Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons

David Thissen, Lynne Steinberg and Daniel Kuang

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2002 27: 77
DOI: 10.3102/10769986027001077

The online version of this article can be found at:
http://jeb.sagepub.com/content/27/1/77

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
Spring 2002, Vol. 27, No. 1, pp. 77–83

Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons

David Thissen
University of North Carolina

Lynne Steinberg
Daniel Kuang
Portland State University

Williams, Jones, and Tukey (1999) showed that a sequential approach to controlling the false discovery rate in multiple comparisons, due to Benjamini and Hochberg (1995), yields much greater power than the widely used Bonferroni technique that limits the familywise Type I error rate. The Benjamini-Hochberg (B-H) procedure has since been adopted for use in reporting results from the National Assessment of Educational Progress (NAEP), as well as in other research applications. This short note illustrates that the B-H procedure is extremely simple to implement using widely available spreadsheet software. Given its easy implementation, it is feasible to include the B-H procedure in introductory instruction in inferential statistics, augmenting or replacing the Bonferroni technique.

Keywords: *Excel, false discovery, multiple comparisons rate*

Williams, Jones, and Tukey (1999) showed that a sequential approach to controlling the false discovery rate in multiple comparisons, due to Benjamini and Hochberg (1995), yields much greater power than the widely used Bonferroni technique that limits the familywise Type I error rate. The Benjamini-Hochberg (B-H) procedure has since been adopted for use in reporting results from the National Assessment of Educational Progress (NAEP) (Braswell, Lutkus, Grigg, Santapau, Tay-Lim, and Johnson, 2001), as well as in other research applications. For example, Steinberg (2001) used the B-H approach with the likelihood-ratio procedure for the evaluation of differential item functioning (*dif*) (Thissen, Steinberg, & Wainer 1988; 1993) to interpret confident direction of *dif* in sets of items, testing *dif* for each item.

We thank Lyle V. Jones for his clear and helpful explanations of the Benjamini-Hochberg procedure and its virtues, as well as for his comments on this article. We are grateful to Valerie S. L. Williams for recovering unpublished computational details for us. Thanks also to Brian Clauser and the editor, Howard Wainer, for additional suggestions clarifying the presentation. Of course, any errors that remain are our own.

Two differences between Williams, Jones, and Tukey's (1999) description of the B-H approach to multiple comparisons and the widely known Bonferroni technique are salient here:

(a) As explained in detail by Williams, Jones, and Tukey (1999, p. 43), α is "a bound on twice the probability of being erroneously confident about the direction of the population comparison." So $\alpha/2$ plays a prominent role, and (b) the B-H approach controls the false discovery rate (FDR): "the average fraction of erroneous assertions among all confident directions asserted" (Williams, Jones, & Tukey, 1999 p. 44). This is different from the familywise error rate controlled by the Bonferroni technique, but it is a more powerful generalization to the context of multiplicity of the conventional test of significance.

Here, we follow Williams, Jones, and Tukey (1999, p. 44) in specifying $\alpha/2$ as an approximate bound, for a given family of comparisons, "on the expected value of the ratio of (a) the number of erroneous declarations of confident differences to (b) the maximum quantity of either the total number of declarations of confidence or 1."

The B-H approach accomplishes control of the FDR by sequentially comparing the observed *p* value for each of a family of multiple test statistics, in order from largest to smallest, to a list of computed B-H critical values [$pB\text{-}H(i)$]. The critical value on the list is computed for each test statistic, indexed by *i*, by linear interpolation between $\alpha/2$ (for the largest observed *p* value) to $(\alpha/2)/m$, where *m* is the family size, for the smallest of the *p* values. The last value is the Bonferroni critical value, so the reason for the gain in power of B-H relative to Bonferroni is clear: In the B-H approach, only the smallest of the *m* observed *p* values is compared to the Bonferroni critical value; all of the other *p* values are compared to less stringent criteria.

