Introduction to Multiple Testing

Jarvis T. Chen (jarvis@hsph.harvard.edu)

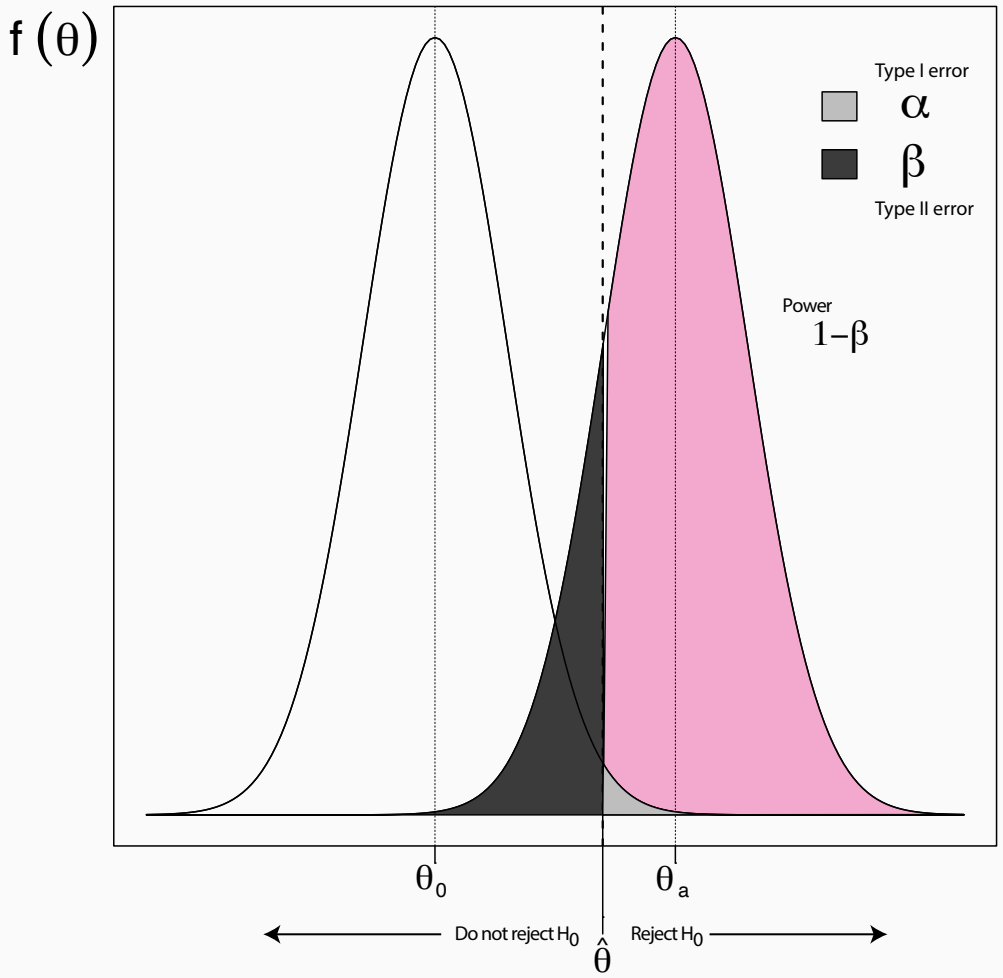
2 May 2023

Harvard T. H. Chan School of Public Health

Outline

- background and concepts
- Family-Wise Error Rate (FWER)
- False Discovery Rate (FDR)
- Other methods
- Conceptual challenges

