Multiple Hypothesis Testing in Stata

Damian Clarke[1]

[1]University of Chile

November 10, 2021
Stata Economics Symposium

Introduction

The State of Empirical Work

Consider a number of facts based on empirical work:

- Most studies consider more than a single dependent variable of interest
- Frequentist hypothesis testing is often quite centrally used (and indeed, comes as default in many methods and their Stata implementations)
- These methods are designed to limit false rejection of a null hypothesis to some small value
- Such error rates are valid test-by-test, but accumulate if we consider multiple tests
- This can be problematic, especially if one views rejection of a single test in a class as 'confirmatory' of some general idea

A Simple Illustrative Simulation

We can try 5000 simulations of the following models in Stata varying K and examening rejection rates...

$$y_i^k = \alpha + \tau Treat_i^k + \varepsilon_i^k \qquad \forall k \in \{1, \ldots, K\}$$

Table: Error Rates, and Error Rates by Class

	Number of Dependent Variables									
	1	2	3	4	5	6	7	8	9	10
Total Tests Rejected	265	569	783	1077	1314	1552	1862	2163	2433	2638
Mean Tests Rejected	0.053	0.057	0.052	0.054	0.053	0.052	0.053	0.054	0.054	0.053
Proportion $\geq$ 1 Rejection	0.053	0.111	0.149	0.197	0.235	0.275	0.320	0.361	0.397	0.414
Proportion $\geq$ 2 Rejection	0.000	0.002	0.007	0.018	0.027	0.034	0.048	0.062	0.079	0.096

What to Do?

There are a number of ways forward to conduct valid inference in cases where multiple hypotheses are tested:

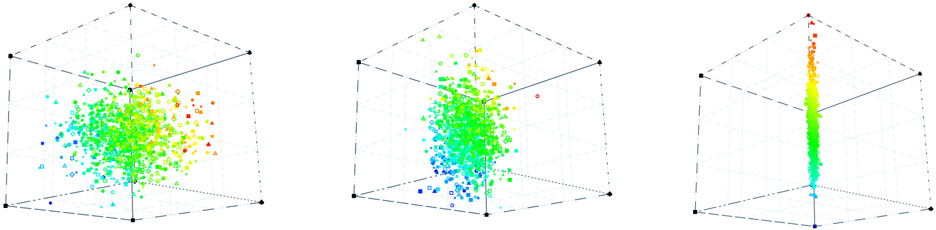
1. Dimension reduction
2. Familywise Error Rate Corrections
3. False Discovery Rate Corrections

This Talk

In this talk I plan to discuss each of these 3 methods and their Stata implementations.

- I *will not* delve deeply into the math here. Many references for this information:
  - *Eg* text-book introductions: Lehmann and Romano (2005), Casella and Berger (2001), and Westfall and Young (1993)
  - Stata Journal papers with background: R. Newson and The ALSPAC Study Team (2003), R. B. Newson (2010), and Clarke, Romano, and Wolf (2020)
- I *will* however work through examples with code
- In the interests of controlling the DGP, this will all be based on simulated data
- This will be specifically tailored to modelling in economics: heavy reliance on regression based framework (OLS, IV, RDD, etc.)
- In general, I will point to papers *and* Stata routines throughout
- Code to follow along: `https://github.com/damiancclarke/multHypStata`

Figure: Correlations (Y_1, Y_2, Y_3)



(a) $\rho = 0$

(b) $\rho = 0.5$

(c) $\rho = 0.99$

Multivariate normals simulated as per Gould (Undated). Stata visualization: Rostam-Afschar and Jessen (2014)

