Tutorial: Economics of Hypothesis Testing

Davide Viviano
Harvard University

May 21, 2024

Illustrative example: clinical trial

- A drug company test the efficacy of a new drug in a clinical trial in a Phase 3 trial. A regulatory agency can impose rules for approval

Illustrative example: clinical trial

- A drug company test the efficacy of a new drug in a clinical trial in a Phase 3 trial. A regulatory agency can impose rules for approval

- Define the null hypothesis H_0
- Collect a sample of patients and test for the efficacy

Illustrative example: clinical trial

- A drug company test the efficacy of a new drug in a clinical trial in a Phase 3 trial. A regulatory agency can impose rules for approval

- Define the null hypothesis H_0
- Collect a sample of patients and test for the efficacy

- How do we design a test-statistic for a given hypothesis?
- When do we reject the desired null hypothesis?
- How do we incorporate costs and benefits in our test?

Some history of hypothesis testing

- Fisher popularized significance test (Fisher, 1955)
    - Consider a null hypothesis and sample (the drug is never effective/sharp null)
    - Report the level of significance (p-value) and with non-significant result draw no conclusions – suspend judgment until further data is available

Some history of hypothesis testing

- Fisher popularized significance test (Fisher, 1955)
  - Consider a null hypothesis and sample (the drug is never effective/sharp null)
  - Report the level of significance (p-value) and with non-significant result draw no conclusions – suspend judgment until further data is available

- Neyman/Pearson popularized hypothesis testing (Neyman and Pearson, 1933)
  - Choose two hypothesis a null (no average effect and alternative)
  - Select the regions of acceptance and rejection
  - Base Type I and Type II error on cost/benefits considerations
  - "Fundamental lemma" $\Rightarrow$ most powerful test for given Type I error

Some history of hypothesis testing

- Fisher popularized significance test (Fisher, 1955)
  - Consider a null hypothesis and sample (the drug is never effective/sharp null)
  - Report the level of significance (p-value) and with non-significant result draw no conclusions – suspend judgment until further data is available

- Neyman/Pearson popularized hypothesis testing (Neyman and Pearson, 1933)
  - Choose two hypothesis a null (no average effect and alternative)
  - Select the regions of acceptance and rejection
  - Base Type I and Type II error on cost/benefits considerations
  - "Fundamental lemma" ⇒ most powerful test for given Type I error

⇒ Inductive vs deductive "In Fisher's view, Neyman-Pearson simply erred in eliminating mental step of modelling because they assumed the situation to already" (Lenhard, 2006)

Current practice in hypothesis testing

Romano and Lehmann (2005) (p. 57):

- "It is customary therefore to assign a bound to the probability of incorrectly rejecting H_0 when it is true and to attempt to minimize the other probability subject to this condition."

Current practice in hypothesis testing

Romano and Lehmann (2005) (p. 57):

- "It is customary therefore to assign a bound to the probability of incorrectly rejecting H_0 when it is true and to attempt to minimize the other probability subject to this condition."

Standard practice has implicit lexicographic preferences

- Control the probability of a mistake under the null (size) first
- Then maximize power

Current practice in hypothesis testing

Romano and Lehmann (2005) (p. 57):

- "It is customary therefore to assign a bound to the probability of incorrectly rejecting H_0 when it is true and to attempt to minimize the other probability subject to this condition."

Standard practice has implicit lexicographic preferences

- Control the probability of a mistake under the null (size) first
- Then maximize power

Q1 How to relate this to a decision problem?

Current practice in hypothesis testing

Romano and Lehmann (2005) (p. 57):

- "It is customary therefore to assign a bound to the probability of incorrectly rejecting H_0 when it is true and to attempt to minimize the other probability subject to this condition."

Standard practice has implicit lexicographic preferences

- Control the probability of a mistake under the null (size) first
- Then maximize power

Q1 How to relate this to a decision problem? How to choose Type I err?

Current practice in hypothesis testing

Romano and Lehmann (2005) (p. 57):

- "It is customary therefore to assign a bound to the probability of incorrectly rejecting H_0 when it is true and to attempt to minimize the other probability subject to this condition."

Standard practice has implicit lexicographic preferences

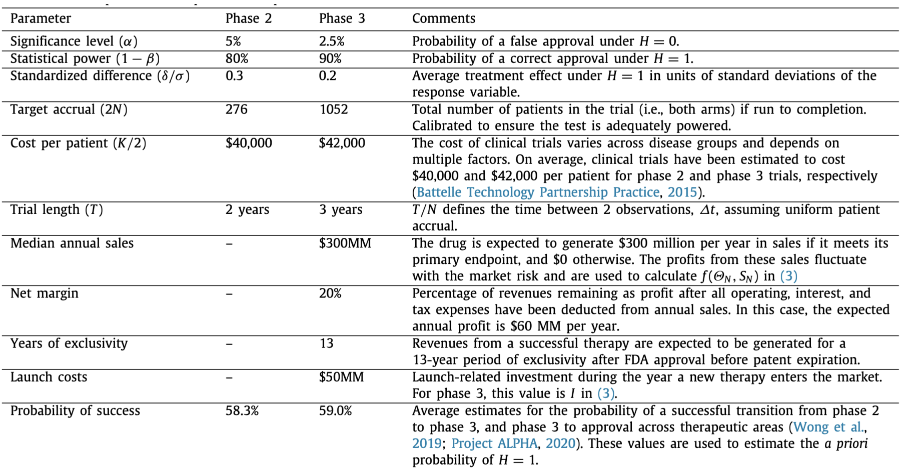
- Control the probability of a mistake under the null (size) first
- Then maximize power

Q1 How to relate this to a decision problem? How to choose Type I err?

Q2 And what if we have multiple hypotheses (decisions)?

“I have met several cases while considering questions of practical expermentation, in which the level of significance $\alpha = 1\%$ proved definitely too stringent. It is the business of the experimenter to choose a proper level any particular case, remembering that the fewer the errors of one kind, more there are of the other” (Neyman and Iwaszkiewicz, 1935)

Clinical trials: example continued (Chaudhuri and Lo, 2020)



Parameter	Phase 2	Phase 3	Comments
Significance level (α)	5%	2.5%	Probability of a false approval under $H = 0$.
Statistical power ($1 - \beta$)	80%	90%	Probability of a correct approval under $H = 1$.
Standardized difference (δ/σ)	0.3	0.2	Average treatment effect under $H = 1$ in units of standard deviations of the response variable.
Target accrual ($2N$)	276	1052	Total number of patients in the trial (i.e., both arms) if run to completion. Calibrated to ensure the test is adequately powered.
Cost per patient ($K/2$)	$40,000	$42,000	The cost of clinical trials varies across disease groups and depends on multiple factors. On average, clinical trials have been estimated to cost $40,000 and $42,000 per patient for phase 2 and phase 3 trials, respectively (Battelle Technology Partnership Practice, 2015).
Trial length (T)	2 years	3 years	T/N defines the time between 2 observations, Δt, assuming uniform patient accrual.
Median annual sales	–	$300MM	The drug is expected to generate $300 million per year in sales if it meets its primary endpoint, and $0 otherwise. The profits from these sales fluctuate with the market risk and are used to calculate $f(\Theta_N, S_N)$ in (3)
Net margin	–	20%	Percentage of revenues remaining as profit after all operating, interest, and tax expenses have been deducted from annual sales. In this case, the expected annual profit is $60 MM per year.
Years of exclusivity	–	13	Revenues from a successful therapy are expected to be generated for a 13-year period of exclusivity after FDA approval before patent expiration.
Launch costs	–	$50MM	Launch-related investment during the year a new therapy enters the market. For phase 3, this value is I in (3).
Probability of success	58.3%	59.0%	Average estimates for the probability of a successful transition from phase 2 to phase 3, and phase 3 to approval across therapeutic areas (Wong et al., 2019; Project ALPHA, 2020). These values are used to estimate the *a priori* probability of $H = 1$.

Content

1. Part 1: Single hypothesis testing

2. Part 2: Multiple hypothesis testing

Outline

- Neyman-Pearson HT framework
- HT as games against nature
- Minimax, Minimax regret and Bayesian decision rules
- Game theoretic interpretations of HT
- Optimal publication decisions

Some useful references

- Ch 1, 3 in Romano and Lehmann (2005),
- Q-values: Storey (2003)
- Some on decision theory: Wald and Wolfowitz (1940), Tetenov (2016), Manski (2004), Isakov et al. (2019), Frankel and Kasy (2022) ...

Neyman-Pearson Hypothesis testing

- Consider data $X \in \mathcal{X}$, with density $f(X; \theta)$ where θ is the parameter of interest (e.g., $X \sim \mathcal{N}(\theta, 1)$)

Neyman-Pearson Hypothesis testing

- Consider data $X \in \mathcal{X}$, with density $f(X;\theta)$ where θ is the parameter of interest (e.g., $X \sim \mathcal{N}(\theta, 1)$)
- We are interested in testing

$$H_0 : \theta \in \Theta_0, \text{ vs } H_1 : \theta \in \Theta_1$$

Neyman-Pearson Hypothesis testing

- Consider data $X \in \mathcal{X}$, with density $f(X;\theta)$ where θ is the parameter of interest (e.g., $X \sim \mathcal{N}(\theta, 1)$)
- We are interested in testing

$$H_0 : \theta \in \Theta_0, \text{ vs } H_1 : \theta \in \Theta_1$$

- A test $\phi : \mathcal{X} \mapsto \{0, 1\}$ is function of data

Neyman-Pearson Hypothesis testing

- Consider data $X \in \mathcal{X}$, with density $f(X;\theta)$ where θ is the parameter of interest (e.g., $X \sim \mathcal{N}(\theta, 1)$)
- We are interested in testing

$$H_0 : \theta \in \Theta_0, \text{ vs } H_1 : \theta \in \Theta_1$$

- A test $\phi : \mathcal{X} \mapsto \{0, 1\}$ is function of data
- The rejection region is defined as

$$R = \Big\{x : \phi(x) = 1\Big\}$$

Type I and Type II Errors

	H_0 True	H_1 True
Choose H_0	correct	Type II error
Choose H_1	Type I error	correct

Type I and Type II Errors

	H_0 True	H_1 True
Choose H_0	correct	Type II error
Choose H_1	Type I error	correct

- Tests have level α (size when exact) if

$$\sup_{\theta \in \Theta_0} \int \phi(x) f(x; \theta) dx \leq \alpha$$

Type I and Type II Errors

	H_0 True	H_1 True
Choose H_0	correct	Type II error
Choose H_1	Type I error	correct

- Tests have level α (size when exact) if

$$\sup_{\theta \in \Theta_0} \int \phi(x) f(x;\theta) dx \leq \alpha$$

- Tests have power $\beta(\theta)$ (note as function of $\theta \in \Theta_1$)

$$\beta(\theta) = \int \phi(x) f(x;\theta) dx$$

Hypothesis testing in practice

Things we typically require in practice

(1) Size control of the test ($\alpha = 0.05$, lexicographic preferences...)

(2) "Sufficient" power subject to size control

Hypothesis testing in practice

Things we typically require in practice

(1) Size control of the test ($\alpha = 0.05$, lexicographic preferences...)

(2) "Sufficient" power subject to size control

Desirable properties we would like from a test

- Unbiased test: a test with size α is unbiased if $\inf_{\theta \in \Theta_1} \beta(\theta) \geq \alpha$
- Consistent test: for a sequence of DGPs $\beta_n(\theta) \to 1, \theta \in \Theta_1$
- Uniformly most powerful (UMP): largest $\beta(\theta)$ for all $\theta \in \Theta_1$ subject to size control (does not always exist)

Hypothesis testing in practice

Things we typically require in practice

(1) Size control of the test ($\alpha = 0.05$, lexicographic preferences...)

