FORUM
FORUM
FORUM

FORUM is intended for new ideas or new ways of interpreting existing information. It provides a chance for suggesting hypotheses and for challenging current thinking on ecological issues. A lighter prose, designed to attract readers, will be permitted. Formal research reports, albeit short, will not be accepted, and all contributions should be concise with a relatively short list of references. A summary is not required.

Arguments for rejecting the sequential Bonferroni in ecological studies

Matthew D. Moran, Dept of Biology, Hendrix College, 1600 Washington Ave., Conway, AR 72032, USA (moran@mercury.hendrix.edu).

Interpretation of results that include multiple statistical tests has been an issue of great concern for some time in the ecological literature. The basic problem is that when multiple tests are undertaken, each at the same significance level (α), the probability of achieving at least one significant result is greater than that significance level (Zaykin et al. 2002). Therefore, there is an increased probability of rejecting a null hypothesis when it would be inappropriate to do so. The typical solution to this problem has been lowering the α values for the table (i.e. establishing a table-wide significance level) and therefore reducing the probability of a spurious result. Specifically, the most common procedure has been the application of the sequential Bonferroni adjustment (Holm 1979, Miller 1981, Rice 1989). Arguments in this essay address the problems of adjusting probability values for tables of multiple statistical tests, and more specifically argue for rejection of the sequential Bonferroni as a solution to this problem.

Since the influential publication of Rice (1989), the sequential Bonferroni correction has become the primary method of addressing the problem of multiple statistical tests in ecological research. The sequential Bonferroni adjusts the table-wide p-value to keep it constant at 0.05, and subsequently reduces the probability of a spurious result. Although other methods exist for addressing tables of multiple statistical tests, the sequential Bonferroni has become the most commonly utilized process. However, this method has several flaws ranging from mathematical to logical to practical that argue for rejecting this method in ecological studies.

Mathematical objections

The sequential Bonferroni adjusts the table-wide Type I error by dividing α (traditionally set at 0.05) by the number of statistical tests (N). The lowest p-value in the table of tests is then rejected (assuming $p < 0.05/N$). One then divides 0.05 by $N - 1$ and examines the table for a remaining p-value less than this new quantity. This process continues until no p-value in the table is less than the calculated adjusted p-value. The major mathematical problem with this method is that it only examines individual p-values of each test, while ignoring the number of statistical tests that are significant. For example, suppose a researcher has 10 individual tests in a table with five of them significant at $p = 0.049$. Using the sequential Bonferroni correction, the maximum p-value to reject the first null hypothesis is 0.005. None fall below that level, so the researcher is forced to fail to reject all null hypotheses. However, the probability of having five significant tests at a p-value of 0.049, and all of these results being due to random chance is very low.

The probability of finding one or more significant results by chance in a table of statistical tests is calculated as:

$$1 - (1 - \alpha)^N$$

which in a table of ten tests would equal 0.40 (Wilkinson 1951). Therefore, a researcher will find at least one test significant due to chance about 40% of the time, which illustrates the argument in favor of the Bonferroni adjustment. The probability of exactly one test being statistically significant due to chance alone can be calculated using a Bernoulli process, where "success" is considered a test that produces a p-value lower than selected a value. This probability is calculated with the equation:

$$p = [N!/(N - K)!K!] \times \alpha^K (1 - \alpha)^{N - K}$$

where $N =$ numbers of tests and $K =$ number of tests below α. With 10 hypothetical tests, the probability of finding exactly one test below $p < 0.05$ is 0.315. How-

ever, in the example described above, the probability of finding five significant tests (at $p < 0.05$) by chance alone is 0.00006 using the Bernoulli equation, which is an extremely small probability. This example illustrates a very important problem of statistical data involving multiple statistical tests: the more individual tests that fall below α, the lower the probability that they are all spurious. It also illustrates the principle that several relatively high p-values can be a stronger indication of significance than one relatively low p-value (Rosenthal 1978). The Bernoulli equation does require independence between tests, which may or may not be a characteristic of ecological experiments.

