———, K. A. SMITH, AND A. L. BLELOCH. 1994. How to model it: problem-solving for the computer age. Burgess Int., Edina, Minn. 206pp.

———, J. D. ROTH, AND K. RALLS. 1995. "Mobbing" in Hawaiian monk seals: the value of simulation modeling in the absence of apparently crucial data. Conserv. Biol. 9:166–174.

TESTER, J. R., A. M. STARFIELD, AND L. E. FRELICH. 1997. Modeling for ecosystem management in Minnesota pine forests. Biol. Conserv. 80:313–324.

TVERSKY, A., AND D. KAHNEMAN. 1974. Judgment under uncertainty: heuristics and biases. Science 185:1124–1131.

WALLACE, W. A. 1994. Ethics in modeling. Oxford Univ. Press, Oxford, U.K. 266pp.

Received 15 June 1996.
Accepted 27 December 1996.

STATISTICAL POWER ANALYSIS IN WILDLIFE RESEARCH

ROBERT J. STEIDL,[1] Oregon Cooperative Wildlife Research Unit, Department of Fisheries and Wildlife, 104 Nash Hall, Oregon State University, Corvallis, OR 97331-3803, USA

JOHN P. HAYES, Department of Forest Science, Oregon State University, Corvallis, OR 97331, USA and Coastal Oregon Productivity Enhancement Program, Hatfield Marine Science Center, Newport, OR 97365, USA

ERIC SCHAUBER,[2] Department of Fisheries and Wildlife, 104 Nash Hall, Oregon State University, Corvallis, OR 97331-3803, USA

Abstract: Statistical power analysis can be used to increase the efficiency of research efforts and to clarify research results. Power analysis is most valuable in the design or planning phases of research efforts. Such prospective (a priori) power analyses can be used to guide research design and to estimate the number of samples necessary to achieve a high probability of detecting biologically significant effects. Retrospective (a posteriori) power analysis has been advocated as a method to increase information about hypothesis tests that were not rejected. However, estimating power for tests of null hypotheses that were not rejected with the effect size observed in the study is incorrect; these power estimates will always be $\leq$0.50 when bias adjusted and have no relation to true power. Therefore, retrospective power estimates based on the observed effect size for hypothesis tests that were not rejected are misleading; retrospective power estimates are only meaningful when based on effect sizes other than the observed effect size, such as those effect sizes hypothesized to be biologically significant. Retrospective power analysis can be used effectively to estimate the number of samples or effect size that would have been necessary for a completed study to have rejected a specific null hypothesis. Simply presenting confidence intervals can provide additional information about null hypotheses that were not rejected, including information about the size of the true effect and whether or not there is adequate evidence to "accept" a null hypothesis as true. We suggest that (1) statistical power analyses be routinely incorporated into research planning efforts to increase their efficiency, (2) confidence intervals be used in lieu of retrospective power analyses for null hypotheses that were not rejected to assess the likely size of the true effect, (3) minimum biologically significant effect sizes be used for all power analyses, and (4) if retrospective power estimates are to be reported, then the α-level, effect sizes, and sample sizes used in calculations must also be reported.

J. WILDL. MANAGE. 61(2):270–279

Key words: confidence intervals, effect size, experimental design, hypothesis testing, power, research design, sample size, statistical inference, statistical power analysis, Type I error, Type II error.

Although the theoretical basis of statistical power was developed decades ago (Tang 1938), power analysis has only recently gained prominence in applied ecological research. Statistical power analysis has been advocated and sometimes used to improve research designs and to facilitate interpretation of statistical results in the applied sciences (Gerrodette 1987, Peterman and Bradford 1987, Peterman 1990, Solow and Steele 1990, Taylor and Gerrodette 1993, Searcy-Bernal 1994, Beier and Cunningham 1996, Hatfield et al. 1996). Failure to consider statistical power when a null hypothesis is not rejected can lead to inappropriate management recommendations (Hayes 1987).

[1] Present address: School of Renewable Natural Resources, 325 Biological Sciences East, University of Arizona, Tucson, AZ 85721, USA.

[2] Present address: Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA.

Table 1. Possible outcomes of statistical hypothesis tests. Probabilities associated with each decision are given in parentheses.

Reality	Decision and result	
	Do not reject null hypothesis	Reject null hypothesis
Null hypothesis is true	Correct (1 − α)	Type I error (α)
Null hypothesis is false	Type II error (β)	Correct (1 − β)

Recently, many journals, including *The Journal of Wildlife Management,* have recommended or required that statistical power be reported routinely. However, statistical power analysis remains unfamiliar to many researchers and sometimes has been misapplied. Our objective is to clarify the role of power analysis in applied research by describing appropriate uses of power analysis, identifying other statistical tools that may be more convenient and appropriate than power analysis, illustrating how statistical power can be used to plan and increase the efficiency of research designs, and suggesting guidelines for reporting the results of retrospective power analyses.

We appreciate the advice by K. P. Burnham who increased the scope and quality of the manuscript. S. DeStefano provided comments on an earlier version.

