Multiple Hypothesis Testing

Problems and Corrections

Julia Clark and Cameron Sells

Methods Workshop
June 2, 2016

Overview

1. What is the multiple comparisons problem (AKA multiple hypothesis testing, multiple inference, multiple significance testing, etc.)?

2. What are potential solutions?

3. How do we implement them in R?

1. The Problem

The Multiple Comparisons Problem

Suppose we have a family of m hypotheses that we are testing at significance level α:

- For any individual test, the probability of a false positive (Type I error) = α
- BUT, the joint probability of one or more T1 error in m independent tests is much higher $= 1 - (1 - \alpha)^m$

e.g., if m = 10 and $\alpha = 0.05$, the probability of at least one false positive is $1 - (1 - 0.05)^{10} \simeq 40\%$

It's technically a problem when …

Testing a **family** of hypotheses simultaneously and making a conclusion about an **individual** hypothesis.

- **Families:** "any collection of inferences for which it is meaningful to take into account some combined measure of error" (Hochberg & Tamhane 1987) —e.g., multiple outcomes, treatment arms, subgroup analyses, interactions, etc.
- Maybe OK if focusing on one hypothesis ***a priori***
- Not OK if picking and choosing which ones to reject

Examples

When in doubt, phrase as a question (are you comparing hypotheses against each other?):

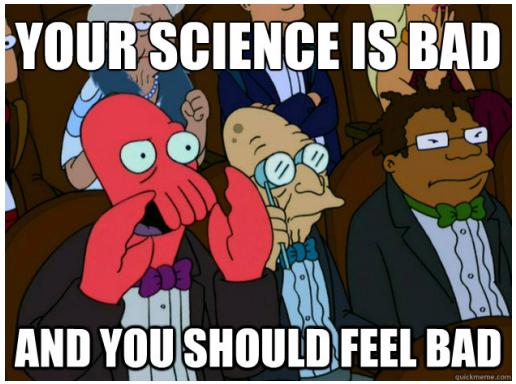
- What is the effect of school vouchers on achievement? {drop out rates, test scores, SATs, college admissions, salary}
- Which treatment affects voter turnout? {A, B, C, D, E}
- Does the treatment effect of GOTV campaign vary by age group? {18–25, 26–35, 36–45, …}

Generally

When testing m hypotheses, $H_1, H_2, \ldots, H_m$, at significance level α, the probability of getting k significant p-values when there are ***no true effects*** is given by:

$$\text{Binom}(k, m, \alpha) = \binom{m}{k} \alpha^k (1 - \alpha)^m$$

Is this a huge problem?



- Yes if you're running models with large families of variables
- Yes if you're cherrypicking a few significant results to fit your theory (p-hacking, etc)
- Less so with smaller m, highly correlated variables (but let us tell you about it anyway!)

2. Some Solutions

Categories of Solutions

(in roughly descending order of conservativeness)

A. Control the family-wise error rate (FWER)

B. Control the false-discovery rate (FDR)

C. Run a simulation

D. Make an index of outcome variables

E. Go Bayesian

F. (Design-based approach)

Recall ...

	Truth	
	$H_0 = T$	$H_0 = F$
Fail to Reject	True Negative	False Negative (T2)
	$Pr = 1 - \alpha$	$Pr = \beta$
Reject	False positive (T1)	True Positive
	$Pr = \alpha$	$Pr = 1 - \beta$

Approaches to dealing with the multiple comparisons problem trade off between TI and T2 errors

A. Controlling the family-wise error rate

FWER $= 1 - (1 - \alpha)^m$, the probability of one or more Type 1 errors in a family of m tests

- "Controling" the FWER means that for any significance level α:

$$\text{FWER} \leq \alpha$$

- Allows us to be $1 - \alpha$ confident that there are no false discoveries in m hypotheses
- BUT loss of power to reject false nulls

