Behavioral Ecology Vol. 15 No. 6: 1044–1045
doi:10.1093/beheco/arh107
Advance Access publication on June 30, 2004

Forum

A farewell to Bonferroni: the problems of low statistical power and publication bias

Shinichi Nakagawa
Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Recently, Jennions and Møller (2003) carried out a meta-analysis on statistical power in the field of behavioral ecology and animal behavior, reviewing 10 leading journals including *Behavioral Ecology*. Their results showed dismayingly low average statistical power (note that a meta-analytic review of statistical power is different from post hoc power analysis as criticized in Hoenig and Heisey, 2001). The statistical power of a null hypothesis (H_o) significance test is the probability that the test will reject H_o when a research hypothesis (H_a) is true. Knowledge of effect size is particularly important for statistical power analysis (for statistical power analysis, see Cohen, 1988; Nakagawa and Foster, in press). There are many kinds of effect size measures available (e.g., Pearson's *r*, Cohen's *d*, Hedges's *g*), but most of these fall into one of two major types, namely the *r* family and the *d* family (Rosenthal, 1994). The *r* family shows the strength of relationship between two variables while the *d* family shows the size of difference between two variables. As a benchmark for research planning and evaluation, Cohen (1988) proposed 'conventional' values for small, medium, and large effects: $r = .10$, .30, and .50 and $d = .20$, .50, and .80, respectively (in the way that *p* values of .05, .01, and .001 are conventional points, although these conventional values of effect size have been criticized; e.g., Rosenthal et al., 2000).

The meta-analysis on statistical power by Jennions and Møller (2003) revealed that, in the field of behavioral ecology and animal behavior, statistical power of less than 20% to detect a small effect and power of less than 50% to detect a medium effect existed. This means, for example, that the average behavioral scientist performing a statistical test has a greater probability of making a Type II error (or β) (i.e., not rejecting H_o when H_o is false; note that statistical power is equals to $1 - \beta$) than if they had flipped a coin, when an experiment effect is of medium size (i.e., $r = .30$, $d = .50$).

Here, I highlight and discuss an implication of this low statistical power on one of the most widely used statistical procedures, Bonferroni correction (Cabin and Mitchell, 2000). Bonferroni corrections are employed to reduce Type I errors (i.e., rejecting H_o when H_o is true) when multiple tests or comparisons are conducted. Two kinds of Bonferroni procedures are commonly used. One is the standard Bonferroni procedure, where a modified significant criterion (α/k where k is the number of statistical tests conducted on given data) is used. The other is the sequential Bonferroni procedure, which was introduced by Holm (1979) and popularized in the field of ecology and evolution by Rice (1989) (see these papers for the procedure). For example, in a recent volume of *Behavioral Ecology* (vol. 13, 2002), nearly one-fifth of papers (23 out of 117) included Bonferroni corrections. Twelve articles employed the standard procedure while 11 articles employed the sequential procedure (10 citing Rice, 1989, and one citing Holm, 1979). A serious problem associated with the standard Bonferroni procedure is a substantial reduction in the statistical power of rejecting an incorrect H_o in each test (e.g., Holm, 1979; Perneger, 1998; Rice, 1989). The sequential Bonferroni procedure also incurs reduction in power, but to a lesser extent (which is the reason that the sequential procedure is used in preference by some researchers; Moran, 2003). Thus, both procedures exacerbate the existing problem of low power, identified by Jennions and Møller (2003).

For example, suppose an experiment where both an experimental group and a control group consist of 30 subjects. After an experimental period, we measure five different variables and conduct a series of *t* tests on each variable. Even prior to applying Bonferroni corrections, the statistical power of each test to detect a medium effect is 61% ($\alpha = .05$), which is less than a recommended acceptable 80% level (Cohen, 1988). In the field of behavioral ecology and animal behavior, it is usually difficult to use large sample sizes (in many cases, $n < 30$) because of practical and ethical reasons (see Still, 1992). When standard Bonferroni corrections are applied, the statistical power of each *t* test drops to as low as 33% (to detect a medium effect at $\alpha/5 = .01$). Although sequential Bonferroni corrections do not reduce the power of the tests to the same extent, on average (33–61% per *t* test), the probability of making a Type II error for some of the tests ($\beta = 1 -$ power, so 39–66%) remains unacceptably high. Furthermore, statistical power would be even lower if we measured more than five variables or if we were interested in detecting a small effect.