BACKGROUND

In the framework of the hypothetico-deductive method (Popper 1962, Romesburg 1981), research hypotheses can never be proven; rather, they can only be disproved (rejected) with the tools of statistical inference. Each time a decision is made about whether to reject a null hypothesis in favor of an alternative, however, there are at least 2 types of errors that can be made (Table 1). First, a null hypothesis that is actually true might be rejected (a Type I error). The rate at which Type I errors will be accepted (α) is typically set by the researcher. In the framework of hypothesis testing, a null hypothesis is considered false and is rejected in favor of an alternative when $P \leq \alpha$. In these instances, results are generally reported as "significant." Second, a null hypothesis that is actually false might not be rejected (a Type II error; Table 1). The probability of a Type II error is denoted as β. Statistical power is equal to 1 − β and is defined as the probability of correctly rejecting a null hypothesis that is false (Sokal and Rohlf 1981:166).

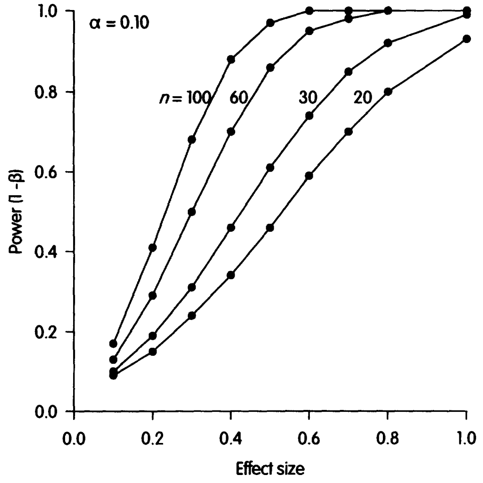

Fig. 1. The relation between power and effect size for 2-sided, 2-sample *t*-tests, α = 0.10, and $n = n_T + n_C$. Increasing sample size for a given effect size and α-level increases statistical power, as does increasing effect size for a given sample size and α-level. Increasing α-level for a given effect size and sample size also increases power (not illustrated).

Power, sample size, α, and effect size are the 4 interrelated components on which statistical hypothesis testing is based (Cohen 1988, The Wildl. Soc. 1995*a*). Each of these components is a function of all the others. Statistical power, therefore, is a function of sample size, α, and effect size. Increasing sample size, α, or effect size always increases power (Cohen 1988; Fig. 1). Effect size is the component of power least familiar to many researchers, but effect size must be specified explicitly to calculate power.

Effect size is sometimes misunderstood because its common usage and statistical meaning are often confused. Therefore, we distinguish "effect" from "effect size" and illustrate the difference in our usage of these terms by comparing the means of some variable between 2 independent populations (Fig. 2). We define effect as the absolute difference between populations in the parameter of interest, or similarly, as the change in the parameter due to application of a treatment: $|\mu_T - \mu_C|$. In Figure 2, effect = 5.5 (μ_T = 3.5 and μ_C = 9.0) for both sets of data. To determine power for a given effect, variance (σ^2) must be incorporated separately into power calculations. We define effect size as the absolute difference between populations in the parameter of interest (i.e., effect) scaled by the within-population standard deviation (σ), $|\mu_T - \mu_C|/\sigma$. Therefore, effect size is

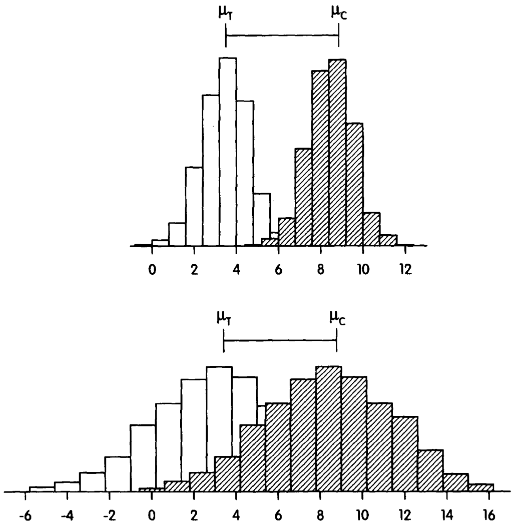

Fig. 2. Frequency distributions for 2 hypothetical populations illustrating the differences between effect and effect size. Effect ($|\mu_T - \mu_C|$) is identical for both sets of distributions; effect size ($|\mu_T - \mu_C|/\sigma$) is smaller for the lower set of distributions.

effect ($3.5 - 9.0 = 5.5$) scaled by standard deviation (if $\sigma = 2.0$, effect size $= 5.5/2.0 = 2.75$). In Figure 2, effect size differs for the 2 sets of data. Establishing a useful effect size when there are >2 groups being compared or for other types of analyses (e.g., regression, categorical models) is considered elsewhere (Cohen 1988, Richardson 1996).

Conceptually, effect can be considered as the degree to which a phenomenon of interest is present, or as the degree to which application of a particular treatment causes a change in a parameter. In applied ecological research, effect should be considered as the minimum response that will be considered biologically significant. For example, to determine if application of an agricultural pesticide reduces a resident population of small mammals by at least 20%, then the relative effect of interest is 20%. With all else equal, power to detect large effects is always greater than power to detect small effects.