Hypothesis Testing



Multiple Testing

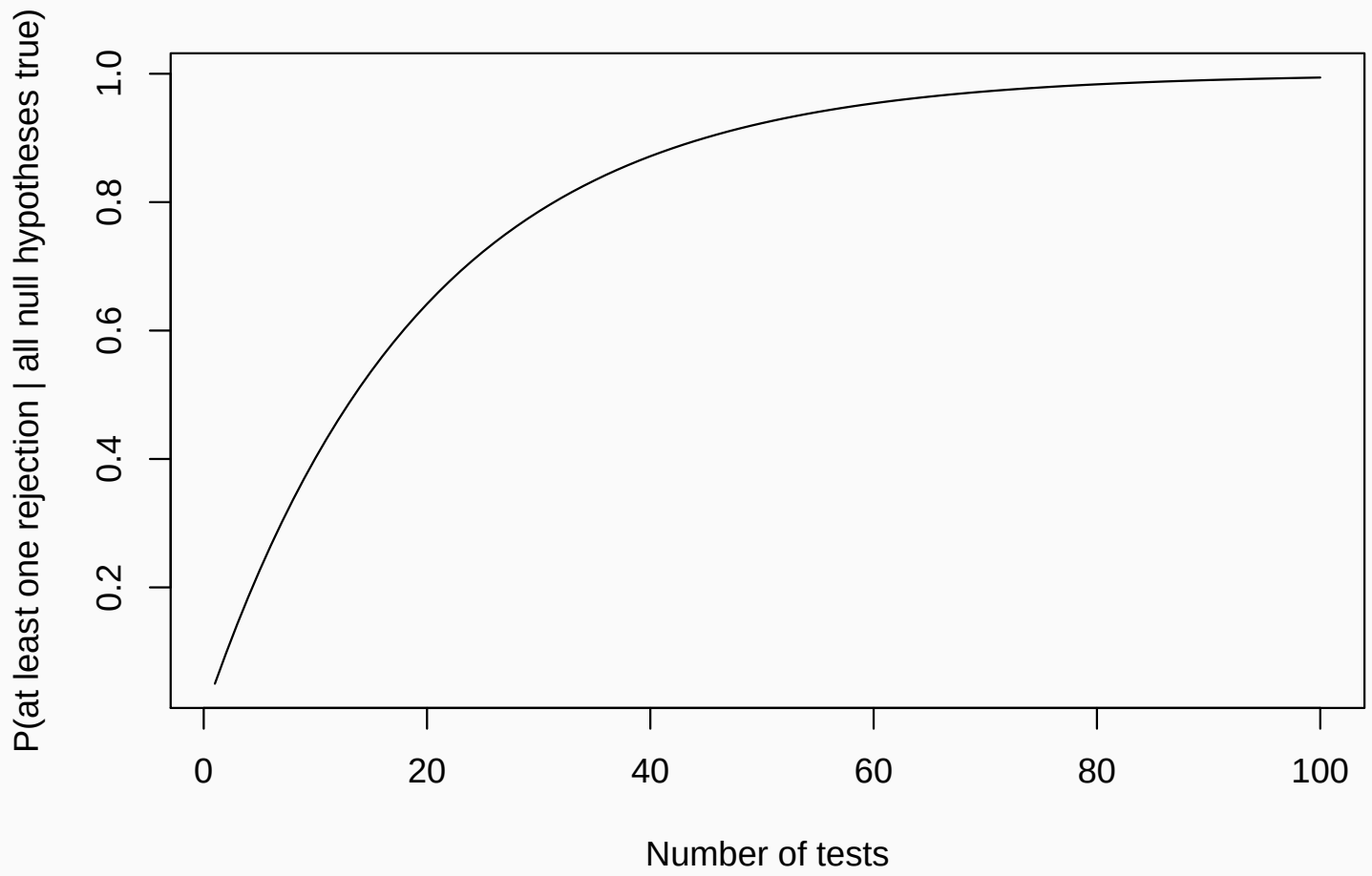
What happens when we test multiple hypotheses?

Example: Let's say that we are performing 10 hypothesis tests and for each the null is true. What is the probability that at least one p-value will be < 0.05?

For a given hypothesis, $P(\text{reject } H^0 | H^0 \text{ true}) = 0.05$. Assuming these are independent,

$$\begin{aligned} P(\text{at least one rejection}|\text{joint null is true}) &= 1 - P(\text{no rejections}|\text{joint null is true}) \\ &= 1 - (1 - 0.05)^{10} \\ &= 0.40 \end{aligned}$$

Multiple Testing



Multiple testing

Multiple testing scenarios include:

- exploring multiple exposures of interest
- exploring multiple outcome variables
- exploring multiple interactions between variables

What Constitutes a Hypothesis?

