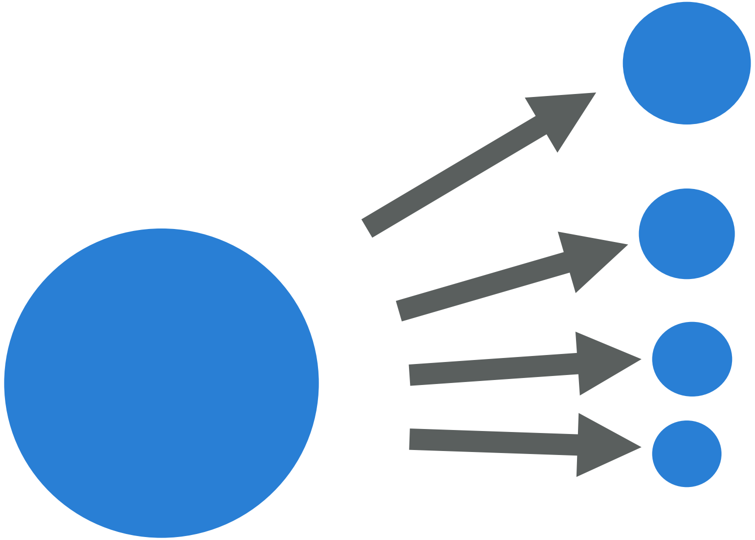
Remaining error budget
or “alpha-wealth”

Online FDR Control : High-Level Picture

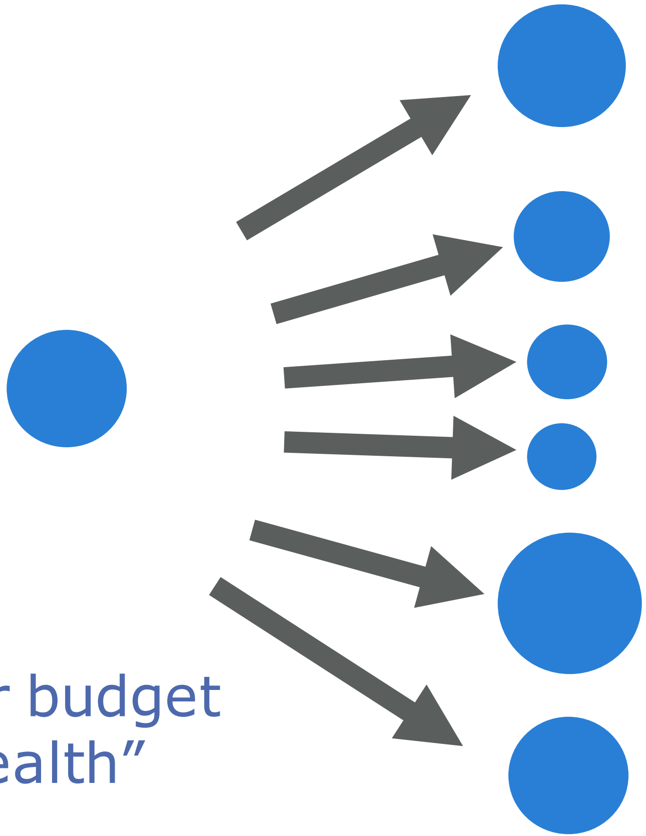


Remaining error budget or “alpha-wealth”

Online FDR Control : High-Level Picture



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 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:

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

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:

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

Two Kinds of Statistical Inference

- Bayesian and Frequentist
- Both inferential frameworks are useful
- It's akin to "waves" vs. "particles" in physics
  - they're both correct in some sense
  - they are complementary in many ways
  - but they also conflict in some serious ways
- Understanding Bayes/frequentist relationships can help you become a real problem solver, not just a person who runs downloads software and runs data analysis procedures

Frequentism

- We want to be able to say that a procedure works "on average"
  - or possibly "with high probability"
- Where does the randomness come from to be able to talk about an "average" or a "probability"?
- The frequentist idea (due to Neyman, Wald, and others) is to assume that we don't just have one dataset, but rather we repeatedly draw datasets independently from the population
  - and the randomness comes from this sampling process
  - for example, that's the meaning of the expectation in going from the FDP to the FDR

Bayesianism

- The idea is to condition on the data and consider the posterior distribution of various unknowns conditional on the data

$$P(\theta \,|\, \text{data}) \propto P(\text{data} \,|\, \theta) P(\theta)$$

- This updates the prior belief into a posterior belief
- A Bayesian doesn't talk about averages over multiple possible data sets; they want to condition on the observed data
- A Bayesian is happy to assign probabilities to things that can't be repeated