Power analysis can be used to improve research design (prospective or a priori) and to provide information about results from completed research efforts (retrospective or a posteriori). Prospective power analysis can help researchers design research efforts that have a high probability of detecting biologically significant effects (i.e., high power). Retrospective power analysis can provide some information about statistical tests in which the null hypothesis was not rejected. Although we illustrate power analysis using parametric statistical methods that focus on evaluating treatment effects (changes in parameters due to a treatment) under the requisite assumptions (sample independence, homogeneity of variance, normally distributed errors), the issues we discuss are relevant to all statistical approaches.

A note on "accepting" null hypotheses.—Hypothesis testing is based on rejecting null hypotheses with a predetermined degree of confidence. When a null hypothesis is not rejected, it is not then appropriate to conclude the null hypothesis to be true (i.e., "accept" the null hypothesis). In practice, however, there are circumstances when it is necessary to decide if a null hypothesis can be considered true. These practical concerns are often why researchers perform retrospective power analyses, and why journal editors request they be reported. We stress, however, that experiments are not designed to prove null hypotheses true; therefore, accepting a null hypothesis as true can never be performed with the same scientific rigor as rejecting a null hypothesis as false. Hence, when we suggest it reasonable to accept a null hypothesis as true, we imply only that the available evidence suggests that, given an established confidence level, the size of the effect observed was too small to be considered of management or biological significance. This point is important because an effect of any size is detectable—no matter how small—if sample sizes are large enough (Johnson 1995).

PROSPECTIVE POWER ANALYSIS

Power Analysis in Research Planning

When a research effort is being planned, power analysis should be used to determine the sample sizes necessary to achieve acceptably high power, or to determine the probability that an effect size of interest will be detected with a certain sample size (Peterman 1990). Determining power prospectively requires that sample size, α, and a biologically meaningful effect size be established. Power then can be computed with a range of values for each of these parameters and for different experimental designs, yielding a series of power curves that indicate the influence each of these parameters has on the statistical power of the planned research effort.

To use power analysis, a study must be designed to detect a particular effect (or effect size). This effect is often the minimum value considered to be of biological or management significance. Typically, researchers prefer to relegate determination of significance to the results of statistical hypothesis tests—if a statistical test results in a significant *P*-value, the result is then considered "significant." This approach is unacceptable, however, because statistical significance and biological significance are not synonymous (Tacha et al. 1982, Yoccoz 1991). Biologically trivial differences may be statistically significant if sample sizes are large and power is high, and biologically important differences may not be statistically significant if power is low (Johnson 1995).

The expected sample variance (σ^2) or coefficient of variation (σ/μ) necessary for prospective power analyses often can be estimated from previous studies. Estimates of these quantities often can be obtained from prior research that was similar to the planned study but was performed in other geographic regions or with other, related taxa. Alternatively, estimates can be obtained from a pilot study. If no previously collected data are available, then a range of probable values can be used, and power curves generated for the likely range of values.

An example of prospective power analysis to determine sample size.—We considered conducting a study to investigate responses of bird populations to snags created in managed forests in Oregon. Abundance of cavity-nesting birds would be estimated with auditory and visual counts at fixed circular plots both before and after snags were created.

We determined statistical power resulting from potential changes in population sizes for the 4 most abundant species that nest in snags in this area, hairy woodpeckers (*Picoides villosus*), brown creepers (*Certhia familiaris*), chestnut-backed chickadees (*Parus rufescens*), and red-breasted nuthatches (*Sitta canadensis*). For each species we estimated power to detect 50, 100, 150, and 200% increases in abundance, using 3–9 replicates of control and treatment stands—the range of replicates that was logistically and economically feasible. We determined power for a repeated-measures analysis of variance, and for this example, set α = 0.05. For variance estimates, we used those reported by Hagar (1992), who collected data in similar habitat using similar techniques in a nearby geographic area.

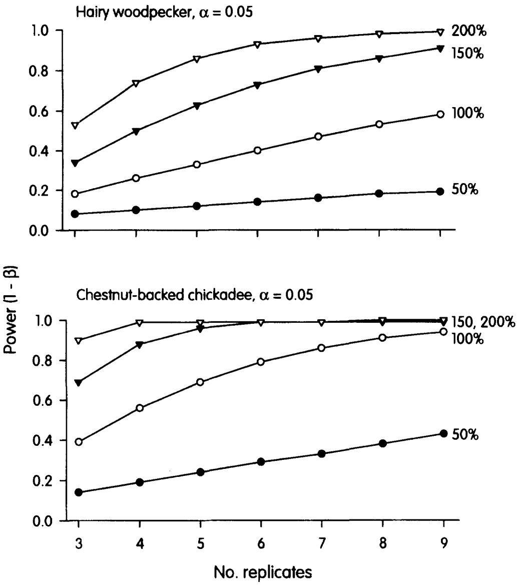

Fig. 3. Results of a prospective power analysis to detect increases in abundance of 50, 100, 150, and 200% for 2 bird species in Oregon, based on a repeated-measures design, 2-sided, with 3–9 replicates, and α = 0.05.