- While it is relatively obvious what the primary outcome and exposures are in RCTs, the same does not hold in epidemiological or social science studies
- With the many data sets available these days, it seems quite natural to just “play around” with the data a bit
- This kind of exploratory analysis is important because it can detect potentially interesting associations, but also increases the risk of false discoveries

Multiple testing

Possibilities when m tests are performed

	Accept Null	Reject Null	
H_0 true	U	V	m_0
H_1 true	T	S	m_1
	$m - R$	R	m

When m tests are performed, the aim is to decide which of the nulls should be rejected.

- m_0 is the number of *true nulls*
- V is the number of *Type I errors*
- T is the number of *Type II errors*

(Note that we only observe m and R).

Family-Wise Error Rate

The Family-Wise Error Rate (FWER) is the probability of making *at least* one Type I error, i.e.

$$Pr(V \geq 1 | H_1 = 0, \ldots, H_m = 0)$$

Intuitively, this seems like a sensible criterion if one has a strong prior belief that all of the null hypotheses are true, since in such a situation making at least one Type I error should be penalized.

In contrast, if one believes that a number of the nulls are likely to be false, then one would be prepared to accept a greater number of Type I errors in exchange for discovering more true associations.

As in all hypothesis testing situations, we will need to trade off Type I and Type II errors.

Family-Wise Error Rate

Let V_i be the event that the ith null is incorrectly rejected, so that, with respect to the table above, V, the random variable representing the number of incorrectly rejected nulls, corresponds to $\cup_{i=1}^{m} V_i$. With a common level for each test α^*, the FWER is

$$\begin{aligned}
\alpha_F &= \Pr(V \geq 1 | H_1 = 0, \ldots, H_m = 0) \\
&= \Pr(\cup_{i=1}^{m} V_i | H_1 = 0, \ldots, H_m = 0) \\
&\leq \sum_{i=1}^{m} \Pr(V_i | H_1 = 0, \ldots, H_m = 0) \\
&= m\alpha^*
\end{aligned}$$

Family-Wise Error Rate

That is, to keep the **overall** FWER at α_F, we want to select α^* so that the overall probability of making at least one Type I error is α_F.

The easiest (and toughest) way to do this is to use the Bonferroni FWER correction, taking

$$\alpha^* = \frac{\alpha_F}{m}$$

to give FWER $\leq \alpha_F$.

Family-Wise Error Rate: Bonferroni method

For example, to control the FWER at a level of $\alpha = 0.05$ with $m = 10$ tests, we would take

$$\begin{aligned} \alpha_F &= \frac{0.05}{10} \\ &= 0.005 \end{aligned}$$

The Bonferroni method is stringent (i.e. conservative in the sense that the bar is set high for rejection) and so results in a loss of power in the usual situation in which the FWER is set to a low value (e.g. 0.05).

