A model of multiple hypothesis testing

Davide Viviano (Harvard)   Kaspar Wüthrich (UCSD)   Paul Niehaus (UCSD)

Multiple hypothesis testing (MHT): Some Recap

- Multiple hypotheses because of multiple treatments, subgroups, or outcomes

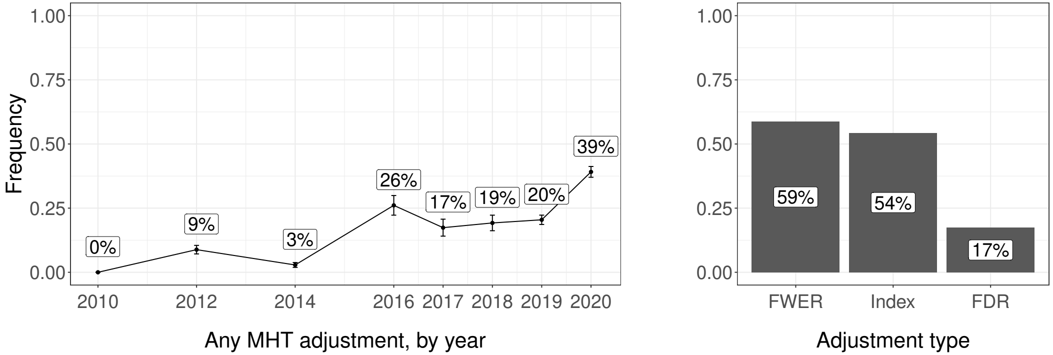
Multiple hypothesis testing (MHT): Some Recap

- Multiple hypotheses because of multiple treatments, subgroups, or outcomes
- Classical motivation for multiple testing adjustments
  - 100 true null hypotheses, mutually independent tests, size = level $= \alpha = 5\%$
  - Probability of rejecting at least one true null hypothesis $= 1 - 0.95^{100} = 0.994$
  - Separate testing generally does not control notions of compound error at 5%.

Multiple hypothesis testing (MHT): Some Recap

- Multiple hypotheses because of multiple treatments, subgroups, or outcomes
- Classical motivation for multiple testing adjustments
  - 100 true null hypotheses, mutually independent tests, size = level $= \alpha = 5\%$
  - Probability of rejecting at least one true null hypothesis $= 1 - 0.95^{100} = 0.994$
  - Separate testing generally does not control notions of compound error at 5%.
- There is substantial variation on the choice of compound error and/or tests
  - Family-wise error rate (FWER): probability of rejecting at least one true null;
  - False discovery rate (FDR): expected proportion of incorrectly rejected null hypotheses;
  - Indexing: aggregate outcomes into a single index [e.g., Anderson (2008)].
- Several algorithms to control compound errors (e.g., Bonferroni correction)

Policy experiments with multiple treatments in Economics: top-5 journals



Heterogeneity in when and how to adjust inference.

This paper

- Study whether/how to adjust inferences in experiments based on the economics

  E.g., nature of multiplicity, research costs, welfare/policy implications of hypothesis testing

This paper

- Study whether/how to adjust inferences in experiments based on the economics

  E.g., nature of multiplicity, research costs, welfare/policy implications of hypothesis testing

- Take into account the researchers' incentives based on three core ideas:
  1. Research is a public good, and policy decisions are influenced by hypothesis tests
  2. Research costs are born privately by the researcher, who decides to experiment
  3. The regulator can ex-ante enforce hypothesis testing protocols

This paper

- Study whether/how to adjust inferences in experiments based on the economics

  E.g., nature of multiplicity, research costs, welfare/policy implications of hypothesis testing

- Take into account the researchers' incentives based on three core ideas:
  1. Research is a public good, and policy decisions are influenced by hypothesis tests
  2. Research costs are born privately by the researcher, who decides to experiment
  3. The regulator can ex-ante enforce hypothesis testing protocols

- **Goal:** characterize protocol that maximizes worst-case welfare over policy effects subject to trade-off (a) motivating research and (b) generating bad policy guidance

This paper

- Study whether/how to adjust inferences in experiments based on the economics

  E.g., nature of multiplicity, research costs, welfare/policy implications of hypothesis testing

- Take into account the researchers' incentives based on three core ideas:
  1. Research is a public good, and policy decisions are influenced by hypothesis tests
  2. Research costs are born privately by the researcher, who decides to experiment
  3. The regulator can ex-ante enforce hypothesis testing protocols

- **Goal:** characterize protocol that maximizes worst-case welfare over policy effects subject to trade-off (a) motivating research and (b) generating bad policy guidance

- Model's assumptions can justify single-hypothesis testing [Tetenov (2016)]

We study MHT for different types of multiplicity

Key feature of our model: hypothesis tests correspond to policy decisions

- **Multiple treatments (or subgroups):** simple mapping betw/ tests and decisions
- **Multiple outcomes**: might or might not interpret as informing multiple decisions
  - Research informs a single policy decision (e.g., whether to scale up an intervention)
  - Research informs multiple heterogeneous policy-makers