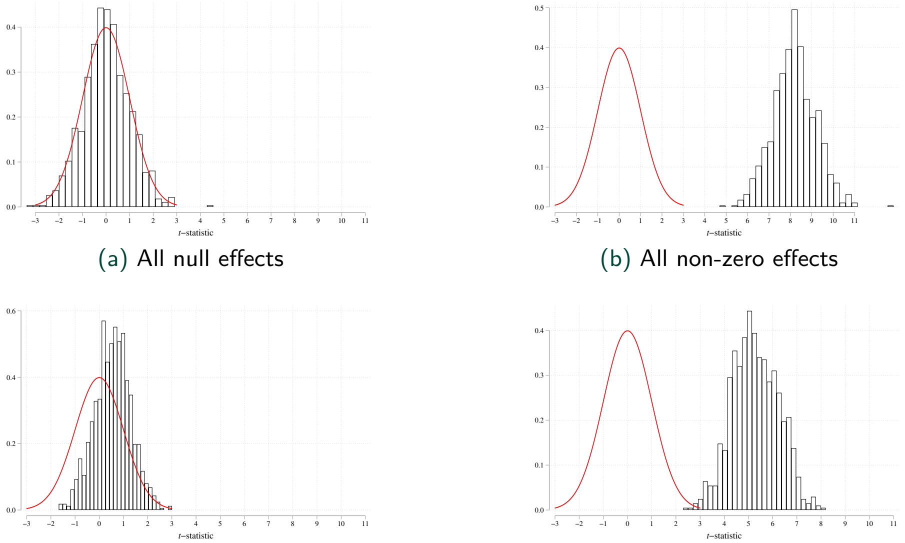
Indexes and Dimension Reduction

Indexes and Dimension Reduction

A simple option if one is concerned about multiple outcomes is to compress the information into a single dimension or index.

- How to go about generating the index and aggregating sources of information is an interesting problem
- Anderson (2008) is a tour of force here (in Stata: Schwab et al. (2020))
- Anderson (2008)'s proposal: "overweight" variables which bring more independent variation to the index
- This is very different to what a principal component analysis would draw out
- This decision is *not* innocuous
- One draw-back here: it pre-supposes some prior about relationships between independent and dependent variables

Figure: Behaviour of the Anderson Index



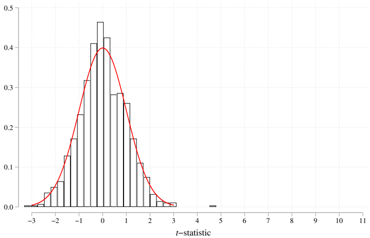
(a) All null effects

(b) All non-zero effects

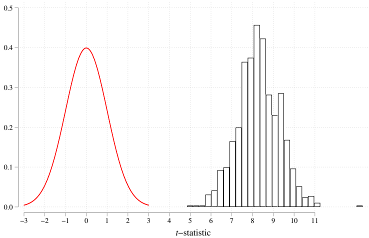
(c) 1 correlated non-zero effect

(d) 1 independent non-zero effect

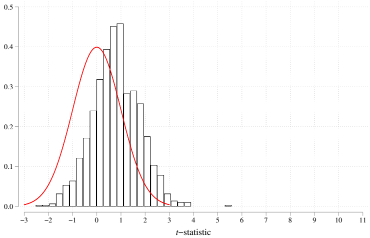
Figure: Behaviour of Principal Component



