The Uses of Statistical Power in Conservation Biology: The Vaquita and Northern Spotted Owl

BARBARA L. TAYLOR*

Department of Biology C-016
University of California San Diego
La Jolla, CA 92093, U.S.A.

TIM GERRODETTE

Southwest Fisheries Science Center
National Marine Fisheries Service
P.O. Box 271
La Jolla, CA 92038, U.S.A.

Abstract: *The consequences of accepting a false null hypothesis can be acute in conservation biology because endangered populations leave little margin for recovery from incorrect management decisions. The concept of statistical power provides a method of estimating the probability of accepting a false null hypothesis. We illustrate how to calculate and interpret statistical power in a conservation context with two examples based on the vaquita (*Phocoena sinus*), an endangered porpoise, and the Northern Spotted Owl (*Strix occidentalis caurina*). The vaquita example shows how to estimate power to detect negative trends in abundance. Power to detect a decline in abundance decreases as populations become smaller, and, for the vaquita, is unacceptably low witin the range of estimated population sizes. Consequently, detection of a decline should not be a necessary criterion for enacting conservation measures for rare species. For the Northern Spotted Owl, estimates of power allow a reinterpretation of results of a previous demographic analysis that concluded the population was stable. We find that even if the owl population had been declining at 4% per year, the probability of detecting the decline was at most 0.64, and probably closer to 0.13; hence, concluding that the population was stable was not justified. Finally, we show how calculations of power can be used to compare different*

Los Usos del poder estadístico en conservación biológica: la vaquita y el búho moteado del Norte

Resumen: *En conservación biológica, las consecuencias de aceptar hipótesis nulas falsas pueden ser muy severas puesto que las poblaciones en peligro de extinción dejan poco margen para revertir el efecto de decisiones incorrectas de manejo. El concepto de poder estadístico provee un método para estimar la probabilidad de aceptar hipótesis nulas falsas. Nosotros ilustramos como calcular e interpretar el poder estadístico en un contexto de conservación con dos ejemplos basados en la vaquita (*Phocoena sinus*), una marsopa en peligro de extinción, y el búho moteado del Norte (*Strix occidentalis caurina*). El ejemplo de la vaquita muestra como estimar el poder para detectar tendencias negativas en abundancia. El poder para detectar una disminución en la abundancia decrece a medida que las poblaciones se hacen mas pequeñas, y en el caso de la vaquita, es inaceptablemente bajo para el rango de tamaños poblacionales estimados. Por consiguiente, la detección de una declinación en el tamaño poblacional no debe ser un criterio necesario para decretar medidas de conservación en especies raras. En el caso del búho moteado del Norte, la estimación del poder permite la reinterpretación de resultados de análisis demográficos previos que concluyeron que la población era estable. Nosotros encontramos que aún si la población del búho moteado a estado declinando un 4% por año, la probabilidad de detectar esta declinación fue de a lo sumo 0.64%, y probablemente más cercana al 0.13%. Por consiguiente, nó se justi-*

[*] Present address: Southwest Fisheries Science Center, National Marine Fisheries Service, P.O. Box 271, La Jolla, CA 92038, U.S.A.

Paper submitted November 27, 1991; revised manuscript accepted June 19, 1992.

methods of monitoring changes in the size of small populations. The optimal method of monitoring Northern Spotted Owl populations may depend both on the size of the study area in relation to the effort expended and on the density of animals. At low densities, a demographic approach can be more powerful than direct estimation of population size through surveys. At higher densities the demographic approach may be more powerful for small populations, but surveys are more powerful for populations larger than about 100 owls. The tradeoff point depends on density but apparently not on rate of decline. Power decreases at low population sizes for both methods because of demographic stochasticity.

ficaba concluir que la población era estable. Finalmente, demostramos como los cálculos de poder pueden ser usados para comparar distintos métodos de monitoreo de cambios en el tamaño de poblaciones pequeñas. El método óptimo de monitoreo de las poblaciones del búho moteado del Norte depende quizas tanto del tamaño del área de estudio en relación con el esfuerzo realizado como de la densidad de los aminales. A bajas densidades, la aproximación demográfica puede ser más poderosa que la estimación directa del tamaño poblacional a partir de evaluaciones. A mayores densidades la aproximación demográfica puede ser más poderosa para poblaciones pequeñas, pero las evaluaciones son más poderosas para poblaciones de mas de 100 búbos. El punto de relación (tradeoff) depende de la densidad pero aparantemente no depende de la tasa de declinación. Para tamaños poblacionales bajos, el poder decrece para ambos métodos debido a la estocasticidad demográfica.

Introduction

Consider the following scenario: a species is declining in abundance, and we have gathered data that may show that a certain pollutant is responsible. We evaluate the data with an appropriate statistical test, but the null hypothesis (of no effect) is not rejected. The result? Without statistically significant evidence that the pollutant is harmful, it is unlikely that any action will be taken to eliminate or reduce the pollutant.

Now consider the following question: If the pollutant does have a harmful effect, what is the probability that we would have detected it? The answer is clearly of central importance, yet this probability, called statistical power, is rarely calculated. The reasons why power has been largely ignored lie partly in the historical development of hypothesis testing and partly in the extra effort required to make power calculations. We believe that a consideration of power is critical in many conservation issues, however, and that every conservation biologist should be familiar with the concept of statistical power. An awareness of statistical power is particularly important in conservation biology because the consequences of incorrect decisions can be severe: the extinction of a species. A medical analogy may be helpful. Consider a medical test that determines whether a patient has some deadly disease. Physicians are properly less concerned with a false positive (concluding that the patient has the disease when she does not) than with a false negative (concluding that the patient does not have the disease when she does). Conservation biologists deal with the health of species and ecosystems and should be similarly concerned with false negatives.

The importance of statistical power is becoming more widely appreciated in many fields of biology. A number of papers in recent years have pointed out the importance of considering power in ecological studies (Quinn & Dunham 1983; Toft & Shea 1983; Rotenberry & Wiens 1985; Peterman 1990*a*). Consideration of statistical power is an integral part of proper experimental and sampling design (see Eberhardt & Thomas [1991] and Andrew & Mapstone [1987] for recent examples). Explicit calculations of power are increasingly being utilized in applied ecology, for example in wildlife ecology (Skalski et al. 1983; Halverson & Teare 1989), insect demography (Solow & Steele 1990), toxicology (Hayes 1987), fisheries (Peterman & Bradford 1987; Peterman 1990*b;* Cyr et al. 1992) marine mammal studies (de la Mare 1984; Holt et al. 1987; Forney et al. 1991), and ecosystem and population monitoring (Skalski & McKenzie 1982; Hinds 1984; Gerrodette 1987, 1991; Green 1989).

