Short Course on Program Evaluation
Multiple Hypothesis Testing

Matias D. Cattaneo

University of Michigan

May 31, 2017

Outline

Gerber and Hopkins (AJPS, 2011)

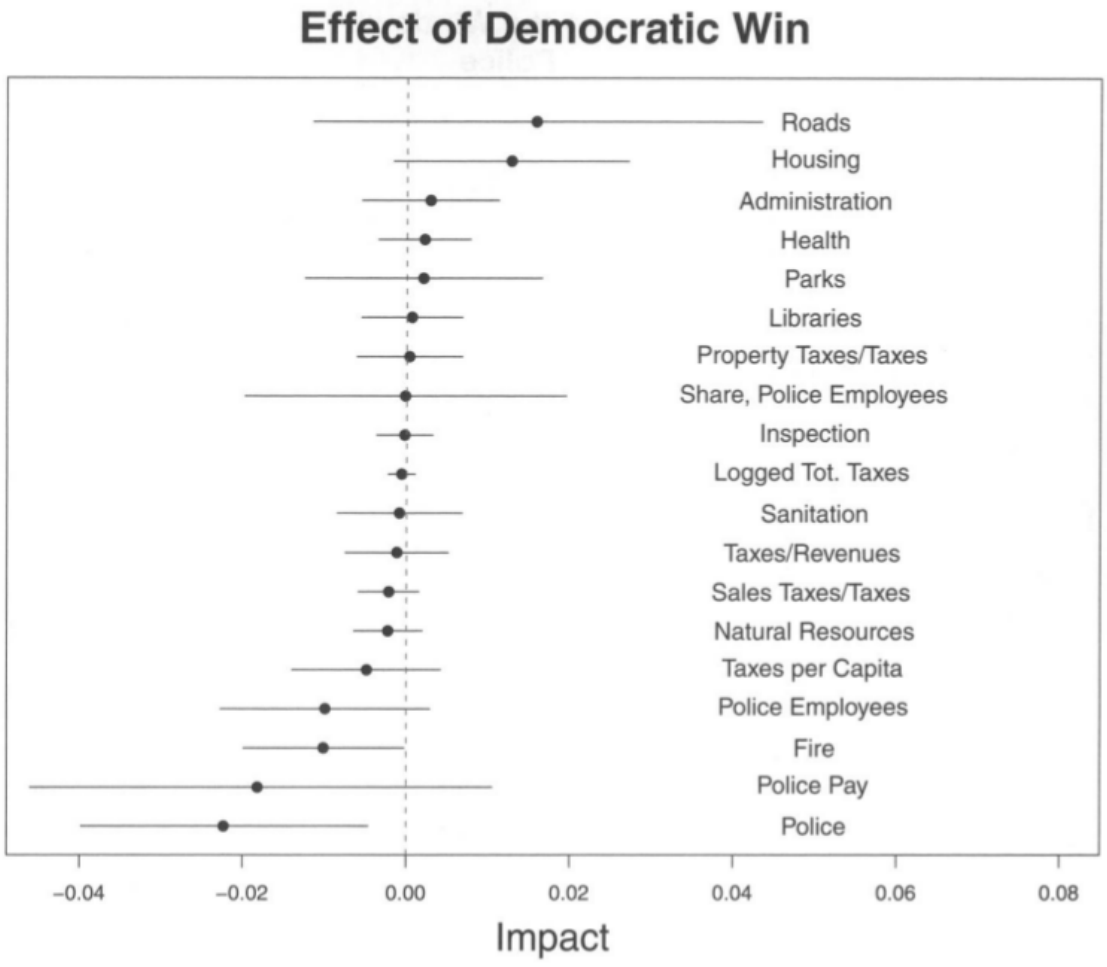
When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy

Elisabeth R. Gerber University of Michigan
Daniel J. Hopkins Georgetown University

U.S. cities are limited in their ability to set policy. Can these constraints mute the impact of mayors' partisanship on policy outcomes? We hypothesize that mayoral partisanship will more strongly affect outcomes in policy areas where there is less shared authority between local, state, and federal governments. To test this hypothesis, we create a novel dataset combining U.S. mayoral election returns from 1990 to 2006 with city fiscal data. Using regression discontinuity design, we find that cities that elect a Democratic mayor spend a smaller share of their budget on public safety, a policy area where local discretion is high, than otherwise similar cities that elect a Republican or an Independent. We find no differences on tax policy, social policy, and other areas that are characterized by significant overlapping authority. These results suggest that models of national policymaking are only partially applicable to U.S. cities. They also have implications for political accountability: mayors may not be able to influence the full range of policies that are nominally local responsibilities.