Family-Wise Error Rate: Šidák correction

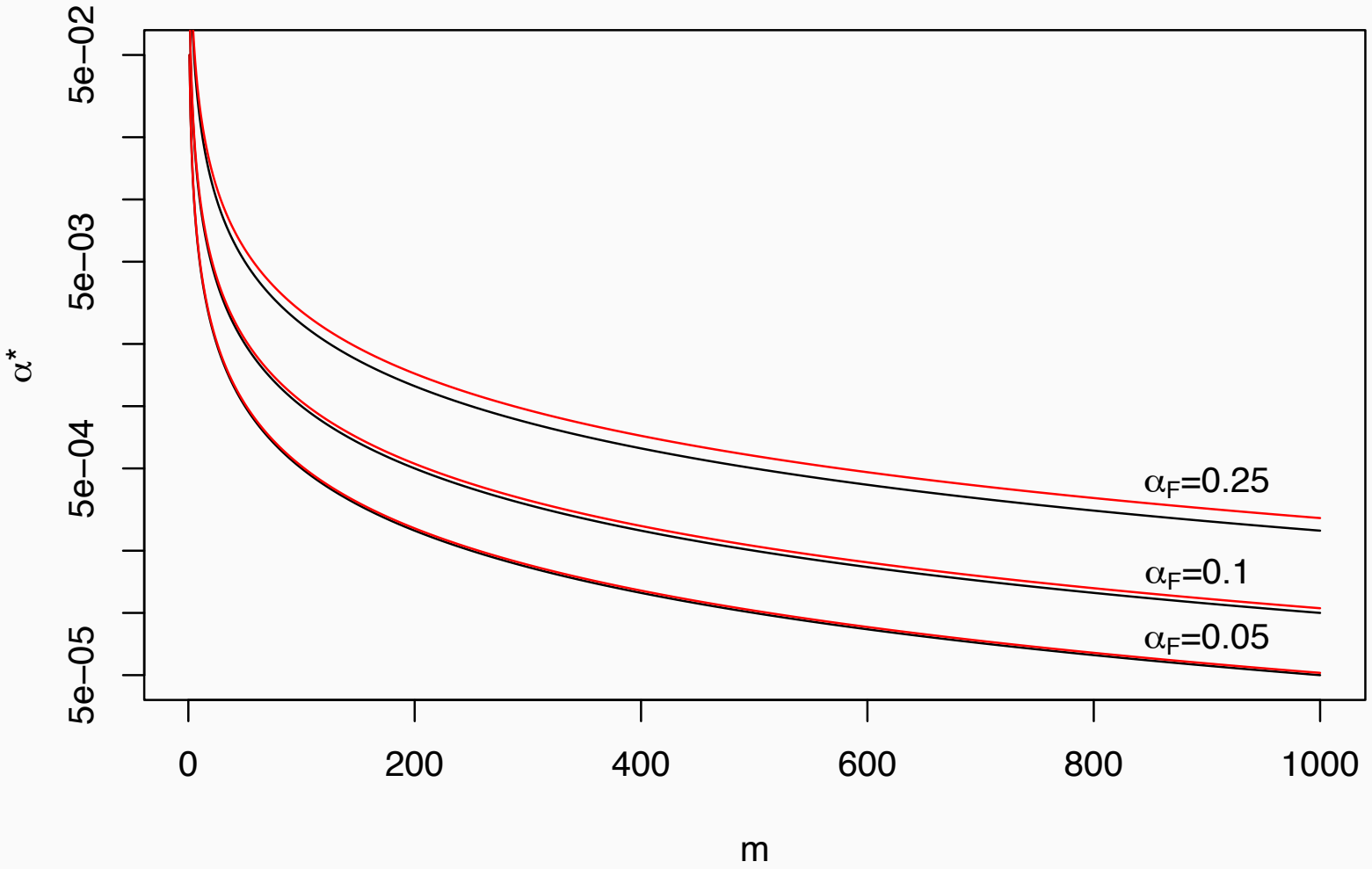
If the test statistics are independent, Šidák (1967) showed that the FWER can be expressed as

$$\begin{aligned}
\Pr(V \geq 1) &= 1 - Pr(V = 0) \\
&= 1 - \Pr(\cap_{i=1}^{m} V_i') \\
&= 1 - \prod_{i=1}^{m} V_i' \\
&= 1 - (1 - \alpha^*)^m
\end{aligned}$$

where V_i' indicates the complement of V_i (i.e. that the ith null is **not** rejected).

Consequently, to achieve FWER $= \alpha_F$, we can take $\alpha^* = 1 - (1 - \alpha_F)^{1/m}$. This is slightly less restrictive than the Bonferroni correction and provides marginally more power.

Family-Wise Error Rate: Šidák correction



Step-Down Procedure: Holm 1979

Holm (1979) proposed to test hypotheses sequentially. The goal is to increase the power of the statistical tests while keeping the FWER under control.

1. Perform the tests in order to obtain their p-values.
2. Order the p-values from smallest (most significant) to largest (least significant).
3. Use Bonferroni or Šidák*for the first hypothesis, for a family of m tests.
4. If the test is not significant, the procedure stops.
5. If the first test is significant, the test with the second smallest p-value is then corrected with a Bonferroni or Šidák correction for a family of $(m - 1)$ tests.
6. Proceed in this manner through the ordered list of tests, stopping when the p-value for the kth test is no longer significant.

