ECON 626: Applied Microeconomics

Lecture 9:

Multiple Test Corrections

Professors: Pamela Jakiela and Owen Ozier

Multiple Hypothesis Testing: The Problem

Consider testing 100 true null hypotheses — how many will rejected?

Multiple Hypothesis Testing: The Problem

Consider testing 100 true null hypotheses — how many will rejected?

	Number of Tests
	1
Test size	0.05
No rejections	0.95
Any rejections	0.05

Multiple Hypothesis Testing: The Problem

Consider testing 100 true null hypotheses — how many will rejected?

	Number of Tests	
	1	2
Test size	0.05	0.05
No rejections	0.95	0.95^2
Any rejections	0.05	$1 - 0.95^2$

Multiple Hypothesis Testing: The Problem

Consider testing 100 true null hypotheses — how many will rejected?

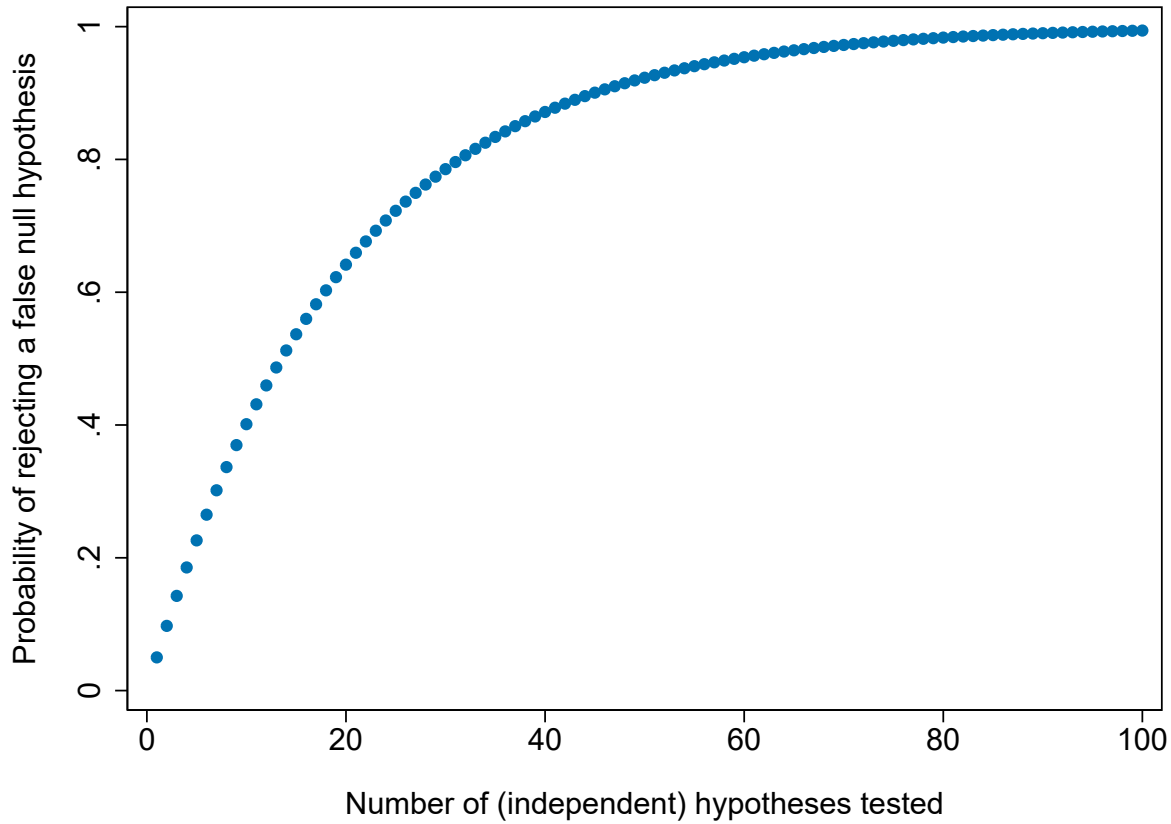
	Number of Tests	
	1	2
Test size	0.05	0.05
No rejections	0.95	0.9025
Any rejections	0.05	0.0975

Multiple Hypothesis Testing: The Problem

Consider testing 100 true null hypotheses — how many will rejected?