Frequentist Hypothesis Testing

- This is what one learns in classical statistics classes
- The basic idea is to specify, via a probability distribution, what data one expects to see under the null hypothesis
  - and similarly for the alternative hypothesis
- One then collects actual data and assesses, via some algorithm, how well the data fit that null distribution
- If the answer is “not so much,” then one rejects the null
- One then proves that such a decision-making algorithm will perform well on average
  - e.g., having a controlled probability of a Type I error

Bayesian Hypothesis Testing

- Has risen, fallen and risen again many times over history
- The basic idea is to specify, via a probability distribution, what data one expects to see under the null hypothesis and similarly for the alternative hypothesis
- One places a prior probability on the null and the alternative
- One now has all the ingredients to compute a conditional probability of the hypothesis given the data

Comparisons

- Bayesian perspective
  - conditional perspective--inferences should be made conditional on the actual observed data, not on possible data one could have observed
  - natural in the setting of a long-term project with a domain expert
  - the optimist---let's make the best use possible of our sophisticated inferential tool
- Frequentist perspective
  - unconditional perspective---inferential procedures should give good answers in repeated use
  - natural in the setting of writing software that will be used by many people for many problems
  - the pessimist--let's protect ourselves against bad decisions given that our inferential procedure is a simplification of reality

Comparisons

- Bayesian perspective
  - conditional perspective--inferences should be made conditional on the actual observed data, not on possible data one could have observed
  - natural in the setting of a long-term project with a domain expert
  - the optimist---let's make the best use possible of our sophisticated inferential tool
- Frequentist perspective
  - unconditional perspective---inferential procedures should give good answers in repeated use
  - natural in the setting of writing software that will be used by many people for many problems
  - the pessimist--let's protect ourselves against bad decisions
- Q: Are "bias" and "variance" frequentist or Bayesian?

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

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

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

Reality \ Decision	0	1
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

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

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

Reality \ Decision	0	1
0	$l(0,0)$	$l(0,1)$
1	$l(1,0)$	$l(1,1)$

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

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

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

Reality \ Decision	0	1
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

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

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

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

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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

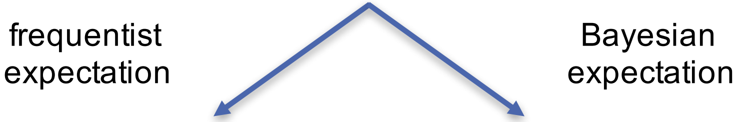
Decision-Theoretic Framework

- 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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

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



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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

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

frequentist expectation

Bayesian expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

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

frequentist expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Bayesian expectation

$$\mathbb{E}[l(\theta, \delta(X)) \mid X]$$

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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

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

frequentist expectation

Bayesian expectation

$$\mathbb{E}_\theta l(\theta, \delta(X)) \qquad \mathbb{E}[l(\theta, \delta(X)) \mid X]$$

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:

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

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

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

frequentist expectation

$$\mathbb{E}_\theta l(\theta, \delta(X))$$

Bayesian expectation

$$\mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Risk Functions

- The frequentist risk:

$$R(\theta) = \mathbb{E}_\theta l(\theta, \delta(X))$$

- The Bayesian posterior risk:

$$\rho(X) = \mathbb{E}[l(\theta, \delta(X)) \mid X]$$

Risk Functions

- The frequentist risk:

$$R(\theta) = \mathbb{E}_\theta l(\theta, \delta(X))$$

- The Bayesian posterior risk:

$$\rho(X) = \mathbb{E}[l(\theta, \delta(X)) \mid X]$$

- A fun bonus exercise: If we take an expectation of $R(\theta)$ with respect to θ, or an expectation of $\rho(X)$ with respect to X, we get a constant known as the "Bayes risk"

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet
- The right model for the truncated data is a truncated Gaussian, and the sample mean is no longer unbiased under the new model

A Small Thought Experiment

- Suppose that you want to estimate the average height of the population in a city
- You take a random sample of 100 people, measure their height X_i and adopt the model $X_i \sim N(\mu, 1)$
- An unbiased estimator of μ is given by $\bar{X}$, the sample mean
  - i.e., the sample mean is a good frequentist estimator
- Now suppose that someone tells you that the measuring device was broken, and anybody over 7 feet tall was recorded as 7 feet
  - but there actually was no one over 7 feet tall; everyone was actually less than 6.5 feet
- The right model for the truncated data is a truncated Gaussian, and the sample mean is no longer unbiased under the new model
- Should you alter your estimate?
  - consider this question from both a Bayesian and frequentist point of view