ExpEcon Methods:
Multiple Hypotheses Corrections

ECON 8877

P.J. Healy

First version thanks to Floyd Carey

Updated 2026-01-28

Multiple Hypothesis Corrections

Multiple Tests

The setup:

- You're going to run m tests (say, $m = 2$)
- Each test on its own has a 0.05 Type-I error
- The game is that you "win" if *at least* one is significant
    - Publish a paper, claim a result, etc.
- Problem: $\uparrow m \Rightarrow \uparrow$ chance you win

Definition
The **Family-Wise Error Rate (FWER)** is the probability that you "win" given that all m null hypotheses are true.

- We want FWER to be 0.05, not each test

Multiple Tests

Suppose you run two tests and "**win**" if one is significant
Each has 0.05 Type-I error. Baseline case: independent tests

Independent tests:

	Accept	Reject	
Accept	0.9025	0.0475	0.95
Reject	0.0475	0.0025	0.05
	0.95	0.05	$FWER = 0.0975$

Šidák correction: use a lower α:

	Accept	Reject	
Accept	$(1-\alpha)^2$	$(1-\alpha)\alpha$	$1-\alpha$
Reject	$(1-\alpha)\alpha$	α^2	α
	$1-\alpha$	α	$FWER = 2\alpha - \alpha^2$

For $FWER = 0.05$ use $\alpha \approx 0.025321$. If you have k tests:
$1-(1-\alpha)^k = 0.05 \Rightarrow \alpha^* = 1-(1-0.05)^{1/k}$ which is $> 0.05/k$

Multiple Tests

Worst-case situation: perfectly negative correlation
Need a bigger correction!

Perfect Negative Correlation:

	Accept	Reject	
Accept	0.90	0.05	0.95
Reject	0.05	0	0.05
	0.95	0.05	$Pr(R) = 0.10$

Bonferroni correction: what's the right α?

	Accept	Reject	
Accept	$1 - 2\alpha$	α	$1 - \alpha$
Reject	α	0	α
	$1 - \alpha$	α	$Pr(R) = 2\alpha$

For $Pr(R) = 0.05$ use $\alpha \approx 0.025$
k tests: $1 - (1 - k\alpha) = 0.05 \Rightarrow k\alpha = 0.05 \Rightarrow \alpha^* = 0.05/k$

Multiple Tests

Best case: perfect positive correlation

Extra tests don't add to the FWER!

Perfect Positive Correlation:

	Accept	Reject	
Accept	0.95	0	0.95
Reject	0	0.05	0.05
	0.95	0.05	$Pr(R) = 0.05$

No correction needed!

	Accept	Reject	
Accept	$1 - \alpha$	0	$1 - \alpha$
Reject	0	α	α
	$1 - \alpha$	α	$Pr(R) = \alpha$

Šidák or Bonferroni would be way too conservative!

The Bonferroni Correction

Setup:

- k tests. Nulls: $H_o^1, \ldots, H_o^k$
- α_f is your adjusted p-value on each
- FWER (Family-Wise Error Rate) is $Pr(R)$ on at least one test

Bonferonni Correction: $\alpha_f = \alpha/k$

- The most popular (and conservative)
- Safe: appropriate regardless of correlation
- Too safe: likely have FWER < 0.05
- Tradeoff: high chance of Type-II error (failure to reject false H_o)

Šidák Correction: $\alpha_f = 1 - (1 - \alpha)^{1/k}$

- Exact correction for independent tests
- In practice, Bonferroni $\approx$ Šidák

The Holm-Bonferroni Correction (Holm, 1979)

- More powerful while keeping FWER ≤ 0.05
- Order the p-values lowest to highest ($p_1 \leq p_2 \leq \ldots \leq p_k$). Will test nulls ($H_0^1, H_0^2, \ldots, H_0^k$) *sequentially*:

The Holm-Bonferroni Correction (Holm, 1979)

- More powerful while keeping FWER ≤ 0.05
- Order the p-values lowest to highest ($p_1 \leq p_2 \leq \ldots \leq p_k$). Will test nulls ($H_0^1, H_0^2, \ldots, H_0^k$) *sequentially*:

1. Is $p_1 < \frac{\alpha}{k}$?
   - Yes: Reject H_0^1 and move on to test H_0^2.
   - No: Do not reject any H_0^i (as in Bonferroni). Stop.
     - Note: $k-1$ tests remaining, so correction increases to $\frac{\alpha}{k-1}$
2. Is $p_2 < \frac{\alpha}{k-1}$?
   - Yes: Reject H_0^2 as well and continue. $k-2$ tests remain.
   - No: Do not reject H_0^2 through H_0^k. Stop.

j. Is $p_j < \frac{\alpha}{k+1-j}$?
   - Yes: Reject H_0^j as well and continue.
   - No. Do not reject H_0^j through H_0^k. Stop.