	Number of Tests		
	1	2	k
Test size	0.05	0.05	0.05
No rejections	0.95	0.9025	0.95^k
Any rejections	0.05	0.0975	$1 - 0.95^k$

Multiple Hypothesis Testing: The Problem



Under the null, probability of rejecting at least on hypothesis increases rapidly with number of independent hypothesis tests

Multiple Hypothesis Testing: The Problem

How can we (credibly) test multiple hypotheses?

Multiple Hypothesis Testing: The Problem

How can we (credibly) test multiple hypotheses?

- What sort of ninny would test 100 hypotheses?

Multiple Hypothesis Testing: The Problem

How can we (credibly) test multiple hypotheses?

- What sort of ninny would test 100 hypotheses?
- Valid reasons for testing many hypotheses:
  - Studies often have 2 or 3 treatment arms (and rightly so!)
  - Difficult to predict which outcomes will be affected
    - Particularly true for secondary hypotheses/treatment effects
  - Different measures of the same outcome often available
  - Heterogeneity in treatment effects (across sub-samples)

Multiple Hypothesis Testing: The Problem

Published empirical papers include a lot of hypothesis tests!

TABLE I
CHARACTERISTICS OF THE SAMPLE

53 papers					1780 regressions	
location	journal	tables	treatment coefficients reported	treatment coefficients unreported	method	covariance
39 field	27 AER	17 1-2	17 2-30	41 0	.67 ols	.25 default
14 lab	26 AEJ	17 3-4	18 32-80	7 1-48	.22 mle	.70 cl/robust
		19 5-8	18 90-260	5 76-744	.11 other	.04 bootstrap
						.02 other

Notes. For papers, numbers reported are number of papers by characteristic. For regressions, numbers reported are the average across papers of the share of regressions within each paper with the noted characteristic.

Source: Young (2019)

Bonferroni Corrections

Most conservative approach is the **Bonferroni method***

- Problem: you wish to test hypotheses $H_1, ...H_k$ using a test size of α
- Solution (of sorts): use a test size of α/k instead
  - **Family-wise error rate (FWER)**: probability of rejecting a true null
  - Bonferroni correction holds FWER below α
  - Bonferroni corrections are too conservative:
    - FWER ≈ 0.04877 when number of independent tests is large
    - Bonferroni corrections can be extremely conservative when tests are not independent (consider example of perfectly correlated tests)

Good news: if you are testing k hypotheses and a Bonferroni correction works (i.e. your results hold up), you don't need the rest of this lecture

*Purportedly developed by Olive Jean Dunn and not, ahem, Carlo Emilio Bonferroni

Bonferroni Corrections

	Number of Tests	
	1	k
Test size (per test)	0.05	α/k
1 - (single) test size	0.95	$1-\alpha/k$
No rejections	0.95	$(1-\alpha/k)^k$
Any rejections	0.05	$1-(1-\alpha/k)^k$

Bonferroni Corrections

	Number of Tests		
	1	2	10
Test size (per test)	0.05	0.025	0.005
1 - (single) test size	0.95	$1 - \alpha/k$	
No rejections	0.95	$(1 - \alpha/k)^k$	
Any rejections	0.05	$1 - (1 - \alpha/k)^k$	

Bonferroni Corrections

	Number of Tests		
	1	2	10
Test size (per test)	0.05	0.025	0.005
1 - (single) test size	0.95	0.975	0.995
No rejections	0.95	$(1-\alpha/k)^k$	
Any rejections	0.05	$1-(1-\alpha/k)^k$	

