Conservation Biology Vol. 11 (1997), No. 1, pages 276-280

RETROSPECTIVE POWER ANALYSIS

Len Thomas
Centre for Applied Conservation Biology, Faculty of Forestry, University of British Columbia, #270-2357 Main Mall, Vancouver, British Columbia, V6T 1Z4, Canada.
lthomas@unixg.ubc.ca

Many papers have appeared in the recent biological literature encouraging us to incorporate statistical power analysis into our hypothesis testing protocol (Peterman 1990; Fairweather 1991; Muller & Benignus 1992; Taylor & Gerrodette 1993; Searcy-Bernal 1994; Thomas & Juanes 1996). The importance of doing a power analysis before beginning a study (prospective power analysis) is universally accepted: such analyses help us to decide how many samples are required to have a good chance of getting unambiguous results. In contrast, the role of power analysis after the data are collected and analyzed (retrospective power analysis) is controversial, as is evidenced by the papers of Reed and Blaustein (1995) and Hayes and Steidl (1997). The controversy is over the use of information from the sample data in retrospective power calculations. As I will show, the type of information used has fundamental implications for the value of such analyses. I compare the approaches to calculating retrospective power, noting the strengths and weaknesses of each, and make general recommendations as to how and when retrospective power analyses should be conducted.

INFORMATION REQUIRED FOR POWER ANALYSIS

The power of a statistical hypothesis test is the probability of rejecting the null hypothesis, given that the null is false and some alternative hypothesis is true. To calculate power we must specify the sample size, α-level, sampling variance and effect size (difference between the null and alternative hypothesis). Many measures of effect size are available (Cohen 1990). These can be divided into two major classes: standardized, dimensionless measures such as correlation coefficients or *d*-values, and raw measures such as the slope in a regression analysis or difference between means. Standardized measures incorporate the sampling variance implicitly and thus remove the need to specify variance when calculating power. Raw effect size measures are generally easier to visualize and interpret.

Before a study begins values for these parameters must be assumed. Prospective power analyses are therefore exploratory in nature, investigating the relationship between

the range of sample sizes that are deemed feasible, effect sizes thought to be biologically important, levels of variance that could exist in the population (usually taken from the literature or from pilot data), and desired levels of α and statistical power. The result is a decision about the sample size and α-level that will be used and the target effect size that will be "detectable" with a given level of statistical power.

After the study is completed and the data analyzed, the perspective changes. The outcome of the statistical test is known: either the null hypothesis was rejected or it was not. If not we may be concerned that the statistical power of the test was low, i.e., that the test had a low probability of rejecting the null hypothesis even if it is false. At this point more information is available with which to calculate power. We know the α-level and sample size used, and the effect size and variance observed in the sample provide an estimate of the effect size and variance in the population. The question arises: should this additional information (particularly the observed effect size and variance) be used to retrospectively calculate the power of the test?

APPROACHES TO RETROSPECTIVE POWER ANALYSIS

Retrospective power analyses can be conducted using both the observed effect size and variance (Reed & Blaustein 1995), only the observed variance (Hayes & Steidl 1997), neither the observed effect size nor variance (Rotenberry & Weins 1985), or avoided completely by computing confidence intervals about the observed effect size (Hayes & Steidl 1997). To illustrate the differences between these approaches I use the example of two hypothetical population monitoring studies (Fig. 1). In both studies we cannot reject the null hypothesis that there is no systematic change in abundance over time, using linear regressions on log-transformed data and an α-level of 0.05 (details of the analysis methods are given in the appendix and SAS programs are available from the author on request[1]). Nevertheless, from the data it appears that population 1 is relatively stable with little annual variation in abundance, whereas population 2 may be declining at a precipitous rate although there is considerable annual variability. What do the different approaches to retrospective power analysis tell us about these results?

(1) Calculate power using the observed effect size and variance.

In the example studies raw effect size and variance can be measured as the size of the trend (slope in the regression) and the residual mean square. Power calculated using

[1] The program, REGPOW.SAS, is available on the word-wide web at http://www.interchg.ubc.ca/cacb/people/lthomas/

these values is 0.11 for population 1 and 0.31 for population 2, well below the level of 0.8 that is often considered adequate (e.g., Cohen 1988). We conclude that both studies had insufficient power to detect the observed effect sizes. One problem with this approach is that such a conclusion is almost inevitable. Both the *p*-value and power are dependent upon the observed effect size and so are inversely related such that tests with high *p*-values tend to have low power and visa-versa. Therefore calculating power using the observed effect size and variance is simply a way of re-stating the statistical significance of the test.