There are two main ways that power calculations can be applied in conservation biology. First, before collecting data, study designs can be evaluated in terms of their ability to yield significant results. How large must samples be? How many years will it take? And (ultimately) how much money must we spend? Calculating power for study designs can help answer these questions. We illustrate this use of power by considering the ability of line-transect surveys to show a decline in abundance of a rare species the vaquita (*Phocoena sinus*), a porpoise. We illustrate evaluation of two monitoring designs with the Northern Spotted Owl (*Strix occidentalis caurina*) by comparing demographic to survey methods for detecting declines in abundance. Second, after data have been collected, calculations of power can help interpret the results, particularly when the null hypothesis has not been rejected. We illustrate this use of power in our second example by evaluating the strength of Lande's

(1988) conclusion that the Northern Spotted Owl was not declining in abundance.

Statistical Hypothesis Testing

The dominant paradigm for hypothesis testing, as described in most introductory textbooks, involves:

(1) choosing null and alternative hypotheses;
(2) devising and carrying out an experiment or sampling program designed to distinguish between the two alternatives;
(3) computing an appropriate statistic that summarizes the property to be compared;
(4) determining whether the observed value of the statistic has a probability of occurrence less than a pre-chosen level of significance α; and
(5) if it does, rejecting the null hypothesis in favor of the alternative, or if it does not, retaining the null hypothesis.

The final step involves a yes/no decision about the falsity of the null hypothesis, and the possible logical outcomes of this procedure are often displayed in the form of a simple table (Table 1). Two types of error are possible. If the null hypothesis is true but is rejected, a Type 1 error occurs with probability α; a correct decision is made with probability $1 - \alpha$. If the null hypothesis is false but is not rejected, a Type 2 error occurs with probability β. Statistical power is the probability that the null hypothesis will be rejected when it is false. Hence, power is the probability that we reach a correct decision when the null hypothesis is false, and is calculated as $1 - \beta$ (see Table 1).

Before proceeding to our examples, two brief comments on this procedure are in order. First, the evaluation of data relative to a significance level α (commonly 0.05) depends on naming a specific null hypothesis, but the alternative hypothesis may be nonspecific. For example, we might have as our null hypothesis H_o: mean of population A = mean of population B, but the alternative may be the non-specific H_A: mean of $A \neq$ mean of B. On the other hand, the calculation of β and power ($1 - \beta$) requires that a specific alternative be given—for example, H_A: mean of A = ½ (mean of B). Power has meaning only in relation to a specific alternative hypothesis, and different alternatives result in different values of power.

Table 1. Possible logical outcomes and types of statistical error when testing a null hypothesis H_o.

	Result of statistical test	
	Do not reject H_o	Reject H_o
H_o is true	Correct decision made with probability $1 - \alpha$	Type 1 error (α) made with probability α
H_o is false	Type 2 error (β) made with probability β	Correct decision made with probability $1 - \beta$ (power)

The power of a test is the probability that H_o *will be rejected when* H_o *is false.*

Second, although the above procedure is well established and our discussion applies strictly within the framework of this procedure, there are other methods of testing statistical hypotheses. We may also decide between two hypotheses on the basis of likelihood ratios (Berger & Wolpert 1985) or Bayesian methods (Box & Tiao 1973; Berger 1988; Howson & Urbach 1989). Barnett (1982) presents a general discussion of statistical inference.

Example 1: *Phocoena Sinus*

The vaquita is a small porpoise that occupies a limited range in the northern Gulf of California, Mexico. The status of the porpoise is listed as endangered by the United States Endangered Species Act. Although very little is known about vaquita, one can unequivocally state that the species is rare. In the first dedicated survey in 1976, only two sightings were made in 1959 km of trackline (Wells et al. 1981). From 1986–1988 a total of 3236 km of boat and aircraft surveys resulted in 51 sightings of 96 individual porpoise (Silber 1990). The surveys were not random, but tended to concentrate in areas with highest sighting probability. Barlow (1986) estimated 50–100 individuals as a rough lower limit for the population, noting that available data could not be used for an upper limit. In September of 1991, experimental aerial surveys were conducted to assess the viability of this method for estimating abundance (Barlow et al. 1993). A single sighting of two animals was made in 1143 km of random transect lines. While estimates from so few data are crude, it is likely that there are fewer than 1000 vaquita remaining. There is, meanwhile, substantial mortality occurring due to gill net fisheries. A conservative estimate of the number of animals killed in gill nets is 102 (Vidal 1990). Of these, 79 have occurred since 1985 and 72 were in nets for *Totoaba macdonaldi*, a large sciaenid fish which is itself endangered.

Are surveys able to tell us if the vaquita population is declining in abundance? To investigate this question we created a simple simulation of a line-transect survey (Appendix 1). The results showed that an intensive survey covering virtually all known vaquita habitat could provide an accurate estimate of population size, but that the precision of that estimate strongly depended on population size (Table 2). On theoretical grounds the

Table 2. Results of simulated line-transect surveys for the vaquita, *Phocoena sinus.*

Actual population size	*Mean estimate of abundance*	*Coefficient of variation of estimate of abundance*
250	253	0.387
500	495	0.283
1000	1015	0.209
2000	2005	0.138
4000	3999	0.100
8000	8010	0.071
16,000	16,020	0.050

Mean and coefficient of variation were computed from 1000 simulated surveys at each population size.

variance of a line-transect estimate was expected to be proportional to abundance (Burnham et al. 1980). This relationship was confirmed when we regressed the coefficient of variation of our simulated abundance estimates (CV) (Table 2, column 3) against the inverse of the square root of population size (N) (Table 2, column 1) to give

$$CV = 6.248\left(\frac{1}{\sqrt{N}}\right) \qquad (r^2 = 0.995, p \ll 0.001). \tag{1}$$

The importance of Equation 1 lies in the fact that our ability to detect a decline in population size depends strongly on the precision of the estimate of population size. As N decreases, CV increases, and the probability that a series of surveys will indicate a significant negative trend decreases. This is a specific example of the more general relationship between power, the size of the "effect" we want to detect (*ES*, for effect size; see Cohen 1988), and the variability (V) in our data. Very roughly, we can summarize the general relationship by

$$Power = f\left(\frac{ES - T_{1\text{-}\alpha}}{V}\right).$$

In words, power is an increasing function of effect size (bigger effects are easier to detect), a decreasing function of the test statistic $T_{1\text{-}\alpha}$, which itself negatively depends on α (thus, higher α leads to higher power), and an inverse function of variability (more variable data mean lower power). In the specific case of a series of surveys, the probability of obtaining a significant negative trend (that is, the power, or 1-β) can be approximated by the following relationship (Gerrodette 1987):

$$z_\alpha + z_\beta \leq \ln\lambda\left(\sqrt{\frac{n^2(n+1)(n-1)}{12\sum_{i=1}^{n}\ln\left(\frac{CV_1^2}{\lambda^{i-1}}+1\right)}}\right) \tag{2}$$

where z_x = the x quantile of the standard normal distribution,
α = the probability of Type 1 error,
β = the probability of Type 2 error,
λ = the factor of decrease between surveys ($0 < \lambda < 1$),
n = the number of surveys, and
CV_1 = the coefficient of variation of the population estimate at the initial population size.