Gerber and Hopkins (AJPS, 2011)

FIGURE 2 OLS Estimates, Effect of a Democratic Victory on Outcomes



Note: Conditional on covariates in Table 1, estimated on multiply imputed data sets with standard errors clustered by city.

Multiple Comparisons

- The problem of *multiple testing* arises when we test many hypotheses
- When we test n hypotheses, the chance of mistakenly rejecting at least one true null hypothesis is high due to chance
- False positive: rejecting the null hypothesis when it is true
- What's the probability of at least one false positive?
- Test m independent hypotheses:

$$\begin{aligned}
\mathbb{P}(\text{Making an error}) &= \alpha \\
\mathbb{P}(\text{Not making an error}) &= 1 - \alpha \\
\mathbb{P}(\text{Not making an error in } m \text{ tests}) &= (1 - \alpha)^m \\
\mathbb{P}(\text{Making at least one error in } m \text{ tests}) &= 1 - (1 - \alpha)^m
\end{aligned}$$

Multiple Comparisons

- Test 20 hypothesis at level $\alpha = 0.05$

$$\begin{aligned}\mathbb{P}(\text{reject at least one hypothesis}) &= 1 - \mathbb{P}(\text{reject no hypotheses}) \\ &= 1 - (1 - 0.05)^{20} \\ &\approx 0.642\end{aligned}$$

- If we perform 20 tests and the null hypothesis is true in all cases, we have a 64% chance of rejecting at least once

- If we perform enough tests, we will reject the null hypothesis at least once !

  - Key problem: False discoveries

Outline

Family-Wise Error Rate

- The Family-wise Error Rate (FWER) is the probability of making at least one mistake (i.e., of rejecting at least one true null hypothesis)
- Two main approaches to control the FWER
  - Single step: all p-values are adjusted equally
  - Sequential: the adjustment to each p-value is adaptive
- Alternatively, consider other notions of multiple-hypothesis testing

Single Step FWER Correction: Bonferroni

- Perform m hypothesis tests: $H_1, H_2, \ldots, H_m$ based on p-values $P_1, P_2, \ldots, P_m$
- If we fix type I error at α for each H_i

$$\begin{aligned}\mathsf{FWER} &= \mathbb{P}\Big(\text{reject } H_1 \text{ or reject } H_2 \text{ or } \ldots \text{ reject } H_m \,\Big|\, H_1, H_2, \ldots, H_m\Big) \\ &= \mathbb{P}\Big(\big\{\text{reject} H_1 | H_1)\big\} \cup \big\{\text{reject} H_2 | H_2)\big\} \cup \ldots \big\{\text{reject} H_m | H_m)\big\}\Big) \\ &\leq \sum_{i=1}^{m} \mathbb{P}(\{\text{reject } H_i | H_i\}) \leq m\alpha\end{aligned}$$

We used: recall that $\mathbb{P}(\cup_i E_i) \leq \Sigma_i \mathbb{P}(E_i)$

- **Bonferroni Correction**: ensures $\mathsf{FWER} \leq \alpha$ by rejecting H_i if $P_i \leq \alpha/m$

$$\mathsf{FWER} \leq \sum_{i=1}^{m} \mathbb{P}\Big(\big\{\text{reject } H_i | H_i\big\}\Big) \leq m\frac{\alpha}{m} = \alpha$$

- Adjusted p-values are $\tilde{P}_i = \min\{m \cdot P_i, 1\}$
- Example: 100 tests, want to control FWER at $\alpha = 0.05$, we reject each H_i if its p-value is less than $0.05/100 = 0.0005$

Single Step FWER Correction: Bonferroni Approach

- Controlling FWER controls the *experiment-wide* Type I error rate
- It is a very conservative procedure: leads to a high probability of Type II errors—i.e., of failing to reject the general null hypothesis when effects do exist
- The general null hypothesis is that all the null hypotheses $H_1, H_2, \ldots, H_m$ are true
- This is appropriate when we want to protect against *any* false positives
- In some cases, this may not be of interest
- For example, in genomics, scientists may want to allow for a certain number of false positives

Refined Bonferroni-Type Corrections

- Perform m hypothesis tests: $H_1, H_2, \ldots, H_m$ based on p-values $P_1, P_2, \ldots, P_m$
- **Holm-Bonferroni Approach**: letting $P_{(j)}$ denote the j-th order statistics (i.e., $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(J)}$):

$$\text{Reject } H_j \qquad \text{iff} \qquad P_{(j)} \leq \frac{\alpha}{m+1-j}, \qquad \text{for all } j = 1, 2, \cdots, m.$$