Two further problems arise because the calculated power is often regarded as an estimate of the "true" power of the test, i.e., the actual probability of rejecting the null hypothesis in the study population. Firstly, it is an over-estimate of the true power and calculating an unbiased estimate is problematic (Wright & O'Brien 1986; Taylor and Muller 1996). For example, in population 1 the mean unbiased estimate of power is < 0.05, which is not admissible because power cannot be lower than the α-level. Secondly, the estimate is often imprecise: the 95% confidence intervals for power are 0.05 - 0.74 for population 1 and 0.05 - 0.90 for population 2. Clearly, calculating power using the observed effect size and variance is uninformative.

(2) Calculate power using a pre-specified effect size and the observed variance.

Here an a priori value of effect size is used, such as the minimum effect size considered biologically significant. For example, suppose a population trend of 0.05 on the log scale (~ 5% per year) would be considered biologically significant in the two study populations. Power to detect a trend of this magnitude is 0.99 in population 1 and 0.07 in population 2. Because the variance is estimated from sample data, it is important to report the precision of the resulting estimate of power. In our example, the 95% confidence intervals for power are 0.81 - 1.00 and 0.06 - 0.10 respectively. (Note these intervals are narrower than in the previous approach because one less parameter is being estimated.) We conclude that the power of the test to detect a biologically significant trend was satisfactory for population 1 (lower confidence limit > 0.8) but not for population 2 (entire confidence interval $<< 0.8$) due to the high variance.

Researchers often find it difficult to specify which effect size should be considered the minimum for biological significance. One alternative is to report power over a range of effect sizes (e.g., Thomas & Juanes 1996). Another is to pre-specify the minimum acceptable level of power and determine the effect size that is "detectable" with that power ("reverse power analysis", Fairweather 1991). In our example the smallest population trend detectable with a power of 0.8 is 0.03 for population 1 (95%

confidence interval 0.02 - 0.05) and 0.32 for population 2 (0.22 - 0.61). In summary, using observed variances but not observed effect sizes is helpful because it allows one to evaluate whether the sample size and α-level were sufficient to have a good chance of detecting a biologically significant effect given the observed level of variation.

(3) Calculate power using a pre-specified standardized effect size.

Using standardized effect size measures avoids the need to specify the sampling variance, so the only information needed from the study is the sample size and α-level. For example, the coefficient of determination (r^2) is a standardized measure of effect size in regression analysis that implicitly incorporates both the trend (raw effect size) and residual mean square (sampling variance). Suppose a population trend that explains 25% or more of the variation in abundance of the example populations is considered biologically significant. The power at this standardized effect size ($r^2 = 0.25$), given 10 years of data and an α-level of 0.05, is 0.34. When the researcher cannot pre-specify a standardized effect size, Cohen (1988) suggested that power be calculated at three levels as implied by the adjectives "small," "medium," and "large". These conventions are widely used in psychology and other disciplines, where a medium standardized effect size may correspond with the median effect size found in psychological research (Sedlmeier & Gigerenzer 1989). For regression analysis, Cohen's standardized effect sizes (f^2-values) translate into r^2-values of 0.02, 0.13, and 0.26, giving powers of 0.07, 0.18, and 0.36 respectively. Reverse power analysis is also possible: the r^2 that would be detectable with a power of 0.8 is 0.53. As with the previous approach, using standardized effect sizes allows one to evaluate the study design. The major disadvantage is that it is much harder to assess the biological significance of a standardized measurement.

(4) Calculate a confidence interval about the observed effect size.

