Bonferroni, Holm, and Hochberg Corrections: Fun Names, Serious Changes to *P* Values

Matthew J. McLaughlin, MD, Kristin L. Sainani, PhD

"If you torture the data long enough, it will confess," said noted economist Ronald Coase, when referring to the problem of multiple testing. In short, if analysts run enough statistical tests, then they will find "statistically significant" results just by chance. By using a *P* value cutoff of .05 for statistical significance, approximately 1 in 20 tests (5%) will be deemed significant when no effects exist. One way to guard against such chance findings is to lower the threshold for statistical significance (or, equivalently, adjust *P* values higher). Although these *P* value adjustments can be complicated to understand and compute, clinical researchers should be aware of 3 simple corrections that can easily be calculated by hand: Bonferroni, Holm, and Hochberg.

WHAT IS A BONFERRONI CORRECTION?

A simple way to avoid chance findings, which also are called type I errors or false positives, is to lower the significance threshold from .05 to .05/k, where k is the number of statistical tests run. Called the Bonferroni correction, this is the simplest and most widely recognized adjustment for multiple testing. For example, if we run 18 tests, the new threshold becomes .05/18 = .00278, such that only approximately 1 in 360 tests (.0278%) will be deemed significant when no effects exist. The Bonferroni correction guarantees that the probability of making any type I errors, across all tests run, is held at 5% or less. This can be proven mathematically. If no effects exist and we run k independent statistical tests (which represents the "worst-case" scenario), then the following is true:

$$\text{Probability of any type 1 errors} = 1 - (1 - [.05/k])^{k} = .05$$

For example, for 18 tests:

$$\text{Probability of any type 1 errors} = 1 - .99972^{18} = .05$$

To apply the Bonferroni correction, researchers can either divide their significance threshold by k or, equivalently, multiply all their *P* values by k. The results from a hypothetical trial that compares stroke patients who received a novel strengthening intervention versus those who received a control intervention are presented in Table 1. The researchers compared 18 different Functional Independence Measures items between the groups, which generated 18 *P* values. Although 9 Functional Independence Measures items had *P* values less than .05, only 3 are significant at the Bonferroni correction cutoff of .00278.

Advantages and Disadvantages of Bonferroni Corrections

The beauty of the Bonferroni correction is its simplicity. It remains one of the easier methods to calculate, which lends itself to lasting use. The drawback is that the Bonferroni correction is overly conservative. For most real-world cases, researchers run statistical tests that are highly overlapping, not independent. For example, if we compare participants who were normal weight with participants who were overweight, overweight to obese, and

M.J.M. Division of Pediatric Rehabilitation Medicine, The Children's Mercy Hospital, Kansas City, MO
Disclosure: nothing to disclose

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, Stanford, CA. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

PM&R
1934-1482/14/$36.00
Printed in U.S.A.

Vol. 6, 544-546, June 2014
http://dx.doi.org/10.1016/j.pmrj.2014.04.006

Table 1. *Results from a hypothetical trial that compared 18 Functional Independence Measures (FIM) between 2 groups by using 3 different corrections for multiple comparisons*

FIM Category	Unadjusted *P* Value	Bonferroni Correction, Cutoff = .00278*	Holm or Hochberg Threshold	Holm Correction*	Hochberg Correction*
Bed-to-chair transfer	.00001	Significant	.05/18 = .00278	Significant	Significant
Toilet transfer	.00002	Significant	.05/17 = .00294	Significant	Significant
Tub and/or shower transfer	.002	Significant	.05/16 = .00313	Significant	Significant
Locomotion	.003		.05/15 = .00333	Significant	Significant
Lower extremity dressing	.0035		.05/14 = .00357	Significant	Significant
Stairs	.0039		.05/13 = .00385		Significant
Upper extremity dressing	.004		.05/12 = .00417		Significant
Bowel management	.01		.05/11 = .00455		
Bladder management	.04		.05/10 = .005		
Social interaction	.13		.05/9 = .00556		
Memory	.23		.05/8 = .00625		
Eating	.24		.05/7 = .00714		
Grooming	.27		.05/6 = .00833		
Comprehension	.54		.05/5 = .01		
Problem solving	.66		.05/4 = .0125		
Expression	.77		.05/3 = .01667		
Toileting	.78		.05/2 = .025		
Bathing	.81		.05/1 = .05		

*The blank areas indicate values that were not statistically significant.

normal weight to obese, this involves 3 statistical tests. But, clearly, if the first 2 comparisons achieve statistical significance, then it is more likely that the third will as well. The Bonferroni correction charges the same penalty for all 3 tests, which essentially penalizes us for the same comparison twice and thus needlessly increases the chances of missing real effects (also known as type II errors or false negatives).