Literature

- Economic analysis of optimal statistical approaches [E.g., Chassang et al. (2012); Tetenov (2016); Spiess (2018); Henry and Ottaviani (2019); Di Tillio et al. (2017); Kasy and Spiess (2023)]
  - We focus on MHT
- Models of scientific communication [E.g., Frankel and Kasy (2022); Andrews and Shapiro (2021); Banerjee et al. (2017)]
  - We relate the structure of the scientific process to MHT
- Work on decision theory and hypothesis testing [E.g., Wald (1950); Robbins (1951); Storey (2003); Lehmann and Romano (2005); Efron (2008)]
  - We provide an economic model with incentives that allows for discriminating between different MHT procedures. We show when MHT is optimal and when it is not.
- Statistical methods for MHT corrections [E.g., Holm (1979); Westfall and Young (1993); Benjamini and Hochberg (1995); Romano et al. (2010)]
  - We provide guidance for choosing appropriate methods

Multiple treatments (interventions or subgroups)

Players and stakeholders' welfare

- **Representative researcher**
  - decides whether to run a pre-specified experiment with J treatments

Players and stakeholders' welfare

- **Representative researcher**
  - decides whether to run a pre-specified experiment with J treatments
  - If she experiments, she draws a vector of statistics $X \sim F_\theta$, where $\theta = (\theta_1, \ldots, \theta_J)^\top$

Players and stakeholders' welfare

- **Representative researcher**
  - decides whether to run a pre-specified experiment with J treatments
  - If she experiments, she draws a vector of statistics $X \sim F_\theta$, where $\theta = (\theta_1, \ldots, \theta_J)^\top$
  - She then reports discoveries/rejections $r(X) = (r_1(X), \ldots, r_J(X))$, where $r_j(X) = 1$ if treatment j is recommended, and $r_j(X) = 0$ otherwise

Players and stakeholders' welfare

- **Representative researcher**
  - decides whether to run a pre-specified experiment with J treatments
  - If she experiments, she draws a vector of statistics $X \sim F_\theta$, where $\theta = (\theta_1, \ldots, \theta_J)^\top$
  - She then reports discoveries/rejections $r(X) = (r_1(X), \ldots, r_J(X))$, where $r_j(X) = 1$ if treatment j is recommended, and $r_j(X) = 0$ otherwise

- **Social planner** chooses a testing procedure r^* to maximize worst-case welfare

  Ex FDA approval agency (or editorial standards in economics)

Players and stakeholders' welfare

- **Representative researcher**
  - decides whether to run a pre-specified experiment with J treatments
  - If she experiments, she draws a vector of statistics $X \sim F_\theta$, where $\theta = (\theta_1, \ldots, \theta_J)^\top$
  - She then reports discoveries/rejections $r(X) = (r_1(X), \ldots, r_J(X))$, where $r_j(X) = 1$ if treatment j is recommended, and $r_j(X) = 0$ otherwise

- **Social planner** chooses a testing procedure r^* to maximize worst-case welfare

  Ex FDA approval agency (or editorial standards in economics)

$\Rightarrow$ **Policy implementation**
  - upon experimentation, additive welfare effects $\theta^\top r(X)$ on stakeholders (no spillovers)
  - In the paper: general model with interactions

Example

Ex **Parameters of interest:** the researcher evaluates J treatments $D_1, \ldots, D_J$ using

$$Y = \theta_1 D_1 + \cdots + \theta_J D_J + \varepsilon$$

- Here: $X = (\hat{\theta}_1, \ldots, \hat{\theta}_J)^\top$, F_θ is the CDF of a $\mathcal{N}(\theta, \Sigma)$ distribution, and Σ is known

Example

Ex **Parameters of interest:** the researcher evaluates J treatments $D_1, \dots, D_J$ using

$$Y = \theta_1 D_1 + \cdots + \theta_J D_J + \varepsilon$$

- Here: $X = (\hat{\theta}_1, \dots, \hat{\theta}_J)^\top$, F_θ is the CDF of a $\mathcal{N}(\theta, \Sigma)$ distribution, and Σ is known

Ex **Testing protocol (t-test):** $r_j(X) = 1\{X_j/\sqrt{\Sigma_{j,j}} > t\}$ for $j = 1, \dots, J$

$\Rightarrow$ In the paper general testing protocol

Game

Stage 1: the social planner, who doesn't know θ, chooses r to maximize worst-case welfare: for $\lambda \geq 0, \pi \in \Pi$

$$r^* \in \arg\max_r \underbrace{\min_{\theta \in \Theta} v_r(\theta)}_{\text{ambiguity aversion}} + \lambda \underbrace{\int e_r(\theta')\pi(\theta')d\theta'}_{\text{subjective utility}}$$

Game

Stage 1: the social planner, who doesn't know θ, chooses r to maximize worst-case welfare: for $\lambda \geq 0, \pi \in \Pi$

$$r^* \in \arg\max_r \underbrace{\min_{\theta \in \Theta} v_r(\theta)}_{\text{ambiguity aversion}} + \lambda \underbrace{\int e_r(\theta')\pi(\theta')d\theta'}_{\text{subjective utility}}$$

where

$$\Big(v_r(\theta), e_r(\theta)\Big) = \begin{cases} \Big(\int \theta^\top r(x) dF_\theta(x), 1\Big) & \text{if the researcher experiments,} \\ & \end{cases}$$

Game

Stage 1: the social planner, who doesn't know θ, chooses r to maximize worst-case welfare: for $\lambda \geq 0, \pi \in \Pi$

$$r^* \in \arg\max_r \underbrace{\min_{\theta \in \Theta} v_r(\theta)}_{\text{ambiguity aversion}} + \lambda \underbrace{\int e_r(\theta')\pi(\theta')d\theta'}_{\text{subjective utility}}$$

where

$$\Big(v_r(\theta), e_r(\theta)\Big) = \begin{cases} \left(\int \theta^\top r(x) dF_\theta(x), 1\right) & \text{if the researcher experiments,} \\ (0,0) & \text{if the researcher doesn't experiment} \end{cases}$$

