How Much Power Is Enough? Against the Development of an Arbitrary Convention for Statistical Power Calculations

Author(s): Julian di Stephano

Source: *Functional Ecology*, Vol. 17, No. 5 (Oct., 2003), pp. 707-709
Published by: British Ecological Society
Stable URL: http://www.jstor.org/stable/3599167
Accessed: 01-05-2018 08:08 UTC

Functional Ecology 2003
17, 707–709

Forum

How much power is enough? Against the development of an arbitrary convention for statistical power calculations

Due largely to the work of Jacob Cohen (1962, 1988), statistical power analysis has become a widely known technique in many fields of experimental science. Promotion of power analysis in ecology (Toft & Shea 1983; Peterman 1990a,b; Fairweather 1991) has led to its increasing use over the last few decades. This trend is encouraging as power analysis is a useful tool in the planning phase of ecological experiments and can also be used after data analysis to improve the interpretation of non-significant results (Quinn & Keough 2002).

Statistical power = $1 - \beta$ where β is the Type II error rate. Power can generally be described by the equation

$$\text{Power} \propto (\text{ES} \times \alpha \times \sqrt{n})/\sigma \quad (1)$$

where ES is the effect size, α is the Type I error rate, n is the sample size and σ is the population standard deviation. A Type I error is the probability of erroneously rejecting the null hypothesis, while a Type II error is the probability of erroneously failing to reject the null hypothesis. Equation 1 (its specific form depends on the statistical model being used) is usually solved for power, sample size or effect size depending on the objectives of the analysis (Fairweather 1991), and can be implemented either before or after data are collected. Whichever way power analysis is used, researchers must often make a decision about what constitutes an acceptable level of power. This is almost always difficult as there is no simple way to decide how much power is enough.

Unfortunately, the definition of adequate statistical power in the ecological literature often appears arbitrary with minimal attention to the context within which individual studies are conducted. Authors frequently state that statistical power is adequate if its value is 0·80 or above (Steidl, Hayes & Schauber 1997; Lougheed, Breault & Lank 1999; Manolis, Andersen & Cuthbert 2000; Strehlow *et al.* 2002), and a convention is fast developing where 'significance and power levels are set ... at 0·05 and 0·80 ... respectively' (Walsh *et al.* 1999). This practice is hereafter referred to as the five-eighty convention.

When the five-eighty convention is used, the probabilities of making Type I and Type II errors are 5% and 20%, respectively. Implicitly, this means that the cost of making a Type I error is considered four times more important than the cost of making a Type II error (Cohen 1988). While in some situations this may be the case, it certainly cannot be assumed. For example, in the fields of impact assessment and conservation biology, making a Type II error will often be more costly than making a Type I error (Peterman 1990b; Taylor & Gerrodette 1993). However, the specific scientific, economic and socio-political context within which research is conducted will influence the relative importance of each type of error (Di Stefano 2001), thus designating research fields in which Type I or Type II errors are more important is problematic. The relative costs of statistical errors need to be considered each time that an experiment or monitoring program is planned.

The need to identify the relative costs of Type I and Type II errors when deciding on an acceptable level of statistical power has been discussed at some length in the literature (Toft & Shea 1983; Peterman 1990a; Peterman & Mgonigle 1992; Mapstone 1995; Keough & Mapstone 1997; Downes *et al.* 2002). Although evaluating the relative costs of Type I and Type II errors is complex and may be influenced by the perspective of the decision maker, the process brings the functional consequences of statistical errors to the fore and thus promotes rational consideration of the scientific, economic or socio-political issues that may be at stake. For example, consider an experiment designed to test the ecological impact of a toxin in a waterway. In this case the cost of making a Type II error (concluding there is no toxic effect when there is one) is arguably greater than the cost of making a Type I error (concluding that there is a toxic effect when one does not exist). A Type I error may result in unnecessary clean up efforts, or the implementation of unjustified fines or other penalties. A Type II error, however, would result in management inaction and (depending on the type and quantity of the toxin) potentially serious environmental damage. Peterman (1990b) provides an example from the field of fisheries ecology where making a Type II error led to management inaction and a subsequent decline in fish stocks.