Thus, the probability of obtaining a significant negative trend depends on the precision of the surveys (CV), how rapidly the population is declining (λ), the number of surveys (n), as well as the significance level of the test (α). Equation 2 assumes that the population is declining exponentially, that line-transect (or similar sighting per unit effort) data are used to estimate abundance, and that a one-tailed test (for a decline) is used. Note that because the relationship between CV and N is included in the derivation of Equation 2, it is necessary only to give the initial CV for any particular power calculation. Also note that, if anything, the approximation represented by Equation 2 *over*estimates power (Gerrodette 1991); this makes the following pessimistic conclusions about our ability to detect trends all the stronger. As applied to the endangered vaquita, the results may be expressed in several ways:

(1) As population size decreases, so does our ability to detect the decrease (Fig. 1A). Five annual surveys are unlikely to detect a 5%/year population decline for any population size less than 3000 vaquita. If we conduct five biennial surveys, the probability of detecting a 5%/year decline in the vaquita population (a 40% decline over the ten-year period) is 0.81 if the initial population is 3000 porpoise but only 0.45 if the initial population is 1000. Even under the most intensive effort shown in Fig. 1A (10 annual surveys), the power of detecting a 5%/year decline is acceptable (if we define acceptable as $\beta \leq \alpha$) only if the vaquita population is larger than 2300 animals. The actual vaquita population is almost certainly less than that (Silber 1990). If the vaquita population size is in the low hundreds of animals, as the best available data indicate, the most likely outcome of *any* surveys will be a nonsignificant trend, even when the population actually is declining.

(2) As population size decreases, the detectable rate of decline (that is, the minimum rate of decline that could be detected with a given amount of survey effort) increases (Fig. 1B). For example, if there were 300 vaquita, even the most intensive survey effort (10 annual surveys) gives a minimum detectable rate of decline of 18%/year (λ = 0.82). This rate implies a reduction of 86%, from 300 to 42 vaquita, during the ten-year study period, which is clearly unacceptable. Less frequent sur-

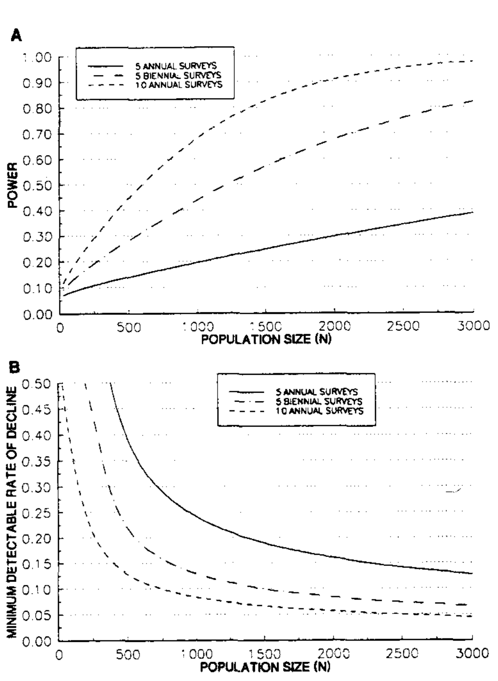

Figure 1. Results of a power analysis for a simulated vaquita survey (Appendix 1). Values are computed from Equation 2 using $\alpha = 0.05$ (1-tailed) for various numbers of surveys (n), as a function of initial population size (N). (A) Power to detect a 5%/year decline ($\lambda = 0.95$). (B) Minimum detectable annual rate of decline ($1 - \lambda$) with high power ($\alpha = \beta = 0.05$).

veys (five biennial surveys) or a shorter study period (five years) result in minimum detectable rates of decline that are even higher (Fig. 1B).

The management implications of this analysis are clear. While there may be important reasons for undertaking vaquita surveys (and we believe there are), determining whether the vaquita population is declining is not one of them. Even worse would be to predicate conservation efforts on whether the surveys indicate a decline. Simply put, if we were to wait for a statistically significant decline before instituting stronger protective measures, the vaquita would probably go extinct first.

Example 2: *Strix Occidentalis Caurina*

Northern Spotted Owls, found in western North America, depend on old-growth forests (Thomas et al. 1990). Concern is prompted because their habitat has been greatly reduced and fragmented by logging, and this habitat loss is expected to continue. Recent studies have demonstrated declines in several owl populations (Thomas et al. 1990). Northern Spotted Owls are long-lived, territorial animals. Because of their relatively sedentary adult life, natural mortality in adults can be accurately assessed by banding studies. Most juveniles are forced to disperse some distance to claim a vacant territory. Estimating juvenile mortality, therefore, has proven difficult. Current estimates place lower and upper bounds for Northern Spotted Owls at 2000 and 6000 individuals (Thomas et al. 1990). Because logging of old-growth forest is continuing, determining the dynamics of the owl population is complex. Current efforts to estimate population growth rates target a snapshot estimate of whether populations are declining while habitat is being destroyed (Anderson et al. 1990). Here we address a simpler question: "Given a static habitat, can we detect a decline in owl abundance?"

Power to Detect a Decline by Demographic Analysis

Several studies have attempted to determine whether Northern Spotted Owl populations were declining by performing a demographic analysis (Lande, 1988; Noon & Biles 1990). As first laid out by Lande (1988), this approach models the population's dynamics as

$$N_t = N_0\lambda^t, \qquad (3)$$

where N_t is population size at time t, and λ is the geometric factor of change. We can estimate λ by solving the characteristic equation using a simplified three-category age structure (Noon & Biles 1990; Thomas et al. 1990) (note that this equation and those used for variance differ from Lande [1988] and Caswell [1989]):

$$\lambda^2 - s\lambda - s_0 s_1 b = 0, \qquad (4)$$

where
s_0 = survival rate from age zero to one,
s_1 = annual survival rate of sub-adults,
s = annual adult survival rate, and
b = annual birth rate.