Our analyses yielded power curves that were generally similar for all species, which indicated that power to detect 50% population increases was low (<0.45) for any number of replicates considered (Fig. 3). Power to detect a 100% increase was acceptable for chestnut-backed chickadees with 8 or 9 replicates (>90%), marginal for brown creepers (≈0.80), but low for the other species (<0.70). Not until effects reached 150 or 200% did power become acceptable (>0.80) for all species, and then only with the highest number of replicates considered feasible. In scenarios such as these, low power does not guarantee that statistically significant results would not be obtained (or vice versa), only that the probability of detecting statistically significant effects will be low. Further, it is possible that variance estimates obtained from the actual study might differ from those used in calculations, resulting in a difference between true and estimated power. Nevertheless, these analyses indicate a low probability of obtaining conclusive results within the range of feasible sample sizes.

Research Design and Power

Many of the choices made when an experiment or survey is being planned influence the power of the research effort. These choices include the range of treatment levels selected, the number and type of experimental units chosen, and how treatments are assigned to experimental units (design; Kuehl 1994). A principal objective in research design should be to maximize efficiency by decreasing experimental error and increasing precision of parameter estimates. Any technique that reduces error will increase statistical power—design is a most important mechanism by which to accomplish this objective.

Even if the maximum number of replicates that can be used is constrained by cost or logistics, the range of treatment levels dictated by study objectives (applying a wider range of treatments could increase the effect and thereby increase power), and the techniques for measuring response variables established, statistical power for a given research effort can often be increased by (1) establishing homogeneous blocks of experimental units, (2) measuring concomitant information, and (3) selecting an efficient experimental design—the manner in which treatments are assigned to experimental units. These and other techniques for decreasing experimental error variance, increasing precision, and therefore increasing statistical power are discussed in texts on research design (Kuehl 1994).

An example will illustrate the gains in power when an efficient experimental design and appropriate statistical model for analysis are used. The effect of recreation on breeding bald eagles (*Haliaeetus leucocephalus*) was investigated by measuring brooding behavior of eagles with people camped at distances of 500 and 100 m from nests (Steidl 1995). Assuming these data were collected with a completely randomized design, the null hypothesis of no difference in the percent day that eagles spent brooding with people camped at these 2 distances could not be rejected at any reasonable α-level with a 2-tailed t-test for independent samples (t = 0.54, 52 df, P = 0.59, observed effect = 4.5%, SE = 4.1). However, power to detect a 20% effect with this design and α = 0.10 was low (0.20), indicating that the results were inconclusive. Eagle nesting behavior changes rapidly as nestlings mature (Steidl 1995), and a completely randomized design did not account for this known source of variability. Instead, a crossover design was used (Jones and Kenward 1989), where both treatment and control were applied in succession to the same experimental unit (nest). This design eliminated variability due to nestling age between nests. The null hypothesis of no difference in behavior between distances was rejected with this approach (t = 2.19, 26 df, P = 0.038), indicating that eagle behavior changed when people camped near nests.

This example illustrates how choice of research design can increase precision and therefore statistical power: the pooled standard deviation for the completely randomized design (29.8) was nearly 3 times as high as the standard deviation for the paired design (10.7), even though sample size for the crossover design was half that of the completely randomized design. Further, this example also illustrates the importance of using a statistical model that is consistent with the research design. Here, the power gained by using an appropriate statistical model for analysis changed the study's conclusion.

RETROSPECTIVE POWER ANALYSIS

When a null hypothesis is not rejected, it has become an increasingly common practice to inquire about the power of the statistical test. This additional information is sought to help distinguish between failing to reject a null hypothesis that was actually true (i.e., no real effect existed), and incorrectly failing to reject a null hypothesis that was actually false (a Type II error was made). If a null hypothesis was not rejected, but the estimated power of the test was high (for the min. biologically significant effect), we might infer that there was no biologically significant effect and contend the null hypothesis to be true. If estimated power was low, however, we would consider the test to be inconclusive. Unfortunately, power has often been estimated incorrectly for null hypotheses not rejected (Hayes and Steidl 1997).

Power estimated with the data used to test the null hypothesis and the observed effect size is meaningless. These retrospective analyses yield no information beyond that provided by the original hypothesis test because both power estimated in this way and the P-value of the statistical test are determined by sample size, α, and the observed effect size. Consequently, power incorrectly estimated this way and the P-value for the test are completely confounded: a hypothesis test that yields a high P-value will always have low estimated power and vice versa (Fig. 4). However, there is no relation between the observed P-value for a hypothesis test that was not rejected and true power. Further, the

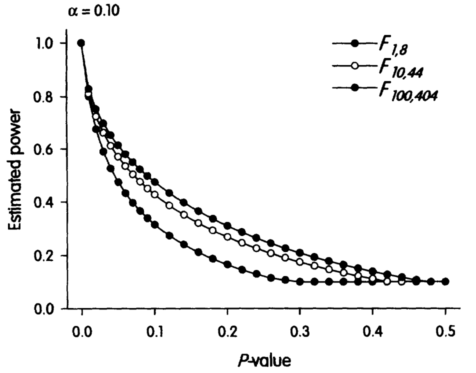

Fig. 4. Relation between estimated power and *P*-value for *F*-tests with different degrees of freedom, incorrectly calculated with observed effects, and $\alpha = 0.10$. There is no relation between the *P*-value for a hypothesis test that was not rejected and true power.

estimated power of any hypothesis test not rejected, properly calculated but based on the observed effect size, will never exceed 0.5 (K. P. Burnham, Colo. Coop. Fish. Wildl. Res. Unit, pers. comm.). Hence, retrospective estimates of power calculated with the observed effect size provide no information about null hypotheses that are not rejected.

