

## COMMENTARIES

### A Note on Statistical Power

L. SCOTT FORBES<sup>1</sup>

Formal hypothesis testing consists of a series of steps: designation of the null ( $H_0$ ) and alternative ( $H_A$ ) hypotheses, collection of data to discriminate between these hypotheses, and then statistical analysis of the data. If the probability of an observed result is less than 5% ( $P < 0.05$ ), we generally accept  $H_A$  and reject  $H_0$ . Conversely, if  $P > 0.05$ , we have failed to reject the null hypothesis. Most analyses stop here. If the null hypothesis is not rejected, however, the statistical power of the test should be evaluated. Power analysis adds another important dimension to hypothesis testing, and failure to evaluate power may result in false inferences. A recent example illustrates these points.

Edwards et al. (1988) presented evidence that a correlation exists between the sex ratios of fledgling Golden Eagles (*Aquila chrysaetos*) and the abundance of food (jackrabbit density), presumably because the larger females have an advantage in sibling competition when food is short. Bortolotti (1989) and Arnold (1989) raised a number of criticisms, many of which Edwards and Collopy (1989) addressed; but, the question of whether there is evidence that sex ratio covaries with food remains. Using a Spearman's rank correlation, Arnold reanalyzed Edwards et al.'s data (their fig. 1 and table 1) and found a near zero correlation ( $r_s = -0.003$ ,  $P = 0.99$ ) between food and sex ratio. Edwards et al. (1988: 795), however, used a different statistical procedure to conclude that sex ratio was highly correlated with jackrabbit density. Such diametrically opposed conclusions are puzzling.

The nonparametric rank correlation Arnold used has two advantages. One, the underlying relationship between  $x$  and  $y$  does not have to be linear (Steel and Torrie 1980). Second, it does not require a bivariate normal distribution as does a parametric correlation, but it still retains 91% of the power of the parametric analog (Elliot 1983). These are important points because ratios are used in the analysis, and the correlation may be curvilinear. Also, it is immaterial whether the shift is from a male-biased to a female-biased sex ratio as Arnold implied, or from a male-biased to an even sex ratio as Edwards et al. (1988: 795) and Edwards and Collopy (1989: 524) state. It is necessary only that the proportion of males increases with the abundance of food.

The statistical procedure of Edwards et al. is more complex, consisting of two steps. They detected cyclicity in jackrabbit density and sex ratio with a 10-

year period. Subsequently they fitted sine curves to both data sets, and both were significant (sex ratio,  $P = 0.041$ ; jackrabbit density,  $P = 0.004$ ). The two sine curves were then compared to determine if sex ratio and jackrabbit density were independent. This was a key step. The authors *failed* to reject the null hypothesis of no difference between the two curves, and on this basis they concluded that sex ratio and jackrabbit density were *highly correlated*.

Herein lies an error, albeit a subtle one, in statistical inference. Failure to reject the null hypothesis ( $H_0$ ) may result from one of two causes. The null hypothesis may indeed be true, or the null hypothesis may be false, but we lack sufficient statistical power to detect a real difference. This is a Type II error. Statistical power is distinct from  $\alpha$  (the probability of rejecting  $H_0$  when  $H_0$  is true, which we set routinely at 0.05 or, more conservatively, at 0.01).  $\beta$  is the probability of making a Type II error, and power is  $1 - \beta$ . When the power of a test is high (close to 1), one is confident of detecting a real difference *if it exists*. Power depends upon the sample size, the magnitude of the difference (if it exists), and the level of  $\alpha$ . As  $\alpha$  becomes more conservative, the probability of a Type II error increases. Power analysis is a valuable but greatly underused statistical tool. It can be used a priori to design experimental or sampling regimes that minimize the probability of a Type II error. It can also be used in post hoc analyses (as was the case here) to evaluate the likelihood of finding a real difference if it existed. In a trivial sense, one can always fail to find a difference if the sample size is small enough. Descriptions of the various uses of statistical power can be found in Toft and Shea (1983), Rotenberry and Wiens (1985), Cohen (1988), and Peterman (1989).

Statistical power is directly relevant to Edwards et al.'s analysis of sex ratios. They made a strong positive assertion in concluding that a correlation exists between sex ratio and jackrabbit density. This correlation was based on a negative result: failure to reject the null hypothesis when the fitted sine curves for sex ratio and jackrabbit density were compared. As Toft and Shea (1983: 624) note: "When strong assertions are based on lack of statistical significance, we should meet the same stringent standards for avoiding Type II error as we now do for avoiding Type I error" ( $\beta$  should also be 0.05). Edwards et al. did not report  $\beta$  and thus have not shown that they could have detected a real difference between the sine curves *if it existed*. Only if  $\beta$  is high is it appropriate to conclude (in a probabilistic sense) that no difference exists when one fails to reject  $H_0$ . Otherwise it is appropriate to conclude only that no evidence of a

<sup>1</sup> Natural Resource Management Program, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

difference was found, not that there is no difference, a critical distinction. Edwards et al.'s assertion of a strong correlation between sex ratio and jackrabbit abundance thus is unjustified; the data are inconclusive. Similarly, one cannot use the result of Arnold's rank correlation (which Arnold interpreted correctly) to assert that there is *no effect* of food on sex ratio, because power once again was not reported. My own analysis indicates that the power of Arnold's test was indeed low,  $<10\%$  or  $\beta > 0.9$ . I suspect that the power of Edwards et al.'s test was similarly low. The fact that the sample size was sufficient to derive significant fits of the sine curves does not ensure high power in the comparison of the fitted curves. Given the unusual nature of the statistic performed, I also suspect a power analysis here might be a bit tricky. Tables of statistical power exist for most routine statistical procedures (see Cohen 1988), but where such tables are not available, power can be estimated through Monte Carlo simulation.

