

•

# Assessing the Reliability of Ecological Monitoring Data: Power Analysis and Alternative Approaches

Lloyd W. Morrison<sup>1</sup>

Heartland Inventory and  
Monitoring Network  
National Park Service  
and

Department of Biology  
Missouri State University  
901 S. National Avenue  
Springfield, MO 65897

•

<sup>1</sup> Corresponding author:  
LloydMorrison@MissouriState.edu;  
417-836-3119

**ABSTRACT:** To identify natural resources in need of conservation, and assess the effectiveness of ongoing management practices, a 'reliable' monitoring program is necessary. It is critical to assess the reliability of our data, and our data analyses, so that we draw the appropriate conclusions regarding the natural resource of interest. One way to evaluate this reliability is through the use of statistical power analysis. Although power analysis may provide valuable insights into the design and results of a study or monitoring program, its misuse may lead to inappropriate conclusions and management actions. This review describes the appropriate use of statistical power analysis in the context of natural areas management, and points out numerous misuses, some of which are not widely recognized. Alternative approaches to traditional power analyses are presented, along with a discussion of their advantages and disadvantages.

*Index terms:* confidence intervals, null hypothesis significance testing, parameter estimation, power analysis, Type II error

## INTRODUCTION

As the world's natural resources are degraded and destroyed, monitoring the health of remaining resources becomes increasingly important. Evaluating the ecological integrity of natural ecosystems has assumed a critical role in the activities of many university researchers, private foundations, and governmental agencies concerned with conservation. A critical issue in monitoring ecological resources is the reliability of the resulting data. All monitoring programs face practical, financial, and logistical constraints, which restrict the amount of information obtained. Given the limited data available, and the amount of inherent variability that characterizes most natural systems, an important question is: How likely are we to detect important changes when such changes exist? Or, perhaps the more relevant question is: How likely are we to fail to detect important changes when such changes are in fact occurring? The reliability of our data, in statistical terms, when hypotheses are tested, is the subject of power analysis.

Over the past two decades, ecologists have been extolled to use power analysis, both when planning studies and interpreting results (e.g., Toft and Shea 1983; Peterman 1990a, b; Thomas and Juanes 1996; Stefano 2001). This review focuses on problems with power analyses in the evaluation of ecological monitoring data, and discusses potential alternative approaches. It has critical implications for how we should approach data analysis and, ultimately, how we should manage natural resources.

## ELEMENTS OF POWER

Statistical power is the probability that a null hypothesis will be rejected if the null hypothesis is in fact false. It is the complement to Type II error ( $\beta$ ). Discussions of the relationships between power and Type I and Type II errors can be found in Toft and Shea (1983) and Peterman (1990a). Statistical power increases with higher values of  $\alpha$  (the probability of making a Type I error), larger sample sizes, and greater effect sizes. The effect size includes the magnitude of difference in the parameter of interest, as well as the variance of the measurements. The variance of the measurements includes both natural variability and measurement error. Thus, for the components comprising effect size, power increases with an increasing magnitude of difference and decreases with increasing natural variability and increasing measurement error. Measurement error is not frequently mentioned in discussions of power (and not easily identified once data are collected), yet anything that can be done to make sampling more accurate will increase power (Steidl et al. 1997; Stoehr 1999; Lenth 2001). If effect size is a standardized measure (e.g., the correlation coefficient,  $r$ ), it is dimensionless, and there is no associated sample variance, although measurement error will still have an influence (Jennions and Moller 2003).

Statistical power does not exist outside of formal hypothesis testing; conversely, it is always a facet of such testing, although it is frequently ignored (e.g., Yoccoz 1991). Power is usually desired to be at least 0.8 ( $\beta = 0.2$ ) (Cohen 1988; Stefano 2003), although values of power as low as 0.5 are

---

sometimes considered acceptable (Murphy and Myers 2004). In practice, it is often difficult to obtain power much greater than 0.8 (i.e., a very large sample size is necessary).

There are two types of power analyses: prospective and retrospective. Prospective power analysis is conducted before a study is initiated, most commonly to determine the sample size necessary to achieve a target power level. It may also be used to evaluate competing sample designs. Retrospective power analysis is conducted after a study is complete, and is only relevant when one has failed to reject the null hypothesis. Retrospective power analysis has frequently been employed to determine the likelihood that the null hypothesis would have been rejected if it was false, given the parameters of the study. This use of retrospective power analysis is invalid, however (see below). An important, yet frequently overlooked point is that prospective and retrospective power for the same study are not necessarily equivalent (see Zumbo and Hubley 1998).

It is possible to make a third type of error, termed a Type III error (Leventhal and Huynh 1996). This occurs when one rejects the null hypothesis and concludes the direction of difference in the population is the same as that in the sample, even though the alternative hypothesis is nondirectional. The actual difference in the population may be in the opposite direction, however, in which case a Type III error has been made. The existence of this potential third type of error has led to a revised definition of power, which is the conditional probability of rejecting the null hypothesis and correctly identifying the true direction of difference (Leventhal and Huynh 1996). When conducting such tests, power will be less using the revised definition than using the traditional definition. The likelihood of making such an error, however, must be very small in most cases .