*Šidák requires that the hypothesis tests are not negatively dependent

Example: Holm Step-Down

$m = 10$

		Traditional		Holm Step Down			
		Accept/Reject		Cutoffs		Accept/Reject	
Test	p-value	Bonferroni	Šidák	Bonferroni	Šidák	Bonferroni	Šidák
1	0.001511	Reject	Reject	0.005000	0.005116	Reject	Reject
2	0.002257	Reject	Reject	0.005556	0.005683	Reject	Reject
3	0.004326	Reject	Reject	0.006250	0.006391	Reject	Reject
4	0.005066	Accept	Reject	0.007143	0.007301	Reject	Reject
5	0.007531	Accept	Accept	0.008333	0.008512	Reject	Reject
6	0.007694	Accept	Accept	0.010000	0.010206	Reject	Reject
7	0.009707	Accept	Accept	0.012500	0.012741	Reject	Reject
8	0.009970	Accept	Accept	0.016667	0.016952	Reject	Reject
9	0.025251	Accept	Accept	0.025000	0.025321	Accept	Reject
10	0.044535	Accept	Accept		0.050000	Accept	Reject

Step-Down Procedure: Holm 1979

As with the Bonferroni and Šidák corrections more generally, Holm's sequential approach is conservative when the tests are not independent.

The approach is also conservative when the number of comparisons becomes very large.

Step-Up Procedure: Hochberg (1988)

1. Perform the tests in order to obtain their p-values.
2. Order the p-values from smallest (most significant) to largest (least significant), $P_{(1)}, \ldots, P_{(m)}$, and let the associate hypotheses be $H_{(1)}, \ldots, H_{(m)}$.
3. For a given α, let R be the largest k such that

$$P_{(k)} \leq \frac{\alpha}{m + 1 - k}$$

4. Reject the null hypotheses $H_{(1)}, \ldots, H_{(R)}$

Holm (Step-Down) vs. Hochberg (Step-Up)

Holm:

- Begin by ordering the p-values, $P_{(1)}, \ldots, P_{(m)}$, from smallest (most significant) to largest (least significant).
- Find the **smallest** k for which

$$P_{(k)} > \frac{\alpha}{m+1-k}$$

- **Reject** $H_{(1)}, \ldots, H_{(k-1)}$ and **accept** $H_{(k)}, \ldots, H_{(m)}$.

Hochberg:

- Begin by ordering the p-values, $P_{(1)}, \ldots, P_{(m)}$, from smallest (most significant) to largest (least significant).
- Find the **largest** k for which

$$P_{(k)} \leq \frac{\alpha}{m+1-k}$$

- **Reject** $H_{(1)}, \ldots, H_{(k)}$ and **accept** $H_{(k+1)}, \ldots, H_{(m)}$.

Step-Down vs. Step-Up

$m = 10$

Test	p-value	Bonferroni $\alpha/(m+1-k)$	Holm Step Down	Hochberg Step Up
1	0.002005	0.005000	Reject	Reject
2	0.004674	0.005556	Reject	Reject
3	0.005387	0.006250	Reject	Reject
4	0.008667	0.007143	Accept	Reject
5	0.009010	0.008333	Accept	Reject
6	0.009942	0.010000	Accept	Reject
7	0.011271	0.012500	Accept	Reject
8	0.018642	0.016667	Accept	Accept
9	0.035487	0.025000	Accept	Accept
10	0.071872	0.050000	Accept	Accept

- Holm: The **smallest** k for which the p-value is greater than $\alpha/(m+1-k)$ is 4, so reject $H_{(1)}, \ldots, H_{(3)}$.
- Hochberg: The **largest** k for which the p-value is less than or equal to $\alpha/(m+1-k)$ is 7, so reject $H_{(1)}, \ldots, H_{(7)}$.