The Bonferroni (Boole) inequality

Say we have m hypotheses, $H_i, \ldots, H_m$. Let $Pr(H_i)$ be the probability that a test of H_i gives a false positive, and $Pr(\cup_{i=1}^{m} H_i)$ be the probability of at least one false positive (i.e., the FWER).

$$
\begin{aligned}
Pr(H_i) \cup Pr(H_j) &= Pr(H_i) + Pr(H_j) - Pr(H_i \cap H_j), \forall_{i \neq j} \\
&= Pr(H_i) + Pr(H_j) \text{ (if disjoint)} \\
&\rightarrow \\
Pr(\cup_{i=1}^{m} H_i) &\leq \sum_{i=1}^{m} Pr(H_i) \\
\text{FWER} &\leq \sum_{i=1}^{m} \alpha \\
\text{FWER} &\leq m\alpha
\end{aligned}
$$

Bonferroni's correction

(classic and impractical)

If the upper bound of the FWER in m tests is $m\alpha$, we can control it by:

- Setting confidence level $\alpha^* = \alpha/m$
- Rejecting when $\hat{p}_i \leq \alpha^*$

Lots of variations, including adding weights to different hypotheses: reject if $\hat{p}_i \leq w_m(\alpha/m)$, where $w_m \geq 0, \sum w_m = 1$.

Bonferroni Example

Let's say we have ten p-values ($m = 10$), and $\alpha = 0.05$. The Bonferroni-adjusted $\alpha^* = 0.05/10 = 0.005$, so

m	$\hat{p}$	B_{sig}?
1	0.001	Y
2	0.003	Y
3	0.005	N
4	0.017	N
5	0.025	N
6	0.034	N
7	0.046	N
8	0.053	N
9	0.160	N
10	0.250	N

Should you use Bonferroni?

(probably not)

- Appropriate for **worst case** scenario: all tests are independent and even one false positive is a big problem
- Easy and simple—same adjustment for all p-values, and can apply to other people's regression tables

…but WAY too conservative when there is dependence (if dependence is perfect, FWER $\rightarrow \alpha$) or if restricting Type I errors isn't your top priority

Holm's method

(an improvement over Bonferroni that's still pretty conservative)

Controls FWER using a **stepwise** (step-up) method:

1. Order p-values $1 \ldots m$ from smallest to largest
2. Find the **smallest** p-value such that

$$p_k > \frac{\alpha}{m + 1 - k}, \text{ where } k \text{ is the p-value index}$$

3. Declare this and all larger p-values ***insignificant***

Holms vs. Bonferroni

Bonferroni: reject when
$p \leq \alpha/m = 0.005$

Holm: no reject when
$p_k > \frac{\alpha}{m+1-k} = \frac{0.05}{10+1-k}$

k	$\hat{p}$	B_{rej}?	H_{value}	H_{rej}?
1	0.001	Y	0.005	Y
2	0.003	Y	0.006	Y
3	0.005	N	0.006	Y
4	0.017	N	0.007	N
5	0.025	N	0.008	N
6	0.034	N	0.010	N
7	0.046	N	0.013	N
8	0.053	N	0.017	N
9	0.160	N	0.025	N
10	0.250	N	0.050	N

Advantages of Holm over Bonferroni

Starts by comparing most significant hypothesis to Bonferroni value:

$$p_k > \frac{\alpha}{m+1-k} \equiv p_k > \frac{\alpha}{m} \text{ for } k = 1$$

If the condition is unmet, this hypothesis is rejected and it moves to the next, $p_k > \frac{\alpha}{m-1} \rightarrow$ with each, it reduces the critical value, gaining power.

$\Longrightarrow$ Hypotheses rejected by Bonferroni are also rejected by Holm + a few more.