It is quite common to have multiple significant statistical tests (at $p < 0.05$), but none with very small p-values. This is particularly true among ecological studies, which often have small numbers of replicates, high variability, and subsequently low statistical power. There are powerful alternatives to performing multiple statistical tests. For instance, one could perform a MANOVA on tables of related response variables to detect overall treatment effects. However, MANOVA designs have decreased power as the number of tests increases, and may eventually become ineffective (von Ende 1993). Therefore, researchers are often left with the option of analyzing results individually. As tables of statistical tests get larger, the sequential Bonferonni makes it more unlikely to find significance by greatly inflating the Type II error, without taking in account the number of statistical tests below the assigned α value. Therefore, although less conservative than the single-step Bonferroni correction, it is still an overly conservative statistical method (Westfall and Young 1993). Other more powerful methods have also been developed (e.g. Sidak inequality, Westfall and Young 1993), although these still cannot address multiple significant results.

Logical objection

The logical concern is that it is not possible to develop a standard way to apply multiple testing procedures to data sets. Should one apply it to a particular table, the entire paper, all the papers in a particular journal issue, or to a lifetime of research? Clearly no one would apply this correction to a lifetime's worth of work because it is an absurd notion, and because the resulting p-value would be so small that one could not reject the null hypothesis for anything (especially if one had a productive career). Although the subjectivity of the sequential Bonferroni has been noted (Cabin and Mitchell 2000), no author has made concrete suggestions regarding when to apply the correction. Because no consensus has developed among ecologists, it continues to be used haphazardly (Cabin and Mitchell 2000). While absolute rules about when to use a particular statistical analysis are not necessarily preferred, general guidelines need to be established. Because the sequential Bonferroni has been utilized for many years without the development of these guidelines, it is unlikely that this issue will be resolved.

Practical objections

In many ways, practical objections are the most important arguments against utilizing the sequential Bonferroni. Ecologists are encouraged to perform detailed analyses of their study systems. For instance, it is common to lump similar species together into trophic levels or guilds for analysis when it may be more appropriate to analyze the response of individual species. Certainly the latter is preferred when possible to acquire the most information from a study. However, as one performs a more detailed analysis (i.e. more statistical tests), the probability of finding a significant result declines if the sequential Bonferroni is applied. In essence, a researcher is punished for performing more work. The irony of the sequential Bonferroni correction (and multiple tests in general) is that as one performs more detailed work, the probability of finding anything significant declines dramatically. It therefore produces a paradox (one could call it a hyper-Red Queen phenomenon): the more research one does, the lower the probability that a significant result is discovered. I am not suggesting that researchers should look at all possible angles in the hopes of finding something significant ("data snooping," Westfall and Young 1993), but detailed studies of ecological systems should be encouraged.

In addition to inhibiting detailed analyses, the sequential Bonferroni makes it more difficult to find significance in diverse communities. These communities are likely to have more individual statistical tests, while in simple communities with only a few species, significance is more easily attained. A methodology that discourages ecologists to study complex systems or inhibits one from performing more detailed study of ecosystems is never desirable. Imagine an experiment that contains 100 response species, and a researcher wishes to study the effect of some treatment on those species. The researcher would have to achieve a p-value < 0.0005 to reject the null hypothesis for any single species. Very few would invest time in an experiment with those odds.

P-values, although important, are not more important than effect sizes and do not replace quality interpretation of results (Yoccoz 1991). I would therefore like to suggest some reasonable guidelines for reporting results for multiple statistical tests. Researchers should report all results including exact p-values. Then, instead

of applying the sequential Bonferroni (or other multiple testing adjustments), researchers should use the accepted $p < 0.05$ cutoff and make reasonable interpretations based on experimental design, power analyses, differences between control and treatment groups, and basic logic. These interpretations along with high quality reviewers and editors will assure the desirable amount of statistical rigor. This will also assure that potentially important findings will not fall into the nonsignificant results category and end up unpublished. This could be true especially for studies that present novel results which could further knowledge within the field. Although overly conservative multiple test adjustments would tend to suppress these data, one should be encouraged to publish these types of results (Lander and Kruglyak 1995).