Retrospective power analyses, however, can be useful in other circumstances. For statistical tests that do not reject the null hypothesis, retrospective power estimates are meaningful if calculated with effect sizes other than the observed effect size (i.e., under a different alternative hypothesis). Let us assume, for example, that for a particular study a treatment will be considered biologically significant if its application yields an effect size of ≥25. Data are collected and the null hypothesis of zero effect is tested and is not rejected. With the data collected, power could be estimated correctly for an effect size of 25 (not the effect size observed with the data). This result will correctly answer the question "What was the estimated power of this study to detect an effect size of 25?" We note, however, that true power (an unknown parameter) always remains unknown and is only estimated with the data available.

Retrospective power analyses can be used to estimate the effect size or sample size that would have been necessary for a study to achieve a particular level of power. For example, with data already collected, the effect or sample sizes that would have been necessary to achieve 80% power can be estimated. The effect size necessary to achieve acceptable power has been called the detectable effect size (Rotenberry and Wiens 1985). After an experiment has been completed, all the components necessary for calculating the necessary effect or sample size have been amassed and determining their values is relatively simple (Cohen 1988). Note, however, that if variance could have been estimated, and sample size and α-level set, then the above retrospective power analyses could (and probably should) have been done *before* any data were actually collected. These are the only retrospective power analyses that we find meaningful.

Results of retrospective power analyses must be interpreted carefully because they answer only specific questions relating to hypothetical scenarios. For example, determining a detectable effect size does not answer the question, "How large an effect might have actually occurred in a study?", and low power to detect a biologically significant effect does not indicate whether or not such an effect actually exists.

Reporting Retrospective Power Analyses

Retrospective power, when estimated for effect sizes other than the observed effect size, can provide information about the potential for Type II errors to be made under a range of alternative hypotheses. However, because power depends on sample size, effect size, and α-level used in calculations, reporting these values is essential for others to evaluate power estimates and to allow power to be compared among different studies of the same phenomenon. When reporting retrospective power estimates, we recommend that researchers report the specified effect size (or effect and variance), α-level, and sample size used in power calculations. Using the above example for bald eagles, we would report the results of the hypothesis test, power, and parameter values used to calculate power ($t = 0.54$, 52 df, $P = 0.59$; power $= 0.20$ for a 20% change in behavior at $\alpha = 0.10$). Note that variance (or CV, SD, SE, etc.) estimates, which are necessary for computing power, should be reported with summary data analyses.

Confidence Intervals In Lieu of Retrospective Power

When power is estimated retrospectively, researchers must recognize that these estimates of true power (again, an unknown parameter) are based only on a single sample, and must be interpreted as such. Confidence intervals, how-

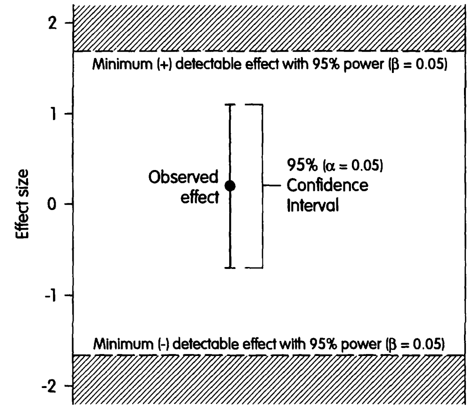
2
1
0
-1
-2
Effect size
Minimum (+) detectable effect with 95% power (β = 0.05)
Observed effect
95% (α = 0.05) Confidence Interval
Minimum (-) detectable effect with 95% power (β = 0.05)

Fig. 5. Range of effect sizes included within the 95% confidence interval for an observed effect size is narrower than the lower limits of effect sizes that are detectable with 95% power (represented by the filled area beyond the dashed lines). In this example, effect size, confidence interval, and power are for a 2-sided *t*-test.

ever, provide a simple, more informative, and preferred alternative to retrospective power analyses. In some fields of research, confidence intervals have been employed as an alternative to significance testing (Greenland 1988, Goodman and Berlin 1994). Similarly, Graybill and Iyer (1994:35) suggest never using hypothesis tests when confidence intervals are available, because confidence intervals are more informative; hypothesis tests, when used alone, can be misleading. Confidence intervals are useful in lieu of retrospective power analyses because the same factors that reduce power, including low sample size, high sample variability, and high α, also increase the width of confidence intervals. Further, confidence intervals provide information about the true size of an effect rather than simply whether or not an effect differed from zero the only information provided by hypothesis tests (The Wildl. Soc. 1995*b*). The range denoted by a 100(1 − α)% confidence interval is narrower than the range between different effect sizes necessary to achieve 100(1 − α)% power (Fig. 5). Therefore, even though a particular study might have lacked the power necessary to detect a specified effect size, the data from that study might indicate that there was low probability that the effect size specified actually existed. Hence, questions about the likely size of true effects can be answered with confidence intervals, not retrospective power analyses.