Bonferroni procedures appear to raise another set of problems. There is no formal consensus for when Bonferroni procedures should be used, even among statisticians (Perneger, 1998). It seems, in some cases, that Bonferroni corrections are applied only when their results remain significant. Some researchers may think that their results are 'more significant' if the results pass the rigor of Bonferroni corrections, although this is logically incorrect (Cohen, 1990, 1994; Yoccoz, 1991). Many researchers are already reluctant to report nonsignificant results (Jennions and Møller, 2002a,b). The wide use of Bonferroni procedures may be aggravating the tendency of researchers not to present nonsignificant results, because presentation of more tests with nonsignificant results may make previously 'significant' results 'nonsignificant' under Bonferroni procedures. The more detailed research (i.e., research measuring more variables) researchers do, the less probability they have of finding significant results. Moran (2003) recently named this paradox as a hyper-Red Queen phenomenon (see the paper for more discussion on problems with the sequential method).

Imagine that we conduct a study where we measure as many relevant variables as possible, 10 variables, for example. We find only two variables statistically significant. Then, what should we do? We could decide to write a paper highlighting these two variables (and not reporting the other eight at all) as if we had hypotheses about the two significant variables in the first place. Subsequently, our paper would be published. Alternatively, we could write a paper including all 10 variables. When the paper is reviewed, referees might tell us that there were no significant results if we had 'appropriately' employed Bonferroni corrections, so that our study would not be advisable for publication. However, the latter paper is scientifically more important than the former paper. For example, if one wants to conduct a meta-analysis to investigate an overall effect in a specific area of study, the latter paper is five times more informative than the former paper. In the long term, statistical significance of particular tests may be of trivial importance (if not always), although, in the short term, it makes papers publishable. Bonferroni procedures may, in part, be preventing the accumulation of knowledge in the field of behavioral ecology and animal behavior, thus hindering the progress of the field as science.

However, it is true that researchers often seem to measure apparently irrelevant variables in their study, unnecessarily expanding a probability of making Type I errors. Therefore, it is understandable that reviewers sometimes have to demand Bonferroni corrections for such studies.

As explained above, the use of Bonferroni procedures further reduce power, increasing a Type II error to unacceptable levels, and they may also contribute to publication bias, hindering the advance of the field. Therefore, the use of Bonferroni corrections and the practice of reviewers demanding Bonferroni procedures should be discouraged (and also, researchers should play their part in carefully selecting relevant variables in their study). These problems probably stem from overemphasis on statistical significance (i.e., p values) in journals rather than more emphasis on practical or biological significance (i.e., effect size) (Cohen, 1990). I recommend routine presentation of observed (standardized) effect size, such as Pearson's r and Cohen's d, in journal articles (r and d are readily convertible; Rosenthal, 1994), as also recommended by Stoehr (1999) and Jennions and Møller (2003) (see these papers for various benefits of reporting effect size; see also Kirk, 1996).

Standardized effect sizes are degrees of experimental effects, which are comparable across studies even with different sample sizes (note that p values do not tell us what degree of experimental effect is present; Cohen, 1990, 1994; Yoccoz, 1991). Thus, reporting observed effect size, along with exact p values, allows readers as well as researchers to evaluate the biological importance (and statistical significance) of results. Although presenting standardized effect size is still not common in ecology and evolution, some journals in other disciplines oblige or strongly recommend researchers to present effect size (e.g., Murphy, 1997; Wilkinson and The Task Force on Statistical Inference, 1999). Furthermore, reporting not only effect sizes but also confidence intervals (CIs) for effect sizes may prove more useful (see Thompson, 2002, and references therein for software that calculates CIs for effect sizes), because the CIs for effect sizes can be used to interpret nonsignificant results (see Hoenig and Heisey, 2001; Colegrave and Ruxton, 2003; Jennions and Møller, 2003; Nakagawa and Foster, in press).

