The importance of statistical power analysis: an example from *Animal Behaviour*

LEN THOMAS* & FRANCIS JUANES†

*Centre for Applied Conservation Biology, University of British Columbia

†Department of Forestry and Wildlife Management and Graduate Program in Organismic and Evolutionary Biology, University of Massachusetts

(Received 16 January 1996; initial acceptance 26 February 1996; final acceptance 30 May 1996; MS. number: SC-1072)

'Power calculations are as important as significance calculations . . .'

Greenwood (1993)

'Greenwood's suggestion that power analysis should be regularly incorporated into experiments will find few dissenters . . .'

Thompson & Neill (1993)

Statistical significance and biological significance are not the same thing. For example, given a large enough sample size, any statistical hypothesis test is likely to be statistically significant, almost regardless of the biological importance of the results. Conversely, when the sample size is small, biologically interesting phenomena may be missed because statistical tests are unlikely to yield statistically significant results. Statistical and biological significance can be linked through the use of statistical power analysis. While power analysis is gaining popularity in many branches of biology (reviewed by Fairweather 1991; Taylor & Gerodette 1993; Searcy-Bernal 1994), it has been largely ignored in others, including animal behaviour, despite attempts to draw it to general attention (Greenwood 1993). The purpose of this article is to reinforce Greenwood's advice by clearly demonstrating the relationship between sample size, biological significance and statistical power and by providing key references to introductory papers, texts and computer software. We use an example from *Animal Behaviour* to illustrate the importance of power analysis and the consequences of ignoring power.

Our example is taken from a recent aquarium study on the willingness of juvenile rainbow trout, *Oncorhynchus mykiss*, to forage under the risk of predation (Johnsson 1993). As a part of the study, the investigator tested the null hypothesis that large and small juvenile trout do not differ in their susceptibility to predation. To test this hypothesis, eight replicate groups of six large and six small juveniles were exposed one by one to a standardized encounter with a predatory adult trout. On average 19 ± 4.9% ($\bar{X}$ ± SE) of the large fish and 45 ± 7.0% of the small fish were killed by the predator. The difference between the two size classes was not statistically significant using a Wilcoxon signed-ranks test and a significance criterion of $\alpha=0.05$ ($T^+=29$, $N=8$, $P=0.15$). Does this mean that the null hypothesis of no difference should be accepted? Not necessarily: another possibility is that there exists a biologically significant difference in susceptibility to predation in the population, but that the test was not sensitive enough to detect it. Statistical power analysis allows the evaluation of these two alternatives.

The statistical power of a test is the probability of getting a statistically significant result, given that the null hypothesis is false. Power is proportional to the sample size, significance criterion (α level) and effect size, and is inversely proportional to the variance in the population. Effect size is a measure of biological significance: it is the difference between the results predicted by the null hypothesis and the actual state of the population being tested. In our example effect size is the difference in probability of predation between size classes. Power analysis can be used to determine whether the experiment had a good chance of producing a statistically significant result if a biologically significant difference existed in the population. In other words, whether the

Correspondence: Len Thomas, Centre for Applied Conservation Biology, Faculty of Forestry, No. 270-2357 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada (email: lthomas@unixg.ubc.ca).

0003–3472/96/100856+04 $25.00/0

Table I. A selection of microcomputer software that can be used to calculate the statistical power of many common tests*

Name	Operating systems supported	Reference
General purpose statistical software that contains power analysis routines		
JMP	Windows, Macintosh	SAS Institute 1995
General purpose statistical software that can be programmed to calculate power		
SAS†	DOS, Windows, OS/2, Macintosh, Unix	SAS Institute 1990
S-PLUS‡	Windows, Unix	StatSci 1995
Stand-alone power analysis software		
GPOWER§	DOS, Macintosh	Erdfelder et al., in press
PASS	Windows	Hintze 1996
nQuery	Windows	Elashoff 1995
Stat-Power	DOS	Bavry 1996
STPLAN**	DOS	Brown et al. 1996
[13 other programs]††	DOS	Reviewed by Goldstein 1989

*A complete list with contact addresses is available at http://www.interchg.ubc.ca/cacb/power.