(2) "Sufficient" power subject to size control

Desirable properties we would like from a test

- Unbiased test: a test with size α is unbiased if $\inf_{\theta \in \Theta_1} \beta(\theta) \geq \alpha$
- Consistent test: for a sequence of DGPs $\beta_n(\theta) \to 1, \theta \in \Theta_1$
- Uniformly most powerful (UMP): largest $\beta(\theta)$ for all $\theta \in \Theta_1$ subject to size control (does not always exist)

Ok...but how do we choose size-α tests?

Karlin-Rubin Theorem (Neyman-Pearson Lemma)

- Consider $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$
- Let $l(x; \theta_1, \theta_0) = f_{\theta_0}(x)/f_{\theta_1}(x)$ be monotonic in x for any $\theta_1 \geq \theta_0$
  - $\Rightarrow$ Typically attained within the exponential family

Karlin-Rubin Theorem (Neyman-Pearson Lemma)

- Consider $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$
- Let $l(x; \theta_1, \theta_0) = f_{\theta_0}(x)/f_{\theta_1}(x)$ be monotonic in x for any $\theta_1 \geq \theta_0$

  $\Rightarrow$ Typically attained within the exponential family
- Take a test

$$\phi_{x^\star}(x) = \begin{cases} 1 \text{ if } x > x^\star \\ 0 \text{ otherwise} \end{cases}, \quad x^\star : \mathbb{E}_{\theta_0}[\phi_{x^\star}(X)] = \alpha$$

- This is the uniformly most powerful test of level α

Karlin-Rubin Theorem (Neyman-Pearson Lemma)

- Consider $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$
- Let $l(x; \theta_1, \theta_0) = f_{\theta_0}(x)/f_{\theta_1}(x)$ be monotonic in x for any $\theta_1 \geq \theta_0$

  $\Rightarrow$ Typically attained within the exponential family
- Take a test

$$\phi_{x^\star}(x) = \begin{cases} 1 \text{ if } x > x^\star \\ 0 \text{ otherwise} \end{cases}, \quad x^\star : \mathbb{E}_{\theta_0}[\phi_{x^\star}(X)] = \alpha$$

- This is the uniformly most powerful test of level α
- Example: normal-shift model Suppose that $X_1, \cdots, X_n \sim_{i.i.d.} \mathcal{N}(\theta, 1)$. Then we can take $\phi(X) = 1\{\sqrt{n}(\bar{X} - \theta_0) > z_{1-\alpha}\}$.

Karlin-Rubin Theorem (Neyman-Pearson Lemma)

- Consider $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$
- Let $l(x; \theta_1, \theta_0) = f_{\theta_0}(x)/f_{\theta_1}(x)$ be monotonic in x for any $\theta_1 \geq \theta_0$

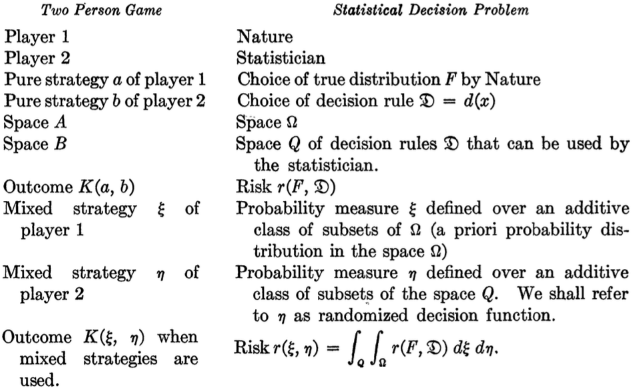
  $\Rightarrow$ Typically attained within the exponential family
- Take a test

$$\phi_{x^\star}(x) = \begin{cases} 1 \text{ if } x > x^\star \\ 0 \text{ otherwise} \end{cases} , \quad x^\star : \mathbb{E}_{\theta_0}[\phi_{x^\star}(X)] = \alpha$$

- This is the uniformly most powerful test of level α
- Example: normal-shift model Suppose that $X_1, \cdots, X_n \sim_{i.i.d.} \mathcal{N}(\theta, 1)$. Then we can take $\phi(X) = 1\{\sqrt{n}(\bar{X} - \theta_0) > z_{1-\alpha}\}$.

$\Rightarrow$ Note however that UMP test does not exist for vector of parameters

Statistical decision problem Wald and Wolfowitz (1940)



Two Person Game	*Statistical Decision Problem*
Player 1	Nature
Player 2	Statistician
Pure strategy a of player 1	Choice of true distribution F by Nature
Pure strategy b of player 2	Choice of decision rule $\mathfrak{D} = d(x)$
Space A	Space Ω
Space B	Space Q of decision rules $\mathfrak{D}$ that can be used by the statistician.
Outcome $K(a, b)$	Risk $r(F, \mathfrak{D})$
Mixed strategy ξ of player 1	Probability measure ξ defined over an additive class of subsets of Ω (a priori probability distribution in the space Ω)
Mixed strategy η of player 2	Probability measure η defined over an additive class of subsets of the space Q. We shall refer to η as randomized decision function.
Outcome $K(\xi, \eta)$ when mixed strategies are used.	Risk $r(\xi, \eta) = \int_Q \int_\Omega r(F, \mathfrak{D})\, d\xi\, d\eta$.

Statistical testing as a decision problem

- Define $\phi(X)$ as a decision function
  - Should we approve the drug tested by the company?

Statistical testing as a decision problem

- Define $\phi(X)$ as a decision function
  - Should we approve the drug tested by the company?

- For a given decision a, define a loss function and corresponding risk

$$a \mapsto L(\theta, a), \quad R(\theta, \phi) = \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Statistical testing as a decision problem

- Define $\phi(X)$ as a decision function
    - Should we approve the drug tested by the company?

- For a given decision a, define a loss function and corresponding risk

$$a \mapsto L(\theta, a), \quad R(\theta, \phi) = \mathbb{E}_\theta[L(\theta, \phi(X))]$$

- Example: Binary loss function

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta \in \Theta_0$	0	K
$H_1 : \theta \notin \Theta_0$	1	0

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\left\{\theta \in \Theta_0\right\}}_{\text{loss from approval}} + \underbrace{(1-\phi(X))1\left\{\theta \notin \Theta_0\right\}}_{\text{loss from status quo}}$$

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\Big\{\theta \in \Theta_0\Big\}}_{\text{loss from approval}} + \underbrace{(1 - \phi(X))1\Big\{\theta \notin \Theta_0\Big\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_{\theta}[L(\theta, \phi(X))]$$

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\left\{\theta \in \Theta_0\right\}}_{\text{loss from approval}} + \underbrace{(1-\phi(X))1\left\{\theta \notin \Theta_0\right\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Under mild regularity conditions

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\left\{\theta \in \Theta_0\right\}}_{\text{loss from approval}} + \underbrace{(1-\phi(X))1\left\{\theta \notin \Theta_0\right\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Under mild regularity conditions

$$\max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))] = \max\Big\{K \underbrace{\max_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\leq \text{size } \alpha},$$

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\left\{\theta \in \Theta_0\right\}}_{\text{loss from approval}} + \underbrace{(1 - \phi(X))1\left\{\theta \notin \Theta_0\right\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Under mild regularity conditions

$$\max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))] = \max\left\{K \underbrace{\max_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\leq \text{size } \alpha}, 1 - \underbrace{\min_{\theta \notin \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\geq \text{size } \alpha}\right\}$$

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\Big\{\theta \in \Theta_0\Big\}}_{\text{loss from approval}} + \underbrace{(1-\phi(X))1\Big\{\theta \notin \Theta_0\Big\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Under mild regularity conditions

$$\max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))] = \max\left\{K \underbrace{\max_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\leq \text{size } \alpha}, 1 - \underbrace{\min_{\theta \notin \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\geq \text{size } \alpha}\right\}$$

$\Rightarrow$ Optimal ϕ chooses $K\alpha = 1 - \alpha \Rightarrow \alpha = 2.5\%$ if $K = 39$

Minimax decision rule

Consider an objective function of the form

$$L(\theta, \phi(X)) = \underbrace{K\phi(X)1\Big\{\theta \in \Theta_0\Big\}}_{\text{loss from approval}} + \underbrace{(1-\phi(X))1\Big\{\theta \notin \Theta_0\Big\}}_{\text{loss from status quo}}$$

We do not know $\theta \Rightarrow$ maximin decision

$$\min_{\phi} \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))]$$

Under mild regularity conditions

$$\max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \phi(X))] = \max\Big\{K \underbrace{\max_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\leq \text{size } \alpha}, 1 - \underbrace{\min_{\theta \notin \Theta_0} \mathbb{E}_\theta[\phi(X)]}_{\geq \text{size } \alpha}\Big\}$$

$\Rightarrow$ Optimal ϕ chooses $K\alpha = 1 - \alpha \Rightarrow \alpha = 2.5\%$ if $K = 39$

$\Rightarrow$ Problem: Minimax rule does not incorporate magnitudes of the effects

Minimax regret: the magnitude matters

$\Rightarrow$ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect

Minimax regret: the magnitude matters

- ⇒ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect
- ⇒ Minimax is conservative ($\phi(X) = 0$)

Minimax regret: the magnitude matters

- ⇒ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect
- ⇒ Minimax is conservative ($\phi(X) = 0$) ⇒ look at minimax regret

$$L(\theta, \phi(X)) = \theta \underbrace{1\{\theta > 0\}}_{\text{oracle decision}} - \theta\phi(X)$$

Minimax regret: the magnitude matters

$\Rightarrow$ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect

$\Rightarrow$ Minimax is conservative ($\phi(X) = 0$) $\Rightarrow$ look at minimax regret

$$L(\theta, \phi(X)) = \theta \underbrace{1\{\theta > 0\}}_{\text{oracle decision}} - \theta\phi(X)$$

- Manski (2004) recommends empirical success rule

$$\hat{\phi}(X) = 1\left\{\bar{X} \geq 0\right\}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = \theta$$

Minimax regret: the magnitude matters

⇒ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect

⇒ Minimax is conservative ($\phi(X) = 0$) ⇒ look at minimax regret

$$L(\theta, \phi(X)) = \theta \underbrace{1\{\theta > 0\}}_{\text{oracle decision}} - \theta\phi(X)$$

- Manski (2004) recommends empirical success rule

$$\hat{\phi}(X) = 1\left\{\bar{X} \geq 0\right\}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = \theta$$

⇒ $\sup_\theta \mathbb{E}_\theta[L(\theta, \hat{\phi}(X))] = \mathcal{O}(1/\sqrt{n})$ (rate is minimax optimal)

Minimax regret: the magnitude matters

$\Rightarrow$ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect

$\Rightarrow$ Minimax is conservative ($\phi(X) = 0$) $\Rightarrow$ look at minimax regret

$$L(\theta, \phi(X)) = \theta \underbrace{1\{\theta > 0\}}_{\text{oracle decision}} - \theta\phi(X)$$

- Manski (2004) recommends empirical success rule

$$\hat{\phi}(X) = 1\Big\{\bar{X} \geq 0\Big\}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = \theta$$

$\Rightarrow$ $\sup_\theta \mathbb{E}_\theta[L(\theta, \hat{\phi}(X))] = \mathcal{O}(1/\sqrt{n})$ (rate is minimax optimal)