HOLM AND HOCHBERG CORRECTIONS: TWO NEWER PLAYERS IN THE STATISTICAL GAME

The Bonferroni method remains the most popular *P* value correction in health studies [1]; in contrast, many researchers are unaware of the Holm and Hochberg procedures [2]. This is unfortunate because the Holm and Hochberg procedures are nearly as easy to understand and implement but are more powerful than the Bonferroni correction. Holm and Hochberg are sometimes called "step-down" and "step-up" Bonferroni procedures, respectively.

The Holm step-down method [3] works by arranging the *P* values from smallest to largest and then by comparing these *P* values to sequentially less conservative significance cutoffs:

1. If the smallest *P* value is greater than or equal to .05/k (where k is the total number of tests run), then the procedure stops and no *P* values are significant. If the smallest *P* value is less than .05/k, then this *P* value is significant and the procedure continues.
2. If the second smallest *P* value is greater than or equal to .05/(k − 1), then the procedure stops and no further *P* values are significant. If the second smallest *P* value is less than .05/(k − 1), then this *P* value is deemed significant and the procedure continues.
3. If the third smallest *P* value is greater than or equal to .05/(k − 2), then the procedure stops and no further *P* values are significant. If the third smallest *P* value is less than .05/(k − 2), then this *P* value is deemed significant and the procedure continues.
4. Keep going as needed.

By this method, the Holm method never rejects fewer hypotheses than the Bonferroni method, but the overall probability of making any type I errors is still kept at or less than .05. For example, for the hypothetical data in Table 1: .00001 < .05/18; .00002 < .05/17; .002 < .05/16; .003 < .05/15; .0035 < .05/14; but .0039 > .05/13. Thus, 5 differences between the intervention and the control groups are deemed significant. This is 2 more than with the Bonferroni correction.

The Hochberg step-up method [4] proceeds in the opposite direction, from the largest to the smallest *P* value:

1. If the largest *P* value is less than .05, then all the *P* values are significant and the procedure stops. Otherwise, the procedure continues.
2. If the second largest *P* value is less than .05/2, then this and all remaining *P* values are significant, and the procedure stops. Otherwise, the procedure continues.
3. If the third largest *P* value is greater than or equal to .05/3, then this and all remaining *P* values are significant. Otherwise, the procedure continues.
4. Keep going as needed.

For example, for the hypothetical data in Table 1: .81 > .05; .78 > .05/2; .77 > .05/3; .66 > .05/4; .54 > .05/5; .27 > .05/6; .24 > .05/7; .23 > .05/8; .13 > .05/9; .04 > .05/10; .01 > .05/11; but .004 < .05/12. So, the smallest 7 *P* values are deemed significant.

In practice, the Hochberg and the Holm procedures usually give similar results. However, the Hochberg procedure can be more inclusive than the Holm procedure, as illustrated here. Although the Hochberg procedure is more powerful than the Holm procedure, it only guarantees that the overall probability of making any type I errors is maintained at .05 or lower if certain assumptions are met. For this reason, some experts recommend using the Holm method unless one is certain of the ramifications of the Hochberg method [5].

P VALUE ADJUSTMENTS AND STATISTICAL POWER

By lowering the threshold for statistical significance, one increases the chance of a type II error (of missing effects) and thus decreases statistical power. Therefore, researchers who are planning to use formal *P* value corrections in their final analyses should factor these into their initial power calculations to avoid being underpowered for important endpoints. Researchers also should consult with a statistician during study planning to determine whether more mathematically sophisticated *P* value adjustments (beyond Bonferroni, Holm, and Hochberg corrections) may be more powerful for their study design. In practice, Bonferroni, Holm, and Hochberg corrections are often cited informally after completing exploratory analyses to help readers gauge whether "significant" *P* values (<.05) likely reflect real effects or just chance findings. In this case, researchers and readers will need to consider multiple pieces of information, including potential effects on statistical power, when making judgments.

CONCLUSION

Applying poststudy corrections to *P* values helps to sort through the murkiness of studies with multiple comparisons. Although the Bonferroni correction is the most widely known and applied, Holm and Hochberg corrections offer more statistical power with similar ease of calculation. The advantage of Bonferroni, Holm, and Hochberg corrections is that they can be implemented by hand. For more computationally complex *P* value corrections, researchers may need to work with a statistician.

REFERENCES

1. Stacey AW, Pouly S, Czyz CN. An analysis of the use of multiple comparison corrections in ophthalmology research. Invest Ophthalmol Vis Sci 2012;53:1830-1834.
2. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: The Bonferroni vs. Holm methods. Am J Public Health 1996;86:726-728.
3. Holm S. A simple sequentially rejective multiple test procedure. Scand Stat J 1979;6:65-70.
4. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometricka 1988;75:800-802.
5. Levin B. Annotation: On the Holm, Simes, and Hochberg multiple testing procedure. Am J Public Health 1996;86:628-629.