#### **RELEVANCE TO HYPOTHESIS TESTING**

Power analysis is inseparably linked with null hypothesis significance testing. The

origin of employing a preset significance threshold to make a binary decision regarding a null hypothesis stems from Neyman and Pearson's work, developed as a sort of quality control for determining acceptable rates of defective products in cost-benefit analyses (e.g., Neyman and Pearson 1967). The purpose of the Neyman-Pearson approach was not to inform belief, but to establish rules for behavior, in which the outcome was some practical yes-no decision (Oakes 1990). In fact, Neyman and Pearson's original papers contained no application in which a scientific hypothesis was of primary interest (Birnbaum 1977). The other major school of thought influencing modern frequentist statistical concepts was the Fisherian approach, in which *P*-values were emphasized, rather than some specific significance threshold. Fisher (1955) rejected Neyman and Pearson's emphasis on power.

Modern hybrid statistical theory combines elements of both the Fisher and Neyman-Pearson schools of thought, at times supplemented with a somewhat Bayesian interpretation of what significance means (Gigerenzer et al. 1989). Yet, this attempt to fuse opposing theories into a single one has generated much confusion regarding even basic statistical concepts (Sedlmeier and Gigerenzer 1989) .

Statistical power must be interpreted relative to the null hypothesis under test. The null hypothesis is usually one of no difference or no effect (also called the nil hypothesis; Cohen 1994). The alternate hypothesis is usually that there exists a non-zero difference or effect, but the magnitude of this effect is not specified. The problem is that a null hypothesis of no difference is usually trivial or known to be false before initiation of a study. This is especially true in ecological situations. No two populations are likely to be exactly the same when compared across space or over time. Given a large enough sample size, even a very small difference can be detected (Utts 1988; McBride et al. 1993; Johnson 1999; Anderson et al. 2000).

This has led some to state, rather cynically, yet accurately, that failure to reject a null hypothesis does not mean no difference

actually exists, but simply that the sample size was too small. Thus, most tests of null hypotheses are actually tests of whether the sample size was large enough (Johnson 1999). This process of testing what is already known has been referred to as "gratuitous" significance testing (Abelson 1997), and such null hypotheses have been referred to as "silly" (Robinson and Wainer 2002). In practice, significance tests are often not taken seriously (Guttman 1985). The real issue is not whether there is any relationship among the variables under study, but the magnitude of the relationship .

#### **PROBLEMS, MISUSES, AND ABUSES**

Throughout this section, I refrain from identifying authors who have misused or abused power analyses in some way. Examples of such misuse are not difficult to find in the literature.

#### **Statistical assumptions and estimation**

Power analyses, like statistical tests in general, are only valid if the appropriate statistical assumptions are met. These assumptions are frequently not met in environmental studies, however, which are notorious for non-normality, heterogeneous error structure, over-dispersion, and spatial-temporal dependency (Fox 2001). Many field studies suffer from pseudoreplication and nonrandomization (Hurlbert 1984; Suter 1996). Hypothesis testing was designed for experimental rather than observational studies.

All valid power analyses require the specification of possible parameter values, which necessarily involves a subjective aspect and includes the possibility of bias. Power analyses for more complex statistical tests (e.g., repeated measures ANOVA) require additional assumptions and the estimation of multiple parameters.

#### **Prospective power analysis**

The primary problem with prospective power analysis is that one must specify

---

the effect size prior to data collection. This quantity is frequently unknown, however, and is usually the reason for gathering the data; if the researcher knew the effect size, he would probably not be conducting the study. This can be a serious problem because power may be very sensitive to small changes in estimated effect size (Rotenberry and Wiens 1985). Even when investigators are able to accurately determine a priori the amount of power associated with their study, they may be able to do little beforehand to change that power level because of financial or logistical constraints.

### Retrospective power analysis

In retrospective power analysis, the focus is usually on determining the power of a test when the null hypothesis was not rejected. The estimated effect size obtained from the data should not be used to calculate statistical power after the analysis, however (Goodman and Berlin 1994; Hayes and Steidl 1997; Steidl et al. 1997; Thomas 1997; Gerard et al. 1998; Hoenig and Heisey 2001; Lenth 2001; Nakagawa and Foster 2004). Statistical power refers to the pre-experiment probability of obtaining a hypothetical group of results; it is not a property of a particular data set (Greenland 1988; Goodman and Berlin 1994).

If a large  $P$ -value is obtained in a study, calculations using the observed effect size will always indicate low power. Observed power and the  $P$ -value are both dependent upon the observed effect size (Nakagawa and Foster 2004). Once one has failed to reject the null hypothesis, power calculated from the observed effect size adds nothing to the interpretation of results (Hoenig and Heisey 2001). This problem is not widely appreciated, and many canned software programs use the effect size determined from the sample to do retrospective power analyses, even though it is inappropriate (Thomas and Krebs 1997).