Although a null hypothesis of "zero effect" is simplest to consider, the null hypothesis of interest is usually whether or not the observed effect was large enough to be considered biologically significant (i.e., affected the system of interest to a degree that merits concern). By expanding null hypotheses beyond the strict statistical sense of zero effect to include all effects that are not biologically significant, it is then possible to evaluate if there is ample evidence to consider the null hypothesis to be true (i.e., to state that the treatment had no biologically significant effect). In general, a confidence interval for the observed effect provides the information necessary to assess whether a null hypothesis can be accepted reliably (The Wildl. Soc. 1995*a*). Using confidence intervals to evaluate the null hypothesis that "the treatment has no biologically significant effect on the parameter of interest" is one approach used in tests of bioequivalence (Metzler 1974, Westlake 1976). Bioequivalence tests, originally developed in pharmacology and gaining increased use in ecological research (Dixon and Garrett 1994, Erickson and McDonald 1995), have also been developed formally for *t*-tests (Hauck and Anderson 1984) and 2 × 2 contingency tables (Dunnett and Gent 1977).

When a null hypothesis is not rejected at some α, the 100(1 − α)% confidence interval for the observed effect always includes values indicating zero effect (e.g., 0 for comparisons of means, 1 for odds ratios), but also denotes the entire range of hypothesized effects that could not be rejected given the available data. Therefore, you can conclude, with 1 − α confidence, that the true effect lies within the range specified by the confidence interval. If the minimum biologically significant effect lies outside the 100(1 − α)% confidence interval for the observed effect, then it is reasonable to conclude the null hypothesis to be true at the specified α-level (Fig. 6). This approach is equivalent to rejecting the null hypothesis that a biologically significant effect occurred. If a portion of the confidence interval for the observed effect includes values considered biologically significant, then the null hypothesis should not be accepted as true and the results should be considered inconclusive. The wider the confidence interval, the more likely it is to include biologically significant effects, rendering the test inconclusive (Fig. 6). In summary, a null hypothesis of no biologically significant effect should be consid-

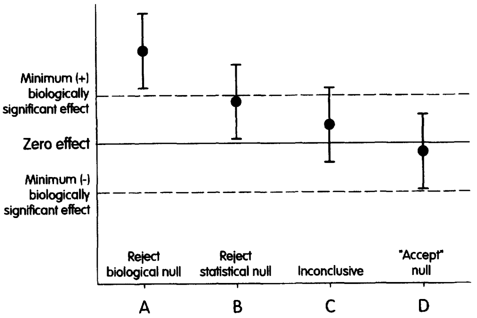
Minimum (+) biologically significant effect
Zero effect
Minimum (-) biologically significant effect
Reject biological null
Reject statistical null
Inconclusive
"Accept" null
A
B
C
D

Fig. 6. Hypothetical observed effects (circles) and their associated 100(1 − α)% confidence intervals. The solid line represents zero effect and dashed lines represent minimum biologically significant effects. In case A, the confidence interval for the estimated effect excludes zero effect and includes only biologically significant effects, so the study is both statistically and biologically significant. In case B, the confidence interval excludes zero effect, so the study is statistically significant; however, the confidence interval also includes values below those thought to be biologically significant, so the study is inconclusive biologically. In case C, the confidence interval includes zero effect and biologically significant effects, so the study is both statistically and biologically inconclusive. In case D, the confidence interval includes zero effect but excludes all effects considered biologically significant, so the "practical" null hypothesis of no biologically significant effect can be accepted with 100(1 − α)% confidence.

ered true only when all biologically significant effects lie outside the confidence interval for the observed effect (Fig. 6).

We illustrate this use of confidence intervals with the above example of bald eagles and a completely randomized design, with $\alpha = 0.10$. We will consider a biologically significant effect as one where the percent day spent brooding changed by 20% between treatment and control distances (control $\bar{x} = 32.6\%$, therefore a 20% change = ±6.6%). The observed effect was 4.5% (SE = 4.1), and the 90% confidence interval for the observed effect (−3.64—12.60%) includes the value for a 20% effect (6.6%). The confidence interval for the observed effect includes values indicating a biologically significant effect; therefore, the null hypothesis should not be accepted as true and the results should be considered inconclusive. However, if instead the 90% confidence interval for the observed effect was 2.98—5.98%, the statistical null hypothesis of zero effect would be rejected, but because this confidence interval did not include the value indicating a biologically significant effect (6.6%), you could conclude the null hypothesis of no biologically significant effect to be true with 90% confidence.

METHODS TO DETERMINE POWER

For many common statistical procedures and experimental designs, tables of power values have been published for a range of effect sizes and α-levels (Tiku 1967, 1972; Beyer 1982, Kraemer and Thiemann 1987, Cohen 1988). Further, dozens of software packages that perform power analyses have been developed recently, many of which provide power estimates for a broader range of statistical procedures than are available in published tables (e.g., Borenstein and Cohen 1988, Hintz 1996). We have found these packages to be a useful means of incorporating power into research planning and analysis. However, power tables or current software are not readily available for several statistical procedures and circumstances. In these instances, efforts required to calculate power range from relatively simple to challenging. Most comprehensive statistical software packages (e.g., SAS, SPSS, S+) include a range of functions for many common statistical distributions that can be used to calculate power. Further, Monte Carlo procedures can be used to generate power estimates (Peterman 1990), especially for nonparametric statistical methods. In instances where retrospective power cannot be readily determined, and a prospective power analysis was not done, we recommend that confidence intervals be used to increase the information about hypothesis tests that are not rejected.