Step-Down vs. Step-Up

As we saw in the table above, Hochberg's procedure is more powerful than Holm's procedure (i.e. leads us to reject the null more).

Both procedures are more powerful than Bonferroni or Šidák and both still ensure control of the FWER.

Note that while the Hochberg procedure is uniformly more powerful than the Holm procedure, the Hochberg procedure requires the hypotheses to be independent or under certain forms of positive dependence, whereas the Holm procedure does not require this asumption.

False Discovery Rate

Rather than just focusing on Type I error, Benjamini and Hochberg (1995) recommend focusing on the relative likelihood of false discoveries. They define the **False Discovery Rate** as the expected proportion of rejected null hypotheses that are erroneously rejected.

The main idea of the FDR is to focus on the relative risk of false positive rather than the absolute risk. Rather than keeping the absolute risk the same, we want to make sure the proportion of false discoveries (among all discoveries) does not exceed a certain threshold.

False Discovery Rate

	Accept Null	Reject Null	
H_0 true	U	V	m_0
H_1 true	T	S	m_1
	$m - R$	R	m

Consider the problem of testing simultaneously m (null) hypotheses, of which m_0 are true. R is the number of hypotheses rejected. The specific m hypotheses are assumed to be known in advance. R is an observable random variable; U, V, S, and T are unobservable random variables.

False Discovery Rate

In terms of these random variables,

- The Per Comparison Error Rate (PCER) is $\mathbb{E}(V/m)$, i.e. the expected proportion of Type I errors if we were to ignore the multiple testing problem.
- Testing individually each hypothesis at level α guarantees that $\mathbb{E}(V/m) \leq \alpha$.
- The FWER is $P(V \geq 1)$.
- Testing individually each hypothesis at level α/m guarantees that $P(V \geq 1) \leq \alpha$.

False Discovery Rate

The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $Q = V/(V+S)$, i.e. the proportion of the rejected null hypotheses which are erroneously rejected. Define $Q = 0$ when $V + S = 0$, since no error of false rejection can be committed.

We define the FDR Q_e to be the expectation of Q,

$$Q_e = \mathbb{E}(Q) = \mathbb{E}[V/(V+S)] = \mathbb{E}(V/R)$$

False Discovery Rate

$$Q_e = \mathbb{E}(Q) = \mathbb{E}[V/(V+S)] = \mathbb{E}(V/R)$$

- Note that if all null hypotheses are true, the FDR is equivalent to the FWER.
- When $m_0 < m$, the FDR is smaller than or equal to the FWER. Any procedure that controls the FWER also controls the FDR. However, if a procedure controls the FDR only, it can be less stringent, and a gain in power may be expected. In particular, the larger the number of non-true null hypotheses is, the larger S tends to be, and so is the difference between the error rates. As a result, the potential for increase in power is larger when more of the hypotheses are non-true.

Benjamini and Hochberg's Procedure for Controlling FDR (1995)

- Let $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$ be the ordered p-values corresponding to $H_1, H_2, \ldots, H_m$. Denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$.
- Define the following Bonferroni-type multiple-testing procedure:

$$\text{Let } k \text{ be the largest } i \text{ for which } P_{(i)} \leq \frac{i}{m}\alpha$$

- Then reject all $H_{(i)}, i = 1, 2, \ldots, k$.
- If no such i exists, then reject no hypotheses.
- Benjamini and Hochberg show that for independent test statistics and for any configuration of false null hypotheses, this procedure controls the FDR at α.

Example: FDR vs. FWER

p-value	Holm Bonferroni	Rejected?	FDR Cutoff	Rejected?
0.004	0.0050			
0.009	0.0056			
0.012	0.0063			
0.017	0.0071			
0.021	0.0083			
0.040	0.0100			
0.051	0.0125			
0.060	0.0167			
0.100	0.0250			
0.400	0.0500			

Example: FDR vs. FWER

p-value	Holm Bonferroni	Rejected?	FDR Cutoff	Rejected?
0.004	0.0050	Reject		
0.009	0.0056	Accept		
0.012	0.0063	Accept		
0.017	0.0071	Accept		
0.021	0.0083	Accept		
0.040	0.0100	Accept		
0.051	0.0125	Accept		
0.060	0.0167	Accept		
0.100	0.0250	Accept		
0.400	0.0500	Accept		

