ORIGINAL CONTRIBUTION

Approaches to the Problem of Type I Error in Multiple Comparisons.

Christopher T. Naugler, M.Sc.

Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, B3H 4H7

The acceptable probability of a given statistical test showing a "false positive" result is termed the type I error. This is generally chosen to be 5%, meaning that one in 20 such tests would be expected to be significant by chance. When authors present a mass of significance tests the question arises as to which are "truly significant" and which represent those "one in twenty" due to chance alone. This paper discusses several approaches to this problem. These include using combined outcome measures, choosing simultaneous inference statistical tests, or applying the Bonferonni correction to a table of p-values.

When numerical analyses are presented in research papers, they are generally said to be "statistically significant" or "not statistically significant". This determination is based on the probability of the obtained result occurring by chance, compared with the percentage of times that a "false positive" result is acceptable. This latter value is also known as the type I error rate. Thus, if the probability of the result occurring by chance is less than the type I error rate, the result is said to be statistically significant. By convention, the type I error rate is generally set at 5%, which is why probabilities of less than 0.05 are considered significant. With this level of type I error, the author is accepting that there is a one in 20 chance that a claimed significant result will be due to chance alone. Problems arise, however, when multiple statistical tests are conducted on the same data (1-5). If, for example, 10 independent tests are conducted on the same data with a type I error of 0.05, the chance of getting at least one significant result is 1-0.9510 or 0.40. Clearly, this is more than the stated type I error rate of 0.05 and could lead to spurious results being reported as statistically significant.

In this paper, I discuss various approaches which I hope will provide both readers and authors with the background necessary to deal intelligently with this problem. However, this is a basic introduction only. Individuals requiring details on the application of specific methods are directed to the references cited in this paper.

COMBINED OUTCOME MEASURES

Perhaps the simplest solution to this problem is to reduce the number of statistical tests performed on the same data by choosing the outcomes judiciously or combining the outcome measures (6-8). For example, is it really necessary for a study to assess clinical response at one week and at two weeks? If the same information could be gained by measuring clinical response at two weeks alone, then the number of statistical tests, and thus the chance of spurious results, could be reduced. Alternatively, outcome measures can be combined to form single variables, which can be either simple or weighted sums of the various outcomes. For example, if a study considered either increased hemoglobin oxygen saturation or decreased respiratory distress as important outcomes, then a single variable that counted either of these as a "positive" response obviate the need to perform multiple tests.

Address correspondence to:

Christopher Naugler, Box 362, Sir Charles Tupper Medical Building,
Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7

MULTIVARIATE TESTS

Researchers with the available resources and expertise can employ statistical tests that examine a number of variables in the same test (simultaneous inference), thus eliminating the potentially inflated type I errors that would occur if the variables were analyzed in separate tests. Some examples of these tests are multiple regressions, discriminant function analyses, principle components analyses and multivariate analyses of variance (MANOVA).

PRELIMINARY OVERALL TESTS

Preliminary overall tests, along with the Bonferonni correction, are known as post hoc methods because they attempt to control false positive results after the experiment has been completed. This approach would be suitable to data in which pair-wise comparisons must be made among a large number of groups. Suppose, for example, that a researcher wanted to see if there was a difference in IQ among the four different classes in a medical school. Instead of testing all of the possible pairwise combinations (e.g. first year versus second year, first year versus third year, etc.) for a total of six different tests, the researcher could first employ a one-way ANOVA (analysis of variance) test to examine differences in all groups simultaneously. Then, if the overall test is not statistically significant, no further tests are done on the data. However, if the overall test shows a significant difference, the researcher can assume that a significant result exists in the data and pair-wise tests can then be conducted using one of the specific post hoc methods such as Fisher's least significant difference, Dunnett's t and Sheffe's post hoc test. For further information on this method refer to O'Brien and Shampo (9)or any standard statistics text.