Recall: Bonferroni also allows only some nulls to be rejected. Same. "Win" if you reject at least one? Then Holm-Bonferroni = Bonferroni.
Can use Šidák version assuming independence: $1-(1-\alpha)^{1/(k+1-j)}$

The Hotchberg Step-Down Procedure

- Holm-Bonferonni: Reject $H_0^1, \ldots, H_0^j$ where j is the smallest index for which $p_{j+1} \geq \frac{\alpha}{k+1-(j+1)}$
  - Reject up to the "first crossing" of the threshold
- Hotchberg procedure: Reject $H_0^1, \ldots, H_0^j$ where j is the largest index for which $p_j \leq \frac{\alpha}{k+1-j}$
  - Reject up to the "last crossing" of the threshold
- Alternatively, first crossing when working top-to-bottom.
- This method is more powerful than the Holm-Bonferroni correction, but it sometimes does not control the FWER (see Dmitrienko et al., 2010 for details).
  - Not valid for negative correlation (worst case)

Balanced Resampling Using Bootstrapping

- How do we know the correlation across tests??
- Can be estimated via resampling methods! Hooray!
- Romano and Wolf (2005,2005,2016)
- This, combined with a “step-down" method like that used in Holm (1979), creates a more powerful correction.
- Furthermore, this method also creates balance, such that all tests contribute equally to error control.
- List et al. (2019) develop version of this correction for experiments where treatment is randomly assigned.

I would use these methods!
Stata: `rwolf` can be downloaded

The Family-Wise Error Rate

- What is the “family” in the Family-Wise Error Rate? What tests should be “combined”?
  - A “family” is (frustratingly) loosely defined, but an intuitive way to think about it is a set of tests whose inference is getting at the same question.
  - An easy experimental example: suppose you have two treatments and a control group, and you want to determine if either of the treatments increased the mean, so you perform two t-tests. Both of those t-tests constitute a family.

When to Use Corrections?

- Some ~~people~~ non-statisticians say we should *never* use them (O'Keefe, 2003; Perneger, 1998; Rothman, 1990)
- Other ~~people~~ non-statisticians say we should *always* use them (Bennett et al., 2009; Goeman & Solari, 2014; Moyé, 1998; Ottenbacher, 1998)
- Still others say we should use them only in exploratory research (Armstrong, 2014; Cramer et al., 2016; Streiner, 2015)
- Finally, some say we should use them only in confirmatory research (Bender & Lange, 2001; Schochet, 2009; Stacey et al., 2012; Tutzauer, 2003; Wason et al., 2014)

When to Use Corrections? (Continued)

- MHT corrections in ExpEcon are fairly rare, but growing.
- List et al. (2019) argue we *should* correct when:

1. Multiple outcomes for a given treatment (eg, GPA, SAT, ACT)
2. Multiple subgroups for given trt (eg, Black, Hispanic, Asian)
3. Multiple treatments (eg, different incentive schemes on effort)

PJ's view: "Can your claim be validated by *any* test being significant?"
Thus, depends on what you're claiming!
Ex: "Giving teachers incentives improves educational outcomes"
vs. "Giving teachers incentives improves GPAs"

If GPA, SAT, and ACT are 3 different hypotheses, then don't correct

Discussion: I write a paper with 20 unrelated hypotheses. MHT??

When to Use Corrections? (Continued)

Fancy language for this idea: Rubin (2021)

- Disjunction testing: "win" if you reject at least once
  - H_0: both green and red jelly beans *do not* cause acne
  - H_1: either green or red jelly beans (or both) cause acne
  - Rule: reject H_0 if either is significant (correction needed!)
- Conjunction testing: "win" only if you reject all
  - H_0: either green or red jelly beans (or both) *do not* cause acne
  - H_1: both green and red jelly beans cause acne
  - Rule: reject H_0 if bother are significant (no correction needed)
- Individual testing: each test has its own "win"
  - H_0^1: green jelly beans do not cause acne
  - H_1^1: green jelly beans cause acne
  - H_0^2: red jelly beans do not cause acne
  - H_1^2: red jelly beans cause acne
  - Rule: two separate tests. No ex-ante "either/both" claim

Conclusion

- Mathematically, there are ways to correct for MHT
- Degree of correction should depend on correlation across tests
  - Resampling methods (Wolf-Romano) are best here.
- The hard question is: when to use them??
  - Be precise in what exactly you're claiming!
  - "educational outcomes" vs. "GPAs"
- Final thought: readers are decent Bayesians. They don't just blindly trust 0.05. Tie your hands with preregistration and report everything. They'll update appropriately.