Estimates of these vital rates are available (Table 3). As Lande notes, because $\lambda^3 b \geq 0$, the real positive solution of this equation must be such that $\lambda \geq s$; that is, the rate of decline cannot be less than the adult survival rate. In other words, if all recruitment into the adult population were to cease and the survival rate of territory-holding adults were to remain constant, λ would be 0.94.

Lande concludes: "The estimated value of $\lambda = 0.961$ is less than twice its standard error from 1.0 and is therefore not significantly different from that for a stable pop-

Table 3. Demographic parameters for the Northern Spotted Owl used by Lande (1988).

Parameter	*Estimate*	*Sample size*
s_0	0.108	179
s_1	0.710	7
s	0.942	69
b	0.240	438
λ	0.961	

Sample size is the number of individuals used to estimate the parameter. We have combined Lande's s_0 (the predispersal survival rate) and s_d (survival rate of dispersers) into a single term for survivorship through the first year of life (s_0), as has been done in subsequent analyses (Anderson et al. 1990; Thomas et al. 1990).

ulation, supporting the contention that the population currently may be near a demographic equilibrium." Although these data cannot reject the null hypothesis using Lande's equations, the data do not support the latter contention. A value for λ of 1.000 cannot be rejected, but the same could be said for $\lambda = 0.920$; it also cannot be rejected. In fact, given Lande's own assumptions about the distribution of λ, $\lambda = 0.920$ is just as likely as $\lambda = 1.000$, and the most likely value is the mean, $\lambda = 0.961$. Lande properly states that the confidence interval on the estimated λ includes 1000, but the data hardly support the contention that $\lambda = 1{,}000$.

To conduct a power analysis, we consider the following question: If $\lambda = 0.961$ (a decline of 4%/year), what is the probability of rejecting a conclusion of a stable population ($\lambda = 1.000$)? We generated a distribution for $\lambda = 0.961$ and $\lambda = 1.000$ (Fig. 2A). Details of the simulations are given in Appendix 2. The histograms in Fig. 2A represent the spread of values for λ that we would expect to obtain if we were to repeat our measurement of Northern Spotted Owl demographic parameters many times, under the assumption that the parameters themselves were constant. Different values of λ result from sampling error in the estimation of demographic parameters. The histograms show that were the true $\lambda = 0.961$, we would reject the hypothesis that $\lambda = 1.000$ for 64% of all estimates of λ (with $\alpha = 0.05$). Power, in other words, is 0.64. Power can be increased up to 0.84 at the cost of accepting an α level as high as 0.25 (Table 4 column with "sampling error only"). In general, though, we have little power to distinguish between these two distributions even though a decline of 4%/year would lead to loss of a third of the population in ten years. Lande's (1988) procedure of computing a confidence interval on the observed λ has even lower power: 0.08 (Table 4). In other words, given a population actually declining at 4%/year, Lande's procedure would conclude that λ was not significantly different from 1.000 92% of the time. This makes weak indeed the claim that the data support $\lambda = 1.000$.

Even this analysis is optimistic, however, because it considers only the sampling error that arose in the estimation of the demographic rates. The rates were estimated by pooling data over years to obtain a single estimate with the variance in mortality calculated from the binomial distribution. It is most likely, however, that owl populations experience environmental variability that translates into year-to-year variability in the demographic parameters and the population growth rate. For the Tawny Owl (*Strix aluco*), a closely related species, owls did not breed in years of low prey abundance, a harsh winter reduced the adult population by half, and there was a clear ceiling on the number of territories, which must limit recruitment (Southern 1970). To generate more realistic distributions of λ for the Northern Spotted Owl, we used the variance of birth and death

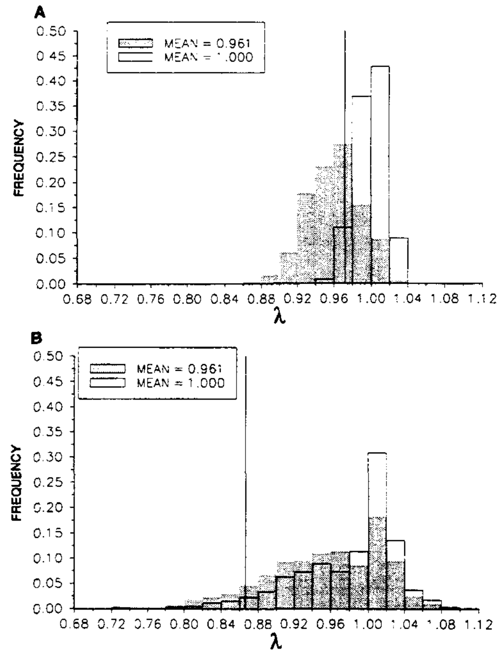

Figure 2. (A) Histograms of 1000 simulations using Lande's (1988) data. Mean data rates are fixed, and variance is due to sampling error (binomial variance). The mean and variance of birth rates (Barrowclough & Coates 1985) also remained fixed. The vertical line is the $\alpha = 0.05$ critical value below which lies the 5% of the unshaded histogram with mean $\lambda = 1.0$. Values less than those would be rejected as not having come from the null distribution. (B) Histograms as in (A), with environmental variance estimated from variance in birth and death parameters estimated from the Tawny Owl.

Table 4. Power ($1 - \beta$) estimated by Monte Carlo simulation to detect a 4%/year decline under two different assumptions about variance: variance is due to sampling error only, and variance is due to environmental variation in addition to sampling error.

	Power	
	Sampling error only	*Including environmental variation*
Simulations with $\alpha = 0.05$	0.644	0.116
Simulations with $\alpha = 0.10$	0.726	0.211
Simulations with $\alpha = 0.25$	0.843	0.432
Lande's Criterion	0.084	0.049

The Lande critical value is the mean plus twice the standard error of λ *as defined by Lande (1988). The distributions are shown in Figs. 2a and 2b.*

rates from the Tawny Owl study in the same Monte Carlo simulations (Appendix 2). The resulting distributions (Fig. 2B) are more similar to each other than when only sampling error is considered (Fig. 2A). This means that the null and alternative hypotheses will be even more difficult to distinguish from each other, and that the power to detect a 4%/year decline ($\lambda = 0.961$) will be dramatically lower (Table 4, column "including environmental variation"). Because of small sample sizes, the likely outcome of the comparison of data from any two years will be an inability to distinguish between estimated parameters, but this does not mean that no environmental variance exists. On the other hand, pooling data over years may lead to unrealistically small variances that give a false picture of the precision of the data. Separation of sampling and environmental variance can be a complicated statistical issue; replication and analysis of model fit can aid in their estimation (see Burnham et al. 1987: Part 4).