For instance, imagine an experiment testing the response of a plant community to grazing. In this hypothetical yet realistic example, five forbs and five grasses are analyzed for differences in biomass between grazed and ungrazed plots (Table 1). The five individual forbs have significantly lower biomass (at $p < 0.05$) in grazed plots while the grasses appear unaffected. The logical conclusion from this example is that the grazing causes a decline in forb biomass but has no effect on grasses. The sequential Bonferroni correction would not allow a researcher to reject any of the null hypotheses, although it is very unlikely that the response pattern of grasses and forbs is due to chance alone. Logical interpretation is clearly in conflict with the sequential Bonferroni conclusion.

Hill (1966, reviewed by Westfall and Young 1993) provides a coherent set of criteria for determining whether significant results are due to real biological effects. These criteria were developed for the medical sciences but are equally applicable to the ecological sciences. Most important for the ecological sciences include 1) whether the results can be reproduced, 2) if the significant results were planned comparisons, and 3) whether the results can satisfy rules of logic and reason. These guidelines are much more effective means of interpretation than overly conservative statistical methods.

Table 1. Results of a hypothetical experiment testing the effects of grazing on ten plant species. Response refers to the difference between grazed (G) and ungrazed (U) plots. * indicates significance at $p < 0.05$.

Plant Species	Response	Significance
Grass # 1	G < U	0.45
Grass # 2	G > U	0.67
Grass # 3	G < U	0.93
Grass # 4	G < U	0.25
Grass # 5	G > U	0.53
Forb # 1	G < U	0.04*
Forb # 2	G < U	0.02*
Forb # 3	G < U	0.03*
Forb # 4	G < U	0.01*
Forb # 5	G < U	0.02*

Conclusions

How to adjust probability values for multiple statistical tests is an issue that will affect ecological research indefinitely. Although the sequential Bonferroni has become the standard method of dealing with this problem, the objections described here make a strong case for rejecting the use of this process in tables of statistical tests. Instead, ecologists should inject some logic into their interpretations of statistical results. For instance, one significant test with a relatively large p-value (e.g. 0.025) in a large table would certainly be suspect. However, many significant results in a table indicate something important is occurring. I will reiterate the very important point that several relatively high p-values are stronger evidence against a null hypothesis than one moderately low value (Rosenthal 1978). In essence, as ecologists collect more data, the probability of finding some spurious results is quite high, but the chance of all the results being spurious is extremely improbable. These spurious results should not be of great concern, as they will not be confirmed in future experiments. The sequential Bonferroni, however, makes it likely that researchers will not publish important results that could open up new avenues of knowledge. Most importantly, as we have come to realize the complexity of ecosystems, we should encourage detailed and complex investigations and not be inhibited by unreasonably low p-values.

Acknowledgements – Thanks to D. H. Wise, C. Spatz, and C. Mouron for reviewing earlier drafts of this manuscript. Two anonymous reviewers greatly improved the final version. Special thanks to D. Lyon and W. E. Snyder for our lengthy discussions of this issue at the 2001 ESA meeting.

References

Cabin, R. J. and Mitchell, R. J. 2000. To Bonferroni or not to Bonferroni: when and how are the questions. – ESA Bulletin 81: 246–248.

Hill, A. B. 1966. Principles of medical statistics. – Oxford Univ. Press.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. – Scand. J. Stat. 6: 65–70.

Lander, E. and Kruglyak, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. – Nat. Genet. 11: 241–247.

Miller, R. G. 1981. Simultaneous statistical inference. – McGraw Hill.

Rice, W. R. 1989. Analyzing tables of statistical tests. – Evolution 43: 223–225.

Rosenthal, R. 1978. Combining results of independent studies. – Psychol. Bull. 85: 185–193.

von Ende, C. N. 1993. Repeated-measures analysis: growth and other time-dependent measures. – In: Scheiner, S. M. and Gurevitch, J. (eds), Design and analysis of ecological experiments. Chapman and Hall, pp. 113–137.

Westfall, P. H. and Young, S. S. 1993. Resampling-based Multiple Testing. – John Wiley & Sons.

Wilkinson, B. 1951. A statistical consideration in psychological research. – Psychol. Bull. 48: 156–158.

Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. – ESA Bulletin 72: 106–111.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. 2002. Truncated product method for combining p-values. – Genet. Epidemiol. 22: 170–185.