ExpEcon Methods: Multiple Hypotheses Corrections

ECON 8877

P.J. Healy

First version thanks to Floyd Carey

Updated 2023-11-17

Multiple Hypothesis Corrections

Multiple Tests

Suppose you run two tests of the same hypothesis.
Each has 0.05 Type-I error.

Independent tests:

	Accept	Reject	
Accept	0.9025	0.0475	0.95
Reject	0.0475	0.0025	0.05
	0.95	0.05	$Pr(R) = 0.0975$

So, 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	$Pr(R) = 2\alpha - \alpha^2$

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

Multiple Tests

Suppose you run two tests of the same hypothesis.
Each has 0.05 Type-I error.

Perfect Negative Correlation:

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

So, use a lower 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

Suppose you run two tests of the same hypothesis.
Each has 0.05 Type-I error.

Perfect Positive Correlation:

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

So, use a lower α:

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

No correction needed!
Using $\alpha^* = 0.05/k$ would be way too conservative!

The Bonferroni Correction

Setup:

- k tests. Nulls: $H_0^1, \ldots, H_0^k$
- 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)
- Under independence $FWER = 1 - (1 - \frac{\alpha}{k})^k \approx \alpha$
- Safe: appropriate even with negative correlation
- Tradeoff: high chance of Type-II error (failure to reject false H_0)

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

- Exact correction for independent tests

The Holm-Bonferroni Correction

- A more powerful (i.e., higher β) correction that still controls the FWER is the Holm-Bonferroni correction (Holm, 1979).
- For this correction, order the p-values in the family from lowest to highest ($p_1 \leq p_2 \leq \ldots \leq p_k$).

The Holm-Bonferroni Correction

- A more powerful (i.e., higher β) correction that still controls the FWER is the Holm-Bonferroni correction (Holm, 1979).
- For this correction, order the p-values in the family from lowest to highest ($p_1 \leq p_2 \leq \ldots \leq p_k$).
- Then follow the algorithm:

1. Is $p_1 < \frac{\alpha}{k}$?
    - No: Do not reject any H_0^i (as in Bonferonni). Stop.
    - Yes: Reject H_0^1 and continue to step 2.
        - Note: There are now $k - 1$ tests remaining.
2. Is $p_2 < \frac{\alpha}{k-1}$?
    - No: Do not reject H_0^2 through H_0^k. Stop.
    - Yes: Reject H_0^2 as well and continue. $k - 2$ tests remain.

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

Can use a Sidak 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

Issues with the Holm and Hotchberg Corrections

- They assume the “worst-case scenario"for the joint distribution of the test statistics (i.e., independence)
- They are not balanced, so that there is the potential for a rejection of H_0 for one test which has a higher unadjusted p-value than another test whose null hypothesis is not rejected.
- Romano and Wolf’s (2010) method deals with these issues and creates a correction that is more powerful than either the Holm or Hotchberg corrections.

Balanced Resampling Using Bootstrapping

- Resampling methods used in Romano and Wolf (2010) can estimate the degree of dependence between the test statistics.
- 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 experimental studies which randomly assign treatments to experimental treatments.

I would use these methods!

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)

- In the economics literature, these corrections are rarely used. However, List et al. (2019) argue that there are 3 scenarios under which experimental economists *should* use some kind of correction:

1. When there are multiple outcomes for a given treatment that researchers wish to analyze for a given treatment
2. When there is heterogeneity or expected heterogeneity in an effect across different subgroups
3. When there are multiple treatments and we wish to compare the effect size relative to a control or the other treatments

When to Use Corrections? (Continued)

- Recently, a paper by Rubin (2021) advocated for correction based on the *type* of multiple testing that occurs.
- The Jelly Bean Example (Munroe, 2011):

1. disjunction (union-intersection) testing: *neither* green jelly beans *nor* red jelly beans causes acne.
2. conjunction (intersection-union) testing: *either* green jelly beans *or* red jelly beans do not cause cane.
3. individual testing: red jelly beans cause do not acne; green jelly beans do not cause acne

Conclusion

- On one hand, allowing researchers to choose which paradigm to use creates an incentive problem
- On the other, we cannot decrease the Type-1 Error probability without increasing the Type-2 Error probability.
- It depends on what the experiments' goals are, the relative importance of Type-1 and Type-2 errors, and ultimately comes down to a few judgment calls.
- Pre-registration forces us to think more deeply about what questions we want to answer and how we'll answer them