Impact Assessment
and Environmental Monitoring:
The Role of Statistical Power Analysis

Bonnie Antcliffe

A Manuscript Report Prepared for the
Canadian Environmental Assessment
Research Council
June 1992

ACKNOWLEDGEMENTS

I wish to thank CEARC for funding this report. This report was also partly funded by a British Columbia Science Council Grant and by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada. I wish to thank my supervisor, Dr. Randall Peterman, for his advice and support and for his ideas that have been incorporated in this report. I also wish to thank Dr. Frank Gobas for helpful discussions. I am grateful to Alan Redenbach and Bryan Kelso at Environment Canada, North Vancouver, B.C. for supporting my research on statistical power analysis through a Graduate Co-op Work Term.

TABLE OF CONTENTS

ABSTRACT

Statistical hypothesis testing is often used in environmental monitoring, and to a lesser extent in impact assessment, to test some null hypothesis (e.g. that there is no effect of a hydroelectric project on fish mortality rate). An important concept in statistical hypothesis testing is statistical power, which is the probability that a monitoring project or impact assessment will correctly detect an effect of a specified magnitude, provided this effect exists. Statistical power analysis methods for designing impact assessments or monitoring programs that have high power are readily available, yet they are rarely used. Instead, designs are often based on historical precedents or other non-statistical criteria. As a result, statistical power is often low for environmental studies, meaning that such studies have little chance of correctly detecting specified effects, even if they actually exist.

The purpose of this report is to illustrate the importance of statistical power analysis and to show its applicability to impact assessment and environmental monitoring. Here, I review statistical power and factors that influence it. I show how statistical power analysis can be used (a *priori*) to improve the design of impact assessments or monitoring programs, and (a *posteriori*) to help interpret the results of past studies that failed to reject some null hypothesis. These uses of statistical power analysis are illustrated by several examples. I make recommendations to routinely apply statistical power analysis and to include statistical power as a criterion to evaluate the effectiveness and efficiency of proposed impact assessments and monitoring projects.

1. INTRODUCTION

In Canada there are no standard protocols describing the methodologies to be used when designing impact assessments for projects under the federal Environmental Assessment Review Process. Practitioners can thus use a variety of approaches to design assessments. Some of these approaches are qualitative (e.g. resource inventories, check lists, matrices), while others are more quantitative (e.g. experimental research, statistical hypothesis testing, and simulation modelling). Several authors have recognized the need to encourage quantitative impact assessments that make use of hypothesis testing and statistically-based study designs (e.g. Gore et al. 1979; Green 1979; Fritz et al. 1980; Beanlands and Duinker 1983). Quantitative approaches, including statistical hypothesis testing, are also encouraged for environmental monitoring, which is an integral part of impact assessment (CEARC 1986).

Statistical power is an important concept relevant to quantitative impact assessments and monitoring programs that utilize statistical hypothesis testing. Statistical power describes the ability of an impact assessment or monitoring project to correctly detect a specified effect, provided this effect exists (Dixon and Massey 1983; Cohen 1988). High-power impact assessments or monitoring programs have the greatest chance of correctly detecting specified effects, if they actually exist. Low-power studies, on the other hand, are flawed because they have little chance of correctly detecting such effects.

The theory of statistical power is well developed and the methods of statistical power analysis for the design and interpretation of statistically-based impact assessment or monitoring projects are readily available (e.g.

Dixon and Massey 1983; Zar 1984; Cohen 1988; Lipsey 1990; Peterman 1990a). Yet, statistical power has received little attention in impact assessment and monitoring. For example, Beanlands and Duinker (1983) fail to mention statistical power in their report on ecological approaches to environmental impact assessment, but their report has become the basic scientific framework for quantitative approaches to impact assessment in Canada (although they do include references that allude to statistical power). As well, very few practitioners actually use statistical power analysis in the design and evaluation of impact assessments or monitoring projects (Green 1989). This lack of application of power analysis is a serious problem because it can lead to lengthy and costly low-power studies that are unlikely to detect even large, ecologically important effects (e.g. Vaughan and Van Winkle 1982; De la Mare 1984; Hayes 1987; Peterman 1990a). Furthermore, many practitioners ignore statistical power when their studies fail to detect effects (Peterman 1990b), which can mislead decision makers in cases where the environmental assessment or monitoring program had only a low probability of detecting an effect anyway.

The purpose of this report is to review the concept of statistical power and to illustrate the role of statistical power analysis in impact assessment and environmental monitoring. I show how a ***priori*** power analysis can be used in the design phase of impact assessments or monitoring programs, and how a ***posterior?*** power analysis can help interpret the results of past studies that failed to reject some null hypothesis. I also make recommendations for practitioners and for the Canadian Environmental Assessment Research Council (CEARC).

2. STATISTICAL POWER

Traditional statistical hypothesis testing (Zar 1984) is a process where data are used to test a null hypothesis (for example, that there is no effect of a hydroelectric project on fish mortality rate). The statistical procedure results in a decision to either reject the null hypothesis (H_O) or not reject it. Since the true state of nature (that is, whether H_O is true or false) is unknown, there are four possible outcomes for a statistical hypothesis test (Table 1). If the true state of nature is that the null hypothesis (H_O) is actually true (i.e. there really is no effect of the project on the mortality rate of fish), then a statistical decision to reject H_O will result in a type I error. Alpha (α) is the acceptable probability for making a type I error. The probability of correctly failing to reject H_O when H_O is actually true, is 1-α. If the true state of nature is that H_O is false (that is, if a true effect exists), then a statistical decision failing to reject H_O will result in a type II error. Beta (β) is the acceptable probability of making such an error. The probability of correctly rejecting H_O when H_O is actually false, is 1-β. Thus, 1-β is statistical power, that is, the probability of correctly detecting a specified effect, if that effect actually exists.

In statistical hypothesis testing, researchers always preset α - the acceptable level for the probability of making a type I error (i.e. the probability of incorrectly concluding that there is an effect when in fact there is no effect). α however, applies only when the H_O is true. Since there is no way of knowing whether H_O is actually true or false, researchers and decision makers should also be concerned with the probability (β) of making a type II error (i.e. the probability of incorrectly concluding that there

is no effect when in fact there really is an effect). This type II error applies when the H_0 is false.