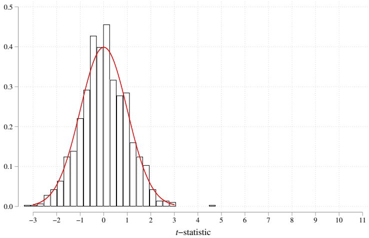
(a) All null effects



(b) All non-zero effects

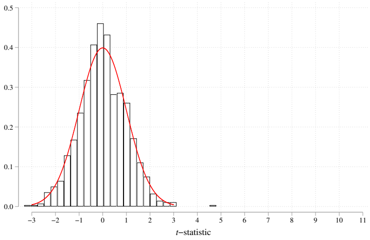


(c) 1 correlated non-zero effect

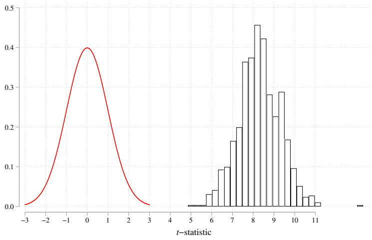


(d) 1 independent non-zero effect

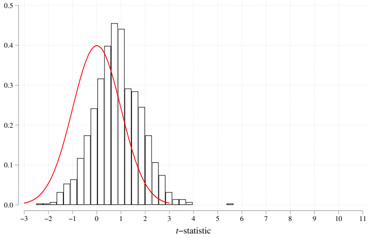
Figure: Behaviour of a Summary Index



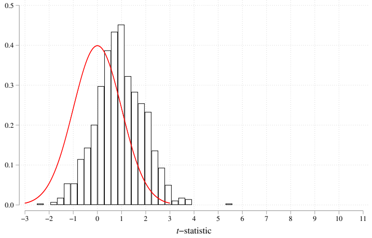
(a) All null effects



(b) All non-zero effects



(c) 1 correlated non-zero effect



(d) 1 independent non-zero effect

Familywise Error Rate Corrections

Familywise Error Rate Corrections

Alternative to aggregation is the consideration of the full family of hypotheses as separate tests: then correction for multiple testing just requires adjusting critical values

- This provides you with more information
- Familywise error rate (FWER) corrections seek to limit limit the probability of falsely rejecting *any* tests across the entire family to α
- Earliest and perhaps most well known of these is Bonferroni (1935)
- These procedures – in particular the early generation models – can be costly in terms of power
- This is a classic trade-off between size and power
- *But* much of the cost in early models comes from restrictive dependence assumptions
- Tremendous recent advances here using (a) step-down, and, especially (b) simulation-based methods

Family Wise Error Rate Corrections

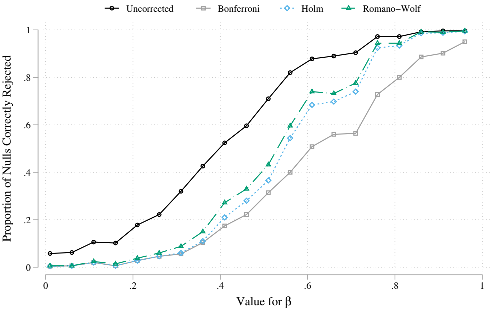
Table: Family Wise Error Rate Corrections in Stata

Correction	Stata Implementation	Note
Bonferroni (1935)	R. B. Newson 2010	1^{st} generation
Holm (1979)	R. B. Newson 2010	Step-down
Westfall and Young (1993)	Reif 2017; Jones, Molitor, and Reif 2019	Step-down, arbitrary dependence
Romano and Wolf (2005)	Clarke 2016; Clarke 2021	Step-down, arbitrary dependence
Method Specific		
List, A. Shaikh, and Xu (2019)	Seidel and Yang Xu 2016	Restricted R-W style implementation
List, A. Shaikh, and Xu (2019)	Steinmayr 2020	Restricted R-W style implementation

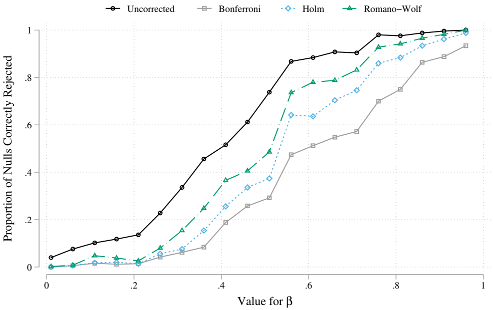
Notes: Full details of algorithms is provided in papers in left-hand column. Help files, github repos or Stata Journal papers (R. B. Newson 2010; Clarke, Romano, and Wolf 2020) provide computational backdground and Stata-specific syntax points. David McKenzie has an extremely useful blog post summary of this: https://blogs.worldbank.org/impactevaluations/updated-overview-multiple-hypothesis-testing-commands-stata