The appropriate use of retrospective power analysis presents a conundrum to the experimenter: one cannot use the best estimate of effect size, which is arguably that obtained from the data at hand, but

the effect size must be estimated based on some other information. Because of this, some have recommended that retrospective power analyses simply never be done (Goodman and Berlin 1994; Gerard et al. 1998). Others concede retrospective power analysis may have "extremely limited" applications, perhaps only in 'meta-analyses' (Hoenig and Heisey 2001). For example, a recent meta-analytic survey of journals focusing on animal behavior by Jennions and Moller (2003) revealed statistical power to be quite low: 0.13-0.16 to detect a 'small' effect and 0.40-0.47 to detect a 'medium' effect ('small' and 'medium' effects as described by Cohen 1988).

One possible solution to the subjectivity involved in estimating effect size is to use approximations obtained from other studies of similar taxa or habitats, or derived from a survey of multiple studies. As an example of the latter, Gibbs et al. (1998) present a table of variability estimates for various plant and animal groups derived from several hundred studies.

Unbiased retrospective power analysis may also be possible by estimating the effect size that would have been necessary for a study to achieve a particular level of power (Steidl et al. 1997; Gerard et al. 1998). This type of power analysis could, and probably should, however, be done in the planning stages of a study.

### Interpretation of the effect size

Some power analyses may be misinterpreted, in that the effect size is perceived as the actual amount of change that can be detected in a population. For example, one may conduct a power analysis and determine that an effect size of 20% can be detected, for a given  $\alpha$  and  $n$ , and with a particular level of power. It would not be appropriate, however, to conclude that a difference of 20% exists in the population if the null hypothesis is one of no difference. One would have to conclude, at the specified level of power, that there either is or is not a nonzero difference in the population of interest (i.e., reject or fail to reject the null hypothesis).

The estimated effect size represents the amount of change or difference in the parameter of interest as a function of the variability *in the sample*, which will allow one to reject the null hypothesis given the levels of  $\alpha$  and  $n$ . It does not represent the magnitude of change or difference in the population. This is clearly evident when one considers that the magnitude of difference in the population is absolute, whereas the effect size in the analysis, for a given level of power, will vary depending upon  $\alpha$  and  $n$  (i.e., with a larger sample size, one could detect a smaller effect size).

### Length of time and effect size necessary to obtain desirable power

A related issue is that power analyses conducted in relation to long-term monitoring programs are often described in terms of the amount of annual change that can be detected over a multi-year period (e.g., a 5% annual rate of change over 10 years). Yet, populations do not change at a constant rate. What these analyses actually measure is the *net* change in population size (e.g., a ~50% net change). Thus, even if change was gradual (e.g., 5% per year), one would not be able to detect a difference after four or five years. Depending upon the methodology used, either a constant small amount of change occurring each year or a major change occurring in a single year would yield the same results as far as power is concerned. Ecologically, however, these scenarios convey very different messages.

Perhaps the most worrisome aspect of many attempts to incorporate power into ecological monitoring programs is that to obtain a desirable level of power, one has to settle for a relatively large effect size, and this may only occur after a considerable period of time. Frequently, the ability to detect an effect size of up to 50% within 10 or 20 years, with power of at least 0.8, appears to represent an acceptable convention. If one follows the convoluted logic involved, and accepts the effect size to be equal to the actual change in the population, the conservation implications are cause for alarm. In a scenario in which one were monitoring threatened or endangered

---

species, for example, up to half of the population would be lost before one would be willing to admit that the population was actually declining (at all). Moreover, reversing a 10- or 20-year trend is likely to be much more difficult than reversing a trend of shorter duration. The use of traditional null hypothesis testing along with power analyses in such cases could easily do more harm than good, and could put an already threatened or endangered population at much greater risk than the use of other analytical methods.

### High population variability or measurement error

If natural variability is high, it simply may not be possible to achieve an acceptable power level. An example of the effect of sampling variability on power is provided by de la Mare (1984), who calculated that, because of large sampling variabilities, whale abundance could decrease by 50% over 20 years, yet the power of testing the null hypothesis of no decline would be only 0.31.

Furthermore, the power to detect a decline in abundance will decrease as populations become smaller if the coefficient of variation of the population estimate increases. If a population becomes small enough, the most likely outcome of any survey methodology will be a nonsignificant trend, even when the population is actually declining (Taylor and Gerrodette 1993). Thus, for the species with the smallest populations (and presumably at the greatest risk), if we wait for a statistically significant decline before instituting stronger protective measures, the species will probably go extinct first.

### What should $\alpha$ and $\beta$ be?

The probability of making a Type I error,  $\alpha$ , is usually set by convention to be 0.05, a value considered by many to be arbitrary (e.g., Peterman 1990a; Mapstone 1995; Cherry 1998; Johnson 1999; Stoehr 1999; Anderson et al. 2000; Stefano 2001; Robinson and Wainer 2002). The existing convention for  $\beta$  is 0.20 (Stefano 2003). This implies that a Type I error is four times more serious than a Type II error.