In general, researchers do not tend to design environmental impact assessments, monitoring programs, or other data gathering procedures in order to generate a low β value; in fact, general practice appears to ignore the implied β (Peterman 1990a). For the reasons discussed below, this leads to large β, which means that power is low (i.e. there is a low probability of correctly detecting specified effects, even when they are present). In order to attain high power, and thus to minimize the probability of a type II error, the desired β should be preset to some low value during the phase of designing a data-gathering procedure that will lead to a statistical hypothesis test. The desired β, however, might not necessarily be set to the same, low level as that for α because the cost of a type II error might not equal the cost of a type I error. Often in environmental monitoring, the costs of a type II error (i.e. the costs of an incorrect conclusion of no effect, such as human health effects or loss of revenues from a fishery) are greater than the costs of a type I error (i.e. the costs of an incorrect conclusion of an effect, such as the installation of a pollution control device) (Toft and Shea 1983; Peterman 1990a). Thus, providing there is no preference for a particular type of error, researchers might set CY and β so that the expected cost (probability of occurrence times the cost if the event occurs) of a type I and type II error are equal (see Peterman 1990a). If researchers want to set α and β based on the sampling costs and the expected costs of each of the four possible outcomes from a statistical test, then formal decision analysis (Raiffa 1968; Parkhurst 1984) can be used.

Factors That Influence Statistical Power

The specific equation for estimating statistical power depends on the statistical test (e.g. t-test, F test, or chi-square test). In general, statistical power is a function of four factors: alpha, the effect size, the sample size, and the sample variance (Dixon and Massey 1983). Below, I describe these four terms in the power equation, along with other factors that influence power.

Alpha (α)

Alpha (CY) is the acceptable probability for a type I error. If, after a statistical analysis of some data, the probability of a type I error (P value derived from the statistical test) is less than α, then H_0 is rejected at an α level of significance, i.e. the result is statistically significant (Cohen 1988). If the P value exceeds α then H_0 is not rejected.

α is inversely related to β, or equivalently, α is positively related to power (power = 1 -β), with all else equal (Dixon and Massey 1983). So as CY increases, the power of the statistical test increases, but at the expense of an increased risk of a type I error. α is, however, almost always set by convention at 0.05 for statistical significance testing.

Effect Size

The effect size is often defined as the magnitude of the real effect, e.g. the true difference between a control and a treatment mean (Lipsey 1990). However, this effect size might not be the effect size that is important, say for biological, social, or economic reasons. Thus, researchers should have in mind the effect size that is of concern (Sharma et al. 1976; Green 1984). A study can then be designed to detect this important effect with high-power. Often there is very little information regarding the importance of effects, hence a range of effect sizes should be considered in any analysis.

Large effects are easier to distinguish against background variation than small effects, with all else equal, and thus large effects increase the chance of showing up as statistically significant in a statistical test. Hence, statistical power also increases as the effect size increases (Cohen 1988). In certain situations, it might be possible to select variables for impact assessments or monitoring projects according to the size of the effect that is of concern. For example, consider a study designed to test the effectiveness of a pollution control device. Dynamic response models (e.g. McKay 1989) can be used to determine which compartment (e.g. water, sediment, fish, benthos, or plants) will respond the fastest to reduced loading of a pollutant (Dr. Frank Gobas, Simon Fraser University, Burnaby, B.C., pers. communic.) That compartment will provide the largest effect size and will thereby be the most detectable (have the highest power).

Sample Size

The sample size influences the sample variance, along with other variables. The larger the sample size, the smaller the sample variance, and as variability decreases, power increases because real effects are easier to distinguish from natural background variability (Cohen 1988; Lipsey 1990). However, larger sample sizes also increase sampling costs and in some cases may lead to pseudoreplication (Hurlbert 1984). Therefore, it is important to note that sample size is only one of the four main factors that determine power.

Sample Variance

Although larger sample sizes decrease the sample variance and hence increase power, there are other ways to decrease the sample variance. These include the reduction of measurement error (by improving sampling or analytical techniques), and the reduction of the variability associated with uncontrolled factors (by using a sampling design amenable to analysis of variance with blocking, or control-treatment pairing) (see McKenzie et al. 1977; Skalski and McKenzie 1982; Millard and Lettenmaier 1986). Note, however, that if a sampling design based on analysis of variance with blocking or pairing is used in a situation where it does not reduce the variance, then the power of the statistical test will be lower than that for the same design without blocking or pairing because blocking or pairing reduces the degrees of freedom (Dixon and Massey 1983). Blocking or pairing should thus be used only when the loss of power from having fewer degrees of freedom is offset by the increase in power from removing the variance associated with extraneous factors.

Other Factors

The directionality of the statistical test can also influence power. Two-tailed tests assess deviations from the null hypothesis in two directions, and are therefore less powerful than unidirectional tests (which assess deviations from the null hypothesis in one direction only) (Cohen 1988). Although one-tailed tests are more powerful, they have no power to detect effects in the direction opposite to that stated by the alternative hypothesis (Cohen 1988).

Statistical power is also influenced by the extent to which the data meet the assumptions of the statistical test. The degree to which power is affected depends on the statistical test and the particular assumption. It is important to note that many statistical tests are extremely robust to violations of their assumptions (Dixon and Massey 1983; Green 1979). Thus, researchers should consider how such violations will affect the power of the test before transforming the data or resorting to non-parametric tests (Green 1979). Non-parametric statistics have fewer assumptions than parametric tests, but they are generally less powerful (Green 1979, Lipsey 1990).

3. STATISTICAL POWER ANALYSIS IN IMPACT ASSESSMENT AND ENVIRONMENTAL MONITORING

Statistical **power analysis is relevant** for all types of impact assessment and monitoring where a statistical test is used to test some null hypothesis (H_O). In general, it is more relevant in environmental monitoring because statistical hypothesis testing is more common there. Although statistically-based study designs are used to a lesser extent in impact assessment in Canada, experimental approaches based on statistical hypothesis testing are increasingly common in impact assessment (e.g. Thomas et al. 1978; McKenzie et al. 1979; Fritz et al. 1980). Thus, in this report I will refer to the applicability of statistical power analysis in environmental studies, which could be impact assessments or monitoring programs.

Power analysis can be applied a ***priori*** in the experimental design of a study and a ***posteriori*** in the interpretation of results when a statistical test fails to reject some H_O (Fig. 1). Below, I explain the role of a ***priori*** and *a* ***posteriori*** power analysis in impact assessment and monitoring. The general concepts presented here are also applicable for statistically-based studies in other disciplines.