Published tables and software packages function as if the values input for effect and variance are hypothetical parameters rather than estimates (i.e., $\mu_T - \mu_C$ rather than $\bar{x}_T - \bar{x}_C$, σ^2 rather than s^2). Effect, variance, and sample size are then combined into a noncentrality parameter, λ, which is a measure of the overall treatment effects in a study, whose form depends on the research design used. When estimates of these parameters are obtained from data in retrospective power calculations, λ is estimated with the original test statistic (e.g., *F*-ratio). This can result in biased estimates of λ which tend to overestimate true power.

CONSEQUENCES OF TYPE I AND TYPE II ERRORS

By setting α at some predetermined level, such as the canonical $\alpha = 0.05$, scientists are making a de facto choice as to the relative importance of Type I and Type II errors, because β increases as α is reduced. Decreasing α can increase β to an unacceptably high level and consequently reduce power to an unacceptably low level. In many circumstances, such as when the costs of environmental effects could be great, the potential risks and consequences associated with making a Type II error may outweigh those associated with Type I errors (Toft and Shea 1983, Hayes 1987, Peterman 1990).

The burden of proof is typically on researchers to "prove" a phenomenon exists by rejecting the null hypothesis that the phenomenon does not exist. This approach implies a willingness to accept the consequences of Type II errors over those of Type I errors. In some situations this approach may be appropriate. However, when there are considerable risks associated with management actions based on the results of hypothesis tests that are not rejected, the consequences of Type II errors can exceed those of Type I errors (Peterman 1990). For example, in the Pacific Northwest, there is a question as to the amount of timber that can be harvested without adverse effects on songbird populations. A relevant null hypothesis might be that a particular level of timber harvest has no effect on the density of songbird populations. In this and similar instances, the null hypothesis might be stated as one of no effect. If an experiment with low statistical power is performed to test this hypothesis, the probability of rejecting the null hypothesis will be low, whether or not the true effect was biologically significant. If songbird populations were adversely affected by a certain level of timber harvest, but forests continued to be managed as if songbirds were not affected because of decisions based on low-power tests, then this Type II error could lead to population declines.

Management actions resulting from hypothesis tests that were not rejected have an underlying, often unrecognized, assumption about the relative costs of Type I and Type II errors that is independent of their true costs (Toft and Shea 1983, Cohen 1988, Peterman 1990). In particular, when $\beta \geq \alpha$, scientists implicitly assume that costs of Type I errors exceed those of Type II errors when their recommendations assume that a null hypothesis that is not rejected is true (Toft and Shea 1983). One approach suggests considering Type II errors as paramount when a decision would result in the loss of unique habitats or species (Shrader-Frechette and McCoy 1993). Other approaches have been suggested by which Type I and II errors can be balanced based on their relative costs (Osenberg et al. 1994).

In general, the framework of hypothesis testing has been largely overused by scientists (Salsburg 1985, Yoccoz 1991), especially in the context of environmental decision making. Hypothesis tests only assess "statistical significance"; "practical importance" may be better evaluated by the use of confidence intervals (Graybill and Iyer 1994:xiii). Reliance on hypothesis testing should be decreased in favor of more informative methods that better evaluate available information, including Bayesian methods (Ellison 1996). In circumstances similar to those outlined above for timber and songbirds, the relevant issue is not whether timber harvest affects songbirds (obviously, there will be an effect on resident songbirds if all timber is cut); rather, the issue is to understand the magnitude of the effect caused by a particular level of harvest. Hypothesis testing should not be the only tool used for decision-making issues, especially where the risk associated with a decision is considerable. In these instances, knowledge of the potential risks and available evidence for each possible decision should guide the decision-making progress.

LITERATURE CITED

BEIER, P., AND S. C. CUNNINGHAM. 1996. Power of track surveys to detect changes in cougar populations. Wildl. Soc. Bull. 24:540–546.

BEYER, W. H. 1982. CRC handbook of tables for probability and statistics. Chem. Rubber Co., Cleveland, Oh. 656pp.

BORENSTEIN, M., AND J. COHEN. 1988. Statistical power analysis. Lawrence Erlbaum Assoc., Hillsdale, N.J. 187pp.

COHEN, J. 1988. Statistical power analysis for the behavioral sciences. Second ed. Lawrence Erlbaum Assoc., Hillsdale, N.J. 567pp.

DUNNETT, C. W., AND M. GENT. 1977. Significance testing to establish equivalence between treatments with special reference to data in the form of 2 × 2 tables. Biometrics 33:593–602.

DIXON, P. M., AND K. A. GARRETT. 1994. Statistical issues for field experimenters. Pages 439–450 *in* R. J. Kendall and T. E. Lacher, Jr., eds. Wildlife toxicology and population modeling: integrated studies of agroecosystems. CRC Press, Boca Raton, Fla.

ELLISON, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Appl. 6:1036–1046.

ERICKSON, W. P., AND L. L. MCDONALD. 1995.