Comparison of Two Methods of Monitoring Population Size

In this final section, we use a power analysis to compare two methods of monitorng Northern Spotted Owls for possible declines in population size. To determine whether a population is declining, we could attempt to determine if $\lambda = 1.0$ from estimates of birth and death rates, as considered in the previous section, or we could attempt to estimate population size directly over several years and to determine whether the estimates indicated a decline over time. We will call the former approach the demographic method and the latter the survey method. The question is, given a fixed amount of effort, which method has the greatest probability of detecting a decline in population size?

Although we use the Northern Spotted Owl as an example, we emphasize that the following comparison is presented as an heuristic example of using power analysis to compare study designs. It shows how different study designs could, before time and money are invested, be evaluated for their ability to yield useful information. It is not intended as a recommendation for the study of any particular owl population, or as a criticism of any past or present owl studies. In particular, our analysis does not consider the nonequilibrium conditions that currently exist due to timber harvest (Lamberson et al., in press).

The comparison of the two methods depends on several assumptions: the amount of time and money available, the probability of detecting an owl from a given distance, and the relation between population size and capture rate. We have attempted to use reasonable values based on past studies (Appendix 2). The details of our results depend on the specific values we have chosen for, say, the amount of banding effort, but this does not detract from the generality of the approach. Because Northern Spotted Owls are territorial, we assume that owls occur at some given density in a potential study area, and thus that the choice of a study area determines the size of the study population.

First, for the demographic method, we simulated the estimation of λ from banding studies and estimated the probability (power) of concluding that $\lambda < 1.0$ for several different true values of λ (Appendix 2). The results show, as expected, that power increases as λ decreases (*solid curves, bottom to top in Fig. 3*). Less obvious is that, for a given λ, power generally declines as the size of the study population increases (solid curves in each graph in Fig. 3). If we have chosen to monitor a large population (area), the proportion of the population captured for banding will be small, the variance of the estimates of birth and death rates and hence, λ will be high, and the ability to reject the null hypothesis that $\lambda = 1.0$ (power) will be low. Thus, we do not want to choose too large a study population relative to the planned banding and capturing effort. However, we also do not want to choose too small a study population. At very small population sizes, power is affected by variability due to stochastic demographic effects. If we choose a very small study population, we may be able to monitor every individual owl, but power decreases because the probability decreases that the actual number of owls surviving will exactly equal the survival probability (left ends of solid lines in Fig. 3). For example, if the adult survival rate is 0.96 and our study population consists of 10 adult owls, it is impossible that 9.6 will survive. The power to detect $\lambda < 1.0$ is therefore maximized at some intermediate value of study population size. For the amount of effort assumed in these simulations, the optimum population size (study area) to choose for a banding study to estimate λ is about 60

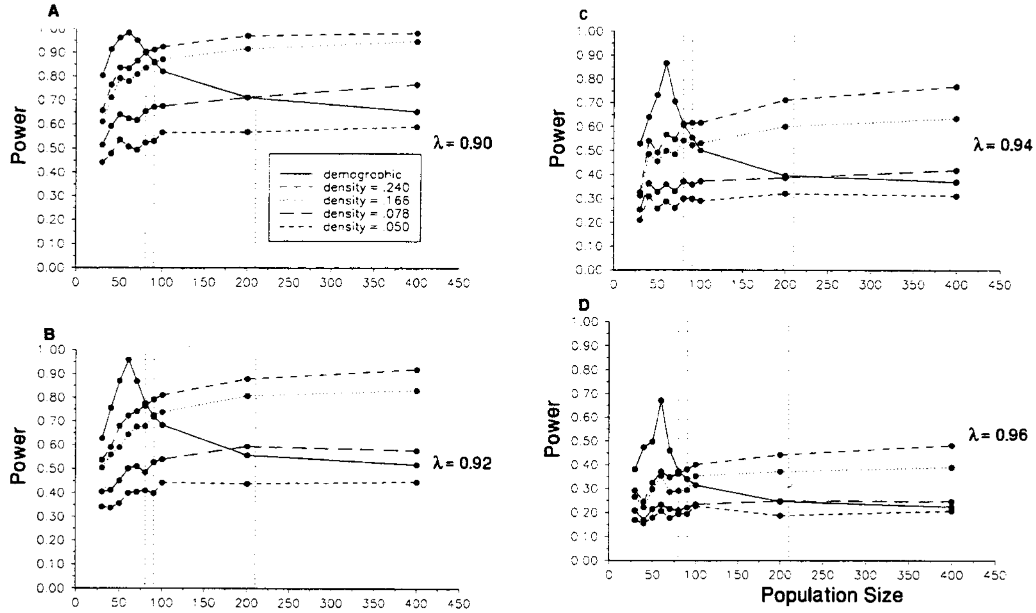

Figure 3. Comparison of power of two methods of monitoring declines in Northern Spotted Owl population size. Solid lines plot power for the demographic approach, and broken lines plot power of line-transect surveys at four owl densities, for the following population growth rates (λ): (A) 0.90, (B) 0.92, (C) 0.94, and (D) 0.96. Dots indicate power calculated from simulations. Connecting lines are linear interpolations. Vertical dotted lines running through figures A–D show that the tradeoff point where power from line-transect techniques exceeds the power from demographic techniques is not affected by the population growth rate.

adult owls (Fig. 3). This is approximately the size of the population chosen for an intensive banding study in northern California (Franklin et al. 1990).

Second, for the survey method, we simulated the estimation of population size from a line-transect survey and estimated the probability (power) of concluding that there was a downward trend in population size over a five-year period (Appendix 2). The results show that, as for the demographic method, power increases as λ decreases, and that power declines at small population sizes due to stochastic demographic effects (dashed lines, bottom to top in Fig. 3). In contrast to the demographic method, however, the power of the survey method does not decline with increasing size of the study population. Power increases with population size up to the point where stochastic demographic effects become negligible and is constant thereafter. Also in contrast to the demographic method, power is an increasing function of owl density (dashed lines in each graph of Fig. 3). These differences occur because the precision of an abundance estimate from a survey depends primarily on the number of animals seen on the survey, and, other things being equal, on the density.