- This is equivalent to one sided hypothesis test with size $\alpha = 50\%$

Minimax regret: the magnitude matters

$\Rightarrow$ Consider a utility function $\theta\phi(X) \Rightarrow \theta =$ treatment effect

$\Rightarrow$ Minimax is conservative ($\phi(X) = 0$) $\Rightarrow$ look at minimax regret

$$L(\theta, \phi(X)) = \theta \underbrace{1\{\theta > 0\}}_{\text{oracle decision}} - \theta\phi(X)$$

- Manski (2004) recommends empirical success rule

$$\hat{\phi}(X) = 1\left\{\bar{X} \geq 0\right\}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = \theta$$

$\Rightarrow$ $\sup_\theta \mathbb{E}_\theta[L(\theta, \hat{\phi}(X))] = \mathcal{O}(1/\sqrt{n})$ (rate is minimax optimal)

- This is equivalent to one sided hypothesis test with size $\alpha = 50\%$

  "It has perhaps not been sufficiently noted that there are decisional situations [...] where, one might say, an insignificant difference is better than no difference at all" (Simon, 1945)

Including asymmetry in the regret function

(i) No clear justification for $\alpha = 5\%$

(ii) No clear understanding of lexicographic preferences over Type I/II err

Including asymmetry in the regret function

(i) No clear justification for $\alpha = 5\%$

(ii) No clear understanding of lexicographic preferences over Type I/II err

$\Rightarrow$ Tetenov (2012) proposes asymmetric regret

Including asymmetry in the regret function

(i) No clear justification for $\alpha = 5\%$

(ii) No clear understanding of lexicographic preferences over Type I/II err

$\Rightarrow$ Tetenov (2012) proposes asymmetric regret

> "To obtain decision rules comparable to tests at conventional levels, the asymmetry factor K has to be much greater than with the 1-K loss function (2). The difference is due to the interaction between the magnitude of errors and their probability. One-sided 5% tests are minimax optimal for K=102, while 1% tests are optimal for K=970. In contrast, a moderate loss aversion coefficient of K=3 would lead to a one-sided 34% test." (Tetenov, 2016)

Including asymmetry in the regret function

(i) No clear justification for $\alpha = 5\%$

(ii) No clear understanding of lexicographic preferences over Type I/II err

$\Rightarrow$ Tetenov (2012) proposes asymmetric regret

> "To obtain decision rules comparable to tests at conventional levels, the asymmetry factor K has to be much greater than with the 1-K loss function (2). The difference is due to the interaction between the magnitude of errors and their probability. One-sided 5% tests are minimax optimal for K=102, while 1% tests are optimal for K=970. In contrast, a moderate loss aversion coefficient of K=3 would lead to a one-sided 34% test." (Tetenov, 2016)

$\Rightarrow$ But what if we use prior information about magnitude of the effects?

Bayesian decision making

- Main issue $\mathbb{E}_\theta[L(\theta, \phi(X))]$ is a function of θ (which we do not know)

Bayesian decision making

- Main issue $\mathbb{E}_\theta[L(\theta, \phi(X))]$ is a function of θ (which we do not know)
- admissibility: ϕ is not admissible if there exists another ϕ' such that $\mathbb{E}_\theta[L(\theta, \phi(X))] \geq \mathbb{E}_\theta[L(\theta, \phi'(X))]$ for all θ and strictly larger for some.

Bayesian decision making

- Main issue $\mathbb{E}_\theta[L(\theta, \phi(X))]$ is a function of θ (which we do not know)
- admissibility: ϕ is not admissible if there exists another ϕ' such that $\mathbb{E}_\theta[L(\theta, \phi(X))] \geq \mathbb{E}_\theta[L(\theta, \phi'(X))]$ for all θ and strictly larger for some.
- Suppose for known prior $\pi(\theta)$ we maximize

$$r(\pi, \phi) = \int \mathbb{E}_\theta[L(\theta, \phi(X))]\pi(\theta)d\theta$$

Bayesian decision making

- Main issue $\mathbb{E}_\theta[L(\theta, \phi(X))]$ is a function of θ (which we do not know)
- admissibility: ϕ is not admissible if there exists another ϕ' such that $\mathbb{E}_\theta[L(\theta, \phi(X))] \geq \mathbb{E}_\theta[L(\theta, \phi'(X))]$ for all θ and strictly larger for some.
- Suppose for known prior $\pi(\theta)$ we maximize

$$r(\pi, \phi) = \int \mathbb{E}_\theta[L(\theta, \phi(X))]\pi(\theta) d\theta$$

$\Rightarrow$ Complete class theorem: If ϕ is admissible, then ϕ is a Bayes decision rule for some prior distribution. [Any Bayes rule is admissible] (!)

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.
- By Bayes thm, for some $m(x)$:

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.
- By Bayes thm, for some $m(x)$:

$$r(\pi, 0) = \int (1 - \pi) f(x|\theta_1)/m(x) dx,$$

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.
- By Bayes thm, for some $m(x)$:

$$r(\pi, 0) = \int (1 - \pi) f(x|\theta_1)/m(x)dx, \quad r(\pi, 1) = K\pi \int f(x|\theta_0)/m(x)dx$$

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.
- By Bayes thm, for some $m(x)$:

$$r(\pi, 0) = \int (1 - \pi) f(x|\theta_1)/m(x) dx, \quad r(\pi, 1) = K\pi \int f(x|\theta_0)/m(x) dx$$

- Optimal rule

Bayes optimal rules with 0/1 loss

World/decision	$\phi(X) = 0$	$\phi(X) = 1$
$H_0 : \theta = \theta_0$	0	K
$H_1 : \theta = \theta_1$	1	0

- Suppose $\theta \in \{\theta_0, \theta_1\}$ with probability $\pi, 1 - \pi$.
- By Bayes thm, for some $m(x)$:

$$r(\pi, 0) = \int (1 - \pi) f(x|\theta_1)/m(x) dx, \quad r(\pi, 1) = K\pi \int f(x|\theta_0)/m(x) dx$$

- Optimal rule

$$\phi(x) = \begin{cases} 1 \text{ if } \frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{(1-\pi)}{K\pi} \\ 0 \text{ otherwise} \end{cases}$$

Clinical trial: example revisited

- Isakov et al. (2019) study Bayesian decision analysis (BDA)

  “for terminal illnesses with no existing therapies such as pancreatic cancer, the standard threshold of 2.5% is substantially more conservative than the BDA-optimal threshold of 23.9% to 27.8%. For relatively less deadly conditions such as prostate cancer, 2.5% is more risk-tolerant or aggressive than the BDA-optimal threshold of 1.2% to 1.5%”

Clinical trial: example revisited

- Isakov et al. (2019) study Bayesian decision analysis (BDA)

  "for terminal illnesses with no existing therapies such as pancreatic cancer, the standard threshold of 2.5% is substantially more conservative than the BDA-optimal threshold of 23.9% to 27.8%. For relatively less deadly conditions such as prostate cancer, 2.5% is more risk-tolerant or aggressive than the BDA-optimal threshold of 1.2% to 1.5%"

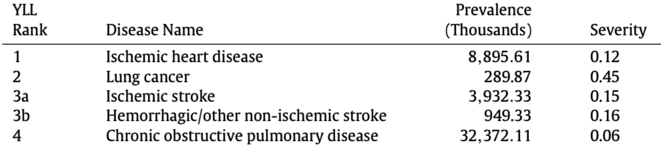
- Authors assume non-informative priors $\pi = 0.5$
- Impose a power constraint on the test of 90%
- Estimate the costs as function of the benefits net of side effects

Clinical trial: example revisited

- Isakov et al. (2019) study Bayesian decision analysis (BDA)

  "for terminal illnesses with no existing therapies such as pancreatic cancer, the standard threshold of 2.5% is substantially more conservative than the BDA-optimal threshold of 23.9% to 27.8%. For relatively less deadly conditions such as prostate cancer, 2.5% is more risk-tolerant or aggressive than the BDA-optimal threshold of 1.2% to 1.5%"

- Authors assume non-informative priors $\pi = 0.5$
- Impose a power constraint on the test of 90%
- Estimate the costs as function of the benefits net of side effects



YLL		Prevalence	
Rank	Disease Name	(Thousands)	Severity
1	Ischemic heart disease	8,895.61	0.12
2	Lung cancer	289.87	0.45
3a	Ischemic stroke	3,932.33	0.15
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16
4	Chronic obstructive pulmonary disease	32,372.11	0.06

Measures of uncertainty about the null

- For t-stat $T(X)$ ad observed t-stat t Fisher suggests p-values

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T \geq t).$$

Measures of uncertainty about the null

- For t-stat $T(X)$ ad observed t-stat t Fisher suggests p-values

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T \geq t).$$

- Storey et al. introduced the q-value [see Storey (2003)]
- Suppose $T(X)|H_0 \sim F_0$, $T(X)|H_1 \sim F_1$, and $P(H_0 \text{ is true}) = \pi$

Measures of uncertainty about the null

- For t-stat $T(X)$ ad observed t-stat t Fisher suggests p-values

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T \geq t).$$

- Storey et al. introduced the q-value [see Storey (2003)]
- Suppose $T(X)|H_0 \sim F_0$, $T(X)|H_1 \sim F_1$, and $P(H_0 \text{ is true}) = \pi$
- The q-value defines the posterior prob that H_0 is true, under rejection

$$P\left(H_0 \text{ is true}\middle| T \geq t\right)$$

Measures of uncertainty about the null

- For t-stat $T(X)$ ad observed t-stat t Fisher suggests p-values

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T \geq t).$$

- Storey et al. introduced the q-value [see Storey (2003)]
- Suppose $T(X)|H_0 \sim F_0$, $T(X)|H_1 \sim F_1$, and $P(H_0 \text{ is true}) = \pi$
- The q-value defines the posterior prob that H_0 is true, under rejection

$$P\left(H_0 \text{ is true} \middle| T \geq t\right)$$

- For observed statistic t, q-value and p-value are define

$$P(H_0 \text{ is true}|T \geq t) = P(T \geq t|H_0 \text{ is true}) \times \frac{P(H_0 \text{ is true})}{P(T \geq t)}$$

Measures of uncertainty about the null

- For t-stat $T(X)$ ad observed t-stat t Fisher suggests p-values

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T \geq t).$$

- Storey et al. introduced the q-value [see Storey (2003)]
- Suppose $T(X)|H_0 \sim F_0$, $T(X)|H_1 \sim F_1$, and $P(H_0 \text{ is true}) = \pi$
- The q-value defines the posterior prob that H_0 is true, under rejection

$$P\left(H_0 \text{ is true} \middle| T \geq t\right)$$

- For observed statistic t, q-value and p-value are define

$$P(H_0 \text{ is true}| T \geq t) = P(T \geq t | H_0 \text{ is true}) \times \frac{P(H_0 \text{ is true})}{P(T \geq t)}$$

⇒ Direct connection to Bayesian decision making

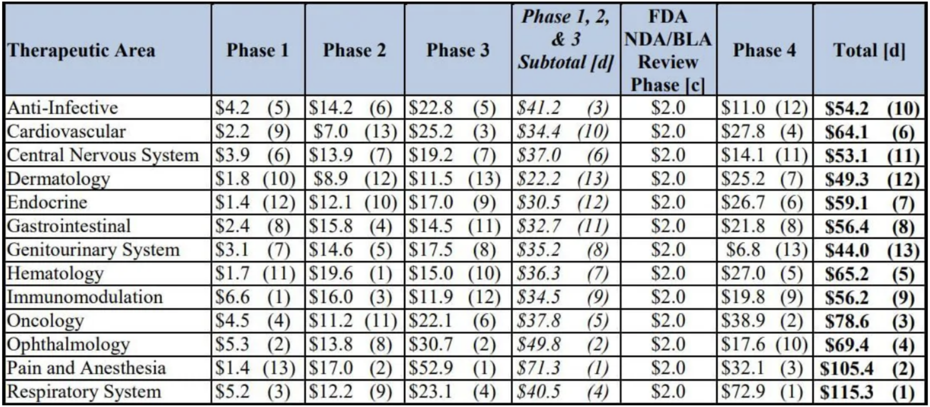
Taking stock

- We have interpreted hypothesis testing as a game against nature
- We have a single decision maker
- The costs and benefits occurring after the decision is taken matter

⇒ Anything we have missed?