In this viewpoint statistical power, like the α-level, is only relevant before the results of the hypothesis test are known (Greenland 1988; Goodman 1994). After the study uncertainty in the results is quantified by calculating a confidence interval around the observed effect size. In the example populations the observed trend on the log scale was -0.01 for population 1 (~ 1% per year decline) with 95% confidence interval -0.03 - +0.01 (~ 3% per year decline - 1% per year increase) and -0.17 for population 2 (~ 15% per year decline) with 95% confidence interval -0.40 - +0.06 (~ 33% per year decline - 7% per year increase). The confidence interval for population 1 does not contain values considered biologically significant (0.05 or greater), so we can be confident there is no

important population trend, whereas for population 2 the confidence interval includes both zero trend and very large declines, indicating the results are inconclusive. Confidence intervals focus on estimation, rather than hypothesis testing, and so provide a useful summary of what the results tell us about the underlying population parameters. Despite their superficial similarity to detectable effect sizes, confidence intervals about observed effect sizes cannot be used to directly evaluate the study design because they do not take the desired level of statistical power into account (Peterman 1990). Nevertheless, the conclusions are often similar (compare the conclusions for the example studies in this section with those in section 2). If confidence intervals are used in presenting the results, then expected confidence interval lengths can usefully be calculated in the planning stages of the study (Greenland 1988; Goodman 1994).

CONCLUSIONS

Different retrospective analyses can yield substantially different information. The appropriate approach, therefore, depends upon the goal of the analysis. Often the goal is simply to quantify our uncertainty in the findings of a study, in which case calculating a confidence interval about the observed effect size is the most straightforward approach. Sometimes the goal is to evaluate the ability of the study to detect a biologically meaningful pattern, for example to determine whether the study meets a pre-specified target or to make comparisons between a number of different studies. In these cases calculating retrospective power (or detectable effect size) can be useful, depending upon how the analysis is done.

Calculating power using observed effect sizes is not helpful because such values are very poor estimates of the actual power of the test given the population effect size, and do not take into account the biological significance of the effect size value used. It is unfortunate that this kind of power analysis is readily available in a number of statistical software packages (e.g., SAS JMP, Sigmastat, SPSS), whereas more informative power analyses are generally harder to perform (Thomas & Krebs 1997). Calculating power using pre-specified effect sizes (or calculating detectable effect size using pre-specified power) is helpful, especially if easily interpreted raw effect size measures are used. Standardized measures may be useful in more complex tests (such as tests for interaction in multi-way ANOVA) where it is hard to specify an intuitive raw measure of effect size. In these cases power analysis may be performed using conventional levels of effect size, such as those proposed by Cohen (1988).

All power calculations should be accompanied by a sensitivity analysis. For power calculations that use assumed values for the effect size or variance, this means

trying a range of plausible values for each variable. Graphs showing how two or more variables interact with one another are particularly valuable (e.g., Peterman 1990; Muller & Benignus 1992; Taylor & Gerrodette 1993; Thomas & Juanes 1996). For power calculations that use values estimated from sample data (such as sampling variance), a confidence interval about the power estimate should be given (Taylor & Muller 1995).

It is important to note that retrospective analyses are no substitute for the proper planning of research (Cohen 1990). Only in the planning stages is it reasonable to change the sampling design, the α-level, or even to completely re-think the goals of the study. Conservation studies are, by their nature, often characterized by small sample sizes and high sampling variation. The appropriate use of power analysis and confidence interval estimation allows us to obtain the most information from our limited resources and to make an honest assessment of what our results do and do not tell us.

ACKNOWLEDGMENTS

I thank Richard Goldstein, Wesley Hochachka, Charles Krebs, Kathy Martin, Keith Muller, David Tait, and an anonymous reviewer for their useful comments. This research was supported through grants from the Natural Sciences and Engineering Research Council of Canada to Kathy Martin and a Canadian Commonwealth Scholarship to Len Thomas.

LITERATURE CITED

Cohen, J. 1988. Statistical power analysis for the behavioral sciences, 2nd Edition. Lawrence Erlbaum, Hillsdale, New Jersey.

Cohen, J. 1990. Things I have learned (so far). American Psychologist **45**:1304-1312.

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

Goodman, S. N. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of Internal Medicine **121**:201-206.

Greenland, S. 1988. On sample-size and power calculations for studies using confidence intervals. American Journal of Epidemiology **128**:231-237.

Hayes, J. P., and R. J. Steidl. Submitted. Statistical power analysis and amphibian population trends. Conservation Biology.

Muller, K. E., and V. A. Benignus. 1992. Increasing scientific power with statistical power. Neurotoxicology and Teratology **14**:211-219.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fish and Aquatic Sciences **47**:2-15.

Reed, J. M., and A. R. Blaustein. 1995. Assessment of "nondeclining" amphibian populations using power analysis. Conservation Biology **9**:1299-1300.

