Physiotherapy Practice and Research 38 (2017) 67–68

The problem of too many hypothesis tests

Gissane, C.

School of Sport, Health and Applied Science, St Mary's University, Twickenham, Middlesex, TW1 4SX, UK.

conor.gissane@stmarys.ac.uk

Phone: +44 (0)20 8240 4228

Fax: +44 (0)20 8240 4255

Research articles frequently report on several significance tests. When multiple hypothesis tests report on a single issue, the P values may not be an accurate guide to significance of a given result.[1] Whenever an investigator conducts a statistical significance test, they could make either a Type I or a Type II error (see box). The risk of making such errors is part of the hypothesis testing process, but it is generally agreed that making a Type I error is more serious than making a Type II error.[2] Normal practice dictates that the chance of making a Type I error is set before beginning the research. The chance of making a Type I error is set as $\alpha = 0.05$, corresponding to the P value where the null hypothesis will either be accepted or rejected. However, the chance of making a Type I error of 0.05 is for one test. But, if the number of tests increases, so does the chance of making a Type I error.

Type I error*: An error made by wrongly rejecting a true null hypothesis (a false-positive error)*.[2]

Type II error: *An error made when falsely accepting a false null hypothesis*.[2]

Multiple tests are common in research.[3] For example, researchers wishing to examine treatment effects on several dependent variables. Similarly, studies sometimes report sub-group analyses after examining the main effects of a study. Both practices increase the number of hypothesis tests and the chance of making a Type I error. If the aim is to reduce or maintain the chance of making a Type I error

at 0.05, researchers need to employ techniques whereby they can adjust the P should they need to conduct multiple tests.[4]

Some methodologists argue that making corrections is necessary,[5, 6] while others regard the adjustment as unnecessary because research allows the comparisons across separate experiments.[7] Multiple tests within a given study are unlikely to be independent, and without adjusting the P values, the chance of declaring a significant relationship between an independent and a dependent variable is greater than the 0.05 level.[1] Also, pure chance dictates that when a P value is set to 0.05, the probability of getting a significant result is one in twenty (0.05), even if a significant result does not exist.[6]

How to go about it

The most often used correction is the Bonferroni correction. It is simple to apply, but is sometimes considered too conservative. It lowers the significance threshold from .05 to .05/k, where k is the number of statistical tests run.[8] Kim et al.,[9] reported on functional instability of the ankle joint. To do so, they reported the results of six significance tests (table 1). Without a controlling for multiple tests, there were four significant difference reported. When using a Bonferroni correction the p values is adjusted by dividing the significance value by the number of tests conducted (0.05/6 = 0.0083). After the correction, the number of significant tests is reduced to three. Truthfully, the correction only influenced the third P value of 0.047, it was never going to alter the original non-significant results. Please note that in spite of the low P values, each of the tests is significant at $P<0.05$.[3]

In spite of its simplicity, the Bonferroni correction is criticised for being too conservative.[6, 10, 11] Other options are available, and some are only a little harder to calculate.[10, 11] The Holm[11] and Hochberg[10] procedures are also more powerful than Bonferonni, and this is attributed to the fact that both are sequential.[12, 13]

The two methods are similar in operation, with Holm being described as a 'step down' technique and Hochberg a 'step up' technique.[8] The Holm calculations are shown in table 2, with the P values arranged from smallest to largest. If the smallest value is greater than 0.05/k (0.05/6 = 0.0083) stop, nothing is significant. If it is less than 0.05/k, it is significant. The process continues with the second smallest p value being compared with 0.05/(k-1) (0.05/5 = 0.001). The procedure continues until a non significant value is found. [8]

The Hochberg procedure works in the opposite direction with the P value arranged from largest to smallest (table 3).[8] If the smallest value is lower than 0.05, all tests are significant, and the process can stop. Otherwise it continues and the second value is compared against 0.05/2 (0.025), then it and all subsequent P values are significant. If it not, the process continues with the third P values compared against 0.05/3 (0.0167). If it significant, so are each of the remaining P values. Using the Kim *et al*.,[9] data, the fourth test was significant against a critical value of 0.0125 (0.05/4). The remaining significance tests are also significant.

All of the methods described will keep the probability of making a Type I error at $P<0.05$. For each of the three corrections, Kim's[9] data shows three significant test results. This is not always the case, Holm[11] and Hochberg[10] usually produce similar results,[8] and usually more significant results than the Bonferroni.

Authors and readers are encouraged to apply corrections for multiple hypothesis testing. Inflated P values are a problem,[14] but controlling for them can present a clearer picture of study effects when multiple tests are presented.[8] The methods described in this paper are simple with all calculations performed on a spreadsheet. Present your findings as clearly as possible, and examine thoroughly the results of others.

References

[1] Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American Journal of Public Health. 1996;86(5):726-8.
[2] Vogt WP. Dictionary of statistics and methodology. Sage Publications, Ltd.; 1993.
[3] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. 1995;310(6973):170.
[4] Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. Controlled Clinical Trials. 2000;21(6):527-39.
[5] Knudson D. Significant and meaningful effects in sports biomechanics research. Sports Biomechanics. 2009;8(1):96-104.
[6] Sainani KL. The problem of multiple testing. PM&R. 2009;1(12):1098-103.
[7] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43-6.
[8] McLaughlin MJ, Sainani KL. Bonferroni, Holm, and Hochberg corrections: fun names, serious changes to p values. PM&R. 2014;6(6):544-6.
[9] Kim Y, Kim E, Song Y, Han D, Richards J. The effects of functional instability of the ankle joint on balance. Physiotherapy Practice and Research. 2015;37(1):3-9.
[10] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800-2.
[11] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979:65-70.
[12] Nakagawa S. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology. 2004;15(6):1044-5.
[13] Rice WR. Analyzing tables of statistical tests. Evolution. 1989;43(1):223-5.
[14] Stacey AW, Pouly S, Czyz CN. An Analysis of the Use of Multiple Comparison Corrections in Ophthalmology ResearchAn Analysis of the Use of Multiple Comparison CorrectionsStacey et al. Investigative ophthalmology & visual science. 2012;53(4):1830-4.

Table 1. Significance results from Kim *et al*.,[9] before and after a Bonferroni correction.

P value	No correction	Bonferroni correction
0.874	Not significant	Not significant
0.074	Not significant	Not significant
0.047	Significant	Not significant
0.007	Significant	Significant
0.001	Significant	Significant
0.001	Significant	Significant

Table 2. Significance results from Kim *et al*.,[9] using the Holm correction.

P value	Holm correction	Critical P value	
0.001	.05/6	0.0083	Significant
0.001	.05/5	0.0100	Significant
0.007	.05/4	0.0125	Significant
0.047	.05/3	0.0167	Not significant
0.074	.05/2	0.0250	Not significant
0.874	.05/1	0.0500	Not significant

Table 3. Significance results from Kim *et al*.,[9] using the Hochberg correction.

P value	Hochberg correction	Critical P value	
0.874	.05/1	0.0500	Not significant
0.074	.05/2	0.0250	Not significant
0.047	.05/3	0.0167	Not significant
0.007	.05/4	0.0125	Significant
0.001	.05/5	0.0100	Significant
0.001	.05/6	0.0083	Significant