Game

Stage 1: the social planner, who doesn't know θ, chooses r to maximize worst-case welfare: for $\lambda \geq 0, \pi \in \Pi$

$$r^* \in \arg\max_r \underbrace{\min_{\theta \in \Theta} v_r(\theta)}_{\text{ambiguity aversion}} + \lambda \underbrace{\int e_r(\theta')\pi(\theta')d\theta'}_{\text{subjective utility}}$$

where

$$\Big(v_r(\theta), e_r(\theta)\Big) = \begin{cases} \Big(\int \theta^\top r(x)dF_\theta(x), 1\Big) & \text{if the researcher experiments,} \\ (0,0) & \text{if the researcher doesn't experiment} \end{cases}$$

Stage 2: given r, the researcher, who knows θ (can be relaxed), experiments if her expected utility $\beta_r(\theta)$ is positive, where

$$\beta_r(\theta) = \underbrace{\int \sum_{j=1}^{J} r_j(x)dF_\theta(x)}_{\substack{\text{benefit from approval} \\ \text{(expected number of rejections)}}} - \underbrace{C(J,\Sigma)}_{\text{research costs relative to benefits}}$$

(I will write $C(J)$ implicit as a function of Σ)

Returning to the FDA example

- Social planner = FDA

Returning to the FDA example

- Social planner = FDA
  - $\Rightarrow$ Ambiguity aversion = “primum non nuocere”

Returning to the FDA example

- Social planner = FDA
  - $\Rightarrow$ Ambiguity aversion = "primum non nuocere"
  - $\Rightarrow$ Utility from experimentation = "the rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out exploratory studies" (Lewis, 1999)

Returning to the FDA example

- Social planner = FDA
  - ⇒ Ambiguity aversion = "primum non nuocere"
  - ⇒ Utility from experimentation = "the rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out exploratory studies" (Lewis, 1999)
- Researcher = firm producing drugs
  - ⇒ Pre-specification typically recommended
  - ⇒ Interpret $\int \sum_j r_j(x) dF_\theta(x)$ as proportional to profits from approval
  - ⇒ Phase 3 trials costs are betw/ 14 - 50 millions \$ (Wong et al., 2014)

Returning to the FDA example

- Social planner = FDA
  - ⇒ Ambiguity aversion = “primum non nuocere”
  - ⇒ Utility from experimentation = “the rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out exploratory studies” (Lewis, 1999)
- Researcher = firm producing drugs
  - ⇒ Pre-specification typically recommended
  - ⇒ Interpret $\int \sum_j r_j(x) dF_\theta(x)$ as proportional to profits from approval
  - ⇒ Phase 3 trials costs are betw/ 14 - 50 millions $ (Wong et al., 2014)
- Some extensions (main conclusions unchanged):
  - Sub-populations have varying size
  - The researcher has a prior over θ
  - Endogenous choice of which treatment to test (and J), but pre-specify

Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \;\; \forall \theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \;\; \forall \theta \in \Theta \setminus \Theta_0$

Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \;\; \forall \theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \;\; \forall \theta \in \Theta \setminus \Theta_0$

  - Connection to (weak) size control ($J = 2$):

    (a) $\iff \underbrace{P(r_1^*(X) = 1|\theta) + P(r_2^*(X) = 1|\theta)}_{\text{benefit from approval}} \leq \underbrace{C(2)}_{\text{costs}} \;\; \forall \theta \in \Theta_0$

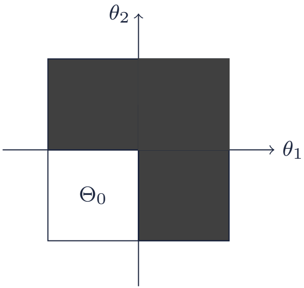
Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \;\; \forall \theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \;\; \forall \theta \in \Theta \setminus \Theta_0$

  - Connection to (weak) size control ($J = 2$):

    (a) $\iff \underbrace{P(r_1^*(X) = 1|\theta) + P(r_2^*(X) = 1|\theta)}_{\text{benefit from approval}} \leq \underbrace{C(2)}_{\text{costs}} \;\; \forall \theta \in \Theta_0$



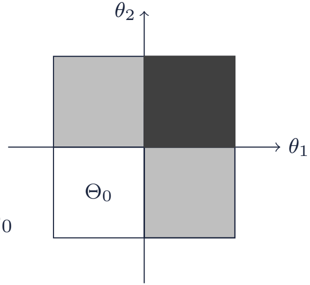
Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \ \ \forall \theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \ \ \forall \theta \in \Theta \setminus \Theta_0$

  - Connection to (weak) size control ($J = 2$):

    (a) $\iff$ $\underbrace{P(r_1^*(X) = 1|\theta) + P(r_2^*(X) = 1|\theta)}_{\text{benefit from approval}} \leq \underbrace{C(2, \Sigma)}_{\text{costs}} \ \ \forall \theta \in \Theta_0$



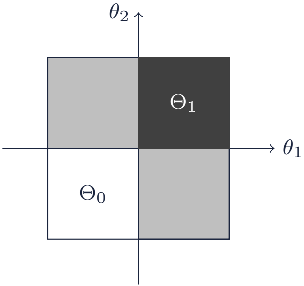
Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \;\; \forall \theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \;\; \forall \theta \in \Theta \setminus \Theta_0$