Tests for bioequivalence of control media and test media in studies of toxicity. Environ. Toxicol. Chem. 14:1247–1256.

GERRODETTE, T. 1987. A power analysis for detecting trends. Ecology 68:1364–1372.

GOODMAN, S. N., AND J. A. BERLIN. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annu. Intern. Med. 121:200–206.

GRAYBILL, F. A., AND H. K. IYER. 1994. Regression analysis: concepts and applications. Duxbury Press, Belmont, Calif. 701pp.

GREENLAND, S. 1988. On sample-size and power calculations for studies using confidence intervals. Am. J. Epidemiol. 128:231–237.

HAGAR, J. C. 1992. Bird communities in commercially thinned and unthinned Douglas-fir stands of western Oregon. M.S. Thesis, Oregon State Univ., Corvallis. 110pp.

HATFIELD, J. S., W. R. GOULD, B. A. HOOVER, M. R. FULLER, AND E. L. LINDQUIST. 1996. Detecting trends in raptor counts: power and Type I error rates of various statistical tests. Wildl. Soc. Bull. 24:505–515.

HAUCK, W. W., AND S. ANDERSON. 1984. A new statistical procedure for testing equivalence in two-group comparative bioavailability studies. J. Pharmacol. Biopharmacol. 12:83–91.

HAYES, J. P. 1987. The positive approach to negative results in toxicology studies. Ecotoxicol. and Environ. Safety 14: 73–77.

———, AND R. J. STEIDL. 1997. Statistical power analysis and amphibian population trends. Conserv. Biol. 11:273–275.

HINTZ, J. 1996. Power analysis and sample size user's guide. NCSS, Kaysville, Ut. 245pp.

JOHNSON, D. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

JONES, B., AND M. G. KENWARD. 1989. Design and analysis of cross-over trials. Chapman and Hall, New York, N.Y. 340pp.

KRAEMER, H. C., AND S. THIEMANN. 1987. How many subjects? Statistical power analysis in research. Sage Publ. Ltd., Newbury Park, Calif. 120pp.

KUEHL, R. O. 1994. Statistical principles of research design and analysis. Duxbury Press, Belmont, Calif. 686pp.

METZLER, C. M. 1974. Bioavailability—a problem in equivalence. Biometrics 30:309–317.

OSENBERG, C. W., R. J. SCHMITT, S. J. HOLBROOK, K. E. ABU-SABA, AND A. R. FLEGAL. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecol. Appl. 4:16–30.

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

———, AND M. J. BRADFORD. 1987. Statistical power of trends in fish abundance. Can. J. Fish. Aquat. Sci. 44:1879–1889.

POPPER, K. R. 1962. Conjectures and refutations. Basic Books, New York, N. Y. 412pp.

RICHARDSON, J. T. E. 1996. Measures of effect size. Behav. Res. Methods Instrum. Comput. 28:12–22.

ROMESBURG, H. C. 1981. Wildlife science: gaining reliable knowledge. J. Wildl. Manage. 45:293–313.

ROTENBERRY, J. T., AND J. A. WIENS. 1985. Statistical power and community-wide patterns. Am. Nat. 125:164–168.

SALSBURG, D. S. 1985. The religion of statistics as practiced in medical journals. Am. Stat. 39:220–223.

SEARCY-BERNAL, R. 1994. Statistical power and aquacultural research. Aquaculture 127:371–388.

SHRADER-FRECHETTE, K. S., AND E. D. MCCOY. 1993. Method in ecology. Cambridge Univ. Press, U.K. 328pp.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry. Second ed. Freeman and Co., San Francisco, Calif. 859pp.

SOLOW, A. R., AND J. H. STEELE. 1990. On sample size, statistical power, and the detection of density dependence. J. Anim. Ecol. 59:1073–1076.

STEIDL, R. J. 1995. Human impacts on the ecology of bald eagles in interior Alaska. Ph.D. Thesis, Oregon State Univ., Corvallis. 155pp.

TACHA, T. C., W. D. WARDE, AND K. P. BURNHAM. 1982. Use and interpretation of statistics in wildlife journals. Wildl. Soc. Bull. 10:355–362.

TANG, P. C. 1938. The power function of the analysis of variance tests with tables and applications of their use. Stat. Res. Memoirs 2:126–149.

TAYLOR, B. L., AND T. GERRODETTE. 1993. The uses of statistical power in conservation biology: the vaquita and northern spotted owl. Conserv. Biol. 7:489–500.

TIKU, M. L. 1967. Tables of the power of the *F*-test. J. Am. Stat. Assoc. 62:525–539.

———. 1972. More tables of the power of the *F*-test. J. Am. Stat. Assoc. 67:709–710.

THE WILDLIFE SOCIETY. 1995*a*. Journal News. J. Wildl. Manage. 59:196–198.

———. 1995*b*. Journal News. J. Wildl. Manage. 59:630.

TOFT, C. A., AND P. J. SHEA. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am. Nat. 122:618–625.

WESTLAKE, W. J. 1976. Symmetrical confidence intervals for bioequivalence trials. Biometrics 32:741–744.

YOCCOZ, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bull. Ecol. Soc. Am. 72:106–111.

Received 8 April 1996.
Accepted 18 November 1996.
Associate Editor: McDonald.