A *Priori* Power Analysis

The steps in designing a study to test some null hypothesis (i.e. objective - questions - hypotheses - model - sampling design - statistical test) are well known concepts in experimental design and are described elsewhere (e.g. Cochran and Cox 1957; Winer 1971; Green 1979, 1984). However, the usefulness of a ***priori*** power analysis in the experimental design of a study is less familiar to many researchers. A ***priori*** power analysis can be used in the design of a study to determine the sample size required to attain high power to detect an important effect, given α and an estimate of the sample variance (e.g. Green **1979,** 1984; Skalski and McKenzie 1982; Bernstein and Zalinski 1983;, Alldredge 1987; Gerrodette 1987). Desired high power should be set to at least 0.8, which means the study design will have at least an 80% chance of correctly detecting the important effect, if this effect actually exists. If desired power is $\geq$ 0.8, then β- the probability of type II error, is $\leq$ 0.2. If researchers want to be as conservative about making a type **II** error as they are about making a type I error, then they can set β equal to CY. Thus, if α is set to 0.05 by convention, then β would be 0.05 and hence desired power (l-beta) would be 0.95. Other approaches to set desired power include balancing the expected costs of type I and type II errors (see Peterman1990a), and decision analysis (Raiffa 1968), which can account for both the sampling costs and the expected costs of the possible outcomes from a statistical test.

Generally, a *priori* power analysis is not used to determine the sample size required to design an environmental impact or monitoring study to have desired high power. Instead, sample sizes are usually set arbitrarily, based on past practices, or by logistical constraints, which has lead to a number of low-power studies because of too few samples to attain high power to detect a specified effect (e.g. McCaughran 1977; Thomas 1977; Vaughan and Van Winkle 1982; Hayes 1987; Peterman 1990a). Low-power studies, however, provide little information about whether the important effect size actually exists because the study had little chance of correctly detecting that effect, if it was present. Only when power is high can researchers be reasonably sure that their methods would have detected an important effect, if it was present.

In order to prevent wasted time and money on low-power impact assessments or monitoring projects that are unlikely to detect postulated effects, researchers should use statistical power analysis to design studies to have high power. High-powered studies tend to make β, the probability of a type II error, low. There are two main reasons for doing everything possible to reduce the chance of making type II errors in environmental studies. First, type II errors (for example, a conclusion that there is no effect of some pollutant on biota when in fact, there really is some detrimental effect), can lead to the unjustified degradation of our environment. Second, type II errors (which might, for example, lead to the unjustified loss of a valuable sport fish species and hence potential revenues, or to health effects associated with consumption of contaminated food sources), are often costly in environmental situations.

In addition to using a *priori* power analysis in the experimental design of a study to determine the sample size required to attain desired high-power, it can also be used to estimate power for studies with a planned sample size (e.g. Gerrodette 1987). If power is estimated to be low for the planned sample size, then alternative designs that might increase power should be considered. Power can be increased by increasing the sample size or the duration of the study, or by decreasing the sample variance. Green (1989) shows that a resampling approach (where the same sites are sampled before and after an impact) can increase power substantially over a reallocation approach (where new sites are sampled after the impact). For estimating sample means, Brumelle et al. (1983) show that composite sampling (where a number of samples are pooled and analyzed as a single sample) provides a smaller estimate of the standard error than does grab sampling (where each sample is analyzed separately); composite sampling could then increase the power of the statistical test. The choice of sampling methods should also be considered because the size and type of sampler and the density of organisms can influence the sample variability and hence statistical power (e.g. Morin 1985).

Another application of power analysis in the experimental design of a study is to determine which variables can be monitored with high power (Dr. R. M. Peterman, Simon Fraser University, Burnaby, B.C., pers. communic.). For example, a ***priori*** power analysis can be applied to a number of variables to determine the sample size required to monitor each variable with high power. If the sample size required for high power cannot be attained for some of those variables due to cost constraints or too few units available to sample, then those variables should not be monitored. An “effects monitoring” strategy for Canada’s east coast (Thomas et al. 1985) recommends that this type of approach be taken so that those variables that are unlikely to be distinguished. from background variation can be eliminated

from the monitoring program at an early stage. Alternatively, new sampling techniques could be considered to improve the probability of detecting important effects in those variables. Furthermore, monitoring all variables may be unnecessary because of the redundant information provided by correlated variables (e.g. Kaesler et al. 1974; MacDonald and Green 1983). Thus, power analysis can be applied to correlated variables to determine which variable(s) can be monitored with high power; this will improve the cost effectiveness of the monitoring program.

Statistical power provides information about the ability of impact assessments or monitoring programs to detect important effects. It should thus be considered, along with other criteria such as proper experimental design, in an a *priori* evaluation of proposed projects (Fig. 1). Studies designed to have low-power might be deemed inefficient because time and money could be wasted on projects that have little chance of detecting an effect of a specified magnitude, if that effect exists (Peterman 1990a). Furthermore, low-power projects might have to be redesigned with high power and repeated later in order to test strongly whether there is an effect (Peterman 1990a). Studies are also inefficient if the sample sizes used are far in excess of that necessary to detect an important effect with high power. Power can also be used as a criterion to evaluate effectiveness of projects because if they have low-power, they have little chance of detecting anything but catastrophic changes. Hence they are ineffective because they provide little or no information about the presence of smaller, yet possibly still important effects. Thus, statistical power analysis should be taken into account in the a *priori* design phase of an environmental impact assessment or monitoring study to ensure that only those projects that are efficient and effective are implemented. A *posteriori* power analysis

(described below) can also be informative by indicating which designs should not be implemented in the future, because they had power that was too low.

A *Posteriori* Power Analysis

Although a *priori* power analysis can be used to design a study to attain desired high power, one cannot always be sure that high power will result. For example, if the actual sample variance is larger than that estimated before the study, or if the sample size required for high power is not attained due to logistical difficulties, then power will be lower than estimated *a priori.* Thus, *a posteriori* power analysis must be applied to calculate the power of a statistical test when analysis of the data gathered does not reject the H_O.

Statistical power needs to be determined *a posteriori* because there are two reasons for failing to reject some H_O. First, a decision to fail to reject the H_O can occur when the true state of nature is that there really is no effect. Second, a failure to reject the H_O can occur when there really is an effect of a specified magnitude present, but the study had low power and hence a low probability of detecting it. Thus, if an *a posteriori* power analysis shows that power is low, then the results of the statistical test are inconclusive for the effect size deemed important because the researcher does not know the real reason for failing to reject the H_O (Peterman and M'Gonigle 1992). When the H_O is not rejected with high power (i.e. ≥ 0.8) for some specified effect size, researchers can be reasonably sure that their methods would have detected the associated size of effect, if that effect was present.