  - Connection to (weak) size control ($J = 2$):

    (a) $\iff$ $\underbrace{P(r_1^*(X) = 1|\theta) + P(r_2^*(X) = 1|\theta)}_{\text{benefit from approval}} \leq \underbrace{C(2)}_{\text{costs}} \;\; \forall \theta \in \Theta_0$



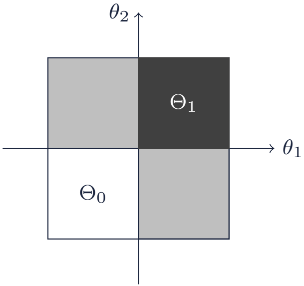
Characterization of maximin protocols ($\lambda = 0$)

- **Proposition:** r^* is maximin optimal if and only if

  (a) $\beta_{r^*}(\theta) \leq 0 \quad \forall\theta \in \Theta_0$ and (b) $v_{r^*}(\theta) \geq 0 \quad \forall\theta \in \Theta \setminus \Theta_0$

  - Connection to (weak) size control ($J = 2$):

  (a) $\iff$ $\underbrace{P(r_1^*(X) = 1|\theta) + P(r_2^*(X) = 1|\theta)}_{\text{benefit from approval}} \leq \underbrace{C(2)}_{\text{costs}} \quad \forall\theta \in \Theta_0$



- **Intuition**
  - When $\theta \in \Theta_0$, research has only downside $\Rightarrow$ keep approval probability low
  - When the cost doesn't depend on J, this condition will be violated for large enough J

Optimal protocols ($\lambda \geq 0$)

- There are many maximin protocols (including $r_j(X) = 0$ for all j).
- **Proposition:** for $J > 1$, no maximin recommendation function leads to higher welfare for all θ than all other maximin recommendation functions (no UMP test)

  $\implies$ have to choose a suitable notion of power

Optimal protocols ($\lambda \geq 0$)

- There are many maximin protocols (including $r_j(X) = 0$ for all j).
- **Proposition:** for $J > 1$, no maximin recommendation function leads to higher welfare for all θ than all other maximin recommendation functions (no UMP test)

  $\implies$ have to choose a suitable notion of power
- **Proposition**: Consider subjective priors $\pi \in \Pi$ with support on $\Theta_1 = [0, 1]^J$. Then

$$r^* \in \arg\max_{r \in \mathcal{R}} \left\{ \min_{\theta \in \Theta} v_r(\theta) + \lambda \int_{\Theta_1} e_r^*(\theta)\pi(\theta)d\theta \right\},$$

  for all $\lambda \geq 0$ and $\pi \in \Pi$ if and only if

Optimal protocols ($\lambda \geq 0$)

- There are many maximin protocols (including $r_j(X) = 0$ for all j).
- **Proposition:** for $J > 1$, no maximin recommendation function leads to higher welfare for all θ than all other maximin recommendation functions (no UMP test)

  $\Longrightarrow$ have to choose a suitable notion of power
- **Proposition**: Consider subjective priors $\pi \in \Pi$ with support on $\Theta_1 = [0, 1]^J$. Then

$$r^* \in \arg\max_{r \in \mathcal{R}} \left\{ \min_{\theta \in \Theta} v_r(\theta) + \lambda \int_{\Theta_1} e_r^*(\theta) \pi(\theta) d\theta \right\},$$

  for all $\lambda \geq 0$ and $\pi \in \Pi$ if and only if

  (i) r^* is maximin optimal ($\Rightarrow$ size control/non-negative welfare)

Optimal protocols ($\lambda \geq 0$)

- There are many maximin protocols (including $r_j(X) = 0$ for all j).
- **Proposition:** for $J > 1$, no maximin recommendation function leads to higher welfare for all θ than all other maximin recommendation functions (no UMP test)

  $\Longrightarrow$ have to choose a suitable notion of power
- **Proposition**: Consider subjective priors $\pi \in \Pi$ with support on $\Theta_1 = [0,1]^J$. Then

$$r^* \in \arg\max_{r \in \mathcal{R}} \left\{ \min_{\theta \in \Theta} v_r(\theta) + \lambda \int_{\Theta_1} e_r^*(\theta) \pi(\theta) d\theta \right\},$$

for all $\lambda \geq 0$ and $\pi \in \Pi$ if and only if

(i) r^* is maximin optimal ($\Rightarrow$ size control/non-negative welfare)

(ii) $\beta_{r^*}(\theta) \geq 0$ for all $\theta \in \Theta_1$ ($\Rightarrow$ unbiased test/power)

Optimal protocols ($\lambda \geq 0$)

- There are many maximin protocols (including $r_j(X) = 0$ for all j).
- **Proposition:** for $J > 1$, no maximin recommendation function leads to higher welfare for all θ than all other maximin recommendation functions (no UMP test)

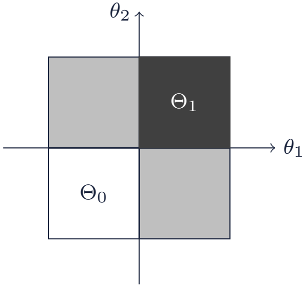
  $\implies$ have to choose a suitable notion of power
- **Proposition**: Consider subjective priors $\pi \in \Pi$ with support on $\Theta_1 = [0,1]^J$. Then

$$r^* \in \arg\max_{r \in \mathcal{R}} \left\{ \min_{\theta \in \Theta} v_r(\theta) + \lambda \int_{\Theta_1} e_r^*(\theta)\pi(\theta)d\theta \right\},$$

for all $\lambda \geq 0$ and $\pi \in \Pi$ if and only if

(i) r^* is maximin optimal ($\Rightarrow$ size control/non-negative welfare)

(ii) $\beta_{r^*}(\theta) \geq 0$ for all $\theta \in \Theta_1$ ($\Rightarrow$ unbiased test/power)



Globally most powerful maximin protocols

- Existence of such a rule?