Example: FDR vs. FWER

p-value	Holm Bonferroni	Rejected?	FDR Cutoff	Rejected?
0.004	0.0050	Reject	0.0050	
0.009	0.0056	Accept	0.0100	
0.012	0.0063	Accept	0.0150	
0.017	0.0071	Accept	0.0200	
0.021	0.0083	Accept	0.0250	
0.040	0.0100	Accept	0.0300	
0.051	0.0125	Accept	0.0350	
0.060	0.0167	Accept	0.0400	
0.100	0.0250	Accept	0.0450	
0.400	0.0500	Accept	0.0500	

Example: FDR vs. FWER

p-value	Holm Bonferroni	Rejected?	FDR Cutoff	Rejected?
0.004	0.0050	Reject	0.0050	Reject
0.009	0.0056	Accept	0.0100	Reject
0.012	0.0063	Accept	0.0150	Reject
0.017	0.0071	Accept	0.0200	Reject
0.021	0.0083	Accept	0.0250	Reject
0.040	0.0100	Accept	0.0300	Accept
0.051	0.0125	Accept	0.0350	Accept
0.060	0.0167	Accept	0.0400	Accept
0.100	0.0250	Accept	0.0450	Accept
0.400	0.0500	Accept	0.0500	Accept

FDR under dependency

- In many multiple testing situations, the variables of interest (and related hypotheses) are often correlated
- Benjamini and Yekutieli (2001) show that the standard Benjamini-Hochberg procedure controls the FDR in families with positively dependent test statistics
- If there is no clear prior regarding dependence or its structure, Benjamini and Yekutieli (2001) propose the following formula to control the FDR

$$P_{(i)} \leq \frac{i}{mc(m)}\alpha$$

noting that this procedure is more conservative.

FDR under dependency

- If tests are independent, then $c(m) = 1$ (i.e. the standard procedure)
- If tests are positively or negatively correlated,

$$c(m) = \sum_{j=1}^{m} \frac{1}{j}$$

- Example: $m = 10$ hypotheses

$$c(m) = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \frac{1}{10} = 2.93$$

$\rightarrow$ Critical p-values become smaller and there will be fewer rejections.

Resampling Class Methods

Westfall and Young 1993

- An alternative to the methods mentioned above is to bootstrap p-value distributions
- One commonly used method is the minP resampling class method, which is recommended for strongly correlated outcomes (Blakesley et al 2009)
- The main idea of this method is to use bootstrap samples to look at the distribution of minimum p-values (across outcomes) in random samples
- The adjusted p-value is then computed by looking up each p-value in the min list

Power Calculations

- In settings where we think that we will consider multiple outcomes, correlation between these outcomes seem likely, but are hard to define *ex ante*
- Given that we want to make sure we have enough power *ex post*, most researchers simply use Bonferroni-corrected α in their power calculations (α/m)
- This often results in rather large sample sizes, but is conservative and still allows for using other adjustments *ex post*

How Can We Reduce False Discovery Risks?

- Report all of the analysis conducted (not just the “significant” ones)
- Publish non-significant results (very important for meta-analyses)
- Develop analysis plan and pre-register them prior to starting analysis
- For RCTs: be clear about analyses outside of primary pre-specified ones

References

- Benjamini Y, Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* 57(1):289–300.
- Benjamini Y, Yekutieli Y. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* 29(4):1165-1188. doi: 10.2307/2674075.
- Blakesley RE, Mazumdar S, Dew MA, et al. (2009) Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research. *Neuropsychology*. 23(2):255-264. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3045855/pdf/nihms268580.pdf
- Fink, G., M. McConnell and S. Vollmer (2014). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness* 6(1):44-57. http://www.tandfonline.com/doi/full/10.1080/19439342.2013.875054
- Šidák ZK. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*. 62(318):626-633. doi:10.1080/01621459.1967.10482935.