Well... firms must decide whether to run experiments

Table 1: Total Per-Study Costs (in $ Millions), by Phase and Therapeutic Area [a]



Therapeutic Area	Phase 1	Phase 2	Phase 3	*Phase 1, 2, & 3 Subtotal [d]*	FDA NDA/BLA Review Phase [c]	Phase 4	Total [d]
Anti-Infective	$4.2 (5)	$14.2 (6)	$22.8 (5)	*$41.2 (3)*	$2.0	$11.0 (12)	**$54.2 (10)**
Cardiovascular	$2.2 (9)	$7.0 (13)	$25.2 (3)	*$34.4 (10)*	$2.0	$27.8 (4)	**$64.1 (6)**
Central Nervous System	$3.9 (6)	$13.9 (7)	$19.2 (7)	*$37.0 (6)*	$2.0	$14.1 (11)	**$53.1 (11)**
Dermatology	$1.8 (10)	$8.9 (12)	$11.5 (13)	*$22.2 (13)*	$2.0	$25.2 (7)	**$49.3 (12)**
Endocrine	$1.4 (12)	$12.1 (10)	$17.0 (9)	*$30.5 (12)*	$2.0	$26.7 (6)	**$59.1 (7)**
Gastrointestinal	$2.4 (8)	$15.8 (4)	$14.5 (11)	*$32.7 (11)*	$2.0	$21.8 (8)	**$56.4 (8)**
Genitourinary System	$3.1 (7)	$14.6 (5)	$17.5 (8)	*$35.2 (8)*	$2.0	$6.8 (13)	**$44.0 (13)**
Hematology	$1.7 (11)	$19.6 (1)	$15.0 (10)	*$36.3 (7)*	$2.0	$27.0 (5)	**$65.2 (5)**
Immunomodulation	$6.6 (1)	$16.0 (3)	$11.9 (12)	*$34.5 (9)*	$2.0	$19.8 (9)	**$56.2 (9)**
Oncology	$4.5 (4)	$11.2 (11)	$22.1 (6)	*$37.8 (5)*	$2.0	$38.9 (2)	**$78.6 (3)**
Ophthalmology	$5.3 (2)	$13.8 (8)	$30.7 (2)	*$49.8 (2)*	$2.0	$17.6 (10)	**$69.4 (4)**
Pain and Anesthesia	$1.4 (13)	$17.0 (2)	$52.9 (1)	*$71.3 (1)*	$2.0	$32.1 (3)	**$105.4 (2)**
Respiratory System	$5.2 (3)	$12.2 (9)	$23.1 (4)	*$40.5 (4)*	$2.0	$72.9 (1)	**$115.3 (1)**

A game-theoretic approach to statistical testing

- Research costs are typically burnt privately
- But research is a public good

⇒ If tests are too conservative, firms will not experiment

A game-theoretic approach to statistical testing

- Research costs are typically burnt privately
- But research is a public good

⇒ If tests are too conservative, firms will not experiment

⇒ Consider two agents (Tetenov, 2016)

  - Regulator (FDA) can ex-ante enforce a statistical testing procedure but does not know treatment effects

A game-theoretic approach to statistical testing

- Research costs are typically burnt privately
- But research is a public good

⇒ If tests are too conservative, firms will not experiment

⇒ Consider two agents (Tetenov, 2016)
  - Regulator (FDA) can ex-ante enforce a statistical testing procedure but does not know treatment effects
  - Drug company can decide to run a pre-specified experiment after observing the regulator's choice and has private info about effects

A game-theoretic approach to statistical testing

- Research costs are typically burnt privately
- But research is a public good

⇒ If tests are too conservative, firms will not experiment

⇒ Consider two agents (Tetenov, 2016)
  - Regulator (FDA) can ex-ante enforce a statistical testing procedure but does not know treatment effects
  - Drug company can decide to run a pre-specified experiment after observing the regulator's choice and has private info about effects

⇒ Principal-agent problem where a rationality constraint can bind

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

- Scenario 2: drug company knows θ (or has prior) but cares about profits, whereas regulator cares about welfare

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

- Scenario 2: drug company knows θ (or has prior) but cares about profits, whereas regulator cares about welfare
  - Maximin solution:

$$\max_{\phi} \min_{\theta} v_\phi(\theta), \quad v_\phi(\theta) = \begin{cases} \underbrace{\mathbb{E}_\theta[\phi(X)]\theta}_{\text{welfare}} \\ \\ \end{cases}$$

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

- Scenario 2: drug company knows θ (or has prior) but cares about profits, whereas regulator cares about welfare
  - Maximin solution:

$$\max_{\phi} \min_{\theta} v_{\phi}(\theta), \quad v_{\phi}(\theta) = \begin{cases} \underbrace{\mathbb{E}_{\theta}[\phi(X)]\theta}_{\text{welfare}} & \text{if } \underbrace{b\mathbb{E}_{\theta}[\phi(X)]}_{\text{firm's profits}} \geq C \\ & \end{cases}$$

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

- Scenario 2: drug company knows θ (or has prior) but cares about profits, whereas regulator cares about welfare
  - Maximin solution:

$$\max_{\phi} \min_{\theta} v_{\phi}(\theta), \quad v_{\phi}(\theta) = \begin{cases} \underbrace{\mathbb{E}_{\theta}[\phi(X)]\theta}_{\text{welfare}} & \text{if } \underbrace{b\mathbb{E}_{\theta}[\phi(X)]}_{\text{firm's profits}} \geq C \\ 0 & \text{otherwise} \end{cases}$$

Different scenarios

- Scenario 1: drug company and regulator have the same incentives/utility and drug company knows θ
  - $\Rightarrow$ Optimal is to impose no constraint on the statistical test
  - $\Rightarrow$ Drug company will self-approve the drug if generates positive effect
  - $\Rightarrow$ No need to run any experiment!

- Scenario 2: drug company knows θ (or has prior) but cares about profits, whereas regulator cares about welfare
  - Maximin solution:

$$\max_{\phi} \min_{\theta} v_\phi(\theta), \quad v_\phi(\theta) = \begin{cases} \underbrace{\mathbb{E}_\theta[\phi(X)]\theta}_{\text{welfare}} & \text{if } \underbrace{b\mathbb{E}_\theta[\phi(X)]}_{\text{firm's profits}} \geq C \\ 0 & \text{otherwise} \end{cases}$$

  - $\Rightarrow$ Optimal rule $\mathbb{E}_\theta[\phi(X)] \leq C/b$ for all $\theta \leq 0$
  - $\Rightarrow$ Tetenov (2016) suggests $C/b = 15\%$.

In summary

- Hypothesis testing is difficult to rationalize
- In a frequentist framework it requires strong asymmetries
- In a Bayesian framework we may want to report posterior probabilities
- In general, measuring costs and benefits is crucial
- And... we should not forget incentives!

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$,

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\Big(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\Big)^2 \phi(X)\right.$$

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\Big(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\Big)^2 \phi(X) + \underbrace{c_p}_{\text{cost of publication}} \phi(X)\right.$$

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\left(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\right)^2 \phi(X) + \underbrace{c_p}_{\text{cost of publication}} \phi(X) + \theta^2(1-\phi(X))\right]$$

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\Big(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\Big)^2 \phi(X) + \underbrace{c_p}_{\text{cost of publication}} \phi(X) + \theta^2(1 - \phi(X))\right]$$

- Optimal ϕ weights costs vs benefits so that

$$\phi(X) = 1\Big\{\frac{X}{\sigma} \geq x^\star\Big\}, \quad X^\star = \frac{\sigma}{\eta^2} + \frac{1}{\sigma} \geq \sqrt{c_p}$$

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\Big(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\Big)^2 \phi(X) + \underbrace{c_p}_{\text{cost of publication}} \phi(X) + \theta^2(1 - \phi(X))\right]$$

- Optimal ϕ weights costs vs benefits so that

$$\phi(X) = 1\Big\{\frac{X}{\sigma} \geq x^\star\Big\}, \quad X^\star = \frac{\sigma}{\eta^2} + \frac{1}{\sigma} \geq \sqrt{c_p}$$

$\Rightarrow$ We want more surprising results for larger publication costs (Frankel and Kasy, 2022)

Connection with optimal publication rules

- Should we also publish "more surprising" results? (consider ϕ as publication decision)
- Abadie (2020) argues non significance is more informative in the limit
- On the other hand: consider $X|\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \eta^2)$, loss:

$$\mathbb{E}\left[\Big(\theta - \underbrace{\mathbb{E}[\theta|X]}_{\text{action of audience}}\Big)^2 \phi(X) + \underbrace{c_p}_{\text{cost of publication}} \phi(X) + \theta^2(1 - \phi(X))\right]$$

- Optimal ϕ weights costs vs benefits so that

$$\phi(X) = 1\Big\{\frac{X}{\sigma} \geq x^\star\Big\}, \quad X^\star = \frac{\sigma}{\eta^2} + \frac{1}{\sigma} \geq \sqrt{c_p}$$

⇒ We want more surprising results for larger publication costs (Frankel and Kasy, 2022)

⇒ But what if researchers are strategic in the choice of the design?

Model overview [Jagadeesan and Viviano, 2024+]

- Three agents: an editor, an audience, and a researcher
- State of the world $\theta_0 \sim \mathcal{N}(0, \eta_0^2)$

Model overview [Jagadeesan and Viviano, 2024+]

- Three agents: an editor, an audience, and a researcher
- State of the world $\theta_0 \sim \mathcal{N}(0, \eta_0^2)$

game timing:

1. researcher chooses which study $\Delta = (\beta_\Delta, S_\Delta)$ to run
   - to maximize chance of publication, net of cost of executing Δ
2. researcher obtains results $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, S_\Delta^2)$
3. editor decides whether to *publish* results
4. audience takes action $a^*(X)$ to minimize expected loss $\mathbb{E}[(a - \theta_0)^2 | X]$

Model overview [Jagadeesan and Viviano, 2024+]

- Three agents: an editor, an audience, and a researcher
- State of the world $\theta_0 \sim \mathcal{N}(0, \eta_0^2)$

game timing:

1. researcher chooses which study $\Delta = (\beta_\Delta, S_\Delta)$ to run
    - to maximize chance of publication, net of cost of executing Δ
2. researcher obtains results $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, S_\Delta^2)$
3. editor decides whether to *publish* results
4. audience takes action $a^*(X)$ to minimize expected loss $\mathbb{E}[(a - \theta_0)^2 | X]$

- editor maximizes audience's welfare net of publication cost c_p

Model overview [Jagadeesan and Viviano, 2024+]

- Three agents: an editor, an audience, and a researcher
- State of the world $\theta_0 \sim \mathcal{N}(0, \eta_0^2)$

game timing:

1. researcher chooses which study $\Delta = (\beta_\Delta, S_\Delta)$ to run without observing X
   - to maximize chance of publication, net of cost of executing Δ
2. researcher obtains results $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, S_\Delta^2)$
3. editor decides whether to *publish* results based on X and Δ
4. audience takes action $a^*(X)$ to minimize expected loss $\mathbb{E}[(a - \theta_0)^2 | X]$

- editor maximizes audience's welfare net of publication cost c_p
- symmetric info case: which research designs to incentivize?