Globally most powerful maximin protocols

- Existence of such a rule? **Proposition:** if $X \sim \mathcal{N}(\theta, \Sigma)$, with $\Sigma_{j,j} = \sigma^2$

$$r_j^*(X) = 1\left\{\frac{X_j}{\sigma} > t_j^*\right\}, \quad \text{where} \quad t_j^* \geq \Phi^{-1}\left(1 - C(J)/J\right), \quad \text{for} \quad j = 1, \dots, J,$$

is maximin optimal. It is maximin and unbiased if and only if $t_j^* = \Phi^{-1}\left(1 - C(J)/J\right)$.

Globally most powerful maximin protocols

- Existence of such a rule? **Proposition:** if $X \sim \mathcal{N}(\theta, \Sigma)$, with $\Sigma_{j,j} = \sigma^2$

$$r_j^*(X) = 1\left\{\frac{X_j}{\sigma} > t_j^*\right\}, \quad \text{where} \quad t_j^* \geq \Phi^{-1}\left(1 - C(J)/J\right), \quad \text{for} \quad j = 1, \dots, J,$$

is maximin optimal. It is maximin and unbiased if and only if $t_j^* = \Phi^{-1}\left(1 - C(J)/J\right)$.

- Comparison to condition for maximin optimality:

$$\underbrace{\sum_{j=1}^{J} P(r_j^*(X) = 1|\theta) \leq C(J) \quad \text{for all } \theta \in \Theta_0}_{\text{maximin (condition (a))}} \quad \text{vs.} \quad \underbrace{\sum_{j=1}^{J} P(r_j^*(X) = 1|\theta) = C(J) \quad \text{for } \theta = 0}_{\text{unbiased and maximin test}}$$

Globally most powerful maximin protocols

- Existence of such a rule? **Proposition:** if $X \sim \mathcal{N}(\theta, \Sigma)$, with $\Sigma_{j,j} = \sigma^2$

$$r_j^*(X) = 1\left\{\frac{X_j}{\sigma} > t_j^*\right\}, \quad \text{where} \quad t_j^* \geq \Phi^{-1}\left(1 - C(J)/J\right), \quad \text{for} \quad j = 1, \ldots, J,$$

  is maximin optimal. It is maximin and unbiased if and only if $t_j^* = \Phi^{-1}\left(1 - C(J)/J\right)$.

- Comparison to condition for maximin optimality:

$$\underbrace{\sum_{j=1}^{J} P(r_j^*(X) = 1|\theta) \leq C(J) \ \text{ for all } \theta \in \Theta_0}_{\text{maximin (condition (a))}} \quad \text{vs.} \quad \underbrace{\sum_{j=1}^{J} P(r_j^*(X) = 1|\theta) = C(J) \ \text{ for } \theta = 0}_{\text{unbiased and maximin test}}$$

$\Rightarrow$ Challenge here to show maximin optimality in mixed orthants

Simple examples of research costs

- Decompose the costs into fixed costs and variable costs: $C(J) = c_f + c_v(J)$
- Optimal level for separate t-tests: $\alpha(J) = (c_f + c_v(J))/J$
- **Examples**

	Cost function	Level	Intuition
Bonferroni	$c_f = \alpha, c_v(J) = 0$	α/J	Adjustment for increased benefits due to false positives
No adjustment	$c_f = 0, c_v(J) = \alpha J$	α	MHT adjustments are "built into" the cost structure

Simple examples of research costs

- Decompose the costs into fixed costs and variable costs: $C(J) = c_f + c_v(J)$
- Optimal level for separate t-tests: $\alpha(J) = (c_f + c_v(J))/J$
- **Examples**

	Cost function	Level	Intuition
Bonferroni	$c_f = \alpha, c_v(J) = 0$	α/J	Adjustment for increased benefits due to false positives
No adjustment	$c_f = 0, c_v(J) = \alpha J$	α	MHT adjustments are "built into" the cost structure

- General MHT adjustment based on relative costs:

$$\alpha(J) = \underbrace{\frac{C(J)/J}{C(1)}}_{\text{adjustment factor}} \times \alpha(1)$$

Additional formal results in the paper

- **Robustness guarantees to misspecified** $\beta_r(\theta)$
  - $\Rightarrow$ maximin optimality using worst-case upper bounds
  - $\Rightarrow$ For example, only need to know $C'(J) \geq C(J)$ for maximin optimality
  - $\Rightarrow$ Optimality for any λ, holds for a restricted class of priors Π