The reason for the algorithmic difference between the B-H approach and the Bonferroni procedure is that the two methods accomplish different goals. The goal of the Bonferroni procedure is to control the familywise error rate at α (often conventionally 0.05). That is, in samples from a population in which the classical null hypothesis is precisely true, in less than $100\alpha\%$ of *families* are *any* comparisons declared significant. By contrast, in the presentation by Williams, Jones, and Tukey (1999) of the B-H approach, the classical null hypothesis that the difference is exactly zero is not considered, because the probability of that event is, itself, zero. The population difference can only be in the same direction as the sample result, or in the opposite direction. The statistical decision to be made is whether the data are sufficient to provide confidence in the direction of the difference. The B-H approach controls the FDR such that it remains less than $\alpha/2$. For a family of comparisons, some number may involve declarations of confident direction; for those comparisons, it is asserted that the probability that the population difference is in the opposite direction from the sample result is less than $\alpha/2$. The relative frequency associated with that probability is the proportion of incorrect assertions among all assertions of confident direction.

Williams, Jones, and Tukey (1999) noted that SAS 6.12 provides computations for the B-H technique in PROC MULTTEST; however, the use of SAS may not be convenient in the context of elementary instruction in statistics, when multiple-comparisons are typically first introduced. This short note illustrates the ease with which the B-H procedure can be implemented using widely available spreadsheet software. The illustrations here use Microsoft® Excel; we understand that Excel may not the best solution for many complex statistical computations, but this is not a complex statistical computation. Implementation with more widely available, easier-to-use software such as Excel should make it feasible to include the B-H technique as early as the first course in statistics; and it should make the procedure more attractive to researchers.

Illustrations

Displays 1 and 2 are augmented reproductions of the parts of Williams, Jones, and Tukey's (1999) Tables 1 and 3 that involve the B-H approach. Step-by-step instructions for the use of Excel to compute the values required for the B-H approach are included in Displays 1 and 2.

DISPLAY 1
Reconstruction of the Benjamini-Hochberg results of Table 1 from Williams, Jones, and Tukey (1999) using Excel.

Step 1: Enter the *t* values and *df*

Step 2a: Enter in cell D2 the formula to compute the *p* value for the *t* [=TDIST(ABS(B2), C2, 1)], and fill the column

Step 2b: Sort columns A–D using column D as the index in descending order

Step 3: Enter in column E index numbers from 1 to the number of test statistics

Step 4: Enter in cell F2 the formula to compute the B-H critical value based on the index [=((6-E2+1)*0.05)/(2*6)]; fill column F

Step 5: Enter in cell G2 a formula to mark with an asterisk rows in which the *p* value is less than the B-H critical value [=IF(D2<F2, "*", " ")], and fill column G

A	B	C	D	E	F	G
Comparison	*t*	*df*	*p* value	Index	B-H critical	value
NE vs West	1.09	30	0.1422	1	0.0250	
Central vs NE	2.44	30	0.0104	2	0.0208	*
SE vs West	–2.51	30	0.0088	3	0.0167	*
Central vs West	3.53	30	0.0007	4	0.0125	*
NE vs SE	3.60	30	0.0006	5	0.0083	*
Central vs SE	6.04	30	0.0000	6	0.0042	*

Note. * Indicate comparisons for which the direction of the difference is confidently interpreted at the $\alpha/2$ level.

DISPLAY 2

Reconstruction of the Benjamini-Hochberg results of Table 3 from Williams, Jones, and Tukey (1999) using Excel.