It is not my intent to denigrate Edwards et al. Many authors (including myself) have been guilty of the same error. Rather I commend them for making one of the few attempts to link interannual variation in food to the occurrence of brood reduction, a valuable contribution simply if it stimulates others to do the same. And, as both Bortolotti and Arnold noted, an intriguing correlation between brood size and sex ratio remains. Instead, my purpose is to encourage avian researchers to make use of statistical power analysis. Ignorance of statistical power is a pervasive and serious problem. In scanning recent issues of *The Auk*, I found no mention of  $\beta$ . In most cases when reporting a negative result, authors correctly stated that they failed to find a significant difference, but on at least 20 occasions in 10 different papers, positive assertions were drawn from negative results. A more thorough search would undoubtedly reveal many more similar examples.

Even when the original researchers correctly interpret a negative result, others frequently cite such results as evidence of no difference. Without a power analysis, such an interpretation is unjustified. For example, studies of avian reproductive costs often fail to detect significant results, but as Nur (1988) noted, the sample sizes are far too low to provide adequate power. It is incorrect to cite such studies as evidence against a cost of reproduction, although this often occurs.

Many field biologists face another problem, that of small sample sizes. In some systems this is, for all practical purposes, unavoidable. For example, acquiring data on the fledging ratios of Golden Eagles is undoubtedly both difficult and costly. Invariably, the power of such analyses is low and setting a conservative  $\alpha$  ensures frequent Type II errors. This seems to be an inefficient use of the information contained within such data sets. Recall, however, that power and  $\alpha$  are inversely related (i.e. we can "purchase"

greater power by relaxing  $\alpha$ ). For example, I found Bortolotti's (1989) reanalysis of Edwards et al.'s (1988) data on brood size and sex ratio most interesting even though the probability value for the result was greater than 0.05 (it was 0.069). Bortolotti reported the exact probability value, and all the relevant information was provided for readers to draw their own inferences from the analysis. Exact probability values are preferable to arbitrary or subjective adjectives such as significant, nonsignificant, weak, or strong. Given the nature of Edwards et al.'s data set, I find a result with  $P = 0.069$  sufficiently interesting (or even one of  $P = 0.1$  for that matter) to warrant publication. And although some readers are bound to ask where one draws the line, I would state only that, when exact probabilities are reported, there is less need to draw any line.

I note that few avian researchers make use of the techniques of power analysis even though they have long been available. Reporting power when positive assertions are drawn from negative results would be a constructive first step. To do so would strengthen our statistical inferences and, ultimately, our science.

I thank Thomas Edwards and Alan H. Brush for comments on the manuscript.

#### LITERATURE CITED

- ARNOLD, T. W. 1989. Sex ratios of fledgling Golden Eagles and jackrabbit densities. *Auk* 106: 521-522.
- BORTOLOTTI, G. R. 1989. Sex ratios of fledgling Golden Eagles. *Auk* 106: 520-521.
- COHEN, J. 1988. *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, New Jersey, Lawrence Erlbaum Associates, Inc.
- EDWARDS, T. C., JR., M. W. COLLOPY, K. STEENHOF, AND M. N. KOCHERT. 1988. Sex ratios of fledgling Golden Eagles. *Auk* 105: 793-796.
- , & M. W. COLLOPY. 1989. Response to G. R. Bortolotti and T. W. Arnold. *Auk* 106: 523-524.
- ELLIOT, J. M. 1983. Some methods for the statistical analysis of samples of benthic invertebrates. *Freshwater Biol. Assoc. Publ.* 25.
- NUR, N. 1988. The cost of reproduction in birds: an examination of the evidence. *Ardea* 76: 155-188.
- PETERMAN, R. M. 1989. Application of statistical power analysis to the Oregon coho salmon (*Oncorhynchus kisutch*) problem. *Can. J. Fish. Aquat. Sci.* 46: 1183-1187.
- ROTEBERRY, J. T., & J. A. WIENS. 1985. Statistical power analysis and community wide patterns. *Am. Nat.* 125: 164-168.
- STEEL, R. G. D., & J. H. TORRIE. 1980. *Principles and procedure of statistics*. New York, McGraw Hill.
- TOFT, C. A., & P. J. SHEA. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. *Am. Nat.* 122: 618-625.

Received 19 September 1989, accepted 17 October 1989.