The methods to use a ***posteriori*** power analysis to calculate power are readily available (e.g. Dixon and Massey 1983; Zar 1984; Cohen 1988), even in computer software (Goldstein 1989). The general approach is to calculate the power of the statistical test for the specified important effect size, α, the sample size used in the study, and the sample variance estimated from the study. It is note-worthy that power is not relevant when a statistical test rejects some H_0 (from Table I). Thus, a ***posteriori*** power analysis applies only when a statistical test fails to reject some H_0 (Cohen 1988).

Statistical power is often low in environmental studies (Peterman 1990a). As well, a review of the toxicology literature indicated that power was high in only 19 out of 688 reports that failed to reject some H_0 (Hayes 1987). Low-power studies are uninformative for the effect size deemed important because they have little chance of ,detecting this important effect, if it actually exists. Such studies will only have a high probability of detecting some much larger effect size, one that may have already generated some severe consequences.

Although the power of a statistical test is necessary for the correct interpretation of the results when they fail to reject some H_0, it is rarely reported. This is often misleading because when the H_0 is not rejected but power is not reported, researchers might conclude that there is no effect. For example, Toft and Shea (1983) have pointed out several cases where a failure to reject H_0 has led researchers to conclude that some effect is absent. These conclusions of no effect are of course reasonable only when power is high, that is, when there is a high probability of correctly detecting an effect of a specified size, if this effect is actually present. Researchers can help prevent invalid conclusions of no effect by routinely reporting statistical power for a specified effect size (or a range of important effect sizes), when a test fails to reject some H_0.

4. EXAMPLES OF POWER ANALYSIS

Although the methods of statistical power analysis are well developed, they are rarely used in impact assessment or monitoring. There are, however, some examples of power analysis in environmental studies, only a few of which are illustrated here. These examples illustrate the variety of ways in which power analysis can be informative in the design and evaluation of impact assessments and monitoring programs. Most of these examples pertain to environmental monitoring because there are few examples of power analysis in impact assessment. However, the methods and concepts of statistical power analysis presented in the examples below are applicable to impact assessments that utilize statistical hypothesis testing.

Designing Studies

Power analysis can be used to help design studies to have the best chance of correctly detecting effects (e.g. trends, violations of environmental standards, or differences between control and treatment means). For example, Gerrodette (1987) applied ***a priori*** power analysis to alternative ways of detecting a time trend with linear regression in the abundance of a California sea otter population. Based on a five-year program, one aerial survey per year would lead to a 72% chance of detecting an increase in sea otter abundance of 10% per year (Fig. 2). Two flights per year would give a 95% chance of detecting a change of 10% per year. To detect an increase in abundance of 6.8% per year with power of 0.95, five flights per year would be required. Although Gerrodette (1987) applied power analysis to detect a trend in abundance, this method can also be used to detect linear trends in other variables such as diversity, productivity, or mortality. This method has been criticized by Link and Hatfield (1990), who report that the power calculations are unreliable because the t-distribution approximates the standard normal distribution only when the sample size is large. In response, Gerrodette (1991) shows that the original method (Gerrodette 1987) is valid when the number of degrees of freedom are large, and when samples are taken at regular intervals in time and space. If data are not taken at equal intervals in time or space, then the method in Gerrodette (1 991) applies.

Another approach to using power analysis to design monitoring programs (or impact assessments) is to use optimization techniques to maximize power, given some other constraints. This method was used by Millard and Lettenmaier (1986), who developed two optimization strategies for designing environmental monitoring programs to detect biological or ecological effects. They show how a monitoring program (based on a factorial treatment design) can be designed to maximize power for a fixed cost. They also show how the sampling costs can be minimized for a given desired power or probability of detecting an effect.

Selecting Variables

Power analysis can also be used to determine which variables should be assessed or monitored. This type of approach was adopted by Lissner et al. (1986), who examined the power of various indices of community structure to detect the effects of oil and gas exploration on the benthic community. Using data from a reconnaissance survey in the Santa Maria Basin and the Western Santa Barbara Channel, California, they found that the power of a two-way analysis of variance to test the interaction between time and location was consistently higher (across all depths and strata) for tests based on community indices (the Gleason diversity, Shannon-Wiener diversity, evenness, total abundance and number of species) than for tests based on abundance of individual species. Thus, in order to have the greatest chance of detecting a change in that community's structure, tests based on community indices would yield the best results.

Caswell and Weinburg (1986) also considered the effects of community structure (species diversity and density) on statistical power for the two-sample t-test. Their results were similar to those of Lissner et al. (1986); tests based on community diversity indices had higher power than those based on the abundance of single-species indicators. Their results showed that the power of the test differed greatly for different indices, with evenness being the most powerful, followed by the Shannon-Wiener index, species richness, and then Simpson's index. They also found that the power of each index increased with density, and at high densities the differences in power among the various indices was reduced. Practitioners could use this type of information to help design high-power impact assessments or monitoring programs by selecting those variables that are likely to have the highest power.

Selecting Sampling Designs and Statistical Models

The choice of a sampling design or statistical model can also be based, in part, on the criterion of statistical power. Skalski and McKenzie (1982) compared control-treatment pairing (CTP) designs and traditional unpaired designs, to see which design will improve the ability to detect the effects of nuclear power plants on benthic and planktonic communities. They found that higher power could be attained using CTP designs because such designs reduce the experimental error associated with the monitoring study. Thus, by evaluating the power of alternative monitoring designs, researchers can select an appropriate design that is likely to have the greatest power, and hence the least sample size required to detect a specified effect with desired high power.

Loftis et al. (1989) evaluated the power of seven statistical tests that are often used to detect trends in water quality variables caused by acidic precipitation. They used Monte Carlo simulation to generate the frequency distributions of these tests because they are unknown and hence analytical formula have not been derived. Their results showed that no one test was the most powerful for all circumstances, but based on the range of variables tested, they recommended the Mann-Kendall test for annual sampling and the Seasonal Kendall or the analysis of covariance test for seasonal sampling.

Power analysis was also used by Hipel et al. (1 986) to help select a statistical test with the best ability to detect trends in time series for environmental management problems. These authors employed Monte Carlo simulation to examine the power of Kendall's tau and the lag-one serial correlation test. They found that the Kendall's tau was more powerful for deterministic trends but for stochastic trends the lag-one serial correlation was more powerful.

Although several authors have used power analysis to select the statistical test that is mostly likely to attain high-power, in some cases the choice of a statistical test will be constrained by the objectives of the study and the sampling design. Thus, a range of plausible statistical tests may not always be possible.