Comparing the two methods in Fig. 3 shows several interesting features. First, for the lowest density of owls considered here (0.050 owls/km^2), the demographic method is always the more powerful design. Thus, were we considering monitoring Northern Spotted Owl populations in a low density area, such as the Olympic peninsula in Washington (Thomas et al. 1990), we should choose the demographic method regardless of size of study area. For higher densities, there is a tradeoff point where the survey method becomes more powerful than the demographic method as study population size increases. The tradeoff point is approximately 80 owls for the highest density (0.240 owls/km^2), 90 owls for the next highest density (0.166 owls/km^2), and 210 owls for the third highest density (0.078 owls/km^2). These tradeoff points do not depend on the actual rate of decline (which is fortunate since this is the quantity we ultimately want to estimate!). Thus, if we were considering monitoring Northern Spotted Owl populations in areas of moderate to high density of owls and the study area was thought to contain at least 100 owls, we should choose the survey method as the more powerful design to detect a population decline.

Conclusion

In conservation biology, as in any scientific research, experiments should be carefully designed to answer the most pressing questions. However, the need for careful experimental design is particularly important in conservation biology because (1) the crisis nature of many situations may not allow time for research to be repeated, (2) money is always in short supply, so it is imperative to use it in a way that will yield the most information, (3) the research activity itself may have some effect on the population, which should be minimized, and (4) the precarious nature of many populations allows little margin to recover from incorrect decisions. An analysis of power is an integral part of good experimental design (Winer 1971). The examples provided here have been chosen to demonstrate how power analysis can allow us to (1) decide whether the proposed research can answer our question, (2) choose among alternate experimental designs, and (3) interpret the results in such a way that is is clear exactly what we can and cannot state given our data.

Although awareness is increasing, statistical power is often ignored in ecological studies (Peterman 1990*a*). A recent review in the field of fisheries biology pointed out that of 408 fisheries papers that reported at least one failure to reject the null hypothesis, only one calculated the probability of making a Type 2 error (Peterman 1990*b*). Our informal survey of past issues of *Conservation Biology* indicate a similar lack of reporting power. We contend that a consideration of power is especially important in conservation biology. Both the vaquita and Northern Spotted Owl examples demonstrate why it is insufficient merely to state that the data failed to reject the null hypothesis. With small populations, failure to reject the null hypothesis may often result from inadequacies in the data rather than from any evidence concerning the falsity of the hypothesis. Such inadequacies may be due to small sample sizes, stochastic demographic effects, or both. In this paper we have particularly illustrated the use of power for detecting changes in population size. However, there are many other situations in conservation biology for which a power analysis is appropriate. For example, consider the problem of defining suitable habitat. This could arise in the context of designing wildlife reserves (what areas are most important?) or in altering habitat for the benefit of rare species (have restoration or mitigation efforts been successful?). We might be comparing abundance, survival rates, behavior, or other characteristics of populations in several areas. In these situations a Type 2 error would lead to the designation of less suitable habitat in a reserve or to the false conclusion that restoration was being successful.

What level of power should we consider acceptable? There is no simple answer to this question. Although there is a generally accepted level of Type 1 error ($\alpha \leq 0.05$), there is no such generally accepted standard for Type 2 error. Furthermore, the relative importance attached to these two kinds of statistical error depends on one's perspective. Consider again the example of the putative pollutant given in the introduction. A manager of a factory producing the pollutant would be most concerned with minimizing Type 1 error—that is, with minimizing the probability of deciding that the pollutant is responsible when it really is not. The result of this incorrect conclusion may be the unnecessary installation of costly equipment. A conservation biologist would also not want to make a Type 1 error, but for a different reason: loss of scientific credibility. However, biologists should be even more concerned about making a Type 2 error—that is, of deciding the pollutant is not responsible when it is—because the result of this incorrect conclusion may be the extinction of the species in question. Because Type 1 and 2 errors result in quite different consequences, weighing their relative costs can be a complex and contentious undertaking. We do not disparage its difficulty. Our point here is that a discussion of the costs cannot proceed without a recognition and calculation of the probability of Type 2 error and its complement, power.

Because of the critical nature of management decisions in conservation biology, we should also consider where the burden of proof should lie. Should scientists be required to show that a population is declining before a negative impact (a direct kill or habitat destruction) can be controlled? One alternative is to require that the party affecting the population show, with high power, that the impact will have no effect before it is allowed (Peterman 1990*a*). A precedent for this approach already exists. Before a new drug is approved, the U.S. Food and Drug Administration puts the burden of proof on the drug industry to show that the drug is *not* harmful (Belsky 1984). Another approach might be to take as our null hypothesis, on the basis of past experience with this or a similar species, that there *will* be an effect, and that the impact cannot be allowed unless this null hypothesis can be rejected. For rare species, such as the vaquita, we have seen that it is inappropriate to require proof of a decline before reductions in the population are halted. An alternative approach may be to require proof that the population is not declining either through survey techniques or by demonstrating that recruitment exceeds removal. Consideration of power may thus cause us to rephrase our hypotheses so that they are appropriate for each conservation problem.

Acknowledgments

This work grew out of discussions of a graduate seminar on the use of matrix population models. We thank the members of this group from the University of California, San Diego, and the Southwest Fisheries Science Center. The paper also benefitted from thoughtful reviews by Jay Barlow, Ted Case, Michael J. Conroy, Doug DeMaster, Michael Gilpin, Daniel Goodman, Edwin O. Green, and Trevor Price. We thank J. B. Jasiunas for assistance in analysis of owl data. We are also grateful to K. P. Burnham for providing status reports on the Northern Spotted Owl. The work of B. Taylor was supported by a National Institute of Health Genetics Training Grant and later by the National Research Council.

Literated Cited

Anderson, D. E., O. J. Rongstad, and W. R. Mytton. 1985. Line transect analysis of raptor abundance along roads. *Bulletin of the Wilderness Society* 13:533–539.

Anderson, D. R., J. Bart, T. C. Edwards, Jr., C. B. Kepler, and E. C. Meslow. 1990. Status review of the Northern Spotted Owl *Strix occidentalis caurina.* U.S. Fish and Wildlife Service, Department of the Interior, Portland, Oregon.

Andrew, M. L., and B. D. Mapstone. 1987. Sampling and the description of spatial pattern in marine ecology. Annual Review of Oceanography & Marine Biology 25:39–90.