Alternatively, if researchers were interested in testing whether a new highly mechanised timber-harvesting technique resulted in better wildlife habitat than a conventional labour-intensive technique, the cost of making a Type I error (concluding that the new technique is better when it is not) may be greater. If the implementation of the new technique resulted in large technology changeover costs and job losses, making a Type I error could place substantial economic pressure on the local human population. A Type II error (concluding that the new technique made no difference when in fact it results in better habitat) may be less important if there were no pressing reason to improve

Correspondence: Forest Science Centre, Water St, Creswick Vic, 3363. Tel: + 6 35321 4259. Fax: + 6 35321 4277. E-mail: julian.distefano@dse.vic.gov.au

wildlife habitat in the area. In other situations, the costs of Type I and Type II errors may be the same. Clearly, appropriate values for α and β (and hence power) should not be fixed, but should vary depending on circumstances specific to each experiment or monitoring program.

An example of this approach involves defining an acceptable ratio of α:β in the planning phase of an experiment, deciding on ideal values for α and power and then computing the sample size required to achieve these values. If achieving this sample size falls beyond the project budget (a common occurrence when experimental units are large or data are inherently variable) new values for α and power can be assigned, but the initial α:β ratio does not change; maintenance of this ratio is important as it preserves the relative cost of Type I and Type II errors throughout the process of sample size determination (Mapstone 1995; Keough & Mapstone 1997; Downes *et al.* 2002). Following this process not only facilitates rational determination of error rates and power but enables researchers to assess the feasibility of their initial objectives. For cases where power is low, raising α within acceptable limits or redesigning the study to increase power (e.g. Foster 2001) will improve the clarity, precision and usefulness of statistical outputs and thus may even increase the likelihood of publication.

Research recently published in *Functional Ecology* serves to illustrate the inappropriate use of the five-eighty convention. Perkins & Speakman (2001) investigated whether measuring the abundance of ^{13}C in the breath of laboratory mice could be used to detect differences in their diet. One of their experiments used ANOVA to detect differences in ^{13}C abundance between three groups of 10 mice fed diets of wheat, maize and mealworm, respectively. Although the ANOVA detected differences between maize and wheat diets and maize and mealworm diets, differences between the wheat and mealworm diets were not detected due to high variance and a small effect size. Using the five-eighty convention, Perkins and Speakman predicted that if the experiment were conducted again, 41 mice per group would be required to establish a statistically significant difference between the wheat and mealworm diets on the basis of ^{13}C abundance.

This calculation was informative because it suggested that, in a future experiment, large numbers of sample animals would be needed to differentiate between diet types if the ^{13}C signature of different diets was similar. However, the use of the five-eighty convention implied that the cost of a Type I error was four times more important than the cost of a Type II error when this was clearly not the case. In the context of a future experiment where diets are unknown, a Type I error would mean detecting a difference between diets when diets were the same, and a Type II error would mean concluding that diets were the same when in fact they were different. Given no further information, it is reasonable to assume that the cost of Type I and Type II errors in an experiment of this nature would be approximately equal. Thus, if α was set at 0·05, power should have been 0·95. If these values were used in the power calculation the number of mice per group rises from 41 to 67. Although the general conclusion (that a large sample of mice would be needed to establish a statistically significant difference in ^{13}C abundance) does not change, a sample size of 67 has a logical basis while a sample size of 41 does not. Theoretically, other combinations of α and power would also be acceptable as long as the α:β ratio remained at 1:1 – it would simply depend on the level of error probability that researchers (and perhaps other stakeholders) were prepared to accept.

Ironically, the increasing use of the five-eighty convention may be due to Jacob Cohen, or rather to the misrepresentation of comments made in his well known power analysis book (Cohen 1988). Cohen proposed the use of the five-eighty convention (Cohen 1988, p. 56), and this has sometimes been seen to legitimise its use in ecological studies (Lougheed *et al.* 1999; Walsh *et al.* 1999). Cohen's comments, however, are made with reference to the field of behavioural psychology where (he suggests) the cost of Type I errors is usually greater than the cost of Type II errors. This is frequently *not* the case in ecological studies. In addition, Cohen suggests the use of the five-eighty convention only when researchers have no other basis for setting the desired level of power. In almost every case, a rational basis for determining an adequate level of power can be gained by considering the relative costs of Type I and Type II errors.