Additionally, if research involves a large number of variables, controlling 'false discovery rate' (FDR; the proportion of rejecting true H_os; cf., Bernoulli equation in Moran, 2003) may be an option rather than controlling the probability of obtaining even one false rejection of H_o (i.e., using Bonferroni corrections), as García (2003) recently highlighted. Since the idea of FDR was first introduced, this powerful method has been improved and rapidly established in many fields that require large series of multiple tests (e.g., analysis of microarray gene expression; for more detail on FDR, see García, 2003, and the references therein). Controlling FDR provides a much better compromise between Type I and Type II errors when multiple testing is necessary (although it is rarely known in the field of ecology and evolution).

To conclude, it is time to make our farewell to Bonferroni procedures and to start reporting effect size and/or confidence intervals for effect size (or other alternatives) in the field of behavioral ecology and animal behavior.

My special thanks go to Wolfgang Forstmeier, discussions with whom have made this article possible and complete. I thank Lisa Pope, Jens Rolff, Ben Hatchwell, and Paul Johnson for their help and comments. I am also grateful to two anonymous reviewers, Nigel G. Yoccoz, and Gunilla Rosenqvist, whose comments have greatly improved the manuscript. The author is supported by the Foundation for Research Science and Technology, New Zealand.

Address correspondence to S. Nakagawa. E-mail: s.nakagawa@sheffield.ac.uk.

Received September 8, 2003; revised February 17, 2004; accepted March 16, 2004.

REFERENCES

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

Cohen J, 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, New Jersey: Erlbaum.

Cohen J, 1990. Things I have learned (so far). Am Psychol 45:1304–1312.

Cohen J, 1994. The earth is round ($p < .05$). Am Psychol 49:997–1003.

Colegrave N, Ruxton GD, 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. Behav Ecol 14:446–450.

García LV, 2003. Controlling the false discovery rate in ecological research. Trends Ecol Evol 18:553–554.

Hoenig JM, Heisey DM, 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55:19–24.

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

Jennions MD, Møller AP, 2002a. Publication bias in ecology and evolution: an empirical assessment using the "trim and fill" method. Biol Rev 77:211–222.

Jennions MD, Møller AP, 2002b. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc R Soc Lond Ser B 269:43–48.

Jennions MD, Møller AP, 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behav Ecol 14:438–445.

Kirk RE, 1996. Practical significance: a concept whose time has come. Educ Psychol Meas 56:746–759.

Moran MD, 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos 100:403–405.

Murphy KR, 1997. Editorial. J Appl Psychol 82:3–5.

Nakagawa S, Foster TM, in press. The case against retrospective power analyses with an introduction to power analysis. Acta Ethol.

Perneger TV, 1998. What's wrong with Bonferroni adjustments. Brit Med J 316:1236–1238.

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

Rosenthal R, 1994. Parametric measures of effect size. In: The handbook of research synthesis (Cooper H, Hedges LV, eds). New York: Sage; 231–244.

Rosenthal R, Rosnow R, Rubin DB, 2000. Contrasts and effect sizes in behavioral research: a correlational approach. Cambridge: Cambridge University Press.

Still AW, 1992. On the number of subjects used in animal behaviour experiments. Anim Behav 30:873–880.

Stoehr AM, 1999. Are significance thresholds appropriate for the study of animal behaviour? Anim Behav 57:F22–F25.

Thompson B, 2002. What future quantitative social science research could look like: confidence intervals for effect sizes. Educ Res 31:25–32.

Wilkinson L, The Task Force on Statistical Inference, 1999. Statistical methods in psychology journals. Am Psychol 54:594–604.

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