Barlow, J. 1986. Factors affecting the recovery of *Phocoena sinus,* the vaquita or Gulf of California harbor porpoise. Administrative Report LJ-86-37. U.S. National Marine Fisheries Service, Southwest Fisheries Center.

Barlow, J. 1988. Harbor porpoise, *Phocoena phocoena,* abundance estimation for California, Oregon, and Washington: I. Ship surveys. Fisheries Bulletin 86:417–432.

Barlow, J., L. Fleisher, K. A. Forney, and O. Maravilla-Chavez. 1993. An experimental aerial survey for vaquita (*Phocoena sinus*) in the northern Gulf of California, Mexico. Marine Mammal Science 9:89–94.

Barnett, V. 1982. Comparative statistical inference. Wiley & Sons, Chichester, England.

Barrowclough, G. F., and S. L. Coates. 1985. The demography and population genetics of owls, with special reference to the conservation of the Spotted Owl (*Strix occidentalis*). Pages 74–85 in R. J. Gutiérrez and B. Carey, editors. Ecology and management of the Spotted Owl in the Pacific Northwest. (Gen Tech Rept PNW-185). Pacific Northwest Forest and Range Experiment Station, USDA Forest Service, Portland, Oregon.

Belsky, M. H. 1984. Environmental policy low in the 1980s: shifting the burden of proof. Ecology Law Quarterly 12:1–88.

Berger, J. O. 1988. Statistical decision theory and Bayesian analysis. Springer Verlag, New York, New York.

Berger, J. O., and R. L. Wolpert. 1985. The likelihood principle. IMS Monograph Series, Vol. 6. Institute of Mathematical Statistics, Hayward, California.

Box, G. E. P., and G. C. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley, Reading, Massachusetts.

Burnham, K. P., D. R. Anderson, and J. L. Laake. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs 72:1–202.

Burnham, K. P., D. R. Anderson, G. C. White, C. Brownie, and K. H. Pollack. 1987. Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society Monograph 5. Bethesda, Maryland.

Caswell, H. 1989. Matrix population models. Sinauer Associates, Sunderland, Massachusetts.

Cohen, J. 1988. Statistical power analsyis for the behavioral sciences. Lawrence Erlbaum, Hillsdale, New Jersey.

Cyr, H., J. A. Downing, S. Lalonde, S. Baines, and M. L. Pace. 1992. Sampling larval fish populations: choice of sample number and size. Transactions of the American Fisheries Society 121:356–368.

de la Mare, W. K. 1984. On the power of catch per unit effort series to detect declines in whale stocks. Report of the International Whaling Commission. 34:655–661.

Eberhardt, L. L., and J. M. Thomas. 1991. Designing environmental field studies. Ecology Monographs 61:53–73.

Forney, K. A., D. A. Hanan, and J. Barlow. 1991. Detecting trends in harbor porpoise abundance from aerial surveys using analysis of covariance. Fisheries Bulletin 89:367–377.

Franklin, A. S., J. P. Ward, R. J. Gutierrez, and G. I. Gould, Jr. 1990. Density of Northern Spotted Owls in northwest California. Journal of Wildlife Management 54:1–10.

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

Gerrodette, T. 1991. Models for power of detecting trends—a reply to Link and Hatfield. Ecology 75:1889–1892.

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

Halverson, T. G., and J. A. Teare. 1989. Carfentanil and overwinter survival in bison: the alternative hypothesis. Journal of Wildlife Diseases 25:448–450.

Hayes, J. P. 1987. The positive approach to negative results in toxicology studies. Ecotoxicological and Environmental Safety 14:73–77.

Hinds, W. T. 1984. Towards monitoring of long-term trends in terrestrial ecosystems. Environmental Conservation 11:11–18.

Holt, R. S., T. Gerrodette, and J. B. Cologne. 1987. Research vessel survey design for monitoring dolphin abundance in the eastern tropical Pacific. U.S. National Marine Fisheries Service Fishery Bulletin 85:435–446.

Howson, C., and P. Urbach. 1989. Scientific reasoning: the Bayesian approach. Open Court, La Salle, Illinois.

Lamberson, R. H., R. McKelvey, B. R. Noon, and C. Voss. 1992. A dynamic analysis of Northern Spotted Owl viability in a fragmented landscape. Conservation Biology 6:505–512.

Lande, R. 1988. Demographic models of the Northern Spotted Owl (*Strix occidentalis caurina*). Oecologia 75:601–607.

Noon, B. R., and C. M. Biles. 1990. Mathematical demography of Spotted Owls in the Pacific Northwest. Journal of Wildlife Management 54:18–27.

Parkhurst, D. F. 1990. Statistical hypothesis tests and statistical power in pure and applied science. Pages 181–201 in G. M. Furstenberg, editor. Acting under uncertainty: multidisciplinary conceptions. Kluwer Academic Publishers, Boston, Massachusetts.

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

Peterman, R. M. 1990*b*. 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.

Quinn, J. F., and A. E. Dunham. 1983. On hypothesis testing in ecology and evolution. American Naturalist 122:602–617.

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

Silber, G. K. 1990. Occurrence and distribution of the vaquita (*Phocoena sinus*) in the northern Gulf of California. Fisheries Bulletin 88:339–346.

Skalski, J. R., and D. H. McKenzie. 1982. A design for aquatic monitoring programs. Journal of Environmental Management 14:237–251.

Skalski, J. R., D. S. Robson, and M. A. Simmons. 1983. Comparative census procedures using single mark-recapture methods. Ecology 64:752–760.

Solow, A. R., and J. H. Steele. 1990. On sample size, statistical power, and the detection of density dependence. Journal of Animal Ecology 59:1073–1076.

Southern, H. N. 1970. The natural control of a population of Tawny Owls (*Strix aluco*). Journal of Zoology 162:197–285.

Thomas, J. W., E. D. Forsman, J. B. Lint, E. C. Meslow, B. R. Noon, and J. Verner. 1990. A conservation strategy for the Northern Spotted Owl. Report of the Interagency Scientific Committee to address the conservation of the Northern Spotted Owl. Portland, Oregon.

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

Vidal, O. 1990. Population biology and exploitation of the vaquita, *Phocoena sinus*. Working paper SC/42/SM24. International Whaling Commission.

Wells, R. S., B. G. Würsig, and K. S. Norris. 1981. A survey of the marine mammals of the upper Gulf of California, Mexico, with an assessment of the status of *Phocoena sinus*. Final report to U.S. Marine Mammal Commission MM1300950-0. NTIS 2881-168791.