THE BONFERRONI CORRECTION

This method relies on adjusting the p-values of multiple tests to "correct for" the number of statistical tests performed (10-12). This approach is becoming increasingly common in medical research papers. Consider a study which presents the results of seven statistical tests on the same data. These could be seven different variables or the same test on seven subsets of the same data. The sequential Bonferroni correction increases the p-values to reflect the number of statistical tests performed. The net effect is to reduce the number of significant tests and hopefully to "weed out" the spurious findings. In its simplest form, the correction is performed by simply multiplying all p-values by the number of tests performed. However, a modification of the standard Bonferroni correction called the sequential Bonferonni technique (13) has increased statistical power and is therefore the preferred method. A sample calculation is given in Table 1. The author using the Bonferonni correction must decide how inclusive the correction should be. That is, should the correction be applied to all p-values within a table, all p-values within an experiment, or all p-values within an entire study. The appropriate level is very much open to interpretation but table-wide corrections appear to be the most common at present.

Table 1: *Applying the sequential Bonferroni correction to a table containing seven p-values. This method uses corrections to the actual p-values while keeping the significance level (type I error) at 0.05. The same result could be achieved by adjusting the acceptable type I error for each test (1). In this example, the number of significant results was reduced from five to two.*

Step 1.	Step 2.	Step 3.
Rank p-values from smallest to largest.	Starting at the top, multiple each p-value by the total number of tests minus the number of tests above it. When the first non-significant result is obtained, all subsequent tests are also non-significant.	Report new p-values from as being corrected for table-wide significance using the sequentia Bonferronni technique.
0.001	x 7 = 0.007 (significant)	p=0.007
0.005	x 6 = 0.030 (significant)	p=0.025
0.020	x 5 = 0.10 (1st non-significant result)	NS
0.041	x 4 = non-significant	NS
0.045	x 3 = non-significant	NS
0.063	x 2 =non-significant	NS
0.082	x 1 = non-significant	NS

NON-INDEPENDENT TESTS

The problem of multiple comparisons is most sinister when the variables or outcomes are statistically independent. In many studies, however, variables are highly correlated and the problem of multiple comparisons becomes less important. Consider the hypothetical example of a study examining the side effects of a certain medication. Separate outcome variables for nausea, vomiting and gastric upset were used and all were found to be marginally statistically significant. Obviously, these three symptoms are likely to occur together and likewise would tend to become significant together. Thus, these variables are not statistically independent, and the chance of spurious results is decreased by an unknown amount depending on the magnitude of dependence among the variables (12). In this example, it could be argued that the multiple comparisons do not increase the risk of spurious significant results.

On the other hand, if this study had examined variables that were probably essentially independent, such as gastrointestinal symptoms, headaches and rashes, then it would be entirely appropriate to apply a Bonferonni correction to the results.

CONCLUSION

The problem of spurious significant results when multiple statistical tests are performed is one that all researchers must consider. Approaches vary from prevention (simultaneous inference, combined outcome measures) to post hoc solutions (preliminary overall tests, Bonferroni corrections). Furthermore, depending how closely correlated the variables are, no correction at all may be necessary. This paper provides an introduction to the options available to researchers to combat this problem.

ACKNOWLEDGEMENT

I wish to thank Dr. Weerasinghe, Director, Biostatistical Consulting Service, Dalhousie University, for helpful comments on an earlier version of the manuscript.

REFERENCES

1. Rice WR. Analyzing tables of statistical tests. *Evolution* 1989;43:223-225.
2. O'Brien PC and Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. *Mayo Clin Proc* 1988;63:813-815.
3. O'Brien PC and Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 4. Performing multiple statistical tests on the same data. *Mayo Clin Proc* 1988;63:1043-1045.
4. Mills JL. Data torturing. *NEJM* 1993;329:1196-1199.
5. Ottenbacher KJ. Statistical conclusion validity. Multiple inferences in rehabilitation research. *Am J Phys Med Rehabil* 1991;70:317-322.
6. Lefkopoulou M and Ryan L. Global tests for multiple binary outcomes. *Biometrics* 1993;49:975
7. Lehmacher W, Wassmer G and Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 1991;47:511-521.
8. O'Brien PC and Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 5. Comparing two therapies with respect to several endpoints. *Mayo Clin Proc* 1988;63:1140-1143.
9. O'Brien PC and Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 2. Comparisons among several therapies. *Mayo Clin Proc* 1988;63:816-820.
10. Tarone RE. A modified Bonferroni method for discrete data. *Biometrics* 1990;46:515-522.
11. Holy S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6:65-70.
12. Bland JM and Altman DG. Multiple significance tests: the Bonferroni Method. *BMJ* 1995;310:170.
13. Holm S. A simple sequentially rejective multiple test procedure. *Scan J Stat* 1979;6:65-70.