Bonferroni Corrections

	Number of Tests		
	1	2	10
Test size (per test)	0.05	0.025	0.005
1 - (single) test size	0.95	0.975	0.995
No rejections	0.95	0.950625	0.951110
Any rejections	0.05	$1-(1-\alpha/k)^k$	

Bonferroni Corrections

	Number of Tests		
	1	2	10
Test size (per test)	0.05	0.025	0.005
1 - (single) test size	0.95	0.975	0.995
No rejections	0.95	0.950625	0.951110
Any rejections	0.05	0.049375	0.048890

Bonferroni Corrections

Most conservative approach is the **Bonferroni method***

- Problem: you wish to test hypotheses $H_1, ...H_k$ using a test size of α
- Solution (of sorts): use a test size of α/k instead
  - **Family-wise error rate (FWER)**: probability of rejecting a false null
  - Bonferroni correction holds FWER below α
  - Bonferroni corrections are too conservative:
    - FWER ≈ 0.04877 when number of independent tests is large
    - Bonferroni corrections can be extremely conservative when tests are not independent (consider example of perfectly correlated tests)

Good news: if you are testing k hypotheses and a Bonferroni correction works (i.e. your results hold up), you don't need the rest of this lecture

*Purportedly developed by Olive Jean Dunn and not, ahem, Carlo Emilio Bonferroni

Stepdown Methods

Holm (1979) proposes a less conservative **stepdown method**:

0. Order k p-values from smallest to largest, $p_{(1)}, p_{(2)}, ..p_{(k)}$
1a. If $p_{(1)} > \alpha/k$, stop. Fail to reject all hypotheses
1b. Reject $H_{(1)}$ if $p_{(1)} < \alpha/k$. Proceed to Step 2.
2a. If $p_{(2)} > \alpha/(k-1)$, stop. Fail to reject all remaining hypotheses.
2b. Reject $H_{(2)}$ if $p_{(2)} < \alpha/(k-1)$. Proceed to Step 3.
...
j. Repeat as needed until you stop rejecting hypotheses because $p_{(j)} > \alpha/(k-(j-1))$ or all k hypotheses have been rejected

More good news: Romano & Wolf (JASA, 2005) state "This procedures holds under arbitrary dependence on the joint distribution of p-values."

Stepdown Methods: Holm vs. Bonferroni

p-value	Bonferroni	Holm
0.010	0.050	0.050
0.010	0.050	0.040
0.015	0.075	0.045
0.050	0.250	0.100
0.100	0.500	0.100

Blue indicates hypotheses that would not be rejected using a test size of $\alpha = 0.05$

Resampling-Based Stepdown Methods

More complicated/powerful bootstrap-based stepdown methods exist

- Examples: Westfall & Young (1993), Romano & Wolf (2005)
- These procedures exploit additional assumptions to increase power (so you don't need them if simpler methods "work" in your setting)
- They are also more computationally-intensive, often including phrases like "efficient computation" or "computationally feasible"
- Approaches use some form of stepdown structure
  - At each step, "accept"/reject decisions use empirical distribution of bootstrapped p-values associated with not-yet-rejected hypotheses
  - Can be modified to generate adjusted p-values

Example: Romano and Wolf (2005)

For each of k hypotheses, let $t_k^{*,m}$ be a resampling-based test statistic, defined for $m = 1, \ldots, M$ bootstrap replications, permutations, etc.

- Test statistics defined so that higher indicates greater significance
- Unadjusted p-value: $\hat{p}_k = \#\{t_k^{*,m} \geq t_k\}/M$

Example: Romano and Wolf (2005)

For each of k hypotheses, let $t_k^{*,m}$ be a resampling-based test statistic, defined for $m = 1, \ldots, M$ bootstrap replications, permutations, etc.