(The fact that power is often not mentioned in many studies implies that frequently a Type II error is not even serious enough to consider!) Traditionally, scientists have guarded against making Type I errors more vehemently than Type II errors because Type I errors have been perceived as more costly. This may be true in pure science, in which accepting a falsehood (committing a Type I error) could lead to much time and money being wasted in investigating a phenomenon that does not exist. The preference for Type I or Type II error varies, however, and traditionally many applied scientists have been more willing to commit Type I than Type II errors (Shrader-Frechette and McCoy 1992).

In the context of an ecological monitoring program, a Type I error might mean that we think a population is in danger of declining when in actuality it is not. The cost would be resources expended in order to mitigate the decline; the population, however, would likely not suffer. If a Type II error was made, however, we might fail to note a real decline in a population. The population could decline to a dangerously low level or even go extinct. The cost would be greater resources necessary to rescue the population from such a low level or, in the extreme case, extinction of the population. Thus, a Type II error may have a much greater cost than a Type I error in ecological monitoring programs.

One way to increase statistical power involves an attempt to balance  $\alpha$  and  $\beta$  in a compromise analysis (Peterman 1990a; Stefano 2001, 2003; Field et al. 2005). If one is able to estimate the cost associated with each error type, it is possible to adjust the ratio of the probability of each type of error to balance the costs. This approach, while logical, will often have the advantage of increased statistical power, as it will usually result in  $\alpha$  being increased. The problem is that it is often difficult to put an objective cost on the two types of error. One may be able to estimate the cost to rescue a declining population, for example, but what is the monetary value of a species that goes extinct?

Decisions regarding the appropriate ratio of  $\alpha$  and  $\beta$  should be made before any data

are collected (Stefano 2001). Field et al. (2004) present a cost function approach for determining optimal  $\alpha$  levels. Mapstone (1995) presents a set of decision rules for setting values of  $\alpha$  and  $\beta$ .

In addition to a priori specification of  $\alpha$ , or a compromise analysis attempting to balance  $\alpha$  and  $\beta$  for a particular application, a third option exists: a priori specification of power. Thus, instead of controlling primarily for Type I error (as is usually done), it is possible to control primarily for power, and use the resulting  $\alpha$  level as determined by the sample size and other parameters of the analysis (as in Lindley et al. 2000). Such an approach would be valuable with threatened or endangered species, in which a Type II error would be associated with extinction, or other cases in which the cost of a Type II error is much greater than the cost of a Type I error.

### Reporting power

True statistical power can never be known exactly; it is a conditional probability, and can only be estimated (Taylor and Muller 1995; Steidl et al. 1997). It is analogous to finding the mean of a large population: we can estimate the mean from a sample, but we do not know the true mean. Although we usually report variances associated with our estimates of mean (i.e., standard deviations, standard errors), rarely is the variability associated with power estimates reported. Yet, confidence intervals of power estimates can be very large (Gerard et al. 1998), and it has been suggested that one should report them whenever power is estimated (Thomas 1997; Hoenig and Heisey 2001). Unlike confidence intervals for other variables, one-sided intervals may frequently be sufficient for power, as one is usually interested in the least amount of power that characterizes a test (Taylor and Muller 1995).

### Statistical vs. Biological Significance

An important consideration in the use of power analysis is the difference between statistical significance and biological significance (Yoccoz 1991; McBride et al. 1993; Reed and Blaustein 1997). A

---

powerful test (large sample size) may detect a very small difference, yet this difference may not be meaningful biologically. Conversely, a test with a small sample size may not yield a significant result even when a large difference exists in the population. As emphasized earlier, statistical significance, and *P*-values, varies with sample size. Biological significance is often subjective.

## ALTERNATIVES TO POWER ANALYSIS

### Why test null hypotheses?

Given the difficulties associated with, and misinterpretations plaguing hypothesis testing and power analysis, why do so many take this approach? Traditional null hypothesis testing has long been the subject of controversy (see Nickerson 2000 for a comprehensive review). For years, statisticians have warned against excessive use of hypothesis tests, and failure to use methods (such as parameter estimation and confidence intervals) that are often simpler and more informative (e.g., Carver 1978, 1993; Guttman 1985; Utts 1988; Yoccoz 1991; Shaver 1993; Cohen 1994; Goodman and Berlin 1994; Kirk 1996; Steidl et al. 1997; Cherry 1998; Gerard et al. 1998; Johnson 1999; Anderson et al. 2000, 2001; Hoenig and Heisey 2001; Colegrave and Ruxton 2003; Nakagawa and Foster 2004). For some unknown reason, ecology has lagged behind other sciences in terms of awareness and discussions of problems associated with null hypothesis significance testing (Anderson et al. 2000).

Perhaps it is the notion that one is not really doing science unless one is testing hypotheses. Modern hypothesis testing has its roots in Popperian inference, which attempts to test hypotheses that can be clearly falsified (Popper 1959), and Platt's (1964) paradigm of strong inference, which emphasizes the formulation of alternate hypotheses. Hypothesis testing is a cornerstone of the hypothetico-deductive scientific method (Johnson 1999). Traditional null hypothesis testing emphasizes making binary decisions regarding some threshold level of significance. This may be necessary for some decision-making processes, but

if our primary goal is to obtain a greater elucidation of a pattern or process, this may not represent the best approach.