B. Controlling the false-discovery rate

(allows more false positives than FWER)

FDR = The expected percent of rejections that are type I errors in a family of tests:

$$E\left[\frac{\text{false rejections}}{\text{total rejections}}\right]$$

FDR vs. FWER

Let V be the number of **false rejections** (T1 errors) and R be the **total number of rejections**:

$$\text{FWER} = Pr(V > 0)$$
$$\text{FDR} = E[V/R]$$

- If ALL null hypotheses are true, $V = R$ and:
  FDR = $E[V/R] = Pr(R > 0) = Pr(V > 0)$ = **FWER**
- If one or more null hypotheses is false, FDR is less conservative than FWER, so you get more power to reject false nulls

Benjamini-Hochberg (BH)

(like Holm with more power)

Controls FDR using a **stepwise** (step-down) method:

1. Order p-values $1 \ldots m$ from smallest to largest
2. Find the **largest** p-value such that

$$p_k \leq \frac{k\alpha}{m}, \text{ where } k \text{ is the p-value index}$$

3. Declare this and all smaller p-values ***significant***

Like FWER, lots of variations (e.g., weighting hypotheses, etc.)

Holms vs. Bonferroni vs. BH

Bonferroni: reject when
$p \leq \alpha/m = 0.005$

Holm: no reject when
$p_k > \frac{\alpha}{m+1-k} = \frac{0.05}{10+1-k}$

BH: reject when
$p_k \leq \frac{k\alpha}{m} = \frac{k(0.05)}{10}$

k	$\hat{p}$	$B_{rej}?$	H_{val}	$H_{rej}?$	BH_{val}	$BH_{rej}?$
1	0.001	Y	0.005	Y	0.005	Y
2	0.003	Y	0.006	Y	0.010	Y
3	0.005	N	0.006	Y	0.015	Y
4	0.017	N	0.007	N	0.020	Y
5	0.025	N	0.008	N	0.025	Y
6	0.034	N	0.010	N	0.030	N
7	0.046	N	0.013	N	0.035	N
8	0.053	N	0.017	N	0.040	N
9	0.160	N	0.025	N	0.045	N
10	0.250	N	0.050	N	0.050	N

Advantages of BH over FWER

- Keeps Type 2 errors as low as possible
- Penalty scales with the number of hypotheses
- Gains in power larger when fewer nulls are true
- BUT mixed results with dependence

C. Simulation

(best at dealing with dependence between tests)

See R implementation up next, also EGAP tools and Anderson (2008).

D. Construct index of outcome variables

(instead of changing p-values, reduce m)

1. Transform variables so "beneficial" effects go in same direction
2. Calculate z-score for each variable (using mean/sd of control group)
3. Combine outcomes into single score:
   A. **Mean effects index:** Sum z-scores
   B. **Inverse-covariance weighted matrix index (ICWMI):** Calculate weighted index using inverted var-cov matrix of z-scores

Should you use an index?

Pros:

- Good for “general effect”
- May gain power (average out random variation in outcome measures)
- ICWMI more efficient (less weight to highly correlated outcomes)

Cons:

- Not as good for theory-building
- For ICWMI, possible to generate negative weights (reversing direction of effect)
- Not helpful for multiple treatment arms, interactions, etc.

E. Bayesian solution

(stop worrying so much about multiple comparisons)

Gelman, Hill & Yajima, 2012:

- Not super concerned about Type 1 error, because rarely believe null is strictly true (i.e., that $H_0 = 0$)
- Real problem isn't multiple comparisons, it's "insufficient modeling of the relationship between the corresponding parameters of the model" $\rightarrow$ use multi-level model to build multiplicity in from the beginning

Read paper for more …

F. Design-based approach

(not a technical fix, but …)

1. **Pre-analysis plans** can help:
   - Clarify number of comparisons and how you will address them (reduce researcher degrees of freedom)
   - Specify "primary" hypotheses (disagreement over this)
2. **Replication** can help root out false positives (and negatives) in previous work

3. R Implementation

References

- Holm (1979)
- Benjamini & Hochberg (1995)
- Benjamini & Hochberg (2000)
- Westfall & Young (1997)
- Kling, Liebman & Katz (2005)
- Romano & Wolf (2005)
- Anderson (2008)
- Gelman et al (2012)
- EGAP tools