Evaluation of Proposed or Existing Studies

Vaughan and Van Winkle (1982) illustrate how a *priori* power analysis can be used to evaluate an existing monitoring program to determine whether it is likely to detect specified effects, provided they exist. Exactly the same methods can be used to evaluate proposed programs. Using historical data, Vaughan and Van Winkle showed that a project that was designed to study the impact of an electric power plant on white perch (*Morone americana*) recruitment in the Hudson River, New York, had a low probability of detecting even large effects. They calculated that with ten years of monitoring data, in order to have a 75% chance of detecting a significant change in recruitment, a 78% reduction in fish recruitment would have to occur (Point A on Fig. 3). Nineteen years of data would be required to have a 50% chance (power) of detecting a 50% reduction in year-class strength (point B on Fig. 3). This corresponds to about half the expected lifetime of the power plants! If a 50% reduction in year-class strength is important to detect, even 100 years of data would not give a greater than 0.55 probability of detecting such an effect! By evaluating proposed studies, managers can help to prevent situations where studies are implemented that have high power to detect only very large effects.

Power analysis can also be used to evaluate existing impact assessments or monitoring projects in order to help modify them to improve the chances of correctly detecting specified effects. For example, Ontario Hydro (Wismer 1990) used power analysis to evaluate an existing program designed to monitor fish impingement mortality on cooling water intake devices at nuclear power plants. A *posteriori* power analysis indicated that the existing monthly sampling schedule, conducted over a five-year period, generated only an 18% chance of detecting a 50% change in May monthly alewife impingement mortality, caused by new devices aimed at reducing mortality. By changing to a weekly sampling program, two years of data will create a 78% chance of detecting the observed rate of annual change in impingement mortality. Power analysis could thus be used to evaluate current monitoring programs to determine whether they are likely to detect the postulated effects of the new devices. If the chances of detecting important effects are shown to be small, then power analysis can provide information about the modifications to the monitoring program required to increase the chances of correctly detecting the effects of concern.

Evaluation of Past Studies

A posterior/ power analysis can be used to evaluate past studies to provide the information that is required to design future studies that will have high power. This strategy was used by the U.S Nuclear Regulatory Commission (NRC) in their review and evaluation of the design of aquatic monitoring programs at nuclear power plants. For example, McKenzie et al. (1 977) reviewed the logarithmic transformed benthic data collected from the Haddam Neck power plant, and showed that a sample size of 100 observations (at each of the control and treatment stations in the preoperational and operational phases) would be required to obtain an 80% chance of detecting changes in benthic densities in the 20 to 55% range, depending on the sample variance. This sample size is larger than that used at nine of the nuclear power plants reviewed by Gore et al. (1 979). Thomas (1977), working with data from the Monticello and Haddam Neck power plants, showed that sample sizes larger than those currently used in most monitoring programs are needed to detect changes of 50% in benthic densities, given the variability commonly found in environmental samples. Thomas (1977) recommends that much of the monitoring effort that currently goes into all nuclear power plants should be focussed on one or two programs so that adequate numbers of samples can be obtained.

The U.S Nuclear Regulatory Commission also reviewed the monitoring data from the San Onofre, Calvert Cliffs, and Pilgrim nuclear power plants to estimate the experimental error (MSE) associated with the control-treatment pairing (CTP) analysis of variance (McKenzie et al. 1979). They found that the MSE for plankton abundance and productivity is relatively site-independent, thereby allowing estimation of the sample size required to provide high power for CTP designs without extensive preliminary sampling to estimate the expected MSE. The MSE for benthic communities, however, was found to be less stable, meaning that this approach would yield less precise estimates of the sample size needed for future monitoring of benthic communities. In situations where a number of studies or other background data are available to estimate the sample variance, this approach may help eliminate the need for expensive, and perhaps time consuming preliminary studies.

5. CONCLUSIONS

Statistical power is an important part of traditional statistical hypothesis testing, a method that is commonly applied in environmental monitoring, and to a lesser extent in impact assessment. In this report, I review statistical power and show how it increases as the sample size, α, and the important effect size increase, and as the sample variance decreases. I illustrate the applicability of statistical power analysis in the design (a *priori')* of monitoring or environmental impact programs and in the interpretation (a *posteriori*) of results when a statistical test fails to reject some null hypothesis. I show how a *priori* power analysis can be used to design studies to have high power, thereby helping to prevent wasted time and money on low-power studies that are unlikely to detect postulated effects. A *priori* power analysis can also save time and money by preventing situations in which unnecessarily large samples are collected to detect an effect of a specified magnitude with desired high power, and by eliminating from the study those variables that are not likely to attain high power to detect specified changes. I show how a *posterior-i* power analysis can be informative when interpreting the results of studies, and when evaluating past results (which can help improve the design of future studies). This report also shows how power analysis can be used to salvage current monitoring approaches that have little chance of detecting important effects, if they are present.

In this report I included examples to illustrate the role of statistical power analysis in impact assessment and environmental monitoring. Most of these examples, however, are from U.S. sources, perhaps because the U.S. has incorporated power analysis into several documents pertaining to the design of impact assessments and monitoring strategies (e.g. Sharma et al. 1976; Wolfe 1978; Fritz et al. 1980). As well, various agencies in the U.S. have published reports that illustrate their application of power analysis in the review and evaluation of completed impact assessments and monitoring projects (e.g. McKenzie et al. 1977; Thomas 1977; Warren-Hicks et al. 1989).

McCaughran, D.A., 1977. The quality of inferences concerning the effects of nuclear power plants on the environment, p. 229-242. *In* W. Van Winkle [ed.] Proceedings of the conference on assessing the effects of power-plant-induced mortality on fish populations. Pergamon Press, New York, NY.

McKay, D. 1989. Modelling the long-term behavior of an organic contaminant in a large lake: application to PCBs in Lake Ontario. J. Great Lakes Res. 15(2): 283-297.

McKenzie, D.H., L.D. Kannberg, K.L. Gore, E.M. Arnold, and D.G. Watson. 1977. Design and analysis of aquatic monitoring programs at nuclear power plants. Battelle Pacific Northwest Laboratories, Richland, WA. NRC-10 PNL-2423. 129 p.

McKenzie, D.H., E.M. Arnold, J.R. Skalski, D.H. Fickeisen, and KS. Baker. 1979. Quantitative assessment of aquatic impacts of power plants. Battelle Pacific Northwest Laboratories, Richland, WA. NUREG/CR-0631 PNL-2891 RE.

Millard, S.P. 1987. Environmental monitoring, statistics, and the law: room for improvement. Am. Stat. 41(4): 249-253.

