Behavioral Ecology Vol. 14 No. 3: 446–450

Forum

Confidence intervals are a more useful complement to nonsignificant tests than are power calculations

Nick Colegrave[a] and Graeme D. Ruxton[b]
[a]ICAPB, University of Edinburgh, Ashworth Laboratory, Kings Building, west Mains Road, Edinburgh, EH9 3JT, UK, and [b]Division of Environmental and Evolutionary Biology IBLS, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, Glasgow, UK

Many leading journals, including *Behavioral Ecology,* emphasize the importance of considering the power of statistical tests in the light of nonsignificant results. However, there is considerable scope for misinterpretation of what this advice actually implies. The common conception among biologists is that a nonsignificant result with low power is not to be relied on, but a nonsignificant result with high power is strong support for the null hypothesis. Here we will draw on a recent paper by Hoenig and Heisey (2001) to explain why use of (post hoc) power analysis actually provides no more information than does the p value itself.

Imagine we are interested in knowing whether the type of bean on which a beetle larva develops affects the size of the beetle as an adult. We raise 10 beetle larvae on black-eyed beans and 10 on mung beans, and we measure their size as adults. The mean size of beetle raised on black-eyed beans is 5.32 mg (± 0.33 SD); on mung beans, 4.95 mg (± 0.56 SD). A t test to compare the mean size of the two groups of beetles yields a nonsignificant p value ($p = 0.09$); what can we conclude?

The reasoning behind recommending a power calculation to help us draw conclusions is that a nonsignificant outcome can occur for two very different reasons: (1) the null hypothesis of no effect is actually true; for example, there really is no difference in the size of beetles raised on the two hosts; and (2) the null hypothesis of no effect is actually false, but a combination of some or all of small effect size, high within-population variability, and small sample sizes prevent the test from being able to detect this effect (e.g., the species of host bean does affect the size of beetles, but we failed to detect this because our sample size was too low). This is what is referred to by statisticians as a type 2 error.

The argument behind recommending a power calculation is that if we get a nonsignificant result but have low power, we cannot discriminate between these two alternatives. In contrast, if we get a nonsignificant result but have high power, then this suggests that the null hypothesis of no-effect really is true.

A power calculation performed on this experiment (and provided by some statistical packages such as SPSS) gives the so-called observed power. That is, it is the probability of rejecting the null hypothesis, assuming that the effect size measured from the samples is the true effect size and the variabilities measured in the samples are identical to the variabilities of the populations from which they are drawn. Using our example, this means that we assume that the real effect of host bean is to change beetle size by 0.37 mg (5.32−4.95 mg) and that the true SD of the populations is 0.46 (the pooled SD of the two samples). These assumptions may not be true, but they are adopted for reasons of convenience if we have no other information on the actual properties of the two populations. We can then calculate the observed power, which in this case is 0.40. Does this estimate of the power of our experiment allow us to conclude anything new?

Hoenig and Heisey (2001) demonstrate that the calculated p value and observed power (often labeled B) are inextricably linked. Indeed, there is a one-to-one correspondence between the p and B values for a given statistical test. Hence, if we already have the p value, calculating B is pointless as it provides no further information. As Hoenig and Heisey put it, "computing the observed power after observing the p value should cause nothing to change about our interpretation of the p value. . . . Higher observed power does not imply stronger evidence for a null hypothesis that is not rejected" (20–21).

So is there nothing extra that we can learn from a post-hoc power analysis? A second way in which power analysis has been frequently used is to determine what is called the detectable effect size. This can be defined as the size that the biological effect would have to be if we are to have a reasonable chance of detecting it with our experimental design. Thus, with the above experiment, we only have a power of 0.40 to detect the observed effect (of a 0.37-mg difference owing to bean type). However, if host bean actually had an effect of 0.62 mg, our experiment would have had a much greater power of 0.81. What do we define as a reasonable chance of detecting a difference? Clearly, this will depend on the purpose of the study, but by convention, a power of about 0.80 is regarded as acceptable for most purposes (see Cohen, 1988). Thus, we can regard 0.62 mg as the detectable effect size of our experiment; if the effect of bean type was equal or greater than this detectable effect size, we would expect to able to detect it 80% of the time with an experiment like the one we performed. This detectable effect size is often then used as a measure of our confidence that the null hypothesis is actually true; the closer the detectable effect size is to zero, the more confident we can be that the true effect size really is zero.

Again, Hoenig and Heisey (2001) point out that this kind of logic can lead to a fundamental paradox that makes inference drawn in this way invalid. To see why, imagine we repeat the experiment described above. Our new experiment has the same sample sizes, and the two groups of beetles have exactly the same mean values as in the first experiment. The only difference is that this time we obtain a p value of 0.21 (compared with 0.09 in the first experiment). What can we conclude from our two experiments? Because the p value is the probability that the null hypothesis is actually true given this data, the first experiment gives us more confidence that the null hypothesis is false than does the second (although in neither case does the p value become so low that we would reject the null hypothesis with any confidence).