†See (1) sample library programs AOVPOWER and POWER; (2) Nemac, 1991; (3) POWERLIB.SAS by Keith E. Muller (ftp://ftp.uga.edu/pub/sas/contrib/cntb0014); (4) UnifyPow. SAS by Ralph G. O'Brien (ftp://ftp.bio.ri.ccf.org/UnifyPow.all)

‡See POWER by Daniel F. Heitjan (ftp://lib.stat.cmu.edu/s/power).

§ftp://ftp.uni-trier.de/pub/pc/msdos/Gpower2i.exe.

**ftp://odin.mdacc.tmc.edu/pub/msdos/stplan41.exe.

††Design Power; EPISTAT; Ex-sample; GAUSS; MSUSTAT; N; NCSS; NONCDIST and SMPLSIZE; Power; PowerPack; SST; Statistical Power Analysis; Systat Design.

experiment had high power, given a biologically significant effect size. What constitutes 'high power' is best judged by the researcher, but conventions of 0.8 and 0.95 have been suggested in the literature (Cohen 1988, page 56; Peterman 1990).

For most common statistical tests, power is easily calculated from tables (e.g. Cohen 1988; Zar 1996) or using statistical computer software (Table I). For more complex tests, and for most non-parametric statistics, pre-fabricated tables are often not available. In these cases, Monte Carlo simulations or bootstrapping techniques (Manly 1991) can be used to estimate power. Hence, in order to examine the power of the Wilcoxon signed-ranks test in our example, we used a simulation.

We modelled the experiment by assuming that each standardized encounter with the predator was an independent trial, with a probability of predation that was dependent on size class alone. We set the mean probability of predation to be the same as that observed in the experiment (0.32, the mean of 0.19 and 0.45), and specified the differences between the two size classes (the effect size) using a parameter in the model. Each run simulated eight replicates of six large and six small fish (same as the experiment), using a pseudo-random number generator to determine whether each fish survived or was predated (function RANUNI in SAS: SAS Institute 1990). As a partial validation, we ran the model with the effect size set equal to the observed difference in probability of predation (0.26, the difference between 0.19 and 0.45). The simulated results were similar, when averaged over a large number of simulations, to those observed in the experiment (mean 19 ± 5.4% of large fish and 45 ± 6.9% of small fish were predated, averaged over 100 000 simulations).

If the model is an accurate representation of the experiment, then the probability of getting a statistically significant result from the model is equal to the probability of getting a statistically significant result in the experiment, that is, the statistical power. The probability of a significant result in the model can be calculated by repeating the simulation a large number of times and computing the proportion of runs that produced significant results. Thus, in order to estimate the statistical power of the Wilcoxon signed-ranks test at a given effect size, we repeated the simulation 100 000 times at that effect size, and recorded the proportion of runs that were statistically significant using the signed-ranks test and a significance criterion of $\alpha = 0.05$. We used 100 000 simulations because this made our estimate of power

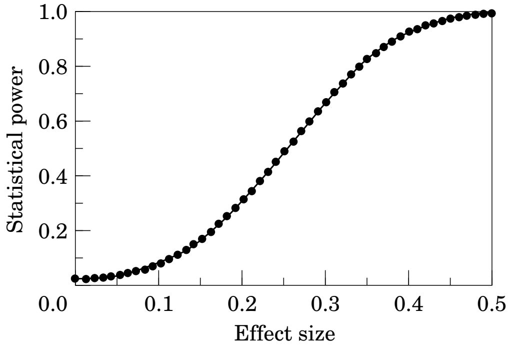

Figure 1. Statistical power to detect a given effect size (difference in probability of predation) with eight replicates of six large and six small juvenile trout using a Wilcoxon signed-ranks test. The mean probability of predation was 0.32 and the significance criterion was α=0.05. Statistical power was estimated using a Monte Carlo simulation based on the susceptibility to predation experiment reported in Johnsson (1993).