### Parameter estimation and confidence intervals

Parameter estimation with confidence intervals may provide more information than hypothesis testing, be more straightforward to interpret, and easier to compute. Parameter estimation emphasizes the magnitude of effects, and the biological significance of the results, rather than making binary decisions (Shaver 1993; Kirk 1996; Stoehr 1999). There is no formal classification of error associated with parameter estimation. It is assumed that the estimate is not accurate, and the width of the confidence interval provides information on the degree of uncertainty (Simberloff 1990). Parameter estimation with confidence intervals is also preferable to hypothesis tests because the former removes the focus from decision-making and allows the reader of a research report to draw his own conclusions (Utts 1988). Narrow confidence intervals are also more resistant to sources of random error than low *P*-values (Poole 2001).

Of course, one may test hypotheses *and* estimate parameters. However, if one first tests a hypothesis and the test results in a rejection decision, and then one proceeds to estimate the parameter, confidence limits should be computed *conditionally* based on the specified outcome of the test (Meeks and D'Agostino 1983). Such conditional confidence intervals will be wider than unconditional confidence intervals computed independently of hypothesis testing. Such conditional estimation is usually discouraged, and confidence intervals conducted in the absence of hypothesis tests are preferred (Meeks and D'Agostino 1983). In fact, parameter estimation makes hypothesis testing unnecessary. Once a confidence interval has been constructed, power calculations generally yield no additional insights (Hoenig and Heisey 2001).

### Non-null hypotheses

It is possible to create a null hypothesis that specifies a value that is biologically

meaningful, rather than using the traditional null hypothesis of no difference or no effect. For example, if a resource manager would be alarmed if a population changed by 20%, one could posit a null hypothesis that the population in question changes by less than 20%. The alternate hypothesis would be that the population changes by 20% or more. In this case, a statistically significant result would also be biologically meaningful. Additionally, one would no longer be testing the nil hypothesis, and it would not likely be obvious beforehand whether the null hypothesis was false. This has been termed 'minimum effect' testing (Murphy and Myers 2004).

Testing such hypotheses, however, requires a different type of sampling distribution than a null hypothesis of no difference or no effect. For example, if the appropriate statistic were an *F*-test, a test for a minimum difference would be based on a non-central *F* distribution, rather than on the central *F* distribution used in testing traditional nil hypotheses. Most tables in statistical texts and most canned software use a sampling distribution consistent with a null hypothesis of no difference or no effect (Murphy and Myers 2004). Additionally, it is more difficult to reject such a null hypothesis: a much larger sample size is necessary to achieve a similar amount of power.

### Equivalence tests

Equivalence tests (sometimes called 'bio-equivalence tests') evaluate interval null hypotheses. This approach is frequently employed in the pharmaceutical industry (e.g., Berger and Hsu 1996), in which emphasis is on the practical similarity of two drugs, rather than an absolute difference (analogous to the difference between biological significance and statistical significance, respectively). In theory, such equivalence tests could be employed in many other fields, replacing traditional null hypothesis tests (McBride et al. 1993; Hauck and Anderson 1996; Hoenig and Heisey 2001; Parkhurst 2001). Yet, there have been very few applications of equivalence testing in the environmental sciences (McBride 1999).

Unfortunately, the mathematics of testing equivalence hypotheses are not straightforward. Many different equivalence tests have been suggested. There is no 'optimal' test, but rather a tradeoff between the Type I error rate, power, and the shape of the rejection region (Dixon and Pechmann 2005). The appropriate tests have not been developed for many potential applications, and controversy exists over which tests should be used (Berger and Hsu 1996; Parkhurst 2001). An equivalence test applicable for trend detection in ecological monitoring programs has recently been proposed, however (Dixon and Pechmann 2005). It should be noted that confidence intervals could also be used to evaluate the types of interval hypotheses associated with equivalence testing (Steiger 2004).

### Reversing the burden of proof

The discussion of power in relation to natural resource conservation begs the question of where the burden of proof should be placed. Traditionally, scientists have usually had to document a statistically significant decline or degradation in the quality of a resource before any conservation actions were put into effect. Yet the sample sizes and sample variability that often characterize ecological studies make rejection of traditional null hypotheses generally difficult (Parkhurst 2001). Using this approach, because of the difficulties of finding 'statistically significant' changes, the resource may decline to a severe extent before anything is done. This is well documented in the field of traditional fisheries management (Peterman and Bradford 1987; Peterman 1990a). One option is to reverse the burden of proof, and require any potential exploiters of natural resources to prove that their activities do not cause damage, rather than assume this is the case until it can be demonstrated otherwise (Peterman and M'Gonigle 1992; Dayton 1998). Fox (2001) introduced a new statistic termed 'environmental power,' which seeks to incorporate a reversal of the burden of proof, and includes a Bayesian aspect.

A related concept is the 'precautionary principle,' developed in the context of

marine pollution, which basically states that potentially damaging pollutants should be reduced or eliminated even when there is no scientific evidence to prove a causal link between emissions and effects (Gray 1990; Johnston and Simmonds 1990; Peterman and M'Gonigle 1992). This principle could apply to the entire spectrum of environmental policy-making, including all human impacts on the environment (Earll 1992; Ortiz 2002).

Another approach to reversing the burden of proof is to conduct 'reverse tests' following the failure to reject a nil hypothesis (Parkhurst 2001). Such a test would allow one to determine whether the data support a conclusion consistent with the null hypothesis being true. A reverse test would serve a similar function as a power analysis, in that it would attempt to differentiate between evidence for lack of an effect versus uncertainty due to factors such as a small sample size or high sample variability. Unfortunately, this approach is not well developed; reverse tests suffer from similar problems as equivalence tests, and have only been applied in a limited number of situations (Parkhurst 2001).

Although reversing the burden of truth has obvious benefits for natural resource protection, scientifically speaking it simply shifts the bias inherent in hypothesis testing from one side to the other (Suter 1996). As scientists, we should strive to eliminate all forms of bias from our analyses.

### The information-theoretic approach

The relatively new information-theoretic approach is an extension of likelihood theory and avoids many of the limitations of statistical hypothesis testing. It focuses on relationships of variables through model selection, as well as estimates of effect size (Burnham and Anderson 1998, 2002; Anderson et al. 2000, 2001). A good set of a priori alternative hypotheses is essential to this approach, however, which requires an in-depth understanding of the biology of the system under study. Frequently, supplying an appropriate and limited set of well-supported models a priori is difficult (Eberhardt 2003). Additionally, if all

models fit poorly, the information-theoretic approach will only select the best of the set of poorly fitting models (Burnham and Anderson 1998). Likelihood also provides no protection against trivial hypotheses (Guthery et al. 2001). Furthermore, researchers using likelihood must make assumptions regarding the probability distributions obtained in nature. If these assumptions are inaccurate, the resulting inferences are unlikely to be useful (Guthery et al. 2001).

### Bayesian statistics

A Bayesian approach (e.g., Ellison 1996; Wade 2000; Dorazio and Johnson 2003) would avoid many of the problems of hypothesis testing and power. Although an overview of Bayesian methods is beyond the scope of this review, it can be noted that they are not widely used because they are often difficult to apply, and many researchers are not comfortable specifying subjective degrees of belief in their hypotheses (Utts 1988). As summed up by Hoenig and Heisey (2001), the "real world of data analysis is for the most part solidly frequentist and will remain so into the foreseeable future."

### Astatistical science

Guthery et al. (2001) contend that the greatest accomplishments in the history of science did not rely on statistical hypothesis testing. Furthermore, they state that science of the "highest order" is possible without the use of statistical hypotheses (but not without research hypotheses). They propose that other, largely astatistical approaches (e.g., graphic depictions of data) should be considered in research programs. Cherry (1998) decries the testing of "obvious hypotheses," and advises that it is not necessary to test every result.

### CONCLUSIONS

How important is statistical power in the long-term monitoring of populations of management concern? If one's primary goal is to answer yes-no type questions, then it is undeniably important. Asking

questions such as “Is the population changing?” are not likely to be informative, however. We know the population is changing. Our focus should be on: (1) the magnitude of this change, (2) the reliability with which we can estimate this magnitude, and (3) the biological relevance of this amount of change. Parameter estimates along with associated confidence intervals can satisfy the first two objectives. Establishing the biological importance will depend upon our knowledge of the natural history of the system, along with the findings of other, related studies. One of the primary recommendations from a workshop on environmental monitoring organized by the Ecological Society of America was that trend studies should focus on description of trends and their uncertainty, rather than hypothesis testing (Olsen et al. 1997).

Ultimately, there is no statistical analysis that can provide a magic number from which it will be obvious what course management should take. Statistics is merely a tool that can provide information about a parameter of interest. Interpretation of this information will always involve a subjective element, and some degree of uncertainty will always exist in any statistical analysis.

Statistics, like other fields of science, evolves, albeit slowly. Many have advocated the abandonment of traditional null hypothesis testing, yet it persists. Others have pointed out the need to consider Type II error in the context of such tests, yet relatively few currently do. Many of those who do attempt to use power analysis ultimately misuse it. In time, however, a new paradigm of statistics may arise, which will allow us to draw more appropriate conclusions from our data. Until this happens, however, we must make the best use of currently employed techniques, taking care not to slip into the pitfalls that have claimed the studies of so many.

## ACKNOWLEDGMENTS

This paper benefited from helpful discussions with Mike DeBacker.

*Lloyd Morrison is a Quantitative Ecologist with the Heartland Network of the National Park Service's Vital Signs Inventory and Monitoring Program, and Adjunct Professor at Missouri State University.*

## LITERATURE CITED

Abelson, R.P. 1997. A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). Pp 117-141 in L.L. Harlow, S.A. Mulaik, and J.H. Steiger, eds., *What If There Were No Significance Tests?* Lawrence Erlbaum, Mahwah, N.J.

Anderson, D.R., K.P. Burnham, and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* 64:912-923.

Anderson, D.R., W.A. Link, D.H. Johnson, and K.P. Burnham. 2001. Suggestions for presenting the results of data analyses. *Journal of Wildlife Management* 65:373-378.

Berger, R.L., and J.C. Hsu. 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 11:283-319.

Birnbaum, A. 1977. The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindsley-Savage argument for Bayesian theory. *Synthese* 36:19-49.

Burnham, K.P., and D.R. Anderson. 1998. *Model Selection and Inference*. Springer-Verlag, New York.

Burnham, K.P., and D.R. Anderson. 2002. *Model Selection and Multi-model Inference*. Springer-Verlag, New York.

Carver, R.P. 1978. The case against statistical significance testing. *Harvard Educational Review* 48:378-399.

Carver, R.P. 1993. The case against statistical significance testing, revisited. *Journal of Experimental Education* 61:287-292.

Cherry, S. 1998. Statistical tests in publications of *The Wildlife Society*. *Wildlife Society Bulletin* 26:947-953.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence-Erlbaum, Hillsdale, N.J.

Cohen, J. 1994. The earth is round ( $p < 0.05$ ). *American Psychologist* 49:997-1003.

Colegrave, N., and G.D. Ruxton. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology*

14:446-450.

Dayton, P.K. 1998. Reversal of the burden of proof in fisheries management. *Science* 279:821-822.

de la Mare, W.K. 1984. On the power of catch per unit effort series to detect declines in whale stocks. *Reports of the International Whaling Commission* 34:655-662.

Dixon, P.M., and J.H.K. Pechmann. 2005. A statistical test to show negligible trend. *Ecology* 86:1751-1756.

Dorazio, R.M., and F.A. Johnson. 2003. Bayesian inference and decision theory--a framework for decision-making in natural resource management. *Ecological Applications* 13:556-563.

Earll, R.C. 1992. Commonsense and the precautionary principle--an environmentalist's perspective. *Marine Pollution Bulletin* 24:182-186.

Eberhardt, L.L. 2003. What should we do about hypothesis testing? *Journal of Wildlife Management* 67:241-247.

Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6:1036-1046.

Field, S.A., A.J. Tyre, N. Jonzen, J.R. Rhodes, and H.P. Possingham. 2004. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters* 7:669-675.

Field, S.A., A.J. Tyre, and H.P. Possingham. 2005. Optimizing allocation of monitoring effort under economic and observational constraints. *Journal of Wildlife Management* 69:473-482.

Fisher, R.A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B* 17:69-78.

Fox, D.R. 2001. Environmental power analysis - a new perspective. *Environmetrics* 12:437-449.

Gerard, P.D., D.R. Smith, and G. Weerakody. 1998. Limits of retrospective power analysis. *Journal of Wildlife Management* 62:801-807.

Gibbs, J.P., S. Droege, and P. Eagle. 1998. Monitoring populations of plants and animals. *BioScience* 48:935-940.

Gigerenzer, G., Z. Swijtink, T. Porter, L.J. Daston, J. Beatty, and L. Kruger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, U.K.

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. *Annals of Internal Medicine* 121:200-206.