- Test statistics defined so that higher indicates greater significance
- Unadjusted p-value: $\hat{p}_k = \#\{t_k^{*,m} \geq t_k\}/M$

To simplify notation, assume hypotheses are ordered: $t_1 \geq t_2 > \ldots \geq t_k$

- For $j = 1, \ldots, k$ and $m = 1, \ldots, M$, define:

$$\max_j^{*,m} = \max\{t_j^{*,m}, t_{j+1}^{*,m}, \ldots, t_k^{*,m}\}$$

Example: Romano and Wolf (2005)

For each of k hypotheses, let $t_k^{*,m}$ be a resampling-based test statistic, defined for $m = 1, \ldots, M$ bootstrap replications, permutations, etc.

- Test statistics defined so that higher indicates greater significance
- Unadjusted p-value: $\hat{p}_k = \#\{t_k^{*,m} \geq t_k\}/M$

To simplify notation, assume hypotheses are ordered: $t_1 \geq t_2 > \ldots \geq t_k$

- For $j = 1, \ldots, k$ and $m = 1, \ldots, M$, define:

$$\max_j^{*,m} = \max\{t_j^{*,m}, t_{j+1}^{*,m}, \ldots, t_k^{*,m}\}$$

Let $\hat{c}(1-\alpha, j)$ denote **empirical quantile** of $\max_j^{*,m}$

- For $\alpha = 0.05$, $j = 2$, $\hat{c}(1-\alpha, 2)$ is value of $\max_2^{*,m}$ at 95$^{\text{th}}$ percentile

Romano-Wolf Algorithm for testing at size α

1. Step 1.

   1.1 Reject all hypotheses with $t_k > \hat{c}(1-\alpha, 1)$

   $\Rightarrow$ Reject H_k if t_k is larger than 95 percent of values of $\max_1^{*,m}$

Romano-Wolf Algorithm for testing at size α

1. Step 1.

   1.1 Reject all hypotheses with $t_k > \hat{c}(1-\alpha, 1)$

   $\Rightarrow$ Reject H_k if t_k is larger than 95 percent of values of $\max_1^{*,m}$

   1.2 Let R_1 denote number of rejected hypotheses

   1.2.1 If $R_1 = 0$, stop — fail to reject all hypotheses

   1.2.2 If $R_1 > 0$, proceed to Step 2

Romano-Wolf Algorithm for testing at size α

1. Step 1.
   - 1.1 Reject all hypotheses with $t_k > \hat{c}(1-\alpha, 1)$
     - $\Rightarrow$ Reject H_k if t_k is larger than 95 percent of values of $\max_1^{*,m}$
   - 1.2 Let R_1 denote number of rejected hypotheses
     - 1.2.1 If $R_1 = 0$, stop — fail to reject all hypotheses
     - 1.2.2 If $R_1 > 0$, proceed to Step 2
2. Steps 2, 3, etc.
   - 2.1 Reject H_k if $t_k > \hat{c}(1-\alpha, R_1 + 1)$
   - 2.2 Define R_2 as total number rejected hypotheses
     - 2.2.1 If $R_2 = R_1$, stop
     - 2.2.2 If $R_2 > R1$, proceed to Step 3, repeating until $R_{j+1} = R_j$

Calculating Romano-Wolf Adjusted p-values

Consider k hypotheses ordered such that $t_1 \geq t_2 > \ldots \geq t_k$

1. Step 1. Calculate initial set of adjusted p-values

$$\hat{p}_k^0 = \#\{\max_k^{*,m} \geq t_k\}/M$$

Calculating Romano-Wolf Adjusted p-values

Consider k hypotheses ordered such that $t_1 \geq t_2 > \ldots \geq t_k$

1. Step 1. Calculate initial set of adjusted p-values

$$\hat{p}_k^0 = \#\{\max_k^{*,m} \geq t_k\}/M$$

2. Step 2. Enforce monotonicity: for $j = 2, \ldots, k$, let

$$\hat{p}_j = \max\{\hat{p}_j^0, \hat{p}_{j-1}\}$$

Calculating Romano-Wolf Adjusted p-values

Consider k hypotheses ordered such that $t_1 \geq t_2 > \ldots \geq t_k$

1. Step 1. Calculate initial set of adjusted p-values

$$\hat{p}_k^0 = \#\{\max_k^{*,m} \geq t_k\}/M$$

2. Step 2. Enforce monotonicity: for $j = 2, \ldots, k$, let

$$\hat{p}_j = \max\{\hat{p}_j^0, \hat{p}_{j-1}\}$$

$\Rightarrow$ The j^{th} adjusted p-value cannot be lower than the $(j-1)^{\text{th}}$ p-value