Model overview [Jagadeesan and Viviano, 2024+]

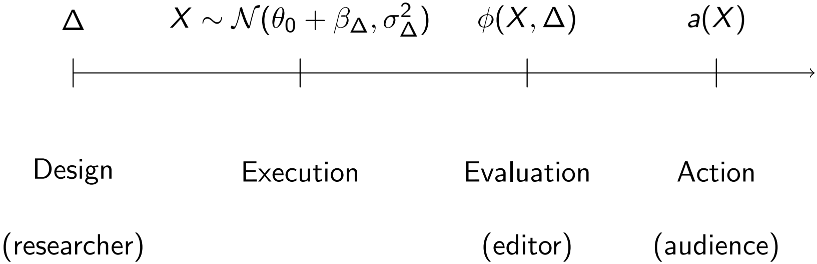
- Three agents: an editor, an audience, and a researcher
- State of the world $\theta_0 \sim \mathcal{N}(0, \eta_0^2)$

game timing:

1. researcher chooses which study $\Delta = (\beta_\Delta, S_\Delta)$ to run based on observing X
    - to maximize chance of publication, net of cost of executing Δ
2. researcher obtains results $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, S_\Delta^2)$
3. editor decides whether to *publish* results based on X only
4. audience takes action $a^*(X)$ to minimize expected loss $\mathbb{E}[(a - \theta_0)^2 | X]$

- editor maximizes audience's welfare net of publication cost c_p
- symmetric info case: which research designs to incentivize?
- asymmetric info case: what form of selective publication to use?

Symmetric information case



Δ $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, \sigma^2_\Delta)$ $\phi(X, \Delta)$ $a(X)$

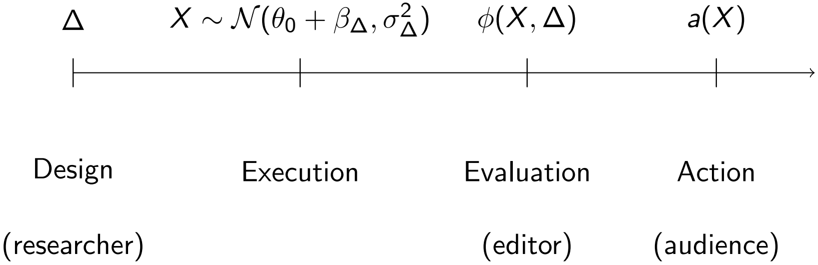
Design (researcher)

Execution

Evaluation (editor)

Action (audience)

Symmetric information case



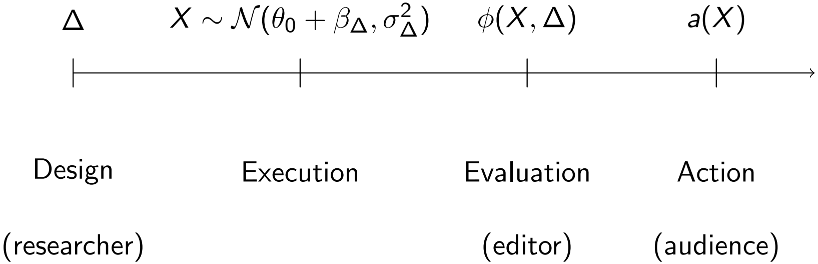
Δ $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, \sigma^2_\Delta)$ $\phi(X, \Delta)$ $a(X)$

Design Execution Evaluation Action

(researcher) (editor) (audience)

- potential designs $\Delta \in \{\text{Experiment}, \text{Observational}\}$

Symmetric information case



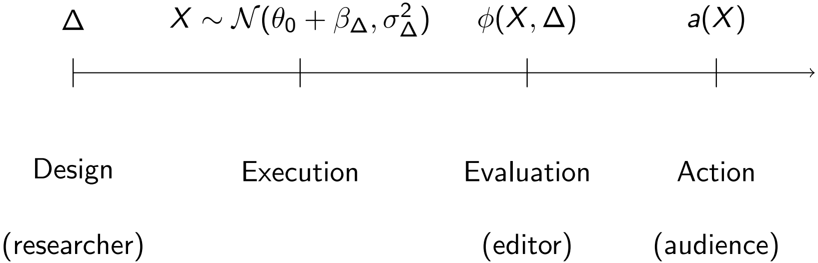
Δ | $X \sim \mathcal{N}(\theta_0 + \beta_\Delta, \sigma_\Delta^2)$ | $\phi(X, \Delta)$ | $a(X)$

Design | Execution | Evaluation | Action

(researcher) | | (editor) | (audience)

- potential designs $\Delta \in \{\text{Experiment}, \text{Observational}\}$
- researcher chooses Δ to maximize $\mathbb{E}_X[\phi(X, \Delta)] - C(\Delta)$

Symmetric information case



Δ

$X \sim \mathcal{N}(\theta_0 + \beta_\Delta, \sigma^2_\Delta)$

$\phi(X, \Delta)$

$a(X)$

Design

(researcher)

Execution

Evaluation

(editor)

Action

(audience)

- potential designs $\Delta \in \{\text{Experiment}, \text{Observational}\}$
- researcher chooses Δ to maximize $\mathbb{E}_X[\phi(X, \Delta)] - C(\Delta)$
- suppose that $C(E) > C(O) = 0$ (costly large-scale experiment)

Which experiment studies to publish?

if the editor is constrained to implement $\Delta =$ Experiment, then optimal publication decision rules satisfy

$$\phi(X, E) = \begin{cases} 1 & \text{if } |X| > X_E^* \\ 0 & \text{if } |X| < X_E^* \end{cases},$$

where

$$X_E^* = \frac{S_E^2 + \eta_0^2}{\eta_0^2} \sqrt{c_p}$$

Which experiment studies to publish?

if the editor is constrained to implement $\Delta = \text{Experiment}$, then optimal publication decision rules satisfy

$$\phi(X, E) = \begin{cases} 1 & \text{if } |X| > X_E^* \\ 0 & \text{if } |X| < X_E^* \end{cases},$$

where

$$X_E^* = \max\left\{ \frac{S_E^2 + \eta_0^2}{\eta_0^2}\sqrt{c_p}, \; \Phi^{-1}(1 - C_E/2)\sqrt{S_E^2 + \eta_0^2} \right\}$$

- intuition: need to make $\mathbb{E}[\phi(X, E)]$ large enough to implement Experiment

Which experiment studies to publish?

if the editor is constrained to implement $\Delta = \text{Experiment}$, then optimal publication decision rules satisfy

$$\phi(X, E) = \begin{cases} 1 & \text{if } |X| > X_E^* \\ 0 & \text{if } |X| < X_E^* \end{cases},$$

where

$$X_E^* = \max\left\{\frac{S_E^2 + \eta_0^2}{\eta_0^2}\sqrt{c_p},\ \Phi^{-1}(1 - C_E/2)\sqrt{S_E^2 + \eta_0^2}\right\}$$

- intuition: need to make $\mathbb{E}[\phi(X, E)]$ large enough to implement Experiment
- relevant if the researcher's IR constraint binds for $\Delta = E$xperiment

General optimal publication rule

Defn the experiment is *cheap* if $C_E \leq 2\Phi\Big(-\frac{1}{\eta_0^2}\sqrt{c_p(S_E^2+\eta_0^2)}\Big)$

- if the experiment is cheap, then optimal publication rules implement the one with $\sim$ lowest mean-squared error

General optimal publication rule

Defn the experiment is *cheap* if $C_E \leq 2\Phi\Big(-\frac{1}{\eta_0^2}\sqrt{c_p(S_E^2+\eta_0^2)}\Big)$

- if the experiment is cheap, then optimal publication rules implement the one with $\sim$ lowest mean-squared error
- if the experiment is expensive and c_p, C_E are sufficiently large then optimal publication rules implement $\Delta = O$bservational

General optimal publication rule

Defn the experiment is *cheap* if $C_E \leq 2\Phi\Big(-\frac{1}{\eta_0^2}\sqrt{c_p(S_E^2+\eta_0^2)}\Big)$

- if the experiment is cheap, then optimal publication rules implement the one with $\sim$ lowest mean-squared error
- if the experiment is expensive and c_p, C_E are sufficiently large then optimal publication rules implement $\Delta = Observational$

Asymmetric info

- Consider a cost $c_d|\beta_\Delta|$ of manipulation

General optimal publication rule

Defn the experiment is *cheap* if $C_E \leq 2\Phi\Big(-\frac{1}{\eta_0^2}\sqrt{c_p(S_E^2+\eta_0^2)}\Big)$

- if the experiment is cheap, then optimal publication rules implement the one with $\sim$ lowest mean-squared error
- if the experiment is expensive and c_p, C_E are sufficiently large then optimal publication rules implement $\Delta = Observational$

Asymmetric info

- Consider a cost $c_d|\beta_\Delta|$ of manipulation
- There exists $X^* \in \Big(\sqrt{c_p} - \frac{1}{c_d}, \sqrt{c_p}\Big)$ such that at optimum

$$\phi(X) = \begin{cases} 0 & \text{if } |X| \leq X^* \\ c_d(|X| - X^*) & \text{if } X^* < |X| < X^* + \frac{1}{c_d} \\ 1 & \text{if } |X| \geq X^* + \frac{1}{c_d} \end{cases}$$

General optimal publication rule

Defn the experiment is *cheap* if $C_E \leq 2\Phi\Big(-\frac{1}{\eta_0^2}\sqrt{c_p(S_E^2+\eta_0^2)}\Big)$

- if the experiment is cheap, then optimal publication rules implement the one with $\sim$ lowest mean-squared error
- if the experiment is expensive and c_p, C_E are sufficiently large then optimal publication rules implement $\Delta = Observational$

Asymmetric info

- Consider a cost $c_d|\beta_\Delta|$ of manipulation
- There exists $X^* \in \Big(\sqrt{c_p}-\frac{1}{c_d}, \sqrt{c_p}\Big)$ such that at optimum

$$\phi(X) = \begin{cases} 0 & \text{if } |X| \leq X^* \\ c_d(|X|-X^*) & \text{if } X^* < |X| < X^* + \frac{1}{c_d} \\ 1 & \text{if } |X| \geq X^* + \frac{1}{c_d} \end{cases}$$

- intuition: more continuous publication decision rule makes it less attractive for the researcher to manipulate the research design

Content

1. Part 1: Single hypothesis testing

2. Part 2: Multiple hypothesis testing

Organization

- Multiple hypothesis testing in economic research
- Family wise error rate and algorithms
- False discovery rate and algorithms
- FDR, q-value and application to detecting firms' discrimination
- Indexing outcomes

Relevant references

- Romano and Lehmann (2005) Ch. 9, List et al. (2019), Benjamini and Hochberg (1995), Kline et al. (2022), Viviano et al. (2021)

Multiple hypothesis testing (MHT)

- Most applied economics papers test multiple hypotheses because there are multiple treatments, subgroups, or outcomes
- Classical motivation for multiple testing adjustments
  - 100 true null hypotheses, mutually independent tests, size 5%
  - Probability of rejecting at least one true null ≈ 1
  - Separate testing does not control compound error at 5%.