FWER Corrections and Improvements in Power

Figure: Simulated Power to Reject False Null Hypothesis



(a) $\rho = 0.25$



(b) $\rho = 0.75$

Figure: How Bootstrap Step-down Procedure Gains Power: $\rho = 0$

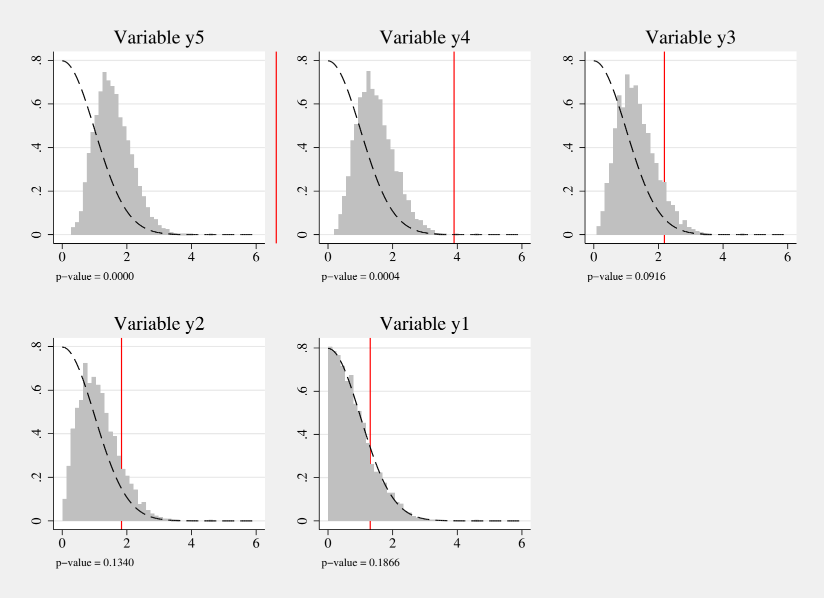


Figure: How Bootstrap Step-down Procedure Gains Power: $\rho = 0.5$

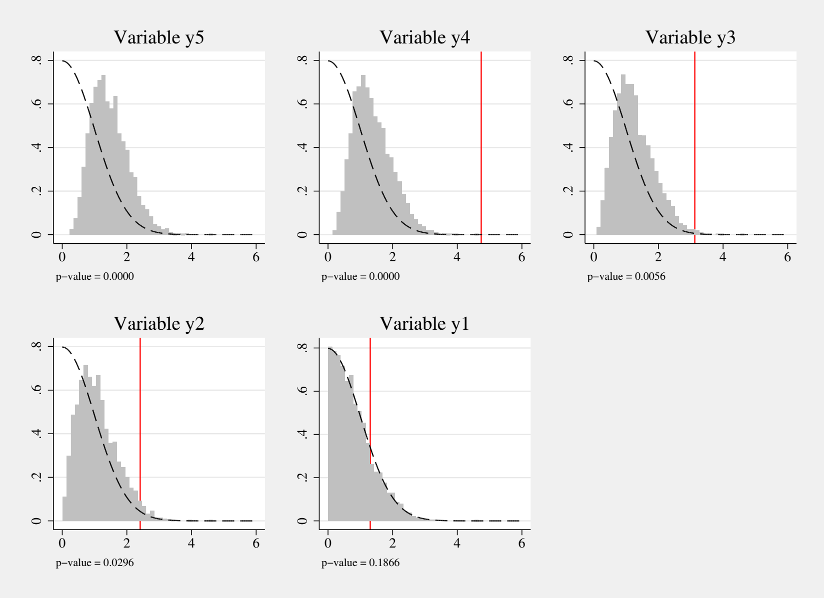
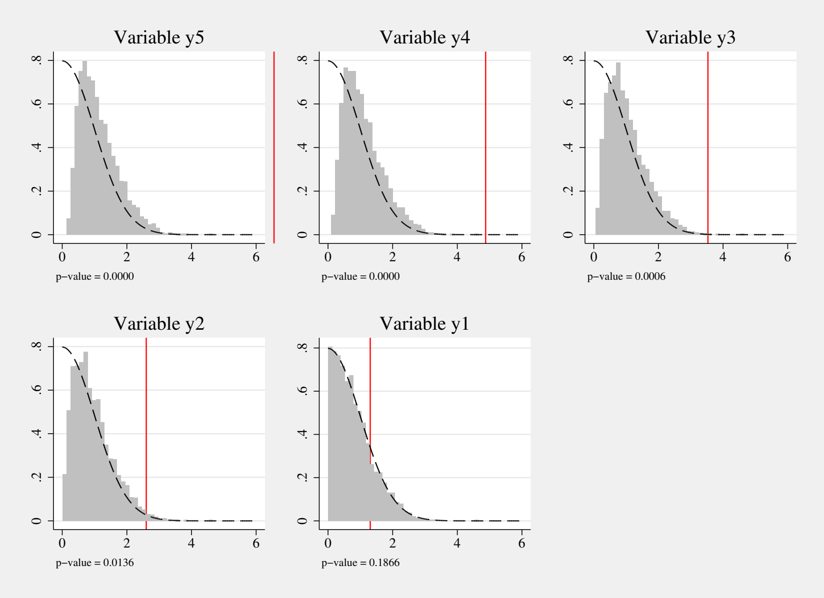