If the results of ecological studies are to be considered rational and logical, the process of data analysis and interpretation itself needs to be as rational and logical as possible. Reverting to baseless conventions does not help to achieve this end. Using the five-eighty convention for power analysis may be appealing because it suggests objectivity and removes the need to make difficult decisions about the relative cost of Type I and Type II errors. Nevertheless, values of α and power have important implications for the interpretation of power analyses and should not be set arbitrarily. Advice by Mapstone (1995), Keough & Mapstone (1997) and others suggesting a logical process for determining these values in a *priori* power calculations is sound and should be followed. Similar logical processes should be applied when calculating power retrospectively.

Acknowledgements

Thanks to Lauren Bennett, Sabine Kasel, Jan Carey and Alan York for commenting on early drafts, and to Tom M^{c}Kenzie for inspiration. Thanks also to Sarah Perkins and John Speakman who facilitated the sample size recalculation using their original data, and to John Hutchinson and another anonymous referee for making valuable suggestions.

References

Cohen, J. (1962) The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology* **69**, 145–153.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum, New Jersey.

Di Stefano, J. (2001) Power analysis and sustainable forest management. *Forest Ecology and Management* **154**, 141–153.

Downes, B.J., Barmuta, L.A., Fairweather, P.G., Faith, D.P., Keough, M.J., Lake, P.S., Mapstone, B.D. & Quinn, G.P. (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge University Press, Cambridge.

Fairweather, P.G. (1991) Statistical power and design requirements for environmental monitoring. *Australian Journal of Marine and Freshwater Research* **42**, 555–567.

Foster, J.R. (2001) Statistical power in forest monitoring. *Forest Ecology and Management* **151**, 211–222.

Keough, M.J. & Mapstone, B.D. (1997) Designing environmental monitoring for pulp mills in Australia. *Water Science and Technology* **35**, 397–404.

Lougheed, L.W., Breault, A. & Lank, D.B. (1999) Estimating statistical power to evaluate ongoing waterfowl population monitoring. *Journal of Wildlife Management* **63**, 1359–1369.

Manolis, J.C., Andersen, D.E. & Cuthbert, F.J. (2000) Patterns in clearcut edge and fragmentation effect studies in northern hardwood–conifer landscapes: retrospective power analysis and Minnesota results. *Wildlife Society Bulletin* **28**, 1088–1101.

Mapstone, B.D. (1995) Scalable decision rules for environmental-impact studies – effect size, Type-I, and Type-II errors. *Ecological Applications* **5**, 401–410.

Perkins, S.E. & Speakman, J.R. (2001) Measuring natural abundance of ^{13}C in respired CO_2: variability and implications for non-invasive dietary analysis. *Functional Ecology* **15**, 791–797.

Peterman, R.M. (1990a) The importance of reporting statistical power: the forest decline and acidic deposition example. *Ecology* **71**, 2024–2027.

Peterman, R.M. (1990b) Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Science* **47**, 2–15.

Peterman, R.M. & Mgonigle, M. (1992) Statistical Power Analysis and the Precautionary Principle. *Marine Pollution Bulletin* **24**, 231–234.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Steidl, R.J., Hayes, J.P. & Schauber, E. (1997) Statistical power analysis in wildlife research. *Journal of Wildlife Management* **61**, 270–279.

Strehlow, K., Bradley, J.S., Davis, J. & Friend, G.R. (2002) Short term impacts of logging on invertebrate communities in jarrah forests in south-west Western Australia. *Forest Ecology and Management* **162**, 165–184.

Taylor, B.L. & Gerrodette, T. (1993) The uses of statistical power in conservation biology – the Vaquita and Northern Spotted Owl. *Conservation Biology* **7**, 489–500.

Toft, C.A. & Shea, P.J. (1983) Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist* **122**, 618–625.

Walsh, H.E., Kidd, M.G., Moum, T. & Friesen, V.L. (1999) Polytomies and the power of phylogenetic inference. *Evolution* **53**, 932–937.

Received 4 February 2003; revised 23 May 2003; accepted 11 June 2003

JULIAN DI STEFANO
Forest Science Centre,
Water St.,
Creswick Vic, 3363,
Australia