Multiple hypothesis testing (MHT)

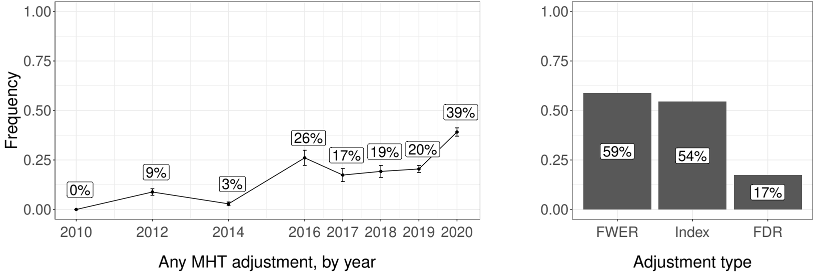
- Most applied economics papers test multiple hypotheses because there are multiple treatments, subgroups, or outcomes

- Classical motivation for multiple testing adjustments
  - 100 true null hypotheses, mutually independent tests, size 5%
  - Probability of rejecting at least one true null ≈ 1
  - Separate testing does not control compound error at 5%.

- There is substantial variation on the choice of compound error
  - Family-wise err rate (FWER): prob of rejecting at least one true null;
  - False discovery rate (FDR): expected prop/ of incorrectly rejected nulls;
  - Indexing for multiple outcomes: aggregate outcome into a single index

MHT in top-5 journals' experiments (Viviano et al., 2021)

- Variation in whether and how inferences are adjusted for MHT



A standard example in economics (Anderson, 2008)

“We begin by limiting the total number of hypotheses being tested.

A standard example in economics (Anderson, 2008)

“We begin by limiting the total number of hypotheses being tested. First, we choose a specific set of outcomes based on a priori notions of importance.

A standard example in economics (Anderson, 2008)

“We begin by limiting the total number of hypotheses being tested. First, we choose a specific set of outcomes based on a priori notions of importance.We then implement summary index tests in three broad outcome areas: preteen, adolescent, and adult. These indexes combine multiple measures to reduce the total number of tests conducted.

A standard example in economics (Anderson, 2008)

“We begin by limiting the total number of hypotheses being tested. First, we choose a specific set of outcomes based on a priori notions of importance.We then implement summary index tests in three broad outcome areas: preteen, adolescent, and adult. These indexes combine multiple measures to reduce the total number of tests conducted. Nevertheless, we still test multiple indexes.

A standard example in economics (Anderson, 2008)

"We begin by limiting the total number of hypotheses being tested. First, we choose a specific set of outcomes based on a priori notions of importance.We then implement summary index tests in three broad outcome areas: preteen, adolescent, and adult. These indexes combine multiple measures to reduce the total number of tests conducted. Nevertheless, we still test multiple indexes.Thus we adjust the p values on the summary index tests to reflect this fact."
(Anderson, 2008)

Clinical trials with multiple subgroups

"When clinically relevant differences in treatment effect are anticipated across age, racial, or ethnic groups, it is important to consider proper clinical study design, sufficient enrollment of subgroups to allow meaningful analysis, and controlling of study-wise Type 1 error for overall and subgroup-specific hypothesis testing, if appropriate and feasible." (Food and Drug Administration, 2019)

Some challenges with MHT

(1) Choice of the test

- Typically no UMP test exist:
  - ⇒ Worst-case power such as (Romano and Wolf, 2005)

$$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

for some "compound power" $\beta(\theta)$

Some challenges with MHT

(1) Choice of the test

- Typically no UMP test exist:
  - ⇒ Worst-case power such as (Romano and Wolf, 2005)

$$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

    for some "compound power" $\beta(\theta)$

  - ⇒ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$

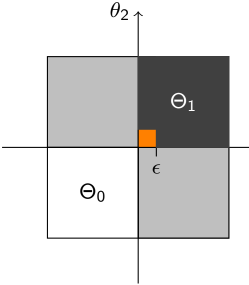
Some challenges with MHT

(1) Choice of the test

- Typically no UMP test exist:
  - ⇒ Worst-case power such as (Romano and Wolf, 2005)

$$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

  for some "compound power" $\beta(\theta)$
  - ⇒ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$



Some challenges with MHT

(1) Choice of the test

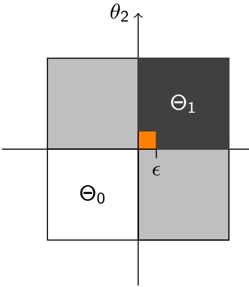
- Typically no UMP test exist:

  ⇒ Worst-case power such as (Romano and Wolf, 2005)

$$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

  for some "compound power" $\beta(\theta)$

  ⇒ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$



(2) Policy implications are often difficult: from a clinical trial

Some challenges with MHT

(1) Choice of the test

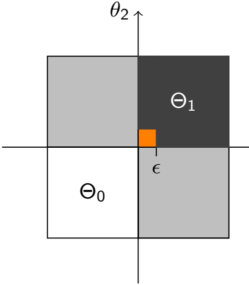
- Typically no UMP test exist:

  $\Rightarrow$ Worst-case power such as (Romano and Wolf, 2005)

  $$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

  for some "compound power" $\beta(\theta)$

  $\Rightarrow$ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$



(2) Policy implications are often difficult: from a clinical trial

"This observed heterogeneity led two regulatory agencies to different assessments.

Some challenges with MHT

(1) Choice of the test

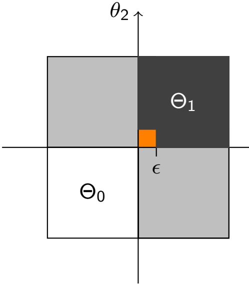
- Typically no UMP test exist:

  ⇒ Worst-case power such as (Romano and Wolf, 2005)

  $$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

  for some "compound power" $\beta(\theta)$

  ⇒ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$



(2) Policy implications are often difficult: from a clinical trial

"This observed heterogeneity led two regulatory agencies to different assessments. The National Institute for Health and Care Excellence (NICE, English and Welsh authority) concluded a clinical benefit for the overall population

Some challenges with MHT

(1) Choice of the test

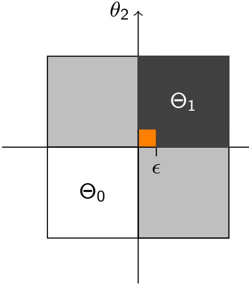
- Typically no UMP test exist:

  ⇒ Worst-case power such as (Romano and Wolf, 2005)

$$\inf_{\theta \in \Theta(\epsilon)} \beta(\theta), \quad \Theta(\epsilon) = \{\theta \geq \epsilon\}$$

  for some "compound power" $\beta(\theta)$

  ⇒ Alternatively WAP $\int \pi(\theta)\beta(\theta)d\theta$



(2) Policy implications are often difficult: from a clinical trial

"This observed heterogeneity led two regulatory agencies to different assessments. The National Institute for Health and Care Excellence (NICE, English and Welsh authority) concluded a clinical benefit for the overall population whereas the Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG, German authority) concluded efficacy only for the most beneficial subgroup of patients" (Tanniou et al., 2016)

MHT: FDR and FWER

Test/truth	Null is true	Null is false	Total
Test is significant	V	S	R
Test is non-significant	U	T	$J - R$
Total	J_0	$J - J_0$	J

FDR Control $\mathbb{E}[V/R]$ or positive control $\mathbb{E}[V/R|R > 1]$

MHT: FDR and FWER

Test/truth	Null is true	Null is false	Total
Test is significant	V	S	R
Test is non-significant	U	T	$J-R$
Total	J_0	$J-J_0$	J

FDR Control $\mathbb{E}[V/R]$ or positive control $\mathbb{E}[V/R|R>1]$

FWE (weak) Control $P\Big(V \geq 1|H_0^1, \cdots, H_0^J\Big)$

MHT: FDR and FWER

Test/truth	Null is true	Null is false	Total
Test is significant	V	S	R
Test is non-significant	U	T	$J-R$
Total	J_0	$J-J_0$	J

FDR Control $\mathbb{E}[V/R]$ or positive control $\mathbb{E}[V/R|R>1]$

FWE (weak) Control $P\Big(V \geq 1|H_0^1, \cdots, H_0^J\Big)$

FWE (strong) Control $P\Big(V \geq 1| \cdot \Big)$ over all combinations of nulls

MHT: FDR and FWER

Test/truth	Null is true	Null is false	Total
Test is significant	V	S	R
Test is non-significant	U	T	$J-R$
Total	J_0	$J-J_0$	J

FDR Control $\mathbb{E}[V/R]$ or positive control $\mathbb{E}[V/R|R>1]$

FWE (weak) Control $P\Big(V\geq 1|H_0^1,\cdots,H_0^J\Big)$

FWE (strong) Control $P\Big(V\geq 1|\cdot\Big)$ over all combinations of nulls

$\Rightarrow$ Under the global null hypothesis $V=R\Rightarrow$ [FDR $=$ weak FWER]

$$P\Big(V\geq 1|H_0^1,\cdots,H_0^J\Big)=\mathbb{E}\left[\frac{V}{R}|H_0^1,\cdots,H_0^J\right]$$

MHT: FDR and FWER

Test/truth	Null is true	Null is false	Total
Test is significant	V	S	R
Test is non-significant	U	T	$J - R$
Total	J_0	$J - J_0$	J

FDR Control $\mathbb{E}[V/R]$ or positive control $\mathbb{E}[V/R|R > 1]$

FWE (weak) Control $P\Big(V \geq 1|H_0^1, \cdots, H_0^J\Big)$

FWE (strong) Control $P\Big(V \geq 1| \cdot \Big)$ over all combinations of nulls

$\Rightarrow$ Under the global null hypothesis $V = R \Rightarrow$ [FDR $=$ weak FWER]

$$P\Big(V \geq 1|H_0^1, \cdots, H_0^J\Big) = \mathbb{E}\left[\frac{V}{R}|H_0^1, \cdots, H_0^J\right]$$

- But strong FWER more conservative than FDR $[V/R \leq 1]$

Family wise error rate: procedures

- Bonferroni: typically conservative

  ⇒ Define $p^{(j)}$ the p-value associated with the j^{th} hypothesis

$$P\left(\cup_{j=1}^{J_0} p^{(j)} \leq \frac{\alpha}{J}\right) \leq \sum_{j=1}^{J_0} P\left(p^{(j)} \leq \frac{\alpha}{J}\right) = \alpha\frac{J_0}{J} \leq \alpha$$

Family wise error rate: procedures

- Bonferroni: typically conservative

  ⇒ Define $p^{(j)}$ the p-value associated with the j^{th} hypothesis

$$P\Big(\cup_{j=1}^{J_0} p^{(j)} \leq \frac{\alpha}{J}\Big) \leq \sum_{j=1}^{J_0} P\Big(p^{(j)} \leq \frac{\alpha}{J}\Big) = \alpha \frac{J_0}{J} \leq \alpha$$

- Step down procedures [E.g., Holm (1979); Romano and Wolf (2005)]
  - Sort p-values in increasing order
  - Reject H_0^j sequentially (based on the order of the p-values)
  - Typically less conservative