Figure: How Bootstrap Step-down Procedure Gains Power: $\rho = 0.9$



False Discovery Rate Corrections

False Discovery Rate Corrections

FWER corrections provide a clear framework for controlling error rates across multiple hypotheses, but in the limit, clearly become too demanding...

- While it is reasonable to aspire to *no* false rejection when K=10, it seems less reasonable when K=100 (or more)
- False discovery rate corrections seek to limit "false discoveries": proportion of null hypotheses rejected that are actually true
- More recent developments, in line with bio-statistic applications with many many tests
- FDR corrections imply that α% of "discoveries" will actually be false
- Typically, FDR corrections have greater power than FWER procedures, at the cost of more false rejections
- In the specific case that all null hypotheses are true, $FDR = FWER$
- Suitability of one or other also depends on subjective consderations of type I vs type II

False Discovery Rate Corrections

Table: False Discovery Rate Corrections in Stata

Correction	Stata Implementation	Note
Benjamini and Hochberg 1995	R. B. Newson 2010	Provides decision rule (accept/reject) for input values of α
Benjamini and Yekutieli 2001	R. B. Newson 2010	Provides decision rule (accept/reject) for input values of α
Benjamini and Yekutieli 2001	Anderson 2008	Directly provides a p-value

Notes: Full details of algorithms are provided in papers in left-hand column. Stata Journal papers (R. B. Newson 2010) or do file documentation provide computational backdground and Stata-specific syntax points. David McKenzie has an extremely useful blog post summary of this: `https://blogs.worldbank.org/impactevaluations/updated-overview-multiple-hypothesis-testing-commands-stata`

False Discovery Rate Corrections

Table: Performance of FDR and FWER Routines

	$\rho = 0$			$\rho = 0.33$			$\rho = 0.67$		
	Pr(A)	Pr(B)	Pr(C)	Pr(A)	Pr(B)	Pr(C)	Pr(A)	Pr(B)	Pr(C)
Naïve	0.294	0.115	0.258	0.264	0.117	0.258	0.174	0.116	0.262
FDR									
Benjamini-Hochberg	0.084	0.042	0.168	0.078	0.046	0.163	0.046	0.043	0.164
Benjamini-Yekutieli	0.018	0.013	0.105	0.022	0.016	0.109	0.020	0.022	0.104
FWER									
Bonferroni	0.028	0.020	0.131	0.024	0.023	0.132	0.024	0.028	0.128
Holm	0.030	0.020	0.135	0.032	0.026	0.135	0.028	0.030	0.132
Westfall-Young	0.032	0.022	0.130	0.032	0.028	0.132	0.040	0.041	0.153
Romano-Wolf	0.030	0.023	0.129	0.036	0.029	0.137	0.038	0.038	0.156

$Pr(A) \equiv Pr(\text{Reject at least 1 true null})$

$Pr(B) \equiv Pr(\text{Rejected null is actually true}|\text{null is rejected})$

$Pr(C) \equiv Pr(\text{Reject null}|\text{null is false})$

Discussion and Conclusion

Discussion and Conclusion

There is a growing, though incomplete, movement towards correctly adjusting for multiple hypothesis testing

- A very small selection of empirical papers implementing these sort of things are:
  - Lee and A. M. Shaikh 2014
  - Gertler et al. 2014
  - Attanasio et al. 2014
- However, this is certainly not universal
- Non-comprehensively, it seems to me like these are quite widely used in pre-specified projects (some values in Viviano, Wuthrich, and Niehaus (2021))
- But much less so in designs where power is an issue: IV, RDD
  - Presumably adaptive designs like RDD could define bandwidths optimally to account for multiple hypothesis correction
- More generally, relates to interesting work on the file drawer problem

References I

Anderson, Michael L (2008). "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". In: *Journal of the American statistical Association* 103.484, pp. 1481–1495.