Rotenberry, J. T., and J. A. Wiens. 1985. Statistical power analysis and community-wide patterns. The American Naturalist **125**:164-168.

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

Sedlmeier, P., and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? Psychological Bulletin **105**:309-316.

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

Taylor, D. J., and K. E. Muller. 1995. Computing confidence bounds for power and sample size of the general linear univariate model. The American Statistician **49**:43-47.

Taylor, D. J., and K. E. Muller. 1996. Bias in linear model power and sample size calculation due to estimating noncentrality. Communications in Statistics, Theory and Methods **25**: 1595-1610.

Thomas, L., and C. Krebs. 1997. A review of statistical power analysis software. Bulletin of the Ecological Society of America **78**: in press.

Thomas, L., and F. Juanes. 1996. The importance of statistical power analysis: an example from Animal Behaviour. Animal Behaviour **52**: 856-859.

Wright, S. P., and R. G. O'Brien. 1988. Power analysis in an enhanced GLM procedure: what it might look like. Pages 1097-1102 in Proceedings of the thirteenth annual SAS users group international conference. SAS Institute, Cary, North Carolina.

APPENDIX - POWER ANALYSIS FOR SIMPLE LINEAR REGRESSION[2]

The following is a brief statement of the equations used to calculate power and confidence intervals for the example studies. For further details, including a more general formulation in the context of generalized linear modeling, see Wright and O'Brien (1988) and Taylor and Muller (1995, 1996).

The statistical model for simple linear regression is

[2] Note - there is a typesetting error in this appendix: the symbol ∃ should be ^ throughout.

$$Y_i = \iota + \kappa X_i \tag{A1}$$

where ι and κ are population parameters for the intercept and slope, and Y_i is the observed value of the dependent variable measured at level X_i of the explanatory variable (i = 1 .. n). In the example studies Y = ln (*abundance*) and X = *year*.

Let $\hat{\kappa}$ represent the estimated value of the slope parameter and $\hat{\sigma}^2$ represent the residual mean square from the regression. Using analysis of variance the null hypothesis H_0: $\kappa = 0$ is tested against the alternative hypothesis HA: $\kappa \neq 0$ with the test statistic

$$F_{obs} = \frac{SSH_{obs}/\nu_1}{\hat{\sigma}^2} \tag{A2}$$

where $SSH_{obs} = \hat{\kappa}^2 \sum x^2$ ($\sum x^2$ is an abbreviation for $\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$) and $Y_1 = 1$. If the assumptions of linear regression are met (errors in Y are independent and normally distributed with equal variance; no errors in X) then the test statistic follows a noncentral F distribution with ν_1 numerator degrees of freedom, ν_2 denominator degrees of freedom ($\nu_2 = n - 2$) and noncentrality parameter

$$\lambda = SSH_{pop}/\sigma^2 \tag{A3}$$

where SSH_{pop} is the (unknown) population sum of squares.

Let $F_F\left(\cdot \mid \nu_1, \nu_2, \lambda\right)$ represent the cumulative distribution function of the noncentral F distribution. The power of the test is

$$\text{power} = 1 - F_F\left(F_{crit} \mid \nu_1, \nu_2, \lambda\right) \tag{A4}$$

where F_{crit} is the 100·(1-α) percentile from a central F distribution with ν_1 and ν_2 degrees of freedom and α is the size of the test (i.e., the α-level). In SAS power for a given value of λ can be computed with the statements (Wright & O'Brien 1988)

```
Fcrit = FINV(1-alpha,df1,df2);
power = 1-PROBF(Fcrit,df1,df2,lambda);.
```
(A5)

(1) Calculating power using the observed effect size and variance

Power can be estimated by substituting

$$\hat{\lambda} = SSH_{obs}/\hat{\sigma}^2 \tag{A6}$$

for λ in A4 or A5. This estimate is positively biased because the expected value of $\hat{\lambda}$ is (Wright & O'Brien 1988)

$$E\left(\hat{\lambda}\right) = \left[\nu_2/\left(\nu_2 - 2\right)\right] \cdot \left[\nu_1 + \lambda\right]. \tag{A7}$$

Therefore, the mean unbiased estimate of power is calculated by substituting

$$\hat{\lambda}_{adj} = \left[\hat{\lambda} \cdot \left(\nu_2 - 2\right)/\nu_2\right] - \nu_1 \tag{A8}$$

for λ in A4 or A5.