Millard, S. P. and D. P. Lettenmaier. 1986. Optimal design of biological sampling programs using analysis of variance. Estuarine, Coastal and Shelf Science 22: 637-656.

Morin, A. 1985. Variability of density estimates and the optimization of sampling programs for stream benthos.' Can. J. Fish. Aquat. Sci. 42: 1530-I 534.

Parkhurst, D.F. 1984. Decision analysis for toxic waste releases. J. Envir. Mgmt. 18: 105-I 30.

Peterman, R.M. 1990a. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. 47(1): 2-15.

Peterman, R.M. 1990b. The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology 71(5): 2024-2027.

Peterman, R.M., and M. M'Gonigle. 1992. Statistical power analysis and the precautionary principle. In press, Marine Pollution Bulletin (May or June 1992).

Raiffa, H. 1968. Decision analysis. Addison-Wesley Publ. Co., Reading, MA. 309 p.

Sharma, R.K., J.D. Buffington, J.T. McFadden [eds.]. 1976. Proceedings of the conference on the biological significance of environmental impacts. U.S. Nuclear Regulatory Commission, Washington, D.C. NR-CONF-002.

In general, little attention has been given to statistical power analysis in Canada (however, the ignorance of statistical power is not restricted to Canada). It is unclear why many researchers have ignored the concept of statistical power. Millard (1987) suggests that some of the reasons for ignoring statistical power in environmental monitoring are the lack of emphasis on statistical power in statistical textbooks and in statistics courses, and the lack of hiring of statisticians in this field. As well, many researchers might be less familiar with statistical power analysis in impact assessment and monitoring because there are few examples, technical manuals, or other documents that illustrate the applicability and methods for environmental studies. Up until recently, researchers also may not have been thinking seriously about the potentially high cost of type II errors (Peterman 1990a, b). Although there are a number of software packages for statistical power analysis, which have been reviewed by Goldstein (1989), many of these are not user friendly and they assume that the user has a working knowledge of the concepts of statistical power.

Despite the lack of attention given to statistical power (and the related concept of type II errors) in impact assessment and environmental monitoring in Canada, this report argues the need to adopt statistical power analysis in our current approaches to impact assessment and monitoring. The routine application of statistical power analysis in both the design and evaluation of studies can improve the state-of-the-art in impact assessment and monitoring by leading to more rigorous tests of statistical hypotheses and to stronger inferences drawn from statistically based-study designs. Furthermore, statistical power analysis can also lead to more cost-effective studies by helping to prevent low-power studies, which provide little information on the effects of concern. We should thus attempt to develop a standard protocol that includes the use of statistical power analysis to design studies to have high power and to help interpret the results of impact assessments or monitoring projects (see recommendations below). In the meantime, we should encourage more rigorous approaches to impact assessment and environmental monitoring by ensuring that practitioners and CEARC follow the recommendations provided below.

6. RECOMMENDATIONS

This study leads to a number of recommendations for statistical power analysis in impact assessment and environmental monitoring. First, I make recommendations for practitioners. Second, I make recommendations for CEARC. These recommendations may also apply to other management agencies involved in impact assessment and monitoring (see Peterman 1990a).

Practitioners

1) Practitioners should use statistical power analysis to determine the sample size required by quantitative impact assessments and monitoring projects to achieve high statistical power. Desired high power should be at least 0.8, however, other approaches (e.g. balancing the expected costs of type I errors and type II errors, or other approaches such as decision analysis) might be used to set desired power.

2) The effect of different estimates of the sample variance and the important effect size on the sample size required to attain desired high power should be explored during the experimental design.

3) Power should be routinely reported for a range of important effect sizes, along with P values and confidence intervals, when a statistical test fails to reject some H_o.

4) When a statistical test fails to reject some H_o, managers should not take action or make recommendations as if there is “no effect” unless the power of the test is high for an effect size deemed important.

CEARC

5) CEARC should encourage and support research towards determining what constitutes an important effect, for biological, economic or other reasons. Impact assessments or monitoring programs should then be designed to detect these important effects with high power.

6) CEARC should take steps to ensure that study designs for impact assessments and monitoring projects are evaluated a ***priori*,** before they are implemented. Projects should be implemented only if they are based on proper experimental and statistical design, including high statistical power. This will help to prevent situations where time and money are wasted on low-power projects. As well, CEARC should not fund research projects unless they are designed to attain high power to detect the effect size of concern.

7) Statistical power should be taken into account, where applicable, in the efficiency and effectiveness criteria used by the Canadian Environmental Assessment Research Council for evaluating impact assessments and monitoring (CEARC 1988).

8) Statistical power analysis should be included in standard protocols for designing quantitative impact assessments and monitoring programs. The role of statistical power analysis in these protocols is described under recommendations 1 to 4 above. In addition, CEARC should include the concept of statistical power and statistical power analysis in other documents that pertain to the methodological aspects of impact assessment or monitoring.

9) CEARC should prepare a technical manual, to illustrate by example the methods of power analysis for a variety of statistical tests that are commonly used in environment impact assessment and monitoring. A manual or guide to statistical power analysis in those fields should also be prepared for managers and decision makers.

BIBLIOGRAPHY

Alldredge, J.R. 1987. Sample size for monitoring of toxic chemical sites. Envir. Monitoring and Assess. 9: 143-154.

Beanlands, G.E., and P.N. Duinker. 1983. An ecological framework for environmental impact assessment in Canada. Institute for Resource and Environmental Studies, Dalhousie University, Halifax, Nova Scotia; and Federal Environmental Assessment Review Office, Hull, Quebec. 132 p.

Bernstein, B.B., and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. J. Envir. Mgmt. 16: 35-43.

Brumelle, S., P. Nemetz, and D. Casey. 1983. Estimating means and variances: the comparative efficiency of composite and grab samples. Envir. Monitoring and Assess. 4: 81-84.

Caswell, H., and Weinberg J.R. 1986. Sample size and sensitivity in the detection of community impact. *In* IEEE Oceans 1986 Conference Proceedings, p. 1040-l 045.

CEARC 1986. Learning from experience: a state-of-the-art review and evaluation of environmental impact assessment audits. Canadian Environmental Assessment and Research Council, Hull, Quebec.

-----. 1988. Evaluating environmental impact assessment: an action prospectus. Canadian Environmental Assessment and Research Council, Hull, Quebec.

Cochran, W.G., and G. M. Cox. 1957. Experimental designs. 2nd ed. John Wiley and Sons, New York. 617 p.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. L. Erlbaum Associates, Hillsdale, NJ. 567 p.