- **Šidák Approach**: assuming $P_1, P_2, \cdots, P_J$ are independent:

$$\text{Reject } H_j \qquad \text{iff} \qquad P_j \leq 1 - (1-\alpha)^{1/J}, \qquad \text{for all } j.$$

Outline

Multiple Testing More Generally

Table: Number of Errors Committed When Testing m null hypotheses

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

- The number of hypotheses tested m is known
- R: number of rejected hypothesis (observed random variable)
- U, V, S and T are unobservable random variables
- R: total number of rejected hypothesis
- V: number of hypotheses incorrectly rejected

Beyond Controlling the FWER

- Recall: FWER is the probability of making at least one mistake (i.e., of rejecting at least one true null hypothesis)
- Using this notation, we can define the FWER as

$$\mathbb{P}(V \geq 1)$$

- Controlling FWER controls *experiment-wide* Type I error rate
- That is, FWER controls probability of rejecting at least one true hypothesis
- We might be interested in controlling other types of mistakes

False Discovery Rate

- The FDR is designed to control the *proportion of false positives among the set of all rejected hypothesis*
- Proportion of the rejected null hypotheses that are erroneously rejected:

$$Q = \frac{V}{V+S} = \frac{V}{R}$$

- Q captures the proportion of errors committed by falsely rejecting null hypotheses
- Q is an unobserved random variable
- The False Discovery Rate (FDR) is defined as the expectation of Q:

$$\mathsf{FDR} = \mathbb{E}\left[\frac{V}{V+S}\right] = \mathbb{E}\left[\frac{V}{R}\right]$$

False Discovery Rate

$$\mathsf{FDR} = \mathbb{E}\left[\frac{V}{V+S}\right] = \mathbb{E}\left[\frac{V}{R}\right]$$

- If all null hypotheses are true, the FDR is equivalent to the FWER:
  - When all null hypotheses are true, $S = 0$ and $V = R$
  - If $V = 0$, $Q = 0$
  - If $V > 0$, $Q = 1$ so $\mathbb{P}(V \geq 1) = \mathbb{E}[Q] = \mathsf{FDR}$
  - Therefore, control of the FDR implies control of the FWER in the weak sense

False Discovery Rate

$$\mathsf{FDR} = \mathbb{E}\left[\frac{V}{V+S}\right] = \mathbb{E}\left[\frac{V}{R}\right]$$

- When $m_0 < m$, the FDR is smaller than or equal to the FWER
  - If $V > 0$ then $V/R \leq 1$, leading to $\mathbf{1}(V \geq 1) \geq Q$
  - Taking expectations on both sides, $\mathbb{P}(V \geq 1) \geq \mathsf{FDR}$
  - If $v > 0$, $Q = 1$ so $\mathbb{P}(V \geq 1) = \mathbb{E}(Q) = \mathsf{FDR}$
  - Therefore, any procedure that controls the FWER controls the FDR
  - But if a procedure controls the FDR only, it can be less stringent and lead to power gains
  - The larger the number of null hypotheses, the larger S tends to be, and the larger the difference between the FDR and the FWER
  - Thus, the potential for power gains is larger when more hypotheses are non-true

Example 1

- Imagine a case when multiple tests are done and an overall decision (i.e., whether to recommend a policy) is made at the end
- Overall decision: whether to recommend a new treatment over a status-quo treatment
- Discoveries $\Rightarrow$ rejections of null hypotheses that claim that the status quo is no better than the new treatment
- We wish to make as many discoveries (which will increase the chances of recommending the new treatment), subject to control of the FDR
- Controlling the probability of *any* error is unnecessarily strict, as a small proportion of errors will not change the validity of the overall conclusion

Example 2

- Imagine a screening problem where multiple potential effects are screened to eliminate the null effects
  - Screen several chemicals for potential drug development
  - Test multiple factors in an experimental design
- We wish to make as many discoveries as possible
- But we want to control the FDR, because too large a fraction of false leads would burden the second phase of the confirmatory analysis

Take-aways: FDR versus FWER

- FWER is the probability of (mistakenly) rejecting at least one true null hypothesis
- FDR is the *proportion* of true null hypothesis mistakenly rejected or the *proportion of false discoveries*