Confidence intervals for power are found by solving (Taylor & Muller 1996)

$$F_F\left(F_{obs}|\nu_1,\nu_2,\hat{\lambda}_U\right)=\alpha_U$$

and

$$F_F\left(F_{obs}|\nu_1,\nu_2,\hat{\lambda}_L\right)=1-\alpha_L \tag{A9}$$

for $\hat{\lambda}_U$ and $\hat{\lambda}_L$, where α_U and α_L are the upper and lower tail probabilities that define the 100·(1 - α_U - α_L) percent confidence interval. If $\hat{\lambda}_C$ ($C = U$ or L) is not defined (i.e., where $F_{obs} < F_F^{-1}(1-\alpha_c|\nu_1,\nu_2,0)$), set $\hat{\lambda}_C$ to 0. In SAS $\hat{\lambda}_U$ and $\hat{\lambda}_L$ can be calculated with the statements

```
lambdaU = max(0,FNONCT(Fsamp,df1,df2,alphaU));
lambdaL = max(0,FNONCT(Fsamp,df1,df2,1-alphaL));.
```
(A10)

(2) Calculating power using a pre-specified effect size and the observed variance

Given a specified slope parameter, κ_{hyp}, calculate $SSH_{hyp}=\kappa_{hyp}^2\sum x^2$. Power is estimated by substituting

$$\hat{\lambda}_{hyp}=SSH_{hyp}/\hat{\sigma}^2 \tag{A11}$$

for λ in A4 or A5. Confidence intervals about this estimate are calculated by substituting (Taylor and Muller 1995)

$$\hat{\lambda}_{hyp,U}=\hat{\lambda}_{hyp}\cdot c_{crit}\left(1-\alpha_U|\nu_2\right)/\nu_2$$

and

$$\hat{\lambda}_{hyp,L}=\hat{\lambda}_{hyp}\cdot c_{crit}\left(\alpha_L|\nu_2\right)/\nu_2 \tag{A12}$$

for λ in A4 or A5, where $c_{crit}(p|\nu)$ is the 100.(p) percentile from a central χ^2 distribution with ν degrees of freedom. In SAS $\hat{\lambda}_{hyp,U}$ and $\hat{\lambda}_{hyp,L}$ can be calculated with the statements

```
lambdaU = lambdah*CINV(1-alphaU,df2)/df2;
lambdaL = lambdah*CINV(alphaL,df2)/df2;.
```
(A13)

Detectable effect size is calculated iteratively, starting with κ_{hyp} = 0 and increasing κ_{hyp} in small increments until the calculated power meets the desired level.

(3) Calculating power using a pre-specified standardized effect size

Given some specified r^2 value, r_{hyp}^2, calculate

$$\lambda_{hyp}=\left[r_{hyp}^2/\left(1-r_{hyp}^2\right)\right]\cdot\nu_2 \tag{A14}$$

and substitute for λ in A4 or A5. Cohen's standardized effect sizes measure, f^2, can be converted to r^2 values using the relationship $f^2=\left[r^2/\left(1-r^2\right)\right]$ (Cohen 1988, Chap. 9).

Detectable effect size is calculated iteratively, starting with $r^2_{hyp} = 0$ and increasing r^2_{hyp} in small increments until the calculated power meets or just exceeds the desired level.

(4) Calculating a confidence interval about the observed effect size

Confidence intervals for the estimated slope, $\hat{\kappa}$, are calculated as

$$\hat{\kappa}_U = \hat{\kappa} + \sqrt{F_{crit}} \cdot \hat{\sigma}_{\hat{\kappa}}$$

and

$$\hat{\kappa}_L = \hat{\kappa} - \sqrt{F_{crit}} \cdot \hat{\sigma}_{\hat{\kappa}}, \qquad \text{(A15)}$$

where $\hat{\sigma}_{\hat{\kappa}} = \sqrt{\hat{\sigma}^2 / \sum x^2}$.

FIGURE 1. Survey data for two hypothetical wildlife populations. Trend lines were fitted using linear regression on log-transformed data. For population 1 ln(*abundance*) = -0.01 *year* + 4.67, RMS (residual mean square) = 0.01, $F_{1,8} = 0.97$, $p = 0.35$, and for population 2 ln(*abundance*) = -0.17 *year* + 5.41, RMS = 0.82, $F_{1,8} = 2.75$, $p = 0.14$.