Additional formal results in the paper

- **Robustness guarantees to misspecified $\beta_r(\theta)$**
  - $\Rightarrow$ maximin optimality using worst-case upper bounds
  - $\Rightarrow$ For example, only need to know $C'(J) \geq C(J)$ for maximin optimality
  - $\Rightarrow$ Optimality for any λ, holds for a restricted class of priors Π
- **Baseline model with interactions in the researcher's benefit ($\beta_r(\theta) = \gamma \int 1\left\{\sum_{j=1}^{J} r_j(x) \geq \kappa\right\} dF_\theta(x) - C(J)$)**
  - $\Rightarrow$ separate t-tests with level depending on J are optimal

Additional formal results in the paper

- **Robustness guarantees to misspecified $\beta_r(\theta)$**
  - $\Rightarrow$ maximin optimality using worst-case upper bounds
  - $\Rightarrow$ For example, only need to know $C'(J) \geq C(J)$ for maximin optimality
  - $\Rightarrow$ Optimality for any λ, holds for a restricted class of priors Π
- **Baseline model with interactions in the researcher's benefit ($\beta_r(\theta) = \gamma \int 1\left\{\sum_{j=1}^{J} r_j(x) \geq \kappa\right\} dF_\theta(x) - C(J)$)**
  - $\Rightarrow$ separate t-tests with level depending on J are optimal

  $+$ **interactions in the welfare effects**
  - $\Rightarrow$ rationalizes weak FWER control between groups of treatments sufficient for approval

Additional formal results in the paper

- **Robustness guarantees to misspecified $\beta_r(\theta)$**
  - $\Rightarrow$ maximin optimality using worst-case upper bounds
  - $\Rightarrow$ For example, only need to know $C'(J) \geq C(J)$ for maximin optimality
  - $\Rightarrow$ Optimality for any λ, holds for a restricted class of priors Π
- **Baseline model with interactions in the researcher's benefit**
  ($\beta_r(\theta) = \gamma \int 1 \left\{ \sum_{j=1}^{J} r_j(x) \geq \kappa \right\} dF_\theta(x) - C(J)$)
  - $\Rightarrow$ separate t-tests with level depending on J are optimal

  $+$ **interactions in the welfare effects**
  - $\Rightarrow$ rationalizes weak FWER control between groups of treatments sufficient for approval
- **Other notions of power**
  - Worst-case power: study a ϵ deviations from the positive orthant
  - Weighted Average Power: no rule most powerful for any choice of the weights

Bringing the model to the data

- Clinical trials
  - use calibration to historical data on costs for subgroup analysis (Sertkaya et al., 2016)
  - Tabulate values with $\alpha(1) = 5\%$, $\alpha(2) = 3.2\%, \alpha(3) = 2.6\%, \alpha(9) = 1.8\%$, ...
  - Find adjustments less stringent than FWER adjustments

Bringing the model to the data

- Clinical trials
  - use calibration to historical data on costs for subgroup analysis (Sertkaya et al., 2016)
  - Tabulate values with $\alpha(1) = 5\%$, $\alpha(2) = 3.2\%, \alpha(3) = 2.6\%, \alpha(9) = 1.8\%$, ...
  - Find adjustments less stringent than FWER adjustments
- Economic studies: unique application using financial data from JPAL
  - Find that financial costs are not invariant to scale, but also do not scale linearly

Bringing the model to the data

- Clinical trials
  - use calibration to historical data on costs for subgroup analysis (Sertkaya et al., 2016)
  - Tabulate values with $\alpha(1) = 5\%$, $\alpha(2) = 3.2\%, \alpha(3) = 2.6\%, \alpha(9) = 1.8\%$, ...
  - Find adjustments less stringent than FWER adjustments
- Economic studies: unique application using financial data from JPAL
  - Find that financial costs are not invariant to scale, but also do not scale linearly

	Sample size ($\alpha = 0.05$)			$\boldsymbol{\alpha}^{\mathbf{0.05}}_{\text{Šidák}}$
$\lvert J\rvert$	100%	50%	200%	0.05
1	0.050	0.047	0.052	0.0500
2	0.027	0.026	0.028	0.0253
3	0.019	0.018	0.020	0.0170
4	0.015	0.014	0.015	0.0127
5	0.012	0.011	0.013	0.0102
6	0.010	0.010	0.011	0.0085
7	0.009	0.008	0.009	0.0073

Extensions: multiple outcomes (one treatment) and additional extensions

Multiple policymakers

- Outcomes $Y = (Y_1, \ldots, Y_G)$ associated with $X = (X_1, \ldots, X_G)$ and $(\theta_1, \cdots, \theta_G)$
- **Example:** for $g = 1, \ldots, G$, the researcher estimates the effect of treatment D on outcome Y_g using the regression model $Y_g = \mu + \theta_g D + \varepsilon_g \Rightarrow X = (\hat{\theta}_1, \ldots, \hat{\theta}_G)^\top$
- G policymakers and uncertainty wrt which policymaker will implement the policy

Multiple policymakers

- Outcomes $Y = (Y_1, \ldots, Y_G)$ associated with $X = (X_1, \ldots, X_G)$ and $(\theta_1, \cdots, \theta_G)$
- **Example:** for $g = 1, \ldots, G$, the researcher estimates the effect of treatment D on outcome Y_g using the regression model $Y_g = \mu + \theta_g D + \varepsilon_g \Rightarrow X = (\hat{\theta}_1, \ldots, \hat{\theta}_G)^\top$
- G policymakers and uncertainty wrt which policymaker will implement the policy
- Welfare for policymaker j is θ_j (each policy-maker cares about one outcome)