FDR Controlling Procedure

- We test $H_1, H_2, \ldots, H_m$ based on p-values $P_1, P_2, \ldots, P_m$
- Let $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$ be the ordered p-values
- Let $H_{(i)}$ the hypothesis corresponding to $P_{(i)}$
- Define the following procedure
  - Let k be the largest i for which $P_{(i)} \leq \frac{i}{m} q^\star$
  - Reject all $H_{(i)} = 1, 2, \ldots, k$
- Result: For independent test statistics and for any configuration of false null hypotheses, this procedure controls the FDR at $q^\star$
- The adjusted p-values—also known as q-values—are defined as

$$q_{(i)} = \tilde{P}_{(i)} = \frac{P_{(i)} \cdot m}{i}$$

$q_{(i)} \Rightarrow$ minimum expected proportion of false positives that can be attained when calling $H_{(i)}$ significant

FDR Controlling Procedure: Example

- We have 14 tests with the following 14 ordered p-values

$$0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344,$$
$$0.0459, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7598, 1.0000$$

- Controlling FWER at $\alpha = 0.05$, we use $0.05/15 = 0.0033$ and reject the three hypotheses corresponding to the three smallest p-values, i.e. $H_{(1)}, H_{(2)}, H_{(3)}$

- To control FWER at $q^\star = 0.05$, compare sequentially each $p_{(i)}$ with $0.05\frac{i}{15}$

- The first p-vaue to satisfy the constraint is $p_{(4)}$:

$$p_{(4)} = 0.0095 \leq 0.05\frac{4}{15} = 0.013$$

- Therefore, we reject the four hypotheses corresponding to the four smallest p-values—$H_{(1)}, H_{(2)}, H_{(3)}, H_{(4)}$—which have p-values ≤ 0.013

- Controlling FDR we reject one more hypothesis than controlling FWER

Multiple Hypothesis Testing: FDR Approach

- Method is based on the idea of combining results of many tests.

- The procedure is as follows:
  1. Set the level $\alpha \in (0, 1)$
  2. Order the p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(J)}$
  3. Plot versus rank (j) and add line $p_{(j)} = \frac{j\alpha}{J}$.
     (Recall that Bonferroni's used $\frac{\alpha}{J}$, so this method is slightly different!)
  4. Let $\ell^* = \max\left\{j : p_{(j)} \leq \frac{j\alpha}{J}\right\}$
  5. Reject hypotheses with $p_{(j)} \leq p_{(\ell^*)}$
  6. In can be shown that if we define

$$FDR = \frac{\#\left\{\text{Hypothesis falsely declared significant}\right\}}{\#\left\{\text{Hypothesis declared significant}\right\}}$$

     then

$$\mathbb{E}_{\mathsf{H}_0}\left[FDR\right] \leq \alpha.$$

Outline

Multiple Hypothesis Testing: Scheffé's Approach (Projection Method)

- Method useful to construct confidence intervals for subsets of parameter $\theta \in \mathbb{R}^d$.
- It is in general (very) conservative, but it is simple and easy to implement.
- *Idea*: consider the case of linear combinations $\mathbf{a}'\theta$, for some constant vector $\mathbf{a} \in \mathbb{R}^d$.
  1. Assume, under the null hypothesis,
  $$(\mathbf{a}'\hat{\theta}_n - \mathbf{a}'\theta)'(\mathbf{a}'\hat{\mathbf{\Sigma}}_n\mathbf{a})^{-1}(\mathbf{a}'\hat{\theta}_n - \mathbf{a}'\theta) \sim F, \qquad F \text{ pivotal distribution}$$
  2. First note that the correct size $1-\alpha$ confidence interval is:
  $$\mathbb{P}_{\mathsf{H}_0}\left[\mathbf{a}'\theta \in \left[\mathbf{a}'\hat{\theta}_n - F_{\alpha/2}\sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_n\mathbf{a}}\ ,\ \mathbf{a}'\hat{\theta}_n - F_{1-\alpha/2}\sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_n\mathbf{a}}\right]\right] = 1-\alpha$$
  3. Consequently, a conservative CI is:
  $$\left[\mathbf{a}'\theta \in \left[\mathbf{a}'\hat{\theta}_n - F_{\alpha/2}\sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_n\mathbf{a}}\ ,\ \mathbf{a}'\hat{\theta}_n - F_{1-\alpha/2}\sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_n\mathbf{a}}\right], \forall \mathbf{a} \in \mathbb{R}^d\right]$$
- This is related to the so-called Projection Method!

Multiple Hypothesis Testing: Scheffé's Approach (Projection Method)