Attanasio, Orazio P, Camila Fernández, Emla O A Fitzsimons, Sally M Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina (2014). "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial". In: *British Medical Journal* 349.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". In: *The Annals of Statistics* 29.4, pp. 1165–1188.

Bonferroni, C. E. (1935). "Il calcolo delle assicurazioni su gruppi di teste". In: *Studi in Onore del Professore Salvatore Ortu Carboni*. Rome, pp. 13–60.

Casella, George and Roger Berger (2001). *Statistical Inference*.

References II

Clarke, Damian (2016). *RWOLF: Stata module to calculate Romano-Wolf stepdown p-values for multiple hypothesis testing*. Statistical Software Components, Boston College Department of Economics.

– (2021). *RWOLF2: Stata module to calculate Romano-Wolf stepdown p-values for multiple hypothesis testing*. Statistical Software Components, Boston College Department of Economics.

Clarke, Damian, Joseph P. Romano, and Michael Wolf (2020). "The Romano–Wolf multiple-hypothesis correction in Stata". In: *The Stata Journal* 20.4, pp. 812–843.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor (2014). "Labor market returns to an early childhood stimulation intervention in Jamaica". In: *Science* 344.6187, pp. 998–1001.

Gould, William (Undated). *Stata 6: Simulating multivariate normal observations*.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70.

Jones, Damon, David Molitor, and Julian Reif (2019). "What do Workplace Wellness Programs do? Evidence from the Illinois Workplace Wellness Study". In: *The Quarterly Journal of Economics* 134.4, pp. 1747–1791.

Lee, Soohyung and Azeem M. Shaikh (2014). "MULTIPLE TESTING AND HETEROGENEOUS TREATMENT EFFECTS: RE-EVALUATING THE EFFECT OF PROGRESA ON SCHOOL ENROLLMENT". In: *Journal of Applied Econometrics* 29.4, pp. 612–626.

Lehmann, E. L. and Joseph P. Romano (2005). *Testing statistical hypotheses*. Third. Springer Texts in Statistics. New York: Springer.

List, J.A., A.M. Shaikh, and Y Xu (2019). "Multiple hypothesis testing in experimental economics". In: *Experimental Economics* 22.1, pp. 773–793.

Newson, R. and The ALSPAC Study Team (2003). "Multiple-test procedures and smile plots". In: *Stata Journal* 3.2, 109–132(24).

Newson, Roger B. (2010). "Frequentist q-values for multiple-test procedures". In: *Stata Journal* 10.4, pp. 568–584.

References IV

Reif, Julian (2017). *WYOUNG: Stata module to perform multiple testing corrections*. Statistical Software Components, Boston College Department of Economics.

Romano, Joseph P. and Michael Wolf (2005). "Stepwise Multiple Testing as Formalized Data Snooping". In: *Econometrica* 73.4, pp. 1237–1282.

Rostam-Afschar, Davud and Robin Jessen (2014). *GRAPH3D: Stata module to draw colored, scalable, rotatable 3D plots*. Statistical Software Components, Boston College Department of Economics.

Schwab, Benjamin, Sarah Janzen, Nicholas P. Magnan, and William M. Thompson (2020). "Constructing a summary index using the standardized inverse-covariance weighted average of indicators". In: *The Stata Journal* 20.4, pp. 952–964.

Seidel, Joseph and Yang Xu (2016). *MHTEXP: Stata module to perform multiple hypothesis testing correction procedure*. Statistical Software Components, Boston College Department of Economics.

Steinmayr, Andreas (2020). *MHTREG: Stata module for multiple hypothesis testing controlling for FWER*. Statistical Software Components, Boston College Department of Economics.

References V

Viviano, Davide, Kaspar Wuthrich, and Paul Niehaus (2021). *(When) should you adjust inferences for multiple hypothesis testing?* arXiv preprint.

Westfall, Peter H. and S. S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley.