IMPROVING RESOURCE MONITORING AND MANAGEMENT THROUGH STATISTICAL POWER

Richard A. Evans, U.S.D.I. National Park Service, Delaware Water Gap National Recreation Area, H.C. 38, Milford, PA 18337. Telephone: (717) 296-6952; email: richard_evans@nps.gov.

ABSTRACT: Park protection requires the ability to recognize the degradation of resources before they are irreparably damaged. Statistical power, in this context, is the ability to detect resource degradation when it occurs. Environmental monitoring studies are instituted on the premise that they will detect resource degradation if it occurs, and decisions are made with the assumption that if no resource degradation has been detected, none has occurred, and management actions are not needed. However, many environmental studies have been shown to have low statistical power; they are incapable of detecting significant resource damage. Scientists and managers should consider statistical power when designing studies and making decisions based on their results. I describe statistical power, and how it can be used to improve designs of studies, interpretation of results, and management decisions.

Introduction

Most environmental research and monitoring studies are conducted with the purpose of describing, explaining, or predicting changes or differences in resources. Statistical power can be defined as the probability of correctly detecting "effects," such as changes or differences in resources. Statistical power analysis is a method of calculating the probability of detecting such effects, provided information about the size of the effect, the amount of variability in the data, and certain other conditions. The importance of power analyses in ecological and environmental studies has been demonstrated by de la Mare (1984), Gerrodette (1987), Peterman (1990), Taylor and Gerrodette (1990), Toft and Shea (1983), and others.

Medical Analogy: It may help to consider a medical analogy to understand the concept and importance of statistical power. Medical diagnostic tests are conducted to determine whether or not there is a health problem, and identify it. Two types of erroneous test results can occur: a "false positive," which indicates a health problem that does not really exist, and a "false negative," which fails to detect a real health problem. Here, statistical power is the probability of detecting a health problem that really exists. Low statistical power is equivalent to a high probability of false negative results. If a potentially life threatening cancer is present, it is crucial that medical tests have very high statistical power to detect it as soon as possible. Similarly, data from environmental studies are used to determine whether or not there is an environmental problem, and identify it. In many cases, as in medicine, it is crucial that the environmental tests have high statistical power, in order to detect real problems.

Implications for "Adaptive Management"

Natural resource agencies in the United States have been encouraged to pursue a path of "adaptive management," in which management policies and actions are adjusted appropriately in response to new scientific information. However, false negative results can produce misguided, "maladaptive" management in two ways. First, there may be a failure to recognize detrimental impacts. For example, in 1983 the International Whaling Commission adopted a policy that allowed existing harvest rates to continue for whale populations that did not show "statistically significant" declines. However, de la Mare (1984) showed that major declines in whale populations (exceeding 50%) could occur with little probability of being recognized with the monitoring data being collected (Peterman 1990). The whale monitoring program was clearly insufficient to inform and guide management, because it had very low statistical power. Whale populations could have been driven nearly to extinction before any problem was recognized. Second, false negative results can lead to a failure to recognize the benefits of conservation programs and management actions. Those beneficial programs and actions may then lose support and be abandoned, forfeiting potential environmental gains.

Hypothesis Testing

Statistical power is one aspect of hypothesis testing, and most environmental research and monitoring involves hypothesis testing, either implicitly or explicitly. The foundation of hypothesis testing are two mutually exclusive statements about conditions in nature: the Null Hypothesis (HO), which typically asserts "no effect" (or a minimal effect that does not exceed some threshold), and the Alternative Hypothesis (HA), which asserts a significant effect. For example:

HO: Water conductivity has not increased (beyond some threshold).

vs. HA: Water conductivity has increased (beyond some threshold).

Although there are numerous forms of hypothesis tests ("t," "F," chi-squared, etc.), the outcomes can always be summarized by the same, simple "truth table" (Table 1). This table contrasts two mutually exclusive states of nature with two mutually exclusive decisions, to produce four possible outcomes. Of the four possible outcomes, two are correct, and two are incorrect. First, consider the case where HO is true: a decision to "accept" HO would be correct, whereas a decision to reject HO would be in error. Historically, scientists have tightly controlled the probability (alpha, or "a") of this kind of error (a false positive or "Type I" error); typically, $a \leq 0.05$ (5%). Second, consider the case where HO is not true; instead, HA is true: now a decision to reject HO would be correct, whereas a decision to "accept" HO would be in error. Historically, scientists have not controlled the probability (beta, or "b") of this second kind of error (a false negative or "Type II" error).

Table 1. The four possible outcomes of a statistical hypothesis test, and their probabilities (in parentheses).

	Decision	
State of Nature	Do not reject Ho	Reject Ho
Null hypothesis (Ho) True	Correct (1-a)	Type I Error (a) "False Positive"
Alternative hypothesis (HA) True	Type II Error (b) "False Negative"	Correct (1-b)= Statistical Power

Monitoring No Better than Flipping a Coin?

Many environmental studies have been shown to have very low power (Peterman 1990). A power analysis of white perch (*Morone americana*) in the Hudson River, New York, revealed that with a=0.05, the probability of detecting a 50% decline in recruitment was only about 50%, even after 100 years of monitoring. Similarly, there was only about a 45% chance of detecting a 7% annual decline in an endangered dolphin population. In another case, Peterman (1989) calculated that the probability of rejecting HO when HA was true was only 20%; there was an 80% probability of erroneously "accepting" HO. In other words, the odds of these studies detecting such effects were no better - or much worse - than simply flipping a coin! Clearly, whenever the decision is to "accept" HO, it is important to know the probability (b) of a false negative. If b is high (corresponding to low power) or is unknown, a failure to reject HO should not be interpreted as proof that HO is true.

Power Relationships

Statistical power is determined by four elements of study design:
(1) b (probability of a false negative): Reducing b increases power (power = 1-b).
(2) a (probability of a false positive): Increasing a increases power (decreases b).
(3) "Effect size:" The minimum size of the effect stated in the alternative hypothesis. Increasing the effect size increases power. For example, the power to detect a 50% change in a breeding bird population would be much higher than the power to detect a 5% change.
(4) Variability: Reducing data variability increases power.

Costs of Errors

The relative "costs" of the two types of errors are implied by the ratio b/a. Historically, a was set at $a \leq 0.05$, but b was uncontrolled. If a=0.05 and b=0.5, b/a = (0.5/0.05)=10; this ratio implies the cost of a false positive error is ten times the cost of a false negative

error. However, the cost of a false negative error is often much greater than the cost of a false positive in environmental management and conservation biology. A false positive usually entails socio-economic costs (regulations, new technology and operational procedures, etc.) but not environmental costs. In contrast, a false negative often entails both the socio-economic costs of a false positive and the additional costs associated with environmental damage (Peterman 1990). In some cases false negatives entail collapse of the resource (e.g. fisheries) or species extinction. Saetersdal (1980) demonstrated that emphasis on maintaining a traditionally low a and ignoring b and power "contributed to the collapse of several North Atlantic and North Sea pelagic fish stocks because large decreases in abundance occurred before strong actions were recommended" to restrict fishing pressure (Peterman 1990, p.10). Adherence to the tradition of $a \leq 0.05$ without consideration of the ratio b/a in environmental studies often amounts to "stacking the deck" against the resource. A more rational approach is to increase a and decrease b to reflect the relative costs of each type of error.

Designing Studies with Power

At least four steps can be taken to increase the statistical power of studies:

(1) Obtain a preliminary estimate of data variability, either from the literature, or by conducting a pilot study. Use that estimate to conduct *a priori* (Peterman 1990) or "prospective" (Thomas 1997) power analysis to determine specific relationships among a, b, effect size, and sample size for your study.

(2) Explicitly consider the relative costs of false positive and false negative errors, and make sure they are appropriately reflected in the ratio b/a. If the cost of a false negative is greater than the cost of a false positive, decrease b and increase a so the ratio is less than one. If information about the relative costs of errors is not available, a rule-of-thumb is to set $a=b \leq 0.2$.

(3) Identify a meaningful effect size, based on biology, public perceptions, etc. Calculate the minimum effect size detectible with specified a, b, variability, and sample size. Adjust a, b, and sample size to ensure reliable detection of meaningful effects.

(4)Minimize the amount of variability in the data. While a certain amount of variability is inherent, variability can be minimized in the following ways: (a) Carefully choose measurement protocols and response variables. Often, certain measurements or response variables will have much lower variability than others, and yet sensitively respond to the environmental issue of concern. (b) Use special sampling designs, such as stratified random sampling or multi-stage sampling. When applied appropriately, these sampling designs can effectively reduce variability without increasing sampling effort. (c) Collect as many samples or measurements as possible; increasing sample size almost always reduces variability.

Examples of the calculations involved in power analysis can be found in Gerrodette (1987), Green (1989), and Peterman (1989), and others. Computer software to perform statistical power analysis is widely available, both commercially and as “freeware” on

the internet. Thomas and Krebs (1997) provide an excellent review of this software.

Interpreting Results with Power

When the null hypothesis is rejected with an acceptably low a, no further analysis is necessary; the alternative hypothesis should be accepted. However, if the null hypothesis is not rejected and b (or power) is not known, an appropriate "retrospective" power analysis will help to interpret the result. A pre-specified "effect-size" should be used in such retrospective power analyses, not the effect measured in the study (Thomas 1997). If the analysis indicates that b is acceptably low (e.g. $b \leq 0.2$; power≥ 0.8), the null hypothesis can be "accepted." In contrast, if b is found to be high (e.g. $b \geq 0.2$; power≤ 0.8), the results are inconclusive.

Making Management Decisions with Power

Management decisions should include consideration of the reliability of study results, and the consequences of errors.

(1) Consider the relative costs of false positive and false negative errors (b/a).

(2) Do not base management decisions on the misguided assumption that there was "no effect" just because a study failed to demonstrate an effect.

(3) Require scientists to report the "retrospective" statistical power of studies that failed to demonstrate an effect.

(4) Require "prospective" power analyses of major studies and monitoring projects.

(5) Redesign or discontinue ongoing studies, and do not commit to new studies, if they have such low statistical power they are unlikely to detect meaningful effects.

(6) Reverse the traditional burden of proof in cases where a false negative error would incur very high cost. Instead of allowing and having to document resource damage before taking action, require those using and threatening resources to show, with high power studies, no resource damage results from their activities.

(7) Consider revising policies, regulations, and standards that were based on the results of studies having very low power.

References

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

Gerrodette, T. 1987. A power analysis for detecting trends. Ecology 68(5): 1364-1372.

Green, R. H. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research 50: 195-205.

Peterman, R. M. 1989. Application of statistical power analysis to the Oregon coho salmon (*Oncorhunchus kisutch*) problem. Canadian Journal of Fisheries and

Aquatic Sciences 46: 1183-1187.

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

Saetersdal, G. 1980. A review of past management of some pelagic fish stocks and its effectiveness. Rapp. P.-V. Reun. Cons. Int. Explor. Mer 177: 505-512.

Taylor, B. L. and T. Gerrodette. 1993. The uses of statistical power in conservation biology: the vaquita and the northern spotted owl. Conservation Biology 7(3): 489-500.

Thomas, L. 1997. Retrospective power analysis. Conservation Biology 11(1): 276-280.

Thomas, L. and C. J. Krebs. 1997. A review of statistical power analysis software. Bulletin of the Ecological Society of America 78(2): *in press*.

Toft, C. A. and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. The American Naturalist 122(5): 618-625.