Family wise error rate: procedures

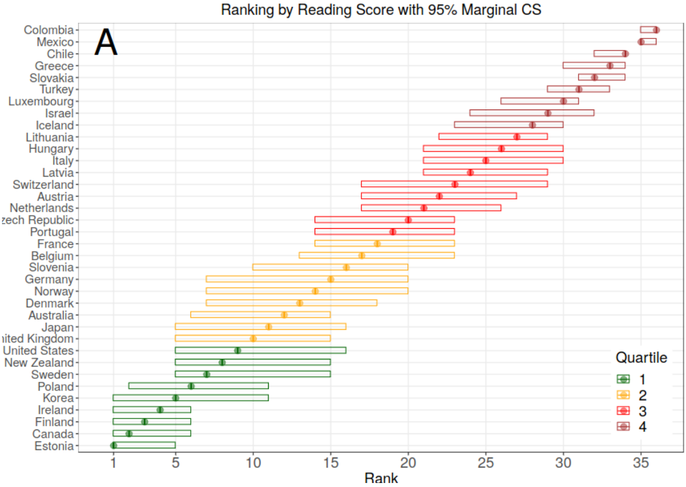
- Bonferroni: typically conservative
  - ⇒ Define $p^{(j)}$ the p-value associated with the j^{th} hypothesis

$$P\Big(\cup_{j=1}^{J_0} p^{(j)} \leq \frac{\alpha}{J}\Big) \leq \sum_{j=1}^{J_0} P\Big(p^{(j)} \leq \frac{\alpha}{J}\Big) = \alpha \frac{J_0}{J} \leq \alpha$$

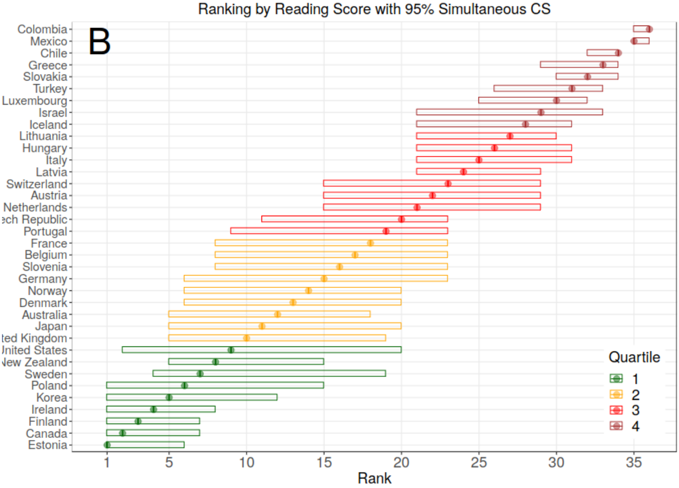
- Step down procedures [E.g., Holm (1979); Romano and Wolf (2005)]
  - Sort p-values in increasing order
  - Reject H_0^j sequentially (based on the order of the p-values)
  - Typically less conservative

- See Westfall and Young (1993), Romano and Wolf (2005), List et al. (2019) for bootstrap-based procedure
  - Adjusts for correlations (look at maximum t-stat)
  - Authors propose to control FWER_Q within a group Q of hypothesis
  - Idea of groups is that hypothesis between groups are not "related"

Marginal CI to reading scores [Mogstad et al. (2020)]



Joint (FWER) CI to reading scores [Mogstad et al. (2020)]



False Discovery Rate

Algorithms

- Benjamini and Hochberg procedure if tests are independent: rank p-values and reject if $p_j \leq \alpha j/J$
- Benjamini Yekutieli for dependence: add additional penalty
- Some recent work also for sequential testing (Robertson et al., 2023)

False Discovery Rate

Algorithms

- Benjamini and Hochberg procedure if tests are independent: rank p-values and reject if $p_j \leq \alpha j/J$
- Benjamini Yekutieli for dependence: add additional penalty
- Some recent work also for sequential testing (Robertson et al., 2023)

Properties

- Less conservative + admits Bayesian interpretation
  - Suppose that we have J hypothesis, each true with $H_i^0 \sim_{i.i.d.} \mathrm{Bern}(\pi)$
  - Tests are distributed $T_i \sim H_i F_0 + (1 - H_i) F_1$
  - Then (p)FDR $= \mathbb{E}[H = 0|\text{hypothesis is rejected}]$! (Storey, 2003)
  - Therefore (p)FDR capture posterior probability of rejecting the null

FDR and detecting discrimination (Kline et al., 2022)

- Kline et al. sent resume to firms f to detect discrimination
- They build a p-value for testing for each firm whether they discriminate (zero "contact-gaps")
- They estimate the distribution of p-values

FDR and detecting discrimination (Kline et al., 2022)

- Kline et al. sent resume to firms f to detect discrimination
- They build a p-value for testing for each firm whether they discriminate (zero "contact-gaps")
- They estimate the distribution of p-values
- Suppose there is a cost of auditing a possibly discriminatory firm $\Rightarrow$ which firm should we audit?

FDR and detecting discrimination (Kline et al., 2022)

- Kline et al. sent resume to firms f to detect discrimination
- They build a p-value for testing for each firm whether they discriminate (zero "contact-gaps")
- They estimate the distribution of p-values

- Suppose there is a cost of auditing a possibly discriminatory firm ⇒ which firm should we audit?
  - ⇒ Find those with smallest q-value

FDR and detecting discrimination (Kline et al., 2022)

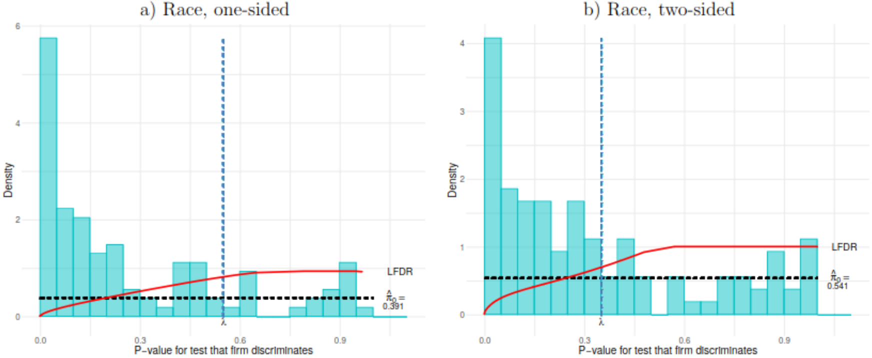
- Kline et al. sent resume to firms f to detect discrimination
- They build a p-value for testing for each firm whether they discriminate (zero "contact-gaps")
- They estimate the distribution of p-values
- Suppose there is a cost of auditing a possibly discriminatory firm $\Rightarrow$ which firm should we audit?
  - $\Rightarrow$ Find those with smallest q-value
  - $\Rightarrow$ 23 firms have q-values less than 0.05 $\Rightarrow$ in expectation only one does not discriminate

FDR and detecting discrimination (Kline et al., 2022)

- Kline et al. sent resume to firms f to detect discrimination
- They build a p-value for testing for each firm whether they discriminate (zero "contact-gaps")
- They estimate the distribution of p-values
- Suppose there is a cost of auditing a possibly discriminatory firm ⇒ which firm should we audit?
  - ⇒ Find those with smallest q-value
  - ⇒ 23 firms have q-values less than 0.05 ⇒ in expectation only one does not discriminate
  - ⇒ Choose q-value to balance benefits/costs of auditing
  - ⇒ This is equivalent to appropriately threshold p-values

FDR control across discriminatory firms Kline et al. (2022)

Figure 10: P-value distributions and local false discovery rates



a) Race, one-sided

b) Race, two-sided

Ok...but what about multiple outcomes

- From a decision-theoretic perspective...tricky
- With multiple treatments – there is mapping betw/ tests and decisions
- With many outcomes, this becomes unclear

Ok...but what about multiple outcomes

- From a decision-theoretic perspective...tricky
- With multiple treatments – there is mapping betw/ tests and decisions
- With many outcomes, this becomes unclear
- Practice in economics:
  - build families of outcomes
  - within each family construct an indexes (typically statistical)
  - Then correct within families (sometimes betw/)

Ok...but what about multiple outcomes

- From a decision-theoretic perspective...tricky
- With multiple treatments – there is mapping betw/ tests and decisions
- With many outcomes, this becomes unclear

- Practice in economics:
    - build families of outcomes
    - within each family construct an indexes (typically statistical)
    - Then correct within families (sometimes betw/)

- FDA multiple end-points (Food and Drug Administration, 2019)
    1. "When Demonstration of Treatment Effects on All of Two or More Distinct Endpoints Is Necessary to Establish Clinical Benefit (Co-Primary Endpoints)"
    2. "When Demonstration of a Treatment Effect on at Least One of Several Primary Endpoints Is Sufficient"

Ok...but what about multiple outcomes

- From a decision-theoretic perspective...tricky
- With multiple treatments – there is mapping betw/ tests and decisions
- With many outcomes, this becomes unclear

- Practice in economics:
  - build families of outcomes
  - within each family construct an indexes (typically statistical)
  - Then correct within families (sometimes betw/)

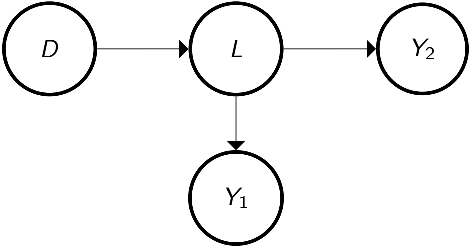
- FDA multiple end-points (Food and Drug Administration, 2019)
  (1) "When Demonstration of Treatment Effects on All of Two or More Distinct Endpoints Is Necessary to Establish Clinical Benefit (Co-Primary Endpoints)"
  (2) "When Demonstration of a Treatment Effect on at Least One of Several Primary Endpoints Is Sufficient"
  ⇒ For (1) α-tests can be conservative and for (2) why separate testing?

Why indexing? An illustrative description

- Consider a binary treatment D and two outcomes (Y_1, Y_2)
- Consider the following model with latent factor L

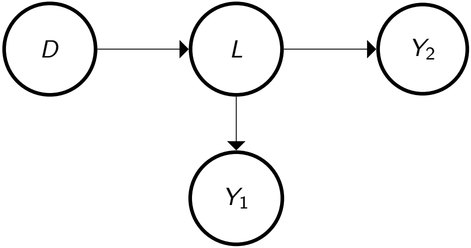
Why indexing? An illustrative description

- Consider a binary treatment D and two outcomes (Y_1, Y_2)
- Consider the following model with latent factor L



Why indexing? An illustrative description

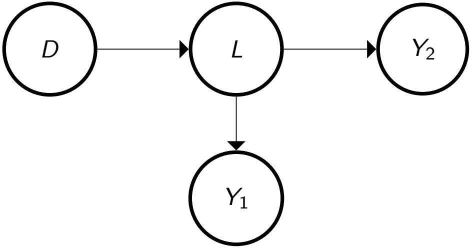
- Consider a binary treatment D and two outcomes (Y_1, Y_2)
- Consider the following model with latent factor L



- Index for outcomes with same factor L (e.g. Ludwig et al. (2017))

Why indexing? An illustrative description

- Consider a binary treatment D and two outcomes (Y_1, Y_2)
- Consider the following model with latent factor L



- Index for outcomes with same factor L (e.g. Ludwig et al. (2017))
- But ... statistical indexing is not always optimal
  - Suppose our utility is $\mathbb{E}_\theta[u(Y_1, Y_2)]$
  - Then I should account for our preferences (Viviano et al., 2021)
  - Some recent examples in Bhatt et al. (2024) and Give Directly

Summary

- We reviewed notions of MHT
- We have connected this to recent works in economics
- We have discussed some of the decision-theoretic interpretations

What is coming next

- Decision-theoretic justification of different notions of compound error?
- How to incorporate incentives of researchers?
- How to incorporate different nature of multiplicity?
- When to adjust inference for multiplicity?