Pros and Cons of Romano-Wolf Approach

Romano-Wolf can be implemented in Stata using `rwolf` command

```
rwolf y1 y2 y3, indepvar(x) controls(c1 c2) reps(250)
```

Pros and Cons of Romano-Wolf Approach

Romano-Wolf can be implemented in Stata using `rwolf` command

```
rwolf y1 y2 y3, indepvar(x) controls(c1 c2) reps(250)
```

Resampling-approach is computationally intensive

- Large data set, large number of hypotheses potentially problematic

Pros and Cons of Romano-Wolf Approach

Romano-Wolf can be implemented in Stata using `rwolf` command

```
rwolf y1 y2 y3, indepvar(x) controls(c1 c2) reps(250)
```

Resampling-approach is computationally intensive

- Large data set, large number of hypotheses potentially problematic

Romano-Wolf provides **strong control** of FWER

- Controls FWER for all combinations of true/false hypotheses
- Limiting FWER when all k hypotheses are true is **weak control**

Pros and Cons of Romano-Wolf Approach

Romano-Wolf can be implemented in Stata using `rwolf` command

```
rwolf y1 y2 y3, indepvar(x) controls(c1 c2) reps(250)
```

Resampling-approach is computationally intensive

- Large data set, large number of hypotheses potentially problematic

Romano-Wolf provides **strong control** of FWER

- Controls FWER for all combinations of true/false hypotheses
- Limiting FWER when all k hypotheses are true is **weak control**
- **Strong control means relatively low statistical power**

Controlling the False Discovery Rate

Anderson (JASA, 2008): "[Family-wise error rate] adjustments become increasingly severe as the number of tests grows — it is inherent in controlling the probability of making a single false rejection."

- Alternative is to tolerate some small number of false positives

Controlling the False Discovery Rate

Anderson (JASA, 2008): "[Family-wise error rate] adjustments become increasingly severe as the number of tests grows — it is inherent in controlling the probability of making a single false rejection."

- Alternative is to tolerate some small number of false positives

The **false discovery rate**: expected proportion of rejections that are Type I errors (i.e. where null was true and should not have been rejected)

Controlling the False Discovery Rate

Anderson (JASA, 2008): "[Family-wise error rate] adjustments become increasingly severe as the number of tests grows — it is inherent in controlling the probability of making a single false rejection."

- Alternative is to tolerate some small number of false positives

The **false discovery rate**: expected proportion of rejections that are Type I errors (i.e. where null was true and should not have been rejected)

- FWER and FDR are identical under the null (all rejections are errors)
- When some null hypotheses are false, FDR adjustments can be less stringent than FWER adjustments (because FDR $<$ FWER)

Controlling the False Discovery Rate

Anderson (JASA, 2008): "[Family-wise error rate] adjustments become increasingly severe as the number of tests grows — it is inherent in controlling the probability of making a single false rejection."

- Alternative is to tolerate some small number of false positives

The **false discovery rate**: expected proportion of rejections that are Type I errors (i.e. where null was true and should not have been rejected)

- FWER and FDR are identical under the null (all rejections are errors)
- When some null hypotheses are false, FDR adjustments can be less stringent than FWER adjustments (because FDR $<$ FWER)

Thought experiment: Let $k = 100$. The first 20 hypotheses are false, and clearly rejected using any approach. What expected number of false rejections you are willing to accept in the remaining set of 80 hypotheses?