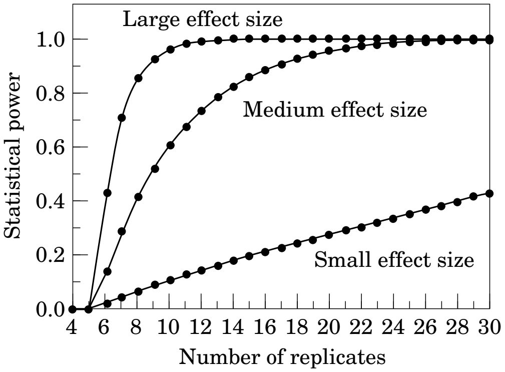

Figure 2. The influence of number of replicates on statistical power to detect small (0.09), medium (0.23) and large (0.36) effect sizes (differences in the probability of predation) between six large and six small trout using a Wilcoxon signed-ranks test. The mean probability of predation was 0.32 and the significance criterion was α=0.05. Statistical power was estimated using a Monte Carlo simulation based on the susceptibility to predation experiment reported in Johnsson (1993).

precise to two decimal places; 1000 should be sufficient where precision to one decimal place is satisfactory (formula 17 in Link & Hatfield 1990).

There are no agreed conventions as to what constitutes a biologically significant effect: this will depend upon the context of the experiment and on the judgement of the researcher. We therefore ran the simulation over a wide range of effect sizes, from 0 to 0.5 (Fig. 1). In the absence of biological intuition, Cohen (1988) has suggested that power be calculated at effect sizes implied by the adjectives small, medium and large, and has provided operational definitions of these for a number of common tests. Applying his definitions for tests of differences in proportions, a small effect translates into an effect size (difference in probability of predation) of 0.09, a medium effect is 0.23 and a large effect is 0.36. The power at these effect sizes was 0.07, 0.41 and 0.85, respectively (Fig. 1).

Our intuition tells us that a difference in probability of predation of 0.23 between two size classes of fish (a medium effect) is undoubtedly biologically significant in the context of the research goal, which was to determine whether the two size classes are taking different risks when exposing themselves to the predator. A power of 0.41 at this effect size means that the experiment had a less than even chance of yielding a statistically significant result if a difference in probability of predation of 0.23 existed in the population. We conclude that there may be a biologically significant difference in susceptibility to predation between size classes of juvenile trout, but that these data were not sufficient to tell. This conclusion is consistent with other studies of piscivorous fish predation, which have shown that smaller prey are generally more vulnerable to predation (reviewed by Juanes 1994). Indeed, in the experimental fish, smaller juveniles were more than twice as likely to be predated as larger juveniles on average although the difference was not statistically significant owing to low power.

Clearly, more data should have been collected. But how much more? We repeated the above simulations using small, medium and large effect sizes and varying the number of replicates from four to 30 (Fig. 2). Given a medium effect size, a statistical power of 0.8 was obtained with 14 replicates (only six more than were taken), and a power of 0.95 required 20 replicates. Analyses such as these, when performed a priori, can help researchers to design experiments that are more likely to lead to conclusive results. Power analysis can be used to calculate the number of samples required to have a good chance of detecting a biologically significant effect size, and to compare the power of different experimental designs. Where it is not possible to take sufficient samples, the power of an experiment can be increased by increasing the significance criterion to greater than the customary level of 0.05 (see discussion by Shrader-Frechette & McCoy 1992).

Our example has shown that power analysis is useful when planning an experiment, and that it is a vital precursor to the interpretation of non-significant results. The quotations that begin this article, which were taken from a recent issue of *Animal Behaviour*, imply that the importance of power analysis is already widely appreciated amongst ethologists. However, a survey of the literature suggests otherwise: of the 359 research articles that appeared in *Animal Behaviour* in 1994 (volumes 47 and 48), 279 included at least one statistically non-significant result but only one (Queller 1994) reported the power of the tests used. Clearly, power analysis is being largely ignored.

Perhaps some ethologists are unfamiliar with the concept of power analysis, or are not convinced of its fundamental importance. We advise these people to read some of the many introductory papers in the recent biological literature (e.g. Peterman 1990; Fairweather 1991; Nemac 1991; Taylor & Gerodette 1993; Searcy-Bernal 1994 and references therein). Others may be unaware of the resources available to calculate power (Table I). We hope that our comments will stimulate researchers to use power analysis as a tool to increase the strength of their inferences, and editors and reviewers to demand that statistical power be reported in all cases where a non-significant result is obtained.