Step 1: Enter differences and standard errors

Step 2: Compute the *t*s, [=B2/C2]; fill column]

Step 3: Enter *df* in column E

Step 4: Enter in cell F2 the formula to compute the *p* value for the first *t* [=TDIST(ABS(D2),E2,1)], and fill the column; sort columns A-F in descending order of F

Step 5: Enter in column G index numbers from 1 to the number of test statistics

Step 6: Enter in cell H2 the formula to compute the B-H critical value based on the index [=((34-G2+1)*0.05)/(2*34)]; fill column H downward

Step 7: Enter in cell I2 a formula to mark with an asterisk rows in which the *p* value is less than the B-H critical value [=IF(F2<H2, "*", " ")], and fill column I

A	B	C	D	E	F	G	H	I
State	Avg92-Avg90	*SE*	*t*	*df*	*p* value	Index	B-H critical value	
GA	–0.323	1.776	–0.182	60	0.4281	1	0.0250	
AR	–0.777	1.485	–0.523	60	0.3014	2	0.0243	
AL	–1.568	2.017	–0.777	60	0.2200	3	0.0235	
NJ	1.565	1.927	0.812	60	0.2100	4	0.0228	
NE	1.334	1.528	0.873	60	0.1930	5	0.0221	
ND	1.526	1.686	0.905	60	0.1844	6	0.0213	
DE	1.374	1.347	1.020	60	0.1558	7	0.0206	
MI	2.215	1.847	1.199	60	0.1176	8	0.0199	
LA	2.637	2.079	1.268	60	0.1048	9	0.0191	
IN	2.149	1.636	1.314	60	0.0969	10	0.0184	
WI	2.801	1.963	1.427	60	0.0794	11	0.0176	
VA	2.859	1.930	1.481	60	0.0719	12	0.0169	
WV	2.331	1.396	1.669	60	0.0501	13	0.0162	
MD	3.399	1.923	1.767	60	0.0411	14	0.0154	
CA	3.777	2.115	1.786	60	0.0396	15	0.0147	
AH	3.466	1.850	1.873	60	0.0330	16	0.0140	
NY	4.893	2.532	1.933	60	0.0290	17	0.0132	
PA	4.303	2.205	1.951	60	0.0279	18	0.0125	
FL	3.784	1.933	1.958	60	0.0274	19	0.0118	
WY	2.226	1.096	2.030	60	0.0234	20	0.0110	
NM	2.334	1.148	2.033	60	0.0233	21	0.0103	
CT	3.204	1.534	2.088	60	0.0205	22	0.0096	
AK	4.181	1.755	2.383	60	0.0102	23	0.0088	
KY	4.327	1.618	2.674	60	0.0048	24	0.0081	*

(*continued*)

DISPLAY 2 (*continued*)
Reconstruction of the Benjamini-Hochberg results of Table 3 from Williams, Jones, and Tukey (1999) using Excel.

A	B	C	D	E	F	G	H	I
State	Avg92-Avg90	*SE*	*t*	*df*	*p* value	Index	B-H critical value	
AZ	4.994	1.851	2.698	60	0.0045	25	0.0074	*
ID	2.956	1.068	2.768	60	0.0037	26	0.0066	*
TX	5.645	1.888	2.990	60	0.0020	27	0.0059	*
CO	4.326	1.389	3.115	60	0.0014	28	0.0051	*
IA	4.811	1.488	3.233	60	0.0010	29	0.0044	*
NH	4.422	1.354	3.266	60	0.0009	30	0.0037	*
NC	7.265	1.587	4.578	60	0.0000	31	0.0029	*
HI	5.550	1.171	4.738	60	0.0000	32	0.0022	*
MN	6.421	1.352	4.748	60	0.0000	33	0.0015	*
RI	5.097	0.948	5.374	60	0.0000	34	0.0007	*

Note. * Reflects the direction of change in the population average between 1990 and 1992 for 11 of the 34 states

Display 1 involves six comparisons among four regional averages for mathematics proficiency from the 1990 NAEP Trial State Assessment (TSA). The six *t* values in column B of Display 1 are computed by dividing the six pairwise differences by estimates of the standard errors of those differences. The *t* values form the starting point for the computation. Those, along with the degrees of freedom (*df*)[1] in column C are used with Excel's built-in function[2] to compute the *p* values in column D[3]. The *p* values are used to sort columns A, B, C, and D in decreasing order of magnitude of the *p* values in column D. The B-H critical values, linearly interpolated between $\alpha/2 = 0.025$ and the Bonferroni critical value, $(\alpha/2)/6 = 0.0042$, are computed in column F, and a decorative column of asterisks is added in column G to indicate the rows in which the observed *p* value is less than the critical value.