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

Dixon, W.J., and F.J. Massey, Jr. 1983. Introduction to statistical analysis. 4th ed. McGraw Hill Book Co., New York. 678 p.

Fritz, ES., P.J. Rago, and I.P. Murarka. 1980. Strategy for assessing impacts of power plants on fish and shellfish populations. U.S. Fish and Wildlife Service, Biological Services Program, National Power Plant Team, FWS/OBS-80/34. 68 p.

Gerrodette, T. 1987. A power analysis for detecting trends. Ecology 68: 1364-l 372.

Gerrodette, T. 1991. Models for power of detecting trends - a reply to Link and Hatfield. Ecology 72(5): 1889-l 892.

Goldstein, Richard. 1989. Power and sample size via MS/PC-DOS computers. Am. Stat. 43(4): 253-260.

Gore, K.L., J.M. Thomas, and D.G. Watson. 1979. Quantitative evaluation of environmental impact assessment, based on aquatic monitoring programs at three nuclear power plants. J. Envir. Mgmt. 8: l-7.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley and Sons, New York, NY. 257 p.

------. 1984. Statistical and non-statistical methods for environmental monitoring studies. Envir. Monitoring and Assess. 4: 293-301.

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

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

Hipel, K.W., A.I. McLeod, and P.K. Fosu. 1986. Empirical power comparisons of some tests for trends. 346-362 p. *In* Shaarawi, A.H. and R.E. ` wiatkowski [ed.]. Developments in Water Science, No. 27. Elsevier, ew York, N.Y.

Hurlbert, S.J. 1984. Pseudoreplication and design of ecological field experiments. Ecol. Monogr. 54: 187-211.

Kaesler, R.L., J. Cairns, Jr., and J.S. Crossman. 1974. Redundancy in data from stream surveys. Water Res. 8: 637-642.

Link, W.A., and J.S. Hatfield. 1990. Power calculations and model selection for trend analysis: a comment. Ecology 71(3): 1217-l 220.

Lipsey, M.W. 1990. Design sensitivity. Statistical power for experimental research. Sage Publications, Newbury Park, California. 207 p.

Lissner, A., C. Phillips, D. Cadien, R. Smith, B. Bernstein, R. Cimberg, T. Kauwling, and W. Anikouchine. 1986. Assessment of long-term changes in biological communities in the Santa Maria Basin and Western Santa Barbara Channel-Phase I. Prepared for U.S. Minerals Management Service (Contract No. 14-l 2-0001-30031) by Science Applications International Corp., La Jolla, California.

Loftis, J.C., R.C. Ward, R.D. Phillips, and C.H. Taylor. 1989. An evaluation of trend detection techniques for use in water quality monitoring programs. U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, OR EPA/600/3-89/037. 139 p.

MacDonald, J.S., and R.H. Green. 1983. Redundancy of variables used to describe importance of prey species in fish diets. Can. J. Fish. Aquat. Sci. 40: 635-637.

sur les besoins du Ministère et de mettre l'accent sur les aspects pratlques et les applications de la science. Il faudrait déterminer dès que possible les attributions et la composition de l'équipe qui effectuera la première évaluation.

Décision

Il est convenu d'utiliser l'option recommandée pour la première évaluation scientifique, qui portera sur la biodiversité.

Une étude de faisabilité sera effectuée sans tarder et le cadre de référence qui en découlera (le mandat précis, les noms des membres de l'équipe d'évaluation et le calendrier de l'évaluation) sera soumis à l'examen du Comité suprême de gestion, à sa réunion d'avril.

La planification des ressources nécessaires aux évaluations scientifiques sera confiée au Comité d'examen interne.

Suite *: A.* Chlsholm

Participants :

L. Good
D. Wetherup
R. Slater
A. Clarke
K. Dawson
B. Emmett
J. Collinson
J. Roszell (pour A. Lefebvre-Anglin)
M. Dorais
K. MacCormack
A. Chisholm
G. Kowalski
H. Lacombe
D. Small
M. Berman
J.C. Dumesnil

E. Norrena (pour le point 1)
V. Shantora (pour le point 1)
M. Balshaw (pour le point 1)
J. Hollins (pour le point 2)

Skalski, J.R. and D. H. McKenzie. 1982. A design for aquatic monitoring programs. J. Envir. Mgmt. 14: 237-251.

Thomas, J.M. 1977. Factors to consider in monitoring programs suggested by statistical analysis of available data, p. 243-255. *In* W. Van Winkle [ed.] Proceedings of the conference on assessing the effects of power-plant-induced mortality on fish populations. May 3-6, 1977. Pergamon Press, New York, NY.

Thomas, J.M., J.A. Mahaffey, K.L. Gore, and D.G. Watson. 1978. Statistical methods used to assess biological impact at nuclear power plants. J. Envir. Mgmt. 7: 269-290.

Thomas, D.J., W.S. Duval, C.S.Johnston, G.S. Lewbel, A. Birdsall, MS. Hutcheson, G.D. Greene, R.A. Buchanan, and J.W. McDonald. 1985. Effects monitoring strategies for Canada's east coast. Environmental Studies Revolving Funds Report Series, No. 005, Ottawa. 88 p.

Toft, C.A., and P.J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am. Nat. 122: 618-625.

Vaughan, D.S. and W. Van Winkle. 1982. Corrected analysis of the ability to detect reductions in year-class strength of the Hudson River white perch (*Morone americana*) population. Can. J. Fish. Aquat. Sci. 39: 782-785.

Warren-Hicks, W., B.R. Parkhurst, S. Baker JR. [eds.]. 1989. Ecological assessment of hazardous waste sites: a field and laboratory reference. U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, OR. EPA/600/3-89/013.

Winer, B.J. 1971. Statistical principles in experimental design. McGraw-Hill, New York, NY.

Wismer, D.A. 1990. Statistical criteria for intake fish impingement protection system performance evaluation. Ontario Hydro Environmental Studies and Assessment Dept. Report No. 90181. 37 p.

Wolfe, D.A. 1978. Marine biological effects of OCS petroleum development: a program review of research supported under the NOAA Outer Continental Shelf Environmental Assessment Program. U.S. Dept. of Commerce, National Oceanic and atmospheric Administration, Environmental Research Laboratories, Boulder, Colorado. 324 p.

Zar, J.H. 1984. Biostatistical analysis. 2nd ed. Prentice-Hall, Englewood Cliffs, NJ. 718 p.

Table 1. Four possible outcomes for a statistical hypothesis test, depending on the true state of nature and the statistical decision. The probability of a certain outcome is given in parentheses. Redrawn with permission from Peterman (1990a).