We thank Christopher Barnard, Brad Blackwell, Joe DeGisi, Jeremy Greenwood, Jay Hestbeck, Scott Hinch, Wes Hochachka, Jörgen Johnsson, James McCormick and Katrine Vogt for their helpful comments. L.T. is supported by a Canadian Commonwealth Scholarship. Copies of the SAS programs used to estimate power are available from L.T.

REFERENCES

Bavry, J. L. 1996. *Statistical Design Analysis Software (Version 3): User's Guide.* Portland, Oregon: QEI Systems.

Brown, B. W., Brauner, C., Chan, A., Gutierrez, D., Herson, J., Lovato, J., Plosley, J. & Russell, K. 1996. *STPLAN: Calculations for Sample Sizes and Related Problems.* Version 4.1. Houston, Texas: University of Texas.

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

Elashoff, J. 1995. *nQuery Users Guide.* Version 1.0. Crosse's Green, Ireland: Statistical Solutions.

Erdfelder, E., Faul, F., & Buchner, A. In press. GPOWER: a general power analysis program. *Behav. Res. Meth., Instr. & Comp.*

Fairweather, P. G. 1991. Statistical power and design requirements for environmental monitoring. *Austral. J. mar. Freshwat. Res.*, **42,** 555–567.

Goldstein, R. 1989. Power and sample size via MS/PC-DOS computers. *Am. Stat.*, **43,** 253–260.

Greenwood, J. J. D. 1993. Statistical power. *Anim. Behav.*, **46,** 1011.

Hintze, J. L. 1996. *PASS Users Guide.* Version 6.0. Kaysville, Utah: NCSS.

Johnsson, J. I. 1993. Big and brave: size selection affects foraging under risk of predation in juvenile rainbow trout, *Oncorhynchus mykiss. Anim. Behav.*, **45,** 1219–1225.

Juanes, F. 1994. What determines prey size selectivity in piscivorous fishes? In: *Theory and Application in Fish Feeding Ecology.* Belle W. Baruch Library in Marine Sciences 18 (Ed. by D. J. Stouder, K. L. Fresh & R. J. Feller), pp. 79–100. Columbia, South Carolina: University of South Carolina Press.

Link, W. A. & Hatfield, J. S. 1990. Power calculations and model selection for trend analysis: a comment. *Ecology,* **77,** 1217–1220.

Manly, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology.* London: Chapman & Hall.

Nemac, A. F. L. 1991. *Power Analysis Handbook for the Design and Analysis of Forestry Trials.* Biometrics Information Handbook 2. Victoria, British Columbia: Ministry of Forests.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. *Can. J. Fish. Aquat. Sci.*, **47,** 2–15.

Queller, D. C. 1994. A method for detecting kin discrimination within natural colonies of social insects. *Anim. Behav.*, **47,** 569–576.

SAS Institute. 1990. *SAS Language: Reference, Version 6.* Cary, North Carolina: SAS Institute.

SAS Institute. 1995. *JMP Statistics and Graphics Guide.* Version 3.1. Cary, North Carolina: SAS Institute.

Searcy-Bernal, R. 1994. Statistical power and aquacultural research. *Aquaculture*, **127,** 371–388.

Shrader-Frechette, K. S. & McCoy, E. D. 1992. Statistics, costs and rationality in ecological inference. *Trends Ecol. Evol.*, **7,** 96–99.

StatSci. 1995. *Statistical Sciences, S-PLUS User's Manual, Version 3.3 for Windows.* Seattle, Washington: Statsci (MathSoft, Inc.).

Taylor, B. L. & Gerodette, T. 1993. The uses of statistical power in conservation biology: the Vaquita and Northern spotted owl. *Conserv. Biol.*, **7,** 489–500.

Thompson, C. F. & Neill, A. F. 1993. Statistical power and accepting the null hypothesis. *Anim. Behav.*, **46,** 1012.

Zar, J. H. 1996. *Biostatistical Analysis.* 3rd edn. London: Prentice-Hall.