Multiple policymakers

- Outcomes $Y = (Y_1, \ldots, Y_G)$ associated with $X = (X_1, \ldots, X_G)$ and $(\theta_1, \cdots, \theta_G)$
- **Example:** for $g = 1, \ldots, G$, the researcher estimates the effect of treatment D on outcome Y_g using the regression model $Y_g = \mu + \theta_g D + \varepsilon_g \Rightarrow X = (\hat{\theta}_1, \ldots, \hat{\theta}_G)^\top$
- G policymakers and uncertainty wrt which policymaker will implement the policy
- Welfare for policymaker j is θ_j (each policy-maker cares about one outcome)
- Researcher reports G tests, one for each policymaker, $r(X) = (r_1(X), \ldots, r_G(X))$

Multiple policymakers

- Outcomes $Y = (Y_1, \ldots, Y_G)$ associated with $X = (X_1, \ldots, X_G)$ and $(\theta_1, \cdots, \theta_G)$
- **Example:** for $g = 1, \ldots, G$, the researcher estimates the effect of treatment D on outcome Y_g using the regression model $Y_g = \mu + \theta_g D + \varepsilon_g \Rightarrow X = (\hat{\theta}_1, \ldots, \hat{\theta}_G)^\top$
- G policymakers and uncertainty wrt which policymaker will implement the policy
- Welfare for policymaker j is θ_j (each policy-maker cares about one outcome)
- Researcher reports G tests, one for each policymaker, $r(X) = (r_1(X), \ldots, r_G(X))$
- Isomorphic to model with multiple treatments, and optimal t-tests

$$r_j^*(X) = 1\left\{\frac{X_j}{\sqrt{\Sigma_{j,j}}} \geq \Phi^{-1}\left(1 - \frac{C(G)}{G}\right)\right\}, \quad \forall j.$$

Single policymaker

- Here $r(X) \in \{0, 1\}$, $\beta_r(\theta) = \int r(x) dF_\theta(x) - C(G)$, and utility $u(\theta) = \theta^\top w^*$

Single policymaker

- Here $r(X) \in \{0, 1\}$, $\beta_r(\theta) = \int r(x) dF_\theta(x) - C(G)$, and utility $u(\theta) = \theta^\top w^*$

$\Rightarrow$ Each X_g measures the impact on an economically distinct concept.

Single policymaker

- Here $r(X) \in \{0, 1\}$, $\beta_r(\theta) = \int r(x) dF_\theta(x) - C(G)$, and utility $u(\theta) = \theta^\top w^*$

$\Rightarrow$ Each X_g measures the impact on an economically distinct concept. Then optimal r^*

$$r^*(X) = 1\left\{\frac{X^\top w^*}{\sqrt{w^{*\top} \Sigma w^*}} > \Phi^{-1}(1 - C(G))\right\},$$

implies economic aggregation (e.g. Bhatt et al. (2024))

Single policymaker

- Here $r(X) \in \{0,1\}$, $\beta_r(\theta) = \int r(x) dF_\theta(x) - C(G)$, and utility $u(\theta) = \theta^\top w^*$

$\Rightarrow$ Each X_g measures the impact on an economically distinct concept. Then optimal r^*

$$r^*(X) = 1\left\{\frac{X^\top w^*}{\sqrt{w^{*\top}\Sigma w^*}} > \Phi^{-1}(1 - C(G))\right\},$$

implies economic aggregation (e.g. Bhatt et al. (2024))

$\Rightarrow$ Each X_g measures the same underlying effect: $\theta_1 = \cdots = \theta_G$.

Single policymaker

- Here $r(X) \in \{0,1\}$, $\beta_r(\theta) = \int r(x) dF_\theta(x) - C(G)$, and utility $u(\theta) = \theta^\top w^*$

$\Rightarrow$ Each X_g measures the impact on an economically distinct concept. Then optimal r^*

$$r^*(X) = 1\left\{\frac{X^\top w^*}{\sqrt{w^{*\top}\Sigma w^*}} > \Phi^{-1}(1 - C(G))\right\},$$

implies economic aggregation (e.g. Bhatt et al. (2024))

$\Rightarrow$ Each X_g measures the same underlying effect: $\theta_1 = \cdots = \theta_G$. Then

$$r^*(X) = 1\left\{\frac{X^\top w^{\min}}{\sqrt{w^{\min\top}\Sigma w^{\min}}} > \Phi^{-1}(1 - C(G))\right\},$$

where $w^{\min}$ minimizes $\sqrt{w^\top \Sigma w}$ st $\sum_g w_g = 1$ (Statistical aggregation)

Recap of our extensions

- Endogenous J (pre-specified by the researcher ex-ante)
- Unknown θ and researcher's prior on θ
- Some benevolent researcher
- Additional forms of interactions
- Alternative notions of power (WAP and local power)
- Variance that might depend on J and heterogeneous variance
- Weighted welfare function
- Two sided tests