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

Appendix 1

Porpoise Simulations

Vaquita (*Phocoena sinus*) are very similar in their sighting characteristics to the harbor porpoise, *P. phocoena*. We therefore used the sighting detection function for harbor porpoise in calm conditions (Beaufort 0 and 1) (Barlow 1988) in our simulations of a ship-based vaquita line-transect survey. Sightability was not affected by group size since vaquita are found only in small groups (Silber 1990); the distribution of group size was taken from Silber's work. We assumed that groups of porpoise were randomly located within their range. The range of the species was considered to be approximately 4900 km^2, which lies between the 20- and 40-meter depth contours in the northern Gulf of California; nearly all sightings of vaquita have been made in this habitat (Silber 1990; Vidal 1990). For complete coverage we set track lines 5 km apart (see part one of the simulation protocol). For the given area this would yield 980 km of survey, which at a survey speed of 15 km hr^{-1} would require approximately eight days of eight hours of survey under perfect conditions. Obtaining these hours would take several weeks, which seemed a likely amount of effort.

To generate statistics for a vaquita population estimate, we repeated the following procedure 1000 times: (1) a distance from the track line was chosen from a uniform distribution from zero to half the distance between transect lines (2500 m); (2) group size was chosen randomly from the group size distribution; (3) simulation population size was incremented; (4) the probability of being sighted at that distance was determined from the sighting detection function; (5) animals seen were added to the abundance estimate; (6) steps 1 to 5 were repeated until the simulation population size equaled N. The procedure was repeated for N = 250, 500, 1000, 2000, 4000, 8000, and 16,000. Number of animals seen was an index of population size. Because sighting conditions were assumed to be constant, the line-transect estimate of porpoise abundance was directly proportional to the number of vaquita seen. The mean and coefficient of variation of abundance in Table 2 were computed from this index of abundance.

The simulations for this example are intentionally simplistic and have not taken into account many sources of error that would be found in a real survey. For example, we allow for no error in estimation of group size, use data only from the best sighting conditions and from a large ship that is likely to be a better sighting platform than will be available for a survey of vaquita. Each of these simplications have reduced variance. The result is a best-case scenario for power to detect a population decline.

Appendix 2

Owl Simulations

The distribution of λ with sampling error only (Fig. 2A) was generated by repeating the following steps 1000 times: (1) each survival rate was calculated by doing n (sample size for that age category) repeats of a trial where a randomly chosen value from a uniform distribution from zero to one determined the fate of the individual according to the survival probability for that age category; (2) the birth rate was determined by finding the mean of n trials (sample size for birth rate) where the birth number was chosen from a normal distribution with the mean b and variance of $1.2b$ (Barrowclough & Coates 1985); (3) the λ for this set of demographic parameters was computed by solving Equation 4. We followed this procedure using the demographic parameters in Table 3 (mean $\lambda = 0.961$), with all parameters multipled by 1/0.961 (mean $\lambda = 1.000$).

The same Monte Carlo techniques were used to generate the distribution of λ with environmental variation (Fig. 2B), except that, for each time-step demographic rates were chosen from normal distributions with the following means and standard deviations: s_0: 0.112, 0.615; s_a: 0.739, 0.265; s: 0.980, 0.080; b: 0.250, 0.640. These parameters were calculated from the Tawny Owl data (Southern 1970). Birth rates were constrained to be non-negative, and survival rates were constrained to lie between zero and one.

Simulations for Fig. 3 produced distributions for both the demographic technique and the line-transect technique. Four possible adult survival rates (s) were chosen: 0.90, 0.92, 0.94, 0.96. Birth rate was held constant and juvenile survival was adjusted to obtain $\lambda = 1.000$. These parameters were used to obtain the distribution for the null hypothesis. The alternate hypothesis assumed no recruitment, that is $s_0 = 0$. Thus the rate of decline was $1 - s$. The following assumptions were used for the line-transect portion: (1) probability of sighting (p) with distance is 0–100m, $p = 1.00$; 101–200m, $p = 0.60$; 201–400m, $p = 0.45$; 401–500m, $p = 0.25$; 501–600m, $p = 0.15$, 601–700m, $p = 0.05$ (based on buteos) (Anderson et al. 1985); (2) densities are estimated from home-range data from radio-tagged owls as presented in Thomas et al. (1990)—for the Olympic peninsula, Washington (0.050 owls/km^2), Washington Western Cascades (0.078 owls/km^2), Oregon Western Cascades (0.166 owls/km^2), and northern California (0.240 owls/km^2, Franklin et al. 1990). Assumptions for the demographic portion were as follows: capture rate = $63.7/N$ or 1.0 for $N <$ 63 (based on capture data in Franklin et al. [1990]); all owls are of equal ease of capture, and capture rate is not dependent on density. Effort was held constant at the level reported in Franklin et al. (1990), which gave a capture probability of 0.91 for a population of approximately 70 owls. The average effort of 400 hours/year was translated into line-transect effort by assuming a survey speed of 1.6 km/hour. For both techniques, it was assumed that only females were counted. The following steps were repeated 10,000 times: (1) the number of owls for a five-year period is determined by allowing each individual to die and/or give birth stochastically (as in previously described simulations); (2) number of owls seen each year is determined stochastically according to the detection function; (3) the log of the number seen each year is regressed linearly against time to obtain the slope, which is the estimated rate of decline; (4) demographic parameters are computed for years 2–5 as (1) survival rate of adults = (adults captured time B)/[(adults captured time A) + (juveniles captured time A)]; (2) survival rate of juveniles = (juveniles captured time B)/(newborns captured time A); (3) birth rate = (newborns captured time B)/(adults captured time A); (5) the means of the four estimated death and birth rates (for years 2–5) are used to calculate λ.

For the case of $\alpha = 0.05$, the critical value is the lower fifth percentile of the null hypothesis distribution. Power was estimated for the alternate (declining) hypothesis by the fraction of statistics less than the critical value. For Fig. 3 each point for the demographic method represents 80,000 simulations, 40,000 each for the null and alternate cases. Each line-transect point represents 20,000 simulations. The demographic method has four times the number of line-transect simulations because density does not affect power for the demographic technique and therefore 20,000 simulations were accumulated for each of four densities used for line-transect estimates.

As with the vaquita, the simulations are intentionally simplistic and dependent on assumptions about capture rate, sightability, etc. The exercise is not intended as a management answer but is presented to demonstrate techniques useful for evaluation of experimental design. The quality of the evaluation can only be as good as the quality of the preliminary data used in the assessment.