	Statistical decision	
State of nature	Reject H_O	Do not reject H_O
H_O true	Type I error (α)	Correct (1-α)
H_O false	Correct (1-β = power)	Type II error (β)

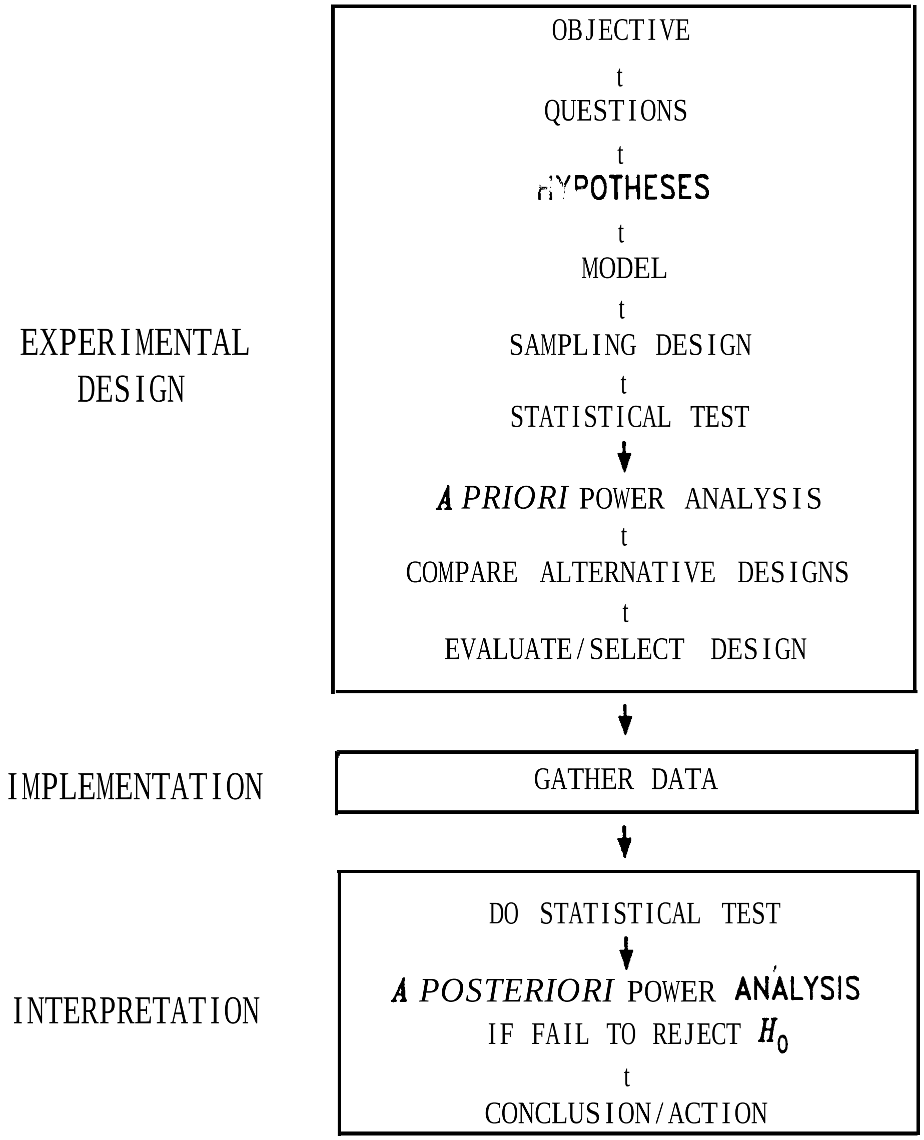

Figure 1. Flowchart showing the use of power analysis in the experimental design (a *priori)* and interpretation (a *posteriori)* of studies designed to test some null hypothesis (H_0).

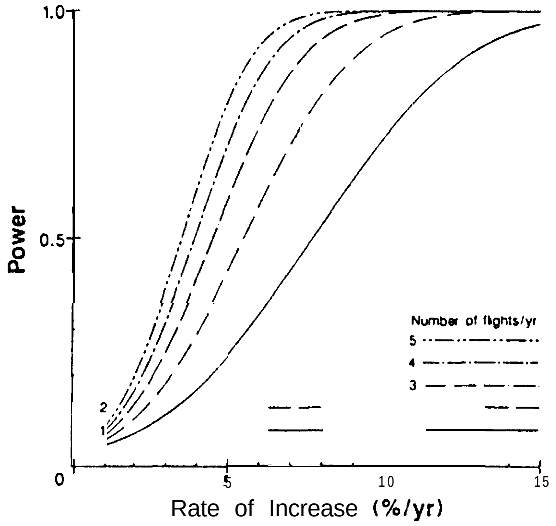

Figure 2. Power curves to detect annual rates of increase in population size of sea otters in central California for various numbers of aerial surveys per year. Reprinted from Gerrodette (1987).

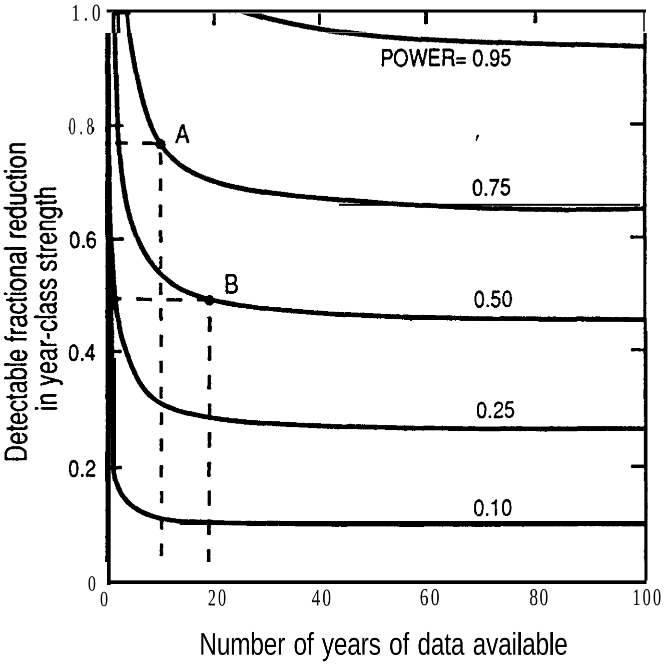

Figure 3. Isopleths of power, or probab.lity or detecting a statistically significant decline in recruitment of white perch in the Hudson River, New York. Each curve represents a common power value that can be attained by various combinations of number of years of data available and the detectable reduction in year-class strength. Reprinted with permission from Peterman (1990a).