- Gray, J.S. 1990. Statistics and the precautionary principle. *Marine Pollution Bulletin* 21:174-176.
- Greenland, S. 1988. On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* 128:231-237.
- Guthery, F.S., J.J. Lusk, and M.J. Peterson. 2001. The fall of the null hypothesis: liabilities and opportunities. *Journal of Wildlife Management* 65:379-384.
- Guttman, L. 1985. The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis* 1:3-10.
- Hauck, W.W., and S. Anderson. 1996. Comment-Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 11:303.
- Hayes, J.P., and R.J. Steidl. 1997. Statistical power analysis and amphibian population trends. *Conservation Biology* 11:273-275.
- Hoening, J.M., and D.M. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 55:19-24.
- Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological monographs* 54:187-211.
- Jennions, M.D., and A.P. Moller. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* 14:438-445.
- Johnson, D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772.
- Johnston, P., and M. Simmonds. 1990. Precautionary principle. *Marine Pollution Bulletin* 21:402.
- Kirk, R.E. 1996. Practical significance: a concept whose time has come. *Educational and Psychological Measurement* 56:746-759.
- Lenth, R.V. 2001. Some practical guidelines for effective sample size determination. *The American Statistician* 55:187-193.
- Leventhal, L., and C. Huynh. 1996. Directional decisions for two-tailed tests: power, error rates, and sample size. *Psychological Methods* 1:278-292.
- Lindley, S.T., M.S. Mohr, and M.H. Prager. 2000. Monitoring protocol for Sacramento River winter Chinook salmon, *Oncorhynchus tshawytscha*, application of statistical power analysis to recovery of an endangered species. *Fishery Bulletin* 98:759-766.
- Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. *Ecological Applications* 5:401-410.
- McBride, G.B. 1999. Equivalence tests can enhance environmental science and management. *Australian and New Zealand Journal of Statistics* 41:19-29.
- McBride, G.B., J.C. Loftis, and N.C. Adkins. 1993. What do significance tests really tell us about the environment? *Environmental Management* 17:423-432.
- Meeks, S.L., and R.B. D'Agostino. 1983. A note on the use of confidence limits following rejection of a null hypothesis. *The American Statistician* 37:134-136.
- Murphy, K.R., and B. Myers. 2004. *Statistical Power Analysis: a Simple and General Model for Traditional and Modern Hypothesis Tests*. Lawrence Erlbaum, Mahwah, N.J.
- Nakagawa, S., and T.M. Foster. 2004. The case against retrospective power analyses with an introduction to power analysis. *Acta Ethologica* 7:103-108.
- Neyman, J., and E.S. Pearson. 1967. *Joint Statistical Papers*. University of California Press, Berkeley.
- Nickerson, R.S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* 5:241-301.
- Oakes, M. 1990. *Statistical Inference*. Epidemiology Resources, Chestnut Hill, Mass.
- Olsen, T., B.P. Hayden, A.M. Ellison, G.W. Oehlert, and S.R. Esterby. 1997. Ecological resource monitoring: change and trend detection workshop report. *Bulletin of the Ecological Society of America* 78:11-13.
- Ortiz, M. 2002. Optimum sample size to detect perturbation effects: the importance of statistical power analysis--a critique. *Marine Ecology* 23:1-9.
- Parkhurst, D.F. 2001. Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. *BioScience* 15:1051-1057.
- Peterman, R.M. 1990a. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Science* 47:2-15.
- Peterman, R.M. 1990b. The importance of reporting statistical power: the forest decline and acidic deposition example. *Ecology* 71:2024-2027.
- Peterman, R.M., and M.J. Bradford. 1987. Statistical power of trends in fish abundance. *Canadian Journal of Fisheries and Aquatic Science* 44:1879-1889.
- Peterman, R.M., and M. M'Gonigle. 1992. Statistical power analysis and the precautionary principle. *Marine Pollution Bulletin* 24:231-234.
- Platt, J.R. 1964. Strong inference. *Science* 146:347-353.
- Poole, C. 2001. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* 12:291-294.
- Popper, K.R. 1959. *The Logic of Scientific Discovery*. Basic Books, New York.
- Reed, J.M., and A.R. Blaustein. 1997. Biologically significant population declines and statistical power. *Conservation Biology* 11:281-282.
- Robinson, D.H., and H. Wainer. 2002. On the past and future of null hypothesis significance testing. *Journal of Wildlife Management* 66:263-271.
- Rotenberry, J.T., and J.A. Wiens. 1985. Statistical power analysis and community-wide patterns. *American Naturalist* 125:164-168.
- Sedlmeier, P., and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309-316.
- Shaver, J.P. 1993. What statistical significance testing is, and what it is not. *Journal of Experimental Education* 61:293-316.
- Shrader-Frechette, K.S., and E.D. McCoy. 1992. Statistics, costs and rationality in ecological inference. *Trends in Ecology and Evolution* 7:96-99.
- Simberloff, D. 1990. Hypotheses, errors, and statistical assumptions. *Herpetologica* 46:351-357.
- Stefano, J.D. 2001. Power analysis and sustainable forest management. *Forest Ecology and Management* 154:141-153.
- Stefano, J.D. 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology* 17:707-709.
- Steidl, R.J., J.P. Hayes, and E. Schaubert. 1997. Statistical power analysis in wildlife research. *Journal of Wildlife Research* 61:270-279.
- Steiger, J.H. 2004. Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods* 9:164-182.
- Stoehr, A.M. 1999. Are significance thresholds appropriate for the study of animal behaviour? *Animal Behaviour* 57:F22-F25.
- Suter, G.W.I. 1996. Abuse of hypothesis testing statistics in ecological risk assessment. *Human and Ecological Risk Assessment* 2:331-347.
- 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.
- Thomas, L. 1997. Retrospective power analysis. *Conservation Biology* 11:276-280.
- Thomas, L., and F. Juanes. 1996. The importance of statistical power analysis: an example from *Animal Behaviour*. *Animal Behaviour* 52:856-859.
- Thomas, L., and C.J. Krebs. 1997. A review of statistical power analysis software. *Bulletin of the Ecological Society of America* 78:126-139.
- Toft, C.A., and P.J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist* 122:618-625.
- Utts, J. 1988. Successful replication versus statistical significance. *The Journal of Parapsychology* 52:305-320.
- Wade, P.R. 2000. Bayesian methods in conservation biology. *Conservation Biology* 14:1308-1316.
- Yoccoz, N.G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72:106-111.
- Zumbo, B.D., and A.M. Hubley. 1998. A note on misconceptions concerning prospective and retrospective power. *The Statistician* 47:385-388.