Controlling the False Discovery Rate

Benjamini & Hochberg (1995) propose an approach to FDR control:

1. Order k p-values from smallest to largest, $p_1, p_2, ..., p_j, ..., p_k$, where j indicates the rank of the p-value for a specific hypothesis
2. Rejecting all p-values with $p_j < qj/k$ yields an expected FDR no higher than q when p-values are independent or positively correlated

Controlling the False Discovery Rate

Benjamini & Hochberg (1995) propose an approach to FDR control:

1. Order k p-values from smallest to largest, $p_1, p_2, ..., p_j, ..., p_k$, where j indicates the rank of the p-value for a specific hypothesis
2. Rejecting all p-values with $p_j < qj/k$ yields an expected FDR no higher than q when p-values are independent or positively correlated

All of the procedures discussed so far modify test sizes ("accept"/reject)

- We often want an adjusted p-value, not a yes/no decision

Controlling the False Discovery Rate

Benjamini & Hochberg (1995) propose an approach to FDR control:

1. Order k p-values from smallest to largest, $p_1, p_2, ..., p_j, ..., p_k$, where j indicates the rank of the p-value for a specific hypothesis
2. Rejecting all p-values with $p_j < qj/k$ yields an expected FDR no higher than q when p-values are independent or positively correlated

All of the procedures discussed so far modify test sizes ("accept"/reject)

- We often want an adjusted p-value, not a yes/no decision

Anderson (2008) proposed intuitive approach to calculating BH q-values:

- Rescale p-values by number of hypotheses / p-value rank
- Adjust for non-monotonicity

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	$\times 5$		
0.002	$\times 5$		
0.040	$\times 5$		
0.041	$\times 5$		
0.099	$\times 5$		

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005		
0.002	0.010		
0.040	0.200		
0.041	0.205		
0.099	0.495		

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	$\times 5$	
0.002	0.010	$\times 4$	
0.040	0.200	$\times 3$	
0.041	0.205	$\times 2$	
0.099	0.495	$\times 1$	

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	$\times 5$	$\times 5/1$
0.002	0.010	$\times 4$	$\times 5/2$
0.040	0.200	$\times 3$	$\times 5/3$
0.041	0.205	$\times 2$	$\times 5/4$
0.099	0.495	$\times 1$	$\times 5/5$

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	$\times 5$	$\times 5$
0.002	0.010	$\times 4$	$\times 2.5$
0.040	0.200	$\times 3$	$\times 1.67$
0.041	0.205	$\times 2$	$\times 1.25$
0.099	0.495	$\times 1$	$\times 1$

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	0.005	0.005
0.002	0.010	$\times 4$	$\times 2.5$
0.040	0.200	$\times 3$	$\times 1.67$
0.041	0.205	$\times 2$	$\times 1.25$
0.099	0.495	$\times 1$	$\times 1$

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	0.005	0.005
0.002	0.010	$\times 4$	$\times 2.5$
0.040	0.200	$\times 3$	$\times 1.67$
0.041	0.205	$\times 2$	$\times 1.25$
0.099	0.495	0.099	0.099

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	0.005	0.005
0.002	0.010	0.008	0.005
0.040	0.200	0.120	0.067
0.041	0.205	0.082	0.051
0.099	0.495	0.099	0.099

Multiple Test Corrections: Example

p-value	Bonferroni	Holm	Anderson
0.001	0.005	0.005	0.005
0.002	0.010	0.008	0.005
0.040	0.200	0.120	0.051
0.041	0.205	0.120	0.051
0.099	0.495	0.120	0.099

Multiple Hypothesis Testing: Summary

Try to avoid testing a large number of hypotheses

- Aggregate your main outcomes into indices (when appropriate)
- Consider pre-specifying "surprising" relationships

Acceptable adjustments differ in complexity, control/power tradeoffs

- Use simple approaches (Bonferroni, Holm) when they work
- Choose more control vs. more power when appropriate

Be suspicious of (your own and others') p-values near significance cutoffs