The five starred lines in Display 1 indicate comparisons for which the direction of the difference is confidently interpreted at the $\alpha/2$ level. That is to say, we are confident at the $\alpha/2$ level that the sample values for the regional differences reflect the direction of the population differences for all comparisons except the first on the list, NE vs West, for which the *t* statistic is only 1.09. In contrast, application of the Bonferroni procedure would mark as significant only the largest three differences (those in the lower three lines of Display 1, with observed *p* values less than the Bonferroni critical value of 0.0042).

Display 2 reproduces the larger example provided by Williams, Jones, and Tukey (1999), and shown in their Table 3. Tabulated are the mean eighth-grade mathematics proficiency changes between 1990 and 1992 for 34 states. In Display 2, we begin with the average differences and the standard errors of those differences in columns B and C, and then use Excel to compute the *t* statistics in column D. Given the value of *df* in column E (60)[4], we let Excel compute the observed

p values, and sort columns A–F in descending order of the *p* values in column F. The observed *p* values are compared to the B-H critical values in column H.

The results in Display 2 agree exactly with those presented by Williams, Jones, and Tukey (1999). The B-H approach permits us to have confidence that the sample difference reflects the direction of change in the population average between 1990 and 1992 for 11 of the 34 states (the states marked with an asterisk in column I in Display 2). By contrast, only four states would be considered significant applying the Bonferroni critical value (0.0007) to all comparisons.

For a more detailed discussion of the results in Displays 1 and 2, and the relative performance of the B-H approach and the Bonferroni procedure, the reader is referred to the original Williams, Jones, and Tukey (1999) source. Our point here is that the B-H procedure is easy to implement in contemporary spreadsheet software—nearly as easy as the single division involved in the Bonferroni technique.

Conclusion

Given its easy implementation, it is feasible to include the B-H procedure in introductory instruction in inferential statistics, augmenting or replacing the Bonferroni technique. Students trained with this more powerful technique should be less likely to use the nearly powerless Bonferroni procedure, or to eschew correction for multiple comparisons entirely, due to a perceived loss of power. Certainly no one can complain that the B-H procedure is too difficult to compute!

Notes

[1]The *df* in Display 1, 30, is derived from the relatively complex Jackknife computations of the standard errors from NAEP TSA (L.V. Jones & V. S. L. Williams, personal communication, September 5, 2001).

[2]Excel includes functions that compute the *p* values for most commonly used statistical distributions.

[3]The *p* value for the first entry in column C of Display 1 differs in the fourth decimal place from that tabulated by Williams, Jones, and Tukey (1999) in their Table 1 due to the use here of the *t* values to two decimal places, instead of the larger number of decimal places they used in computation.

[4]Again, the value for *df* is derived with the Jackknife estimates of the standard errors.

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*, 289–300.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B., & Johnson, M. (2001). *The nation's report card: Mathematics 2000*. NCES 2001-517. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81*, 332–342.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 147–169). Hillsdale, NJ: Erlbaum,.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Williams, V. S. L., Jones, L. V., & Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42–69.

Authors

DAVID THISSEN is Professor, Department of Psychology, University of North Carolina, Chapel Hill, CB#3270 Davie Hall, Chapel Hill, NC 27599-3270; dthissen@email.unc.edu. He specializes in quantitative psychology and psychological measurement.

LYNNE STEINBERG is Associate Professor, Department of Psychology, Portland State University, PO Box 751, Portland OR 97207-0751; steinbergl@pdx.edu. She specializes in quantitative psychology, and social and personality measurement.

DANIEL KUANG is a doctoral student in the Systems Science and Applied Psychology Program at Portland State University, PO Box 751, Portland OR 97207-0751. He specializes in quantitative psychology and psychological measurement.

Manuscript Received December 2001
Revision Received December 2001
Accepted December 2001