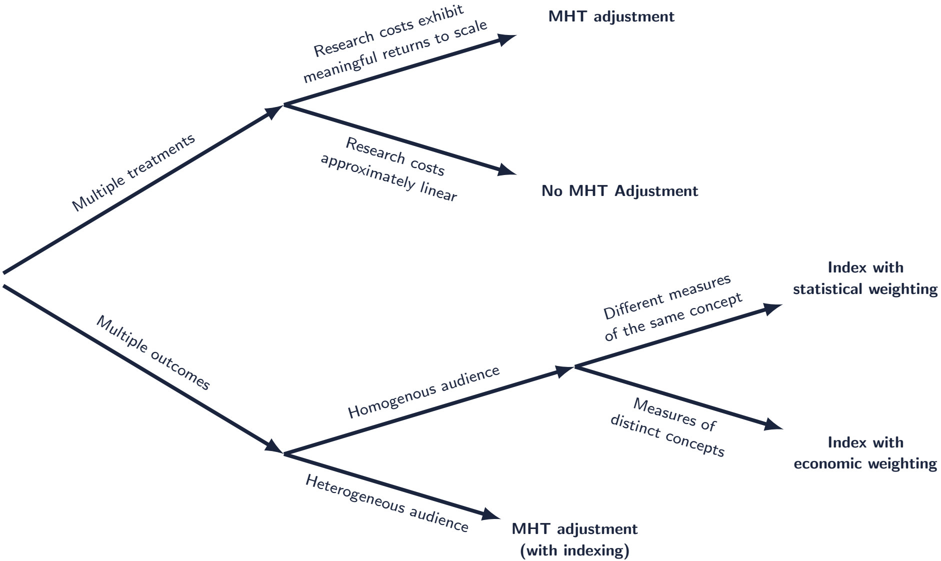
Conclusions

Something we have learned working on this project

- Scale economies should guide our choice for MHT adjustments:
  - With large returns to scale we need more stringent MHT adjustments (e.g., Bonferroni)
  - In the extreme case where research costs scale linearly no adjustment is required
  - We think practice is between these two cases:
    - data from clinical trial suggesting less stringent thresholds than those suggested by FDA
    - data from JPAL suggesting adjustments similar to Sidak corrections

$\Rightarrow$ Ask to report proxies for costs and be explicit about welfare function

Conclusions



Thank you!

Questions? Thoughts? Comments? Please reach out:

dviviano@fas.harvard.edu, kwuthrich@ucsd.edu,
pniehaus@ucsd.edu

References

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.

Andrews, I., Shapiro, J.M., 2021. A model of scientific communication. Econometrica 89, 2117–2142.

Banerjee, A., Chassang, S., Montero, S., Snowberg, E., 2017. A theory of experimenters. Technical Report. National Bureau of Economic Research.

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.

Chassang, S., Padro I Miquel, G., Snowberg, E., 2012. Selective trials: A principal-agent approach to randomized controlled experiments. American Economic Review 102, 1279–1309. URL: https://www.aeaweb.org/articles?id=10.1257/aer.102.4.1279, doi:10.1257/aer.102.4.1279.

Di Tillio, A., Ottaviani, M., Sørensen, P.N., 2017. Persuasion bias in science: Can economics help? The Economic Journal 127, F266–F304.

Efron, B., 2008. Simultaneous inference: when should hypothesis testing problems be combined? Annals of Applied Statistics 2, 197–223.

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

Henry, E., Ottaviani, M., 2019. Research and the approval process: the organization of persuasion. American Economic Review 109, 911–55.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Kasy, M., Spiess, J., 2023. Optimal pre-analysis plans: Statistical decisions subject to implementability .

Lehmann, E.L., Romano, J.P., 2005. Testing statistical hypotheses. Springer Science & Business Media.

Lewis, J.A., 1999. Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. Statistics in medicine 18, 1903–1942.

Robbins, H., 1951. Asymptotically subminimax solutions of compound statistical decision problems, in: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, The Regents of the University of California.

Romano, J.P., Shaikh, A.M., Wolf, M., 2010. Hypothesis testing in econometrics. Annual Review of Economics 2, 75–104.

Sertkaya, A., Wong, H.H., Jessup, A., Beleche, T., 2016. Key cost drivers of pharmaceutical clinical trials in the United States. Clinical Trials 13(2), 117–126.

Spiess, J., 2018. Optimal estimation when researcher and social preferences are misaligned. Working Paper.

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

Tetenov, A., 2016. An economic theory of statistical testing. Working Paper.

Wald, A., 1950. Statistical decision functions. Wiley.

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

Wong, H.H., Jessup, A., Sertkaya, A., Birkenbach, A., Berlind, A., Eyraud, J., 2014. Examination of clinical trial costs and barriers for drug development final. Office of the Assistant Secretary for Planning and Evaluation, US Department of Health & Human Services , 1–92.

Digression: Can we justify the FDR?

- Suppose that $u_j(\theta) = \theta_j$. FDR is optimal if

$$\beta_r(\theta) = \int \left[\sum_{j=1}^{J} \frac{1\{\theta_j < 0\} r_j(x)}{\sum_{j=1}^{J} r_j(x)} \cdot 1 \left\{ \sum_{j=1}^{J} r_j(x) > 0 \right\} \right] dF_\theta(x) - C(J)$$

- Researcher is malevolent: her utility is increasing in the number false discoveries

Digression: Can we justify the FDR?

- Suppose that $u_j(\theta) = \theta_j$. FDR is optimal if

$$\beta_r(\theta) = \int \left[\sum_{j=1}^{J} \frac{1\left\{\theta_j < 0\right\} r_j(x)}{\sum_{j=1}^{J} r_j(x)} \cdot 1\left\{ \sum_{j=1}^{J} r_j(x) > 0 \right\} \right] dF_\theta(x) - C(J)$$

- Researcher is malevolent: her utility is increasing in the number false discoveries

- FDR does not arise as a natural solution in our frequentist maximin framework
- Complementarities betw/ Bayesian [Storey (2003)] and frequentist