Now suppose that we determine the detectable effect size for this second experiment. The fact that the sample sizes and effect sizes are unchanged, but the p value is higher in the second experiment implies that the variation in the second experiment is also higher than in the first, and so its minimum detectable effect size will be greater (i.e., the effect would have to be larger for us to have confidence in detecting it). Because the detectable effect size of the first experiment is closer to zero than the second, this suggests that we have more confidence from the first experiment that the null hypothesis

Note: this could happen if the pooled standard deviation was higher, by chance, in this second experiment.

This is NOT the correct definition of *p*. Their subsequent argument is not effected by this error.

is actually true than from the second. This is exactly the opposite conclusion to that drawn from the *p* values. Thus, using the detectable effect size of an experiment that produces a nonsignificant result to infer something about the probability that the null hypothesis is indeed true is a flawed endeavor.

Detectable effect sizes have also been used to give some idea of the maximum biological effect that can be supported by the data. The argument goes, "because my experiment had a detectable effect size of 0.62 mg, I can be confident that although there may be some effect of host bean, it is smaller than 0.62 mg, otherwise I would have detected it". Such figures are also often compared with a priori expectations of the size of a biologically interesting or important effect size. However, even this use of power analysis is generally not helpful and runs into the same problems discussed above. If our suggestion below with regard to confidence intervals is adopted, such analysis is also not necessary.

Is there something we can use instead of observed power to get any idea of how we should interpret nonsignificant results? We would recommend quoting a confidence interval for the effect size, whether or not the *p* value was above or below 0.05. As has been argued so many times (see Johnson, 1999), the strong dichotomy in some people's minds between "significant" *p* values below 0.05 and "nonsignificant" ones above 0.05 is false. What we are interested in is the description of the possible effect sizes that are supported by the data that we have, and the possible effect sizes that are not supported. Confidence intervals are most simple and efficient way to convey this.

How does a confidence interval help us to interpret nonsignificant results? If the test was nonsignificant, then the confidence interval for effect size will span zero. However the breadth of that confidence interval gives an indication of the likelihood of the real effect size being zero (or at least very small). We return to our example *t* tests. Imagine that one of us tested the difference in weight between beetles raised in black-eye and mung beans by using a *t* test and found a *p* value above 0.05 and a 95% confidence interval for the weight difference of (−0.07–0.81 mg). Imagine now that the other of us performed a similar experiment and calculation and got a nonsignificant *p* value and a 95% confidence interval of (−0.59–1.33 mg). The first confidence interval is narrower and more consistent with the null hypothesis of no effect actually being true than the second.

If this seems a little too touchy-feely, then this idea can be formalized. However, this requires us to introduce equivalence testing. In this case, we have a null hypothesis that the effect size is actually greater than some defined value Δ. Schuirmann (1987) argues that if a $1-2\alpha$ confidence interval lies entirely between $-\Delta$ and Δ, then we can reject the null hypothesis at the α level. For the examples above, we are dealing with 95% confidence intervals, so $\alpha = 0.025$. Hence, for the first confidence interval of (−0.07–0.81 mg), we can reject the hypothesis that the difference in weight caused by rearing conditions is greater than 0.81 mg at the 2.5% level. For the other confidence interval, we can reject the hypothesis that larval bean makes a difference of greater than 1.33 mg at the 2.5% level. Thus, the confidence limit of largest in magnitude gives us an estimate of the maximum effect size that is supported by our data.

The reservations about the use of post-hoc power analysis do not of course apply to the other use of power analysis—to determine the optimal size and design of a planned study (Cohen, 1988; Lipsey, 1990). For this purpose, power analysis is still a powerful tool, and we recommend that its use in this role be encouraged.

Hence, our general advice is that we should adopt confidence intervals for effect sizes more widely, to encourage us to think more about the range of effect sizes that are supported by the data and those that are not and think less about *p* values.

Address correspondence to N. Colegrave. E-mail: n.colegrave@ed.ac.uk.

Received 21 June 2002; revised 29 September 2002; accepted 1 October 2002.

REFERENCES

Cohen J, 1988. Statistical power analysis for the behavioral sciences, 2nd ed. New York: Lawrence Erlbaum Associates.

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

Johnson DH, 1999. The insignificance of statistical significance testing. J Wildl Manage 63:763–772.

Lipsey MW, 1990 Design sensitivity: statistical power for experimental research. New York: Sage.

Schuirmann DJ, 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of bioavailability. J Pharmacokinet Biopharm 15:657–680.