  E.g. Should we adjust inference across all papers we write?

What is coming next

- Decision-theoretic justification of different notions of compound error?
- How to incorporate incentives of researchers?
- How to incorporate different nature of multiplicity?
- When to adjust inference for multiplicity?
  E.g. Should we adjust inference across all papers we write?

Tomorrow I will present a model of multiple hypothesis testing addressing some of these questions in a joint work with Wuthrich and Niehaus

What is coming next

- Decision-theoretic justification of different notions of compound error?
- How to incorporate incentives of researchers?
- How to incorporate different nature of multiplicity?
- When to adjust inference for multiplicity?

  E.g. Should we adjust inference across all papers we write?

Tomorrow I will present a model of multiple hypothesis testing addressing some of these questions in a joint work with Wuthrich and Niehaus

Thanks!

Abadie, A., 2020. Statistical nonsignificance in empirical economics. American Economic Review: Insights 2, 193–208.

Anderson, M.L., 2008. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. Journal of the American statistical Association 103, 1481–1495.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

Bhatt, M.P., Heller, S.B., Kapustin, M., Bertrand, M., Blattman, C., 2024. Predicting and preventing gun violence: An experimental evaluation of READI chicago. The Quarterly Journal of Economics 139, 1–56.

Chaudhuri, S.E., Lo, A.W., 2020. Financially adaptive clinical trials via option pricing analysis. Journal of econometrics , 105026.

Fisher, R., 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society Series B: Statistical Methodology 17, 69–78.

Food and Drug Administration, 2019. Demonstrating substantial evidence of effectiveness for human drug and biological products guidance for industry. https://www.fda.gov/media/133660/download.

Frankel, A., Kasy, M., 2022. Which findings should be published? American Economic Journal: Microeconomics 14, 1–38.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics , 65–70.

Isakov, L., Lo, A.W., Montazerhodjat, V., 2019. Is the fda too conservative or too aggressive?: A bayesian decision analysis of clinical trial design. Journal of econometrics 211, 117–136.

Kline, P., Rose, E.K., Walters, C.R., 2022. Systemic discrimination among large us employers. The Quarterly Journal of Economics 137, 1963–2036.

Lenhard, J., 2006. Models and statistical inference: The controversy between fisher and neyman–pearson. The British journal for the philosophy of science .

List, J.A., Shaikh, A.M., Xu, Y., 2019. Multiple hypothesis testing in experimental economics. Experimental Economics 22, 773–793.

Ludwig, J., Mullainathan, S., Spiess, J., 2017. Machine-learning tests for effects on multiple outcomes. arXiv preprint arXiv:1707.01473 .

Manski, C.F., 2004. Statistical treatment rules for heterogeneous populations. Econometrica 72, 1221–1246.

Mogstad, M., Romano, J.P., Shaikh, A., Wilhelm, D., 2020. Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. Technical Report. National Bureau of Economic Research.

Neyman, J., Iwaszkiewicz, K., 1935. Statistical problems in agricultural experimentation. Supplement to the Journal of the Royal Statistical Society 2, 107–180.

Neyman, J., Pearson, E.S., 1933. Ix. on the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231, 289–337.

Robertson, D.S., Choodari-Oskooei, B., Dimairo, M., Flight, L., Pallmann, P., Jaki, T., 2023. Point estimation for adaptive trial designs i: A methodological review. Statistics in medicine 42, 122–145.

Romano, J.P., Lehmann, E., 2005. Testing statistical hypotheses.

Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237–1282.

Simon, H.A., 1945. Statistical tests as a basis for "yes-no" choices. Journal of the American Statistical Association 40, 80–84.

Storey, J.D., 2003. The positive false discovery rate: a bayesian interpretation and the q-value. The annals of statistics 31, 2013–2035.

Tanniou, J., Van Der Tweel, I., Teerenstra, S., Roes, K.C., 2016. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. BMC medical research methodology 16, 1–15.

Tetenov, A., 2012. Statistical treatment choice based on asymmetric minimax regret criteria. Journal of Econometrics 166, 157–165.

Tetenov, A., 2016. An economic theory of statistical testing. Technical Report. cemmap working paper.

Viviano, D., Wuthrich, K., Niehaus, P., 2021. (when) should you adjust inferences for multiple hypothesis testing? arXiv preprint arXiv:2104.13367 .

Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11, 147–162.

Westfall, P.H., Young, S.S., 1993. Resampling-based multiple testing: Examples and methods for p-value adjustment. volume 279. John Wiley & Sons.

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Proof First, note that if we falsely reject $k-1$ hypothesis, it must be that $k-1 \leq J - J_0$. Therefore $\frac{1}{J-k+1} \leq \frac{1}{J_0}$.

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Proof First, note that if we falsely reject $k-1$ hypothesis, it must be that $k-1 \leq J - J_0$. Therefore $\frac{1}{J-k+1} \leq \frac{1}{J_0}$.

$$P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J+1-j}\right)$$

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Proof First, note that if we falsely reject $k-1$ hypothesis, it must be that $k - 1 \leq J - J_0$. Therefore $\frac{1}{J-k+1} \leq \frac{1}{J_0}$.

$$P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J+1-j}\right) \leq P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J_0}\right)$$

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Proof First, note that if we falsely reject $k-1$ hypothesis, it must be that $k-1 \leq J - J_0$. Therefore $\frac{1}{J-k+1} \leq \frac{1}{J_0}$.

$$P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J+1-j}\right) \leq P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J_0}\right) \leq \sum_{j \in J_0} P\left(p_{(j)} \leq \frac{\alpha}{J_0}\right)$$

Step down procedures for FWER: Holm's (1979)

- Order p-values $p_{(1)}, p_{(2)}, \cdots$
- Let k the maximal index so that $p_{(k)} \geq \frac{\alpha}{J+1-k}$
- Reject all null for $k' < k$

Thm Holm's procedure control FWER at level α

Proof First, note that if we falsely reject $k-1$ hypothesis, it must be that $k-1 \leq J - J_0$. Therefore $\frac{1}{J-k+1} \leq \frac{1}{J_0}$.

$$P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J+1-j}\right) \leq P\left(\cup_{j \in J_0} p_{(j)} \leq \frac{\alpha}{J_0}\right) \leq \sum_{j \in J_0} P\left(p_{(j)} \leq \frac{\alpha}{J_0}\right)$$

$\Rightarrow$ No assumption on the dependence because using union bound

Resampling (Romano and Wolf, 2005)

- Romano and Wolf (and Westfall and Young (1993)) replace the union bound by using the resampling method to get the correlation

Resampling (Romano and Wolf, 2005)

- Romano and Wolf (and Westfall and Young (1993)) replace the union bound by using the resampling method to get the correlation

- The idea is to use the test stat $\hat{\theta}_j/\hat{\sigma}_j$, and reject sequentially
  - Rank test stat from largest to smallest
  - Define $\hat{c}(j, R)$ the critical value of largest test-stat j after having reject R hypotheses (computed via re-sampling)
  - Reject sequentially based on stat

FDR and q value (Storey, 2003)

Th Consider J independent test, and hypotheses (T_i, H_i), $T_i \sim H_i F_0 + (1 - H_i) F_1$, each true if $H_i = 1$, $H_i \sim_{i.i.d.} \mathrm{Bern}(\pi)$. Consider a rejection region Γ. Then

$$\mathrm{(p)FDR}(\Gamma) = \mathbb{E}[H = 0 | T \in \Gamma]$$

FDR and q value (Storey, 2003)

Th Consider J independent test, and hypotheses (T_i, H_i), $T_i \sim H_i F_0 + (1 - H_i) F_1$, each true if $H_i = 1$, $H_i \sim_{i.i.d.} \mathrm{Bern}(\pi)$. Consider a rejection region Γ. Then

$$\mathrm{(p)FDR}(\Gamma) = \mathbb{E}[H = 0 | T \in \Gamma]$$

Proof Let $p_k = P(R(\Gamma) = k | R(\Gamma) > 0)$

$$\mathrm{(p)FDR}(\Gamma) = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{R(\Gamma)} | R(\Gamma) = k\Big] p_k$$

FDR and q value (Storey, 2003)

Th Consider J independent test, and hypotheses (T_i, H_i), $T_i \sim H_i F_0 + (1 - H_i) F_1$, each true if $H_i = 1$, $H_i \sim_{i.i.d.} \mathrm{Bern}(\pi)$. Consider a rejection region Γ. Then

$$\mathrm{(p)FDR}(\Gamma) = \mathbb{E}[H = 0 | T \in \Gamma]$$

Proof Let $p_k = P(R(\Gamma) = k | R(\Gamma) > 0)$

$$\mathrm{(p)FDR}(\Gamma) = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{R(\Gamma)} | R(\Gamma) = k\Big] p_k = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{k} | R(\Gamma) = k\Big] p_k$$

FDR and q value (Storey, 2003)

Th Consider J independent test, and hypotheses (T_i, H_i), $T_i \sim H_i F_0 + (1 - H_i) F_1$, each true if $H_i = 1$, $H_i \sim_{i.i.d.} \mathrm{Bern}(\pi)$. Consider a rejection region Γ. Then

$$\mathrm{(p)FDR}(\Gamma) = \mathbb{E}[H = 0 | T \in \Gamma]$$

Proof Let $p_k = P(R(\Gamma) = k | R(\Gamma) > 0)$

$$\mathrm{(p)FDR}(\Gamma) = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{R(\Gamma)} | R(\Gamma) = k\Big] p_k = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{k} | R(\Gamma) = k\Big] p_k$$

$$= \sum_{k=1}^{J} \frac{1}{k} \mathbb{E}\Big[1\{T_k \in \Gamma\} 1\{H_k = 0\} | T_{1:k} \in \Gamma, T_{(k+1):J} \notin \Gamma\Big] p_k$$

FDR and q value (Storey, 2003)

Th Consider J independent test, and hypotheses (T_i, H_i), $T_i \sim H_i F_0 + (1 - H_i) F_1$, each true if $H_i = 1$, $H_i \sim_{i.i.d.} \mathrm{Bern}(\pi)$. Consider a rejection region Γ. Then

$$\mathrm{(p)FDR}(\Gamma) = \mathbb{E}[H = 0 | T \in \Gamma]$$

Proof Let $p_k = P(R(\Gamma) = k | R(\Gamma) > 0)$

$$\begin{aligned}\mathrm{(p)FDR}(\Gamma) &= \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{R(\Gamma)} | R(\Gamma) = k\Big] p_k = \sum_{k=1}^{J} \mathbb{E}\Big[\frac{V(\Gamma)}{k} | R(\Gamma) = k\Big] p_k \\ &= \sum_{k=1}^{J} \frac{1}{k} \mathbb{E}\Big[1\{T_k \in \Gamma\} 1\{H_k = 0\} | T_{1:k} \in \Gamma, T_{(k+1):J} \not\in \Gamma\Big] p_k \\ &= \mathbb{E}\Big[1\{H_i = 0\} | T_i \in \Gamma\Big] \sum_{k=1}^{J} p_k = \mathbb{E}[H = 0 | T \in \Gamma]\end{aligned}$$